Eur. Phys. J. C (2023) 83:1125
https://doi.org/10.1140/epjc/s10052-023-11949-2

THE EUROPEAN PHYSICAL JOURNAL C

Review

50 Years of quantum chromodynamics

Introduction and Review

Franz Gross[1,2,a], Eberhard Klempt[3,b], Stanley J. Brodsky[4], Andrzej J. Buras[5], Volker D. Burkert[1], Gudrun Heinrich[6], Karl Jakobs[7], Curtis A. Meyer[8], Kostas Orginos[1,2], Michael Strickland[9], Johanna Stachel[10], Giulia Zanderighi[11,12], Nora Brambilla[5,12,13], Peter Braun-Munzinger[10,14], Daniel Britzger[11], Simon Capstick[15], Tom Cohen[16], Volker Crede[15], Martha Constantinou[17], Christine Davies[18], Luigi Del Debbio[19], Achim Denig[20], Carleton DeTar[21], Alexandre Deur[1], Yuri Dokshitzer[22,23], Hans Günter Dosch[10], Jozef Dudek[1,2], Monica Dunford[24], Evgeny Epelbaum[25], Miguel A. Escobedo[26], Harald Fritzsch[27], Kenji Fukushima[28], Paolo Gambino[11,29], Dag Gillberg[30,31], Steven Gottlieb[32], Per Grafstrom[33,34], Massimiliano Grazzini[35], Boris Grube[1], Alexey Guskov[36], Toru Iijima[37], Xiangdong Ji[16], Frithjof Karsch[38], Stefan Kluth[11], John B. Kogut[39,40], Frank Krauss[41], Shunzo Kumano[42,43], Derek Leinweber[44], Heinrich Leutwyler[45], Hai-Bo Li[46,47], Yang Li[48], Bogdan Malaescu[49], Chiara Mariotti[50], Pieter Maris[51], Simone Marzani[52], Wally Melnitchouk[1], Johan Messchendorp[53], Harvey Meyer[20], Ryan Edward Mitchell[54], Chandan Mondal[55], Frank Nerling[53,56,57], Sebastian Neubert[3], Marco Pappagallo[58], Saori Pastore[59], José R. Peláez[60], Andrew Puckett[61], Jianwei Qiu[1,2], Klaus Rabbertz[33,62], Alberto Ramos[63], Patrizia Rossi[1,64], Anar Rustamov[53,65], Andreas Schäfer[66], Stefan Scherer[67], Matthias Schindler[68], Steven Schramm[69], Mikhail Shifman[70], Edward Shuryak[71], Torbjörn Sjöstrand[72], George Sterman[73], Iain W. Stewart[74], Joachim Stroth[53,56,57], Eric Swanson[75], Guy F. de Téramond[76], Ulrike Thoma[3], Antonio Vairo[77], Danny van Dyk[41], James Vary[51], Javier Virto[78,79], Marcel Vos[80], Christian Weiss[1], Markus Wobisch[81], Sau Lan Wu[82], Christopher Young[83], Feng Yuan[84], Xingbo Zhao[55], Xiaorong Zhou[48]

[1] Thomas Jefferson National Accelerator Facility, 12000 Jefferson Avenue, Newport News, VA 23606, USA
[2] Department of Physics, William and Mary, Williamsburg, VA 23187, USA
[3] Helmholtz-Institut für Strahlen- und Kernphysik, Universität Bonn, Nußallee 14-16, 53115 Bonn, Germany
[4] Theoretical Physics, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
[5] Institute for Advanced Study, Technische Universität München, Lichtenbergstraße 2a, 85748 Garching b. München, Germany
[6] Institut für Theoretische Physik, Karlsruher Institut für Technologie (KIT), 76128 Karlsruhe, Germany
[7] Physikalisches Institut, Universität Freiburg, 79104 Freiburg, Germany
[8] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[9] Department of Physics, Kent State University, 800 E Summit St, Kent, OH 44240, USA
[10] Physikalisches Institut, Universität Heidelberg, 69120 Heidelberg, Germany
[11] Max-Planck-Institut für Physik, Föhringer Ring 6, 80805 Munich, Germany
[12] Physik Department, Technische Universität München, James-Franck-Straße 1, 85748 Garching b. München, Germany
[13] Munich Data Science Institute, Technische Universität München, Walther-von-Dyck-Straße-10, 85748 Garching b. München, Germany
[14] Extreme Matter Institute EMMI, GSI, 64291 Darmstadt, Germany
[15] Department of Physics, Florida State University, Tallahassee, FL 32306, USA
[16] Department of Physics, University of Maryland, College Park, MD 20742, USA
[17] Physics Department, Temple University, 1925 N. 12th Street, Philadelphia, PA 19122, USA
[18] School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, UK
[19] Higgs Centre for Theoretical Physics, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3FD, UK
[20] PRISMA + Cluster of Excellence and Institut für Kernphysik and Helmholtz Institute Mainz, Johannes Gutenberg University Mainz, 55128 Mainz, Germany
[21] Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA
[22] Riga Technical University Center of High Energy Physics and Accelerator Technologies, Riga, Latvia
[23] Petersburg Nuclear Physics Institute, Gatchina, Russia
[24] Kirchhoff-Institut für Physik, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany
[25] Institut für Theoretische Physik II, Ruhr-Universität Bochum, 44780 Bochum, Germany
[26] Instituto Galego de Física de Altas Enería (IGFAE), Universidade de Santiago de Compostela, 15782 Galicia, Spain

Published online: 12 December 2023

[27] Department für Physik der Universität München, Theresienstraße 37, 80333 Munich, Germany
[28] School of Science, University of Tokyo, Bunkyo, Tokyo 113-8654, Japan
[29] Dipartimento di Fisica, Università di Torino and INFN, Sezione di Torino, Via Pietro Giuria 1, 10125 Turin, Italy
[30] Department of Physics, Carlton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada
[31] Department of Physics, Lund University, Lund, Sweden
[32] Department of Physics, Indiana University, Bloomington, IN 47405, USA
[33] CERN, Geneva, Switzerland
[34] Dipartimento di Fisica, Università di Bologna, 40126 Bologna, Italy
[35] Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[36] Joint Institute for Nuclear Research, 141980 Dubna, Moscow Region, Russia
[37] Kobayashi-Maskawa Institute (KMI)/Graduate School of Science Nagoya University, Furocho, Chikusa Ward, Nagoya, Aichi 464-8601, Japan
[38] Physics Department, Bielefeld University, 33615 Bielefeld, Germany
[39] Department of Energy, Division of High Energy Physics, Washington, DC 20585, USA
[40] Department of Physics-TQHN, University of Maryland, 82 Regents Drive, College Park, MD 20742, USA
[41] Institute for Particle Physics Phenomenology, Physics Department, Durham University, Durham DH1 3LE, UK
[42] Department of Mathematics, Physics, and Computer Science, Faculty of Science, Japan Women's University, 2-8-1 Mejirodai, Bunkyo-ku, Tokyo 112-8681, Japan
[43] Theory Center, Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organization (KEK), 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[44] Centre for the Subatomic Structure of Matter (CSSM), Department of Physics, The University of Adelaide, Adelaide, SA 5005, Australia
[45] Albert Einstein Center for Fundamental Physics, Institute for Theoretical Physics, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
[46] Institute of High Energy Physics, Beijing 100049, People's Republic of China
[47] University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[48] University of Science and Technology of China, No. 96, JinZhai Road, Baohe District, Hefei, Anhui 230026, People's Republic of China
[49] LPNHE, Sorbonne Université, Université de Paris Cité, CNRS/IN2P3, 75252 Paris, France
[50] INFN, Sezione di Torino, Via Pietro Giuria 1, 10125 Turin, Italy
[51] Deptarment of Physics and Astronomy, Iowa State University, Ames, IA 50011, USA
[52] Dipartimento di Física, Università di Genova and INFN, Sezione di Genova, Via Dodecaneso 33, 16146 Genoa, Italy
[53] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstraße 1, 64291 Darmstadt, Germany
[54] Department of Physics, Indiana University Bloomington, 107 S. Indiana Avenue, Bloomington, IN 47405, USA
[55] Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou, Gansu 730000, People's Republic of China
[56] Helmholtz Forschungsakademie Hessen für FAIR (HFHF), GSI Helmholtzzentrum für Schwerionenforschung, Campus Frankfurt, Frankfurt, Germany
[57] Goethe Universität, Institut für Kernphysik, Max-von-Laue-Str. 1, 60438 Frankfurt, Germany
[58] Dipartimento Interateneo di Fisica, Università di Bari and INFN, Sezione di Bari, Via Amendola 173, 70125 Bari, Italy
[59] Department of Physics and McDonnell Center for the Space Sciences, Washington University in Saint Louis, Saint Louis, MO 63130, USA
[60] Departamento de Física Teórica and IPARCOS, Universidad Complutense, 28040 Madrid, Spain
[61] University of Connecticut, Storrs, CT 06269, USA
[62] ETP, KIT, Postfach 6980, 76128 Karlsruhe, Germany
[63] IFIC (UVEG/CSIC) Valencia, C. del Catedrático José Beltrán 2, 46980 Paterna, Spain
[64] INFN, Laboratori Nazionali di Frascati, 00044 Frascati, Italy
[65] National Nuclear Research Center, 1000 Baku, Azerbaijan
[66] Institut für Theoretische Physik, Universität Regensburg, 93040 Regensburg, Germany
[67] Institut für Kernphysik, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany
[68] Department of Physics and Astronomy, University of South Carolina, Columbia, SC 29208, USA
[69] Département de Physique Nucléaire et Corpusculaire, Université de Genève, 1205 Geneva, Switzerland
[70] School of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA
[71] Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA
[72] Department of Astronomy and Theoretical Physics, Lund University, Box 43, 221 00 Lund, Sweden
[73] C. N. Yang Institute for Theoretical Physics and Department of Physics and Astronomy Stony Brook University, Stony Brook, New York 11794, USA
[74] Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[75] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA
[76] Laboratorio de Física Teórica y Computacional, Universidad de Costa Rica, 11501 San José, Costa Rica
[77] Physik Department, Technische Universität München, James-Franck-Straße 1, 85748 Garching b. München, Germany
[78] Departament de Física Quántica i Astrofísica, Universitat de Barcelona, Martí i Franqués 1, 08028 Barcelona, Catalunya, Spain
[79] Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona, Martí i Franqués 1, 08028 Barcelona, Catalunya, Spain
[80] IFIC (UVEG/CSIC) Valencia, 46980 Paterna, Spain
[81] Departmant of Physics, Louisiana Tech University, 201 Mayfield Ave, Ruston, LA 71272, USA
[82] Department of Physics, University of Wisconsin, Madison, WI 53706, USA
[83] Institute of Physics, Albert Ludwig University of Freiburg, Freiburg im Breisgau, Germany
[84] Nuclear Science Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA

Received: 1 December 2022 / Accepted: 21 August 2023
© The Author(s) 2023

Abstract Quantum Chromodynamics, the theory of quarks and gluons, whose interactions can be described by a local SU(3) gauge symmetry with charges called "color quantum numbers", is reviewed; the goal of this review is to provide advanced Ph.D. students a comprehensive handbook, helpful for their research. When QCD was "discovered" 50 years ago, the idea that quarks could exist, but not be observed, left most physicists unconvinced. Then, with the discovery of charmonium in 1974 and the explanation of its excited states using the Cornell potential, consisting of the sum of a Coulomb-like attraction and a long range linear confining potential, the theory was suddenly widely accepted. This paradigm shift is now referred to as the *November revolution*. It had been anticipated by the observation of scaling in deep inelastic scattering, and was followed by the discovery of gluons in three-jet events. The parameters of QCD include the running coupling constant, $\alpha_s(Q^2)$, that varies with the energy scale Q^2 characterising the interaction, and six quark masses. QCD cannot be solved analytically, at least not yet, and the large value of α_s at low momentum transfers limits perturbative calculations to the high-energy region where $Q^2 \gg \Lambda^2_{\rm QCD} \simeq (250\ {\rm MeV})^2$. Lattice QCD (LQCD), numerical calculations on a discretized space-time lattice, is discussed in detail, the dynamics of the QCD vacuum is visualized, and the expected spectra of mesons and baryons are displayed. Progress in lattice calculations of the structure of nucleons and of quantities related to the phase diagram of dense and hot (or cold) hadronic matter are reviewed. Methods and examples of how to calculate hadronic corrections to weak matrix elements on a lattice are outlined. The wide variety of analytical approximations currently in use, and the accuracy of these approximations, are reviewed. These methods range from the Bethe–Salpeter, Dyson–Schwinger coupled relativistic equations, which are formulated in both Minkowski or Euclidean spaces, to expansions of multi-quark states in a set of basis functions using light-front coordinates, to the AdS/QCD method that imbeds 4-dimensional QCD in a 5-dimensional deSitter space, allowing confinement and spontaneous chiral symmetry breaking to be described in a novel way. Models that assume the number of colors is very large, i.e. make use of the large N_c-limit, give unique insights. Many other techniques that are tailored to specific problems, such as perturbative expansions for high energy scattering or approximate calculations using the operator product expansion are discussed. The very powerful effective field theory techniques that are successful for low energy nuclear systems (chiral effective theory), or for non-relativistic systems involving heavy quarks, or the treatment of gluon exchanges between energetic, collinear partons encountered in jets, are discussed. The spectroscopy of mesons and baryons has played an important historical role in the development of QCD. The famous X,Y,Z states – and the discovery of pentaquarks – have revolutionized hadron spectroscopy; their status and interpretation are reviewed as well as recent progress in the identification of glueballs and hybrids in light-meson spectroscopy. These exotic states add to the spectrum of expected $q\bar{q}$ mesons and qqq baryons. The progress in understanding excitations of light and heavy baryons is discussed. The nucleon as the lightest baryon is discussed extensively, its form factors, its partonic structure and the status of the attempt to determine a three-dimensional picture of the parton distribution. An experimental program to study the phase diagram of QCD at high temperature and density started with fixed target experiments in various laboratories in the second half of the 1980s, and then, in this century, with colliders. QCD thermodynamics at high temperature became accessible to LQCD, and numerical results on chiral and deconfinement transitions and properties of the deconfined and chirally restored form of strongly interacting matter, called the Quark–Gluon Plasma (QGP), have become very precise by now. These results can now be confronted with experimental data that are sensitive to the nature of the phase transition. There is clear evidence that the QGP phase is created. This phase of QCD matter can already be characterized by some properties that indicate, within a temperature range of a few times the pseudocritical temperature, the medium behaves like a near ideal liquid. Experimental observables are presented that demonstrate deconfinement. High and ultrahigh density QCD matter at moderate and low temperatures shows interesting features and new phases that are of astrophysical relevance. They are reviewed here and some of the astrophysical implications are discussed. Perturbative QCD and methods to describe the different aspects of scattering processes are discussed. The primary parton–parton scattering in a collision is calculated in perturbative QCD with increasing complexity. The radiation of soft gluons can spoil the perturbative convergence, this can be cured by resummation techniques, which are also described here. Realistic descriptions of QCD scattering events need to model the cascade of quark and gluon splittings until hadron formation sets in, which is done by parton showers. The full event simulation can be performed with Monte Carlo event

Harald Fritzsch: Deceased.

Stanley J. Brodsky, Andrzej J. Buras, Volker D. Burkert, Gudrun Heinrich, Karl Jakobs, Curtis A. Meyer, Kostas Orginos, Michael Strickland, Johanna Stachel, Giulia Zanderighi: Convenor.

[a] e-mail: flgros@wm.edu

[b] e-mail: klempt@hiskp.uni-bonn.de (corresponding author)

generators, which simulate the full chain from the hard interaction to the hadronic final states, including the modelling of non-perturbative components. The contribution of the LEP experiments (and of earlier collider experiments) to the study of jets is reviewed. Correlations between jets and the shape of jets had allowed the collaborations to determine the "color factors" – invariants of the SU(3) color group governing the strength of quark–gluon and gluon–gluon interactions. The calculated jet production rates (using perturbative QCD) are shown to agree precisely with data, for jet energies spanning more than five orders of magnitude. The production of jets recoiling against a vector boson, $W^{\pm}$ or Z, is shown to be well understood. The discovery of the Higgs boson was certainly an important milestone in the development of high-energy physics. The couplings of the Higgs boson to massive vector bosons and fermions that have been measured so far support its interpretation as mass-generating boson as predicted by the Standard Model. The study of the Higgs boson recoiling against hadronic jets (without or with heavy flavors) or against vector bosons is also highlighted. Apart from the description of hard interactions taking place at high energies, the understanding of "soft QCD" is also very important. In this respect, Pomeron – and Odderon – exchange, soft and hard diffraction are discussed. Weak decays of quarks and leptons, the quark mixing matrix and the anomalous magnetic moment of the muon are processes which are governed by weak interactions. However, corrections by strong interactions are important, and these are reviewed. As the measured values are incompatible with (most of) the predictions, the question arises: are these discrepancies first hints for New Physics beyond the Standard Model? This volume concludes with a description of future facilities or important upgrades of existing facilities which improve their luminosity by orders of magnitude. The best is yet to come!

Contents

Preface

Quantum Chromodynamics or QCD was developed and defined over a brief period from 1972–1973. One of us (EK) wrote an article early in 2021 on the scalar glueball and searched the literature to find where glueballs were first mentioned. This was at the 16th International Conference on High-Energy Physics (ICHEP 72). In the winter of 2021/2022 he thought it was time to prepare a volume dedicated to 50 Years of QCD. He got approval from the EJPC, and asked FG to join the effort. Here is the result.

It's been quite an adventure to guide and prepare this volume. From the start it was to be published as a single article, organized and edited by the two coeditors, with integrated contributions from invited scientists familiar with all aspects of the subject. Our initial outline included only eight sections, but as we got advice from our conveners and early contributors, the number of sections grew to the 14 you see here, and in some cases the number of subsections in each section also grew. The subject is both beautiful and vast, and keeping this volume "limited" in length was a real challenge.

Our goal was to prepare a volume for young Ph.Ds and postdocs that could serve as a readable resource and introduction to specialties outside of their own field of research – a shortcut to acquiring the broad familiarity that usually takes time to acquire. We also invited our contributors to reflect on how they developed their ideas/insights, usually discouraged in scientific articles. We believe that what has resulted is truly unique.

The volume begins with the personal reflections of two scientists who were contributors to the foundations of QCD (Sect. 1), and follows with three early developments that quickly showed that QCD as on the right track (Sect. 2). Prominent among these was the "November revolution," where the discovery and explanation of the states of charmonium lead quickly to the Cornell potential and an early description of why quarks could not be seen, convincing many doubters that quarks were real.

After establishing that the QCD fine structure constant, α_s, is too large at hadronic scales for perturbation theory to work (Sect. 3), we describe in some detail Lattice QCD (Sect. 4), believed now to be the only method that can give *exact* predictions for QCD (with numerical errors, of course, which are decreasing rapidly as the computations and computers improve). Unfortunately, Lattice QCD does not give much of an intuitive picture of how the physics works, so approximate analytic methods are needed (and will probably always be needed) and these are summarized in Sects. 5 and 6, including effective field theories, a powerful tool with many applications. Perhaps some day we will have exact analytic solutions, but not today.

From there our account turns to experimental manifestations of QCD (with theoretical support), starting with the exploration of the QCD phase diagram in heavy ion collisions and in dense matter (Sect. 7), followed by the study of mesons (Sect. 8) and baryons (Sect. 9) that reveal the existence of "exotic" states like glueballs, hybrids, hadronic molecules, and tetra- and pentaquarks. A special focus is given to the nucleon and its structure (Sect. 10). Then, collisions at high energies are discussed, from the hard scattering of two partons followed by their hadronization (Sect. 11); the production and identification of jets of particles culminated in the discovery of the Higgs boson and measurements of its properties (Sect. 12); weak decays, precision analyses of the quark mixing matrix, and the anomalous magnetic moment of muons that show the first hints of New Physics beyond the Standard Model (Sect. 13). The volume concludes with a brief account of experimental projects under construction or already funded (Sect. 14). We do not discuss the many exciting theoretical or experimental ideas that are currently in the drawing board, or as theorists sometimes say, on the "second sheet" (when they are joking about wonderful ideas still in an imaginative state). These we save for the next volume!

It has been a great experience for us to work on this volume; we hope you will find some pleasure in skimming through it.

1 Theoretical Foundations

Conveners:
Franz Gross and Eberhard Klempt

It was a long path before the principles of Quantum Chromodynamics could be formulated. With the discovery of electrons by Thompson in 1897, protons by Rutherford in 1917, neutrons by Chadwick in 1932, and the prediction of neutrinos by Pauli in 1930, the basis was completed of what we now call "the first generation of elementary particles". With pions as mediators of the strong interaction – proposed by Yukawa in 1935 and confirmed by Powell and his collaborators in 1947, a consistent picture of particles and their interactions, gravitational, electromagnetic, weak and strong, seemed to have emerged. Only the muon, discovered in 1936, was superfluous. It could not be an excited state of the electron, there was no $\mu \to e\gamma$ decay. Hence Isidor I. Rabi asked: "Who ordered that"? Nowadays, the muon has a well defined place as member of the second lepton family in analogy to the electron in the first lepton family.

But then, the number of particles grew rapidly: the charged Kaon was discovered, the Λ and the Σ. Some particles like the $\Delta(1232)$ baryon were found to be extremely short-lived. More and more resonances were found, their number started to explode. Attempts to break up protons or neutrons into "truly elementary" constituents by bombardment of protons with energetic particles failed.

A theory to understand the zoo of particles was missing. "Nuclear democracy" was declared: all particles were supposed to be "elementary" and be formed by forces arising from the exchange of these particles. Reactions were studied within S-matrix theory, Regge pole analysis, dispersion relation and other theorems derived in function theory. A field theory of strong interactions was thought to be impossible.

The early development or "discovery" of QCD proceeded in three steps: the first step was the quark model by Gell-Mann and Zweig which allowed the zoo of particles to be organized into multiplets under the SU(3) symmetry when baryons were thought of as composed of three quarks and mesons of a quark and an antiquark. The Pauli principle then required a quark property for which Gell-Mann coined the name color. The second step was the idea that color could be the charge of strong interaction, and that colored quarks would interact by the exchange of gluons carrying color themselves.

One problem remained: in spite of excessive searches, no quarks were observed even when nucleons were bombarded with projectiles of the highest energy available at that time. This problem was solved in the breakthrough papers of Gross, Wilczek and Politzer demonstrating that the observation of quasi-free quarks (*asymptotic freedom*) in deep inelastic scattering is compatible with strong confining forces (*infrared slavery*).

This first section contains two personal accounts of the early development or "discovery" of QCD. Leutwyler's contribution starts with a broad picture of the chaotic state of "theories" of the strong interactions in the 1960s and carries us through to the present day. He describes how many thought field theory could not work for the description of "nuclear forces." They thought the use of dispersion relations and unitarity would provide a better approach, but now we know that these are only useful tools. His discussion of how the exact and approximate symmetries of QCD lead to an understanding of the mass scales of the quarks shows how much the development of QCD and the standard model have brought order out of chaos, and have led to a deep understanding of the physics.

The second contribution by Fritzsch gives a more focused and personal account of how some issues that had to be surmounted before QCD became the accepted theory of the strong forces. He describes several arguments that led them to the necessity for three colors of quarks (and the SU(3) color symmetry). He reminds us that QCD and the existence of quarks did not become widely accepted until the discovery of the J/ψ, among the topics discussed in the following Sect. 2.

Both of these accounts of the history and the physics are exciting to read, and a broad introduction to this volume. We hope you will enjoy them as much as we have.

1.1 The strong interaction[1]

Heinrich Leutwyler

1.1.1 Beginnings

The discovery of the neutron in 1932 [2] may be viewed as the birth of the strong interaction: it indicated that the nuclei consist of protons and neutrons and hence the presence of a force that holds them together, strong enough to counteract the electromagnetic repulsion. Immediately thereafter, Heisenberg introduced the notion of isospin as a symmetry of the strong interaction, in order to explain why proton and neutron nearly have the same mass [3]. In 1935, Yukawa pointed out that the nuclear force could be generated by the exchange of a hypothetical spinless particle, provided its mass is intermediate between the masses of proton and electron – a *meson* [4]. Today, we know that such a particle indeed exists: Yukawa predicted the pion. Stueckelberg pursued similar ideas, but was mainly thinking about particles of spin 1, in analogy with the particle that mediates the electromagnetic interaction [5].

In the thirties and fourties of the last century, the understanding of the force between two nucleons made consider-

[1] The present section is an extended version of my lecture notes *On the history of the strong interaction* [1].

able progress, in the framework of nonrelativistic potential models. These are much more flexible than quantum field theories. Suitable potentials that are attractive at large distances but repulsive at short distances do yield a decent understanding of nuclear structure: Paris potential, Bonn potential, shell model of the nucleus. In this framework, nuclear reactions, in particular the processes responsible for the luminosity of the sun, stellar structure, α-decay and related matters were well understood more than 60 years ago.

These phenomena concern interactions among nucleons with small relative velocities. Experimentally, it had become possible to explore relativistic collisions, but a description in terms of nonrelativistic potentials cannot cover these. In the period between 1935 and 1965, many attempts at formulating a theory of the strong interaction based on elementary fields for baryons and mesons were made. In particular, uncountable PhD theses were written, based on local interactions of the Yukawa type, using perturbation theory to analyze them. The coupling constants invariably turned out to be numerically large, indicating that the neglect of the higher order contributions was not justified. Absolutely nothing worked even half way.

Although there was considerable progress in understanding the general principles of quantum field theory (Lorentz invariance, unitarity, crossing symmetry, causality, analyticity, dispersion relations, CPT theorem, spin and statistics) as well as in renormalization theory, faith in quantum field theory was in decline, even concerning QED (Landau pole). To many, the renormalization procedure – needed to arrive at physically meaningful results – looked suspicious, and it appeared doubtful that the strong interaction could at all be described by means of a local quantum field theory. Some suggested that this framework should be replaced by S-matrix theory – heated debates concerning this suggestion took place at the time [6]. Regge poles were considered as a promising alternative to the quantum fields (the Veneziano model is born in 1968 [7]). Sixty years ago, when I completed my studies, the quantum field theory of the strong interaction consisted of a collection of beliefs, prejudices and assumptions. Quite a few of these turned out to be wrong.

1.1.2 Flavor symmetries

Symmetries that extend isospin to a larger Lie group provided the first hints towards an understanding of the structure underneath the strong interaction phenomena. The introduction of the strangeness quantum number and the Gell-Mann–Nishijima formula [8,9] was a significant step in this direction. Goldberger and Treiman [10] then showed that the *axial vector current* plays an important role, not only in the weak interaction (the pion-to-vacuum matrix element of this current – the pion decay constant F_π – determines the rate of the weak decay $\pi \to \mu\nu$) but also in the context of the strong interaction: the nucleon matrix element of the axial vector current, g_A, determines the strength of the interaction between pions and nucleons:

$$g_{\pi N} = g_A M_N / F_\pi .$$

At low energies, the main characteristic of the strong interaction is that the energy gap is small: the lightest state occurring in the eigenvalue spectrum of the Hamiltonian is the pion, with[2] $M_\pi \simeq 135$ MeV, small compared to the mass of the proton, $M_p \simeq 938$ MeV. In 1960, Nambu found out why that is so: it has to do with a hidden, approximate, continuous symmetry [11]. Since some of its generators carry negative parity, it is referred to as a *chiral symmetry*. For this symmetry to be consistent with observation, it is essential that an analog of spontaneous magnetization occurs in particle physics: for dynamical reasons, the state of lowest energy – the vacuum – is not symmetric under chiral transformations. Consequently, the symmetry cannot be seen in the spectrum of the theory: it is *hidden* or *spontaneously broken*. Nambu realized that the spontaneous breakdown of a continuous symmetry entails massless particles analogous to the spin waves of a magnet and concluded that the pions must play this role. If the strong interaction was strictly invariant under chiral symmetry, there would be no energy gap at all – the pions would be massless.[3] Conversely, since the pions are not massless, chiral symmetry cannot be exact – unlike isospin, which at that time was taken to be an exact symmetry of the strong interaction. The spectrum does have an energy gap because chiral symmetry is not exact: the pions are not massless, only light. In fact, they represent the lightest strongly interacting particles that can be exchanged between two nucleons. This is why, at large distances, the potential between two nucleons is correctly described by the Yukawa formula.

The discovery of the *Eightfold Way* by Gell-Mann and Ne'eman paved the way to an understanding of the mass pattern of the baryons and mesons [13,14]. Like chiral symmetry, the group SU(3) that underlies the Eightfold Way represents an approximate symmetry: the spectrum of the mesons and baryons does not consist of degenerate multiplets of this group. The splitting between the energy levels, however, does exhibit a pattern that can be understood in terms of the assumption that the part of the Hamiltonian that breaks the symmetry transforms in a simple way. This led to the Gell-Mann–Okubo formula [14,15] and to a prediction for the mass of the Ω^-, a member of the baryon decuplet which was still missing, but was soon confirmed experimentally, at the predicted place [16].

[2] I am using natural units where $\hbar = c = 1$.

[3] A precise formulation of this statement, known as the *Goldstone theorem*, was given later [12].

1.1.3 Quark model

In 1964, Gell-Mann [17] and Zweig [18] pointed out that the observed pattern of baryons can qualitatively be understood on the basis of the assumption that these particles are bound states built with three constituents, while the spectrum of the mesons indicates that they contain only two of these. Zweig called the constituents "aces". Gell-Mann coined the term "quarks", which is now commonly accepted. The Quark Model gradually evolved into a very simple and successful semi-quantitative framework, but gave rise to a fundamental puzzle: why do the constituents not show up in experiment? For this reason, the existence of the quarks was considered doubtful: "Such particles [quarks] presumably are not real but we may use them in our field theory anyway …" [19]. Quarks were treated like the veal used to prepare a pheasant in the royal french cuisine: the pheasant was baked between two slices of veal, which were then discarded (or left for the less royal members of the court). Conceptually, this was a shaky cuisine.

If the flavor symmetries are important, why are they not exact? Gell-Mann found a beautiful explanation: *current algebra* [14,19]. The charges form an exact algebra even if they do not commute with the Hamiltonian and the framework can be extended to the corresponding currents, irrespective of whether or not they are conserved. Adler and Weisberger showed that current algebra can be tested with the sum rule that follows from the nucleon matrix element of the commutator of two axial vector charges [20,21]. Weinberg then demonstrated that even the strength of the interaction among the pions can be understood on the basis of current algebra: the $\pi\pi$ scattering lengths can be predicted in terms of the pion decay constant [22].

1.1.4 Behavior at short distances

Bjorken had pointed out that if the nucleons contain point-like constituents, then the ep cross section should obey scaling laws in the deep inelastic region [23]. Indeed, the scattering experiments carried out by the MIT-SLAC collaboration in 1968/69 did show experimental evidence for such constituents. Feynman called these *partons*, leaving it open whether they were the quarks or something else. For an account of the experimental developments, see the Nobel lectures of Taylor, Kendall and Friedman [24–26]. The comparison of the data on νp and $\bar{\nu} p$ scattering from Gargamelle [27,28] with the MIT-SLAC results confirmed that the partons indeed have fractional charges compatible with the predicted charges of quarks, $+\frac{2}{3}e$ and $-\frac{1}{3}e$. The evaluation of a sum rule for the momenta of the charged partons showed that (in the infinite momentum frame) half of the proton momentum is carried by neutral partons; now we know that these are gluons. Later, the CDHS collaboration also demonstrated that the quarks do have spin $s = 1/2$ while the gluons have spin $s = 1$ [29].

The operator product expansion turned out to be a very useful tool for the short distance analysis of the theory – the title of the paper where it was introduced [30], "Non-lagrangian models of current algebra," reflects the general skepticism towards Lagrangian quantum field theory that I mentioned in Sect. 1.1.1.

1.1.5 Color

The Quark Model was difficult to reconcile with the spin-statistics theorem which implies that particles of spin $\frac{1}{2}$ must obey Fermi statistics. Greenberg proposed that the quarks obey neither Fermi-statistics nor Bose-statistics, but "para-statistics of order three" [31]. The proposal amounts to the introduction of a new internal quantum number. Indeed, Bogolyubov, Struminsky and Tavkhelidze [32], Han and Nambu [33] and Miyamoto [34] independently pointed out that some of the problems encountered in the quark model disappear if the u, d and s quarks occur in 3 states. Gell-Mann coined the term "color" for the new quantum number.

One of the possibilities considered for the interaction that binds the quarks together was an abelian gauge field analogous to the e.m. field, but this gave rise to problems, because the field would then interfere with the other degrees of freedom. Fritzsch and Gell-Mann pointed out that if the gluons carry color, then the empirical observation that quarks appear to be confined might also apply to them: the spectrum of the theory might exclusively contain color neutral states [35].

In his lectures at the Schladming Winter School in 1972 [36], Gell-Mann thoroughly discussed the role of the quarks and gluons: theorists had to navigate between Scylla and Charybdis, trying to abstract neither too much nor too little from models built with these objects. The basic tool at that time was *current algebra on the light cone*. He invited me to visit Caltech. I did that during three months in the spring break of 1973 and spent an extremely interesting period there. The personal recollections of Harald Fritzsch (see Sect. 1.2) describe the developments that finally led to *Quantum Chromodynamics*.

As it was known already that the electromagnetic and weak interactions are mediated by gauge fields, the idea that color might be a local symmetry as well does not appear as far fetched. The main problem at the time was that for a gauge field theory to describe the hadrons and their interaction, it had to be fundamentally different from the quantum field theories encountered in nature so far: all of these, including the electroweak theory, have the spectrum indicated by the degrees of freedom occurring in the Lagrangian: photons, leptons, intermediate bosons, … The proposal can only make sense if this need not be so, that is if the spectrum of physical states in a quantum field theory can differ from the spectrum

of the fields needed to formulate it: gluons and quarks in the Lagrangian, hadrons in the spectrum. This looked like wishful thinking. How come that color is confined while electric charge is free?

1.1.6 Electromagnetic interaction

The final form of the laws obeyed by the electromagnetic field was found by Maxwell, around 1860 – these laws survived relativity and quantum theory, unharmed. Fock pointed out that the Schrödinger equation for electrons in an electromagnetic field,

$$\frac{1}{i}\frac{\partial\psi}{\partial t} - \frac{1}{2m_e^2}(\vec{\nabla} + i\,e\vec{A})^2\psi - e\,\varphi\,\psi = 0, \tag{1.1}$$

is invariant under a group of local transformations:

$$\vec{A}'(x) = \vec{A}(x) + \vec{\nabla}\alpha(x), \qquad \varphi'(x) = \varphi(x) - \frac{\partial\alpha(x)}{\partial t}$$
$$\psi(x)' = e^{-ie\alpha(x)}\,\psi(x), \tag{1.2}$$

in the sense that the fields $\vec{A}', \varphi', \psi'$ describe the same physical situation as $\vec{A}, \varphi, \psi$ [37]. Weyl termed these *gauge transformations* (with gauge group U(1) in this case). In fact, the electromagnetic interaction is fully characterized by symmetry with respect to this group: gauge invariance is the crucial property of this interaction.

I illustrate the statement with the core of Quantum Electrodynamics: photons and electrons. Gauge invariance allows only two free parameters in the Lagrangian of this system: e, m_e. Moreover, only one of these is dimensionless: $e^2/4\pi = 1/137.035\,999\,084\,(21)$. U(1) symmetry and renormalizability fully determine the properties of the e.m. interaction, except for this number, which so far still remains unexplained.

1.1.7 Nonabelian gauge fields

Kaluza [38] and Klein [39] had shown that a 5-dimensional Riemann space with a metric that is independent of the fifth coordinate is equivalent to a 4-dimensional world with *gravity*, a *gauge field* and a *scalar field*. In this framework, gauge transformations amount to a shift in the fifth direction: $x^{5'} = x^5 + \alpha(\vec{x}, t)$. In geometric terms, a metric space of this type is characterized by a group of isometries: the geometry remains the same along certain directions, indicated by Killing vectors. In the case of the 5-dimensional spaces considered by Kaluza and Klein, the isometry group is the abelian group U(1). The fifth dimension can be compactified to a circle – U(1) then generates motions on this circle. A particularly attractive feature of this theory is that it can explain the quantization of the electric charge: fields living on such a manifold necessarily carry integer multiples of a basic charge unit.

Pauli noticed that the Kaluza-Klein scenario admits a natural generalization to higher dimensions, where larger isometry groups find place. Riemann spaces of dimension > 5 admit nonabelian isometry groups that reduce the system to a 4-dimensional one with *gravity*, *nonabelian gauge fields* and several *scalar fields*. Pauli was motivated by the isospin symmetry of the meson-nucleon interaction and focused attention on a Riemann space of dimension 6, with isometry group SU(2).

Pauli did not publish the idea that the strong interaction might arise in this way, because he was convinced that the quanta of a gauge field are massless: gauge invariance does not allow one to put a mass term into the Lagrangian. He concluded that the forces mediated by gauge fields are necessarily of long range and can therefore not mediate the strong interaction, which is known to be of short range. More details concerning Pauli's thoughts can be found in [40]. The paper of Yang and Mills appeared in 1954 [41]. Ronald Shaw, a student of Salam, independently formulated nonabelian gauge field theory in his PhD thesis [42]. Ten years later, Higgs [43], Brout and Englert [44] and Guralnik, Hagen and Kibble [45] showed that Pauli's objection is not valid in general: in the presence of scalar fields, gauge fields can pick up mass, so that forces mediated by gauge fields can be of short range. The work of Glashow [46], Weinberg [47] and Salam [48] then demonstrated that nonabelian gauge fields are relevant for physics: the framework discovered by Higgs et al. does accommodate a realistic description of the e.m. and weak interactions.

1.1.8 Asymptotic freedom

Already in 1965, Vanyashin and Terentyev [49] found that the renormalization of the electric charge of a vector field is of opposite sign to the one of the electron. In the language of SU(2) gauge field theory, their result implies that the β-function is negative at one loop.

The first correct calculation of the β-function of a nonabelian gauge field theory was carried out by Khriplovich, for the case of SU(2), relevant for the electroweak interaction [50]. He found that β is negative and concluded that the interaction becomes weak at short distance. In his PhD thesis, 't Hooft performed the calculation of the β-function for an arbitrary gauge group, including the interaction with fermions and Higgs scalars [51,52]. He demonstrated that the theory is renormalizable and confirmed that, unless there are too many fermions or scalars, the β-function is negative at small coupling.

In 1973, Gross and Wilczek [53] and Politzer [54] discussed the consequences of a negative β-function and suggested that this might explain Bjorken scaling, which had been observed at SLAC in 1969. They pointed out that QCD

predicts specific modifications of the scaling laws. In the meantime, there is strong experimental evidence for these.

1.1.9 Arguments in favor of QCD

The reasons for proposing QCD as a theory of the strong interaction are discussed in [55]. The idea that the observed spectrum of particles can fully be understood on the basis of a theory built with quarks and gluons still looked rather questionable and was accordingly formulated in cautious terms. In the abstract, for instance, we pointed out that "…there are several advantages in abstracting properties of hadrons and their currents from a Yang–Mills gauge model based on colored quarks and color octet gluons." Before the paper was completed, the papers by Gross, Wilczek and Politzer quoted above circulated as preprints – they are quoted and asymptotic freedom is given as argument #4 in favor of QCD. Also, important open questions were pointed out, in particular, the U(1) problem.

Many considered QCD a wild speculation. On the other hand, several papers concerning gauge field theories that include the strong interaction appeared around the same time, for instance [56,57].

1.1.10 November revolution

The discovery of the J/ψ was announced simultaneously at Brookhaven and SLAC, on November 11, 1974. Three days later, the observation was confirmed at ADONE, Frascati and ten days later, the ψ' was found at SLAC, where subsequently many further related states were discovered. We now know that these are bound states formed with the c-quark and its antiparticle which is comparatively heavy and that there are two further, even heavier quarks: b and t.

At sufficiently high energies, quarks and gluons do manifest themselves as jets. Like the neutrini, they have left their theoretical place of birth and can now be seen flying around like ordinary, observable particles. Gradually, particle physicists abandoned their outposts in no man's and no woman's land, returned to the quantum fields and resumed discussion in the good old *Gasthaus zu Lagrange*, a term coined by Jost. The theoretical framework that describes the strong, electromagnetic and weak interactions in terms of gauge fields, leptons, quarks and scalar fields is now referred to as the Standard Model – this framework clarified the picture enormously.[4]

1.1.11 Quantum chromodynamics

If the electroweak gauge fields as well as the leptons and the scalars are dropped, the Lagrangian of the Standard Model reduces to QCD:

$$\mathcal{L}_{QCD} = -\frac{1}{4} F^A_{\mu\nu} F^{A\,\mu\nu} + i\bar{q}\gamma^\mu(\partial_\mu + i g_s \frac{1}{2}\lambda^A \mathcal{A}^A_\mu) q \\ -\bar{q}_R \mathcal{M} q_L - \bar{q}_L \mathcal{M}^\dagger q_R - \theta\,\omega. \tag{1.3}$$

The gluons are described by the gauge field $\mathcal{A}^A_\mu$, which belongs to the color group $\mathrm{SU_c}(3)$ and g_s is the corresponding coupling constant. The field strength tensor $F^A_{\mu\nu}$ is defined by

$$F^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu - g_s f_{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu, \tag{1.4}$$

where the symbol f_{ABC} denotes the structure constants of SU(3). The quarks transform according to the fundamental representation of $\mathrm{SU_c}(3)$. The compact notation used in (1.3) suppresses the labels for flavor, color and spin: the various quark flavors are represented by Dirac fields, $q = \{u, d, s, c, b, t\}$ and $q_R = \frac{1}{2}(1+\gamma_5)q$, $q_L = \frac{1}{2}(1-\gamma_5)q$ are their right- and left-handed components. The field $u(x)$, for instance, contains 3×4 components. While the 3×3 Gell-Mann matrices λ^A act on the color label and satisfy the commutation relation

$$[\lambda^A, \lambda^B] = 2i f_{ABC} \lambda^C, \tag{1.5}$$

the Dirac matrices γ^μ operate on the spin index. The mass matrix $\mathcal{M}$, on the other hand, acts in flavor space. Its form depends on the choice of the quark field basis. If the right- and left-handed fields are subject to independent rotations, $q_R \to V_R q_R$, $q_L \to V_L q_L$, where $V_R, V_L \in \mathrm{U}(N_f)$ represent $N_f \times N_f$ matrices acting on the quark flavor, the quark mass matrix is replaced by $\mathcal{M} \to V_R^\dagger \mathcal{M} V_L$. This freedom can be used to not only diagonalize $\mathcal{M}$, but to ensure that the eigenvalues are real, nonnegative and ordered according to $0 \le m_u \le m_d \le \cdots \le m_t$.

As it is the case with electrodynamics, gauge invariance fully determines the form of the chromodynamic interaction. The main difference between QED and QCD arises from the fact that the corresponding gauge groups, U(1) and SU(3), are different. While the structure constants of U(1) vanish because this is an abelian group, those of $\mathrm{SU_c}(3)$ are different from zero. For this reason, gauge invariance implies that the Lagrangian contains terms involving three or four gluon fields: in contrast to the photons, which interact among themselves only via the exchange of charged particles, the gluons would interact even if quarks did not exist.

[4] Indeed, the success of this theory is amazing: Gauge fields are renormalizable in four dimensions, but it looks unlikely that the Standard Model is valid much beyond the explored energy range. Presumably it represents an effective theory. There is no reason, however, for an effective theory to be renormalizable. One of the most puzzling aspects of the Standard Model is that it is able to account for such a broad range of phenomena that are characterized by very different scales within one and the same renormalizable theory.

The Lagrangian in Eq. (1.3) includes a parity-violating term proportional to the winding number density,

$$\omega = \frac{g_s^2}{32\pi^2} F^A_{\mu\nu} \tilde{F}^{A\,\mu\nu}, \tag{1.6}$$

where $\tilde{F}^{A\,\mu\nu} \equiv \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} F^A_{\rho\sigma}$ is the dual of the field strength. The constant θ is referred to as the vacuum angle. Since ω can be represented as a derivative, $\omega = \partial_\mu f^\mu$, the θ-term looks irrelevant: only the integral over the Lagrangian counts, so that the contribution from this term is determined by the behaviour of the gauge field at the boundary of space-time. In the case of QED, where renormalizability allows the presence of an analogous term, quantities of physical interest are indeed unaffected by such a contribution, but for QCD, this is not the case. Even at the classical level, nonabelian gauge fields can form instantons, which minimize the Euclidean action for a given nonzero winding number $\nu = \int d^4x\, \omega$.

The θ-term did not play a significant role in the developments that led to QCD. Indeed, neither the QCD Lagrangian specified in Ref. [55], nor the discussion of the origins of QCD in Sect. 1.2 of the present review involve such a term. Also, there is no experimental evidence indicating that the strong interaction might violate parity. In the present understanding of QCD, however, the θ-term plays a central role, because it is intimately related to an important property of QCD: the Ward identity obeyed by the singlet axial current contains an anomaly proportional to ω. An immediate consequence of this identity is that the change of the quark field basis considered above entails a change not only of the quark mass matrix, but also of the vacuum angle. Quite apart from that, the anomaly very strongly affects the physics of the strong interaction, in particular the spectrum of the theory – some of the implications are briefly discussed below.

1.1.12 Theoretical paradise

In order to briefly discuss some of the basic properties of QCD, let me turn off the electroweak interaction, treat the three light quarks as massless and the remaining ones as infinitely heavy:

$$m_u = m_d = m_s = 0, \qquad m_c = m_b = m_t = \infty. \tag{1.7}$$

The Lagrangian then contains a single parameter: the coupling constant g_s, which may be viewed as the net color of a quark. Unlike an electron, a quark cannot be isolated from the rest of the world – its color g_s depends on the radius of the region considered. According to perturbation theory, the color contained in a sphere of radius r grows logarithmically with the radius[5]:

$$\alpha_s \equiv \frac{g_s^2}{4\pi} = \frac{2\pi}{9\,|\ln(r\,\Lambda)|}. \tag{1.8}$$

Although the classical Lagrangian of massless QCD does not contain any dimensionful parameter, the corresponding quantum field theory does: the strength of the interaction cannot be characterized by a number, but by a dimensionful quantity, the intrinsic scale Λ.

The phenomenon is referred to as *dimensional transmutation*. In perturbation theory, it manifests itself through the occurrence of divergences – contrary to what many quantum field theorists thought for many years, the divergences do not represent a disease, but are intimately connected with the structure of the theory. They are a consequence of the fact that a quantum field theory does not inherit all of the properties of the corresponding classical field theory. In the case of massless Chromodynamics, the classical Lagrangian does not contain any dimensionful constants and hence remains invariant under a change of scale. This property, which is referred to as conformal invariance, does not survive quantization, however. Indeed, it is crucial for Quantum Chromodynamics to be consistent with what is known about the strong interaction that this theory does have an intrinsic scale.

Massless QCD is how theories should be: the Lagrangian does not contain a single dimensionless parameter. In principle, the values of all quantities of physical interest are predicted without the need to tune parameters (the numerical value of the mass of the proton in kilogram units cannot be calculated, of course, because that number depends on what is meant by a kilogram, but the mass spectrum, the width of the resonances, the cross sections, the form factors, … can be calculated in a parameter free manner from the mass of the proton, at least in principle).

1.1.13 Symmetries of massless QCD

The couplings of the u-, d- and s-quarks to the gauge field are identical. In the *chiral limit*, where the masses are set equal to zero, there is no difference at all – the Lagrangian is symmetric under SU(3) rotations in flavor space. Indeed, there is more symmetry: for massless fermions, the right- and left-handed components can be subject to independent flavor rotations. The Lagrangian of QCD with three massless flavors is invariant under $\mathrm{SU}(3)_R \times \mathrm{SU}(3)_L$. QCD thus explains the presence of the mysterious chiral symmetry discovered by Nambu: an exact symmetry of this type is present if some of the quarks are massless.

Nambu had conjectured that chiral symmetry breaks down spontaneously. Can it be demonstrated that the symmetry group $\mathrm{SU}(3)_R \times \mathrm{SU}(3)_L$ of the Lagrangian of massless QCD

[5] The formula only holds if the radius is small, $r\,\Lambda \ll 1$.

spontaneously break down to the subgroup $SU(3)_{R+L}$? To my knowledge an analytic proof is not available, but the work done on the lattice demonstrates beyond any doubt that this does happen. In particular, for $m_u = m_d = m_s$, the states do form degenerate multiplets of $SU(3)_{R+L}$ and, in the limit $m_u, m_d, m_s \to 0$, the pseudoscalar octet does become massless, as required by the Goldstone theorem.

1.1.14 Quark masses

The 8 lightest mesons, $\pi^+, \pi^0, \pi^-, K^+, K^0, \bar{K}^0, K^-, \eta$, do have the quantum numbers of the Nambu–Goldstone bosons, but massless they are not. The reason is that we are not living in the paradise described above: the light quark masses are different from zero. Accordingly, the Lagrangian of QCD is only approximately invariant under chiral rotations, to the extent that the symmetry breaking parameters m_u, m_d, m_s are small. Since they differ, the multiplets split. In particular, the Nambu–Goldstone bosons pick up mass.

Even before the discovery of QCD, attempts at estimating the masses of the quarks were made. In particular, nonrelativistic bound state models for mesons and baryons where constructed. In these models, the proton mass is dominated by the sum of the masses of its constituents: $m_u + m_u + m_d \simeq m_p$, $m_u \simeq m_d \simeq 300\,\text{MeV}$.

With the discovery of QCD, the mass of the quarks became an unambiguous concept: the quark masses occur in the Lagrangian of the theory. Treating the mass term as a perturbation, one finds that the expansion of $m^2_{\pi^+}$ in powers of m_u, m_d, m_s starts with $m^2_{\pi^+} = (m_u + m_d)B_0 + \cdots$ The constant B_0 also determines the first term in the expansion of the square of the kaon masses: $m^2_{K^+} = (m_u + m_s)B_0 + \cdots$, $m^2_{K^0} = (m_d + m_s)B_0 + \cdots$ Since the kaons are significantly heavier than the pions, these relations imply that m_s must be large compared to m_u, m_d.

The first crude estimate of the quark masses within QCD relied on a model for the wave functions of π, K, ρ, which was based on SU(6) (spin-flavor-symmetry) and led to $B_0 \simeq \frac{3}{2} m_\rho F_\rho / F_\pi$. Numerically, this yields $B_0 \simeq 1.8\,\text{GeV}$. For the mean mass of the two lightest quarks, $m_{ud} \equiv \frac{1}{2}(m_u + m_d)$, this estimate implies $m_{ud} \simeq 5\,\text{MeV}$, while the mass of the strange quark becomes $m_s \simeq 135\,\text{MeV}$ [58]. Similar mass patterns were found earlier, within the Nambu–Jona–Lasinio model [59] or on the basis of sum rules [60].

1.1.15 Breaking of isospin symmetry

From the time Heisenberg had introduced isospin symmetry, it was taken for granted that the strong interaction strictly conserves isospin. QCD does have this symmetry if and only if $m_u = m_d$. If that condition were met, the mass difference between proton and neutron would be due exclusively to the e.m. interaction. This immediately gives rise to a qualitative problem: why is the charged particle, the proton, lighter than its neutral partner, the neutron?

The Cottingham formula [61] states that the leading contribution of the e.m. interaction to the mass of a particle is determined by the cross section for electron scattering on this particle. We evaluated the formula on the basis of Bjorken scaling and of the experimental data for electron scattering on protons and neutrons available at the time. Since we found that the electromagnetic self energy of the proton is larger than the one of the neutron, we concluded that the strong interaction does not conserve isospin: even if the e.m. interaction is turned off, m_u must be different from m_d. In fact, the first crude estimate for the masses of the light quarks [62],

$$m_u \simeq 4\,\text{MeV}, \quad m_d \simeq 7\,\text{MeV}, \quad m_s \simeq 135\,\text{MeV}, \tag{1.9}$$

indicated that m_d must be almost twice as large as m_u.

It took quite a while before this bizarre pattern was generally accepted. The Dashen theorem [63] states that, in a world where the quarks are massless, the e.m. self energies of the kaons and pions obey the relation $m^{2\,em}_{K^+} - m^{2\,em}_{K^0} = m^{2\,em}_{\pi^+} - m^{2\,em}_{\pi^0}$. If the mass differences were dominated by the e.m. interaction, the charged kaon would be heavier than the neutral one. Hence the mass difference between the kaons cannot be due to the electromagnetic interaction, either. The estimates for the quark mass ratios obtained with the Dashen theorem confirm the above pattern [64].

1.1.16 Approximate symmetries are natural in QCD

At first sight, the fact that m_u strongly differs from m_d is puzzling: if this is so, why is isospin such a good quantum number? The key observation here is the one discussed in Sect. 1.1.12: QCD has an intrinsic scale, Λ. For isospin to represent an approximate symmetry, it is not necessary that $m_d - m_u$ is small compared to $m_u + m_d$. It suffices that the symmetry breaking parameter is small compared to the intrinsic scale, $m_d - m_u \ll \Lambda$.

In the case of the eightfold way, the symmetry breaking parameters are the differences between the masses of the three light quarks. If they are small compared to the intrinsic scale of QCD, then the Green functions, masses, form factors, cross sections … are approximately invariant under the group $SU(3)_{R+L}$. Isospin is an even better symmetry, because the relevant symmetry breaking parameter is smaller, $m_d - m_u \ll m_s - m_u$. The fact that $m^2_{\pi^+}$ is small compared to $m^2_{K^+}$ implies $m_u + m_d \ll m_u + m_s$. Hence all three light quark masses must be small compared to the scale of QCD.

In the framework of QCD, the presence of an approximate chiral symmetry group of the form $SU(3)_R \times SU(3)_L$ thus has a very simple explanation: it so happens that the masses of u, d and s are small. We do not know why, but there is no doubt that

this is so. The quark masses represent a perturbation, which in first approximation can be neglected – in first approximation, the world is the paradise described above.

1.1.17 Ratios of quark masses

The confinement of color implies that the masses of the quarks cannot be identified by means of the four-momentum of a one-particle state – the spectrum of the theory does not contain such states. As parameters occurring in the Lagrangian, they need to be renormalized and the renormalized mass depends on the regularization used to set up the theory. In the $\overline{\rm MS}$ scheme [65–67], they depend on the running scale – only their ratios represent physical quantities. Among the three lightest quarks, there are two independent mass ratios, which it is convenient to identify with

$$S = \frac{m_s}{m_{ud}}, \qquad R = \frac{m_s - m_{ud}}{m_d - m_u}, \tag{1.10}$$

where $m_{ud} \equiv \frac{1}{2}(m_u + m_d)$.

Since the isospin breaking effects due to the e.m. interaction are not negligible, the physical masses of the Goldstone boson octet must be distinguished from their masses in QCD, i.e. in the absence of the electroweak interactions. I denote the latter by $\hat{m}_P$ and use the symbol $\hat{m}_K$ for the mean square kaon mass in QCD, $\hat{m}_K^2 \equiv \frac{1}{2}(\hat{m}_{K^+}^2 + \hat{m}_{K^0}^2)$. The fact that the expansion of the square of the Goldstone boson masses in powers of m_u, m_d, m_s starts with a linear term implies that, in the chiral limit, their ratios are determined by R and S. In particular, the expansion of the ratios of $\hat{m}_{\pi^+}^2, \hat{m}_{K^+}^2$ and $\hat{m}_{K^0}^2$ starts with

$$\frac{2\hat{m}_K^2}{\hat{m}_{\pi^+}^2} = (S+1)\{1 + \Delta_S\}, \tag{1.11}$$

$$\frac{\hat{m}_K^2 - \hat{m}_{\pi^+}^2}{\hat{m}_{K^0}^2 - \hat{m}_{K^+}^2} = R\{1 + \Delta_R\}, \tag{1.12}$$

where Δ_S as well as Δ_R vanish in the chiral limit – they represent corrections of $O(\mathcal{M})$. The left hand sides only involve the masses of π^+, K^+ and K^0. Invariance of QCD under charge conjugation implies that the masses of π^-, K^- and $\bar{K}^0$ coincide with these. There are low energy theorems analogous to (1.11), (1.12), involving the remaining members of the octet, π^0 and η, but these are more complicated because the states $|\pi^0\rangle$ and $|\eta\rangle$ undergo mixing {at leading order, chiral symmetry implies that the mixing angle is given by $\tan(2\theta) = \sqrt{3}/2R$}. In the isospin limit, $\{m_u = m_d, e = 0\}$, the masses of π^0 and π^+ coincide and $\hat{m}_\eta$ obeys the Gell-Mann–Okubo formula, $(\hat{m}_\eta^2 - \hat{m}_K^2)/(\hat{m}_K^2 - \hat{m}_\pi^2) = \frac{1}{3}\{1 + O(\mathcal{M})\}$.

While the accuracy to which S can be determined on the lattice is amazing, the uncertainty in R is larger by almost an order of magnitude [68]:

$$S = 27.42(12), \qquad R = 38.1(1.5). \tag{1.13}$$

The reason is that R concerns isospin breaking effects. The contributions arising from QED are not negligible at this precision and since the e.m. interaction is of long range, it is more difficult to simulate on a lattice.

The difference shows up even more clearly in the corrections. The available lattice results [68] lead to $\Delta_S = -0.055(6)$, indicating that the low energy theorem (1.11) picks up remarkably small corrections from higher orders of the quark mass expansion. Those occurring in the Gell-Mann–Okubo formula are also known to be very small. The number $\Delta_R = -0.016(57)$ obtained from the available lattice results is also small, but the uncertainty is so large that even the sign of the correction remains open.

The quantities Δ_S, Δ_R exclusively concern QCD and could be determined to high precision with available methods, in the framework of $N_f = 1 + 1 + 1$: three flavors of different mass. For isospin breaking quantities, the available results come with a large error because they do not concern QCD alone but are obtained from a calculation of the physical masses, so that the e.m. interaction cannot be ignored. A precise calculation of $\hat{m}_{\pi^+}$, $\hat{m}_{K^+}$, $\hat{m}_{K^0}$ within lattice QCD would be of considerable interest as it would allow to subject a venerable low energy theorem for the quark mass ratio $Q^2 \equiv (m_s^2 - m_{ud}^2)/(m_d^2 - m_u^2)$ [69] to a stringent test. The theorem implies that the leading contributions to Δ_R and Δ_S are equal in magnitude, but opposite in sign: $\Delta_R = -\Delta_S + O(\mathcal{M}^2)$ [70]. The available numbers are consistent with this relation but far from accurate enough to allow a significant test. There is no doubt that the leading terms dominate if the quark masses are taken small enough, but since the estimates for Δ_R and Δ_S obtained at the physical values of the quark masses turn out to be unusually small, it is conceivable that the corrections of $O(\mathcal{M}^2)$ are of comparable magnitude. For $m_u = m_d$, the masses of the Goldstone bosons have been worked out to NNLO of Chiral Perturbation Theory [71]. An extension of these results to $\hat{m}_{\pi^+}, \hat{m}_{K^+}$, $\hat{m}_{K^0}$ for $m_u \neq m_d$ should be within reach and would allow a much more precise lattice determination of Δ_R.

1.1.18 U(1) anomaly, CP-problem

Even before the discovery of QCD, it was known that, in the presence of vector fields, the Ward identities for axial currents contain anomalies [72–74]. In particular, an external e.m. field generates an anomaly in the conservation law for the axial current $\bar{u}\gamma^\mu\gamma_5 u - \bar{d}\gamma^\mu\gamma_5 d$. The anomaly implies a low energy theorem for the decay $\pi^0 \to \gamma + \gamma$, which states that, to leading order in the expansion in powers of the momenta and for $m_u = m_d = 0$, the transition amplitude is

determined by F_π, i.e. by the same quantity that determines the rate of the decay $\pi^+ \to \mu + \nu_\mu$.

In QCD, the conservation law for the singlet axial current contains an anomaly,

$$\partial_\mu(\bar{q}\gamma^\mu\gamma_5 q) = 2i\,\bar{q}\mathcal{M}\gamma_5 q + 2N_f\,\omega, \tag{1.14}$$

where N_f is the number of flavors and ω is specified in (1.6). The phenomenon plays a crucial role because it implies that even if the quark mass matrix $\mathcal{M}$ is set equal to zero, the singlet axial charge is not conserved. Hence the symmetry group of QCD with 3 massless flavors is $\mathrm{SU(3)_R \times SU(3)_L \times U(1)_{R+L}}$, not $\mathrm{U(3)_R \times U(3)_L}$. QCD is not invariant under the chiral transformations generated by the remaining factor, $\mathrm{U(1)_{R\text{-}L}}$. This is why the paradise described above contains 8 rather than 9 massless Goldstone bosons.

The factor $\mathrm{U(1)_{R\text{-}L}}$ changes the phase of the right-handed components of all quark fields by the same angle, $q'_{\mathrm{R}} = e^{i\beta}q_{\mathrm{R}}$, while the left-handed components are subject to the opposite transformation: $q'_{\mathrm{L}} = e^{-i\beta}q_{\mathrm{L}}$. This change of basis can be compensated by modifying the quark mass matrix with $\mathcal{M}' = e^{2i\beta}\mathcal{M}$, but in view of the anomaly, the operation does not represent a symmetry of the system. The relation (1.14) shows, however, that current conservation is not lost entirely – it only gets modified. In fact, if the above change of the quark mass matrix is accompanied by a simultaneous change of the vacuum angle, $\theta' = \theta - 2\beta$, the physics does remain the same. Note that, starting from an arbitrary mass matrix, a change of basis involving the factor $\mathrm{U(1)_{R\text{-}L}}$ is needed to arrive at the convention where $\mathcal{M}$ is diagonal with real eigenvalues. In that convention, the vacuum angle does have physical significance – otherwise only the product $e^{i\theta}\mathcal{M}$ counts.

The Lagrangian of QCD is invariant under charge conjugation, but the term $-\theta\,\omega$ has negative parity. Accordingly, unless θ is very small, there is no explanation for the fact that CP-violating quantities such as the electric dipole moment of the neutron are too small to have shown up in experiment. This is referred to as the strong CP-problem.

There is a theoretical solution of this puzzle: if the lightest quark were massless, $m_u = 0$, QCD would conserve CP. The Dirac field of the u-quark can then be subject to the chiral transformation $u'_{\mathrm{R}} = e^{i\beta}u_{\mathrm{R}}$, $u'_{\mathrm{L}} = e^{-i\beta}u_{\mathrm{L}}$ without changing the quark mass matrix. As discussed above, the physics remains the same, provided the vacuum angle is modified accordingly. This shows that if one of the quarks were massless, the vacuum angle would become irrelevant. It would then be legitimate to set $\theta = 0$, so that the Lagrangian becomes manifestly CP-invariant.

This 'solution', however, is fake. If m_u were equal to zero, the ratio R would be related to S by $R = \frac{1}{2}(S-1)$. The very accurate value for S in Eq. (1.13) would imply $R = 13.21(6)$, more than 16 standard deviations away from the result quoted for R.

1.1.19 QCD as part of the standard model

In the Standard Model, the vacuum contains a condensate of Higgs bosons. At low energies, the manner in which the various other degrees of freedom interact with these plays the key role. Since they do not have color and are electrically neutral, their condensate is transparent for gluons and photons. The gauge bosons $W^\pm$, Z that mediate the weak interaction, as well as the leptons and quarks do interact with the condensate: photons and gluons remain massless, all other particles occurring in the Standard Model are hindered in moving through the condensate and hence pick up mass. In cold matter only the lightest degrees of freedom survive: photons, gluons, electrons, u- and d-quarks – all other particles are unstable, decay and manifest their presence only indirectly, through quantum fluctuations.

At low energies, the Standard Model boils down to a remarkably simple theory: QCD + QED. The Lagrangian only contains the coupling constants g_s, e, θ and the masses of the quarks and leptons as free parameters, but describes the laws of nature relevant at low energies to breathtaking precision. The gluons and the photons represent the gauge fields that belong to color and electric charge, respectively. Color is confined, but electric charge is not: while electrons can move around freely, quarks and gluons form color neutral bound states – mesons, baryons, nuclei.

The structure of the atoms is governed by QED because the e.m. interaction is of long range. In particular, their size is of the order of the Bohr radius, $a_B = 4\pi/e^2 m_e$, which only involves the mass of the electron and the coupling constant e. The mass of the atoms, on the other hand, is dominated by the energy of the gluons and quarks that are bound in the nucleus. It is of the order of the scale Λ_{QCD}, which characterizes the value of g_s in a renormalization group invariant manner. Evidently, the sum of the charges of the quarks contained in the nucleus also matters, as it determines the number of electrons that can be bound to it. The mass of the quarks, on the other hand, plays an important role only in so far as it makes the proton the lightest baryon – the world would look rather different if the neutron was lighter …

The properties of the interaction among the quarks and gluons does not significantly affect the structure of the atoms, but from the theoretical point of view, the gauge field theory that describes it, QCD, is the most remarkable part of the Standard Model. In fact, it represents the first non-trivial quantum field theory that is internally consistent in four-space-time dimensions. In contrast to QED or to the Higgs sector, QCD is asymptotically free. The behavior of the quark and gluon fields at very short distances is under control. A cutoff is needed to set the theory up, but it can unambiguously be removed. In principle, all of the physical quantities of interest are determined by the renormalization group invariant quark mass matrix, by the vacuum angle θ and a

scale. In the basis where the quark mass matrix is diagonal and real, the vacuum angle is tiny. We do not know why this is so, nor do we understand the bizarre pattern of eigenvalues.

1.2 The origins of QCD

Harald Fritzsch

Murray Gell-Mann and I started to collaborate in October 1970. We considered the results of the experiments on deep inelastic scattering at the Stanford Linear Accelerator Center. James Bjorken had predicted, using current algebra, that the cross sections showed at large values of the virtual photon mass and the energy transfer to the nucleon a scaling behavior, i.e. the cross section is a function of the ratio x, where x is the ratio of the square of the virtual photon mass to the energy transfer to the nucleon, multiplied with the nucleon mass. This ratio x varies from zero to one.

Since in the scaling region the cross sections were determined by the commutator of two electromagnetic currents at nearly lightlike distances, Gell-Mann and I assumed, that this commutator near the light cone is given by the free quark model. Thus the Bjorken scaling followed from this assumption.

The interaction between the quarks was assumed not to be present near the light cone. The cross section in the deep inelastic region determined the distribution functions of the three quarks and antiquarks, which are given by the proton matrix element of the commutator of the electromagnetic current.

In the free quark model the commutator near the light cone is given by a singular function, multiplied by a bilocal function of quark fields [75]. The matrix elements of these bilocal operators determined the quark distribution functions of the nucleon. The integral of the quark distribution functions gives the contribution of all the quark momenta to the nucleon momentum.

Gell-Mann and I expected that the momentum sumrule of the proton constituents should be +1. However, it turned out that the integral

$$\int_0^1 x\left[u(x)+\overline{u}(x)+d(x)+\overline{d}(x)+s(x)+\overline{s}(x)\right]\mathrm{d}x \\ (0.52\pm 0.03). \tag{1.15}$$

was found to be only ≈ 0.5 [28], thus indicating that besides the charged partons there must exist also neutral partons in the proton (see also "Behavior at short distances" 1.1.4 in the preceding contribution). This observation was the first indication that the strong interactions are described by a gauge theory. In such a theory there would be besides the quarks and antiquarks also neutral gluons.

Afterwards Gell-Mann and I considered several problems of the quark theory. The Ω^- particle was a bound state of three strange quarks. The three spin vectors of the quarks were symmetricals arranged, and the space wave function was symmetric, since the Ω^- is the ground state of three strange quarks. Thus an interchange of two strange quarks was symmetric, but according to the Pauli principle it should be antisymmetric.

Another problem was related to the electromagnetic decay of the neutral pion. The decay rate, calculated in the quark model, is much smaller than the observed decay rate, only about 1/9 of the observed rate.

We also studied the cross section for the reaction electron–positron annihilation into hadrons. The ratio R of the cross section for hadron production and the cross section for the production of a muon pair can be calculated in the quark model. It is given by the sum of the squares of the electric charges of the three quarks, i.e. 2/3. But according to the experiments at the CEA accelerator at Harvard university this ratio was about three times larger: $R \simeq 2$.

To solve theses problems, Murray Gell-Mann, William Bardeen and I introduced for the quarks a new quantum number, which we called "color". Each quark is described by a red, a green and a blue quark. The three colors can be transformed by the color group $SU(3)_C$, which is assumed to be an exact symmetry. Measurable quantities, e.g. cross sections or the wave functions of hadrons, are color singlets.

The quark wave function ψ_Ω of the Ω^- is also a color singlet:

$$\psi_\Omega \simeq (rgb - grb + brg - rbg + gbr - bgr). \tag{1.16}$$

This wave function is antisymmetric under the exchange of two quarks – there is no problem with the Pauli principle. The quark wave functions of mesons are also color singlets:

$$\psi_{\mathrm{meson}} \simeq (\bar{r}r + \bar{g}g + \bar{b}b). \tag{1.17}$$

The decay amplitude for the neutral pion decay is three times larger, if the quarks are colored. Thus the decay rate is nine times larger and agrees with the observed decay rate [76]. The ratio R for electron–positron annihilation, given by the sum of the squares of the quark charges, is now also three times larger: $R \simeq 2$. Thus the introduction of the color quantum number solved the three problems mentioned above.

The color quantum number also explains why mesons are quark–antiquark bound states and baryons are three quark bound states, since they must be color singlets. Thus the mesons and baryons could be considered to be "white" states, since a particular color cannot be seen from the outside – the color quantum number is only relevant inside the mesons and baryons.

In the spring of 1972 Gell-Mann and I tried to understand why a colored quark cannot be observed – it is confined inside a baryon or meson or inside an atomic nucleus. We considered to use the color symmetry group as a gauge group. The gauge bosons of such a gauge theory would be color

octets. I proposed to call these gauge bosons "chromons", but Gell-Mann insisted to call them "gluons", mixing the English language and the Greek language.

We called this new gauge theory "Quantum Chromodynamics" (QCD). The Lagrangian of QCD is [35,55]:

$$\mathcal{L} = \overline{q}\left[i\gamma^\mu\left(\partial_\mu + ig_s\frac{\lambda^A}{2}\mathcal{A}^A_\mu\right) - m\right]q - \frac{1}{4}F^A_{\mu\nu}F^{A\,\mu\nu}, \tag{1.18}$$

where the λ^A are the Gell-Mann matrices, and

$$F^A_{\mu\nu} = \partial_\mu\mathcal{A}^A_\nu - \partial_\nu\mathcal{A}^A_\mu - g_s f_{ABC}\mathcal{A}^B_\mu\mathcal{A}^C_\nu. \tag{1.19}$$

f_{ABC} are called SU(3) structure constants. This Lagrangian is very similar to the Lagrangian of Quantum Electrodynamics. The electromagnetic field is replaced by the eight gluon fields $\mathcal{A}^A$, the electron mass by the quark mass, and the charge e is replaced by the strong coupling g_s. The strong interaction constant is defined by $\alpha_s = g_s^2/4\pi$.

However, the big difference between Quantum Electrodynamics and Quantum Chromodynamics is the presence of the $\mathcal{A}^2$ term in $F^A_{\mu\nu}$, not present in Quantum Electrodynamics. This term shows that a gluon interacts not only with a quark, but also with another gluon, and gives rise to 3- and 4-gluon couplings.

The quark masses, which appear in the Lagrangian of QCD, are not the masses of free quarks, but the masses, relevant inside the hadrons. The masses of the quarks depend on the energy scale. They are large at small energies and small at high energies. Here are the typical masses for the up-quark, the down-quark and the strange quark at the energy given by the mass of the Z-boson, $M_Z \simeq 91.2\,\text{GeV}$:

$$m_u \simeq 1.2\,\text{MeV},\ m_d \simeq 2.2\,\text{MeV},\ m_s \simeq 53\,\text{MeV}. \tag{1.20}$$

These masses describe the symmetry breaking of the $SU(3)_\text{F}$ flavor group. Interesting is the violation of the isospin symmetry. The down quark is heavier than the up quark. For this reason the neutron is heavier than the proton, and the proton is stable. If there would be no isospin violation, i.e. $m_u = m_d$, the proton would be heavier than the neutron due to the electromagnetic self-energy and it would decay into the neutron – life would not be possible.

Gell-Mann and I assumed that the interaction in QCD is zero at light-like distances. The light cone current algebra, which we had discussed in Ref. [75], would not be changed. The confinement of colored states, i.e. the quarks and the gluons, would be due to the interaction at long distances.

Soon we realized that our assumption, that there is no interaction near the light-cone, was not correct. David Gross, Frank Wilczek and, independently, David Politzer calculated this interaction, which is the interaction, given by the Lagrangian, but near the light-cone the relevant coupling constant is not zero, but only very small.

The QCD Lagrangian describes a theory, which is asymptotically free. At small distances the interaction is very small, at large distances the interaction is strong. Thus the coupling constant is not constant, but a function of the energy. The sliding of the coupling constant g_s as a function of the renormalization mass μ is given by the beta-function $\beta(g_s)$:

$$\mu\frac{\mathrm{d}}{\mathrm{d}\mu}g_s(\mu) = \beta(g_s). \tag{1.21}$$

This beta function is positive for many theories, for example quantum electrodynamics. The fine structure constant α is at the energy of 100 GeV about 10% larger than at low energies.

The beta function can be calculated in perturbation theory. One finds for QCD:

$$\mu\frac{\mathrm{d}}{\mathrm{d}\mu}g_s(\mu) \simeq -\frac{1}{16\pi^2}\left(11 - \frac{2}{3}n_f\right)g_s^3(\mu). \tag{1.22}$$

Here the coefficient "11" describes the contribution of the gluons to the beta function. The asymptotic freedom of QCD is due to this coefficient – it is related to the self-interactions of the gluons. The number n_f is the number of the different quark flavors. For the three quarks up, down and strange one has $n_f = 3$.

In QCD one can describe the energy dependence of the coupling constant by introducing a scale parameter Λ:

$$\alpha_s(\mu^2) \simeq \frac{4\pi}{\left(11 - \frac{2}{3}n_f\right)\ln\left(\frac{\mu^2}{\Lambda^2}\right)}. \tag{1.23}$$

This scale parameter has been measured by many experiments (see Sect. 3.2):

$$\Lambda = (332 \pm 17)\,\text{MeV}. \tag{1.24}$$

In experiments one has measured the scale dependence of the coupling constant. It agrees very well with the theoretical prediction. We also mention the value of the coupling constant at the mass of the Z-boson, where it was possible to measure the coupling constant rather precisely (see Sect. 3.2):

$$\alpha_s = 0.1181 \pm 0.0011. \tag{1.25}$$

In QCD, Bjorken scaling in deep inelastic scattering is not an exact property of the strong interactions. The quark distribution functions change slowly at high energies. This change can be calculated in perturbation theory (see Sect. 2.3). The results agree rather well with the experimental results. Also the gluon distribution function $g(x)$ has been measured. Since the gluons and the quarks contribute to the momentum of a high energy proton, the following sum rule must be obeyed:

$$\int_0^1 x\left[g(x) + u(x) + \overline{u}(x) + d(x) + \overline{d}(x) + s(x) + \overline{s}(x)\right]\mathrm{d}x = 1. \tag{1.26}$$

Using the scale parameter Λ, one can in principle calculate many properties of the strong interactions, for example the masses of the hadrons like the proton mass: $m_p = \text{const} \times \Lambda$. The proton mass depends also on the quark masses, however the up and down quark masses are very small and can be neglected. The calculations of the hadron masses are complicated and are often carried out by discretizing space and time (see Sect. 4 on Lattice QCD).

In QCD one can also change the three quark masses. For example we can assume that the three quark masses are zero. In this case the flavor group $SU(3)_F \times SU(3)_F$ would be unbroken. The three pions, the four K-mesons and the η – meson would be massless and the eight vector mesons would have the same mass. There is not a ninth massless pseudoscalar meson, since the singlet axial current has an anomaly:

$$\partial_\mu \left(\bar{u}\gamma^\mu\gamma_5 u + \bar{d}\gamma^\mu\gamma_5 d + \bar{s}\gamma^\mu\gamma_5 s\right) = const \times g_s^2\, \epsilon^{\mu\nu\rho\sigma} F^A_{\mu\nu} F^A_{\rho\sigma}. \quad (1.27)$$

where $\epsilon^{\mu\nu\rho\sigma}$ is the totally antisymmetric tensor. In Ref. [35] Gell-Mann and I also studied what happens if the quarks are removed from the QCD Lagrangian. In this case only the eight gluons are present. At low energies there would be a discrete spectrum of particles, which consist of gluons – the glue mesons, gluonium particles or glueball (see Sect. 8.4). If the three quarks are introduced, the glue mesons would mix with the quark–antiquark mesons. The experimentalists have thus far not clearly identified a glue meson. Presumably in nature there are only mixtures of glue mesons and quark–antiquark mesons. But there might be mesons, which are essentially glue mesons, since the mixing is very small for these mesons.

It is useful to consider the theory of QCD with just one heavy quark Q. The ground-state meson in this hypothetical case would be a quark–antiquark bound state (see Sects. 8.1, 8.6). The effective potential between the quark and its antiquark at small distances would be a Coulomb potential proportional to $1/r$, where r is the distance between the quark and the antiquark. However, at large distances the self-interaction of the gluons becomes important. The gluonic field lines at large distances do not spread out as in electrodynamics. Instead, they attract each other. Thus the quark and the antiquark are connected by a string of gluonic field lines. The force between the quark and the antiquark is constant, i.e. it does not decrease as in electrodynamics. The heavy quarks are confined.

In the annihilation of electrons and positrons at very high energies it has been possible to test the theory of quantum chromodynamics rather precisely. If an electron and a positron collide, a quark and an antiquark are produced. The two quarks move away from each other almost with the speed of light. Since the two quarks do not exist as free particles, they fragment into two jets of hadrons, mostly pions. These particles form two narrow jets. These jets have been observed since 1979 at the collider at DESY, later at the LEP-collider at CERN. Sometimes a quark emits a high energy gluon, which also fragments into hadrons. Thus three jets are produced, two quark jets and one gluon jet. Such three jet events have been observed since 1979 at DESY, later at CERN (see Sect. 2.2).

Now we consider high energy collisions of atomic nuclei, for example collisions of lead nuclei. Such collisions are studied at the Relativistic Heavy Ion Collider (RHIC) in Brookhaven, at Fermilab and at the LHC in CERN. In such collisions a new state of matter is produced for a short time, a quark–gluon-plasma. Astrophysicists assume that such a plasma exists also for a long time near the center of a large neutron star (see Sect. 7.1).

Right after the Big Bang the matter was a quark–gluon-plasma. During the expansion of the universe the plasma changed later into a gas of protons and neutrons (see Sect. 7.2).

In the fall of 1973 I was convinced, that Gell-Mann and I had discovered the correct theory of the strong interactions: Quantum Chromodynamics. Almost every day I discussed this theory with Richard Feynman, and he also thought that it was correct. In 1974 Feynman gave lectures on QCD. But Gell-Mann still thought that the true theory of the strong interactions should be a theory based on strings.

In the years after 1973 it became clear that QCD is the correct theory of the strong interactions. I was proud that I had contributed to the birth of this theory, which is now a major part of the Standard Theory of particle physics.

2 Experimental foundations

Conveners:
Eberhard Klempt and Franz Gross

Quantum Chromodynamics or QCD: What a gorgeous theory! You start with free colored quarks. You request invariance with respect to the exchange of colors at any time and any space point, and the quarks interact. That is all that QCD requires (see Sect. 1). QCD is based on a simple Lagrangian but embodies an extremely rich phenomenology which is still being explored. Nowadays, QCD is the accepted theory of the strong interaction and is used as a "working horse" to interpret experimental data. In the early days, however, the realms of perturbative and nonperturbative approaches were not understood, radiative corrections were not applied, and QCD was not uncontested: Still in 1979, five leading theoreticians at CERN, de Rújula, Ellis, Petronzio, Peparata, and Scott presented a "theatre" discussion in five acts at the Inter-

national School of Subnuclear Physics on "Point like Structures inside and outside hadrons" in Erice in which achievements and failures of perturbative QCD were discussed [77]. In 1992, however, in a workshop at Aachen [78], QCD was grown up "from a rather fragile construction of ideas into an actual microscopic quantum field theory of the strong interactions" [79].

In this section, milestones are discussed which convinced even sceptical physicists of the quark model and of the new theory. Important first steps to verify the quark model and QCD were already discussed by Leutwyler in first section (see Sect. 1.1).

A breakthrough was achieved in the *November revolution:* Charmonium was discovered at SLAC, the c-quark was shown to exist, the GIM mechanism (proposed by Glashow, Iliopoulos and Maiani in 1964 [80]) explaining the absence of neutral currents in weak interactions found experimental confirmation. John B. Kogut's contribution remembers the excitement in these days. A new spectroscopy came into life, many new resonances were discovered, some of them with completely unexpected properties that are still studied today, both experimentally (see Sect. 8.5) and theoretically (see Sect. 8.6).

One year later, the τ-lepton [81] was discovered (later also its neutrino [82]), the b-quark and the rich bottomonium spectrum [83]. Schaile and Zerwas [84] determined the weak isospin of the b-quark and established the b and t quarks as members of the third generation before the t-quark – completing the third family of fermions – was discovered [85]. The need for a third family had already been claimed by Kobayashi and Maskawa to explain CP violation in K decays [86].

San Lau Wu recalls her personal contributions to the discovery of gluons at DESY where events were found in which e^+e^- annihilate into three bunches of particles, three jets. The three jets were interpreted as processes in which the two quarks – observed as jets – radiate off a gluon which manifests itself as the third jet.

The evidence for the correctness of QCD grew rapidly. A huge activity was started at the SPS at CERN and elsewhere performing QCD analyses exploiting the Altarelli-Parisi equations [87], now called DGLAP equations. At that time, nobody in the western countries had realized the important contributions of Gribov and his school.

Yuri Dokshitzer – the "D" in DGLAP – reminds us of the most important steps. Scaling, observed already in 1972, proved the existence of interaction centers – called partons by Feynman – inside of nucleons. In the meantime, elastic and inelastic scattering off nucleons has grown to an industry supplying us with a detailed view of internal structure of nucleons (see Sect. 10). From the ratio of the cross sections for e^+e^- annihilation into hadrons over that for $\mu^+\mu^-$ the number of colors $N_c = 3$ was deduced. And the strong interaction constant α_s was shown to decrease with momentum transfer opening QCD to perturbative approaches (see Sect. 3.2). Dokshitzer introduces many basic concepts like jet finding algorithms, evolution, divergences and resummation, which will be discussed in more detail in Sect. 11.

2.1 Discovery of heavy mesons as bound states of heavy quarks

John B. Kogut

2.1.1 SLAC, light quarks and deep inelastic scattering

Many physicists and accelerators contributed to the establishment of the Standard model. But two accelerators were particularly important to US-based researchers. They were the 2-mile Linac and the 80 m diameter electron–positron ring, SPEAR (Stanford Positron–Electron Asymmetric Rings), of the Stanford Linear Accelerator Center (SLAC), Fig. 1. The Linac, which was built under the direction of SLAC's first director, W. Panofsky ("PIEP"), and started operations in 1965, discovered the light constituents of the protons, the u, d and s quarks, by measuring the inclusive deep inelastic cross section of $e^- + p \to e^{-'} + X$. The deep inelastic scattering program was critical to the founding of Quantum Chromodynamics (QCD) and is discussed extensively elsewhere in this journal review.

When I arrived at SLAC as an incoming graduate student in 1967, theoretical research revolved around Bjorken ("bj") scaling, and the parton model of bj and Feynman. One of the tools of the trade was the Infinite Momentum Frame (IMF). D. Soper, bj and I put the IMF on a firm foundation by quantizing Quantum Electrodynamics (QED) on the light cone [88]. This work initiated the program of light cone formulations of field theories (later called light front quantization by some advocates) that will be reviewed in Sect. 5.3. Later, S. Berman, bj and I developed the parton picture of the final states of inclusive processes involving large momentum transfers [89] and introduced parton fragmentation functions. This work had to address the mysterious phenomenon of quark confinement, the fact that quarks were "observed" when their properties were measured in deep inelastic processes, but no quarks were found isolated in the debris of the collisions. Although considerable progress has been made and many field theoretic mechanisms have been studied and proposed, especially in the context of Lattice formulations of QCD, the quark confinement problem remains open. It certainly was on many physicists' minds in the early days.

2.1.2 Charmonium and The November Revolution

Several years later, during the summer of 1974, experimentalists from SLAC presented some intriguing data from the

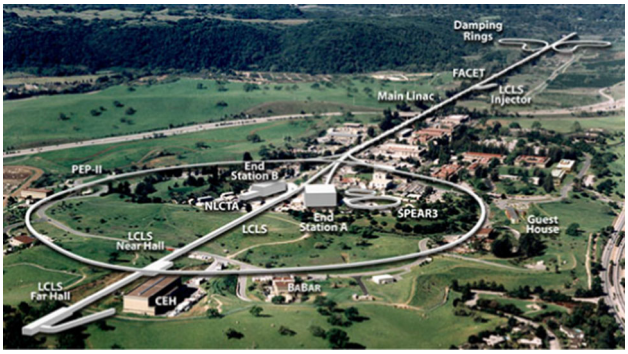


Fig. 1 Aerial view of SLAC, 2020: The Linac, SPEAR and their descendents

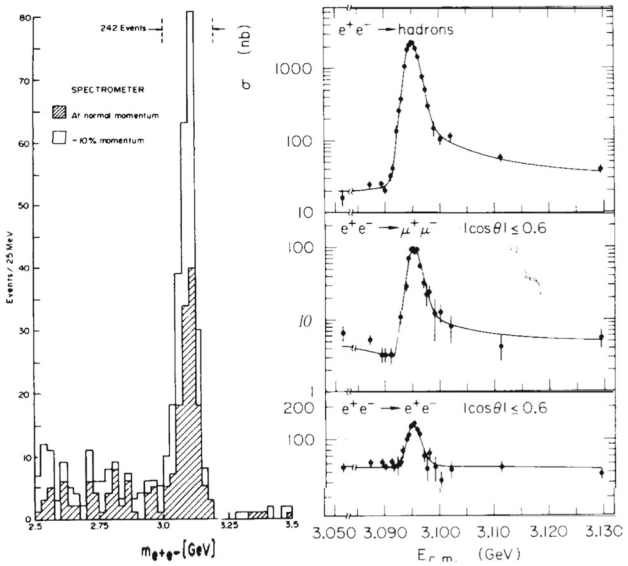


Fig. 2 The discovery of the J at BNL [94] and of the ψ at SLAC [95]

earliest runs of their very new electron–positron collider, SPEAR. A later section of this article will sketch the history of electron–positron colliders at SLAC since these machines were so central to the establishment of the Standard Model. The data of the summer of 1974 focused on the ratio R,

$$R = \frac{\sigma_{e^+e^-\to\text{hadrons}}}{\sigma_{e^+e^-\to\mu^+\mu^-}} \tag{2.1}$$

which, when plotted against the CM energy, showed a high, broad peak around 3.0–3.5 GeV. This suggested new interactions in the reaction's direct channel. One popular speculation was that a new quark threshold had been reached. A notable paper [90] stated that the reported broad peak in R should be accompanied by narrow resonant peaks at slightly lower energies. On November 11, 1974, SPEAR announced such a narrow peak at an energy 3.105 GeV [91] with an electronic width of $\Gamma_e \approx 5.5$ keV. Brookhaven also found this state in proton–proton collisions in fixed target experiments at the Alternating Gradient Synchrotron (AGS) [92] but that experiment didn't have the resolution of the relatively clean electron–positron collisions at SPEAR to measure its narrow width (see Fig. 2). With the news of a narrow state at 3.105 GeV, the high energy theory community exploded with speculations. The charmed quark hypothesis was just one of many competitors. Recall that the 1960–1970s was an era of discovery of many strong interaction states that were described by non-field theoretic approaches to high energy physics, such as Regge poles, bootstraps, etc. The field was stunned again two weeks later, on November 25, 1974, when SPEAR announced a second narrow peak at energy 3.695 GeV [93]! This challenged all the speculations circulating worldwide. The charm hypothesis was the most appealing to myself and collaborators since we were students of deep inelastic scattering and local field theory. The charm hypothesis was critical to the phenomenology of the electroweak sector of the Standard Model: the four quark model of u, d, s and c quarks solved the problem of neutral strangeness changing weak currents (the GIM mechanism [80]) of the three quark model. In addition, for the cancellations of the GIM mechanism to work effectively, the charm quark could not be too heavy. There were estimates that its mass $m_c \le 2.0-2.5$ GeV which put it inside the interesting range to explain the new resonances. In fact, the conventional quark model of mesons and baryons predicted that the charmed meson threshold of the SPEAR experiment, the minimum energy to produce two free charmed mesons, each consisting of a charmed quark and a light (u, d, or s) quark or anti-quark, should be $M_c = 2m_c + 0.7$ GeV. Since the second state at 3.695 GeV was very narrow, M_c had to be above 3.695 GeV. So, if m_c lay in the range 1.5–2.0 GeV, the charm hypothesis appeared to be compatible with all the known data. The only "fly in the ointment" was that SPEAR had not announced the discovery of charmed mesons above 3.695 GeV. Nervous charm enthusiasts worried that maybe the charm idea was flawed!

Following Ref. [90], the new states were tentatively called "charmonium", in analogy to positronium. Then the 3.105 state would be the 1^3S_1 state of a c and $\bar{c}$, and the 3.695 would be the 2^3S_1. S-waves were required so that the c and $\bar{c}$ would couple directly to the virtual photon created in the direct channel when the electron and positron annihilated. I recall that when these ideas were first discussed, many researchers sought to understand positronium better and ran off to their physics libraries and read Schwinger's classic works on the subject! Positronium spectroscopy had been calculated in great detail. This was possible because the static electron–positron interaction potential was just Coulomb's law. One needed the generalization of this interaction potential to strong interactions, QCD, to repeat those exercises for charmonium. At short distances it was plausible to assume a Coulomb-like formula with the fine structure

constant replaced by $\alpha_s = g^2/4\pi$, where g is the strong coupling constant of QCD. In fact, g should be the running coupling, a scale dependent quantity, and α_s should be small, say $\alpha_s \sim 0.2$ for mass scales of $\sim 2\,\text{GeV}$ to accommodate the success of the parton model in deep inelastic scattering where experiments suggested that the parton distribution functions satisfy Bjorken scaling to good approximation for $Q^2 \approx 2{-}3\,\text{GeV}^2$. Next, one needed the potential at intermediate distances, where the $c\bar{c}$ pair feels the QCD forces of confinement but the system is below the charm threshold so that screening by light quarks is not yet active. Studies of model field theories of confinement [96] and the lattice version of QCD [97,98] led to the idea that chromo-electric flux tubes form in this kinematic region and lead to a linear confining potential between heavy colored quarks. These ideas lead to the static $c\bar{c}$ potential [99],

$$V(r) = -\frac{\alpha_s}{r}\left\{1 - \frac{r^2}{a^2}\right\} \tag{2.2}$$

where a sets the scale of the linear potential. The need for the linear term in Eq. (2.2) was actually compelling in the original data. The ratio of the squares of the wave functions of the two charmonium states at the origin was called

$$\eta = \left|\frac{\psi(1^3S_1; r=0)}{\psi(2^3S_1; r=0)}\right|^2 = \frac{3.105}{3.695}\frac{\Gamma_e(3105)}{\Gamma_e(3695)} \approx 1.4{-}1.7 \tag{2.3}$$

where we related the wave functions at the origin to the electronic width of each state and used early data to evaluate η. What do the values 1.4–1.7 imply about the potential? One can check that for a harmonic potential $\eta = 2/3$, for a linear potential $\eta = 1$ and for a Coulomb potential $\eta = 8$. So, to accommodate Eq. (2.3), a combination of a linear confining potential and Coulomb potential was preferred. In Ref. [99] the parameters in the potential (α_s, a) were determined from the experimental data of the day by solving the radial Schrodinger equation and imposing the constraints: 1. The mass difference between the two charmonium states is 0.59 GeV, 2. $\Gamma_e(3105) = 5.5\,\text{keV}$, 3. m_c should lie between 1.5 and 2.0 GeV, and 4. α_s should be between 0.2 and 0.3. At this point the authors of Ref. [99] needed a convenient computer program to solve the radial Schrodinger equation with a potential of the form Eq. (2.2). Luckily, we had access to a skilled computational physicist with a trove of software programs! That computational physicist was K. G. Wilson who used numerical methods to teach undergraduate quantum mechanics. Remember that this was 1974 when universities had computer centers with IBM mainframes driven by punch cards! A good fit was found with his program for $m_c = 1.6\,\text{GeV}$, $\alpha_s = 0.2$, and $a = 2\,\text{fm}$. It was important to check that these parameters led to a non-relativistic description of the charmonium bound states. In fact, the average velocity-squared of the charmed quarks in the bound states was computed to be $(v/c)^2 \leq 1/25$. The bound states of the $c\bar{c}$ system that resulted are shown in Fig. 3.

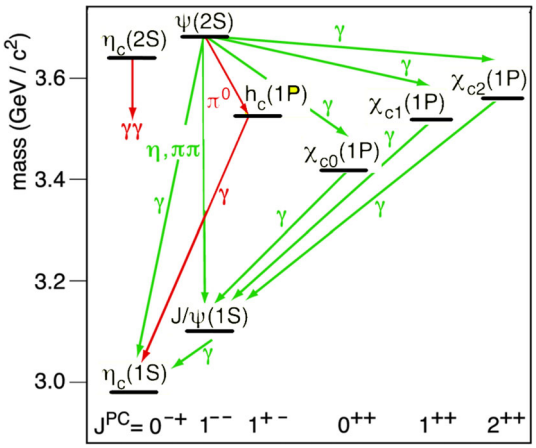


Fig. 3 Charmonium spectroscopy. Note the P-waves ${}^3P_J$ and the radiative transitions

The most relevant result in Fig. 3 was the existence of the P wave states that lie between the 3.105 and 3.695 GeV states. For a pure Coulomb potential the P wave states would be degenerate with the 3.695 state. However, for a linear potential, the 2^3S_1 state resides at higher energy than the P wave states, as shown in the figure, because the 2^3S_1 has a radial node. The existence of these states led to the main point of Ref. [99]: there are additional states which could be found experimentally at SPEAR and they constitute strong, new evidence for the charm hypothesis! Strong E1, electric dipole, transitions would produce monochromatic photons when the 3.695 state decays to one of the P waves and then additional monochromatic photons should appear when each P wave decays to the 3.105 state! These monochromatic photons should be "easy" to find at SPEAR because it had a 4π general purpose detector, the Mark I. The energies of the P waves and the strengths of the E1 transitions followed from the wave functions found from the radial Schrodinger equation. These results were catalogued in Ref. [99] and were refined in later more ambitious publications. Of course, the wave functions and the radiative transition rates depend much more sensitively on the parameters in the potential than the energies of the P waves themselves. In any case, the predictions of Ref. [99] were reasonable guides for the experimental program which discovered the states and the radiative transitions in 1976, the same year that the charmed D mesons were also identified in the final states of the electron–positron collisions! Many more predictions and calculations were presented in Ref. [99]] and in similar works done by other groups [100]. Some of these points will be discussed in later chapters in this journal review. In addition, more sophisticated potentials than Eq. (2.2) were eventually stud-

ied. Tensor interactions, fine and hyperfine interactions were added in, and their effects are shown in some of the splittings in Fig. 3 (Refs. [99,100]). And the influence of the nearby threshold at M_c on the bound states was also accounted for. All of these developments did not change the main thrust of Ref. [99]: the existence of the P wave states and their radiative transitions were special to the charm quark interpretation of the SPEAR experiment and gave additional motivation to the early acceptance of the Standard Model.

2.1.3 Electron–positron colliders at Stanford

Now let's change the viewpoint of this article and turn to the accelerator physicists and the experimentalists at SPEAR. There is a cliché that behind every invention there is a visionary. In the case of electron–positron colliders, one of the field's several visionaries was definitely Gerry O'Neil. Other visionaries were Burt Richter and Martin Perl. Professor O'Neil taught me physics in college, but he was more interested in building accelerators to collide electrons and positrons head-on in their center of mass frame to create pure electromagnetic energy and search for new states of matter. I recall that he traveled to Novosibirsk, where a collider was being constructed, several times during a one semester undergraduate course on modern physics. Upon each return he "debriefed" his class on the progress of his efforts. In 1965 Gerry O'Neil and others from Princeton and Stanford built two 300 MeV electron storage rings in the High Energy Physics Laboratory (HEPL) at Stanford. These rings resulted in electron–electron collisions which successfully increased the limits of validity of Quantum Electrodynamics. However, it was basically a "single experiment" machine, so during construction Gerry and his collaborators also sketched an outline of a 3 GeV electron–positron colliding beam facility. These ideas evolved into the blueprints for the famous SPEAR collider at SLAC. To many persons' surprise, just as electron–positron collider ideas were gaining traction, Gerry's visionary ideas moved in a different direction: to outer space projects, such as a permanent space station in an earth orbit. He left the fledgling field of colliders just as it was about to yield great discoveries!

The construction of SPEAR began in 1970 under the direction of Burt Richter and John Rees, and it was completed quickly (in 20 months, four months ahead of schedule) in 1972 and at modest cost. The final SPEAR design was the result of several revisions, forced on the group by budget restrictions and engineering considerations. During one of the revisions, the two planned rings for the electrons and positrons became one and SPEAR was no longer asymmetric. Nonetheless, the inventors kept the appealing name "SPEAR"!

Wolfgang Panofsky was still the Director of SLAC and had lobbied the US Congress and the funding agency, the Atomic Energy Commission (AEC), the predecessor of the Department of Energy, to fund the construction of SPEAR as a federal project. However, there were many projects competing with SPEAR at the time, and it did not achieve federal project status. However, Panofsky and Richter did not want to delay its construction, so the AEC allowed SPEAR to be built using ordinary laboratory operating funds! This meant that it had to be done cheaply. Some have estimated the cost between 2 to 5 million dollars. So, the usual idea of having the accelerator constructed underground within an enclosed building had to be abandoned. SPEAR was built outside on a parking lot (Fig. 4), with concrete blocks providing the shielding!

Fig. 4 The 80 m SPEAR Ring in a parking lot at SLAC. The photo also shows the separate e^+ and e^- beam lines and the detector hall

Of course, the accelerator needed a detector or two at its beam intersection regions. Richter and others formed a Berkeley/Stanford team to design and build a multipurpose detection system surrounding one of the SPEAR interaction regions (Fig. 5). The result was the Magnetic Detector or Mark I. This was the first 4π general purpose detector. It proved crucial in the coming discovery process. Other detector designs with limited angular apertures would have suffered from the relatively low statistics of the early machines and wouldn't have operated as well with diverse final states consisting of photons, leptons and various light mesons.

It was clear at the time that electron–positron colliders had many attractive properties: 1. All the energy of the beams goes into creating new particles, unlike fixed target machines, 2. The beams consist of pointlike particles, so the interactions are simple and clean theoretically. However, they suffer from one limitation: radiation losses. However, it turned out that "one man's problem is another man's opportunity". From the beginning, several Stanford faculty members realized SPEAR's potential to produce useful synchrotron radiation, so they asked Panofsky and Richter to devise a way to form an

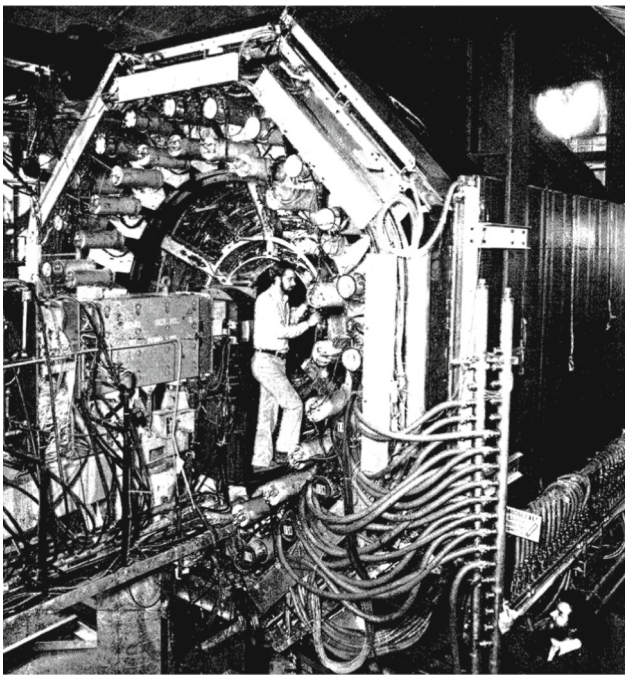


Fig. 5 Photograph of the wide angle SLAC-LBL Mark I detector in 1974

X-ray beam out of SPEAR. The X-ray synchrotron radiation emitted by the circulating beams in the machine was much higher in intensity, by a factor of 10 to 100, than any other facility in the world. It could be used for imaging and structural analysis in many areas of research, from semiconductor materials to protein molecules. So, Richter's team attached an extra vacuum chamber to SPEAR and made provision for a hole in the shielding wall for the beamline. This was the start of The Stanford Synchrotron Radiation Project (SSRP). Even though it began as a parasitic operation, synchrotron radiation represented an unparalleled opportunity!

Richter also saw the SLAC Linac as a light source. These ideas led to the invention and development of a undulator so that the Linac's electron beam could become the source for the most intense Free electron Laser (FEL) on the planet. The LCLS (Linear Collider Light Source) was born in 2009. It has led to revolutions in our understanding of the temporal dynamics of atoms, molecules and condensed matter systems. This is another story which we can't cover here, but it is amusing to understand that a "problem" with circular colliders grew into a new generation of accelerator facilities!

2.1.4 The revolution begins

In the spring of 1973, SPEAR began to gather high-energy physics data. By the next year, the machine was measuring very erratic but generally much larger than expected values of R, Eq. (2.1), the ratio of hadron production to lepton production. These early measurements were done with wide energy resolution, several hundred MeV, to produce and measure many interactions and final state particles. But there were "inconsistencies" in the data: small changes in the beam energies sometimes led to large changes in the observed value of R. These were the first signs of a new particle, which Richter's team called the "ψ". "Nobody dreamed that there was any state, particle, that was as narrow in width as the W turned out to be," said Richter in 2003. "So the first question was what the hell was wrong with the apparatus, is there something wrong with the computers, is there something wrong with the data taking?" [101]. No-one could find any such errors, and some researchers on the Mark I collaboration pushed to rescan the region. In fact, by this time, SPEAR had been upgraded and Robert Hofstadter, who was running an experiment at SPEAR's other detector, wanted to move on to higher energies. Finally, Richter decided to go ahead with rechecking the anomalous results, but only for one weekend in November 1974.

Fig. 6 The SPEAR Control Room during the Big Night. SLAC and LBL physicists analyzing the raw data

2.1.5 Minute-by-minute developments in the SPEAR control room

"During the night of 9–10 November, the hunt began, changing the beam energies in 0.5 MeV steps. By 11.00 a.m. Sunday morning the new particle had been unequivocally found. A set of cross section measurements around 3.1 GeV showed that the probability of interaction jumped by a factor of ten from 20 to 200 nanobarns. In a state of euphoria, the champagne was cracked open and the team began celebrating an important discovery. While Gerson Goldhaber retired to write up the findings 'on-line' for immediate publication, Fig. 6, it was decided to polish up the data by going slowly over the resonance again. The beams were nudged from 1.55 to 1.57 MeV and everything went crazy. The interaction probability soared higher; from around 20 nanobarns

the cross section jumped to 2000 nanobarns and the detector was flooded with events producing hadrons. Pief Panofsky, the Director of SLAC, paced around the control room invoking the Deity in utter amazement at what was being seen. This heavy particle, displaying such extraordinary stability, they called 'ψ' and they announced it in a paper beginning with the words 'We have observed a very sharp peak'. Within hours of the SPEAR measurements, the telephone wires across the Atlantic were humming as information, enquiries and rumors were exchanged" [102].

Just two weeks later, the scene repeated itself, except at a higher energy, 3.695 GeV. And the next S-wave charmonium state was found. Physicists around the country had "befriended" various members of the SLAC/LBL group in the control room by now and news of the new resonance spread across the country within minutes. I heard about it in an early morning phone call with bj. I also learned that it was he who had suggested the high resolution scans in energy that led to both discoveries!

SPEAR meanwhile continued to yield breakthroughs. In 1976 the P-waves were discovered through their radiative transitions [103] and charmed mesons [104] were found above threshold as well.

Those were the days!

2.1.6 The path forward. Hamiltonian lattice gauge theory and statistical field theory

After Ref. [99] was published, it was time to move on to more fundamental considerations. I believed that the most important implications of the potential model had been made and working through additional details was less important. Instead, there were major challenges in developing an approach to QCD that would lead to systematic, potentially exact, predictions of the theory. This was the thrust of Wilson's lattice formulation of QCD [97], which will be discussed at length in Sect. 4 below. A Hamiltonian version of the theory [98] was also developed because it emphasized 1. The spectroscopy of the theory, and 2. The quantum character of the states. An added bonus of this development was a new formulation for strongly coupled systems for applications to condensed matter physics [105]. This development mirrors the past of SPEAR: SPEAR started out by establishing the Standard model of high energy physics, and now is pushing the frontiers of imaging, free electron lasers and quantum systems. In parallel, the lattice Hamiltonian form of strongly coupled gauge theories is playing a role in the development of Quantum Information Systems that may lead to new quantum computers and quantum detectors. These subjects are now the central themes in a new generation of studies and workshops on quantum physics [106,107]. References [98] and [105], which were originally conceived for QCD, are proving useful here, and are, in fact, among the most cited publications in the 48 year history of lattice gauge theory. Perhaps, these contributions will inspire the next generation of theorists who will push the frontiers of strongly coupled gauge theories into the next era.

2.2 Experimental discovery of gluons

Sau Lan Wu

2.2.1 Yang–Mills non-Abelian gauge particles

It was in 1954 when Chen Ning Yang and Robert Mills, who was a graduate student, shared the same office at the Brookhaven National Laboratory and developed their non-Abelian gauge theory. Their office was shared with another famous physicist Burton Richter, who was also a graduate student at that time. Almost exactly 25 years later, the first Yang–Mills non-Abelian gauge particle was observed at the German National Laboratory called Deutsches Elektronen-Synchrotron (DESY). Here are some of the interesting dates. The idea of Yang and Mills was first presented at the April 1954 meeting in Washington, DC of the American Physical Society and the full Yang–Mills paper was submitted for publication on June 28, 1954 [41]. The first public announcement for the experimental discovery of the first Yang–Mills gauge particle was made at the Neutrino 79 conference on June 18–22, 1979 [108], and the first full paper was received for publication on August 29, 1979 [109].

The word "gluon" was originally introduced by Murray Gell-Mann to designate a hypothetical neutral vector field [14] coupled strongly to the baryon current, without reference to color. Since then, the meaning of this word has changed: nowadays, this word "gluon" is used exclusively to mean the Yang–Mills non-Abelian gauge particle for strong interactions.

2.2.2 Harvard to M.I.T. to Wisconsin

After being awarded my Ph.D. degree at Harvard University, Samuel S. S. Ting of M.I.T. kindly offered me a postdoctoral position in his group. A few years later, I felt that, for the development of my career in physics, it was time for me to get a faculty position. Sam then helped me to look for a faculty position at the University of Michigan, where he received his own doctoral degree. I got into contact with Michael Longo, a professor of physics there, and he was very supportive. Therefore I applied to the University of Michigan. Since thanks to Longo I got on the so-called short list of candidates, I was invited to go to Ann Arbor for an interview.

In the meantime, I contacted David Cline, a professor of the University of Wisconsin I had met before. David told me that he would forward my name to Ugo Camerini, a colleague of his at Wisconsin. I contacted Ugo. Shortly before

my scheduled interview at Ann Arbor, I got a telegram from the University of Michigan saying that the position had been given to somebody else. I hesitated about going to that interview, but my friends told me that I should nevertheless keep the appointment. In the meantime, I got an invitation from the University of Wisconsin for an interview. Thus I traveled from Europe for an interview at Michigan first, and then continued to Wisconsin for another one.

I remember very well that, when I had the interview at the University of Wisconsin in Madison, Don Reeder took me out to dinner at an Italian restaurant close to the University Square and we had a very nice discussion. Don was at that time not only a Professor of Physics but also the Principal Investigator for the funding of experimental high-energy physics. Afterwards, I met with a number of faculty members in high-energy physics, and they were all very supportive. Again through the effort of Cline, I also got an offer from Fermilab. I had to make a decision, and I finally chose the University of Wisconsin. It was one of the best decisions I have made.

2.2.3 DESY

After becoming an assistant professor at the University of Wisconsin-Madison in 1977, I had to make the decision of what important problem in physics to tackle. Once again, I got wise advice from David Cline, who had helped me so much. He told me: "Sau Lan, you do not need to work with anybody, and you have no boss. You are your own boss, and you decide what to work on." At that time, the Department of Energy gave one lump sum of money to the University of Wisconsin for the faculty members in experimental high-energy physics to share. From this funding, Don Reeder gave me the positions of three post-docs and one graduate student.

I spent the first months of my assistant professorship thinking about what physics to work on.

At that time, we knew of four quarks: the up quark, the down quark, the strange quark, and the newly discovered charm quark from the J/ψ, which has led to the Nobel Prize for Sam Ting and Burt Richter. The immediate and important question is: how do these quarks interact with each? For this, we knew very little at that time besides that this interaction is likely to be mediated by a Yang–Mills non-Abelian gauge particle – the gluon. In other words, while the electromagnetic interaction is transmitted by the photon, which is an Abelian gauge particle, this additional interaction is transmitted by a Yang–Mills non-Abelian gauge particle.

Indirect indication of gluons had been first given by deep inelastic electron scattering and neutrino scattering. The results of the SLAC-MIT deep inelastic scattering experiment [110–113] on the Callan–Gross sum rule were inconsistent with parton models that involved only quarks. The neutrino data from Gargamelle [114] showed that 50% of the nucleon momentum is carried by isoscalar partons or gluons. Further indirect evidence for gluons was provided by the observation of scale breaking in deep inelastic scattering [115–117]. The very extensive neutrino scattering data from BEBC and CDHS Collaborations [118–120] at CERN made it feasible to determine the distribution functions of the quark and gluon by comparison with what was expected from QCD, and it was found that the gluon distribution function is sizeable. This information about the gluon is interesting but indirect. The discovery of the gluon requires direct observation.

During my first year as an assistant professor at the University of Wisconsin, I was fascinated by the Yang–Mills non-Abelian gauge theory. This was to be contrasted with the experimental situation at that time: while photons were everywhere in the detectors, no Yang–Mills gauge particle had been observed in any experiment.

From these considerations, I formulated the following problem for myself: how could I discover experimentally the first Yang–Mills gauge particle?

From previous experience with electron accelerators and proton accelerators at DESY and BNL, it was soon clear to me that the experimental discovery of the first Yang–Mills gauge particle was more likely at an electron machine rather than a proton machine. At that time, two electron–positron colliding beam accelerators were being built: PEP at SLAC and PETRA at DESY; after visiting both SLAC and DESY, I decided that PETRA was a better choice for me.

At PETRA (Positron–Electron Tandem Ring Accelerator), there were five experiments: CELLO, JADE, MARK J, PLUTO, TASSO. I approached first the PLUTO Collaboration and then the JADE Collaboration, but nothing worked out. Then my luck changed completely: I ran into Björn Wiik, one of the two co-spokesman of the TASSO Collaboration, the other one being Günter Wolf. Björn asked me what I was doing; when I told him my situation, he was surprised and said to me: "Come to see me in my office this afternoon." When I went to his office, he asked me: "Why don't you join the TASSO Collaboration instead?" I said that I would love to do that. Björn said that he would talk to Günter and also to Paul Söding, a senior physicist in TASSO, and let me know. Thanks to Björn, this was how I became a member of the TASSO Collaboration at DESY. All three of them, Björn, Günter, and Paul, are excellent physicists.

After becoming a member of the TASSO Collaboration, the physics problem that I formulated for myself took on a concrete form: how could I discover experimentally the first Yang–Mills gauge particle with the TASSO detector?

A feature of the TASSO detector is the two-arm spectrometer, which leads to the name TASSO – Two – Arm Spectrometer SOlenoid. The end view of this detector, i.e., the view along the beam pipe of the completed detector, is shown in Fig. 7. When TASSO was first moved into the

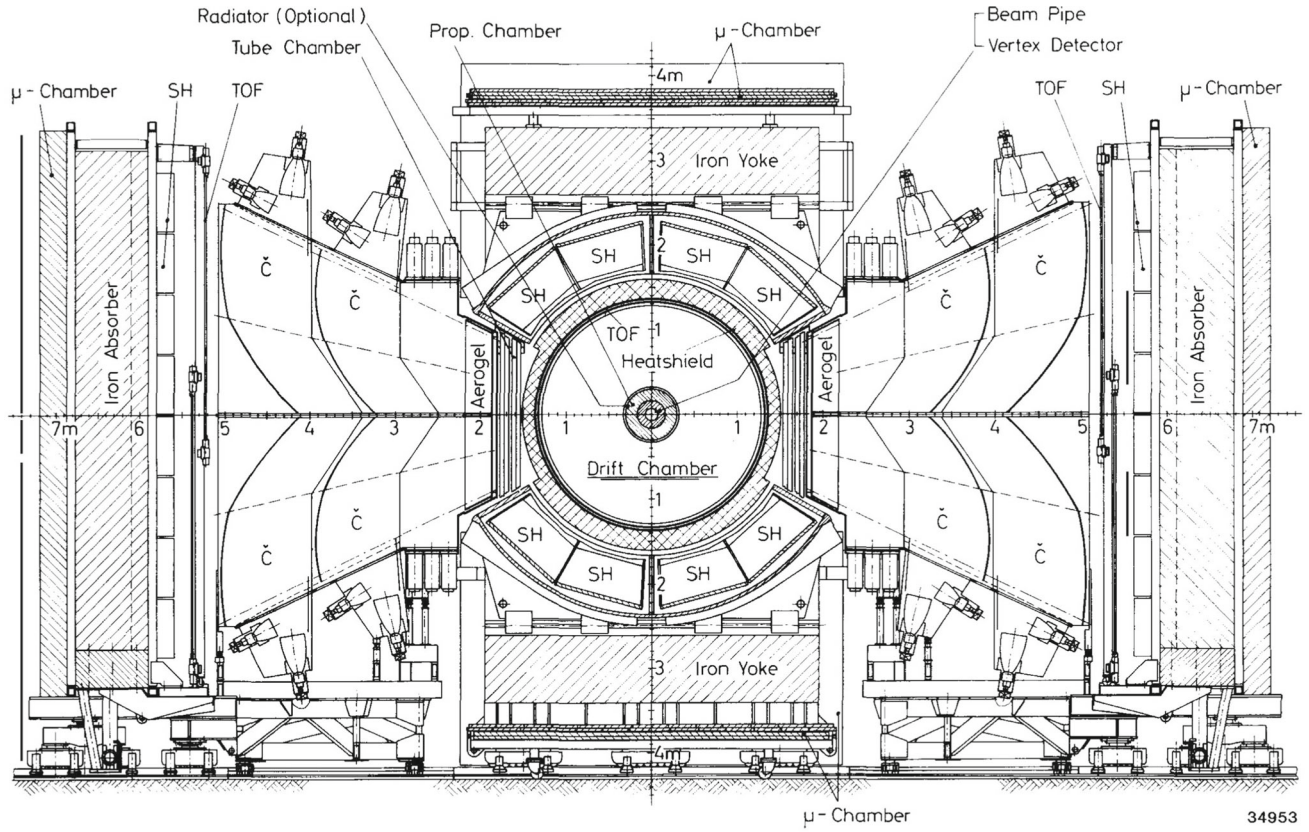


Fig. 7 End view of the TASSO detector

PETRA beams in 1978, not all of the detector components shown in Fig. 7 were in working order. For my purpose of the experimental discovery of the first Yang–Mills non-Abelian gauge particle, the most important component of the TASSO detector was the drift chamber, which was already functioning properly.

2.2.4 Three-jet events

One of the simplest ways to produce a photon – the Abelian gauge particle for electromagnetic interactions – is through electron bremsstrahlung process, i.e.,

$$e\, e \rightarrow e\, e\, \gamma.$$

Ellis, Gaillard and Ross had suggested that hard gluons should be emitted by quarks via bremsstrahlung in analogy with the radiation of electromagnetic bremsstrah-lung [121]

$$e^+e^- \rightarrow q\, \bar{q}\, g$$

where q is the quark of Gell-Mann [17] and Zweig [18], and g is the gluon – the Yang–Mills non-Abelian gauge particle for strong interactions.

This seemed to be the way to discover the gluon experimentally, but I faced the following two major problems.

(1) How can these production processes $e^+e^- \rightarrow q\bar{q}g$ be found in the TASSO detector?
(2) How high does the center-of-mass e^+e^- energy have to be for this process to be seen clearly?

A couple of years before I became a faculty member at the University of Wisconsin, the production process

$$e^+e^- \rightarrow q\bar{q}$$

was observed at the SPEAR e^+e^- collider at SLAC [122]. In the MARK I detector at SPEAR, both the quark q and the anti-quark $\bar{q}$ were observed as jets, i.e., groups of particles moving in nearly the same direction. With this experimental information from MARK I, I had to make my best guess as to how the gluon bremsstrahlung process $e^+e^- \rightarrow q\bar{q}g$ would look like in the TASSO detector. Since the gluon is the Yang–Mills non-Abelian gauge particle for strong interactions, it is itself a source for gluon fields. It therefore seemed reasonable to believe that the gluon in the gluon bremsstrahlung process would be seen in the detector also as a jet, just like the quark and the antiquark.

Therefore the gluon bremsstrahlung process $e^+e^- \rightarrow q\bar{q}g$ leads to three-jet events.

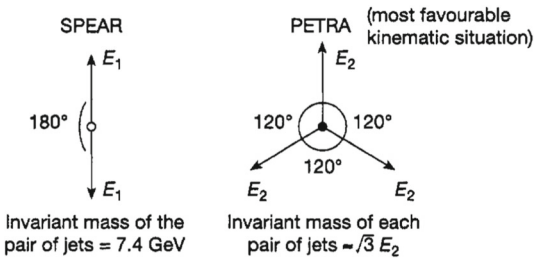


Fig. 8 Two-jet and three-jet configurations at SPEAR and PETRA respectively

Using the SPEAR information on the quark jets from the process, $e^+e^- \to q\bar{q}$, I convinced myself that three-jet events, if they were produced, could be detected once the PETRA energy went above three times the SPEAR energy i.e., $3 \times 7.4 \sim 22\,\text{GeV}$. The arguments were as follows:

Figure 8 shows a comparison of the two-jet configuration at SPEAR with the most favorable kinematic situation of the three-jet configuration at PETRA. If the two invariant masses are taken to be the same, i.e., $\sqrt{3}E_2 \approx 7.4\,\text{GeV}$, then the total energy of the three jets is $3E_2 \approx 13\,\text{GeV}$, which must be further increased because each jet has to be narrower than the SPEAR jets. This additional factor is estimated to be $180°/120° = 1.5$, leading to about 20 GeV. Phase space considerations further increase this energy to about 22 GeV.

This answers the question (2) above.

This estimate of 22 GeV was very encouraging because PETRA was expected to exceed it soon; indeed, it provided the main impetus for me to continue the project to discover the first Yang–Mills non-Abelian gauge particle.

At the same time, I had to address the question (1) above: how could I find three-jet events at PETRA? I made a number of false starts until I realized the power of the following simple observation. By energy–momentum conservation, the two jets in $e^+e^- \to q\bar{q}$ must be back-to-back. Similarly, the three jets in $e^+e^- \to q\bar{q}g$ must be coplanar. Therefore, the search for the three jets can be carried out in the two-dimensional event plane, the plane formed by the momenta of q, $\bar{q}$ and g. A few pages of my notes written in June 1978 and further historical details can be found in Ref. [123].

The procedure of mine did not identify which jet would be the gluon. Still, this procedure has a number of desirable features.

- First, all three jet axes are determined, and they are in the same plane. This is the feature that played a central role in the later determination of the spin of the gluon.
- Secondly, particle identification is not needed.
- Thirdly, the computer time is moderate for the "slow" computers at that time even when all the measured momenta are used.
- Finally, it is not necessary to have the momenta of all the produced particles; it is only necessary to have at least one momentum from each of the three jets. Thus, for example, my procedure works well even when no neutral particles are included.

This last advantage is important, and it is the reason why this procedure is a good match to the TASSO detector at the time of the PETRA turn-on.

I had Georg Zobernig as my post-doc; he was and is excellent in working with computers. My procedure of identifying the three-jet events in order to discover the gluon, programmed by Zobernig on an IBM 370/168 computer, was ready before the turn-on of PETRA in September of 1978. For that time in 1978, the programming was highly non-trivial. In his later publications, he has used the name Haimo Zobernig.

2.2.5 *Discovery of the gluon*

When we had obtained data for center-of-mass energies of 13 GeV and 17 GeV, Zobernig and I looked for three-jet events. It was not until just before the Neutrino 79 (International Conference on Neutrino, Weak Interactions and Cosmology at Bergen, Norway) in the late spring of 1979 that we started to obtain data at the higher center-of-mass energy of 27.4 GeV. We found one clear three-jet event from a total of 40 hadronic events at this center-of-mass energy. This first three-jet event of PETRA, as seen in the event plane, is shown in Fig. 9. When this event was found, Wiik had already left Hamburg to go to the Bergen Conference. Therefore, during the weekend before the conference, I took the display produced by my procedure for this event to Norway to meet Wiik at his house near Bergen. During this weekend, I also telephoned Günter Wolf at his home in Hamburg and told him of the finding. Wiik showed the event in his plenary talk "First Results from PETRA", acknowledging that it was my work with Zobernig by putting our names on his transparency of the three-jet event, and referred to me for questions. Donald Perkins of Oxford University took this offer and challenged me by wanting to see all forty TASSO events. I showed him all forty events, and, after we had spent some time together studying the events, he was convinced.

With these three-jet events, the question is: what are the three jets? Since quarks are fermions, and two fermions (electron and positron) cannot become three fermions, it immediately follows that these three jets cannot all be quarks and antiquarks. In other words, *a new particle has been discovered*.

The earliest papers related to the PETRA three-jet events are Refs. [108,109,124,125] all by members of the TASSO Collaboration, and *TASSO Note 84*, June 26, 1979 (by Sau Lan Wu and Haimo Zobernig). Reference [124] provides the

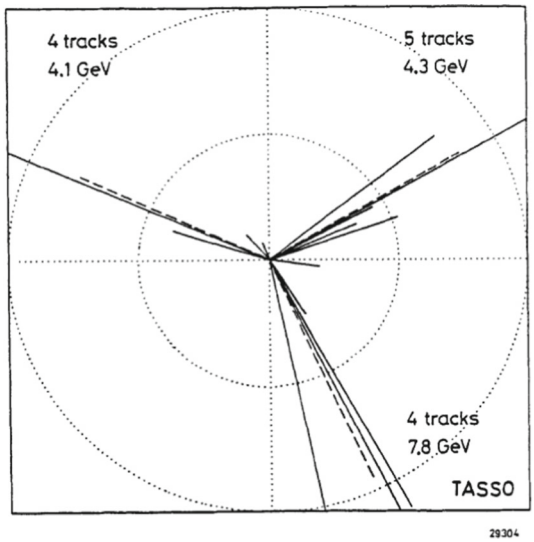


Fig. 9 The first three-jet event from electron–positron annihilation, as viewed in the event plane. It has three well separated jets [108]

method of analysis used in the four later papers, which all give experimental results.

Very shortly afterwards, the other experiments at PETRA – JADE, MARK J, and PLUTO Collaborations – published their own three-jet analyses. Their early papers related to the PETRA three-jet events are Refs. [126–128], and their results all confirm the earlier ones of TASSO. Since this discovery of the gluon was the highlight of the 1979 Lepton-Photon Conference at Fermi National Accelerator Laboratory (FNAL), Leon Lederman, Director of FNAL, called a press conference on the discovery of the gluon. Recent reviews of the discovery of gluons including further studies of jets can be found in Refs. [129,130].

Because of the discovery of the gluon by the TASSO Collaboration, Söding, Wiik, Wolf, and I were awarded the 1995 European Physical Society High Energy and Particle Physics Prize. With my leading role in this discovery, I was chosen to give the acceptance speech at the EPS award ceremony.

This was how the first Yang–Mills non-Abelian gauge particle was discovered experimentally at DESY, Hamburg, Germany in the spring of 1979, a quarter of a century after the original paper of Chen Ning Yang and Robert Mills. Four years later, the second and third Yang–Mills non-Abelian gauge particles – the W and Z – were discovered at CERN by the UA1 and UA2 Collaborations [131–134].

The experimental discovery of these Yang–Mills non-Abelian gauge particles points to another prophetic feature of the original paper of Yang and Mills [41]: the mass of the first Yang–Mills gauge particle has been found to be nearly zero, while those of the second and third Yang–Mills gauge particle are quite high – about 80 GeV for the W and 91 GeV for the Z. The relevant sentence in the original paper [41] is the following: "We have therefore not been able to conclude anything about the mass of the **b** quantum." For further comments on this point, see pp. 19–21 of [135].

2.2.6 *Some later developments*

The discovery of the gluon in 1979 was not only the discovery of a new elementary particle, but also the first elementary boson that has been seen experimentally as a jet. Indeed, it is so far the ONLY elementary boson seen this way. In principle, a scalar quark would share this property, but no scalar quark has ever been observed in any experiment.

The discovery of such a new type of elementary particle is guaranteed to lead to subsequent new understanding of fundamental physics, both experimental and theoretical. Here I will discuss one of the of the most important experimental consequences of this 1979 discovery of the gluon; the role it plays in the 2012 discovery of the Higgs particle.

An important theoretical topic, the very recent understanding of the quark–gluon coupling constant g_s, is discussed in considerable detail in Sect. 3, and briefly in my Summary and Outlook, Sect. 2.2.8.

2.2.7 *Role of gluon in the discovery of the Higgs particle [43–45]*

Since the gluon is the Yang–Mills gauge particle for strong interactions, to a good approximation a proton consists of a number of gluons in addition to two u quarks, one d quark, and some sea-quarks. Since the coupling of the Higgs particle to any elementary particle is proportional to its mass, there is little coupling between the Higgs particle and these constituents of the proton. Instead, some heavy particle needs to be produced in a proton–proton collision, for example at LHC, and is then used to couple to the Higgs particle. Among all the known elementary particles, the top quark t, with a mass of 173 GeV/c^2, is the heaviest [136,137].

The top quark, which may be virtual, is produced predominantly together with an anti-top quark or an anti-bottom quark [138]. Since the top quark has a charge of +2/3 and is a color triplet, such pairs can be produced by

(a) a photon: $\gamma \rightarrow t\bar{t}$;
(b) a Z: $Z \rightarrow t\bar{t}$;
(c) a W: $W^+ \rightarrow t\bar{b}$; or
(d) a g: $g \rightarrow t\bar{t}$.

As discussed in the preceding paragraph, there is no photon, or Z, or W as a constituent of the proton. Since, on the other hand, there are gluons in the proton, (d) is by far the most important production process for the top quark.

Because of color conservation – the gluon has color but not the Higgs particle – the top and anti-top pair produced by a

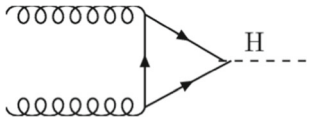


Fig. 10 Feynman diagram for the Higgs (H) production by gluon–gluon fusion (also called gluon fusion)

gluon cannot annihilate into a Higgs particle. In order for this annihilation into a Higgs particle to occur, it is necessary for the top or the anti-top quark to interact with a second gluon to change its color content. It is therefore necessary to involve two gluons, one each from the protons of the two opposing beams of LHC, and we are led to the diagram of Fig. 10 for Higgs production. This production process is called "gluon–gluon fusion" (also called "gluon fusion"). As expected from the large mass of the top quark, this gluon–gluon fusion is by far the most important Higgs production process, and shows the central role played by the gluon in the discovery of the Higgs particle in 2012.

The percentage of this gluon–gluon fusion contribution to the Higgs production cross section depends on the mass of the Higgs particle. For the actual mass of the Higgs particle, the gluon contributes, through this gluon–gluon fusion process, about 90% of Higgs production at the Large Hadron Collider. A more dramatic, but perhaps unfair, way of saying the same is that, if there were no gluon, the Higgs particle could not have been discovered for years!

2.2.8 Summary and outlook

One of the most influential papers in theoretical physics during the second half of the twentieth century – very likely the most important and influential one – is that of Yang and Mills published in 1954 [41]. The importance of this paper on the non-Abelian quantum gauge theory is due to that (1) it presents a completely new idea, and (2) it points out the direction for the later development of the understanding of particle physics.

Twenty five years later, in 1979, the first such particle – the Yang–Mills non-Abelian gauge particle for strong interactions, later called the gluon, even though this word "gluon" refers originally to a different proposed particle – was experimentally discovered with the TASSO Collaboration at the German Laboratory DESY [108,109].

Another 33 years later in 2012, this gluon played a central role in the discovery of the Higgs particle [43–45] by the ATLAS Collaboration [139] and the CMS Collaboration [140] at CERN: this Higgs particle is produced predominantly through gluon fusion, i.e., the fusion of one gluon from one proton beam with another gluon from the opposing proton beam.

As soon as the gluon was discovered in 1979, the obvious question was immediately raised: What determines the strength of the gluon-quark coupling constant? I have kept this important question in my mind for 40 years. The conventional answer is discussed in Sect. 3 below, but I have a novel idea about how the Standard Model might be modified to determine g_s. I refer to this idea as the "basic standard model." It is discussed in two unpublished notes [141,142].

2.3 Successes of perturbative QCD

Yuri Dokshitzer

Fifty years is a long time, though not for a theory as ambitious as QCD. To cover all the pQCD applications would be mission impossible. There are many review papers, both topical and anniversary, some good, some excellent. My review is biased, focusing on issues that I personally find important and/or entertaining.

QCD?

Sure. It is undoubtedly the true microscopic theory of hadrons and their interactions. Whether it deserves a status of a well formulated Quantum Field Theory (QFT) is another matter. QCD is an ultimate proof of non-maliciousness of the God of physics. This theory is as amazing as it is embarrassing, in enabling us to predict so much while understanding so little.

Perturbative?

A perturbative (PT) approach means casting an answer as power series in a small expansion parameter. By calculating more terms of the series one aims at increasing accuracy of a theoretical prediction. The quark–gluon dynamics does offer such parameter: the QCD coupling. At small distances it becomes reasonably small thanks to asymptotic freedom, inviting us to draw and calculate Feynman diagrams for interacting quark and gluon fields.

Successes?

Countless experimental findings speak loudly and clearly in favor of pQCD. However, until the color confinement problem is solved, we have to invent hypotheses and build models linking quark–gluon dynamics and the hadron world. It is useful to keep this in mind when what is commonly referred to as a *QCD prediction* confronts reality.

By trial-and-error we learn.

2.3.1 pQCD: domain of interest

The name of the pQCD kingdom is Hard Processes. We call "hard" any process involving hadrons where the energy–momentum that color objects exchange or acquire from (transfer to) colorless fields is much larger than the confinement scale $\mathcal{O}(\Lambda_{\rm QCD})$. Classical examples are e^+e^- annihilation into hadrons, Deep Inelastic lepton–hadron Scattering

(DIS), or the Drell–Yan process of production in hadron collisions of massive lepton pairs or any other heavy colorless objects like $W^{\pm}$, Z^0, H bosons. To the same family belong production of heavy quarks and their bound states, as well as large-p_T photons and hadron jets.

Heavy quarks are often thought to be more friendly towards pQCD than their light siblings. This is true, but not because a massive quark couples to the gluon field more weakly than a massless one. The QCD interaction strength is universal, as a matter of principle. An internal structure of a D meson is as non-perturbative (NP) as that of K or π. At the same time, heavy quarks are typically produced with relatively large transverse momenta $p_T \sim m_Q$ and are closer to one another inside the $Q\bar{Q}$ bound states. This is what actually explains that *friendliness* motto.

Sometimes pQCD applies even to light hadrons. This occurs when a hadron is put under a condition forcing its valence quarks to sit tight in order to hide their color. Small-size configurations dominate when an *initial state* hadron, in spite of having experienced a hit with large momentum transfer, is forbidden to break up and is asked to scatter elastically. Alternatively, a hadron can be squeezed by demanding its exclusive production in the *final state*.

This class of phenomena goes under the name of *color transparency*. Diffractive dissociation of an energetic pion on a nuclear target is a bright example. Normally a big nucleus would absorb the projectile. However, if an incident pion happens to be in a squeezed state, its valence quarks act as a small-size color dipole. Its interaction with the medium weakens and the pion gets a chance to penetrate the nucleus, defying the exponential attenuation wisdom. What one finds behind the target then is a pair of quark jets, because the probability for such a $q\bar{q}$ configuration to return back into a normal pion state is too small to be counted on.

Also pQCD unexpectedly finds its place in the hA (AA) interaction environment where multiple scattering of a projectile effectively pushes up the characteristic hardness scale, $\langle k_T^2 \rangle \propto A^{1/3}$, putting interesting physics like induced gluon radiation or jet quenching under pQCD control.

Whatever the hardness of the process, it is hadrons, not quarks or gluons, that hit the detectors. This makes the applicability of the pQCD approach, even to hard processes, far from obvious. One relies on plausible arguments (completeness, duality) and tries to learn from inclusive hadron observables that are *less vulnerable* to our ignorance about confinement.

2.3.2 pQCD: domain of applicability

The main lesson we learned from confronting QCD expectations with reality is quite encouraging. The strong interaction that is supposed to hold color bearers inside hadrons turns out to be not so strong, if you think about it. The strong color force gets easily screened at large distances by light quarks that pop up from the vacuum. We have not yet mastered this mechanism quantitatively. Meanwhile, the very fact that the confinement happens to be "soft" dramatically enlarges the pQCD playing ground.

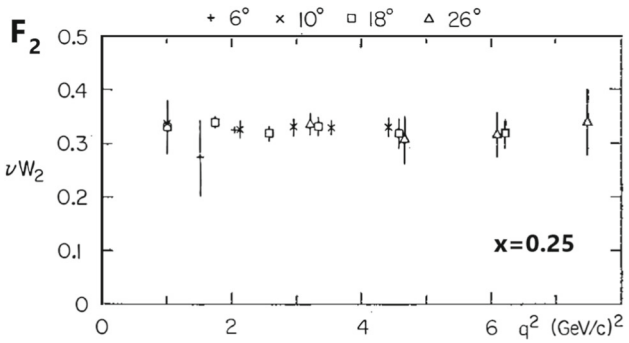


Fig. 11 DIS structure function $F_2 = \nu W_2$ *precociously* scales with momentum transfer q^2 [112]

Precocious pQCD

The parton model [143] pictured electron–nucleon interactions as *elastic* scattering of an incident electron that transfers, via virtual photon exchange, momentum q to a point-like constituent of the target hadron – a parton. Inelasticity of the ep collision is characterized by a dimensionless Lorentz-invariant parameter $x = -q^2/2(q \cdot P)$ which determines an invariant mass W of the final hadronic system: $W^2 - M_P^2 = 2(q \cdot P)(1 - x)$. The physical meaning of the Bjorken variable x becomes transparent in a reference frame where the virtual photon has zero energy component, $q_0 = 0$, and collides with the proton head-on (Breit frame). Here x becomes a fraction of the large proton momentum P carried by the hit parton ($p_{\text{part}} \simeq x P$).

This picture culminated in the Bjorken hypothesis: that the probability of finding a given parton inside the nucleon is independent from the momentum transfer q^2. The Bjorken scaling was expected to hold *asymptotically*, that is when $|q^2|$ is so large as to ensure insignificance of any re-interaction between constituents. In the Bjorken limit $|q^2| \to \infty$ the elastic ep cross section dies out, while proton breakup into large-mass hadron systems dominates: hence *Deep* and *Inelastic*.

The first SLAC–MIT observation of DIS sent a striking message. Defying expectation, the scaling regime manifested itself surprisingly early, right above 1 GeV momentum transfer, as shown in Fig. 11. Charged constituents (read: quarks), probed with better than 0.2 fm resolution, behaved as free objects. And 50 years later they still do.

Another evidence in favor of *precocious freedom* comes from e^+e^- annihilation into hadrons which provides the cleanest environment for exploring QCD. Here all the murky hadron dynamics is restricted to the final state, and we can

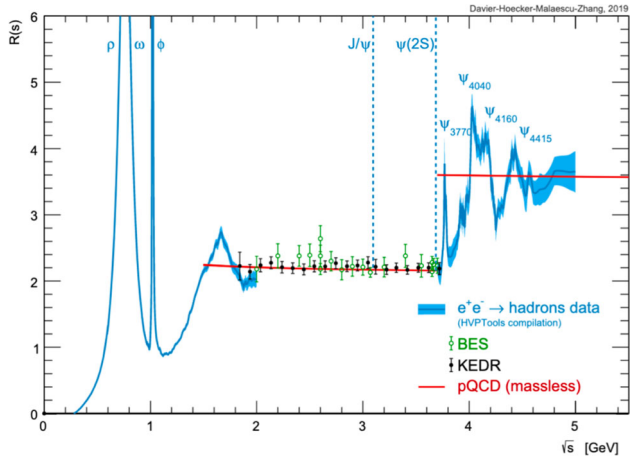


Fig. 12 In e^+e^- annihilation, a quark and an antiquark born with momenta above 1 GeV fly away as free partons

watch what happens to a pair of bare quarks created in the annihilation point and moving apart with light speed.

Figure 12 shows the total hadroproduction cross section, normalized by the QED cross section $e^+e^- \to \mu^+\mu^-$ as a function of annihilation energy $\sqrt{s} = 2E_q$. We see that first the quark and the antiquark interact in the final state producing hadron resonances (vector mesons ρ, ω, ϕ). As soon as the quark energy exceeds 1 GeV, the stormy sea calms down abruptly and turns into still waters. Quarks with larger energies forget about one another and behave as free particles. They separate unimpeded and develop their private multi-hadron images – jets. (The story repeats above the charm threshold.)

This plot contains more than a mere counting of the number of families of colored quarks,

$$R(s) = \frac{\sigma_{e^+e^-\to\text{hadr.}}}{\sigma_{e^+e^-\to\mu^+\mu^-}}, \qquad R_{\text{q.m.}} = N_c \sum_f e_f^2.$$

Notice a slight non-linearity of the pQCD red line in Fig. 12. Its origin – a QCD correction to the annihilation cross section due to gluon radiation:

$$\frac{R(s)}{R_{\text{q.m.}}(s)} = 1 + \frac{3C_F}{4}\frac{\alpha_s(s)}{\pi} + \cdots,$$

where $C_F = (N_c^2-1)/2N_c = 4/3$ is the quark "color charge" (quadratic Casimir operator of the fundamental representation of the $SU(N_c)$ group). The running coupling effect timidly winks at us.

Tau-lepton as a pQCD blessing.
Even at smaller momentum scales, pQCD can be successfully applied. It suffices to "properly place your eyes",[6] that is to choose the right question to ask.

An amusing and practically important example of precocious pQCD control is provided by hadronic decays of the τ-lepton. Given lepton-quark universality of the weak interaction, by simply counting degrees of freedom one would expect

$$R_\tau = \frac{\tau \to \nu_\tau + \text{hadrons}}{\tau \to \nu_\tau + e^-\bar{\nu}_e} = N_c = 3.$$

Experimentally it is 20% higher: $R_\tau \simeq 3.64$. Quite a serious discrepancy. We could refuse to discuss it by presenting a legitimate excuse: the lepton mass $m_\tau \simeq 1.78$ GeV is too small for pQCD to apply.

Meantime, there is a more constructive way to address this discrepancy. In the spirit of the Bloom–Gilman duality idea that has emerged in the DIS context [144], it is tempting to explore whether hadron and quark languages would complement each other. The lepton τ decays via many hadronic channels with squared invariant mass $s = (P_\tau - p_\nu)^2 = m_\tau(m_\tau - 2E_\nu)$ ranging from $m_\pi^2 \simeq 0$ all the way up to m_τ^2. Summing over all hadron states and integrating over s one has a good chance to mimic the QCD prediction, should there be one.

On the QCD side, since the gluon interaction does not discriminate quark flavors ($W^- \to d\bar{u}, s\bar{u}$ in place of $\gamma^* \to u\bar{u}$ or $d\bar{d}$), formation of the final state via virtual W^- is no different from that in the e^+e^- annihilation case. This allows one to express the pQCD correction to the branching ratio B_h via $\alpha_s(m_\tau^2)$ – the strong coupling at the tau-mass scale. Moreover, by employing the Shifman–Vainshtein–Zakharov (SVZ a.k.a. ITEP) sum rules (discussed in Sect. 5.7) designed to match theoretical quark–gluon calculations with hadron phenomenology via dispersion relations [145], it was possible to prove that the NP contributions are negligible [146] being suppressed as a high power of the τ mass, $(\Lambda/m_\tau)^6$ [147].

The creator has chosen the τ mass wisely. It lies conveniently inside a window where $\alpha_s(m_\tau^2)$ is *sufficiently large* as to make pQCD correction significant and well visible, and at the same time *not too large* to undermine the PT treatment. This resulted in [148,149]

$$\alpha_s(m_\tau^2) = 0.345 \pm 0.010,$$

which value is three times larger that the reference QCD coupling at the Z-boson scale, $\alpha_s(M_Z^2)$, and is indispensable as a lever arm for visualizing asymptotic freedom, see Sect. 3 of this volume.

2.3.3 (p)QCD: precursors and hints

QCD inherited quite a dossier of puzzles from the constituent quark model. It is worth recalling certain successes of the pre-QCD quark picture of hadrons, some of which are short of a miracle.

[6] M. B. Voloshin.

Inheritance

Among the first dynamical applications of the constituent quark model of hadrons were the 2-to-3 ratio of the total πp and pp cross sections [150] and an intriguingly simple additive pattern of magnetic moments of baryons (see [151] and references therein). In these and many other phenomena, well before QCD, quarks already demanded to be treated as independent quasi-free entities.

Probably the most amusing example of such inheritance is the so-called quark (or, more precisely, "constituent") counting rule [152,153]. It links the exponent of the energy fall-off of large-angle $a + b \to c + d$ scattering cross sections with the number of "constituents" of the participating (initial and final) particles: $N = n_a + n_b + n_c + n_d$:

$$\frac{d\sigma}{dt} \propto s^{N-2}, \quad -t/s = \mathcal{O}(1). \tag{2.4}$$

A chilling example of this scaling law is provided by the process of photo-disintegration of a deuteron [154]. The scaling Eq. (2.4) with $N = 13$ holds in the photon energy interval $1\text{GeV} < E_\gamma < 4\text{GeV}$ while the cross section falls by whopping six orders of magnitude! (see [155] to enjoy the picture).

Hints.

An approximate constancy of the total hadron–hadron scattering cross sections hinted at the presence of a vector field ($J = 1$) as a strong interaction mediator. Invention of gluons inspired a model of the Pomeron as a two-gluon t-channel exchange [156–158]). It was a little while before the Low-Nussinov Pomeron picture was confirmed and extended by rigorous analysis of high-energy scattering in a non-Abelian QFT [159], to become known as the BFKL Pomeron.

Another early benefit that QCD has offered was an (at least qualitative) explanation of the famous Okubo–Zweig–Iizuka (OZI) rule. It postulated that interacting hadrons do not mind exchanging constituent quarks but hate to allow a quark and its antiquark that are present in the initial state to annihilate. The OZI rule was forged to explain unwillingness of ϕ mesons ($\phi = s\bar{s}$) to decay into light u, d-built mesons. According to QCD, annihilation of a $q\bar{q}$ pair that constitutes a vector meson has to proceed via 3-gluons, so that the decay width becomes small: $\Gamma/M \propto \alpha_s^3$. It may look too brave to rely on the asymptotic freedom concept at scales as small as $M_\phi/2 = \mathcal{O}(0.5\,\text{GeV})$. However, bound states of heavier quarks ($J/\psi = c\bar{c}$ and $\Upsilon = b\bar{b}$ families) can be related with their QED counterpart – the C-odd e^+e^- bound state – orthopositronium [90]. Constructing the ratio of the widths of hadronic and radiative decays

$$J/\psi,\ \Upsilon \to ggg \to X, \quad J/\psi,\ \Upsilon \to \gamma gg \to \gamma + X,$$

one arrives at a reasonable quantitative estimate of the QCD coupling at m_c and m_b scales, correspondingly. Here gluons manifested themselves as mediators of the strong interaction.

Gluons as hidden constituents of the proton also showed up indirectly in DIS as electrically neutral matter that carries about a half of the energy–momentum of the fast proton.

The last but not the least: the nature of multi-particle production in the processes involving hadrons also necessitates the presence of a vector field as interaction mediator.

Indeed, the bulk of inelastic high energy hadron–hadron collisions was long known to produce multi-particle final states with hadrons having finite transverse momenta and distributed uniformly in rapidity. In 1968 Gribov considered a fast proton with large energy $E \gg M_p$ fluctuating into a system of $\ln E$ quasi-real particles as an s-channel image of the t-channel vacuum pole (a.k.a. Pomeron) exchange [160]. Feynman has reverted the picture by prescribing $\ln E$ hadron multiplicity to a fragmenting quark with energy E [161].

A uniform rapidity plateau is the key attribute of vector particles, hinting at gluon radiation underlying production of hadrons.

2.3.4 pQCD: modus operandi

Massless gluons and quarks are treated by pQCD as if they were photons and electrons. This is clearly not a nice thing to do. In the QED case electrons and photons are legitimate QFT objects. They know how to propagate freely, have a definite relation between energy and momentum and therefore can be prescribed a physical (measurable) mass. Causality and unitary unequivocally dictate the analytic structure of their respective Green functions and interaction amplitudes in general.

Quarks and gluons don't have this luxury. Being well aware of this complication, pQCD ignores it in a hope to be considered innocent until proven guilty.

Renormalization: scale

To calculate probability of radiation, a gluon is put on mass-shell, $k^2 = 0$, as if it were a photon. Intensity of photon radiation is proportional to the fine structure constant $\alpha_{\text{e.m.}} \simeq 1/137.04$, whichever the process and its hardness. The on-mass-shell value of the QED coupling is a measurable quantity that determines multitude of macroscopic electromagnetic phenomena.

In QCD, on the contrary, the on-mass-shell coupling $\alpha_s(0)$ is undefinable simply because "on-mass-shell gluon" is an oxymoron, as is "on-mass-shell quark". One has to choose some sufficiently large momentum scale $\mu_R \gg \Lambda_{\text{QCD}}$ and employ $\alpha_s(\mu_R^2)$ as an expansion parameter to construct the PT series. This is called the *renormalization scale*.

The dependence of α_s on μ_R (hence, *running coupling*) is governed by the β-function, see Sect. 3.1. The first two coefficients β_0 and β_1 of the Taylor series of $\beta(\alpha_s)$ are driven by the ultraviolet (UV) behavior of the theory. Their values are universal, while $\beta_{n\geq 2}$ depend on the way α_s is defined.

Obviously, physical observables should not dependent on the choice of μ_R. This enforces, through *running*, a definite μ_R-dependence of the coefficients of higher order terms of the series, starting from the next-to-leading (NLO) one. In practice only a few terms of PT expansion are known for a given observable (say, Born + NLO + NNLO, with N^3LO becoming available in certain cases). One puts the residual μ_R-dependence of truncated series to a good use. By varying μ_R in some interval (conventionally, between $\mu_R/2$ and $2\mu_R$) one gets an ad hoc estimate of theoretical uncertainty due to unknown higher orders.

Renormalization: scheme

By slightly lowering the dimension of the world, $4 \to D = 4–2\epsilon$, one trades UV divergences for singularities in ϵ to tame the misbehaving integrals. Logarithmic UV-divergences of loop integrals that renormalize the coupling translate then into a pole at $\epsilon = 0$. By dropping it ("minimal subtraction") together with a boring constant (an artifact of the trick) one arrives at a finite answer – the $\overline{\rm MS}$ coupling. Dimensional regularization (DREG) [162,163] is a gentle procedure in that it respects and preserves internal symmetries of the problem (with gauge invariance the first to name).[7] Being well suited for multi-loop calculations, the $\overline{\rm MS}$ scheme has become the standard of the trade.

Alternatively, one can introduce α_s directly from a physical observable without bothering about the UV problem [165]. Called *effective couplings*, many have been suggested since, emerging from e^+e^- hadronic annihilation data [166], the Bjorken sum rule [167], static heavy quark interaction potential [168] or intensity of dipole gluon radiation off non-relativistic heavy quarks [169], etc. Effective couplings can be related to one another via the $\overline{\rm MS}$ expansions (see, e.g. [170]).

There is one scheme that deserves special credit. Known as Monte Carlo (MC), Catani–Marchesini–Webber (CMW), bremsstrahlung, or simply "physical scheme", it first appeared implemented in the HERWIG MC parton cascades generator [171] and rediscovered in the context of an optimized pQCD description of inclusive heavy quark fragmentation functions [172]. The same coupling shows up in the anomalous dimension of a cusped Wilson line (Polyakov anomalous dimension) [173,174].

This scheme adds to the $\overline{\rm MS}$ coupling a definite $\mathcal{O}(\alpha_s^2)$ piece that keeps emerging in a multitude of observables. Among them the behavior of DIS parton distributions (pdf) and jet fragmentation functions (ff) in the quasi-elastic limit $1-x \ll 1$, threshold effects, quark and gluon Sudakov form factors and Regge trajectories, etc.

[7] When dealing with a sypersymmetric dynamics, one has to sharpen the DREG tool to preserve the fermion–boson symmetry. This is achieved by turning to the *dimensional reduction* (DRED) [164].

The reason is simple: it is the scheme that defines the coupling by the radiation intensity of gluons with relatively small energies. Radiation of *soft gluons* is classical by nature. In accord with the Low theorem it is fully determined by the classical trajectory of the charge (be it electromagnetic or color one) and is insensitive to quantum properties of the particle that caries it [175].

Infrared-finite coupling

The QCD coupling grows with distance and becomes infinitely large at some point. This is true both at the one-loop level (β_0) where it develops a simple pole, c.f. (1.8), and in the two-loop approximation when one takes into account the β_1 term in the running of the coupling. This is often referred to as the Landau pole/singularity in memory of the discovery 70 years ago by L. Landau and collaborators of the explosive behavior of running coupling in the context of QED.

Beyond the two loops, however, the situation changes. With the sign of β_2 depending on the scheme, some effective charges at this level stop suffering from the Landau singularity and instead freeze in the origin [166,176,177]. Actually, this freezing is as much an artifact as the Landau pole itself. To unambiguously define α_s and establish its behavior at small momenta is inconceivable without cracking the confinement problem.

At the same time, the very supposition that $\alpha_s(k^2)$ is finite for any $k^2 \geq 0$ (more accurately, is *integrable* over the infra-red domain) enhances the predictive power of pQCD. The Parisi–Petronzio analysis of the differential distribution of Drell–Yan pairs with very small transverse momenta q_T and large invariant masses $q^2 \gg q_T^2$ provided a key example [178]. It was enough to assume that such a "good coupling" existed to get a PT prediction that actually *did not depend* on details of its behavior in the origin and agreed with the data.

Assuming the existence of a dispersion relation made it possible to quantify the leading power-suppressed NP contributions by expressing their magnitude via momentum integrals of the "good coupling" over the NP domain [179]. This approach proved to be especially productive in the realm of *jet shapes*, the majority of which suffer from significant $1/Q$ hadronization corrections [180] (see [155] for details).

2.3.5 Partons and jets

The word *jets* appeared (though only once!) in a monumental parton-model study of inclusive production of a nucleon in e^+e^- (the process related by crossing with lepton–nucleon DIS) [181].

The picture of quark jets has been elaborated [182], and the Feynman conjecture implemented as a working hypothesis to characterize the final state structure of hadroproduction

processes with large transverse momenta [89]. In a footnote the authors remarked: "*The question of the ultimate fate of the fractional charge may be a difficulty of the quark–parton model*". Since "*the important longitudinal distances in configuration space for electroproduction*" increase linearly with energy and may become macroscopically large [183] "*This may imply that the active parton tends to travel a considerable distance without interaction before disintegrating into a jet of hadrons. Thus, there can be a separation of fractional charge over large distances in configuration space as well as momentum space*". The footnote ended with a prophetic remark: "*However, this does not mean that partons must "backflow" that distance to provide the necessary neutralization of fractional charge. This can be accomplished, for example, by a polarization current created by parton–antiparton pairs created from the vacuum by the field of the active parton*".

The worry was answered 5 years later when, with the advent of QCD, responsibility for confining fractional charges has been laid upon color.

In 1974 Kogut and Susskind came up with a picture of a flux tube (color string) that connects the quarks. With the color field strength increasing with quark separation, a chain of successive vacuum breakups, $q \to q + q'\bar{q}' \to (q\bar{q}')_{\text{meson}} + q' \to$ etc, contained fractional charges, together with the open color, inside colorless hadrons.

The authors have also remarked that hard gluon bremsstrahlung off the $q\bar{q}$ pair *may be expected* to give rise to three-jet events in the e^+e^- annihilation into hadrons.

The time had come for pQCD to face the challenge.

Gluon jets

To unequivocally confirm QCD's claim to an honorable place of the theory of strong interactions, gluons had to be found manifesting as true particles.

Section 2.2 is devoted to the groundbreaking discovery of 3-jet $e^+e^- \to q\bar{q}g$ events. We'll stay on the theory side and peek into a seminal paper that set up the 3-jet quest [121]. What a shaky ground the authors were pushing off back in 1976! Quote:

- no direct experimental evidence yet exists for gluons (except possibly the fact that not all the nucleon's momentum is carried by known quark constituents),
- there is no direct evidence for asymptotic freedom (though there may be some deviations from scaling in DIS at high Q^2),
- fashion sets $\alpha_s(Q)$ to lie between 0.2 and 1 for $Q^2 \sim 10\,\text{GeV}^2$.

The authors professed *coplanar* structure of the final state, cross section scaling in $x_T = 2p_T/Q$, verified *asymptotic 2-jetness*, and rightly guessed a 10% fraction of 3-jet events.

Moreover, they drew a picture with two hadron chains stemming from the gluon fragmentation and remarked, without much ado:

> Looking at [this] one might naively expect more hadrons to be produced in gluon fragmentation than in quark fragmentation, and therefore that $f(x)$ for gluons should be more concentrated at low x.

That is, higher hadron multiplicity and softer energy spectrum in a gluon jet as compared to quark one. This little picture became a precursor of the Lund model interpretation of a gluon as a "kink" on the color string connecting the separating quark and antiquark [184].

IRCS ideology

In 1977 Sterman and Weinberg drew an image of two-jet events as opposite cones of angular size δ containing all but a small fraction ϵ of the total annihilation energy.

In the Born approximation, $e^+e^- \to q\bar{q}$, the back to back quarks fit in with unit probability. In the next order in α_s there emerge a negative virtual correction to $\sigma_{q\bar{q}}$ and a new 3-particle production cross section $\sigma_{q\bar{q}+g}$, both infinite. However, the collinear divergence at $k_T \propto \Theta \to 0$ (present in all logarithmic QFTs with massless fields) and the soft divergence, $k_0 \to 0$ (specific for vector gluons and photons), cancel in the sum, leaving behind a finite correction $\propto \alpha_s \ln\delta \ln\epsilon$.

The SW construction became the first hadron observable that, after the total cross section $\sigma_{\text{tot}}(e^+e^- \to X)$, enjoyed the power of the Bloch–Nordsieck theorem. An ideology of Infrared-and-Collinear Stability (IRCS) was born:

> If radiative corrections to a given observable happen to be free from collinear and soft gluon divergences and thus the result is finite, feel free to confront the PT answer directly with experiment, without worrying about NP hadronization effects.

The flag got hoisted over the boot camp from where pQCD went on a rampage to conquer multiple production of hadrons in hard interactions: "*the detailed results of perturbation theory for production of arbitrary numbers of quarks and gluons can be reinterpreted in quantum chromodynamics as predictions for the production of jets*" [185].

Defining and finding

A narrow bunch of hadrons is not good enough: one needs an operational definition in order to deal with jets, to predict, study and work with them. There emerged two major threads: 1) to look for a set of cones (of certain angular size) that would embed the final-state hadrons in an optimal way and 2) to look for a pair of particles closest in the momentum space and (if judged close enough) join them into one, thus recursively reducing an ensemble of N hadrons to a few clusters – jets.

The original JADE clustering jet finder used an invariant mass of the pair as closeness measure. It did well experimentally, but did not satisfy theorists. By the time when the Workshop on Jet Studies at LEP and HERA was taking place in Durham in 1990, theorists became too greedy. To deal with respectful IRCS observables (which JADE finder's output are) was no longer enough for them.

Jet rates suffer from (or enjoy, up to you) large double-logarithmic corrections, and theorists were eager to make all-order resummed predictions. And the JADE finder did not allow that because of a weird way it was dealing with small-momenta particles (soft gluons). At a brainstorm session a proposal from the audience was made to replace the invariant mass *distance measure* $m_{ik}^2 \simeq 2E_iE_k(1-\cos\Theta_{ik})$ by the relative transverse momentum $k_T^2 \simeq 2\min\{E_i, E_k\}(1-\cos\Theta_{ik})$ to cure the problem. The next morning Siegfried Bethke who spent a sleepless night testing the new idea came up with encouraging news: the k_T measure did well in yielding jets less affected by hadronization.

First reported in the summary of the Hard QCD working group [186], the "Durham" algorithm [187] has got a "Geneva" cousin [188], and then "Cambridge" [189] and "Aachen" [190] fraternal twins that have further reduced hadronization effects. The k_T-algorithm, generalized to DIS and hadron–hadron collisions [191], allowed theorists to produce all-order resummed expressions for the jet rates in e^+e^- and elsewhere.

For 15 years or so the clustering algorithms lagged behind the cone-based ones. And for a good reason: N^3 operations needed to sort out a final state containing N particles. Given that in the pp environment (not mentioning pA and AA) multiplicities are large, this made clustering procedures impractical.

The tables turned when an ingenious application of combinatorial geometry to the momenta clustering problem by Cacciari and Salam has reduced the calculation load down to $N\ln N$. Development of the "fast-k_T" clustering procedure permitted to analyze large multiplicity final states "in no time" [192]. This was especially welcome since all then-known cone-based finders were caught red-handed at violating the IRCS demand one way or another.

A long and turbulent history of competing jet-finders has terminated with invention of the "anti-k_T" jet finding algorithm [193]. It came in time – right before the start of the LHC operation. It satisfied both theorists (as pQCD-friendly, IRCS respecting) and experimenters (fast and producing aesthetically pleasant roundish jets), and has established itself as the main (if not only) tool of the trade since. A full coverage of *Jetography* can be found in an excellent review [194].

Heavy quark jet

QCD expected the jets initiated by heavy quarks Q to have a hole in the forward direction – *dead cone* of the size $\Theta_0 \simeq m_Q/E$. Indirect consequences of this specific feature have been experimentally confirmed a while ago: Q loses little energy (leading particle effect), light hadron multiplicity in a Q-jet is reduced by a constant, $N_q(E) - N_Q(E) \simeq N_q(m_Q)$ [172].

A direct observation of the dead cone by the ALICE was recently reported in *Nature* [195].

2.3.6 Many jets, some loops

To construct a scattering amplitude at leading order (LO: Born approximation with the minimal power of the coupling constant) one sums up topologically different tree diagrams, each of which is a product of internal Feynman propagators and vertices. Momenta of all internal lines are fixed by kinematics so that no integration is involved. Because of heavier combinatorics and more complicated color structure, the complexity of the scattering amplitude increases with the number of external legs (read: jets).

Loops and divergences

QCD jets have become an indispensable tool for collider experiments in search for new physics. It is imperative to know the yield and structure of multi-jet final states with the best accuracy possible. One has to go beyond the Born approximation and calculate, step by step, higher order corrections. A virtual correction (VC) generates a loop along with an integration over the 4-momentum flowing through the loop. With loops in the game, complexity of the task rises to all new level.

UV divergences being dealt with, VC is still divergent in the collinear and soft corners of the integration space. But so is the inclusive (integrated) cross section of the same order in α_s. This time, due to real emission (RE) of an infinitely soft gluon or a collinear 2-parton configuration in the final state phase space. Combining VC with RE one gets rid of *almost all* divergences. The surviving collinear divergences hide into initial state pdf (and ff, should there be hadrons explicitly registered in the final state). Apart from that, the answer is finite. However it is difficult to get by subtracting infinities. One needs to regularize VC and RE separately and consistently or, better still, to perform subtraction at the level of the integrand to avoid divergences altogether.

NLO

Early NLO studies sent a rather disturbing message: large corrections were found both to Drell–Yan [196] and large-p_T production [197] putting under question the very applicability of the PT approach. There is a good physical reason why those corrections turned out to be alarmingly large. I will hide it from you for lack of space-time. One way or another the initial shock was mitigated and a systematic attack on the NLO started.

The method that has been proposed for e^+e^- annihilation, used DREG to deal with the VC+RE problem [198]. An idea to employ the notion of color dipoles to accurately treat collinear and soft singularities and cancel them at the integrand level gave more flexibility and allowed to construct a popular general purpose scheme for calculating the NLO jet cross sections in any hard process [199].

L-loop VCs are given by $4L$-dimensional Feynman integrals. They are analytic functions of external momentum invariants and can be reduced to a finite set of basic scalar integrals.

The problem has been fully solved for $L = 1$ [200]. This means that today all NLO amplitudes are known (with 6-gluon scattering marking the present-day complexity limit) [201]. Parton showers have been promoted to NLO as well [202].

NNLO

Since 2015, the number of important processes controlled in the following order of pQCD (NNLO) has been steadily increasing. In the bibliography titles of the Les Houches 2019 Summary [203] *next-to-next-to*, or *NNLO* appears 155 times. Drell–Yan/Higgs [204–207] and semi-inclusive DIS [208, 209] allowed to peek into N^3LO.

Just enjoy the names that appear in the $N^{\geq 2}LO$ context: CoLoRFulNNLO and Projection-to-Born methods, Nested soft-collinear and N-jettiness subtractions. A Shakespearean review [210] discusses pros and contras of DREG vs. subtraction regularization.

Mathematical aspects

To calculate Feynman integrals analytically is notoriously hard. General techniques for attacking loop amplitudes were listed and demonstrated in 1996 and are being used since: *spinor helicity formalism, color decompositions, supersymmetry, string theory, factorization and unitarity* [211].

The Loop-Tree duality approach (LTD) was initiated [212] and later generalized to become Four-dimensional Unsubtraction (FDU) [213].

Proceedings of the topical Florence workshop (cunningly named WorkStop/ThinkStart) [214] link to 200+ articles that cover the basics and the progress.

An all-in assault [215] resulted in an astonishing symbiosis of theoretical physics and pure mathematics. Particle theorists, maybe already familiar with integrability, now have to learn twisted cohomology groups, Hopf algebra, algebraic number theory and other scary things.

2.3.7 Resummation and evolution

Art of expansion

Series in α_s can behave well, as for $R(e^+e^-)$, or look troubling as is the case of diphoton production, where moving from NLO to NNLO changes the cross section by 50% [216].

In fact, independent of the observable, PT series in QFT are *asymptotic*, so that beyond $N^{1/\alpha}LO$ things are bound to go haywire. This was not much trouble for QED, but it should be kept in mind for QCD, where the number of reliable terms in the expansion may be not so large.

Examining *how violently* a specific series diverges, hints at how much the NP physics affects a given observable (infrared renormalons [217]).

Resummation

Often α_s acquires one or even two enhancement factors: $\alpha_s \ln Q^2$ (SL), $\alpha_s \ln^2 Q^2$ (DL), and the PT expansion fails. When this happens, in order to get a reliable approximation one has to collect enhanced contributions and sum them in all orders. The Sterman-Weinberg 2-jet cross section acquires DLogs because of a veto imposed on accompanying gluon radiation. The Q_T-spectrum of a Drell–Yan pair or of a hadron registered in the current fragmentation of DIS in the kinematical region $Q_T \ll Q$, and an almost back-to-back energy–energy correlation in e^+e^- were the first examples of inclusive observables which, in spite of not being subject to any explicit veto, are still affected by DLogs [218,219].

In all these cases the origin of one of the logs is soft gluon radiation which is relatively easy to control. This makes resummation of DL-enhanced contributions straightforward and gives rise to Sudakov form factors. Quark and gluon form factors manifest themselves in a multitude of observables characterized by the presence of two different momentum scales. In particular, in distributions of various jet shapes, jet rates, etc.

It is important to emphasize that the very possibility of an all-order resummation depends on whether the operational definition of jets corresponds to the dynamics of the QCD parton multiplication picture (k_T-algorithms vs. JADE, as discussed above).

All-order resummation of single-logarithmic contributions (SLogs) becomes mandatory when we deal either with quasi-collinear configurations of partons with comparable energies (DGLAP physics) or with ensembles of soft gluons at large angles with respect to energetic emitters (radiative corrections to parton scattering amplitudes). In both cases particles involved are strongly ordered in *transverse momenta*.

Factorization

A particle with the smallest k_T in the game *factors out*, in a sense that a singular contribution comes only from its attachment to an external leg. Generalization of the Low theorem from $\omega \ll m_e$ to arbitrary photon energies [220] was a precursor of the QCD k_T/collinear factorization.

Another arbitrary scale enters: factorization scale μ_F. It sets a conventional border between PT and NP ingredients of the problem. In IRCS observables μ_F gets replaced by a variable related to resolution, rendering two well separated

physical scales. For example, $yQ^2 \ll Q^2$ for jet rates, $(1-T)Q^2 \ll Q^2$ for the differential thrust distribution, etc.

Whenever there is *Factorization*, one can carry out *Resummation*, and interpret the results in terms of *Evolution* and corresponding *Evolution Equations*.

A few examples of the application of this idea, both well-known and lesser-known.

KL

The Kirschner–Lipatov equation resums DLogs in parton scattering amplitudes with quark exchange in the t-channel [221]. Such amplitudes fall as the energy s increases, and higher-order DL contributions decelerate this fall. These DLog effects are inherently different from the DLog effects due to accompanying soft gluon radiation (Sudakov form factors).

By isolating the virtual particle with the lowest k_T in the Feynman graph, and using gauge invariance and the unitarity relation, one can form the kernel of the evolution equation for the partial wave amplitudes, with $\ln k_T$ as the "evolution time".

KOS

Kidonakis, Oderda and Sterman have set the quest of resummation of SL radiative corrections to $2 \to 2$ QCD parton scattering amplitudes [222]. In QCD it becomes a multi-channel problem, since each gluon emission (either virtual or real) changes the color state of a parton pair. For gluon–gluon scattering, the anomalous dimension is a $6 \otimes 6$ matrix (for the general $SU(N_c)$ case; which reduces to $5 \otimes 5$ for $SU(3)$). It depends on the scattering angle and, obviously, on the rank of the color group, N_c. Three of the eigenvalues of the anomalous dimension matrix are proportional to N_c, and thus respect the so-called *Casimir scaling* (the perturbative expansion running in N_c), see e.g. [223]. The N_c-dependence of the other three eigenvalues is more involved [224]. They solve the cubic equation whose coefficients exhibit a weird symmetry between the number of colors and the scattering angle:

$$N_c \iff \pm\frac{\ln(s^2/tu)}{\ln(t/u)}.$$

This symmetry can hardly be accidental, but its origin remains a mystery.

ERBL

The ERBL equation applies to exclusive high-Q^2 reactions involving mesons and baryons, e.g. electromagnetic pion form factor [225,226] or photo- (electro-) production of vector mesons like J/ψ [227,228]. Separate components of the valence quark wave function (distribution amplitude) acquire different log Q behavior – anomalous dimensions. The dominant component in the $Q^2 \to \infty$ limit is called the asymptotic wave function: $\psi_\pi(z) \propto z(1-z)$ with z the longitudinal momentum of the fast pion carried by a quark.

It is manifest in the distribution of energy between the two quark jets stemming from diffractive dissociation of a pion in πA collisions [229].

DGLAP

The parton model implied limited transverse momenta. In logarithmic QFTs, instead, k_T^2 are broadly distributed up to the external momentum transfer scale Q^2, resulting in violation of the Bjorken scaling. The first systematic analysis of DIS structure functions and e^+e^- fragmentation functions was carried out in the Leading Logarithmic Approximation (LLA) based on selection of enhanced contributions in each order of PT series, $\sum_n C_n(x)(g^2 \log Q^2)^n$, in the framework of then-known QFT models [230,231].

In 1974 the results were recast in the language of pdf evolving via Markov chain of independent $1 \to 2$ parton splittings [232].

In 1977 arrived the QCD parton dynamics whose name was eventually settled as DGLAP [87,233]. It was received with enthusiasm and gave rise to a host of new ideas: jet calculus, preconfinement, parton showers, to name a few.

With anomalous dimensions now known in 3 loops [234,235], DGLAP does its job, predicting pdf evolution due to space-like cascades. Thanks to factorization, they describe the flux of initial-state partons as an input for any hard lepton–hadron or hadron–hadron interaction. The same universality applies to the final state (time-like cascades).

Parton cascades

Partons have space-like momenta ($k^2 < 0$) in the initial state cascades; in the final state they are time-like ($k^2 > 0$). In the LLA, parton splitting functions in space-like (S) and time-like kinematics (T) are the same: $P_{ba}^{(S)}(z) = P_{ba}^{(T)}(z)$, and so are the anomalous dimensions – Mellin image of $P(z)$. Beyond LLA $P^{(T)}(z)$ departs from $P^{(S)}$ acquiring, in particular, $(\alpha_s \ln^2 z)^k$ terms in $\mathrm{N}^k\mathrm{LL}$.

Originally, the picture of QCD partons was treating the Bjorken/Feynman variable x as being of the order one. Then $\alpha_s \ln^2 z = \mathcal{O}(\alpha_s) \ll 1$ and causes no trouble. However, when x gets *parametrically* small so that $\alpha_s \ln^2 x \sim 1$, an entire tower of these enhanced terms has to be resummed.

This can be achieved by modifying the "time" in the evolution equation from $\ln k_T$ to $\ln \Theta$. In other words, by replacing the k_T-ordered cascades (S) by ordering of successive splitting angles (T). Angular Ordering (AO) takes care of destructive soft-gluon interference and affects particle production.

BFKL

The BFKL equation [159,236] was derived in the LLA in $g^2 \ln s = \mathcal{O}(1)$ to predict high-energy behavior of scattering amplitudes in Yang–Mills theory.

Gluons *reggeize* (spin of a t-channel gluon becomes effectively t-dependent, $J = J_g(t)$). In the vacuum channel ladder diagrams dominate with two Low–Nussinov gluons, now *reggeized*, connected by multiple gluon rungs strongly

ordered in rapidity (*multiregge kinematics*). This yielded the growing total cross section $\sigma_{\rm tot} \propto s^{c\alpha_s}$. The NLL correction lowered the exponent. A power-like energy growth contradicts the asymptotic Froissart theorem, $\sigma_{\rm tot} \leq A \ln^2 s$, but at available energies is legitimate. A need to rescue s-channel unitarity ignited new ideas and, correspondingly, equations: McLerran–Venugopalan Color Glass Condensate model of high-energy saturation (CGC), Balitsky–Kovchegov (BK) and Jalilian-Marian, Iancu, McLerran, Weigert, Leonidov and Kovner (JIMWLK) equations, for references [237].

The true problem is that the high energy scattering does not belong to the pQCD jurisdiction. This is not a hard process as long as no large-k_T scale is involved. As a result, the "BFKL Pomeron" is sensitive to the behavior of the coupling in the NP domain [238,239]. Strictly speaking it would be safer to apply to compact projectiles like bound states of heavy quarks, say $\sigma_{\rm tot}^{J/\psi\ J/\psi}(s)$.

Triggering a jet with $p_T \sim Q$ in DIS target fragmentation region should expose the BFKL dynamics (Mueller–Navelet jets [240]). DGLAP evolution gets suppressed over a large rapidity interval, leaving room for PT-controlled BFKL growth. Experimental data are not yet conclusive [241].

Applied to DIS, BFKL predicts a steep growth of pdf in the $x = Q^2/s \to 0$ limit, equivalent to $s \to \infty$. With DGLAP having its own way of making pdf rise, the two are difficult to disentangle.

BFKL vs. DGLAP

The meaning of *evolution* in the two cases is essentially different. Action $d/d\ln k_T^2$, dynamics in x (DGLAP), vs. action $d/d\ln(1/x)$, dynamics in $\vec{k}_T$ (BFKL). The kernel of the DGLAP evolution equation is a function of the longitudinal momentum $P_{ba}(x)$, the BFKL kernel lives in the plane of transverse momenta $K(\vec{k}_T, \vec{q}_T)$. Eigenvalues of DGLAP are anomalous dimensions; the spectrum of BFKL – Regge trajectories. The origin of DGLAP evolution is the k_T-factorization [242]; BFKL rests upon t-channel unitarity. In spite of all the difference the two are intimately related [243].

2.3.8 Soft gluons and LPHD

It is soft gluon radiation that bears responsibility for faster-than-logarithmic growth of particle multiplicities in hard processes.

Hadron energy spectra in jets brought an exotic fruit. It was not poisonous, but still not easy to digest.

Inside jet

LEP [244], HERA [245] and Tevatron have found that the shape of single-inclusive energy spectra of all-charged hadrons (dominated by pions) is mathematically similar to that predicted by pQCD for soft gluons [246]. And this in spite of the fact that the characteristic *hump* that the spectrum develops because of soft-gluon coherence was situated as low as 1 GeV at LEP (and well below at TASSO energies).

CDF studies proved the origin of the hump due to parton cascading (as opposed to nonrelativistic finite mass effects) [247] and confirmed the pQCD expectation that the particle yield scales with maximal k_T of partons, $E_{\rm jet} \sin \Theta_c$, with Θ_c the half-angle of the jet cone [248].

Inter-jet particles

Studies of hadron flows *in-between jets* added insult to injury. The message here is even more surprising. Information about the color structure of the ensemble of hard partons that form the jets is transmitted to pions with energies of 200–300 MeV, which make up the bulk of the hadrons produced away from the jets ("QCD Radiophysics") [249]. For example, a comparison of the hadron yield in the direction transverse to the 3-jet-event plane with the pQCD prediction of the soft gluon radiation pattern [250], yielded an independent measurement of the ratio of quark and gluon color charges [251], competing with results from hard gluon physics (scaling violation and 4-jet rates) [149].

From a theory standpoint, this similarity was not entirely unexpected. There was a premonition based on a semi-classical analysis of the structure of parton cascades in the configuration space which concluded that when the time comes for a given parton to hadronize, other partons are too far away, leaving no chance for cross-talk [252].

Local Parton–Hadron Duality (LPHD) as a Nature-approved supplement to pQCD sends a powerful message to the future quantitative theory of confinement: the Poynting vector of the color field should translate into Poynting vector of the hadron matter practically undamaged.

2.3.9 Conclusions

There are a number of pQCD-related stories I have left untold.

Why did it take almost 20 years for the inclusive energy–energy correlation in $e^+e^- \to h_1h_2X$, believed to be the most reliable IRCS pQCD prediction, to agree with the experimental data?

Why did the discovery of angular ordering – so important for understanding the coherent nature of particle production – remain unpublished for a long time?

What would make you submit to *Phys. Lett.* an article under the *wrong title* [253]?

How is it that a specific jet shape distribution turns out to be *narrower* than that of the underlying parton ensemble, in spite of usual smearing at the hadronization stage?

How tragic was a misprint in Ref. [254]?

I am confident that by the time *QCD-60* gets published, there will be many more pQCD success stories to tell, in addition to anecdotes.

3 Fundamental constants

Conveners:
Eberhard Klempt and Giulia Zanderighi

The previous two sections reviewed the early history of QCD. Early experiments had provided first support for the quark model, which was definitely established when the charmonium states were discovered. The new theory of strong interaction seemed to account for the non-observation of free quarks (*infrared slavery*) and for possibility to understand deep inelastic lepton scattering off nucleons (*asymptotic freedom*). The following sections turn the focus to major aspects of the development of the *theory* of QCD. Of course, no theoretical discussion can neglect comparisons with experimental data, but we will return to a systematic review of the experimental data and how they compare with QCD predictions only in Sect. 8. Those who want to jump directly to the data might choose to proceed to Sect. 8, and return to these earlier sections when needed.

The masses of the six quarks and the strong interaction constant g_s or $\alpha_s = \frac{g_s^2}{4\pi}$ are fundamental constants of QCD. The masses and α_s are called "constants" even though they depend on the momentum transfer at which they are probed. Quark masses have the additional complication that there are no free quarks for which masses could be determined from experiment directly. Section 3.1 reviews how the quark masses are defined and renormalized, and describes briefly what measurements are compared with lattice predictions in order to determine the values of these fundamental parameters. Some quark masses are very small and others are very large. For the different quark masses, special techniques have been developed such as Effective Field Theory or Heavy Quark Symmetry. These will be discussed later.

Section 3.2 reviews recent determinations of α_s and discusses systematic uncertainties and the procedure used by the current Particle Data Group (PDG) to obtain the world average value of α_s. As it turns out, α_s runs; it is small at high momentum transfer q^2 and large at low q^2. A precise knowledge of this coupling constant is needed to predict any background process in high-energy collisions and to achieve precision in the calculation of signal processes. Later in this review, analytic approximations to QCD are discussed that allow for an extension of α_s determinations to very low q^2 where α_s seems to saturate at $\alpha_s \approx 3$ (Sect. 5.5).

The editors decided to place the determinations of quark masses and α_s early in the volume in order to emphasize that the size of $\alpha_s(Q^2)$ at low Q^2 means that perturbation theory cannot work at the modest values of Q^2 characteristic of matter in its ground state. Nonperturbative methods will be required. Some may prefer to read Sect. 3.1 after the discussion of LQCD (in Sect. 4) and Sect. 3.2 after discussion of the measurements presented in Sect. 12.

3.1 Lattice determination of α_s and quark masses

Luigi Del Debbio and Alberto Ramos

Lattice QCD provides a first-principles, non-perturbative description of the strong interaction in the Standard Model (see Sect. 4.1). Current state of the art simulations include sea quark effects, electromagnetic interactions, and isospin breaking, yielding accurate predictions for low-energy hadronic quantities that are not accessible in perturbation theory.

By discretizing space-time in a cubic lattice with spacing a, lattice QCD provides a non-perturbative regularization of QCD. Moreover this formulation is amenable to numerical simulations using Monte Carlo methods. A key ingredient in any lattice calculation consists in removing the regulator (i.e. taking the continuum limit $a \to 0$). This requires to tune the bare parameters of the lattice QCD action (n_f bare quark masses in lattice units am_i, and the bare coupling g_0) in order to reproduce some hadronic input. Note that since the input of any simulation are dimensionless quantities, only dimensionless predictions can be made. Typically one uses meson masses (π, K and D in case that the simulation includes the charm quark) in units of a reference hadronic quantity to fix the values of the bare quark masses. The reference quantity, usually the mass of the omega baryon M_Ω or the π/K meson decay constants f_π, f_K is the quantity used to *set the scale*: all dimensionless predictions are computed in units of this reference scale. This tuning of the bare parameters in favor of physical observables constitutes the *renormalization* of the theory. Once this process is carried out one can make solid predictions for many other hadronic quantities, and also determine the values of the fundamental parameters of QCD. All in all, quark masses are computed in units of the reference scale. The running of the strong coupling is also computed at energy scales measured in units of the same reference scale. Using as input the experimental value of this reference scale (M_Ω, f_π, f_K or any other convenient choice), one can quote physical dimensionfull predictions.[8] In this way Lattice QCD is able to connect the experimentally observed hadron spectrum (meson and baryon masses) with the fundamental quark masses and strong coupling.

Here we address conceptually how the fundamental parameters of QCD are extracted from Lattice QCD computations, and what are the dominating sources of uncertainty. We will also comment on a few recent results. For a detailed overview on lattice determinations of the strong coupling, we point the reader to the recent review [255]. An exhaustive and critical list of lattice determinations both of quark

[8] The interested reader can consult the section on scale setting in the review [255] and in the 2021 FLAG document [256] for a more detailed discussion.

masses and the strong coupling is available in the excellent FLAG review [256].

3.1.1 The scale of the strong interactions

It is convenient to frame the determination of the strong coupling constant as a determination of the intrinsic scale of QCD. We start from an observable P that depends on a single scale μ (i.e. $P(\mu)$). Ideally this observable should be easy to determine from numerical lattice simulations and with a known perturbative expansion. As we will see later there are several possibilities. Once an observable is chosen, it can be used to define a renormalization scheme (renormalized coupling with *Minimal Subtraction*) via

$$\bar{g}_s^2(\mu) \propto P(\mu), \tag{3.1}$$

where the proportionality factor (a simple normalization) is fixed by

$$\bar{g}_s^2(\mu) \overset{\mu\to\infty}{\sim} \bar{g}^2_{\overline{\text{MS}}}(\mu) \tag{3.2}$$

with $\bar{g}^2_{\overline{\text{MS}}}(\mu) \equiv (4\pi)\alpha_{\overline{\text{MS}}}(\mu)$. It is convenient to work in mass independent renormalization schemes (i.e. the observable $P(\mu)$ is defined in the chiral limit $m_q = 0$). In these schemes the energy dependence of the coupling $\bar{g}_s(\mu)$ is described by the renormalization group (RG) function that has a known perturbative expansion

$$\beta_s(\bar{g}) = \mu\frac{d}{d\mu}\bar{g}_s(\mu) \overset{\bar{g}\to 0}{\sim} -\bar{g}_s^3\sum_{k=0}^{\infty} b_k \bar{g}_s^{2k}, \tag{3.3}$$

where the first two perturbative coefficients

$$b_0 = \frac{1}{(4\pi)^2}\left(11 - \frac{2n_f}{3}\right), \tag{3.4a}$$

$$b_1 = \frac{1}{(4\pi)^4}\left(102 - \frac{38n_f}{3}\right), \tag{3.4b}$$

n_f is the number of fermions in the fundamental representation (i.e. quarks). Different renormalization schemes are related perturbatively by

$$\bar{g}^2_{s'}(\mu) \overset{\bar{g}_s\to 0}{\sim} \bar{g}_s^2(\mu) + c_{ss'}\bar{g}_s^4(\mu) + \cdots. \tag{3.5}$$

It is easy to check that the first two coefficients of the β-function Eq. (3.4) are invariant under such changes of scheme (i.e. they are *scheme independent*).

Integrating the evolution equation (3.3) yields

$$\log\frac{\mu_1}{\mu_2} = \int_{\bar{g}_1}^{\bar{g}_2}\frac{\mathrm{d}x}{\beta_s(x)}, \tag{3.6}$$

where $\bar{g}_1 = \bar{g}_s(\mu_1)$ and $\bar{g}_1 = \bar{g}_s(\mu_1)$. The integral can be rewritten as

$$\int_{\bar{g}_1}^{\bar{g}_2}\frac{\mathrm{d}x}{\beta_s(x)} = \frac{1}{2b_0}\left(\frac{1}{\bar{g}_1^2} - \frac{1}{\bar{g}_2^2}\right) + \frac{b_1}{b_0^2}\log\frac{\bar{g}_1}{\bar{g}_2} + \int_{\bar{g}_1}^{\bar{g}_2}\mathrm{d}x\left[\frac{1}{\beta_s(x)} + \frac{1}{b_0x^3} - \frac{b_1}{b_0^2x}\right]. \tag{3.7}$$

Note that given the asymptotic form of the β_s function (Eq. 3.3), the original integral in Eq. (3.6) is divergent when either $\bar{g}_1 \to 0$ or $\bar{g}_2 \to 0$. On the other hand the integral in Eq. (3.7) is finite in these limits (cf. the integrand is $\mathcal{O}(x)$). This observation allows us to split the integral in Eq. (3.7) as $\int_{\bar{g}_1}^{\bar{g}_2} = \int_{\bar{g}_1}^{0} + \int_0^{\bar{g}_2}$ and write Eq. (3.6) in the following way

$$\log\mu_1 - \frac{1}{2b_0\bar{g}_1^2} - \frac{b_1}{b_0^2}\log\bar{g}_1 \tag{3.8}$$

$$+ \int_0^{\bar{g}_1}\mathrm{d}x\left[\frac{1}{\beta_s(x)} + \frac{1}{b_0x^3} - \frac{b_1}{b_0^2x}\right] = \tag{3.9}$$

$$\log\mu_2 - \frac{1}{2b_0\bar{g}_2^2} - \frac{b_1}{b_0^2}\log\bar{g}_2 \tag{3.10}$$

$$+ \int_0^{\bar{g}_2}\mathrm{d}x\left[\frac{1}{\beta_s(x)} + \frac{1}{b_0x^3} - \frac{b_1}{b_0^2x}\right]. \tag{3.11}$$

Note that this last equation claims that a function of μ_1 (the left hand side) is equal to a function of μ_2 (the right hand side). The only solution is that both are constant. The constant is defined to be $\log\Lambda_s$ and we can write

$$\Lambda_s = \mu\left[b_0\bar{g}_s^2(\mu)\right]^{-\frac{b_1}{2b_0^2}} e^{-\frac{1}{2b_0\bar{g}_s^2(\mu)}} \times \exp\left\{-\int_0^{\bar{g}(\mu)}\mathrm{d}x\left[\frac{1}{\beta_s(x)} + \frac{1}{b_0x^3} - \frac{b_1}{b_0^2x}\right]\right\}. \tag{3.12}$$

Note that the integration of the renormalization group equation here is exact, valid beyond perturbation theory. The combination on the right-hand side of Eq. (3.12) has units of mass, and is independent of μ. It is called the Λ-parameter and can be understood as the *intrinsic scale* of QCD. It is a free parameter, which provides a boundary condition for the evolution equation of the coupling.

Determining Λ is equivalent to determining the coupling constant. It is customary to report the value of $\alpha_s(M_Z^2)$ in the $\overline{\text{MS}}$ scheme, however the latter can be used together with the perturbative expansion of the beta function to compute the Λ-parameter. While the two pictures are clearly equivalent, there are some advantages in focussing on Λ as the main character of our story:

- It makes clear that the determination of the strong coupling constant really amounts to the determination of one energy scale.
- Although the Λ-parameter depends on the renormalization scheme. The relation between Λ-parameters in two

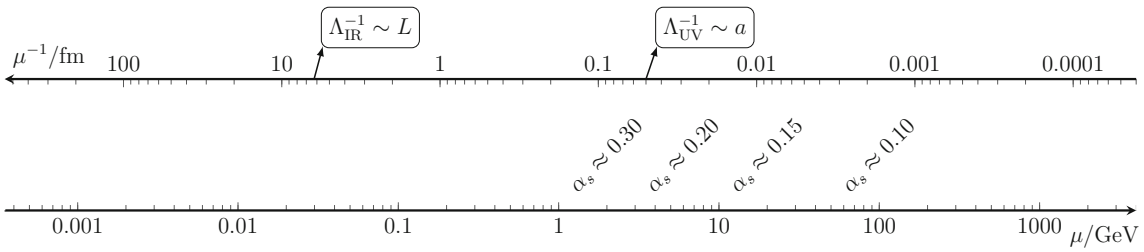


Fig. 13 Any quantity determined in a lattice simulation must be determined at energy scales between the intrinsic UV cutoff of a few GeV (given by the lattice spacing $\Lambda_{\rm UV} \sim a^{-1}$) and the IR cutoff (given by the volume simulated $\Lambda_{\rm IR} \sim L^{-1}$. On a typical state of the art simulations these scales are a few GeV and a few dozen MeV respectively

different schemes is exactly given by a one-loop computation. In order to see this we recall that by convention couplings in different schemes are normalized so that they agree to leading order (cf. Eq. 3.2). This implies that renormalized couplings in two schemes s and s' are related perturbatively by

$$\bar{g}_{s'}^2(\mu) \overset{\bar{g}_s \to 0}{\sim} \bar{g}_s^2(\mu) + c_{ss'}\bar{g}_s^4(\mu) + \cdots, \tag{3.13}$$

with $c_{ss'}$ a finite number. This implies that the relation

$$\frac{\Lambda_{s'}}{\Lambda_s} = \exp\left(\frac{-c_{ss'}}{2b_0}\right) \tag{3.14}$$

is exact.

- The Λ-parameter is defined non-perturbatively. Even for schemes that are intrinsically defined in a perturbative context: $\overline{\rm MS}$ is a "perturbative scheme", but $\Lambda_{\overline{\rm MS}}$ is a meaningful quantity beyond perturbation theory thanks to Eq. (3.14).
- Even if the actual precision in the determination of the strong coupling looks impressive ($\approx 0.7\%$), this amounts to a determination of the Λ-parameter with approximately a 4% uncertainty. In particular some sub-percent effects (QED and isospin breaking corrections) are subdominant for lattice extractions of the strong coupling.

3.1.2 Challenges in extractions of the strong coupling

The extraction of the Λ-parameter in units of a well determined hadronic scale $\mu_{\rm had}$ (like the proton mass) via Eq. (3.12) requires the knowledge of the β-function in the scheme of choice, $\beta_s(x)$, for values $x \in [0, g_s(\mu_{\rm had})]$. Although in principle Lattice QCD can determine the running of $g_s(\mu)$ at any energy scale (it is just the scale dependence of the observable $O(\mu)$ in Eq. (3.1)), computational constraints impose that a typical lattice simulation can only resolve a certain range of scales. In particular if we want to describe hadronic physics, we can reach at most scales $\mu_{\rm PT} \sim 2{-}5$ GeV (see Fig. 13). For this reason, any lattice QCD extraction of the strong coupling uses the perturbative expansion

$$O(\mu) = \sum_{k=0}^{n_{\rm PT}} c_k \alpha_{\overline{\rm MS}}^k(\mu) + \mathcal{O}\left(\alpha_{\overline{\rm MS}}^{n+1}(\mu)\right) + \mathcal{O}\left(\frac{\Lambda^p}{\mu^p}\right), \tag{3.15}$$

The known perturbative coefficients c_i ($i = 1, \ldots, n_{\rm PT}$) together with the known 5-loop running of the beta function allow us to estimate the high-energy contribution

$$\int_0^{\bar{g}(\mu_{\rm PT})} {\rm d}x \left[\frac{1}{\beta_s(x)} + \frac{1}{b_0 x^3} - \frac{b_1}{b_0^2 x}\right], \tag{3.16}$$

to $\Lambda_{\overline{\rm MS}}$. It is worthwhile to emphasize a few subtleties involved in this procedure. Since we only know a few terms in the perturbative expansion of the observable, the missing higher orders are a source of systematic error in the determination of Λ. In fact it is easy to convince oneself that it introduces uncertainties of order

$$\mathcal{O}(\bar{g}^{2n_{\rm PT}}(\mu_{\rm PT})). \tag{3.17}$$

A further source of systematic error comes from non-perturbative (power corrections) to the perturbative expansion. These corrections are suppressed as

$$\mathcal{O}\left(\frac{\Lambda^p}{\mu^p}\right).$$

Both sources of systematic effect can be eliminated by just pushing $\mu_{\rm PT}$ to a high enough scale, but with data only available in a limited range of energies it is challenging to estimate the size of these corrections. Moreover, the perturbative corrections $O(\alpha_{\overline{\rm MS}}^{n_{\rm PT}+1}(\mu))$ decrease very slowly (i.e. logarithmically) with the scale $\mu_{\rm PT}$. This makes reducing perturbative uncertainties an *exponentially* difficult problem.

The window problem

The need to use low energies to determine the Λ parameter in terms of a known, precise, hadronic input, is at odds with the need to reach large energy scales where perturbation theory is applicable with high enough accuracy. This is usually referred to as the window problem. In practice scales of a few GeV are reached and the estimates of perturbative uncertainties remain the main source of error in most lattice calculations. Reference [255] estimates that perturbative

uncertainties alone amount to about 1–2% error in $\alpha_s(M_Z)$ for any method that suffers from the window problem.

Dedicated approaches
There exists however a known solution to overcome this intrinsic difficulty, and it comes under the name of finite size scaling [257]. The idea consists in decoupling the simulations where the hadronic input is determined from the simulations used to determine the running of the coupling. Each simulation can only resolve a limited range of scales, but a recursive procedure called finite size scaling allows us to relate the energy scales resolved in different simulations (see below for more details). Another recent proposal [258] does not provide a complete solution to the window problem, but reduces the problem substantially. In particular in this approach we are only concerned with power corrections, that decrease much faster with the energy scale than perturbative ones.

3.1.3 Lattice observables

There is a wide variety of lattice observables that are used for a determination of the strong coupling. This rich landscape allows for multiple independent determinations, providing a robust cross-check of the methodologies. Here we want to emphasise the broad range of observables. For a full review and combination of the results we refer the reader to Refs. [255] and [256].

We first review the dedicated strategies that aim at solving (or ameliorating) the window problem with a dedicated approach. Typically they require dedicated simulations and the uncertainties are statistically dominated.

Finite size scaling.
An ingenious solution to the window problem is obtained by separating the RG evolution, resolving only a limited range of scales in each single simulation, and adopting a recursive procedure to connect different simulations. The main idea is to use a finite-volume renormalization scheme, where the renormalization scale is identified with the inverse volume of the lattice. The renormalized coupling, denoted here as

$$\bar{g}^2_{\rm SF}(\mu), \qquad (\mu = 1/L), \tag{3.18}$$

is extracted from observables computed in Monte Carlo simulations. The running of the coupling is encoded in the so-called *step scaling* function,

$$\sigma_s(u) = \bar{g}^2(\mu)\Big|_{\bar{g}^2(\mu/s)=u}, \qquad (\mu = 1/L), \tag{3.19}$$

which yields the value of the renormalized coupling at the scale μ as a function of its value at the scale μ/s, where s a scaling factor. The step scaling function is evaluated numerically by computing $\bar{g}^2_{\rm SF}(\mu)$ on pairs of lattices of size L and sL. Thereby multiple simulations on physical volumes much smaller than the typical hadronic scales are used to compute the non-perturbative evolution of the coupling from a hadronic scale $\mu_{\rm had}$ up to a high-energy scale, $\mu_{\rm PT}$, where the matching with perturbation theory is fully under control. While these volumes are too small to study hadronic physics, they are perfectly suitable to study the RG flow of the coupling. The only experimental input needed in this procedure is *one* dimensionful quantity that needs to be compared to one lattice measurement in a large volume in order to set the scale in physical units. It is interesting to remark that the strong coupling constant in this approach is determined from just *one* experimental dimensionful quantity. Perturbation theory is only used at scales larger than the perturbative scale, $\mu_{\rm PT} = s^n \mu_{\rm had}$. This scale can be made (almost) arbitrarily large with a modest (but dedicated) computational effort (typically $\mu_{\rm PT} \sim 100\,\text{GeV}$). Finally it is worthwhile to emphasise that for the determinations based on finite size methods, the main source of uncertainty is statistical rather than systematic. Dedicated simulations will allow further improvements.

Heavy quark decoupling
Recently a new way to ameliorate the window problem has been proposed [258] (see also the review [259]). It is well known that the QCD coupling with n_l massless quarks and n_h heavy quarks (with mass $M \gg \Lambda$), can be matched using perturbation theory to the coupling of QCD with n_l massless quarks. This matching is done in perturbation theory to high order and the perturbative and non-perturbative corrections are very small. These perturbative decoupling relations can also be understood as relations between the Λ parameters with $n_l + n_h$ flavors and the Λ parameter with n_l flavors

$$\frac{\Lambda^{(n_l)}}{\Lambda^{(n_l+n_h)}} = P_{nl,n_l+n_h}(M/\Lambda). \tag{3.20}$$

Both perturbative and non-perturbative uncertainties are very small in these relations even for quark masses of the order of the charm [260].

The main point of this new proposal consists in simulating n_f fictitious heavy quarks. Since all quarks are heavy, a coupling computed in this setup $\bar{g}^{(n_f)}(\mu, M)$ is, up to heavy mass corrections just a pure gauge coupling

$$\frac{\bar{g}^{(n_f)}(\mu, M)}{g^{(0)}(\mu)} \overset{\Lambda/M\to 0}{\sim} 1 + \mathcal{O}(M^{-2}), \tag{3.21}$$

where $\mathcal{O}(M^{-2})$ represent corrections that can be $(M/\mu)^2$ or $(M/\Lambda)^2$. Conversely we can declare that both couplings are the same at slightly different values of the scale

$$\bar{g}^{(n_f)}(\mu^{(n_f)}, M) = g^{(0)}(\mu^{(0)}), \tag{3.22}$$

implying the relation between scales

$$\frac{\mu^{(n_f)}}{\mu^{(0)}} \overset{\Lambda/M\to 0}{\sim} 1 + \mathcal{O}(M^{-2}). \tag{3.23}$$

Together with the basic definition of the Λ parameter Eq. (3.12), this last relation allows immediately to write a relation between Λ parameters

$$\frac{\Lambda^{(0)}}{\mu^{(0)}} \overset{\Lambda/M\to 0}{\sim} P_{0,nf}(M/\Lambda)\frac{\Lambda^{(n_f)}}{\mu^{(n_f)}} + \mathcal{O}(M^{-2}) \tag{3.24}$$

This strategy allows us to determine the n_f-flavor Λ-parameter from the pure gauge one. One only needs the values of a massive coupling with either three or four flavors in order to apply the matching condition Eq. (3.22), and the pure gauge Λ-parameter

$$\Lambda^{(n_f)} = \lim_{M\to\infty} \frac{\mu^{(n_f)}}{P_{0,nf}(M/\Lambda)} \times \frac{\Lambda^{(0)}}{\mu^{(0)}} \tag{3.25}$$

The limit of infinite mass ensures that all corrections (both perturbative and power corrections) vanish, which makes this an exact relation.

Although this strategy does not completely solve the window problem, the slowly decreasing perturbative uncertainties are only present in the pure gauge determination of $\frac{\Lambda^{(0)}}{\mu^{(0)}}$. Note that the pure gauge theory is much more tractable: simulations are much cheaper, algorithms are better, and the step scaling strategy is much more straightforward. Perturbative uncertainties in the decoupling of heavy quarks are negligible, and only the power corrections $\mathcal{O}(M^{-2})$ have to be dealt with. A recent publication [261], using quarks in the range 2–12 GeV shows that precise results can be obtained with this strategy. The uncertainty is still dominated by statistical uncertainties, and in fact a substantial part of it comes from the pure gauge running, which can be further reduced.

Now we move to strategies that suffer from the window problem described above. In all these methods the uncertainties are dominated typically by the uncertainties associated with the truncation of the perturbative expansion Eq. (3.15), or the cutoff effects arising from the difficulty in performing a continuum extrapolation for quantities defined at scales of a few lattice spacings.

Ghost-ghost-gluon vertex
The QCD vertices are computed numerically and compared to their perturbative expansion. As the field correlators involved are not gauge-invariant, these calculations require a gauge-fixing procedure, which has potential extra uncertainties due to Gribov copies [262,263]. Non-perturbative corrections and lattice cutoff effects are sizeable in the regime of current simulations (see [255] for a review.).

Static potential
The interaction between static quarks is known to high orders in perturbation theory, and the data seems to follow to perturbative prediction down to scales of the order of 1.5 GeV. The main drawback comes from the fact that the observable is not IR-safe, which leads to the resummation of soft and ultra-soft divergences, and hence the introduction of an extra soft scale in the problem.

Heavy-quark correlators
The pseudoscalar density correlators are defined as

$$G(x_0) = a^6(am_0)^2 \sum_{\mathbf{x}} \langle \overline{\psi}\gamma_5\psi(\mathbf{x}, x_0)\, \overline{\psi}\gamma_5\psi(\mathbf{0}, 0)\rangle. \tag{3.26}$$

Note that after summing over all spatial sites on the right-hand side, the correlator only depends on x_0. The normalization is fixed by multiplying the field correlator by the factor $a^6(am_0)^2$. Their moments have a well-defined perturbative expansion in powers of the strong coupling constant. These correlators are computed in lattice simulations, which yield a good statistical precision on the final result. The main drawback of this approach is the large cutoff effects that affect the quantities used. It is very challenging indeed to explore energy scales larger than the physical charm quark mass $m_c \sim 1.4\,\text{GeV}$, which is clearly not in the perturbative regime. The recent work in Ref. [264] explores different energy scales in the range $\bar{m}_c - 3\bar{m}_c$, but the continuum extrapolation is very challenging already at $\mu \gtrsim 2m_c$.

Wilson loops
The expectation values of Wilson loops of multiple sizes $m \times n$ are computed at the scale of the lattice cutoff $1/a$. While these quantities are not extrapolated to their continuum limit, they can be computed in bare lattice perturbation theory. The perturbative series can then be translated into an expansion in the remormalized coupling $\alpha_{\overline{\text{MS}}}(\mu)$. The typical scale for these observables is $\mu \sim 1/a$. Unfortunately the known perturbative orders are not sufficient to describe the data and several coefficients of the expansion need to be fitted. While the statistical uncertainty of these determinations is excellent, they are plagued by the systematic errors due to the perturbative truncation.

Hadron vacuum polarization (HVP)
The strong coupling constant can be extracted from the correlators of vector and axial vector currents:

$$V^a_\mu(x) = \bar{\psi}_a\gamma_\mu\psi_a(x),$$
$$A^a_\mu(x) = \bar{\psi}_a\gamma_5\gamma_\mu\psi_a(x),$$

after a decomposition in Fourier space (with $J_\mu = V_\mu, A_\mu$)

$$\int \mathrm{d}^4x\, e^{\imath px}\, \langle J^a_\mu(x) J^a_\nu(0)\rangle$$
$$= (\delta_{\mu\nu}p^2 - p_\mu p_\nu)\Pi^{(1)}_J(p^2) - p_\mu p_\nu \Pi^{(0)}_J(p^2).$$

The quantity

$$\Pi(p^2)$$
$$= \Pi^{(0)}_V(p^2) + \Pi^{(1)}_V(p^2) + \Pi^{(0)}_A(p^2) + \Pi^{(1)}_A(p^2)$$

is dimensionless and has a perturbative expansion

$$\Pi(p^2) \overset{p\to\infty}{\sim} c_0 + \sum_{k=1}^{4} c_k(s)\alpha_{\overline{\rm MS}}^k(\mu) + \mathcal{O}(\alpha_{\overline{\rm MS}}^5), \quad (s = p/\mu)$$

known up to 5-loops. The constant term $c_0(s)$ is divergent, so that the strong coupling is usually extracted from the difference $\Pi(p^2) - \Pi(p^2_{\rm ref})$, or the Adler function

$$D(p^2) = p^2 \frac{{\rm d}\Pi(p^2)}{{\rm d}p^2}. \tag{3.27}$$

The main issue with extractions based on the HVP is that power corrections are significant even for large momenta [265]. Reference [266] pushes the determination to high energies, so that the data can be described without any power corrections, but then cutoff effects become larger and the window of scales to obtain the strong coupling decreases.

Dirac spectral density

The density of the eigenvalues of the Dirac operator,

$$\rho(\lambda) = \frac{1}{V}\left\langle \sum_k [\delta(\lambda - \imath\lambda_k) + \delta(\lambda + \imath\lambda_k)] \right\rangle, \tag{3.28}$$

has recently been used to determine the strong coupling via its perturbative expansion

$$\rho(\lambda) = \frac{3\lambda^3}{4\pi^2}\Big(1 - \rho_1(s)\alpha_{\overline{\rm MS}}(\mu) - \rho_2(s)\alpha_{\overline{\rm MS}}^2(\mu) - \rho_3(s)\alpha_{\overline{\rm MS}}^3(\mu) + \mathcal{O}(\alpha_{\overline{\rm MS}}^4)\Big), \quad (s = \mu/\lambda).$$

The extraction of the spectral density is usually performed at very low energy scales in order to keep the discretization effects under control. Recent work [267] imposes a cut $a\lambda < 0.5$ in order to avoid a substantial deviation from the continuum result. This restricts the energy scales that can be reached with their data-set (with lattice spacings $a^{-1} = 2.5, 3.6$ and $4.5\,\text{GeV}$) to $\lambda < 1.2\,\text{GeV}$.

3.1.4 Determinations of the quark masses

Because of confinement, only color-neutral states are observed as physical states and therefore the quark masses cannot be measured directly in experiments. On the other hand lattice QCD offers a unique opportunity to determine these quantities. In fact the n_f bare quark masses appearing as parameters in the lattice QCD action have to be tuned using n_f physical observables in order to make any meaningful prediction. Once this tuning is performed, we only need to renormalize its values to some convenient scheme. The scale dependence of renormalized values for quark masses in mass-independent renormalization schemes is described by the mass anomalous dimension, $\gamma(\bar{g})$, which only depends on the gauge coupling and obeys the RG equation

$$\mu\frac{d}{d\mu}\bar{m}_i(\mu) = \gamma(\bar{g})\bar{m}_i(\mu) \overset{\bar{g}\to 0}{\sim} -\bar{g}^2\sum_{k=0}^{\infty} d_k \bar{g}^{2k}, \tag{3.29}$$

where the leading perturbative coefficient

$$d_0 = \frac{1}{(4\pi)^2}\left(11 - \frac{2n_f}{3}\right) \tag{3.30}$$

is scheme-independent. As for the coupling, the quark masses are defined in a given renormalization scheme and at a given renormalization scale; the conventional practice is to quote a value for the masses in the $\overline{\rm MS}$ scheme, $\bar{m}_{\overline{\rm MS}}(\mu)$ (with $\mu = 2$ GeV for light quarks), but as in the case of the coupling we find more natural to work with renormalization group invariant (RGI) quantities. The evolution equation, Eq. (3.29), can be integrated exactly to yield

$$M_i = \bar{m}_i(\mu)\left(2b_0\bar{g}(\mu)^2\right)^{-d_0/(2b_0)} \times \exp\left\{-\int_0^{\bar{g}(\mu)} {\rm d}x \left[\frac{\gamma(x)}{\beta(x)} - \frac{d_0}{b_0 x}\right]\right\}. \tag{3.31}$$

Once again we can think of the RGI mass M_i as a scale-independent energy that specifies the boundary condition for the mass evolution and hence fully determines the renormalized mass at all energies. An additional benefit of quoting RGI quark masses is that they are scheme independent (and therefore well defined beyond perturbation theory). On the other hand, the determination of RGI quark masses requires the knowledge of the evolution of the coupling. Given that the current precision of the Λ parameter (about 4%) is much lower than the precision of quark masses at low energies (about 1%), the values of quark masses at a few GeV are much more precise than their RGI counterparts. Note however that this usually means that perturbation theory has been used at a few GeV, and all the caveats about the use of perturbation theory at medium energies raised in the previous section are also applicable here; the determination of quark masses is also plagued by a window problem. However from a practical point of view, the perturbative uncertainties in this case seem to be much better behaved than in the case of the extractions of the strong coupling.

Nowadays the most precise results available in the FLAG review [256] for light and heavy quark masses are obtained in the isosymmetric limit. There are few subtleties involved in these extractions; they originate from the fact that experimental inputs include QED and isospin-breaking corrections, while these effects are not included in the lattice simulations. These effects are small but they are relevant at the level of precision of state-of-the-art lattice computations. Ideally one would like to subtract the isospin breaking corrections from

the experimental data. The problem is that electromagnetic interactions affect the RG functions (both $\beta(\bar{g})$ and $\gamma(\bar{g})$) with $\mathcal{O}(\alpha_{\rm EM})$ contributions: quarks with different electric charges (like the u and d quarks) run differently. QED makes the isospin symmetric point ill defined. Even if we impose $\bar{m}_u(\mu) = \bar{m}_d(\mu)$ at $\mu = 2\,\text{GeV}$, the u and d quarks will be non-degenerate at another generic renormalization scale. Since the subtraction of isospin breaking corrections depends on the definition of the isospin symmetric limit, it is clear that there are (small) ambiguities whatever convention one chooses. The FLAG review [256] contains a detailed discussion both in the quark mass section and in the scale setting section about this particular issue, and the reader is encouraged to consult it for more details.

We end this introduction by emphasizing that the inclusion of the leading QED and strong isospin breaking corrections (including quark loop effects) is an active area of research in lattice QCD. Results with a first principles description of the standard model at low energies, including QCD, QED and strong isospin breaking, are rapidly becoming the new standard for lattice computations where this level of precision is required.

3.1.5 Quark mass definitions

Here we consider the determination of quark masses in QCD alone (i.e. a sensible definition of the isospin symmetric point has been made). Quark currents play a central role in QCD. In particular, the axial current and pseudo scalar density

$$A^a_\mu(x) = \bar{\psi}(x)\gamma_\mu\gamma_5\frac{\sigma^a}{2}\psi(x), \tag{3.32}$$

$$P^a(x) = \bar{\psi}(x)\gamma_5\frac{\sigma^a}{2}\psi(x), \tag{3.33}$$

are expected, in the continuum, to obey the PCAC relation

$$\partial_\mu A^a_\mu(x) = m P^a(x). \tag{3.34}$$

This relation is often used to *define* renormalized quark masses. The reason is that we expect the same relation to hold in the lattice regularized theory after renormalization and up to cutoff effects.[9] The axial current and pseudo scalar density are renormalized multiplicatively

$$(A_R)^a_\mu(x) = Z_A A^a_\mu(x), \tag{3.35}$$

$$(P_R)^a(x) = Z_P(\mu) P^a(x). \tag{3.36}$$

Note that the axial current renormalization factor is scale independent. Quark masses are also expected to renormalize multiplicatively $\bar{m}(\mu) = Z_m(\mu)m_0$, leading to the lattice version of the PCAC relation

$$\partial_\mu A^a_\mu(x) = \frac{2Z_m(\mu)Z_P(\mu)}{Z_A} m_0 P^a(x). \tag{3.37}$$

This relation allows to determine the renormalized quark masses via the relation

$$\bar{m}(\mu) = Z_m(\mu)m_0 = \frac{Z_A\langle\partial_\mu A^a_\mu(x)O_{\rm ext}\rangle}{Z_P(\mu)\langle P^a(x)O_{\rm ext}\rangle}, \tag{3.38}$$

with much freedom to choose the probe $O_{\rm ext}$. Note that the *running* of the quark masses is given by the scale-dependent renormalization factor $Z_P(\mu)$. There are several methods to determine it on the lattice. Most recent works use nonperturbative renormalization schemes.

RI-(S)MOM schemes

These renormalization schemes are conceptually very similar to the one used in perturbation theory. There exist several possibilities, but all are based on imposing a suitable renormalization condition to some Green functions with external momenta playing the role of the renormalization scale. In principle the renormalization scheme is formulated in infinite volume and at zero mass. In this setup the connection with perturbation theory is known to high accuracy (up to 4-loops), but this setup cannot be simulated directly on the lattice, so the infinite volume and zero mass limit require a dedicated study. In particular these methods suffer from a window problem (the impossibility to keep the volume large and at the same time have access to high energy scales where perturbation theory can be trusted).

Finite volume schemes

In these schemes the renormalization condition is imposed in a finite volume L, which plays the role of the renormalization scale (i.e. $\mu \sim 1/L$). With a smart choice of boundary conditions one can directly simulate massless quarks. Contact with perturbation theory is typically only known up to 2-loops, but using the techniques of finite size scaling, this matching can be performed at very high energies (i.e. 100 GeV), where perturbative uncertainties are negligible.

3.1.6 Approaches for heavy quarks

Heavy quarks are difficult to simulate on the lattice. The reason is that in order to have discretization errors under control, the lattice cutoff a^{-1} has to be much larger than all other scales considered in the problem. In particular we require $am \ll 1$. The lattice community has typically dealt with this problem using an effective description for the heavy quarks (see for example Refs. [288] and [289]). This topic is beyond the scope of this review. Here instead we will focus on some recent works that use a relativistic formulation for the heavy quarks. In particular the recent work [287] uses the expansion of a heavy-light meson mass M_{hl} as a function of

[9] Depending on the type of fermion formulation used and other details, the cutoff effects can be $\mathcal{O}(a)$ or $\mathcal{O}(a^2)$. In practice most lattice determinations nowadays choose to eliminate the linear effects in a.

the heavy quark pole mass m_h

$$M_{hl} = m_h + \bar{\Lambda} + \frac{\mu_\pi - \mu_G(m_h)}{2m_h} + \mathcal{O}(1/m_h^2). \tag{3.39}$$

Here $\bar{\Lambda}$ is the binding energy, $\mu_\pi/2m_h$ is the kinetic energy and $\mu_G(m_h)$ is the hyperfine energy. This relation allows to fit meson masses to the heavy quark pole mass, and therefore to determine it by using the perturbative relation

$$m_h \sim \bar{m}_{\overline{\rm MS}} \left(1 + \sum_{k=0}^{\infty} r_n \alpha^{n+1}(\bar{m}_{\overline{\rm MS}})\right). \tag{3.40}$$

The problem of this approach is that the pole mass has a terribly behaved perturbative expansion. In fact

$$r_n = (2b_0)^n \Gamma(n + 1 + b_1/(2b_0^2)). \tag{3.41}$$

Reference [287] uses instead the minimal renormalon subtraction scheme, that has better PT properties.

Making a long story short, heavy-light meson masses are directly related to quark masses, without the need of any non-perturbative renormalization. This approach is used to determine the b meson mass. Masses of other quarks are extracted from appropriate quark mass ratios, that do not need the determination of any renormalization constant.

It has to be pointed out that the heavy quark masses used in this work are often of the order of the lattice UV cutoff, i.e. $aM \sim 1$, and that the direct connection between heavy-light meson masses and quark masses depends on the application of a particular resummed perturbative relation at relatively low energy scales. Despite these caveats, it is clear that this work has looked into the future by simulating relativistic heavy quarks close to the b meson mass.

3.1.7 Conclusions

We conclude this section by summarising briefly the status of the determinations of the fundamental parameters of the SM from lattice QCD.

With the advent of dynamical quark simulations and new methods for non-perturbative renormalization, lattice QCD determinations of the strong coupling and quark masses have become both very accurate and very precise. Even if numerical simulations do not qualify as a proof, many of us believe that these computations have fulfilled the dream of connecting the fundamental quark masses and strong coupling to the well measured spectra of hadrons from first principles.

There are two challenges that lattice QCD computations face in this game. On one hand the strong coupling and quark masses are useful when quoted in the $\overline{\rm MS}$-scheme, requiring to make contact with perturbation theory while most lattice simulations are performed to explore hadronic low energy scales. On the other hand experimental input (hadron masses) have electromagnetic and strong isospin breaking corrections, while most lattice QCD simulations are performed in the isospin symmetric limit.

The window problem

Connecting the perturbative and hadronic regimes of QCD is hard. These two scales are separated by a large gap in energy scales, due to the logarithmic running of the strong coupling with the renormalization scale. It is very challenging to accommodate these disparate scales in a single lattice simulation, and if one insists on doing so, compromises have to be made and perturbation theory has to be used at a few GeV.

Isospin breaking corrections

The simulation of electromagnetism on the lattice poses its own challenges (see [290] for a review), related to the description of charged states in presence of long range interactions. The simulation of non-degenerate light quarks is also numerically challenging. These facts explain that most lattice computations are performed in the isospin symmetric limit.

The lattice community has made great progress on these fronts in recent years. The window problem has a known solution since the early 1990s: finite size scaling [257]. It has been applied to $N_{\rm f} = 0, 2, 3, 4$ QCD [291–293] and to the determination of quark masses [294–296], but these determinations traditionally produced results for the strong coupling with large statistical uncertainties. Thanks to recent developments [297], finite size scaling studies can achieve a subpercent level of precision in the strong coupling [298]. These techniques have also been applied to the determination of quark masses [270,296,299]. Finite size scaling has been for a long time the only solution to the window problem, until a new method based on decoupling of heavy quarks has been proposed [258]. This new method largely reduces the window problem and recent results show that the strong coupling can also be determined using these techniques with a sub-percent precision [261]. This strategy has not yet been applied to the determination of quark masses, but the method should also lead to precise determinations of the running of quark masses.

With the advent of dynamical fermion simulations the precision of lattice determinations of quark masses has rapidly reached a very mature status. Renormalization is nowadays performed in a fully non-perturbative way, and using different strategies. Although contact with perturbation theory has to be made, and in principle there is also a window problem present in the extraction of quark masses, perturbative uncertainties in this case seem to be much better behaved than in the case of extractions of the strong coupling. All in all, at the current level of precision the presence of electromagnetism and strong isospin in nature is the main factor limiting the precision of many lattice computations. But the field evolves very quickly and there exist several lattice computations of the individual light quark masses m_u, m_d that directly com-

Table 1 FLAG averages of the RGI quark masses in MeV for the u, d, s, c and b quarks (see [256]). Several works contribute to these averages [264,264,268–273,273,273–275,275,275–287] computed with either $N_f = 2+1$ or $N_f = 2+1+1$ lattice simulations with about a percent precision for all different quark masses

	$M_u^{\rm RGI}$	$M_d^{\rm RGI}$	$M_s^{\rm RGI}$	$M_c^{\rm RGI}$	$M_b^{\rm RGI}$	[MeV]
$N_f = 2+1$	3.15(13)	6.49(14)	128(2)	1526(17)	6881(63)	
$N_f = 2+1+1$	2.97(11)	6.53(11)	129.7(1.5)	1520(22)	6934(58)	

pute the QED effects in the quenched approximation. We are convinced that unquenched results will follow soon, and isospin breaking corrections will be applied to the determinations of all quark masses.

Only 15 years after the first lattice QCD simulations with dynamical quarks, lattice QCD has been able to determine from first principles the strong coupling with a 0.7% error. Quark masses are determined with a percent error (see Table 1), and soon these computations will include full isospin breaking corrections. The implications of these calculations are far reaching in constraining the SM description of physical phenomena. Lattice determinations of α_s are the most precise (see the next section). The FLAG average based on $N_f = 2+1+1$ simulations of the u quark mass is $M_u^{\rm RGI} = 2.97(11)$ MeV (based on the works [269,287], see also [268]). This value, derived from first principles of QCD, disfavors a popular solution to the strong CP problem (a massless u quark) by 30 standard deviations.

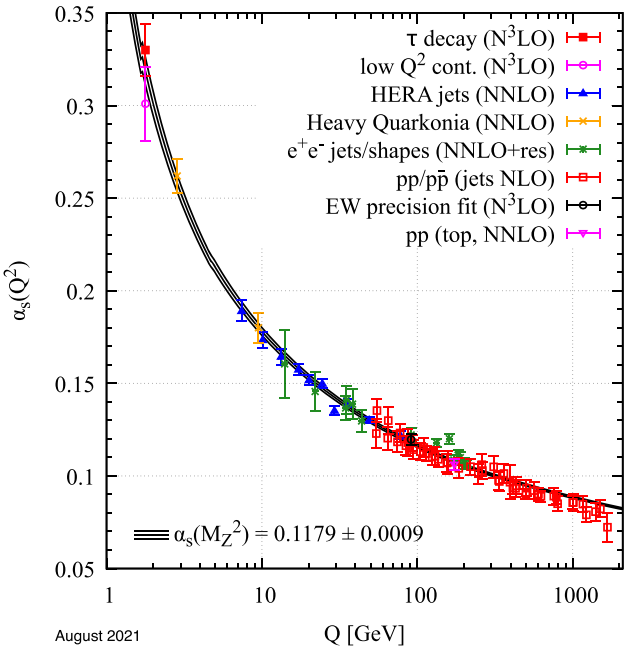


Fig. 14 Measurements of the coupling constant α_s, as a function of the energy scale Q. The level of precision of the perturbative prediction used in the measurement of α_s is indicated in brackets (*NLO* next-to-leading order, *NNLO* next-to-next-to-leading order, *NNLO+res.* NNLO matched to a resummed calculation, N^3LO next-to-next-to-leading order). Figure taken from Ref. [300]

3.2 The strong-interaction coupling constant

Giulia Zanderighi

3.2.1 The world average determination of α_s

We summarize here the current procedure used in the PDG [300] to obtain the world average value of $\alpha_s(M_Z^2)$ and its uncertainty, and we discuss future prospects for its improvement.

Preliminary considerations

All observables involving the strong interaction depend on the value of the strong coupling constant. This implies that a number of different observables can be used to determine the coupling constant, provided that a suitable theoretical prediction is available for that observable. Figure 14 presents values for α_s derived from different observables. The following considerations are used to assess if a particular observable is suitable for use in the determination of the strong coupling constant:

– The observable's sensitivity to α_s as compared to the experimental precision. For example, for the e^+e^- cross section to hadrons (e.g. the R ratio), QCD effects are only a small correction, since the perturbative series starts at order α_s^0, but the experimental precision is high. Three-jet production, or event shapes, in e^+e^- annihilation are directly sensitive to α_s since they start at order α_s. Four- and five-jet cross-sections start at α_s^2 and α_s^3 respectively, and hence are very sensitive to α_s. However, the precision of the measurements deteriorates as the number of jets involved increases.
– The accuracy of the perturbative prediction, or equivalently of the relation between α_s and the value of the observable.

The minimal requirement is generally considered to be an NLO prediction. The PDG imposes now that at least NNLO accurate predictions be available. In certain cases where phase space restrictions require it, fixed-order predictions are supplemented with resummation. An improved perturbative accuracy is necessary to guar-

antee that the theoretical uncertainty is assessed in a robust way.
- The size of non-perturbative effects. Sufficiently inclusive quantities, like the e^+e^- cross section to hadrons, have small non-perturbative contributions $\sim \Lambda^4/Q^4$. Other quantities, such as event-shape distributions, typically have contributions $\sim \Lambda/Q$. All other aspects being equivalent, observables with smaller non-perturbative corrections are preferable.
- The scale at which the measurement is performed. An uncertainty δ on a measurement of $\alpha_s(Q^2)$, at a scale Q, translates to an uncertainty $\delta' = \alpha_s^2(M_Z^2)/\,\alpha_s^2(Q^2) \times \delta$ on $\alpha_s(M_Z^2)$. For example, this enhances the already important impact of precise low-Q measurements, such as from τ decays, in combinations performed at the M_Z scale.

The PDG determination of α_s first separates measurements into a number of different categories, then calculates an average for each category. This average is then used as an input to the world average. The PDG procedure requires:

1. a specification of the conditions that a determination of α_s should fulfill in order to be included in the average;
2. a specification of the separations of the different extractions of $\alpha_s(M_Z^2)$ into the separate categories;
3. a specification of the procedure within each category to compute the average and its uncertainty;
4. a specification of the manner in which the different subaverages and their uncertainties are combined to determine the final value of $\alpha_s(M_Z^2)$ and its uncertainty.

Details of the PDG averaging procedure
In the following, we summarize the procedure adopted in the last edition of the PDG [300]. There, the selection of results from which to determine the world average value of $\alpha_s(M_Z^2)$ is restricted to those that satisfy a well defined set of criteria. These are that the fit should be

1. accompanied by reliable estimates of all experimental and theoretical uncertainties;
2. based on the most complete perturbative QCD predictions of at least next-to-next-to leading order (NNLO) accuracy;
3. published in a peer-reviewed journal at the time of writing of the PDG report.

Note that the second condition to some extent follows from the first. In fact, determinations of the strong coupling from observables in e^+e^- involving e.g. five or more jets are very sensitive to α_s, and could provide additional constraints. However, these observables are currently described only at leading order (LO) or next-to-leading order (NLO), and the determination of the theoretical uncertainty is thus considered not sufficiently robust. It is also important to note that some determinations are included in the PDG, but the uncertainty quoted in the relevant publications is increased by the PDG authors to fulfill the first condition. Similarly, in some cases the central value used in the PDG differs from the one quoted in some publications, but can be extracted from the analysis performed in that work.

Categories of observables
All observables used in the determination of $\alpha_s(M_Z^2)$ in the PDG averaging procedure are classified in the following categories

- "Hadronic τ decays and low Q^2 continuum" (τ decays and low Q^2): the coupling constant is here determined at the τ mass, therefore once it is evolved up to the Z mass the uncertainty shrinks. Perturbative calculations for τ decays are available at N^{3}LO, however there are different approaches to treat the perturbative and non-perturbative contributions that result in significant differences. These discrepancies are currently the limiting factor in reducing the uncertainty in this category.
- "Heavy quarkonia decays" ($Q\bar{Q}$ bound states): calculations are available at NNLO and N^{3}LO.
- "PDF fits" (PDF fits): this category includes both global PDF fits and analyses of singlet and non-singlet structure functions. To quantify the theory uncertainty, half of the difference between results obtained with NNLO and NLO predictions is added in quadrature.
- "Hadronic final states of e^+e^- annihilations" (e^+e^- jets and shapes): these fits use measurements at PETRA and LEP. Non-perturbative corrections are important, going as Λ/Q and can be estimated either via Monte Carlo simulations or analytic modeling.
- "Observables from hadron-induced collisions" (had-ron colliders): NNLO calculations for $t\bar{t}$ or jet production at both the LHC and HERA, and Z+jet production at the LHC have allowed measurements for these processes to be used in α_s determinations. An important open question is whether a simultaneous PDF and α_s fit has to be carried out in order to avoid a potential bias.
- "Electroweak precision fit" (electroweak): α_s determinations are averaged from electroweak fits to data from the Tevatron, LHC, LEP and the SLC. These fits rely on the strict validity of the Standard Model.
- "Lattice": the average determined by the FLAG group in 2019 [301] from an input of 8 determinations was used in the last PDG determination; the subsequent 2021 α_s average is very consistent with that of 2019.

Detailed information about which observables are included in the different categories can be found in Ref. [300].

Table 2 PDG average of the categories of observables. These are the final input to the world average of α_s

Category	$\alpha_s(M_Z^2)$
τ decays and low Q^2	0.1178 ± 0.0019
$Q\bar{Q}$ bound states	0.1181 ± 0.0037
PDF fits	0.1162 ± 0.0020
e^+e^- jets and shapes	0.1171 ± 0.0031
Hadron colliders	0.1165 ± 0.0028
Electroweak	0.1208 ± 0.0028
Lattice	0.1182 ± 0.0008

Average and uncertainty in each category

In order to calculate the world average value of $\alpha_s(M_Z^2)$, a preliminary step of pre-averaging results within each category listed in Sect. 3.2.1 is carried out. For each sub-field, except for the "Lattice" category, the *unweighted average* of all selected results is taken as the pre-average value of $\alpha_s(M_Z^2)$, and the unweighted average of the quoted uncertainties is assigned to be the respective overall error of this pre-average. An unweighted average is used to avoid the situation in which individual measurements, which may be in tension with other measurements and may have underestimated uncertainties, can considerably affect the determination of the strong coupling in a given category. As an example, the determination of $\alpha_s(M_Z^2)$ from e^+e^- jets and shapes currently averages ten determinations and arrives at $\alpha_s(M_Z^2) = 0.1171 \pm 0.0031$. Since two determinations [302,303], both based on a similar theoretical framework, arrive at a small value of $\alpha_s(M_Z^2)$ and have a very small uncertainty, if one were to perform a weighted average one would arrive at $\alpha_s(M_Z^2)$ from e^+e^- jets and shapes of $\alpha_s(M_Z^2) = 0.1155 \pm 0.0006$, which is not compatible with the current world average. This would, in fact, considerably change the world average because of the very small uncertainties. The current procedure is instead robust against $\alpha_s(M_Z^2)$ determinations that are outliers with small uncertainties as compared to the other determinations in the same category. For the "Lattice QCD" (lattice) sub-field, the PDG adopts the LAG2019 average value and uncertainty for this sub-field [301]. FLAG2019 also requires strict conditions on its own for a determination to be included in their average, which are in line with those used in the PDG. The results of the averages of the categories are given in Table 2. From the table, it is clear that determinations from different categories are compatible with each other and accordingly can be combined to give rise to a final average.

Final average

Since the six sub-fields (excluding lattice) are largely independent of each other, the PDG determines a non-lattice world average value using a standard 'χ^2 *averaging*' method. This results in the final average of the six categories of

$$\alpha_s(M_Z^2) = 0.1175 \pm 0.0010, \quad \text{(without lattice)}, \quad (3.42)$$

which is fully compatible with the lattice determination. In a last step the PDG performs an unweighted average of the values and uncertainties of $\alpha_s(M_Z^2)$ from the non-lattice result and the lattice result presented in the FLAG2019 report, which results in the final average of

$$\alpha_s(M_Z^2) = 0.1179 \pm 0.0009, \quad \text{(final average)}. \quad (3.43)$$

Performing a weighted average of all seven categories would instead give rise to $\alpha_s(M_Z^2) = 0.1180 \pm 0.0006$. The PDG uncertainty is instead more conservative and about 50% larger. These final results are summarized in Fig. 15.

3.2.2 Outlook

Despite the numerous determinations of the strong coupling constant, it remains to date the least well-known gauge coupling, with an uncertainty of about 1%. Still it is a remarkable success that all determinations from all categories agree well with each other, all within about one sigma. Future improvements are likely to be driven by those categories which today have the smallest uncertainties, i.e. lattice determinations, τ decays and low Q^2 measurements.

As far as the category "τ decays and low Q^2 measurements" are concerned, it is important to mention that the uncertainty quoted in the latter category includes the difference in the extractions that are obtained using contour improved perturbation theory (referred to as CIPT) and fixed order perturbation theory (FOPT). Recent arguments suggest that FOPT are to be preferred, see also dedicated discussions on this point in Ref. [304]. If this is confirmed, the value of $\alpha_s(M_Z^2)$ in this category would shift slightly to lower values, and would allow one to quote a reduced theoretical uncertainty since this additional source of uncertainty would be completely removed. Further improvements could also come from a better understanding of non-perturbative effects.

Important progress is also expected in the category "e^+e^- jets and shapes", where the calculation of power corrections in the 3-jet region [180,305] could have a sizeable impact, and improve fits of the coupling from event shapes. In fact, in current determinations that rely on an analytic computation of non-perturbative power corrections, these calculations are performed in the two-jet limit and applied to the kinematic region used in the fits where events typically have an additional hard emission, i.e. to three-jet configurations. A treatment of these corrections in the three-jet region is now possible, at least for some observables and the impact of this improved treatment of non-perturbative effects on $\alpha_s(M_Z^2)$ in this category is eagerly awaited.

As far as the hadron collider category is concerned, it is an open question if it is always preferred to fit $\alpha_s(M_Z^2)$ and

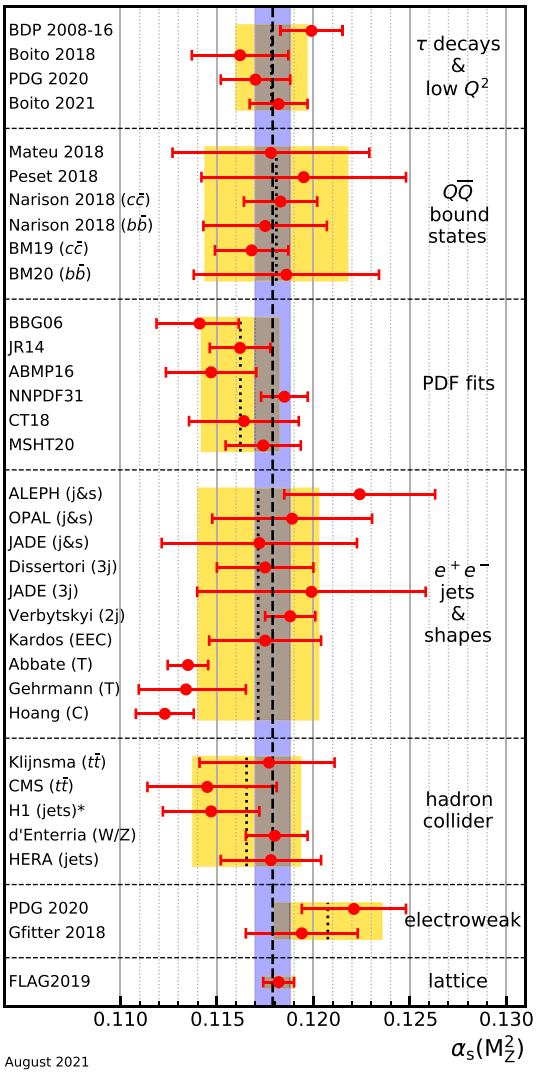


Fig. 15 Summary of the determinations of $\alpha_s(M_Z^2)$ from the seven sub-fields used in the PDG [300], as discussed in the text. The yellow (light shaded) bands and dotted lines indicate the pre-average values of each sub-field. The dashed line and blue (dark shaded) band represent the final world average value of $\alpha_s(M_Z^2)$. The '*' symbol within the "hadron colliders" sub-field indicates a determination including a simultaneous fit of parton distribution functions. All other "hadron collider" determinations instead use a set of parton distribution functions as input to the fit. Figure taken from Ref. [300]

the parton distribution functions simultaneously and how to best deal with correlations between the parton distribution function parameters and $\alpha_s(M_Z^2)$ in the cases where the fit is not performed simultaneously. In view of many more NNLO results to come and many more data from the LHC, we can expect theoretical work and advances in addressing this question. Ratios of cross sections are less sensitive to parton distribution functions and therefore could be considered more suitable to extract α_s. For instance, NNLO predictions for 3-jet production will enable to perform fits of $\alpha_s(M_Z^2)$ from ratios with at least partial cancellation of some uncertainties. It is not clear whether this reduction in uncertainty also holds for the PDF dependence of such ratio predictions. Moreover, for predictions of ratios of cross sections, the natural central scale choice in numerator and denominator are in general not the same. Data from the "hadron collider category" at high Q^2 will also be crucial to test the running of the coupling to highest energy scales. Such tests are important since heavy states that couple strongly could modify the running of α_s at high Q^2.

Finally, it is important to mention that the recent years have seen remarkable advances in the determination of $\alpha_s(M_Z^2)$ from lattice calculations, also thanks to the FLAG effort which imposes strict quality criteria for lattice determinations to be included in the FLAG average. This is now the single most precise result of all categories included in the PDG and agrees remarkably well (both in terms of central value and uncertainty) with the PDG world average of α_s without lattice data. Further improvements from lattice calculations are also expected in the coming decade. Given all the progress to be expected in the coming years in various aspects and categories, a determination of α_s with sub-percent precision seems finally within reach.

4 Lattice QCD

Conveners:
Kostas Orginos and Franz Gross

The previous sections have shown how early measurements, phenomenology, and theoretical arguments lead to the discovery of the QCD Lagrangian, and how the parameters in the Lagrangian, the QCD fine structure constant α_s and the quark masses m_h, could be fixed from experiment.

With this section, we begin a systematic study of the theory (and provide the discussion needed to understand Sect. 3). Since α_s is not small at energy scales appropriate to the study of cool nuclear matter, a non-perturbative method is needed. Lattice QCD (LQCD) is currently the only known way to obtain accurate, non-perturbative QCD predictions. Since this method is both complicated and not well covered in most textbooks, this section presents a detailed, systematic study of LQCD.

LQCD began in 1974, shortly after Quantum Chromodynamics was established, when Kenneth Wilson published a seminal paper in which he formulated the theory on a space-time lattice. This formulation had profound implications. It preserved the gauge invariance of the theory while regulating ultraviolet divergences and providing a definition of QCD as the continuum limit of the lattice theory. However, one may argue that the most crucial implication was the fact that it offered a pathway to non-perturbative computations. Quantities such as the spectrum of stable hadrons, decay constants, and Parton distribution functions to name a few, could now

in principle be computed from the fundamental theory without the need for uncontrolled approximations. In the beginning, however, this formulation of QCD lent itself to a different type of analytic computations such as the strong coupling expansion where Wilson showed that color charges are indeed confined in the strong coupling limit.

Numerical investigations of Lattice QCD (LQCD) started a few years later with the pioneering work of Michael Creutz in 1980. There for the first time, the SU(2) pure Yang–Mills theory was investigated using Monte Carlo methods. Subsequently, many groups around the world started studying Lattice QCD, developed methods and algorithms, and investigated the efficacy of the available computer hardware for numerical calculations in LQCD. Although the fundamental principles of such calculations were clear, it was evident from the beginning that the computational cost for achieving phenomenologically relevant results was enormous. In addition, the limitations of Euclidean time formulation as well as the computational limitations imposed by finite volume and lattice spacing made it clear that computational power alone will not be enough. Therefore, intense theoretical research to develop methods and algorithms started in the 1980s. Together with that effort, many groups devoted efforts to designing custom-made supercomputers that were best suited for the problem at hand. The idea of a massively parallel computer to solve scientific problems seemed at odds at the time with the vector machines that defined the commercially available high-performance computers. Yet in the 1990s the rise of massively parallel computers, commercial or custom-made, led to major advances in LQCD. The new century brought a combination of powerful supercomputers, sophisticated numerical techniques, and advanced theoretical approaches that allowed for the first time to compute physical quantities at phenomenologically relevant accuracy.

Lattice QCD is now an established field that can provide results at unprecedented accuracy and can help move forward our fundamental understanding of particle physics. The impact of lattice QCD computations on strong interaction physics is evident throughout this volume. Nearly every section contains references to landmark lattice QCD computations. In this section, a brief introduction to the formulation of lattice QCD is given by Gottlieb, followed by De Tar's review of the basic LQCD algorithms. Leinweber discusses the structure of the QCD vacuum as it emerges from numerical experiments. Karsch reviews computations at non-zero temperatures and densities relevant to understanding quark–gluon plasma physics.

The discussion then continues with a focus on applications. Dudek reviews hadron spectroscopy with emphasis on finite volume methods that allow for the extraction of scattering amplitudes from Euclidean time correlation functions. Constantinou/Orginos discuss computations of the nucleon structure including modern approaches that allow for the extraction of momentum-fraction-dependent distributions from Euclidean time computations. Finally, Davies reviews computations for Weak matrix element computations which play a central role in the experimental program for probing physics beyond the standard model (BSM).

4.1 Lattice field theory

Steven Gottlieb

4.1.1 Introduction

In perturbative quantum field theories loop integrals lead to infinities. To deal with these infinities, a regularization scheme must be introduced. Examples of regularization schemes are Pauli–Villars modification of particle propagators and dimensional regularization in which the number of space-time dimensions of the system becomes a variable. After regularization, calculations no longer suffer from infinities, but they do depend on a new parameter specific to the regularization scheme, e.g., Λ a large mass in the Pauli–Villars scheme, or $\epsilon = 4 - d$ in the case of dimensional regularization. Since physical results should be independent of the regularization scheme, a renormalization procedure is introduced so that the so-called bare parameters of the theory depend on Λ or ϵ in such a way that physical observables do not as there is a cancellation between the regularization dependence of the bare parameters and those of the loop integrals.

In lattice field theories (LFTs), the theory is modified so that (in finite volume) there are no longer an infinite number of degrees of freedom. For instance, in a scalar field theory instead of a real or complex value of the field at each of the infinite points of space time, there are only a finite number of real or complex degrees of freedom defined on a hypercubic grid of space time points. In this case, the parameter that characterizes the regulator is the distance between the space time points called the lattice spacing, usually denoted a. Usually, periodic boundary conditions in space and antiperiodic boundary conditions in time are used. As we will see in more detail below, the field can be Fourier transformed and in momentum space there is a maximum momentum as each component of the momentum is in the range $-\pi/a < p_i \leq \pi/a$. In a finite volume, there is also a minimum spacing between allowable momenta components that serves as an infrared regulator. To summarize, the lattice field theory regularizes the theory by introducing a maximum momentum, and the renormalization program is implemented by requiring that physical quantities be independent of the lattice spacing as $a \to 0$. Also, since the lattice theory only has hypercubic and not full rotational symmetry, we must demonstrate that the latter is restored for distances much larger than a.

Actions for a free scalar theory

To see how LFTs work, let's start with a free scalar field theory in the continuum, transform it to a Euclidean field theory and then put it on a lattice. Start with the Lagrangian density

$$\mathcal{L}(x) = \frac{1}{2}[\partial_\mu \phi(x)\partial^\mu \phi(x) - m^2\phi(x)^2] \tag{4.1}$$

and the action

$$\begin{aligned} S &= \int dt L = \int dt \int d^3x \mathcal{L}(x) \\ &= \int d^4x \frac{1}{2}[\partial_\mu \phi(x)\partial^\mu \phi(x) - m^2\phi(x)^2] \\ &= \int d^4x \frac{1}{2}[\partial_t \phi(x)\partial_t \phi(x) \\ &\quad - \nabla\phi(x)\cdot\nabla\phi(x) - m^2\phi(x)^2] \end{aligned} \tag{4.2}$$

where $\phi(x)$ is the scalar field and m is its mass. The Feynman path integral is defined as

$$Z = \int [d\phi] \exp\{iS\}, \tag{4.3}$$

where $[d\phi]$ denotes the integration measure of all possible fields $\phi(x)$. To Euclideanize the theory let $t \to -i\tau$ which changes the sign of the time derivative term in the Lagrangian density. It also adds a factor $-i$ because of the change of integration variable in the action. So, the Euclidean action is defined to be

$$\begin{aligned} S_E = \int d^3x d\tau \frac{1}{2}[\partial_\tau \phi(x)\partial_\tau \phi(x) \\ + \nabla\phi(x)\cdot\nabla\phi(x) + m^2\phi(x)^2], \end{aligned} \tag{4.4}$$

and the path integral becomes

$$Z = \int [d\phi] exp\{-S_E\}. \tag{4.5}$$

At this point, it is traditional to rename τ to t, the time variable with which we started, or let $\tau = x_4$. In any case, the field ϕ is defined on a 4-dimension Euclidean domain, S_E is positive definite, and this looks like a partition function of a statistical mechanical system. The transformation to Euclidean time allows us to use the importance sampling techniques of statistical mechanics (Monte Carlo methods) introduced in the next section.

To convert to a lattice theory, introduce a spacing a between the points of a hypercubic grid, so the lattice field ϕ_n is defined on a discrete set of points $n = (n_1, n_2, n_3, n_4)$ in R^4 and $x = an$. Typically, work is done in a finite volume so that n_i is an integer between 0 and $N_i - 1$, where N_i is the extent of the lattice in the i-th direction. The derivatives must be replaced by a finite difference approximation. There is more than one way to do this. Pretending for the moment that ϕ depends only on a single variable x, a forward difference is defined by

$$\Delta^+\phi(x) = \frac{\phi(x+a) - \phi(x)}{a}. \tag{4.6}$$

Taylor expanding $\phi(x+a)$ gives

$$\Delta^+\phi(x) = \phi'(x) + \frac{a}{2}\phi''(x) + \cdots. \tag{4.7}$$

Note that the symmetric finite difference operator

$$\Delta^S\phi(x) = \frac{\phi(x+a) - \phi(x-a)}{2a} \tag{4.8}$$

$$= \phi'(x) + \frac{a^2}{6}\phi'''(x) + \cdots \tag{4.9}$$

is a much better approximation of the continuum derivative since the correction is second order in the small lattice spacing a.

Actions for a gauge invariant scalar theory with a ϕ^4-type interaction

To introduce gauge invariance, change the real scalar field to a complex field, and introduce a ϕ^4-type interaction term

$$S = \int d^4x[\partial_\mu \phi^*(x)\partial^\mu \phi(x) - m^2\phi^*(x)\phi(x) \tag{4.10}$$

$$- \lambda(\phi^*(x)\phi(x))^2]. \tag{4.11}$$

A global gauge transformation is just a change $\phi \to \phi' = \Omega\phi$ where Ω is complex phase factor, $\Omega = \exp i\theta$, with θ a real number independent of x. The action is clearly invariant under this gauge transformation since $(\phi')^* = \Omega^*\phi^*$ and for every factor of Ω coming from transforming ϕ, there is a corresponding factor of Ω^* from transforming ϕ^*. A cubic term in the action would break this gauge invariance.

To generalize to local gauge invariance, allow θ to become a function of x. The mass and interaction terms are clearly still invariant because they only depend on x. However, the first term with derivatives transforms in a non-trivial way.

$$\begin{aligned} \partial_\mu\phi'(x) &= \partial_\mu(\Omega(x)\phi(x)) \\ &= (\partial_\mu\Omega(x))\phi(x) + \Omega(x)(\partial_\mu\phi(x)). \end{aligned} \tag{4.12}$$

To handle the extra term depending on $\partial_\mu\Omega(x)$, define a *covariant derivative* D_μ that has the property

$$D'_\mu\phi'(x) = \Omega(x)D_\mu\phi(x), \tag{4.13}$$

so that the covariant derivative $D_\mu\phi(x)$ transforms under a gauge transformation the same way that $\phi(x)$ does. To accomplish this, introduce a vector field $A_\mu(x)$, and define the covariant derivative to be

$$D_\mu = \partial_\mu + ieA_\mu. \tag{4.14}$$

Using this definition in (4.13) gives the constraint

$$(\partial_\mu + ieA'_\mu)(\Omega(x)\phi(x)) = \Omega(x)(\partial_\mu + ieA_\mu)\phi(x). \tag{4.15}$$

Requiring that this hold for any field $\phi(x)$ gives the gauge transformation for the field A_μ

$$A'_\mu \Omega = \Omega A_\mu + \frac{i}{e}\partial_\mu \Omega. \tag{4.16}$$

This derivation has preserved the order of the terms, so that this equation will hold even for non-Abelian theories in which Ω is a matrix. Solving for A'_μ in this most general case gives

$$A'_\mu = \Omega A_\mu \Omega^{-1} + \frac{i}{e}(\partial_\mu \Omega)\Omega^{-1}. \tag{4.17}$$

For the Abelian theory, this reduces to

$$A'_\mu = A_\mu - \frac{1}{e}\partial_\mu \theta. \tag{4.18}$$

This all works out very nicely in the continuum theory. Wilson's brilliant insight [97] was to define the lattice theory not with variables from the gauge algebra, but with variables that are elements of the gauge group, denoted $U(n, m)$. These are called *link variables*. or parallel transporters because they allow the comparison of a field at one point on the lattice with a neighboring point in a gauge covariant way. If $U(n, m)$ is associated with the link connecting nearest neighbor points n and m, then

$$U(m, n) = U^\dagger(n, m) = U^{-1}(n, m) \tag{4.19}$$

where the second identity follows from the fact that U is a unitary matrix. So, defining $U_{n\mu} = U(n, n + \hat{\mu})$, Eq. (4.19) shows that $U(n + \hat{\mu}, n) = U^\dagger_{n\mu}$.

We want the product $U_{n\mu}\phi_{n+\hat{\mu}}$ to transform under a gauge transformation the same way that the field does at the point n. In other words, under a gauge transformation $U \to U'$ and $\phi_n \to \phi'_n = \Omega_n \phi_n$, so we must have

$$U'_{n\mu}\phi'_{n+\hat{\mu}} = \Omega_n U_{n\mu}\phi_{n+\hat{\mu}}. \tag{4.20}$$

Since $\phi'_{n+\hat{\mu}} = \Omega_{n+\hat{\mu}}\phi_{n+\hat{\mu}}$, this implies

$$U'_{n\mu} = \Omega_n U_{n\mu}\Omega^{-1}_{n+\hat{\mu}}. \tag{4.21}$$

Hence, the products of link variables along a path transform as Ω_n if the left-most point is n and Ω^{-1}_m, if the right-most point is m. With suitable products of link variables, we can transport a field as far as we wish and have it transform as a variable that 'lives' at the left-most point in the product.

The difference $U_{n\mu}\phi_{n+\hat{\mu}} - \phi_n$ transforms in a gauge covariant way, since under a gauge transformation it picks up a factor of Ω_n. The relationship between the group element $U_{n\mu}$ and the gauge field $A_\mu(x)$ that takes a value in the Lie algebra is

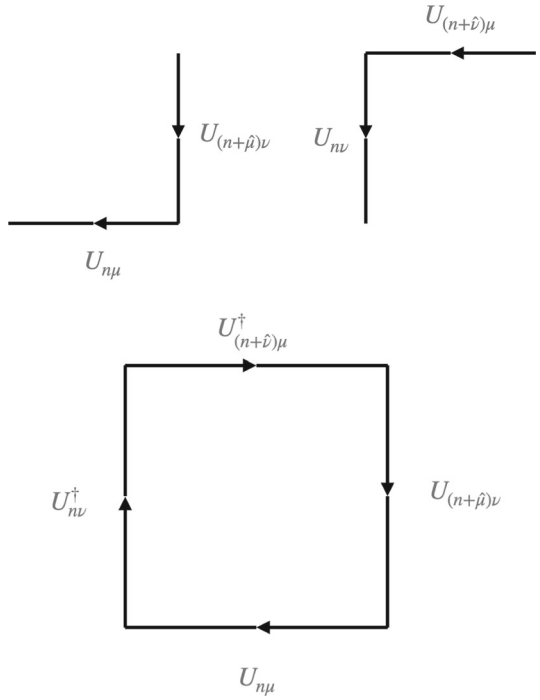


Fig. 16 Top: The two paths that contribute to $[D_\mu, D_\nu]$. The μ-direction is the horizontal axis, and ν is vertical. On the left, we have $D_\mu D_\nu$, on the right $D_\nu D_\mu$ Bottom: The links of a plaquette whose lower-left corner is site n, with directions same as in top figure

$$\begin{aligned} U_{n\mu} &= \mathcal{P}\exp\left\{ie\int_{an}^{an+a\hat{\mu}} dy_\nu\, A_\nu(y)\right\} \\ &= \exp\left\{iea\Big[A_\mu(an + a\hat{\mu}/2) + \frac{a^2}{24}\partial^2_\mu A_\mu(an + a\hat{\mu}/2) \right. \right. \\ &\quad \left. \left. + \cdots\Big]\right\} = 1 + iaeA_\mu(an + a\hat{\mu}/2) + \cdots. \end{aligned} \tag{4.22}$$

where the lattice spacing a is shown explicitly, and it is natural to relate the link variable U to the (continuum) gauge field at the midpoint of the link. Note that in Wilson's original work, the lattice gauge field variables are $A_{n\mu}$, i.e., they are labeled by the left-hand site of the link.

Having defined the covariant derivative, the field strength tensor can be calculated. In the continuum:

$$F_{\mu\nu} = ie[D_\mu, D_\nu] \tag{4.23}$$

where the square brackets denote the commutator. This formula also holds in the case of non-Abelian gauge theory for which $A_\mu(x)$ is a matrix in the Lie algebra of the gauge group.

On the lattice, the covariant derivative involves parallel transport from a neighboring site (Fig. 16). Since there are two covariant derivatives a field is transported from two sites away:

$$\begin{aligned} D_\mu D_\nu \phi_n = U_{n\mu}(U_{(n+\hat{\mu})\nu}\phi_{n+\hat{\mu}+\hat{\nu}} \\ -\phi_{n+\hat{\mu}}) - U_{n\nu}\phi_{n+\hat{\nu}} - \phi_n \end{aligned}$$

$$= U_{n\mu}U_{(n+\hat{\mu})\nu}\phi_{n+\hat{\mu}+\hat{\nu}} - U_{n\mu}\phi_{n+\hat{\mu}} - U_{n\nu}\phi_{n+\hat{\nu}} - \phi_n. \quad (4.24)$$

The last three terms are symmetric under the interchange of μ and ν, so only the first term contributes to the commutator. Thus, the field strength tensor is the difference between the product of the two two-link paths connecting sites n and $n+\hat{\mu}+\hat{\nu}$. In one path we move in the μ direction first and in the other we move in the ν direction first. Also, the field strength tensor is gauge covariant as the common endpoints of the two paths determine how $F_{n\mu\nu}$ transforms:

$$F'_{n\mu\nu} = \Omega_n F_{n\mu\nu}\Omega^{-1}_{n+\hat{\mu}+\hat{\nu}}. \quad (4.25)$$

In the continuum the gauge action is proportional to $F^2_{\mu\nu}$, and is gauge invariant. Having just determined that $F_{\mu\nu}$ can be expressed in terms of a two-link path, we might expect that a four-link path would yield $F^2_{\mu\nu}$. It is easy to construct gauge invariant products of links. If we take the trace of the product of links along any closed path, it will be gauge invariant. The only closed four-link paths are those around the elementary squares of the lattice. The term plaquette is sometimes used to refer to the elementary squares of a hypercubic lattice. The plaquette is also used to refer to the product of the four link matrices around the square, or to the trace of this matrix. The context should make clear whether the author is referring to a shape, a matrix, or a number. Here the plaquette $U_{n\mu\nu}$ will be the trace of the product of the four links

$$U_{n\mu\nu} = \mathrm{Tr}(U_{n\mu}U_{(n+\hat{\mu})\nu}U^\dagger_{(n+\hat{\nu})\mu}U^\dagger_{n\nu}) \quad (4.26)$$

The Wilson plaquette gauge action is defined as

$$S_W = \frac{1}{g^2}\sum_n\sum_{\mu\neq\nu}(3 - \mathrm{Re}\, U_{n\mu\nu}). \quad (4.27)$$

Actions for fermions

In the continuum, the free fermion action S_F is given by:

$$S_F = \int d^4x\bar{\psi}(x)(i\gamma^\mu\partial_\mu - m)\psi(x), \quad (4.28)$$

where the gamma matrices obey $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$. Going through the transformation to Euclidean space time, we introduce the Euclidean gamma matrices $\gamma^E_4 = \gamma^0$ and $\gamma^E_i = -i\gamma^i$. These gamma matrices obey $\{\gamma^E_\mu, \gamma^E_\nu\} = 2\delta_{\mu\nu}$. The Euclidean action is given by

$$S^E_F = \int d^4x\bar{\psi}(x)(\gamma^E_\mu\partial_\mu + m)\psi(x). \quad (4.29)$$

We simplify notation below by dropping the superscript E on the Euclidean gamma matrices. To include the interaction with the gauge field, the ordinary partial derivative in Eq. (4.28) is replaced by the covariant derivative. For S^E_F, a gauge covariant finite difference approximation is used

$$\partial_\mu\psi(x) \to \frac{1}{2a}(U_{n\mu}\psi_{n+\hat{\mu}} - U^\dagger_{(n-\hat{\mu})\mu}\psi_{n-\hat{\mu}}) \quad (4.30)$$

which is the analog of Δ^S introduced in Eq. (4.8). This action is called the naive fermion action, and we are about to see that it suffers from the so-called "fermion doubling problem."

To explore this, consider the case of a free fermion, so the link variables may be replaced by the unit matrix. Going to momentum space, let

$$\psi_n = \sum_p e^{(iap\cdot n)}\psi(p). \quad (4.31)$$

On the lattice there is maximum value for each momentum component because if $ap_\mu = 2\pi$ then the exponential will always be always be the same as for $p_\mu = 0$. Thus, the momentum components can be restricted to be less than $(2\pi)/a$ or more symmetrically,

$$-\frac{\pi}{a} < p_\mu \leq \frac{\pi}{a}. \quad (4.32)$$

Because of the periodic boundary conditions on a lattice of finite extent, say L in each direction, there is another restriction that $ap_\mu L = 2\pi j$ for some integer j. Thus the allowable momentum components are restricted to $(2\pi j)/(aL)$, so for finite L the lattice provides an infrared as well as an ultraviolet cutoff. However, as L goes to infinity, the momentum becomes a continuous variable, and in this case Eq. (4.31) becomes

$$\psi_n = \int_{-\pi/a}^{\pi/a} d^4p\, e^{(iap\cdot n)}\psi(p). \quad (4.33)$$

The fact that $\bar{\psi}$ and ψ are displaced from each other on the lattice results in factors of $\exp \pm ip_\mu a$. The final result for the Euclidean action, written in momentum space, is

$$S^E_F = \int d^4p\left[\frac{i}{a}\sum_\mu \bar{\psi}(p)\gamma_\mu \sin(p_\mu a)\psi(p) + m\bar{\psi}(p)\psi(p)\right] = \int d^4p\ \bar{\psi}(p)\, S^{-1}(p)\,\psi(p), \quad (4.34)$$

The fermion doubling problem

At this point, most authors go on to solve for the free quark propagator and examine the pole structure. Let's just look at the current expression and compare with the continuum. When $p_\mu a$ is small, we may approximate $\sin(p_\mu a) \to p_\mu a$ so the factor of a^{-1} before the sum is cancelled and this looks a lot like $i\not{p} + m$. As p_μ continues to grow toward $\pi/(2a)$, the sin function flattens out and then starts to return to zero at $p_\mu = \pi/a$. That means at the end of the Brillouin zone, there is again a region were there is linear dependence on the momentum. More concretely, let $p_\mu = \pi/a - k$ and note that $\sin(p_\mu a) = \sin(ka)$. We also need the region where $p_\mu = -\pi/a + k$ to have a region in momentum space just like the one at the origin. Since any component of p can be near zero, or at the edge of the Brillouin zone there are 2^4 regions in momentum space where the action takes the form

of a free action. We wanted one fermion and we wound up with 16! This is the crux of the doubling problem.

In his Erice lectures, Wilson provided a fix [306]. He added to the action a higher dimensional term, the lattice Laplacian, multiplied by the lattice spacing. This term vanishes as $a \to 0$. The covariant version of the second derivative ∇_μ^2 is defined

$$\nabla_\mu^2 \psi_n = \frac{1}{a^2}\Big(U_{n\mu}\psi_{n+\hat{\mu}} + U^\dagger_{(n-\hat{\mu})\mu}\psi_{n-\hat{\mu}} - 2\psi_n\Big). \quad (4.35)$$

The Wilson fermion action is therefore

$$\begin{aligned} S_W^F &= S_{naive} - \frac{ar}{2}\sum_x \bar{\psi}(x)\sum_\mu \nabla_\mu^2\psi(x) \\ &= \bar{\psi} M_W(m)\psi, \end{aligned} \quad (4.36)$$

where r is a free parameter, usually set to $r = 1$, and S_{naive} is given by Eq. (4.29) after substituting Eq. (4.30). Fourier transforming, the free inverse propagator now is

$$aS^{-1}(p) = i\sum_\mu \gamma_\mu \sin(ap_\mu) + am - r\sum_\mu \big(\cos(ap_\mu) - 1\big). \quad (4.37)$$

The last term, proportional to r, vanishes near $p = 0$, but near the edge of the Brillouin zone $\cos(ap_\mu) = -1$ and the doublers, with n momentum components $p_\mu = \pm\pi/a$, now attain masses $m + 2nr/a$, and only one fermion, with $p \approx 0$, remains light. The Wilson term cures the doubling problem, but the action with $m = 0$ no longer has a chiral symmetry so there is an additive mass renormalization, and we must fine tune the parameters to determine where the fermion mass vanishes. The Wilson fermion action has errors $\mathcal{O}(a)$.

An important property of the Wilson Dirac operator is its γ_5 Hermiticity. That is

$$M_W^\dagger(m) = \gamma_5 M_W(m)\gamma_5. \quad (4.38)$$

We will see in the next section that $\det M_W(m)$, the fermion determinant, arises from integrating over the fermion fields. A consequence of γ_5 Hermiticity is that $\det M_W^\dagger(m) = \det M_W(m)$. If a theory has two equal mass fermions, the fermion determinant will be positive (semi-) definite as

$$\det(M_W(m)M_W(m)) = \det(M_W^\dagger(m)M_W(m)). \quad (4.39)$$

In addition to the dimension-5 operator Wilson introduced, there is a second operator introduced by Sheikholeslami and Wohlert [307] that can be adjusted to reduce the error to $\mathcal{O}(a^2)$. The operator is the lattice analog of $\bar{\psi}(x)\sigma_{\mu\nu}F_{\mu\nu}(x)\psi(x)$ where $\sigma_{\mu\nu} = \frac{i}{2}[\gamma_\mu, \gamma_\nu]$ is the commutator of the γ matrices and $F_{\mu\nu}(x)$ is the field strength tensor defined in Eq. (4.23). Previously, we were considering electromagnetism, but the same formula applies to non-Abelian theories if we replace e by g, the coupling constant for the non-Abelian group. A lattice expression for the field strength tensor can be constructed from four suitably oriented (uncontracted) plaquettes surrounding site n. This has come to be known as the *clover action* because the four plaquettes look like a four-leaf clover and clover is easy to spell. Thus, the Sheikholeslami-Wohlert or clover term in the action is

$$S_{SW} = \frac{iag}{4} c_{SW} \sum_{n,\mu,\nu} \bar{\psi}_n \sigma_{\mu\nu} \mathcal{F}_{n\mu\nu}\psi_n, \quad (4.40)$$

where $\mathcal{F}_{n\mu\nu}$ is the clover-like term discussed above. The coefficient c_{SW} can be tuned either perturbatively [308,309], or better yet, non-perturbatively [310,311]. The addition of the clover term is an example of an improvement program introduced by Symanzik [312,313].

A number of collaborations generate ensembles of gauge configurations using the Wilson-clover action. The scientific output from the expense of creating these configurations is greatly enhanced by sharing them for complementary investigations. The CLS, HSC, PACS, and QCDSF Collaborations are among those generating ensembles. Reference [314] describes the ensembles generated by a dozen collaborations and their plans to share them as presented at Lattice 2022. This paper also covers several of the other quark actions discussed below.

4.1.2 Twisted mass quarks

One issue with the Wilson formulation is that for small mass, it is possible to encounter so-called 'exceptional configurations' for which it is very difficult, if not impossible, to construct the quark propagator [315]. This was particularly an issue in the quenched approximation in which the fermion determinant is neglected. It can also slow down generation of configurations with dynamical quarks. For a theory with two light flavors, such as u and d, the twisted mass operator was invented to ensure that the fermion determinant is positive definite [316]. If the lattice Dirac operator is $D + m$, then

$$D_{\text{twist}} = D + m + i\mu\gamma_5\tau_3, \quad (4.41)$$

where τ_3 operates on the two flavors of quarks. Then $\det D_{\text{twist}} = \det((D+m)^\dagger(D+m) + \mu^2)$. So, as long as μ is non-zero, $\det D_{\text{twist}}$ is positive and exceptional configurations are avoided. This action has been used by the European Twisted Mass Collaboration for over 15 years. The collaboration is now the Extended Twisted Mass Collaboration as there are non-European members.

4.1.3 Staggered quarks

Staggered quarks are an alternative to Wilson quarks that reduce the degree of doubling and retain some of the chiral properties of the continuum theory [98,317–319]. One must be careful in reading the literature since some authors use x_0

for the time coordinate and others use x_4. This can have consequences for the field redefinition essential to the reduction in the number of fermions. Here we adopt the conventions in Refs. [320] and [321] rather than those in Ref. [322]. The key simplification is to rearrange the Dirac components at each site of the lattice in such a way that the action can be seen as comprised of four non-interacting fields. In this way, we may retain a single field component at each site and the doubling is reduced from 16 to 4. Initially, it was thought that this could be interpreted as four flavors or quarks, say, u, d, s, and c, but the modern interpretation is that each flavor has four 'tastes.' Tastes are not physical, so we must take a fourth root of the fermion determinant for each quark, and must be careful in constructing hadron operators to avoid mixing tastes as physical operators should really be constructed from a single taste. In the continuum limit, taste breaking vanishes so operators with mixed tastes should become degenerate with single taste operators. The rooting procedure and its validity is quite a technical subject. We refer the interested reader to Sec. III.C of Ref. [322] for a detailed discussion with references to the original literature.

Define a local redefinition of the Dirac components of the quark field by $\psi_n = \Omega_n \psi'_n$ and $\bar{\psi}_n = \bar{\psi}'_n \Omega_n^\dagger$. The 4×4 matrix Ω_n is defined as

$$\Omega_n = \gamma_0^{n_0} \gamma_1^{n_1} \gamma_2^{n_2} \gamma_3^{n_3}. \tag{4.42}$$

This may appear more complicated than it really is. Note that as $\gamma_\mu^2 = 1$, each gamma matrix appears in Ω_n only when the corresponding coordinate is odd. There are only 16 distinct values for Ω_n, and if we translate two sites in any direction, we have the same matrix. We will see that staggered quarks are naturally defined on 2^4 sub-hypercubes of the lattice. The gamma matrices are unitary and Hermitian, so

$$\Omega_n^\dagger \gamma_\mu \Omega_{n+\hat{\mu}} = (-1)^{n_0+\cdots+n_{\mu-1}} \equiv \alpha_\mu(n). \tag{4.43}$$

The hopping term in the naive fermion action

$$\bar{\psi}_n \gamma_\mu U_{n\mu} \psi_{n+\hat{\mu}} \tag{4.44}$$

is transformed into

$$\bar{\psi}'_n \Omega_n^\dagger \gamma_\mu U_{n\mu} \Omega_{n+\hat{\mu}} \psi_{n+\hat{\mu}} = \bar{\psi}'_n \alpha_\mu(n) U_{n\mu} \psi_{n+\hat{\mu}}. \tag{4.45}$$

The same factor appears in the hopping term that involves $\psi_{n-\hat{\mu}}$ since $\Omega_{n+\hat{\mu}} = \Omega_{n-\hat{\mu}}$ as the two sites differ by two units in the μ-direction. The gamma matrices have disappeared, and we are left with a unit matrix in Dirac index space, so there are four equivalent non-interacting components ψ'_n. We may discard three of the four components and write the staggered action in terms of a single component field χ.

$$S_{\text{stag}} = \frac{1}{2a} \sum_{n,\mu} \bar{\chi}_n \alpha_\mu(n) [U_{n\mu} \chi_{n+\hat{\mu}} - U^\dagger_{(n-\hat{\mu})\mu} \chi_{n-\hat{\mu}}] + m \sum_n \bar{\chi}_n \chi_n. \tag{4.46}$$

As mentioned above, because $\alpha_\mu(n)$ is periodic in each direction with period two, it is possible, perhaps natural, to interpret the 16 components on the sites of each 2^4 as the components of four Dirac spinors, i.e., the four tastes.

For the free theory, the four tastes can be expressed in the following way. Let y be a 4-component integer valued vector labeling the hypercubes. Let η be a four component vector whose components may only take the value 0 or 1. That is, η labels the 16 sites of a hypercube. For each hypercube y, the sites of the original lattice take the values $2y + \eta$ for one of the 16 values of η. Let α be a Dirac component index and a be a taste label. Both α and a range between 1 and 4. We have

$$\psi_y^{\alpha a} = \frac{1}{8} \sum_\eta \Omega_\eta^{\alpha a} \chi_{2y+\eta}. \tag{4.47}$$

This is not gauge covariant since we are adding together χ values from different lattice sites, so in the interacting case, χ at each site must be multiplied by suitable parallel transporter to move it to the origin of the hypercube. In practice, one really does not have to worry about this.

For Wilson quarks the action was improved by adding the clover term. For staggered quarks there have been similar advances by improving the action. For the most simple staggered action, the errors are $\mathcal{O}(a^2)$. Naik [323] introduced a 3-link hopping term. The gauge action was also improved by adding 2×1 rectangles, and 6-link terms that circle a 3-dimensional cube, sometimes called the bent chair diagram, known as the Lüscher–Weisz gauge action [324,325]. These terms are depicted in the top of Fig. 17. Essential benefits come from averaging or smearing the gauge fields in the 1-link hopping terms. These smearings are designed to reduce taste symmetry breaking. There have been two major rounds of these improvements, the first is known as the asqtad action [326–330] and the second is known as the highly improved staggered quark or HISQ action [331]. The paths for the fermion link smearings are shown in the bottom of Fig. 17. The HISQ action employs two levels of smearing. Reference [322] details the asqtad and HISQ actions and provides many physics results using the former action. The MILC Collaboration generates HISQ ensembles that are also used by the Fermilab Lattice and HPQCD Collaborations, and others. These improvements make the coding more complicated and require more floating point operations on a fixed grid size, but the payoffs can be enormous as the errors for the same lattice spacing are significantly reduced with the improved actions. If, say, an improvement would allow one to work at twice the lattice spacing as without the improvement there would be a significant reduction in computer time as halving the lattice spacing increases the work by a factor of 32 or more.

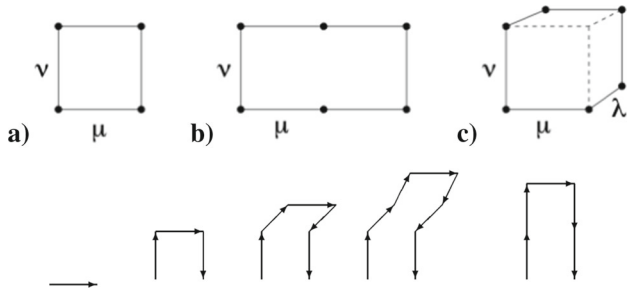


Fig. 17 Top: Loops that are included in the gauge action for asqtad and HISQ quarks. Bottom: For the asqtad action, the one-link hopping term in the naive staggered quark action is replaced by a combination of 1-, 3-, 5-, and 7-link smearings. The right-most 5-link term is known as the Lepage term. This figure is adopted from Ref. [330]. There is also a straight 3-link term known as the Naik term. For the HISQ action, there are two levels of smearing, but no additional paths are involved

4.1.4 Improving chiral symmetry

When the quark mass vanishes, the theory contains an important continuous symmetry known as chiral symmetry. The dynamical breaking of this symmetry is responsible for the pions being so light. The Wilson action explicitly breaks this symmetry, and the staggered actions discussed above only maintain some of the symmetry. However, there are other lattice actions that have much better chiral symmetry. These include the domain wall and overlap actions.

In the continuum, chiral symmetry follows from the fact that γ_5 anticommutes with the kinetic operator $D = \not{D}$. In 1982, Ginsparg and Wilson [335] considered the consequences of a generalized lattice chiral symmetry which is currently expressed as

$$D\gamma_5 + \gamma_5 D = aD\gamma_5 D. \tag{4.48}$$

Note the factor of the lattice spacing a on the RHS. As $a \to 0$, we restore chiral symmetry; however, even at non-zero a there is a more complicated chiral symmetry for operators that obey Eq. (4.48).

$$\psi \to \psi' = \exp\left(i\alpha\gamma_5\left(1 - \frac{a}{2}D\right)\right)\psi \tag{4.49}$$

with a similar expression for $\bar{\psi}$. As $a \to 0$, Eq. (4.49) approaches the usual expression for a chiral rotation, but at non-zero lattice spacing the transformation is more complicated as D is a finite difference operator. Reference [335] was not heavily cited until 1998 when the overlap operator that was developed by Narayanan and Neuberger [336–339] was shown to obey Eq. (4.48) [340]. If A obeys γ_5 Hermiticity, let $H = \gamma_5 A$, then

$$D_{ov} = \frac{1}{a}(1 + \gamma_5 \,\mathrm{sign}[H]) \tag{4.50}$$

defines the overlap operator. An alternative expression is

$$D_{ov} = \frac{1}{a}(1 + \gamma_5 H(HH)^{-1/2}). \tag{4.51}$$

A suitable choice for A is $D_W(0) - r$, with $0 < r < 2$. Numerically, it is difficult to compute the sign function or the inverse square root of a matrix. The χQCD Collaboration uses overlap fermions.

Two other papers from 1998 were also important in reviving interest in the Ginsparg–Wilson (GW) relation. In Ref. [341], Hasenfratz, Laliena, and Niedermayer showed that the fixed point action obeys the GW relation. Luscher demonstrated that the GW relation leads to an exact chiral symmetry even at non-zero lattice spacing [342].

In 1992, Kaplan introduced domain wall fermions in which chiral modes are bound to a defect in a 5-dimensional (5D) theory [343]. The theory was further developed by Shamir [344,345], and Furman and Shamir [346]. We adapt here the notation of Ref. [347]. Points in the five dimensional lattice are labeled by m in the four dimensional space and r in the 5th dimension, with $r = 0, \ldots, N_5 - 1$. The 5D fermion field is $\Psi(m, s)$. The 5D Dirac operator consists of two parts:

$$D^{\mathrm{dw}}(n, s; m, r) = \delta_{s,r} D(n; m) + \delta_{n,m} D_5^{\mathrm{dw}}(s, r). \tag{4.52}$$

The first term can be an ordinary Wilson operator with a modified mass:

$$D(n; m) = (4 - M_5)\delta_{n,m} - \frac{1}{2}\sum_{\mu=\pm1}^{\pm4}(1 - \gamma_\mu)U_{n;m}\delta_{n+\hat{\mu},m} \tag{4.53}$$

where we use notation $U_{n;m}$ to avoid having to specify hermitian conjugation for negative directions. Using $P_\pm = (1 \pm \gamma_5)/2$,

$$\begin{aligned} D_5^{\mathrm{dw}}(s; r) = {} & \delta_{s,r} - (1 - \delta_{s,N_5-1})P_-\delta_{s+1,r} \\ & -(1 - \delta_{s,0})P_+\delta_{s-1,r} + m(P_-\delta_{s,N_5-1}\delta_{0,r} \\ & +P_+\delta_{s,0}\delta_{N_5-1,r}). \end{aligned} \tag{4.54}$$

The physical 4D fields come from the boundaries of the 5D field:

$$\psi(n) = P_-\Psi(n, 0) + P_+\Psi(n, N_5 - 1). \tag{4.55}$$

Domain wall fermions are used extensively for dynamical quarks, especially by the RBC/UKQCD and JLQCD Collaborations.

4.1.5 Continuum limit

To control systematic errors it is crucial to tune the quark masses to their physical value, to have a volume that is large enough to avoid finite volume errors, and to take the limit

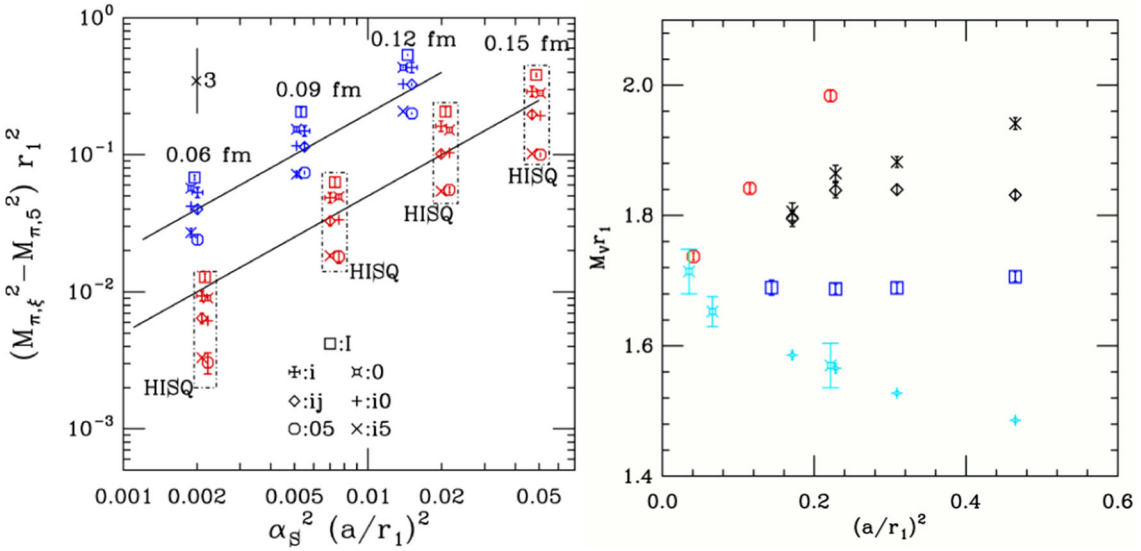


Fig. 18 Left: We show how taste breaking of pseudoscalar mesons decreases as the lattice spacing is reduced for two types of improved staggered quarks asqtad (blue) and HISQ (red). Different symbols denote different taste mesons. The quantity plotted is the difference of squared mesons masses for the plotted meson mass (ξ) and the Goldstone taste combination (γ_5). The horizontal axis is the $\alpha_s^2 a^2$ in units determined by the heavy quark potential r_1. Taste symmetry is restored in the continuum limit and taste breaking is much smaller for HISQ than for asqtad. See Ref. [332] for details. Right; The ρ meson mass as a function of lattice spacing for multiple actions shows a common continuum limit, but some actions have much more gentle lattice spacing dependence than others. Red octagons are unimproved staggered fermions with Wilson gauge action, diamonds are unimproved staggered fermions with Symanzik improved gauge action, crosses are Naik fermions and blue squares are asqtad fermions, both with Symanzik improved gauge action. For comparison we also show in light blue tadpole clover improved Wilson fermions with Wilson gauge action [333] (fancy squares) and with Symanzik improved gauge action [334] (fancy diamonds). (See Ref. [322] for details)

$a \to 0$. In the early days, it was too expensive to use physically light u and d quarks, so one also had to use chiral perturbation theory to extrapolate to those quark masses. Because QCD has the property of asymptotic freedom, the coupling constant goes to zero as the cutoff goes to infinity. On the lattice, the inverse lattice spacing plays the role of the cutoff. By dimensional transmutation, instead of expressing physical results in terms of the coupling, we do it in terms of the lattice spacing. In the left panel of Fig. 18, we show how taste breaking decreases as $a \to 0$ in accord with expected behavior for both asqtad and HISQ quarks. This also clearly shows that taste breaking is much smaller for HISQ (as it was designed with that in mind). In the right panel, we show how the ρ meson mass depends on the lattice spacing. Some of these results are rather old, and some are in the quenched approximation; however, the point to be made is that different ways of putting quarks on the lattice have the same continuum limit, although the rate at which they approach that limit will vary.

Modern calculations use multiple lattice spacings to control the continuum limit. Results using various quark actions are compared by the Flavor Lattice Averaging Group [256]. Calculations must use at least three lattice spacings to satisfy the quality criteria. Some calculations use five or six lattice spacings and can span a range as wide as about 0.15 fm to 0.03 fm. There is strong evidence from many different physical quantities that different quark actions agree in the continuum limit. Differences between calculations with and without a dynamical charm quark tend to be quite small. See Sects. 4.5 and 4.7 for results comparing different actions.

4.1.6 Further reading

I have made no attempt at a historically accurate account of lattice QCD, and due to space limitations much has been left out. Here I list some books on the topic. As far as I know, "Quarks, gluons and lattices" by Creutz is the first monograph[348]. Creutz also edited "Quantum Fields on the Computer," which covers scalar and Yukawa theories in addition to QCD [349]. Proceedings from the 1989 TASI summer school edited by DeGrand and Toussaint [350] was an early essential reference. Books by Rothe [351] and by Montvay and Munster [352] appeared in the early 1990s. The former is now in its fourth edition and is available online via open access. Since 2000, at least three books have been published. Authors are Smit [321]; DeGrand and DeTar [320]; and Gattringer and Lang [347].

4.1.7 Personal remarks

In 1975, I had the opportunity to take my first European physics trip when I attended the Erice summer school in

Sicily. Little did I know as I listened to Ken Wilson lecture on quark confinement and lattice gauge theory how profoundly his work would impact my own. (As an undergraduate, I only remember talking to Wilson once when he kindly gave me advice on which graduate schools I should apply to.) I recall being quite friendly with Michael Creutz during the school. Claudio Rebbi was one of the lecturers. I have had many great interactions with both of them. I was in awe of seeing Paul Dirac walking quietly around the school. Tom De Grand would later become a collaborator. Sidney Coleman, of course, gave a great series of lectures.

During my postdoc at Argonne, Creutz was kind enough to send me a printed copy of his code. I had a great title for a paper: "Looking for Glue in SU(2)." Unfortunately, I didn't really know anything about glueballs, so I did not pursue that. While at Fermilab, I was looking for something new and worked on the Migdal–Kadanoff recursion relations with Khalil Bitar and Cosmas Zachos (who I had known when he was an undergrad and I was in graduate school). Don Weingarten visited and I started my career of Monte Carlo lattice calculations (using SU(3) not SU(2)). Hank Thacker, Paul Mackenzie, Weingarten, and I used some of the VAX computers at Fermilab for our calculations to examine ρ decay. We worked on a $6^2 \times 12 \times 18$ lattice and had so few configurations we joked that we knew each one by name. For no good reason, I still have some of the magnetic tapes on which we stored the configurations. This project continued when I moved to UC San Diego. A year later, my grad school housemate Doug Toussaint arrived as an assistant professor. I started working with him, and more senior people such as Bob Sugar and Julius Kuti. A few years later the MILC Collaboration started, and I would like to mention fellow founding members Claude Bernard and Carleton DeTar. Lattice gauge theory has been my life ever since then.

4.2 Monte-Carlo methods

Carleton DeTar

4.2.1 Introduction

In 1980, Michael Creutz pioneered the numerical simulation of lattice QCD [353,354] with studies of Wilson's lattice formulation of SU(2) Yang–Mills theory. This feasibility study started a vast enterprise devoted to "solving" QCD in the nonperturbative regime. Later on, as computing power grew, it became possible to include quarks, thus bringing simulations in contact with reality. This subsection introduces basic methods for carrying out the numerical simulation of lattice QCD using Monte Carlo methods. It concludes with a mention of ongoing improvements.

4.2.2 Lattice path integration

Partition function

The most widely used strategy for numerical simulation of QCD starts from a Feynman path integral formulation [355], which is based on the partition function

$$Z = \int [dU d\overline{\psi} d\psi] \exp[-S(U, \overline{\psi}, \psi)] \tag{4.56}$$

where

$$S(U, \overline{\psi}, \psi) = S_W + \overline{\psi} M \psi \tag{4.57}$$

is the Euclidean action for the lattice SU(3) gauge field U and quark field ψ, as defined in Sect. 4.1. For simplicity here, we treat only one quark flavor, and we suppress the color (c), vector (μ), and spatial (n) indices on $U_{cc',\mu}(n)$ and the color, spin (α), and spatial indices on $\psi_{c,\alpha}(n)$. Note that for lattice volume V (number of sites) there are $4V$ SU(3) matrices denoted by U and V spin/color vector fields denoted by ψ.

The integration over the gauge links U is done over the classical SU(3) gauge field $U_\mu(n)$ on each lattice link. We use the invariant Haar measure $dU_\mu(n)$ on each link. (We won't need it, but there is an Euler-angle representation of the measure [356].) The integration is done without gauge fixing. Since the action S is gauge invariant and the gauge group is compact, the integral over gauge choices is finite. In the Feynman formulation, fermion fields, in particular ψ, must be anticommuting Grassmann variables. This assures that they obey Fermi-Dirac statistics. It would be challenging to treat them directly in a computer simulation, but, fortunately, they can be integrated out using only the identities listed below and their analogs, leaving expressions involving only the classical gauge field. For a few more details, see Ref. [320].

In a Euclidean spacetime with finite time extent T, the quantity Z in Eq. (4.56) is the thermal partition function for the theory defined by the action S with hamiltonian $\mathcal{H}$ at inverse temperature $\beta = T$. Thus

$$Z(\beta) = \mathrm{Tr} \exp(-\beta \mathcal{H}). \tag{4.58}$$

The zero temperature limit corresponds to an infinite time extent.

Grassman calculus

We need three important identities from the Grassmann calculus:

$$\int [d\overline{\psi} d\psi] \exp(-\overline{\psi} M \psi) = \det M \tag{4.59}$$

$$\int [d\overline{\psi} d\psi] \overline{\psi}_{c,\alpha}(n) \psi_{c',\alpha'}(n') \exp(-\overline{\psi} M \psi) = M^{-1}_{c,\alpha;c',\alpha'}(n, n') \det M \tag{4.60}$$

$$\int [d\overline{\psi} d\psi] \psi_{c,\alpha}(n) \overline{\psi}_{c',\alpha'}(n') \overline{\psi}_{d,\beta}(m) \psi_{d',\beta'}(m')$$
$$\exp(-\overline{\psi} M \psi) = [M^{-1}_{c,\alpha;c',\alpha'}(n,n') M^{-1}_{d,\beta;d',\beta'}(m,m')$$
$$- M^{-1}_{c,\alpha;d',\beta'}(n,m') M^{-1}_{d,\beta;c',\alpha'}(m,n')] \det M \tag{4.61}$$

The inverse of the fermion matrix M is the fermion propagator. We see that each $\overline{\psi}$, ψ pair in the integrand contributes a fermion propagator. All pairings can occur, as in the last example. The minus sign there arises from the anticommuting property of the fields.

Observables

Physical quantities are defined in terms of observables $\mathcal{O}(U, \overline{\psi}, \psi)$ constructed from the variables U, $\overline{\psi}$, and ψ. To obtain the expectation value of the observable, we calculate

$$\langle \mathcal{O} \rangle = Z^{-1} \int [dU d\overline{\psi} d\psi] \mathcal{O}(U, \overline{\psi}, \psi) \exp[-S(U, \overline{\psi}, \psi] \tag{4.62}$$

Meson propagator

For example, we might want to determine the mass of a pseudoscalar meson. To do so we work with an "operator" that "creates" or "destroys" the meson:

$$\mathcal{O}_{PS}(\mathbf{p}, t) = \sum_{\mathbf{r}} \exp(i\mathbf{p} \cdot \mathbf{r}) \overline{\psi}(\mathbf{r}, t) \gamma_5 \psi(\mathbf{r}, t), \tag{4.63}$$

where $\mathbf{p}$ is the momentum. Note that if we replace the Grassmann field with a quantum field, the same operator in quantum field theory would create or destroy the meson. The sum over spatial sites $\mathbf{r} = (x, y, z)$ for fixed t and $\mathbf{p}$ gives a meson of momentum $\mathbf{p}$ at Euclidean time t. To obtain the mass, we calculate at zero momentum and large $|t' - t|$

$$C_{PS}(t', t) = \langle \mathcal{O}_{PS}(\mathbf{0}, t') \mathcal{O}_{PS}(\mathbf{0}, t) \rangle$$
$$= z^2_{PS} \exp[-M_{PS}|t' - t|] \tag{4.64}$$

where z_{PS} is the amplitude and M_{PS} is the meson mass. In effect, we are creating the meson at time t and destroying it at time t'. The meson propagates between these times. In Minkowski space the meson propagator would be proportional to the phase factor $\exp[-iM_{PS}|t' - t|]$. In Euclidean space here, it falls exponentially in the time separation at a rate controlled by the mass M_{PS}. This expression is strictly valid only for large time separations $|t' - t|$. At smaller separations, we would get additional, higher-mass contributions.

The meson interpolating operators are sometimes called "source" and "sink". Which is which depends on the point of view, since they can serve a dual purpose.

Integrating out the fermion fields

Let's examine the expectation value in Eq. (4.61) in more detail. Note that we can integrate out the fermion fields exactly by making use of the identities in Eqs. (4.59) and (4.61). When we insert the product of two interpolating operators from Eq. (4.63) into Eq. (4.61) we get a product of two

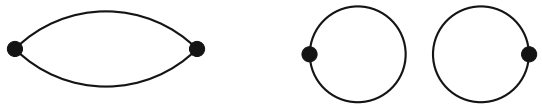


Fig. 19 Quark line connected and disconnected diagrams

Grassmann fields $\overline{\psi}$ and two Grassmann fields ψ. According to Eq. (4.61), we get

$$C_{PS}(t', t)$$
$$= \Big\langle \sum_{\mathbf{r};\mathbf{r}'} \{ \mathrm{Tr}_{cs}\, \gamma_5 M^{-1}(\mathbf{r}, t; , \mathbf{r}', t') \gamma_5 M^{-1}(\mathbf{r}', t'; \mathbf{r}, t))$$
$$- \mathrm{Tr}_{cs}\, \gamma_5 M^{-1}(\mathbf{r}, t; \mathbf{r}, t)\, \mathrm{Tr}_{cs}\, \gamma_5 M^{-1}(\mathbf{r}', t'); \mathbf{r}', t') \} \Big\rangle_G \tag{4.65}$$

where Tr_{cs} denotes a trace over color and spin indices and we have defined, for any function E of the gauge field,

$$\langle E \rangle_G = Z^{-1} \int [dU] E \exp[-S_W] \det M. \tag{4.66}$$

With the fermion fields integrated out, the Feynman path integrals now involve only integration over the classical gauge field, which is amenable to numeric integration. The meson propagator in Eq. (4.65) has two terms that are represented diagrammatically in the two panels of Fig. 19. We call the two contributions quark-line "connected" and quark-line "disconnected". The loop in the disconnected diagram represents the annihilation of the quark and antiquark in the interpolating operator. It contributes only if the meson is a flavor singlet. With the addition of flavor we would find that the pion does not have this term.

Form of the meson propagator

On a finite lattice with Euclidean time extent T and the usual periodic/antiperiodic boundary conditions on the fields, the meson correlation function $C_{PS}(t', t)$ gets another contribution as the meson propagates in the opposite direction from the source, exploiting the periodic/antiperiodic boundary condition, and arriving again at the sink. The distance traveled in Euclidean time is now $T - |t' - t|$. Thus we have

$$C_{PS}(t', t) = z^2_{PS} \exp[-M_{PS}|t' - t|]$$
$$+ z^2_{PS} \exp[-M_{PS}(T - |t' - t|)]. \tag{4.67}$$

Figure 20 illustrates the result of a calculation of the pion propagator showing both forward and backward propagation.

Decay constant

The amplitude z_{PS} is proportional to the meson decay constant f_{PS}:

$$z_{PS} = Z_{PS} f_{PS} \tag{4.68}$$

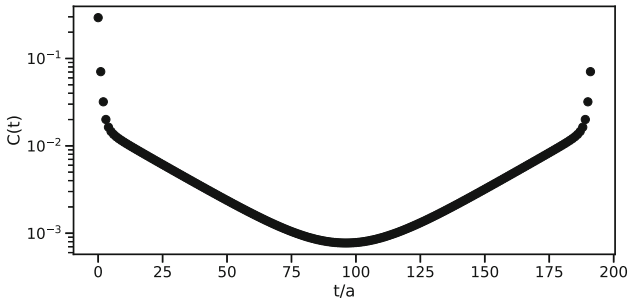


Fig. 20 Zero momentum pion propagator at lattice spacing $a = 0.06$ fm as a function of Euclidean time expressed in units of the lattice spacing (courtesy William Jay). The source is at $t = 0$. In this case $T = 192a$

where Z_{PS} is a renormalization constant ("matching factor") that relates the lattice interpolating operator to a physical continuum interpolating operator.

Form factor

Form factors give information about hadron structure and decay. Here we illustrate the construction of the electromagnetic form factor for the meson illustrated above. We calculate the three-point function

$$C_\mu(t, \mathbf{q}, t', t'') = \langle \mathcal{O}_{PS}(t, \mathbf{q}) J_\mu(t', \mathbf{q}) \mathcal{O}_{PS}(t'', \mathbf{0}) \rangle, \quad (4.69)$$

where $J_\mu(t', \mathbf{q})$ is the current density projected onto spatial three momentum $\mathbf{q}$ and $\mathbf{0}$ denotes zero momentum. For simplicity we have chosen zero momentum for the meson interpolating operator at time t'', and we have enforced three-momentum conservation. (In Euclidean space-time, we don't have energy conservation, but the meson propagators are on shell.)

Form of the three-point function

For $t \ll t' \ll t''$ the three-point function has the form

$$C_\mu(t, \mathbf{q}, t', t'') = z_{PS}(\mathbf{q}) Z_V z_{PS}(\mathbf{0}) \exp(-M_{PS}|t' - t|) \times F_\mu(\mathbf{q}) \exp(-M_{PS}|t'' - t'|) \quad (4.70)$$

where $F_\mu(\mathbf{q})$ is the desired form factor. The current renormalization constant Z_V matches the lattice current J_μ to the continuum current.

Integrating out the fermion fields

There are a variety of choices for the current density. We could work with the conserved lattice Noether current. Or we could work with a "local" current

$$J_\mu(\mathbf{r}, t) = Q \overline{\psi}(\mathbf{r}, t) \gamma_\mu \psi(\mathbf{r}, t). \quad (4.71)$$

where Q is the charge. This current is not conserved at nonzero lattice spacing, but with suitable renormalization, it should give the same result as the conserved current in the continuum limit. We use the local current here for simplicity.

We integrate out the fermion fields following the same steps as for the meson propagator. We display, here, only the quark-line-connected contribution:

$$C_\mu(t, \mathbf{q}, t', t'') = \sum_{\mathbf{r}, \mathbf{r}', \mathbf{r}''} \exp(-i\mathbf{r} \cdot \mathbf{q}) \exp(i\mathbf{r}' \cdot \mathbf{q}) \times \Big\langle \mathrm{Tr}_{cs}\, \gamma_5 M^{-1}(\mathbf{r}, t; , \mathbf{r}', t') \gamma_\mu M^{-1}(\mathbf{r}', t'; \mathbf{r}'', t'')) \times \gamma_5 M^{-1}(\mathbf{r}'', t'', \mathbf{r}, t) \Big\rangle_G \quad (4.72)$$

The quark-line structure is the closed loop diagrammed in the left panel of Fig. 19.

4.2.3 Monte Carlo methods

Importance sampling

The path integral in Eq. (4.66) involves integration over so many variables that Monte-Carlo importance sampling becomes the only method of choice. A single point in the domain of integration is specified by the gauge field values U on each link – called a gauge field configuration. The integrand is sampled over random gauge-field configurations with probability density P of encountering a given configuration U. If the sampling is designed so that P is proportional to the integrand weight factor

$$P \propto \exp[-S_W] \det M], \quad (4.73)$$

then in an ensemble of such gauge configurations U_i for $i = 1, \ldots, N$, the expectation value of an observable E is simply the ensemble average in the limit $N \to \infty$.

$$\langle E \rangle = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E(U_i). \quad (4.74)$$

Of course, the weight factor must be positive definite in order to be treated as a probability density. This is usually the case, but there are important exceptions. One can use the same path-integral formalism to treat a grand-canonical ensemble of fermions at nonzero fermion number (or flavor) density; see Sect. 4.4. In this case the fermion determinant acquires a complex phase (the so-called "sign problem") that obviates a probabilistic treatment.

Markov chain

There are various methods for generating such an ensemble. They all involve creating a Markov chain of gauge configurations U_i, i.e., a sequence generated by a stochastic rule that takes the previous configuration U and produces a new configuration U'. The Markov chain proceeds from an arbitrary starting configuration. With a properly devised stochastic rule, after a sufficient number of steps the probability distribution approaches the desired distribution of Eq. (4.73). We say that the distribution has "reached equilibrium". Of course we must also take care that the distribution is "ergodic" in the

sense that all important regions of the integrand are included in the ensemble – that the distribution isn't "frozen" around one local minimum of the effective action at the expense of other equally important minima.

Heatbath algorithm

The heatbath algorithm runs through the lattice updating each gauge link, one at a time. For the gauge link matrix $U_\mu(n)$, the integrand weight is regarded as defining a probability distribution $R[U_\mu(n)]$ for the gauge link being updated. One chooses a new gauge link $U_\mu(n)'$ from that distribution and then moves on to the next gauge link. The name "heat bath" comes from early studies of $SU(2)$ pure gauge theory in which the effective action was proportional to a coupling constant that could be interpreted as an inverse Monte-Carlo temperature (not to be confused with the temperature of the partition function). So the update was analogous to exposing each link to a heat bath of that temperature. The heat bath method has fallen into disuse in lattice QCD now that more calculations include fermions, because the fermion determinant has a nontrivial dependence on the gauge links, which makes selecting a new link matrix $U_\mu(n)$ from a local probability distribution $R[U_\mu(n)]$ too expensive to implement.

Metropolis–Hastings algorithm

A classic method for generating the desired Markov chain uses the algorithm of Metropolis et al. and Hastings [357], usually abbreviated as the "Metropolis" algorithm. It works with a general class of stochastic rules for proposing a new gauge configuration U' and then either accepts or rejects the new configuration based on a criterion designed to lead asymptotically to the desired ensemble:

- Propose a new configuration U' with probability $Q(U' \leftarrow U)$. The transition must satisfy the reversibility condition:

$$Q(U \leftarrow U') = Q(U' \leftarrow U). \tag{4.75}$$

  Also, it must be possible after some number of steps to reach any configuration with nonzero probability.
- Choose a random number λ distributed uniformly on $[0, 1]$.
- If the proposed change decreases the effective action $\Delta S_{\text{eff}} = S_{\text{eff}}(U') - S_{\text{eff}}(U) < 0$ then accept the change.
- Otherwise, accept the change if $\exp[-\Delta S_{\text{eff}}] > \lambda$. Otherwise, reject it.

The transition process defined by $Q(U' \leftarrow U)$ is quite general, which makes the algorithm particularly useful.

4.2.4 Molecular dynamics

By far the most common present-day method for generating the Markov chain uses a "molecular dynamics" method. We illustrate it for a scalar field ϕ with path-integral partition function

$$Z = \int [d\phi] \exp[-S(\phi)]. \tag{4.76}$$

We pair a dummy "momentum" $p(n)$ with the field $\phi(n)$ on each site of the lattice and rewrite the partition function as

$$Z' = \int [dp][d\phi] \exp[-p^2/2 - S(\phi)]. \tag{4.77}$$

The momentum integral is trivial and results in an immaterial constant factor. We then take a lesson from classical statistical mechanics and observe that this partition function describes a statistical ensemble of "particles" of unit mass, one per lattice site, and unit temperature $kT = 1$ moving in an interacting "potential" $S(\phi)$. The ensemble is microcanonical with total energy

$$E_{\text{tot}} = p^2/2 + S(\phi). \tag{4.78}$$

The Hamilton equations of motion are, as usual,

$$d\phi(n)/d\tau = p(n) \tag{4.79}$$

$$dp(n)/d\tau = -\partial S/\partial\phi(n), \tag{4.80}$$

where τ is a fictitious "Monte Carlo time". We then observe that if the system is large and the interactions are nontrivial, the classical motion of the system will lead to a Maxwell-Boltzmann distribution in the coordinates ϕ given by

$$P(\phi) \propto \exp[-S(\phi)]. \tag{4.81}$$

In standard practice, one chooses an arbitrary starting field configuration ϕ and sets the initial momenta according to the Gaussian distribution $\exp[-p^2/2]$. Using a numerical integrator, one integrates the equations of motion over some time interval $\Delta\tau$, at which time one saves an "updated" configuration ϕ_i. Thus the Markov chain is defined by the values of ϕ at a series of time intervals or a series of what are called "molecular dynamics trajectories".

Refreshed and hybrid Monte Carlo

The total energy E_{tot} is constant over a given trajectory. But it has no particular physical significance. To improve coverage of phase space it is common, at the beginning of each trajectory, to "refresh" the momenta p by drawing new values from their Gaussian distribution. Thus each trajectory starts in a new direction with a new total energy, but the coordinates ϕ are kept continuous.

Another common variation of the method combines refreshed molecular dynamics with the Metropolis et al. method. This combination is called "hybrid Monte Carlo" [358]. That is, one starts a trajectory with coordinates p, ϕ. At the end of the trajectory, one has coordinates p', ϕ'. The transition $\phi' \leftarrow \phi$ is taken as a Metropolis move. The randomness in the refreshed initial momentum p makes the move

stochastic. Time-reversal invariance in τ assures detailed balance. If a trajectory is rejected, one reverts to the coordinate ϕ at the beginning of the trajectory, selects a new stochastic momentum, and tries again. The hybrid scheme helps compensate for possible inaccuracies in the numerical integration scheme, since it absolves many sins.

Autocorrelations

Markov chains have inherent correlations between successive members. These "autocorrelations" are undesirable, because they reduce the statistical independence of terms in the ensemble averages of Eq. (4.74) that give expectation values of physical observables. Autocorrelation is especially a concern with methods that make a series of small changes in the field configuration. With refreshed molecular dynamics one can adjust the trajectory length $\Delta\tau$ to help reduce correlations between successive terms ϕ_i. One might expect that longer trajectories are better in this regard, but the "molecular motion" can contain cycles that bring parts of the system close to their original values. With hybrid schemes, longer trajectories can lead to lower Metropolis acceptance, which impedes progress. Shorter trajectories suffer from greater autocorrelation. Thus there is usually an optimum choice for the trajectory length that needs to be found empirically.

Molecular dynamics for the gauge field

The methods described above for a scalar field carry over to the SU(3) gauge field U. The gauge momentum, actually associated with the vector potential, $A_\mu(n)$, is given by a traceless antihermitian 3x3 matrix $H_\mu(n)$ for each $U_\mu(n)$. The molecular dynamics hamiltonian is, then,

$$\mathcal{H} = \frac{1}{2}\sum_{n,\mu} \mathrm{Tr}\, H_\mu(n)^2 + S_{\mathrm{eff}} \tag{4.82}$$

where we recall that

$$S_{\mathrm{eff}} = S_W + \ln\det[M] \tag{4.83}$$

To remain an SU(3) matrix, the equation of motion for $U_\mu(n)$ must be

$$dU_\mu(n)/d\tau = i H_\mu(n) U_\mu(n). \tag{4.84}$$

The equation of motion for $H_\mu(n)$ can be found by requiring that the molecular dynamics Hamiltonian $\mathcal{H}$ remain constant in molecular-dynamics time [359]. For the sake of pedagogy, we first ignore the fermion determinant and consider the unimproved SU(3) gauge theory; see Sect. 4.1:

$$S_W = \frac{\beta}{6}\sum_{n,\mu\neq\nu}[3 - \mathrm{Re}\,\mathrm{Tr}\, U_{\mu\nu}(n)] \tag{4.85}$$

where $U_{\mu\nu}(n)$ is the plaquette product in the $\mu\nu$ plane with corner at site n. The plaquette can also be written as

$$\sum_\nu \mathrm{Re} U_{\mu\nu}(n) = U_\mu(n)V_\mu(n) + V_\mu(n)^\dagger U_\mu(n)^\dagger \tag{4.86}$$

where $V_\mu(n)$ is the sum of all "staples" attached to the link $U_\mu(n)$. Armed with this notation, we can write

$$0 = \dot{\mathcal{H}} = \sum_{n,\mu}\mathrm{Tr}\left[\dot{H}_\mu(n)H_\mu(n) + \frac{\beta}{6}\left(\dot{U}_\mu(n)V_\mu(n) + h.c.\right)\right], \tag{4.87}$$

and, using Eq. (4.84), we get

$$0 = \sum_{n,\mu}\mathrm{Tr}\left[\dot{H}_\mu(n)H_\mu(n) + \frac{\beta}{6}\left(i H_\mu(n)U_\mu(n)V_\mu(n) - h.c\right)\right], \tag{4.88}$$

or

$$0 = \sum_{n,\mu}\mathrm{Tr}\, H_\mu(n)[\dot{H}_\mu(n) + i F_\mu(n)] \tag{4.89}$$

where the gauge force is

$$F_\mu(n) = -\frac{\beta}{6}(U_\mu(n)V_\mu(n) - h.c.). \tag{4.90}$$

Since $H_\mu(n)$ in Eq. (4.89) is traceless the expression in brackets must be proportional to the identity matrix cI. But if it is to remain traceless, we must have $c = 0$. So, finally, we get

$$i\dot{H}_\mu(n) = F_\mu(n) = -\frac{\beta}{3}U_\mu(n)V_\mu(n)|_{\mathrm{TA}}, \tag{4.91}$$

where TA denotes the traceless, antihermitian part. The Eqs. (4.84) and (4.91) form the basis for molecular dynamics evolution of the pure gauge theory.

Spectrum of the Dirac matrix

The Dirac matrix has the form (see Sect. 4.1)

$$M(U) = m + D(U). \tag{4.92}$$

where m is the quark mass. For all fermion formulations in common use today, the operator D satisfies "γ_5 hermiticity", namely

$$D^\dagger = \gamma_5 D \gamma_5 \tag{4.93}$$

for some definition of γ_5. (For brevity, we drop the (U) dependence of M and D in the following.) This implies that the complex eigenvalues of D appear in complex conjugate pairs. Thus we can write the fermion determinant as

$$\det M = \prod_{\mathrm{Im}\lambda_i = 0, i}(m + \lambda_i)\prod_{\mathrm{Im}\lambda_i > 0}(m^2 + |\lambda_i|^2). \tag{4.94}$$

In order for $\det M$ to serve as a probability weight, it must be real and positive definite. Indeed, for all but the Wilson and clover actions, the real parts of the eigenvalues are nonnegative. For domain-wall and Wilson fermions, the eigenvalues λ_i populate an ellipse in the right-half plane with voids, as illustrated in Figs. 21 and 22. For staggered fermions, they lie entirely on the imaginary axis (not shown). For overlap fermions, they lie on a circle in the right-half complex plane tangent to the imaginary axis (also not shown). For Wilson

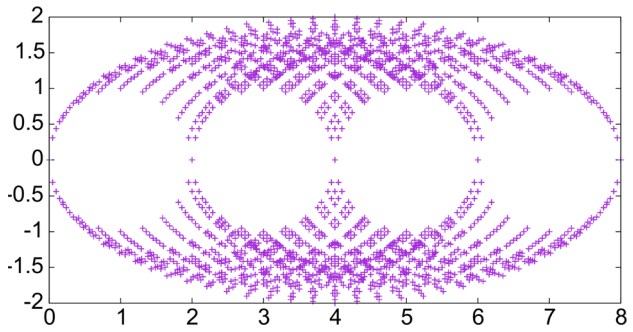


Fig. 21 The spectrum of the free Wilson Dirac operator for a massless quark [360]

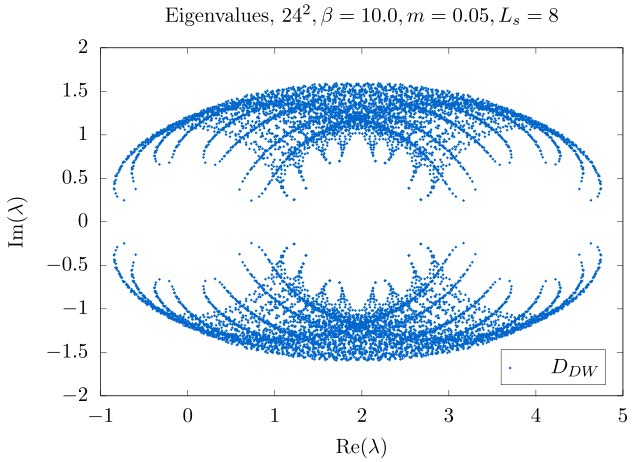


Fig. 22 The spectrum of the domain wall operator at quark mass 0.05. [361]

and clover fermions, they appear mostly in the right-half plane, but real, negative eigenvalues are possible, depending on the gauge configuration U. The eigenvalues of M are $m + \lambda_i$, so negative λ_i usually causes trouble for M only for light quarks. As the lattice spacing is decreased, negative real parts become less frequent. Twisted-mass fermions (see Sect. 4.1) do not have this problem at maximal twist [362].

The Φ algorithm

The fermion determinant in Eq. (4.66) can be cast in a form compatible with the molecular-dynamics treatment of the gauge field. Perhaps the simplest approach is the "Φ algorithm". We introduce a complex lattice scalar field Φ, often called a pseudofermion field, and first try

$$\det M = \int d\Phi d\Phi^* \exp[-\Phi^* M^{-1} \Phi]. \tag{4.95}$$

This works as long as the eigenvalues of M have positive definite real parts. However, this form is awkward to implement. A more convenient form works with the normal operator $M^\dagger M$. From γ_5 hermiticity we have

$$\det M = \det[\gamma_5 M^\dagger \gamma_5] \tag{4.96}$$

$$\det M^2 = \det[M^\dagger M] \tag{4.97}$$

so

$$\det M^2 = \int d\Phi d\Phi^* \exp\{-\Phi^* [M^\dagger M]^{-1} \Phi\}. \tag{4.98}$$

Now the integral is always well defined for nonzero quark mass, but the square doubles the number of fermions. That could be acceptable if we are simulating up and down quarks in the isospin symmetric limit ($m_u = m_d$), but it would be a bad approximation for the other quarks. Remedies are discussed below. We continue with this form.

Molecular dynamics with fermions

To simulate Eq. (4.98), we note that the integrand has the form $\exp(-R^\dagger R)$ for $R = M^{-1}\Phi$, so if we draw R from a Gaussian distribution, then $\Phi = MR$ is distributed according to the desired weight.

The Φ algorithm begins a short trajectory by constructing $\Phi = MR$ for a given starting gauge field U. The gauge field is then evolved with Φ fixed. The force exerted on the gauge field in Eq. (4.91) acquires a new contribution, namely, the fermion force:

$$\begin{aligned} i F_{F,\mu n} &= \frac{\partial}{i \partial A_{\mu,n}} \Phi^* [M^\dagger M]^{-1} \Phi \\ &= X^* \frac{\partial}{i \partial A_{\mu,n}} (M^\dagger M) X, \end{aligned} \tag{4.99}$$

where $M^\dagger M X = \Phi$. Typically, one refreshes the gauge momentum, evolves the gauge field at the initial fixed value of Φ, and then repeats.

Rational function approximation

As we saw above, working with the normal operator $M^\dagger M$ doubles the number of fermion species. To eliminate the doubling, we should replace $M^\dagger M$ with $\sqrt{M^\dagger M}$. Similarly, for staggered fermions, we start with four tastes per flavor, which suggests $(M^\dagger M)^{1/8}$. With staggered fermions, the normal operator is checkerboard block-diagonal, so restricting the calculation to even lattice sites eliminates the normal-operator doubling. We then want $(M^\dagger M)^{1/4}|_{\text{even}}$.

Such fractional powers are difficult to implement. A now commonly used remedy introduces a rational-function approximation for the fractional power [363]. Expanded in terms of its poles, the rational function approximation for a real function $f(x)$ of real x has the form

$$f(x) \approx r(x) = \sum_{i=1}^{N} \frac{\alpha_i}{x - \beta_i}, \tag{4.100}$$

where α_i and β_i are parameters of the rational function, and N is a suitably high order. The approximation deteriorates for small x. It is designed to work over an interval $[x_{\min}, x_{\max}]$. The smaller $x_{\min}$ or the finer the desired accuracy, the larger

the needed order N. The Zolotarev method [364] is widely used to obtain an efficient set of parameters α_n and β_n.

We note that $M^\dagger M = D^\dagger D + m^2$ for mass m. It is convenient to treat this expression as a function of $x = D^\dagger D$. So to apply the rational function approximation, we write

$$(M^\dagger M)^h \approx r_h(D^\dagger D) = \sum_{i=1}^{N} \frac{\alpha_{h,i}}{D^\dagger D - \beta_{h,i}}, \tag{4.101}$$

where we have labeled the coefficients of the expansion with the desired power h. So, finally, we have

$$\det(M^\dagger M)^h \approx \int d\Phi d\Phi^* \exp[-\Phi^* r_h(D^\dagger D)]\Phi] \tag{4.102}$$

To implement the Φ algorithm with fractional power h,

$$\int d\Phi d\Phi^* \exp\{-\Phi^*(M^\dagger M)^h \Phi\}, \tag{4.103}$$

we choose Gaussian random R and calculate

$$\Phi = [M^\dagger M]^{-h/2} R \tag{4.104}$$

using a rational function approximation $r_{-h/2}(D^\dagger D)$. Then we calculate the fermion force with

$$\begin{aligned} i F_{F,\mu n} &= \frac{\partial}{i\partial A_{\mu,n}} \Phi^*(M^\dagger M)^h \Phi \\ &= \Phi^* \frac{\partial}{i\partial A_{\mu,n}} r_h(D^\dagger D)\Phi \qquad (4.105) \\ &= \sum_i X_i^* \alpha_{h,i} \frac{\partial}{i\partial A_{\mu,n}} [M^\dagger M] X_i, \qquad (4.106) \end{aligned}$$

where $X_i = [D^\dagger D - \beta_{h,i}]^{-1}\Phi$. Here the rational function parameters are appropriate for r_h. The X_i are obtained using a multishift conjugate-gradient solver.

Multiple flavors

The rational function approximation can be extended to handle the products of determinants that arise with multiple flavors. For example, suppose we are simulating two degenerate light quarks (up and down) $m_l = m_u = m_d$ and one strange quark m_s. We use f to distinguish the flavors in the fermion matrix M_f. After integrating out the Grassmann fields, the fermion integrand becomes

$$\det(M_l^\dagger M_l) \det(M_s^\dagger M_s)^{1/2}. \tag{4.107}$$

We could simulate this product by introducing a separate pseudofermion field for each flavor and proceeding as we did for a single flavor for each contribution. However, we can also simulate it using just one pseudofermion field:

$$\int d\Phi d\Phi^* \exp\{-\Phi^*(M_l^\dagger M_l)^{-1}(M_s^\dagger M_s)^{-1/2}\Phi\} \tag{4.108}$$

We construct a rational function that approximates the entire product.

$$(M_l^\dagger M_l)^{-1}(M_s^\dagger M_s)^{-1/2} = r_{-1,-1/2}(D^\dagger D), \tag{4.109}$$

where we have added more labels to $r(x)$. The Φ algorithm is otherwise similar to that of the single-flavor case.

4.2.5 Improvements

Hasenbusch term

One popular and effective improvement [365] introduces a "preconditioning" determinant, a "Hasenbusch term", with moderately large mass m_x together with its compensating inverse, for example, as

$$(M_l^\dagger M_l)^{-1}(M_s^\dagger M_s)^{-1/2}(M_x^\dagger M_x)^{3/2}(M_x^\dagger M_x)^{-3/2}. \tag{4.110}$$

The first three factors are then assigned a single pseudofermion field and approximated with a single rational function, and the fourth factor is assigned a separate pseudofermion field with a separate rational function. The Hasenbusch term tends to reduce the condition number of the product operator, thus reducing the needed rational function order and the associated computation time. The last (compensating) factor also has a lower condition number because of the larger mass.

Multigrid solvers

To evaluate the rational function in Eq. (4.106) requires solving a large linear system. As the lattice spacing decreases, the condition number of the linear system grows, making the conventional conjugate-gradient calculation more costly. This "critical slowing down" can be mitigated by using an adaptive geometric multigrid solver instead [366,367]. So far the benefits of using multigrid solvers for gauge-field evolution have been demonstrated only for the Wilson-clover action [368]. Algorithms for multigrid solvers for staggered fermions [369] and domain-wall fermions [361,370] are newer, so it remains to be seen whether they will lead to improvements in molecular dynamics evolution for those fermion formulations as well.

Accelerating molecular dynamics

As the lattice spacing decreases, the gauge-field evolution slows, and it gets trapped in a subset of gauge configurations with the same total topological charge. Thus it takes more computational time to obtain a new, statistically uncorrelated gauge configuration. Long-distance decorrelation is slower than short-distance. This observation suggests Fourier transforming Hamilton's equation for the gauge momentum,

$$i d H_\mu(n) = F_\mu(n) d\tau \tag{4.111}$$

to (coordinate) momentum space, and, instead of using a common time step $d\tau$ for each momentum component, consider using a larger time step for the low-momentum modes

[371] to move them farther. This method never proved effective enough to use in full-scale simulation. Modern versions of the Fourier acceleration scheme are under investigation. See, for example, [372].

Trivializing map
If we can find an invertible map of the gauge field U to a new field V,

$$U = \mathcal{F}(V), \tag{4.112}$$

such that the Jacobian of the transformation cancels the gauge action:

$$\det[\partial U_\mu(n)(V)/\partial V_\nu(m)]\exp[-S(U)] = 1, \tag{4.113}$$

then the path integral becomes trivial [373–375]. Lüscher describes this as a "map to the strong-coupling limit" and discusses possible maps for the pure gauge action. Of course, finding such a map is entirely nontrivial, but if one can at least find one that moves the action partially toward strong coupling, then one could construct a hybrid Monte Carlo scheme that updates the gauge field according to the recipe

$$U \to V \to V' \to U' \tag{4.114}$$

where the $V \to V'$ step uses standard gauge evolution for the transformed gauge field V. This stronger coupling evolution would suffer less from critical slowing down. Recently, there have been efforts to find such a map using machine-learning methods. See, for example, Ref. [376].

4.2.6 Personal remarks

I first learned about the lattice formulation of QCD and its virtues when Ken Wilson gave a seminar at the MIT Center for Theoretical Physics around the time he was developing his lattice formulation. I was quite impressed with how easily confinement, in the form of an area law for Wilson loops, emerged in the strong-coupling regime. But I wasn't as brave or savvy as Creutz in proceeding to develop numerical methods for working out the nonperturbative consequences of Wilson's formulation. I didn't turn to numerical lattice calculations until shortly after Creutz's seminal papers. For the rest of my career, I have enjoyed participating in and contributing to the remarkable progress in this field. As a graduate student schooled in the analytic S-matrix and bootstrap, I was pleased when I could make a strong-interaction prediction to an accuracy of 25%, based on phenomenological considerations. There was always the inevitable doubt about the validity of the methods. Today, in some cases, we are able to obtain per mille accuracy for aome hadronic properties. Furthermore, we have little doubt that our results are a correct prediction of the Standard Model, since our methods are grounded in first-principles. That has been enormously satisfying.

4.3 Vacuum structure and confinement

Derek Leinweber

4.3.1 Introduction

The self interactions of gluons make the empty vacuum unstable to the formation of quark and gluon field configurations which permeate spacetime. These ground-state QCD-vacuum field configurations form the foundation of matter. Lattice QCD simulations enable first principles explorations of this nontrivial vacuum field structure.

These gluon field configurations form the foundation of every lattice QCD calculation. Each field configuration on its own contains a rich diversity of emergent nonperturbative structure. It is the process of averaging over thousands of field configurations that restores the translational invariance of the vacuum. Each field configuration with its own rich structure is uncorrelated with other configurations considered in the averaging process.

Deep insight into the mechanisms giving rise to the observed quantum phenomena can be obtained through the visualization of these complex scientific data sets constructed in Lattice QCD simulations, insights that would otherwise remain hidden in the typical gigabyte data sets of modern quantum field theory.

The essential, fundamentally-important, nonperturbative features of the QCD vacuum fields are: the dynamical generation of mass through chiral symmetry breaking, and the confinement of quarks. But what are the fundamental mechanisms of QCD that underpin these phenomena? What aspect of the QCD vacuum causes quarks to be confined? Which aspect is responsible for dynamical mass generation? Do the underlying mechanisms share a common origin?

In this brief review, we will address these questions in a chronological manner to convey the progress in developing an understanding of the essential mechanisms underpinning the phenomena of QCD.

4.3.2 Nonperturbative vacuum structure

Among the earliest of vacuum-structure visualizations are images of the Euclidean action density, or energy density

$$S_E(\vec{x},t) = \frac{1}{2}\,F^{ab}_{\mu\nu}(\vec{x},t)\,F^{ba}_{\mu\nu}(\vec{x},t), \tag{4.115}$$

$$= \mathrm{Tr}\left(\vec{E}^2(\vec{x},t) + \vec{B}^2(\vec{x},t)\right), \tag{4.116}$$

where $F^{ab}_{\mu\nu}$ is the Euclidean field strength tensor

$$F^{ab}_{\mu\nu} = \partial_\mu A^{ab}_\nu - \partial_\nu A^{ab}_\mu + ig[A^{ab}_\mu, A^{ba}_\nu], \tag{4.117}$$

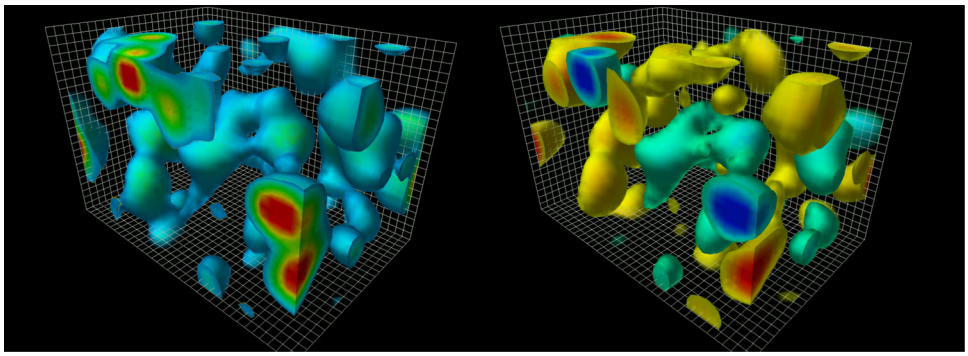


Fig. 23 Frames from the animation of Ref. [377] illustrating the Euclidean action density or energy density of Eq. (4.116) (left) and the corresponding topological charge density of Eq. (4.118) (right) at an instant in time. The spatial volume is approximately 2.4 by 2.4 by 3.6 fm

with color indices $a,\ b = 1,\ 2,\ 3$. The corresponding topological charge density proportional to $\vec{E}(\vec{x},t)\cdot\vec{B}(\vec{x},t)$

$$q(\vec{x},t) = \frac{g^2}{32\,\pi^2}\,\epsilon_{\mu\nu\rho\sigma}\,F^{ab}_{\mu\nu}(\vec{x},t)\,F^{ba}_{\rho\sigma}(\vec{x},t), \tag{4.118}$$

is also of interest as it characterizes the profile of instantons, nontrivial solutions of the classical Yang–Mills equations, discussed in further detail in Sect. 5.11.4.

Reference [378] provides one of the earliest observations of instanton-like objects in lattice gauge-field configurations. Here cooling with the standard Wilson action was used to suppress short-distance field fluctuations enabling the observation of long-distance structures.

However a problem with the use of the standard Wilson action or even the $\mathcal{O}(a^2)$-improved plaquette plus rectangle action is that the lattice action of an instanton can be reduced by shrinking the size of an instanton [379] through lattice-spacing errors. Instantons shrink under cooling with these lattice actions and "fall through the lattice." This led to the development of highly-improved actions [380,381] eliminating errors to $\mathcal{O}(a^4)$ or even over-improved actions where improvement terms are tuned to stabilize instantons, ensuring their stability under smoothing algorithms [379,382].

The results presented in this section are based on pure SU(3) gluon fields created with the standard Wilson action at $\beta = 6.0$ on a $24^3\times 36$ lattice with a lattice spacing, $a \simeq 0.1$ fm. The first coordinate of the Euclidean lattice was used for the time axis creating a $24^2 \times 36$ spatial volume. It is these calculations [383] that captured the attention of Prof. Frank Wilczek as he prepared his 2004 Nobel Prize lecture. Reference [384] provides a link to the *QCD Lava Lamp* animation that appeared in his Nobel Lecture [385]. In support of the Nobel Lecture a web page incorporating the best algorithms and visualization techniques of the time was created [386]. Parallel spatially-uniform $\mathcal{O}(a^4)$-improved smoothing algorithms [387] and an $\mathcal{O}(a^4)$-improved lattice field strength tensor [380] were formulated to accurately retain and present the long-distance nonperturbative properties of the ground-state vacuum fields. These images and animations [377,386] have since appeared in popular-science publications, leading YouTube channels [388,389], etc. [390].

Figure 23 displays two frames from the animation of Ref. [377]. Here 25 sweeps of three-loop, mean-field, $\mathcal{O}(a^4)$-improved cooling has been applied. Areas of high energy density are rendered in red and regions of moderate energy density are rendered in blue. The lowest energy densities are not rendered such that one can see into the volume. Similarly the topological charge density has regions of positive density rendered in red through yellow and regions of negative density rendered blue through cyan. While instanton-like objects are manifest, current research is examining the extent to which instanton-dyon degrees of freedom [391], i.e. fractionally charged regions, can be observed within these field configurations.

To directly view the the eight chromo-electric and eight chromo-magnetic gluon fields composing the vacuum, one must select a gauge. Figure 24 presents a stereoscopic illustration of one of the chromo-magnetic fields in Landau gauge [392]. Here the color and length of the arrows describe the magnitude of the vector fields. Animations of the fields are also available [377].

To see the 3D image of Figs. 24 and 31, try the following:

1. If you are viewing the image on a monitor, ensure the image width is 12 to 13 cm.
2. Bring your eyes very close to one of the image pairs.
3. Close your eyes and relax.
4. Open your eyes and allow the (blurry) images to line up. Tilting your head from side to side will move the images vertically.
5. Move back slowly until your eyes are able to focus. There's no need to cross your eyes!

With its lattice implementation of chiral symmetry, the overlap-Dirac operator provided a new approach to the exploration of the nonperturbative structure of the vacuum without resorting to smoothing algorithms [341]. Here low-lying Dirac eigenmode densities could be used to construct the topological charge density with the level of smoothness inversely related to the number of low-lying modes one considers [393]. Strong correlation with the instanton-like objects observed via smoothing algorithms was observed.

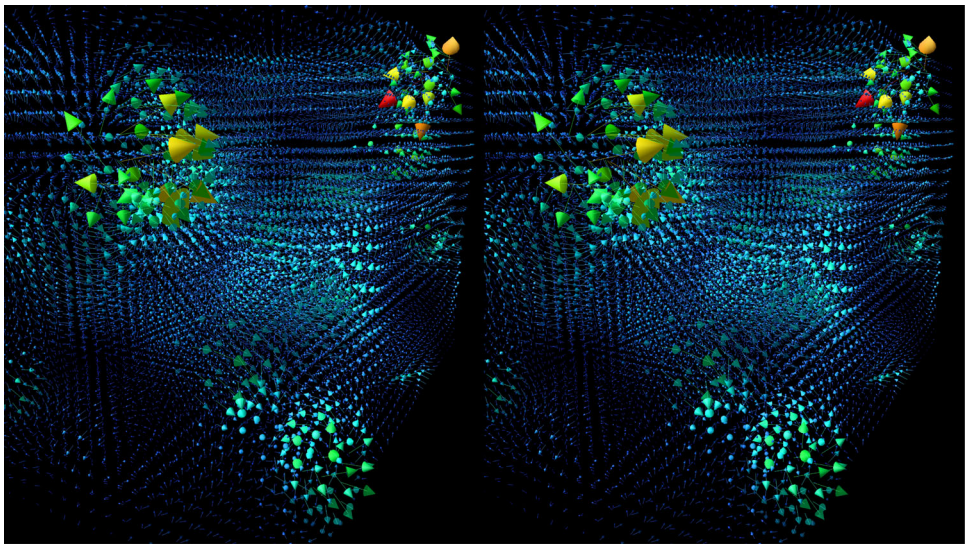


Fig. 24 Stereoscopic image of one of the eight chromo-magnetic fields composing the nontrivial vacuum of QCD. Hints for stereoscopic viewing are provided in the text

The zero modes are chiral and are distributed across topological charge regions of a unique sign. Low-lying eigenmode densities are also highly correlated with the topological structures revealed under smoothing [393]. These correlations between gluonic and fermionic structures expose the dynamics underpinning dynamical chiral symmetry breaking and the origin of mass.

The manner in which the topological charge density is rendered can lead to rather different views on the nature of how topological charge is distributed in the vacuum. Figure 25 illustrates two different renderings of the same topological charge density. The sheet-like structure associated with the sign-changing nature of the topological charge density correlator $\langle\, q(0)\, q(x)\,\rangle$ [394,395] is manifest when all magnitudes of the topological charge density are rendered down to zero. This is the celebrated sheet-like structure of the topological charge density [396]. However, when the rendering is restricted to larger values, one reveals a more lumpy structure with regions of significant coherent topological charge density.

More recently explorations of correlations between QCD phenomena and QED phenomena have commenced drawing on QCD+QED lattice simulations [290,397,398]. First results [399,400] and links to associated animations are reported in Ref. [390].

4.3.3 Center cluster structure of QCD vacuum fields

Further insight into the structure of QCD vacuum fields, their temperature dependence, and their evolution under Monte-Carlo evolution can be obtained through the consideration of the local Polyakov loop. The expectation value of the Polyakov loop is related to the finite temperature phase transition in QCD. It has an expectation value of zero in the confined phase and becomes nonzero in the deconfined phase.

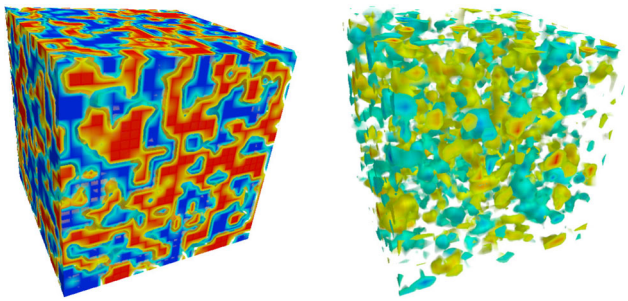


Fig. 25 The short-distance sheet-like structure of the vacuum is made apparent in the left-hand illustration by rendering all magnitudes of the topological charge density down to zero. Negative charge density is rendered green to blue, and positive charge density is yellow to red. The same data is rendered in the right-hand plot, this time only rendering the regions having large topological charge density, revealing a structure of topological lumps

The local Polyakov loop is the traced gauge-invariant product of time-oriented gauge links around the time extent of the lattice at each spatial point

$$L(\vec{x}) = \mathrm{Tr} \prod_{t=1}^{N_t} U_4(t, \vec{x}) = \rho(\vec{x})\, e^{i\phi(\vec{x})}, \tag{4.119}$$

Here, U_4 is the time-oriented link variable on a lattice with lattice spacing a, given by

$$U_\mu(x) = P \exp\left(ig \int_x^{x+\hat{\mu}a} dx^\mu A_\mu(x) \right). \tag{4.120}$$

Center clusters [401,402] are defined in terms of $L(\vec{x})$. They are regions of space where the local Polyakov loop prefers a single complex phase associated with the center of SU(3). The deconfinement transition occurs through the growth of a center cluster.

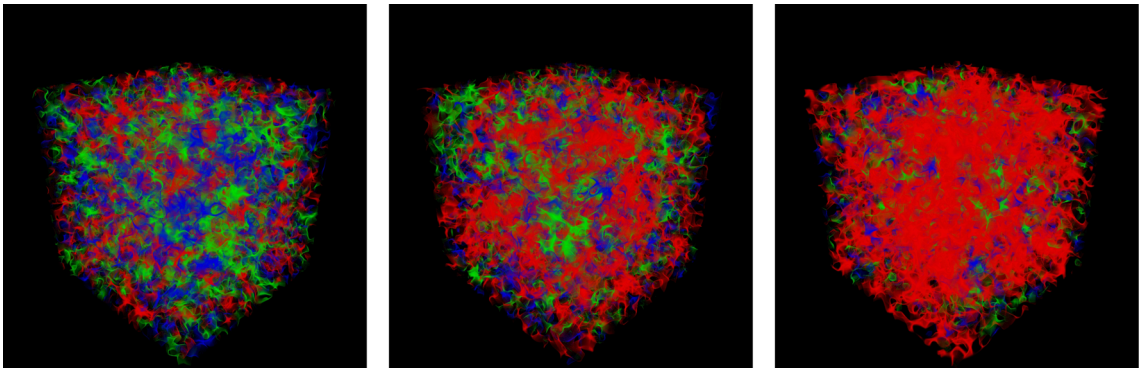


Fig. 26 Center clusters on a gauge field configuration at $T = 0.89(1)\,T_C$ (left), $T = 1.14(2)\,T_C$ (middle), and $T = 1.36(2)\,T_C$ (right), This rendering from Ref. [402] is based on the proximity of the local Polyakov loop phase, $\phi(\vec{x})$, to one of the three center phases of SU(3). The length of each side of the cubic volume is 2.4 fm. The percolation of the red phase in the middle and right-hand plots illustrates the deconfinement of quarks above T_C

In the final expression of Eq. (4.119), the local Polyakov loop is decomposed into a phase, $\phi(\vec{x})$ and a magnitude, $\rho(\vec{x})$. Both the proximity of the phase to one of the cube-roots of one and the magnitude are considered in visualizing the structure of the center domains of the gluon field. In either case, the most proximal cube root of one to the phase is indicated by the use of color.

In Ref. [402] an anisotropic gauge action was used to explore the evolution of coherent center domains in the gluon field under both temperature and the Hybrid Monte Carlo (HMC) update algorithm. To investigate the larger-scale behavior of the clusters, small scale noise is removed from the visualization by performing four sweeps of stout-link smearing [403] prior to calculating the Polyakov loops.

In Fig. 26, clusters are rendered where the phase $\phi(\vec{x})$ is within a small window around each center phase, and the rest of the volume is rendered transparent. Within these coherent center domains, color-singlet quark–antiquark pairs or three-quark triplets have a finite energy and are spatially correlated. Thus, these fundamental domains govern the size of the quark cores of hadrons. As one domain dominates the vacuum above the critical temperature, the correlation length diverges and quarks become deconfined.

The evolution of these clusters with HMC simulation time is presented in Ref. [405], showing how center clusters are slowly moving with correlations in the center clusters persisting for approximately 5 seconds corresponding to 25 HMC trajectories. The temperature dependence of the center-cluster structure is also explored in these animations where a single phase eventually dominates above the critical temperature, as illustrated in Fig. 26.

4.3.4 Flux tubes in QCD ground-state vacuum fields

Early seminal work on the static quark potential considered the transverse fluctuations of confining strings connecting static quark–antiquark pairs in non-Abelian gauge theories [406–408]. In the large L limit one finds a logarithmic relationship between the flux-tube width, σ, and its length, L with $\sigma^2(L/2) \sim \sigma_0^2 \ln(L/\lambda)$. The string model describes the divergence of the width as the flux-tube length $L \to \infty$ as arising from the large quantum mechanical fluctuations of a thin *bare* flux tube connecting the quark with the antiquark. This prediction was observed in a precise manner in a (2 + 1)-dimensional SU(2) Yang–Mills theory lattice calculation using a multi-level algorithm [409].

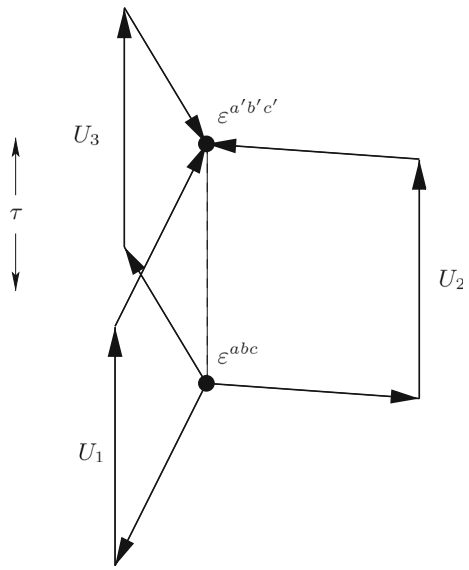


Fig. 27 Gauge-link paths for three static quark propagators, U_1, U_2, and U_3, are connected in a gauge-invariant manner via spatially smeared link paths. ε^{abc} and $\varepsilon^{a'b'c'}$ provide color anti-symmetrisation at the source and sink respectively, while τ indicates evolution of the three-quark system in Euclidean time [404]

Flux tubes in the QCD-vacuum fields of lattice QCD are revealed by examining the correlation between ground-state field properties and the positions of static quarks within the

fields. One begins with the standard approach of connecting static quark propagators by spatial-link paths in a gauge invariant manner. For mesonic systems, this is the standard Wilson loop. However, for baryonic systems one needs the structure illustrated in Fig. 27. The spatial link paths are typically broadened through a smearing algorithm to approximate the shape of the flux tube and thus obtain better overlap with the ground state potential of interest. While early calculations tuned the amount of smearing to provide optimal overlap with the ground state, more modern approaches create a basis of smeared sources and solve the generalized eigenvalue problem [410–412] to obtain the optimal combination of sources. The static quark propagators are constructed from time directed link products at fixed spatial coordinate, $\prod_i U_t(\vec{x}, t_i)$, using the untouched "thin" links of the gauge configuration.

The correlation of the gluon field with the static quark positions is characterized by the gauge-invariant Euclidean action density $S_E(\vec{y}, t)$ observed at spatial coordinate $\vec{y}$ and Euclidean time t measured relative to the origin of the three-quark Wilson loop. For the results presented herein, the action density is calculated using the highly-improved $\mathcal{O}(a^4)$ three-loop improved lattice field-strength tensor [380] on four-sweep APE-smeared gauge links [404].

Defining the quark positions as $\vec{r}_1$, $\vec{r}_2$ and $\vec{r}_3$ relative to the origin of the three-quark Wilson loop, and denoting the Euclidean time extent of the loop by τ, one evaluates the following correlation function

$$C(\vec{y}; \vec{r}_1, \vec{r}_2, \vec{r}_3; \tau) = \frac{\langle W_{3Q}(\vec{r}_1, \vec{r}_2, \vec{r}_3; \tau)\, S_E(\vec{y}, \tau/2)\rangle}{\langle W_{3Q}(\vec{r}_1, \vec{r}_2, \vec{r}_3; \tau)\rangle \langle S_E(\vec{y}, \tau/2)\rangle}, \tag{4.121}$$

where $\langle \cdots \rangle$ denotes averaging over configurations and translational/reflection/rotational lattice symmetries [404]. Note that the correlation is examined at the midpoint in the time evolution of the static quark propagation to ensure the three quark state has relaxed to its ground state form. For fixed quark positions and Euclidean time, C is a scalar field in three dimensions.

This measure has the advantage of being positive definite, eliminating any sign ambiguity on whether vacuum field fluctuations are enhanced or suppressed in the presence of static quarks. The correlation, C, is generally less than 1, signaling the expulsion of vacuum fluctuations from the interior of heavy-quark hadrons. In other words, flux tubes represent the suppression of the vacuum field fluctuations that form the foundation of matter.

Figure 28 provides an illustration of the correlation $C(\vec{y})$. For values of $\vec{y}$ well away from the quark positions $\vec{r}_i$, there are no correlations and $C \to 1$. As the separation between the quark–antiquark pair changes, the flux tube of Fig. 28 (top) gets longer, but the diameter of the tube and the depth of the expulsion remain approximately constant. As it costs energy to expel the vacuum field fluctuations, the confinement potential grows linearly as the quark separation increases.

Of historical significance was the endeavor to determine whether baryon flux tubes are Y-shape or Δ-shape (empty triangle) in nature. For the latter, the expectation was two-body tube-like structures around the edge of the three-quark system would dominate. Quantitative analyses of the static quark potential and the distribution of flux tubes led to a consensus [413] that the distribution is Y shape for large quark separations more than ~ 0.5 fm from the system center with the observation of filled Δ shapes at shorter-distance separations. The Y-shape ground state localizes at the Steiner point which minimizes the total string length.

The characteristic sizes of the flux-tube and node were quantified in Ref. [404]. The ground state flux-tube radius is ~ 0.4 fm with vacuum-field fluctuations suppressed by 7%. The node connecting the flux tubes is larger at 0.5 fm with a suppression of the vacuum action at 8%.

It is also of interest to consider flux-tube dynamics. Nontrivial flux-tube dynamics give rise to hybrid quarkonium states where excited gluon fields give rise to excited potentials between a static quark–antiquark pair. The energy spectrum of the excited gluon field was summarized in Refs. [414,415]. With the static potentials determined via lattice simulations, the spectrum of conventional and hybrid quarkonium states were found to be in good agreement with the spin-averaged experimental measurements of bottomonium states [414].

4.3.5 Flux tube string breaking in QCD

With the advent of numerical simulations incorporating the dynamics of light fermion loops in the QCD vacuum, the observation of flux-tube breaking or string breaking was keenly anticipated. The idea is that for increasing quark separations, eventually there would be enough energy in the flux tube joining the two static b quarks that it would become energetically favorable to break the string through the creation of a light quark–antiquark pair and the formation of two B mesons. Even to this day, this *implicit* form of string breaking has yet to be observed. The difficulty lies in the extraordinarily poor overlap of the two-B meson state with the spatial flux-tube operators used to create the string state.

This situation is in contrast to explorations of the structure of the $\Lambda(1405)$ baryon, where lattice-QCD calculations of the quark-sector contributions to the baryon magnetic moment indicate a molecular meson–baryon structure [416,417]. Here a three-quark operator carrying the quantum numbers of the $\Lambda(1405)$ have *implicitly* excited quark–antiquark pairs to form the five-quark molecule.

In the absence of implicit string breaking, Bali et al. [418] led the breakthrough in observing string breaking in QCD

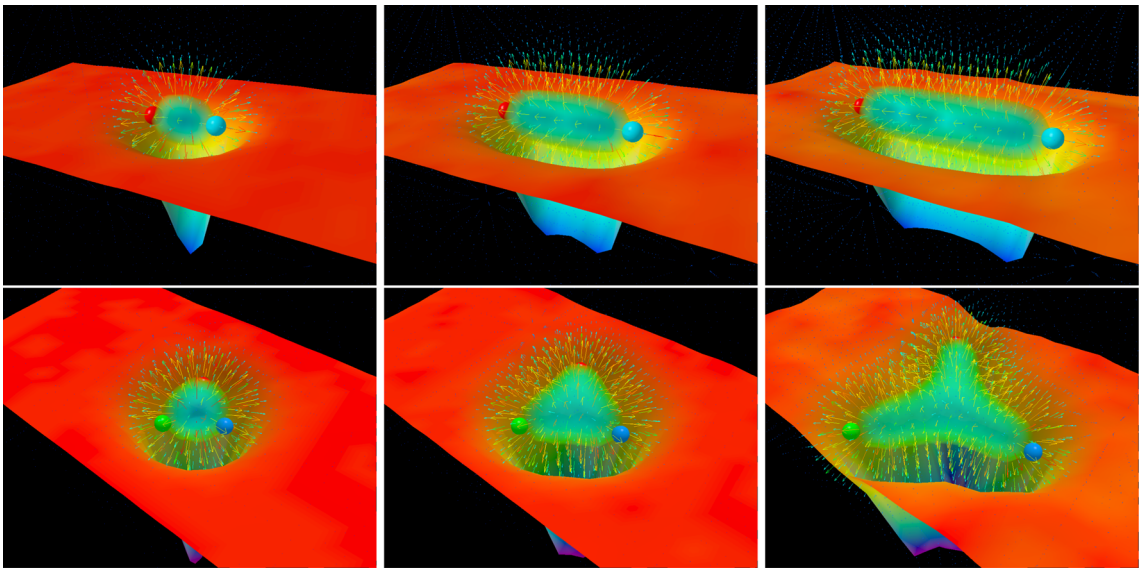


Fig. 28 The suppression of QCD vacuum fields, as represented by the energy density, from the region between a quark–antiquark meson (top) or three-quark baryon (bottom). Quark positions are illustrated by the colored spheres. The separation of the quarks in the meson are 0.50 fm (left), 1.00 fm (middle), and 1.50 fm (right). The baryon frames illustrate the spherical cavity (or bag) observed at small quark separations of 0.27 fm from the center (left), the development of a filled-Δ shape at moderate separations of 0.42 fm (middle) and finally the emergence of a Y-shape flux tube (right) at large quark separations of 0.72 fm from the system center [404]. The surface plot illustrates the reduction of the vacuum energy density in a plane passing through the centers of the quarks. The vector field illustrates the gradient of this reduction. The tube joining the quarks reveals the positions in space where the vacuum energy density is maximally expelled and corresponds to the "flux tube" of QCD

via a variational method with explicit B-meson operators. These interpolating fields mix with the traditional flux-tube operators in a matrix of correlation functions. Upon solving for the energy eigenstates, mixed states with their associated avoided level crossings are observed.

Following the notation of Ref. [418], the calculation proceeds as follows. The $Q\overline{Q}$ static quark operator connected with an optimized spatially smeared flux-tube operator $V_t(\mathbf{r}, \mathbf{0})$ from position $\mathbf{0}$ to $\mathbf{r}$ at Euclidean time t is

$$\overline{Q}_{(\mathbf{r},t)} \frac{\boldsymbol{\gamma} \cdot \mathbf{r}}{r} V_t(\mathbf{r}, \mathbf{0})\, Q_{(\mathbf{0},t)}, \tag{4.122}$$

where $\boldsymbol{\gamma} \cdot \mathbf{r}/r$ selects the spin-symmetric state to be combined with the symmetric gluonic string $V_t(\mathbf{r}, \mathbf{0})$, enabling mixing with two pseudoscalar B mesons. Note, the anti-symmetric spin-combination is obtained via $\boldsymbol{\gamma} \cdot \mathbf{r}/r \to \gamma_5$ and yields the same energy levels, as both spin cases are calculated from the same Wilson loop.

Similarly, the $B\overline{B}$ meson interpolating field for a pseudoscalar $\overline{B}$ meson at $\mathbf{r}$ and a B meson at $\mathbf{0}$ at Euclidean time t is

$$\overline{Q}_{(\mathbf{r},t)}\, \gamma_5\, q^i_{(\mathbf{r},t)}\, \bar{q}^i_{(\mathbf{0},t)}\, \gamma_5\, Q_{(\mathbf{0},t)}, \tag{4.123}$$

where $q^i_{(\mathbf{r},t)}$ annihilates the light-quark flavor, i. The four elements of the correlation matrix are obtained from the four combinations of these two operators.

Contracting the heavy-quark operators in the standard flux-tube operators provides

$$\left[\overline{Q}_{(\mathbf{r},t)} \frac{\boldsymbol{\gamma} \cdot \mathbf{r}}{r} V_t(\mathbf{r}, \mathbf{0})\, Q_{(\mathbf{0},t)}\right]^\dagger \overline{Q}_{(\mathbf{r},0)} \frac{\boldsymbol{\gamma} \cdot \mathbf{r}}{r} V_0(\mathbf{r}, \mathbf{0})\, Q_{(\mathbf{0},0)}$$
$$= 2\ \mathrm{tr}\left\{V_t^\dagger(\mathbf{r}, \mathbf{0})\, U_{\mathbf{r}}(t, 0)\, V_0(\mathbf{r}, \mathbf{0})\, U_{\mathbf{0}}^\dagger(t, 0)\right\} \equiv \square \tag{4.124}$$

where the heavy-quark mass dependence has been suppressed for simplicity. Here $U_{\mathbf{r}}(t, 0)$ denotes the product of time-oriented links at the position $\mathbf{r}$ from time 0 to t and the trace is over color indices. This is the standard Wilson loop depicted by the $\mathbf{r}$ (horizontally) by t (vertically) rectangle in Eq. (4.124).

Similarly, contracting out the quark field operators in the mixed correlator provides

$$\overline{Q}_{(\mathbf{0},t)}\, \gamma_5\, q^i_{(\mathbf{0},t)}\, \bar{q}^i_{(\mathbf{r},t)}\, \gamma_5\, Q_{(\mathbf{r},t)}\ \overline{Q}_{(\mathbf{r},0)} \frac{\boldsymbol{\gamma} \cdot \mathbf{r}}{r} V_0(\mathbf{r}, 0)\, Q_{(\mathbf{0},0)}$$
$$\equiv \square\!\!\sim\!\!\square = \square\!\!\sim\!\!\square \tag{4.125}$$

where the wavy line depicts a light quark operator. Finally, contraction of the quark operators in the $\overline{B}B$ correlator provides

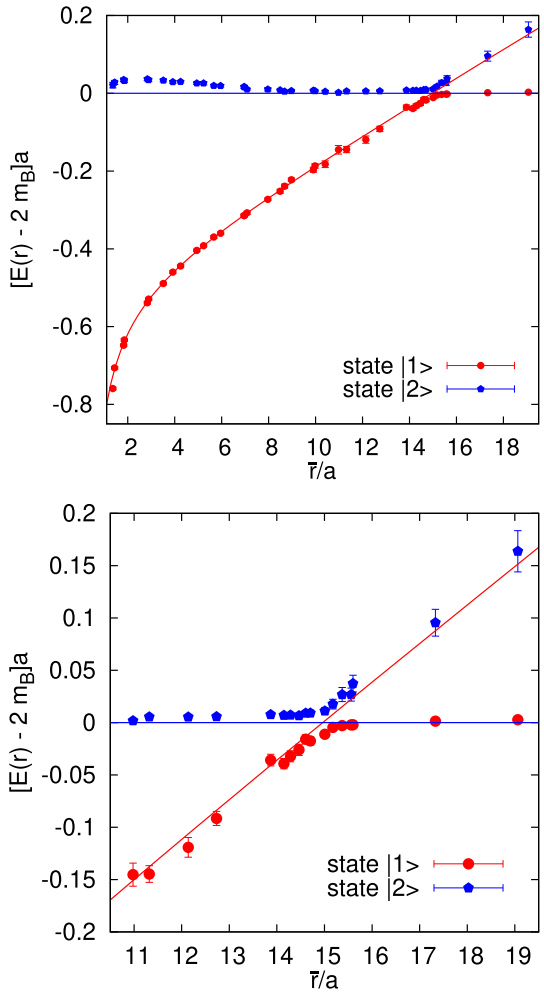


Fig. 29 From Ref. [418], the two energy levels obtained in the variational analysis are plotted as a function of the static quark–antiquark separation $\bar{r}/a$ with lattice spacing $a \approx 0.083$ fm. Energy values are relative to twice the mass of the B-meson, $2\,m_B$ (horizontal line). The curve corresponds to the three parameter fit of $E_1(r) = V_0 + \sigma r - e/r$, for $0.2\,\text{fm} \leq \bar{r} \leq 0.9\,\text{fm} < r_c$ with $r_c \approx 15\,a \approx 1.25$ fm.. The bottom plot zooms into the avoided level crossing

$$\overline{\mathcal{Q}}_{(\mathbf{0},t)}\,\gamma_5\,q^i_{(\mathbf{0},t)}\,\bar{q}^i_{(\mathbf{r},t)}\,\gamma_5\,\mathcal{Q}_{(\mathbf{r},t)}\;\overline{\mathcal{Q}}_{(\mathbf{r},0)}\,\gamma_5\,q^j_{(\mathbf{r},0)}\,\bar{q}^j_{(\mathbf{0},0)}\,\gamma_5\,\mathcal{Q}_{(\mathbf{0},0)}$$

$$\equiv \left(\delta_{ij}\;\; - \;\;\right). \tag{4.126}$$

Considering n_f fermion flavors, one finally arrives at the correlation matrix

$$C(t) = \left(\begin{array}{cc} & \sqrt{n_f} \\ \sqrt{n_f} & -n_f \; + \end{array} \right). \tag{4.127}$$

Calculation of the light-quark propagators demands the use of all-to-all techniques. Reference [418] used a truncated eigenmode approach, complemented by a stochastic estimator technique, improved by hopping parameter acceleration. Through the use of a tuned flux-tube operator and tuned smeared-local quark propagators in the meson operators, the correlation matrix is parameterized in terms of two low lying energy eigenstates and solved.

Figure 29 illustrates the two energy levels obtained in the $n_f = 2$ analysis of Ref. [418]. Remarkably, the region of mixing is small and the energy associated with the mixing is subtle. The analysis has since been extended to $2 + 1$ light+strange fermion flavors in Ref. [419] where both B and B_s mesons participate in the mixing.

These results reflect the diverse nature of these two states. Indeed with so little overlap between the two states away from the avoided crossing region, a string-oriented system may evolve such that it maintains the string structure at very large separations [420]. In this "sudden approximation," the system evolves along the red lines of Fig. 29 providing a pathway to extraordinarily high energy excitations. The subsequent decay is considered "adiabatic" [420] where hadrons then follow the energy-eigenstate curves and split into fragments.

4.3.6 Impact of dynamical fermions on vacuum field structure

With the advent of full QCD simulations incorporating the effects of light dynamical-fermion flavors, attention turned to understanding how these light fermion loops in the vacuum changed the QCD ground-state structure. Drawing on gauge fields from the MILC collaboration [421,422], advances in instanton-preserving smoothing algorithms [382] were deployed to reveal the impact of dynamical fermions on the topological charge density of the gauge fields [395].

The MILC simulations were performed using a one-loop Symanzik improved gauge action and an improved Kogut-Susskind quark action. Using the static quark potential, the lattice spacings were determined and tuned to be the same in all the runs to better expose differences due to dynamical fermions. At large distances, screening of the string tension was observed for light dynamical flavors [421,422].

Figure 30 illustrates the topological-charge densities revealed following four sweeps of over-improved stout-link smearing [395]. The top illustration from quenched QCD, is qualitatively different from the lower illustration for a $2 + 1$ flavor dynamical-fermion configuration.[10] The zero modes

[10] In the top illustration, one can see through the bulk of the topological charge distribution and observe the white background and the dotted lattice grid lines. This is not the case in the lower illustration where the topological charge fills out the space. Only a sprinkling of white space is observed. The quark-mass dependence of the dynamical-fermion illustration is subtle [395] indicating that the qualitative differences in the distributions comes about through the introduction of dynamical fermions in generating the configurations through

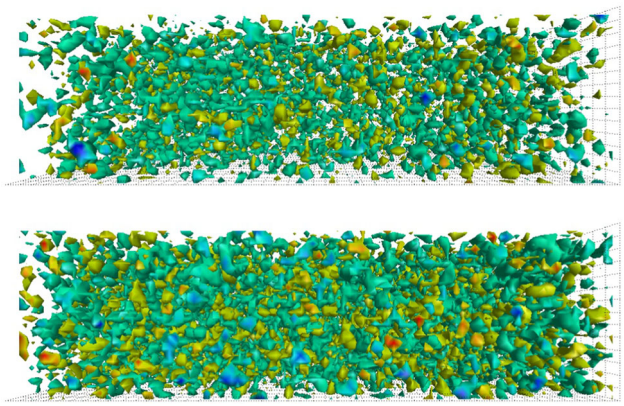


Fig. 30 The topological charge density from Ref. [395] for the quenched (top) and the light-quark dynamical ensemble from the MILC Collaboration [421,422], with dynamical masses of $am_{u,d} = 0.0062$, $am_{=}0.031$

associated with well-separated topological objects act to suppress the fermion determinant, such that the top configuration is improbable in full QCD. In the full-QCD simulations, the topological objects grow in size and number [395] to suppress the zero modes.

4.3.7 Center vortex structure of QCD vacuum fields

The essential, fundamentally-important, nonperturbative features of the QCD vacuum fields are the dynamical generation of mass through chiral symmetry breaking, and the confinement of quarks. But what is the fundamental mechanism of QCD that underpins these phenomena?

One of the most promising candidates is the center vortex perspective of QCD vacuum structure. While the ideas of a center-vortex dominated vacuum were laid down long ago [423–425], it wasn't until 1997 when Jeff Greensite, Manfried Faber, et al. demonstrated that lattice QCD techniques could be used to explore the importance of these ideas [426–431]. Indeed by the end of the millennium, the field had attracted broad interest with a comprehensive review in 2003 [432].

This perspective describes the nature of the nontrivial vacuum in terms of the most fundamental center of the gauge group. Herein our focus is on the $SU(3)$ gauge group where center vortices are characterized by the three center phases, $\sqrt[3]{1}$.

Footnote 10 Continued
Monte-Carlo methods. Not only are the objects in the quenched simulation further apart, a statistical analysis indicates there are fewer objects and the objects themselves are smaller in size when compared with the dynamical fermion distributions [395]. The physics underpinning these differences in the topological charge density distributions can be understood in terms of the modes of the Dirac operator generated by these distributions.

By identifying center vortices within the ground-state fields and then removing them, a deep understanding of their contributions has been developed. Removal of center vortices from the ground-state fields results in a loss of dynamical mass generation and restoration of chiral symmetry [433–435], a loss of the string tension [436–439], a suppression of the infrared enhancement in the Landau-gauge gluon propagator [437,440–442], and the possibility that gluons are no longer confined [442].

One can also examine the role of the center vortices alone. Remarkably, center vortices produce both a linear static quark potential [436,438,439,443,444] and infrared enhancement in the Landau-gauge gluon propagator [441,442]. The planar vortex density of center-vortex degrees of freedom scales with the lattice spacing providing a well defined continuum limit [436]. These results elucidate strong connections between center vortices and confinement.

A connection between center vortices and instantons was identified through gauge-field smoothing [444]. An understanding of the phenomena linking these degrees of freedom was illustrated in Ref. [445]. In addition, center vortices have been shown to give rise to mass splitting in the low-lying hadron spectrum [433,434,446].

Still, the picture in pure $SU(3)$ gauge theory is not perfect. The vortex-only string tension obtained from pure Yang–Mills lattice studies has been consistently shown to be about $\sim 60\%$ of the full string tension. Moreover, upon removal of center vortices the gluon propagator showed a remnant of infrared enhancement [441]. In short, within the pure gauge sector, the removal of long-distance non-perturbative effects via center-vortex removal is not perfect.

Understanding the impact of dynamical fermions on the center-vortex structure of QCD ground-state fields is a contemporary focus of the center-vortex field [435,438,439,442,447,448]. Herein, changes in the microscopic structure of the vortex fields associated with the inclusion of dynamical fermions are illustrated. The introduction of dynamical fermions brings the phenomenology of center vortices much closer to a perfect encapsulation of the salient features of QCD, confinement and dynamical mass generation through chiral symmetry breaking.

As such, it is interesting to ask, what do these center-vortex structures look like? To this end, we present visualizations of center vortices as identified on lattice gauge-field configurations. Some of these visualizations are presented as stereoscopic images. See the instructions provided in Sect. 4.3.2 for help in viewing these images.

Center vortex identification

Center vortices are identified through a gauge fixing procedure designed to bring the lattice link variables as close as possible to the identity matrix multiplied by a phase equal to one of the three cube-roots of 1. Here, the original

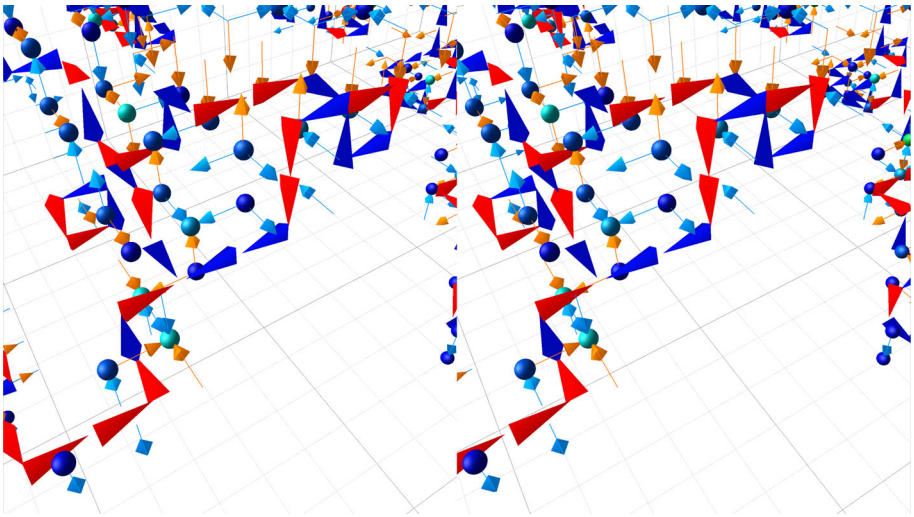


Fig. 31 Stereoscopic image of center vortices as identified on the lattice from Ref. [447]. Vortex features including vortex lines (jets), branching points (3-jet combinations), crossing points (4 jets), indicator links (arrows) and singular points (spheres) are described in the text

Monte-Carlo generated configurations are considered. They are gauge transformed directly to Maximal Center Gauge [436,449,450]. This brings the lattice link variables $U_\mu(x)$ close to the center elements of $SU(3)$,

$$Z = \exp\left(\frac{2\pi i}{3}\, n\right) \mathbf{I}, \tag{4.128}$$

with $n = -1$, 0, or 1 enumerating the three cube roots of 1 such that the special property of $SU(3)$ matrices, $\det(Z) = 1$, is satisfied. One considers gauge transformations Ω such that,

$$\sum_{x,\mu} \left|\mathrm{tr}\, U^{\Omega}_{\mu}(x)\right|^2 \overset{\Omega}{\rightarrow} \max, \tag{4.129}$$

and then projects the link variables to the center

$$U_\mu(x) \rightarrow Z_\mu(x) \text{ where } Z_\mu(x) = \exp\left(\frac{2\pi i}{3}\, n_\mu(x)\right) \mathbf{I}. \tag{4.130}$$

Here, n has been promoted to a field, $n_\mu(x)$, taking a value of $-1, 0$, or 1 for each link variable on the lattice. In this way, the gluon field, $U_\mu(x)$, is characterized by the most fundamental aspect of the $SU(3)$ link variable, the center, $Z_\mu(x)$. In the projection step, eight degrees of freedom are reduced to one of the three center phases. This "vortex-only" field, $Z_\mu(x)$, can be examined to learn the extent to which center vortices alone capture the essence of nonperturbative QCD.

The product of these center-projected links, $Z_\mu(x)$, around an elementary 1×1 square (plaquette) on the lattice also produces a centre element of $SU(3)$. The value describes the center charge associated with that plaquette

$$z = \prod_{\square} Z_\mu(x) = \exp\left(2\pi i\, \frac{m}{3}\right), \; m = -1, 0, \text{ or } 1. \tag{4.131}$$

The most common value observed has $m = 0$ indicating that no centre charge pierces the plaquette. However, values of $m = \pm 1$ indicate that the center line of an extended three-dimensional vortex pierces that plaquette.

The complete center-line of an extended vortex is identified by tracing the presence of nontrivial center charge, $m = \pm 1$, through the spatial lattice. Figure 31 exhibits rich emergent structure in the nonperturbative QCD ground-state fields in a stereoscopic image. Here a 3D slice of the 4D space-time lattice is being considered at fixed time. Features include:

Vortex Lines:
The plaquettes with nontrivial center charge, characterized by $m = +1$ or -1, are plotted as jets piercing the center of the plaquette. Both the orientation and color of the jets reflect the value of the non-trivial center charge. Using a right-hand rule for the direction, plaquettes with $m = +1$ are illustrated by blue jets in the forward direction, and plaquettes with $m = -1$ are illustrated by red jets in the backward direction. Thus, the jets show the directed flow of $m = +1$ center charge, $z = e^{2\pi i/3}$, through spatial plaquettes. They are analogous to the line running down the center of a vortex in a fluid.

Vortices are somewhat correlated with the positions of significant topological charge density, but not in a strong manner [445]. However, the percolation of vortex structure is significant and the removal of these vortices destroys most instanton-like objects.

Branching Points or Monopoles:
In $SU(3)$ gauge theory, three vortex lines can merge into or emerge from a single point. Their prevalence is surprising, as is their correlation with topological charge density [445].

Vortex Sheet Indicator Links:
As the vortex line moves through time, it creates a vortex sheet in 4D spacetime. This movement is illustrated by arrows along the links of the lattice (shown as cyan and orange

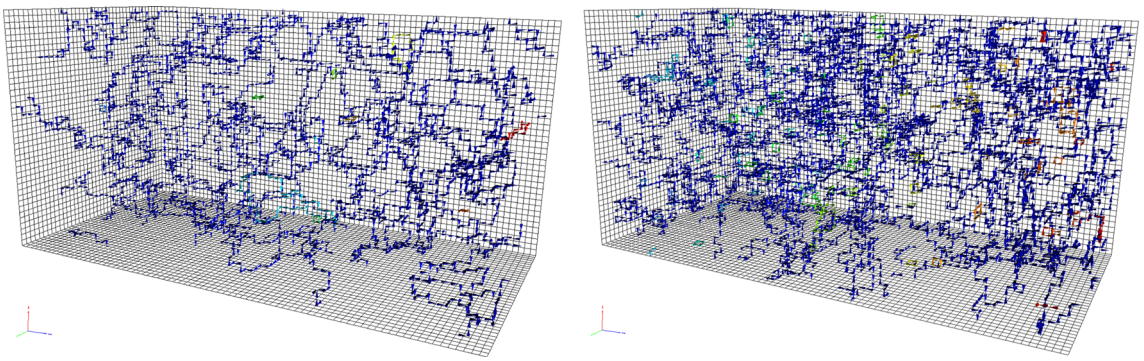


Fig. 32 From Ref. [448], the center-vortex structure of a ground-state vacuum field configuration in pure SU(3) gauge theory (left) is compared with a field configuration in dynamical 2+1 flavor QCD corresponding to $m_\pi = 156$ MeV (right). The flow of +1 center charge through the gauge fields is illustrated by the jets. Blue jets are used to illustrate the single percolating vortex structure, while other colors illustrate smaller structures

arrows in Fig. 31) indicating center charge flowing through space-time plaquettes in the suppressed time direction.

Singular Points:
When the vortex sheet spans all four space-time dimensions, it can generate topological charge. Lattice sites with this property are called singular points [431,451–453] and are illustrated by spheres. The sphere color indicates the number of times the sheet adjacent to a point can generate a topological charge contribution [445].

Reference [448] presents the first results demonstrating the impact of dynamical fermions on the center-vortex structure of QCD ground-state fields. There matched lattices were considered, one in pure-gauge and the other a $2+1$-flavor dynamical-fermion lattice from the PACS-CS Collaboration [454]. These $32^3 \times 64$ lattice ensembles employ a renormalisation-group improved Iwasaki gauge action and non-perturbatively $\mathcal{O}(a)$-improved Wilson quarks, with $C_{\rm SW} = 1.715$.

The lightest u- and d-quark-mass ensemble identified by a pion mass of 156 MeV [454] is presented here. The scale is set using the Sommer parameter [455] with $r_0 = 0.4921$ fm providing a lattice spacing of $a = 0.0933$ fm [454]. A matched $32^3 \times 64$ pure-gauge ensemble using the same improved Iwasaki gauge action with a Sommer-scale spacing of $a = 0.100$ fm was created [448] to enable comparisons with the PACS-CS ensembles.

The center-vortex structure of pure-gauge and dynamical fermion ground-state vacuum fields is illustrated in Fig. 32 from Ref. [448], where interactive 3D plots of this structure which can be activated in Adobe Reader. The impact of dynamical fermions on the center-vortex structure is much more significant than that discussed in Sect. 4.3.6.

In both illustrations, the vortex structure is dominated by a single large percolating structure. Whereas small loops will tend to pierce a Wilson loop twice with zero effect, it is this extended structure that gives rise to a net vortex piercing of a Wilson loop and the generation of an area law associated with confinement. These two illustrations are representative of the ensemble in that the vortex structure is typically dominated by a single large percolating cluster.

Closer inspection reveals a continuous flow of center charge, often emerging or converging to monopole or anti-monopole vertices where three jets emerge from or converge to a point. These are referred to as branching points, as a $+1$ center charge flowing out of a vertex is equivalent to $+2$ center charge flowing into the vertex and subsequently branching to two $+1$ jets flowing out of the vertex.

With the introduction of dynamical fermions, the structure becomes more complex, both in the abundance of vortices and branching points. The average number of vortices composing the primary cluster in these $32^2 \times 64$ spatial slices roughly doubles from $\sim$ 3000 vortices in the pure gauge theory to $\sim$ 6000 in full QCD. Still, there are $32^2 \times 64 \times 3 = 196{,}608$ spatial plaquettes on these lattices and thus the presence of a vortex is a relatively rare occurrence.

By counting the number of vortices between branching points one discovers the distribution is exponential, indicating a constant branching probability. This probability is higher in full QCD by a ratio of $\sim 3/2$.

With an understanding of the impact of dynamical-fermion degrees of freedom on the center-vortex structure of ground-state vacuum fields, attention has turned to understanding the impact on confinement. In a variational analysis of standard Wilson loops composed of several spatially-smeared sources to isolate the ground state potential, the static quark potential has been calculated on three ensembles including the original untouched links, $U_\mu(x)$, the vortex-only links, $Z_\mu(x)$, and vortex-removed links, $Z^\dagger_\mu(x)\, U_\mu(x)$ [442] where the multiplication of the conjugate of the centre-projected field ensure all plaquettes have $z = 0$.

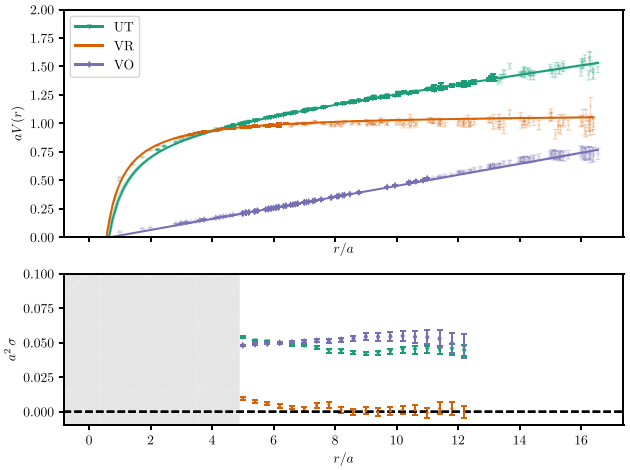


Fig. 33 The static quark potential, as presented in Ref. [438], calculated on the vortex-modified dynamical-fermion ensemble, corresponding to a pion mass of 156 MeV. The lower plot shows the local slope from linear fits of the potentials in the upper plot over a forward-looking window from r to $r + 4a$

For the original untouched configurations, the static quark potential is expected to follow a Cornell potential

$$V(r) = V_0 - \frac{\alpha}{r} + \sigma\, r. \tag{4.132}$$

As center vortices are anticipated to encapsulate the nonperturbative long-range physics, the vortex-only results should give rise to a linearly rising potential. On the other hand, the vortex-removed results are expected to capture the short-range Coulomb behavior. Figure 33 from Ref. [442] illustrates the static quark potentials obtained from these three ensembles for the dynamical $2 + 1$-flavor ensemble with a pion mass of 156 MeV [454].

Qualitatively, center vortices account for the long-distance physics. The removal of center vortices completely removes the confinement potential. And while the vortex-only string tension is typically 60 % of the original string tension in the pure gauge sector, the introduction of dynamical fermions has improved the vortex-only phenomenology significantly. Vortices alone capture both the screening of the pure-gauge string tension and the full string tension of the original untouched ensemble. This result is associated with the significant modification of the center-vortex structure of ground-state vacuum fields induced by dynamical fermions.

The improved separation of perturbative and nonperturbative physics through the consideration of vortex-removed and vortex-only ensembles in full QCD is also manifest in the nonperturbative gluon propagator [442]. This time vortex removal removes the infrared enhancement of the gluon propagator, leaving a tree-level structure. Indeed the vortex-removed Euclidean correlator remains positive definite, admitting the possibility of a positive-definite spectral density associated with free gluons. The vortex-only ensembles capture the infrared enhancement of the gluon propagator and the screening of this enhancement in full QCD [442].

Similarly, dynamical mass generation in the nonperturbative quark propagator is suppressed under vortex removal in full QCD while the vortex-only ensemble provides dynamical mass generation [435]. While explicit chiral symmetry breaking through the quark mass, leaves a remnant of dynamical mass generation, it is anticipated that for sufficiently light current quark masses, chiral symmetry will be restored [434] and dynamical mass generation will be completely eliminated in the vortex-removed theory.

In summary, center-vortex structure is complex. Each ground-state configuration is dominated by a long-distance percolating center-vortex structure. In $SU(3)$ gauge field theory, a proliferation of branching points is observed, with further enhancement as light dynamical fermion degrees of freedom are introduced in simulating QCD. There is an approximate doubling in the number of nontrivial center charges in the percolating vortex structure as one goes from the pure-gauge theory to full QCD. Increased complexity in the vortex paths is also observed as the number of branching points is significantly increased with the introduction of dynamical fermions. In short, dynamical-fermion degrees of freedom radically alter the center-vortex structure of the ground-state vacuum fields. This change in structure acts to improve the phenomenology of center vortices better reproducing the string tension, dynamical mass generation and better removing nonperturbative physics under vortex removal. This represents a significant advance in the ability of center vortices to capture the salient nonperturbative features of QCD.

4.3.8 Summary

In the 50 years following the advent of QCD, the complexity of the nontrivial QCD vacuum has been exposed. Many theoretical ideas have been created and developed to explain the salient features of this nontrivial vacuum and their exploration continues. Numerical experiments within the realm of lattice QCD have been particularly useful in testing the veracity of the theoretical ideas proposed. Today, these numerical experiments are exploring the ideas of instanton-dyons and center-vortices as the essential features of QCD vacuum structure, confining color and dynamically generating mass through dynamical chiral symmetry breaking. The results are fascinating, and encourage further exploration of the essence of QCD vacuum structure.

4.4 QCD at non-zero temperature and density

Frithjof Karsch

4.4.1 QCD thermodynamics on Euclidean lattices

The path integral formulation of QCD can easily be applied to cases of non-vanishing temperature (T) and other external control parameters, e.g. the chemical potentials (μ_f) that couple to the conserved currents of quark-flavor number.

Using the lattice regularization scheme of QCD, introduced by K. Wilson [97], QCD thermodynamics is formulated on Euclidean space-time lattices of size $N_\sigma^3 N_\tau$ where, for a given lattice spacing (a), the lattice extent in Euclidean time controls the temperature $T = 1/N_\tau a$ and the spatial extent is related to the volume of the thermodynamic system, $V = (N_\sigma a)^3$. The chemical potentials enter directly in the fermion matrices, M_f, which arise from the QCD Lagrangian after integrating out the fermion fields.

Bulk thermodynamics can then be derived from the lattice regularized partition function,

$$Z = \int \prod_{x_0=1}^{N_\tau} \prod_{x_i=1}^{N_\sigma} \prod_{\nu=\hat{0}}^{3} \mathcal{D}U_{x,\hat{\nu}}\, e^{-S_G} \times \prod_{f=u,d,s..} \det M_f(m_f, \mu_f), \tag{4.133}$$

where $x = (x_0, \vec{x})$ labels the sites of the 4-dimensional lattice, S_G denotes the gluonic part of the Euclidean action, which is expressed in terms of $SU(3)$ matrices $U_{x,\hat{\nu}}$ and M_f is the fermion matrix for quark flavor f. It is a function of quark mass, m_f and flavor chemical potential $\hat{\mu}_f \equiv \mu_f/T$. Basic bulk thermodynamic observables (equation of state, susceptibilities, etc.) can then be obtained from the logarithm of the partition function, Z, which defines the pressure, P, as

$$P/T = \frac{1}{V} \ln Z(T, V, \vec{\mu}, \vec{m}) \,. \tag{4.134}$$

Applying standard thermodynamic relations one obtains other observables of interest; e.g. the energy density is related to the trace anomaly of the energy–momentum tensor, $\Theta^{\mu\mu}$,

$$\frac{\Theta^{\mu\mu}}{T^4} = \frac{\epsilon - 3P}{T^4} \equiv T \frac{\partial P/T^4}{\partial T} \,, \tag{4.135}$$

and the conserved charge densities are obtained as,

$$\frac{n_X}{T^3} = \frac{\partial P/T^4}{\partial \hat{\mu}_X} \quad , \quad X = B,\ Q,\ S \,. \tag{4.136}$$

While the framework of lattice QCD provides easy access to QCD thermodynamics at vanishing values of the chemical potentials, major difficulties arise at $\mu_f \neq 0$. The fermion determinants, $\det M_f(m_f, \mu_f)$, are no longer positive definite when the real part of the chemical potential is non-zero, $\mathrm{Re}\hat{\mu}_f \neq 0$. This includes the physically relevant case of strictly real chemical potentials. The presence of a complex valued integrand in the path integral makes the application of standard Monte Carlo techniques, which rely on a probabilistic interpretation of integration measures, impossible. The two most common approaches to circumvent this problem are to either (i) perform numerical calculations at imaginary values of the chemical potential, $\hat{\mu}_f^2 < 0$ [456,457], or to (ii) perform Taylor series expansions around $\hat{\mu}_f = 0$ [458,459]. In the former case numerical results need to be analytically continued to real values of μ_f. In the latter case the QCD partition function is written as,

$$P/T^4 = \frac{1}{VT^3} \ln Z(T, V, \vec{\mu}) = \sum_{i,j,k=0}^{\infty} \frac{\chi_{ijk}^{BQS}}{i!j!k!} \hat{\mu}_B^i \hat{\mu}_Q^j \hat{\mu}_S^k \,, \tag{4.137}$$

with $\chi_{000}^{BQS} \equiv P(T, V, \vec{0})/T^4$ and expansion coefficients,

$$\chi_{ijk}^{BQS}(T) = \left. \frac{\partial P/T^4}{\partial \hat{\mu}_B^i \partial \hat{\mu}_Q^j \partial \hat{\mu}_S^k} \right|_{\hat{\mu}=0} \,, \tag{4.138}$$

can be determined in Monte Carlo simulations performed at $\hat{\mu}_X = 0$.

The phase structure of QCD can be explored using suitable observables that are sensitive to the spontaneous breaking and the eventual restoration of global symmetries. They can act as order parameters in certain limits of the parameter space spanned by the quark masses. In QCD exact symmetries exist either in the chiral limit, i.e. at vanishing values of n_f quark masses, or for infinitely heavy quarks, i.e. in pure $SU(N_c)$ gauge theories, with N_c denoting the number of colors.

In order to probe the restoration of the global chiral symmetries one analyzes the chiral condensate and its susceptibilities,

$$\langle \bar{\chi}\chi \rangle_f = \frac{T}{V} \frac{\partial}{\partial m_f} \ln Z = \frac{T}{V} \langle \mathrm{Tr} M_f^{-1} \rangle \,, \tag{4.139}$$

$$\chi_m^{fg} = \frac{\partial \langle \bar{\chi}\chi \rangle_f}{\partial m_g} \,, \quad \chi_t^f = T \frac{\partial \langle \bar{\chi}\chi \rangle_f}{\partial T} \,. \tag{4.140}$$

The former is an order parameter for the restoration of the $SU(n_f)_L \times SU(n_f)_R$ chiral flavor symmetry of QCD and distinguishes, in the limit of vanishing quark masses, a symmetry broken phase at low temperature from a chiral symmetry restored phase at high temperature,

$$\lim_{m_\ell \to 0} \langle \bar{\chi}\chi \rangle_\ell \begin{cases} > 0 & T < T_\chi \\ = 0 & T \geq T_\chi \end{cases} \,. \tag{4.141}$$

Similarly one considers the Polyakov loop $\langle L \rangle$ and its susceptibility χ_L,

$$\langle L \rangle = \frac{1}{N_\sigma^3} \left\langle \sum_{\vec{x}} \mathrm{Tr} L_{\vec{x}} \right\rangle \,, \quad L_{\vec{x}} = \prod_{x_0=1}^{N_\tau} U_{(x_0,\vec{x}),\hat{0}} \,,$$

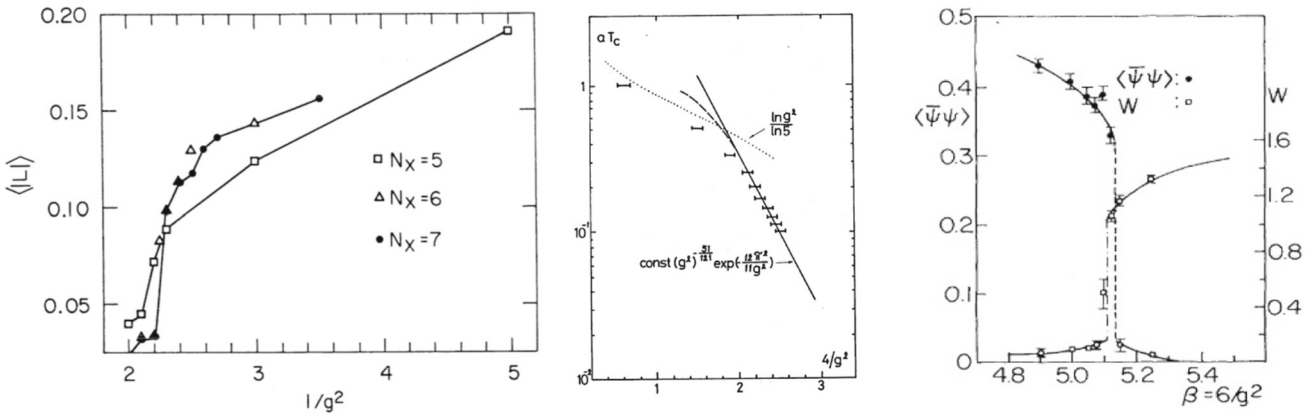


Fig. 34 First evidence for the existence of a deconfinement phase transition in $SU(2)$ gauge theories using the Polyakov loop expectation value as an order parameter (left) [460] and a first extrapolation of the phase transition temperature to the continuum limit (middle) [461] . The right hand figure shows a first comparison of the temperature dependence of the Polyakov loop ($W \equiv \langle |L| \rangle$) and chiral condensate ($\langle \bar{\psi}\psi \rangle$) order parameters in a SU(3) gauge theory [462]

$$\chi_L = N_\sigma^3 \left(\langle L^2 \rangle - \langle L \rangle^2 \right) , \tag{4.142}$$

to probe the breaking and restoration of the global $Z(N_c)$ center symmetry of pure $SU(N_c)$ gauge theories; i.e. $SU(N_c)$ gauge theories at finite temperature, formulated on Euclidean lattices, are invariant under global rotation of all temporal gauge field variables, $U_{\vec{x},\hat{0}} \rightarrow z U_{\vec{x},\hat{0}}$, with $z \in Z(N_c)$. The Polyakov loop expectation value vanishes as long as this center symmetry is not spontaneously broken.

The Polyakov loop expectation value also reflects the long distance behavior of Polyakov loop correlation functions,

$$|\langle L \rangle|^2 \equiv \lim_{|\vec{x}| \to \infty} G_L(\vec{x}) \begin{cases} = 0 \Leftrightarrow F_q = \infty, & T \leq T_d \\ > 0 \Leftrightarrow F_q < \infty, & T > T_d \end{cases} \tag{4.143}$$

where

$$G_L(\vec{x}) = \mathrm{e}^{-F_{\bar{q}q}(\vec{x},T)/T} = \langle \mathrm{Tr} L_{\vec{0}} \mathrm{Tr} L_{\vec{0}}^\dagger \rangle \tag{4.144}$$

is the correlation function of two Polyakov loops. It denotes the change in free energy (excess free energy, $F_{\bar{q}q}$), that is due to the presence of two static quark sources introduced in a thermal medium. At zero temperature this free energy reduces to the potential between static quark sources.

At least in the case of pure gauge theories this provides a connection between the confinement-deconfinement phase transition and the breaking of a global symmetry, the $Z(N_c)$ center symmetry of the $SU(N_c)$ gauge group. This symmetry, however, is explicitly broken in the presence of dynamical quarks with mass $m_f < \infty$. Unlike chiral symmetry restoration, deconfinement thus is not expected to be related to a phase transition in QCD with physical quark masses. Nonetheless, the consequences of deconfinement, related to the dissolution of hadronic bound states, becomes clearly visible in many thermodynamic observables.

4.4.2 *Early lattice QCD calculations at non-zero temperature*

Almost immediately after the formulation of QCD as the theory of strong interaction physics, its consequences for strong interaction matter at non-zero temperature were examined [463,464]. It rapidly became obvious that fundamental properties of QCD, confinement and asymptotic freedom on the one hand [464,465], and chiral symmetry breaking on the other hand [466], are likely to trigger a phase transition in strong interaction matter that separates a phase being dominated by hadrons as the relevant degrees of freedom from that of almost free quarks and gluons. The notion of a quark–gluon plasma was coined at that time [467].

Soon after these early, conceptually important developments it was realized that the formulation of QCD on discrete space-time lattices, which was introduced by K. Wilson as a regularization scheme in QCD [97], also provides a powerful framework for the analysis of non-perturbative properties of strong interaction matter through Monte-Carlo simulations [353]. This led to a first determination of a phase transition temperature in $SU(2)$ [460,461] and $SU(3)$ [462,468,469] gauge theories, and a first determination of the equation of state of purely gluonic matter [470,471]. The interplay between deconfinement on the one hand and chiral symmetry restoration on the other hand also was studied [462] early on and the question whether or not these two aspects of QCD may lead to two distinct phase transitions in QCD has been considered ever since. Some results from these first lattice QCD studies of the thermodynamics of strong interaction matter are shown in Fig. 34.

At physical values of the quark masses, neither deconfinement nor the effective restoration of chiral symmetry leads to a true phase transition. Still the transition from the low temperature hadronic to the high temperature partonic phase of

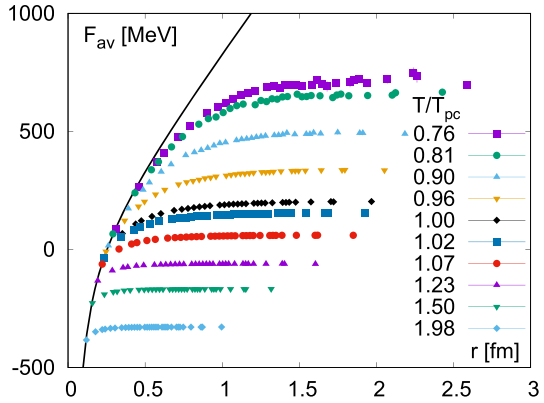


Fig. 35 The so-called color averaged, heavy quark free energy ($F_{\rm av} \equiv F_{\bar{q}q}$) in the vicinity of the pseudo-critical transition temperature (T_{pc}) in 2-flavor QCD [472]. Results shown cover a temperature range from $T/T_{pc} \simeq 0.75$ to $T/T_{pc} \simeq 2$

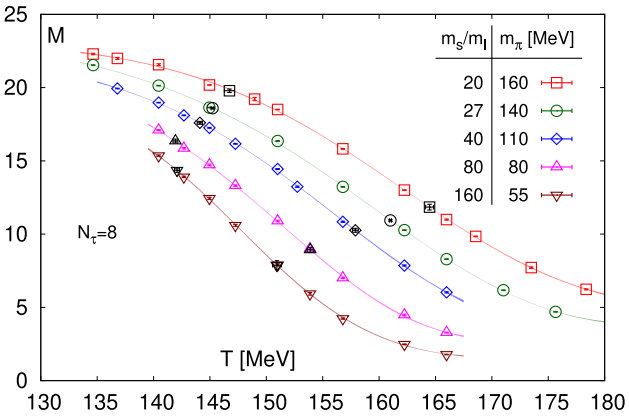


Fig. 36 Quark mass dependence of chiral order parameter, M, defined in Eq. 4.146 for QCD with two degenerate light quark masses and a strange quark mass tuned to its physical value. Shown are results from calculations on lattices with temporal extent $N_\tau = 8$ performed for several values of the light quark masses [478,479]. The light quark masses, m_ℓ, are expressed in units of the strange quark mass, $H = m_\ell/m_s$. In the figure we give $1/H = m_s/m_\ell$ together with the corresponding values of the Goldstone pion mass

QCD is clearly visible in the pseudo-critical behavior of the heavy quark free energy and the chiral condensate respectively. Some recent results on these observables, obtained in simulations of QCD with light, dynamical quark degrees of freedom, are shown in Figs. 35 and 36.

4.4.3 Global symmetries and the QCD phase diagram

The early studies of QCD thermodynamics made it clear that universality arguments and renormalization group techniques, successfully developed in condensed matter physics and applied in statistical physics to the analysis of phase transitions, also can be carried over to the analysis of the phase structure of quantum field theories [473,474]. The renormalization group based arguments for the existence of a second order phase transition in the universality class of the 3-d Ising model in a $SU(2)$ gauge theory, and a first order transition for the $SU(3)$ color group of QCD [475] have been confirmed by detailed lattice QCD calculations [476,477].

In the presence of n_f light, dynamical quarks, distinguished by a flavor quantum number, it is the chiral symmetry of QCD that triggers the occurrence of phase transitions [466]. In addition to a global $U(1)$ symmetry that reflects the conservation of baryon number and is unbroken at all temperatures and densities, the massless QCD Lagrangian is invariant under the symmetry group

$$U(1)_A \times SU(n_f)_L \times SU(n_f)_R \; . \tag{4.145}$$

The $SU(n_f)_L \times SU(n_f)_R$ symmetry corresponds to chiral rotations of n_f massless quark fields in flavor space. This symmetry is spontaneously broken at low temperatures, giving rise to $n_f^2 - 1$ massless Goldstone modes, which for $n_f = 2$ are the three light pions of QCD. They have a non-vanishing mass only because of the explicit breaking of chiral symmetry by a mass term in the QCD Lagrangian that couples to the chiral order parameter field $\bar{\chi}_f \chi_f$. The axial $U(1)_A$ group corresponds to global rotations of quark fields for a given flavor f. Although it is an exact symmetry of the classical Lagrangian, it is explicitly broken in the quantized theory. This explicit breaking of a global symmetry, arising from fluctuations on the quantum level, is known as the $U(1)_A$ anomaly.

The renormalization group based analysis of the chiral phase transition, performed by Pisarski and Wilczek [466], made it clear that the chiral phase transition is sensitive to the number of light quark flavors that become massless. Furthermore, it has been argued in [466] that the order of the transition may be sensitive to the magnitude of the axial anomaly at non-zero temperature, which is closely related to the temperature dependence of topological non-trivial field configurations.

Although it was generally expected that the chiral phase transition in 3-flavor QCD becomes a first order phase transition in the chiral limit [466], there is currently no direct evidence for this from lattice QCD calculations. In fact, the current understanding is that the chiral phase transition is second order for all $n_f \leq 6$ [480].

In Fig. 37 (top) we show the original version of the QCD phase diagram in the plane of two degenerate light (m_ℓ) and strange (m_s) quark masses, proposed in 1990 [481], together with an updated version from 2021 [480]. Here m_ℓ denotes the two degenerate up and down quark masses, $m_\ell \equiv m_u = m_d$. This sketch of our current understanding of the 3-flavor phase diagram also is supported by the increasing evidence for a non-singular crossover transition in QCD with physical light and strange quark masses and the absence of any evidence for a first order phase transition at lighter-

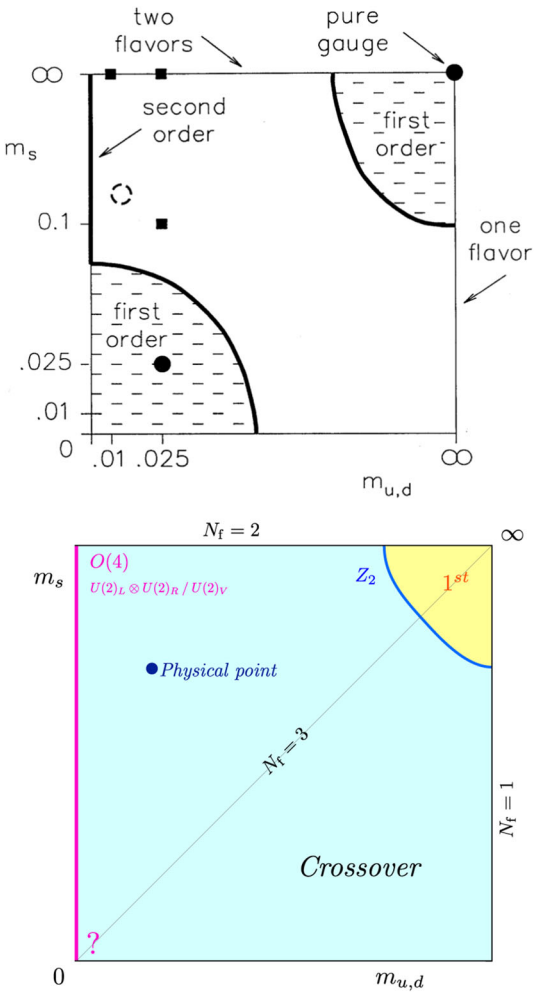


Fig. 37 Sketch of the phase diagram of QCD in the plane of degenerate, light up and down quark masses and a strange quark mass (Columbia plot). The figure shows the original version from 1990 [481] (top) and an updated version from 2021 [480] (bottom)

than-physical values of the light and strange quark masses [480,482]. In the chiral limit, i.e. for vanishing up and down quark masses,[11] a second order phase transition will then occur.

4.4.4 *The chiral phase transition at vanishing chemical potential*

The occurrence of the chiral phase transition is signaled by the vanishing of the light quark chiral condensate. In order to remove multiplicative and additive divergences in $\langle \bar{\chi}\chi\rangle_\ell$ one considers instead the order parameter M which is a combination of light and strange quark condensates,

$$M = 2\left(m_s\langle\bar{\psi}\psi\rangle_\ell - m_\ell\langle\bar{\psi}\psi\rangle_s\right)/f_K^4 \ , \tag{4.146}$$

[11] Lattice QCD studies of the (2+1)-flavor phase diagram generally are performed with degenerate up and down quark masses.

and its derivative with respect to the light quark masses, i.e. the chiral susceptibility χ_M

$$\chi_M = m_s\left(\frac{\partial M}{\partial m_u} + \frac{\partial M}{\partial m_d}\right)_{m_u=m_d\equiv m_\ell} \ . \tag{4.147}$$

Here the kaon decay constant $f_K = 156.1(9)/\sqrt{2}$ MeV, has been used to introduce a dimensionless order parameter. The scaling behavior of M and χ_M, have been used to characterize the chiral phase transition,

$$M \ \underset{m_\ell}{\sim} \ 0 \quad \begin{cases} A\left(\frac{T_c^0-T}{T_c^0}\right)^\beta, & T < T_c^0 \\ 0 & T \geq T_c^0 \end{cases} \tag{4.148}$$

$$\chi_M \ \underset{m_\ell}{\sim} \ 0 \quad \begin{cases} \infty, & T \leq T_c^0 \\ C\left(\frac{T-T_c^0}{T_c^0}\right)^{-\gamma}, & T > T_c^0 \end{cases} \tag{4.149}$$

where β, γ are critical exponents.

We note that the low temperature behavior of the order parameter susceptibility, χ_M, is quite different from that known, for instance, in the 3-d Ising model. The susceptibility diverges in the massless limit at all values of the temperature, $T \leq T_c^0$. This is a consequence of the breaking of a continuous rather than a discrete symmetry. The former gives rise to Goldstone modes, the pions in QCD, which contribute to the chiral condensate and as such to the order parameter M, i.e.,

$$M \sim a(T)\sqrt{m_\ell} \ , \quad T < T_c^0 \ . \tag{4.150}$$

As a consequence the chiral susceptibility diverges below T_c^0, $\chi_M \sim 1/\sqrt{m_\ell}$, while at T_c^0 its divergence is controlled by the critical exponent $\delta = 1 + \gamma/\beta$,

$$\chi_M \sim \begin{cases} H^{-1/2} & T < T_\chi \\ H^{1/\delta-1} & T = T_\chi \end{cases} , \tag{4.151}$$

with $H = m_\ell/m_s$. As $1 - 1/\delta > 1/2$ in all relevant universality classes χ_M develops a pronounced peak at small, but non-zero values of the quark masses,

$$\chi_M^{peak} \equiv \chi_M(T_{pc}(H)) \sim H^{1/\delta-1} \ , \ H = m_\ell/m_s \ . \tag{4.152}$$

The location of such a peak in either χ_M or similarly in $T\partial M/\partial T$, defines pseudo-critical temperatures, $T_{pc}(H)$, which converge to the unique chiral phase transition, T_c^0, at $H = 0$. Some results on the quark mass dependence of M and χ_M are shown in Figs. 36 and 38, respectively. A scaling analysis of these observables, performed in [478], led to the determination of the chiral phase transition temperature [478],

$$T_c^0 = 132^{+3}_{-6}\ \text{MeV}. \tag{4.153}$$

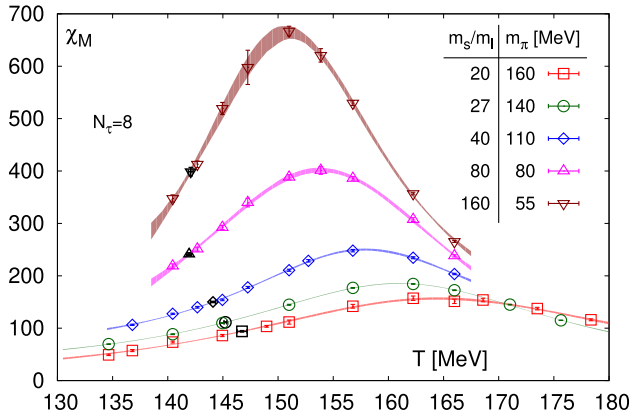


Fig. 38 Same as Fig. 36 but for the chiral susceptibility

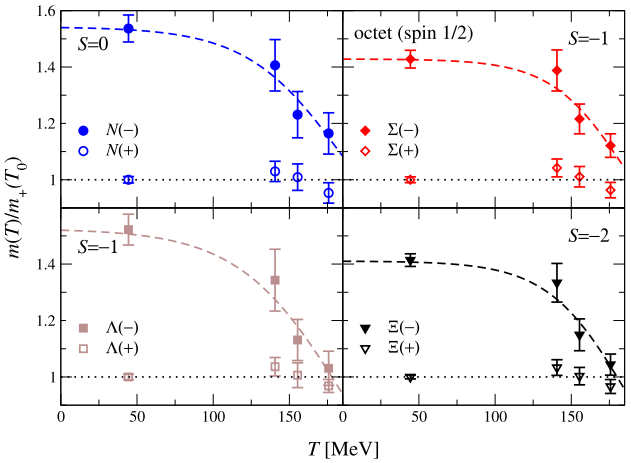


Fig. 39 Temperature dependence of masses of parity partners in the baryon octet [491]

Similar results have also been obtained in [483] where a quite different discretization scheme for the fermion sector of QCD has been used.

For physical light and strange quark masses, corresponding to $H \simeq 1/27$, one finds as a pseudo-critical temperature [484],

$$T_{pc} = 156.5(1.5)\ \text{MeV}\ , \tag{4.154}$$

which is in good agreement with other determinations of pseudo-critical temperatures in $(2+1)$-flavor QCD [485–487].

The chiral symmetry group $SU(2)_L \times SU(2)_R$ is isomorphic to the rotation group $O(4)$. It thus is expected that the chiral phase transition for two vanishing light quark masses is in the same universality class as 3-d, $O(4)$ symmetric spin models. In fact, the rapid rise of χ_M, shown in Fig. 38, is consistent with a critical exponent in this universality class, $\delta = 4.824$ [488]. However, a precise determination of this exponent in 2-flavor QCD is not yet possible. This leaves open the possibility for other symmetry breaking patterns and other universality classes playing a role in the chiral limit of 2-flavor QCD [489]. In fact, the discussion of such possibilities is closely related to the yet unsettled question concerning the influence of the axial $U(1)_A$ symmetry on the chiral phase transition. For a recent review on this question see, for instance [490].

Thermal masses and screening masses

The restoration of symmetries is reflected also in the modification of the hadron spectrum at non-zero temperature. Interactions in a thermal medium lead to modifications of resonance peaks that can modify the location of maxima and the width of spectral functions that control properties of hadron correlation functions. This gives rise to so-called thermal masses as well as thermal screening masses that control the long-distance behavior of hadron correlation functions in Euclidean time and spatial directions, respectively.

A consequence of $U(1)_A$ breaking in the vacuum or at low temperature is that masses of hadronic states that are related to each other through a $U(1)_A$ transformation differ, while they become identical, or close to each other, when the $U(1)_A$ symmetry is effectively restored. This is easily seen to happen at high temperature. The crucial question, of relevance for the QCD phase transition, however, is to which extent $U(1)_A$ symmetry breaking is reduced, or already disappeared at the chiral phase transition temperature. Settling this question requires the analysis of observables sensitive to $U(1)_A$ breaking close to T_c^0 and for smaller-than-physical light quark masses.

The calculation of in-medium modifications of hadron masses is difficult, but has been attempted for quark masses close to their physical values [491]. Results for the temperature dependence of the mass-splitting of parity partners in the baryon octet [491] are shown in Fig. 39. These results suggest a strong temperature dependence of the negative parity states while the positive parity partners are not sensitive to temperature changes. At T_{pc} the masses of parity partners are almost degenerate.

More easily accessible are so-called screening masses, which also are obtained from ordinary hadron correlation functions and can be analyzed close to the chiral limit. Rather than analyzing the long-distance behavior of hadron correlation functions in Euclidean time, one extracts a so-called screening mass from the long-distance behavior in one of the spatial directions [494,495]. Finite temperature meson screening correlators, projected onto lowest Matsubara frequency of a bosonic state, $p_0 \equiv \omega_0 = 0$, and zero transverse momentum, $\mathbf{p}_\perp \equiv (p_x, p_y) = 0$, are defined by

$$G_\Gamma(z,T) = \int_0^\beta d\tau \int dxdy \left\langle \mathcal{M}_\Gamma(\vec{r},\tau)\overline{\mathcal{M}_\Gamma}(\vec{0},0)\right\rangle \overset{z\to\infty}{\sim} \mathrm{e}^{-m_\Gamma(T)z}\ ,\quad \vec{r}\equiv(x,y,z)\ , \tag{4.155}$$

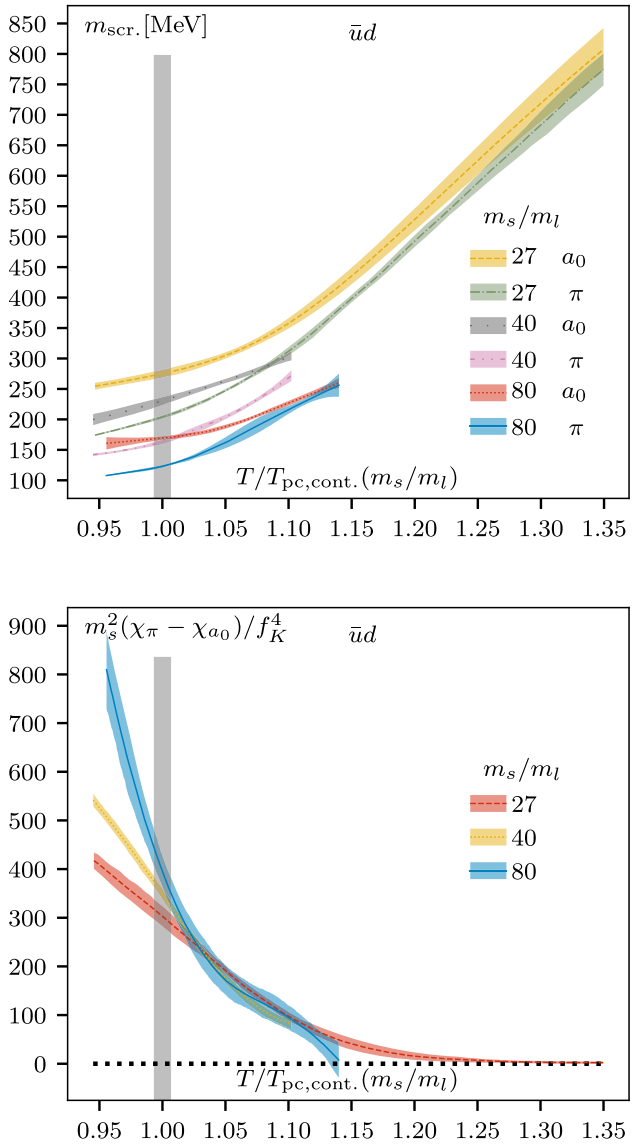


Fig. 40 Screening masses (top) and the related susceptibilities (bottom) of scalar and pseudo-scalar mesons [492,493]

where $\mathcal{M}_\Gamma \equiv \bar{\psi}\Gamma\psi$ is a meson operator that projects onto a quantum number channel that is selected through an appropriate choice of Γ-matrices [492,494]. At large distances this permits the extraction of the screening mass, m_Γ, in the quantum number channel selected by Γ from the exponential fall-off of these correlation functions. In Fig. 40 (left) we show results for the scalar and pseudo-scalar screening masses obtained in (2+1)-flavor QCD calculations for different values of the light to strange quark mass ratio. The integrated correlation functions define susceptibilities in these quantum number channels, which also should be degenerate, if $U(1)_A$ is effectively restored. Both observables seem to suggest that there remains a significant remnant of $U(1)_A$ breaking at the chiral phase transition temperature, T_c^0, which however reduces quickly above the chiral transition and gives rise to an effective restoration of $U(1)_A$ at $T \simeq 1.1T_c^0$.

In the region $T > T_c^0$ the difference between pseudo-scalar and scalar susceptibilities is related to the so-called disconnected part, χ_{dis}, of the chiral susceptibility, $\chi_M = \chi_{dis} + \chi_{con}$, with

$$\chi_{dis} = \frac{1}{4N_\tau N_\sigma^3}\left(\langle(\mathrm{Tr}M_\ell^{-1})^2\rangle - \langle \mathrm{Tr}M_\ell^{-1}\rangle^2\right) , \tag{4.156}$$

$$\chi_{con} = \frac{1}{2N_\tau N_\sigma^3}\langle \mathrm{Tr}M_\ell^{-2}\rangle . \tag{4.157}$$

While the disconnected chiral susceptibility can in general be expressed in terms of an integral over the quark mass derivative of the eigenvalue density [496], $\rho(\lambda)$, of the fermion matrix M_f, it is directly related to an integral over $\rho(\lambda)$ in the chirally symmetric high temperature phase,

$$\chi_{dis} = \int_0^\infty d\lambda\, \rho(\lambda)\frac{2m_\ell^2}{(\lambda^2+m_\ell^2)^2} . \tag{4.158}$$

In the chiral symmetric phase the density of vanishing eigenvalues, $\rho(0)$, vanishes. In order for χ_{dis} to be nonetheless non-zero in the chiral limit, the density of near-zero eigenvalues needs to converge to a non-vanishing value (δ-function) at $\lambda = 0$ in the limit $m_\ell \to 0$ and $V \to \infty$. Controlling the various limits involved and also taking into account that the pseudo-critical transition temperature, $T_{pc}(H)$, has a sizeable quark mass dependence is difficult. Nonetheless, studies of the temperature dependence of the eigenvalue density of the Dirac matrix are crucial for a detailed understanding of the influence of the $U(1)_A$ anomaly on the QCD phase transition. Not surprisingly, it turns out that at non-zero values of the lattice spacing the spectrum of low lying eigenvalues is quite sensitive to the fermion discretization scheme. Using fermions with good chirality even at non-zero lattice spacing seems to be advantageous, although after having performed the extrapolation to the chiral limit, they should lead to results identical with those obtained, e.g. within the staggered discretization scheme. Current results are ambiguous. We show in Fig. 41 results from a calculation of eigenvalue distributions obtained from calculations with dynamical overlap fermions [497,498]. These calculations provide evidence for a large density of near-zero eigenvalues and a non-zero eigenvalue density, possibly building up at $\lambda = 0$. This is in contrast to calculations performed with domain wall fermions [499] as well as so-called partially quenched calculations that use the overlap fermion operator to calculate eigenvalue distributions on gauge field configurations generated with dynamical staggered fermions [500]. Obviously this subtle aspect of the chiral phase transition is not yet resolved and the analysis of $U(1)_A$ restoration will remain to be a central topic in finite temperature QCD in the years to come.

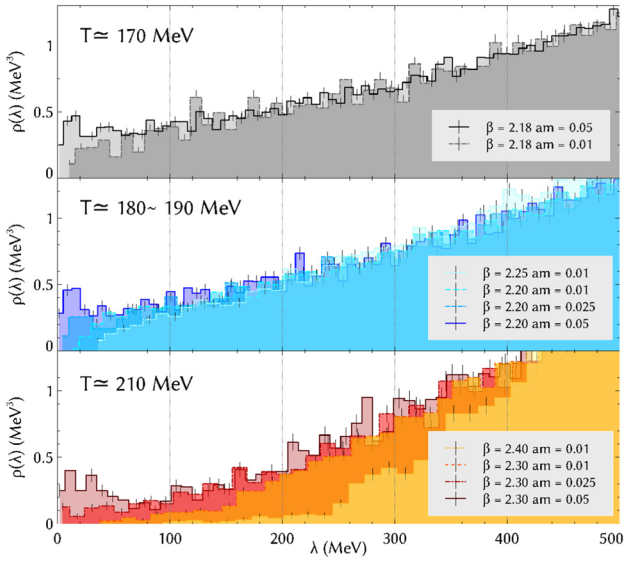


Fig. 41 Eigenvalue density of the overlap fermion matrix obtained in calculations with dynamical overlap fermions [497]

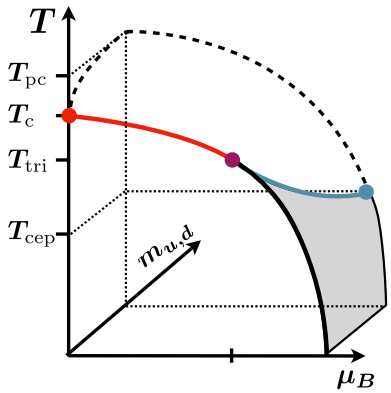


Fig. 42 Sketch of a possible QCD phase diagram in the space of temperature (T), baryon chemical potential (μ_B) and light quark masses ($m_{u,d}$)

4.4.5 *The chiral phase transition at non-vanishing chemical potential*

In the studies of QCD at non-vanishing baryon chemical potential the search for the existence of a second order phase transition at physical values of the quark masses, the critical end point (CEP), finds particular attention. It separates the crossover regime at small values of the chemical potential from a region of first order phase transitions, which is predicted in many phenomenological models to exist at high density. The CEP is searched for extensively in heavy ion experiments and, if confirmed, would provide a solid prediction for the existence of first order phase transitions in dense stellar matter, e.g. in neutron stars.

The dependence of the transition temperature on the chemical potentials, e.g. $T_{pc}(\mu_B)$, can be deduced from the μ_B-dependent shift of the peak in the chiral susceptibility. At non-vanishing values of the baryon chemical potential, μ_B, the QCD phase transition temperature in the chiral limit as well as the region of pseudo-critical behavior in QCD with its physical quark mass values shifts to smaller values of the temperature. This shift has been determined in calculations with imaginary values of the chemical potentials as well as from Taylor series expansions of the order parameter M and its susceptibility χ_M. Using a Taylor series ansatz for $T_{pc}(\mu_B)$,

$$T_{pc}(\mu_B) = T_{pc}^0 \left(1 - \kappa_2^B \left(\frac{\mu_B}{T_c^0}\right)^2 - \kappa_4^B \left(\frac{\mu_B}{T_c^0}\right)^4\right) \quad (4.159)$$

one finds for the curvature coefficients $\kappa_2^B \simeq 0.012$ while the next correction is consistent with zero in all current studies, e.g. $\kappa_4^B = 0.00032(67)$ [487]. The pseudo-critical temperature T_{pc} at physical values of the light and strange quark masses thus drops to about 150 MeV at $\mu_B \simeq 2T_{pc}$. This is still considerably larger than the chiral phase transition temperature, T_c^0, determined at $\mu_B = 0$. As various model calculations [504,505] suggest that the CEP at non-zero μ_B is located at a temperature below T_c^0 one thus needs to get access to thermodynamics at large chemical potentials. Assuming that the curvature of the pseudo-critical line does not change drastically at large values of the chemical potentials, our current understanding of the QCD phase diagram in the m_ℓ-T-μ_B space (see Fig. 42) suggests that a possible CEP in the phase diagram may exist only at a temperature,

$$T^{CEP}(\mu_B^{CEP}) < 130 \text{ MeV}, \ \mu_B^{CEP} > 400 \text{ MeV} . \quad (4.160)$$

Reaching the region $\mu_B/T > 3$ is a major challenge for any of the currently used approaches in lattice QCD calculations as well as for collider based heavy ion experiments that search for the CEP.

4.4.6 *Equation of state of strongly interacting matter*

The equation of state (EoS) of strongly interacting matter, i.e. the pressure and its derivatives with respect to temperature and chemical potentials provides the basic information on the phase structure of QCD. It is of central importance not only for the analysis of critical behavior in QCD but also for the analysis of experimental results on strong interaction thermodynamics that are obtained in relativistic heavy ion collision experiments.

At vanishing values of the chemical potentials the QCD EoS is well controlled and consistent results for pressure, energy and entropy densities, as well as derived observables such as the speed of sound or specific heat, have been obtained by several groups [501,502]. We show results for some of these observables in Fig. 43. The figure on the right shows the square of the speed of sound, c_s^2, as function of the energy density. It can be seen that c_s^2 has a minimum in

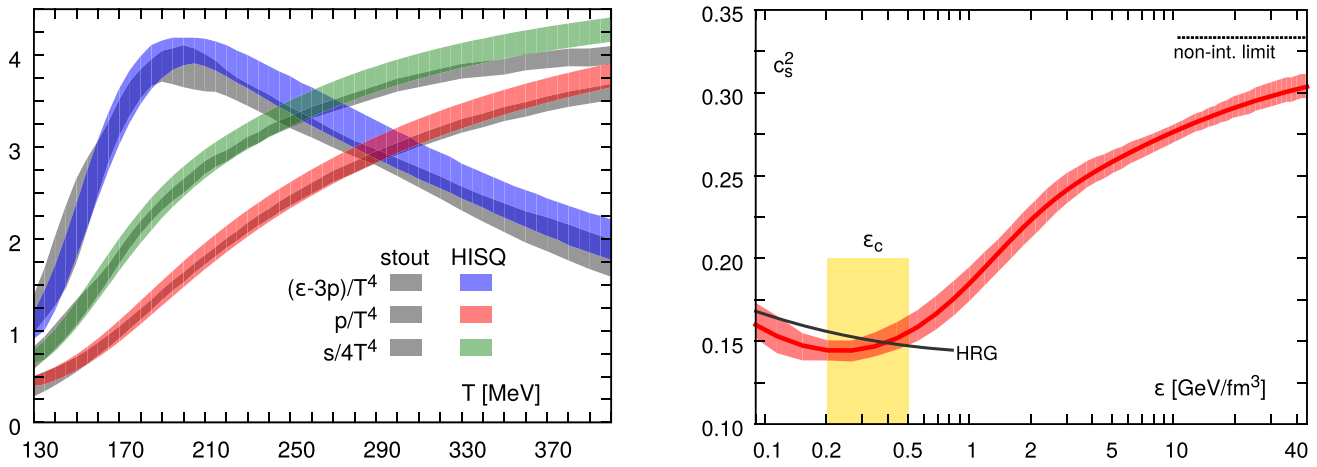


Fig. 43 Left: Pressure, energy and entropy densities in (2+1)-flavor QCD at vanishing chemical potential. The figure is taken from [501]. Also shown in the figure are results obtained with the stout discretization scheme for staggered fermions [502]. Right: The speed of sound as function of energy density

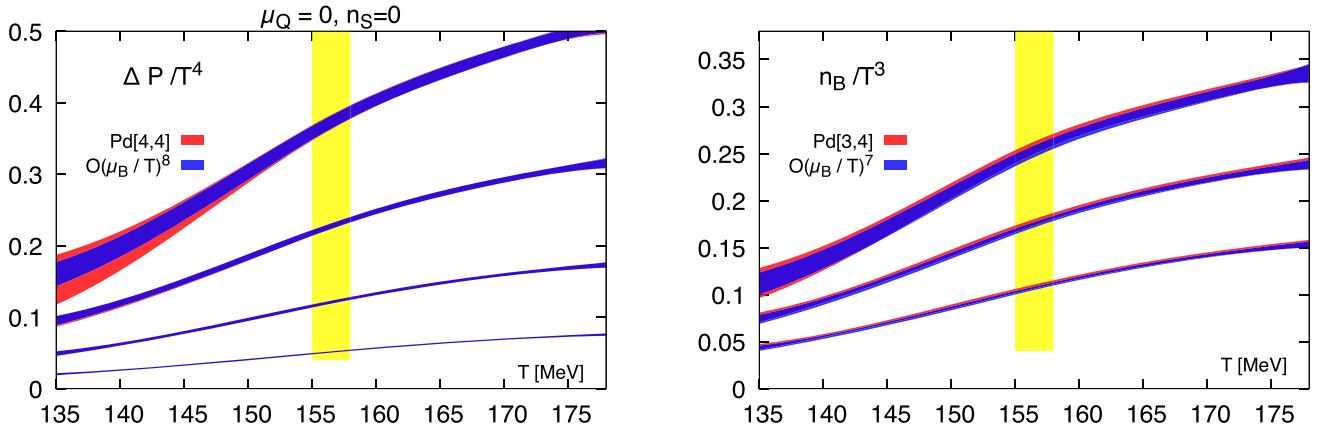


Fig. 44 μ_B-dependent contribution to the pressure (left) and net baryon number density (right) in (2+1)-flavor QCD at several values of the baryon chemical potential chemical potential, $\mu/T_B = 1.0, 1.5, 2.0, 2.5$, (bottom to top) and for $\hat{\mu}_B = 2.0$. Shown are results from Taylor expansion up to eighth order in $\hat{\mu}_B$ in the pressure series for isospin symmetric ($\mu_Q = 0$) strangeness neutral ($n_S = 0$) matter and corresponding Padé approximants obtained from these Taylor expansion coefficients. The figures are taken from [503]

the transition region, sometimes called the softest point of the QCD EoS [506]. The energy density in the vicinity of the pseudo-critical temperature ($T_{pc} \simeq 155$ MeV) is found to be,

$$\epsilon_c \simeq (350 \pm 150)\ \mathrm{MeV/fm}^3\ , \tag{4.161}$$

which is compatible with the energy density of the nucleon, $m_N/(4\pi r_N^3/3)$ for nucleon radii in the range $r_N = (0.8-1)$ fm. Also shown in the top figure is the trace of the energy–momentum tensor, $(\epsilon - 3P)/T^4$. Its deviation from zero gives some hint to the relevance of interactions in the medium (for an ideal gas as well as to leading order in high temperature perturbation theory one has $\epsilon = 3P$). Not unexpected this is largest close to the transition region and decreases only slowly in the high temperature regime. This large deviations from ideal gas or perturbative behavior is seen in many observables at temperature $T_{pc} < T < 2T_{pc}$.

Calculations of the equation of state as a function of T and μ_B have been performed using direct simulations at imaginary chemical potentials, which then get analytically continued to real values of the chemical potentials [507], as well as calculations using up to eighth order Taylor expansions in μ_B [503]. Results of such calculations agree well for $\mu_B/T \leq (2-2.5)$. In Fig. 44 we show results for the μ_B-dependent contribution to the pressure and net baryon number density. Comparing Fig. 44 (left) with Fig. 43 (left) shows that at $\mu_B/T \simeq 2$ and $T \simeq T_{pc}$ the pressure increases by about 30%, which is due to the increase in number of baryons in the medium.

At larger values of the baryon chemical potential the Taylor series will not convergence due to the presence of either

poles in the complex μ_B-plane or a real pole, that may correspond to the searched for CEP. The occurrence of poles in the complex plane also generates problems for the analytic continuation of results obtained in simulations at imaginary values of μ_B as a suitable ansatz for the continuation needs to be found. Many approaches to improve over straightforward Taylor series approaches or simulations at imaginary chemical potential are currently being discussed [508–511]. In the context of Taylor expansions a natural way to proceed is to use Padé approximants, which provide a resummation of the Taylor series and reproduce this series, when expanded for small μ_B [503,512]. Results from [4,4] and [3,4] Padé approximants for the pressure and number density series, respectively, are also shown in Fig. 44. The good agreement with the Taylor series for $\mu_B/T \leq 2.5$ gives confidence in the validity of the Taylor series results and once more seems to rule out the occurrence of a CEP in this parameter range.

4.4.7 Outlook

Achieving better control over the influence of the axial anomaly on the QCD phase transition in the chiral limit at vanishing chemical potentials and getting better control over the dependence of the QCD phase diagram at large non-zero values of the chemical potentials certainly are the two largest challenges in studies of QCD thermodynamics for the next decade.

4.5 Spectrum computations

Jozef Dudek

4.5.1 Motivation for hadron spectroscopy

Many decades of experimental data collection has lead to a compendium of observed hadrons [513], most of which are short-lived resonances. The job of hadron spectroscopy is to understand the patterns in the spectrum, such as the distribution of states by spin, parity and flavor, and which decays are preferred by which states. These patterns are typically interpreted in terms of models or 'pictures' of hadron structure in which e.g. certain mesons are assigned status as $q\bar{q}$, as glueballs, as hybrids, as higher quark Fock states, or as molecular states of lighter hadrons [514].

For a long time, simplified dynamical models whose connection to QCD is often obscure have dominated the field, and through these significant intuition has been developed, but in recent years lattice QCD has matured to the level where it can address the physics of excited hadrons directly. Using this tool we aim to build an understanding of how QCD binds quarks and gluons into hadrons from first principles.

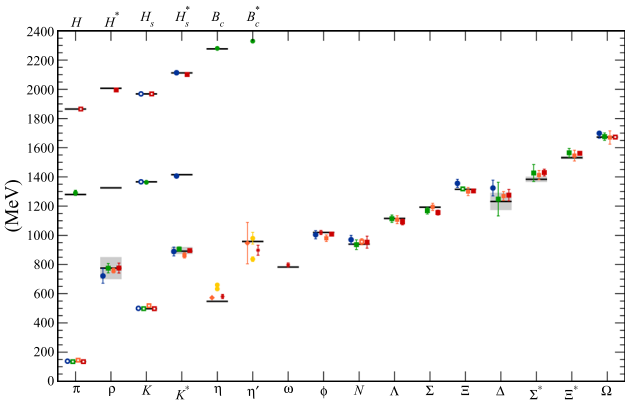


Fig. 45 Summary of hadron spectrum calculations taken from Ref. [515]. Different symbol shapes indicate different quark discretizations, while the colors (red, orange, green, blue) indicate an increasing level of systematic control in the calculation. *b*-flavored meson masses are shifted down by 4000 MeV

4.5.2 Precise mass determination for stable hadrons

As described in Sect. 4.2, hadron masses can be determined from the large time behavior of two-point correlation functions utilizing operators with the quantum numbers of hadrons constructed from quark and gluon fields. These correlation functions are calculated using quark propagators computed with a particular choice of discretization of the QCD action, and particular values of parameters which set the lattice spacing and the quark masses. When seeking precise determination of hadron masses, one can calculate with several quark masses and lattice spacings, and attempt to extrapolate to the physical limit where the quark mass takes its true value and where the lattice spacing becomes zero.

Figure 45 (taken from Ref. [515]) summarizes a number of efforts in this direction, showing the masses for low-lying mesons and baryons constructed from light, strange, charm and bottom quarks, comparing the computed values to measured values. Clear agreement is observed for many stable or nearly-stable hadrons. With increasing levels of precision on the mass estimates, the role of small effects like QED become important, and in recent years, these too have been estimated (e.g. Refs. [516–519]) .

4.5.3 Expanding the scope of lattice spectroscopy

There are relatively few calculations in which hadron masses have been determined with a somewhat complete study of systematics, and they have been largely restricted to those situations where only a single completely-connected Wick contraction features in the relevant correlation function, and where the state of interest is the lightest with a given quantum number.

Examples which require something beyond this include *isoscalar mesons* in which quark–antiquark annihilation diagrams must be computed. Conventional propagator techniques cannot handle these diagrams, and while various stochastic techniques have been used, it was the introduction of the *distillation* approach [520] which not only opened up isoscalar meson spectroscopy but also the determination of multiple excited states.

Distillation is in essence a quark-field smearing implementation, where the smearing operator,

$$\Box(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} v_i(\mathbf{x})\, v_i^\dagger(\mathbf{y}),$$

is constructed from a limited number of low-lying eigenvectors of the gauge-covariant spatial Laplacian,

$$-\nabla^2 v_i(\mathbf{x}) = \lambda_i\, v_i(\mathbf{x}).$$

All quark fields in hadron interpolating operators are smeared by this operator, enhancing overlap with low-lying states. The unique advantage of this approach though is the way that the outer-product nature of the smearing operator allows a *factorization* of correlation functions into objects describing hadron operators, independent of objects called "perambulators" describing quark propagation,

$$\tau_{ij}(t,t') = \sum_{\mathbf{x},\mathbf{y}} v_i^\dagger(\mathbf{x}) M^{-1}(\mathbf{x},t;\mathbf{y},t') v_j(\mathbf{y}).$$

Annihilation contributions can be handled straightforwardly using timeslice-to-timeslice perambulators, $\tau_{ij}(t,t)$.

The factorization within distillation allows for massive re-use of the propagation objects, so that the inversion time cost of building a set of perambulators is amortized over a huge number of subsequent calculations.[12] In the context of determining excited states, it allows for the computation of many correlation functions using a large basis of interpolating operators.

While in principle any single correlation function

$$C(t,0) = \sum_n a_n\, e^{-M_n t}$$

contains information about the entire excited spectrum, $\{M_n\}$, in practice determining the spectrum by fitting subleading time-dependence is highly unstable. It is obvious for example, that degenerate or near-degenerate states cannot be distinguished by their time-dependence alone. A much more powerful approach makes use of orthogonality – if one considers a large basis of hadron interpolating operators all with the same overall quantum numbers, we expect there to be one linear combination that most effectively produces the ground state, another that produces the first-excited state and so on. It is straightforward to show that if one forms the *matrix* of two-point correlation functions

$$C_{ij}(t) = \langle 0|\mathcal{O}_i(t)\, \mathcal{O}_j^\dagger(0)|0\rangle,$$

with a basis of operators $\{\mathcal{O}_i\}_{i=1\ldots N}$, the optimal combinations correspond to the eigenvectors of the *generalized eigenvalue problem*,

$$C(t)\, v_n = \lambda_n(t,t_0)\, C(t_0)\, v_n,$$

where the eigenvalues give access to the corresponding mass or energy spectrum, $\lambda_n(t,t_0) \sim e^{-E_n(t-t_0)}$. This approach is typically referred to as *variational analysis* [410–412].

An example of a large basis of operators with the quantum numbers of mesons is the one presented in Ref [521], where smeared quark field bilinears featuring up to three gauge-covariant derivatives are used [523–526]. In order to respect the reduced rotational symmetry of the cubic lattice, operators of definite J^P are *subduced* into irreducible representations (irreps) of the cubic symmetry. Using a basis like this, with the variational analysis approach presented above, can lead to results like those shown in Fig. 46. The extracted spectrum shows many of the systematics of the experimental meson spectrum such as the J^{PC} ordering of states and the presence of an "OZI-rule" in the hidden-light/hidden-strange composition of isoscalar mesons (dominantly ideal flavor mixing except for a few notable exceptions like 0^{-+}). Also present in these extracted spectra are mesons with *exotic* $J^{PC} = 1^{-+}, 0^{+-}, 2^{+-}$, i.e. those not accessible to just a $q\bar{q}$ pair. Examining which interpolating operators are the largest components in the optimal operators for these states, we observe the presence of non-trivial gluonic structures, and it is natural to interpret these states as *hybrid mesons*. Non-exotic J^{PC} states high in the spectrum are also observed to have these gluonic operator overlaps (states outlined in orange in Fig. 46), and this leads to an identification of the lightest supermultiplet of hybrid mesons [527], ruling out certain previously reasonable models.

A closely related calculation using a large basis of operators with baryon quantum numbers appeared in Refs. [528,529], with the spectra for $N^\star$ (isospin-1/2) and $\Delta^\star$ (isospin-3/2) excitations shown in Fig. 47.

The calculation presented in Fig. 46 was performed with a light quark mass heavier than physical, and at a single lattice spacing, and as such the results cannot be treated as precise, or suitable for direct comparison to experiment. But in the case of excited spectroscopy, precision is not the main aim, rather the intent is to build an understanding of the systematics of the hadron spectrum having a direct connection to QCD. In fact there is a more relevant problem with these results – they do not reflect the complete physics of excited states which lie above hadronic decay thresholds – these states should be

[12] There is also a stochastic implementation of distillation [522], which is argued to have a better cost-scaling with the volume of the lattice, but at the cost of somewhat less flexibility in re-use.

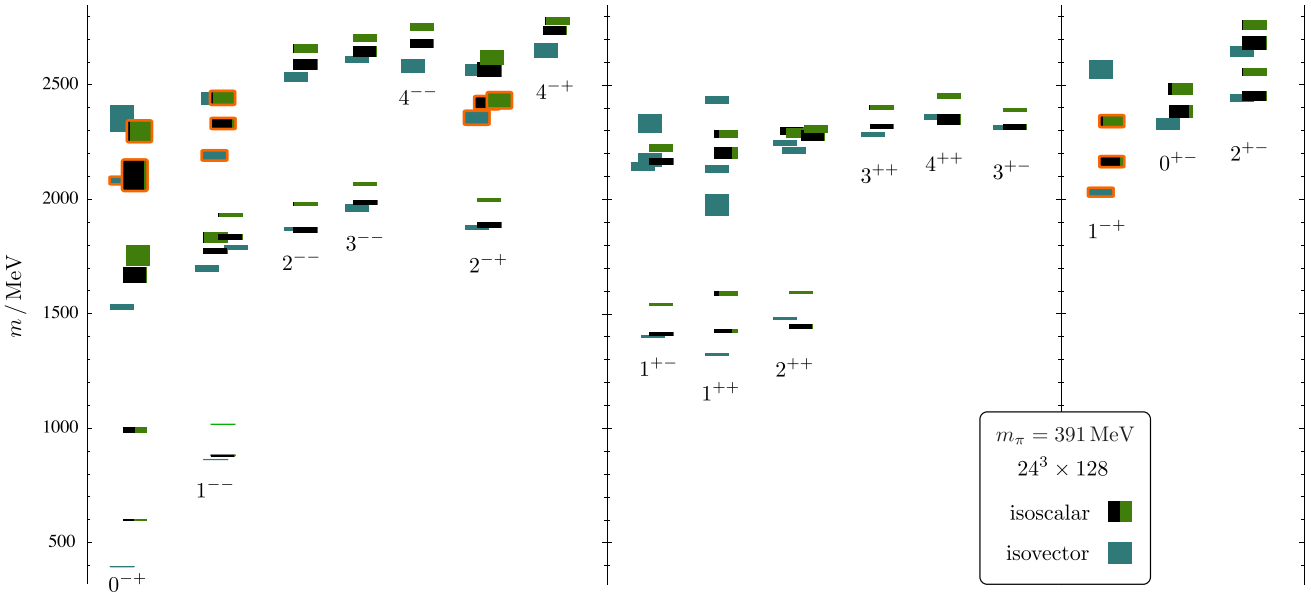


Fig. 46 Spectrum of excited mesons extracted from lattice QCD calculation with heavier that physical light quarks. States labelled by their J^{PC}. Vertical height of each box represents the statistical uncertainty. Isoscalar meson boxes show the hidden-light (black) versus hidden-strange (green) composition. States with orange outlines have large overlap with operators featuring the chromomagnetic field, suggesting an identification as the lightest supermultiplet of *hybrid mesons*. Taken from Ref. [521]

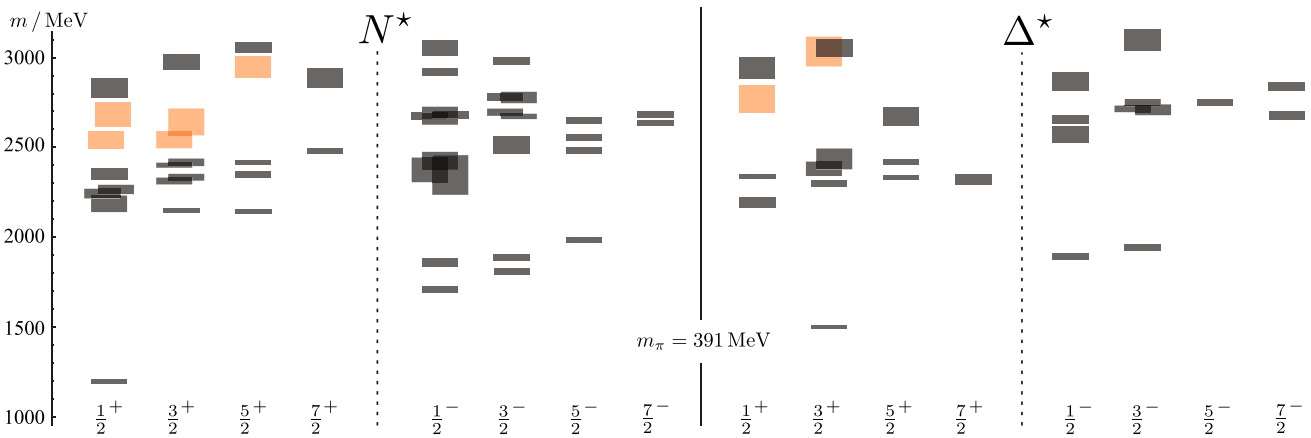


Fig. 47 Spectrum of excited baryons extracted from lattice QCD calculation with heavier than physical light quarks. States labelled by their J^P. Vertical height of each box represents the statistical uncertainty. States colored orange have large overlap with operators featuring the chromomagnetic field, suggesting an identification as the lightest supermultiplet of *hybrid baryons*. Taken from Refs. [528,529]

unstable *resonances*, and resonances are not simply characterized by a mass.

4.5.4 Resonances and the finite-volume approach to scattering

The simplest context in which resonances appear is *elastic hadron–hadron scattering* in which the initial and final states are identical, and the amplitude can be expanded in partial-waves. Resonances of definite spin appear as enhancements in a single partial-wave in the continuous energy spectrum, and formally may be associated with pole singularities at complex values of the scattering energy.

In a finite spatial volume, such as that provided by the lattice, there can be no continuous energy spectrum, and instead only a discrete spectrum, but it is easy to see that this spectrum should be volume-dependent and sensitive to the infinite-volume scattering amplitudes. This can be illustrated in one-dimensional quantum mechanics [320] – a finite-length of L can be implemented by applying periodic boundary conditions to a scattering wavefunction and its derivative. This leads to a quantization condition on possible allowed

momenta, $p_n = \frac{2\pi}{L} n - \frac{2}{L}\delta(p_n)$, where $\delta(p)$ is the elastic phase-shift that describes scattering.

This observation is the core principle behind the lattice QCD approach to scattering. If the discrete spectrum of states in the finite spatial volume defined by the lattice can be obtained, it can be used to provide a set of constraints on the energy dependence of scattering amplitudes.

The analogous formalism for relativistic scattering in three spatial dimensions was derived in Refs. [530,531], and has been extended many times to now be in a form that is applicable to any number of coupled channels of two-body scattering (see the review, Ref. [532]). One way of writing this *quantization condition* is

$$\det\left[\mathbf{1} + i\boldsymbol{\rho}(E)\,\mathbf{t}(E)\left(\mathbf{1} + i\boldsymbol{\mathcal{M}}(E,L)\right)\right] = 0, \tag{4.162}$$

where the scattering t-matrix is a dense matrix in the space of scattering channels, but block diagonal in angular momentum, ℓ, while the matrix $\boldsymbol{\mathcal{M}}$, which features known functions (of essentially kinematic origin) of energy and box-size, is block-diagonal in channels, but dense in ℓ.

The presence of multiple ℓ in the quantization condition is an important complicating factor that reflects the fact that the basis of *partial waves* of definite ℓ, in which one naturally expands scattering, is not respected by the reduced rotational symmetry of the cubic boundary of the lattice. The angular momentum barrier at low energies ensures that in practice only a small finite number of ℓ values need to be considered.

Equation (4.162) can be interpreted as follows: if one knew the scattering amplitudes $\mathbf{t}(E)$, one would seek to find all the zero-crossings of the determinant function for a fixed value of L, and these would determine the finite-volume spectrum, $E_n(L)$, corresponding to this scattering amplitude. Of course in practice, lattice QCD will supply the discrete finite-volume spectra and one must work backwards to find the corresponding $\mathbf{t}(E)$.

One situation in which this is relatively straightforward is when we are in an energy region where only *elastic scattering* is kinematically allowed, and where one partial wave, ℓ, is dominant. In this case Eq. 4.162 reduces to the simple form $\cot\delta_\ell(E) = \mathcal{M}_{\ell,\ell}(E,L)$. In this case, given a lattice QCD determined finite-volume energy E, one simply plugs into the right-hand-side to obtain a value of the scattering phase-shift at that energy. If enough finite-volume energies are determined, in one or more lattice volumes, the energy dependence of $\delta_\ell(E)$ can be mapped out.

So the job of lattice QCD computation in studies of resonances is to provide accurate discrete finite-volume spectra. In order for calculations to resolve the *full* discrete spectrum of states (as opposed to the limited set described in the previous section) it proves necessary to include in the basis of operators a set which resemble pairs of mesons. These "meson–meson-like" operators are typically constructed from a product of two quark-bilinears, with each one being projected into a definite momentum. The important difference with respect to the single quark-bilinear operators described in the previous section, is that the "meson–meson-like" operators sample the entire spatial volume, causing them to have a much enhanced overlap with finite-volume eigenstates resembling a pair of mesons.

A basis of "meson–meson-like" operators can be constructed [533–535] and a natural guide to which are required in any given calculation comes from a *non-interacting energy* associated with each such operator. For example, operators resembling a pair of pions with $\ell = 0$ can be constructed as $\sum_{\hat{\mathbf{p}}} \mathcal{O}_\pi(\mathbf{p})\mathcal{O}_\pi(\mathbf{-p})$ where $\mathcal{O}_\pi(\mathbf{p})$ is a quark bilinear with the quantum numbers of a pion, and where the sum is over directions of momentum allowed on a cubic lattice. These operators naturally have a non-interacting energy $E_{\text{n.r.}} = 2\sqrt{m_\pi^2 + \mathbf{p}^2}$ associated with them that corresponds to the energy a state interpolated by this operator would have if there were no residual pion–pion interactions. Because there *are* interactions, the actual energy spectrum will differ from this, but it should be clear that operators with non-interacting energy far above the energy region under consideration will not need to be included.

Adding "meson–meson-like" operators to the basis increases the variety of Wick diagrams that need to be evaluated, and in general diagrams including quark–antiquark annihilation are present. Distillation is a very powerful tool to evaluate these diagrams using previously computed perambulators, without the need to make further approximations, or to introduce noise through stochastic approaches.

4.5.5 Elastic meson–meson scattering

An example of the approach described in the previous section is presented in Fig. 48 which shows the P-wave of $\pi\pi$ scattering with isospin–1. The calculation, done with light-quark masses such that the pion mass is 391 MeV, computed the finite-volume spectrum in three lattice volumes. The panels on the left show the spectra in the rest frame ([000]) and several frames in which the $\pi\pi$ system has a net momentum $\mathbf{P} = \frac{2\pi}{L}[n_x n_y n_z]$. Each discrete energy is used to obtain a value of δ_1 at the same energy, and these are plotted in the right panel, where the behavior is clearly that of a narrow resonance. The energy dependence can then be fitted using a Breit–Wigner or other suitable amplitude parameterization from which the mass and width of the ρ resonance can be determined.

Calculations like this one, of the ρ resonance, have become mainstream within the lattice community [533,535–547],[13] and the vector K^* resonance in $K\pi$ scattering is similar (although in this case one has to deal with the effect of

[13] One calculation has considered the ρ in $\pi\pi$ scattering using two lattice spacings [544], finding no statistically significant differences.

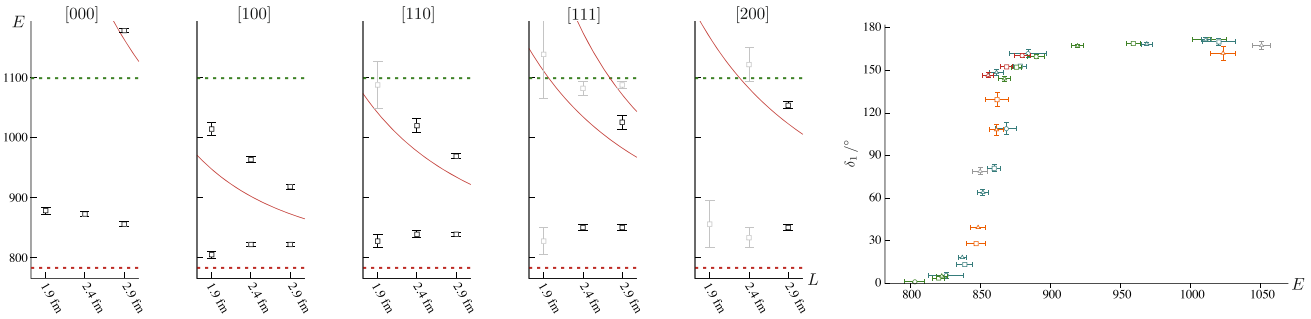


Fig. 48 Isospin–1 $\pi\pi$ scattering with $J^P = 1^-$ from lattice QCD with $m_\pi = 391$ MeV taken from Ref. [533]. Left five panels show discrete spectrum of states in three lattice volumes, for five values of total $\pi\pi$ momentum. Red curves indicate the non-interacting $\pi\pi$ energies, and the green dashed line shows the $K\bar{K}$ threshold where scattering ceases to be elastic. Rightmost panel shows the P-wave elastic scattering phase-shift determined using the discrete spectrum points which is observed to correspond to a narrow ρ resonance

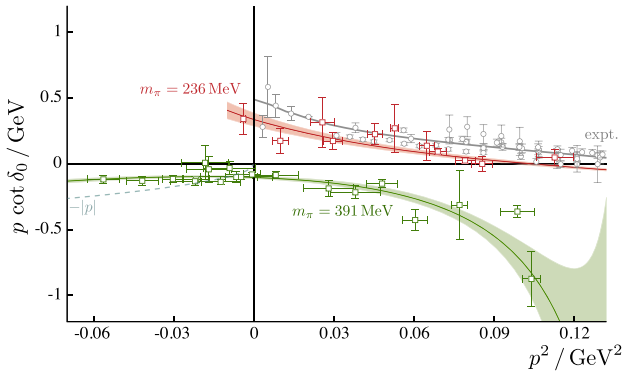


Fig. 49 Isospin–0 $\pi\pi$ scattering with $J^P = 0^+$ from lattice QCD at two pion masses taken from Ref. [558]. Intersection of $p \cot \delta_0$ with $-|p|$ indicates the presence of a bound-state σ at the heavier pion mass which is not present at the lower pion mass, or in experiment, where a broad resonance is believed to be present

S-wave scattering in moving frames) [541,548–552]. The elastic scattering amplitudes do not need to be resonant for this approach to be used, an example is $\pi\pi$ scattering with isospin–2 where the relatively weak effects can be resolved [534,542,553–557].

Pion–pion scattering with isospin–0 has received less attention [557–559]. In order to evaluate the relevant correlation functions, many diagrams featuring $q\bar{q}$ annihilation are required. One example calculation [558] that made use of distillation to evaluate all these diagrams is summarized in Fig. 49, where a function of the phase-shift as a function of energy is shown for calculations at two different light quark masses. The behavior at the heavier quark mass is that of a system featuring a *stable bound state*, while at the lower quark mass, which much more closely resembles the experimental data, there appears to be a *broad resonance*. These results provide the first signs within QCD of the quark mass evolution of the σ meson.

Scattering of mesons featuring charm or bottom quarks can be studied using the same technology [561–573]. Relatively few calculations have so far attempted to determine meson–baryon scattering and the baryonic resonances therein [574–576], largely because of the increased computational cost of such efforts above what is required for meson–meson scattering, and the fact that the lowest-lying resonance, the $\Delta(1232)$, only becomes unstable for decay to $N\pi$ at relatively low light-quark masses.

4.5.6 *Coupled-channel scattering*

The bulk of experimentally observed hadron resonances can decay into more than one hadronic final state, and as such can be considered to be resonances in *coupled-channel* scattering. Coupled-channel scattering (in a particular partial wave) can be described by a t-matrix, $t_{ij}(E)$, where the indices i, j run over hadronic channels, e.g. $\pi\pi$, $K\bar{K}$ ....

Equation (4.162) controls how the discrete spectrum in a finite volume is related to the t-matrix, but practical use of this equation when lattice QCD-obtained finite-volume spectra are in hand requires some thought. It is not possible to work energy-level by energy-level as we did for elastic scattering, as the t-matrix contains multiple unknowns at each energy. Rather, a successful approach has been to *parameterize* the energy-dependence of $\mathbf{t}(E)$, and to attempt to describe the entire finite-volume spectrum using this parameterization. A χ^2 can be defined which quantifies the difference between the finite-volume spectrum obtained from solving Eq. (4.162) for a particular parameterization and the lattice QCD obtained spectrum. This χ^2 can be minimized by varying the free parameters to obtain a best fit.

In order to carry this out, it is necessary to construct appropriate parameterizations of $\mathbf{t}(E)$ which must include all kinematically open channels in the energy region being considered. They must also exactly respect *two-body unitarity* which is implicit in Eq. (4.162). A rather general framework

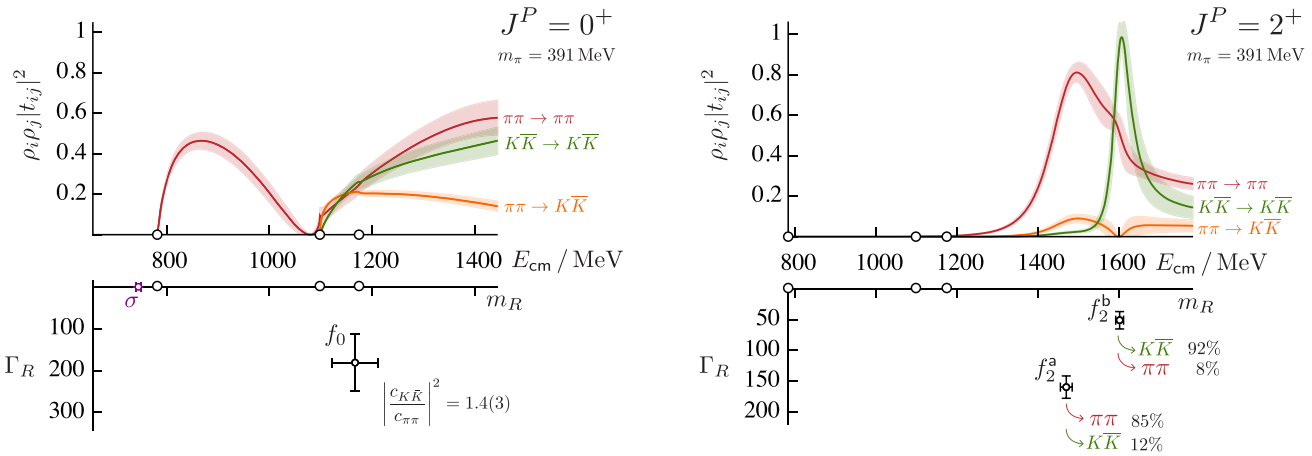


Fig. 50 Coupled $\pi\pi$, $K\bar{K}$ scattering (also $\eta\eta$, not shown) computed on three lattice volumes with $m_\pi = 391$ MeV. Taken from Ref. [560]. Lower panels show resonance pole locations found by analytically continuing into the complex energy plane. In $J^P = 0^+$ case, ratio of couplings of f_0 resonance to $\pi\pi$, $K\bar{K}$ given. In $J^P = 2^+$ case, *branching fractions* of two resonances to $\pi\pi$, $K\bar{K}$ final states are given

to achieve this is to use parameterizations of the K-matrix, which is flexible enough to handle both resonant and non-resonant cases in any number of channels.

The first lattice QCD calculation of coupled-channel scattering considered the πK, ηK system which was found to be almost decoupled, with resonances appearing coupled only to πK [577,578]. Since then there has been a steady stream of calculations of meson–meson scattering of gradually increasing complexity [535,560,566,579–584].

An example of what can be extracted from lattice QCD for coupled-channel scattering is shown in Fig. 50, taken from Ref. [560]. In this calculation of coupled $\pi\pi$, $K\bar{K}$, $\eta\eta$ scattering, performed with 391 MeV pions, the finite volume spectrum was found in three lattice volumes and several moving frames, leading to 57 energy levels to constrain the S-wave t-matrix and 36 levels to constrain the D-wave.

We observe a highly non-trivial energy-dependence in the S-wave where a broad enhancement at low energies is followed by a *dip* in the $\pi\pi \to \pi\pi$ amplitude at the $K\bar{K}$ threshold, while amplitudes leading to a $K\bar{K}$ final state turn on rapidly at threshold. While this energy dependence does not "by-eye" immediately suggest a simple resonance interpretation, the t-matrix can be analytically continued to complex energies, and two poles are found: one lies below $\pi\pi$ threshold and corresponds to the stable σ discussed earlier, while the second lies close to the $K\bar{K}$ threshold, and might be associated with the experimental $f_0(980)$ resonance (which also appears as a sharp dip in $\pi\pi$ scattering). This resonance pole has large couplings to both $\pi\pi$ and $K\bar{K}$. These results prove to be robust to variations in the detailed form of the amplitude parameterization.

The D-wave result reflects more closely our intuitive picture of resonances, with two bumps appearing, associated to two pole singularities. The lighter state dominantly couples to $\pi\pi$, and a heavier narrower state is dominantly coupled to $K\bar{K}$, a situation that is very similar to the experimental $f_2(1270)$, $f_2'(1525)$ states. The selective final state couplings reflect the 'OZI-rule' emerging dynamically from a non-perturbative calculation if we interpret the lighter state as dominantly $u\bar{u} + d\bar{d}$ and the heavier as dominantly $s\bar{s}$.

A different complication can occur when the scattering hadrons have non-zero spin. In this case, the same total J^P can be constructed by more than one hadron-spin, orbital angular momentum combination. For example, if one scatters a vector ω meson from a pion, $J^P = 1^+$ can be constructed from $\ell = 0$ or from $\ell = 2$, or using the spectroscopic notation, ${}^3S_1$, ${}^3D_1$. In this case, even if $\pi\omega$ is the only channel accessible, one still has a system of coupled-partial-waves, and a two-dimensional t-matrix.

A version of Eq. (4.162) still holds in such situations, and once again, provided enough energy levels can be computed in lattice QCD to provide sufficient constraint, the t-matrix can be determined. An example is shown in Fig. 51 where coupled $\pi\omega$, $\pi\phi$ scattering was studied with pions of mass 391 MeV. With light quarks as heavy as this, the ω and ϕ mesons are absolutely stable. A clear resonant behavior is observed which can be associated with the experimental $b_1(1235)$ state, and the couplings at the pole yield a value for the D/S amplitude ratio, a quantity that has been measured previously (references are listed in Ref. [513]).

The coupled channel technology has also been applied to scattering systems with charmed mesons [566,584], and recently, for the first time to a scattering system housing an exotic J^{PC} resonance believed to be a *hybrid meson* [582]. For meson resonances having decays only to one or more two-body final states, rigorous study within lattice QCD is

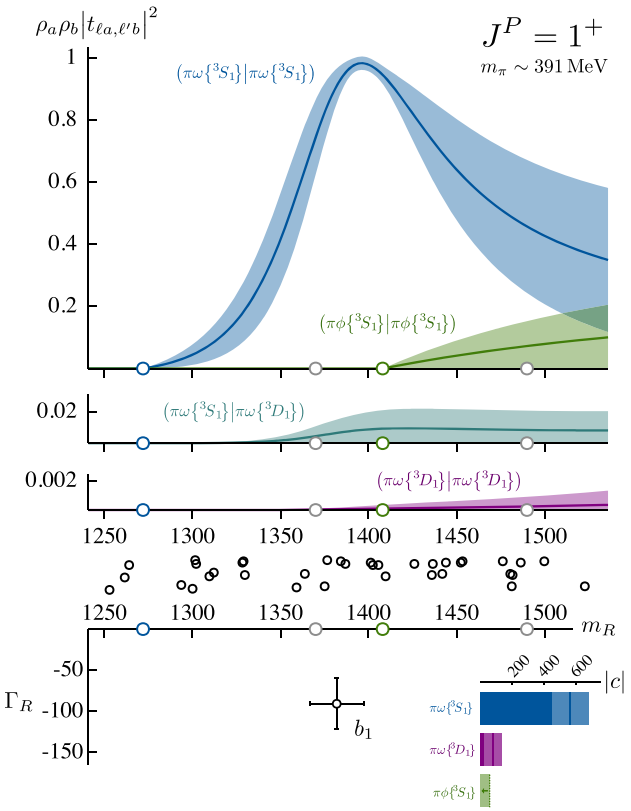


Fig. 51 Coupled $\pi\omega$, $\pi\phi$ scattering, with $\pi\omega$ in coupled partial waves, $^3S_1$, $^3D_1$, computed on three lattice volumes with $m_\pi = 391$ MeV. Taken from Ref. [581]. b_1 resonance pole and coupling to channels shown in bottom panel

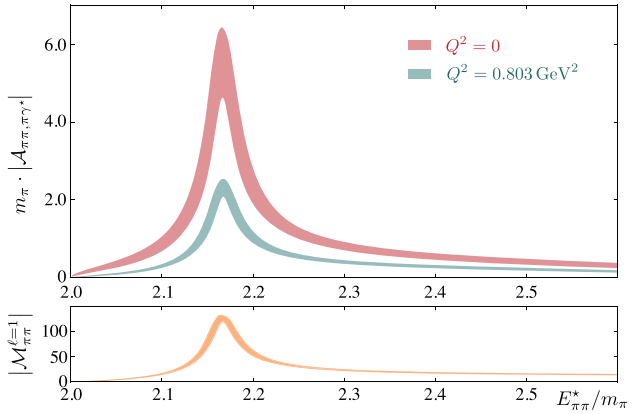


Fig. 52 Upper panel shows the transition amplitude for $\pi\gamma \to \pi\pi$ with $J^P = 1^-$ computed from a lattice QCD calculation with $m_\pi = 391$ MeV for two sample values of the photon virtuality. The lower panel shows the corresponding $\pi\pi \to \pi\pi$ elastic scattering amplitude. Taken from Ref. [586]

today a reality, with observables being the mass and width of the resonance, as well as the couplings to decay channels, all of which follow from scattering amplitudes. Going beyond this, more information about resonances can be obtained if we generalize away from scattering to also consider processes in which an external current probes the system.

4.5.7 Beyond scattering

An extension of the finite-volume formalism allows us to study systems in which a stable hadron emits or absorbs an electroweak current and transitions into a pair of strongly-interacting hadrons which may resonate. Applications include semileptonic heavy flavor decays with resonances in the final state, e.g. $B \to \ell^+\ell^- K^*$ where the K^* decays to $K\pi$. To date the only application of this technology has been to a simpler reaction, $\gamma\pi \to \pi\pi$, where the final state features the ρ resonance [585–587]. The approach requires first the determination of the $\pi\pi$ elastic scattering amplitude as described earlier, followed by computations of three-point correlation functions, from which transition matrix elements are extracted. The effect of the finite-volume is encoded in a correction to the normalization of the $\pi\pi$ state [588–590] that requires knowledge of the scattering amplitude. Figure 52 illustrates one result of such a calculation, showing the transition matrix element for $\pi\gamma \to \pi\pi$ (for two sample values of photon virtuality) along with the elastic $\pi\pi$ scattering amplitude – the clear ρ resonance is present in both.

As well as computation of experimentally measurable processes (such as the heavy flavor decays), this approach also allows us to compute in lattice QCD quantities that cannot be easily accessed in experiment. For example, analytically continuing the transition amplitude obtained above to the ρ resonance pole, one obtains a *resonance* transition form-factor $\rho \to \pi\gamma^*$, whose virtuality dependence can be used to infer structural information about the ρ. A recent extension of the finite-volume formalism [592] to be able to handle processes like $\pi\pi\gamma \to \pi\pi$ will allow us to compute the true *resonance form-factors*.

4.5.8 The three-hadron frontier and other challenges

The progress reported above in the two-hadron sector has opened up the world of hadron resonance spectroscopy to first principles study using lattice QCD, but to go further an extension in formalism is required. The applicability of Eq. (4.162) is limited to energies below the lowest *three-hadron threshold*, and this is particularly constraining as the light quark mass is decreased and the threshold for $\pi\pi\pi$ becomes very low, lower than the mass of most interesting resonances.

Development of finite-volume formalism to extend into the three-body sector has been underway for some time, making use of several approaches to three-body scattering, and they are now converging to a consensus, as reviewed in Ref. [593]. The resulting formalism is, as one might expect, significantly more complicated than in the two-body case, but

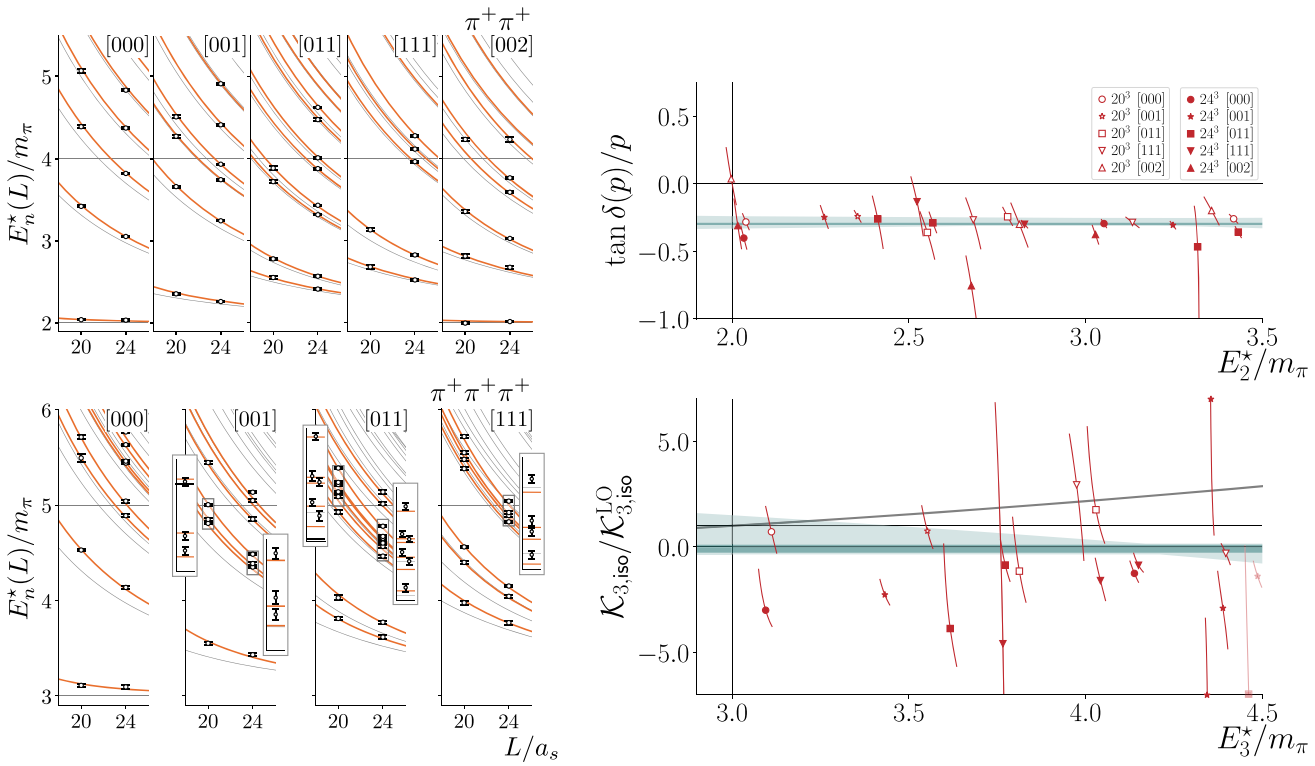


Fig. 53 A lattice QCD determination of the spectra in two volumes of isospin–2 $\pi\pi$ and isospin–3 $\pi\pi\pi$ with $m_\pi = 391$ MeV. Orange curves show a description of these spectra using two-body and three-body finite-volume formalism with the amplitudes shown on the right. Taken from Ref. [591]

the essential idea is still the same – the input from lattice is a set of discrete energy levels, now computed in channels with the quantum numbers of a three-hadron system.

The lattice QCD determinations of the finite-volume spectra follow a similar pattern to those described above, including now operators resembling systems of three-hadrons, but these are relatively straightforward to compute. The first investigations have focussed mostly on systems of maximal isospin [591,594–598], e.g. $\pi\pi\pi$ with isospin–3, where there are no resonances either in the three-hadron system, nor in the two-hadron subsystems.

An example is presented in Fig. 53 where we see discrete lattice QCD energy levels in two volumes for the $\pi\pi$ isospin–2 system and the $\pi\pi\pi$ isospin–3 system. These spectra can be described by two-body and three-body scattering amplitudes propagated through the finite-volume formalism, as shown by the orange curves. The amplitudes, as shown on the right of the figure (see the paper for the definition of the quantities plotted), are essentially structureless as expected in this non-resonant system. With proof-of-principle calculations like this one now done, the field is moving towards cases in which there are resonances, either in two-body subchannels, or in the three-body system, or both.

4.5.9 Summary

The progress in applying lattice to problems in hadron spectroscopy, as illustrated in this volume, suggests we have the beginnings of a rigorous foundation for the subfield, grounding it in first-principles QCD. The experimental hadron spectrum is already well studied, and there is a considerable corpus of model-based understanding, with which the lattice effort has to catch up. But already, with examples like the hybrid meson spectrum, lattice calculations are resolving long-standing conflicts. The ability to resolve excited hadrons as they truly are, as unstable resonances, makes a more direct connection to experiment possible, and the fact that calculations are possible of quantities which cannot be easily reached in experiments, like resonance form-factors, provides an opportunity to explore the internal structure of states that are otherwise poorly understood.

4.6 Hadron structure

Martha Constantinou and K. Orginos

The structure of the nucleon has been a central component to the development of QCD. Fundamental properties of strong interactions, such as asymptotic freedom, were discovered while trying to unravel the nature of the nucleon.

Hofstadter's elastic electron scattering experiments [599] discovered the first indications of a complex structure inside the proton. Later on, Deep Inelastic Scattering (DIS) discovered that partons, the constituents of the nucleon, are nearly free at short distances and led to the discovery of asymptotic freedom. Confinement, the fact that partons cannot break free from a hadron, is also a property of strong interactions that emerges from the study of hadronic structure. It was asymptotic freedom that eventually convinced theorists that QCD can describe the rich phenomenology of strong interactions.

Since its first exploration more than half a century ago, hadronic structure continues to be studied intensely both experimentally and theoretically. Theoretical studies include computations of various hadronic properties using lattice QCD, which offers a powerful non-perturbative, and systematically improvable way of computing fundamental properties of hadrons. This section summarizes the current status of lattice QCD calculations relevant to hadron structure. We start from simple observables such as nucleon charges which are important matrix elements for searches for physics beyond the standard model. We then proceed to a review of computations of nucleon form factors which are observables that give us information about the low energy structure of the hadron. Finally, we discuss modern methods for obtaining distribution functions from lattice QCD. Parton distribution functions are the simplest of such observables, which are relevant to understanding high-energy scattering experiments and give us a one-dimensional picture of the hadron. Generalized parton distribution functions (GPDs) and Transverse Momentum dependent distributions (TMDs) and their determination from lattice QCD will also be discussed.

4.6.1 Nucleon charges

Nucleon matrix elements of local quark bi-linear operators of the form $\mathcal{O}_{\Gamma,\tau}(t) = \bar{q}(t)\Gamma\tau q(t)$ define the nucleon charges. Here Γ is a general spin matrix and τ a flavor matrix. Isovector charges are obtained when $\tau = \tau_3$ the diagonal flavor Pauli matrix, while flavor diagonal charges are defined with an appropriate choice of τ that selects individual flavors. Nuclear matrix elements are obtained through computations of three-point functions of the form

$$C^{s,s'}_{\Gamma,\tau}(t',t) = \langle N^s(t)\mathcal{O}_{\Gamma,\tau}(t')\overline{N}^{s'}(0)\rangle, \tag{4.163}$$

where $N^s(t)$ is a nucleon interpolating field at time t, with helicity s and projected to zero momentum. Typical nucleon interpolating fields can be written as $\sum_{abc,ijk}\epsilon_{abc}C^s_{ijk}q^a_i q^b_j q^c_k$ with C^s_{ijk} appropriate weights. For a discussion of how these weights are obtained, see Ref. [600]. In the limit of $t \gg t' \gg 0$ the above correlator can be written as

$$C^{s,s'}_{\Gamma,\tau}(t',t) = z^s(z^{s'})^*\langle s|\mathcal{O}_{\Gamma,\tau}|s'\rangle e^{-M_N t} \tag{4.164}$$

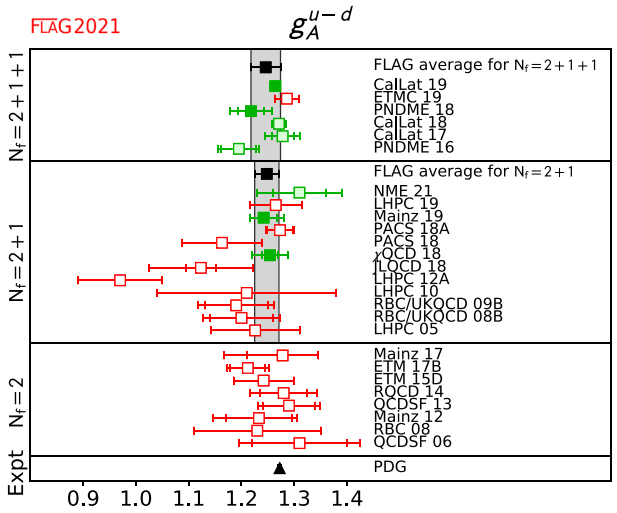


Fig. 54 Lattice QCD determinations of the isovector axial charge compared to the experimental world average is taken from PDG. Figure from Ref. [256], and reprinted based on the arXiv distribution license

where $\langle s|\mathcal{O}_{\Gamma,\tau}|s'\rangle$ is the desired nucleon matrix element and M_N is the nucleon mass and z^s is the overlap factor $\langle 0|N^s|s\rangle$. Using appropriate fitting procedure together with a nucleon two-point function

$$C(t) = \langle N^s(t)\overline{N}^s(0)\rangle = z^s(z^s)^* e^{-M_N t} \tag{4.165}$$

one obtains the desired matrix element. In general, these matrix elements require renormalization to obtain the matrix element at a given scale μ in a particular renormalization scheme. For a review of various methods used in lattice QCD to renormalize quark bi-linear operators, we refer the reader to Ref. [601]. Following this procedure, the nucleon charges have been obtained from lattice QCD. The isovector and flavor diagonal charges are essential quantities that, together with experimental observation, can constrain Beyond the Standard Model (BSM) theories. Therefore a significant effort in lattice QCD has been devoted to precise computations of the nucleon charges.

Establishing the lattice formulation of QCD requires that experimentally well-known quantities are correctly reproduced from numerical simulations. The axial charge of the nucleon, g_A, falls under this category and has been under investigation for several years. The field exhibits tremendous progress and among the highlights is the calculation of g_A with controlled statistical uncertainties. The Flavor Lattice Averaging Group (FLAG) periodically reviews lattice results on several quantities, including g_A, and produces the FLAG averages. In Fig. 54, we provide a summary plot of lattice calculations [256] demonstrating that lattice results have improved in accuracy over the years and recent calculations at the physical point agree with the experimental average.

The overall progress stimulated an intense activity in the field of hadron structure with the study of a large class of observables, some of which are known experimentally, but many that are still unexplored or difficult to measure [256, 602]. The investigations include nucleon charges such as the tensor and scalar and form factors for mesons and baryons. Selected results with simulations at physical quark masses can be found in Refs. [256,602].

4.6.2 Nucleon form factors

The Nucleon form factors are important properties of the nucleons that are essential for understanding their interactions in low-energy scattering experiments. They convey information about the internal structure of the hadron and their response to external probes, such as electromagnetic and weak currents. Properties such as the internal distribution of electric currents and charge and the size of the hadron can be deduced from electromagnetic form factors. Axial form factors describe the response of the hadron to external weak interaction probes. Future experiments, such as DUNE at Fermilab [603] and Hyper-Kamiokande [604], that aim to understand the properties of neutrinos, will require precise knowledge of the Nucleon axial form factors in order to achieve the precision they aim for. Therefore, lattice QCD computations of the Nucleon form factors are deemed essential and are vigorously pursued by several groups at this point. Advances in lattice QCD methods and computer hardware make such computations possible with sufficient precision to impact phenomenology [605].

Nucleon form factors are matrix element computations that require 3-point function computations

$$C^{s,s'}_{\Gamma,\tau}(t',t;\vec{p},\vec{p}') = \langle N^s(\vec{p},t)\mathcal{O}_{\Gamma,\tau}(t')\overline{N}^{s'}(\vec{p}',0)\rangle, \quad (4.166)$$

where $\vec{p}'$, $\vec{p}$ are the initial and final momenta of the hadrons. In the limit of $t \gg t' \gg 0$, the above correlator can be written as the matrix element associated with the form factor, which emerges as:

$$C^{s,s'}_{\Gamma,\tau}(t',t;\vec{p},\vec{p}') = z(p)^s z(p')^{s'*} e^{-E(p)(t-t')} \times \langle s,\vec{p}|\mathcal{O}_{\Gamma,\tau}|s',\vec{p}'\rangle e^{-E(p')t'} \quad (4.167)$$

where $E(p)$ is the energy of the nucleon with momentum p and z^s is the overlap factor $\langle 0|N^s(\vec{p})|s,\vec{p}\rangle$. The matrix element $\langle s,\vec{p}|\mathcal{O}_{\Gamma,\tau}|s',\vec{p}'\rangle$ is related to the appropriate form factor for the operator $\mathcal{O}_{\Gamma,\tau}$ and is extracted with appropriate fitting methodology (see Refs. [605–607] for details of some of these methods).

In the case of the electromagnetic form factor where $\Gamma = \gamma_\mu$ and the flavor matrix combines the flavors of quarks with their appropriate charges, the matrix element is

$$\langle s,\vec{p}|\sum_f e_f \bar{q}_f \gamma_\mu q_f|s',\vec{p}'\rangle = \bar{U}(\vec{p})\left[F_1(Q^2) + \frac{i\sigma_{\mu\nu}}{2M}q^\nu F_2(Q^2)\right]U(\vec{p}') \quad (4.168)$$

where $U(\vec{p})$ is the spinor associated with the nucleon, $q_\mu = p_\mu - p'_\mu$, $Q^2 = -q^2$, and F_1, F_2 the two Lorentz invariant Dirac and Pauli form factors. The electric and magnetic form factors are defined as

$$G_E(Q^2) = F_1(Q^2) - \frac{Q^2}{4M^2}F_2(Q^2)$$
$$G_M(Q^2) = F_1(Q^2) + F_2(Q^2). \quad (4.169)$$

With these form factors we can define the charge radius $\langle r_E^2\rangle$ and the magnetic radius $\langle r_M^2\rangle$ of the nucleon as

$$\langle r_E^2\rangle = -6\frac{dG_E(Q^2)}{dQ^2}\Big|_{Q^2=0}$$
$$\langle r_M^2\rangle = -\frac{6}{G_M(0)}\frac{dG_M(Q^2)}{dQ^2}\Big|_{Q^2=0}. \quad (4.170)$$

Because of the finite volume in lattice QCD computations, the form factors are only known on a set of discrete points. The full Q^2 dependence is recovered by fitting the data points to particular phenomenologically motivated forms. The simplest such form is the dipole:

$$F_{\text{dipole}}(Q^2) = \frac{r_F}{\left(1+\frac{Q^2}{M_F^2}\right)^2}, \quad (4.171)$$

where r_F is the residue and M_F^2 is a mass parameter associated with the form factor at hand. This simple parametrization works well for the lattice calculations that are typically restricted to low Q^2. Recently the z-expansion [608] given by

$$F(Q^2) = \sum_{k=0}^{\infty} a_k\, z(Q^2)^k, \quad (4.172)$$

with

$$z(Q^2) = \frac{\sqrt{t_{\text{cut}}+Q^2} - \sqrt{t_{\text{cut}}-t_0}}{\sqrt{t_{\text{cut}}+Q^2} + \sqrt{t_{\text{cut}}-t_0}}, \quad (4.173)$$

has been employed for a more flexible parametrization. The position of the cut, t_{cut}, is the time-like kinematic threshold for particle production associated with the current whose form factor is discussed. The parameter t_0 is the point in Q^2 that is mapped to $z = 0$ and is chosen for convenience.

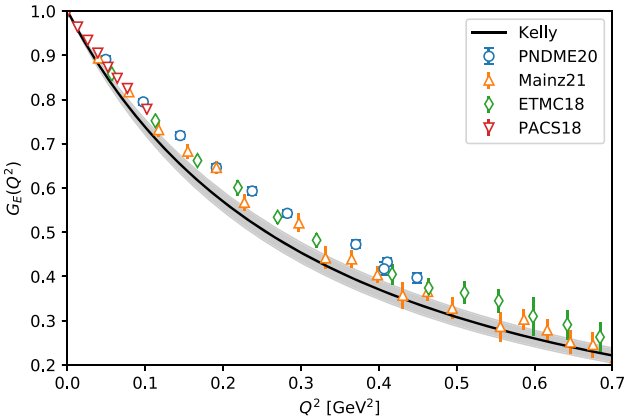


Fig. 55 Status of recent lattice QCD results for the isovector nucleon electric form factor in comparison with the Kelly parametrization of experimental results (figure from Ref. [609]). Reprinted under the terms of the Creative Commons Attribution 4.0 International license

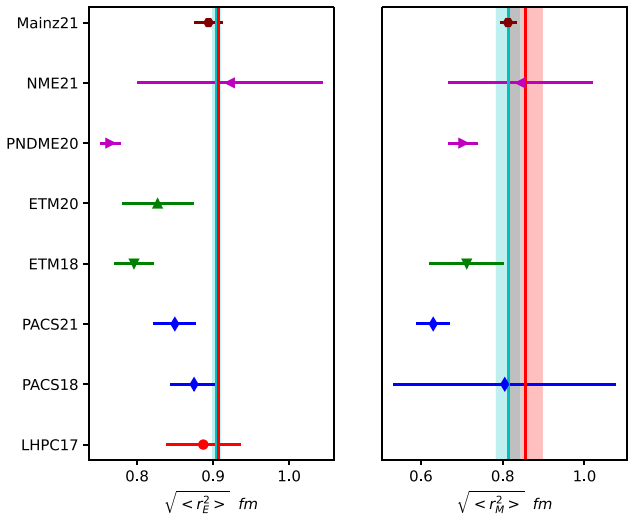


Fig. 56 Lattice results for charge magnetic radii of the nucleon. The vertical bands are the estimates from experiment (see text for details)

Multiple lattice QCD collaborations have recently computed the nucleon vector form factors. Several lattice collaborations have recently computed the isovector electric form factor (i.e., the difference between the proton and the neutron form factors). After many years of study of various systematics involved, we now have computations with physical quark masses, careful analysis of excited state contamination of the ground state matrix element, and large enough volumes to avoid finite volume effects. In Fig. 55, the lattice data together with the Kelly parametrization [610] of experimental results are presented. The lattice data of PNDME20 [607] are plotted as blue circles, the Mainz21 [611] data are the orange triangles, the ETMC18 [612] data are the green diamonds, and the PACS18 [613] data are the red triangles. All these calculations are performed with different methodologies and approaches in treating excited state effects and varying fermion actions in both the sea and the valence sectors. PNDME20 uses the HISQ action in the sea sector and smeared Clover action in the valence sector. The ETMC18 calculations use the twisted mass action. Both the Mainz21 and the PACS18 collaborations use Clover fermion actions. Clearly, there are some tensions between various collaborations that will be resolved in future, more refined calculations. However, it should be noted that there is a fairly good agreement between the state-of-the-art calculations and experiment.

Lattice QCD computations of the form factors can lead to the determination of the radii of the nucleon. In addition, direct methods of determining the nucleon radii also exist. Lattice QCD calculation results for the magnetic and the charge isovector radius of the nucleon are presented in Fig. 56. In this figure, the magenta right triangles are PNDME20 [607] using the mixed actions with Clover on HISQ, and the green triangles are from ETM18/20 [612,614] using the twisted-mass action. Calculations using the Clover fermion action are represented by the maroon octagons [611] from Mainz21, the blue diamonds from PACS18/20 [613,615], the red circle is from LHPC17, and the magenta left triangles from NME21 [606]. Note that results from [614,615] are obtained with methods that directly estimate the slope of the form factor at $Q^2 = 0$. The vertical bands represent the phenomenological values for the radii obtained from the experiment by combining data from the proton and the neutron. In particular the isovector charge r_E^{iv} and magnetic r_M^{iv} radii are given by

$$r_E^{iv} = \sqrt{r_{Ep}^2 - r_{En}^2} \quad , \quad r_M^{iv} = \sqrt{\frac{\mu_p r_{Mp}^2 - \mu_n r_{Mn}^2}{\mu_p - \mu_n}}, \tag{4.174}$$

where r_{Ep}^2, r_{En}^2 are the proton and neutron charge radii, r_{Mp}^2, r_{Mn}^2 are the proton and neutron magnetic radii, and μ_p, μ_n are the proton and neutron magnetic moments. By combining results exclusively from the particle data group (PDG) [616] we obtain the red bands. For the charge radius, the cyan band is obtained by using the CODATA2018 value for the proton charge radius and the neutron charge radius from the recent work in [617]. The cyan band for the magnetic radius was obtained using the proton radius obtained by [618] and the rest of the needed quantities from PDG.

In the case of the isovector axial form factors, one can take τ^+ as the flavor matrix and $\Gamma = i\gamma_5\gamma_\mu$ and the resulting matrix element is

$$\langle s, \vec{p}|\bar{q}(i\gamma_5\gamma_\mu\tau^+)q|s', \vec{p}'\rangle = \bar{U}(\vec{p})\left[F_A(Q^2) + \frac{q_\mu}{M}\gamma_5 F_P(Q^2)\right]U(\vec{p}'), \tag{4.175}$$

with F_A and F_P being the corresponding invariant form factors. In Fig. 57, recent lattice QCD computations of the

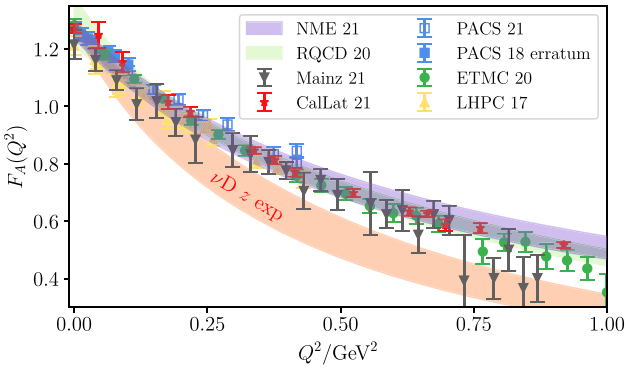


Fig. 57 Lattice QCD results for the nucleon axial form factor compared to the experimental results from neutrino deuteron scattering. Figure from Ref. [605] and reprinted under the terms of the Creative Commons Attribution 4.0 International license

axial form factor are presented. The red band denotes the parametrized experimental results from neutrino deuteron scattering [619]. The purple band are results from NME21 [606], and the green band are results from RQCD20 [620], where continuum, chiral and finite volume extrapolations have been performed. The rest contain results [613,615,621–624] from a few ensembles and are presented as points without the interpolating curves. It is clear that although there is tension between lattice QCD results and experiment, lattice QCD calculations are consistent with each other. As it is argued in Ref. [605] that lattice QCD calculations of axial nucleon form factors may play an essential role in future experiments and thus help us better understand neutrino physics.

4.6.3 Partonic structure

Information on the internal structure of hadrons is obtained through their partonic content, particularly parton PDFs, GPDs, and TMDs (see Sect. 10). These quantities are light-cone correlation functions and cannot be calculated using the Euclidean formulation of lattice QCD due to the rotation $t \to i\tau$. The most common avenue to proceed is to calculate Mellin moments of distribution functions, which provide partial information on distribution functions.

Lattice QCD calculations have focused on proton charges, vector, and axial form factors, that are, the first Mellin moments of PDFs and GPDs, respectively. There are also limited studies of the scalar and tensor charges, as well as the second Mellin moments of PDFs and GPDs.

In theory, one can use a large number of Mellin moments to reconstruct the parton distributions using an operator product expansion (OPE). Practically, a proper and exact reconstruction is not possible due to the challenges of calculating reliably high moments; the signal-to-noise rapidly decreases, and an unavoidable power-law mixing occurs beyond the fourth moment [625–629]. Therefore, alternative methods are needed to obtain the x dependence of distribution functions from a Euclidean formulation. The realization that matrix elements of momentum-boosted hadrons coupled with bilinear non-local operators can be related to light-cone distributions has transformed the field of PDF, GPDs, and TMDs calculations. The pioneering method of Large-Momentum Effective Theory (LaMET) that uses the aforementioned non-local operators has renewed the interest of the community to access the x dependence of parton distributions. Over the years, there have been several methods proposed: a technique based on the hadronic tensor [630–632], auxiliary quark field approaches [633–635], a method to obtain high Mellin moments using smeared operators [636], LaMET [637,638], pseudo-ITD [639], current–current correlators [640–642], and a method based on OPE [643].

In this review, we highlight selected results demonstrating the field's progress. More details can be found in the recent reviews [644–648].

Isovector PDFs

The isovector leading-twist PDFs have been the most well-studied and serve as a benchmark of the various methodologies to extract x dependence from lattice data. Results with ensembles at physical quark masses have already been obtained for the unpolarized [649–652], helicity [649,650,653] and transversity [650,654,655] PDFs for the proton. Here we focus on the unpolarized case that has the most results allowing comparison between different methods and lattice formulations. The work of Ref. [650] uses a twisted-mass fermions ensemble with physical pion mass and employs the quasi-PDFs method. The lattice spacing is about 0.09 fm, and the nucleon momentum boost is up to 1.4 GeV. The unpolarized PDF of Ref. [651] has been obtained using the pseudo-ITD framework on three clover Wilson ensembles with pion mass 172, 278, and 358 MeV; a chiral extrapolation has been performed to get the physical point. The pseudo-ITD methodology computes the Lorentz invariant amplitudes that contribute to the non-local matrix element and isolates the amplitude that contains the leading twist contribution. This amplitude is a function of the so-called Ioffe time ν, which is the Fourier-dual of the momentum fraction x [656–658]. The analysis of [651] includes lattice data up to Ioffe time $\nu = 8$ for the near-physical mass ensemble. Finally, the work of Ref. [652] extends and reanalyzes the data of Ref. [650] within the pseudo-ITD framework with up to $\nu = 8$. Having three independent calculations of the unpolarized PDF allows one to compare them and understand potentially systematic effects related to the method and computational setup. Such a comparison can be found in Ref. [651], which we include in

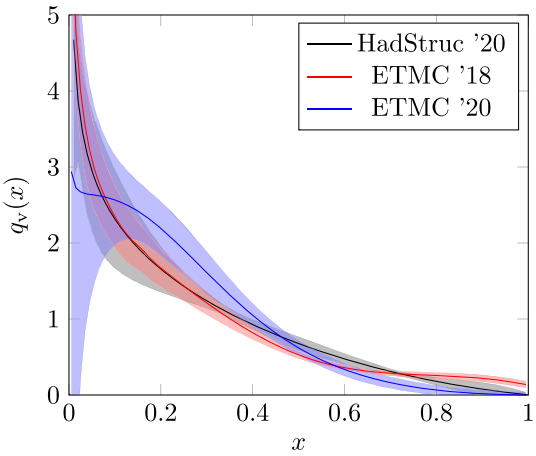


Fig. 58 Lattice results for the unpolarized PDF using quasi-PDFs [650] (red band) and pseudo-ITDs from Ref. [651] (gray band) and Ref. [652] (blue band). Plot from Ref. [651]. Reprinted under the terms of the Creative Commons Attribution 4.0 International license

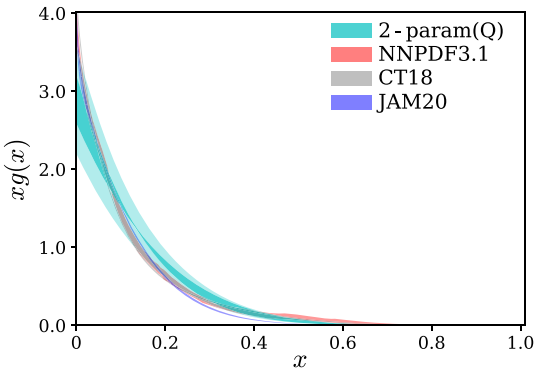


Fig. 59 Lattice QCD results on the gluon PDF from Ref. [659] (cyan band) compared to estimates from global analyses [663–665]. Reprinted under the terms of the Creative Commons Attribution 4.0 International license

Fig. 58. A good agreement is observed between the different calculations, which is very encouraging, as each methodology may suffer from different systematic effects.

Gluon PDFs

In general, gluon contributions are limitedly studied due to the enhanced gauge noise, the involvement of disconnected diagrams, and the challenges in the non-perturbative renormalization. In the case of x-dependent gluon PDFs, the renormalization cancels out using the pseudo-ITD method, which is a significant advantage. Recently, there have been calculations of the gluon PDF for the proton and the pion using the pseudo-ITD method [659,660]. Reference [659] presents a calculation using clover fermions at a pion mass $m_\pi = 358$ MeV. One novelty of the calculation is the use of the momentum-smeared distillation technique [661] to suppress gauge noise. The work also employs Jacobi polynomials to reconstruct the x dependence of the distribution [662]. The main results are shown in Fig. 59. The work of Ref. [660] presents a calculation of the gluon PDF for the pion using two HISQ coarse ensembles ($a = 0.12, 0.15$ fm) and pion masses $m_\pi = 220,\ 310,\ 690$ MeV. While the current status of gluon PDFs is exploratory, the available results are promising.

Individual quark PDFs

Calculations of individual-quark PDFs are challenging due to the involvement of disconnected diagrams that increases the statistical fluctuations of the correlators. The flavor decomposition of quark PDFs is interesting in its own right but is also needed to form the flavor-singlet combination to eliminate mixing with the gluon PDF. The mixing holds only for the unpolarized and helicity cases; there is no gluon transversity. Furthermore, the strange and charm quark PDFs are more susceptible to mixing as they enter the sea sector from gluon splitting. The effect of mixing is expected to be smaller for the light quarks that appear in the valence sector of the proton. The individual light quark unpolarized, helicity, and transversity PDFs were calculated in Refs. [666,667] using an ensemble of twisted mass fermions at $m_\pi = 260$ MeV. The work shows that disconnected contributions to the unpolarized and transversity PDFs are tiny and can be neglected. However, calculations at the physical value of the quark masses are needed to confirm this. References [666,667] include the strange quark contributions, which may have increased systematic effects due to the mixing with the gluon PDFs. The same holds for Ref. [668] (clover on HISQ, $m_\pi = 220,\ 310,\ 690$ MeV), which calculates the strange and charm quark PDFs for the proton.

GPDs

Another progress for lattice QCD is related to calculating x-dependent GPDs. These are computationally more expensive than PDFs due to the momentum transfer between the initial and final hadronic states. The momentum transfer must be equally split between the initial and final states, as the GPDs are defined in the symmetric frame; such a frame is computationally costly, preventing the extraction of GPDs for a dense set of values of t. A novel approach that related light-cone GPDs to Lorentz-invariant amplitudes has been recently proposed [669]. First results on the proton unpolarized and helicity GPDs have been obtained using the quasi-distribution approach [666]. The calculation is performed on a 260 MeV pion mass ensemble of twisted mass fermions. The work was extended for the chiral odd twist-2 GPDs in Ref. [670]. In Fig. 60, we compare the three types of GPDs for zero and nonzero skewness. As can be seen, the introduction of nonzero skewness leads to the appearance of a nontrivial ERBL region. Another calculation of the unpolarized GPDs can be found in Ref. [671], which was originally reported in a non-symmetric frame similar to the one used for frame-independent form factors.

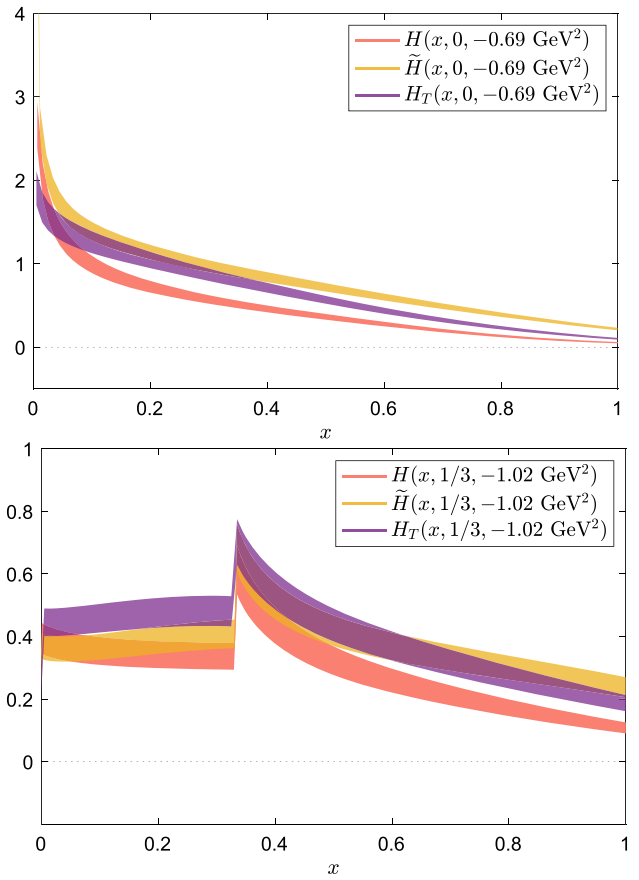


Fig. 60 Top: H, $\widetilde{H}$, H_T GPDs $t = -0.69\,\mathrm{GeV}^2$, $\xi = 0$. Bottom: H, $\widetilde{H}$, H_T GPDs $t = -1.02\,\mathrm{GeV}^2$, $\xi = 1/3$. The unpolarized, helicity, and transversity data are shown with red, yellow, and purple bands, respectively. Figure from Ref. [670] and reprinted under the terms of the Creative Commons Attribution 4.0 International license

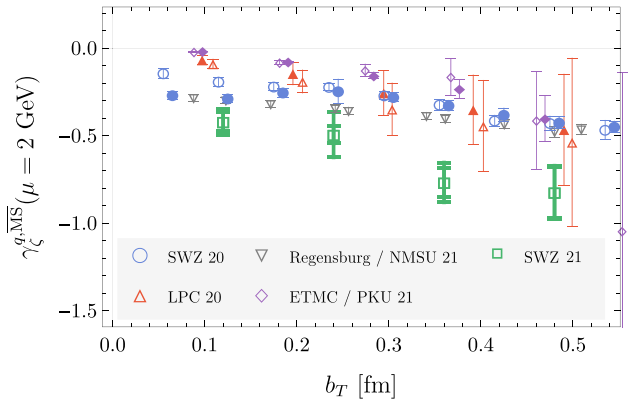


Fig. 61 Lattice QCD determinations of the Collins–Soper evolution kernel obtained from Ref. [672] (SWZ 20), Ref. [673] (LPC 20), Ref. [674] (Regensburg/NMSU 21), and Ref. [675] (ETMC/PKU 21), and Ref. [676] (SWZ 21). Figure adapted from Ref. [676] and reprinted under the terms of the Creative Commons Attribution 4.0 International license

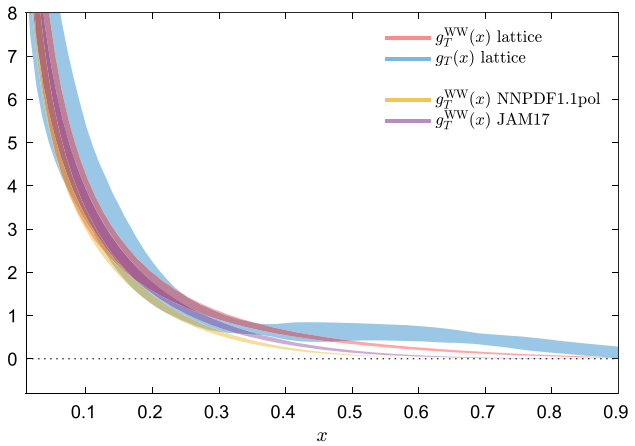


Fig. 62 The Wandzura–Wilczek approximation for g_T. Figure from Ref. [677] and reprinted under the terms of the Creative Commons Attribution 4.0 International license

TMDs

Unlike PDFs and GPDs, TMDs contain, in addition, rapidity divergences that require regularization. The regulator is encapsulated within the so-called soft function. The evolution in rapidity of the soft function can be studied separately through the Collins–Soper (CS) kernel. Aspects of the soft function are actively studied in lattice QCD [668,672–676,678], which is the ideal formulation as the soft-function is a non-perturbative quantity. A summary plot for the CS kernel is shown in Fig. 61.

Higher-twist

One of the latest developments in extracting x-dependent distribution functions is the exploration of twist-3 PDFs and GPDs that contain information on quark–gluon–quark correlations [679]. They are also related to the transverse force acting on transversely polarized quarks [680] and to the nuclear electric dipole moments [681]. First exploratory studies of twist-3 PDFs $e(x)$, $g_T(x)$, and $h_L(x)$ can be found in Refs. [677,682–684], with numerical results for $g_T(x)$ and $h_L(x)$. An interesting investigation of twist-3 PDFs is the Wandzura–Wilczek (WW) approximation [685] according to which the twist-3 g_T can be fully determined by its twist-2 counterpart, g_1. The WW approximation can also be studied for h_L. In Fig. 62 one can see g_T^{WW}, demonstrating that the approximation holds in some regions of x, but an overall violation of up to 40% is permitted. Note that the 2-parton twist-3 PDFs mix with quark–gluon–quark correlations and the mixing should be addressed within the matching kernel [686,687].

4.6.4 *Outlook*

Since the early days of lattice QCD in the 1980s, hadron structure calculations have been pursued vigorously. Over the years, the methods used to perform these calculations have improved steadily, and the Monte Carlo methods for sampling the QCD vacuum have reached the degree of effi-

ciency required for such computations. Furthermore, computer hardware has now reached the Exaflop era. As a result, calculations for hadron structure are now achieving unprecedented precision in some cases (ex., nucleon charges). In other cases, new horizons open up, such as the ability to compute the momentum fraction x-dependence of distribution functions. In the future, lattice QCD computations of hadronic structure will continue to improve and provide us with the theoretical input needed to understand strong interaction physics better.

4.7 Weak matrix elements

Christine Davies

Quarks have the special property that they experience all of the fundamental forces in the Standard Model. As well as exchanging the gluons that keep them confined into hadrons, quarks can also occasionally emit weak interaction W bosons or QED photons. Because W and γ have no color charge they escape cleanly from the hadron, carrying valuable information about the structure of the bound state. This structure is determined by strong interaction physics and so predictions from QCD can be tested against experimental information on these processes. The number of different quark flavors, and hadrons constructed from them, makes a rich mine for lattice QCD to work in.

In the bigger picture of the Standard Model we need to determine accurately the couplings between quarks and the W boson given by the elements of the CKM matrix ([688], Sect. 13.2). This programme is a crucial ingredient in constraining the possibilities for new physics beyond the Standard Model. However, quarks are not free particles when they emit W bosons. The experimental measurement of appropriate hadronic weak decay rates allows us to determine CKM elements but only if, as discussed above, we have understood the strong interaction physics that confines the quarks through calculation of the appropriate *hadronic* matrix elements of the weak current. As we will see below, some of the experimental information for weak (and electromagnetic) decay rates is very accurate and correspondingly accurate theoretical calculations in QCD are needed to make the most of it. These have always been a high priority for lattice QCD. Results have improved over time to the point where uncertainties are now below 1% in some cases. We will discuss the current status below, and briefly mention developments that will lead to improvements in future.

4.7.1 Decay constants

Decay constants are the hadronic parameters that encode the amplitude for finding the valence quark and anti-quark of a meson at the same point. This is then the parameter that is needed to determine the rate of annihilation of mesons

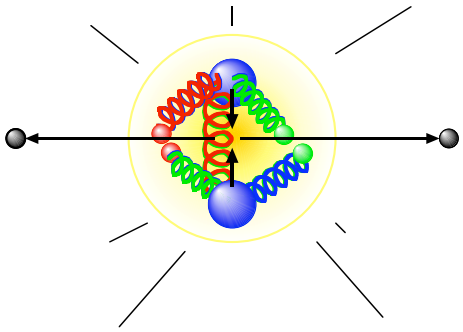


Fig. 63 Schematic diagram of a meson annihilation to leptons via the coupling of the valence quark–antiquark pair to a W or γ. The decay constant parameterises the amplitude to find the quark and antiquark at a point, the key hadronic information needed to determine the annihilation rate

with appropriate flavor quantum numbers to a W or γ (see Fig. 63). For a pseudoscalar meson the decay constant, f, is defined from the vacuum to meson matrix element of the axial current. For meson P of quark content $a\bar{b}$

$$\langle 0|\bar{a}\gamma_\mu\gamma_5 b|P(p)\rangle \equiv f_P\, p_\mu. \tag{4.176}$$

For a meson at rest, applying the partially-conserved axial current (PCAC) relation $\partial_\mu A^\mu = (m_a + m_b)P_s$ to relate axial-vector and pseudoscalar currents gives

$$(m_a + m_b)\langle 0|\bar{a}\gamma_5 b|P(\vec{p}=0)\rangle = f_P M_P^2, \tag{4.177}$$

where M_P is the meson mass.

In lattice QCD the matrix element on the l.h.s. of Eq. (4.176) or (4.177) is obtained from the two-point correlation function between source and sink $a\bar{b}$ currents (Sect. 4.2) with Euclidean time separation, t, between them. The two-point function has contributions, exponential in t, from a tower of $a\bar{b}$ mesons. The exponential corresponding to the ground-state (lowest mass) meson dominates at large t and this is the meson for which the parameters of the fit, the amplitude and mass, are most precisely determined. This mirrors experiment, where accurate meson weak or electromagnetic annihilation rates are possible when strong-interaction decay channels are heavily suppressed (not usually true for excited states). Note, however, that lattice QCD can determine f for mesons which do not have the flavor quantum numbers to annihilate to W or γ – these results are still useful in other contexts.

The fit to the two-point function $C(t)$ gives both the ground-state meson mass and its decay constant. The contribution of the ground-state to $C(t)$ is

$$C(t) = a_0(e^{-M_0 t} + e^{-M_0(T-t)}) + \cdots. \tag{4.178}$$

Here T is the lattice time extent and ... represents contributions from higher mass states. M_0 is the ground-state meson mass and the amplitude a_0 is given by

$$a_0 = (\langle 0|J|P_0\rangle)^2/(2M_0) \tag{4.179}$$

where J is the current used at the source and sink of $C(t)$. The decay constant for P_0 can then be obtained from a_0 using Eq. (4.176) or (4.177) as appropriate for J.

Decay constants for light pseudoscalar mesons (f_π and f_K) have been calculable in lattice QCD with errors at the few percent level since 2004 [689]. This was one of the first calculations to be done once ensembles of gluon field configurations were available (from the MILC collaboration) that included u, d and s sea quarks with multiple values of the lattice spacing and light enough u/d quarks for a reasonably well-controlled extrapolation to the physical continuum limit.

To achieve a small uncertainty in the result for the ground-state meson mass and decay constant it is important to have a large sample of correlators (to achieve small statistical errors) at multiple values of the lattice spacing using a discretisation of the QCD action with small discretisation errors (Sects. 4.2, 4.1). An accurate determination of the lattice spacing (to convert $C(t)$'s fit parameters from lattice units to GeV) is needed. Attention must also be paid to the effect of the finite-volume of the lattice on the π and K. Finite-volume (and discretisation) effects are incorporated into the chiral perturbation theory [690] used to fit the results as a function of u/d quark mass (or M_π) to extrapolate to the continuum limit with physical quark masses. M_π is used to fix the average u/d quark mass (u and d are taken to be degenerate in almost all calculations) and the physical value appropriate to a calculation in which the quark electric charges are ignored is the experimental value of the π^0 mass. M_K fixes the s quark mass and the physical value used is an average of the masses of K^0 and K^+ with an allowance for QED effects [689].

For decay constants a further important consideration is the normalisation of the axial vector current that appears in Eq. (4.176) so that it matches that of the continuum QCD current. For lattice QCD actions that have an exact PCAC relation (such as asqtad staggered quarks used in [689]) no renormalisation is needed. Rather than use the partially conserved axial current (which is a complicated point-split construction) it is easiest to use the pseudoscalar current, which is local, and calculate the decay constant directly from Eq. (4.177). The quark masses that appear in this expression are then the bare lattice quark masses being used in the calculation.

The key physics importance of the lattice QCD calculations of f_π and f_K is in determining the rate for π^+/K^+ annihilation to a W boson, which can be measured accurately in experiment. The annihilation rate for meson P with appropriate quark flavor quantum numbers is

$$\Gamma(P \to \ell\overline{\nu}) = \frac{G_F^2|V_{ab}|^2}{8\pi} f_P^2 m_\ell^2 M_P \left(1 - \frac{m_\ell^2}{M_P^2}\right)^2 \tag{4.180}$$

up to well-studied QED corrections. Only the A of the $V-A$ weak interaction contributes in this case, so that $\Gamma \propto f_P^2$. V_{ab} is the appropriate CKM element; this can be determined from the experimental measurement of Γ given a value for f_P from lattice QCD.

Several systematic errors are reduced in an analysis of the ratio of widths for K and π [693]. This enables the ratio $|V_{us}|/|V_{ud}|$ to be determined and converted to a result for $|V_{us}|$ using accurate $|V_{ud}|$ values from super-allowed nuclear β decay [616]. Lattice QCD calculations have then largely concentrated on determining the ratio f_K/f_π, equivalent to fixing the lattice spacing from f_π. Following a great deal of work by the lattice community, current day results have improved to the point where the uncertainty on f_{K^+}/f_{π^+} is reduced to 0.2%. The recent FLAG review [256] quotes an average of

$$f_{K^+}/f_{\pi^+} = 1.1932(21), \quad n_f = 2+1+1 \tag{4.181}$$

from lattice QCD results that include u, d, s and c quarks in the sea obtained in Refs. [692,694–696]. The average is dominated by the result from the Fermilab Lattice/MILC collaborations [692]. The lattice calculations now include an analysis of the impact of the u/d mass difference; work is ongoing to analyse QED effects on the lattice [697].

Heavier pseudoscalar mesons also annihilate to Ws, giving access to other CKM elements. For example, the rate for $B \to \ell\overline{\nu}$ depends on $|V_{ub}|$ and f_B. The experimental determination of the decay rates is harder and they currently have larger uncertainties than for K and π [616]. On the lattice QCD side the heavier masses of the c and b quarks increase discretisation errors, since they take the form of powers of ma for quark mass m. To counteract this lattice QCD theorists must improve the discretisation of the QCD (Dirac) action to increase the power of ma (for $ma < 1$) with which these errors first appear. A very successful action in this regard is the Highly Improved Staggered Quark (HISQ) action [331] developed by the HPQCD collaboration, with tree-level discretisation errors starting at $(ma)^4$.

This discretisation allowed the first 1% accurate calculations for charmed meson decay constants [703]. The current state-of-the-art results are from the Fermilab Lattice/MILC collaborations using HISQ quarks and have 0.3% uncertainties [692]. The dominant uncertainty in the values of V_{cs} and V_{cd} from meson leptonic decay is then from the experimental decay rate [616].

For b quarks discretisation errors are even more of a headache. During the 1990s methods were developed that exploited the nonrelativistic nature of the b quark in its bound states, thus removing the b quark mass as a dynamical scale (so that discretisation errors instead depend on the much smaller scales of the b quark kinetic energy and momentum). These approaches are based on the discretisation onto a lattice of Heavy Quark Effective theory (HQET) [704] (for 'heavy-

light' hadrons) and of non-relativistic QCD (NRQCD) [289] (applicable also to heavyonium). It was also shown that the large-mass limit of the clover-improved Wilson quark action [307] could be interpreted as a nonrelativistic effective theory [705]. A limitation of these formalisms is the need to normalise the weak current to match that of continuum QCD; this requires challenging calculations in lattice QCD perturbation theory and has only been done through $\mathcal{O}(\alpha_s)$ [706–708]. The ETM collaboration developed a ratio approach [709] to interpolate between results for quark masses around c using the twisted mass quark formalism [316] and the infinite-mass (static) limit. These methods have been able to achieve a 2% uncertainty on B decay constants [709,710].

As increased computational power could be exploited to generate gluon field configurations with finer values of the lattice spacing, alternative methods became available. The MILC collaboration led the way including $2+1$ flavors of asqtad sea quarks with a range of lattice spacing values down to $a = 0.044$ fm. On these lattices the HPQCD collaboration showed that b quarks could be treated with the relativistic HISQ formalism (with its absolute current normalisation) if calculations were done for a range of quark masses $> m_c$ and a range of lattice spacing values [691]. Fig. 64 shows the lattice results for the heavy-strange meson along with the joint fit of the dependence on the heavy meson mass and the lattice spacing. This enables a curve for the dependence of the decay constant on the heavy meson mass to be obtained in the continuum limit, from which the decay constant for the B_s meson can be read. At the same time the dependence on heavy meson mass becomes clear; $f_{D_s} > f_{B_s}$ but only by about 10%, rather less than the leading order result from HQET, $f_P\sqrt{M_P} =$ constant [711] would suggest. The Fermilab Lattice/MILC collaborations have now extended this to B mesons and including $2+1+1$ flavors of HISQ sea quarks for uncertainties on f_B and f_{B_s} below 1% [692].

The SU(3)-isospin-breaking ratio of decay constants, f_{P_s}/f_P, is calculated to better than 0.4% in Ref. [692] with results summarized in Fig. 65. The ratios are all close to 1.2 but there are small and significant differences as the mesons increase in mass from K/π to B_s/B.

Vector mesons with appropriate quark flavor quantum numbers can also annihilate to leptons via a W boson. Although the decay rate is not suppressed by lepton masses in that case (because of the meson spin) it is nevertheless hard to see experimentally because it is overwhelmed by the QED radiative decay $V \to P\gamma$; it may be possible in future for the D_s^* [712]. The vector leptonic decay proceeds through the vector piece of the weak current and is determined by the corresponding vector decay constant. The lattice QCD vector current must again be normalized to match continuum QCD. Although in principle a conserved vector current can be used, it is easier to use a local vector current and renormalise it. There are a number of techniques to do this (Sect. 4.2). The

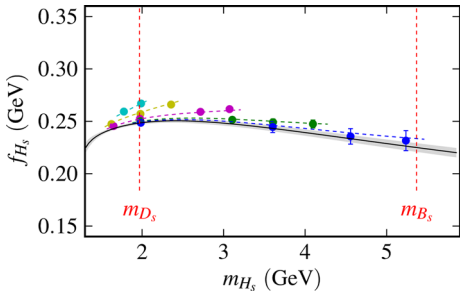


Fig. 64 The decay constant of the heavy-strange pseudoscalar meson as a function of its mass from lattice QCD calculations [691] using the HISQ action [331]. Points with different colors are results for different lattice spacing values, with smaller lattice spacings having more reach to heavier masses. The grey curve is the continuum limit of an HQET-inspired fit to the results including discretisation effects. The result for f_{B_s} can be read off at the mass of the B_s meson

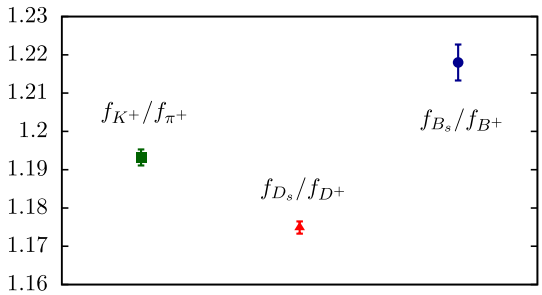


Fig. 65 SU(3)-isospin breaking ratios of decay constants from lattice QCD. f_K/f_π is from Eq. (4.181) [68], other results from Ref. [692]

ratio of vector to pseudoscalar decay constants for heavy-light mesons has been calculated using NRQCD [700] (with perturbative renormalisation [713]) and using twisted-mass quarks [699] (using a MOM scheme [714]). Interestingly it is found that the ratio of f_V/f_P is larger than 1 for D mesons and less than 1 for B mesons. Reference [699] gives 1.078(36) for f_{D^*}/f_D and 0.958(22) for f_{B^*}/f_B.

Vector $q\bar{q}$ mesons can annihilate to $\ell\bar{\ell}$ via a γ, and such decay rates have been determined experimentally to better than 2% for heavyonium mesons [616]. This provides an excellent opportunity for accurate comparison of lattice QCD and experiment for a decay rate free from CKM elements since

$$\Gamma(V \to \ell^+\ell^-) = \frac{4\pi\alpha^2 e_q^2}{3}\frac{f_V^2}{M_V}, \tag{4.182}$$

with e_q the valence quark electric charge in units of e. Results for $f_{J/\psi}$ [279] and f_Υ [702] calculated with HISQ quarks, normalized via an SMOM scheme [715,716] show good agreement with values inferred from the experimental decay rates, providing a solid underpinning for the other decay constants being discussed here.

Figure 66 summarises the values of meson decay constants that are well-determined in lattice QCD, arranged by

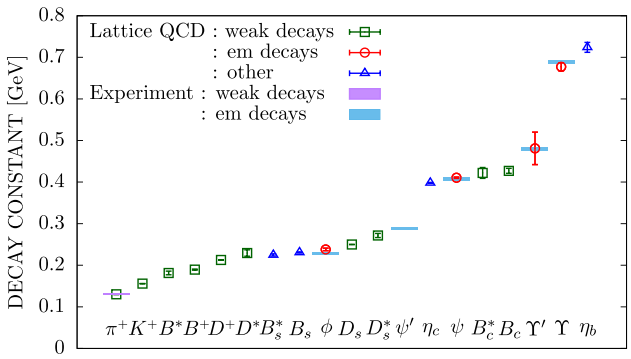


Fig. 66 Summary of meson decay constant values calculated in lattice QCD and arranged in order of their size. Points with error bars use different symbols for values needed to determine weak or QED leptonic decay rates or those not linked to any simple decay mode. The decay constants inferred from experimental values for QED leptonic decay are given by blue bands. For weak decays, experimental results must be combined with lattice QCD to obtain CKM elements; f_π can be inferred from the π^+ leptonic rate taking $|V_{ud}|$ from nuclear β decay [616] and is shown by a purple band at 130.56(14) MeV. The lattice QCD result for f_π comes from RBC/UKQCD [285], using the Ω baryon mass to fix the lattice spacing. Other results shown use f_π to determine the lattice spacing, and so do not give a value for that quantity. f_K is taken from Eq. 4.181; f_B, f_D, f_{D_s} from Ref. [692]; f_{B_c} [698], $f_{D^*_{(s)}}/f_{D_{(s)}}$ and $f_{B^*_{(s)}}/f_{B_{(s)}}$ [699]; $f_{B^*_c}/f_{B_c}$ [700]; f_ϕ [701] and charmonium and bottomonium results [279,281,702]

value order. It does not include values for mesons, such as the ρ or K^*, that have a large decay width from a strong-interaction decay mode (Sect. 4.5). Notice that the range of decay constant values, from $f_{\pi^+} = 130.2(9)$ MeV [285] to $f_{\eta_b} = 724(12)$ MeV [702] is much smaller than the range of meson masses. As discussed above, decay constants reflect meson internal structure set by momenta inside the bound state rather than quark masses. For mesons containing u/d quarks the range of variation is even smaller, less than a factor of two from f_π to $f_{D^+} = 212.7(6)$ MeV [692], and the ordering is not intuitively obvious. Results are shown for decay constants relevant to weak leptonic decays (where comparison to experimental results yields a determination of the relevant CKM element) as well as those relevant to QED leptonic decays (where direct comparison to experimental rates is possible). It also includes decay constants that cannot be simply related to a decay process, but which nevertheless help to fill in the 'big picture' that we now have from lattice QCD for these simple matrix elements.

4.7.2 Mixing matrix elements and bag parameters

A fascinating phenomenon for neutral K and B mesons is that of 'oscillations', induced by the tiny weak interaction coupling between the mesons and their antiparticles. For exact CP invariance the eigenstates of the Hamiltonian are then $+/-$ combinations of the strong-interaction P^0 and $\overline{P}^0$ states, analogous to the eigenstates of two weakly coupled pendulums. An initial P^0 beam, created by a strong-interaction process, is equivalent to setting one pendulum swinging. At later times it becomes clear that the other pendulum is swinging/$\overline{P}^0$ is present (from interrogating the beam via suitable decay processes). The oscillation frequency is set by the eigenstate mass difference ΔM_P and can be measured very precisely in experiment. The coupling is a second-order weak process with the short-distance contribution given by the 'box diagram' of Fig. 67. As such it is sensitive to new physics that can be tested with accurate matrix elements for the box diagram between P^0 and $\overline{P}^0$, calculated in lattice QCD.

At the hadronic mass scales of the lattice the box diagram shrinks to an effective 4-quark operator (multiplied by a Wilson coefficient). For the SM case, the 'left-left' operator is

$$\mathcal{O}^{(1)} = \left[\overline{h}^\alpha \gamma_\mu (1-\gamma_5)\ell^\alpha\right]\left[\overline{h}^\beta \gamma_\mu (1-\gamma_5)\ell^\beta\right]. \tag{4.183}$$

h is either s or b and α/β are color indices. Matrix elements of further (BSM) 4-quark operators have also been calculated, see Ref. [256].

The matrix element of Eq. (4.183) between P^0 and $\overline{P}^0$, having a hadron on either end, is much harder to determine in lattice QCD than a decay constant, so results are not as mature and have larger uncertainties. The renormalisation of the 4-quark operator to match continuum QCD is also more challenging. Results are most usefully presented in terms of 'bag parameters' by removing factors of masses and decay constants from the matrix elements that would appear in the 'vacuum saturation approximation', i.e. inserting $|0\rangle\langle 0|$ between the two halves of the 4-quark operator. For $\mathcal{O}^{(1)}$ this gives [717]

$$\langle P^0|\mathcal{O}^{(1)}|\overline{P}^0\rangle = \frac{8}{3} f_P^2 M_P^2 B_P^{(1)}(\mu) \tag{4.184}$$

where the leftover 'fudge factor', B_P, is the bag parameter. It is normally quoted in the $\overline{\text{MS}}$ scheme; note its scale-dependence. Historically the assumption was then made that $B \approx 1$ but lattice QCD can achieve a much better result than this.

The bag parameter is often converted from $B^{(1)}(\mu)$ to its renormalisation-group-invariant (RGI) value,

$$\hat{B}^{(1)} = c_{\text{RGI}} B^{(1)}(\mu) \tag{4.185}$$

where c_{RGI} is calculated to two-loops in perturbative QCD [256] and takes values 1.369 for B_K (when $\mu = 2$ GeV) and 1.516 for B_B (when $\mu = m_b$)).

Reference [256] quotes an average for $\hat{B}_K^{(1)} = 0.7625(97)$ as an average of several lattice QCD results [285,718–720] using different lattice QCD actions and renormalisation approaches with $n_f = 2+1$ sea quarks; Ref. [721] gives an $n_f = 2+1+1$ result. B meson results are less accu-

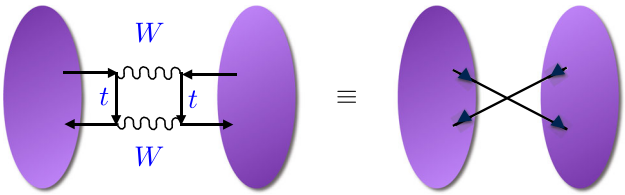


Fig. 67 Schematic diagram of the short-distance contribution to neutral meson mixing via the 'box diagram' (left) involving W bosons and top quarks. The matrix element that must be calculated in lattice QCD is that of the equivalent 4-quark operator (right)

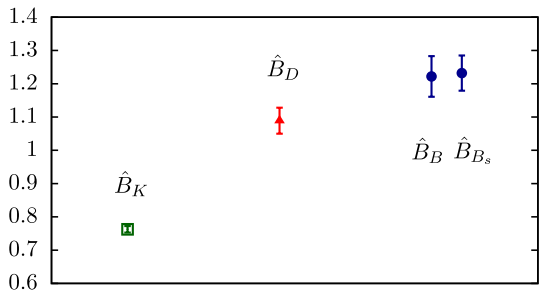


Fig. 68 A comparison of RGI bag parameters from lattice QCD for K^0, D^0, B^0 and B_s, showing significant deviations from the naive vacuum saturation approximation estimates of 1 and a trend with meson mass

rate, because of a significantly worse signal/noise problem in the determination of the correlation functions [722]; direct determination of B_B rather than $\mathcal{O}^{(1)}$ matrix elements cancels discretisation and light quark mass effects, however. Results including $n_f = 2+1+1$ sea quarks are available from HPQCD using NRQCD b quarks (with $\mathcal{O}(\alpha_s)$ renormalisation [723]), giving $\hat{B}^{(1)}_{B_d} = 1.222(61)$ and $\hat{B}^{(1)}_{B_s} = 1.232(53)$) [722]; $n_f = 2+1$ results using other lattice QCD actions are given in [724–726]. Note that $\hat{B}_{B_s}/\hat{B}_{B_d}$ is consistent with 1 (1.008(25) from [722]), showing that the SU(3)-breaking in the 4-quark matrix elements is entirely that of the decay constants.

Figure 68 compares the results for $\hat{B}$, including a value for $\hat{B}_D$ [721] that lies between $\hat{B}_K$ and $\hat{B}_B$. The D^0 box diagram is mediated by down-type quarks and is expected to contribute only a small part of ΔM_D, dominated by long-distance contributions. The short-distance results can be used to constrain new physics, however, see Ref. [727].

For B/B_s mesons the box diagram with top quarks of Fig. 67 dominates mixing (since $V_{tb} \approx 1$) so that

$$\Delta M_q = \frac{G_F^2 M_W^2 M_{B_q}}{6\pi^2} S_0(x_t)\eta_{2B} \left|V_{tq}^* V_{tb}\right|^2 f_{B_q}^2 \hat{B}^{(1)}_{B_q}, \quad (4.186)$$

and lattice QCD results for the bag parameters can be combined with (the very accurate) experimental results for the oscillation frequency to determine CKM elements $|V_{ts}|$ and $|V_{td}|$ that multiply the effective 4-quark operator. Agreement is seen within 2σ with CKM values from tree-level weak decays and unitarity [722].

For K oscillations the situation is more complicated because of sizeable long-distance contributions to ΔM_K involving u- and c-mediated contributions. At the same time analysis of $K \to \pi\pi$ amplitudes [256] is also needed to determine the direct and indirect CP-violation parameters, ϵ' and ϵ that describe the CP-properties of the mass eigenstates and their decays. These are very hard calculations that have required the development of new techniques, and results are still at a fairly early stage, e.g. often only available at one value of the lattice spacing. The RBC/UKQCD collaboration has led the way here, exploiting the excellent chiral properties of the domain-wall quark action. They have calculated the amplitude A_2 to the isospin 2 two-pion state (the $\Delta I = 3/2$ amplitude) [728] and the amplitude A_0 to the isospin 0 state ($\Delta I = 1/2$) [729]. This enables a result of Ref. [729]

$$\epsilon'/\epsilon = 21.7(8.4) \times 10^{-4} \quad (4.187)$$

in good agreement with experiment ($16.6(2.3)\times 10^{-4}$), suggesting no violation of the CKM paradigm at this level of accuracy. At the same time the lattice QCD results provide some insight into the observed $\Delta I = 1/2$ rule by which A_0 exceeds A_2 by a factor of 20. A factor of 2 is provided by perturbative QCD corrections to the coefficients of the appropriate 4-quark operators; lattice QCD shows that the other factor of 10 arises from the fact that, contrary to naive expectations, the contributions from different color contractions of the dominant operator tend to cancel in A_2 and reinforce each other in A_0 [729,730]. The development of methods to determine the long-distance contributions to ΔM_K [731] are also aimed at long-distance contributions to $K^+ \to \pi^+\nu\overline{\nu}$ and $K \to \pi\ell^+\ell^-$.

Future improvements here require improved renormalisation techniques for lattice 4-quark operators. Gradient flow methods look promising here, see e.g. Ref. [732].

4.7.3 Form factors

Semileptonic weak decays of hadrons in which one of the constituent quarks changes flavor and the virtual W boson emitted is seen as a $\ell\overline{\nu}_l$ pair (see Fig. 69) provide a huge range of possibilities for determining CKM elements and understanding hadron structure. The hadronic parameters that control the rate of these processes are known as form factors and they are functions of q^2, the squared 4-momentum transfer from parent hadron, ϱ, to child, χ. The kinematic range of q^2 is from $q^2_{\rm max} = (M_\varrho - M_\chi)^2$ (where ℓ and $\overline{\nu}_\ell$ have maximum back-to-back momentum in the ϱ rest-frame) to 0 (where χ and the $\ell\overline{\nu}_\ell$ pair are back-to-back). The form factors are largest at $q^2_{\rm max}$ and fall towards $q^2 = 0$, reflecting the internal momentum transfer via gluon exchange necessary to achieve the final state configuration.

Form factors are defined from matrix elements between ϱ and χ of weak currents. The simplest situation is when

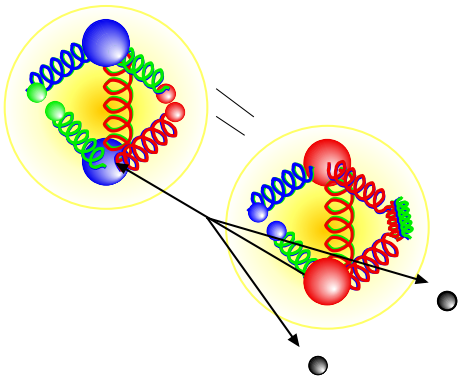


Fig. 69 Schematic diagram of a meson to meson semileptonic decay. The hadronic information needed to determine the rate is parameterized by form factors

both ϱ and χ are pseudoscalar mesons. In that case only the vector current and vector form factor, $f_+(q^2)$, contribute to the decay rate for $\varrho \to \chi \ell \overline{\nu}$ for zero lepton mass, with m_ℓ-dependent corrections from the scalar form factor, $f_0(q^2)$. We have

$$\frac{d\Gamma}{dq^2} = \frac{G_F^2}{24\pi^3}|V_{ab}|^2(1-\epsilon)^2 \times \left[|\vec{p}_\chi|^3(1+\frac{\epsilon}{2})|f_+(q^2)|^2 + |\vec{p}_\chi| M_\varrho^2 \left(1-\frac{M_\chi^2}{M_\varrho^2}\right)^2 \frac{3\epsilon}{8}|f_0(q^2)|^2\right] \tag{4.188}$$

for quark transition $a \to b$, $\epsilon = m_\ell^2/q^2$, and $\vec{p}_\chi$ is the 3-momentum of child χ in ϱ's rest frame. The form factors are defined from matrix elements

$$\langle \chi | V^\mu | \varrho \rangle = f_+^{\varrho\to\chi}(q^2)\left[p_\varrho^\mu + p_\chi^\mu - \frac{M_\varrho^2 - M_\chi^2}{q^2}q^\mu\right] + f_0^{\varrho\to\chi}(q^2)\frac{M_\varrho^2 - M_\chi^2}{q^2}q^\mu, \tag{4.189}$$

$$\langle \chi | S | \varrho \rangle = \frac{M_\varrho^2 - M_\chi^2}{m_a - m_b} f_0^{\varrho\to\chi}(q^2), \tag{4.190}$$

with kinematic constraint $f_+(0) = f_0(0)$. Equation (4.190) makes use of the partially conserved vector current relation $\partial_\mu V^\mu = (m_a - m_b)S$ that means f_0 is correctly normalized in lattice QCD [733]. The renormalisation factor, Z_V, for the vector current can then be determined by, for example, matching $f_0(q^2_{\rm max})$ from Eqs. (4.189) and (4.190) (see Ref. [734]).

To determine the form factors in lattice QCD requires the calculation of three-point correlation functions with appropriate source and sink operators for parent and child hadrons, and a current insertion at an intermediate time between them. Usually the parent hadron is taken to be at rest on the lattice and different spatial momenta are given to the child to map out the q^2 range. Fitting the three-point correlation function simultaneously with the two-point correlation functions for parent and child allows the parent-to-child matrix elements to be determined and Eqs. (4.189) and (4.190) applied. To obtain form factors in the continuum limit, interpolation in q^2 and extrapolation to $a = 0$ and physical quark masses is needed. Modern calculations (see, for example, Ref. [734]) transform q^2 into a region within the unit circle in z-space and then apply a polynomial fit in z that allows for discretisation effects and mistuning of quark masses.

The channel $K \to \pi \ell \overline{\nu}$ is a key for the determination of V_{us}. The q^2 range for this decay is very small and so conventionally experiment accounts for the q^2 dependence of Eq. (4.188) and gives the final result as a value for $|V_{us}|f_+(0)$. Combining charged and neutral meson decay rates with QED radiative and strong-isospin-breaking corrections gives a result with 0.2% accuracy : $V_{us} f_+(0) = 0.21635(39)(3)$ [735], where the first, dominant, error is from the experiment. The 0.2% accuracy is now also available from lattice QCD with 2+1+1 flavors. Reference [256] gives $f_+(0) = 0.9698(17)$ from averaging [736,737]. The two lattice QCD calculations take contrasting approaches. Reference [736] determines $f_+(q^2)$ and $f_0(q^2)$, interpolating to $q^2 = 0$ and testing q^2 dependence against experiment; Ref. [737] tunes to $q^2 = 0$ using twisted boundary conditions [738] and calculates $f_0(0)$ since this needs no renormalisation. The result for V_{us} from $K \to \pi \ell \overline{\nu}$ then shows an intriguing 3σ tension with CKM first row unitarity [735] and 2.5σ tension with V_{us} from $K \to \ell \overline{\nu}$ [616].

D meson decays (to K or π) have a larger q^2 range and experimental data is available in bins of q^2. This provides the opportunity to test the q^2-dependence predicted by QCD against experiment as well as to determine V_{cs} and V_{cd}. Figure 70 shows how this is done [734]. The upper plot shows the determination of the f_+ and f_0 form factors and the lower plot shows the result of determining V_{cs} bin-by-bin in q^2 using Eq. (4.188). A good fit is obtained to a constant with $V_{cs} = 0.9663(80)$, with errors from lattice QCD, experiment and QED corrections making similar contributions to the total uncertainty. See Ref. [739] for a determination of V_{cd} using lattice QCD $D \to \pi$ form factors.

The semileptonic decays of B mesons have a huge potential in searches for new physics as well as in giving access to key CKM elements V_{ub} and V_{cb}. Form factors for these decays are challenging for lattice QCD, however, because the large b quark mass means a large q^2 range. To reach $q^2 = 0$ the child spatial momentum must approximate $M_B/2$. Large values of $a|\vec{p}|$ induce poor signal/noise in correlation functions as well as discretisation effects, so early lattice QCD calculations worked close to $q^2_{\rm max}$ with nonrelativistic formalisms for the b quark.

To determine V_{ub} from $B \to \pi \ell \overline{\nu}$, Ref. [256] performs a joint fit to lattice form factor results from Refs. [741,742] (which use different variants of the improved Wilson action for the b quark and different light quarks) and experimen-

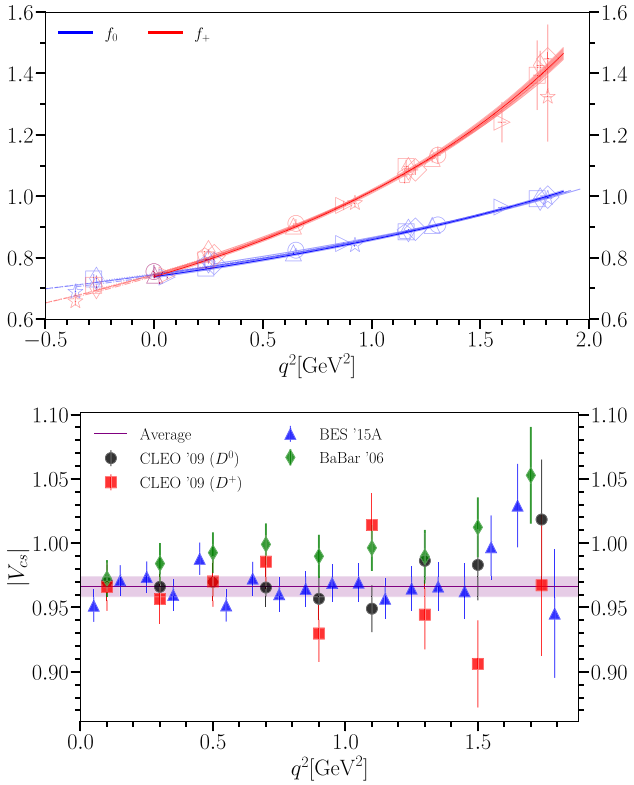


Fig. 70 (Upper) Data points show lattice QCD results at multiple values of q^2 and multiple lattice spacings. Blue and red curves show the final determination of f_0 and f_+ in the continuum limit at physical quark masses. (Lower) Bin-by-bin values of V_{cs} from combining these form factors with experimental data. The constancy of V_{cs} shows that the q^2 dependence predicted by QCD matches that of experiment [734]

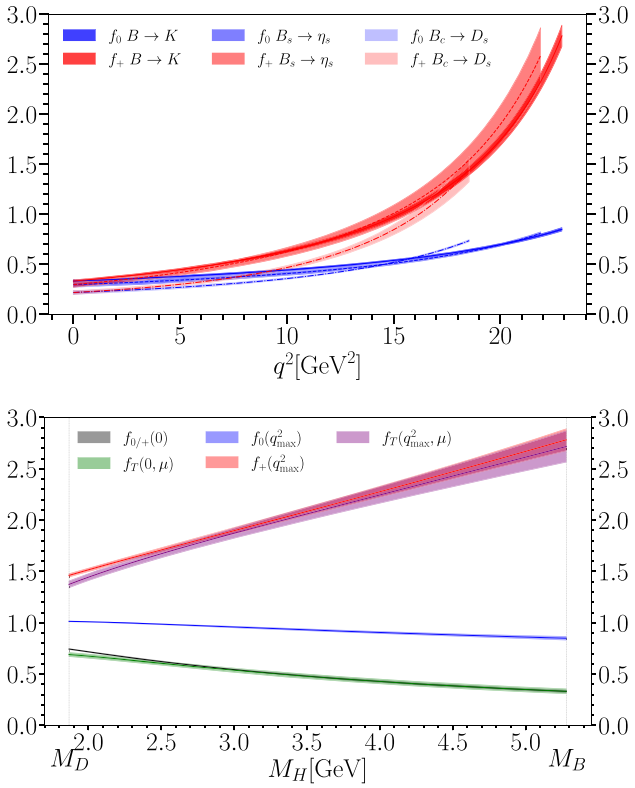


Fig. 71 (Upper) Comparison of $b \to s$ form factors for meson transitions with different spectator quarks. Increasing the spectator mass to that of c quarks reduces the form factors at low q^2 values [740]. (Lower) The dependence on heavy meson mass, M_H, of the form factors for $H \to K$ decay at $q^2_{\rm max}$ and $q^2 = 0$. Notice the slow downward drift at $q^2 = 0$ and for $f_0(q^2_{\rm max})$ as H varies from D to B, but much stronger variation upwards for f_+ and f_T (the tensor form factor) at $q^2_{\rm max}$ (remembering that $q^2_{\rm max}$ depends on M_H)

tal data from BaBar and Belle, leaving V_{ub} as a parameter. Such a fit allows experimental information on q^2-dependence to constrain the lattice results. The value for V_{ub} obtained, $3.74(17) \times 10^{-3}$ is 1.7σ lower than that obtained from inclusive $b \to u$ determinations that do not specify the final state hadron.

The transitions $b \to c$ have also shown a persistent tension between inclusive and exclusive results. Here the preferred exclusive method is to use $B \to D^*$ decay. Although a pseudoscalar to vector transition is more complicated, with 4 form factors, only the axial vector A_1 form factor contributes at $q^2_{\rm max}$. Lattice QCD therefore initially concentrated on this point [743,744]. Now it has become clear that the q^2-dependence of the differential rate must be understood from the lattice QCD side and form factors have been calculated by the Fermilab Lattice/MILC collaboration [745] that cover more of the q^2 range using their improved-Wilson action for both b and c. This does not resolve the inclusive/exclusive V_{cb} tension but points the way to improved future analyses.

Recent B form factors have been calculated using relativistic formalisms that can make use of nonperturbative current normalisation techniques discussed for Eqs. (4.189) and (4.190). They obtain results for multiple heavy quark masses and lattice spacings and fit to obtain results for B mesons in the continuum limit in a similar way to that for decay constants in Fig. 64. Calculations include HPQCD's form factors for $B_s \to D_s$ [746], $B_s \to D_s^*$ [747] and $B \to K$ [740] using HISQ quarks and JLQCD's form factors for $B \to \pi$ using domain-wall quarks [748]. This is likely to be the way forward for the future.

It is important to remember that QCD provides a smooth connection between different form factors as we change the mass for one or other of the participating quarks. In this way lattice QCD can provide 'a big picture' for form factors. Figure 71 shows this connection for different spectator (not part of the weak current) quarks for the $b \to s$ transition. It also shows results for $H \to K$ decay where H is a meson containing a heavy quark with mass varying from c to b [740].

Future calculations will improve B form factor uncertainties to the 1% level [749] for the increased datasets planned from LHC and Belle II. New developments include tech-

niques for inclusive B decays [750] and for handling final-state mesons that decay strongly (e.g. for $B \to K^* \ell \overline{\nu}$ analysis) [589]. An important focus will be improving lattice calculations needed to understand 'B anomalies' seen, for example in ratios of branching fractions to different flavors of leptons and differential rates for flavor-changing neutral current $b \to s$ transitions (e.g. $B \to K \ell^+ \ell^-$) that proceed through loops in the SM.

The lattice QCD calculation of form factors for weak decays of baryons is still in its infancy, because of the extra challenges provided by the poorer signal-noise. The nucleon axial coupling, g_A, has been a particular focus of attention and is discussed in Sect. 4.6. A notable success has been the use of lattice QCD form factors for $\Lambda_b \to \Lambda_c$ and $\Lambda_b \to \Lambda$ [751] to determine V_{ub}/V_{cb} by LHCb [752]. This is clearly a developing area for the future.

5 Approximate QCD

Conveners:
Stanley J. Brodsky and Franz Gross

The previous sections have introduced the QCD Lagrangian and shown how to make numerical predictions using LQCD. These predictions are subject to numerical uncertainties, but otherwise use the Lagrangian fully, without truncations, and can be systematically improved by going to smaller and smaller lattice spacings; in this sense they are sometimes referred to as "exact" calculations.

The disadvantage of LQCD predictions, however, is that they usually do not give much insight into the physical processes involved – they do not help us "understand" how certain physical properties emerge from QCD. Analytical solutions usually lead to this understanding, even though even after 50 years we still have no method to solve the equations of QCD analytically!

In situations where the momentum transfers are large, so that the coupling constant α_s is small, we can use perturbation theory to gain physical insight. But even then, higher order corrections will have loops with low momenta and large values of α_s so that these terms can only be estimated. For "cold" nuclear matter under normal conditions, the only analytical approaches today are to develop theoretical models which usually are tailored to treating some part of the problem carefully, and lumping other parts into "constants" which must be fit to data. Study of a variety of these models will be the subject of the next two sections of this volume.

At the heart of all modern models are quarks. Early models of mesons and baryons assigned (constituent) masses to the u and d quarks of 300-350 MeV, and $\sim$500 MeV to the strange-quark mass, in sharp contrast to the (current) quark masses that enter the QCD Lagrangian (see Sect. 3.1). Nevertheless, quark models met with considerable success and are still used as benchmarks when data on spectroscopy are interpreted. These models are reviewed in Sect. 5.1 (see also Sect. 8 for mesons and Sect. 9 for baryons). The section then moves on to a discussion of the Bethe Salpeter (BS) and Dyson Schwinger (DS) equations (Sect. 5.2), where quark–gluon interactions are treated microscopically, much as pion–nucleon interactions were described in an earlier era. Here the multiple interactions make it impossible to treat them all systematically, and the equations must be truncated, introducing approximations with an accuracy that is sometimes hard to estimate. Light front coordinates are the preferred way to describe multi-quark systems, and Sect. 5.3 describes methods for expanding multi-quark quark wave functions in a light front basis that avoids some of the issues with the microscopic description, but also requires truncations of the expansion to a finite number of basis states.

These methods handle the confinement of quarks in different ways with very different assumptions. In Sect. 5.4, recent developments based on superconformal quantum mechanics, light-front quantization, and its holographic embedding in a higher-dimension classical gravity theory, known as AdS/QCD, have led to new analytic insights into the non-perturbative structure and dynamics of hadrons in physical spacetime, such as color confinement and chiral symmetry breaking. This contribution is followed by a short discussion (Sect. 5.5) of the model dependence of predictions of the behavior of the strong fine structure constant, $\alpha_s(Q^2)$ at small Q, where it becomes large. This discussion complements and completes the discussions of $\alpha_s(Q^2)$ in Sect. 3. Next, the interesting features that can be drawn from the study of QCD with a large number of colors, and the solvable 't Hooft model, are reviewed in Sect. 5.6.

The next four contributions in this section discuss approximations that treat specific issues: the use of sum rules based on the operator product expansion (OPE) to explain properties of mesons and other physical quantities (Sect. 5.7); approximations that work for high energy reactions which can be factorized into reaction specific high energy parts that can be computed perturbatively and low energy, reaction independent parts expressed in terms of unknown functions that are extracted from many experiments (Sect. 5.8); the power counting rules that describe the behavior of exclusive processes at very high energy (Sect. 5.9); and finally the possibility of new hidden color states, i.e. virtual colored degrees of freedom occupied by groups of quarks at short distances (Sect. 5.10).

Finally, a theoretical discussion of the complexity of the QCD vacuum needed to understand confinement and chiral symmetry breaking is presented in Sect. 5.11. This discussion is complementary to the Lattice discussion of the same topic, Sect. 4.3.

This Sect. 5 covers a very wide range of topics, but as you will see from what follows, is only part of the theoretical tool

box developed to "solve" a theory based on a Lagrangian that can be written in one line!

5.1 Quark models

Eric Swanson

> "It is more important to have the right degrees of freedom moving at the wrong speed, than the wrong degrees of freedom moving at the right speed."
> – Gabriel Karl, as frequently quoted by Nathan Isgur.

5.1.1 Early quark models

The phrase "quark model" originally meant something like the "quark idea", referring to the introduction of quarks as the elements of the fundamental representation of $SU_F(3)$ by Gell-Mann and Zweig in 1964 [17,18,753]. Gell-Mann initially avoided attributing physical reality to the quark concept, and it was others, such as Dalitz [754], Becchi and Morpugo [755], Rubinstein, Scheck and Socolow [756], and Lipkin and Scheck [757] who developed the idea into a viable and predictive model in the sense we use now. That this was not a simple task is illustrated by a famous line from Kokkedee's review of the quark model, "The quark model should …not be taken for more than it is, namely, the tentative and simplistic expression of an as yet obscure dynamics underlying the hadronic world" [758].

Kokkedee's pessimism was not misplaced. The inability to observe free quarks was originally explained by assuming that they had very high masses. The existence of relatively light hadrons then implied that the interquark binding force was "ultra-strong", which in turn requires relativistic and nonperturbative techniques. These technical problems were further exacerbated by the "statistics problem", wherein bound states of fermions must be antisymmetric. Thus, for example, the Δ^{++} requires an antisymmetric spatial wavefunction, in contrast with expectations for a low lying state. No satisfactory solution to the problem was found, in spite of the great contortions theorists invented.

Nevertheless, a few determined individuals persisted with the notion that quarks are "real". Early computations drew from long tradition in nuclear physics [755,759,760] and tended to focus on electroweak transitions since the couplings are weak and the effects of unknown spatial wavefunctions can be ignored (in magnetic dipole transitions) or simply modelled (in electric dipole transitions). These computations typically assumed nonrelativistic dynamics, factorized spatial wavefunctions, and electroweak currents coupling directly to quarks. The state of the art was formalized in a classic paper from 1967 by van Royen and Weisskopf, which placed the topic on firm footing (even though the quark model problems remained unresolved) [761]. By 1969, Copely, Karl, and Obryk had brought the quark model to a high level of predictiveness, introducing explicit simple harmonic oscillator wavefunctions and a "constituent" quark mass of roughly one third the proton mass, in line with its modern value [762].

5.1.2 QCD-improved quark models

It is no surprise that the advent of QCD revolutionized the conceptualization and application of the quark model, releasing a flood of research. QCD, of course, is the theory of hadrons; thus the quark model was no longer the first and final word for hadronic properties, and it quickly evolved into its current role as a computationally feasible model for QCD in the strong coupling regime.

Already by 1975 (November 1974), Appelquist and Politzer famously applied QCD to the R ratio (proportional to the cross section for e^+e^- to hadrons) and noted that ladder exchanges of gluons should give rise to "orthocharmonium" (the J/ψ) and "paracharmonium" (the η_c) states [90]. This was the time of the "November revolution" described in Sect. 2.1 above. These notions were greatly expanded by De Rujula, Georgi, and Glashow, who argued that one gluon exchange should dominate the short-distance quark interaction and that it explained a wealth of experimental data, concluding that "The naive quark model, supplemented by color gauge theory, asymptotic freedom, and infrared slavery, is turning out to be not so naive, and more than just a model." [763]. In fact the results were successful enough that the authors initiated *and ended* the field in the same paper, declaring,

> Not until many of these predicted charmed states are discovered and measured can the subject of hadron spectroscopy join its distinguished colleagues, atomic and nuclear spectroscopy, as subjects certainly worthy of continued study, but understood (at some level) in principle.

Needless to say, such proclamations seem premature to modern eyes!

Amongst the first to join the fray were Isgur and Karl, who wrote a complete model Hamiltonian for baryons, assuming nonrelativistic dynamics, a quadratic confinement potential, and short distance spin-dependence as given by one gluon exchange [764]. (For a full discussion of baryon quark models, see Sect. 9.1) The resulting reasonably complete description of the low lying baryon spectrum and its properties caused a sensation, as it was realized that comprehensive and quantitative computations of hadronic properties were possible. However, there was a price to be paid: the good results were obtained only upon neglecting the spin–orbit interaction arising from one gluon exchange. It is, of course, difficult

to argue in favor of one aspect of perturbative QCD while neglecting another! By way of defense, Isgur and Karl noted that the confinement interaction should contribute Thomas precession spin–orbit interactions, even though it is spin-independent, and that the long range spin–orbit interaction tends to cancel that due to one gluon exchange.

The issue of the spin-dependence of the long range (confinement) interaction reappeared in a nearly contemporary and seemingly disconnected area. At issue was the Dirac structure of a (presumed) relativistic long range two-body interaction for quarks,

$$1/2 \iint J(x)K(x-y)J(y),$$

where the current is written as $J = \bar{\psi}\Gamma\psi$, ψ is a quark field, and Γ is a four-by-four Dirac matrix. In 1978, Schnitzer realized that the masses of several newly discovered charmonia and bottomonia permitted settling the issue in favor of a scalar ($\Gamma = 1$) confinement interaction [765,766].

Of course, assuming that the interaction between quarks is specified by a current–current operator yields more than spin-dependence – it also gives the amplitude for quark pair creation, and therefore opens the field of strong hadronic transitions to investigation. (Such investigations actually date to the beginnings of the quark model, starting with Micu's hypothesis that quark pairs are produced in a spin-triplet angular-momentum-one state [767,768].)

In 1978, Eichten *et al.* produced the most famous version of such a model, the "Cornell model" (first introduced in 1975), in an ambitious attempt to understand the properties of charmonia, including their coupling to the open charm continuum [769]. Pragmatism forced compromise: the Cornell group had to assume a color density current to obtain agreement with the – by now well-established – one-gluon-exchange short-range structure of the quark interaction, and in disagreement with the decay model of Micu (which is admittedly a guess) and Schnitzer's scalar confinement. Nevertheless, the model is well-constrained and does admirably well in predicting a wealth of charmonium properties.

By 1985 the field had progressed enough that comprehensive models capable of describing all mesons and baryons were being attempted. The most famous of these is that due to Godfrey and Isgur (mesons) and Capstick and Isgur (baryons) [770,771]. The model has much in common with earlier ones such as Ref. [772]. The model assumes relativistic quark kinematics, the full one-gluon-exchange short-range interaction, and a scalar confinement interaction (including its spin–orbit relativistic correction). All interactions were convoluted over a Gaussian to ameliorate the strength of the short range terms (which are not legal operators in quantum mechanics).

A model of the running strong coupling was used because there is strong evidence that weaker spin-dependent interactions are required for heavier quarks. The possibility of quark annihilation in isoscalar channels was allowed by including a phenomenological term. The model was "relativized" by including factors of $(m/E)^{\nu}$, where ν is a model parameter, in various matrix elements. Finally, additional factors of meson and quark mass were introduced to certain rates to bring their form into alignment with low energy theorems. The resulting masses, strong decays, and electroweak transitions have served as a benchmark in hadronic physics over the intervening 37 years.

5.1.3 Bag models

The advent of QCD raised the possibility of inventing field-theoretic models of hadrons. The opportunity was seized first by Ken Johnson, who drew an analogy to bubble nucleation in first-order phase transitions to imagine a hadron as perturbative fields confined to a vacuum bubble of size about 1 fm. The resulting model, developed with colleagues in 1974, became known as the "MIT bag model" [773]. The starting point is a postulated nontrivial QCD vacuum that exerts a pressure (described by the constant B) on a region of trivial space-time (called the "bag"). The model Hamiltonian is

$$L_{\text{bag}} = (L_{QCD} - B)\,\theta(bag) \tag{5.1}$$

where θ is zero outside the bag region. Because the action involves an integration over a finite region of space, the location of the bag surface is itself a dynamical field, related through the Euler–Lagrange equations to the quark and gluon fields by a complicated, nonlinear expression. As a result quantization is very difficult and semiclassical approximations are used to study the system. In particular, the "static bag approximation" is made, wherein quarks and gluons are presumed to be confined to a region of a given radius (it is possible to make more complicated models where small oscillations in the bag surface are permitted). The resulting equations of motion describe free fields subject to cavity boundary conditions, which can be obtained by summing cavity modes.

Almost simultaneously, similar ideas were being explored at Stanford, giving rise to the "SLAC bag model" [774]. In this case a scalar field played a role similar to the bag. Symmetry breaking in the scalar vacuum served to confine quarks to a small region where the scalar field exhibits soliton-like behavior. However, this implies that quarks are confined to a spherical shell, which contradicts experiment [775]. A subsequent model, called the "soliton bag model", is able to avoid this feature while interpolating the MIT and SLAC bag models [776]. Many variant bag models have been developed over the years that seek to address various shortcomings. For example, the MIT, SLAC, and soliton models all violate chiral symmetry. This can be overcome by explicitly intro-

ducing pion fields [777,778] or topological features [779]. Other models will be discussed below.

A number of advantages of bag models are apparent: hadrons are bound systems of relativistic quarks and gluons, obey asymptotic freedom automatically, are confined to regions of order 1 fm in size, and respect color gauge invariance. These benefits spurred a large theoretical effort in hadronic modelling that lasted through the 1980s, and continues at a reduced level to the present. Unfortunately, the complexity of the model introduces a number of conceptual and technical difficulties. The cavity approximation, for example, is not translationally invariant and no projection onto momentum eigenstates exits. This has the practical demerit of introducing undesired center-of-mass degrees of freedom to the problem. Quark and gluon propagators can be formed by summing over appropriate cavity modes, but in practice this is difficult, and evaluating Feynman diagrams is technically cumbersome [775]. For example, self-energy diagrams are difficult to evaluate and are often ignored. Similarly, the expectation value of the bag Hamiltonian has a sum over zero point energies that diverges. Renormalizing this quantity is subtle, and the zero point energy is often replaced with a simple model. Lastly, the rigid cavity gives rise to spurious states that must be identified.

Early MIT bag-model computations contained three parameters, the bag constant, the gauge coupling, and a zero-point energy parameter. Fits to the ρ, N, and Δ masses then fixed these constants. Unfortunately the resulting value for the strong coupling was $\alpha_S \approx 2.2$, which gives spin splittings that are too large in other hadrons. The resulting phenomenology is often of poor quality; for example, an early calculation of P-wave masses gives disappointing results [780]. Bag model phenomenology is clearly geared toward light hadrons. Heavy quark states, on the other hand, are surely described by nonrelativistic kinematics, a string-like confinement mechanism, and a value of the strong coupling that is set by $\alpha_S(m_Q)$. These features can be incorporated by allowing the bag to distort into a tube shape (in practice the distortion is small) and refitting the model parameters [781]. The resulting model does a reasonable job with the low lying charmonium and bottomonium vectors, predicts a $J/\psi - \eta_c$ splitting of 180 MeV (the measured value is 113 MeV), and $J^{PC} = 1^{-\pm}$ charmonium (bottomonium) hybrids at mass of approximately 4.0 (10.49) GeV.

One of the great advantages of bag models is that they made it clear that states incorporating gluonic degrees of freedom (glueballs and hybrids) should be considered seriously. Early contributions to the theory of glueballs include Jaffe and Johnson [782], who examined many novel states in the model, and Barnes, Close, and Monaghan, who computed spin-dependent mass shifts in the glueball spectrum [783]. These shifts are very large when common model parameters are used, giving glueball masses of $m(0^{++}) = 100$ MeV, $m(0^{-+}) = 400$ MeV, and $m(2^{++}) = 1300$ MeV, all of which are in strong disagreement with modern lattice values [780].

Studies of hybrid ($q\bar{q}g$) mesons originated in the MIT bag model [784] only a few years after the advent of both QCD and bag models, thereby raising interest in these novel states and highlighting the unusual ("exotic") quantum numbers that are available to these systems. Early computations in the MIT bag model worked to first order and focussed on light hybrid mesons [785,786], obtaining, for example, a $J^{PC} = 1^{-+}$ light hybrid mass of 1300 MeV [787].

Problems with complexity and fidelity have caused bag models to largely fall out of favor as descriptions of hadrons. They do, however, continue to find applications in models of strongly interacting matter or other complex hadronic systems.

Table 3 Diquark quantum numbers

J^P	Color	Flavor
0^+	$\bar{\mathbf{3}}$	$\bar{\mathbf{3}}$
1^+	$\bar{\mathbf{3}}$	$\mathbf{6}$
0^-	$\bar{\mathbf{3}}$	$\mathbf{6}$
1^-	$\bar{\mathbf{3}}$	$\bar{\mathbf{3}}$

5.1.4 Diquark models

Two quarks in a baryon experience a (perturbative) mutual attraction that is one half of the strength of that between a quark and an antiquark in a meson. If the third quark is isolated in some sense, it is fruitful to consider this quark–quark state as a compact object, called a *diquark*. More generally, a diquark is any system of two quarks considered collectively. The idea is already mentioned by Gell-Mann in 1964 [17] and was introduced in Refs. [789] and [790] as a way to reduce three-body dynamics to the computationally simpler two-body dynamics.

In general a pair of quarks, denoted $[qq]$ in the following, can form $\bar{\mathbf{3}}$ and $\mathbf{6}$ color states, with the former being antisymmetric and the latter being symmetric under quark interchange. Because a pair of quarks in the $\mathbf{6}$ representation has a (perturbative) repulsive interaction ($+\alpha_S/(6r)$), diquarks are only considered in the $\bar{\mathbf{3}}$ representation. In this case, possible quantum numbers for $[qq']$ are as listed in Table 3. The first two of these entries are often called "good" and "bad" diquarks respectively [788].

An early application of diquarks was to the description of light baryons [791]. The primary effect is a reduction in the number of degrees of freedom compared to a "symmetric" quark model, with commensurate decrease in the complexity of the excitation spectrum. For example, a symmetric quark model will feature orbital excitations in two relative coordi-

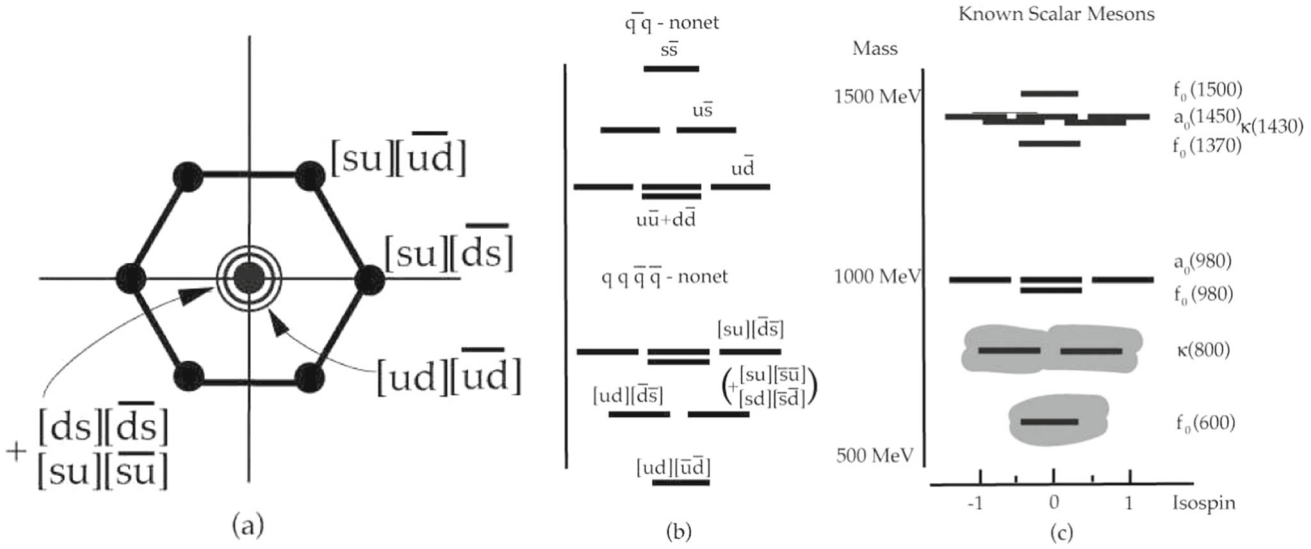


Fig. 72 **a** Quark content of a diquark–antidiquark nonet. **b** Mass levels of ideally mixed $q\bar{q}$ nonet and diquark–antidiquark nonet. **c** Light scalar mesons. The shaded region indicates large widths. Figure from Ref. [788]

nates (often taken to be the Jacobi coordinates $\vec{\rho}$ and $\vec{\lambda}$) while a quark–diquark bound state can only have orbital excitations in a single relative coordinate. Recent progress in the field is telling us that this simple diagnostic is incompatible with our knowledge of the excited baryon spectrum: one degree of freedom is not sufficient to explain the richness of of the spectrum of light-quark baryons (see Sect. 9.2).

Light baryons do experience flavor-dependent correlations, which might be attributed to diquarks. For example, a neutron will have a negative charge radius because the d quarks are in a spin-one state and are repelled via the hyperfine interaction, leaving the positive u quark in the center (on average). Similarly, diquark overlaps (denoted by I) affect static observables like the ratio of magnetic moments and the ratio of axial and vector couplings:

$$\frac{\mu_p}{\mu_n} = -\frac{4+5I}{2+4I}, \qquad \left|\frac{G_A}{G_V}\right| = \frac{2+3I}{2+I}. \tag{5.2}$$

Unfortunately, the additional freedom (represented by I) does not permit a simultaneous fit to the experimental values of -1.46 and 1.25, respectively [791].

At a more formal level, the similarity of light quarks makes it difficult to separate one quark from the other two. In the extreme case of identical quarks, antisymmetrization of the state implies that such a separation is not feasible. This was noted long ago by Lichtenberg [792], who suggested including exchange forces to accommodate transitions of the form $q[qq] \to [qq]q$. Of course this implies that the diquark can no longer be thought of a simple quasiparticle, but is rather something with internal structure that can be modified and excited.

Perhaps the most famous application of light diquarks is a model of the scalar mesons. In the 1970s Jaffe noted that a good diquark and a good antidiquark naturally make a scalar nonet of states, as shown in Fig. 72a. This nonet forms a spectrum as shown in panel (b) with counting that contrasts strongly with the "normal" $q\bar{q}$ scheme, shown at the top of panel (b). Remarkably this scheme agrees with the observed spectrum, as shown in panel (c) [788]. This ostensibly simple observation has a long and somewhat controversial history, as general acceptance of the existence of the light scalar mesons $f_0(600)$ and κ has waxed and waned over the years.

More recently, the diquark simplification has been applied to Bethe–Salpeter approaches to the baryon spectrum with some success [793]. The concept has also found support in lattice computations that see evidence for the good light diquark [794].

The discovery of the $X(3872)$ prompted a surge in modelling of exotic hadronics, and led to renewed interest in diquarks. A prominent model, due to Maiani and collaborators [795], advocated that the $X(3872)$ is a $J^P = 1^+$ double diquark state with composition $[cq]_1[\bar{c}\bar{q}]_0 + [cq]_0[\bar{c}\bar{q}]_1$. This assignment sets the mass of the open charm diquark, $m_{[cq]} = 1933$ MeV, and implies a rich spectrum of exotic states. A novel prediction of the model is that *two* neutral vector exotic states should exist with a mass difference of approximately 8 MeV. Focussing on flavor quantum numbers, these are mixtures of $[cu][\bar{c}\bar{u}]$ and $[cd][\bar{c}\bar{d}]$. Amongst others, scalar states are predicted at 3723 MeV and 3832 MeV. In spite of the explanatory power of the model, and reasonable agreement with properties of the $X(3872)$, none of these additional states have been observed. (For a complete discussion of this issue, see Sect. 8.5.2.)

Notwithstanding the checkered history of the diquark model, it must become relevant as quark masses become

much greater than the QCD scale, Λ. In this case the quarks will sit deeply in a Coulombic well, are compact, and are described well by perturbative gluon exchange. It is widely believed that bottom quarks are sufficiently heavy for these phenomena to occur. If a pair of bottom quarks forms a hadron with light degrees of freedom (such as light quarks or gluons), then it is reasonable to model the bottom quarks as a $[bb]$ diquark, and this expectation becomes rigorous as the heavy quark mass becomes very large.

A consequence of this concerns spin splittings in heavy-light mesons and baryons, as first observed by Savage and Wise [796]. In the following Q represents a quark with mass larger than the QCD scale, Λ (thus b, c), while q represents a quark with mass much less than Λ. The latter then refers to u and d quarks. The strange quark is ambiguous in this classification, and is sometimes grouped with the light quarks, and sometimes with heavy quarks. In practice heavy quark symmetries only become clear at the bottom mass and higher, while light quark (chiral) symmetry applies well to u and d quarks, and fairly well to s quarks.

Heavy quark spin degrees of freedom interact via their color dipole moments, which permits relating spin splittings in QQq baryons and $\bar{Q}'q$ states, with a relationship given by

$$m_{\Sigma^*(Q)} - m_{\Sigma(Q)} = \frac{3}{2}\frac{m_{Q'}}{m_Q}\left(\frac{\alpha_S(m_Q)}{\alpha_S(m_{Q'})}\right)^{9/33-2n_f} \times \left(m_{V(Q')} - m_{P(Q')}\right). \tag{5.3}$$

Here V and P refer to vector and pseudoscalar mesons, while Σ^* and Σ refer to ground state and spin-excited QQq baryons.

A slightly more model-dependent application establishes that the heavy $J^P = 1^+$ $ud\bar{b}\bar{b}$ tetraquark state must be strongly bound. The argument relies on the spin splittings, $\Sigma_b - \Lambda_b$ and $\Xi_b' - \Xi_b$, which indicate that the $(\bar{\mathbf{3}}_F, 0, \bar{\mathbf{3}}_c)$ light diquark lies approximately 100 MeV below the spin-averaged light diquark mass. This diquark interacts with a b meson with quantum numbers $(\mathbf{1}_F, \frac{1}{2}, \mathbf{3}_c)$ to produce the relevant baryons. As argued above, and verified by small $B^* - B$ and $\Sigma_b^* - \Sigma_b$ mass splittings, the heavy (di)quark spin must decouple from the light degrees of freedom. Thus a light diquark has a similar mass when coupled to a heavy $[\bar{b}\bar{b}]$ diquark. Since the heavy diquark has quantum numbers $(\mathbf{3}_F, 0, \mathbf{3}_c)$, the $[ud][\bar{b}\bar{b}]$ tetraquark has quantum numbers $I = 0, 1/2$ and $J^P = 1^+$. Recent lattice field theory computations have proven these expectations correct [797].

Diquarks continue to find application in a variety of areas: reducing the daunting complexity that arises in Bethe–Salpeter equations for many-quark systems, Sect. 5.2, the operator-product expansion, Sect. 5.7, instanton vacuum modelling, Sect. 5.11, heavy quark effective field theory, Sect. 6.1, models of quark matter, Sect. 7.2, tetraquark models, Sect. 8.5, baryons, Sects. 9.1, 9.2, 9.4, and models of hadronization, Sect. 11.4.

5.1.5 Current developments

The advent of new theoretical tools and the discovery of many novel hadrons have fueled the continued development of the constituent quark model. Amongst the latter are the $X(3872)$ that strongly hints at $qq\bar{q}\bar{q}$ structure and the importance of coupling mesons to the meson–meson continuum. Strong evidence for states consisting of $qqqq\bar{q}$, called "pentaquarks", also exists. At the same time, the maturation of lattice field theory has permitted the theoretical exploration of many nonperturbative hadronic properties and novel states involving glue, such as glueballs and hybrids. Such studies also inform the development of refined quark models that are capable of describing an ever greater range of phenomena. The development of effective field theory and its application to hadronic physics has also greatly expanded and strengthened the base upon which quark models are developed. Finally, field-theoretic nonperturbative methods, such as those based on the Schwinger–Dyson and Bethe–Salpeter methods, have served to expand the understanding and purview of quark models.

These new tools have helped to clarify several longstanding issues in the field. For example, it is well-known that the pion is anomalously light because it is the pseudo-Goldstone boson of QCD, reflecting the (broken) near chiral symmetry of the theory. Alternatively, the pion is light in quark models because the hyperfine interaction drives its mass well below that of the rho meson. The size of this mass splitting is *infinite* according to the one gluon exchange interaction (because it is proportional to $\delta(r)$)! In practice the hyperfine operator is smeared, which introduces a smearing parameter that can be fit to obtain the pion mass. This is hardly a satisfactory situation! In spite of this, Isgur has argued that the smooth evolution of hyperfine splitting from bottomonium to light quarks (Fig. 73) is a sign that the formalism is correct [798]. How these views can be made consistent is demonstrated in a specific model in Ref. [799], wherein it is shown how chiral symmetry breaking induced by a nontrivial vacuum and an effective hyperfine interaction mesh in a smooth fashion. Further insight is gained from the Schwinger–Dyson formalism, which convincingly demonstrates that chiral symmetry breaking gives rise to both a light pion and a dynamical quark mass that can be interpreted as the constituent quark [800].

Recent results from the lattice and other theoretical analyses indicate that long-held notions are likely incorrect. For example, scalar confinement cannot be correct–it has been known since the 1980s that a confining scalar $q\bar{q}$ interaction implies an *anti-confining* qqq interaction because of the lack of an antiquark line. (This disaster was avoided in, for example, the Godfrey–Isgur and Capstick–Isgur mod-

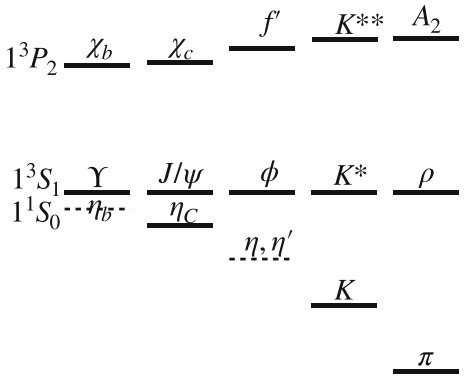


Fig. 73 "A graphic illustration of the universality of meson dynamics". Figure taken from the original [798]

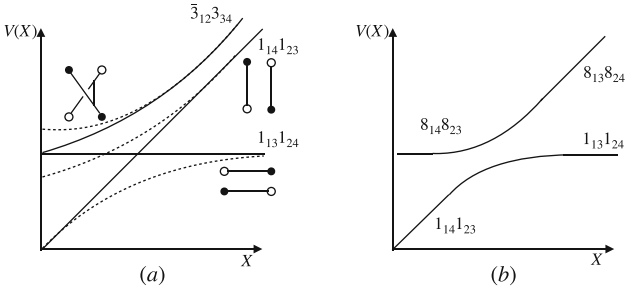


Fig. 74 "The adiabatic potentials of the flux tube model (**a**) and of the $\vec{F}_i \cdot \vec{F}_j$ potential model (**b**) for two $qq\bar{q}\bar{q}$ geometries." Figure taken from the original [798]

els by simply applying an extra sign.) The problem appears again in attempts at inducing chiral symmetry breaking in model field theories, where it is learned that scalar confinement interactions do not lead to a stable BCS-like vacuum [801]. In fact, it is not clear at all that the long range quark interaction need be described by a two-body interaction of the sort given above; QCD is much more complicated than this simple model [802].

Recent computations in lattice field theory have essentially settled the matter. This work relies on the model-independent expansion of the quark interaction in terms of nonperturbative matrix elements of gluonic operators [802,803], which are evaluated numerically. The results disagree strongly with an assumed scalar long range interaction. They do agree in large part with a Dirac vector interaction, with the exceptions that the hyperfine interaction resembles a smeared delta function and the spin orbit interactions have effective string tensions that are reduced by a factor of approximately 77% [804]. The picture emerging is that perturbative gluon exchange dominates the interaction at very short distances (less than 0.1 fm) and an effective vector-like interaction dominates at intermediate ranges. At long range (greater than 1 fm), one must saturate gluon exchange with a sum over hybrid intermediate states. This brings in the nonperturbative matrix elements of chromoelectric and chromomagnetic fields (mentioned above) that give rise to the nontrivial structure seen in lattice field theory. It is somewhat ironic that early enthusiasm for perturbative gluon exchange has evolved in this fashion!

Other quark model lore from the 1980s has been swept away in a similar fashion. For example, the Godfrey–Isgur computation of meson decay to $\gamma\gamma$ employed a perturbative amplitude with a "mock meson" correction factor involving the meson mass. More sophisticated computations, where the amplitude is computed with relativistic quark currents and a sum over intermediate states is made, reveal good agreement with data and no need for artificial factors [805].

5.1.6 Open problems

One of the major goals in modern quark modelling is incorporating the effects of nonperturbative gluonic degrees of freedom, which, of course, permits modelling glueballs and hybrid hadrons. Outright guesses from the past have been superseded by a body of lattice explorations of gluonic properties. Among these are the spectrum of adiabatic gluonic excitations [414,806], the gluelump (bound states of gluons and a static adjoint color source) spectrum [410,807–809], and properties of charmonium hybrids [810–812]. Of particular interest is the confirmation that the heavy quark multiplet structure anticipated in Ref. [414] is reflected in the charmonium spectrum [810]. It is interesting, and very suggestive, that this multiplet structure can be reproduced by degrees of freedom consisting of a quark, an antiquark, and an effective axial gluon with quantum numbers $J^{PC} = 1^{+-}$ [527], pointing the way to possible future models.

The advent of compelling experimental evidence for tetra- and pentaquark states has heightened interest in modelling multiquark hadrons. This is an old field, which in the past suffered from sufficiently poor computations that Isgur dubbed it the "multiquark fiasco" [798]. Many technical problems were present in these calculations, but the chief physics problem is the nature of the quark interaction when more than three quarks are present. The issue, for example, is that a $qq\bar{q}\bar{q}$ can separate into two meson–meson channels and that the gluonic degrees of freedom should experience adiabatic surface crossing when transitioning between these configurations (see Fig. 74). Thus new gluonic interactions are necessarily introduced in multiquark states. Of course, one could always model these as a sum of two-body interactions with a perturbative color structure, but this seems unlikely to be viable. A widely accepted model of multiquark gluodynamics does not exist yet, and is urgently needed.

Multiquark states necessarily couple to systems of mesons and baryons, which makes it incumbent on modellers to understand the effects of coupled channels on hadronic properties. This requires knowing the effective quark pair opera-

tor. A common model, dating to 1969, has already been mentioned [767], but this can surely be improved. As a result, existing models of Fock sector mixing remain speculative. The problem is not amenable to effective field theory, so progress will likely rely on input from lattice field theory. Progress is urgent since channel coupling effects are expected to be important in many sectors of the spectrum, including the perpetually enigmatic light scalars mesons, and all states near thresholds, such as the $X(3872)$, the P_c pentaquarks, and the Z_c and Z_b states.

It is perhaps a surprise that a model dating back nearly 60 years remains an active field of research. Such are the mysteries of QCD. On thing is certain: the quark model remains the de facto standard by which hadrons are interpreted.

5.2 DS/BS equations

Franz Gross and Pieter Maris

5.2.1 Introduction

In this section we look at two closely related approaches to treating the strong interactions that existed before 1972, and remained very useful, even after the onset of QCD. One of these originated with papers by Dyson (1949) [813] and Schwinger (1951) [814,815], referred to as the Dyson–Schwinger equations (DSEs), and the second is the well known Bethe–Salpeter equation (BSE) [816], introduced in 1951.[14]

In general, the DSEs form an infinite set of coupled integral equations for the Green's functions G_n of a quantum field theory.[15] These equations are exact, but in practical calculations this set has to be truncated. The equations can be derived formally from the matrix elements of the Lagrangian density (as was done in the original papers), or in the path-integral formalism using functional derivatives [823], but Feynman diagrams can be used to provide a simple, pictorial way to understand them. Using QED as an example,[16] Fig. 75 shows the exact DSEs needed to describe the self-energy of each fermion, and the dressed $\overline{\psi}_i \gamma^\mu \psi_i A_\mu$ vertex Γ^μ_i.

The fermion–antifermion scattering amplitude G_4 of the two different fermions can be written as a series of interactions shown in the upper line of Fig. 76. Here the kernel K is the sum of *irreducible* contributions to the off-shell scattering (i.e.diagrams that cannot be reduced by drawing an internal line through the diagram that intersects only the two fermions). The infinite series of *iterations* of the irreducible diagrams (each referred to as *reducible* because it can be cut into two pieces by an internal line which intersects only the two particles), connected by dressed propagators, is then summed by the equation shown on the left-hand side (LHS) of the middle line. This is the DSE[17] for the scattering amplitude G_4. If a bound state exists, it shows up as pole in this scattering amplitude, as illustrated on the right-hand side (RHS) of the second line in Fig. 76. The BSE for the Bound State Amplitude (BSA) or vertex function, Γ (shown in the bottom line), has the same kernel as G_4. Figure 77 shows contributions to the QED kernel up to order g^6. There is no known way to sum these contributions in closed form.

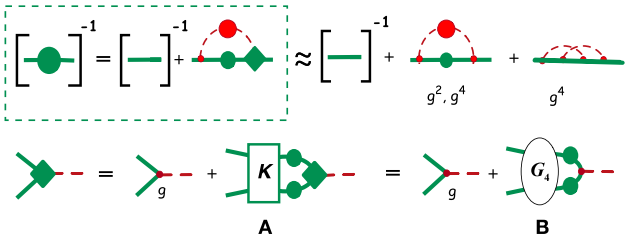


Fig. 75 Top row: The exact DSE for the inverse dressed fermion propagator (in the dotted box), and its approximation to 4th order in QED. Bottom row: two versions of the exact DSE for the dressed QED vertex Γ_μ (green diamond): Diagram (A) in terms of the $q\bar{q}$ irreducible kernel K, and (B) in terms of the full scattering amplitude G_4. The thick green (dashed red) lines are the fermion (photon), solid green (red) circles are the fermion (photon) self energies so that a fully dressed propagator is a green (red) line with a green (red) circle; and small red dots label the point coupling γ_μ and have no structure (renormalization constants are ignored here)

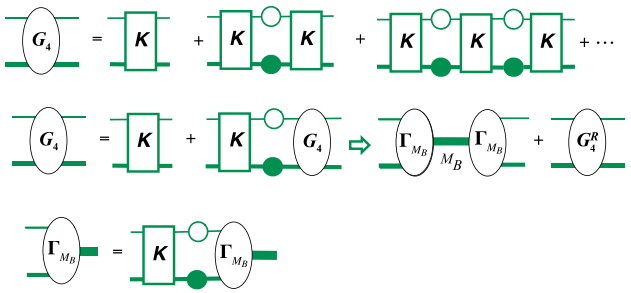


Fig. 76 Diagrammatic representation of the BSE propagator for two unequal mass particles $m_1 > m_2$. The first line represents the iteration of an *irreducible* kernel $\mathcal{K}$, which is summed by the BSE (first part of the second line). If the propagator has a pole, then the BSE vertex function satisfies the homogeneous BSE shown in the last line

The bound state BSE

As an example, the BSE for a $q\bar{Q}$ bound state in QED is

[14] Although the BSE can be used to describe scattering, this seminal paper was entitled *A relativistic equation for bound state problems*, particularly serendipitous for applications to QCD, where all physical states are bound states of quarks, antiquarks and gluons.

[15] For recent reviews of the DSEs in the context of QCD and hadron physics, see Refs. [800,817–822]

[16] When applied to QCD, with the photon replaced by a gluon, additional terms, such as the 3-gluon vertex, must be added.

[17] For two particle scattering, DSE and BSE are used interchangeably.

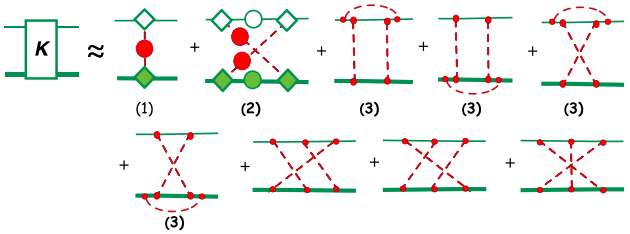


Fig. 77 Diagrammatic representation of the BSE kernel up to 6th order in g. Diagram (2) is the dressed xbox diagram and diagrams (3) are irreducible photon dressings of the box and xbox

$$\Gamma_{M_B}(p;\widehat{P}) = \int \frac{d^4k}{(2\pi)^4} K_{ij}(p,k;\widehat{P})\mathcal{O}_i\chi_{M_B}(k;\widehat{P})\mathcal{O}_j$$
$$\to 4\pi\alpha\int \frac{d^4k}{(2\pi)^4} D_{\nu\mu}(p,k;\widehat{P})\gamma^\nu\chi_{M_B}(k;\widehat{P})\gamma^\mu, \quad (5.4)$$

where $\chi_{M_B}(k;\widehat{P})$ is the BS wave function

$$\chi_{M_B}(k;\widehat{P}) = S_2(k_2)\Gamma_{M_B}(k;\widehat{P})S_1(k_1), \quad (5.5)$$

with $S_i(k_i)$ the dressed propagator for particle i and $\widehat{P}^2 = M_B^2$. The first line is exact, with the kernel written in the general form $K = K_{ij}\mathcal{O}_i \otimes \mathcal{O}_j$[18]; the second line is the *ladder truncation* with the kernel describing one photon exchange only, so $K \to 4\pi\alpha D_{\nu\mu}\gamma^\nu \otimes \gamma^\nu$. Dirac indices have been suppressed, and the four-momentum of the incoming Q is $p_1 = p - (1-\eta)\widehat{P}$ and of the outgoing q is $p_2 = p + \eta\widehat{P}$, reflecting the fact that the total momentum $\widehat{P}$ is conserved in relativistic equations. The physical observables do not depend on the choice of η, and the natural choice for mesons with equal-mass constituents (like a pion) is $\eta = \frac{1}{2}$. The canonical normalization condition for the BSE bound state vertex function can be derived directly from the inhomogeneous BSE (see e.g. Refs. [823,824]).

Very soon after the BSE was introduced, Wick [825] showed that the equation could be transformed from Minkowski space to Euclidean space by rotating the time component to the imaginary axis $\{t,\mathbf{r}\} \to \{i\tau,\mathbf{r}\}$ (now referred to as a Wick rotation). Building on Wick's results, Cutkosky [826] found all the exact solutions to the bound state BSE in *ladder* truncation for a scalar theory of the $\chi^2\phi$ type where the exchange particle ϕ is massless. The solutions are symmetric under the $O(4)$ symmetry group, and hence have the same degeneracy as the nonrelativistic hydrogen atom. Some of the solutions correspond to excitations in the time direction that have no nonrelativistic analogues. Furthermore, these solutions have a negative norm (at least in QED and QCD), and are therefore unphysical. As far as we know, no other analytic solutions have been found, but in the last 25 years accurate solutions of the BSE in ladder truncation have been obtained numerically for both scalar and fermionic systems, discussed below.

Several facts about the BSE are sometimes overlooked:

- The equation shown in Fig. 76 is exact, but only if the *exact* kernel *and* self energies are known.
- All applications of the BSE are therefore approximations using an approximate kernel and self-energies.
- In addition to Eq. (5.4), which is a homogeneous equation, there is also a canonical normalization condition for the BSA; one should not normalize the BSA to just any seemingly convenient observable.

Methods to solve the BSE in Minkowski metric

Due to the presence of poles in both the constituent propagators and in the kernel (coming from the exchanged bosons), it is highly nontrivial to solve the BSE numerically in Minkowski metric, even in ladder truncation. There are two methods to investigate the BSE directly in Minkowski metric, both dating back to the late 60s: the Covariant Spectator Theory (CST) which we discuss in Sect. 5.2.2, and use of the Nakanishi representation (also known as Perturbation Theory Integral Representation) [827].

The Nakanishi representation for the BSA is a spectral representation in which the singularities that arise from the poles in the propagators are isolated, allowing the BSE to be reduced to an integral equation for the (non singular) spectral function. This has been done initially for scalar field theory [828], and subsequently for fermion–antifermion bound states [829–831]. The obtained BSAs have been benchmarked against direct numerical solutions of the ladder BSE for Euclidean (spacelike) relative momenta.

Recently, the scalar BSE in ladder truncation has also been investigated in Minkowski metric by starting in the Euclidean formulation and rotating the p_4 axis to ip_0 (i.e. undoing the Wick rotation numerically) [832], and by using contour deformations in order to avoid singularities [833]. These methods give, within numerical precision, the same results for the BSA in the timelike region as the Nakanishi representation.

Connection to the light-front wavefunction

The use of the light-front (LF, first referred to as the infinite momentum frame) was introduced by Weinberg in 1966 [834], and the technique was developed very extensively in the 1980s by Lepage and Brodsky [226] and many others. It is now a standard method for describing the structure of hadrons and calculating a range of observables. Application of this technique will be extensively discussed in Sect. 5.3. Use of the LF is not manifestly rotationally invariant, but this can be handled by imposing the so-called angular conditions; see, for example, Ref. [835].

The LF wave function can be obtained from $\chi(p;P)$ by integrating over $p^- = p^0 - p^3$, leaving $p^0 + p^3 \equiv xP^+$

[18] Each of the operators $\mathcal{O}_i$ describes the structure of the dressed vertices, including possibilities like those illustrated in diagrams (2) and (3) of Fig. 77.

and $\mathbf{p}_\perp = \{p^x, p^y\}$ as independent variables. It turns out that the LF wave function, $\psi(x, p_\perp)$, is only nonzero for $0 < x < 1$, and vanishes outside this range, even though p^+ runs from minus infinity to plus infinity. This has been confirmed numerically for scalar theories in ladder truncation.

Instead of solving the BSE in Minkowski metric, and then projecting onto the light-front, one can also reconstruct the LF wave function (or e.g. parton distributions) from their moments, which can be evaluated directly from the BS wave function [836–838]. One caveat to keep in mind is that the BSE is typically solved in covariant gauges; the most commonly used gauge in the literature is the Landau gauge, though other gauges such as Feynman gauge are also used. On the other hand, the LF wave function is usually investigated in LF gauge.[19] This makes it nontrivial to compare LF wave functions obtained from the explicitly covariant BSE to LF wavefunctions obtained within a LF approach.

5.2.2 The covariant spectator theory (CST)

The CST, which is related, but not identical, to the BSE, can be obtained from the BSE if the internal loop energy is evaluated keeping only the pole contribution from the heaviest particle [839].[20] If $m_+ = m_1 > m_2 = m_-$, $\eta = \frac{1}{2}$, treating particle 1 as outgoing, and working in the rest frame where $P = \{W, \mathbf{0}\}$, then $p_1 = \frac{1}{2}P - p$, and the one-channel CST equation can be obtained from (5.4) using the prescription[21]

$$\Gamma(p; P) = -i \int \frac{d^4k}{(2\pi)^4} \frac{F(p,k;P)}{d_+(k) d_-(k)} \to \int \frac{d^3k}{(2\pi)^3 2E_k^+} \left[\frac{F(p,\widehat{k};P)}{\delta_m^2 + W(2E_k^+ - W)} \right], \quad (5.6)$$

where $d_\pm(k) = m_\pm^2 - (k \mp \frac{1}{2}P)^2 - i\epsilon$, F is any covariant function, $\widehat{k}_1 = \frac{1}{2}W - \widehat{k} = \{E_k^+, \mathbf{k}\}$, $(E_k^+)^2 = m_+^2 + \mathbf{k}^2$, so that $(\widehat{k}_+)^2 = m_+^2$, and $\delta_m^2 = m_-^2 - m_+^2$. The CST equation is covariant in three dimensional space, and, unlike the LF, is rotationally invariant. The major motivation for the use of CST equations is that they have a smooth nonrelativistic limit, and in a few cases their ladder approximation is more accurate than the ladder approximation to the BSE. Their major disadvantage is that their kernels can be singular, and the treatment of these introduces an additional level of phenomenology (see below).

In scalar field theories when $m_1 \to \infty$, it has been shown that the sum of all ladders and crossed ladders (the *generalized* ladder sum) is given by the solution of the CST equation with *only* the one-boson-exchange (OBE) kernel (see Refs. [824] and [839]).[22] This is referred to as the *cancellation theorem*.

While the complete cancellation holds only in an exceptional case, partial cancellations occur for other cases. Using the Feynman–Schwinger representation [840], it is possible to calculate the exact result for the generalized ladder sum without vertex or self-energy corrections. For scalar theories where $m_1 = m_2 \neq \infty$ and the exchanged mass $\mu = 0.15\, m$ [841], the BSE in ladder approximation gives only about one-quarter of the correct binding energy (at large coupling), while the one-channel CST equation, also in ladder approximation, gives a little more that half the correct result. The OBE approximation in the light-front approach gives the same result as the BSE in ladder approximation [842].[23] Another approach, the equal-time (ET) favored by Tjon [844] is slightly better than the CST, but only the CST (to our knowledge) uses the same two-body scattering amplitude in both the two-body and three-body systems. In a later paper [845], it was shown that the contributions of all self-energies and vertex corrections for scalar QED are very small, so that in this case the generalized ladders dominate (and are well approximated by the CST and ET). These remarkable results apply only to scalar theories, so the main justification for the use of the CST must rest on its simple nonrelativistic limit.

It turns out that the one-body CST prescription (5.6) must be generalized if it is to be used for all cases including $m_- = m_+$ and $W \to 0$. To treat these limits successfully, all four k_0 poles from the two fermion propagators must be included. There are two poles in the upper half k_0 plane ($r = -$) and two in the lower half ($r = +$), and if $s = \pm$ denotes the poles from particles $m_\pm$, then they can all be denoted by $k_{0r}^s = rE_k^s + \frac{1}{2}sW - ir\epsilon$. Since the contour can be closed in *either* half plane (but not both), we average over the two choices. This gives the new prescription

$$\Gamma(p; P) \to \frac{1}{2} \sum_{s,r} \int \frac{d^3k}{(2\pi)^3 2E_k^s} \left[\frac{F(p,\widehat{k}_r^s;P)}{s\delta_m^2 - W(2rsE_k^s + W)} \right], \quad (5.7)$$

where $\widehat{k}_r^s = \{k_{0r}^s, \mathbf{k}\}$ with $(E_k^\pm)^2 = m_\pm^2 + \mathbf{k}^2$.

The sum on the RHS of this equation has four terms, and substituting the four values $p \to \widehat{p}_r^s$ into the LHS gives four coupled equations.[24] As discussed above, only one channel

[19] For further discussion of the LF calculations and an explanation of the LF gauge see Sect. 5.3.

[20] This is sometimes referred to as "restricting the particle to its mass shell."

[21] With our choice of momenta, this is obtained by closing the k_0 contour in the upper half plane and keeping only the positive energy pole of particle 1, at $k_0 = \frac{1}{2}W - E_k^+$.

[22] In other words, when $m_1 \to \infty$, the CST in OBE approximation gives the same result as the BSE for a kernel containing *all* irreducible crossed ladders.

[23] This is not true for three-body systems, due to contributions with two (or more) exchange bosons in flight, which are included in the ladder BSE, but not in the OBE approximation on the light-front [843].

[24] This should be considered the correct form for the CST in all cases, but often some of the channels can be ignored.

is needed when $m_+ \to \infty$. When the particles are identical, symmetry under interchange requires that the equation transform into itself when $p_1 \leftrightarrow p_2$, or $k_1 \leftrightarrow k_2$, and looking at k^s_{0r} shows that this requires (if P is not small) at least the channels where $\{r, s\} = \{+, +\}$ and $\{-.-\}$, so that $rs = +$ in both cases. Looking at (5.7), it is clear that it is symmetric under this transformation, remembering that for identical particles, $\delta^2_m = 0$ and $E^+_k = E^-_k$. Finally, when W is small, there will be a singularity at $W = 0$ unless all four channels are kept.

Unfortunately, when a OBE kernel connects a channel with particle 1 on-shell to a channel with particle 2 on-shell, the kernel will develop singularities. These are discussed in detail in Ref. [846], but the preferred way to remove them was only developed recently.[25]

Nuclear physics applications of the CST

The two-channel CST has been used to give a high precision fit to the np scattering data below 350 MeV ($\chi^2 = 1.12$ using only 15 parameters [846]), to explain the deuteron form factors (giving a quadrupole moment within 1% of its experimental value [848]), and to study the three nucleon system. All of these studies were done with two models. The simplest and most successful one uses a covariant OBE kernel consisting of the exchange of 6 mesons: π, η, σ_0 and σ_1 (scalar mesons with isospin 0 and 1), and ρ and ω. An interesting feature of these OBE models is that they include an off-shell coupling for the σ mesons of the form

$$\Lambda_\sigma(p,k) = g_\sigma - \nu_\sigma\Big[1 - \frac{\not{p} + \not{k}}{2m}\Big], \tag{5.8}$$

where the term proportional to ν_σ will give zero when the nucleons are on shell (with $\not{p} \to m \leftarrow \not{k}$). As it turns out (see below), this off-shell coupling is very important to the success of the model.

In the early days before the advent of QCD and powerful computers, the study of three nucleon systems posed special problems. The Alt–Grassberger–Sandhas equations [849], developed in 1967, introduced a systematic procedure for finding the solutions of n-body problems from the solution of the $n-1$ body problem. Examples of early papers working directly with the the three nucleon equation are found in Ref. [850], which presents solutions with realistic potentials, and Ref. [851], which solves the 3-body BSE with separable kernels.

The three-body CST equation given in Ref. [853] was used to compute the triton binding energy [854], and the three-nucleon form factor [855,856]. During these studies a remarkable discovery was made: the best fits to the np data require $\nu_\sigma \neq 0$,[26] and the same value of ν_σ also gives the best fit to the triton binding energy! This shows that three body-forces are not needed to explain this observable. This discovery, first found in 1996 [854], is shown with the latest (and best) fits in Fig. 78. It is a robust result that has continued to hold as the fits were improved, and is still not understood.

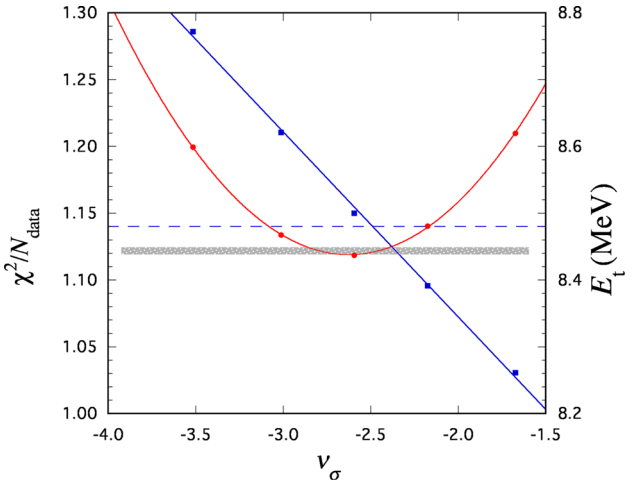


Fig. 78 The red line (left-hand scale) shows how χ^2 varies with ν_σ, with the best fit at $\nu_\sigma \simeq -2.6$. The blue line (right-hand scale) shows the (linear) variation of the triton binding energy with ν_σ, with the best fit also at $\nu_\sigma \simeq -2.6$. (From Ref. [852].)

Meson spectrum in the CST

In the CST treatment, mesons are $q\overline{q}$ bound states with one quark confined to its mass-shell. States like the ρ, where $m_\rho > 2m_q$, could have both the quark and anti-quark on-shell at the same time unless the interaction forbids it. Fortunately, the structure of the CST equations permits an attractive relativistic generalization of linear confinement. This definition of confinement was first introduced in 1991 [857], and in 1999 it was shown explicitly that the confining interaction does indeed guarantee that meson vertex functions are zero when both quark and antiquark are on shell [858]. Subsequently, an improved definition [859] was found. For any smooth S-state function $\phi(p)$ the action of the linear confinement kernel is

$$\Big\langle V_L\phi\Big\rangle(p_1) = -\int_k \frac{m}{E_k}\frac{8\pi\sigma[\phi(\widehat{k}_1) - \phi(\widehat{p}_R)]}{(p_1 - \widehat{k}_1)^4} \tag{5.9}$$

where the spin dependence, and the form factors that provide convergence at large momenta have been omitted, σ is the string tension, p_1 and $\widehat{k}_1$ are the momenta of particle 1, $\widehat{k}^2_1 = m^2_1$ is on-shell, and $\widehat{p}_R$ is chosen to reduce the singu-

[25] **FG**: These singularities troubled me for years. They are integrable, giving finite results, but only with the method described in Ref. [847] do I feel the problem is fully under theoretical control.

[26] **FG**: Originally we (Stadler and I) tried to fit the np data without the off-shell coupling, and got the very high χ^2 that an extrapolation of the curve shown in Fig. 78 suggests. Only after a frantic attempt to do better did we discover the importance of ν_σ. Later, we were surprised to realize that the same mechanism also gave the correct triton binding energy.

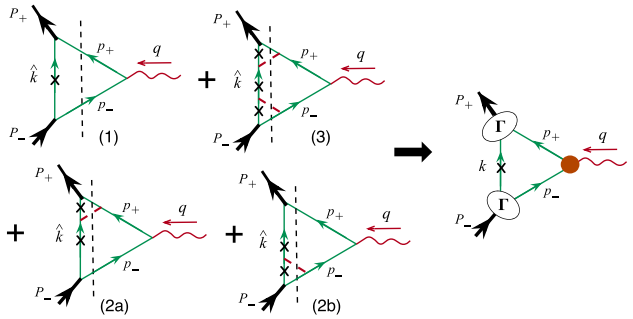


Fig. 79 The form factor of a bound state (meson or deuteron). Left panel: the four diagrams that give the lowest anomalous thresholds for the dispersion integrals, with dispersion cuts shown by the dashed lines. Note the multiple spectators on shell. The dashed red line represents the exchanged particles that bind the state. Right panel: diagram showing the triangle (or impulse) contribution expressed in terms of the BS vertex function Γ and dressed current (large red dot), needed to ensure gauge invariance (only in a CST calculation is the spectator on-shell)

larity at $(p_1 - \widehat{k}_1)^4 = 0$ to an integrable principal value (for details see Ref. [859]). Extension of this definition to states with non-zero angular momentum is discussed in Ref. [860]. Using this confining kernel, together with a phenomenological constant plus a one-gluon-exchange (OGE) contribution in a 1-channel CST equation, gives a good account of the spectrum of heavy-heavy and heavy-light mesons [861,862], as shown in Fig. 88.[27] The 4-channel CST equation also provides a good description of the pion consistent with the axial-vector Ward–Takahashi identity (AV-WTI) [859].

The origin of the CST – **FG**
My involvement with this subject began in 1960 when nucleons and pions were thought to be fundamental particles and S-matrix theory was believed to be the best way to tackle the strong interactions. For my Ph.D. it was suggested that I look at the deuteron electromagnetic form factor. The upshot of my study lead to the realization that the form factor was dominated by a large number of meson-exchange processes, the first four of which are shown in the left panel of Fig. 79,[28] and that these were best calculated by introducing a new equation that would sum these contributions – the one-channel CST equation.

If the internal propagators in the triangle diagram (right panel of Fig. 79) are dressed by form factors, then the off-shell nucleon current must also be dressed in order to ensure gauge invariance.[29]

5.2.3 DSE for the quark propagator

We now turn to a discussion of the DSE for the quark propagator. The exact equation for the quark propagator is shown in the upper left-hand box in Fig. 75. In Euclidean metric ($\{\gamma_\mu, \gamma_\nu\} = 2\delta_{\mu\nu}$, $\gamma_\mu^\dagger = \gamma_\mu$ and $a \cdot b = \sum_{i=1}^4 a_i b_i$) it is given by

$$S(p)^{-1} = i\ \not{p}\, Z_2 + m_q(\mu)\, Z_4 + Z_1\, g^2 \int \frac{d^4k}{(2\pi)^4}\, D_{\mu\nu}(q)\, \gamma_\mu \frac{\lambda^i}{2}\, S(k)\, \frac{\lambda^i}{2} \Gamma_\nu(k, p), \tag{5.10}$$

where $D_{\mu\nu}(q = k - p)$ is the renormalized dressed gluon propagator, and $\Gamma_\nu(k, p)$ is the renormalized dressed quark–gluon vertex. The solution of Eq. (5.10) can be written as

$$S(p) = \frac{1}{i\not{p} A(p^2) + B(p^2)} = \frac{Z(p^2)}{i\not{p} + M(p^2)}, \tag{5.11}$$

renormalized according to $S(p)^{-1} = i\not{p} + m_q(\mu)$ at a sufficiently large spacelike μ^2, with $m_q(\mu)$ the current quark mass at the scale μ. For divergent integrals a translationally-invariant regularization is necessary. Note that in the chiral limit, the current quark mass $m_q(\mu)$ is absent from Eq. (5.10) and there is no mass renormalization.

The most commonly used truncation is the rainbow truncation (analogous to the ladder truncation to the BSE discussed above), in which the dressed gluon propagator and the quark–gluon vertex are replaced by their bare counter-parts, with a model effective running coupling

$$Z_1 g^2 D_{\mu\nu}(q)\gamma_\mu \otimes \Gamma_\nu(k, p) \to 4\pi\, \alpha_s(q^2)\, D^{\text{free}}_{\mu\nu}(q)\gamma_\mu \otimes \gamma_\nu. \tag{5.12}$$

This truncation is the first term in a systematic expansion [866,867]; furthermore, the preferred gauge for the fermion

[27] Since both the light front and CST are relativistic wave functions depending on only three variables, it has long been thought that, perhaps, they can be transformed into one other. The basis for such a comparison might be based on a connection between one of the components of the CST internal momentum (take p_z for example) and the LF momentum fraction x, and a good candidate is $E_p + p_z = yD_0$, where D_0 is the energy of the bound state, and $y = x$. This transformation suggests an equivalence in some cases [863], but since $0 \le y \le \infty$, it is clear that $y \ne x$. Our conclusion is that CST and LF wave functions seem to describe the physics differently.

[28] A novel feature of the dispersion integrals describing these processes is the presence of anomalous thresholds starting at $s_i < 4m^2$. The imaginary part of the dispersion integral in the anomalous region (from s_i to $4m^2$) is given entirely by the contributions from these diagrams when the four-momentum of *all* the spectators are on shell. For diagram

Footnote 28 Continued

(1) this threshold is at

$$s_0 = \frac{M_B^2}{m^2}\left(4m^2 - M_B^2\right) \approxeq 16m\epsilon$$

where $\epsilon = 2m - M_B$ is the binding energy [864]. For diagrams (2a) and (2b), one additional spectator is on shell, and for diagram (3), two additional spectators are on shell. The thresholds for these diagrams are larger than s_0 but still much less than the normal threshold of $4m^2$.

[29] D. O. Riska and I constructed such a current [865], which is used in all CST calculations. This current plays a role analogous to the BC or CP currents discussed below.

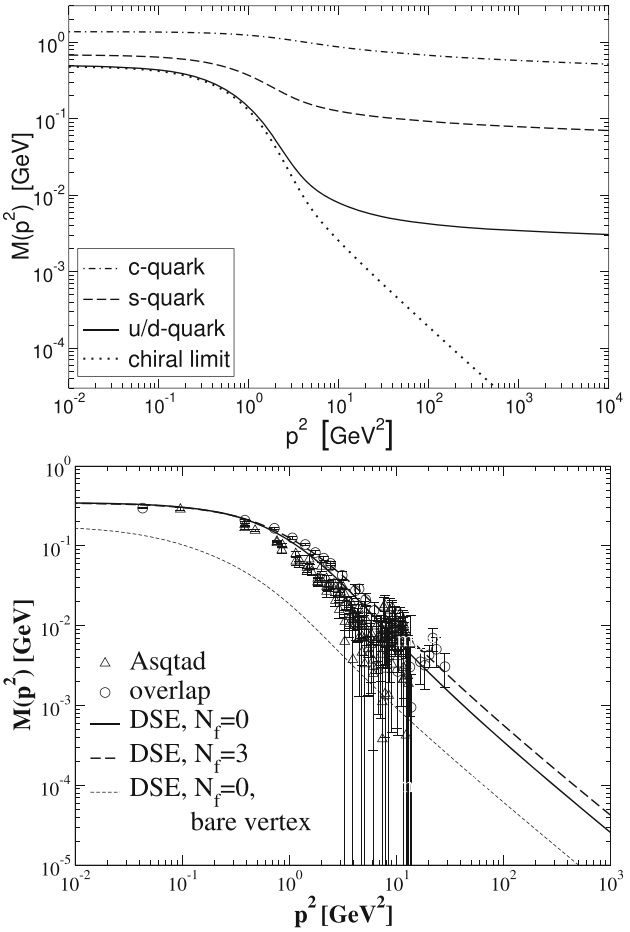


Fig. 80 Dynamical quark mass function $M(p^2)$ for spacelike momenta: using the rainbow truncation with the Maris–Tandy model [868] (top, adapted from [869]), and from quenched ($N_f = 0$) and unquenched ($N_f = 3$ chiral quarks) DSEs using the CP vertex [870], as well as results obtained with a bare quark–gluon vertex, compared to quenched lattice data in the overlap [871] and Asqtad [872] formulations (bottom, adapted from [873])

DSE is Landau gauge, which has the advantage that asymptotically, $Z(p^2) \to 1$.

By choosing a suitable model for the effective running coupling α_s that reduces asymptotically to leading-order perturbation theory, realistic quark mass functions, as shown in Fig. 80, are obtained. In particular, with a nonzero current quark mass, the dynamical mass function behaves at large p^2 like

$$M(p^2) \simeq \frac{\widehat{m}}{\left(\ln\left[p/\Lambda_{\rm QCD}\right]\right)^{\gamma_m}}, \quad \gamma_m = \frac{12}{11N_c - 2N_f}, \quad (5.13)$$

with the anomalous mass dimension, γ_m, in agreement with perturbation theory. In the chiral limit this model gives a nontrivial solution for the mass function that falls off like a power-law, modified by logarithmic corrections [874]

$$M_{\rm chiral}(p^2) \simeq \frac{2\pi^2\gamma_m}{3} \frac{-\langle\bar{q}q\rangle^0}{p^2\left(\ln\left[p/\Lambda_{\rm QCD}\right]\right)^{1-\gamma_m}}, \quad (5.14)$$

with $\langle\bar{q}q\rangle^0$ the quark condensate, in agreement with the Operator Product Expansion [875].

Of course, quantitative details of the quark propagator functions in the infrared region do depend on the truncation. The bottom panel of Fig. 80 shows the quark mass function $M(p^2)$ of the quark propagator in the chiral limit, obtained from the coupled quark, ghost, and gluon DSEs using the Curtis-Pennington (CP) vertex[30] suitably generalized for use in a non-Abelian QFT [870]. Qualitatively, these results agree with the quark mass functions shown in the top panel (both in the chiral limit, and with a nonzero current quark mass), though quantitatively they clearly do depend on both the details of the effective interaction and the vertex Ansatz.

Do real quark mass poles exist?

Knowledge of the behavior of the quark propagator in the complex momentum plane is necessary not only to solve the BSE at the bound state mass pole, but also because of possible connections to confinement, the CST, and the LF wave function. In QED, we know that real mass-poles must exist on the time-like axis, but early DSE studies of the fermion propagator in ladder truncation suggested the existence of complex "mass-like" singularities instead of real mass-poles at timelike momenta [877–879]. The absence of a mass-pole in the fermion propagator on the timelike axis would prevent the fermion from being on-shell, and could be an indication of confinement [880,881].[31] More recently however, is has been shown that, with proper regularization of poten-

[30] The CP vertex [876] is a nonperturbative Ansatz for the electron–photon vertex that satisfies the Ward–Takahashi Identity.

[31] **PM**: My interest in the fermion DSE started with my Masters research in the late 80s, with the question whether or not there was a dynamical mass generation in (2+1)-dimensional QED. In addition to dynamical chiral symmetry breaking, QED_3 also exhibits confinement; these two features make it an illustrative toy model for QCD. Consistent treatment of the photon propagator turns out to be crucial in QED_3: in the quenched approximation (no fermion loops, and hence no vacuum polarization), there is a logarithmically rising potential between a fermion and anti-fermion. This logarithmically confining potential persist in the presence of massive fermion loops in the vacuum polarization, but with massless fermions, this confining potential disappears. With the coupled DSEs for the fermion and photon propagator, it was found that there is a critical number of fermion flavors of about $N_f \sim 3$ to 4, below which there is both dynamical mass generation and a confining potential. Furthermore, it was found that in the presence of the logarithmically confining potential, the fermion propagator exhibits a pair of 'mass-like' singularities at complex conjugate momenta in the complex momentum plane, whereas in the absence of this logarithmically confining potential, the fermion propagator appears to have a real mass-pole at timelike momenta, as one would expect based on perturbation theory.

tially divergent integrals (e.g. using Pauli–Villars), that at least in weak-coupling quenched QED, the DSE for the electron propagator has the expected analytic structure, namely a mass-pole in the timelike region. This was obtained both in Feynman gauge and in Landau gauge; and using two independent numerical methods, explicitly rotating the spacelike region to the timelike region and using the Nakanishi formalism [882].

In QCD however, quarks (and gluons) are confined, and the quark propagator need not have a mass-pole at timelike momenta. A convenient way to study this is to use the Schwinger function, $\Delta(t)$, defined by

$$\begin{aligned}\Delta_{s,v}(t) &= \int d^3x \int \frac{d^4p}{(2\pi)^4} e^{i(tp_4+\vec{x}\cdot\vec{p})}\sigma_{s,v}(p^2) \\ &= \frac{1}{\pi}\int_0^\infty dp_4 \cos(t\ p_4)\sigma_{s,v}(p_4^2)\ \geq 0 \end{aligned} \tag{5.15}$$

where $\sigma_{s,v}(p^2)$ is the scalar or vector part of the dressed quark propagator,

$$S(p) = i\,\not{p}\,\sigma_v(p^2) + \sigma_s(p^2)\ . \tag{5.16}$$

For a propagator with a real mass-pole in the timelike region, this Schwinger function falls off like an exponential. In contrast, a propagator with a pair of complex-conjugate mass-like singularities, the Schwinger function is not positive-definite and exhibits an oscillatory behavior

$$\Delta(t) \sim e^{-a\,t}\ \cos(bt+\delta). \tag{5.17}$$

In Ref. [873] a striking qualitative difference between the use of a bare quark–gluon vertex and the BC [883] or CP vertex was found: with a bare vertex, the Schwinger function behaves like a pair of complex-conjugate mass-like poles for the quark propagator, whereas the results with the BC and the CP vertex behave like a real mass-pole in the timelike region. Qualitatively similar results were found employing different models for the effective running coupling, including (3+1) dimensional QED. The existence of a pair of complex-conjugate mass-like singularities in the DSE solutions of the dressed quark propagator in rainbow truncation was also confirmed by direct analytic continuation of the quark DSE into the complex-momentum plane; the obtained real and imaginary parts of these singularities agree with those extracted from the Schwinger function. Whether or not confinement is realized through the absence of mass-like singularities on the real timelike axis remains to be seen. Note that these results are not inconsistent with the CST, which assumes the existence of real quark mass poles.

5.2.4 Pions: Goldstone bosons of QCD

Pions, and to some extent also kaons, are the pseudo-Goldstone bosons of QCD: in the chiral limit, $m_q = 0$, chiral symmetry is broken dynamically, which implies the existence of massless Goldstone bosons. In the flavor SU(2) chiral limit, there are three Goldstone bosons (three pions); and in the flavor SU(3) chiral limit, there would be eight Goldstone bosons. In the real world, the up, down, and strange quarks are not massless, but have a small current quark masses; in addition, one of the eight 'would-be' Goldstone bosons mixes with the isoscalar pseudoscalar meson (which is massive due to the axial anomaly) to form the η and η'. This explains qualitatively why the three pions and four kaons are so much lighter than all other mesons, among other things. Therefore, in order to describe pions (and kaons), any truncation has to respect all constraints coming from chiral symmetry. Furthermore, it implies that the pion BSA is closely related to the (dynamically generated) scalar part of the quark self-energy, which can be made explicit by using the AV-WTI [884].

The axial-vector vertex Γ_5^μ satisfies a DSE as illustrated in the second row of Fig. 75, with an inhomogeneous term $\gamma^5\gamma^\mu$. But even without solving the DSE, one can relate this vertex directly to the dressed quark propagators via the AV-WTI

$$\begin{aligned}P_\mu\Gamma_5^\mu(p;P) &= S^{-1}(p_2)\gamma_5 + \gamma_5 S^{-1}(p_1) \\ &\quad - 2\,m_q(\mu)\,\Gamma_5(p;P),\end{aligned} \tag{5.18}$$

where $\Gamma_5(p;P)$ is the pseudoscalar vertex, which also satisfies a DSE as shown in Fig. 75, with inhomogeneous term γ^5. This can be compared to the more familiar vector WTI for the quark–photon vertex (which satisfies the same DSE with inhomogeneous term γ^μ),

$$P_\mu\Gamma^\mu(p;P) = S^{-1}(p_2) - S^{-1}(p_1) \tag{5.19}$$

which ensures electromagnetic current conservation.

Meson poles in the quark–antiquark scattering amplitude, G_4, also appear in these vertices, depending on their quantum numbers. For the quark–photon vertex this automatically leads to Vector Meson Dominance (VMD), a model for the coupling of photons to hadrons that predates QCD [885] (see below). In the case of the axial-vector vertex, near a pseudoscalar meson pole at $\widehat{P}^2 = -M_{\rm PS}^2$, we have[32]

$$\begin{aligned}\Gamma_5^\mu(p;P) &\approx \frac{\Gamma_{\rm PS}(p;\widehat{P})}{P^2+M_{\rm PS}^2} Z_2 N_c \int \frac{d^4k}{(2\pi)^4} {\rm Tr}[\chi_{\rm PS}(k;\widehat{P})\,\gamma_5\,\gamma_\mu] \\ &= \frac{\Gamma_{\rm PS}(p;\widehat{P})}{P^2+M_{\rm PS}^2} f_{\rm PS}\widehat{P}^\mu\end{aligned} \tag{5.20}$$

with $f_{\rm PS}$ the pseudoscalar decay constant, which governs the coupling of a pseudoscalar meson to the axial-vector current.

Similarly, pseudoscalar mesons appear as poles in the pseudoscalar vertex, and near $\widehat{P}^2 = -M_{\rm PS}^2$ this vertex

[32] Remember we are using Euclidean metric here.

behaves as

$$\Gamma_5(p;\,P) \approx \frac{\Gamma_{\rm PS}(p;\,\widehat{P})}{P^2+M_{\rm PS}^2} Z_4 N_c \int \frac{d^4k}{(2\pi)^4} {\rm Tr}[\chi_{\rm PS}(k;\,\widehat{P})\,\gamma_5] = \frac{\Gamma_{\rm PS}(p;\,\widehat{P})}{P^2+M_{\rm PS}^2} r_{\rm PS}(\mu) \tag{5.21}$$

with $r_{\rm PS}(\mu)$ the (renormalization-scale dependent) residue in the pseudoscalar channel. The AV-WTI relates the residues at these poles

$$f_{\rm PS}\, M_{\rm PS}^2 = -2\, m_q(\mu)\, r_{\rm PS}(\mu), \tag{5.22}$$

which holds for any pseudoscalar meson. Therefore, in the chiral limit, $m_q(\mu) = 0$, either $f_{\rm PS}$ or $M_{\rm PS}$ must be zero. (If they are both zero, chiral symmetry will not be dynamically broken; see below.)

Furthermore, expanding the AV-WTI in powers of $M_{\rm PS}^2$ in the chiral limit, $m_q(\mu) = 0$, and using the most general Dirac decomposition of $\Gamma_{\rm PS}$[33]

$$\Gamma_{\rm PS}(k;\,\widehat{P}) = \gamma_5\big[i\,E + \widehat{\not{P}}\,F + \not{k}\,G + \sigma_{\mu\nu}\,k_\mu \widehat{P}_\nu\,H\big] \tag{5.23}$$

one finds, to leading order in $M_{\rm PS}$,

$$f_{\rm PS} E(p;0) = B(p^2) \tag{5.24}$$

where $B(p)$ is the scalar part of the quark self-energy.

Thus, if chiral symmetry is dynamically broken, that is, if $m_q(\mu) = 0$ but $B(p^2) \neq 0$, $f_{\rm PS}$ is nonzero, see Eq. (5.24), and pions necessarily emerge as massless Goldstone bosons, see Eq. (5.22). Furthermore, the pseudoscalar component of the pion BSA is proportionally to the (dynamically generated) scalar self-energy of the quarks. In addition, the AV-WTI implies that the decay constant of excited pions (which necessarily have nonzero mass) has to vanish in the chiral limit.

These relations are exact, and the asymptotic behavior of the canonical pion BSA component can be obtained from the asymptotic behavior of the mass functions shown in Fig. 80. The same asymptotic behavior of the canonical BSA component also holds with nonzero current quark masses; as well as for excited pseudoscalar mesons.

Finally, with the definition of $r_{\rm PS}$ implicitly given in Eq. (5.21) and the relation (5.24), we arrive at the well-known Gell-Mann–Oakes–Renner relation

$$f_\pi^2\, m_\pi^2 = 2\, m_q(\mu)\, \langle\bar{q}q\rangle^\mu_{\rm chiral}, \tag{5.25}$$

with the chiral condensate

$$\langle\bar{q}q\rangle^\mu_{\rm chiral} = Z_4 N_c \int \frac{d^4k}{(2\pi)^4} \frac{4\,B_{\rm chiral}(k^2)}{k^2 A^2(k^2) + B^2_{\rm chiral}(k^2)}. \tag{5.26}$$

[33] Here E, F, G, and H scalar functions of k^2 and $k\cdot\widehat{P}$; for equal-mass mesons with $\eta = \frac{1}{2}$, the functions E, F, and H are even in $k\cdot\widehat{P}$, whereas G is odd in $k\cdot\widehat{P}$.

Note that the renormalization scale dependence of the current quark mass, $m_q(\mu)$, exactly cancels that of the chiral condensate.

5.2.5 Mesons in rainbow-ladder (RL) truncation

Different types of mesons, such as pseudoscalar (pions, kaons) or vector mesons (ρ, ϕ), are obtained by considering the most general Dirac and flavor (isospin) structure for the meson of interest, and solving the BSE, Eq. (5.4), at the bound state pole.[34]

To obtain practical solutions from the exact BSE, Eq. (5.4), the kernel K must be truncated; furthermore, one needs to approximate the dressed quark propagators. The most commonly used truncation is the ladder truncation, in which the BSE kernel K in Eq. (5.4) is replaced by an one-gluon exchange (or, in the case of QED, a one-photon exchange)

$$K_{ij}(p,k;\,\widehat{P})\mathcal{O}_i\otimes\mathcal{O}_j \;\to 4\pi\,\alpha(q^2)\,D^{\rm free}_{\mu\nu}(q)\tfrac{\lambda^i}{2}\gamma_\mu\otimes\tfrac{\lambda^i}{2}\gamma_\nu, \tag{5.27}$$

with a model for the effective running coupling $\alpha(q^2)$. Here we use the ladder truncation, in combination with quark propagators that are the solution of the DSE in rainbow truncation – hence we refer to it as the Rainbow-Ladder (RL) truncation.

The resulting approximate BSE is solved numerically, starting from the Euclidean metric, and analytically continuing $\widehat{P}^2$ to negative values while keeping the integration variable Euclidean. This leads to complex momenta for the quark propagators, which is trivial with bare constituent propagators; it is also well-defined and straightforward to implement for (nonperturbatively) dressed propagators as long as there are no singularities in either the (dressed) propagators or the model for the effective interaction over a well-defined domain in the complex momentum plane, depending on the meson mass, choice of η, and choice of frame[35] – though one may have to solve the quark DSE numerically over this domain.

In the previous section we showed in detail that the pion is the Goldstone boson associated with chiral symmetry breaking; it becomes massless in the chiral limit; and its canonical BSA component is given by the scalar self-energy of the quark. The ladder truncation by itself, in combination with bare propagators, does not preserve these features of the pion.

[34] The bound state mass is not known a priori; therefore one has to vary $\widehat{P}^2$ until one finds a solution. This is most conveniently done by introducing a fictitious eigenvalue λ in front of the LHS of Eq. (5.4) to turn it into an eigenvalue problem, and search for a solution with $\lambda = 1$ by varying $\widehat{P}^2$.

[35] The BSA is frame independent, but in Euclidean metric, $k\cdot P$ is purely imaginary in the restframe (remember $\widehat{P}^2$ is negative), and becomes generally complex in a moving frame. It has been shown that physical observables are indeed frame independent by solving the BSE in RL truncation explicitly in a moving frame [886].

However, the RL truncation with consistent dressed quark propagators does preserve the Goldstone nature of the pion, which one can prove analytically using Eq. (5.18) and performing a shift in integration variables.[36]

The RL truncation has been used extensively over the past 25 years, not only for pions, but also for other quantum numbers, and both for light systems, heavy systems, and heavy-light systems. A commonly used model for the interaction is the Maris–Tandy model [868]. This model is finite in the infrared region, with sufficient strength for dynamical chiral symmetry breaking, and agrees perfectly with pQCD for $q^2 > 25\,\mathrm{GeV}^2$. The dynamical mass function of the up/down quarks, strange quarks, and charm quarks were shown in Fig. 80.

For the light pseudoscalar and vector mesons, consisting of u, d, and s quarks, we find excellent agreement with the experimental data, not only for the spectrum, but also for the decay constants. For the charmed mesons (both charmonium, and heavy-light systems) we also find agreement with experiment, within our numerical precision which is dominated by the need to solve the quark propagator over a large domain in the complex momentum plane. Results for axial-vector and scalar mesons are much less in agreement with experiment, but it is known that leading-order corrections to the RL truncation are significantly larger in the axial-vector and scalar channels than in the pseudoscalar and vector channels. Furthermore, the scalar mesons are notoriously difficult to describe, and are likely to have a significant 4-quark content (in particular the broad σ meson, if it can be called a meson).

Meson form factors and scattering

With the BSA we can evaluate a range of other physical observables. We have already mentioned the electroweak decay constant, but more interesting are processes with three external probes such as mesons and/or photons. Consider the elastic form factor of a meson: the right panel of Fig. 79 shows the coupling of a photon to a meson in impulse approximation. One can show analytically that if one considers the dressed quark–photon vertex as the solution of its inhomogeneous BSE using the same RL kernel as for the quark propagators and the meson BSE, current conservation is automatically guaranteed. Another advantage of using such a dressed quark–photon vertex, instead of a bare vertex, is that vector meson poles will automatically appear as poles at $Q^2 = -M_V^2$ in the dressed vertex; thus, VMD is unambiguously included in this approach [887].

A practical challenge is that at least one of the mesons in Fig. 79 has to be in a moving frame. For small values of Q^2 one can use a Taylor expansion of the BSA in the rest frame, but explicitly solving the BSE in a moving frame greatly improves the accessible domain in Q^2 and reduces numerical uncertainties associated with e.g. a Taylor expansion. Figure 81 shows the predictions from the Maris–Tandy model in RL truncation for the pion elastic form factor, which are in perfect agreement with the data. For comparison, we also include a simple VMD model, as well as a recent LF calculation [888] discussed in more detail in Sect. 5.3.

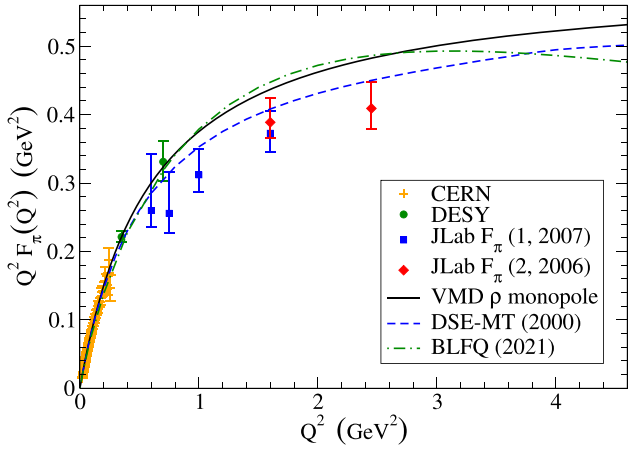


Fig. 81 Spacelike pion form factor: comparison between experiment and a VMD model, DSE in RL truncation [869,887], and a recent LF calculation [888]. For the experimental data, see Refs. [889,890] and references therein

Similar diagrams can and have also been used for electroweak transition form factors and the anomalous $\pi^0 \to 2\gamma$ process [891]. One finds generally good agreement with experimental data, thanks to the fact that this approach satisfies all constraints coming from electromagnetic current conservation, chiral symmetry, and dynamical chiral symmetry breaking; furthermore, it includes unambiguously VMD effects, and it also agrees with perturbative QCD at large momenta. This is not to say that there are no short-comings in this approach: obviously there is physics beyond the RL truncation that is important, some of which are discussed below.

More challenging are scattering observables involving four external mesons and/or electroweak probes. Based on the success of describing form factors in impulse approximation, one might consider just the box diagram with dressed vertices and propagators for such processes. However, it turns out that this is insufficient, and does not reproduce the expected results for e.g. $\pi\pi$ scattering or $\gamma\,3\pi$ coupling – which are both constrained by chiral symmetry. For a consistent description of scattering observables involving four external probes, one needs to include the same RL kernel inside the box diagram as well, resummed to all orders, as indicated in Fig. 82. With these ladder diagrams added to the box diagram, it has been shown explicitly that both the anomalous $\gamma\,3\pi$ process [892] and $\pi\pi$ scattering [893] are

[36] Hence the need for translationally-invariant regularization of potentially divergent integrals – this is also necessary for ensuring current conservation in electromagnetic interactions.

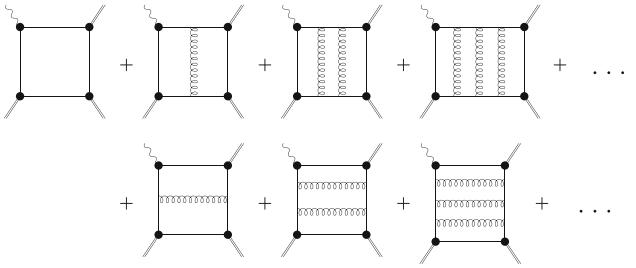


Fig. 82 RL truncation for $\gamma\, 3\pi$ consistent with chiral symmetry and electromagnetic current conservations: quark propagators, vertices and box-diagram all dressed with the same RL kernel (adapted from [892])

in perfect agreement with chiral symmetry and electromagnetic current conservation. The same approach can in principle also be used for other processes, involving other mesons, and it would be very interesting to extend this approach in the future to e.g. Compton scattering on hadrons, as well as pion–nucleon scattering.

5.2.6 Beyond the RL approximation

Over the past two decades significant progress has been made in improving the RL truncation while preserving the relevant vector and axial-vector WTIs [867,894,895]. Although the details of these investigations differ, the general conclusion is that corrections beyond RL are relatively small in the pseudoscalar and vector channels, but can be significantly larger in the axial-vectors and scalar channels. This makes it understandable why the pseudoscalar and vector meson masses and decay constants are in such good agreement with data, but at the same time an accurate description of mesons with other quantum numbers requires going beyond RL.

One of the more promising methods to go beyond the RL truncation is based on the n-Particle Irreducible (n-PI) effective action, in particular the 2-PI and 3-PI effective action up to 3 loops [822,896]. This generally leads to coupled integral equations for the quark, gluon, and ghost propagators, the quark–gluon vertex (and possibly other vertices), and possibly higher n-point functions. Computationally, solving these coupled sets of integral equations in multiple variables is significantly more complicated and time consuming than the RL truncation, but with current (and future) computational resources, the resulting integral equations can be solved for selected cases. The spectrum obtained for the light mesons (including the axial-vector mesons) is in good agreement with available data, see Fig. 83; the only obvious disagreement is in the scalar channel, where pion loops play an important role.

Higher Fock components

Although the RL truncation appears to be quite successful for a range of meson observables, it has its limitations. Consider the pion form factor: Fig. 81 shows this form factor in

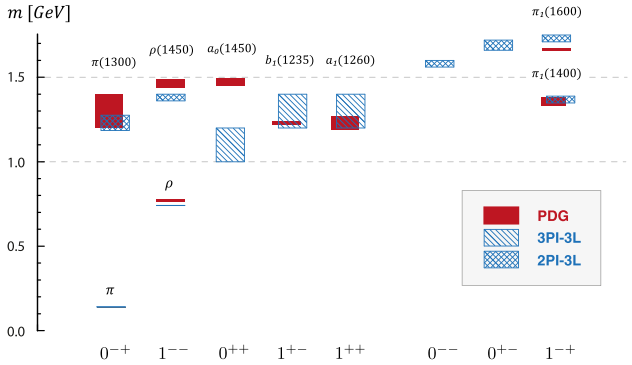


Fig. 83 Light meson spectrum beyond RL truncation (Figure adapted from [896])

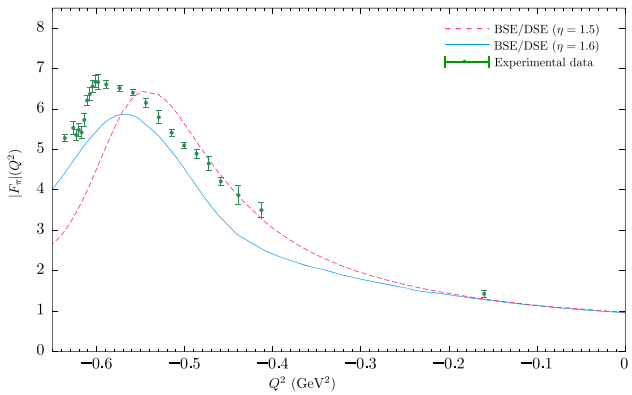


Fig. 84 Pion form factor with pion loops in the timelike region (figure adapted from [897])

the spacelike region, but we can also extend these calculations to the timelike region. In the timelike region, we find a pole at $Q^2 = -M_\rho^2$; exactly as one would expect, because we already know that the homogeneous BSE for the vector channel has a solution at $\widehat{P}^2 = -M_\rho^2$. However, this pole is above the 2π threshold – in the real world, this pole is shifted to the second Reimann sheet, and there is a resonance peak with non-zero width at $Q^2 = -M_\rho^2$. Indeed, incorporating pion loops in the dressed quark–photon vertex in the timelike region changes the vector-meson pole to a resonance peak, and the resulting form factor is in good agreement with the data [897], see Fig. 84. Although the center of the peak is slightly shifted compared to the data, the peak height and width are in good agreement with the data in the timelike region

Similarly, pion loops are likely to be important for the scalar mesons, which can be included by incorporating configurations with two quarks and two anti-quarks in the BSE. This leads to a set of coupled equations between the usual quark–antiquark components, as well as 'meson–meson' contributions and 'diquark–diquark' contributions. This has recently been implemented for the scalar channel [898], which reveals that the σ meson is indeed dominated by two-

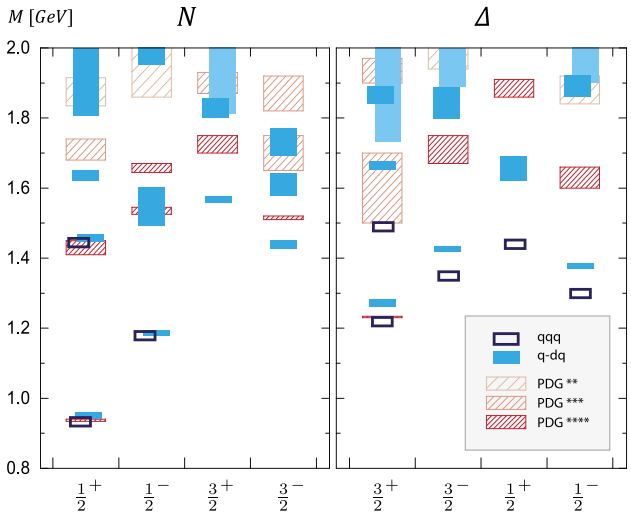


Fig. 85 Baryon spectrum in RL truncation in the quark–diquark picture (blue bars) and as three-quark bound state (open boxes), compared to experimental spectrum (figure adapted from [899])

pion contributions, as one might expect. This approach will also be very useful to investigate exotic mesons, tetraquarks, and in the future also pentaquarks, all within the same framework.

5.2.7 Baryons

The notion of diquarks has been around for almost as long as QCD; see e.g. Ref. [791] and Sect. 5.1.4. Initial DSE studies of baryons were therefore formulated in terms of bound states of a quark and a diquark; specifically a scalar and an axialvector diquark.

However, as described in Sect. 5.2.2, three fermion states can also be described by a 3-body BSE, and in recent years there has been significant progress in describing and understanding baryons as three-quark bound states using the DSE with essentially the same RL approximations as used for the mesons. An effective interaction is modeled using the Dirac structure of a one-gluon exchange between two quarks, see Eq. (5.27), in combination with consistent nonperturbatively dressed quark propagators. Figure 85 shows the calculated spectrum for nucleon and Delta resonances together with the experimental spectrum. The results for the ground state nucleons, as well as their radial excitations, are in fair agreement with experiment, both in the quark–diquark and the three-quark bound state pictures. For the other quantum numbers we see noticable differences between the quark–diquark and three-quark bound state results (and note that not all quantum numbers have been done as a three-quark bound state). The obtained bound state amplitudes can be used for the evaluation of nucleon form factors, see e.g. [899] and references therein, analogous to the calculation of the pion form factor discussed earlier.

5.2.8 Conclusions

At the energy scales of mesons and barons, nonperturbative methods are needed, and the DSEs and BSE (or the CST) work very well. The main shortcoming of these methods is that the kernels needed to solve for the self energies of wave functions are unknown, and must be modeled. The combination of the ladder (L) truncation of the BSE with the closely related rainbow (R) truncation of the DSE for self energies are reasonably successful, in particular in describing chiral symmetry breaking and the role of the pion as the Goldstone boson of QCD. The few calculations beyond the RL truncation that exist show that the additional effects are not large, except in particular spin-isospin channels.

We expect this technique to develop in the years ahead and to remain an attractive method for theoretical study of QCD.

5.3 Light-front quantization

James Vary, Yang Li, Chandan Mondal and Xingbo Zhao

In this section, we discuss non-perturbative light-front Hamiltonian quantization methods. We primarily focus on introducing the Hamiltonians for QED and QCD derived in the light-cone gauge (for extensive reviews, see Refs [900,901]). We introduce methods of solution and results for mesons and baryons. We focus on the Discretized Light Cone Quantization (DLCQ) and Basis Light Front Quantization (BLFQ) methods due to their ability to include gluons and sea quarks dynamically.

Light-front quantization is the natural language for describing the partonic degrees of freedom of QCD at high energies. This connection has been extensively exploited in phenomenological approaches to hard inclusive and exclusive processes (see Sects. 5.8, 5.9). In these approaches, instead of solving the QCD dynamics, the symmetries and properties of QCD are employed to construct phenomenological partonic amplitudes or densities on the light front.

Before introducing specific light-front Hamiltonian methods of solution, let us recap the key concepts of the light-front Hamiltonian approach that spring from Dirac's formulation of Poincare' invariant quantum frameworks [902]. Our choice of light-front variables can be summarized in relation to equal-time variables by introducing

$$P=(P^0+P^3,\, P^0-P^3,\, P^\perp)=\left(P^+,\, \frac{M^2+(P^\perp)^2}{P^+},\, P^\perp\right),$$

where P and M represent the 4-momentum and mass of the hadron, respectively. For the hadron's constituents (quarks, antiquarks, gluons), which we refer to as partons, we adopt $p_i^\perp$ as the transverse momentum of the ith parton, $x_i = p_i^+/P^+$ is its longitudinal momentum fraction, λ_i is its

light-front helicity [903], and roman alphabet subscripts run through the partons of the hadron.

The Hamiltonian eigenvalue problem for the mass-squared eigenstates and their associated light-front wave functions (LFWFs) begins with defining the light-front Schrödinger equation for the system's eigenstates. Taking $P^\perp = 0$ and $H = P^+P^-$

$$H|P, \Lambda\rangle = M^2|P, \Lambda\rangle \tag{5.28}$$

where Λ is the hadron's light-front helicity and H contains kinetic, interaction and Lagrange multiplier terms

$$H = \sum_i \frac{p_i^{\perp 2} + m_i^2}{x_i} + H_{int} + \lambda_{CM} H_{CM}. \tag{5.29}$$

Here, the sum is over all partons and m_i is the mass of the ith parton. The role of the Lagrange multiplier term ensures factorization of the state vector's transverse component into an internal, boost invariant, component times a center of mass (CM) component [904].

We note that this eigenvalue problem applies to systems with arbitrary baryon number so that, for example, it applies to atomic nuclei as well. An eigenstate of a system can be written in terms of a Fock-space expansion over sectors with N-partons as

$$|P, \Lambda\rangle = \sum_N \sum_{\lambda_1,\ldots,\lambda_N} \int \frac{\prod_{i=1}^N \mathrm{d}x_i \mathrm{d}p_i^\perp}{\left[2(2\pi)^N\right]^2 \sqrt{x_1 x_N}} \delta\left(1 - \sum_{i=1}^N x_i\right) \times \delta^2\left(\sum_{i=1}^N p_i^\perp\right) \psi^\Lambda_{\{\lambda_i\}_N}(\{p_i\}_N)|\{\lambda_i, p_i\}_N\rangle, \tag{5.30}$$

where $\psi^\Lambda_{\lambda_1,\ldots,\lambda_N}(p_1, \ldots, p_N)$ is the light-front helicity amplitude for each component. Each of the multi-parton basis states $|\{\lambda_i, p_i\}_N\rangle$ is defined as a properly normalized string of N fermion, anti-fermion and gluon creation operators acting on the vacuum. Eq. (5.30) is schematic since, for fixed N, there can be many subcases with the same net fermion number. We note that the kinetic term in Eq. (5.29) is diagonal in this multi-parton basis. In the following sections, we introduce the discretized and basis function alternatives to Eq. (5.30).

For gauge theories, a traditional approach is to adopt the light-front gauge, $A^+ = 0$, and to reduce the Hamiltonian to the minimum number of dynamical degrees of freedom using constraint equations. For QED and QCD this produces the H_{int} term of Eq. (5.29) expressed in terms of Pauli spinors with the boson-fermion vertices (QED and QCD) as well as boson-boson vertices (QCD only). In addition to these vertices, the gauge-fixing and reduction procedures lead to higher-order instantaneous interactions which manifest divergences. The resulting 3(7) vertices for QED [88] (QCD [905,906]) are deceptively simple and are shown in Fig. 86.

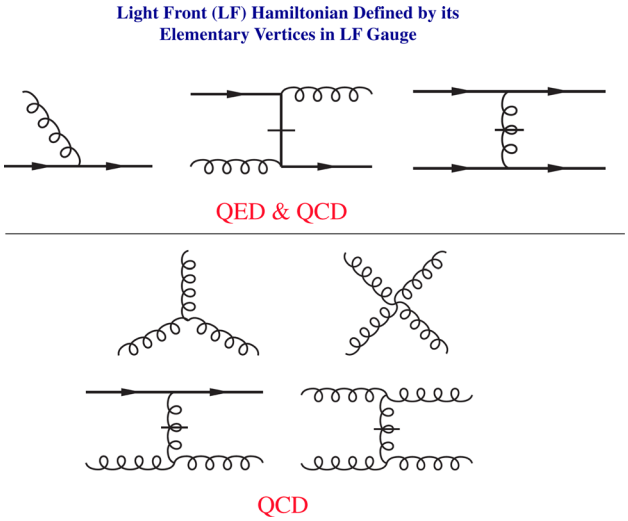


Fig. 86 Vertices appearing in the LF Hamiltonian term H_{int} of Eq. (5.29) upon choosing the LF gauge $A^+ = 0$ for QED [88] and additional vertices for QCD [905,906]. See [900] for a recent review. Solid lines represent fermions (vertices with antifermions are obtained by reversing a fermion line) and wavy lines represent gauge bosons. A graph that includes a fermion or boson with a horizontal line through it represents an instantaneous interaction term. Though one LF time ordering is pictured (increasing LF time flows to the right), all allowed LF time orderings are included in H_{int}. Thus, for example, an incoming line can be switched to an outgoing line at any vertex and vice-versa

Like its Lagrangian counterpart, Hamiltonian field theory needs to be regularized and renormalized. Dimensional regularization is only available for perturbative calculations. In non-perturbative solutions, the invariant mass cutoff and the Pauli–Villars regularization are often adopted. Since non-perturbative eigenvalue problems have to be solved numerically, finite discretization schemes are also needed. One can choose to use the discretization to define the regularization. DLCQ and BLFQ are such schemes. Alternatively, the discretization can be used purely as the numerical method. The problem remains to take the continuum limit. Thanks to the kinematical nature of the light-front boosts, cluster decomposition remains available in the continuum scheme. Hence perturbative type renormalization can be extended to this scheme, as realized in Fock sector dependent renormalization [907].

The similarity renormalization group (SRG) approach is another non-perturbative approach based on Wilson's renormalization group evolution [908,909]. Thanks to asymptotic freedom, the SRG transformation can be evaluated perturbatively up to some scale, say, a few GeV. Different schemes were designed for implementing SRG, notably the Bloch–Wilson formulation [910] and the renormalization group procedure for effective particles (RGPEP [911]). An RGPEP effective Hamiltonian for heavy flavor hadrons is derived using a gluon mass ansatz [912]. In the gluon sector, it

successfully reproduces asymptotic freedom in the 3-gluon effective vertex [913].

The Fock space expansion (Eq. 5.30) provides the most straightforward representation of the eigenvalue problem Eq. (5.28). Within this basis, the eigenvalue equation becomes an infinite tower of coupled integral equations. The integrations can be evaluated using standard numerical techniques; however, truncation is needed to obtain numerical solutions. The situation is similar to the Dyson–Schwinger/Bethe–Salpeter equations approach in the covariant formulation (see Sect. 5.2). The light-front Tamm–Dancoff approximation (LFTDA) truncates the Fock sections in terms of the particle number [914]. LFTDA can be systematically renormalized using the Fock sector dependent renormalization [907]. This was used to investigate various theories within few-body truncation (see Ref. [901] for a review). Typically, the convergence of the Fock sector expansion can be checked numerically [915], although the numerical complexity increases dramatically as the number of Fock sectors increases. The light-front coupled cluster method was proposed to improve the convergence and pathology associated with the hard Fock sector truncation by adopting a coherent basis [916].

Another major development in light-front quantization is the discovery of the remarkable connection between light-front dynamics, its holographic mapping to gravity in a higher-dimensional anti-de Sitter (AdS) space, and conformal quantum mechanics, known as light-front holography (LFH). This approach introduces a remarkably simple yet universal confining potential, which underlines the various phenomenological applications in light-front QCD. See Sect. 5.4 for details.

5.3.1 Discretized light-cone quantization

While lattice calculations (see Sect. 4) solve QCD in Euclidean spacetime, DLCQ formulates the problem directly in Minkowski spacetime using a discretized momentum basis (see Ref. [900] and references therein).

In DLCQ, one defines a mesh in momentum space that corresponds to standing waves in a box of length L in each transverse direction and a similar set of modes in the longitudinal direction. Either periodic or anti-periodic boundary conditions are applied. Early applications of DLCQ to gauge theories included solving QED for positronium at strong coupling [917]. Similarly, early successes include solving QCD in 1+1 dimensions [918]. Moving to QCD in 3+1 dimensions with DLCQ revealed formal and numerical challenges but produced many valuable results as reviewed in Ref. [901].

A hybrid light-front DLCQ/lattice formulation was introduced and employed to evaluate parton distribution functions for a sample set of meson states over a range of coupling strengths [919,920]. These applications of DLCQ motivated the quest for an approach that both preserves the LF kinematic symmetries and provides a computational path with improved numerical efficiency.

5.3.2 Basis light front quantization

The quest to develop LF Hamiltonian approaches in Minkowski-space that retain all available kinematic symmetries began with adoption of basis function methods for solving light front wave equations [921]. Later, the BLFQ approach [922] was introduced to treat gauge theory Hamiltonians using basis-functions that satisfied very general mathematical conditions and respected the LF kinematic symmetries. In addition, the BLFQ framework is well-suited for a longer-term goal of developing basis functions that approximated anticipated dynamical features of QCD such as confinement and chiral symmetry breaking for applications to hadron spectra. Such basis functions have the promise of facilitating convergence in non-perturbative LF QCD calculations.

In BLFQ, one introduces an alternative to the momentum space representation of the LF eigenstate presented in Eq. (5.30). Instead of working with LF plane waves, BLFQ introduces a superposition of orthonormal N-parton Fock space states expressed as independent partons in some convenient orthonormal single-parton basis. That is, we replace the conventional quantization in terms of LF plane waves with LF quantization in modes of a solvable single-parton LF Schrödinger equation akin to Eq. (5.28). Thus, the LF many-parton basis states can be written as strings of fermion, antifermion and boson creation operators that populate independent modes of the single-parton LF Schrödinger equation. All applications described below elect the 2D Harmonic Oscillator for the transverse modes owing to the ability to preserve transverse boost invariance. This choice is further motivated by holographic light-front QCD (see Sect. 5.4 for details) and has been our default choice for practical calculations. For the longitudinal modes there have been a number of choices including DLCQ. In principle, the basis is arbitrary within general mathematical restrictions so convenience and numerical efficiency are the key drivers for the choices represented in applications to date.

Let us label the set of quantum numbers for each single-parton mode with a lower-case Greek letter. This Greek label symbolizes the collection of all space-spin-color-flavor degrees of freedom of a single parton in QCD. Fermion and boson single-parton states are orthonormal and complete. Their creation operators satisfy the conventional anti-commutation (commutation) relations for fermions (bosons). In BLFQ, an eigenstate of a system can then be written in terms of a Fock-space expansion over sectors with N-partons as

$$|P, \Lambda\rangle = \sum_N \sum_{\{\alpha_i\}_N} \psi^{\Lambda}_{\{\alpha_i\}_N} |\{\alpha_i\}_N, \Lambda\rangle, \tag{5.31}$$

where the inner sum includes all allowed configurations of N-partons satisfying global symmetry constraints such as baryon number, charge, total helicity projection on the x^- direction, total LF momentum, flavor, etc. For states with two or more bosons, an additional factor is applied to maintain normalization when bosons occupy the same mode.

Up to this point, the Hamiltonian eigenvalue problem of Eq. (5.28) is infinite dimensional in both the number of single-parton modes and the number of Fock sectors. With a well-chosen BLFQ single-parton basis (see Sect. 5.3.7 for recent advances) and the vertices of QCD from Fig. 86, one hopes to achieve reasonable bound state properties with practical cutoffs in these sums suitable for low-resolution applications of QCD for spectra, electroweak transitions, form factors at low-Q^2, etc.

5.3.3 BLFQ with QED applications

Early applications of BLFQ aimed at solving strong coupling QED problems in order to establish computational techniques and validate BLFQ for achieving converged results in agreement with other methods. These test cases were demanding since they employed the transverse 2D harmonic oscillator and DLCQ for the longitudinal direction to form a basis space that, while suitable for bound state problems in QCD, is far from ideal for these QED applications.

The first application successfully solved for the electron anomalous magnetic moment in an external 2D harmonic trap and took the limit of removing the trap to verify agreement with the well-known Schwinger result [923]. For this application, the first and second vertices in Fig. 86 are included and sector-dependent renormalization [907] was successfully employed.

The next major advance successfully calculated the electron anomalous magnetic moment directly in free space and at the physical coupling [924,925] using the same LF Hamiltonian and renormalization procedures as Ref. [923] except that the instantaneous vertex was omitted. The demands on the numerical procedures increased dramatically due, in large part, to the slow convergence rate with increasing basis cutoff. The extrapolated result agrees with the Schwinger result to within 0.06% which approximately corresponds to the level of agreement expected between a non-perturbative and a perturbative calculation.

Moving ahead from these early applications, the goals of BLFQ were extended to evaluate additional observables familiar to hadronic physics using the resulting LFWFs. In particular, the BLFQ approach was applied to evaluate the GPDs [926] and the TMDs of the dressed electron [927]. In

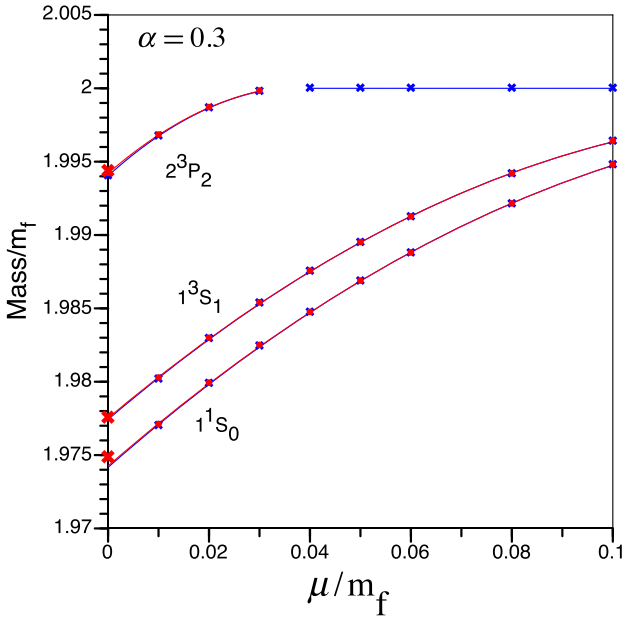


Fig. 87 Positronium spectrum extracted from a BLFQ calculation of QED with an unphysically large coupling $\alpha = 0.3$ [928]. The positronium masses are expressed in terms of the electron mass m_f. The photon mass, μ, serves as an infrared regulator. The positronium states are labeled by the spectroscopic notation $N\,^{2S+1}L_J$. The $O(\alpha^4)$ perturbative results are marked by red crosses on the vertical axis [929]. The blue crosses are obtained from extrapolating $N_{\max} \to \infty$ at fixed and sufficiently large K. For comparison, the results with extrapolated K are shown in solid red disks. The blue and red curves are second order polynomials used to fit and extrapolate the regulator μ to zero

all cases, the non-perturbative BLFQ results compared favorably with results from perturbation theory at weak coupling.

The next major application was to solve for the low-lying spectrum of positronium at strong coupling ($\alpha = 0.3$) in the valence space of the electron and the positron using a derived effective interaction [932]. The application of BLFQ to positronium adopted the LF effective one-photon exchange interaction of Ref. [933], where they achieved a delicate cancellation of the instantaneous photon interaction term through a suitable choice of energy denominators in second order perturbation theory. These calculations were performed in the fermion single-particle basis with the 2D transverse harmonic oscillator and DLCQ for the longitudinal basis. Convergence was achieved directly in K and by extrapolation in $N_{\max}$, the regulators introduced above. The results for the lowest bound states of positronium as a function of the photon regulator mass are shown in Fig. 87. At zero regulator mass, one obtains good agreement with results from perturbation theory. The resulting LFWFs were employed to demonstrate methods of calculating GPDs [934] and reveal relativistic effects in strongly-coupled positronium.

More recently, the BLFQ approach has been successfully applied to solve for the structure of the photon [935]. The basis space consists of the photon sector and the electron–

positron sector so that only the first interaction term from Fig. 86 is retained in solving the Hamiltonian eigenvalue problem of Eq. (5.28). The basis space is defined as for the positronium application above with the addition of the Fock sector for the photon as a single-particle state. Factorization of the CM motion from the LFWFs is addressed using the Lagrange multiplier term in Eq. (5.29) as was accomplished in Ref. [928]. Using sector-dependent renormalization, one achieves the real photon eigenstate to be massless as desired.

The LFWFs obtained for the massless photon are therefore a superposition of a bare photon and an electron–positron pair. These LFWFs provide non-trivial Transverse Momentum Distributions (TMDs) and Parton Distribution Functions (PDFs) which are, in principle, experimentally measurable. Ref [935] provides BLFQ results for TMDs and PDFs in addition to comparisons with results from perturbation theory showing reasonable agreement is obtained as expected.

5.3.4 BLFQ for QCD with effective interactions

The high-precision results from the BLFQ treatment of QED problems (Sect. 5.3.3) provide an avenue to treat the one-gluon-exchange interaction between fermions in QCD (H_{OGE}), which is the dominant short-distance physics for hadrons. The confining interaction from Light-Front Holography (Sect. 5.4), supplemented by a convenient form for confinement in the longitudinal direction, form the long-distance part of the physics (H_{con}). The short distance and long distance terms then lead to the total LF effective interaction, $H_{int} = H_{con} + H_{OGE}$. Similar to the nuclear Shell Model, the solvable part of the Hamiltonian can be chosen to be the kinetic energy plus the confining interaction, $H_0 = H_{kin} + H_{con}$, to implement LFH, augmented with longitudinal confinement, in the zeroth order.

The first application was to compute the spectra and wave functions of heavy quarkonia [930,942]. Figure 88 shows the charmonium and bottomonium spectra obtained from BLFQ. Two parameters, the quark mass and the confining strength, were tuned to fit the available experimental measurements, resulting an r.m.s. deviation of the masses about 40 MeV in each system.

The obtained LFWFs were used to evaluate a wide range of observables, including the decay constants [930,942], light-cone distribution amplitudes [930], form factors [943], radiative transitions [936,938,944], semi-leptonic transitions [945], parton distributions [939] and GPDs [943]. Figure 89 shows the BLFQ results of the charmonium dilepton (for vector mesons, e.g. J/ψ) or diphoton (for the rest) widths in combination with the masses [936], and compared with the available experiments as well as other theoretical approaches whenever available. Figure 90 shows the diphoton transition form factor of η_c from BLFQ, and compared with the *BABAR* measurement. The M1 widths of the radiative tran-

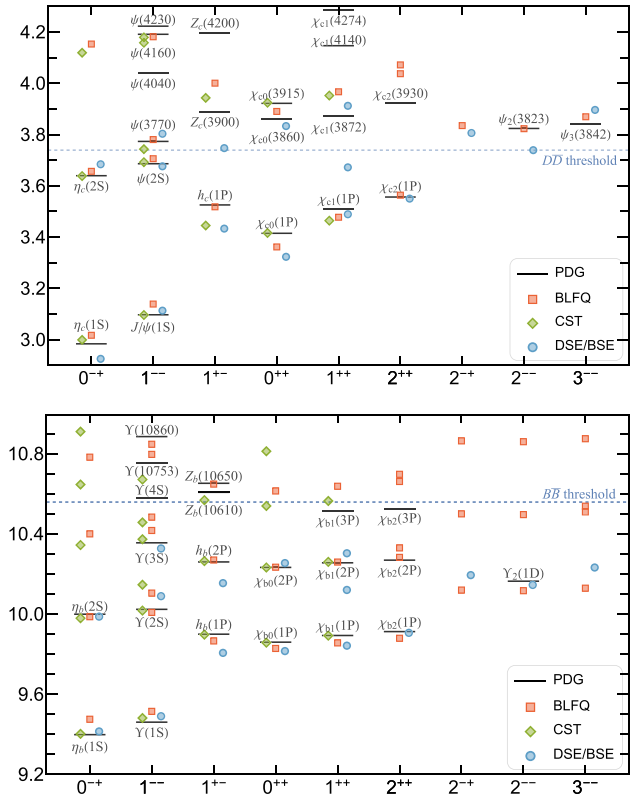


Fig. 88 Charmonium (upper panel) and bottomonium (lower panel) spectra obtained from BLFQ [930], CST [861] and DSE/BSE [931] and compared with the PDG data [616]. See also Sect. 5.2. The vertical axis is the hadron mass in GeV. The horizontal axis is the quantum numbers J^{PC}, where J is the total spin, P, C are the parity and charge conjugation, respectively

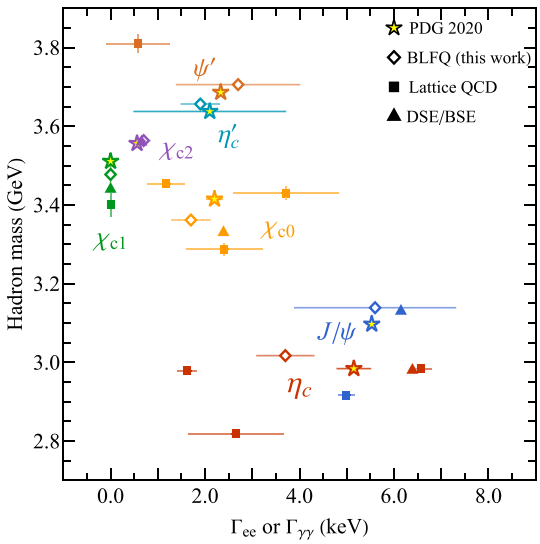


Fig. 89 The BLFQ predictions of the charmonium dilepton (for the vectors) or diphoton (for the rest) widths in combination with the mass spectrum. The experimental data as compiled by the PDG are shown in stars. Lattice and DSE/BSE predictions are shown for comparison (see Ref. [936] and the references therein)

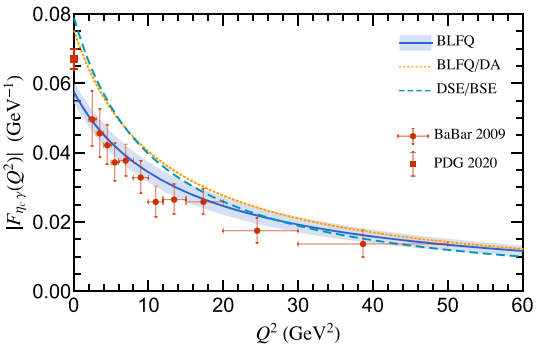


Fig. 90 The singly-virtual two-photon transition form factors of η_c from BLFQ as compared with the BABAR measurement [937] and the predictions from DSE/BSE. The BLFQ/DA result is obtained from pQCD predictions with LCDA obtained from the BLFQ light-front wave functions. The TFF at $Q^2 = 0$ is extracted from the diphoton width. See Ref. [936] and the references therein

sitions across the heavy quarkonium systems are shown in Fig. 91, and compared with the PDG values. The PDF of the hadron at the initial scale μ_0 can be obtained by integrating out the transverse momentum. The PDFs of η_c obtained from BLFQ are shown in Fig. 92.

Applications to heavy-light quarkonia have also been achieved [945–948]. Here, the bottomonia and charmonia results were used to determine the quark masses and the confining strength was calculated using the relationship of heavy-quark effective theory as the r.m.s. of the strengths from the corresponding pure flavor systems. This led to successful applications to the spectra, decay constants and other properties of mixed flavor heavy quarkonia without adjustable parameters.

A major step forward was to apply BLFQ with effective interactions to light mesons [949–952]. In addition to the confining interactions as well as the one-gluon-exchange interaction, a Nambu–Jona–Lasinio (NJL) interaction was incorporated to generate the well-known ρ-π splitting [949]. The obtained LFWFs were used to investigate the partonic structures of the pion. The pion PDF from BLFQ with the effective interactions including the NJL interaction is shown in the top panel of Fig. 93 where the PDF is compared with the PDF from BLFQ calculations that include one dynamical gluon (see Sect. 5.3.5).

More recently, the BLFQ formalism has been successfully applied to solving for the structure of the nucleon [941,954,955] as well as Λ, Λ_c, and their isospin triplet baryons, i.e, Σ^0, Σ^+, Σ^- and Σ_c^0, Σ_c^+, Σ_c^{++} [956]. The investigated observables include the electromagnetic and axial form factors, transverse densities, PDFs, GPDs, radii, axial and tensor charges of the baryons. The electromagnetic form factors of the nucleons are compared with the experimental data as well as other approaches in Fig. 250 in Sect. 10.1. Overall, the theoretical predictions are in good agreement with the experimental measurement for the proton, while the neutron results somewhat deviate from experimental data. The neutron's charge form factor falls well below the data at low Q^2, where both experimental and theoretical uncertainties are large. The magnetic moment of the nucleon is related to the nucleon magnetic form factor at $Q^2 = 0$. We obtained the magnetic moment of the nucleon close to the recent lattice QCD results as shown in Table 4. From the electromagnetic form factors, one can also compute the electromagnetic radii of the nucleon. We summarize our predictions in Table 4. These results are in reasonable agreement with experiment (see Sect. 10.1). Figure 94 shows the nucleon axial form factor (see Sect. 10 for details), $G_A = G_A^u - G_A^d$ as a function of Q^2, while the contributions from up and down quarks to $G_A(Q^2)$ are also displayed. Our results are compared with the available data from (anti)neutrino scattering off protons or nuclei and charged pion electroproduction experiments and the lattice QCD simulations. Considering the experimental uncertainties and our treatment of the BLFQ uncertainties, we found good agreement with experiment.

At $Q^2 = 0$, the axial form factor is identified as the axial charge, $g_A = G_A(0)$. Our prediction, presented in Table 4, is somewhat higher than the extracted data. This discrepancy suggests the need to incorporate higher Fock sectors, which have a significant effect on the quark contribution to the nucleon spin. The corresponding axial radius r_A is in excellent agreement with the extracted data from the analysis of neutrino-nucleon scattering experiments [619,957].

At leading twist, the complete spin structure of the nucleon is explained in terms of three independent PDFs, namely, the unpolarized, the helicity, and the transversity. The obtained LFWFs were also used to evaluate these leading twist quark PDFs. Figure 95 (pink bands) shows the unpolarized PDFs of the valence quarks at $\mu^2 = 10$ GeV^2 for valence-only space results [941] compared with the global fits. The error bands in our PDFs are due to the 10% uncertainties in the initial scale $\mu_0^2 = 0.195 \pm 0.020$ and the coupling constant α_s. Our unpolarized valence PDFs for both the up and the down quarks agree well with the global fits. According to the Drell–Yan–West relation [958,959], at large scale the valence quark distributions fall off at large x as $(1-x)^p$, where p denotes the number of valence quarks and for the nucleon $p = 3$. In our BLFQ approach, we observed that the up quark unpolarized PDF falls off at large x as $(1-x)^{2.99}$, whereas for the down quark the PDF goes as $(1-x)^{3.24}$. These are in accord with the Drell–Yan–West relation and favour the perturbative QCD prediction [960].

The helicity PDFs are displayed in Fig. 96 (upper panel: pink bands), at the scale $\mu^2 = 3$ GeV^2, for the up and down quarks in the proton. Our BLFQ predictions are compared with the measured data from COMPASS [961]. We found that our down quark helicity PDF agrees reasonably well

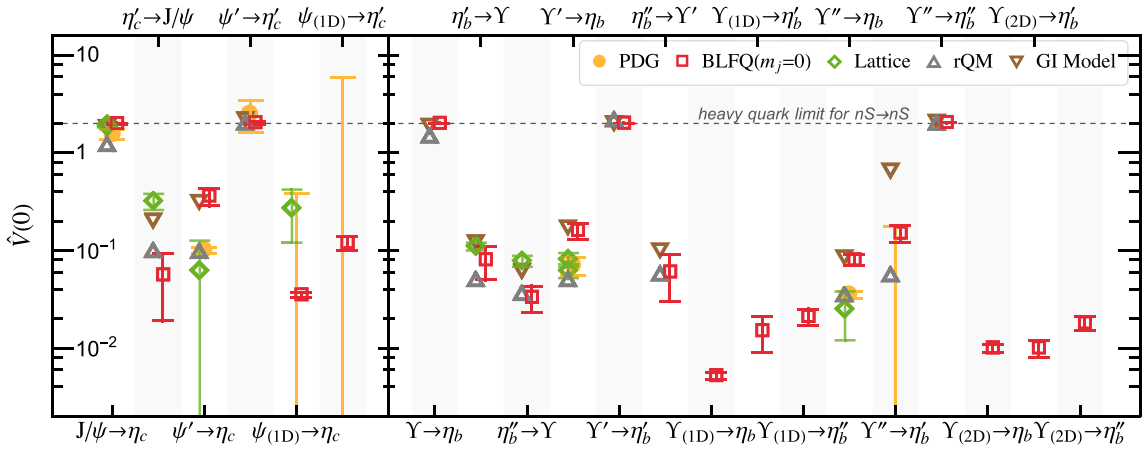


Fig. 91 M1 transition form factor at $Q^2 = 0$ for charmonia and bottomonia obtained from BLFQ and compared with several theoretical predictions as well as the experimental data (see Ref. [938] and the references therein)

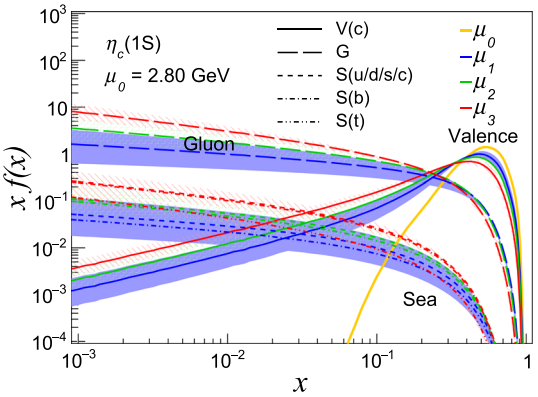


Fig. 92 The PDFs of $\eta_c(1S)$ obtained from BLFQ [939]. The bands represent the range of the distributions for the initial scales $\mu_0 = m_q$ to $2\mu_h$. The lines with different color correspond to the different final scales: $\mu_1 = 20$ GeV (blue), $\mu_2 = 80$ GeV (green), and $\mu_3 = 1500$ GeV (red). The solid, thick long-dashed, dashed, dashed-dot, and dashed double-dot lines represent the x-PDFs of the valence quark, gluon, sea quark ($u/d/s/c$), sea quark (b), and sea quark (t), respectively

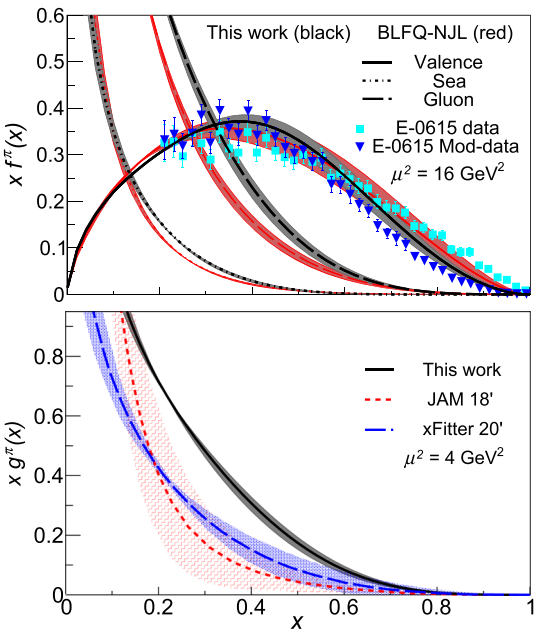


Fig. 93 The PDFs of the pion from BLFQ including one dynamical gluon labeled as "This work" [888]. Upper panel: the black lines are the BLFQ results evolved from the initial scale (0.34 ± 0.03 GeV^2) using the NNLO DGLAP equations to the experimental scale of 16 GeV^2. The red lines correspond to BLFQ-NJL predictions [940]. Results are compared with the original analysis of the FNAL-E615 experiment data and with its reanalysis (E615 Mod-data). Lower panel: the BLFQ result for the pion gluon PDF at $\mu^2 = 4$ GeV^2 is compared with the global fits, JAM and xFitter. See Ref. [888] and the references therein for details

with the experimental data from COMPASS [961]. For the up quark, the $g_1(x)$ solved in the valence-only space is however overestimated at low x, whereas it tends to agree with the data above $x \sim 0.25$ regime.

The obtained LFWFs were also employed to compute the valence quark GPDs for zero skewness [941] and to study quark angular momentum densities inside the proton [963]. The helicity non-flip unpolarized GPD in impact parameter space, $\mathcal{H}^q(x, b_\perp)$, can be interpreted as the number density of quarks with longitudinal momentum fraction x at a given transverse distance $b_\perp$ in the nucleon [964]. One can then define the x dependent squared radius of the quark density in the transverse plane as [962]:

$$\langle b_\perp^2 \rangle^q(x) = \frac{\int d^2\vec{b}_\perp b_\perp^2 \mathcal{H}^q(x, b_\perp)}{\int d^2\vec{b}_\perp \mathcal{H}^q(x, b_\perp)}. \tag{5.32}$$

Figure 97 shows the x-dependent squared radius of the proton, $\langle b_\perp^2 \rangle(x) = 2e_u \langle b_\perp^2 \rangle^u(x) + e_d \langle b_\perp^2 \rangle^d(x)$ and compares the BLFQ prediction with the available extracted data within the range $0.05 \lesssim x \lesssim 0.2$ from the DVCS process [962]. As

Table 4 The electromagnetic properties (magnetic moments in units of nuclear magnetons and radii in units of fm), axial charge, axial radius, tensor charge, and the first moment of transversity PDFs. The BLFQ results are compared with the data extracted from experiments and the lattice QCD simulations (see Ref. [941] and the references therein)

Quantity	BLFQ	Experiments	Lattice
$\mu_{\rm p}$	2.44(3)	2.79	2.43(9)
$\mu_{\rm n}$	−1.40(3)	−1.91	−1.54(6)
$r_{\rm E}^{\rm p}$	0.802(40)	0.833(10)	0.742(13)
$r_{\rm M}^{\rm p}$	0.834(29)	0.851(26)	0.710(26)
$(r_{\rm E}^{\rm n})^2$	−0.033(198)	−0.116(2)	−0.074(16)
$r_{\rm E}^{\rm n}$	0.861(20)	0.864(9)	0.716(29)
$g_{\rm A}^{u}$	1.16(4)	0.82(7)	0.830(26)
$g_{\rm A}^{d}$	−0.248(27)	−0.45(7)	−0.386(16)
$g_{\rm A}^{u-d}$	1.41(6)	1.2723(23)	1.237(74)
$r_{\rm A}$	0.680(70)	0.667(12)	0.512(34)
$g_{\rm T}^{u}$	0.94(15)	0.39(15)	0.784(28)
$g_{\rm T}^{d}$	−0.20(4)	−0.25(20)	−0.204(11)
$\langle x\rangle_{\rm T}^{u-d}$	0.229(48)	–	0.203(24)

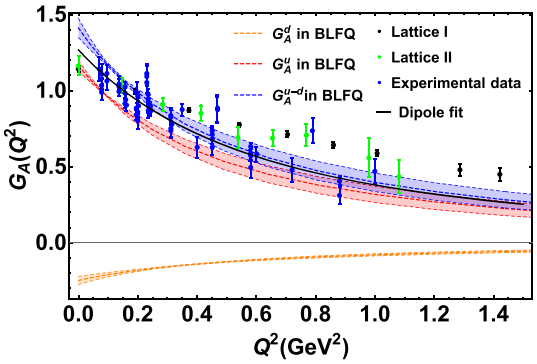


Fig. 94 The axial form factors $G_{\rm A} = G_{\rm A}^u - G_A^d$ and $G_{\rm A}^u$, $G_{\rm A}^d$ as the function of Q^2 from BLFQ. The blue band ($G_{\rm A}$), pink band ($G_{\rm A}^u$), and orange band ($G_{\rm A}^d$) are the BLFQ results, which are compared with the experimental measurements as well as the lattice results. The black line represents the dipole fit of the experimental data. See Ref. [941] and the references therein

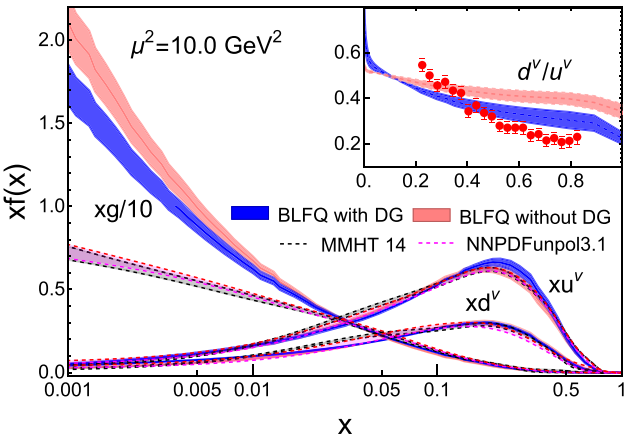


Fig. 95 The unpolarized valence quark and gluon PDFs of the proton. The BLFQ results (blue bands: obtained with one dynamical gluon; pink bands: obtained from a light-front effective Hamiltonian based on only a valence Fock representation [941]) are compared with the NNPDF3.1 and MMHT global fits. (The inset) the ratio of the valence quark PDFs is compared with the extracted data from JLab MARATHON experiment. See Ref. [953] and the references therein

can be seen from Fig. 97, the BLFQ prediction for $\langle b_\perp^2\rangle(x)$ is consistent with the extracted data. We also evaluated the proton's transverse squared radius [962]

$$\langle b_\perp^2\rangle = \sum_q e_q \int_0^1 dx\, f^q(x)\, \langle b_\perp^2\rangle^q(x). \tag{5.33}$$

In our BLFQ approach, we obtained the squared radius of the proton, $\langle b_\perp^2\rangle = 0.40 \pm 0.04$ fm^2, close to the experimental data [962]: $\langle b_\perp^2\rangle_{\rm exp} = 0.43 \pm 0.01$ fm^2.

BLFQ has been recently applied to investigate the all-charm tetraquark system [965]. The results suggest that the lowest two-charm-two-anticharm state is not a tightly bound tetraquark. In particular, the lowest tetraquark mass extrapolated to the continuum limit in longitudinal resolution K lies above the extrapolated threshold for two separated mesons.

5.3.5 BLFQ beyond the valence Fock sector

In this section, we review more recent applications of BLFQ with the inclusion of dynamical gauge degrees of freedom: to positronium at strong coupling ($\alpha = 0.3$) with one dynamical photon earlier in DLCQ [966] and now in BLFQ [967,968]; to mesons with one dynamical gluon [888] and to the proton with one dynamical gluon [953].

For the BLFQ application to QED, the positronium system with one dynamical photon presents valuable challenges with respect to non-perturbative renormalization [967–969]. The dynamics of the single fermion system must first be obtained and then embedded in the positronium system with consistent counting of the basis space quanta. That is, within a given Fock sector of positronium and within a given configuration, the distribution of quanta for that configuration dictates the renormalized mass of the fermion to be applied and the basis space in which that mass was determined. With this dynamical approach, the leading self-energy divergence is taken into account which opens a path to proceed to larger basis spaces.

Going beyond the leading Fock component for QCD, BLFQ has been successfully employed to solve the unflavored light mesons and nucleon with one dynamical gluon [888,953]. In particular, we adopted an effective light-front Hamiltonian and solved for their mass eigenvalues and eigenstates at the scales suitable for low-resolution probes. Our Hamiltonian incorporates light-front QCD interactions [900]

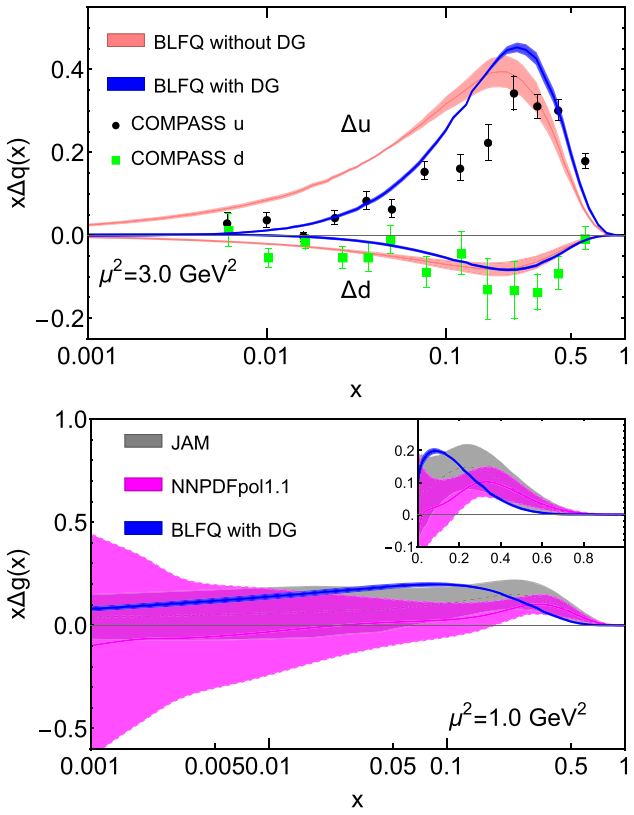


Fig. 96 Upper panel: the helicity PDFs for the valence quarks and the gluon in the proton. We compare BLFQ predictions (blue bands: obtained with one dynamical gluon [953]; pink bands: obtained from a light-front effective Hamiltonian based on only a valence Fock representation [941]) with data from COMPASS Collaboration [961]. Lower panel: the gluon helicity PDF in the proton. We compare the BLFQ prediction (blue bands) with global analyses by JAM (gray band) and NNPDFpol1.1 (magenta band). The inset shows the gluon helicity PDF on a linear scale. See Ref. [953] and the reference therein

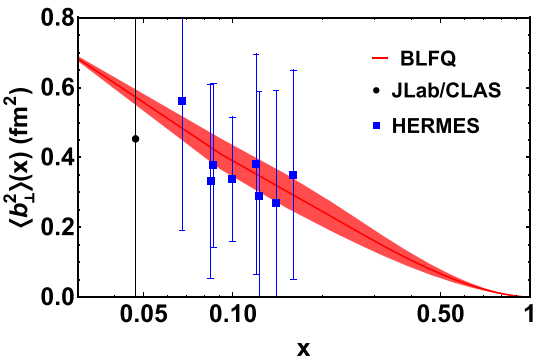


Fig. 97 x-dependence of $\langle b_\perp^2 \rangle$ for quarks in the proton from BLFQ [941]. The line corresponds to the BLFQ predictions and the band indicates its uncertainty. The data points are taken from Ref. [962]

relevant to constituent $|q\bar{q}\rangle$ and $|q\bar{q}g\rangle$ Fock sectors of the mesons and $|qqq\rangle$ and $|qqqg\rangle$ Fock sectors of the nucleon with a complementary 3D confinement [942]. By solving this Hamiltonian in the leading two Fock components and fitting the constituent parton masses and coupling constants as the model parameters [888], we obtained a good quality description of light meson mass spectroscopy[888].

We computed the pion electromagnetic form factor and the PDFs from our Hamiltonian's LFWFs. The BLFQ prediction of the electromagnetic form factor of the charged pion is compared with the experimental data in Fig. 81 in Sect. 5.2. Figure 93 shows our results for the pion PDFs and compares the valence quark distribution after QCD evolution with the data from the E615 experiment as well as the reanalysis of the E615 experiment. The pion PDFs previously obtained in BLFQ-NJL model [940,949,951] based on a valence Fock representation have also been included for comparison. The error bands in our evolved PDFs are manifested from an adopted 10% uncertainty in our initial scale, $\mu_0^2 = 0.34 \pm 0.03\ \text{GeV}^2$, which we determined by requiring the result after evolution to generate the total first moments of the valence quark and the valence antiquark distributions from the global QCD analysis, $\langle x \rangle_{\text{valence}} = 0.48 \pm 0.01$ at $\mu^2 = 5\ \text{GeV}^2$ [970]. We found a good agreement between our prediction for the pion valence quark PDF and the reanalyzed E615 data, while the BLFQ-NJL model favours the original E615 data.

The lower panel of Fig. 93 shows the gluon PDF in the pion. Including one dynamical gluon, the gluon density in the pion significantly increases compared to that in the BLFQ-NJL model as well as to the global fits [971]. The BLFQ-NJL model is based on the pion valence Fock component and gluons are produced solely from the scale evolution. However, the model, which includes a dynamical gluon at the initial scale, results in a larger gluon PDF at large-x (> 0.2) after scale evolution.

We produced the unpolarized and polarized valence quark and gluon distributions in the proton using the resulting LFWFs for the proton with one dynamical gluon. We evolved our initial PDFs from the model scale, $\mu_0^2 = 0.23 \sim 0.25\ \text{GeV}^2$, to the relevant experimental scales. The blue bands in Figures 95 and 96 show our results for the proton unpolarized and polarized PDFs, respectively. We obtained a good consistency between our prediction for the valence quark PDFs and the global fits. The ratio $d^v(x)/u^v(x)$ reasonably agrees with the extracted data from the MARATHON experiment at JLab [972]. At the endpoint, we predicted that $\lim_{x\to 1} d^v/u^v = 0.225 \pm 0.025$.

We found that the down quark unpolarized PDF falls off at large x as $(1-x)^{3.5\pm0.1}$, whereas for the up quark PDF exhibits $(1-x)^{3.2\pm0.1}$. These findings support the perturbative QCD prediction [960]. We observed that the gluon PDF is suppressed at small-x and shifts towards the global fits [664,973] with the addition of a dynamical gluon, whereas the PDF for $x > 0.05$ agrees with the global fits.

Our helicity PDFs for both the up and down quarks (Fig. 96: upper panel) are reasonably consistent with the experimental data from COMPASS [961]. We noticed that

the up quark polarized PDF improves significantly at small-x region with the treatment for the nucleon with dynamical gluon. We observed a fair agreement between our prediction for the gluon helicity PDF (Fig. 96: lower panel) and the global analyses by the JAM [974] and the NNPDF Collaborations [975]. Note that there still remain huge uncertainties both in the large-x region and especially in the small-x region, where even the sign is uncertain. [976]. The partonic spin contributions to the proton spin are given by the first moment of the polarized PDFs. We found that the gluon carries 26% of the proton spin [953], which is likely to increase when more dynamical gluons are included.

5.3.6 Full BLFQ

The applications of BLFQ to hadron structures demonstrated so far have adopted explicit Fock sector truncations. The incorporation of a dynamical gauge boson (Sect. 5.3.5) has shown promising improvements in comparison with valence Fock sector only. A major next step is the full BLFQ [922], in which the Hamiltonians are solved non-perturbatively with basis regulators only and without additional Fock space truncation. The elimination of the additional Fock space truncation positions BLFQ on the path to a genuine ab initio approach to QCD. Initial applications which qualify as full BLFQ include solving scalar 1+1 D field theories without Fock space truncation [977].

The full BLFQ is posed as a quantum many-body problem while the number of partons is not fixed. The single-particle harmonic oscillator basis with the longitudinal discretized momentum basis is the preferred choice of basis, together with the $N_{\max}$-K regularization,

$$\begin{aligned} &\sum_i \left[2n_i + |m_i| + 1\right] \leq N_{\max}, \\ &\sum_i p_i^+ = \frac{2\pi K}{L}. \end{aligned} \tag{5.34}$$

As such, all kinematical symmetries of the LFQCD Hamiltonian, including the the factorization of the center-of-mass motion, are preserved in the many-body Hilbert space. This basis corresponds to a pair of soft IR and UV resolutions and a collinear resolution,

$$b^2/(N_{\max} - 1) \lesssim \sum_i \frac{k_{i\perp}^2}{x_i} \lesssim b^2(N_{\max} - 1), \tag{5.35}$$

$$\Delta x \gtrsim K^{-1}. \tag{5.36}$$

Here, $b = \sqrt{P^+\Omega}$. P^+ is the longitudinal momentum of the bound state. Ω is the scale parameter of the transverse harmonic oscillator functions. Note that, if zero modes are omitted as is conventional, the $N_{\max}$-K regularization renders the number of partons finite, and no further Fock sector truncation is needed.

A fundamental challenge of the full BLFQ is the exponential increase of the dimensionality of the Hilbert space, $\dim \mathcal{H} = N^{dN}$, $(N = \max\{N_{\max}, K\})$, a property shared by all strong coupling non-perturbative quantum many-body problems. Nevertheless, meaningful results may be achievable with continuing advances in high-performance computers at and beyond exascale (10^{18} floating point operations per second). On the other hand, future quantum computers offer the promise to provide supremacy over even the best high-performance computers, in particular for non-perturbative quantum many-body problems such as posed by full BLFQ [978].

5.3.7 BLFQ with chiral symmetry breaking

Due to the light quark mass, $m_{\{u,d\}} \ll \Lambda_{\rm QCD}$, chiral symmetry plays an important role on the light meson spectrum and structures. In particular, the pion is the Goldstone boson of the spontaneously broken chiral symmetry. Formally, chiral symmetry implies a partially conserved axial-vector current (PCAC). In BSE, this relation leads to a set of relations between the pion Bethe–Salpeter amplitudes and the quark self-energy (see Sect. 5.2). Recently, it was revealed that PCAC also leads to a chiral sum rule for the pion LFWFs [979].

It was long pointed out that chiral symmetry breaking in LFQCD is manifested in a different way from the instant form (see Ref. [980] for a recent review). In the instant form, chiral symmetry breaking is due to the condensate of quark–antiquark, viz. $\langle \bar{q}q \rangle \neq 0$. The light-front vacuum is trivial due to the positivity of the longitudinal momenta. Therefore, the vacuum condensate on the light front can only happen through the zero modes. Indeed, the wee parton condensate is long conjectured to be the mechanism for symmetry breaking on the light front, which is supported by 1+1D theories and has shown to be a useful starting point for BLFQ applications [981].

On the other hand, the axial charge on the light front annihilates the light-front vacuum, $Q_5|0\rangle = 0$, which suggests that the chiral condensate should be encoded within the hadron LFWFs [982]. One of the traces of the chiral symmetry breaking in the pion LFWFs is the chiral sum rule [979]. Taking advantage of light-front holography, this sum rule has been shown to be also consistent with the chiral symmetry breaking in AdS/QCD.

5.3.8 Nonperturbative reactions in BLFQ

One major advantage of the Hamiltonian formalism of quantum field theory is that it allows for tracking time evolution of quantum field configurations in real time.

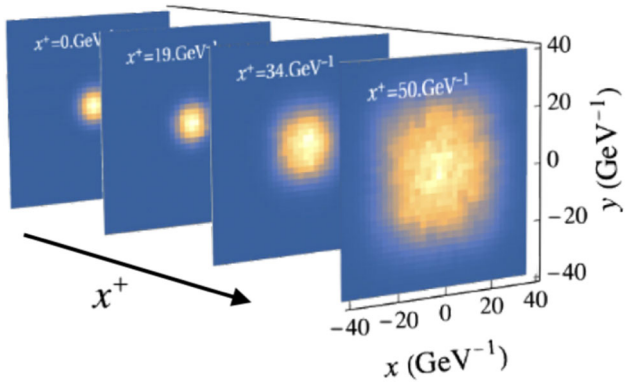


Fig. 98 The evolution of the transverse density of a quark within a classical color field of a heavy nucleus at different light-front time x^+. The four "snapshots" are from Ref. [985]

As an extension of BLFQ, the time-dependent Basis Light-front Quantization (tBLFQ) has been developed as a time-dependent nonperturbative approach to quantum field theory [983]. In tBLFQ the light-front Schrödinger equation is solved to simulate the time evolution of quantum field configurations:

$$i\frac{\partial}{\partial x^+}|\psi;x^+\rangle = \frac{1}{2}P^-(x^+)|\psi;x^+\rangle, \tag{5.37}$$

where $|\psi;x^+\rangle$ represents the quantum field system under consideration and $P^-(x^+)$ is the light-front Hamiltonian, which includes the interactions among the fields under consideration. The tBLFQ approach is suitable for studying particle evolution in a strong and possibly time-dependent background field. The tBLFQ approach motivated a nonperturbative approach simulating nuclear reactions in low energy nuclear physics, named time-dependent Basis Function (tBF) [984].

One of the major goals of tBLFQ is to understand the nonperturbative dynamics in QCD, as in hadron scattering. The investigations of quark scattering with a nucleus constitute a first step toward this goal. In Ref. [985], tBLFQ is employed to simulate the scattering of an ultrarelativistic quark off a heavy nucleus at high energies. The color glass condensate, a classical effective theory of QCD, is adopted as a model for the color field of the heavy nucleus. The results can significantly reduce the theoretical uncertainties in the small $p_\perp$ region of the differential cross section which has important implications for the phenomenology of the hadron-nucleus and deep inelastic scattering at high energies. One important feature of tBLFQ is that it allows one to take "snapshots" of the system at intermediate times of the evolution, which provide physical insights into the nonperturbative mechanism in time-dependent processes. For example, Fig. 98 shows the evolution of the probability distribution of a quark in the transverse direction at different light-front times x^+. In Ref. [987] a calculation is performed in an extended Fock space where one dynamical gluon is included, which paves the way for studying partons' radiational energy loss in nuclear matter [988].

In addition to the applications in QCD, tBLFQ has also been employed to study various nonperturbative processes in strong field QED [983,989–992].

The tBLFQ approach can be further improved in three directions: (i) increase in the level of complexity and realism of the background field: (ii) expansion to reaction processes of a wider class; (iii) the expansion of the Fock space in the description of quantum field configurations. While this can lead to more accurate simulation of dynamical processes, it will dramatically increase the required computational resources. Therefore, it is desirable to explore numerical algorithms for tBLFQ on next-generation advanced computational platforms.

5.3.9 Comparisons between BLFQ and BSE

The similarities and differences of the light-front and the BSE (see Sect. 5.2) approaches motivate a direct comparison of the amplitudes obtained from these two approaches [862]. Figure 88 shows the comparison of quarkonia spectra obtained from BLFQ and CST. In both approaches, the model parameters were fixed by fitting to the experimentally measured quarkonia masses. Then, the obtained wave functions were used to compute physical observables. Figure 99 compares the axial-vector LFWFs obtained from BLFQ and CST. The Brodsky–Huang–Lepage prescription [986] was used to convert the CST amplitude to the LFWFs [862]. Qualitatively, the wavefunctions are similar. However, some spin components show different characteristics due to the different implementation of discrete symmetries, which can be discerned in high-energy exclusive processes.

5.4 AdS/QCD and light-front holography

Stanley J. Brodsky, Guy F. de Téramond, and Hans Günter Dosch

5.4.1 Introduction

In spite of the important progress of Euclidean lattice QCD [97] and other nonperturbative approaches, a basic understanding of fundamental features of hadron physics from first principles, such as the mechanism of color confinement and the origin of the hadron mass scale, as well as general features of hadron structure, spectroscopy and dynamics, have remained among the most important unsolved challenges of the last 50 years in particle physics. Furthermore, other essential properties of the strong interactions, which were manifest in dual models and developed before QCD, are also not explicit properties of the QCD Lagrangian.

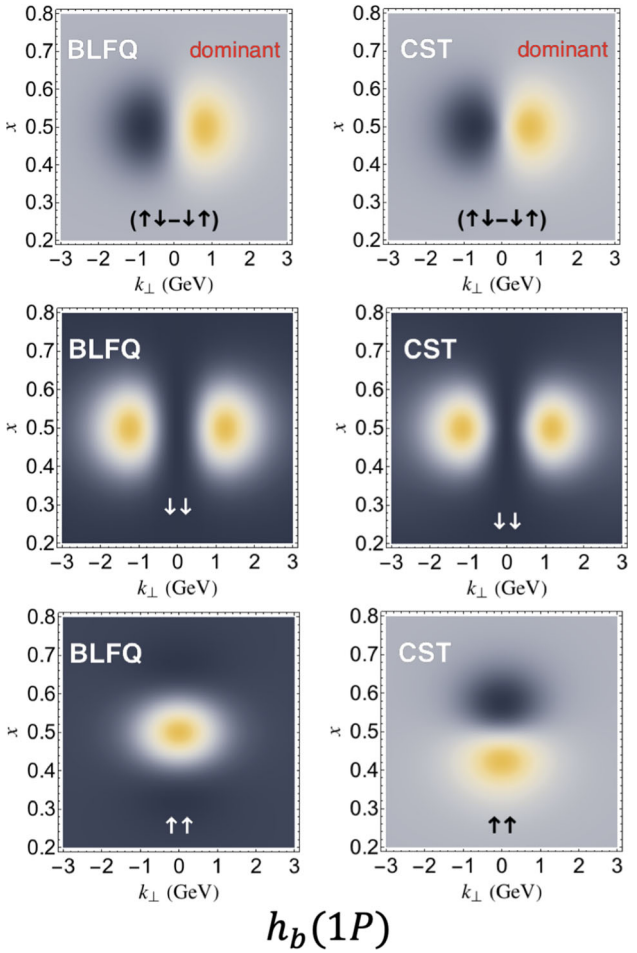


Fig. 99 Comparison of selected LFWFs of h_b obtained from BLFQ and CST [862]. The latter were converted from the BSA with the Brodsky–Huang–Lepage prescription [986]. The non-relativistic dominant spin components from both approaches (top panels) are qualitatively the same. However, subdominant LFWFs may appear dramatically different, some of which are in leading twist (bottom panels). This can be tested in high-energy exclusive processes. The discrepancy stems from the different implementation of the discrete symmetries on the light cone in BLFQ and CST

Recent theoretical developments for understanding features of hadronic physics are based on AdS/CFT – the correspondence between classical gravity in a higher-dimensional anti-de Sitter (AdS) space and conformal field theories (CFT) in physical space-time [993–995]. AdS/CFT has provided a semiclassical approximation for strongly-coupled quantum field theories, giving physical insights into non-perturbative dynamics. In practice, the AdS/CFT duality provides an effective weakly coupled description in a $(d+1)$-dimensional AdS_{d+1} space in terms of a flat d-dimensional superconformal, strongly coupled quantum field theory defined on the AdS asymptotic boundary, the physical four-dimensional Minkowski spacetime, where boundary conditions are imposed [996]. This is illustrated in Fig. 100

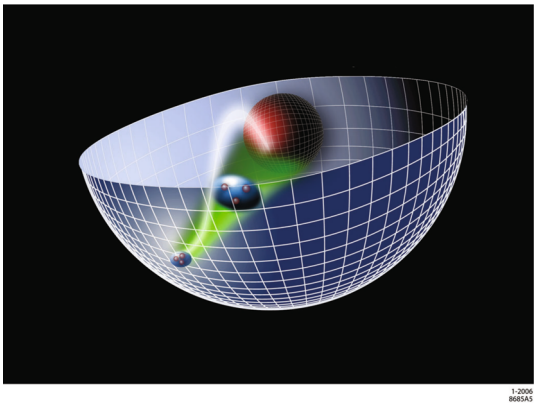


Fig. 100 This figure attempts to show how different values of the AdS holographic coordinate z correspond to different scales at which the proton is examined. Events at short distances in the ultraviolet happen near the four-dimensional AdS boundary (large circumference). The red inner sphere represents large distance infrared events where AdS is modified to model confinement. The green cone represents the warping of AdS space and is due to its negative curvature. A proton (blue ball with seeds) evolves from a small size near the ultraviolet boundary to larger sizes as it propagates towards the infrared region of AdS as perceived by an observer in physical Minkowski space

for $d = 4$, where the asymptotic surface of the 5-dimensional AdS_5 space is the physical four-dimensional Minkowski spacetime.

Anti-de Sitter AdS_{d+1} is the maximally symmetric $d+1$ space with negative constant curvature and a d-dimensional flat spacetime boundary. In Poincaré coordinates $x^M = \left(x^0, x^1, \ldots, x^{d-1}, z\right)$, where the asymptotic border of AdS space is given by $z = 0$. The line element is

$$\begin{aligned} ds^2 &= g_{MN} dx^M dx^N \\ &= \frac{R^2}{z^2}\left(\eta_{\mu\nu} dx^\mu dx^\nu - dz^2\right), \end{aligned} \tag{5.38}$$

where $\eta_{\mu\nu}$ is the usual Minkowski metric in d dimensions, and R is the AdS radius. The group of transformations leaving the AdS_{d+1} metric invariant, the isometry group $SO(2,d)$, has dimension $(d+1)(d+2)/2$. Five-dimensional anti-de Sitter space AdS_5 has thus 15 isometries, which induce in the Minkowski-space boundary the symmetry under the conformal group with 15 generators in four dimensions: 6 Lorentz transformations plus 4 spacetime translations plus 4 special conformal transformations plus 1 dilatation [997]. This conformal symmetry implies that there can be no scale in the boundary theory and therefore no discrete spectrum.

The relation between the dilatation symmetry and the symmetries in AdS_5 can be seen directly from the AdS metric since (5.38) is invariant under a dilatation of all coordinates: A dilatation of the Minkowski coordinates $x^\mu \to e^\sigma x^\mu$ is compensated by a dilatation of the holographic variable

$z \to e^{\sigma} z$. Therefore, the variable z acts like a scaling variable in Minkowski space: different values of z correspond to different energy scales at which a measurement is made. As a result, short spacetime intervals map to the boundary in AdS space-time near $z = 0$. This corresponds to the ultraviolet (UV) region of AdS space. On the other hand, a large four-dimensional object of confinement dimensions $1/\Lambda^2_{\rm QCD}$ maps to the large infrared (IR) region of AdS space $z \sim 1/\Lambda_{\rm QCD}$. Thus, in order to incorporate confinement in the gravity dual the conformal invariance must be broken by modifying AdS space in the large z IR region. For example, a simple way to obtain confinement and discrete normalizable modes (Fig. 100) is to introduce a sharp cut-off at the IR border $z_0 \sim 1/\Lambda_{QCD}$, as in the "hard-wall" model of Ref. [998].

In general, one can deform the original AdS background geometry, giving rise to a less symmetric gravity dual. This approach provides useful tools for constructing dual gravity models in higher dimensions which incorporates confinement and basic QCD properties in physical spacetime. The resulting gauge/gravity duality is broadly known as the AdS/QCD correspondence, or simply holographic QCD, which has become an extensive field of research. The extent to which the full theory of QCD can be described in such a framework remains unclear. It has become clear, however, that holographic models motivated by the AdS/CFT correspondence can capture essential features of QCD and may give important insights into how QCD works. Different models can be derived via a top-down approach from brane configurations in string theory, as well as from more phenomenological bottom-up models, which are not constrained by string theory, and are therefore more flexible for incorporating key aspects of QCD. The best known example of the first category is the Witten–Sakai–Sugimoto model [999], which contains vector mesons and pions in its spectrum arising from the breaking of chiral symmetry. Conversely, in the bottom-up hard-wall model of Refs. [1000,1001], the global $SU(2) \times SU(2)$ chiral symmetry of QCD becomes a gauge invariant symmetry on the gravity side. The AdS/QCD model of Refs. [1000,1001] has also been extended by using the "soft-wall" model introduced in Ref. [1002] in order to reproduce the observed linearity of Regge trajectories.

A third approach to AdS/QCD, *holographic light-front QCD* (HLFQCD) [1003], is based on the holographic embedding of Dirac's relativistic *front form* of dynamics [902] into AdS space. In the front form, the initial surface is the tangent plane to the light cone $x^0 + x^3 = 0$, the null plane, thus without reference to a specific Lorentz frame, in contrast with the usual *instant form* where quantization is defined at $x^0 = 0$. This precise mapping between semiclassical LF Hamiltonian equations in QCD and wave equations in AdS space, [1004] leads to relativistic wave equations in physical space-time (similar to the Schrödinger or Dirac wave equations in atomic physics) and provides an effective computational framework of hadron structure and dynamics [1003].[37]

A remarkable property of HLFQCD is the embodiment of a superconformal algebraic structure which not only introduces a mass scale within the algebra, but also determines the interaction completely [1010–1015].[38] Further extensions of HLFQCD provide nontrivial interconnections between the dynamics of form factors and quark and gluon distributions [1019–1021] with pre-QCD nonperturbative approaches such as Regge theory and the Veneziano model.

In this section we give an overview of relevant aspects of the holographic embedding of QCD quantized in the light front, with an emphasis on the underlying superconformal structure for hadron spectroscopy and hadron duality for amplitude dynamics. Introductory reviews are given in Refs. [1003,1022–1024]. Other reviews describing distinct approaches and aspects of holographic QCD in the context of the gauge/gravity correspondence in addition to Refs. [996,999], are given in Refs. [1025–1027] and in the book [1028], with applications to other topics such as holographic renormalization group flows, QCD at finite temperature and density, hydrodynamics and strongly coupled condensed matter systems.[39]

5.4.2 *Semiclassical approximation to light-front QCD*

A semiclassical approximation to QCD has been obtained using light-front (LF) physics, where the quantization surface is the null plane, $x^+ = x^0 + x^3 = 0$ [902]. Evolution in LF time x^+ is given by the Hamiltonian equation [900]

$$i\frac{\partial}{\partial x^+}|\psi\rangle = P^-|\psi\rangle, \quad P^-|\psi\rangle = \frac{\mathbf{P}_\perp^2 + M^2}{P^+}|\psi\rangle, \tag{5.39}$$

[37] The origins of the light-front holographic approach can be traced back to the original article of Polchinski and Strassler [998], where the exclusive hard-scattering counting rules [153,1005], a property of hadrons in physical spacetime, can be derived from the warped geometry of five-dimensional AdS_5 space. Indeed, one can show that a precise mapping between the hadron form factors in AdS space [1006] and physical spacetime [958,959] can be carried out for an arbitrary number of quark constituents [1007–1009]. The key holographic feature is the identification of the invariant transverse impact variable ζ for the n-parton bound state in physical 3+1 spacetime with the holographic variable z, the fifth dimension of AdS.

[38] The idea to apply an effective supersymmetry to hadron physics is certainly not new [1016–1018], but failed to account for the special role of the pion. In contrast, in the HLFQCD approach, the zero-energy eigenmode of the superconformal quantum mechanical equations is identified with the lightest meson which has no baryonic supersymmetric partner.

[39] Hadron models from effective string configurations in holographic 5-dimensional AdS backgrounds [1029] are useful to describe multiple quark configurations including heavy quarks. See Refs. [1030,1031] and references therein.

for a hadron with 4-momentum $P = (P^+, P^-, \mathbf{P}_\perp)$, $P^\pm = P^0 \pm P^3$, where P^- is a dynamical generator and P^+ and $\mathbf{P}_\perp$ are kinematical. The simple structure of the LF vacuum allows a quantum-mechanical probabilistic interpretation of hadron states in terms of the eigenfunctions of the LF Hamiltonian equation (5.39) in a constituent particle basis, $|\psi\rangle = \sum_n \psi_n |n\rangle$, similar to usual Schrödinger equation. The LF wave functions (LFWFs), ψ_n, underlie the physical properties of hadrons in terms of their quark and gluon degrees of freedom. For a $q\bar{q}$ bound state we factor out the longitudinal $X(x)$ and orbital $e^{iL\theta}$ dependence from ψ,

$$\psi(x, \zeta, \theta) = e^{iL\theta} X(x) \frac{\phi(\zeta)}{\sqrt{2\pi\zeta}}, \tag{5.40}$$

where $\zeta^2 = x(1-x)\mathbf{b}_\perp^2$ is the invariant transverse separation between two quarks, with $\mathbf{b}_\perp$, the relative impact variable, conjugate to the relative transverse momentum $\mathbf{k}_\perp$ with longitudinal momentum fraction x. In the ultrarelativistic zero-quark mass limit the invariant LF Hamiltonian $P_\mu P^\mu |\psi\rangle = M^2 |\psi\rangle$, with $P^2 = P^+P^- - \mathbf{P}_\perp^2$ can be systematically reduced to the wave equation [1004]:

$$\left(-\frac{d^2}{d\zeta^2} - \frac{1-4L^2}{4\zeta^2} + U(\zeta)\right)\phi(\zeta) = M^2\phi(\zeta), \tag{5.41}$$

where the effective potential U comprises all interactions, including those from higher Fock states. The critical value of the LF orbital angular momentum $L = 0$ corresponds to the lowest possible solution. The LF equation (5.41) is relativistic and frame-independent; It has a similar structure to wave equations in AdS provided that one identifies $\zeta = z$, the holographic variable [1004].

5.4.3 Higher integer-spin wave equations in AdS

We start with the AdS action for a tensor-J field $\Phi_J = \Phi_{N_1 \ldots N_J}$ in the presence of a dilaton profile $\varphi(z)$ responsible for the confinement dynamics

$$S = \int d^d x\, dz \sqrt{g}\, e^{\varphi(z)} \left(D_M \Phi_J D^M \Phi_J - \mu^2 \Phi_J^2\right), \tag{5.42}$$

where g is the determinant of the metric tensor g_{MN}, μ is the AdS mass and D_M is the covariant derivative which includes the affine connection.[40],[41] The variation of the AdS action leads to the wave equation

$$\left[-\frac{z^{d-1-2J}}{e^{\varphi(z)}}\partial_z\left(\frac{e^{\varphi(z)}}{z^{d-1-2J}}\partial_z\right) + \frac{(\mu R)^2}{z^2}\right]\Phi_J(z) = M^2\Phi_J(z), \tag{5.43}$$

after a redefinition of the AdS mass μ, plus kinematical constraints to eliminate lower spin from the symmetric tensor $\Phi_{N_1 \ldots N_J}$ [1032]. By substituting

$$\Phi_J(z) = z^{(d-1)/2-J} e^{-\varphi(z)/2} \phi_J(z) \tag{5.44}$$

in (5.43), we find the semiclassical light-front wave equation (5.41) with

$$U_J(\zeta) = \frac{1}{2}\varphi''(\zeta) + \frac{1}{4}\varphi'(\zeta)^2 + \frac{2J-d+1}{2\zeta}\varphi'(\zeta), \tag{5.45}$$

as long as $\zeta = z$. The precise mapping allows us to write the LF confinement potential U in terms of the dilaton profile which modifies the IR region of AdS space to incorporate confinement [1003], while keeping the theory conformal invariant in the ultraviolet boundary of AdS, namely $\varphi(z) \to 0$ for $z \to 0$. The separation of kinematic and dynamic components allows us to determine the mass function in the AdS action in terms of physical kinematic quantities with the AdS mass-radius $(\mu R)^2 = L^2 - (d/2 - J)^2$ and d, the number of transverse coordinates [1004,1032], consistent with the AdS stability bound [1033].

5.4.4 Higher half-integer-spin wave equations in AdS

A similar derivation follows from the Rarita–Schwinger action for a spinor field $\Psi_J \equiv \Psi_{N_1 \ldots N_{J-1/2}}$ in AdS [1032] for half-integral spin J. In this case, however, the dilaton term does not lead to an interaction [1034], and an effective Yukawa-type coupling to a potential V in the action has to be introduced instead [1035–1037]:

$$S = \int d^d x\, dz \sqrt{g}\, \bar{\Psi}_J \left(i\Gamma^A e_A^M D_M - \mu + \frac{z}{R}V(z)\right)\Psi_J, \tag{5.46}$$

where e_A^M is the vielbein and the covariant derivative D_M on a spinor field includes the affine connection and the spin connection. The tangent space Dirac matrices obey the usual anticommutation relations $\{\Gamma^A, \Gamma^B\} = 2\eta^{AB}$. The variation of the AdS action leads to a system of linear differential equations which is equivalent to the second order equations [1032]

$$\left(-\frac{d^2}{d\zeta^2} - \frac{1-4L^2}{4\zeta^2} + U^+(\zeta)\right)\psi_+ = M^2\psi_+, \tag{5.47}$$

[40] The affine connection, the vielbein and the spin connection are important elements in curved spaces, particularly if higher spins are involved. A brief introduction, useful for actual computations in AdS space, is given in Appendix A of Ref. [1003].

[41] In the holographic approach the gluon field emerges as a constituent of the spin-2 metric field g_{MN} in AdS, which is dual to the Pomeron in the 4-dimensional physical space (see Sect. 5.4.15).

$$\left(-\frac{d^2}{d\zeta^2}-\frac{1-4(L+1)^2}{4\zeta^2}+U^-(\zeta)\right)\psi_-=M^2\psi_-, \tag{5.48}$$

with $\zeta = z$, $|\mu R| = L + 1/2$ and equal probability $\int d\zeta\, \psi_+^2(\zeta)^2 = \int d\zeta\, \psi_-^2(\zeta)$. The semiclassical LF wave equations for ψ_+ and ψ_- correspond to LF orbital angular momentum L and $L+1$ with

$$U^{\pm}(\zeta) = V^2(\zeta) \pm V'(\zeta) + \frac{1+2L}{\zeta}V(\zeta), \tag{5.49}$$

a J-independent potential, in agreement with the observed degeneracy in the baryon spectrum.

5.4.5 Superconformal algebraic structure and emergence of a mass scale

Embedding light-front physics in a higher dimension gravity theory leads to important insights into the nonperturbative structure of bound state equations in QCD for arbitrary spin, but it does not answer the question of how the effective confinement dynamics is actually determined, and how it can be related to the symmetries of QCD itself. An important clue, however, comes from the realization that the potential $V(\zeta)$ in Eq. (5.49) plays the role of the superpotential in supersymmetric (SUSY) quantum mechanics (QM) [1038].

Supersymmetric QM is based on a graded Lie algebra consisting of two anticommuting supercharges Q and $Q^\dagger$, $\{Q, Q\} = \{Q^\dagger, Q^\dagger\} = 0$, which commute with the Hamiltonian $H = \frac{1}{2}\{Q, Q^\dagger\}$, $[Q, H] = [Q^\dagger, H] = 0$. If the state $|E\rangle$ is an eigenstate with energy E, $H|E\rangle = E|E\rangle$, then, it follows from the commutation relations that the state $Q^\dagger|E\rangle$ is degenerate with the state $|E\rangle$ for $E \neq 0$, but for $E = 0$ we have $Q^\dagger|E = 0\rangle = 0$, namely the zero mode has no supersymmetric partner [1038]; a key result for deriving the supermultiplet structure and the pattern of the hadron spectrum.

Following Ref. [1011] we consider the scale-deformed supercharge operator $R_\lambda = Q + \lambda S$, with $K = \frac{1}{2}\{S, S^\dagger\}$ the generator of special conformal transformations. The generator R_λ is also nilpotent, $\{R_\lambda, R_\lambda\} = \{R_\lambda^\dagger, R_\lambda^\dagger\} = 0$, and gives rise to a new scale-dependent Hamiltonian G, $G = \frac{1}{2}\{R_\lambda, R_\lambda^\dagger\}$, which also closes under the graded algebra, $[R_\lambda, G] = [R_\lambda^\dagger, G] = 0$. The new supercharge R_λ has the matrix representation

$$R_\lambda = \begin{pmatrix} 0 & r_\lambda \\ 0 & 0 \end{pmatrix}, \quad R_\lambda^\dagger = \begin{pmatrix} 0 & 0 \\ r_\lambda^\dagger & 0 \end{pmatrix}, \tag{5.50}$$

with $r_\lambda = -\partial_x + \frac{f}{x} + \lambda x$, $r_\lambda^\dagger = \partial_x + \frac{f}{x} + \lambda x$. The parameter f is dimensionless and λ has the dimension of $[M^2]$, and thus, a mass scale is introduced in the Hamiltonian without leaving the conformal group. The Hamiltonian equation $G|E\rangle = E|E\rangle$ leads to the wave equations

$$\left[-\frac{d^2}{dx^2}-\frac{1-4(f+\frac{1}{2})^2}{4x^2}+\lambda^2x^2+2\lambda\left(f-\tfrac{1}{2}\right)\right]\phi_+ = E\phi_+, \tag{5.51}$$

$$\left[-\frac{d^2}{dx^2}-\frac{1-4(f-\frac{1}{2})^2}{4x^2}+\lambda^2x^2+2\lambda\left(f+\tfrac{1}{2}\right)\right]\phi_- = E\phi_-, \tag{5.52}$$

which have the same structure as the Euler–Lagrange equations obtained from the AdS/CFT correspondence, but here, the form of the LF confinement potential, λ^2x^2, as well as the constant terms in the potential are completely fixed by the superconformal symmetry [1014,1015].

5.4.6 Light-front mapping and baryons

Upon mapping (5.51) and (5.52) to the semiclassical LF wave equations (5.47) and (5.48) using the substitutions $x \mapsto \zeta$, $E \mapsto M^2$, $f \mapsto L + \frac{1}{2}$, $\phi_+ \mapsto \psi_-$ and $\phi_- \mapsto \psi_+$, we find

$$U^+ = \lambda^2\zeta^2 + 2\lambda(L+1), \tag{5.53}$$

$$U^- = \lambda^2\zeta^2 + 2\lambda L, \tag{5.54}$$

for the confinement potential for baryons [1014]. The solution of the LF wave equations for this potential gives the eigenfunctions

$$\psi_+(\zeta) \sim \zeta^{\frac{1}{2}+L}e^{-\lambda\zeta^2/2}L_n^L(\lambda\zeta^2) \tag{5.55}$$

$$\psi_-(\zeta) \sim \zeta^{\frac{3}{2}+L}e^{-\lambda\zeta^2/2}L_n^{L+1}(\lambda\zeta^2) \tag{5.56}$$

with eigenvalues $M^2 = 4\lambda(n + L + 1)$. The polynomials $L_n^L(x)$ are associated Laguerre polynomials, where the radial quantum number n counts the number of nodes in the wave function. We compare in Fig. 101 the model predictions with the measured values for the positive parity nucleons [513] for $\sqrt{\lambda} = 0.485$ GeV.

5.4.7 Superconformal meson–baryon symmetry

Superconformal quantum mechanics also leads to a connection between mesons and baryons [1015] underlying the $SU(3)_C$ representation properties, since a diquark cluster can be in the same color representation as an antiquark, namely $\bar{3} \in 3 \times 3$. The specific connection follows from the substitution $x \mapsto \zeta$, $E \mapsto M^2$, $\lambda \mapsto \lambda_B = \lambda_M$, $f \mapsto L_M - \frac{1}{2} = L_B + \frac{1}{2}$, $\phi_+ \mapsto \phi_M$ and $\phi_2 \mapsto \phi_B$ in the superconformal equations (5.51) and (5.52). We find the LF meson (M) –

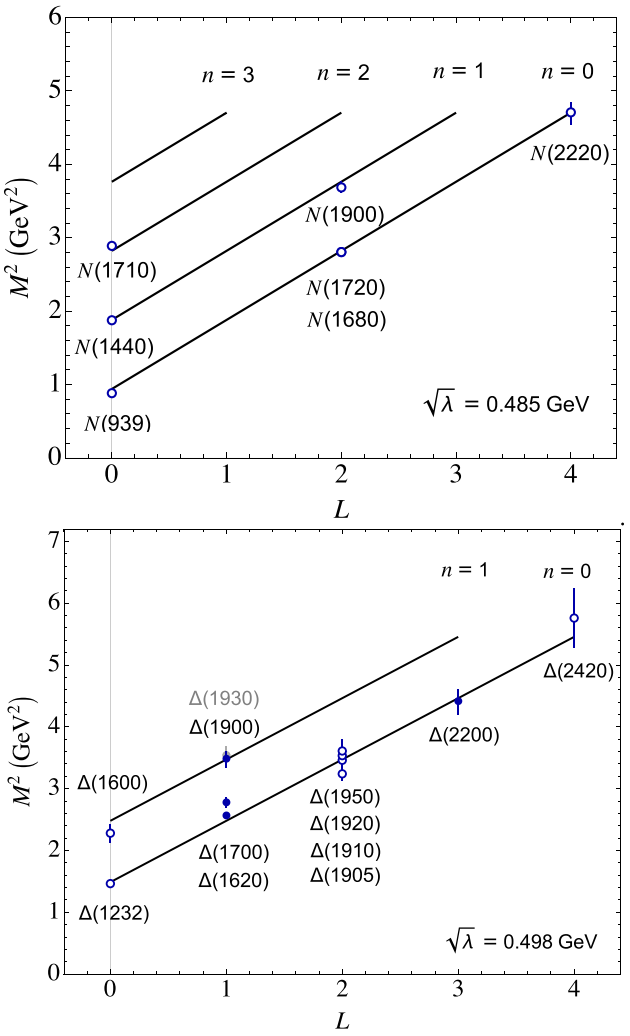


Fig. 101 Model predictions for the orbital and radial positive-parity nucleons (up) and positive and negative parity Δ families (down) compared with the data from Ref. [513]. The values of $\sqrt{\lambda}$ are $\sqrt{\lambda} = 0.485$ GeV for nucleons and $\sqrt{\lambda} = 0.498$ GeV for the deltas

baryon (B) bound-state equations

$$\left(-\frac{d^2}{d\zeta^2} - \frac{1-4L_M^2}{4\zeta^2} + U_M\right)\phi_M = M^2\,\phi_M, \tag{5.57}$$

$$\left(-\frac{d^2}{d\zeta^2} - \frac{1-4L_B^2}{4\zeta^2} + U_B\right)\phi_B = M^2\,\phi_B, \tag{5.58}$$

with the confinement potentials

$$U_M = \lambda_M^2\,\zeta^2 + 2\lambda_M(L_M - 1), \tag{5.59}$$

$$U_B = \lambda_B^2\,\zeta^2 + 2\lambda_B(L_B + 1). \tag{5.60}$$

The superconformal structure imposes the condition $\lambda = \lambda_M = \lambda_B$ and the remarkable relation $L_M = L_B + 1$, where L_M is the LF angular momentum between the quark and antiquark in the meson, and L_B between the active quark and spectator cluster in the baryon. Likewise, the equality of the Regge slopes embodies the equivalence of the $3_C - \bar{3}_C$ color interaction in the $q\bar{q}$ meson with the $3_C - \bar{3}_C$ interaction between the quark and diquark cluster in the baryon. The mass spectrum from (5.57) and (5.58) is

$$M_M^2 = 4\lambda(n + L_M) \text{ and } M_B^2 = 4\lambda(n + L_B + 1). \tag{5.61}$$

The pion has a special role as the unique state of zero mass and, since $L_M = 0$, it does not have a baryon partner.

AdS space is effectively modified in the IR by the dilaton profile in Eq. (5.42), while retaining conformal invariance in the UV (near the boundary of AdS space): It leads to the effective confinement potential $U(z)$ in Eq. (5.45). The dilaton profile can be determined from the superconformal algebra by integrating Eq. (5.45) for the effective potential (5.59). We obtain $\varphi(z) = \lambda z^2$. The dilaton is uniquely determined, provided that it depends only on the modification of AdS space [1039].

5.4.8 Spin interaction and diquark clusters

Embedding the LF bound-state equations in AdS space allows us to extend the superconformal Hamiltonian to include the spin–spin interaction, a problem not defined in the chiral limit by standard procedures. Since the dilaton profile $\varphi(z) = \lambda z^2$ is valid for arbitrary J, it leads to the additional term $2\lambda\mathcal{S}$ in the LF Hamiltonian for mesons and baryons, $G = \frac{1}{2}\{R_\lambda, R_\lambda^\dagger\} + 2\lambda\mathcal{S}$, which maintains the meson–baryon supersymmetry [1040]. The spin $\mathcal{S} = 0, 1$, is the total internal spin of the meson, or the spin of the diquark cluster of the baryon partner. The effect of the spin term is an overall shift of the quadratic mass,

$$M_M^2 = 4\lambda(n + L_M) + 2\lambda\mathcal{S}, \tag{5.62}$$

$$M_B^2 = 4\lambda(n + L_B + 1) + 2\lambda\mathcal{S}, \tag{5.63}$$

as depicted in Fig. 102 for the spectra of the ρ mesons and Δ baryons by shifting one unit the value of L_B [1015]. This shift leads to a degeneracy of meson and baryons states, a property known as the MacDowell symmetry [1041,1042].

For the Δ baryons the total internal spin S is related to the diquark cluster spin $\mathcal{S}$ by $S = \mathcal{S} + \frac{1}{2}(-1)^L$, and therefore, positive and negative Δ baryons have the same diquark spin, $\mathcal{S} = 1$. As a result, all the Δ baryons lie, for a given n, on the same Regge trajectory, as shown in Fig. 101. Plus parity nucleons are assigned $\mathcal{S} = 0$ and are well described by the holographic model as shown in Fig. 101. For negative parity nucleons both $\mathcal{S} = 0$ and $\mathcal{S} = 1$ are possible, but their precise comparison with data is not as successful as for the Δ baryons and positive parity nucleons.

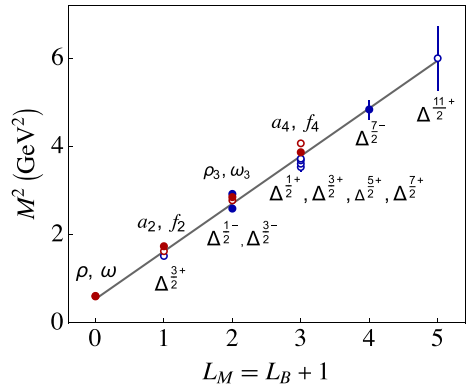


Fig. 102 Supersymmetric vector meson and Δ partners from Ref. [1015]. The experimental values of M^2 from Ref. [513] are plotted vs $L_M = L_B + 1$ for $\sqrt{\lambda} \simeq 0.5$ GeV. The ρ and ω mesons have no baryonic partner, since it would imply a negative value of L_B

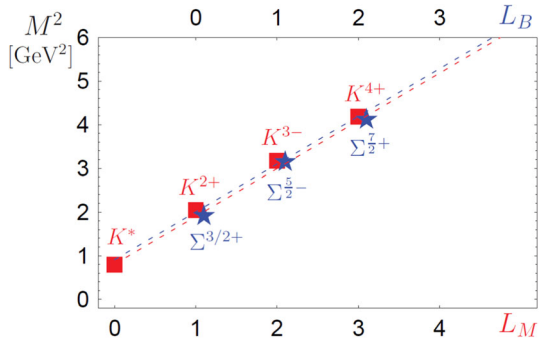


Fig. 103 The K^* and Σ^* trajectories from supersymmetric HLFQCD in Ref. [1044] with $\sqrt{\lambda} = 0.51$ GeV. The error bars are smaller that the symbols in the figure and were not included

5.4.9 Inclusion of quark masses and longitudinal dynamics

Finite quark masses break conformal invariance and pose a special challenge for all AdS/CFT approaches since the dual quantum field theory is inherently conformal. In the usual formulation of bottom-up holographic models one identifies quark mass and chiral condensates as coefficients of a scalar background field $X_0(z)$ in AdS space [1000,1001]. A heuristic way to take into account the occurrence of quark mass terms, is to include the quark mass dependence in the invariant mass squared which controls the off-shell dependence of the LF wave function [1003,1043]. This substitution leads, upon exponentiation, to a natural factorization of the transverse, $\phi(\zeta)$, and the longitudinal, $\chi(x)$, wave functions in (5.40), where $\chi(x) = x^{-1/2}(1-x)^{-1/2}X(x)$. For hadrons with quark masses m_i, one finds for the longitudinal wave functions and the quadratic mass corrections [1003,1040,1043]

$$\chi_{\rm IM}(x) = \mathcal{N} \exp\Big(-\frac{1}{2\lambda}\sum_i \frac{m_i^2}{x_i}\Big), \tag{5.64}$$

$$\Delta M^2 = \int dx\, \delta\Big(\sum_i x_i - 1\Big) \sum_i \frac{m_i^2}{x_i}\, \chi^2_{\rm IM}(x), \tag{5.65}$$

where $\mathcal{N}$ is a normalization factor and (IM) refers to the invariant mass LFWF.

The effective quark masses can be obtained by comparing the holographic results with the observed pseudoscalar masses. One obtains $m_q = 0.046$ GeV for the light quark mass and $m_s = 0.350$ GeV for the strange mass, with values between the Lagrangian and the constituent masses [1003,1040,1043]. The analysis has been consistently applied to the radial and orbital excitation spectra of the light meson and baryon families, giving the value $\sqrt{\lambda} = 0.523 \pm 0.024$ GeV [1040]. The comparison of the predicted K^* and Σ^* trajectories with experiment shown in Fig. 103 is a clear example of the validity of the supersymmetric meson–baryon connection including light quark masses. Starting with Ref. [1045], the application of the light-front holographic wave functions to diffraction physics has also been successful.

For heavy quarks the mass breaking effects are large. The underlying hadronic supersymmetry, however, is still compatible with the holographic approach and gives remarkable connections across the entire spectrum of light and heavy-light hadrons [1039,1044]. In particular, the lowest mass meson of every family has no baryon partner, conforming to the SUSY mechanism. Compatibility with heavy quark symmetry [1039,1044,1046–1050] predicts a dependence of the holographic mass scale λ on the quark mass.[42]

The extension of the LF holographic framework to incorporate longitudinal dynamics and chiral symmetry breaking, inspired in the original work of 't Hooft [1053], has recently attracted much interest [942,981,1051,1054–1064]; however, in contrast with the transverse dynamics, the longitudinal confinement potential is not uniquely determined by the symmetries of the model.

5.4.10 Completing the supersymmetric hadron multiplet

Besides mesons and baryons, the supersymmetric multiplet $\Phi = \{\phi_M, \phi_B^+, \phi_B^-, \phi_T\}$ contains a further bosonic partner, a tetraquark, which, follows from the action of the SUSY operator $R_\lambda^\dagger$ (5.50) on the negative-chirality component of a baryon [1040], as illustrated in Fig. 104. A clear example is the SUSY positive parity J^P multiplet $2^+, \frac{3}{2}^+, 1^+$ of states $f_2(1270)$, $\Delta(1232)$, $a_1(1260)$ where the a_1 is interpreted as a tetraquark.

Unfortunately, it is difficult to disentangle conventional hadronic quark states from exotic ones and, therefore, no clear-cut identification of tetraquarks for light hadrons, or

[42] For a relation with linear confinement see Refs. [1051,1052].

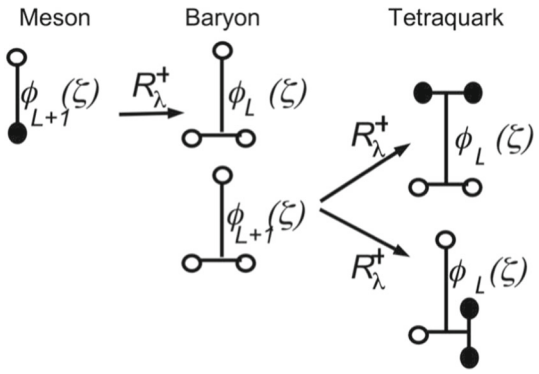


Fig. 104 The meson–baryon–tetraquark supersymmetric 4-plet $\{\phi_M, \phi_B^+, \phi_B^-, \phi_T\}$ follows from the two step action of the supercharge operator $R_\lambda^\dagger$ (5.50): $\bar{3} \to 3 \times 3$ on the pion, followed by $3 \to \bar{3} \times \bar{3}$ on the negative chirality component of the nucleon

Table 5 Predicted masses for double heavy bosons from Ref. [1050]. Exotics which are predicted to be stable under strong interactions are marked by $^{(!)}$

Quark content	J^P	Predicted mass [MeV]	Strong decay	Threshold [MeV]
$cq\overline{cq}$	0^+	3660	$\eta_c\pi\pi$	3270
$cc\overline{qq}^{(!)}$	1^+	3870	D^*D	3880
$bq\overline{bq}$	0^+	10020	$\eta_b\pi\pi$	9680
$bb\overline{qq}^{(!)}$	1^+	10230	B^*B	10800
$bc\overline{qq}^{(!)}$	0^+	6810	BD	7150

hadrons with hidden charm or beauty, is possible [1040, 1049, 1065]. The situation is, however, more favorable for tetraquarks with open charm and beauty which may be stable under strong interactions and therefore easily identified [1066]. In Table 5, the computed masses from Ref. [1050] are presented. Our prediction [1050] for a doubly charmed stable boson T_{cc} with a mass of 3870 MeV (second row) has been observed at LHCb a year later at 3875 MeV [1067], and it is a member of the positive parity J^P multiplet $2^+, \frac{3}{2}^+, 1^+$ of states $\chi_{c2}(3565)$, $\Xi_{cc}(3770)$, $T_{cc}(3875)$. The occurrence of stable doubly beautiful tetraquarks and those with charm and beauty is well established, see also Ref. [1066].

5.4.11 *Holographic QCD and Veneziano amplitudes*

The hadronic mass spectrum (5.61), which follows from the scale deformed superconformal equations (5.51) and (5.52), shows remarkable features which were essential ingredients to the pre-QCD physics of strong interactions. Starting from the S-matrix, Chew and Frautschi [1068] proposed to extend the concept of Regge trajectories [1069], $\alpha(t) = \alpha_0 + \alpha' t$, also to positive t-values. It led to a quadratic mass spectra, linear in the angular momentum, just as the spectra of Eq. (5.61). The analogy goes further: Veneziano [7] constructed a hadronic scattering amplitude

$$A(s,t) \sim B(1-\alpha(s), 1-\alpha(t)), \tag{5.66}$$

based on Euler's Beta function $B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$, which incorporates the duality in strong interactions [1070] and linear Regge trajectories. It is easy to see that this amplitude leads to particle poles at masses exactly matching Eq. (5.61), if one identifies the slope of the trajectory with the scale λ: $\alpha' = 1/4\lambda$. In fact, from the analytic structure of the Beta function, particle poles appear at each value where $\alpha(t)$ is a negative integer. This leads to "Regge-daughter trajectories", which are identified with the radial excitations numbered by the integer n in (5.61). But there is an important difference in the theoretical foundation: in the Veneziano approach, linear trajectories were assumed to exist, whereas here they are a consequence of the model, especially of the superconformal model, where the Regge intercept α_0 is also determined, and expressed in terms of quark masses.

5.4.12 *Electromagnetic form factors in holographic QCD*

Holographic QCD incorporates important elements for the study of hadronic form factors, such as the connection between the twist of the hadron to the fall-off of its current matrix elements for large Q^2, and important aspects of vector meson dominance which are relevant at lower energies. The expression for the electromagnetic (EM) form factor (FF) in AdS space has been given by Polchinski and Strassler [1006]

$$F(Q^2) = \int \frac{dz}{z^3}\, V(Q^2, z)\, \Phi^2(z), \tag{5.67}$$

in their influential article describing deep inelastic scattering (DIS) using the gauge/gravity correspondence.[43] It is written as the overlap of a normalizable mode $\Phi(z)$, representing a bound-state wave function in AdS for the initial and final states, with a non-normalizable solution $V(Q^2, z)$ of the wave equation (5.43) for a spin one conserved current in AdS, with $\mu = 0$ and $M^2 \to -Q^2$. The bulk-to boundary propagator, $V(Q^2, z)$ carries momentum $Q^2 = -t > 0$ from the external EM current. A precise mapping can be carried out to physical spacetime provided that the invariant transverse impact variable ζ for an arbitrary number of quarks is identified with the holographic variable z [1007].

For the soft-wall model (SWM) of Ref. [1002] $\Phi_\tau(z) \sim z^\tau e^{-\lambda z^2/2}$, and $V(Q^2, z)$ is given in terms of the Tricomi function, $V(Q^2, z) \sim U(Q^2/4\lambda, 0, \lambda z^2)$. It corresponds to a conserved vector current with vanishing mass $\mu = 0$ in

[43] For recent DIS studies examining various holographic QCD models see Ref. [1071] and references therein.

AdS. The result for the FF [1003] can be brought into the form of an Euler Beta function

$$F_\tau^{SWM}(t) \sim B(\tau - 1, 1 - t/4\lambda). \tag{5.68}$$

It generates a series of poles located at $M_n^2 = 4\lambda(n+1)$, and thus to the Regge intercept $\alpha_0 = 0$ [1072]. Therefore, one has to perform a pole shift [1003,1019,1020,1073] in the expression (5.68) in order to bring the analytical structure of the FF in accordance with the spectra predicted by HLFQCD, which is in perfect agreement with observations. This shift leads to [1019]

$$F_\tau^{HLF}(t) \sim B\big(\tau - 1,\ 1/2 - t/4\lambda\big), \tag{5.69}$$

for the EM form factors in HLFQCD.

5.4.13 *Form factors in dual models and holographic QCD*

In a model extending the duality concept incorporated in Eq. (5.66) to reactions involving external currents, Ademollo and Del Giudice [1074], and Landshoff and Polkinghorne [1075], proposed a a Veneziano-like amplitude

$$F_\gamma(t) \sim B\big(\gamma, 1 - \alpha_\rho(t)\big), \tag{5.70}$$

in order to describe the electromagnetic FF; here $\alpha_\rho(t)$ is the Regge trajectory of the ρ meson which couples to the quark current in the hadron, and the parameter γ controls the rate of decrease of the FF. In fact, from Stirling's formula we find the asymptotic behavior $F_\gamma(Q^2) \sim \big(1/Q^2\big)^\gamma$ for large $Q^2 = -t$.

In LF QCD the parameter γ has a well defined interpretation. To see this, we compare the asymptotic expression for $F_\gamma(Q^2)$ with the result from hard scattering counting rules at large Q^2 [153], $F_\tau(Q^2) \sim \big(1/Q^2\big)^{\tau-1}$, where the twist τ is the number of constituents N in a given Fock component of the hadron. Thus, one has to choose in Eq. (5.70) $\gamma = \tau - 1$, in order to incorporate the scaling counting rule. This brings us to our final result for the analytical expression of the electromagnetic FF in the extended duality model [1019]

$$F_\tau(t) = \frac{1}{N_\tau} B\left(\tau - 1, 1 - \alpha(t)\right), \tag{5.71}$$

with $N_\tau = B(\tau - 1, 1 - \alpha(0))$, a remarkable expression which incorporates, at tree level, both the nonperturbative pole structure of the form factor and the hard scattering behavior.

For $\tau = N$, the number of constituents, the FF (5.71) is an $N - 1$ product of poles located at [1003]

$$-Q^2 = t = M_n^2 = \frac{1}{\alpha'}\big(n + 1 - \alpha(0)\big) > 0. \tag{5.72}$$

It generates the radial excitation spectrum of the exchanged vector mesons in the t-channel. For example, the ρ trajectory has Regge intercept $\alpha_0 = 1/2$ and slope $\alpha' \equiv 1/4\lambda$, with $\lambda \simeq (0.5\ \text{GeV})^2$. Thus $M_n^2 = 4\lambda(n + \frac{1}{2})$, corresponding to the ρ vector meson and its radial excitations for $n = 0, 1, 2, \ldots, \tau - 2$ in agreement with Eq. (5.62). In general, the hadron wave function is a superposition of an infinite number of Fock components, and thus the full form factor should be written as a superposition $F(Q^2) = \sum_\tau C_\tau F_\tau(Q^2)$, with $\sum_\tau C_\tau = 1$, if all possible states are included. In practice, one expects a rapid convergence in the number of poles, with a dominant contribution from the ρ vector meson plus contributions from the higher resonances ρ', ρ'', …, etc.

As a simple example, consider the valence contribution to the nucleon EM (spin non-flip) Dirac form factors by writing the flavor FFs as

$$F^u(t) = \left(2 - \frac{r}{3}\right) F_3(t) + \frac{r}{3} F_4(t), \tag{5.73}$$

$$F^d(t) = \left(1 - \frac{2r}{3}\right) F_3(t) + \frac{2r}{3} F_4(t), \tag{5.74}$$

where $F_\tau(Q^2)$ is given by (5.71). The holographic constraint of equal probability for nucleon states with LF orbital angular momentum L and $L+1$ (Sect. 5.4.4) determines the value $r = 3/2$, since the probability of the total quark spin along the plus z-direction for $L = 0$ (twist 3) should be identical to the probability of having total quark spin along the minus z-direction for $L = 1$. Actually, the values found in the recent analysis in Ref. [1020] deviate by $10 \sim 15$ % for the u-flavor FF and remain almost identical for the d quark in the valence approximation. This leads to the results show in Fig. 105 for the nucleon isospin FF combination, $F^{I=0,1} = F_p(t) \pm F_n(t)$, where we compare the model predictions with the analysis of Ye *et al.* [1076]. Detailed studies show the importance of higher (large distance meson cloud) Fock components for the spin-flip Pauli FF [1073].

5.4.14 *Quark distribution functions and the exclusive–inclusive connection*

The mathematical structure of the Veneziano-type FFs (5.71), not only incorporates the hard scattering amplitude's dependence on twist, but it also gives important insights into the structure of the parton distributions since it becomes possible to include the Regge behavior at small values of x, as well as the exclusive–inclusive connection [958,1078] at large values of the longitudinal momentum x [1019]. In fact, the relation between the behavior of the structure function near $x = 1$ with the falloff of the FF at large t, described in the article of Landshoff and Polkinghorne [1075], is very close to the Drell–Yan "exclusive–inclusive" connection, also formulated in 1970 [958].

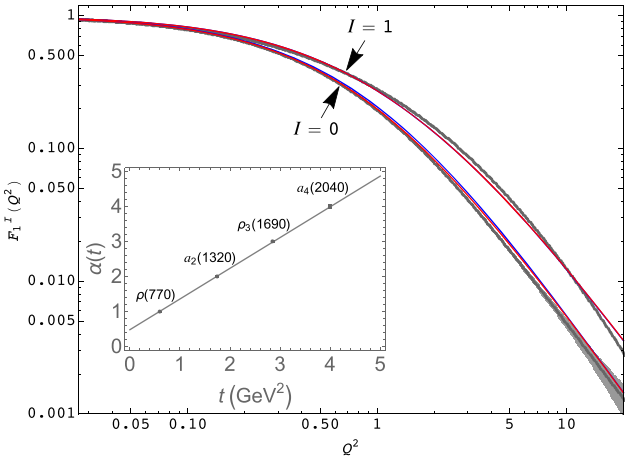


Fig. 105 The LFHQCD prediction for the $I = 0, 1$ isospin combinations of nucleon factors from Ref. [1020] is compared with the z-expansion data analysis of Ye et al. [1076] (grey band): (blue line) valence contribution only, (red line) including $u\bar{u}$ and $d\bar{d}$ pairs. The inset from Ref. [1077] represents the ρ Regge trajectory in Eq. (5.71) for $\sqrt{\lambda} = 0.534$ GeV and $\alpha_\rho(0) = 0.483$

Using the integral representation of the Beta function, the FF (5.71) can be expressed in a reparametrization invariant form

$$F_\tau(t) = \frac{1}{N_\tau}\int_0^1 dx\, w'(x) w(x)^{-\alpha(t)}\,[1-w(x)]^{\tau-2}\,. \tag{5.75}$$

The trajectory $\alpha(t)$ of the vector current can be computed within the superconformal LF holographic framework, and the intercept, $\alpha(0)$, incorporates the quark masses [1014,1015]. The function $w(x)$ is a flavor-independent function with $w(0) = 0$, $w(1) = 1$ and $w'(x) \geq 0$.

The flavor FF can be written in terms of its generalized parton distribution (GPD) [1079–1081], $H^q(x,t) \equiv H^q(x, \xi = 0, t)$, at zero skewness, ξ,

$$\begin{aligned} F^q(t) &= \int_0^1 dx\, H^q(x,t) \\ &= \int_0^1 dx\, q(x) \exp\left[t f(x)\right], \end{aligned} \tag{5.76}$$

with the profile function, $f(x)$, and the particle distribution function (PDF), $q_\tau(x)$, both determined by $w(x)$:

$$f(x) = \frac{1}{4\lambda}\log\left(\frac{1}{w(x)}\right), \tag{5.77}$$

$$q_\tau(x) = \frac{1}{N_\tau} w'(x) w(x)^{-\alpha(0)} [1-w(x)]^{\tau-2}, \tag{5.78}$$

with $\alpha' = 1/4\lambda$. Boundary conditions follow from the Regge behavior at $x \to 0$, $w(x) \sim x$, and at $x \to 1$ from the exclusive–inclusive counting rule [958,1078], $q_\tau(x) \sim (1-x)^{2\tau-3}$, which fixes $w'(1) = 0$. A simple ansatz for $w(x)$, $w(x) = x^{1-x}\,\exp(-a(1-x)^2)$, fulfills all conditions mentioned above. The flavor independent parameter a has the value $a \simeq 0.5$ [1019].

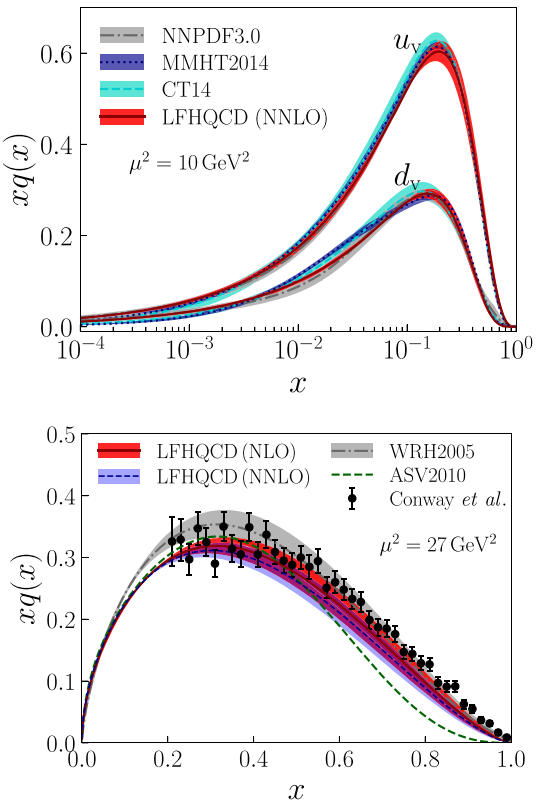


Fig. 106 Comparison for $xq(x)$ in HLFQCD with global fits from [1019]. Up: proton valence approximation (red band). Data analysis from MMHT2014 (blue bands) [973], CT14 [1082] (cyan bands), and NNPDF3.0 (grey bands) [1083]. Down: pion results (red and light blue bands). NLO global fits from [1084,1085] (gray band and green curve) and the LO data extraction [1086]. HLFQCD results are evolved from the initial scale $\mu_0 \simeq 1$ GeV at NLO and NNLO

Using the expression (5.76) at $t = 0$ and Eqs. (5.73–5.74), we obtain for the unpolarized quark distributions in the valence approximation

$$u_v(x) = \left(2 - \frac{r}{3}\right) q_3(x) + \frac{r}{3} q_4(x), \tag{5.79}$$

$$d_v(x) = \left(1 - \frac{2r}{3}\right) q_3(x) + \frac{2r}{3} q_4(x), \tag{5.80}$$

with normalization $\int dx\, u_v(x) = 2$ and $\int dx\, d_v(x) = 1$. The PDF $q_\tau(x)$ is given by (5.78) and $r = 3/2$. Our PDF results for the nucleon, Eqs. (5.79–5.80), and for the pion [1019], are compared with the global data analysis in Fig. 106. If the reparametrization function $w(x)$ is fixed by the nucleon PDFs, then the pion PDF is a prediction. pQCD evolution is performed from an initial scale determined from $\mu_0 \simeq 1$ GeV

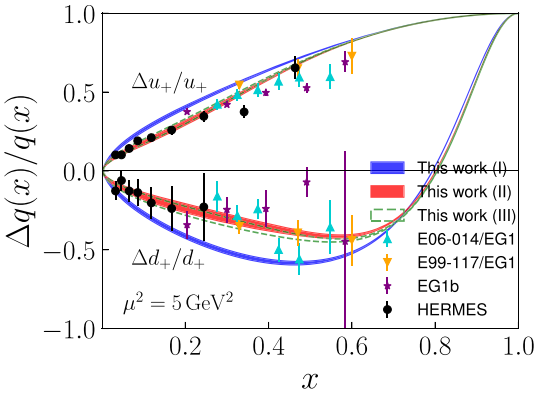


Fig. 107 HLFQCD predictions from Ref. [1020] for the quark helicity asymmetry ratio $\Delta q_+/q_+$, $q_+ = q + \bar{q}$, are compared with existing data. The blue band is the valence contribution, the red band includes $q\bar{q}$ components and the dashed green band also includes the intrinsic sea contribution

from the soft-hard matching procedure described in Ref. [1087]. Our result for the pion PDF in Fig. 106 is in good agreement with the data analysis in Ref. [1084] and consistent with the nucleon global fit through the GPD universality introduced in [1019]. It leads to a $1-x$ falloff, in contrast with the $(1-x)^2$ pQCD result at large-x [1085,1088], an issue much debated recently [837,1089,1090].

An analysis of the polarized quark distribution in the proton has been performed in Ref. [1020], assuming the Veneziano-type FF (5.71), with the separation of chiralities from the axial current. The model predictions for the ratio of polarized to unpolarized quark distribution functions is compared with available data in Fig. 107.

Another application of the LF holographic ideas is the computation of the intrinsic charm-anticharm asymmetry in the proton [1091], $c(x) - \bar{c}(x) = \sum_\tau c_\tau \big(q_\tau(x) - q_{\tau+1}(x)\big)$, with $\int_0^1 dx[c(x)-\bar{c}(x)] = 0$. The normalization of the charm form factor was computed using lattice QCD [1091], and the J/ψ trajectory in the GPDs from HLFQCD and heavy quark effective theory [1049]. A similar procedure was used to determine the intrinsic strange–antistrange asymmetry in the proton with the Regge trajectory in the holographic expressions corresponding to the ϕ meson current [1077], and most recently to study color transparency in nuclei [1092] (see Sect. 5.9), and to model the EMC effect in various nuclei [1093].

5.4.15 *Gravitational form factors, gluon distributions and the Pomeron trajectory*

Gravitational form factors (GFFs) are the hadronic matrix elements of the energy momentum tensor (EMT) and describe the coupling of a hadron to the graviton, thus providing key information on the dynamics of quarks and gluons within hadrons. In holographic QCD Pomeron exchange is identified as the graviton of the dual AdS theory [1094–1099]. The Pomeron couples as a rank-two tensor to hadrons and interacts strongly with gluons. Since we are interested in obtaining the intrinsic gluon distribution in the nucleon, we use the soft Pomeron of Donnachie and Landshoff [1100] with the Regge trajectory $\alpha_P(t) = \alpha_P(0) + \alpha'_P t$, with intercept $\alpha_P(0) \simeq 1.08$ and slope $\alpha'_P \simeq 0.25\,\mathrm{GeV}^{-2}$ [513].

To actually compute the GFF one considers the perturbation of the gravity action by an arbitrary external source at the AdS asymptotic boundary which propagates inside AdS space and couples to the EMT [1009,1101]. In analogy to the EM FF (5.67), the spin non-flip GFF $A(t)$ is written as the overlap of a normalizable mode $\Phi(z)$, representing a bound-state wave function, with a non-normalizable mode $H(Q^2, z)$, the bulk-to-boundary propagator, corresponding to the gravitational current in AdS. We obtain [1009,1101]

$$A(t) = \int \frac{dz}{z^3}\, H(Q^2, z)\, \Phi^2(z). \tag{5.81}$$

For the soft-wall profile introduced in Ref. [1002], the propagator in AdS, $H(Q^2, z)$, is also given by a Tricomi function [1003,1101], $H(Q^2, z) \sim U\left(Q^2/4\lambda_g, -1, \lambda_g z^2\right)$. The effective physical scale λ_g is the scale of the Pomeron, $\lambda_g = 1/4\alpha'_P \simeq 1\,\mathrm{GeV}^2$, which couples to the constituent gluon over a distance $\sqrt{\alpha'_P} \sim 1/\sqrt{4\lambda_g}$, described by the wave function $\Phi^g_\tau(z) \sim z^\tau e^{-\lambda_g z^2/2}$. Our final result is [1021]

$$A^g_\tau(Q^2) = \frac{1}{N_\tau} B\left(\tau - 1, 2 - \alpha_P(Q^2)\right), \tag{5.82}$$

with $N_\tau = B\left(\tau - 1, 2 - \alpha_P(0)\right)$. As for the EM FF, in writing (5.82) we have also shifted the Pomeron intercept to its physical value $\alpha_P(0) \approx 1$, since the holographic result (5.81) leads to a zero intercept. For integer twist, the GFF (5.82) is expressed as a product of $\tau - 1$ timelike poles located at

$$-Q^2 = M_n^2 = \frac{1}{\alpha'_P}\left(n + 2 - \alpha_P(0)\right), \tag{5.83}$$

the radial excitation spectrum of the spin-two Pomeron. The lowest state in this trajectory, the 2^{++}, has the mass $M \simeq 1.92\,\mathrm{GeV}$, compared with the lattice results $M \simeq (2.15-2.4)$ GeV [513].[44] We notice that Eq. (5.82) is the Veneziano amplitude of the FF for a spin-two current [1074,1075].

The lowest twist contributions to the GFF corresponds to the $\tau = 4$ Fock state $|uudg\rangle$ in the proton and the $\tau = 3$ component $|u\bar{d}g\rangle$ in the pion, both containing an intrinsic gluon. The results for $A^g(t)$ are compared in Fig. 108 with recent

[44] There exist many computations of glueballs in top-down holographic models, see for example, [1102]; and also in bottom-up models starting from [1103]. For a recent computation, see for example [1104], and references therein.

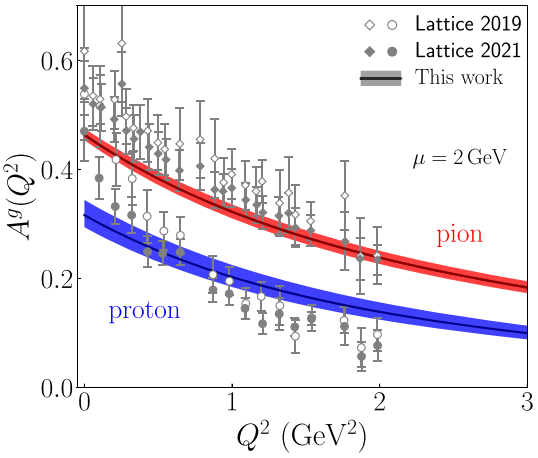


Fig. 108 Gluon gravitational form factor $A^g(Q^2)$ of the proton (blue) and the pion (red) in comparison with lattice QCD computations [1105, 1106]. The value $A^g(0)$ corresponds to the momentum fraction carried by gluons at the scale $\mu = 2\,\text{GeV}$. The bands indicate the uncertainty of λ_g by ±5% and the normalization from the momentum sum rule

lattice computations. We find for the gluon mass squared radius $\langle r_g^2\rangle_p = 2.93/\lambda_g = (0.34\,\text{fm})^2$ for the proton and $\langle r_g^2\rangle_\pi = 2.41/\lambda_g = (0.31\,\text{fm})^2$ for the pion. The model predictions in Fig. 108 have no free parameters [1021].

The intrinsic gluon distributions in the proton and the pion can be determined from the gravitational form factor (5.82) following the same procedure used in Sect. 5.4.14. The results are given in [1021] and agree very well with the data analysis from [663,664,971,1107,1108]. The model uncertainties for large x-values are smaller than those from the phenomenological analysis.

By using the gauge/gravity duality a simultaneous description of the BFKL hard Pomeron [159,236,1109] and the soft Regge domain has been proposed in Ref. [1094]. This model, however, did not solve the problem of the large difference of intercept values between both Pomerons. Using the scale dependence of the gluon distribution functions, our results give strong support to a single Pomeron with a scale dependent intercept [1110], which was proposed in Ref. [1111] in order to explain the diffractive scattering data at LHC energies [1112,1113].

5.4.16 Summary

Holographic light front QCD is a nonperturbative analytic approach to hadron physics. It originates from the precise mapping of light front expressions of form factors in AdS space for an arbitrary number of partons [1007]. The holographic embedding in AdS also leads to semiclassical relativistic wave equations, similar to the Schrödinger equation in atomic physics, for arbitrary integer or half-integer spin [1004,1032]. The model embodies an underlying superconformal algebraic structure from $SU(3)_C$ color symmetry: It is responsible for the introduction of a mass scale within the superconformal group, and determines the effective confinement potential–It is not supersymmetric QCD, a theory which includes squarks and gluinos, but an effective hadronic supersymmetry. There is a zero eigenmode which is identified with the pion: It is massless in the chiral limit. The new framework leads to relations between the Regge trajectories of mesons, baryons, and tetraquarks. It also incorporates features of pre QCD, such as Veneziano model and Regge theory. Further extensions incorporate the exclusive–inclusive connection in QCD and provide nontrivial relations between hadron form factors and quark and gluon distributions. Measurements of the strong coupling in the nonperturbative domain [1114] are remarkably consistent with the predicted form in holographic QCD [177], a relevant issue in QCD which is discussed in the next Sect. 5.5. Holographic light front QCD has led to significant advances in understanding hadron phenomena by incorporating emerging QCD properties in an effective computational framework of hadron structure.

5.5 The nonperturbative strong coupling

Alexandre Deur

The perturbative framework of QCD (pQCD) has been remarkably successful in describing the interactions between the fundamental constituents of hadrons in high energy experiments, thus establishing QCD as the theory of the strong force at small distances [300]. Most of nature's strong force phenomena, however, are governed by large-distance, nonperturbative physics [783,1115–1119] where the methods of pQCD are not applicable. The Landau pole at low-energies in the running of the QCD coupling is an example of the expected failure of perturbation theory as the coupling increases. A nonperturbative treatment is necessary and allows us to define renormalization scheme dependent coupling constants.

Studying $\alpha_s(\mu)$ at low energy has been challenging: not only do nonperturbative calculations represent a difficult problem to solve, but more generally, we only know in the pQCD framework how to relate the α_s calculated in different schemes. Worst, there is no obvious prescription of how to define the coupling. One reason why a variety of definitions is possible is that $\alpha_s(\mu)$ need not be an observable. In fact, in most approaches – including the standard pQCD treatment – it is not an observable. For example α_s^{pQCD} depends on the choice of renormalization scheme, generally taken to be $\overline{\text{MS}}$. Such arbitrary dependence on a human convention shows that $\alpha_s(\mu)$ is not an observable. In addition, the quark–gluon, 3-gluon, 4-gluon or ghost-gluon vertices may have different couplings,[45] i.e., several couplings, with distinct magni-

[45] When needed, we will use superscripts to qualify the coupling. For examples, $\alpha_s^{\text{pQCD},\overline{\text{MS}}}$ is the perturbative coupling in the $\overline{\text{MS}}$ scheme, or

tudes as well as differing μ-dependence, may be necessary to characterize QCD. This happens because the Slavnov–Taylor Identities (STI) [1120,1121], the QCD equivalent of QED's Ward-Takahashi relation [1122,1123], may not hold under certain choices of gauges and renormalization schemes, such as the MOM scheme. With the oft-used $\overline{\text{MS}}$ scheme, the STI hold, viz,

$$\alpha_s^{\text{qg},\overline{\text{MS}}} = \alpha_s^{\text{3g},\overline{\text{MS}}} = \alpha_s^{\text{4g},\overline{\text{MS}}} = \alpha_s^{\text{gh},\overline{\text{MS}}}$$

but $\overline{\text{MS}}$ is not practical for nonperturbative methods such as Lattice QCD and in the nonperturbative domain, the difference between

$$\alpha_s^{\text{qg,MOM}}, \alpha_s^{\text{3g,MOM}}, \alpha_s^{\text{4g,MOM}}, \alpha_s^{\text{gh,MOM}}$$

is conspicuous. A wholly different approach is to define $\alpha_s(\mu)$ to be an observable [165], in analogy with the observable QED coupling α [1124], but while this circumvents the issues of breaking the STI and of scheme and gauge dependence, the prescription is rarely used in pQCD.

Many definitions of α_s have been considered, resulting in a range of values of $\alpha_s(\mu \ll \Lambda_s)$ from 0 to ∞, generating much confusion. Adding to this is the fact that, unlike the high-energy domain where pQCD rules, there is no obviously superior method to study the nonperturbative behavior of $\alpha_s(\mu)$. This is, of course, due to the challenge of solving QCD nonperturbatively. All major non-perturbative approaches, See Sects. 4, 5.2, 5.4 have been used (with the conspicuous exception of chiral effective field theory, Sect. 6.2, since its hadronic degrees of freedom do not couple with α_s) as well as many models. These methods using different type of approximations, and the models being not directly based on QCD's Lagrangian or its symmetries, results have often differed. Yet a number of studies have converged toward a fruitful definition of $\alpha_s(\mu)$ which allows us to account for low energy phenomenology [1125,1126]. Before describing it, we will first recall in broad brushstrokes the history of this endeavor, referring only to pioneering attempts and not the important body of subsequent works that clarified and refined these attempts.

Soon after the advent of QCD, it was realized that $\alpha_s(\mu)$ may display a plateau when $\mu \to 0$ (it is said to *freeze* at low μ) [1127–1129], viz, the β function of QCD, Eq. (1.21) may obey $\beta(\mu \to 0) \to 0$. The actual freezing value $\alpha_s(0)$ was debated and ranged from typically 0.5 to 5 [1125]. A pioneering and influential work in this context is due to Cornwall [1130] who used the Dyson–Schwinger equations (DSE), the gluon self-energy and initiated a method (the *Pinch* technique, PT) that allows to obtain gauge-independent results. The ensuing coupling $\alpha_s^{\text{gse,PT}}$ displays a freezing behavior in qualitative agreement with quark models (e.g., Ref. [770] and Sect. 5.1) and quarkonium spectrum models, e.g., Ref. [99].

A freezing of $\alpha_s(\mu)$ was by no mean the only proposal: others reasoned that it should diverge as $1/\mu^2$ [1131], that it should monotonically increase with $1/\mu$, but without diverging [1132], or that it should vanish as $\mu \to 0$ [172,1133]. In all these cases, $\beta(\mu \to 0) \neq 0$. As we alluded to, multiple reasons caused these widely varying expectations [1125]: differences in the basic definition of $\alpha_s(\mu)$; choice of vertex used to compute it; calculation artifacts from approximations (e.g., discretization in lattice QCD or truncation prescription for the DSE and other functional methods); choice of gauge and renormalization scheme;[46] or the existence of multiple solutions to the QCD equations providing $\alpha_s(\mu)$ without a decisive argument on which one is realized in nature. A prominent example is the *decoupling* [1133] and the *scaling* [1137] solutions that yield a vanishing or a freezing $\alpha_s^{\text{gh,MOM}}$, respectively. Functional methods and lattice QCD have produced both solutions, albeit well-controlled lattice calculations appear to yield only the decoupling solution. In these calculations α_s^{gh}, called the *Taylor coupling* [1120], is most often used because it is the simplest coupling that can be computed from QCD correlation functions.

It is generally believed, after much discussion, that the decoupling solution is the one realized in nature, which simply means that in the particular gauges where ghost fields appear, the gluon and ghost fields decouple at low μ. This is an important finding regarding the behavior of gluons and ghosts but it does not directly illuminate the strength of the strong force at low energy. Besides using correlation functions, other prevalent approaches to define $\alpha_s(\mu)$ are effective charges [165] and analytic approaches [179,1132] – both methods promote $\alpha_s(\mu)$ to an observable quantity – or direct use of phenomenology, for example, using constituent quark models, the $Q\overline{Q}$ potential or the hadronic spectrum [99,770,1131,1138–1141]. Like the DSE, that must choose a truncation prescription, or like lattice QCD with space-time discretization, other methods also use approximations

$\alpha_s^{\text{qg,MOM}}$, $\alpha_s^{\text{3g,MOM}}$, $\alpha_s^{\text{4g,MOM}}$ and $\alpha_s^{\text{gh,MOM}}$ are the couplings for the quark–gluon, 3-gluon, 4-gluon or ghost-gluon vertices, respectively, computed in the MOM scheme.

[46] Methods which optimize the perturbative series by removing renormalization scale ambiguities have rendered this issue negligible. A first-principle method, the BLM procedure [1134], follows by observing that in QED, only the vacuum polarization contributions to the photon propagator cause the coupling to run. Analogously for any pQCD series for an observable, the BLM method absorbs all β-terms in the pQCD series into the QCD running coupling; the resulting series coefficients match the corresponding "conformal" series with $\beta = 0$. The resulting scale-fixed series is free of the renormalon divergence (Sects. 2.3.7 and 5.7.5) and are scheme invariant. The different Q^2 scales for α_s that appear at each order of the series characterize the virtuality of the propagators in the amplitude, as in QED. In fact, the BLM method reduces to Gell-Mann-Low scale setting in the Abelian limit $N_C \to 0$. An analogous procedure applies to the running quark mass. The BLM method is systematically extended to NLO using the Principle of Maximum Conformality [1135]. See reviews [1087,1126,1136].

or/and include model-dependencies. While the systematic effects arising from the approximations or modeling are typically not well controlled, the spread of results arising from methods with very distinct approximations allows for a better understanding of the methods' uncertainties.

After many studies and developments, of which the aforesaid narrative is too a laconic cartoon, a coupling was identified and computed using a formalism guarantying that the STI hold in the nonperturbative domain [1142]. Therefore, QCD is here characterized by a single coupling, independent of the choice of vertex or process used to define it (process-independent, PI). In addition, the Pinch technique [1130] is used to guaranty gauge-independence. The calculation, using either the DSE or lattice QCD results on correlation functions, yields a coupling $\alpha_s^{\rm PI,Pinch}$ in agreement with the phenomenological coupling [1114,1144,1145] derived from the Bjorken sum rule [23] using the effective charge (EC) method [165], $\alpha_s^{\rm EC,g_1}$, and with $\alpha_s^{\rm AdS/QCD}$ obtained using AdS/QCD [177,1087,1126], See Sect. 5.4.

The latter is derived starting from the observation that for strongly coupled systems with a gravity dual, the radial direction z in the bulk can be associated with the energy scale of the boundary theory [1148]: Large values of μ correspond to small values of z near the high-momentum conformal boundary of AdS, $\mu \sim 1/z$. Conversely, large-z distances in the low-momentum region of AdS correspond to low energy scales in the physical theory. The dilaton factor $\exp(\varphi(z))$ is a measure of the departure from conformality at the asymptotic AdS boundary, $z \to 0$, and should grow for large values of z, signaling confinement: It acts as an effective coupling in AdS space. We can use the procedure introduced in [177] to obtain the μ dependence of $\alpha_s^{\rm AdS/QCD}$ from the Hankel transform of the dilaton factor [177]

$$\alpha_s^{\rm AdS/QCD}(\mu) \sim \int_0^\infty z dz J_0(z\mu) e^{-\lambda z^2} \sim e^{-\mu^2/4\lambda}, \qquad (5.84)$$

where the overall normalization is not provided within AdS/QCD. The freezing value of the effective coupling $\alpha_s^{\rm EC,g_1}(0) = \pi$ is used. The dilaton profile λz^2 is determined by the superconformal structure (Sect. 5.4.7). The transition between the predicted Gaussian form (5.84) and the log behavior expected from pQCD is determined from the matching of the perturbative and nonperturbative couplings and their derivatives for $\sqrt{\lambda} = 0.534$ GeV. The specific matching allows us to determine the perturbative QCD scale Λ in terms of the hadronic mass scale $\sqrt{\lambda}$ [1149] for any choice of renormalization scheme, including the $\overline{MS}$ scheme [1087].

The couplings $\alpha_s^{\rm AdS/QCD}$, $\alpha_s^{\rm PI,Pinch}$ and $\alpha_s^{\rm EC,g_1}$ are shown in Fig. 109. When compared in the same renormalization scheme, they agree reasonably well with earlier determinations, such as $\alpha_s^{\rm gse,PT}$ or that of the Godfrey–Isgur quark model [770], see Sect. 5.1 and Refs. [1087,1125]. The couplings in Fig. 109 are in close agreement and have been used in the derivation of many crucial nonperturbative quantities, including the QCD scale $\Lambda_s^{\overline{\rm MS}}$ [1150], as well as elastic and transition form factors [1151–1153], parton distributions (including generalized ones) [836,1019,1154–1157], the hadron mass spectrum [1149,1158], or the pion decay constant [1158].

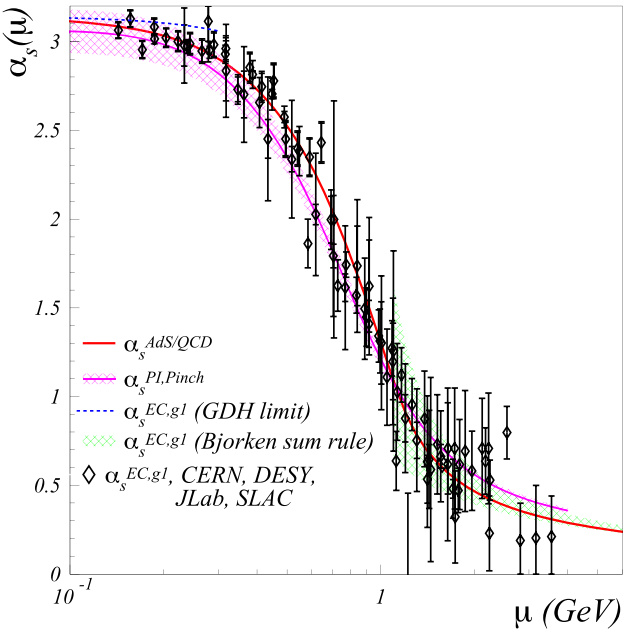


Fig. 109 Nonperturbative strong couplings calculated with the holographic QCD framework ($\alpha_s^{\rm AdS/QCD}$, red line) [177], and Dyson–Schwinger formalism using the lattice determinations of correlations functions ($\alpha_s^{\rm PI,Pinch}$, magenta band) [1142,1143]. The experimental extractions of $\alpha_s^{\rm EC,g_1}$ [1114,1144,1145] following the effective charge definition [165] are shown by the symbols. The green band and the dashed line is $\alpha_s^{\rm EC,g_1}$ deduced from the Bjorken [23] and Gerasimov–Drell–Hearn [1146,1147] sum rules, respectively

In summary, several definitions of the strong coupling in the nonperturbative domain are possible. Most are scheme and gauge dependent. They tend to vanish as $\mu \to 0$ in a non-freezing behavior, viz the QCD β-function itself does not vanish. This informs us on how quark, gluon and ghost fields interact at low energy in the chosen scheme, but does not directly provide a universal parameter reflecting QCD's strength. In contrast, a set of calculations [177,1142] and phenomenological extractions [1114,1144,1145] based on the effective charge prescription [165], following that of QED [1124], provide observable couplings that agree which each other. The consistency of these various approaches in determining a single coupling

$$\alpha_s^{\rm SE,g_1} \simeq \alpha_s^{\rm AdS/QCD} \simeq \alpha_s^{\rm PI,Pinch}$$

and its success in computing a wide range of nonperturbative quantities suggest that a compelling candidate for a

canonical nonperturbative QCD coupling has been identified. It freezes at low energy, a satisfactory behavior since in the nonperturbative domain, the coupling should be finite and non-vanishing, determined by the physics of color confinement, and its scale parameter should be set by a typical hadronic mass, e.g., that of the nucleon. An infrared fixed point is in fact a natural consequence of color confinement: since the propagators of the colored fields have a maximum wavelength, all loop integrals in the computation of the gluon self-energy should decouple at $Q^2 \to 0$ [1159].

5.6 The 't Hooft model and large N QCD

Tom Cohen

In 1973 the QCD Lagrangian was first written down [55]. In the same year, the one-loop function was calculated [53,54, 1160,1161] indicating that that the theory was asymptotically free, but also implying that the coupling constant grew at low momenta. This meant that perturbation theory in the coupling à la QED is inapplicable for low momentum observables such hadron masses, charge radii and the like. The following year 't Hooft [1162] proposed an entirely new expansion for the theory – an expansion in $1/N_c$ where N_c is the number of colors – which, it was hoped, would allow for a systematic computation of these observables.

While the dream of using the $1/N_c$ expansion to compute these quantities directly for QCD in 3+1 dimensions has been elusive, the $1/N_c$ expansion and the associated large N_c limit have played a significant role during the past half century in at least three ways: they have provided a tool for the theoretical exploration of models beyond QCD, including most famously, the AdS/CFT connection [993,996] for $\mathcal{N} = 4$ super Yang–Mills; they have provided a qualitative and occasionally semi-quantitative tool to understand a significant amount of phenomenology (for example in Ref. [1163]); and, they have provided an organizing principle for deciding which terms should be large in phenomenological models or effective field theory treatments (for example in Ref. [1164]).

The underlying idea of the $1/N_c$ expansion is that three is sufficiently large so that a multicolored world with arbitrarily many colors is sufficiently close to the physical world – at least for some observables of interest – that the $N_c \to \infty$ world is a good starting point for an expansion and that systematic $1/N_c$ corrections are controllable. This section will provide an elementary introduction to the large N_c limit and $1/N_c$ expansion with an emphasis on the underlying foundational ideas of the subject. An excellent review of these foundational ideas can be found in Sidney Coleman's Erice lectures [1165]; a more modern review of the large N_c limit and $1/N_c$ expansion for field theories with an emphasis on lattice results can be found in Ref. [1166], while a review of large N_c baryon spectroscopy can be found in Ref. [1167].

5.6.1 Large N_c scaling

The keys to 't Hooft's analysis [1162] are two related insights. The first is that a smooth large N_c limit depends on the QCD coupling, g, scaling with N_c as

$$g^2 = \lambda / N_c \tag{5.85}$$

where λ is independent of N_c. Superficially, this might seem like a weak coupling limit that justifies standard perturbation theory. However, it does not: color factors in loops grow with N_c and can compensate for the small coupling. The second key insight was related to the color factors in loops. 't Hooft developed a clever double line notation for gluons that allows one to easily analyze the scaling behavior of Feynman diagrams. The notation exploits the fact that gluons are in the adjoint representation: they are associated with color matrices with two indices, one carrying a fundamental color and the other an anti-fundamental color. Thus if one ignores the fact that the matrices are traceless (a $1/N_c^2$ effect), the color carried by a gluon propagator is identical to that of a quark line side-by side with an anti-quark line. For the purposes of counting color factors at leading order in $1/N_c$ – and for that purpose only – it is legitimate to replace gluon propagators in Feynman diagrams with parallel quark–antiquark lines. A closed loop of fundamental or antifundamental color in a diagram corresponds to one factor of N_c since there are N_c fundamental colors.

Armed with this, it is straightforward to deduce the following asymptotic scaling behavior for connected diagrams with no external lines:

- Planar connected diagrams of gluons (diagrams in which, except at vertices gluon lines, do not cross when written in a plane) with no external lines grow asymptotically with N_c as N_c^2.
- A diagram containing a non-planar gluon line reduces the asymptotic N_c scaling of a planar diagram by a factor of N_c^{-2}. Multiple non-planar gluons reduce the N_c counting by a factor of at least N_c^{-2} per non-planar gluon.
- A planar diagram that contains quark loops that form the boundary of the diagram, reduces the asymptotic N_c scaling by a factor of N_c^{-1} per quark loop relative to a purely gluonic diagram. Quark loops that cannot be written in this form reduce the N_c scaling by larger amounts.

Note that planar diagrams containing gluons can still be very complicated and can contain arbitrarily many gluon propagators. The fact that planar diagrams of gluons generically scale as N_c^2 can be understood in the following way: a closed loop consisting of a single gluon line scales as N_c^2: in double line form, it has two loops. Any planar diagram of gluons can be constructed starting from this single gluon loop: simply add planar gluons to it one-by-one until one

has the diagram of interest. It is easy to see that any planar gluon added to a previous planar diagram in this construction adds one additional color loop (for a factor of N_c) but also two factors of the coupling constant g at the vertices where the new gluons couple to the old diagram; since $g^2 \sim 1/N_c$ this cancels the additional color loop factor preserving the asymptotic scaling as N_c^2. By inductive reasoning, it is clear that all diagrams of this class diagrams scale asymptotically as N_c^2.

The fact that adding a non-planar diagram reduces the scaling by a factor of N_c^{-2} can be understood in a similar way. If one starts with a planar diagram of gluons and adds a non-planar gluon to it, the number of color loop factors decreases by one for a suppression factor of $1/N_c$ while two additional factors of g must be added for another factor of $1/N_c$. Thus for example a diagram with a single non-planar gluon will scale asymptotically as N_c^0.

Similarly the scaling of diagrams containing quark loops that form the boundary of the diagram can be understood by noting that such diagrams can always be obtained starting from a planar diagram of gluons and then inserting a quark loop into a gluon propagator. Doing so does not change the number of color loop factors but adds two coupling coupling constants for each quark loop which together scale as N_c^{-1} per quark loop.

The scaling rules for diagrams allow one to deduce the asymptotic scaling for the properties of glueballs and mesons [1162,1168]. This can be done via the study of correlators of local gauge-invariant operators, J, that carry the quantum numbers of the glueballs or mesons of interest. For concreteness, consider J to be a quark bilinear such as $J = \overline{q}q$ for the case of mesons (where for simplicity spin and flavor will be neglected in the discussion as they do not affect the N_c scaling) and an operator such as $J = F^a_{\mu\nu}F^{a\,\mu\nu}$ for the case of glueballs. The correlator can be obtained by inserting these operators into closed loop diagrams. Doing so does not alter the leading N_c scaling of the diagram. Thus if one is studying correlators carrying glueball quantum numbers, then the leading diagrams scale as N_c^2; similarly if one is studying a correlator carrying meson quantum numbers, then one needs to have a quark loop in the diagram and the leading diagram scales as N_c.

Consider the two-point correlation function:

$$\begin{aligned} \Pi_J(q^2) &= -i\int d^4x e^{-iq_\mu x^\mu}\langle T\,(J(x)J(0))\rangle \\ &= \int ds \frac{\rho(s)}{q^2 - s + i\epsilon} \end{aligned} \tag{5.86}$$

where up to an overall factor $\rho(s)$, the spectral density, is given by the imaginary part of the correlator. It scales with N_c in the same way as the correlator. The contributions to the spectral density from a given diagram can be extracted from its imaginary part. Moreover, cutting a diagram at various points between the sources reveals the gluon and quark contributions to the imaginary part, which by construction will form color singlet combinations. Using the double line notation, it is easy to see that no matter where the diagrams are cut between the sources, at leading order in $1/N_c$ all of the quark and gluons indices contract into a single color singlet combination – i.e. one that cannot be broken into multiple color singlet combinations.

If additionally one assumes confinement in the most basic sense that all asymptotic states are color singlets, this means that at leading order in the $1/N_c$ expansion, the operator J creates single hadron states. By matching the N_c counting of the leading diagram to the propagation of a single hadron one sees that

$$\begin{aligned} \langle \text{meson}|J_{\text{meson}}|\text{vac}\rangle &\sim N_c \\ \langle \text{glueball}|J_{\text{glueball}}|\text{vac}\rangle &\sim N_c^2 \end{aligned} \tag{5.87}$$

With this one can deduce numerous properties [1162, 1168] of QCD as a theory of hadrons by matching correlators at the quark–gluon level to descriptions at the hadronic level. One finds:

- The masses of mesons and glueballs become independent of N_c as $N_c \to \infty$.
- Mesons and glueballs become stable as $N_c \to \infty$.
- The physics of mesons and glueballs can be described by an effective tree-level theory with vertices that scale at leading order as $N_c^{2-n_g-\frac{1}{2}n_m}$, where n_m and n_g are the number of meson lines and gluon lines respectively at the vertex. This implies that
  1. Interactions between these hadrons are weak.
  2. Meson decay amplitudes scale asymptotically as $N_c^{-1/2}$ and their widths as N_c^{-1}. Glueball decay amplitudes scale as N_c^{-1} and their widths as N_c^{-2}
  3. Meson–meson scattering amplitudes scale as $1/N_c$. Glueball-glueball scattering amplitudes scale as $1/N_c^2$, while meson–glueball scattering amplitudes scales as $1/N_c$.
  4. In decays of hadrons into mesons, when all else is equal, processes with fewer mesons as decay products are favored by powers of N_c. Thus for example the partial decay width of a meson into a ρ-meson and a pion scales as $1/N_c$ while the rate into three pions directly scales as $1/N_c^2$.
- There are an infinite number of distinct mesons for each quantum number. This can be seen by matching the correlator to one at large space-like q^2 which can be computed perturbatively [1168].
- Quantum number exotic hybrid mesons (states whose quantum numbers cannot be obtained as a quark–antiquark state in a simple quark model but require at

least one additional gluon) behave like ordinary mesons in N_c scaling [1169]. At large N_c they are narrow, there are an infinite number of them for any quantum number and their interactions with each other and with other mesons and glueballs scale according to the same rules as ordinary mesons.
– The OZI rule [18,1170,1171] becomes exact in the large N_c limit.
This implies glueball-meson mixing is suppressed.
– Tetraquark states do not exist at large N_c [1168,1172].

These properties can be viewed as predictions of QCD: they specify which quantities are dominant assuming that the large N_c world is a reasonable proxy for our world. But, at best they make qualitative predictions since the coefficients multiplying the leading terms in the expansion are not specified by this analysis. Moreover in the physical world N_c is only three so one might expect that assuming dominance of the leading-order predictions of the $1/N_c$ expansion would at best be a crude description of the phenomenology. In addition, the extent to which the phenomenology is qualitatively described by the leading-order behavior depends on the observable in question.

In the meson sector the large N_c world might well be considered as a crude but recognizable caricature of much of the observed $N_c = 3$ phenomenology, at least for meson constructed from light quarks. There are numerous mesons that are comparatively narrow – with widths much smaller than masses. There are often several identified mesons in a single spin-isospin-parity channels; presumably the number of identifiable meson would increase if N_c were made larger. The OZI rule is typically well satisfied phenomenologically; indeed it was proposed based on phenomenological grounds before the formulation of QCD [18,1170,1171].

While many qualitative aspects aspects of meson physics can be deduced from the behavior of the theory at large N_c, there are observables in the meson sector for which subleading effects are sufficiently large that the leading behavior in a $1/N_c$ expansion does not describe the physical world even qualitatively. For example, the would-be nonet of pseudo-Goldstone bosons; a nonet would exist if the OZI-rule held – as it does at large N_c. However experimentally there is an octet split from a much heavier η' meson. Of course this splitting is related to topology and the axial anomaly citetHooft:1986ooh, but in a large N_c world these effects would be suppressed by an overall factor of N_c^{-2} [1173]. The fact that in the physical world the splitting is large shows that the large N_c world is quite different from ours for this observable .

In fact, there are large classes of observables for which the the large N_c world appears to be quite different from the $N_c = 3$ world. At large N_c there should be a very large number of species of narrow glueballs that are weakly mixed with mesons. However, in the physical of world of $N_c = 3$ there are comparatively few glueball candidates [616] and the evidence for such states is typically somewhat murky, either because the evidence of the resonance is weak or because of mixing with ordinary mesons makes their "glueball" status unclear. Indeed, the identification of a resonance as a glueball may depend on there being an "extra" isoscalar state compared to what one expects from a naive quark model. Nevertheless, large N_c analysis of glueballs is of value at a theoretical level and to a limited extent also acts to inform phenomenology: by providing a regime in which narrow, weakly mixed glueballs must exist, minimally it demonstrates that there is nothing in the basic structure of gauge theories containing both light quark and gluon degrees of freedom that forbids the existence of glueball states.

In a a similar way, the spectrum of quantum number exotic hybrid mesons in nature look quite different than in a large N_c world: there are few candidates for such hadrons carrying light quarks quantum numbers [616]. Moreover, the evidence for those candidates is also typically murky due to inconclusive evidence for a resonance. Again the large N_c analysis demonstrates that there is nothing in the basic structure of gauge theories forbidding hybrid mesons. Large N_c also predicts that there should not be resonances in tetraquark channels. However, a clear signal for a quantum number exotic tetraquark has recently been found [1067]. As it happens this state is associated with heavy quarks – it is a doubly charmed state – and it is easy to see that heavy quark limit and the large N_c limits are not expected to commute for these channels. If one increased N_c while keeping the quark masses fixed, it is expected that this state would disappear.

5.6.2 The 't Hooft model

The N_c scaling rules presented above can be thought of as predictions about the physical world, but only in a qualitative sense – and they fail, even qualitatively, for many observables. Initially it was hoped that the $1/N_c$ expansion could be used as the basis of a quantitative treatment that was largely analytic at low order, in much the same as a expansion in α provided a quantitative treatment of QED. However, for QCD in 3+1 dimensions this has not worked out: even at lowest order in the expansion, the theory has proved to be intractable. Interestingly, however, QCD in 1+1 dimension, the so-called 't Hooft model [1053] was solved (initially for one flavor) at leading order in the expansion in the early 1970s.

Note that the large N_c scaling arguments given above did not depend on QCD being in 3+1 dimensions; they should hold in 1+1 dimension as well. Thus, one can use the explicit solutions of the 't Hooft model as a way to check the self-consistency of these rules.

The 't Hooft model has a critical property in common with QCD – confinement. It is useful to recall, however, that the

mechanism of confinement in 1+1 dimensions is very different than in 3+1 dimensions. It occurs for a rather trivial reason – electric flux lines cannot spread out and thus even electrodynamics is confining in 1+1 dimension. The physics of gauge fields in 1+1 dimensions is also very simple: the field strength tensor, $F_{\mu\nu}$, has an electric component $E = F_{01} = -F_{10}$ but no magnetic component. Thus in 1+1 dimensional QED, the gauge field is not associated with a propagating photon; the Euler–Lagrange equation for the gauge field is not dynamical, but simply an equation of constraint fixing the electric field from the charge density $j_0 = \overline{\psi}\gamma_0\psi$. This is because the Gauss law (plus some conditions at infinity) fully determines the electric field. Something completely analogous occurs in the 't Hooft model.

While the color electric field in QCD in 1+1 dimension can be fixed given the color charge density of the quarks, the gauge field, A_μ, itself depends on making a gauge choice. Certain gauges, such as the axial gauge of $A_1 = 0$ or the light-cone gauge have a particular useful property: they automatically suppress gluon–gluon couplings. In the axial gauge this is clear since all of the nonlinear terms involve products of A_1 and A_0. Gluon–gluon couplings vanish in the light-cone gauge for similar reasons. Since it is these non-linear couplings that make QCD complicated, QCD in 1+1 dimensions greatly simplifies.

The 't Hooft model simplifies further at leading order in the $1/N_c$ expansion. The leading diagrams for the j_μ correlator (which carries meson quantum numbers) are planar with a single quark loop bounding the diagram. This means that no gluon lines can either cross (due to the large N_c constraint) nor interact (due to the lack of gluon–gluon interactions. Accordingly the correlator is given by the so-called rainbow-ladder approximation: each quark propagator has a self-energy given by the sum of rainbow diagrams, while the interactions between quark lines is the the sum of ladder diagrams. The sum of these diagrams can be reduced analytically to integral equation between spinor-valued objects.

These simplify further into simple integral equations if one uses the light-cone gauge, which is based on light-cone coordinates:

$$x^{\pm} = \frac{x^0 \pm x^1}{\sqrt{2}} \tag{5.88}$$

and has a metric given by $g^{+-} = g^{-+} = 1$ and $g^{++} = g^{--} = 0$. The light cone gauge condition is

$$A_- = A^+ = 0\,; \tag{5.89}$$

among other things it has the virtue of being Lorentz invariant.

At leading order in large N_c the spectral function for this correlator is expected to be saturated by arbitrarily narrow meson states. Since the explicit form of the correlator is calculable, one can develop a light-cone Bethe–Salpeter type eigenvalue equation for μ^2 the meson mass, and $\psi(K)$, the light-cone Bethe–Salpeter amplitude for the meson. It is given in terms of a light-cone momentum, K appropriately scaled so that $\psi(K)$ vanishes at $K = 0$ and $K = 1$ and $\psi(K)$ is only defined for $0 \le K \le 1$. It is given by the following integral equation

$$\begin{aligned}\mu_n^2\psi_n(K) = &\frac{m^2 - \frac{g^2}{\pi}}{K(1-K)}\psi_n(K) \\ &- \frac{g^2}{\pi}\int_0^1 dK' \frac{\mathcal{P}}{(K-K')^2}\psi_n(K')\end{aligned} \tag{5.90}$$

where $\mathcal{P}$ indicates principal value, μ_n is the meson mass for the n_{th} meson, $\psi_n(K)$ is the Bethe–Salpeter amplitude for that state, m is the quark mass and g is the coupling constant (which has dimensions of mass in 1+1 dimensions).

While there is no analytic solution to this integral eigenvalue equation, it can easily be solved numerically to give the meson spectrum for the model. Note that N_c is not present in this expression showing self-consistently that meson masses are independent of N_c at large N_c as deduced from general scaling rules.

The fact that $\psi(K)$ vanishes at the $K = 0$ and $K = 1$ implies that the spectrum will be discrete – there are no solutions corresponding to two free quarks; the model correctly incorporates confinement. It is easy to show that for all values of m and g, μ^2 is real. This shows self-consistently that mesons are stable at large N_c and verifies the general analysis discussed above. Moreover it can be shown that μ^2 is always positive, showing that no matter how large the coupling, g, there are no tachyonic states that would signal an instability.

For asymptotically large values of n, it is easy to find the eigenvectors and Bethe–Salpeter amplitudes:

$$\mu_n = g^2\pi n, \quad \psi(K) = \sin(n\pi K). \tag{5.91}$$

This asymptomatic form shows that solutions exist for arbitrarily high n, indicating the self-consistency of the large N_c analysis, which predicted that there are an infinite number of mesons at large N_c.

The limit of zero quark mass in the 't Hooft model at large N_c is interesting as it provides an opportunity to study chiral symmetry and its spontaneous breaking [1174]. The regime in which chiral symmetry breaking takes place requires that care be taken in the ordering of limits. One must take the limit of $N_c \to \infty$ (with the 't Hooft coupling, g^2N_c, held fixed), prior to the $m \to 0$ limit. This limiting procedure insures that the ratio $\frac{g}{m}$ goes to zero in the combined limit. In this limit, it can be shown [1174], that chiral condensate is given by

$$\langle\overline{q}q\rangle = -N_c\sqrt{\frac{g^2N_c}{12\pi}}\,. \tag{5.92}$$

Thus the 't Hooft model provides a simple illustration of how chiral symmetry breaking can work in a gauge theory.

However, the nature of spontaneous chiral symmetry breaking in the 't Hooft model is rather subtle. Note that the spontaneous breaking of chiral symmetry is a violation of Coleman's theorem [1175] which rules out spontaneous symmetry breaking of a continuous symmetry for theories in 1+1 dimensions. Thus, spontaneous chiral symmetry breaking seems paradoxical.

The resolution of the paradox was provided by Witten [1176] in his analysis of an analogous problem: spontaneous chiral symmetry breaking in the Thirring model at large N_c. It turns out that that the spontaneous chiral symmetry breaking is an artifact of working at infinitely large N_c from the outset; it is absent for any finite N_c, no matter how large. Thus, as the large N_c limit is approached the condensate is always strictly zero and there are no Goldstone bosons. However, the theory is in a Berezinski–Kosterlitz–Thouless phase [1177,1178] in which the symmetry is "almost broken" and correlation functions of $\overline{q}q$ behave in a nontrivial way. For space-like separations

$$\langle T[\overline{q}q(x,t)\overline{q}q(0,0)]\rangle \sim (x^2 - t^2)^{\frac{\text{const}}{N_c}} \ . \tag{5.93}$$

One sees that for any finite N_c correlation functions $\overline{q}q$ fall off at large distance and thus do not saturate as they would if a condensate had formed. However, they also do not fall off exponentially as they would if $\overline{q}q$ created massive particles. Instead, there are long-range correlations: the correlation functions fall as a power law with distance. Moreover, the power depends on N_c in such a way that it goes to zero at infinite N_c. Thus if one takes N_c to be infinite at the outset, the systems acts as though spontaneous symmetry breaking had occurred.

The large N_c properties of glueballs deduced earlier cannot be checked in the 't Hooft model for a very simple reason: in 1+1 dimension there are no glueballs.

5.6.3 Baryons

Of course mesons, glueballs and hybrids are not the only hadrons, there are also baryons. Unfortunately, the direct study of correlation functions via diagrammatic methods as was done for meson and glueballs does not work for baryons. This is for an obvious reason: a baryon contains (at least) N_c quarks so that the number of quark lines in diagrams must grow with N_c. Among other things, this destroys the dominance of planar diagrams.

Witten argued that one can deduce the correct scaling behavior of large N_c baryons by first considering the case in which all of the quarks are heavy (with masses much larger than the QCD scale) [1168]. In that situation, quark–antiquark pairs are suppressed and the propagation of quarks is non-relativistic. At the most trivial level, it ought to be apparent that in this regime $M_{\text{baryon}} \approx N_c M_Q$ where M_Q is the mass of a heavy quark: the dominant term in the mass of a nonrelativistic system is the mass of the constituents and the baryon contains N_c quarks. Thus the mass of the baryon scaling of the baryon mass with N_c is

$$M_{\text{baryon}} \sim N_c \ . \tag{5.94}$$

Of course this result is from the leading term in a combined expansion built around the heavy quark and large N_c limits with the heavy quark limit taken first; one might worry that the limits do not commute for the baryon mass. However, it is straightforward to see that subleading terms in a $1/M_Q$ expansion of the baryon mass also have a leading-order term in the N_c expansion that scales like N_c. This suggests that this scaling could be general and hold independently of the quark mass. To see how this comes about, recall that in a heavy quark expansion for the baryon mass, the leading term – the direct quark mass contribution – is essentially not dynamical; the dominant subleading terms overall are the leading dynamical ones. The effective heavy quark lagrangian includes a nonrelativistic kinetic energy for the quarks and a color-Coulomb interaction between them. Witten [1168] demonstrated that at large N_c, the Hartree mean-field approximation to the non-relativistic color Coulomb problem becomes exact. In the Hartree approximation, correlations are neglected and each quark sits in an effective 1-body potential derived from interactions with the other $N_c - 1$ quarks (which sit in the ground state of the same potential).

Since the color-Coulomb interaction between two quarks has two factors of the coupling constant g, it scales as $1/N_c$. The mean-field Hamiltonian between one quark and the remainder has that quark interacting with $N_c - 1$ quarks and interactions add coherently. Thus, the mean-field Hamiltonian scales as $(N_c - 1)/N_c$ and at asymptotically large N_c becomes independent of N_c. The one-body equation for a single quark is then independent of N_c at large N_c and the quark's ground state wave function is also independent of N_c. This means that the spatial extent of the Hartree potential is itself independent of N_c. The contribution of the kinetic energy to the mass scales as is N_c since there are N_c quarks. The potential energy contributes $\frac{1}{2}N_c\langle V_{\text{Hartree}}\rangle$, where $\langle V_{\text{Hartree}}\rangle$ is the expectation value of the mean-field potential for a single quark; the factor of $\frac{1}{2}$ is because the interaction energy in a pair of quarks is split between them. Thus the direct quark mass term, the kinetic energy term and interaction term all scale linearly with N_c, strongly suggesting that $M_{\text{baryon}} \sim N_c$ independent of the quark mass.

Moreover there is a very powerful argument from Witten [1168] that the results deduced from this mean-field behavior should persist when the quarks are light. Formally one would need to start with a relativistic many-body equation for bound states – a type of Bethe–Salpeter equation gener-

alized to many particles – and show that the analog of the Hartree approximation becomes exact in the large N_c limit. While that would be technically quite complicated, it seems apparent that all of the scaling from the Hartree approximation for heavy quarks should go through provided that irreducible n-body interactions between quarks scales as N_c^{n-1}. If this is true it is easy to see that the analog of the Hartree potential will be independent of N_c: at asymptotically large N_c: there are N_c 2-body interactions that each scale as $1/N_c$, N_c^2 three-body interactions that each scale as $1/N_c^2$, N_c^3 four body-interactions that each scale as $1/N_c^3$ and so forth. Each of these has a net contribution that is independent of N_c indicating that this generalized mean-field interaction for a single quark is independent of N_c. Moreover demonstrating that n-body interactions between quarks scales as N_c^{n-1} is straightforward using diagrammatic arguments similar to those used for the glueball and meson sectors.

Using this Hartree picture it is possible to deduce [1168] the asymptotic scaling of numerous baryon properties:

- Ground state baryon masses scale asymptotically as N_c.
- The size of ground state baryons generically is independent of N_c. Explicitly this means that form factors of external currents for baryons (such as electric factors) generically scale as $N_c^0 f(q^2/N_c^0)$; for q^2 of order N_c^0 the form factor is independent of N_c. This in turn means the moments of distributions (which are related to derivatives of form factors) such as $\langle r^2\rangle$, $\langle r^4\rangle$ are independent of N_c at large N_c.
- Generic couplings between a ground state baryon and n mesons scale as $N_c^{1-n/2}$. Among other things this means that
  1. Meson–baryon couplings scale generically as $N_c^{1/2}$.
  2. Meson–baryon scattering amplitudes are generically independent of N_c for large N_c
- Couplings between a meson, a ground state baryon and an excited baryon are generically independent of N_c and excited baryons have widths that are independent of N_c. Unlike in the glueball and meson sectors, these states are not narrow at large N_c, nor can you can conclude that there an infinite number of them.

Witten observed [1168] an interesting pattern to the scaling properties for baryons given above. They scale asymptotically with $1/N_c$ in the same way as analogous properties of solitons scale a with coupling constants squared. This insight lead to a renaissance of interest [1179–1181] in the Skyrme model [1182] as a model for baryons.

The scaling laws given above are generic. Spin and flavor considerations may act to suppress certain couplings below these generic results. Moreover, for the case of two or more degenerate flavors, the notion of "ground state baryon" becomes a bit involved. Both of these issues are related to an emergent spin-flavor symmetry – a symmetry that is not manifest in the QCD lagrangian but emerges at large N_c. In general, this symmetry is a contracted $SU(2N_f)$ where N_f is the number of degenerate light flavors – it reduces to $SU(4)$ if one considers the up and down quarks to be effectively degenerate and the strange quark much heavier.

An initial hint that a new symmetry beyond mere isospin symmetry was emergent at large N_c could be seen in the 2-flavor Skyrme model [1179], treated classically (with requantized collective coordinates to restore broken symmetries). This treatment corresponds to leading order in the $1/N_c$ expansion. It was found that rather than having the nucleon as the sole ground state, one had a tower of states with $I = J$ (the first two being the nucleon ($I = \frac{1}{2}, J = \frac{1}{2}$) and the Δ ($I = \frac{3}{2}, J = \frac{3}{2}$) with the levels in the tower degenerate at leading order in $1/N_c$ [1179]; the splittings can be shown to be $\mathcal{O}(N_c^{-1})$. Moreover, it was found that the ratios of the values of certain observables held independently of the parameters of the model or even the precise form of the Skyrme Lagrangian [1183]. It was realized that this behavior was not a property of Skyrme models per se but rather reflected an underlying symmetry of baryons [1184–1186].

The symmetry can be seen to be required for the consistency [1185] of large N_c scaling provided that the pion–nucleon coupling scales with N_c generically – i.e. as $N_c^{1/2}$. With this scaling, the Born approximation for pion–nucleon nucleon would scale linearly with N_c. However, unitarity constrains the scattering amplitudes to scale no faster than N_c^0. Clearly, something must cancel the Born amplitude in any channel where the meson–baryon coupling scales generically. In the case of scalar-isoscalar mesons, it is easy to show that the heavy mass of the baryon at large N_c implies that at leading order, the contribution of the cross-Born diagram cancels the contribution of the Born diagram. However, pions are derivatively coupled and hence couple to the spin of the nucleon and are isovectors so they also couple to the isospin. The various components of spin do not commute with each other and similarly with the various components of isospin and, as a result, the cancellation between the Born and cross-Born contributions to $\pi - N$ scattering appears to be spoiled. However, the cancellation between the Born and cross-Born contributions at the level of pion–nucleon scattering will be restored provided that the Δ is treated as being degenerate (at this order) with the nucleon and the ratio of $g_{\pi N\Delta}$ (the transition coupling between the pion the nucleon and the Δ) is taken to be a prescribed number times $g_{\pi NN}$ [1185]. Applying the same logic to the process $\pi + N \to \pi + \Delta$, requires $g_{\pi\Delta\Delta}$ to be a fixed multiple of $g_{\pi N\Delta}$. At this order in $1/N_c$, the Δ and the nucleon are degenerate and the Δ should be treated as stable. Thus one can legitimately consider $\pi - \Delta$ scattering. Applying

the same logic, one deduces the existence of a degenerate $I = \frac{5}{2}, J = \frac{5}{2}$ baryon and so forth generating a tower of states that become degenerate at large N_c. Presumably the the nucleon and Δ correspond approximately to the observed states in the $N = 3$ world, while the $I = \frac{5}{2}, J = \frac{5}{2}$ is a large N_c artifact.

It is possible to show that the structure described above is encoded in a contracted $SU(4)$ Lie algebra for two-flavor QCD. The fixed ratio of the coupling constants are given by the Clebsch–Gordan coefficients of the group. The same logic that gives rise to the contracted $SU(4)$ symmetry, gives a contracted $SU(6)$ for 3-flavor QCD to the extent that one can approximate the strange quark as being nearly degenerate with the up and down quarks [1186]. Moreover, it is possible to show that for certain observables the leading corrections to the the contracted $SU(2N_f)$ symmetry is of order $1/N_c^2$ rather than $1/N_c$ [1187]. This fact allows one to make some semi-quantitative predictions based on the emergent symmetry encoded in the large N_c limit for baryons. A good example of this are the mass relations of Ref. [1163].

5.6.4 Nucleon–nucleon interactions and nuclear physics

The study of nucleon–nucleon interactions is complicated for kinematical reasons associated with the large nucleon mass. There are two kinematic regimes of interest: one in which the momentum transfers are independent of N_c and the other in which the momentum transfers are of order N_c – i.e. in which the velocities are independent of N_c. Physical observables associated with nucleon–nucleon scattering do not have a smooth large N_c in the regime in which momentum transfers are of order N_c^0, but an analysis based on a time-dependent Hartree picture suggests that some scattering observables will have smooth large N_c limits [1168] in the regime of momentum transfers of order N_c. These observables do not include many standard scattering observables such as phase shifts; the ones that have smooth limits appear to be those in which one follows the bulk flow of quantities of interest [1188]. Presumably the total cross-section also has a smooth limit [1189]. There is some predictive power for the spin and flavor dependence of such observables owing to the contracted $SU(4)$ symmetry [1188,1189].

In the regime in which momentum transfers are of order unity – the regime of relevance to nuclear structure – the logic of Ref. [1168] implies that the nucleon–nucleon interaction strength is of order N_c, which is formally of the same order as the nucleon mass, while its range is independent of N_c. This implies that nuclear matter would be crystalline at large N_c, with nucleons constrained to be near the minimum of the potential from other nucleons. This is radically different from what is seen nature, suggesting that a $1/N_c$ expansion around the large N_c limit is not a useful approach to nuclear structure. Interestingly, however, if one focuses solely on the spin-flavor structure of the nucleon–nucleon potential – a quantity that is not directly physical – there is a hierarchy in the strength of various spin-flavor contributions. This hierarchy is qualitatively similar to what one would obtain from the contracted $SU(4)$ spin-flavor symmetry of large N_c QCD [1190,1191]. This behavior is consistent with what one would expect if the nucleon–nucleon force was described via meson exchanges, as has been typically done in nuclear physics. Since the overall potential strength at the one-meson exchange level is large in some channels, consistency requires subtle cancellations when multiple-meson exchange are included. Such cancellations naturally occur due to the contracted $SU(4)$ symmetry [1192].

5.6.5 Other large N_c limits

The large N_c limit of QCD is an extrapolation from our world at $N_c = 3$ to a large N_c world. However, that extrapolation is not unique. The standard approach discussed above involves keeping the number of flavors fixed while letting N_c go to infinity. However, there is an alternative, the Venziano limit [1193] in which the ratio of the number of colors to the number of flavors is held fixed as $N_c \to \infty$. The large N_c world for these two limits are quite different.

There is yet another large N_c limit that exploits the fact that at $N_c = 3$, the representation for fundamental color and for the antisymmetric combination of two anti-fundamental colors are identical (i.e. r is indistinguishable from $\left(\overline{g}\overline{b} - \overline{b}\overline{g}\right)/\sqrt{2}$). However quarks with fundamental color and with two-index antisymmetric color extrapolate to large N_c quite differently – there are N_c distinct quark colors for the former and $N_c(N_c - 1)/2 \sim N_c^2$ for the latter.

The large N_c limit based on quarks in the two-index antisymmetric representation, denoted QCD(AS), has remarkable formal connections to supersymmetric QCD [1194–1196]. Phenomenologically, QCD(AS) has scaling of meson properties with N_c similar to those of glueballs; one important difference between QCD(AS) at large N_c and the conventional large N_c limit is that in QCD(AS) quantum number exotic tetraquarks are not forbidden; indeed, they are required [1197]. The description of baryons for QCD(AS) is in analogy to Witten's but a somewhat new type of analysis is required [1198]. Formally, the predictions for baryon spectroscopy are distinct in QCD(AS) and QCD with quarks in the fundamental representation [1199], but phenomenological predictions for both expansions work to the order expected in describing real world data.

5.7 OPE-based sum rules

SVZ sum rules, $\frac{1}{M_Q}$ expansion and all that
Mikhail Shifman

5.7.1 Preamble

Rewind to autumn of 1971. I am a student at ITEP in Moscow, working on my Masters degree. The famous paper of Gerhard 't Hooft [52] was published in Nuclear Physics in October, but neither myself nor anybody else in ITEP immediately noticed this ground-breaking publication. At that time I did not even know what Yang–Mills theories meant. Now, when I think of the inception of QCD, the memories of this paper and its sequel [51] (issued in December of 1971) always come to my mind. For me, psychologically this was the beginning of the QCD era.

To give an idea of the scientific atmosphere at that time (1972) I looked through the Proceedings of the 1972 International Conference On High-Energy Physics [1200]. Theoretical talks were devoted to dual models (a precursor to string theory), deep inelastic scattering and Bjorken scaling, current algebra, $e^+e^- \to$ hadrons, etc. In three talks – by Zumino, Bjorken and Ben Lee – the Weinberg-Salam model (a precursor to the present-day Standard Model) was reviewed.[47] Ben Lee was the only person to refer to 't Hooft's publications [51,52]. The last talk of the conference summarizing its major topics was delivered by Murray Gell-Mann. In this talk Gell-Mann discusses, in particular, whether quarks are physical objects or abstract mathematical constructs. Most interesting for us is his analysis of the $\pi^0 \to 2\gamma$ decay. Gell-Mann notes that if quarks are fermions then the theoretically predicted amplitude is a factor of 3 lower than the corresponding experimental result, but makes no statement of the inevitability of the quark color.[48]

In October 1972 I was accepted to the ITEP graduate school. My first paper on deep inelastic scattering in the Weinberg-Salam model was completed in early 1973; simultaneously, I started studying Yang–Mills theories (in particular, the Faddeev–Popov quantization [1201][49]) in earnest. At the same time, somewhere far away, behind the Iron Curtain, Callan and Gross searched for a theory with an ultraviolet fixed point at zero. In July of 1973 Coleman and Gross submitted to PRL a paper asserting that "no renormalizable field theory that consisted of theories with arbitrary Yukawa, scalar or Abelian gauge interactions could be asymptotically free" [1202]. Damn Iron Curtain! If Gross asked anyone from the ITEP Theory Department he would have obtained the answer right away. The above theorem was known to the ITEP theorists from the Landau time. For brevity I will refer to it as the Landau theorem, although it was established by his students rather than Landau himself. The general reason why this theorem holds was also known – the Källen–Lehman (KL) representation of the polarization operator plus unitarity.

An explanatory remark concerning the Landau theorem might be helpful here. For asymptotic freedom to take place the first coefficient of the β function must be *negative*. The sign of the one-loop graphs which determine the coupling constant renormalization is in one-to-one correspondence with the sign of their imaginary parts (this is due to the dispersion KL representation for these graphs). Unitarity implies the positivity of the imaginary parts which inevitably leads to the *positive* first coefficients in the β functions in renormalizable four-dimensional field theories based on arbitrary Yukawa, scalar or Abelian gauge interactions. This situation is that of the Landau zero charge in the infrared rather than asymptotic freedom. In Yang–Mills theories in physical ghost-free gauges some graphs have no imaginary parts which paves the way to asymptotic freedom (see e.g. [1203]).

In fact, it is quite incomprehensible why asymptotic freedom had not been discovered at ITEP after 't Hooft's 1971 publication. In Ref. [1203] the reader can find a narrative about this historical curiosity.

May 1973 should be viewed as the discovery of asymptotic freedom [53,54]. That's when the breakthrough papers of Gross, Wilczek and Politzer were submitted – simultaneously – to PRL. David Gross recollects [1202]:

> *We completed the calculation in a spurt of activity. At one point a sign error in one term convinced us that [Yang–Mills] theory was, as expected, non-asymptotically free. As I sat down to put it together and to write up our results, I caught the error. At almost the same time Politzer finished his calculation and we compared, through Sidney, our results. The agreement was satisfying.*

It took a few extra months for QCD to take off as *the* theory of strong interactions. The events of the summer of 1973 that led to the birth of QCD are described by H. Leutwyler

[47] There is a curious anecdote I heard later: In December 1979, after the Glashow–Weinberg–Salam Nobel Prize ceremony, a program was aired on Swedish radio. At some point, Weinberg quoted a phrase from the Bible. Salam remarked that it exists in the Quran too, to which Weinberg reacted: "Yes, but we published it earlier!"

[48] For me personally the following remark in his talk was a good lesson for the rest of my career: "Last year the rate of $K_L^0 \to \mu^+\mu^-$ decay was reported to be lower than allowed by unitarity unless fantastic hypotheses are concocted. Now the matter has become experimentally controversial." Alas…concocting fantastic hypotheses was the core of my Masters thesis.

[49] A longer and more comprehensible version appeared in Russian as Kiev preprint ITP 67-36. In the beginning of the 1970s, it was translated in English by B. Lee (NAL-THY-57, 1972). Apparently, in [52], [51] 't Hooft used the short version while I could use the longer one.

in Sect. 1.1 of this Volume. To my mind, the final acceptance came with the November Revolution of 1974 – the discovery of J/ψ and its theoretical interpretation as orthocharmonium.[50] In the fall of 1973 we submitted a paper [1205] explaining why the Landau theorem in four dimensions fails only in Yang–Mills theory.

QCD and its relatives are special because QCD is the theory of *nature*. QCD is strongly coupled in the infrared domain where it is impossible to treat it quasiclassically – perturbation theory fails even qualitatively. It does not capture the drastic rearrangement of the vacuum structure related to confinement. The Lagrangian is defined at short distances in terms of gluons and quarks, while at large distances of the order of $\gtrsim \Lambda_{\rm QCD}^{-1}$ (where $\Lambda_{\rm QCD}$ is the dynamical scale of QCD, which I will refer to as Λ below) we deal with hadrons, e.g. pions, ρ mesons, protons, etc. Certainly, the latter are connected with quarks and gluons in a divine way, but this connection is highly nonlinear and non-local; even now, 50 years later, the full analytic solution of QCD is absent.

Non-perturbative methods were desperately needed.

5.7.2 *Inception of non-perturbative methods*

Four years before QCD Ken Wilson published a breakthrough paper [30] on the operator product expansion (OPE) whose pivotal role in the subsequent development of HEP theory was not fully appreciated until much later. What is now usually referred to as Wilsonian renormalization group (RG), or Wilsonian RG flow, grew from this paper. The Wilsonian paradigm of separation of scales in quantum theory was especially suitable for asymptotically free theories. Wilson's formulation makes no reference to perturbation theory, it has a general nature and is applicable in the non-perturbative regime too. The focus of Wilson's work was on statistical physics, where the program is also known as the block-spin approach. Starting from microscopic degrees of freedom at the shortest distances a, one "roughens" them, step by step, by constructing a sequence of effective (composite) degrees of freedom at distances $2a$, $4a$, $8a$, and so on. At each given step i one constructs an effective Hamiltonian, which fully accounts for dynamics at distances shorter than a_i in the coefficient functions.

QCD required a number of specifications and adjustments. Indeed, the UV fixed point in QCD is at $\alpha_s = 0$; hence, the approach to this fixed point at short distances is very slow, logarithmic rather than power-like, characteristic for the $\alpha_s \neq 0$ fixed point. In fact, it is not the critical regime at the UV fixed point *per se* we are interested in but rather the regime of approach to this critical point. Moreover, it was not realized that (in addition to the dynamical scale Λ) the heavy quarks provide an extra scale – the heavy quark mass m_Q – which must be included in OPE where necessary.

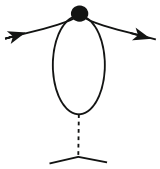


Fig. 110 The penguin mechanism in flavor-changing decays. Any of three heavy quarks c, b or t can appear in the loop

Surprisingly, in high-energy physics of the 1970s the framework of OPE was narrowed down to a very limited setting. On the theoretical side, it was discussed almost exclusively in perturbation theory. On the practical side, its applications were mostly narrowed down to deep inelastic scattering, where it was customary to work in the *leading*-twist approximation.

The fact that the UV fixed point is at zero makes OPE both more simple and more complicated than in the general case. On one hand, the anomalous dimensions of all composite local operators which might be relevant in the given problem scale only logarithmically. On the other hand, slow (logarithmic) fall off of "tails" instead of desired power-like makes analytic separation of scales technically difficult.

I believe that we – Arkady Vainshtein, Valentin Zakharov and myself – were the first to start constructing a QCD version of OPE. The first step in this direction was undertaken in 1974 in the problem of strangeness-changing weak decays [1206, 1207] (currently known as the penguin mechanism in flavor-changing decays). A mystery of $\Delta I = \frac{1}{2}$ enhancement in K decays had been known for years (for a review see [1208] and Sect. 13.3). A suggestion of how one could apply OPE to solve this puzzle was already present in Wilson's paper [30]. Wilson naturally lacked particular details of QCD. The first attempt to implement Wilson's idea in QCD was made in [1209,1210]. Although these papers were inspirational, they missed the issue of a "new" OPE needed for QCD realities. Seemingly, we were the first to address this challenge, more exactly two of its features: mixed quark–gluon operators (in [1206,1207] we introduced

$$O_{\rm peng} = \bar{s}_L \gamma_\mu \left(\mathcal{D}_\nu G^{\mu\nu}\right) d_L$$

which is purely $\Delta I = \frac{1}{2}$) and coefficients logarithmically depending on the charmed (i.e. heavy at that time) quark mass. Currently, c, b, t quark masses appear in the penguin operators (illustrated in Fig. 110), the latter two being genuinely heavy. Through equations of motion the operator $O_{\rm peng}$ reduces to a four-quark operator but its chiral structure is different from conventional, namely, it contains both left-handed and right-handed quark fields since $\mathcal{D}_\nu G^{\mu\nu} \sim \sum_q \bar{q}\gamma^\mu q$. Combined with another revolutionary

[50] I should also mention a highly motivating argument due to S. Weinberg who proved [1204] that (in the absence of the U(1) current gluon anomaly) $m_{\eta'} \leq \sqrt{3} m_\pi$. This argument seemingly was discussed during ICHEP 74 in July 1974.

Table 6 The lowest-dimension operators in OPE. Γ is a generic notation for combinations of the Dirac γ matrices

Normal dim	3	4	5	6	6
Operator	$O_q = \bar{q}q$	$O_G = G^2_{\mu\nu}$	$O_{qG} = \bar{q}\sigma^{\mu\nu}G_{\mu\nu}q$	$O_{4q} = (\bar{q}\Gamma q)^2$	$O_{3G} = GGG$

finding of QCD, the extraordinary smallness of the u and d quark masses, $m_{u,d} \sim 5\,\text{MeV}$ (see Sect. 1.1.15), the mixed chiral structure of the emerging four-fermion operator provides the desired enhancement of the $\Delta I = \frac{1}{2}$ amplitude. It took us over 2 years to fight a succession of referees for publication of Ref. [1207]. One after another, they would repeat that mixed-chirality four-fermion operator in the considered theory was complete nonsense. Currently, the penguin mechanism in flavor changing weak transitions is a basic theoretic element for a large variety of such decays. As Vainshtein put it [1208], "Penguins spread out but have not yet landed."

Systematic studies of Wilsonian OPE in QCD can be traced back to the summer of 1977 – that is when the gluon condensate O_G (see Table 6) was first introduced [1211]. Vacuum expectation values of other gluon and quark operators were introduced in Ref. [145], which allowed one to analyze a large number of vacuum two- and three-point functions, with quite nontrivial results for masses, coupling constants, magnetic moments and other static characteristics of practically all low-lying hadronic states of mesons and baryons. A consistent Wilsonian approach requires an auxiliary normalization point μ which plays the role of a regulating parameter separating hard contributions included in the coefficient functions and soft contributions residing in local operators occurring in the expansion. The degree of locality is regulated by the same parameter. "Hard" versus "soft" means coming from the distances shorter than μ^{-1} in the former case and larger than μ^{-1} in the latter.

After setting the foundation of OPE in QCD [145] we were repeatedly returning to elaboration of various issues, in particular, in the following works: [1212], [1213], and [1214].

5.7.3 SVZ sum rules. Concepts

The 1998 review [1213] summarizes for the reader foundations of the Shifman–Vainshtein–Zakharov (SVZ) sum rules in a pedagogical manner. At short distances QCD is the theory of quarks and gluons. Yang–Mills theory of gluons confines. This means that if you have a heavy probe quark and an antiquark at a large separation, a flux tube with a constant tension develops between them, preventing their "individual" existence. In the absence of the probe quarks, the flux tube can form closed contours interpreted as glueballs. This phenomenon is also referred to in the literature as the area law or the dual Meißner effect. Until 1994 the above picture was the statement of faith. In 1994 Seiberg and Witten found an analytic proof [1215,1216] of the dual Meißner effect in $\mathcal{N} = 2$ super-Yang–Mills.[51] The Seiberg-Witten solution does *not* apply to QCD, rather to its distant relative. The real world QCD, with quarks, in fact has no area law (the genuine confinement is absent) since the flux tubes break through the quark–antiquark pair creation. Moreover, light quarks are condensed, leading to a spontaneous breaking of chiral symmetry, a phenomenon shaping the properties of the low-lying hadronic states, both mesonic and baryonic. The need to analytically understand these properties from first principles led us to the development of the SVZ method.

The quarks comprising the low-lying hadronic states, e.g. classical mesons or baryons, are not that far from each other, on average. The distance between them is of order of Λ^{-1}. Under these circumstances, the string-like chromoelectric flux tubes, connecting well-separated color charges, do not develop and details of their structure are not relevant. Furthermore, the valence quark pair injected in the vacuum – or three quarks in the case of baryons – perturb it only slightly. Then we do not need the full machinery of the QCD strings[52] to approximately describe the properties of the low-lying states. Their basic parameters depend on how the valence quarks of which they are built interact with typical vacuum field fluctuations.

We endowed the QCD vacuum with various condensates – approximately a half-dozen of them – in the hope that this set would be sufficient to describe a huge variety of the low-lying hadrons, mesons and baryons. The original set included[53] the gluon condensate $G^2_{\mu\nu}$, the quark condensate $\bar{q}q$, the mixed condensate $\bar{q}\sigma^{\mu\nu}G_{\mu\nu}q$, various four-quark condensates $\bar{q}\Gamma q\bar{q}\Gamma q$, and a few others (see Table 6). Later this set had to be expanded to address such problems as, say, the magnetic moments of baryons.

Our task was to determine the regularities and parameters of the classical mesons and baryons from a limited set of the vacuum condensates. Figure 111 graphically demonstrates the SVZ concept. On the theoretical side, an appropriate n-point function is calculated as an operator product expansion truncated at a certain order. In most problems only condensates up to dimension 6 (Table 6) are retained. In the "experimental" part the lowest-lying meson (or baryon) is singled

[51] More exactly, confinement through the flux tube formation was proven in the low-energy limit of this theory upon adding a small deformation term breaking $\mathcal{N} = 2$ down to $\mathcal{N} = 1$.

[52] Still unknown.

[53] A meticulous writer would have used the notation $\langle G^2_{\mu\nu}\rangle$, etc. but I will omit bra and ket symbols where there is no menace of confusion.

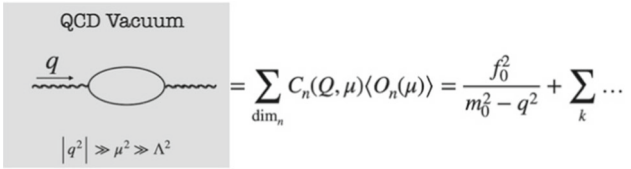


Fig. 111 A two-point correlation function in the QCD vacuum. The left side is the OPE sum with a finite number of the lowest-dimension operators ordered according to their normal dimensions. The right side is the sum over mesons with the appropriate quantum numbers. The ground state in the given channel is singled out. The excited states are accounted for in the quasiclassical approximation. We define a positive variable $Q^2 = -q^2$ and a sliding μ^2 parameter used as a separation parameter in OPE. For better convergence a Borel transformation is applied as explained below

out, while all higher states are represented in the quasiclassical approximation. Above an effective "threshold" s_0, where the spectral density becomes smooth, we apply quark–hadron duality to replace it by a perturbative spectral density. Then the parameter s_0 is fitted along with the parameters of the lowest lying state – its mass and residue.

Acting in this way, one can determine the parameters f_0 and m_0 defined in Fig. 111 and their analogs in other problems. Of course, without invoking the entire infinite set of condensates one can only expect to obtain the hadronic parameters in an admittedly approximate manner.

5.7.4 Borelization

Analyzing the sum rules displayed in Fig. 111 we realized that their predictive power was limited – summation on both sides of the equation does not converge fast enough. On the right-hand side the contribution of high excitations is too large – the lowest lying states are "screened" – because the weight factors fall off rather slowly. Likewise, to achieve reasonable accuracy on the left-hand side one would need to add operators other than those collected in Table 6. At that time we knew next to nothing about higher-dimension operators, of dimension $\gtrsim 7$. The Borel transform came to the rescue.

The Borel transformation is a device well-known in mathematics. If one has a function $f(x)$ expandable in the Taylor series, $f(x) = x\sum_n a_n x^n$ with the coefficients a_n which do not fall off sufficiently fast, one can instead introduce its Borel transform

$$\mathcal{B}f = x\sum_n \frac{a_n}{n!}x^n \tag{5.95}$$

and then, if needed, reconstruct f.[54]

[54] The Borel transform is closely related to the Laplace transform.

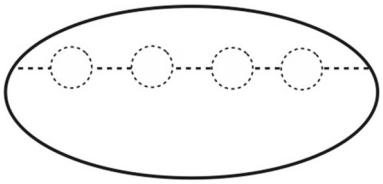


Fig. 112 Graph showing four loops renormalizing a gluon line (represented by the dotted line). A renormalon is the sum over n of such diagrams with n loops

If we apply this procedure to the sum rule in Fig. 111 we obtain for a given hadronic state i

$$\begin{aligned}\mathcal{B}\frac{f_i^2}{m_i^2+Q^2} &= \mathcal{B}\frac{f_i^2}{Q^2}\sum_n(-1)^n\left[\frac{m_i^2}{Q^2}\right]^n \\ &\to \frac{f_i^2}{Q^2}\sum_n\frac{(-1)^n}{n!}\left[\frac{m_i^2}{Q^2}\right]^n \\ &= \frac{f_i^2}{Q^2}\exp\left(-\frac{m_i^2}{Q^2}\right) \\ &\to \frac{f_i^2}{M^2}\exp\left(-\frac{m_i^2}{M^2}\right)\end{aligned} \tag{5.96}$$

where, in the final step (for historical reasons), I replaced Q^2 by a Borel parameter M^2. If M^2 can be chosen sufficiently small, higher excitations are *exponentially* suppressed.

Simultaneously, we improve the convergence of OPE on the left-hand side by applying the same operator $\mathcal{B}$. If the operator $\langle O_n\rangle$ has dimension $2d_n$, then the Borell transformation of the left hand side yields

$$\mathcal{B}\sum_n\frac{1}{(Q^2)^{d_n}}\langle O_n\rangle \to \sum_n\frac{1}{(d_n-1)!}\frac{1}{(M^2)^{d_n}}\langle O_n\rangle, \tag{5.97}$$

where I have again replaced Q^2 by the Borel parameter M^2. Since the expansion (5.97) goes in inverse powers of M^2, it is necessary to keep M^2 large enough. The two requirements on M^2 seem contradictory. However, for all "typical" resonances, such as say ρ mesons, they can be met simultaneously [145,1217,1218] in a certain "window." The only exception is the $J^P = 0^\pm$ channel. There are special reasons why $0^\pm$ mesons are exceptional, see [1219].

5.7.5 Practical version of OPE

At the early stages of the SVZ program the QCD practitioners often did not fully understood the concept of scale separation in the Wilsonian OPE. It was generally believed that the coefficients are fully determined by perturbation theory while non-perturbative effects appear only in the OPE

operators.[55] This belief led to inconsistencies which revealed themselves e.g. in the issue of renormalons (see below). A set of graphs represented by renormalons is constructed from a single gluon exchange by inserting any number of loops in the gluon line like beads in a necklace (Ref. [1220]). Being treated *formally* this contribution, shown in Fig. 112, diverges factorially at high orders. I vividly remember that after the first seminar on SVZ in 1978 Eugene Bogomol'nyi asked me each time we met: "Look, how can you speak of power corrections in the n-point functions at large Q^2 if even the perturbative expansion (i.e. the expansion in $1/\log(Q^2/\Lambda^2)$) is not well defined? Isn't it inconsistent?" I must admit that at that time my answer to Eugene was somewhat evasive.

The basic principle of Wilson's OPE – the scale separation principle – is "soft versus hard" rather than "perturbative versus non-perturbative." Being defined in this way the condensates are explicitly μ dependent. All physical quantities are certainly μ independent; the normalization point dependence of the condensates is compensated by that of the coefficient functions – see Fig. 111.

The problem of renormalons disappears once we introduce the normalization point μ. With $\mu \gg \Lambda$, there is no factorial divergence in high orders of perturbations theory. Renormalons conspire with the gluon condensates to produce, taken together, a well-defined OPE. The modern construction goes under the name of the "renormalon conspiracy"; it is explained in detail in my review [1214]. I hasten to add, though, that the renormalons acquire a life of their own in those cases in which OPE does not exist. Qualitatively, they can shed light on scaling dimensions of non-perturbative effects. The most clear-cut example of this type is the so-called "pole mass of the heavy quarks" [1221,1222] and its relation to a theoretically well-defined mass parameter [1223].

In some two-dimensional solvable models exact OPE can be constructed which explicitly demonstrates the μ dependence of both the coefficient functions and the condensates in the Wilsonian paradigm and its cancellation in the physical quantities (for a recent study see e.g. [1224]). Needless to say, if QCD was exactly solved we would have no need in the SVZ sum rules.

We had to settle for a reasonable compromise, known as *the practical version of OPE*. In the practical version we calculate the coefficient functions perturbatively keeping a limited number of loop corrections. The condensate series is truncated too. The condensates are not calculated from first principles; rather a limited set is determined from independent data.

The practical version is useful in applications only provided μ^2 can be made small enough to ensure that the "perturbative" contributions to the condensates are much smaller than their genuine (mostly non-perturbative) values. At the same time, $\alpha_s(\mu^2)/\pi$ must be small enough for the expansion in the coefficients to make sense. The existence of such "μ^2 window" is not granted *a priori* and is a very fortunate feature of QCD. We did observe this feature empirically in almost all low-lying hadrons [1225,1226].[56] At the same time, we identified certain exceptional channels revealing unforeseen nuances in hadronic physics [1219].

5.7.6 *Implementation of the idea and results*

After the strategic idea of quark and gluon interaction with the vacuum medium became clear we delved into the uncharted waters of microscopic hadronic physics. Remember, in 1977 nobody could imagine that basic hadronic parameters for at least some hadrons could be analytically calculated, at least approximately. As a show-case example we chose the most typical mesons, ρ and ϕ, to calculate their couplings to the electromagnetic current and masses. The agreement of our results with experiment was better than we could *a priori* expect. At first we were discouraged by a "wrong" sign of the gluon condensate term in the theoretical part of the appropriate SVZ sum rule. We suddenly understood that this sign could be compensated by the four-quark condensate – a real breakthrough. In November of 1977 we published a short letter [1211] which still missed a number of elements (e.g. Borelization) developed and incorporated later, one by one. We worked at a feverish pace for the entire academic year, accumulating a large number of results for the hadronic parameters. All low-lying meson resonances built from the u, d, s quarks and gluons were studied and their static properties determined from SVZ: masses, coupling constants, charge radii, ρ–ω mixing, and so on, with unprecedented success. In summer of 1978, inspired by our progress, we prepared a number of preprints (I think, eight of them simultaneously[57]) and submitted to ICHEP-78 in Tokyo. Clearly none of us were allowed to travel to Tokyo to present our results.

I cannot help mentioning an incident that occurred in the spring of 1978 when we were mostly done with this work. The episode may have been funny were it not so nerve-wracking. When we decided that the calculational stage of the work was over, I collected all my drafts (hundreds of sheets of paper with derivations and math expressions), I organized them in proper order, selected all expressions we might have needed for the final draft of the paper and the future work, meticu-

[55] Unfortunately, this misconception lasted through the 1980s and was visible in the literature even in the 1990s and later.

[56] Theoretical understanding of the roots of this phenomenon remains unclear. Seemingly, it has no known analogues in two-dimensional models.

[57] In the journal publication they were combined in three articles occupying the whole issue of Nucl. Phys. B147, №5, 1979.

lously rewrote them in a voluminous notebook (remember, we had no access to photocopying machines), destroyed the original drafts, put the notebook in my briefcase and went home. It was about midnight, and I was so exhausted that I fell asleep while on the metro train. A loud voice announcing my stop awoke me, and I jumped out of the train, leaving the briefcase were it was, on the seat. By the time I realized what had happened the train was gone, and gone with it forever my calculations …I have never recovered my briefcase with the precious notebook…After a few agonizing days it became clear that the necessary formulas and expressions had to be recovered anew. Fortunately, Vainshtein and Zakharov had kept many of their own derivations. Vainshtein never throws away anything as a matter of principle. Therefore, the problem was to dig out "informative" sheets of paper from the "noise" (this was hindered by the fact that Vainshtein was in Novosibirsk while we were in Moscow). Part of my drafts survived in the drawers of a huge desk that I had inherited from V. Sudakov. Better still, many crucial calculations were discussed so many times by us, over and over again, that I remembered them by heart. Nevertheless, I think it took a couple of uneasy weeks to reconstruct in full the contents of the lost notebook.

The SVZ method was further developed by many followers (e.g. the so-called light-cone sum rules for form-factors), see [1227] and [1228]. A broad picture of the hadronic world was obtained by the 1980s and later [1229]. Today the pioneering SVZ paper is cited 6000+ times. Until 1990s, when lattice QCD based on numeric calculations, started approaching its maturity, the SVZ method was the main tool for analyzing static hadronic properties.

5.7.7 *Reliability and predictive power*

The SVZ method is admittedly approximate. Yet, it is not a model in the sense that it cannot be arbitrarily bent to accommodate "wrong" data. It is instructive to narrate here the story of an alleged discovery of an alleged "paracharmonium" referred to as $X(2.83)$ in January of 1977 [1230]. It was widely believed then that $X(2.83)$ was the 0^- ground sate of $\bar{c}c$ quarks, η_c. If this was the case the mass difference between J/ψ and η_c would be close to 270 MeV. Shortly after, the interpretation of $X(2.83)$ as η_c was categorically ruled out by the SVZ analysis [1231] which *predicted* that the above mass difference must be 100 ± 30 MeV. Two years later, a new experiment [1232] negated the existence of the $X(2.83)$ state. In the very same experiment the genuine paracharmonium was observed at 2.98 ± 0.01 GeV, in perfect agreement with [1231]. For us this was a triumph and a lesson – if one believes in a theory one should stand for it!

5.7.8 *OPE-based construction of heavy quark mass expansion*

In the 1980s and early 1990s OPE was generalized to cover theoretical studies of mixed heavy-light hadrons, i.e. those built from light, q, and heavy, Q, flavors. In the 1990s those who used $1/m_Q$ expansion in theoretical analysis of $Q\bar{q}$ and Qqq systems numbered in the hundreds. A large range of practical physics problem related to $Q\bar{q}$ and Qqq systems were solved. Lattice analyses of such systems even now remain hindered, and in many instances the $1/m_Q$ expansion remains the only reliable theoretical method.

As I have mentioned in the second paragraph of Sect. 5.7.2, heavy quarks in QCD introduce an extra scale, m_Q. To qualify as a heavy quark Q the corresponding mass term m_Q must be much larger than $\Lambda_{\rm QCD}$. The charmed quark c can be considered as heavy only with some reservations while b and t are *bona fide* heavy quarks. The hadrons composed from one heavy quark Q, a light antiquark $\bar{q}$, or a "diquark" qq, plus a gluon cloud (which also contains light quark–antiquark pairs) – let us call them H_Q – can be treated in the framework of OPE. The role of the cloud is, of course, to keep all the above objects together, in a colorless bound state. The light component of H_Q, its light cloud, has a complicated structure; the soft modes of the light fields are strongly coupled and strongly fluctuate. Basically, the only fact which we know for sure is that the light cloud is indeed light; typical excitation frequencies are of order of Λ. One can try to visualize the light cloud as a soft medium.[58] The heavy quark Q is then submerged in this medium. The latter circumstance allows one to develop a formalism similar to SVZ in which the soft QCD vacuum medium is replaced by that of the light cloud. As a result, an OPE-based expansion in powers of $1/m_Q$ emerges (see Fig. 113). When heavy quarks are in soft medium the heavy quark–antiquark pair creation does not occur and the field-theoretic description of the heavy quark becomes redundant. A large "mechanical" part in the x dependence of $Q(x)$ can be *a priori* isolated, $Q(x) = \exp(-im_Q t)\tilde{Q}(x)$. The reduced bispinor field $\tilde{Q}(x)$ describes a residual heavy quark motion inside the soft cloud; the heavy quark mass appears only in the form of powers of $1/m_Q$ (first noted in 1982).

Comprehensive reviews on the OPE-based heavy quark theory exist [711,1223,1235,1236]. There the reader will find exhaustive lists of references to original publications. Therefore, in my presentation below I will be brief, with a focus on a historical aspect, as I remember it, and limit myself to a few selected references.

In the early 1980s abundant data on the meson and baryon H_Q states started to appear. Theoretical understanding of the total decay rates beyond the free-quark calculations became a

[58] Hard gluons do play a role too. They have to be taken into account in the coefficient functions as will be mentioned In Sect. 5.7.10.

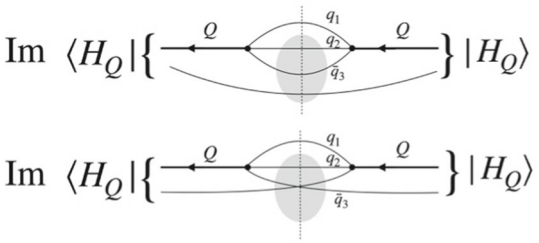


Fig. 113 $1/m_Q$ expansion for a H_Q weak inclusive decay rate (see Eq. (5.99)). Depicted are two operators, the leading $\bar{Q}Q$ and a subleading $(\bar{Q}q_3)(\bar{q}_3 Q)$. Both are sandwiched between the heavy hadron states $\langle H_Q|$ and $|H_Q\rangle$ and the decay rate is determined by the imaginary part. The grey area depicts the soft quark–gluon cloud. Adapted from Refs. [1233,1234]

major goal. This challenge paved the way to the beginning of the $1/m_Q$ expansion in H_Q hadron physics in the mid 1980s. The decay rate into an inclusive final state f can be written in terms of the imaginary part of a forward scattering operator (the so-called transition operator) evaluated to second order in the weak interactions [1233,1234],

$$\mathrm{Im}\hat{T}(Q\to f\to Q)=\mathrm{Im}\int d^4x\; i\,T\left(\mathcal{L}_W(x)\mathcal{L}_W^{\dagger}(0)\right) \quad (5.98)$$

where T denotes the time ordered product and $\mathcal{L}_W$ is the relevant weak Lagrangian at the normalization point $\mu\sim m_Q$. The factor $\exp(-im_Q t)$ mentioned above is implicit in Eq. (5.98). Descending to $\mu\ll m_Q$ one arrives at the OPE expansion

$$\begin{aligned}\Gamma(H_Q\to f)&=G_F^2|V_{\mathrm{CKM}}|^2 m_Q^5\sum_i \tilde{c}_i^{(f)}(\mu)\frac{\langle H_Q|O_i|H_Q\rangle_\mu}{2M_{H_Q}}\\ &\propto\Bigg[c_3^{(f)}(\mu)\frac{\langle H_Q|\bar{Q}Q|H_Q\rangle_{(\mu)}}{2M_{H_Q}}\\ &+c_5^{(f)}(\mu)m_Q^{-2}\frac{\langle H_Q|\bar{Q}\frac{i}{2}\sigma GQ|H_Q\rangle_{(\mu)}}{2M_{H_Q}}\\ &+\sum_i c_{6,i}^{(f)}(\mu)m_Q^{-3}\frac{\langle H_Q|(\bar{Q}\Gamma_i q)(\bar{q}\Gamma_i Q)|H_Q\rangle_{(\mu)}}{2M_{H_Q}}\\ &+\mathcal{O}(1/m_Q^4)+\cdots\Bigg],\end{aligned} \quad (5.99)$$

where Γ_i represent various combinations of the Dirac γ matrices, see also Table 6. In SVZ we dealt with the vacuum expectation values of relevant operators while in the heavy quark physics the relevant operators are sandwiched between H_Q states.

5.7.9 Applications

The expansion (5.99) allowed us to obtain [1233,1234] the first *quantitative* predictions for the hierarchies of the lifetimes of $Q\bar{q}$ mesons and Qqq baryons (Q was either c or b quark) in the mid-1980s – another spectacular success of the OPE-based methods. The dramatic story of η_c narrated in Sect. 5.7.7 repeated itself. With the advancement of experiment in the late 1990s, a drastic disagreement was allegedly detected in the ratio $\tau(\Lambda_b)/\tau(B_d)_{\mathrm{exp}}=0.77\pm0.05$ compared to the theoretical prediction

$$\tau(\Lambda_b)/\tau(B_d)_{\mathrm{theor}}=0.9\pm0.03$$

(e.g. [1223]). In the 2010s the Λ_b lifetime was remeasured shifting the above experimental ratio up to 0.93 ± 0.05. Hurrah!

In the mid-1980s, at the time of the initial theoretical studies of the H_c and H_b lifetime hierarchies [1233,1234], next to nothing was known about heavy baryons. Since then enormous efforts were invested in improving theoretical accuracy both in *mesons and baryons* in particular by including higher-dimension operators in the inverse heavy quark mass expansion and higher-order α_s terms in the OPE coefficients. The status of the Inverse Heavy Quark Mass Expansion (IHQME) for H_Q lifetimes as of 2014 was presented in the review [1237]. The advances reported there and in more recent years cover more precise determination of the matrix elements of four-quark operators via HQET sum rules [1238], calculations of the higher α_s corrections, in particular, α_s^3 corrections to the semileptonic b quark decay [1239], the first determination of the Darwin coefficient for non-leptonic decays [1240,1241], etc. Comparison with the current set of data on $\tau(H_c)$ can be found in [1242]. In this context I should also mention an impressive publication [1243] (see also references therein) which, in addition to a comprehensive review of the OPE-based analysis of the H_c lifetimes, acquaints the reader with a dramatic story of the singly charmed *baryon* hierarchy. Indeed, according to PDG-2018 the lifetime of Ω_c^0 is 69 ± 12 fs while PDG-2020 yields $\tau(\Omega_c^0)=268\pm24\pm10$ fs! The jump in the Ω_c^0 lifetime by a factor of 3 to 4 compared to the previous measurements was reported by LHCb [1244–1246].[59] With these new data the observed hierarchy of lifetimes changes: Ω_c^0 moves from the first place (the shortest living H_c baryon) to the third. The question arises whether the OPE-based theory can explain the current experimental situation $\tau(\Xi^0)<\tau(\Lambda_c^+)<\tau(\Omega_c^0)<\tau(\Xi^+)$. In [1243] it is argued that the answer is "yes, it is possible" (see Fig. 5 in [1243]) provided one takes into account $1/m_c^4$ contributions due to four-quark operators and α_s corrections in the appropriate coefficient functions.[60]

[59] Of course, this could happen only because (presumably) statistical and/or systematic errors in the previous measurements were grossly underestimated. It is also curious to note that 30 years ago Blok and I argued [1247, Secs. 4.2 and 6] that Ω_c^0 could be the longest living singly charmed baryon due to its ss spin-1 diquark structure.

[60] The four-quark operators introduced in [1233,1234] responsible for the Pauli interference yield corrections $O(1/m_c^3)$, see Eq. (5.99). The authors of [1243] go beyond this set.

I should emphasize that the theoretical accuracy in the H_c family is limited by the fact that the expansion parameter $\Lambda_{\rm QCD}/m_c$ is not small enough. Even including sub-leading contributions will hardly provide us with high-precision theoretical predictions. For H_c states IHQME at best provides us with a semi-quantitative guide. On the other hand, in the theory of H_b decays one expects much better precision.

5.7.10 Around 1990s and beyond

(1) Heavy quark symmetry when $\mathbf{m_Q \to \infty}$

The light-cloud interpretation as in Fig. 113 immediately implies that at zero recoil the (appropriately normalized) $B \to D$ formfactors reduce to unity. This is called the "small velocity (SV) limit theorem" [1248,1249]. The above"unification" is similar to the vector charge non-renormalization theorem at zero momentum transfer, say, for the $\bar{u}\gamma^\mu d$ current. The D and B masses are very far from each other. One has to subtract the mechanical part of the heavy quark mass in order to see that all dynamical parameters are insensitive to the substitution $Q_1 \leftrightarrow Q_2$ in the limit $m_{Q_{1,2}} \to \infty$, with the SV limit ensuing at zero recoil. Perhaps, this is the reason why it was discovered so late. The next step was made by Isgur and Wise who generalized this symmetry of the zero-recoil point by virtue of the Isgur-Wise function [1250,1251].

(2) HQET

Heavy quark effective theory which emerged in the 1990s [704,1252] formalizes and automates a number of aspects of the generic $1/m_Q$ expansion. In fact, it immediately follows from the construction similar to (5.99). Simplified rules of behavior proved to be very helpful for QCD practitioners in the subsequent development of various applications. In HQET the reduced field $\tilde{Q}$ is treated quantum-mechanically, its non-relativistic nature is built in, and the normalization point μ is $\ll m_Q$ from the very beginning.[61] Applying the Dirac equation to eliminate small (lower) components in favor of the large components it is easy to derive the expansion of $\mathcal{L}^0_{\rm heavy}$, up to terms $1/m_Q^2$,

$$\begin{aligned}\mathcal{L}^0_{\rm heavy} &= \bar{Q}(i\not{D} - m_Q)Q \\ &= \bar{Q}\frac{1+\gamma_0}{2}\left(1+\frac{(\boldsymbol{\sigma\pi})^2}{8m_Q^2}\right)\left[\pi_0 - \frac{1}{2m_Q}(\boldsymbol{\pi\sigma})^2 - \right. \\ &\left. - \frac{1}{8m_Q^2}\left(-(\vec{D}\vec{E}) + 2\boldsymbol{\sigma}\cdot\vec{E}\times\boldsymbol{\pi}\right)\right] \\ &\times\left(1+\frac{(\boldsymbol{\sigma\pi})^2}{8m_Q^2}\right)\frac{1+\gamma_0}{2}\,Q + \mathcal{O}\left(\frac{1}{m_Q^3}\right),\end{aligned} \tag{5.100}$$

where $\boldsymbol{\sigma}$ denote the Pauli matrices and

$$(\boldsymbol{\pi\sigma})^2 = \boldsymbol{\pi}^2 + \boldsymbol{\sigma}\vec{B},$$

$\vec{E}$ and $\vec{B}$ denote the background chromoelectric and chromomagnetic fields, respectively. Moreover, the operator π_μ is defined through

$$\begin{aligned} iD_\mu Q(x) &= e^{-im_Q v_\mu x_\mu}\left(m_Q v_\mu + iD_\mu\right)\tilde{Q}(x) \\ &\equiv e^{-im_Q v_\mu x_\mu}\left(m_Q v_\mu + \pi_\mu\right)\tilde{Q}(x) \end{aligned} \tag{5.101}$$

where v_μ is the heavy quark four-velocity. The set of operators presented in (5.100) plays the same basic role in $1/m_Q$ expansion as the set in Table 6 in SVZ sum rules.

In the remainder of this section I will briefly mention some classic problems with heavy quarks which were successfully solved in the given paradigm.

(3) CGG/BUV theorem

Up to order $1/m_Q^2$ all inclusive decay widths of the H_Q mesons coincide with the parton model results for the Q decay [1253,1254],

$$\Gamma = \Gamma_0\left(1 - \frac{\mu_\pi^2}{2m_Q^2}\right), \quad \mu_\pi^2 = \frac{1}{2M_{H_Q}}\langle H_Q|\bar{Q}\vec{\pi}^2 Q|H_Q\rangle \tag{5.102}$$

where Γ_0 is the parton model result. There are no corrections $O(1/m_Q)$. This is known as the CGG/BUV theorem.

(4) Spectra and line shapes

Lepton spectra in semileptonic H_Q decays were derived in [1255]. The leading corrections arising at the $1/m_Q$ level were completely expressed in terms of the difference in the mass of H_Q and Q. Nontrivial effects appearing at the order $1/m_Q^2$ were shown to affect mainly the endpoint region; they are different for meson and baryon decays as well as for beauty and charm decays.

The theory of the line shape in H_Q decays, such as $B \to X_s\gamma$ where X_s denotes the inclusive hadronic state with the s quark, resembles that of the Mössbauer effect. It is absolutely remarkable that for 10 years there were no attempts to treat the spectra and line shapes along essentially the same lines as it had been done in deep inelastic scattering (DIS) in the 1970s. Realization of this fact came only in 1994; technical implementation of the idea was carried out in [1256,1257], and [1258].

(5) Hard gluons

Hard-gluon contributions special for the heavy quark theory result in powers of the logarithms $\alpha_s \log\left(m_Q/\mu\right)$. They determine the coefficients c_i in Eq. (5.99) through the anomalous dimensions of the corresponding operators. They were

[61] I personally prefer to consider the heavy quark expansions directly in *full QCD* in the framework of the Wilson OPE bypassing the intermediate stage of HQET.

discovered in [1259,1260] and were called the *hybrid* logarithms. In HQET they are referred to as matching logarithms.
(6) In conclusion
Concluding the heavy quark portion I should add that Kolya Uraltsev (1957–2013), one of the major contributors in heavy quark theory died in 2013 at the peak of his creative abilities (see [1220]).

Concerning the OPE-based methods in QCD in general, I would like to make an apology to the many authors whose works have not been directly cited. The size limitations are severe. The appropriate references are given in the review papers listed in the text above.

Just for the record, a couple of reviews which are tangentially connected to the topic of the present article are given in Refs. [1261] and [1262].

5.7.11 Recent developments unrelated to the OPE-based methods

Quantum field theories from the same class as QCD are now experiencing dramatic changes and rapid advances in a deeper understanding of anomalies. I want to mention two crucial papers: [1263] and [1264]. The latter demonstrates that at $\theta = \pi$ there is a discrete 't Hooft anomaly involving time reversal and the center symmetry. It follows that at $\theta = \pi$ the vacuum cannot be a trivial non-degenerate gapped state.

5.8 Factorization and spin asymmetries

Jianwei Qiu

5.8.1 QCD factorization

Hadrons, such as the proton, neutron and pion, are relativistic bound states of strongly interacting quarks and gluons of QCD. Without being able to see any quark or gluon directly in isolation, owing to the color confinement of QCD, it has been an unprecedented intellectual challenge to explore and quantify the internal structure of hadrons in terms of their constituents, quarks and gluons, and the emergence of hadrons from quarks or gluons. Actually, the QCD color interaction is so strong at a typical hadronic scale $\mathcal{O}(1/R)$ with a hadron radius $R \sim 1$ fm that any scattering cross section with identified hadron(s) cannot be calculated fully in QCD perturbation theory.

QCD factorization [242] has been developed to describe high energy hadronic scattering with a large momentum transfer $Q \gg 1/R \sim \Lambda_{\rm QCD}$ by taking the advantage of the asymptotic freedom of QCD by which the color interaction becomes weaker and calculable perturbatively at short distances. QCD factorization provides a controllable and consistent way to *approximate* QCD contributions to *good* or factorizable hadronic cross sections by demonstrating

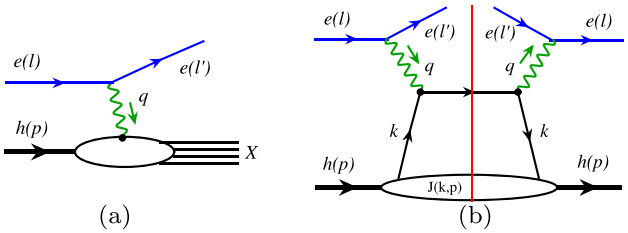


Fig. 114 **a** Sketch for scattering amplitude of inclusive DIS. **b** Leading order contribution to inclusive DIS cross section in its cut diagram notation

- all process-dependent nonperturbative contributions to these *good* cross sections are suppressed by powers of $\Lambda_{\rm QCD}/Q$, which could be neglected if the hard scale Q is sufficiently large,
- all factorizable nonperturbative contributions are process independent, representing the characteristics of identified hadron(s), and
- the process dependence of factorizable contributions is perturbatively calculable from partonic scattering at the short-distance.

With our ability to calculate the process-dependent short distance partonic scatterings perturbatively at the hard scale Q, the prediction of QCD factorization follows when cross sections with different hard scatterings but the same nonperturbative long-distance effect of identified hadron are compared. QCD Factorization also supplies physical content to these perturbatively uncalculable, but universal long-distance effects of identified hadrons by matching them to hadronic matrix elements of active quark and/or gluon operators, which could be interpreted as parton distribution or correlation functions of the identified hadrons, and allows them to be measured experimentally or by numerical simulation.

Inclusive scattering with one identified hadron
The deeply inelastic scattering (DIS) between a lepton e of momentum l and a hadron h of momentum p, $e(l)+h(p) \to e(l')+X$, as shown in Fig. 114a where l' is scattered lepton momentum and X represents all possible final states, is an inclusive scattering with one identified hadron. With a large momentum transfer, $q = l-l'$ and $Q \equiv \sqrt{-q^2} \gg \Lambda_{\rm QCD}$, the DIS experiment at SLAC in 1969 discovered the point-like spin-1/2 partons/quarks inside a proton [110], which helped the discovery and formulation of QCD.

For inclusive DIS with two characteristic scales: $Q(\gg \Lambda_{\rm QCD})$ and $\Lambda_{\rm QCD}$, QCD factorization is to consistently separate QCD dynamics taking place at these two distinctive scales by examining scattering amplitudes in terms of general

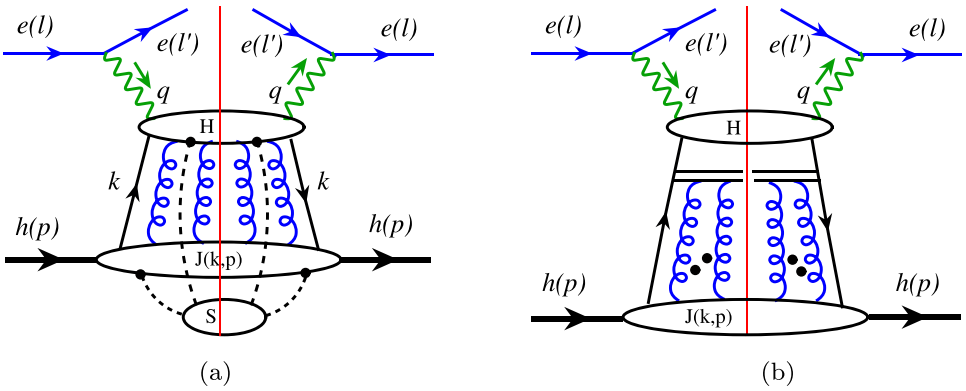


Fig. 115 **a** Pinch surface for inclusive DIS with collinear and longitudinally polarized gluons (curly lines) and soft gluons (dashed lines). **b** Leading power factorized contribution to inclusive DIS with all collinear and longitudinally polarized gluons detached from the hard part H and reconnected to the gauge links

properties of Feynman diagrams in QCD perturbation theory. This leads to a factorization formalism, which is an approximation up to corrections suppressed in powers of $\Lambda_{\rm QCD}/Q$. For example, considering the leading order (LO) contribution to the inclusive DIS, as presented in Fig. 114b in its cut diagram notation, graphical contributions to the cross sections are represented by the scattering amplitude to the left of the final state cut (the red thin line) and the complex conjugate amplitude to the right. The scattering $\hat{\sigma}^{\rm LO}(Q,k)$ between the lepton of momentum l and a quark (or a parton) of momentum k, is taking place at the hard scale Q, while the dynamics describing the quark inside the hadron, $J(k,p)$, is at the hadronic scale $1/R \sim \Lambda_{\rm QCD}$. The validity of such perturbative QCD factorization requires the suppression of quantum interference between the dynamics taking place at these two different momentum scales. This in turn requires that the dominant contributions to the factorized formalism should necessarily come from the phase space where the active parton(s) linking the dynamics at two different scales are forced onto their mass shells, and are consequently long-lived compared to the time scale of the hard collision at the scale Q. This requirement is naturally satisfied for the LO contribution in Fig. 114b,

$$\begin{aligned}\sigma_{\rm DIS}^{\rm LO} &\propto \int d^4k \left[\hat{\sigma}^{\rm LO}(Q,k)\frac{1}{k^2+i\varepsilon}J(k,p)\frac{1}{k^2-i\varepsilon}\right]\\ &\approx \int \frac{dk^+}{2k^+}d^2k_T\,\hat{\sigma}^{\rm LO}(Q,\hat{k})\\ &\quad\times\int dk^2\frac{1}{k^2+i\varepsilon}J(k,p)\frac{1}{k^2-i\varepsilon}+\mathcal{O}\left[\frac{\Lambda_{\rm QCD}^2}{Q^2}\right]\end{aligned} \tag{5.103}$$

where the light-quark mass was neglected, and the active quark of momentum k is perturbatively pinched to be on-shell, $k^2 \approx \hat{k}^2 = 0$ with

$$\hat{k} = \left(k^+, \frac{k_T^2}{2k^+}, \vec{k}_T\right)$$

in the notation of light-cone coordinates, leading to a factorization formalism (see Eq. (5.103)) with all perturbatively pinched poles absorbed into the nonperturbative function of the identified hadron.

However, beyond the LO inclusive DIS, all internal loop-momentum integrals to any scattering amplitude are defined by contours in complex momentum space, and it is only at momentum configurations where some subset of loop momenta are pinched that the contours are forced to or near mass-shell poles that correspond to long-distance behavior. The importance of such *pinched surfaces* in multidimensional momentum space was identified in the Libby-Sterman analysis [1265,1266] that categorized all loop momenta into three groups: hard, collinear, and soft, along with the *reduced diagrams* by contracting off-shell lines to points, from which factorization formalisms can be derived. As shown in Fig. 115a for inclusive DIS, the identified hadron is associated with a group of collinear parton lines, and at the leading power, one physically polarized collinear parton plus infinite longitudinally polarized collinear gluons (curly lines) link the identified hadron to the hard part, H, in which all parton lines are off-shell by the hard scale Q. At the same time, the soft gluon lines (dashed lines in Fig. 115a) can attach to both the hard and collinear lines of the identified hadron. Since all parton propagators in H are off-shell by Q, a soft gluon attachment to any of these lines in H is necessarily to increase the number of off-shell propagators in H, and effectively suppresses the hard part by an inverse power of Q, making the contribution power suppressed. Therefore, we do not need to consider soft contributions to the inclusive DIS cross section at the leading power in $1/Q$ expansion.

The collinear and longitudinally polarized gluons have their polarization vectors proportional to their momenta in a covariant gauge. By applying the Ward Identity, all attachments of collinear and longitudinally polarized gluons to the hard part H can be detached and reconnected to the gauge link pointing to the "−" light-cone direction if the identified hadron is moving in the "+" light-cone direction [242,1267], as sketched in Fig. 115b. After taking the leading power contribution from the spinor trace of the active quark line in Fig. 115b [1267,1268], the inclusive DIS cross section at the leading power can be factorized as [1269–1271]

$$E'\frac{d\sigma^{\rm DIS}_{eh\to eX}}{d^3l'}(l,p;l') = \sum_{f=q,\bar{q},g}\int dx\ \phi_{f/h}(x,\mu^2) \times E'\frac{d\hat{\sigma}_{ef\to eX}}{d^3l'}(l,\hat{k};l',\mu^2) + \mathcal{O}\left[\frac{\Lambda^2_{\rm QCD}}{Q^2}\right] \tag{5.104}$$

where $\hat{k} \equiv xp^+$, $l'_T \sim Q \gg \Lambda_{\rm QCD}$, and $E'd\hat{\sigma}_{ef\to eX}/d^3l'$ is the short-distance part of DIS cross section on a parton state of flavor f and collinear momentum fraction x of the colliding hadron, with its long-distance contributions to the cross section systematically absorbed into the nonperturbative functions $\phi_{f/h}(x,\mu^2)$, which are defined in terms of hadronic matrix elements of active parton operators [1272]. For example, for an unpolarized active quark,

$$\phi_{q/h}(x,\mu^2) = \int\frac{d\xi^-}{2\pi}e^{ixp^+\xi^-}\langle h(p)|\overline{\psi}_q(0)\frac{\gamma^+}{2} \times\mathcal{W}_{[0,\xi^-]}\psi_q(\xi^-)|h(p)\rangle, \tag{5.105}$$

where $\mathcal{W}_{[0,\xi^-]} = \mathcal{P}\exp\left[ig\int_0^{\xi^-}d\eta^- A^+(\eta^-)\right]$ is the gauge link. The $\phi_{f/h}(x,\mu^2)$ carries nonperturbative information of the identified hadron, and is referred as an universal parton distribution function (PDF) for finding a parton of flavor f inside a colliding hadron h, carrying its momentum fraction x, probed at a hard factorization scale $\mu \sim Q$. PDFs are discussed in more detail in Sect. 10.2.

With the precise definition of $\phi_{f/h}(x,\mu^2)$, the QCD factorization formalism, such as the one in Eq. (5.104), provides a systematic way to calculate the short-distance partonic scattering, $E'd\hat{\sigma}_{ef\to eX}/d^3l'$, in QCD perturbation theory. By applying the factorization formalism in Eq. (5.104) to a parton state of flavor f, $|h(p)\rangle \to |f(p)\rangle$, we can use perturbation theory to calculate

$$E'\frac{d\hat{\sigma}_{ef\to eX}}{d^3l'}$$

order-by-order in powers of the strong coupling constant α_s by perturbatively calculating the DIS cross section on a parton of flavor f on the left of Eq. (5.104), and PDFs of the same parton on the right with the collinear divergence regularized. QCD factorization ensures that the regularized collinear divergence of the partonic scattering cross section on the left-hand-side of Eq. (5.104) will be exactly cancelled by the regularized collinear divergence of the PDFs of the same parton on the right [1267].

The inclusive DIS cross section can be physically measured in experiments and should not depend on how we describe it in terms of QCD factorization, or the choice of factorization scale μ. That is, we require

$$d\sigma_{eh\to eX}/d\log\mu^2 = 0,$$

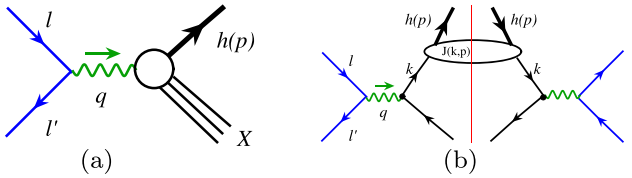


Fig. 116 **a** Sketch for scattering amplitude of inclusive single hadron production in high energy e^+e^- collisions. **b** Leading order contribution to inclusive single hadron production in its cut diagram notation

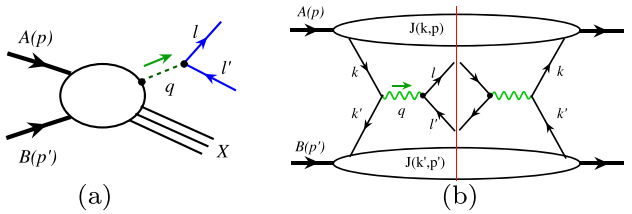


Fig. 117 **a** Sketch for scattering amplitude of Drell–Yan production of a massive lepton pair. **b** Leading order contribution to the Drell–Yan cross section in its cut diagram notation

which implies evolution equations of PDFs, known as the DGLAP equations [87,230,232,233]

$$\frac{d\phi_{f/h}(x,\mu^2)}{d\log\mu^2} = \sum_{f'}\int_x^1\frac{dx'}{x'}P_{f/f'}\left(\frac{x}{x'},\alpha_s(\mu^2)\right) \times\phi_{f'/h}(x',\mu^2) \tag{5.106}$$

where the evolution kernels $P_{f/f'}(x/x',\alpha_s(\mu^2))$ are calculable in perturbative QCD when the strong coupling constant $\alpha_s(\mu)$ is sufficiently small [234,235]. Although PDFs are nonperturbative, their factorization scale dependence is a QCD prediction, which has been confirmed to great accuracy [663,664].

Another example of a factorizable inclusive cross section with one identified hadron is single-inclusive hadron production in high energy electron–positron collision, $e^-(l) + e^+(l') \to h(p) + X$ with an observed hadron energy $E_p \gg \Lambda_{\rm QCD}$, as sketched in Fig. 116a. Like the inclusive DIS in Eq. (5.103), the active parton momentum k, in Fig. 116b, linking the hard e^+e^- annihilation that produces this active parton and describes how it hadronizes into the observed hadron, is perturbatively pinched to its mass-shell, which is necessary for the factorization. For the leading power contribution beyond the LO in Fig. 116b, similar to inclusive DIS, we do not need to worry about soft interactions between the hard part and the collinear partons along the direction of the produced hadron. By applying the Ward Identity, in the same way as in the factorization of inclusive DIS, the attachment of collinear and longitudinally polarized gluons from the observed hadron to the hard part, H, can be detached and reconnected to the gauge link to become a part of the nonperturbative, but universal, fragmentation functions (FFs) of

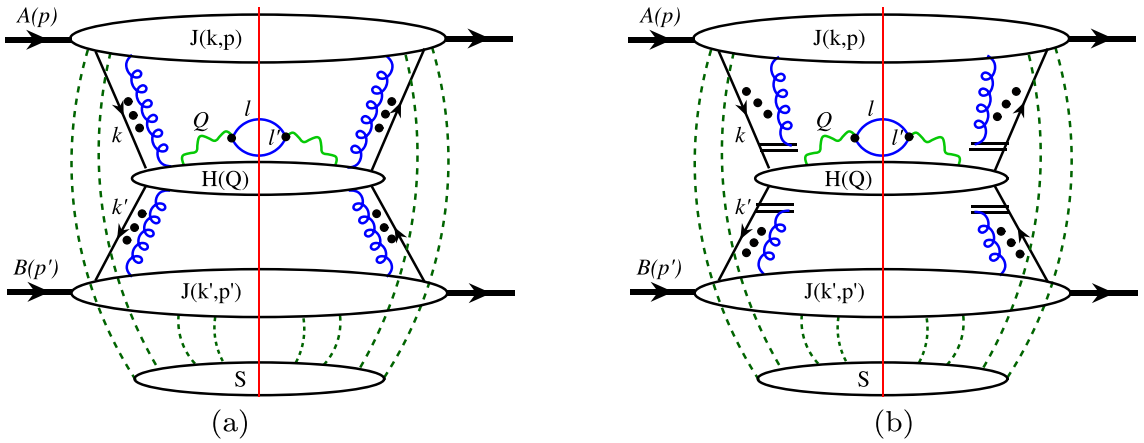


Fig. 118 **a** Sketch for the leading QCD pinch surface for Drell–Yan production of lepton pair with collinear and longitudinally polarized gluons in curly lines and soft gluons in dashed lines. **b** QCD contribution to Drell–Yan process with all collinear and longitudinally polarized gluons detached from the hard part and reconnected to the gauge lines

the identified hadron, leading to the factorization formalism,

$$E_p \frac{d\sigma_{e^+e^-\to hX}}{d^3p}(s,\, p) = \sum_f \int \frac{dz}{z^2} D_{h/f}(z,\mu^2) \times E_k \frac{d\hat{\sigma}_{e^+e^-\to \hat{k}X}}{d^3\hat{k}}(s,\hat{k},\mu^2) + \mathcal{O}\left[\frac{\Lambda^2_{\rm QCD}}{Q^2}\right] \tag{5.107}$$

where active parton momentum is $\hat{k} = p/z$,

$$\sqrt{s} = \sqrt{(l+l')^2}$$

is the collision energy, and $D_{h/f}(z,\mu^2)$ is the FF to find a hadron h emerged from a produced parton of flavor f while carrying the parton's momentum fraction z [1272]. The fact that such a physical cross section should not depend on how we factorized implies evolution equations for the FFs, like DGLAP for PDFs.

Extracting the universal PDFs and FFs from experimental data – exploiting the QCD factorization formalisms which involve one identified hadron in Eqs. (5.104) and (5.107) – is a challenging inverse problem. Although the scale dependence of PDFs and FFs is a prediction of QCD dynamics, measurements of such cross sections with one identified hadron are not sufficient to disentangle the flavor and momentum fraction dependence of all PDFs and FFs, which is necessary for the predictive power of the QCD factorization approach to describe high energy hadronic cross sections.

Inclusive scattering with two identified hadrons

The Drell–Yan (DY) production of lepton pairs via a vector boson in hadron–hadron collisions, $A(p) + B(p') \to V(q) + X$ with $V(q)[= \gamma^*, W/Z, H^0, \ldots] \to l + l'$, as sketched in Fig. 117a, is an ideal example of the study of QCD factorization for inclusive observables with two identified hadrons [242].

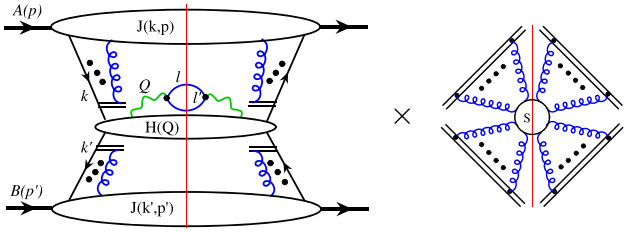


Fig. 119 Sketch for factorized Drell–Yan production of a massive lepton pair at the leading power with all soft gluon interactions factorized into a multiplicative soft factor

From the LO contribution in Fig. 117b, both active partons (quark or antiquark) of momentum k and k' coming from colliding hadrons $A(p)$ and $B(p')$, respectively, are perturbatively pinched to their mass-shell, which is necessary for being able to factorize the nonperturbative hadronic information of colliding hadrons from the hard collision to produce the massive lepton pairs. Beyond the LO, each colliding hadron is associated with a group of collinear partons, and for the leading power contribution, only one physically polarized active parton plus infinite collinear and longitudinally polarized gluons from each hadron should attach to the hard part, H, with the remaining collinear partons forming a (spectator) jet function, which is the same as the inclusive scattering with one identified hadron. The key difference for QCD factorization of inclusive scattering with two identified hadrons from that with one hadron, according to the Libby-Sterman analysis [1265,1266], is the soft interaction between the collinear partons of two different hadrons, as shown by the dashed lines in Fig. 118a. Still the soft interaction between the collinear partons and the hard part can be neglected when calculating the leading power contributions. However, these long-distance soft interactions between

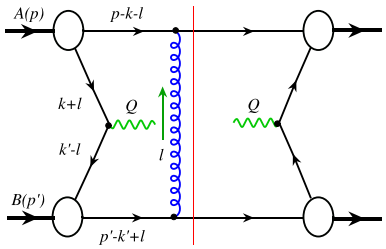


Fig. 120 Sample diagram responsible for soft gluon interaction to have its momentum pinched in the Glauber region

hadrons have the potential to break the universality of the factorizable nonperturbative contribution from each identified hadron, and invalidate the predictive power of the QCD factorization approach for studying hadronic cross sections with identified hadrons.

When the colliding hadrons $A(p)$ and $B(p')$ are moving in the $+z$ and $-z$ direction, respectively, the factorization of collinear and longitudinally polarized gluons from the hard part H is effectively the same as what was done for the case of single identified hadron. Since collinear and longitudinally polarized gluons have their polarization vectors proportional to their momenta in a covariant gauge, by applying the Ward Identity all collinear and longitudinally polarized gluons from hadron $A(p)$ can be detached from the hard part and reconnected to the gauge link in the "$-$" light-cone direction, while those from hadron $B(p')$ can be reconnected to the gauge link in the "$+$" light-cone direction, as sketched in Fig. 118b.

In order to achieve the factorization, we need to get rid of the soft gluon interactions, the dashed lines in Fig. 118b. If we scale collinear parton momenta from colliding hadron A, $k_i = (k_i^+, k_i^-, k_i^T) \sim (1, \lambda^2, \lambda)Q$ with $\lambda \sim \mathcal{O}(\Lambda_{\rm QCD}/Q)$, we maintain $k_i^2 \sim \mathcal{O}(\lambda^2 Q^2) \to 0$ as the loop momenta approach to the pinch surface. If we can choose soft-gluon loop momenta to have the scaling behavior, $l_s \sim (\lambda_s, \lambda_s, \lambda_s)Q$, where $\lambda_s \sim \lambda^2$ (or λ) so that all components vanish at the same rate. We then have $(k_i + l_s)^2 \sim 2k_i^+ l_s^- \sim \mathcal{O}(\lambda^2/Q^2)$. That is, in a covariant gauge we only need to keep the "$-$" components of the soft gluon momenta flowing into the jet of collinear partons from the colliding hadron A. Correspondingly, the Lorentz indices connecting to the soft gluons from the jet function $J(k, p)$ of hadron A will be in the "+" direction. Therefore, we can use the Ward Identity to detach the soft gluons from the jet of collinear partons from colliding hadron A and reconnect them into a gauge link or an eikonal line. Applying the same reasoning with the role of the "$\pm$" components switched, we can detach all soft gluon interactions to the jet of collinear partons from colliding hadron B, and to factorize all soft gluon interactions with two colliding hadrons into an overall soft factor, as shown in Fig. 119.

However, this factorization can fail if the soft gluon momenta are trapped in the Glauber region. In this region the "$\pm$" components of the soft gluons are small compared to their transverse components, i.e.$l_s^\pm / l_s^T \sim \mathcal{O}(\lambda)$, so that we cannot neglect the transverse components, keeping only one "$+$" or "$-$" components [242]. It is the soft-gluon interaction between the spectators of two colliding hadrons that can trap the $\pm$ components of the soft gluon momenta in the Glauber region. For example, in Fig. 120, the pair of propagators of momenta, $p-k-l$ and $k+l$, pinches the "$-$" component of l to be, $l^- \propto l_T^2$, while the pair of propagators of momenta, $p'-k'+l$ and $k'-l$, pinches the "$+$" component of l to be, $l^+ \propto l_T^2$, such that the soft gluon interaction between two jets of collinear partons from the colliding hadrons is pinched in the Glauber region; in this case the leading soft gluon interactions could break the universality of PDFs and the predictive power of the QCD factorization approach.

Removal of the trapped Glauber gluons might be the most difficult part of the QCD factorization proof [242]. It was achieved in three key steps: (1) all poles in one-half plane cancel after summing over all final-states (no more pinched poles), (2) all $l_s^\pm$-type integrations can be deformed out of the trapped soft region, and (3) all leading-power spectator interactions can be factorized and summed into an overall unitary soft factor of gauge links (or eikonal lines) as argued above and shown in Fig. 119. The soft factor is process independent and made of four gauge links, along the light-cone directions conjugated to the directions of two incoming hadrons in the scattering amplitude, and the two in the complex conjugate scattering amplitude, respectively. For the collinear factorization, the soft factor = 1 due to the unitarity, and we have the corresponding factorization formalism for inclusive Drell–Yan production at the leading power,

$$\frac{d\sigma^{(\rm DY)}_{A+B\to ll'+X}}{dQ^2 dy} = \sum_{ff'} \int dx\, dx'\, \phi_{f/A}(x,\mu)\, \phi_{f'/B}(x',\mu) \times \frac{d\hat{\sigma}_{f+f'\to ll'+X}(x,x',\mu,\alpha_s)}{dQ^2 dy} + \mathcal{O}\left[\frac{\Lambda^2_{\rm QCD}}{Q^2}\right], \tag{5.108}$$

where $\sum_{ff'}$ runs over all parton flavors including quark and antiquark, as well as gluon.

To help separate the flavor dependence of PDFs, the lepton–hadron semi-inclusive DIS (SIDIS), $e(l) + h(p) \to e(l') + h'(p') + X$, as shown in Fig. 121a, is another example of QCD factorization with two identified hadrons. From the LO contribution in Fig. 121b, both active partons of momentum k and k' are perturbatively pinched to their mass-shell, leading to a potential factorization of PDF from colliding hadron and FF of the fragmenting parton to the observed hadron. Beyond the LO, like the Drell–Yan process, there could be soft interactions between the jet of collinear par-

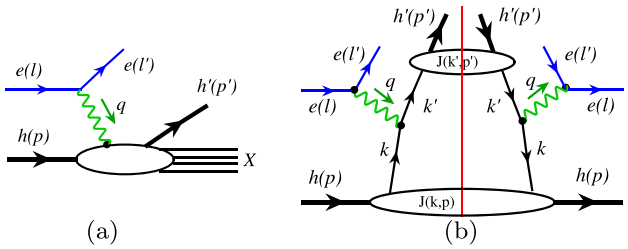


Fig. 121 **a** Sketch for scattering amplitude of lepton–hadron SIDIS. **b** Leading order contribution to SIDIS cross section in its cut diagram notation

tons of the hadron h and the jet of collinear partons along the direction of observed hadron h'.

Integrating over the transverse momentum of the observed final-state hadron to keep the SIDIS as a process with a single hard scale Q, and following the same factorization arguments for inclusive Drell–Yan processes, the SIDIS cross section can be factorized as

$$E'\frac{d\sigma^{\rm SIDIS}_{eh\to eh'X}}{d^3l'dz}(l,\,p;\,l',\,z) = \sum_{f,f'=q,\bar{q},g}\int dz'\,dx\;D_{h'/f'}(z',\,\mu^2)\,\phi_{f/h}(x,\,\mu^2) \times E'\frac{d\hat{\sigma}_{ef\to ef'X}}{d^3l'dz'}(l,\,\hat{k};\,l',\,z',\,\mu^2) + \mathcal{O}\left[\frac{\Lambda^2_{\rm QCD}}{Q^2}\right] \tag{5.109}$$

where $\hat{k} = xp$, $z' = p'/k'$ and $z = p\cdot p'/p\cdot q$.

Inclusive jet production in hadronic collisions: $A(p) + B(p') \to \sum_j J_j(p_j) + X$ is another observable with two identified hadrons although many hadrons were measured in the final-state when jets were constructed. When final-state jets are well-separated, the cross section for jets with large transverse energy has the same factorized formula as that in Eq. (5.108) except the perturbatively calculated hard part, $\hat{\sigma}_{ff'\to ll'X}$ is replaced by the corresponding short-distance hard part, $\hat{\sigma}_{ff'\to \rm Jet}$ [1273]:

$$\frac{d\sigma^{\rm (Jet)}_{A+B\to \rm Jet+X}}{dp_T dy} = \sum_{ff'}\int dx\,dx'\,\phi_{f/A}(x,\mu)\,\phi_{f'/B}(x',\mu) \times\frac{d\hat{\sigma}_{f+f'\to \rm Jet+X}(x,x',\mu,\alpha_s)}{dp_T dy} = \sum_{ff'}\int dx\,dx'\,\phi_{f/A}(x,\mu)\,\phi_{f'/B}(x',\mu) \times\left[\sum_c\int\frac{dz}{z}J_c(z,\,p_T R,\,\mu)\frac{d\hat{\sigma}_{f+f'\to c+X}}{dp_{cT}dy_c} + \tilde{\sigma}(p_T,\,y)\right] \tag{5.110}$$

where p_T and y are the transverse momentum and rapidity of the observed jet, respectively. Like all perturbatively calculable hard parts of QCD factorization, the hard part for the jet production, $\hat{\sigma}_{ff'\to \rm Jet}$ is process-dependent, depending on whether the jet is produced in hadron–hadron or lepton–hadron collisions, as well as the choice of the jet algorithms. In Eq. (5.110), the process-dependent short-distance hard part for the jet production was reorganized into a process-independent jet function, J_c from a single parton of flavor c, leaving all process-dependence into the production of this parton, $\hat{\sigma}_{f+f'\to c+X}$ and $\tilde{\sigma}(p_T,\,y)$ which might be neglected if logarithms of the jet production dominates [1274].

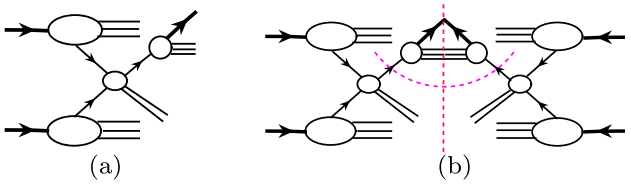


Fig. 122 **a** Sketch for scattering amplitude of hadronic production of single hadron at large transverse momentum. **b** Its contribution to the cross section in the cut diagram notation

Inclusive scattering with three identified hadrons

Inclusive single hadron production at large transverse momentum p_T in hadronic collisions: $A(p)+B(p')\to h(p_h)+X$ is a well-measured observable involving three identified hadrons, as shown in Fig. 122. Due to the additional identified hadron in the final-state, the unitarity sum of final-state hadrons used to prove the factorization of DY-type two-hadron observables needs to be modified.

Luckily, because of the large p_T of the observed final-state hadron, the potentially dangerous gluon interactions between the observed hadron and the spectators of colliding hadrons are suppressed by the power of $1/p_T$, and the leading power pQCD factorization does hold [1275],

$$\frac{d\sigma_{AB\to hX}(p,\,p,\,p_h)}{dy\,dp_T^2} = \sum_{f,f',c}\int\frac{dz}{z^2}\,dx\,dx'\,D_{h/c}(z,\,\mu^2) \times\phi_{f/A}(x,\,\mu^2)\,\phi_{f'/B}(x',\,\mu^2) \times\frac{d\hat{\sigma}_{ff'\to cX}(x,\,x',\,p_c = p_h/z)}{dy_c\,dp_{cT}^2}. \tag{5.111}$$

With proper PDFs and FFs, the NLO pQCD calculations for single hadron production gave an excellent description of RHIC data [1276]. However, the same formalism consistently underestimates the production rate at the fixed target energies [1277]. It was shown that high order corrections at the fixed target energies are very important, and the threshold resummation significantly improves the comparison between the theory and experimental data [1278].

QCD global analysis and predictive power

Much of the predictive power of QCD factorization for cross sections involving identified hadron(s) relies on the universality of the PDFs and/or FFs and our ability to solve the inverse problem to demonstrate the existence of one set of

PDFs and FFs that are capable of describing all data of *good* (e.g. factorizable) cross sections with properly calculated short-distance partonic scattering cross sections in QCD perturbation theory.

The QCD global analysis is a program to test the consistency of QCD factorization by fitting all existing data from high energy scatterings with universal PDFs and/or FFs and corresponding factorization formalisms, from which the best set of PDFs and/or FFs will be extracted. The QCD factorization formalism has been extremely successful in interpreting high energy experimental data from all facilities around the world, covering many orders in kinematic reach in both parton momentum fraction x and momentum transfer of the hard collision Q, and as large as 15 orders of magnitude in difference in the size of observed scattering cross sections, which is a great success story of QCD and the Standard Model at high energy. It has given us the confidence and the tools to discover the Higgs particle in proton–proton collisions [139,140], and to search for new physics [1279].

QCD factorization for two-scale observables

The hard probe with a single large momentum transfer Q ($\gg 1/R$) is so localized in space that it is not very sensitive to the details of confined three-dimensional (3D) internal structure of the colliding hadron, in which a confined parton should have a characteristic transverse momentum scale $\langle k_T\rangle \sim 1/R \ll Q$ and an uncertainty in transverse position $\langle b_T\rangle \sim R \gg 1/Q$. Recently, new and more precise data are becoming available for *two-scale* observables with a hard scale Q to localize the collision to probe the partonic nature of quarks and gluons along with a soft scale to be sensitive to the dynamics taking place at $\mathcal{O}(1/R)$. At the same time, theory has made major progresses in the development of QCD factorization formalism for two types of two-scale observables, distinguished by their inclusive or exclusive nature, which enables quantitative matching between the measurements of such two-scale observables and the 3D internal partonic structure of a colliding hadron.

For inclusive two-scale observables, one well-studied example is the Drell–Yan production of a massive boson that decays into a pair of measured leptons in hadron–hadron collisions as a function of the pair's invariant mass Q and transverse momentum q_T in the Lab frame [1280]. When $Q \gg q_T \gtrsim 1/R$, the measured transverse momentum of the pair is sensitive to the transverse momenta of the two colliding partons before they annihilate into the massive boson, providing the opportunity to extract the information on the active parton's transverse motion at the hard collision, which is encoded in transverse momentum dependent (TMD) PDFs (or simply, TMDs), $\phi_{f/h}(x, k_T, \mu^2)$ [1280].

Like PDFs, TMDs are universal distribution functions describing how a quark (or gluon) with a momentum fraction x and transverse momentum k_T interacts with a colliding hadron of momentum p with $xp \sim \mu \sim Q \gg k_T$. Another well-studied example is the SIDIS when the produced hadron is almost back-to-back to the scattered lepton in the Lab frame, or in the Breit frame, the transverse momentum of the produced hadron p_{hT} is much smaller than the hard scale Q [1281,1282].

A necessary condition for QCD factorization of observables with identified hadron(s) is that the active parton linking the process-dependent short-distance dynamics and the process-independent nonperturbative physics of identified hadron(s) is perturbatively pinched to its mass-shell so that it is long-lived compared to the time scale of the hard collision. In this case the quantum interference between the perturbatively calculable hard collisions at the hard scale Q and the process-independent part of leading nonperturbative information of the identified hadron(s) is strongly suppressed by the power of $\Lambda_{\rm QCD}/Q$. The pinch does not require the active parton's momentum to be collinear to the hadron momentum. The necessary condition is satisfied if the active parton momentum has a transverse component with $\langle k_T\rangle \ll Q$; the same condition that should be satisfied by the TMD factorization of Drell–Yan and SIDIS process for the leading power contribution in q_T/Q or p_{hT}/Q, respectively. Although this condition is not necessarily sufficient, the TMD factorization for Drell–Yan process at the leading power of $q_T/Q \to 0$ was justified [1267,1280], and the same for the SIDIS at leading power of p_{hT}/Q [1283–1285]. More discussion on the impact of TMD factorization for the spin asymmetries will be given in Sect. 5.8.2.

Without breaking the colliding hadron, the exclusive observables could provide different aspects of the hadron's internal structure. Exclusive lepton–nucleon scattering with a virtual photon of invariant mass $Q \gg 1/R$ could provide various two-scale observables, such as the deeply virtual Compton scattering (DVCS) [1286], where the hard scale is Q and the soft scale is $t \equiv (p-p')^2$. When $Q \gg \sqrt{|t|}$, such two-scale exclusive processes are dominated by the exchange of an active $q\bar{q}$ or gg pair and can be systematically treated using the QCD factorization approach; factorized in terms of generalized PDFs or GPDs [1287–1290]. Recently, a new class of *single diffractive hard exclusive processes (SDHEP)* was introduced [1291,1292]. This approach is not only sufficiently generic to cover all known processes for extracting GPDs, but also well-motivated for the search of new processes for the study of GPDs. It was demonstrated that many of those new processes can be factorized in terms of GPDs and could provide better sensitivity to the parton momentum fraction x dependence of GPDs.

5.8.2 Spin asymmetries

A measured cross section is always a positive and classical probability even though its underlying dynamics could

be sensitive to quantum effects. On the other hand, a spin asymmetry, defined to be proportional to a difference of two cross sections with one (or more) spin vector(s) flipped, can probe QCD dynamics that a spin-averaged cross section is not sensitive to, and provide a better chance to explore the dynamics of quantum effects. It also provides opportunities to explore the origin of proton spin by carrying out scattering experiments with polarized protons.

Quark and gluon contributions to proton spin

The leading power collinear factorization formalisms can also apply to asymmetries of cross sections between two longitudinally polarized particles [1267]. Instead of measuring nonperturbative PDFs of a hadron, the double longitudinal spin asymmetry

$$A_{LL} \equiv \frac{\Delta\sigma}{\sigma} = \frac{\sigma(++) - \sigma(+,-)}{\sigma(+,+) + \sigma(+,-)}, \tag{5.112}$$

where $\pm$ indicates the helicity of the active parton compared to the longitudinal spin direction of the colliding particle, is sensitive to the active parton's helicity distribution inside a polarized colliding hadron. The double longitudinal spin-dependent cross sections, $\Delta\sigma$ is given by the same factorization formalisms introduced in the Sect. 5.8.1 with the spin-averaged collinear PDFs replaced by corresponding helicity distributions,

$$\begin{aligned}\phi_{f/h}(x,\mu^2) \to\ & \Delta\phi_{f/h}(x,\mu^2) \\ =\ & \frac{1}{2}\Big[\phi_{+/+}(x,\mu^2) - \phi_{-/+}(x,\mu^2)\Big].\end{aligned}$$

The same leading power collinear factorization formalisms introduced in the Sect. 5.8.1 can also apply to parity violating single longitudinal spin asymmetries of cross sections between one unpolarized and one longitudinally polarized particles,

$$A_L \equiv \frac{\sigma(+) - \sigma(-)}{\sigma(+) + \sigma(-)}. \tag{5.113}$$

The single longitudinal spin-dependent cross section, $\Delta\sigma = \sigma(+) - \sigma(-)$ with spin direction of the polarized parton flipped is also given by the same factorization formalisms by replacing one of the spin-averaged collinear PDFs, corresponding to the hadron that is replaced by a polarized colliding particle, by corresponding helicity distribution. With the flavor sensitivities of the weak interaction, the single longitudinal spin asymmetries measured by the RHIC spin program have provided important information on the flavor separation of quark helicity distributions [1276,1293].

The double and single longitudinal spin asymmetries, defined in Eqs. (5.112) and (5.113), respectively, have been studied in both hadron–hadron collisions at RHIC [1276] and lepton–hadron collisions [1294,1295], and will be a major program at the future EIC [1293].

After over 30 years since the discoveries made by the EMC collaboration, many polarized experiments have been carried out worldwide, the RHIC spin program in particular. From the range of momentum fraction x accessible by existing experimental data, we learned that the proton spin gets about 30% from quark helicity and 40% from gluon helicity. The rest could come from the region of x that we have not been able to explore and/or from the orbital or transverse motion of quarks and gluons inside the bound proton [1293]. (See the discussion in Sect. 10.3.)

Double transverse-spin asymmetries

The double transverse spin asymmetries are,

$$A_{NN} = \frac{\sigma(\uparrow,\uparrow) - \sigma(\uparrow,\downarrow)}{\sigma(\uparrow,\uparrow) + \sigma(\uparrow,\downarrow)},$$

where $\uparrow$ and $\downarrow$ indicate the direction of spin vectors transverse to the momentum direction of the colliding particles. Since QCD factorization requires that the factorized short-distance dynamics is not sensitive to the details of hadronic physics, the spin asymmetries are proportional to the difference of hadronic matrix elements of parton fields with the hadron spin flipped,

$$\begin{aligned}A \propto\ & \sigma(Q,\vec{s}) - \sigma(Q,-\vec{s}) \\ \propto\ & \langle p,\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,\vec{s}\rangle - \langle p,-\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,-\vec{s}\rangle.\end{aligned} \tag{5.114}$$

The parity and time-reversal invariance of QCD requires

$$\begin{aligned}&\langle p,\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,\vec{s}\rangle \\ &\quad = \langle p,-\vec{s}|\mathcal{P}\mathcal{T}\mathcal{O}^\dagger(\psi_q, A_g^\mu)\mathcal{T}^{-1}\mathcal{P}^{-1}|p,-\vec{s}\rangle.\end{aligned} \tag{5.115}$$

Therefore, only partonic operators $\mathcal{O}(\psi_q, A_g^\mu)$ satisfying

$$\begin{aligned}&\langle p,-\vec{s}|\mathcal{P}\mathcal{T}\mathcal{O}^\dagger(\psi_q, A_g^\mu)\mathcal{T}^{-1}\mathcal{P}^{-1}|p,-\vec{s}\rangle \\ &\quad = \pm\langle p,-\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,-\vec{s}\rangle\end{aligned} \tag{5.116}$$

or

$$\langle p,\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,\vec{s}\rangle = \pm\langle p,-\vec{s}|\mathcal{O}(\psi_q, A_g^\mu)|p,-\vec{s}\rangle \tag{5.117}$$

contribute to the factorizable spin asymmetries. Those operators that lead to a "+" sign should contribute to spin-averaged cross sections, while those lead to a "−" sign should contribute to spin asymmetries. Only the leading twist quark operator that defines the quark transversity distribution $\delta q(x,\mu^2)$

$$\delta q(x,\mu^2) = \overline{\psi}_q(0)\gamma^+\gamma^\perp\gamma_5\psi_q(\xi^-),$$

(or $h_1(x,\mu^2)$), is relevant to the double transverse spin asymmetries of observables with a single large momentum transfer Q in proton–proton collisions of transversely polarized protons.

The QCD factorization for the leading power contribution to the Drell–Yan production of a massive lepton pair in a collision with two transversely polarized protons should follow the same arguments that led to those in Fig. 119. Here all collinear and longitudinally polarized gluons factorized into gauge links, and soft gluon interactions are factorized into an overall soft-factor. The factorization of spinor traces of the Fermion lines needs to be modified to reflect the transverse-spin projector $\gamma^{\pm}\gamma^{\perp}\gamma_5$ (where $\pm$ indicates the two possibilities due to two colliding hadrons) instead of the $\gamma^{\pm}$ and $\gamma^{\pm}\gamma_5$ for unpolarized and longitudinally polarized active quarks. Therefore, the QCD factorization formalism for the numerator of the double transverse-spin asymmetries is the same as that in Eq. (5.108), except the unpolarized PDFs are replaced by the quark transversity distributions of various flavors (no gluon transversity distribution in a spin-1/2 transversely polarized proton), and the hard part is calculated with $\gamma^{\pm}\gamma^{\perp}\gamma_5$ spin projection for transversely polarized quarks. The collinear transversity distribution has the same definition as the quark distribution in Eq. (5.105) with the quark operator replaced by

$$\frac{1}{2}\overline{\psi}_q(0)\gamma^{+}\gamma^{\perp}\gamma_5\mathcal{W}_{[0,\xi^-]}\psi_q(\xi^-)$$

and the unpolarized hadron state $|h(p)\rangle$ is replaced by a transversely polarized hadron state $|h(p), \vec{s}_{\perp}\rangle$.

Single transverse-spin asymmetries

The transverse single-spin asymmetry (SSA),

$$A_N \equiv \frac{\sigma(s_T) - \sigma(-s_T)}{\sigma(s_T) + \sigma(-s_T)},$$

is defined as the ratio of the difference and the sum of the cross sections when the spin of one of the identified hadron s_T is flipped. Two complementary QCD-based approaches have been proposed to analyze the physics behind the measured SSAs: (1) the TMD factorization approach [1281,1282,1296–1299], and (2) the collinear factorization approach [1300–1308].

In the TMD factorization approach, the asymmetry was attributed to the spin and transverse momentum correlation between the identified hadron and the active parton, and represented by the TMD parton distribution or fragmentation function. For example, the Sivers effect [1281] describes how hadron spin influences the parton's transverse motion inside a transversely polarized hadron, while the Collins effect [1282] describes how the parton's transverse spin affects its hadronization.

The TMD factorization approach is more suitable for evaluating the SSAs of scattering processes with two observed and very different momentum scales: $Q_1 \gg Q_2 \gtrsim \Lambda_{\rm QCD}$ where Q_1 is the hard scale while Q_2 is a soft scale sensitive to the active parton's transverse motion or momentum. For example, the Drell–Yan lepton pair production when $Q \gg q_T$ is a process that can be studied in terms of the TMD factorization [1267]. In addition, the SIDIS when the transverse momentum of observed final-state hadron $p_h \ll Q$ in the photon–hadron Breit frame is an ideal observable for studying A_N, since the leading power contribution to the TMD factorization of SIDIS is known to be valid [1267,1283]. Although the A_N in SIDIS can receive contribution from various sources, including the Sivers effect (Sivers function $f_{1T}^{\perp}$) and Collins effect (Collins function $H_1^{\perp}$), as well as contribution from the pretzelosity distribution $h_{1T}^{\perp}$ [1284], it is the choice of angular modulation that allows us to separate these three sources of contributions in SIDIS,

$$A_N^{\rm Sivers} \propto \langle \sin(\phi_h - \phi_s)\rangle_{UT} \propto f_{1T}^{\perp} \otimes D \tag{5.118}$$

$$A_N^{\rm Collins} \propto \langle \sin(\phi_h + \phi_s)\rangle_{UT} \propto h_1 \otimes H_1^{\perp} \tag{5.119}$$

$$A_N^{\rm Pretzelosity} \propto \langle \sin(3\phi_h - \phi_s)\rangle_{UT} \propto h_{1T}^{\perp} \otimes H_1^{\perp} \tag{5.120}$$

where D is the normal unpolarized FF, the subscript "UT" stands for unpolarized lepton and transversely polarized hadron, ϕ_h is an angle between the leptonic plane and the hadronic plane in SIDIS and ϕ_s is the angle between the hadron transverse spin vector and the leptonic plane.

The predictive power of TMD factorization leads one to expect that the TMDs will be process-independent. However, it was found that the Sivers function measured in SIDIS and that in Drell–Yan process could differ by a sign. Such simple and generalized universality should preserve the predictive power of TMD factorization approach. Theoretically, such sign change can be better verified from the operator definition of the Sivers function. The quark Sivers function is defined as the spin-dependent part of the TMD parton distributions [1297,1309],

$$f_{q/h\uparrow}(x, k_{\perp}, s_{\perp}) = \int \frac{dy^- d^2 y_{\perp}}{(2\pi)^3} e^{ixp^+y^-} e^{-i\vec{k}_{\perp}\cdot\vec{y}_{\perp}} \times \langle p, s_{\perp}|\bar{\psi}(0) W_{[0,y]}\psi(y)|p, s_{\perp}\rangle|_{y^+=0}, \tag{5.121}$$

where $W_{[0,y]}$ is the gauge link for the leading power initial- and final-state interactions between the struck parton and the spectators or the remnant of the polarized hadron. The form of the gauge links including the phase of the interactions depends on the color flow of the scattering process and is process dependent. Luckily, the parity and time-reversal invariance of QCD removes almost all process dependence of the TMDs. By applying Eq. (5.115) to the matrix element in Eq. (5.121), we have

$$f_{q/h\uparrow}^{\rm SIDIS}(x, k_{\perp}, S_{\perp}) = f_{q/h\uparrow}^{\rm DY}(x, k_{\perp}, -S_{\perp}). \tag{5.122}$$

Therefore, the Sivers function has an opposite sign in SIDIS and DY [1307,1310]. Experimentally, it is important to verify such a relationship.

In the collinear factorization approach, all active partons' transverse momenta are integrated into the collinear distributions, and the explicit spin-transverse momentum correlation in the TMD approach is now included in the high-twist collinear parton distributions or fragmentation functions. Since the massless quark in short-distance hard collisions cannot flip the spin in QCD, the SSAs in the collinear factorization approach are generated by quantum interference between a scattering amplitude with one active parton and an amplitude with two active partons. The necessary spin-flip for SSAs is achieved by angular momentum flip between single active parton state and the state of two active partons. Such nonperturbative effect is represented by twist-3 collinear parton distributions or fragmentation functions, which has no probability interpretation, and the spin flip was made possible by QCD color Lorentz force [1301,1302]. The collinear factorization approach is more relevant to the SSAs of scattering cross sections with a single hard scale $Q \gg \Lambda_{\rm QCD}$. The validity of QCD factorization for SSA in the collinear factorization approach requires study of the collinear factorization beyond the leading power (or twist-2) contribution.

It was demonstrated that QCD factorization works for the first sub-leading power contribution to the hadronic cross section, but, not beyond [1311]. That is, QCD factorization should work for the $1/Q^2$ power correction to inclusive and unpolarized Drell–Yan cross section [1312], $1/p_T^2$ corrections to unpolarized single high-p_T particle production in hadron–hadron collisions [1313], and $1/p_T$ power correction to single high-p_T particle production in hadron–hadron collisions with one of them transversely polarized [1301–1303,1314]. It is the QCD factorization for the $1/p_T$ power correction to single high transverse momentum p_T particle production in hadron–hadron collisions with one of them transversely polarized that enables the systematic collinear factorization approach to study A_N. For example, the SSA of single high-p_T hadron production in hadronic collisions, $A(p, s_T) + B(p') \to h(P_h) + X$, can be factorized [1301,1303]

$$A_N(s_T) \propto T^{(3)}(x, x, s_T) \otimes \hat{\sigma} \otimes D_f(z) + \delta q(x, s_T) \otimes \hat{\sigma}_D \otimes D^{(3)}(z, z) + \ldots, \quad (5.123)$$

where $T^{(3)}$ and $D^{(3)}$ are twist-3 three-parton correlation functions and fragmentation functions, respectively, and δq (or h_1) is the leading power transversity distribution, with "..." representing a small contributions [1315]. Various extractions of $T^{(3)}$ and $D^{(3)}$ from experimental data have been carried out [1304,1316].

The SSA is a physically measured quantity and should not depend on how we describe it from QCD factorization or the choice of factorization scheme or scale, which leads to evolution equations of factorized nonperturbative distributions or twist-3 quark–gluon correlation functions relevant to the SSA [1317]. A complete set of the correlation functions was generated by inserting (1) the field operator $\int dy_1^- \left[i S_{T\rho}\, i\epsilon_T^{\rho\sigma} F_\sigma{}^+(y_1^-) \right]$ into the matrix element of twist-2 PDFs, and (2) the operator $\int dy_1^- \left[i S_T^\sigma F_\sigma{}^+(y_1^-) \right]$ into the matrix element of twist-2 helicity distributions [1317]. A close set of evolution equations of these twist-3 correlation functions as well as the leading order evolution kernels were derived [1317–1319].

Although the two approaches each have their own kinematic domain of validity, they are consistent with each other in the perturbative regime to which they both apply [1320,1321].

5.9 Exclusive processes in QCD

George Sterman

5.9.1 *Exclusive amplitudes for hadrons: geometry and counting rules*

The analysis of exclusive reactions played a role in the development of quantum chromodynamics, and became a subject of ongoing research within QCD. This section reviews some of the early history, landmark developments and ongoing research in this lively topic, concentrating on wide-angle scattering. The reader is referred especially to the preceding contribution on factorization in cross sections, to Sect. 10 on the structure of the nucleon and Sect. 11 on QCD at high energy for closely related subject matter.

Prehistory

For many years, exclusive reactions were the language of experimental strong interaction physics at accelerators. In such reactions, up to low GeV energies (BeV at the time), new resonances were found, whose quantum numbers were revealed in the analysis of their decays. As energies increased, the analysis of exclusive reactions gave rise to theoretical advances like Regge theory, and the Veneziano amplitude [7], resulting eventually in string theory. Around the same time, the quark model for hadron spectroscopy was developed.

With the advent of multi-GeV hadronic and leptonic accelerators, any nonforward exclusive final state became a small part of the cross section. Nevertheless, if we assume that elastic scattering results directly from pairwise scattering amplitudes for constituent quarks, simple counting combined with the optical theorem leads to successful predictions on the ratios of total cross sections [1322]. Other pioneering concepts introduce a geometrical picture of colliding hadrons, whose interactions extend over their entire overlap during the scattering [1323]. This picture is agnostic on the dynamical nature of the strong interactions that mediate momentum

transfer. The dual amplitudes of Ref. [7] are exponentially suppressed for fixed-angle scattering, and indeed, exponential fall-off in $|t|$ is characteristic of near-forward cross sections at high energy [1324]. For $|t|$ in the range of a few GeV, however, this decrease moderates to a power. This, along with the observation of power-law fall-off for form factors [599] suggested that fixed-angle amplitudes might, indeed must, reflect a point-like substructure for nucleons and mesons. This section will review some of the guiding developments in this area, which grew along with QCD, and which continue to shape contemporary theoretical and experimental programs.

Hadrons in the language of partons

Hadrons are bound states, whose fine-grained properties are nonperturbative, yet based in the interactions of the quarks and gluons that appear in the Lagrangian density of QCD. To describe how partons can mediate the scattering of hadrons, we introduce a Fock space picture of the hadronic state with on-shell momentum, in terms of $P^+ = (1/\sqrt{2})(P^0 + P^3)$, mass m_H and spin s_H, as [900]

$$|H, P^+, s_H\rangle = \sum_{F_H} c_F \,|\left\{f_i, x_i, \mathbf{k}_{i,\perp}, \lambda_i\right\} F_H\rangle, \qquad (5.124)$$

where the infinite sum is over partonic Fock states, F_H, each consisting of a set of constituents, $\{f_i \ldots\}$, labelled by flavors, f_i, by the fraction x_i of $\vec{P}_H$, transverse momenta $\mathbf{k}_{i,\perp}$ and helicity λ_i. In QCD, the Fock states are labelled as well by the manner in which the colors of constituents combine to form color singlets. From these states, in principle, we can construct any of the universal quantities of perturbative QCD that can be written as expectation values of the hadronic state, including collinear and transverse momentum parton distributions. Here, however, we will for the most part make use of only the valence state, $F_{\rm val}$, with three constituents for a nucleon, two for a meson. Of course, we assume that $c_{F_{\rm val}}$ is nonzero in Eq. (5.124). The Fock state formalism puts this approximation in context, pointing the way to systematic expansions.

Constituent counting.

Influenced by the success of the parton model applied to quarks, and assuming a constituent expansion like the one just described, Brodsky and Farrar [1325], and Matveev, Muradian and Tavkhelidze [1005] realized that under broad assumptions on the strong interactions, the behavior in momentum transfer of a wide range of exclusive processes can be summarized by a simple rule, which goes under the name of quark, or more generally constituent, counting. We can see how this works by considering the very high-energy elastic scattering of two hadrons, in the first instance assumed to consist of a fixed set of "valence" partons, specified by the quark model ([uud] for the proton, for example), moving within a limited region of space, which we can think of a sphere of radius R_H for hadron H.

Following the intuitive analysis of partons in deep-inelastic scattering, we imagine that hadrons can be thought of as Lorentz contracted and time dilated. Large momentum transfer requires all n_i valence (anti-)quarks of the initial-state hadrons i to arrive within a region of area $1/Q^2$, where Q is the momentum transfer. Now the hadrons don't know they are going to collide, so we assume their partons are more or less randomly scattered about within the areas of their Lorentz-contracted wave functions. Then the likelihood for them all to be within this small area is of order

$$\left(\frac{1}{Q^2} \times \frac{1}{\pi R_H^2}\right)^{n_i - 1}$$

for each hadron of radius R_H. But this must also be true of both incoming and outgoing states, so that their wave functions may overlap.

At the moment of collision, we don't have to make an assumption on the details of the hard scattering that redirects the partons, but we assume that otherwise the amplitude is a function only of the scattering angle. Then, at fixed t/s (that is fixed center of mass scattering angle), we find the quark counting rules of Refs. [1005] and [1325],

$$\frac{d\sigma}{dt} = \frac{f(t/s)}{s^2} \left(\frac{1}{s\,\pi R_H^2}\right)^{\sum_{i=1}^{4}(n_i - 1)} . \qquad (5.125)$$

Figure 123 illustrates the scales involved, and the system just before and after the hard scattering. This relation provides a set of predictions for power-behavior, for example $d\sigma_{pp\to pp}/dt \propto s^{-10}$, which are generally successful [1327]. The determination of normalizations would require, of course, control over the short-distance interactions of the constituents, to which we will return below. For applications of these ideas to nuclei, see Sect. 5.10.

Quark exchange, spin and transparency.

Before going further into the technical status of exclusive amplitudes, it is natural to observe several fundamental consequences of this picture. First, assuming that the integrals over fractional momenta are insensitive to the endpoints, the rules of quark counting follow immediately by dimensional counting in the (in principle) calculable partonic scattering amplitudes. The picture is quite general, and applies as well to lepton–hadron elastic scattering. The constituent rules then determine the power behavior of hadronic form factors in momentum transfer, Q: Q^{-2} for mesons and Q^{-4} for baryons. In all processes, any scattering mediated by larger numbers of constituents is power-suppressed.

In the scattering of hadrons, there are generally many ways in which quarks can flow from the initial to the final state. Almost all of these describe quark exchange, whether

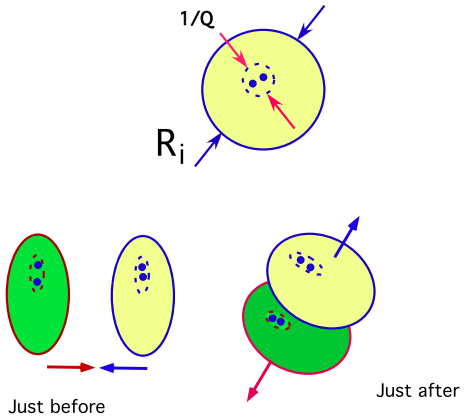


Fig. 123 The geometry of constituent counting for $\pi - \pi$ scattering ($n_i = 2$). The top represents the pion in a valence state that can contribute to an exclusive reaction, as seen along the collision axis by an oncoming hadron. From Ref. [1326]

in elastic scattering like $\pi^+ p \to \pi^+ p$, but especially for charge-exchange exclusive processes, like $\pi^- p \to \pi^0 n$. The valence Fock states described above, considered as functions of the transverse momenta of the constituents, can be used to construct a picture of $2 \to 2$ exclusive amplitudes based on the overlaps of incoming and outgoing states. These considerations lead to a variety of quite successful predictions for dependence on momentum transfer [1328]. A particularly striking example is the difference between proton–proton and antiproton–proton scattering, where the latter provides no opportunity for quark exchange. The ratio of these cross sections is about forty to one [1328].

For hadrons with light-quark valence structure (pions, nucleons) we anticipate that the scatterings will be computed with zero quark masses. Then, in any theory based on the exchange of vector gluons, the helicities of the quarks are conserved, and since the scattering is in valence states at small transverse sizes, the helicities of the valence states directly determine the spins of the external hadrons. This feature leads to many predictions for amplitudes in which spins are prepared and measured [1327]. Unlike constituent counting rules, however, predictions for spin more often fail; for the example of proton–proton scattering, see Ref. [1329].

Finally, specializing to color-singlet hadrons in a theory with colored quarks, another fundamental prediction of this picture is *transparency* [1330], which refers to predictions for exclusive hard-scattering in nuclei. On the one hand, exclusive scattering emerges only from valence parton configurations, with all partons in a small regions of coordinate space. On the other hand, at high energies, the lifetime of such a virtual state is dilated by a large factor. Thus we anticipate that *both* the incoming and outgoing hadrons in an exclusive reaction propagate as effectively point-like particles through the surrounding medium, in particular, through a nucleus. For proton–nucleon elastic scattering with momentum transfer Q, the incoming proton must be in a state of effective area $1/Q^2$ on its way into the nucleus, and will be invisible to the color fields of nucleons it encounters, whose partons are typically spread out over scales of the order of the proton's radius. Only when it encounters a constituent nucleon that happens to be in a corresponding tiny valence state can it undergo elastic scattering, producing again a pair of "stealth" nucleons that are just as invisible on the way out. While the amplitude for this to happen remains just as small as for free proton–proton or proton–neutron scattering, it is not suppressed by initial- or final-state interactions, in contrast to most cross sections on nuclei. These considerations are summarized in the elegant prediction for scattering on a nucleus of atomic number Z,

$$\frac{d\sigma}{dt}[p + Z \to p + p + (Z-1)] \overset{s\to\infty,\, t/s \text{ fixed}}{\to} Z\frac{d\sigma}{dt}[p + p \to p + p]\,. \tag{5.126}$$

This is the case, at least asymptotically, and the manner in which asymptotic behavior is reached for varied elastic reactions is a subject of ongoing experimental (see for example, Refs. [229,1331]) and theoretical investigation [1332,1333].

Splitting the hard scattering: Landshoff mechanism. Without further assumptions, the same geometric – partonic considerations sketched above can lead to an alternative picture and prediction for asymptotic behavior, first formulated by Landshoff [1334]. To be specific, let's consider meson–meson elastic scattering ($n_i = n_f = 2$). Then, instead of a single short-distance scattering involving all four incoming and outgoing partons, we imagine two independent hard scatterings of parton pairs, each resulting in two pairs of partons travelling in the same direction, and forming the outgoing mesons. The geometric picture is shown in Fig. 124. We assume that the separation b between the short-distance collisions of individual pairs of partons is generically of order R_H, the hadronic radius[62] Relative to the strict short-distance picture of Fig. 123, this reaction is enhanced by the ratio $R_H/(1/Q) = R_H Q$ in the amplitude for mesons, which is the ratio of the scale of the hard scattering to the size of the overlap between the hadrons, as shown in the figure. Similarly, there is an enhancement of $(R_H Q)^2$ for baryons, for which

$$\frac{d\sigma}{dt} = \frac{f(t/s)}{s^2}\left(\frac{1}{s\,\pi R_H^2}\right)^6. \tag{5.127}$$

In the forward region with a still-large momentum transfer, $s \gg -t \gg \Lambda_{\rm QCD}$, we anticipate a factor $1/Q^2 \sim 1/t$ for each hard scattering, and we find

[62] We will come back to this assumption below.

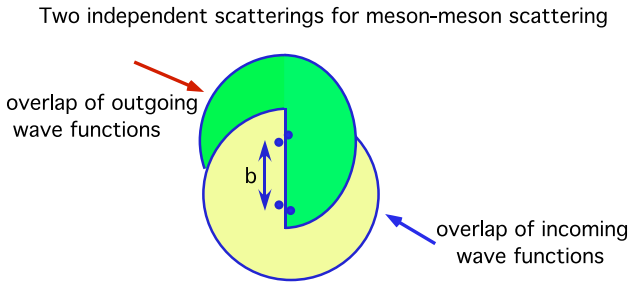


Fig. 124 Geometric enhancement in the Landshoff mechanism. The pairs of colliding partons (within each pair, one from each colliding hadron) are separated by distance b. Within each pair, partons are separated by a much smaller distance of order $1/Q$. From Ref. [1326]

$$\frac{d\sigma}{dt} = \frac{f(0)}{t^2}\left(\frac{1}{t\,\pi R_H^2}\right)^6. \tag{5.128}$$

Experimentally, at wide angles, data appear to prefer the direct counting behavior of Eq. (5.125), but at large t and even higher s, a behavior like Eq. (5.128) is observed [1335,1336].

5.9.2 *Computing hard exclusive amplitudes in QCD*

The considerations described above are based in the parton model, although they are a significant step beyond the classical parton model results, because the hard scattering is itself a strong interaction. With these concepts in hand, the next great step was to apply field theoretic analysis to elastic scattering, relying on asymptotic freedom to calculate short distance interactions where large momenta are exchanged, and on ideas of factorization to separate the dynamics binding each hadron from the short distance scattering and from each other. Before we review this landmark analysis for exclusive processes with hadrons, it is useful to touch on elastic scattering amplitudes for partons. These, of course, are not directly physical, but they play an important role in the factorized hadronic analysis that follows, and also in other areas, particularly jet cross sections.

Partons: exclusive amplitudes in QCD.
We consider partonic scattering amplitudes at "wide angles", labelling the combination of incoming and outgoing (massless) partons and their momenta as f,

$$f:\ f_1(p_1) + f_2(p_2) \to f_3(p_3) + f_4(p_4) + \cdots + f_{n+2}(p_{n+2}). \tag{5.129}$$

To define such an amplitude in perturbation theory requires the regulation of infrared singularities associated with the virtual states that include zero-momentum lines and/or lines collinear to the external particles. This is conventionally done by dimensional regularization, that is, by treating the number of dimensions as a parameter, $d = 4 - 2\varepsilon$, and continuing ε away from zero. Starting at one loop, infrared singularities manifest themselves as poles in ε, generally two per loop. Despite the growing order of the poles, the amplitude can be written in a factorized form, [1337–1339]

$$\mathcal{M}_L^{[f]}\Big(v_i, \frac{Q^2}{\mu^2}, \alpha_s(\mu^2), \varepsilon\Big) = \prod_{i\in f} J^{[i]}\Big(\frac{Q'^2}{\mu^2}, \alpha_s(\mu^2), \varepsilon\Big) \times S_{LI}^{[f]}\Big(v_i, \frac{Q^2}{\mu^2}, \alpha_s(\mu^2), \varepsilon\Big) H_I^{[f]}\Big(\beta_i, \frac{Q^2}{\mu^2}, \alpha_s(\mu^2), s\Big). \tag{5.130}$$

In this expression, the functions $J^{[i]}$ contain all poles in ε due to virtual lines collinear to the velocities, denoted v_i ($v_i^2 = 0$) of the massless external partons i. These infrared poles are universal among the amplitudes of different partonic scattering processes. That is, they only depend on whether or not the external parton is an (anti)quark or gluon. The infrared factors diverge very rapidly as $\varepsilon \to 0$, that is, in four dimensions. Many details can be found in Ref. [1340], but to get an idea of the strength of the infrared singularities, it is sufficient to see leading poles of the two-loop exponent of a jet function, given in terms of its expansion in terms of anomalous $\gamma_K^{[i]}$,

$$J^{[i]}\left(\frac{Q^2}{\mu^2}, \alpha_s(\mu^2), \varepsilon\right) \sim \exp\left\{-\left(\frac{\alpha_s}{8\pi}\right)\left(\frac{1}{\varepsilon^2}\,\gamma_K^{[i]\,(1)}\right) + \left(\frac{\alpha_s}{\pi}\right)^2\left[\frac{\beta_0}{8}\,\frac{1}{\varepsilon^2}\,\frac{3}{4\varepsilon}\gamma_K^{[i](1)} - \frac{1}{2}\left(\frac{\gamma_K^{[i](2)}}{4\varepsilon^2}\right)\right] + \cdots\right\}. \tag{5.131}$$

Here $\gamma_K^{[i]} = \sum_n \gamma_K^{[i](n)}(\alpha_s/\pi)^n$ is the coefficient of the $1/[1-x]_+$ term of the DGLAP evolution kernel for parton i, often denoted $A_i(\alpha_s)$, with $\gamma_K^{[q](1)} = C_F$, and β_0 is the lowest-order coefficient of the QCD beta function. The analysis that leads to the exponentiation of double infrared poles for partonic amplitudes relies on enhancements of radiation by accelerated massless charged particles at low angle and energy in gauge theories. The systematic treatment of these effects often goes by the name "Sudakov resummation", a term we will encounter below when we return to the Landshoff mechanism.

In Eq. (5.130), $S_{LI}^{[f]}$ is a matrix in the space of color exchanges, labelled by color tensors L and I (for example, octet or singlet exchange), which contains the remaining poles, all due to virtual lines with vanishing momenta. The soft matrix, $S_{LI}^{[f]}$ also has an expression in terms of calculable "soft" anomalous dimensions, which have wide uses in inclusive as well as exclusive cross sections. The remaining set of functions, $H_I^{[f]}$ are free of infrared poles and contain all dependence on momentum transfers.

Hadrons: factorization and evolution for form factors and exclusive amplitudes.

Historically, the analysis of hadronic exclusive amplitudes in QCD predated that for partonic amplitudes just discussed. This was possible because in these amplitudes external particles are, by construction, color singlets. We assume that the picture given above for quark counting still applies, that the elastic amplitudes result from redirecting valence quarks and antiquarks into collinear configurations in the final state, and that those configurations are color singlets. Then purely soft, as opposed to collinear, singularities disappear. Comparing to the partonic amplitude, Eq. (5.130), we derive an expression for the hadronic amplitude without a soft matrix, and with dimensionally-regularized jet functions replaced by hadronic wave functions [225,226,1270]. A short-distance, hard-scattering function denoted H describes the short-distance scattering of n_i valence quarks/antiquarks from each external hadron, i. The general form, in this case for $2 \to 2$ scattering, is

$$\mathcal{M}(s,t;\lambda_i) = \int \prod_{i=1}^{4} [dx_i]\, \phi_i(x_{i,m},\lambda_i,\mu) \times H\left(\frac{x_{i,n}x_{j,m}p_i\cdot p_j}{\mu^2};\lambda_i\right). \tag{5.132}$$

In contrast to partonic scattering, which describes the short-distance scattering of a single physical parton for each direction, hadronic wave functions, $\phi_i(x_{i,m},\lambda_i,\mu)$, depend on how their valence partons share the momentum of their external hadron, labelled by fractions $x_{i,m}$, $\sum_m x_{i,m} = 1$. Hadronic helicities, labelled by λ_i, determine spin projections for the quark constituents of the valence state. The integrals over fractional momenta are denoted (here, for baryons) by the notation,

$$[dx_i] = dx_{i,1}dx_{i,2}dx_{i,3}\,\delta\left(1-\sum_{n=1}^{3}x_{i,n}\right). \tag{5.133}$$

The factorization requires the choice of a factorization scale, μ, which is naturally of the order of the renormalization scale for the matrix element that defines the wave functions $\phi(x_i,\lambda_i,\mu)$. A representative example is for π^+, whose wave function is the matrix element of the valence quark operators that absorb an up quark and an anti-down quark, between the single-pion state and the QCD vacuum. In this case, defining $x_1 = 1 - x_2 \equiv x$ as the fraction of the up quark, the expression (in a physical gauge) is

$$\phi_\pi(x,\mu) = p\cdot n \int_{-\infty}^{\infty}\frac{d\lambda}{4\pi}\, e^{i(xp)\cdot(\lambda n)} \times\langle 0|\bar{d}(0)\,\frac{n\cdot\gamma\,\gamma_5}{2\sqrt{2n_c}}\,u(\lambda n)|\pi^+(p)\rangle, \tag{5.134}$$

where the vector n^μ is light-like and oppositely directed to the pion's momentum p^μ, and n_c is the number of colors. The matrix element requires renormalization because its fields are separated by a light-like distance, proportional to n^μ.

We note the many similarities between the exclusive amplitudes Eq. (5.132) and factorized forms of inclusive cross sections in deep-inelastic and hadron–hadron scattering. The role of wave functions here is played by parton distributions there, and in both cases there is a convolution in partonic momentum fraction(s). In both cases also, the presence of a factorization scale, μ, implies evolution equations, there for parton distributions and here for wave functions,

$$\mu\frac{\partial}{\partial\mu}\,\phi(x,\mu) = \int_0^1 dy\, V(x,y,\alpha_s(\mu))\,\phi(y,\mu). \tag{5.135}$$

The evolution kernel $V(x,y,\alpha_s)$ incorporates cancellations between constituent self-energies and diagrams with gluons exchanged between constituents. In general, the factorization scale is proportional to the momentum transfer, and these evolution equations make it possible to extrapolate wave functions (and parton distributions) from one scale to another. While space does not allow a review of the kernel and the solutions of these equations here, an especially beautiful consequence of the particular evolution equations for pion wave functions is that at asymptotically large μ the wave functions approach known, fixed, finite expressions,

$$\lim_{\mu\to\infty}\phi_\pi(x,\mu) = \frac{3f_\pi}{\sqrt{n_c}}\,x(1-x), \tag{5.136}$$

where f_π is the pion decay constant and again n_c the number of colors (3 for QCD of course). Again, this is a consequence of the detailed nature of the kernel in the evolution equation, (5.135), which follows in turn from the underlying factorization for hard exclusive processes, Eq. (5.132).

Exceptional momentum configurations.
In their original form, the factorized amplitudes of Eq. (5.132) apply to a very wide set of processes, including elastic form factors for pions and mesons, for which the external leptons can be counted as if they were hadrons with a single parton. Like any such factorized expression, however, its predictive power depends on its stability under higher-order corrections. Of particular interest are the limits where one fractional momenta x_i approaches unity and the others vanish, a configuration for elastic scattering often referred to as the Feynman mechanism (see Lecture 29 of Ref. [1341]). Noting the example of Eq. (5.136), we generally expect, and in case of pions in the valence state can prove, that wave functions vanish sufficiently rapidly in these limits to preserve the stability of the factorized amplitude in Eq. (5.132). The onset of this limit is not easy to determine, however, and has been the subject of discussion in the literature. For form factors particularly, alternative treatments based on dispersion relations and QCD sum rules, provide an alternative picture for currently accessible momentum transfers [1342]. The situation for baryonic wave functions is even more complex, because the Feynman mechanism is not suppressed at fixed orders [1343]. At high momentum transfers, this may

be resolved by higher-order corrections [1344] (see below), but phenomenological analyses based on the Feynman mechanism are also of interest [1345].

Another point of concern is the Landshoff mechanism identified above, in which subsets of the partons scatter elastically at different points in the space transverse to the beam directions, as in Fig. 124. This process is actually lower order in α_s, but more importantly it is sensitive to the transverse structure of the external hadrons, that is, on information that is not included in the wave functions discussed above. However, the resummation of higher-order QCD corrections shows that large transverse separations are suppressed, returning us to expectations very similar to those of Eq. (5.132).

Sudakov resummation and asymptotic behavior.
As we have seen in Fig. 124 and Eq. (5.127), the Landshoff enhancement to inclusive amplitudes is due to the assumed possibility of separating hard scatterings between subsets of valence partons. As noted above, to estimate the enhancement we assume that the separation is generically of the order of the hadronic radius. The analysis through Sudakov resummation follows from the observation that the separation of partonic hard scatterings in an overall hadronic exclusive amplitude requires the scattering of isolated non-singlet color charges without radiation. In isolation, these accelerated charges would result in infrared singularities, as in Eq. (5.131) above, which would make the amplitude vanish in four dimensions. In our case, however, the outgoing configurations of the scattered partons are almost collinear, and the divergences (infrared poles) cancel. The larger the separation b between the hard scatterings, however, the larger the finite remainder. The result is that any process with separated hard scatterings is suppressed relative to the acceleration of locally singlet charge configurations, which shows that the assumption of separated hard scatterings among pairs of partons made in our analysis of the Landshoff mechanism was not in fact warranted.

The observations above, which are the basis of transparency, can be quantified, by treating the distance between the hard scatterings in Fig. 124 as an impact factor, b, conjugate to transverse momentum. An analysis treating both transverse and longitudinal momenta of quarks leads to a factorized expression for hadronic scattering amplitude in terms of a wave function that depends on both the quark transverse momentum and longitudinal momentum fraction. As with the classic form, Eq. (5.132), there is a close analogy to parton distributions encountered in inclusive cross sections, in this case transverse momentum distributions (TMDs). The necessary wave functions generalize the light-cone matrix elements like Eq. (5.134) by displacing the fields in transverse (impact parameter) directions relative to the opposite-moving light cone.

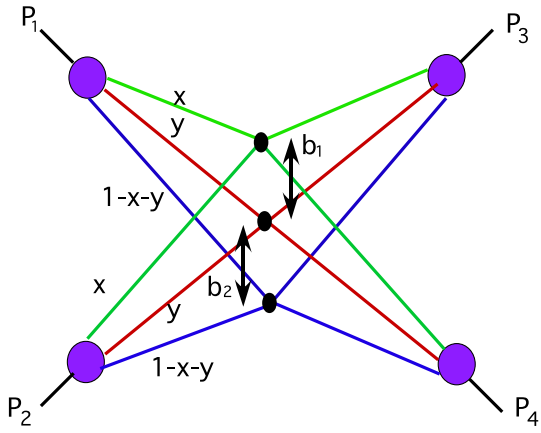


Fig. 125 Transverse separations in a multiple hard scattering. Note that the eight potentially independent integrals over momentum fractions are replaced by only two integrals, the same for each external hadron. From Ref. [1326]

This factorization in impact parameter space requires a soft matrix, which ties together soft radiation from the two (or three) separated hard scatterings in Fig. 124. Referring to the diagram in Fig. 125 for a baryonic exclusive process, we anticipate a perturbative suppression whenever the distances between hard scatterings, b_1 and b_2 in the figure, increase beyond the scale of the momentum transfer. For this process, we note that all four partons external to each hard scattering must carry the same momentum fraction. So the eight integrals over momentum fractions are reduced to two, which we label x and y here.

The form of factorization corresponding to Fig. 125 is then given at scattering angle θ and momentum transfer Q by [1346]

$$\begin{aligned}\mathcal{M}(s,t) = &\frac{1}{2\pi^2 \sin^2\theta}\sum_f \int_0^1 dx dy\, \theta(1-x-y)\\ &\times \int db_1 db_2\, \mathrm{Tr}_{\mathrm{color}}\left[U(b_i Q) H^1 H^2 H^3\right]\\ &\times \prod_{i=1,2,3,4} \mathcal{R}_i(x,y,b_1,b_2), \end{aligned} \tag{5.137}$$

where the color Trace $\left[U(b_i Q) H^1 H^2 H^3\right]$ ties color together and includes ϵ_{abc} for colors of three quarks, with possible color exchange in each hard scattering,

$$H^i(x_i p_1, x_i p_2, x_i p_3, x_i p_4) \sim 1/(x_i Q)^2.$$

In Eq. (5.137) we may define $x_1 = x$, $x_2 = y$ and $x_3 = 1 - x - y$.

The wave functions, $\mathcal{R}(x, y, b_1, b_2)$ drive the suppression of large b_i, and behave as

$$\mathcal{R}_i(x, y, b_i) \sim \phi_i(x, y, b_1, b_2, \mu \sim 1/\langle b\rangle)$$

$$\times \exp\left[-\frac{\alpha_s}{\pi}\gamma_K^{[q][1]}\sum_{a=1}^{3}\ln^2\left(\frac{1}{x_a Q b_a}\right)\right], \tag{5.138}$$

where $\gamma_K^{[q]}$ is the same anomalous dimension as for the quark jets in the partonic amplitude, Eq. (5.131). The $\phi_i(x, y, 1/\langle b\rangle)$ are normal partonic wave functions of the form encountered above, now evaluated at a renormalization scale set by the inverse of the average impact parameter spacing between the hard scatterings. The exponential suppression by double logarithms of b in Eq. (5.138) is the result of the systematic treatment of states with soft and collinear virtual radiation, and is thus an example of Sudakov resummation [1347]. It forces the impact parameters to vanish, on a scale that is for all intents and purposes of order $1/\sqrt{-t}$. Combined with the $1/t$ behaviors of the three partonic hard scatterings, the full amplitude behaves as nearly $1/t^4$, consistent with the original constituent counting rules of Eq. (5.125). The momentum transfer at which this behavior sets in, however, may be quite large, especially given the factors of x and y, which are always less than unity, in the arguments of logarithms.

5.9.3 Toward the future

The true asymptotic behavior of many exclusive reactions in QCD is by now well characterized, but much remains to be understood. In particular, it is not fully clear to what extent the success of constituent counting rules provides us with a quantitative understanding of the normalizations of amplitudes at accessible momentum transfers, and when to expect predictions based on helicity conservation and transparency to apply. Progress in these directions will be part of the future of QCD, a future in which the gap between partonic and hadronic degrees of freedom is bridged.

5.10 Hidden color

Alexandre Deur

Nuclear physics is one of the first rungs of the complexity ladder rising from our current fundamental understanding of Nature in terms of the Standard Model. The effective degrees of freedom (d.o.f.) that emerge in nuclear physics are the hadrons, namely nucleons, mesons and their excited states. Yet, effective theories are intrinsically limited, their effective d.o.f. being insufficient to account for peculiar phenomena, e.g., diffraction for geometrical optics. Then, more fundamental d.o.f. are necessary. Likewise, certain nuclear phenomena are not reducible to hadronic d.o.f. and either partonic d.o.f. or new effective d.o.f. are necessary. Hidden color (HC) is such a phenomenon. In conventional nuclear physics, a nucleus – such as the deuteron[63] – is effectively a bound state of individual nucleons. However, at the more fundamental level of QCD, the nuclear eigenstate can also have additional multi-quark Fock states which have zero color overall, but do not cluster as a collection of nucleons. These Fock states represent the HC d.o.f. of nuclei.

The possibility of HC d.o.f. [1348–1353] arises from observing that the representation of color singlet multihadron systems allows for colored cluster (C_c, colored "hadrons") components, e.g., a red-red-blue cluster bound to a green-green-blue cluster contributing to the deuteron wavefunction. Such a configuration can equivalently be reexpressed as a sum of singlet components, but without well-defined clustering properties since a given valence quark has a substantial probability to belong to any of the singlet states. Therefore, regardless of what (equivalent) representation is preferred, it cannot be expressed with singlet hadronic clusters, i.e., colorless hadronic d.o.f. This is HC. Clearly, HC goes beyond traditional nuclear physics but is a natural expectation of the underlying theory, QCD. HC predicts nuclear states not describable with usual hadronic d.o.f. but with multiquark wavefunctions, e.g., 6-quark singlet states, or singlet systems made of C_c. The latter perspective renders intuitive that HC states are short-distance binding configurations.

For example, in a hadronic basis of nucleon N, Δ and C_c d.o.f. (for simplicity we ignore other N^* isobars contributions), the deuteron is a sum of NN, $\Delta\Delta$ and C_cC_c components, the latter dominating at short distance, viz, large Q^2 [1329]:

$$|D\rangle = |NN\rangle + |\Delta\Delta\rangle + |C_cC_c\rangle$$

with

$$|NN\rangle = \tfrac{1}{3}|[6]\{33\}\rangle + \tfrac{2}{3}|[42]\{33\}\rangle - \tfrac{2}{3}|[42]\{51\}\rangle, \tag{5.139}$$

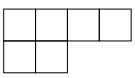
$$|\Delta\Delta\rangle = \sqrt{\tfrac{4}{45}}|[6]\{33\}\rangle + \sqrt{\tfrac{16}{45}}|[42]\{33\}\rangle + \sqrt{\tfrac{25}{45}}|[42]\{51\}\rangle, \tag{5.140}$$

$$|C_cC_c\rangle = \sqrt{\tfrac{4}{5}}|[6]\{33\}\rangle - \sqrt{\tfrac{1}{5}}|[42]\{33\}\rangle, \tag{5.141}$$

where [ ] and { } label respectively the orbital and spin-isospin symmetries which are characterized by the bracketed number in the usual Young tableau way, e.g.,

$[6] \equiv$ 

signifies 6 quarks in s-shell, or

$[42] \equiv$ 

is for 4 quarks in s-shell and 2 in p-shell [1354]. For $Q^2 \to \infty$, [6] dominates over [42]. Thus, the deuteron state is

[63] Throughout this section, deuteron is used as example of nuclear system, but the discussion is generic to multi-nucleon systems.

[6]{33} symmetric (and totally antisymmetric overall), from which 4/5 comes from the HC component, Eq. (5.141). The 80% dominance of HC at large Q^2 is therefore expected to control elastic scattering off the deuteron in this limit. In fact, the ratio of the reduced deuteron form factor (i.e., normalized to the nucleon form factor squared) to that of the pion is about 15% for Q^2 of a few GeV^2, indicating 15% of HC in $|D\rangle$ at this scale [1329]. That $|NN\rangle$ and $|\Delta\Delta\rangle$ nearly vanish at large Q^2 means that two singlet hadrons tend to not be found close to each others, i.e., the traditional (viz, between singlet hadrons) nuclear force is repulsive at short distance. The rise with Q^2 of [6] over [42] tells us that the components of $|D\rangle$ behave differently with Q^2. Their evolutions come from gluon exchange and were calculated in Refs. [1355–1357]. It was shown that the singlet pn state of the deuteron prevalent at small Q^2 evolves into 5 states: itself and 4 HC states.

The number of HC states quickly increases with the mass number A of the system. For $A = 1$ there is 1 singlet state and no HC state:

$$3 \otimes 3 \otimes 3 = 10 \oplus 8 \oplus 8 \oplus 1,$$

the last being the color singlet, the nucleon. For the deuteron, $A = 2$ and

$$\begin{aligned} &3 \otimes 3 \otimes 3 \otimes 3 \otimes 3 \otimes 3 \\ &\quad = 28 \oplus 5(35) \oplus 9(27) \oplus 15(10). \\ &\qquad \oplus 16(8) \oplus 5(10^*) \oplus 5(1), \end{aligned}$$

with the 5 last states 5(1) being the singlet states. Since there can be only one singlet state made of colorless 3-quark clusters – the traditional pn (or isobars) state – the four remaining singlet states are HC states. For $A = 3$, there are 41 HC states [1358]. Calculating strictly within QCD the Q^2-evolution of nuclear amplitudes is presently not possible: Just $|D\rangle$ at leading order involves millions of Feynman graphs. Using a hadronic effective QFT is not helpful because adding the HC d.o.f. negates the theory predicability [1358]. A solution is to use the *reduced nuclear amplitude* technique [1348,1359]. Based on LF QCD [900,1360], it models nuclear scattering amplitudes that obey QCD counting rules [153] (Sect. 5.9) and gauge invariance. The method neglects nuclear binding so that a nucleus is modeled as a cluster of collinear hadrons. Thus, the nuclear LFWF factorizes as a product of LFWF of nucleons in the nucleus times those of quarks in a nucleon: $\psi_A = \psi_{N/A} \prod_N \psi_{q/N}$, with the convenient LFWF probabilistic interpretation of the Fock states retained.

What are the possible signals for HC? An intuitive one is the yield ratio $(\gamma d \to \Delta^{++}\Delta^-)/(\gamma d \to pn)$; if $|D\rangle$ contained only a state of two weakly bound singlet hadrons,

$$\textcircled{\scriptstyle d\,u\,u}\ \textcircled{\scriptstyle d\,d\,u},$$

it would not break into a

$$\textcircled{\scriptstyle u\,u\,u}\textcircled{\scriptstyle d\,d\,d} = \Delta^{++}\Delta^-.$$

However, a 6-quark $|uuuddd\rangle$ state can well split into Δ^{++} and Δ^-.

There are other possible HC signatures [1329]: the dominance of HC at short distances makes large angle Compton scattering and pion photoproduction off the deuteron prime channels to search for HC. In electron scattering, the deuteron form factor at large Q^2 should not be explainable with hadronic d.o.f. Likewise, the deuteron inclusive tensor spin structure function b_1, a leading-twist quantity, is expected to be especially sensitive to HC [1361]. Short range correlation (SRC) measurements can also provide a signal for HC as they probe the 2-nucleon potential at short distance. Thus, SRC data should be sensitive to the repulsion expected by HC and signaled by the vanishing of the $|NN\rangle$ and $|\Delta\Delta\rangle$ components. The quasi-elastic reaction (to access large x) at high Q^2 resolves the nucleons of a nucleus and provides the SRC of nucleon pairs. The ratio of pn over pp pairs was found to be 5 times larger than the standard hadronic expectation [1362,1363]. This may stem from the repulsive core of the 2-nucleon potential. Furthermore, the measurement of the strength of 3-nucleon correlations in $A > 2$ nuclei indicates that their contribution is larger in heavy nuclei than initially expected, suggestive of the rapid increase of number of HC states with A. A challenge with SRC measurements is the fast Q^2 fall-off of form factors, so one may alternatively study, also at large Q^2 and high x, the behavior of inclusive structure functions which should obey in that regime the QCD dimensional counting rules based on the number n_s of spectator partons [153] (see Sect. 5.9),

$$xF(x) \sim (1 - x/2)^{2n_s - 1}.$$

In the maximum $x \to 2$ limit for the deuteron, $n_s = 5$ for HC (6-quark system) but $n_s = 2$ without a dominant HC state. HC evidence may come from indirect observations: without HC, the only process binding hadrons not sharing covalent quarks is glueball exchange. HC provides additional processes [1355] which may be necessary to explain the structure of neutron stars [1364,1365].

HC may have already been observed. We mentioned the SRC observations and that the deuteron form factor normalized to the nucleon form factor squared is 15% that of the pion. The $I(J^P) = 0(3^+)$ of the well-established $d^*(2380)$ (or D_{03}) p-n resonance [1366–1374] compellingly suggests that it is a 6-quark system with dominant HC [1375–1380]. Furthermore, while its dynamical decay properties can also be explained by a $\Delta\Delta$ state, the narrow 70 MeV width of the d^* is 3 times smaller than expected for the $\Delta\Delta$ but agrees with a HC state. References [1381,1382] reviewed recently the $d^*(2380)$ properties. Similarly, the narrow de-excitation of ^{4}He* through e^+e^- emission seen at ATOMKI [1383] can be understood as the ^{4}He nucleus being excited into a 12-quark HC state made of 6 colored ud pairs (hexadiquark) [1384]: it was shown that the ATMOKI anomaly cannot be

accounted for by standard electromagnetic decay without producing first a HC state [1385]. The latter also explains the unusually strong binding of the ^{4}He nucleus. Another possible observation of HC comes the b_1 data from HERMES [1386]. They are positive for $x < 0.1$ but appear to become negative around $x \simeq 0.3$, which is expected of a 6-quark HC state [1361].

These signals each hint at the existence of hidden-color degrees of freedom. By reaching higher x and Q^2, the 12 GeV upgrade of JLab and the future EIC [1293] will provide the opportunity to confirm this fundamental feature of QCD.

5.11 Color confinement, chiral symmetry breaking, and gauge topology

Edward Shuryak

5.11.1 Overview

Nontrivial topological structures of non-Abelian gauge fields were discovered in the 1970s, starting with the 't Hooft–Polyakov monopole [1387,1388] and Belavin–Polyakov–Schwartz–Tyupkin (BPST) instanton [1389]. These two sets of objects were soon related to two main nonperturbative phenomena – *confinement* and *chiral symmetry breaking*.

Confinement was connected to the so called "dual superconductor" model [1387,1390]. This model suggests that magnetically charged monopoles can form a Bose–Einstein condensate, which expels color-electric fields into flux tubes, like a condensate of Cooper pairs in superconductors expels magnetic fields into Abrikosov flux tubes.

Chiral symmetry breaking is connected to instantons, which describe vacuum tunneling between topological barriers. These have fermionic bound states – technically called *zero modes*. In the QCD vacuum the density of these states is high enough, so that they are "collectivized" into *quark condensates* $\langle \bar{q}q \rangle \neq 0$. This condensate breaks the $SU(N_f)_A \, X \, U(1)_A$ chiral symmetry of massless QCD.

For decades, theory and phenomenology of monopoles and instantons were developed separately, but in the last two decades, following a breakthrough paper by Kraan and van Baal [1391] studies of $deconfinement$ and *chiral symmetry restoration* phase transitions, based on new semiclassical objects, called *instanton-monopoles* or *instanton-dyons* lead to a united quantitative description of both phase transitions, in QCD and even in its "deformed" versions.

5.11.2 Color confinement and deconfinement

Discovery of QCD 50 years ago put into motion many important developments in the 1970s. Asymptotic freedom led to a weak coupling regime at small distances and a flourishing "perturbative QCD" describing hard processes. Going in the opposite direction (small momenta or large distance, also called "infrared" or IR), one finds growing QCD coupling. In pure gauge theories the potential energy of a static quark and antiquark pair grows linearly with increasing separation, $V(R_{q\bar{q}}) \sim \sigma R_{q\bar{q}}$. Therefore, with a finite amount of energy one cannot separate color charges: they are "confined". Furthermore, all electric fields are expelled from the vacuum and get confined as well, into so called "electric flux tubes" (also known as "QCD strings"). Their "tension" (energy per length) is $\sigma \approx 1\, GeV/\text{fm}$. In QCD with dynamical quarks, a new $q\bar{q}$ pair can be created, breaking the flux tube into two. Yet it is still true that any objects with nonzero color charge – such as quarks and gluons – do not exist as independent physical objects in the QCD vacuum. This is one of the definitions of "color confinement."

This attractive picture of course needed to be tested. K. Wilson [97] promoted the statement about a linear potential to a more abstract mathematical form: the vacuum expectation value of the Wilson line

$$W = \left\langle \frac{1}{N_c} Tr\, P \exp\left(i \int_C dx_\mu A^a_\mu T^a \right) \right\rangle, \tag{5.142}$$

over some contour C of sufficiently large size with color gauge fields. Here T^a are color algebra generators, and P exp means products of exponents along a given contour C. Wilson's criterium states that in confining theories

$$W = \exp[-\sigma * Area(C)] \tag{5.143}$$

falls exponentially with the area of a surface inclosed by the contour C. If it is a rectangular contour $T * L$ in 0–1 plane, the $area = T * L$ and σ is then identified with the string tension. The very first numerical studies of non-Abelian gauge theory on the lattice, by M. Creutz [354] indeed found that the area law holds for large enough loops, and that σ is indeed physical, that is it has correct dependence on the coupling as dictated by asymptotic freedom. (Needless to say, numerical evidence is not taken for a proof by mathematically inclined folks, and an analytic proof is still missing. A million dollar prize for such a proof still waits to be awarded.)

In Quantum Electrodynamics (QED) charge renormalization makes the coupling $larger$ at small distances (large momenta transfers or UV limit), but $small$ at large distances, which is explained by very intuitive "vacuum polarization" picture, in which virtual e^+e^- pairs screen the charges. Screening of the charges by a QED medium – e.g. plasma of the Sun – is well known and tested.

One may now ask what happens in a "QCD medium". Asymptotic freedom tells us that, contrary to QED, at small distances the coupling $decreases$. But what would happen at large distances? Calculation of the polarization tensor [1392] had shown that, like in QED, the medium $screens$ the charges. Therefore, at high enough temperature the interaction becomes weak at all distances. Therefore hot/dense

QCD matter must be in a phase called a Quark–Gluon Plasma (QGP). It is the "normal phase" of QCD in which fields in the QCD Lagrangian – quarks and gluons – correspond to quasiparticles which move relatively freely. It must be distinct from the QCD vacuum and low-T hadronic phase, as there is no place for confinement, chiral condensate and other nonperturbative phenomena there. The "confining" QCD vacuum and the QGP must therefore be separated by a phase transition: and it is indeed seen in experiment and lattice studies, which now put the critical temperature at $T_{\rm deconfinement} \approx 155\,{\rm MeV}$.

As discussed in detail in section on symmetries of QCD, at vanishing quark masses it has additional *chiral symmetries* . Without mass terms, in the Lagrangian the left and right-polarized components do not directly interact with each other and independent flavor rotations become possible. Such doubled flavor symmetry can be decomposed into a $vector$ (the sum) and the axial (the L-R difference) symmetries. One of them, called axial $SU(N_f)_A$ symmetry ($N_f = 3$ is the number of light quark flavors, u, d, s), is *spontaneously broken* in the QCD vacuum, which possesses a nonzero *quark condensate* $\langle \bar{q} q \rangle \neq 0$. The melting (disappearance) of this condensate should happen at another transition $T > T_{chiral}$. Although in various settings $T_{\rm deconfinement} \neq T_{\rm chiral}$, in QCD they seem to coincide, again based on numerical lattice evidence.

Another chiral symmetry called $U(1)_A$ is broken by the *quantum anomaly* and is not actually a symmetry at all. ("Anomaly" means that while it is a symmetry of the Lagrangian, it is not a symmetry of the quantum partition function.)

5.11.3 Electric–magnetic duality and monopoles

Already our brief discussion above should have convinced the reader that the QCD vacuum is quite complicated, with one outstanding feature being the expulsion of color-electric fields into the flux tube. Already, in the 1970s [1390,1393,1394], an analogy between this phenomenon and an expulsion of magnetic fields from superconductors lead to the so called *"dual superconductor"* model of confinement.

In superconductors of the second kind there exist the so called magnetic flux tubes or $fluxons$. Magnetic fields are confined inside the tubes because of solenoidal (super)current of Cooper pairs on their surface. QCD flux tubes transfer flux of $electric$ field instead. The word "dual" is used indicating that one has to interchange electric and magnetic fields. If so, the current in the solenoid needs to be magnetic. What can it be made of?

The apparent asymmetry of Maxwellian electrodynamics bothered theorists since late 1800s: can one allow magnetic charges, by adding a nonzero r.h.s. to the $\nabla \cdot \vec{B}$ equation? An interesting motion for a set of electric and magnetic charges was predicted by J.J. Thomson and H. Poincare. With discovery of quantum mechanics, Dirac [1395] famously observed that if they exist, then consistency of the theory requires that the product of electric and magnetic coupling be quantized. As he emphasized, the existence of one monopole in the Universe would be enough to demand quantization of all electric charges, an empirical fact to which no other explanation existed. QED magnetic monopoles have been looked for in exceedingly more sensitive experiments, but so far none have been found.

Yet certain Non-Abelian gauge theories with adjoint scalars do possess solitonic magnetic monopole solutions of the equations of motion, as discovered independently by ′t Hooft and Polyakov [1387,1388] . Their prominent feature is that their *magnetic charges* comply with earlier ideas by Dirac about special conditions, making "invisible" Dirac strings and allowing coexistence of magnetic and electric charges in quantum settings. Here we cannot give justice to the explicit solution and its properties: the interested reader can find a detailed pedagogical description in books such as [1396]. Now, monopoles made of glue and scalars are bosons, so at low enough temperature their ensemble should undergo Bose–Einstein Condensation (BEC). If that happens, a "magnetically charged" monopole condensate would expel the (color)electric field into *electric confining flux tubes*, and explain confinement!

Seiberg and Witten [1215] have given an analytic proof in theories with more than one supersymmetry (which possess the needed adjoint scalars). They were able to get the exact dependence of the effective electric coupling on the vacuum expectation value (VEV) of the scalar $g^2(\langle\phi\rangle)$. When the VEV is large, the theory is similar to electroweak theory, with gluons and gluinos being light and weakly interacting, and monopoles very heavy. When the VEV decreases, the coupling increases to $O(1)$, and magnetic monopoles and dyons (particles with both electric and magnetic charges) have masses comparable to that of gluons and gluinos. Finally, near certain singular points the electric coupling goes infinitely strong, with gluons and gluinos much heavier than monopoles. An effective description in this regime is dual QED describing magnetic interactions of light monopoles. The remarkable fact is that opposite motion of electric and magnetic couplings follows exactly the "consistency condition" of QED $g_{electric} \cdot g_{magnetic} = const$ pointed out by Dirac [1395] nearly a century ago!

All this is very beautiful, creating significant theoretical activity at the turn of the century, but we need to return to QCD. It does *not* have adjoint scalar fields, so one cannot directly build 't Hooft–Polyakov monopoles. However, by special procedures, it was possible to identify monopoles

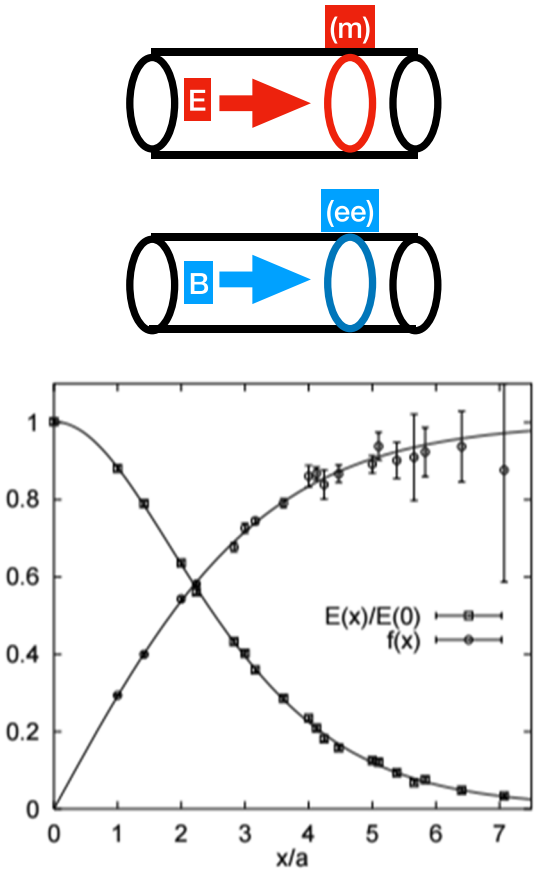


Fig. 126 Upper panel: QCD electric flux tube in QCD vacuum (upper) and magnetic flux tube in superconductor (lower). The current rotated around is made of monopoles (upper) and Cooper pairs (lower), respectively. Lower panel: plot shows the lattice data on the distribution of the electric field strength (squares) and the monopole Bose condensate (discs) in cylindrical coordinates versus the distance in the transverse plane. As one can see, the field is maximal at the center where the monopole condensate vanishes. The flux tube is generated by two static quark–antiquark external sources (not shown). The lines correspond to a solution to (dual) Ginzburg–Landau equations

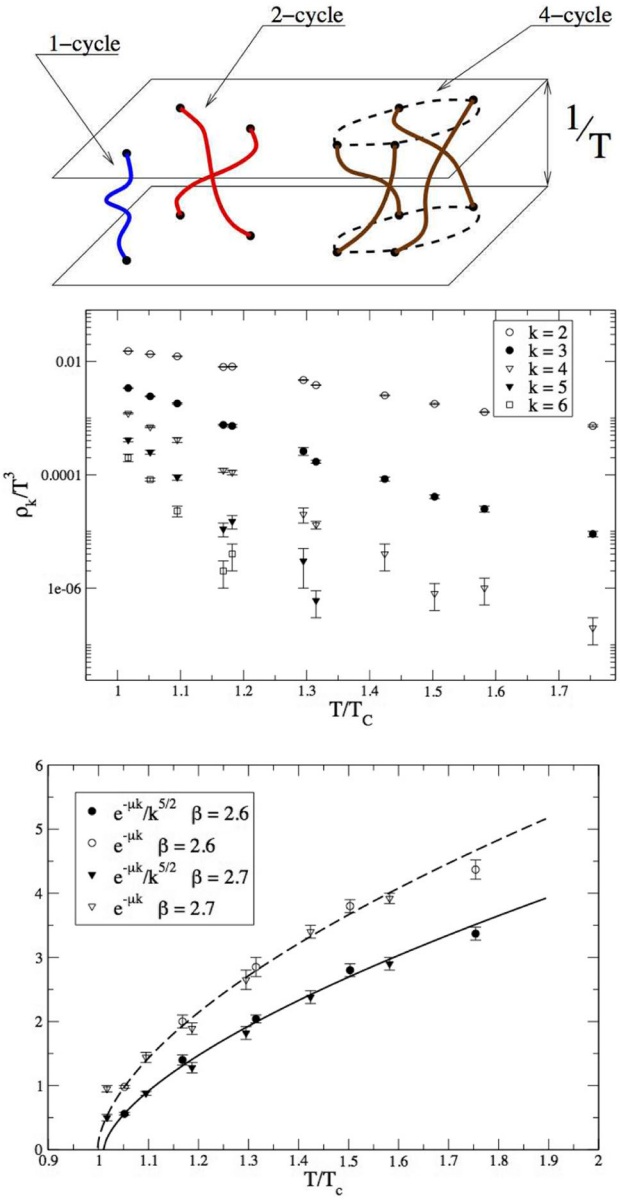


Fig. 127 Top: Example of paths of 7 identical particles which undergo a permutation made up of a 1-cycle, a 2-cycle and a 4-cycle. Middle: Normalized densities of $k-quark$ clusters ρ_k/T^3 as a function of temperature in units of critical temperature T_c. Bottom: Effective chemical potential $\mu_{\rm eff}(T)$ versus temperature, is shown to vanish exactly at the critical temperature defined by thermodynamics

on the lattice, and locate their paths and correlations. It was observed, in particular, that these monopoles do indeed rotate around the confining flux tubes, producing solenoidal magnetic currents needed to stabilize them. The picture turns out to be a *dual copy* (meaning interchange electric $\leftrightarrow$ magnetic) to well known magnetic flux tubes in superconductors. Figure 126 (displaying the result of lattice simulations summarized in the review [1397]) shows the distribution of the electric field and magnetic monopole condensate in a plane transverse to the electric flux tube. Furthermore, it has been shown [1398] that BEC phase transition of monopoles does coincide with the *deconfinement transition* at finite temperature T_c of (pure gauge) theories.

Ensembles of monopoles in QCD were studied, with important applications. Monopole correlations reveal Coulomb-like forces between monopoles [1399], with their charges "running" in the direction opposite to that of electric charges [1400], exactly as predicted by Dirac! It has been shown [1401] that monopoles also play important role in deconfined QGP phase at $T > T_c$: in particular they dominate jet quenching in quark–gluon plasmas created in heavy ion collisions, and explain unusually small viscosities observed.

The idea of Bose-clusters is explained in the top pane of Fig. 127: identical bosons may have "periodic paths" in which some number k of them exchange places. Such clusters

are widely known to the community doing many body path integral simulations for bosons, e.g. liquid He^4. Feynman argued that in order for the statistical sum to be singular at T_c, the sum over k must diverge. In other words, one may see how the probability to observe k-clusters P_k grows as $T \to T_c$ from above. In Fig. 127(middle) from [1398] one sees the corresponding data for the cluster density. Their dependence on k was fitted by two expressions, $\rho_k \sim \exp\left(-k\mu_{\rm eff}(T)\right)/k^{5/2}$ or the same without the $k^{-5/2}$ factor, to show that the critical T is not sensitive to these details of the fit. The effective chemical potential $\mu_{\rm eff}(T)$ plotted versus temperature in the bottom panel of Fig. 127. vanishes exactly at the deconfinement temperature $T = T_c$ (defined by different methods). This means that monopoles indeed undergo Bose–Einstein condensation at exactly $T = T_c$.

5.11.4 Topological landscape

Magnetic monopoles were only the first of the solitons (solutions to nonlinear classical equations of motion, stable in the sector with fixed topology). In fact there exist a whole zoo of them, even in pure gauge theory without any scalar fields.

Gauge symmetry of QCD allow transformations of fields with arbitrary $SU(3)$ matrices $\Omega(x)$, with arbitrary dependence on space-time point x. Those matrices can be divided into topologically distinct classes. Introducing the Chern–Simons number N_{CS} [1402] for the gauge potentials

$$N_{CS} \equiv \frac{\epsilon^{\alpha\beta\gamma}}{16\pi^2}\int d^3x\left(A^a_\alpha\partial_\beta A^a_\gamma + \frac{1}{3}\epsilon^{abc}A^a_\alpha A^b_\beta A^c_\gamma\right), \tag{5.144}$$

one may prove that if it is an integer, then the gauge configuration with minimal energy is "pure gauge", the field strength $G^a_{\mu\nu} = 0$ and the minimal energy is zero. Thus the values of N_{CS} numerate "classical vacua" with different topologies.

Yet when N_{CS} is in between these integers, the field strength and the minimal energy is *nonzero*. This creates a "topological landscape", an infinite sequence of classical vacua separated by *barriers*, see Fig. 128. By minimizing the energy at fixed N_{CS} (and r.m.s. size ρ) of the configurations, one can derive [1403] the shape of this barrier in parametric form. The configuration energy and Chern–Simons number are expressed in terms of a parameter κ as follows

$$U_{\rm min}(\kappa,\rho) = (1-\kappa^2)^2\frac{3\pi^2}{g^2\rho},$$
$$N_{CS}(\kappa) = \frac{1}{4}{\rm sign}(\kappa)(1-|\kappa|)^2(2+|\kappa|). \tag{5.145}$$

The value $\kappa = 0$ corresponds to the top of the barrier: this configuration is called the "sphaleron" (which in Greek means "ready to fall"). It is a solution of the classical equations of motion, a magnetic ball in which field lines of $\vec{B}^a$ ($a = 1, 2, 3$ since it is restricted to the SU(2) subgroup of

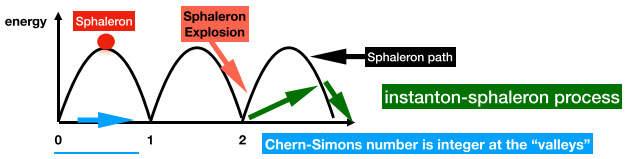


Fig. 128 The "topological landscape": minimal potential energy $U_{\rm min}$ (in units of $1/g^2\rho$) versus the Chern–Simons number N_{CS}. Valleys at integer values are separated by barriers. The terminology and arrows are described in the text

SU(3)) rotate around the x, y, z axes. Since it corresponds to an energy maximum (rather than minimum), a small perturbation would force it to fall down the barrier profile: this process (also studied analytically and numerically) is called "the sphaleron explosion". (We indicated it on the right side of Fig. 128 by red downward arrow.)

Sphalerons were originally discovered in electroweak theory [1404,1405]: in this case the sphaleron energy is very large, about 8 TeV. There were long debates whether those can be produced at LHC or future colliders: so far not a single event of this kind has been observed. Production of sphaleron-like hadronic clusters with various sizes and masses, in pp collisions at RHIC and LHC, are under consideration, see more in review [1406]. Green arrows on the r.h.s. of Fig. 128 indicate the instanton-sphaleron process in which vacuum is excited to a "turning point" magnetic configuration at the side of the barrier, from which it explodes (rolls downward).

Quantum mechanics allows potential barriers to be *penetrable* due to "tunneling". So, at any energy, even zero, *tunneling events* occur, changing N_{CS} spontaneously. Under the barrier the potential energy is larger than the total, and the kinetic energy is negative $E - U = K < 0$. Since it is proportional to momentum squared $K \sim \pi^2$, the motion should occur with imaginary momentum. That lead to the idea to describe this motion in imaginary time $\tau = it$, or Euclidean space-time. Explicit solutions describing tunneling have been found [1389], and are known as the BPST *instantons* (indicated by the horizontal blue line on the left of Fig. 128). To find them one assumes the solution is spherically symmetric in 4-d, and can be described by scalar trial radial function f, with

$$gA^a_\mu = \eta_{a\mu\nu}\partial_\nu F(y), \quad F(y) = 2\int_0^{\xi(y)} d\xi' f(\xi') \tag{5.146}$$

with $\xi = \ln(x^2/\rho^2)$ and η the 't Hooft symbol defined by

$$\eta_{a\mu\nu} = \begin{cases} \epsilon_{a\mu\nu} & \mu,\ \nu = 1,\ 2,\ 3, \\ \delta_{a\mu} & \nu = 4, \\ -\delta_{a\nu} & \mu = 4. \end{cases} \tag{5.147}$$

We also define $\overline{\eta}_{a\mu\nu}$ by changing the sign of the last two equations. Putting this expression into the gauge Lagrangian

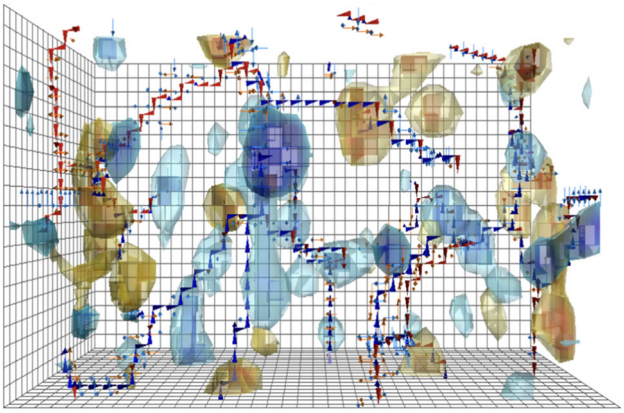


Fig. 129 QCD lattice configuration under "deep cooling": blue and yellow regions are locations of instantons and anti-instantons. One can also see a few magnetic flux tubes

one finds that it takes the form

$$L_{\rm eff} = \int d\xi \left[\frac{\dot{f}^2}{2} + 2f^2(1-f)^2 \right] \tag{5.148}$$

where the dot is the derivative with respect to ξ. This corresponds to the motion of a particle in a double-well potential. Note that, since $L = K - U$, the sign in front of the potential is $inverted$, giving two maxima rather than minima. The instanton solution is the one "sliding" from one maximum, at $\xi = 0$, to the other at $\xi = 1$.

As an individual instanton is basically a 4d ball of $G^a_{\mu\nu}$ fields, the gauge field vacuum (in Euclidean time) can be described by an ensemble of instantons and antiinstantons (those with $\bar{\eta}^a_{\mu\nu}$). The so called instanton liquid model (ILM) [1407] concluded that the instanton size and density

$$\rho = \frac{1}{3}\,{\rm fm}, \qquad n_{I+\bar{I}} = \frac{1}{R^4} = 1\,{\rm fm}^{-4} \tag{5.149}$$

led to chiral symmetry breaking, reproducing parameters of chiral perturbation theory and pion properties. Note that the 4d ball volume is $\pi^2\rho^4/2$, and the diluteness $n_{I+\bar{I}}\,\pi^2\rho^4/2 \sim 1/20 \ll 1$ of the ensemble is quite small. Yet, they are interacting with each other strongly, thus the use of the word "liquid" in the name. Many years later, numerical simulations on the lattice have shown what it looks like, see Fig. 129 from [1408] . Technically, this is a lattice gauge field "deeply cooled" (with the action minimized) which removes gluons but keeps the gauge topology intact. One can find more on lattice topology in Sect. 4.3.2.

5.11.5 *Instantons bind quarks, and by this generates chiral dynamics*

G. 't Hooft [1409] has found that instantons bind massless fermions at zero energy. Technically, these are solutions of the Dirac equation in the instanton field, called *fermionic zero modes*. The Pauli principle applies, and each instanton (a gauge field ball) binds one of each light quark, u, d, s. Therefore an "instanton liquid" contains "collectivized" light quark states. It is analogous to a ensemble of atoms: while each has its own electrons, at a finite density of atoms, these electrons can be in different phases, e.g. "insulating" or "conducting," depending on whether collective electron states do or do not have nonzero density of states on the Fermi surface. Similarly, an ensemble of instantons can have a spectrum of Dirac eigenvalues λ, either $with$ or $without$ a gap at $\lambda = 0$: in the latter case (analogous to a conductor) the chiral symmetry is spontaneously broken. With the ILM parameters mentioned above, one can prove that this is indeed the case in the QCD vacuum, and in fact it correctly reproduces the density of Dirac eigenvalues at zero (proportional to "vacuum quark condensate") $\langle\bar{q}q\rangle \approx -(240\,MeV)^3$ known from phenomenology.

This physics can be described in different, simpler terms. Massless quark fields in QCD have left and right-polarized components which, according to the QCD Lagrangian, have independent flavor symmetries. Yet, as quarks get dressed by nontrivial vacuum fields they may get mixed together so that the quarks develop nonzero "constituent quark masses" $M_{\rm eff} \sim 350$–400 MeV. The nucleon mass is about $3M_{\rm eff}$: so the phenomenon of chiral symmetry breaking explains the "mystery of our mass".

Furthermore, gauge theory in Euclidean time can naturally describe the properties at finite temperature T: just define τ to be on a circle with a circumference $\hbar/T$ (known as Matsubara time). Then the instanton solution can easily be made periodic. Although zero fermionic modes are still there at any T, the spectrum changes at $T > T_\chi$ and the Dirac eigenvalue spectrum contains a nonzero gap, and chiral symmetry gets $unbroken$ at high T. For a review on the chiral dynamics induced by an interacting ensemble of instantons see [1410].

5.11.6 *QCD correlation functions: from quarks to mesons and baryons*

Physics of QCD correlation functions using the so called QCD sum rule method and lattice numerical simulations is described in other sections. For a general pedagogical review see e.g. Ref. [1411]. At small distances between the operators the natural description is provided by perturbative diagrams, defined in terms of quarks and gluons. At large distances they are described in terms of the lowest hadrons with appropriate quantum numbers.

Of great interest however is their behavior at intermediate distances, at which a transition from one language to another takes place. As summarized in Ref. [1411], using diagrams with a $single$ instanton one can explain the scale of this transition in "problematic" channels. In particular, it is attraction

in the pion channel and repulsion in η', attraction for scalar glueball and repulsion for pseudoscalar one, etc.

Furthermore, experimentally known correlation functions were quantitatively reproduced by the interacting instanton liquid model even at large distances, first for many mesonic channels [1412,1413] and subsequently for baryonic correlators [1414]. As a result, the predictive power of the model has been explored in substantial depth. Many of the coupling constants and hadronic masses were calculated, with good agreement with experiment and lattice. (This was shown to be the case, in spite of the fact that instanton models did *not* explain confinement.)

Subsequent calculations of baryonic correlators [1414] have revealed further surprising facts. In the instanton vacuum the nucleon was shown to be made of a "constituent quark" plus a deeply bound *diquark*, with a mass nearly the same as that of constituent quarks. On the other hand, decuplet baryons (like Δ^{++}) had shown no such diquarks, remaining weakly bound set of three constituent quarks. To my knowledge, this was the first dynamical explanation of deeply bound scalar diquarks. Deeply bound scalar diquarks are a direct consequence of the 't Hooft Lagrangian, a mechanism that is also shared by the Nambu–Jona–Lasinio interaction [1415], but ignored for a long time. This subsequently lead to the realization that diquarks can become Cooper pairs in dense quark matter; see [1416] for a review on "color superconductivity".

5.11.7 *Instanton-dyons lead to semiclassical theory of the deconfinement and chiral transitions*

We have described *monopoles* and *instantons*, and have shown how they can help us understand such important nonperturbative properties as *confinement* and *chiral symmetry breaking*, respectively. Yet neither of them were able to describe both of them in a natural way.

This was achieved only during the last decade, using what are called *instanton-dyons* (kind of a hybrid of these two topological animals, also known as instanton-monopoles). Technically, if they are far from each other, they can be described as monopoles, which use the A_0 component of the gauge field instead of the adjoint scalar of the Georgi-Glashow model, involved in 't Hooft–Polyakov monopole construction. When they overlap, they can still be followed analytically. When their centers happen to be at the same spatial point, their superposition turns out to be nothing else but the well known instanton [1391,1417]!

A hybrid often inherits good properties of both parents – but maybe some bad properties as well. In order to sort these out, we need to start explaining from special role of A_0 in the finite-temperature theory. We have mentioned that finite temperature theory is defined on a circle $\tau \in C^1$ with the Matsubara period. In such cases there exist a phenomenon known

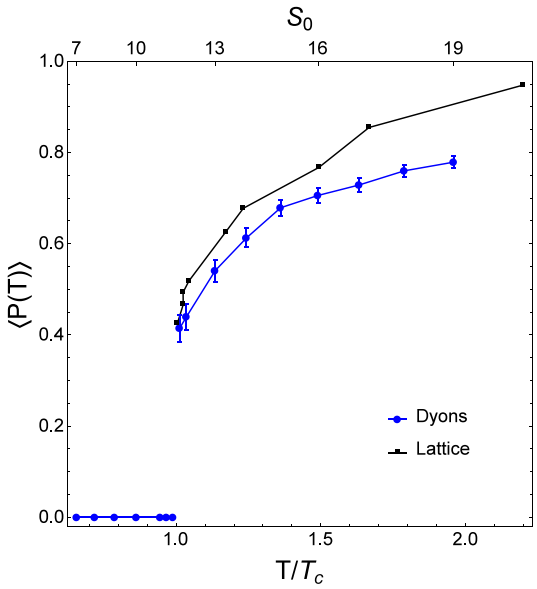


Fig. 130 Temperature dependence of the mean Polyakov line in pure $SU(3)$ gauge theory, from lattice and instanton-dyon statistical simulations, displays a clear first order phase transition in which $\langle P \rangle$ jumps from zero below T_c to a finite value in the quark–gluon plasma phase at high T

in mathematics as "holonomy": there are non-contractable contours. The so called *Polyakov line*

$$P = P \exp\left[i \int_C d\tau A_0^a T^a \right] \tag{5.150}$$

(T^a is a color generator) is a gauge invariant operator. (Because A_0 must be periodic on (Euclidean time circle) C^1, its gauge factors cancel out.) Therefore, if it has certain values, it cannot be undone and thus, at finite T, one cannot use the $A_0 = 0$ gauge. And indeed, the average of P has some well defined expectation value $\langle P(T) \rangle$, extensively studied on the lattice (see Fig. 130). Since it is a unitary $SU(3)$ matrix, it can be defined by three eigenvalues $\exp(i\mu_i)$, $i = 1, 2, 3$. The phases μ_i are called *holonomies* $\mu_i(T)$: they prescribe the magnitude of the fields $A_0^a(T)$. Physically $\langle P(T) \rangle \sim exp[-F_Q/T]$ is related to the free energy of a static quark: in the confining phase the latter is infinite and $\langle P(T) \rangle = 0$ while in quark–gluon plasma phase it is finite and $\langle P(T) \rangle \neq 0$: so it is the *order parameter* of deconfinement.

Recognizing that A_0 may have a nonzero constant value all over the system, which cannot be gauged away and is thus physical, one has to look for solutions of the YM equations at finite temperatures which at distance $\vec{r} \to \infty$ go to such values of A_0. (Rather complicated) solutions of this type [1391] for instantons were found, and it was recognized (only after its actions were plotted) that it describes a continuous deformation, from one spherical instanton into N_c indepen-

dent bumps. If $N_c = 3$, one can follow how the triplet of *instanton-dyons* is born!

Now let us summarize their properties. Like instantons, they are (anti) selfdual $\vec{E} = \pm\vec{B}$ and live in Euclidean space time. So, they are not really particles, since they do not exist in the Minkowski world. Like instantons, they have nonzero topological charges $Q \sim \int d^4x(\vec{E}\cdot\vec{B})$. Unlike instantons, however, those charges are *not quantized to integers*: $Q_i, i = 1,2,3$ can take any values, except that their sum is still $|\sum_1^3 Q_i| = 1$. These Q_i (equal to their actions S_i) are proportional to differences $\nu_i = \mu_{i+1} - \mu_i$ of the eigenvalues of the Polyakov line (the holonomies). So, $S_i = \nu_i S$ where the coefficient is the "instanton action"

$$S = \frac{8\pi^2}{g(T)^2} = \left(\frac{11}{3}N_c - \frac{2}{3}N_f\right)log\left(\frac{T}{\Lambda_{QCD}}\right)$$

Non-integer Q_i is only possible because they inherited properties of another parent, the magnetic monopoles. These objects are connected by Dirac strings (this connection undoes the topological classification theorems which require that the fields be smooth at infinities.) They are called "dyons" because a magnetic charge plus a selfduality implies also presence of an *electric* charge (although real only in the Euclidean world and thus not quite physical).

Before we can proceed, we need to clarify one more puzzle related to fermionic zero modes of instanton-dyons . An instanton has *one* fermionic zero mode, and if it gets split into three instanton-dyons , one may ask how this zero mode be shared between them. The answer, also due to van Baal and collaborators, is that the zero mode is centered near *one of the three*: which one depends on the interrelation between holonomy phases μ_i and quark periodicity phases called z_f where index f means flavor, $u, d, s \ldots$. Further details on instanton-dyons, their interaction and fermionic zero modes can be found in references mentioned.

This information should be sufficient to understand how one can "hunt" for these objects on the lattice. One method is "cooling" of vacuum fields, like that used in Fig. 129. Better still is "constrained minimization" [1418] preserving the value of $\langle P\rangle$: it revealed selfdual clusters of topological charges which integrate to non-integer values. But the best is the "fermionic filter", developed by Gattringer et al. and Ilgenfritz et al., based on the zero modes of the quark Dirac operator. In Fig. 131 we show an example from [1419,1420] in which it was used. QCD simulations with realistic masses, performed by large collaborations on supercomputers, provide a set of configurations to these authors. These calculations are especially expensive since they use the so called *domain wall fermions* providing very accurate chiral symmetry of lattice fermions. Yet Larsen et al. used even better ones, called *overlap fermions*, for which chiral symmetry is exact even for finite lattices (without the continuum limit

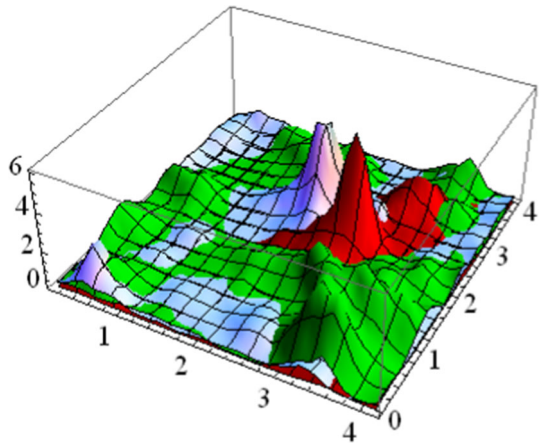


Fig. 131 Space slice of density of an exact zero mode from QCD simulation at $T = T_c$. The three colors refer to dyons of three different types

$a \to 0$ taken). Those possess *exact* zero modes $\lambda = 0$ and configurations have exactly integer topological charges.

Figure 131 shows a typical landscape of the zero mode densities $|\psi_0(x)|^2$ in two spatial dimensions. Red, blue and green colors show those for three different fermionic periodicity phases, identifying three instanton – dyon types (for $N_c = 3$) that they want to locate. The peaks correspond to locations and sizes of the individual zero modes in these field configurations. One can be convinced that the peaks are instanton-dyons because their shapes are well described by analytic formulae as derived by van Baal and collaborators, within a few percent accuracy. Furthermore, this is true not only for well separated ones, but also for overlapping ones! The gauge field configurations are for T a little bit above deconfinement T_c, in a quark–gluon plasma possessing zillions of thermal quarks and gluons: and yet, the instanton-dyons are apparently undisturbed by them! (For clarity: we do not mean here the *values of the topological charge Q or number of zero modes*, protected by certain mathematical theorems. The observed space-time shapes of the Dirac eigenmodes are not protected by any theorems known to us.)

Previous works however have not analyzed the "topological clusters", the situations in which two or three dyons overlap strongly. The Kraan–van Baal solution allows to study these cases, and good agreement was also found in the numerical analysis of instanton-dyon ensembles in [1419,1420]. The *semiclassical description* of zero and near-zero Dirac modes on the lattice is quite accurate, at least in terms of the zero mode shapes. While the very existence of zero modes was required by topological theorems, good correspondence of their shapes (in physical thermal vacuum versus pure semiclassical dyons) was a good unexpected news.

This (and many similar plots) extracted from simulations of the QCD vacuum should convince the reader that instanton-dyons are well identified objects, in terms of which

one can try to describe the underlying gauge field configurations. If so, perhaps a dream being alive for half a century since 1970s to develop consistent semiclassical theory of deconfinement and chiral transitions can still be realized.

Following this idea, ensembles of instanton-dyons were studied by a number of methods, including the mean field (solving certain "gap equations") or straightforward statistical Monte-Carlo simulations. Those were performed first for the SU(2) gauge theory [1421], then for the SU(3) [391], first without dynamical fermions, then with them [1422,1423]. In Fig. 130 we have shown one example of a comparison between a semiclassical instanton-dyon ensemble and lattice simulations. We cannot present here other results of these works but just state that they compare well with the location and properties of QCD phase transitions which we now know from lattice simulations. Note that those works were done on laptops or ordinary PCs, not supercomputers, and yet the number of topological objects in them are counted by hundreds, while very expensive lattice simulations have only few of them (as one can see from the example above).

Furthermore, in such studies people used not just QCD, but also two types of "QCD deformations". One of them adds extra operators with powers of the Polyakov line to the gauge action. By changing their strength one can affect the location and strength of the deconfinement phase transition. Another type of QCD deformation makes quarks obeying modified periodicity condition on the Matsubara circle, making quark statistics to be intermediate between fermions and bosons. This deformation affects the location and strength of the chiral phase transition. What these deformations tell us is that these two phase transitions should not generically be coincident, as they are in QCD. Again, one can apply such deformations on the lattice or in the instanton – dyon semiclassical theory, and compare the results. So far, the agreement between them is quite good, which is encouraging.

5.11.8 Conclusions and discussion

The main thrust of this section is to convince the reader that *topological solitons* play an important role in the understanding of such nonperturbative phenomena in QCD as *confinement* and *chiral symmetry breaking* in vacuum, as well as *deconfinement* and *chiral symmetry restoration* at high temperatures. A wider view on that should include the *deformed versions* of QCD, or even other gauge theories, electroweak or supersymmetric theories.

It would be nice to have just *one* type of those: but in fact the history of the field we followed in this section included (at least) three: *the particle-monopoles, instantons and instanton-dyons*. All of those were found on the lattice, by different "filters", and were shown to be strongly correlated with certain physical phenomena we would like to understand.

The $particle-monopole$ behavior convincingly shows that confinement is a Bose–Einstein condensation, explaining both the confining flux tubes and their disappearance at high T.

The $instantons$ have fermionic clouds bound to them, and their "collectivization" into a "conductor" without a gap explains how a "quark condensate" is formed, the physics of massless pions, and (unlike earlier theories) why η' is so heavy. They explain the value of the "constituent quark mass" as well as that of the nucleon (and thus ourselves). While instanton ensembles do not explain confinement, they do have most of the lowest mesons and baryons (nucleon included) as bound states.

The *instanton-dyons* (being a hybrid of the first two) connect topology with $holonomy$ (the Polyakov line, or nonzero A_0 VEV, in Euclidean formulation) in a way, which produces a nice semiclassical theory of both deconfinement and chiral transition. It was shown to work quantitatively, not only for QCD, but for its deformations as well.

Taken together, those facts and observations are impressive. The reader is reminded that they constitute the result of five decades of work by multiple theorists. But still, the reader is perhaps a bit confused by the very richness of the story told. One would probably prefer a simpler and more uniform picture.

Such feelings are shared by some active participants in this process, and some light at the end of the tunnel is, in fact, now showing. At the end of the section, let us briefly describe these later developments.

It started with Ref. [1424], using the well controlled setting of the most-supersymmetric gauge theory, with $\mathcal{N}=4$ supersymmetries. This theory has adjoint scalars and 't Hooft–Polyakov monopoles as classical solutions, and the partition function in terms of these monopoles can be calculated. The same theory in a R^3C^1 setting (preserving supersymmetries) also has holonomies and instanton-dyons, and the partition function written in terms of these was calculated as well. The two expressions are completely different, one better converges at small and another at large radius of the circle C^1. Nevertheless, as Dorey et al. observed, using Poisson summation formula, both produce *the same answer* for the statistical sum! This unexpected result was called "the Poisson duality". The importance of this paper was not noticed promptly. Indeed, such duality is very nontrivial: it is enough to remember that monopoles are particles moving in Minkowskian space-times, while instanton-dyons can only be defined in Euclidean periodic formulation. And yet, they apparently describe the same dynamics!

In fact this phenomenon has nothing to do with supersymmetry or gauge theories, and is present in a much broader domain. In Ref. [1425], the existence of the *Poisson duality* was demonstrated for a simple quantum mechanical rotator. The duality means that a partition sum can either be