LEARNING LOW-ORDER APPROXIMATIONS OF THE QUARK PROPAGATOR IN LATTICE QCD

Anonymous authors
Paper under double-blind review

ABSTRACT

The hopping expansion is an important approximation of the quark propagator in lattice quantum chromodynamics, but it diverges for light quarks, which limits the scope of strong-coupling and finite-density simulations. We present an artificial scientific discovery framework using gauge-equivariant neural networks in the Clifford basis to extract interpretable approximations. Our approach yields improved low-order approximations that outperform the hopping expansion at identical computational cost. In the parameter regime where the hopping expansion fails, our approach identifies convergent behavior, thus suggesting a novel series expansion. The discovered approximations generalize to unseen gauge fields and volumes, enabling immediate implementation as plug-in replacements and a new theoretical understanding of quark-propagator expansions.

1 INTRODUCTION

Lattice quantum chromodynamics (QCD) is currently the most precise non-perturbative method to compute theoretical predictions for many particle-physics processes (Aoki et al., 2026). These predictions also enable the search for new physics beyond the Standard Model, including the nature of dark matter (Kronfeld, 2012), the muon anomalous magnetic moment (Aliberti et al., 2025), or the origin of neutrino masses (NPLQCD Collaboration et al., 2024). However, obtaining these predictions is computationally very expensive, using roughly $20\,\%$ of the compute available at scientific data centers (Suarez et al., 2025; Camps et al., 2025).

In lattice QCD space-time is discretized to enable computer simulations (Wilson, 1974). An important element of these simulations is the calculation of propagators, which describe how quarks propagate from one space-time point to another. Evaluating these propagators amounts to solving the Dirac equation, which on the discretized space-time lattice becomes a large, sparse linear system of equations. The corresponding matrix is called the Dirac operator or matrix.

A prominent method to obtain approximate solutions of the Dirac equation is the hopping expansion (Hasenfratz & Hasenfratz, 1981), which is a power-series expansion with a finite radius of convergence. Because of its simplicity and practicality, the hopping expansion is widely used in lattice QCD, e.g., as an algorithmic tool for error reduction in Hybrid Monte Carlo simulations (Hasenbusch, 2018), to enable heavy dynamical quark simulations and calculations in the strong-coupling regime (Wakabayashi et al., 2022; Philipsen, 2021; Langelage & Philipsen, 2010; Nakamura et al., 2016), to address sign problems at nonzero baryon chemical potential (Goy et al., 2017), and to power reweighting methods for finite-density simulations (Barbour et al., 1988; Bornyakov, 2017).

While the hopping expansion is the optimal series expansion inside its radius of convergence, low-order truncations may not result in optimal approximations of the quark propagator. For the physically interesting case of light quarks the expansion diverges (Wakabayashi et al., 2022), which currently restricts strong-coupling effective theories (Philipsen, 2021) and finite-density methods (Goy et al., 2017) to heavy quarks.

In this work, we investigate improvements to low-order truncations of the hopping expansion using machine learning, specifically through the lens of neurosymbolic AI (Bhuyan et al., 2024). Rather than using neural networks as black-box predictors, we employ them as discovery engines that identify improved approximations with interpretable coefficients. Our architecture is grounded in geometric deep learning (Bronstein et al., 2021), using gauge-equivariant neural networks that re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

spect the fundamental $\mathrm{SU}(3)$ gauge symmetry of QCD by construction (Favoni et al., 2022), in analogy to how group-equivariant networks in computer vision respect rotational and translational invariance (Cohen & Welling, 2016). The main contributions of our work are:

(1) **Artificial Scientific Discovery Framework.** We introduce neural-network architectures that operate directly in the Clifford basis of $\mathbb{C}^{4\times4}$ in order to obtain model weights that are directly interpretable. From the model weights, we extract expansion coefficients as a plug-in replacement for applications that currently use the hopping expansion.

(2) **Improved low-order approximations.** Within the convergence radius of the hopping expansion, we find improved low-order approximations of the quark propagator that outperform the hopping expansion at the same computational cost.

(3) **Evidence of a novel expansion.** Outside the hopping expansion's convergence radius, our models discover convergent approximations, indicating a novel series expansion.

(4) **Transferability.** We show that our improved approximations generalize perfectly (within statistics) to unseen gauge fields drawn from the same probability distribution and to a different lattice volume, indicating that they capture general properties of the Wilson Dirac operator. We also show good transferability to gauge fields following a different probability distribution, in particular in the regime where the hopping expansion fails to converge.

2 Related Work

Our work lies at the intersection of gauge-equivariant neural networks for lattice QCD, machine-learning approaches to the Dirac equation, and artificial scientific discovery via symbolic regression.

Gauge-equivariant neural networks. Neural networks that respect group symmetries were first introduced in Cohen & Welling (2016) for generic groups and adapted to lattice gauge theories in Favoni et al. (2022), who introduced lattice gauge-equivariant convolutional neural networks. Lehner & Wettig (2023; 2024) showed that gauge-equivariant neural networks can learn preconditioners for the Wilson-clover Dirac operator, requiring minimal retraining across gauge configurations of the same ensemble. Our work builds directly on the gauge-equivariant architecture of Pfahler et al. (2026), which also added long-range interactions to improve the preconditioner near the critical mass. While these works focus on accelerating iterative solvers, we use a similar architecture to learn improved low-order expansions of the quark propagator.

Machine learning for lattice QCD. Beyond preconditioning, machine learning has been applied to several aspects of lattice QCD, including flow-based sampling to mitigate critical slowing down (Kanwar et al., 2020) and variance reduction for correlation functions (Abbott et al., 2026). A recent review is given by Boyda et al. (2022). Zhang et al. (2026) used machine learning to discover analytical formulas from data obtained via lattice QCD. However, to our knowledge, no prior work uses machine learning to discover low-order approximations of the quark propagator.

Symbolic regression and artificial scientific discovery. Our work relies heavily on symbolic regression (Dong & Zhong, 2025), in which a variety of approaches is being used, including deep-learning methods (Petersen et al., 2021), transformers (Biggio et al., 2020), genetic programming (Mundhenk et al., 2021), and equation learners (Martius & Lampert, 2016). Since much of the structure of the expansions is known a priori due to underlying symmetries, we use the equation-learner approach introduced by Martius & Lampert (2016) and adapt it to include the mathematical operations required in our application. In physics, finding formulas that fit noisy data is a central problem, and many efforts have been made to automate this task. While early work mostly focused on genetic-programming approaches, such as Eureqa (Schmidt & Lipson, 2009) or AI Feynman (Udrescu & Tegmark, 2020), many more approaches are used nowadays, see Wetzel et al. (2025) for a review. A related recent work by Kitazawa & Wada (2026) uses human-AI collaboration to provide more efficient algorithmic implementations of higher-order truncations of the hopping expansion, but does not tune coefficients.

3 PROBLEM FORMULATION

3.1 DIRAC OPERATOR AND HOPPING EXPANSION

In lattice QCD, continuous space-time is discretized on a space-time lattice, turning quantum field theoretical problems into numerically tractable quantum mechanics problems. One fundamental object on the lattice is the quark field, which lives on the lattice sites and typically has 4 spin and 3 color degrees of freedom per lattice site. Another object is the gauge field, which connects neighboring lattice sites and acts on the color degrees of freedom of the quark field. On each link on the lattice, the gauge field is represented by a matrix in $\mathrm{SU}(3)$, the group of complex three-dimensional matrices with determinant equal to 1.

We consider a four-dimensional lattice with L_x, L_y, L_z, L_t lattice sites. On this lattice, we define quark fields $\boldsymbol{a} \in \mathbb{C}^{V\times 4\times 3}$, where $V = L_x \times L_y \times L_z \times L_t$, and gauge fields $\boldsymbol{U}_\mu \in \mathrm{SU}(3)^V$, where $\mu = 1, 2, 3, 4$ is a space-time dimension. We use the shorthand notation $\boldsymbol{x}$ for the space-time indices x, y, z, t. We further define gauge transformations $\boldsymbol{\Omega} \in \mathrm{SU}(3)^V$ that act on the color degrees of freedom and transform the quark and gauge fields according to

$$\boldsymbol{a}(\boldsymbol{x}, s) \to \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{a}(\boldsymbol{x}, s)\,, \qquad \boldsymbol{U}_\mu(\boldsymbol{x}) \to \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{U}_\mu(\boldsymbol{x})\boldsymbol{\Omega}^\dagger(\boldsymbol{x} + \boldsymbol{e}^{(\mu)})\,. \tag{1}$$

Here, s is a spin index and $\boldsymbol{e}^{(\mu)}$ is the unit-vector in the μ-direction. QCD obeys a gauge symmetry, which implies that physical observables (constructed from $\boldsymbol{a}$ and $\boldsymbol{U}$) remain invariant under eq. (1). This will become important when we construct our neural networks in section 4.1 below.

Linear combinations such as $\boldsymbol{a}(\boldsymbol{x}) + \boldsymbol{a}(\boldsymbol{x} + \boldsymbol{e}^{(\mu)})$ are incompatible with this gauge symmetry. In order to allow for information transfer between neighboring lattice sites while respecting the gauge symmetry, we introduce the hop operator $\boldsymbol{H}_{\pm\mu} \in \mathbb{C}^{V\times 3\times 3}$ through its action on a quark field $\boldsymbol{a}$,

$$\boldsymbol{H}_{+\mu}\boldsymbol{a}(\boldsymbol{x}, s) = \boldsymbol{U}_\mu^\dagger(\boldsymbol{x} - \boldsymbol{e}^{(\mu)})\boldsymbol{a}(\boldsymbol{x} - \boldsymbol{e}^{(\mu)}, s)\,, \tag{2a}$$

$$\boldsymbol{H}_{-\mu}\boldsymbol{a}(\boldsymbol{x}, s) = \boldsymbol{U}_\mu(\boldsymbol{x})\boldsymbol{a}(\boldsymbol{x} + \boldsymbol{e}^{(\mu)}, s)\,. \tag{2b}$$

Under a gauge transformation, $\boldsymbol{H}_{\pm\mu}\boldsymbol{a}(\boldsymbol{x}, s) \to \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{H}_{\pm\mu}\boldsymbol{a}(\boldsymbol{x}, s)$. This is known as gauge equivariance. Linear combinations such as $\boldsymbol{a}(\boldsymbol{x}) + \boldsymbol{H}_{\pm\mu}\boldsymbol{a}(\boldsymbol{x})$ are compatible with the gauge symmetry.

By applying multiple hop operators in sequence, we can transport information over longer distances. We denote a path p by a list of pairs, where each pair $(\pm\mu, d)$ consists of the space-time direction $\pm\mu$ and the distance $d \in \mathbb{N}$ to indicate a transport in direction $\pm\mu$ by a distance of d. A given pair $h = (\pm\mu, d)$ translates to an operator $\boldsymbol{H}_h = (\boldsymbol{H}_{\pm\mu})^d$. The parallel-transport operator $\boldsymbol{T}_p$ that applies a transport along a path $p = [h_1, \ldots, h_n]$ can then be written as

$$\boldsymbol{T}_p = \boldsymbol{H}_{h_n} \ldots \boldsymbol{H}_{h_1}\,. \tag{3}$$

The Wilson Dirac equation is the linear equation $\boldsymbol{D}\boldsymbol{a} = \boldsymbol{b}$, where $\boldsymbol{D}$ is the Wilson Dirac operator introduced by Wilson (1974),

$$\boldsymbol{D} = (m + 4)\boldsymbol{I} - \frac{1}{2}\sum_{\mu=1}^{4}\left((\boldsymbol{I}_4 + \gamma_\mu)\boldsymbol{H}_{+\mu} + (\boldsymbol{I}_4 - \gamma_\mu)\boldsymbol{H}_{-\mu}\right). \tag{4}$$

Here, $m \in \mathbb{R}$ is the mass parameter, $\gamma_\mu \in \mathbb{C}^{4\times 4}$ are the Euclidean gamma matrices satisfying the anticommutation relation $\gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 2\delta_{\mu\nu}$, and $\boldsymbol{I}_4, \gamma_\mu$ act on the spin degrees of freedom.

Note that the mass parameter m in eq. (4) is not the physical quark mass, which is only recovered after renormalization. Light quark masses (for up, down, and strange quark) correspond to mass parameters $m < 0$.

In order to approximate solutions to the Wilson Dirac equation, we need approximations to $\boldsymbol{D}^{-1}$. In this work, we consider the hopping expansion. We rewrite the Wilson Dirac operator in the form

$$\boldsymbol{D} = \frac{1}{2\kappa}\left(\boldsymbol{I} - \kappa\sum_{\mu=1}^{4}\left((\boldsymbol{I}_4 + \gamma_\mu)\boldsymbol{H}_{+\mu} + (\boldsymbol{I}_4 - \gamma_\mu)\boldsymbol{H}_{-\mu}\right)\right) = \frac{1}{2\kappa}\left(\boldsymbol{I} - \kappa\boldsymbol{H}\right) \tag{5}$$

with $\kappa = \frac{1}{2(m+4)}$ and the hopping term $\boldsymbol{H}$. Since $\boldsymbol{H}_{\pm\mu}$ is unitary, its spectral norm is 1. So is the spectral norm of γ_μ. The spectral norm of $\boldsymbol{H}$ is therefore bounded by 8. Thus, there is a critical

$\kappa_h > 1/8$ (or $m_h < 0$) such that for $\kappa < \kappa_h$ (or equivalently $m > m_h$), $\boldsymbol{D}^{-1}$ can be expanded as a series in powers of κ, resulting in the hopping expansion (Hasenfratz & Hasenfratz, 1981)

$$\boldsymbol{D}^{-1} = 2\kappa \sum_{\ell=0}^{\infty} \kappa^{\ell} \boldsymbol{H}^{\ell} \,. \tag{6}$$

In applications, the series has to be truncated at a given $\ell = n$. With this truncation, it is unclear whether the hopping expansion is still the optimal approximation or if there are other approximations that require the same computational cost but provide a better approximation of $\boldsymbol{D}^{-1}$.

Furthermore, it is unclear whether a different series expansion of the quark propagator can be found in the regime $m < m_h$, where the hopping expansion does not converge anymore (Wakabayashi et al., 2022). Strong-coupling effective theories (Philipsen, 2021) and finite-density methods (Goy et al., 2017) are currently restricted to heavy quarks because of this convergence limit.

3.2 Objective

In order to investigate the limitations and possible improvements of truncated hopping expansions, we quantify how well an operator $\boldsymbol{M}$ approximates the inverse of the Dirac operator $\boldsymbol{D}$ by the error

$$\varepsilon(\boldsymbol{M}, \boldsymbol{D}) = \frac{\|\boldsymbol{M}\boldsymbol{D} - \boldsymbol{I}\|_F^2}{\|\boldsymbol{I}\|_F^2} \,, \tag{7}$$

where $\| \bullet \|_F$ denotes the Frobenius norm, and the denominator is introduced to remove the dependence on the lattice volume. Evaluating the Frobenius norm is expensive since the dimension of the matrices involved is $N = L_x L_y L_z L_t \cdot 4 \cdot 3$. We therefore rewrite the Frobenius norm as an expectation value over random normally distributed and normalized quark fields $\mathbf{v}$ and obtain

$$\varepsilon(\boldsymbol{M}, \boldsymbol{D}) = \mathbb{E}\left(\|\boldsymbol{M}\boldsymbol{D}\mathbf{v} - \mathbf{v}\|^2\right) , \tag{8}$$

see appendix A, with $\| \bullet \|$ denoting the vector norm.

In this work, we train machine-learning models to learn operators $\boldsymbol{M}$ with minimal $\varepsilon(\boldsymbol{M}, \boldsymbol{D})$ under a fixed computational budget to search for improved low-order quark propagator approximations.

4 Methods

4.1 Gauge-Equivariant Neural Networks

The Dirac operator transforms under gauge transformations as $\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y}) \to \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{\Omega}^{\dagger}(\boldsymbol{y})$, and the same needs to be true for its inverse. We have seen that linear combinations of hop operators are compatible with the gauge symmetry, so we can build approximations to $\boldsymbol{D}^{-1}$ from such linear combinations. A model built in this way does not need to learn the gauge symmetry, which is already baked into its architecture. Therefore, the amount of training needed is significantly reduced.

We introduce two types of layers for our models, Clifford layer (C) and hopping layer (H). Clifford layers take n quark fields $\boldsymbol{a}_j$ and construct m linear combinations $\boldsymbol{b}_i$ from them,

$$\boldsymbol{b}_i(\boldsymbol{x}) \overset{\text{C}}{=} \sum_{j=1}^{n} \sum_{k=1}^{16} w_{ijk} \boldsymbol{\Gamma}_k \boldsymbol{a}_j(\boldsymbol{x}) \quad \text{for } i = 1, \ldots, m \,, \tag{9}$$

where the $w_{ijk} \in \mathbb{C}$ are trainable layer weights and $\boldsymbol{\Gamma}_k$ is the Clifford basis of $\mathbb{C}^{4\times4}$. [1] This parametrization allows for the direct interpretation of the model weights and leads to faster training, as the model does not have to learn the structure of the $\boldsymbol{\Gamma}_k$, see appendix B.

The hopping layer is a special case of the parallel-transport layer in Pfahler et al. (2026). It takes as input 9 quark fields $\boldsymbol{a}_i$ and applies hop operators to them to construct 9 output quark fields $\boldsymbol{b}_i$,

$$\boldsymbol{b}_i(\boldsymbol{x}) \overset{\text{H}}{=} \begin{cases} \boldsymbol{a}_i(\boldsymbol{x}) \,, & i = 1 \,, \\ \boldsymbol{H}_{+(i-1)} \boldsymbol{a}_i(\boldsymbol{x}) \,, & i = 2, 3, 4, 5 \,, \\ \boldsymbol{H}_{-(i-5)} \boldsymbol{a}_i(\boldsymbol{x}) \,, & i = 6, 7, 8, 9 \,. \end{cases} \tag{10}$$

[1] The Clifford basis consists of $\boldsymbol{I}_4$, γ_μ, $\frac{i}{2}[\gamma_\mu, \gamma_\nu]$ for $\mu < \nu$, $i\gamma_\mu\gamma_5$, and γ_5 with $\gamma_5 = \gamma_1\gamma_2\gamma_3\gamma_4$ and i the imaginary unit.

Here, $i = 1$ corresponds to the identity, $i = 2, 3, 4, 5$ correspond to hops in the forward direction along every dimension, and $i = 6, 7, 8, 9$ correspond to backward hops. By alternating H and C layers, we can build arbitrarily deep neural networks, see fig. 1. We call a model built from these layers an HC model. Such a model applies the same set of hop operators per H layer as one step of the hopping expansion. Since the hop operations are the computationally expensive part of this calculation,[2] applying an ℓ-layer HC model[3] has the same computational cost as applying an ℓ-step truncated hopping expansion. Note that both the hopping expansion and polynomials in $\boldsymbol{D}$ can be realized by appropriate choices of the model weights.

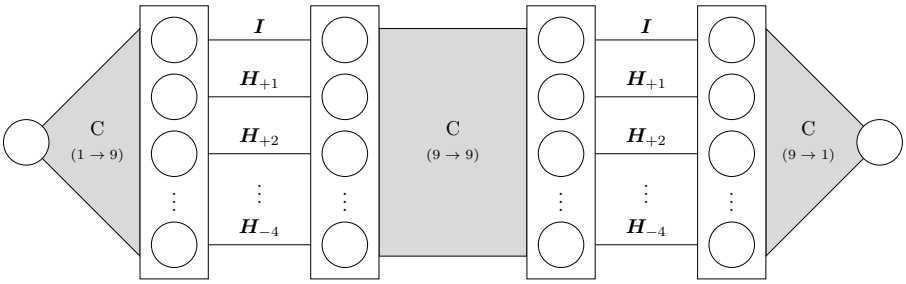


Figure 1: Sketch of the architecture of an HC model. Circles represent quark fields, gray areas represent C layers, and horizontal lines represent the different operations inside the H layer. To construct deeper models, more pairs of H and C($9 \to 9$) layers can be added in the middle.

4.2 Restricted Model

We also introduce a second model architecture that is more closely related to the hopping expansion. The fundamental layer used for this model is the restricted hopping layer (R), adapted from the hopping term $\boldsymbol{H}$ in eq. (5). It takes as input a single quark field $\boldsymbol{a}$ and outputs a single quark field $\boldsymbol{b}$,

$$\boldsymbol{b}(\boldsymbol{x}) \overset{\mathrm{R}}{=} \sum_{\mu=1}^{4} \big((w_{\boldsymbol{I}} \boldsymbol{I}_4 + w_\gamma \gamma_\mu)\boldsymbol{H}_{+\mu} + (w_{\boldsymbol{I}} \boldsymbol{I}_4 - w_\gamma \gamma_\mu)\boldsymbol{H}_{-\mu}\big)\boldsymbol{a}(\boldsymbol{x})\,, \tag{11}$$

with two trainable weights $w_{\boldsymbol{I}}$ and w_γ. A restricted model with n restricted hopping layers is defined similarly to how the hopping expansion is defined via the hopping term,

$$\boldsymbol{M} = w \sum_{\ell=0}^{n} \prod_{\ell'=1}^{\ell} \mathrm{R}_{\ell'}\,, \tag{12}$$

see eq. (6). Note that we allow for different weights in each of the R layers, and that there is one additional trainable weight that scales the entire formula, similar to the 2κ factor in eq. (6).

The HC model is a generalization of this restricted model. While the HC model is more expressive, the restricted model is easier to train, and its weights are easier to analyze. We will see the advantages of both models in section 5.

4.3 Training Setup

The Wilson Dirac operator $\boldsymbol{D}$ depends on the gauge field $\boldsymbol{U}$ and the mass parameter m. Physical observables are obtained as an average over gauge configurations for fixed m. In training, we therefore keep the mass fixed and use a number $N_{\boldsymbol{U}}$ of different gauge fields to avoid overfitting to a single gauge field. Our loss function is

$$C(\boldsymbol{M}, m) = \frac{1}{N_{\boldsymbol{U}}} \sum_{i=1}^{N_{\boldsymbol{U}}} \|\boldsymbol{M}_{\boldsymbol{U}_i} \boldsymbol{D}_{\boldsymbol{U}_i, m} \mathbf{v}_i - \mathbf{v}_i\|^2 \tag{13}$$

[2]Lattice QCD calculations are typically performed by parallelizing over the lattice. Hop operations need communication between the parallel processes, which is why they dominate runtime.

[3]When referring to the number of layers of an HC model, we mean the number of H layers in the model.

for random normally distributed and normalized $\mathbf{v}_i$, see eq. (8). The subscripts on $\boldsymbol{M}$ and $\boldsymbol{D}$ indicate the gauge field and mass parameter for which these operators are evaluated. In practice, we used $N_{\boldsymbol{U}} = 4$ and obtained agreement between training loss and test loss within statistics.

Preliminary measurements indicated that the expectation value in eq. (8) is sufficiently stable, with a worst observed coefficient of variation of $1.63\,\%$. This justifies the use of a single sample[4] $\mathbf{v}$ in eq. (13) per gauge configuration and training step. The gauge fields used for training are taken from the ensemble in Knüttel (2025a), which is an ensemble of gauge configurations on an $8^3 \times 16$ lattice, created in the quenched approximation and using the Wilson action at coupling parameter $\beta = 6$.

We trained the models using the Adam optimizer (Kingma & Ba, 2017) with standard hyperparameters, i.e., learning rate 10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Every 100 training iterations, we calculate a test score on gauge fields not used for training, and terminate when the test score changes by less than $0.1\,\%$, or after a maximum of $10\,000$ training steps. We initialize the model weights such that the initial model acts as an identity for $m < -0.5$, and such that it acts as the hopping expansion for $m > -0.5$. In both cases, we add small complex random noise with standard deviation 10^{-3} to all model parameters. This weight initialization led to stable training for all mass-parameter values.

4.4 Weight Analysis and Scientific Discovery

After training a model, we perform downstream analysis of the model weights to compare our models and the hopping expansion. To this end, we rewrite the model in terms of coefficients $c_{p,k}$ of parallel transports along different paths, i.e.,

$$\boldsymbol{M} = \sum_{\text{Paths } p} \sum_{k=1}^{16} c_{p,k} \boldsymbol{\Gamma}_k \boldsymbol{T}_p \,. \tag{14}$$

The hopping expansion can be expressed in the same form with some $c_{p,k} = 0$, which allows for a direct comparison of the model and the hopping expansion. If the model learns a $c_{p,k} \neq 0$ that is zero in the hopping expansion, it has discovered a new functional form beyond the hopping expansion.

In principle, we could have started with a model of the form eq. (14) and fitted the $c_{p,k}$ via linear regression. However, in this case the number of paths, and thus the number of $c_{p,k}$ to be fitted, grows exponentially with the number of layers, rendering this approach infeasible for many layers.

5 Experiments

5.1 Setup

We train both restricted and HC models using the training setup described in section 4.3. In order to get a good overview of the model behavior in different parameter regimes, we train models for 36 masses between -5 and 2 in steps of 0.2. To assess the convergence behavior of the low-order approximations obtained using our models, we train models with different numbers of layers.

For testing, we evaluate the approximation quality of eq. (8) on four gauge fields not used in training on each of two different lattice volumes, $8^3 \times 16$ and $16^3 \times 32$, both generated as described above. Tests using different probability distributions for the gauge fields[5] were also done, see appendix C. The $8^3 \times 16$ gauge fields are taken from Knüttel (2025a), and the $16^3 \times 32$ gauge fields from Knüttel (2025b). For each gauge field, we use 100 random quark fields to evaluate $\varepsilon(\boldsymbol{M}, \boldsymbol{D})$ and report mean and standard deviation across all measurements.

5.2 Approximation Quality

We first investigate the approximation quality of the models and compare it to the approximation quality of the hopping expansion, and to that of GMRES (Saad, 2003). To obtain an approximation of $\boldsymbol{D}^{-1}\boldsymbol{a}$ via GMRES, we solve $\boldsymbol{D}\boldsymbol{x} = \boldsymbol{a}$ with n steps using $\mathbf{0}$ as initial guess, where n matches the number of layers of the models and the order of the hopping expansion. As GMRES adaptively

[4] On the lattice volume $8^3 \times 16$ used in training, a single sample of $\mathbf{v}$ contains $196\,607$ random numbers.

[5] Specifically, we considered dynamical (or unquenched) gauge fields and Haar-distributed gauge fields.

builds an optimal polynomial approximation of $\boldsymbol{D}^{-1}$ for each right-hand side $\boldsymbol{a}$, it provides a lower bound on the approximation error achievable by any fixed polynomial in $\boldsymbol{D}$. Figure 2 shows the approximation quality for 1- and 4-layer models and compares it to 1 and 4 steps of both the hopping expansion and GMRES.

In fig. 2 we only show the larger $16^3 \times 32$ volume. In fact, the approximation quality of our models is independent of the lattice volume within statistics (see appendix D), even though all training was done on the smaller $8^3 \times 16$ volume. This is because our model architecture is similar to convolutional architectures and was designed to allow for transfer to larger volumes. An investigation of the dependence on the probability distribution of the gauge fields is provided in appendix C.

For all mass parameters, the models systematically outperform the hopping expansion as well as GMRES, with a mass-averaged reduction in approximation error of $43\,\%$ ($55\,\%$ for $m < m_h$, $23\,\%$ for $m > m_h$) for 1-layer models compared to the best baseline (hopping expansion or GMRES). For 4-layer models, the improvement increases to $65\,\%$ ($m < m_h$: $78\,\%$, $m > m_h$: $41\,\%$) for HC models and to $54\,\%$ ($m < m_h$: $64\,\%$, $m > m_h$: $37\,\%$) for restricted models. This is because the models are a generalization both of the hopping expansion and of polynomials in $\boldsymbol{D}$, and this additional freedom enables better approximations of $\boldsymbol{D}^{-1}$. In the regime $m > m_h$ where the hopping expansion is valid, the models learn the same structure as the hopping expansion (and thus HC and restricted models perform similarly here), but the additional freedom of layer-dependent weights leads to better low-order approximations. This tuning is partially dependent on the underlying probability distribution of the gauge fields, so a new model should be trained when switching to gauge fields generated with a significantly different action (for details see appendix C).

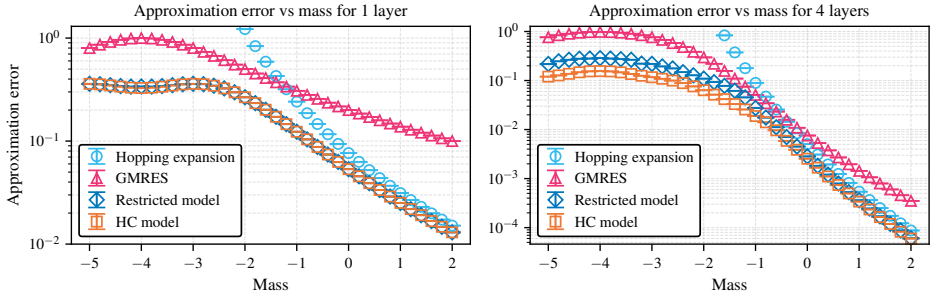


Figure 2: Approximation error of 1-layer models compared to 1 step of the hopping expansion and GMRES (left), and of 4-layer models compared to 4 steps of the hopping expansion and GMRES (right). The models systematically outperform the hopping expansion and GMRES across all mass parameters at the same computational cost (determined by the number of hop operations).

For $m < m_h$, i.e., outside of the convergence radius of the hopping expansion, we observe that our models significantly outperform both the hopping expansion and GMRES, and this effect remains when applying a model that was trained with gauge fields generated from a different action (see appendix C). This is a first indication that the models learn novel expansions of the quark propagator beyond the hopping expansion and polynomials of $\boldsymbol{D}$. To investigate this further, we consider the dependence of the approximation quality on the number of layers of the model. For $m \gtrsim -0.5$ the test score no longer improves reliably beyond 6 layers, which we attribute to the many possibilities of good low-order approximations in this regime. For $m \lesssim -0.5$ the test score improves up to the 16 layers we have investigated. In fig. 3 we observe that for large mass parameter, all approximations show geometric convergence. For small mass parameter, the models show geometric convergence up to 16 layers, while the hopping expansion clearly diverges and GMRES stagnates. Appendix E investigates this in more detail and shows that the hopping expansion fails to converge for $m < -1.2$ and GMRES stagnates for $m \lesssim -3$, while the models show geometric convergence for all masses considered.

To summarize, we observe that both model architectures outperform the hopping expansion and GMRES significantly. In the regime $m \lesssim -0.5$, the HC models perform best and show signs of finding alternative expansions that converge where the hopping expansion does not and GMRES stagnates. In the regime $m \gtrsim -0.5$, both models show similar performance and learn improved truncations of the hopping expansion. In the following, we analyze the formulas obtained in each regime. We restrict ourselves to the HC model for small m and to the restricted model for large m.

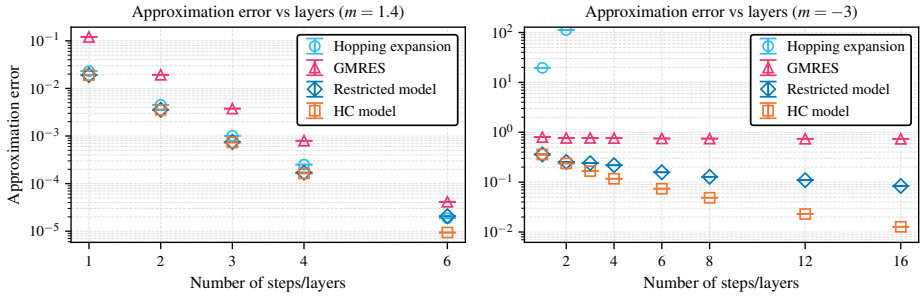


Figure 3: Dependence of approximation error on the number of layers (for models) or steps (for the hopping expansion and GMRES), for mass parameters $m = 1.4$ (left) and $m = -3$ (right).

5.3 Low-order approximations and novel expansions

By analyzing the model weights as described in section 4.4, we can extract what the models learned. In the resulting eq. (14), we first focus on selected $c_{p,k}$ to get an overview of what the models learned, before providing the learned formulas.

We consider the mass dependence of $c_{p,k}$ for a path p with a fixed path length. We first investigate the coefficient of the path $[]$ with length 0, i.e., applying no hop operator to the input field. In the hopping expansion we have $c_{[],1} = 2\kappa = \frac{1}{m+4}$ and $c_{[],k} = 0$ for $k \neq 1$. Figure 4 shows the dependence of $\mathrm{Re}(c_{[],k})$ on the mass parameter for 4-layer HC and restricted models, and for the hopping expansion. $\mathrm{Im}(c_{[],k})$ is zero within statistics. For $m \gtrsim -0.5$, the models clearly recover the hopping expansion, while for $m \lesssim -0.5$, they systematically deviate from it. Similar behavior can be seen for models with different numbers of layers, other paths, and other $\mathbf{\Gamma}_k$, see appendix F.

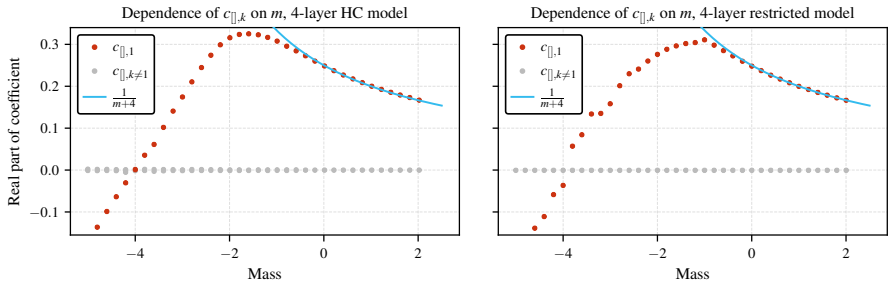


Figure 4: Dependence of $\mathrm{Re}(c_{[],k})$ on the mass parameter, for 4-layer HC models (left) and 4-layer restricted models (right). The blue line shows the expected $\mathrm{Re}(c_{[],1})$ according to the hopping expansion, while the symbols show $\mathrm{Re}(c_{[],k})$ as learned by the models.

We now investigate the low-order approximations obtained in the regime where the hopping expansion is valid. In section 5.2, we saw that in this regime it is sufficient to consider the restricted models. For mass parameter $m = 1$, the n-th order hopping expansion is

$$D^{-1} = 0.2 \sum_{\ell=0}^{n} \sum_{\mu=1}^{4} \left((0.1\boldsymbol{I}_4 + 0.1\gamma_\mu)\boldsymbol{H}_{+\mu} + (0.1\boldsymbol{I}_4 - 0.1\gamma_\mu)\boldsymbol{H}_{-\mu} \right)^{\ell} + \mathcal{O}(\boldsymbol{H}^{n+1}) \,. \tag{15}$$

From the restricted models, we obtain low-order approximations of the quark propagator which improve upon the hopping expansion, such as[6]

$$D^{-1} \approx 0.193386(29) \Big(\boldsymbol{I} + \sum_{\mu=1}^{4} (0.078651(22)\boldsymbol{I}_4 + 0.08896(8)\gamma_\mu)\boldsymbol{H}_{+\mu} + (0.078651(22)\boldsymbol{I}_4 - 0.08896(8)\gamma_\mu)\boldsymbol{H}_{-\mu} \Big) \,. \tag{16}$$

[6]Coefficients are averaged over 5 training runs using different seeds. The standard deviation in the last digits is reported in parentheses.

We only show the learned formula for the 1-layer model. Formulas for deeper models or for different masses are given in appendix G. From the standard deviations, we observe that the formulas differ from the hopping expansion in a statistically significant way, and we have already seen in section 5.2 that they lead to a better approximation quality than the hopping expansion or GMRES. These formulas can be implemented in any framework, providing an improvement in approximation quality in any application where a truncated hopping expansion is currently used.

Finally, we explore the structure of the approximations obtained by HC models in the regime where the hopping expansion diverges. By writing the HC models and the hopping expansion in the form of eq. (14), we can compare the coefficients $c_{p,k}$. Figure 5 shows two of these coefficients and how they change with the number of layers. The left plot shows the coefficient $c_{[],1}$, which differs significantly ($0.1882(23)$ for 16 layers) from the hopping-expansion value of 1 while being very stable as a function of the number of layers. In the right plot, we show the coefficient $c_{[(\pm\mu,1),(\pm\nu,1),(\mp\mu,1)],1}$ of the staple[7] as an example of a coefficient that is zero in the hopping expansion but nonzero in the expansion found by our model. The coefficient again stabilizes when increasing the number of layers, this time reaching $-0.0132(9)$ for 16 layers. The fact that the staple coefficient is nonzero shows that the HC models have learned a structurally new formula. Plots for other coefficients and other mass parameters are shown in appendix H. There, we encounter a case in which the HC models learn a coefficient to be zero which was nonzero in the hopping expansion.

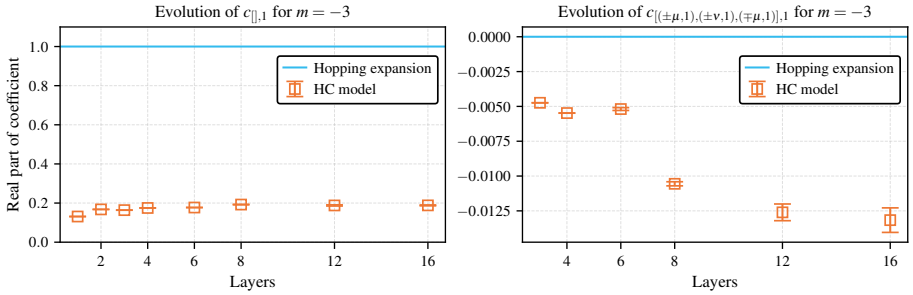


Figure 5: Evolution of the coefficients $c_{[],1}$ (left) and $c_{[(\pm\mu,1),(\pm\nu,1),(\mp\mu,1)],1}$ (right) at $m = -3$ for HC models, in the latter case averaged over all $\pm\mu$ and $\pm\nu$.

6 Discussion

Our models outperform the hopping expansion. They also generalize to unseen gauge fields and to different lattice volumes. This is made possible by realizing that the hopping expansion is overly restrictive and more freedom can be introduced without additional computational cost. In the regime where the hopping expansion converges, this freedom is used by our models to systematically change the coefficients of the paths in order to obtain better approximations of the quark propagators. Outside of this regime, at masses $m \lesssim -0.5$, the models exploit this freedom even more, showing signs of converging to a previously unknown series expansion of the quark propagator. The convergence regime of this new expansion seems to cover at least the mass-parameter regime $-5 < m \lesssim -0.5$. For larger masses, the models recover the hopping expansion, but with modified coefficients that give a better approximation quality at the same order. Appendix F and fig. 4 suggest that analytic formulas for the mass dependence of the coefficients can be found, but this needs further work.

7 Conclusion

In this work, we used machine-learning methods to improve upon existing approximations of the inverse of a large sparse linear operator, the discretized Dirac operator. We found evidence of a previously unknown series expansion in the parameter region where the known expansion does not converge. This demonstrates that machine learning can lead to both algorithmic improvements as well as new insights into well-studied problems in theoretical physics and applied mathematics.

[7] It is called staple because of the shape of the path. The coefficient is identical within statistics for all combinations of $\pm\mu$ and $\pm\nu$. In fig. 5 we have averaged over all such combinations.

AI USE STATEMENT

In this work, we used generative AI tools to help with the code implementation of the method and the result-generating pipeline. We have not used generative AI tools for any other tasks with required disclosure. Additionally, we used generative AI tools to analyze existing literature and to help in polishing the paper's writing. For checking for correctness of presentation and implementation, we used an AI agent to reproduce the results of the paper, given only a paper draft as a reference. We have reviewed and tested all code written by AI, and we have performed our own literature analysis independently of AI analysis. We take responsibility for the final content of this work, including all published code and references produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All gauge fields used for training, as well as all training and analysis scripts, are available at anonymous.4open.science/r/low-order-propagators. The repository also includes a pipeline to produce the figures used in the paper and generate the required data automatically.

AUTHOR CONTRIBUTIONS

Author contributions are omitted to ensure anonymity. They will be added after acceptance.

ACKNOWLEDGMENTS

Acknowledgements are omitted to ensure anonymity. They will be added after acceptance.

REFERENCES

Ryan Abbott, Denis Boyda, Yang Fu, Daniel C. Hackett, Gurtej Kanwar, Fernando Romero-López, Phiala E. Shanahan, and Julian M. Urban. Variance reduction in lattice QCD observables via normalizing flows. *arXiv Preprint*, March 2026. doi: 10.48550/arXiv.2603.02984.

R. Aliberti et al. The anomalous magnetic moment of the muon in the Standard Model: An update. 1143(CERN-TH-2025-101, FERMILAB-PUB-25-0344-T, INT-PUB-25-015, IPARCOS-UCM-25-029, KEK Preprint 2025-22, LTH 1403, MITP-25-037, UWThPh 2025-15, UWThPh 2025-15, ZU-TH 37/25, IPARCOS-UCM-25-029):1–158, 2025. doi: 10.1016/j.physrep.2025.08.002.

Y. Aoki, T. Blum, S. Collins, L. Del Debbio, M. Della Morte, P. Dimopoulos, X. Feng, M. Golterman, Steven Gottlieb, R. Gupta, G. Herdoíza, P. Hernandez, A. Jüttner, T. Kaneko, E. Lunghi, S. Meinel, C. Monahan, A. Nicholson, T. Onogi, P. Petreczky, A. Portelli, A. Ramos, S. R. Sharpe, J. N. Simone, S. Sint, R. Sommer, N. Tantalo, R. Van de Water, A. Vaquero, U. Wenger, and H. Wittig. FLAG Review 2024. *Physical Review D*, 113(1):014508, January 2026. ISSN 2470-0010, 2470-0029. doi: 10.1103/nfzp-p5dn.

I. M. Barbour, C. T. H. Davies, and Z. Sabeur. Lattice QCD at finite density. *Physics Letters B*, 215 (3):567–572, December 1988. ISSN 0370-2693. doi: 10.1016/0370-2693(88)91361-5.

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and T. P. Singh. Neuro-symbolic artificial intelligence: A survey. *Neural Computing and Applications*, 36(21):12809–12844, July 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-09960-z.

Luca Biggio, Tommaso Bendinelli, Aurelien Lucchi, and Giambattista Parascandolo. A Seq2Seq approach to Symbolic Regression. In *Learning Meets Combinatorial Algorithms at NeurIPS2020*, November 2020.

V. G. Bornyakov. New approach to canonical partition functions computation in Nf=2 lattice QCD at finite baryon density. *Physical Review D*, 95(9), 2017. doi: 10.1103/PhysRevD.95.094506.

Denis Boyda, Salvatore Calì, Sam Foreman, Lena Funcke, Daniel C. Hackett, Yin Lin, Gert Aarts, Andrei Alexandru, Xiao-Yong Jin, Biagio Lucini, and Phiala E. Shanahan. Applications of Machine Learning to Lattice Quantum Field Theory. *arXiv Preprint*, February 2022. doi: 10.48550/arXiv.2202.05838.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv Preprint*, May 2021. doi: 10.48550/arXiv.2104.13478.

Daan Camps, Ermal Rrapaj, Katherine Klymko, Hyeongjin Kim, Kevin Gott, Siva Darbha, Jan Balewski, Brian Austin, and Nicholas J. Wright. Quantum Computing Technology Roadmaps and Capability Assessment for Scientific Computing – An analysis of use cases from the NERSC workload. *arXiv Preprint*, September 2025. doi: 10.48550/arXiv.2509.09882.

Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2990–2999. PMLR, June 2016.

Junlan Dong and Jinghui Zhong. Recent Advances in Symbolic Regression. *ACM Computing Surveys*, 57(11):290:1–290:37, June 2025. ISSN 0360-0300. doi: 10.1145/3735634.

Matteo Favoni, Andreas Ipp, David I. Müller, and Daniel Schuh. Lattice Gauge Equivariant Convolutional Neural Networks. *Physical Review Letters*, 128(3):032003, January 2022. doi: 10.1103/PhysRevLett.128.032003.

V. A. Goy, V. Bornyakov, D. Boyda, A. Molochkov, A. Nakamura, A. Nikolaev, and V. Zakharov. Sign problem in finite density lattice QCD. *Progress of Theoretical and Experimental Physics*, 2017(3):031D01, March 2017. ISSN 2050-3911. doi: 10.1093/ptep/ptx018.

Martin Hasenbusch. Exploiting the hopping parameter expansion in the hybrid Monte Carlo simulation of lattice QCD with two degenerate flavors of Wilson fermions. *Physical Review D*, 97(11):114512, June 2018. doi: 10.1103/PhysRevD.97.114512.

Anna Hasenfratz and Peter Hasenfratz. Monte Carlo simulation combined with a hopping parameter expansion. *Physics Letters B*, 104(6):489–493, September 1981. ISSN 0370-2693. doi: 10.1016/0370-2693(81)90521-9.

Gurtej Kanwar, Michael S. Albergo, Denis Boyda, Kyle Cranmer, Daniel C. Hackett, Sébastien Racanière, Danilo Jimenez Rezende, and Phiala E. Shanahan. Equivariant Flow-Based Sampling for Lattice Gauge Theory. *Physical Review Letters*, 125(12):121601, September 2020. doi: 10.1103/PhysRevLett.125.121601.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv Preprint*, January 2017. doi: 10.48550/arXiv.1412.6980.

Masakiyo Kitazawa and Tatsuya Wada. Higher-order hopping-parameter expansion by human-AI collaboration. *arXiv Preprint*, August 2026. doi: 10.48550/arXiv.2606.31492.

Daniel Knüttel. ILDG Gauge Field Ensemble ldg/rqcd/qcdML/ens_001, March 2025a. Dataset. doi: 10.5281/zenodo.15018324.

Daniel Knüttel. ILDG Gauge Field Ensemble mc://ldg/rqcd/qcdML/ens_004_selected, October 2025b. Dataset. doi: 10.5281/zenodo.17494830.

Andreas S. Kronfeld. Twenty-First Century Lattice Gauge Theory: Results from the Quantum Chromodynamics Lagrangian. *Annual Review of Nuclear and Particle Science*, 62:265–284, 2012. ISSN 1545-4134. doi: 10.1146/annurev-nucl-102711-094942.

Jens Langelage and Owe Philipsen. The pressure of strong coupling lattice QCD with heavy quarks, the hadron resonance gas and the large N limit. *Journal of High Energy Physics*, 2010(4):55, April 2010. ISSN 1029-8479. doi: 10.1007/JHEP04(2010)055.

C. Lehner and T. Wettig. Gauge-equivariant neural networks as preconditioners in lattice QCD. *Physical Review D*, 108(3):034503, August 2023. doi: 10.1103/PhysRevD.108.034503.

C. Lehner and T. Wettig. Gauge-equivariant pooling layers for preconditioners in lattice QCD. *Physical Review D*, 110(3):034517, August 2024. doi: 10.1103/PhysRevD.110.034517.

Georg Martius and Christoph H. Lampert. Extrapolation and learning equations. *arXiv Preprint*, October 2016. doi: 10.48550/arXiv.1610.02995.

T. Nathan Mundhenk, Mikel Landajuela, Ruben Glatt, Claudio P. Santiago, Daniel M. Faissol, and Brenden K. Petersen. Symbolic Regression via Neural-Guided Genetic Programming Population Seeding. *arXiv Preprint*, November 2021. doi: 10.48550/arXiv.2111.00053.

Atsushi Nakamura, Shotaro Oka, and Yusuke Taniguchi. QCD phase transition at real chemical potential with canonical approach. *Journal of High Energy Physics*, 2016(2):54, February 2016. ISSN 1029-8479. doi: 10.1007/JHEP02(2016)054.

NPLQCD Collaboration, Zohreh Davoudi, William Detmold, Zhenghao Fu, Anthony V. Grebe, William Jay, David Murphy, Patrick Oare, Phiala E. Shanahan, and Michael L. Wagman. Long-distance nuclear matrix elements for neutrinoless double-beta decay from lattice QCD. *Physical Review D*, 109(11):114514, June 2024. doi: 10.1103/PhysRevD.109.114514.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv Preprint*, April 2021. doi: 10.48550/arXiv.1912.04871.

Simon Pfahler, Daniel Knüttel, Christoph Lehner, and Tilo Wettig. A novel gauge-equivariant neural network architecture for preconditioners in lattice QCD. In *Proceedings of The 42nd International Symposium on Lattice Field Theory — PoS(LATTICE2025)*, volume 518, pp. 027. SISSA Medialab, April 2026. doi: 10.22323/1.518.0027.

Owe Philipsen. Strong coupling methods in QCD thermodynamics. *Indian Journal of Physics*, 95 (8):1599–1611, August 2021. ISSN 0974-9845. doi: 10.1007/s12648-021-02164-4.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, April 2003. ISBN 978-0-89871-534-7.

Michael Schmidt and Hod Lipson. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923):81–85, April 2009. doi: 10.1126/science.1165893.

Estela Suarez, Jorge Amaya, Martin Frank, Oliver Freyermuth, Maria Girone, Bartosz Kostrzewa, and Susanne Pfalzner. Energy efficiency trends in HPC: What high-energy and astrophysicists need to know. *Frontiers in Physics*, 13, April 2025. ISSN 2296-424X. doi: 10.3389/fphy.2025.1542474.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, April 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aay2631.

Naoki Wakabayashi, Shinji Ejiri, Kazuyuki Kanaya, and Masakiyo Kitazawa. Scope and convergence of the hopping parameter expansion in finite-temperature quantum chromodynamics with heavy quarks around the critical point. *Progress of Theoretical and Experimental Physics*, 2022 (3):033B05, March 2022. ISSN 2050-3911. doi: 10.1093/ptep/ptac019.

Sebastian Johann Wetzel, Seungwoong Ha, Raban Iten, Miriam Klopotek, and Ziming Liu. Interpretable Machine Learning in Physics: A Review. *arXiv Preprint*, March 2025. doi: 10.48550/arXiv.2503.23616.

Kenneth G. Wilson. Confinement of quarks. *Physical Review D*, 10(8):2445–2459, October 1974. doi: 10.1103/PhysRevD.10.2445.

Wei-Jie Zhang, Zhenyu Zhang, Jifeng Hu, Bing-Nan Lu, Jin-Yi Pang, and Qian Wang. Machine Learning Unveils Finite-volume Energy Shifts in Three-body System. *arXiv Preprint*, August 2026. doi: 10.48550/arXiv.2608.28700.

A Approximation Error via Expectation Values

The Frobenius norm can be written as an expectation value over random normally distributed and normalized quark fields $\mathbf{v}$,

$$\|\boldsymbol{A}\|_F^2 = N \cdot \mathbb{E}\left(\|\boldsymbol{A}\mathbf{v}\|^2\right). \tag{17}$$

To derive equation eq. (8), we compute

$$\begin{aligned}\mathbb{E}\left(\|\boldsymbol{M}\boldsymbol{D}\mathbf{v}-\mathbf{v}\|^2\right) &= \mathbb{E}\left((\boldsymbol{M}\boldsymbol{D}\mathbf{v}-\mathbf{v})^\dagger(\boldsymbol{M}\boldsymbol{D}\mathbf{v}-\mathbf{v})\right)\\ &= \mathbb{E}\left(\mathbf{v}^\dagger(\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})^\dagger(\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})\mathbf{v}\right)\\ &= \mathbb{E}\left(\mathrm{tr}\left((\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})^\dagger(\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})\mathbf{v}\mathbf{v}^\dagger\right)\right)\\ &= \mathrm{tr}\left((\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})^\dagger(\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})\mathbb{E}(\mathbf{v}\mathbf{v}^\dagger)\right)\\ &= \mathrm{tr}\left((\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})^\dagger(\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I})\right)\frac{1}{N}\\ &= \frac{\|\boldsymbol{M}\boldsymbol{D}-\boldsymbol{I}\|_F^2}{\|\boldsymbol{I}\|_F^2}\,. \end{aligned} \tag{18}$$

In the penultimate step, we have used that $\mathbf{v}$ is normalized, i.e., the elements of $\mathbf{v}$ are drawn from a complex normal distribution of mean 0 and variance $1/N$.

B Effect of Clifford-Weight Parametrization

We investigate the effect of the weight parametrization in the Clifford layer of eq. (9) on the training speed. For comparison, we also train models with the linear layer (L) defined in Pfahler et al. (2026),

$$\boldsymbol{b}_i(\boldsymbol{x}) \overset{\mathrm{L}}{=} \sum_{j=1}^{n} \boldsymbol{W}_{ij}\boldsymbol{a}_j(\boldsymbol{x})\,, \tag{19}$$

where $\boldsymbol{W} \in \mathbb{C}^{4\times 4}$ are the trainable layer weights. Both the L and C layer have the same expressivity, as they just parametrize the complex 4×4 matrix differently.

To test their training behavior, we train HC models with 1 and 4 layers, and compare them to HL-models, which are HC models with the C layers replaced by L layers. Training is done as described in section 4.3. In fig. 6, we can see that both model types reach the same final score, but the HC models converge faster.

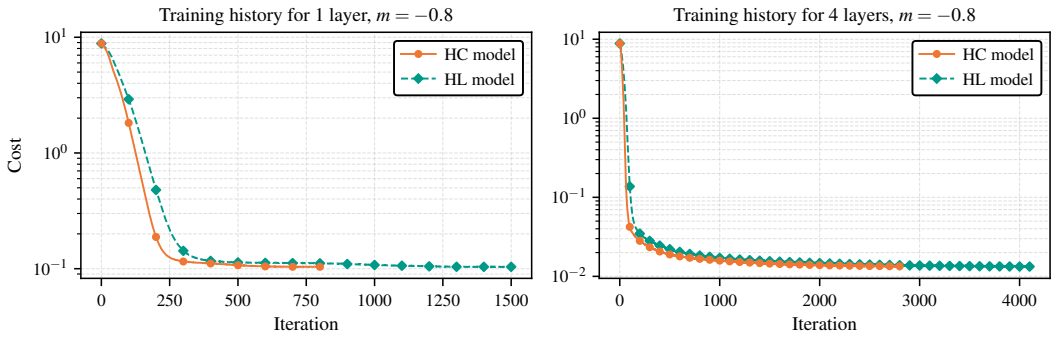


Figure 6: Training (lines) and test (symbols) scores during the training process of HC and HL models.

C Action Dependence

We can use different actions for the underlying probability distributions of our gauge fields. In this appendix, we compare results obtained for three different distributions: Quenched gauge fields generated with the Wilson gauge action at $\beta = 6$, dynamical gauge fields generated with the Wilson gauge action at $\beta = 5.3$ and two flavors of dynamical fermions with the Wilson-clover Dirac operator at $c_{\text{SW}} = 1$ and $m = -0.5$, and random Haar-distributed gauge fields corresponding to $\beta = 0$.

Figure 7 shows that the underlying probability distribution does matter and models perform best when trained and applied on gauge fields with the same probability distribution. However, only a slight performance degradation is observed applying a model trained on one type of gauge configurations to another type. Especially in the small-m regime, all models still outperform both the hopping expansion and GMRES even when applied on gauge fields from a different distribution.

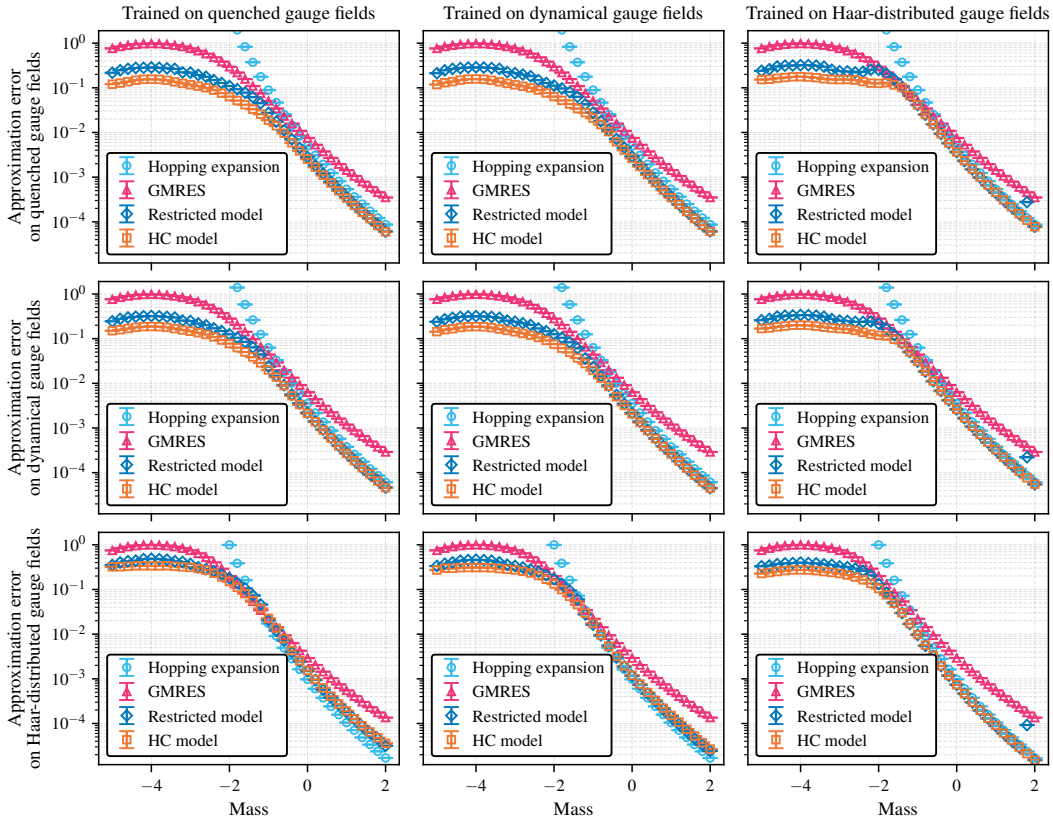


Figure 7: Approximation error of 4-layer HC and restricted models trained on quenched gauge fields (left column), dynamical gauge fields (middle column), and random Haar-distributed gauge fields (right column), applied on quenched (upper row), dynamical (middle row), and random Haar-distributed (lower row) gauge fields. The hopping expansion and GMRES only differ for different rows.

D VOLUME DEPENDENCE

When calculating the approximation error of our models on different lattice volumes, we observe that our models generalize trivially to a different volume than they are trained on, see fig. 8. This generalization is expected since the model is structurally similar to a convolutional neural network.

The only noticeable difference in the data is that on the larger volume, the errors are typically smaller. As mentioned before, this is because the volume provides additional statistics, so the approximation error has a smaller error bar with the same number of samples to calculate eq. (8). The effect is barely visible in the plots, but apparent upon inspection of the numbers.

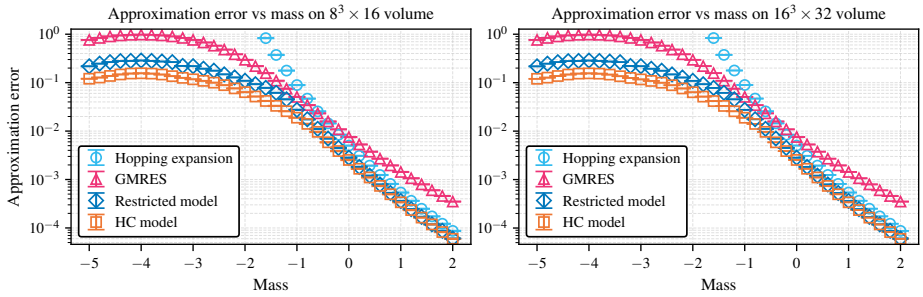


Figure 8: Approximation error of 4-layer HC and restricted models, compared to the hopping expansion and GMRES, for an $8^3 \times 16$ (left) and a $16^3 \times 32$ (right) lattice.

E RATE OF CONVERGENCE

The hopping expansion is a geometric series, i.e., the approximation error scales with the order n of the expansion (denoted by HE_n) as

$$\varepsilon(\mathrm{HE}_n, \boldsymbol{D}) = \frac{\|\mathrm{HE}_n \boldsymbol{D} - \boldsymbol{I}\|_F^2}{\|\boldsymbol{I}\|_F^2} = \frac{\kappa^{2n+2} \|\boldsymbol{H}^{n+1}\|_F^2}{12V} \,. \tag{20}$$

To study the convergence behavior, we fit the n-dependence of ε to a curve $a \cdot b^n$, where the parameter b is the rate of convergence. Figure 9 shows this rate of convergence for the hopping expansion, the HC and the restricted models, using data from models with 1, 2, 3, 4, and 6 layers for each mass-parameter value. For $m < -0.5$, models with 12 and 16 layers were additionally used for the fit. We observe that the hopping expansion diverges (i.e., $b \geq 1$) for $m \lesssim -1.2$ and GMRES stagnates (i.e., $b \approx 1$) for $m \lesssim -3$, while the trained models converge (i.e., $b < 1$) up to the smallest masses tested.

We also observe that while our models provide improved low-order approximations for all masses (see section 5.2), the convergence of GMRES is faster for $m \gtrsim -1$, since it can adapt to the quark field $\boldsymbol{a}$ for which we approximate $\boldsymbol{D}^{-1}\boldsymbol{a}$. For $m \lesssim -1$, our models show faster convergence than any of the baselines, adding to the evidence that they approach a previously unknown convergent expansion.

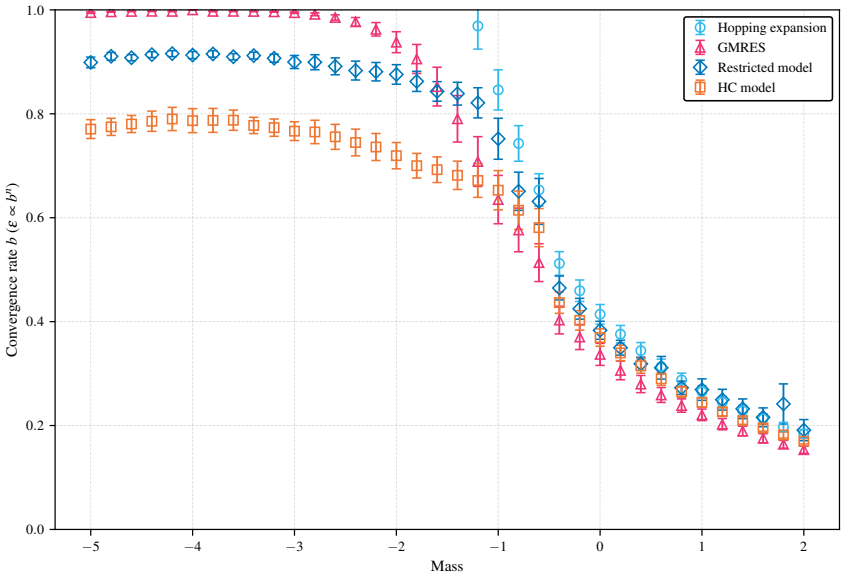


Figure 9: Rate of convergence of the model approximations, the hopping expansion, and GMRES as a function of the mass parameter.

F MASS DEPENDENCE OF COEFFICIENTS

We show additional results for the mass dependence of the coefficients obtained from model weights, for HC models only. Restricted models show similar behavior, but the results are somewhat less smooth in the small-m regime where the restricted models do not perform well. Figures 10 to 12 show the mass dependence of the coefficients $c_{p,k}$ for p with length zero, one, and two, respectively. In all plots, it can be seen that the coefficients change smoothly as a function of the mass, indicating that analytic formulas for the mass dependence of the coefficients can be found.

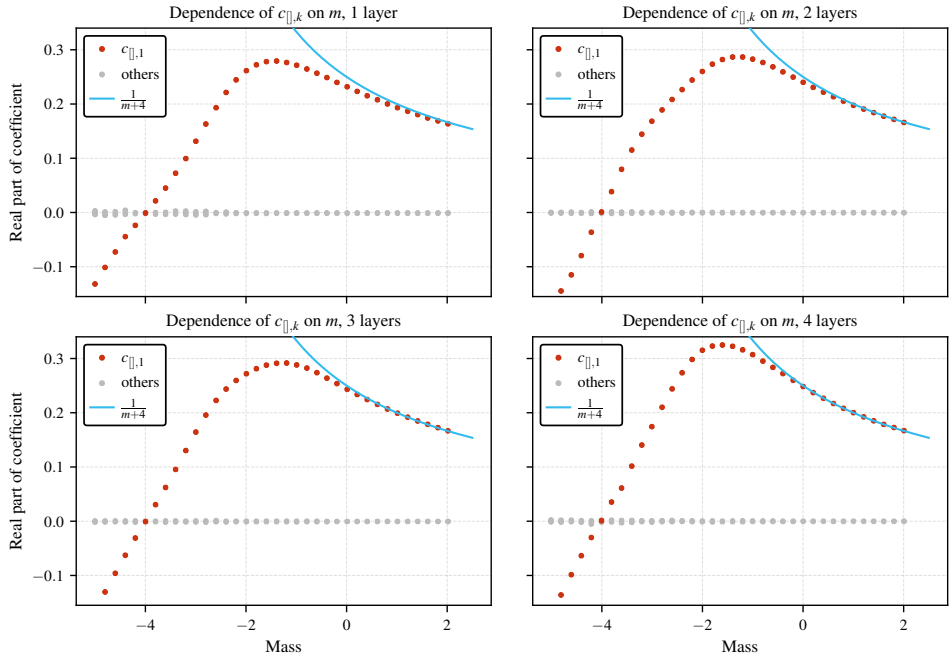


Figure 10: Dependence of the real part of coefficients of length-zero paths on the mass for HC models with different numbers of layers.

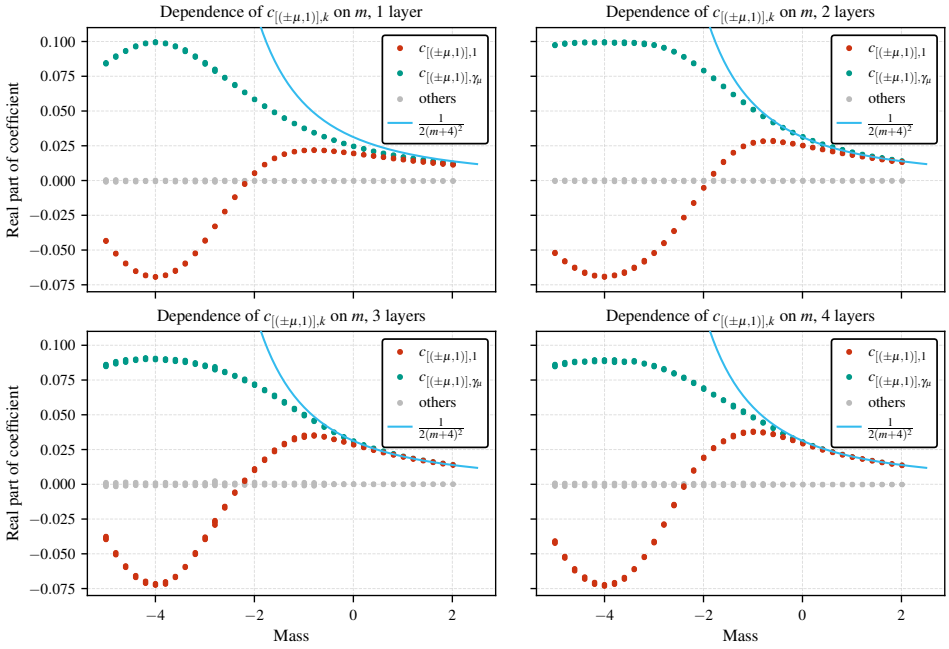


Figure 11: Same as fig. 10, but for paths of length one.

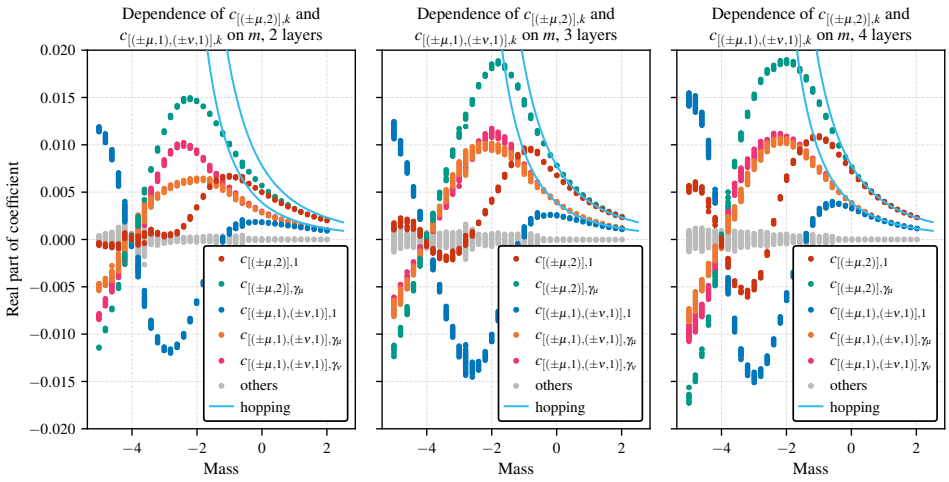


Figure 12: Same as fig. 10, but for paths of length two.

G IMPROVED LOW-ORDER APPROXIMATION FORMULAS

We provide formulas for improved low-order approximations obtained from the restricted models for mass parameters $m = 0, 1, 2$ and $n = 1, 2, 3, 4, 6$ layers. All formulas have the form

$$D^{-1} = w \sum_{\ell=0}^{n} \prod_{\ell'=1}^{\ell} \sum_{\mu=1}^{4} \left((w_{\boldsymbol{I}}^{(\ell')} \boldsymbol{I}_4 + w_{\gamma}^{(\ell')} \gamma_\mu) \boldsymbol{H}_{+\mu} + (w_{\boldsymbol{I}}^{(\ell')} \boldsymbol{I}_4 - w_{\gamma}^{(\ell')} \gamma_\mu) \boldsymbol{H}_{-\mu} \right) + \mathcal{O}(\boldsymbol{H}^{n+1}) \,, \tag{21}$$

and we provide the coefficients w, $w_{\boldsymbol{I}}^{(\ell)}$, and $w_{\gamma}^{(\ell)}$ in tables 1 to 3. For other masses, these formulas can be obtained in a straightforward manner using the code provided as supplementary material.

Table 1: Coefficients of the restricted models for mass parameter $m = 0$ and n layers. The standard deviation in the last digits is reported in parentheses. The parenthesized numbers in the header row indicate the value of the weights in the hopping expansion. The imaginary parts of the weights are not displayed if they are consistent with zero within statistics.

n	ℓ	w (0.25)	$w_{\boldsymbol{I}}^{(\ell)}$ (0.125)	$w_{\gamma}^{(\ell)}$ (0.125)
1	1	$0.231536(19)$	$0.085127(16)$	$0.10732(7)$
2	1	$0.25081(5)$	$0.09250(7)$	$0.11711(5)$
	2		$0.09267(7)$	$0.11688(6)$
3	1	$0.2542(11)$	$0.1017(12)$	$0.114(9)$
	2		$0.1014(13)$	$0.122(7)$
	3		$0.1022(9)$	$0.121(9)$
4	1	$0.25239(21) - 0.0009(7)i$	$0.106(4)$	$0.122(9)$
	2		$0.1094(27)$	$0.127(7)$
	3		$0.107(4)+0.0035(14)i$	$0.123(8) - 0.0040(16)i$
	4		$0.107(4)$	$0.121(10)$
6	1	$0.2507(4)$	$0.1067(25)$	$0.128(14)$
	2		$0.109(11)$	$0.131(5)$
	3		$0.115(6)$	$0.132(13)$
	4		$0.1181(14)$	$0.138(7)$
	5		$0.109(11)$	$0.1360(24)$
	6		$0.111(8)$	$0.134(4)$

Table 2: Same as table 1, but for $m = 1$.

n	ℓ	w (0.2)	$w_{\boldsymbol{I}}^{(\ell)}$ (0.1)	$w_{\gamma}^{(\ell)}$ (0.1)
1	1	$0.193386(29)$	$0.078651(22)$	$0.08896(8)$
2	1	$0.200858(6)$	$0.084093(16)$	$0.09601(5)$
	2		$0.084121(23)$	$0.095910(27)$
3	1	$0.2011(5)$	$0.0908(17)+0.00025(18)i$	$0.098(5) - 0.00027(20)i$
	2		$0.0888(18)$	$0.094(6)$
	3		$0.0907(16)$	$0.098(5)$
4	1	$0.20045(5)$	$0.0909(8)$	$0.0936(13)$
	2		$0.0966(9)$	$0.1026(8)$
	3		$0.092(4)$	$0.099(4)$
	4		$0.0937(24)$	$0.100(6)$
6	1	$0.20017(11)$	$0.082(13)$	$0.085(17)$
	2		$0.092(10)$	$0.094(13)$
	3		$0.081(13)$	$0.083(17)$
	4		$0.089(15)$	$0.094(15)$
	5		$0.087(13)$	$0.093(13)$
	6		$0.091(14)$	$0.094(16)$

Table 3: Same as table 1, but for $m = 2$.

n	ℓ	w (0.166…)	$w_{\boldsymbol{I}}^{(\ell)}$ (0.0833…)	$w_{\gamma}^{(\ell)}$ (0.0833…)
1	1	$0.163918(21)$	$0.070676(16)$	$0.07619(6)$
2	1	$0.1671247(30)$	$0.074569(30)$	$0.080976(9)$
	2		$0.074544(21)$	$0.080985(16)$
3	1	$0.16711(20)$	$0.0785(19)$	$0.080(4)$
	2		$0.0776(11)$	$0.079(5)$
	3		$0.0793(6)$	$0.0835(7)$
4	1	$0.16684(7)$	$0.074(16)$	$0.073(14)$
	2		$0.076(15)$	$0.076(14)$
	3		$0.074(17)$	$0.075(15)$
	4		$0.075(14)$	$0.075(14)$
6	1	$0.166697(29)$	$0.074(5)$	$0.076(5)$
	2		$0.073(13)$	$0.075(14)$
	3		$0.055(11)-0.0006(5)i$	$0.055(11)+0.0006(5)i$
	4		$0.061(18)$	$0.064(19)$
	5		$0.068(18)$	$0.071(17)$
	6		$0.073(5)$	$0.075(6)$

H COEFFICIENT EVOLUTION

We show the evolution of the coefficients $c_{p,k}$ in eq. (14) for HC and restricted models as a function of the number of layers, up to 8 layers. The coefficients $c_{[],1}$ are shown in fig. 13, and the coefficients $c_{[(\pm\mu,1)],1}$ and $c_{[(\pm\mu,1)],\gamma_\mu}$ are shown in figs. 14 and 15. As an example for a coefficient that is zero in the hopping expansion but nonzero for our models in the small-m regime, we show the staple coefficient $c_{[(\pm\mu,1),(\pm\nu,1),(\mp\mu,1)],1}$ in fig. 16. Finally, in fig. 17 we show an example for a coefficient that is nonzero in the hopping expansion, but zero for our models in the small-m regime, i.e., $c_{[(\pm\mu,3)],1}$.

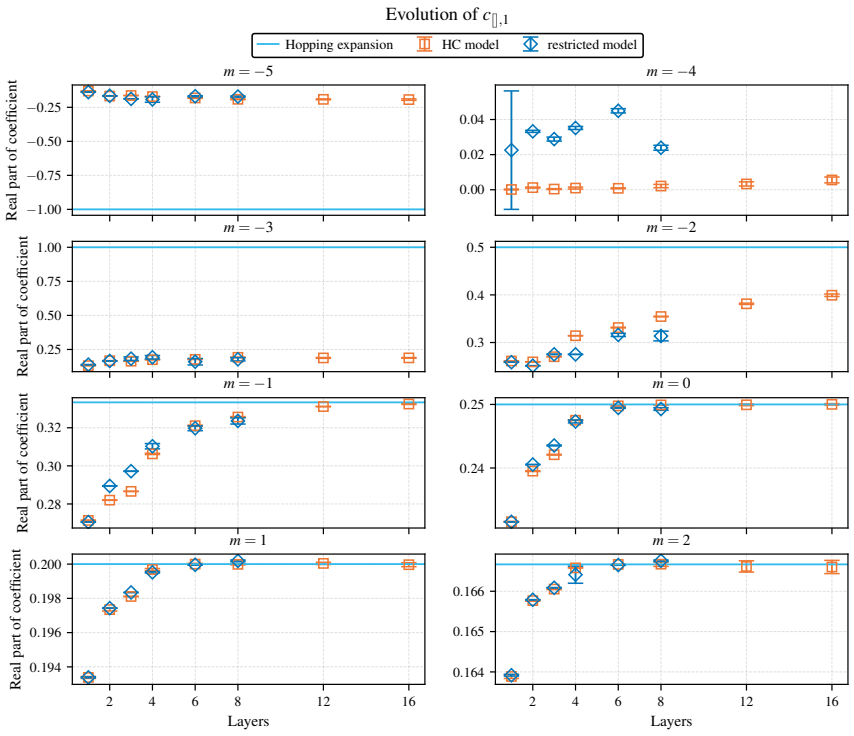


Figure 13: Evolution of the coefficient $c_{[],1}$ as a function of the number of layers for HC models at different masses.

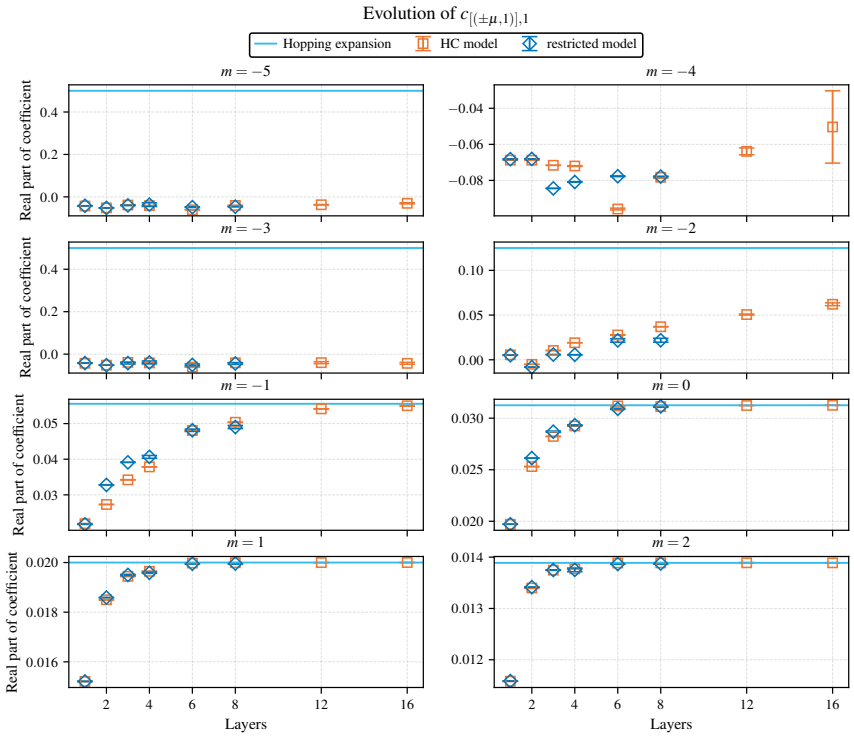


Figure 14: Same as fig. 13, but for $c_{[(\pm\mu,1)],1}$, averaged over $\pm\mu$.

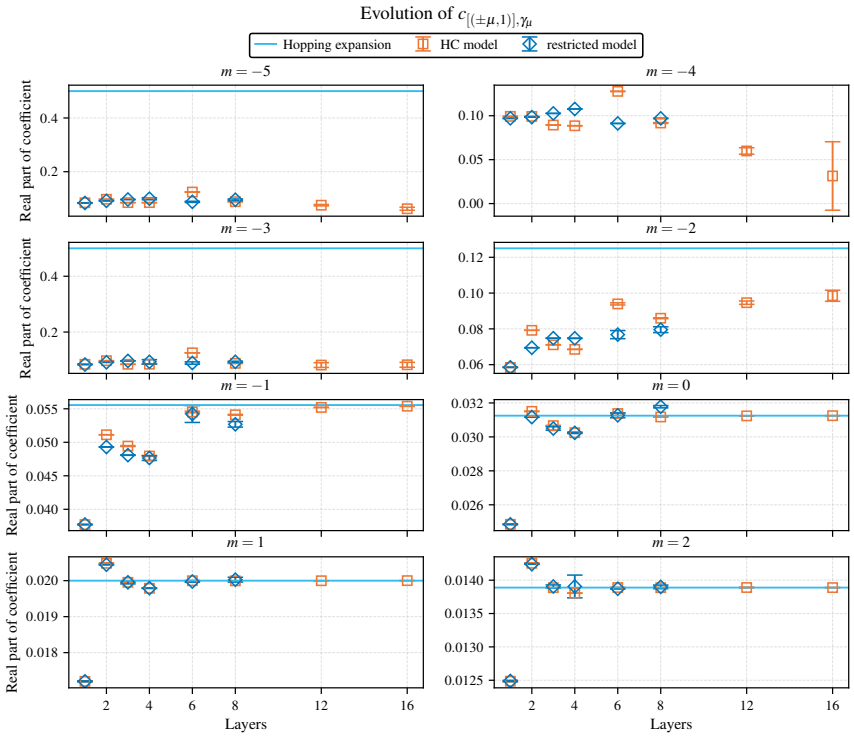


Figure 15: Same as fig. 13, but for $c_{[(\pm\mu,1)],\gamma_\mu}$, averaged over $\pm\mu$.

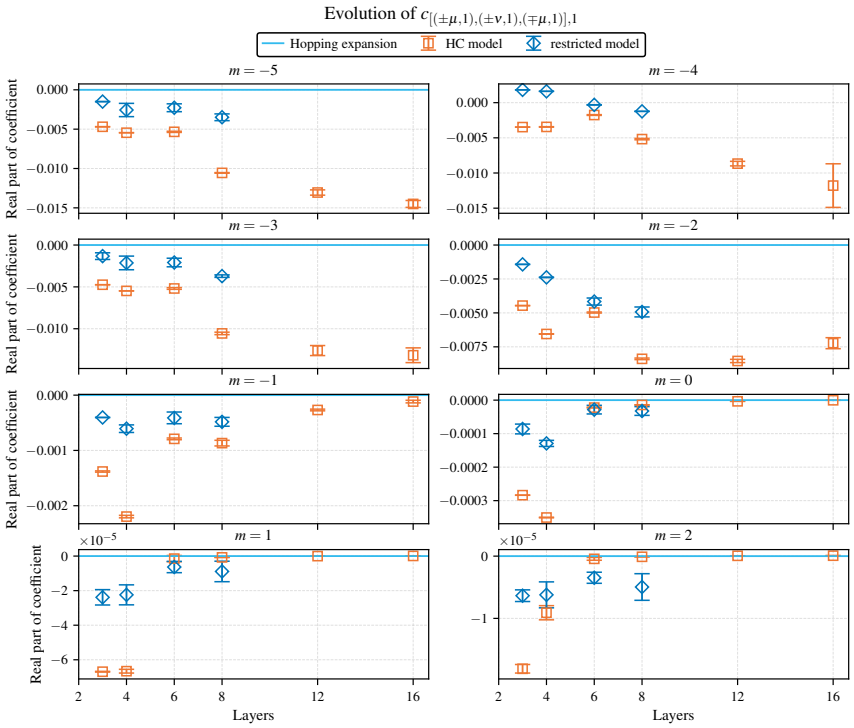


Figure 16: Same as fig. 13, but for $c_{[(\pm\mu,1),(\pm\nu,1),(\mp\mu,1)],1}$, averaged over $\pm\mu$ and $\pm\nu$.

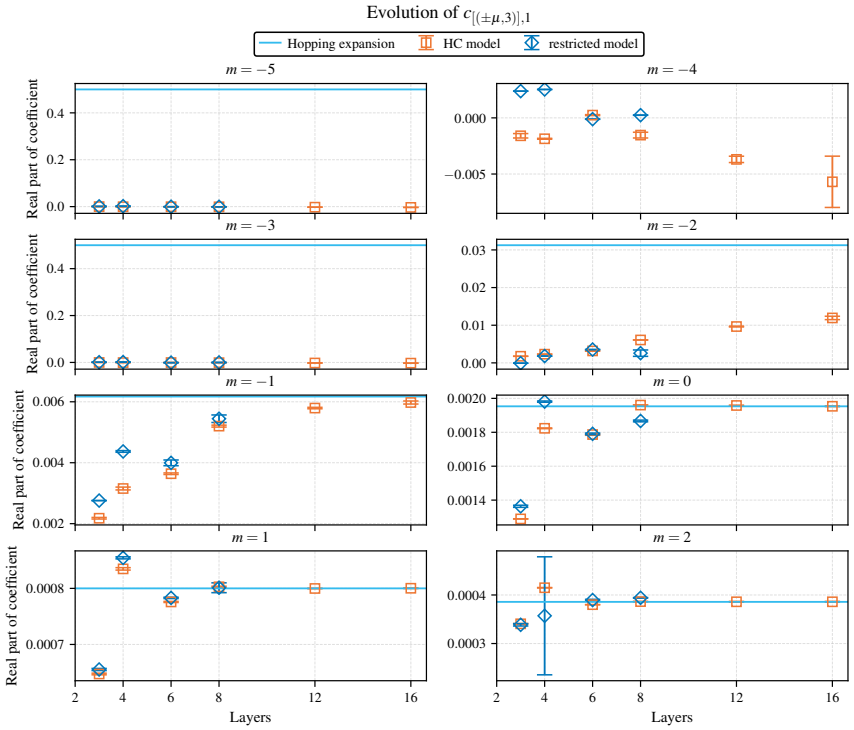


Figure 17: Same as fig. 13, but for $c_{[(\pm\mu,3)],1}$, averaged over $\pm\mu$.