PROCEEDINGS OF SCIENCE

Sampling QCD field configurations with gauge-equivariant flow models

Ryan Abbott,[a,b] **Michael S. Albergo,**[c] **Aleksandar Botev,**[g] **Denis Boyda,**[a,b,d] **Kyle Cranmer,**[c,e] **Daniel C. Hackett,**[a,b] **Gurtej Kanwar,**[a,b,f] **Alexander G. D. G. Matthews,**[g] **Sébastien Racanière,**[g] **Ali Razavi,**[g] **Danilo J. Rezende,**[g] **Fernando Romero-López,**[a,b] **Phiala E. Shanahan**[a,b,*] **and Julian M. Urban**[a,b,h]

[a] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[b] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[c] *Center for Cosmology and Particle Physics, New York University, New York, NY 10003, USA*
[d] *Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont IL 60439, USA*
[e] *Physics Department, University of Wisconsin-Madison, Madison, WI 53706, USA.*
[f] *Albert Einstein Center, Institute for Theoretical Physics, University of Bern, 3012 Bern, Switzerland*
[g] *DeepMind, London, UK*
[h] *Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany*

E-mail: phiala@mit.edu

Machine learning methods based on normalizing flows have been shown to address important challenges, such as critical slowing-down and topological freezing, in the sampling of gauge field configurations in simple lattice field theories. A critical question is whether this success will translate to studies of QCD. This Proceedings presents a status update on advances in this area. In particular, it is illustrated how recently developed algorithmic components may be combined to construct flow-based sampling algorithms for QCD in four dimensions. The prospects and challenges for future use of this approach in at-scale applications are summarized.

The 39th International Symposium on Lattice Field Theory, LATTICE2022 8th-13th August 2022, Bonn Germany

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

PoS(LATTICE2022)036

source: https://doi.org/10.48350/190365 | downloaded: 9.5.2024

1. Introduction

The rapid development of machine learning, and in particular deep learning, over the last decade has spawned a new wave of algorithms for computational physics [1–3]. In lattice quantum field theory, custom machine learning tools are being developed to accelerate almost every step of the computational workflow [4], from configuration generation [5–46] to calculation or design of observables [5, 18, 47–59] and various aspects of analysis [60–71], while maintaining rigorous guarantees of exactness.

One particular approach to accelerating gauge field generation via machine learning is based on the use of flow transformations as trivializing maps [72–75], as outlined in Sec. 2. Bespoke flow architectures tailored to this application have been constructed, including gauge-equivariant architectures designed for Abelian and non-Abelian gauge field theories [30, 31], and methods to incorporate fermions [32, 46]. Proof-of-principle implementations have demonstrated success in ameliorating important challenges such as critical slowing-down and topological freezing, as well as the computation of thermodynamic quantities, in theories ranging from scalar field theory [5, 28, 35, 39, 76] through Yukawa theory [32], to the Schwinger model [9, 43] and SU(3) gauge field theory with fermions in two dimensions [46].

The following sections outline how recent developments may be combined to construct flow transformations to enable sampling of QCD gauge field configurations, and present the first numerical demonstration of this technique applied to lattice QCD.

2. Flow transformations of lattice field configurations

In the machine learning lexicon, a flow transformation [73–75] is a diffeomorphism between manifolds, defined with a large number of free parameters such that it can be optimized, or trained, to some objective. Precisely, a flow f maps samples z from a 'prior' or base distribution $r(z)$ to samples $\varphi = f(z)$ distributed according to

$$q(\varphi) = r(z)|\det \partial f(z)/\partial z|^{-1} \, . \tag{1}$$

The free parameters of f are optimized such that $q(\varphi)$ approximates a 'target' probability distribution $p(\varphi)$, i.e., $q(\varphi) \simeq p(\varphi)$. Typically this is achieved by minimization of a 'loss function' that quantifies the difference between $q(\varphi)$ and $p(\varphi)$, such as the Kullback-Leibler (KL) divergence [77] or related measures. It is often convenient to define the loss function such that it can be computed stochastically using only samples from $r(z)$ (known as 'self-training' as no 'training data', i.e., no samples from $p(\varphi)$, are needed). Given a flow model f, samples from $p(\varphi)$ can be obtained by using samples drawn from the model distribution $q(\varphi)$ as proposals in the independence Metropolis algorithm [78–80], or by reweighting.

Applications of flow models to lattice field theory are discussed in Refs. [5, 7–9, 23, 25, 27–44]. In the context of gauge field configuration generation, φ (and z) are field configurations, and the goal is to sample efficiently from $p(\varphi) = \frac{1}{\mathcal{Z}}e^{-S(\varphi)}$, where $S(\varphi)$ is the Euclidean action of the theory and $\mathcal{Z}$ is a normalization constant. In the most direct approach, $r(z)$ can be taken as a trivial distribution (e.g., an independent uniform distribution over the Haar measure on each link for a gauge theory), and f implements an approximation of a trivializing map [72]. Naturally, many

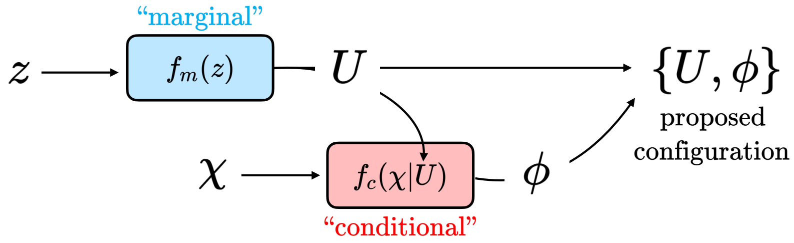


Figure 1: Sketch of the workflow of a joint model, reproduced from Ref. [46]. The notation z and χ denotes samples from the base distribution, transformed via the marginal and conditional models $f_m(z)$ and $f_c(z|U)$ to produce the gauge and pseudofermion fields, U and ϕ. The proposed configuration is distributed as $q(\phi, U) = q(U)q(\phi|U)$, in terms of the marginal and conditional model probability densities $q(U)$ and $q(\phi|U)$.

other choices are possible and may be practical for particular applications; for example, $r(z)$ may represent the distribution of configurations of a pure-gauge theory that can be sampled with the Heat Bath algorithm [81–83], or a theory with a different action or with different parameters from $p(\varphi)$. Alternatively, rather than using the flow directly for sampling, the entire construction may be embedded as a component of a more general or hybrid sampling algorithm [7–9]. Variations of flow transformations may also be used in other applications in lattice field theory, such as contour deformation and density-of-states approaches to ameliorating the sign problem [42, 57, 84, 85].

Naturally, the efficiency of sampling gauge field configurations using a flow transformation, regardless of where the flow construction appears in the sampling algorithm, will depend both on the level of optimization that has been achieved and on the flexibility, or expressivity, of the function f itself. Using architectures that respect important features of the target distribution, such as gauge symmetry [31], can be critical to achieving practical success.

3. Numerical demonstration

Combining the developments of Refs. [31, 46], all of the necessary components are in place to apply flow-based sampling of lattice gauge field configurations to QCD. This section outlines the first demonstration of this application.

For illustration, an architecture is optimized for flow-based sampling of QCD with $N_f = 2$ fermion flavors in 4D, with lattice volume 4^4, strong coupling $\beta = 1$, and $\kappa = 0.1$. As in Ref. [46], the flow is composed of two separate flow models—marginal and conditional submodels—to produce samples of the gauge and pseudofermion degrees of freedom. The resulting joint model structure is illustrated in Fig. 1. As detailed in Ref. [46], an efficient way to use joint models is to draw multiple pseudofermion samples for fixed gauge fields and average the resulting weight factors for the Metropolis accept-reject or reweighting procedure. This yields more precise estimators of the determinant of the Dirac operator; in the limit of an infinite number of pseudofermion samples, this would give the same results as using the marginal flow model with an exact evaluation of the determinant of the Dirac operator.

PoS(LATTICE2022)036

3.1 Model definition

Marginal model

The marginal model uses a Haar-uniform prior distribution and consists of 48 gauge-equivariant spline coupling layers [31], with convolutional neural networks acting in the spectral flow to define the spline parameters. Each spectral flow is parametrized by two convolutional neural networks, each made of four convolutional layers with kernel size 3. The two networks are structurally identical except for the number of input channels, which is determined by the number of plaquettes and concatenated input (see [31] for details). The two intermediate convolutions each have 32 input and output channels. The last layer has 32 input channels and 49 output channels to parametrize 16 spline knots and an offset. The "Leaky ReLU" activation function was used in intermediate layers, but no activation is applied to the final output. Spatially separated convolutions are used; i.e., the four-dimensional convolutional neural networks consist of blocks, where every block is the sequential application of a one-dimensional convolution along each of the four dimensions, with a bias for only the last one. The number of channels between each one-dimensional convolution is set to the number of output channels. The weights and biases of the convolutions comprise all free parameters of the architecture. The gauge-invariant inputs to the inner spectral flow are the real and imaginary parts of the traces of all frozen plaquettes. Masking patterns are generated with the algorithm for higher-dimensional masks presented in Ref. [86], using the parameters *spatial-dims* = [4, 4, 4, 4], *mask-orientations* = [0], *width* = 4, *active-phase* = 0, *orientations-to-shift* = [0, 1, 2, 3], and *shifts* = [1, 2, 1, 1]. These parameters are varied across flow layers using the alternation scheme that always sets *mask-orientations*[0] to *orientations-to-shift*[0], also described in Ref. [86].

The marginal model is constructed and trained prior to the conditional model being trained. The convolution weights are first initialized using the Xavier normal [87] scheme with gain parameter 0.5, and biases are set to zero. After initialization, self-training proceeds by minimizing stochastic estimates of the "reverse" KL loss [28] computed with flow model samples, as discussed above. Gradients of the fermion determinant are estimated stochastically as described in Appendix C of Ref. [32]. During training, the Dirac operator is regulated as $(DD^{\dagger} + \mu_0)$, with $\mu_0 = 10^{-5}$; no regulator is used in evaluation. The marginal model is trained for 21k steps with batch size 512. The initial learning rate is 10^{-3}, and decreases by a factor 0.8 every 10k steps. All optimization is performed with the Adam optimizer [88] with $\epsilon = 10^{-2}$. Clipping is applied to the norm and value of gradients [89]; the maximum norm value was set to 10, and maximum value to 0.1.

Conditional model

The conditional model uses a Gaussian prior and consists of 36 pseudofermion layers built from parallel transport convolutional networks (PTCN), defined in Ref. [46]. The masking pattern rotates through the spin projectors $\frac{1}{2}(1 \pm \gamma_\mu)$ for $\mu \in \{0, 1, 2, 3, 5\}$ along with even/odd spatial projectors. In the notation of Ref. [46], the PTCNs use $n_{PT} = 6$ and $H = 4$. The context function of each PTCN is a convolutional neural network—as described in the Appendix B.2 of Ref. [46]—and it has six input channels, namely $\sum_{\mu\neq\nu}$, $\mathrm{Re}\,\mathrm{Tr}\, P_{\mu\nu}$, $\sum_{\mu\neq\nu} \mathrm{Im}\,\mathrm{Tr}\, P_{\mu\nu}$, I_0, I_1, I_2, and I_3 where $P_{\mu\nu}(x)$ is the plaquette and $I_i(x) = x_i \bmod 4$. The part of the context function that is shared among all PTCNs has 8 hidden channels, with "ELU" used as the inner activation function. The final activation function is "tanh". All layers have biases and use kernel size 3.

PoS(LATTICE2022)036

The weights in the networks defining the conditional model architecture are initialized with the Xavier normal [87] scheme with a gain of 0.3. The biases are initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.05. Conditional model training then proceeds by self-training the joint model for gauge and pseudofermion fields, similar to Ref. [46], except here the marginal model is kept fixed. The conditional model is optimized in 3 phases: for 16k steps with batch size 640, learning rate 10^{-4}; for 28k steps with batch size 640, learning rate 5×10^{-5}, stepped to 2.5×10^{-5} after 20k steps; for 17k steps with batch size 800, learning rate 2.5×10^{-5}. During training, the Dirac operator is regulated as described above for the marginal model. The Adam optimizer [88] is used with the same settings as for the training of the marginal model: $\epsilon = 10^{-2}$, and gradient clipping to norm/value with values 10/0.1 is applied, as for the marginal model.

3.2 Results

The quality of the trained model can be assessed using various measures, including the acceptance rate of samples used as proposals in independence Metropolis or the Effective Sample Size per configuration (ESS), defined from N samples as [44] ESS $= \frac{1}{N}(\sum_{i=1}^{N} w_i)^2 / \sum_{i=1}^{N} w_i^2$, with $w_i = p(\varphi_i)/q(\varphi_i)$. ESS increases with model quality, with ESS $\in [1/N, 1]$. With training as detailed above, the model achieves an ESS of approximately 0.1 with 512 pseudofermion draws. For comparison, reweighting directly from the base distribution to the target gives an ESS of $\lesssim 10^{-4}$, estimated using 65k Haar-uniform configurations and exact evaluation of the determinant in the action.

Comparing observables computed on an ensemble sampled from the trained flow model with those computed on an ensemble prepared using Hybrid Monte Carlo (HMC) at the same parameters shows statistical consistency between the algorithms, as is expected. Precisely, an ensemble of 65k configurations is sampled from the trained flow model; given the ESS of approximately 0.1, this corresponds to approximately 6.5k independent samples. Measurements on each configuration are reweighted when computing observables. For comparison, 65k configurations are generated for the same parameters using HMC as implemented in the `chroma` library [90]. Trajectories of length $\tau = 0.5$ are integrated with five leapfrog steps each, yielding an acceptance rate of $\sim 95\%$. Configurations are generated in 8 independent streams; the initial 500 trajectories in each are discarded for equilibration. The ensemble is thinned by a factor of 10 to remove autocorrelations, for a total of 6.5k independent samples. Figure 2 collates a comparison of several observables computed on both ensembles:

- Plaquette:

  $$P = \frac{1}{N_c}\frac{1}{L^4}\sum_{x}\sum_{\mu<\nu} \mathrm{Re}\,\mathrm{tr}\, P_{\mu\nu}(x), \tag{2}$$

  where $N_c = 3$ is the number of colors, $L = 4$ is the extent of the lattice geometry, and $P_{\mu\nu}$ denotes the 1×1 Wilson loop which extends in the μ and ν directions;

- Polyakov loop:

  $$L = \frac{1}{L^3}\sum_{\vec{x}} \mathrm{tr} \prod_{x_0} U_0(x_0, \vec{x}), \tag{3}$$

  where U_0 is the gauge link in the time direction;

PoS(LATTICE2022)036

- Pion correlation function:

$$C_\pi(x_0) = -\sum_{\vec{x}} \langle [\bar{u}\gamma_5 d](x_0, \vec{x}) [\bar{d}\gamma_5 u](0, \vec{0}) \rangle, \tag{4}$$

  measured using point sources;

- Topological charge:

$$Q = \frac{1}{16\pi^2} \sum_x \epsilon_{\mu\nu\rho\sigma} F_{\mu\nu}(x) F_{\rho\sigma}(x), \tag{5}$$

  where $\epsilon_{\mu\nu\rho\sigma}$ is the Levi-Civita tensor and $F_{\mu\nu}$ is defined through the clover definition. The topological charge is computed on ensembles subject to Wilson flow [72] to flow time $t/a^2 = 4$ in order to reveal the structure of the distribution.

Clearly, statistically consistent results are obtained with flow-based sampling as compared with HMC. The small discrepancies in some bins are consistent with statistical fluctuations, as illustrated in Fig. 3, which displays the distribution over all observables and all bins of residuals of the flow results relative to the HMC benchmark, as compared with the χ^2 distribution.

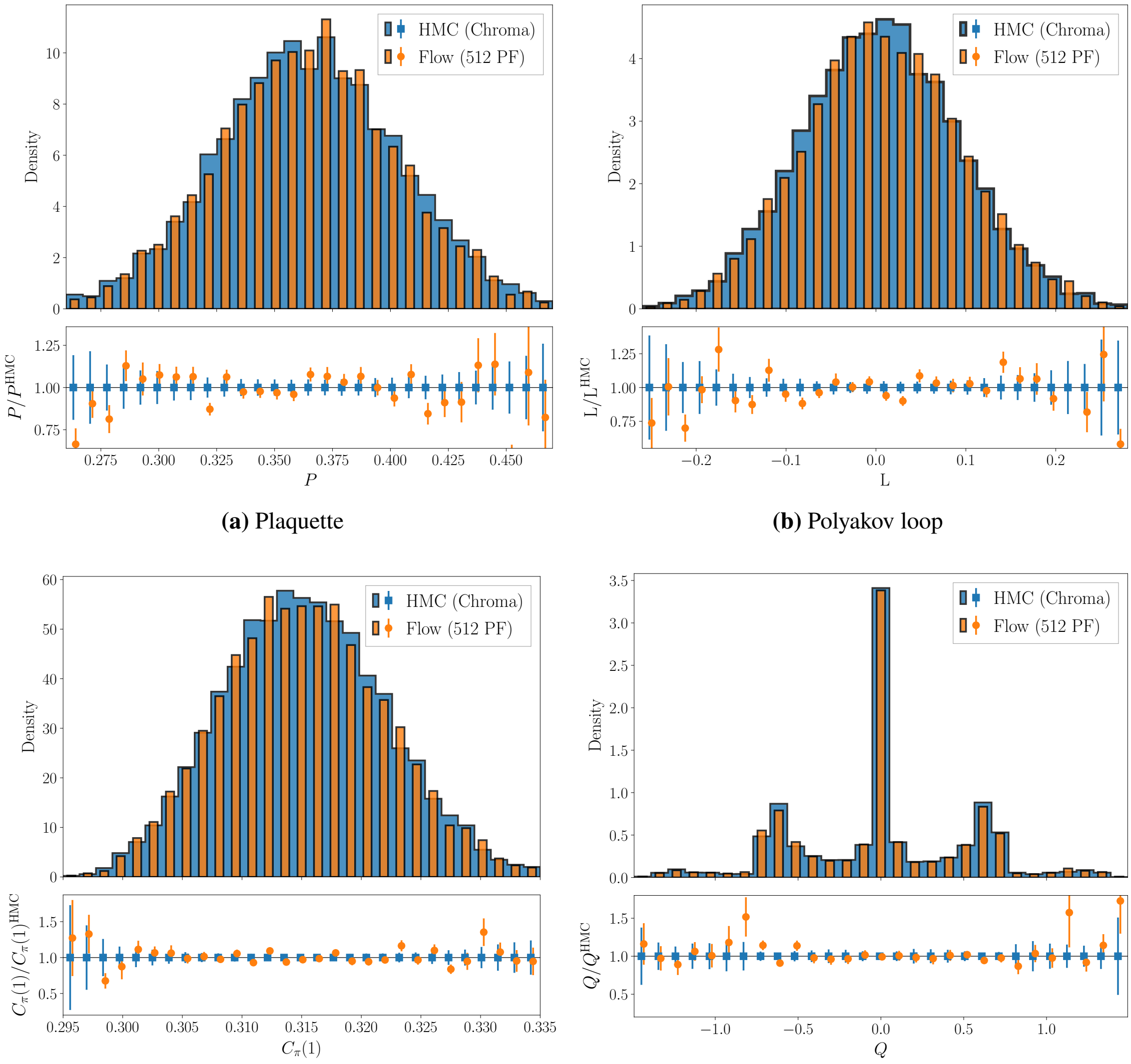
4. Outlook

Section 3 demonstrates the basic technology for the application of flow-based sampling algorithms to lattice QCD. The key question, naturally, is whether any such approach will provide a computational advantage over HMC at the scale of state-of-the-art studies. As discussed at length in Ref. [91], this is a challenging question that can only be answered through experimental investigation.

The key difficulty in assessing the viability of flow-based sampling for lattice QCD is that the effectiveness of the approach depends on a very large number of choices, few of which have been explored to date. In particular, different flow architectures and training approaches yield not only different model quality and hence algorithm performance, but different scaling of that quality across parameter space. Moreover, there is a complicated trade-off between investment in training, and hence model quality, and the resulting acceleration of sampling; the optimal choice of stopping condition for optimization depends on how many configurations, at how many different sets of parameters, are ultimately desired, and may involve artificially truncating training of the model. To give just one example of the complexity of this question; incorporating the latest advances in flow model architecture design, flow-based sampling of scalar field theory configurations can be achieved that is at least an order of magnitude more efficient than the first implementation of the approach in Ref. [28]. For QCD in particular, while a first demonstration has now been achieved, the exploration of the space of suitable architectures has barely begun, and significant advances can be expected through further development.

Complicating the picture further, it is very likely that the first practical, at-scale applications of flow transformations in lattice QCD will not be in the context of direct sampling from the model, but via one of the many possible hybrid algorithms that can be developed to exploit the same or similar flow architectures, as noted in the introduction. Naturally, each such approach presents its own set of advantages and challenges with regard to scaling, which must be assessed independently.

PoS(LATTICE2022)036



(a) Plaquette

(b) Polyakov loop

(c) Pion correlation function at $x_0 = 1$

(d) Topological charge at $t/a^2 = 4$

Figure 2: Density histograms comparing various observables computed on ensembles of configurations sampled from the flow model, and generated via HMC, with equivalent statistics as described in the text. The lower panel of each subfigure shows the ratio of the counts in each bin of the histogram divided by the central value of the counts obtained in the same bin using the HMC ensemble; the uncertainties are estimated from 1000 bootstrap ensembles.

Despite the unknowns, what is nevertheless evident is that the practical application of flow-based sampling in lattice QCD will require the feat of engineering a custom machine learning architecture to a similar scale as those routinely used in industrial applications. Moreover, the deployment of that model will likely require tight coupling with traditional lattice QCD codes on high-performance computing resources. Despite the clear challenge this presents, it is also clear that much stands to be gained; in particular, the approach could enable efficient ensemble generation at fine lattice spacings where HMC suffers from effects such as topological freezing and critical slowing-down, and offers clear advantages in studies of thermodynamic observables by providing

PoS(LATTICE2022)036

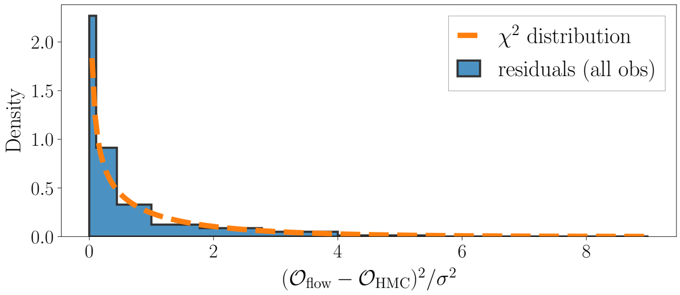


Figure 3: Distribution of the residuals of all the bins in the four subfigures in Fig. 2. Uncertainties of the results obtained with the flow and HMC ensembles are added in quadrature. The dashed line illustrates a χ^2 distribution with 1 degree of freedom. Binning is chosen to be uniform in the square root of the x-axis.

direct access to the partition function. As a single trained model may be used to generate ensembles efficiently across parameter space, one can imagine using the same core model components to generate the diverse ensembles required for the control over the continuum limit in precision flavour physics applications, elongated lattice geometries or anisotropic field configurations for thermodynamics or spectroscopy, and high statistics to beat down signal-to-noise challenges in nuclear physics applications. If they can be achieved, optimized flow models for QCD are thus likely to become a valuable community resource, shared and re-used extensively for applications across lattice QCD.

Acknowledgements

RA, DCH, GK, FRL, and PES are supported in part by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, under grant Contract Number DE-SC0011090. DCH, FRL and PES thank the Institute for Nuclear Theory at the University of Washington for its kind hospitality and stimulating research environment. This research was supported in part by the INT's U.S. Department of Energy grant No. DE-FG02-00ER41132. PES is additionally supported by the National Science Foundation under EAGER grant 2035015, by the U.S. DOE Early Career Award DE-SC0021006, by a NEC research award, and by the Carl G and Shirley Sontheimer Research Fund. GK is additionally supported by the Schweizerischer Nationalfonds. KC and MSA are supported by the National Science Foundation under the award PHY-2141336. MSA thanks the Flatiron Institute for their hospitality. DB is supported by the Argonne Leadership Computing Facility, which is a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster), the Collaborative Research Centre SFB 1225 (ISOQUANT), and the U.S. National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This work is associated with an ALCF Aurora Early Science Program project, and used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center [92] for providing HPC resources that

PoS(LATTICE2022)036

have contributed to the research results reported within this paper. Numerical experiments and data analysis used PyTorch [93], JAX [94], Haiku [95], Horovod [96], NumPy [97], SciPy [98], CHROMA [90], and QUDA [99]. Figures were produced using matplotlib [100].

References

[1] J. Carrasquilla *Adv. Phys. X* **5** no. 1, (2020) 1797528, arXiv:2003.11040 [physics.comp-ph].

[2] A. Boehnlein *et al.* arXiv:2112.02309 [nucl-th].

[3] A. Dawid *et al.* 4, 2022. arXiv:2204.04198 [quant-ph].

[4] D. Boyda *et al.* in *2022 Snowmass Summer Study*. 2, 2022. arXiv:2202.05838 [hep-lat].

[5] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati *Phys. Rev. Lett.* **126** no. 3, (2021) 032001, arXiv:2007.07115 [hep-lat].

[6] M. Medvidovic, J. Carrasquilla, L. E. Hayward, and B. Kulchytskyy arXiv:2012.01442 [cond-mat.dis-nn].

[7] S. Foreman, X.-Y. Jin, and J. C. Osborn in *9th International Conference on Learning Representations*. 5, 2021. arXiv:2105.03418 [hep-lat].

[8] S. Foreman, T. Izubuchi, L. Jin, X.-Y. Jin, J. C. Osborn, and A. Tomiya in *38th International Symposium on Lattice Field Theory*. Dec, 2021. arXiv:2112.01586 [cs.LG].

[9] J. Finkenrath arXiv:2201.02216 [hep-lat].

[10] L. Wang *Phys. Rev. E* **96** (Nov, 2017) 051301. https://link.aps.org/doi/10.1103/PhysRevE.96.051301.

[11] L. Huang and L. Wang *Physical Review B* **95** no. 3, (Jan, 2017) –. http://dx.doi.org/10.1103/PhysRevB.95.035105.

[12] J. Song, S. Zhao, and S. Ermon in *Advances in Neural Information Processing Systems*, pp. 5140–5150. 2017.

[13] A. Tanaka and A. Tomiya arXiv:1712.03893 [hep-lat].

[14] D. Levy, M. D. Hoffman, and J. Sohl-Dickstein arXiv:1711.09268 [stat.ML].

[15] J. M. Pawlowski and J. M. Urban *Mach. Learn. Sci. Tech.* **1** (2020) 045011, arXiv:1811.03533 [hep-lat].

[16] G. Cossu, L. Del Debbio, T. Giani, A. Khamseh, and M. Wilson *Phys. Rev. B* **100** no. 6, (2019) 064304, arXiv:1810.11503 [physics.comp-ph].

PoS(LATTICE2022)036

[17] D. Wu, L. Wang, and P. Zhang *Phys. Rev. Lett.* **122** (Feb, 2019) 080602. https://link.aps.org/doi/10.1103/PhysRevLett.122.080602.

[18] D. Bachtis, G. Aarts, and B. Lucini *Phys. Rev. E* **102** no. 3, (2020) 033303, arXiv:2004.14341 [cond-mat.stat-mech].

[19] Y. Nagai, A. Tanaka, and A. Tomiya arXiv:2010.11900 [hep-lat].

[20] A. Tomiya and Y. Nagai arXiv:2103.11965 [hep-lat].

[21] D. Bachtis, G. Aarts, F. Di Renzo, and B. Lucini *Phys. Rev. Lett.* **128** no. 8, (2022) 081603, arXiv:2107.00466 [hep-lat].

[22] D. Wu, R. Rossi, and G. Carleo *Phys. Rev. Res.* **3** no. 4, (2021) L042024, arXiv:2105.05650 [cond-mat.stat-mech].

[23] M. Gerdes, P. de Haan, C. Rainone, R. Bondesan, and M. C. N. Cheng arXiv:2207.00283 [hep-lat].

[24] A. Alexandru, P. F. Bedaque, H. Lamm, and S. Lawrence *Phys. Rev. D* **96** no. 9, (2017) 094505, arXiv:1709.01971 [hep-lat].

[25] S. Lawrence and Y. Yamauchi *Phys. Rev. D* **103** no. 11, (2021) 114509, arXiv:2101.05755 [hep-lat].

[26] J.-L. Wynen, E. Berkowitz, S. Krieg, T. Luu, and J. Ostmeyer *Phys. Rev. B* **103** no. 12, (2021) 125153, arXiv:2006.11221 [cond-mat.str-el].

[27] S.-H. Li and L. Wang *Phys. Rev. Lett.* **121** (Dec, 2018) 260601. https://link.aps.org/doi/10.1103/PhysRevLett.121.260601.

[28] M. S. Albergo, G. Kanwar, and P. E. Shanahan *Phys. Rev. D* **100** no. 3, (2019) 034515, arXiv:1904.12072 [hep-lat].

[29] K. A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller, and P. Kessel *Phys. Rev. E* **101** no. 2, (2020) 023304, arXiv:1910.13496 [cond-mat.stat-mech].

[30] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan *Phys. Rev. Lett.* **125** no. 12, (2020) 121601, arXiv:2003.06413 [hep-lat].

[31] D. Boyda, G. Kanwar, S. Racanière, D. J. Rezende, M. S. Albergo, K. Cranmer, D. C. Hackett, and P. E. Shanahan *Phys. Rev. D* **103** no. 7, (2021) 074504, arXiv:2008.05456 [hep-lat].

[32] M. S. Albergo, G. Kanwar, S. Racanière, D. J. Rezende, J. M. Urban, D. Boyda, K. Cranmer, D. C. Hackett, and P. E. Shanahan *Phys. Rev. D* **104** no. 11, (2021) 114507, arXiv:2106.05934 [hep-lat].

PoS(LATTICE2022)036

[33] D. C. Hackett, C.-C. Hsieh, M. S. Albergo, D. Boyda, J.-W. Chen, K.-F. Chen, K. Cranmer, G. Kanwar, and P. E. Shanahan arXiv:2107.00734 [hep-lat].

[34] M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden *Proc. Nat. Acad. Sci.* **119** no. 10, (2022) e2109420119, arXiv:2105.12603 [physics.data-an].

[35] L. Del Debbio, J. M. Rossney, and M. Wilson arXiv:2105.12481 [hep-lat].

[36] S. Foreman, X.-Y. Jin, and J. C. Osborn *PoS* **LATTICE2021** (2022) 508, arXiv:2112.01582 [hep-lat].

[37] X.-Y. Jin *PoS* **LATTICE2021** (2022) 600, arXiv:2201.01862 [hep-lat].

[38] P. de Haan, C. Rainone, M. C. N. Cheng, and R. Bondesan arXiv:2110.02673 [cs.LG].

[39] A. Singha, D. Chakrabarti, and V. Arora arXiv:2207.00980 [hep-lat].

[40] A. G. D. G. Matthews, M. Arbel, D. J. Rezende, and A. Doucet arXiv:2201.13117 [stat.ML].

[41] M. Caselle, E. Cellini, A. Nada, and M. Panero arXiv:2201.08862 [hep-lat].

[42] J. M. Pawlowski and J. M. Urban arXiv:2203.01243 [hep-lat].

[43] M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, G. Kanwar, S. Racanière, D. J. Rezende, F. Romero-López, P. E. Shanahan, and J. M. Urban arXiv:2202.11712 [hep-lat].

[44] M. S. Albergo, D. Boyda, D. C. Hackett, G. Kanwar, K. Cranmer, S. Racanière, D. J. Rezende, and P. E. Shanahan arXiv:2101.08176 [hep-lat].

[45] M. Favoni, A. Ipp, D. I. Müller, and D. Schuh *Phys. Rev. Lett.* **128** no. 3, (2022) 032003, arXiv:2012.12901 [hep-lat].

[46] R. Abbott *et al.* arXiv:2207.08945 [hep-lat].

[47] D. L. Boyda, M. N. Chernodub, N. V. Gerasimeniuk, V. A. Goy, S. D. Liubimov, and A. V. Molochkov *Phys. Rev. D* **103** no. 1, (2021) 014509, arXiv:2009.10971 [hep-lat].

[48] D. Bachtis, G. Aarts, and B. Lucini *Phys. Rev. Res.* **3** no. 1, (2021) 013134, arXiv:2010.00054 [hep-lat].

[49] A. Palermo, L. Anderlini, M. P. Lombardo, A. Y. Kotov, and A. Trunin *PoS* **LATTICE2021** (2022) 030, arXiv:2111.05216 [hep-lat].

[50] D. R. Tan, J. H. Peng, Y. H. Tseng, and F. J. Jiang *Eur. Phys. J. Plus* **136** no. 11, (2021) 1116, arXiv:2103.10846 [cond-mat.dis-nn].

[51] C.-D. Li, D.-R. Tan, and F.-J. Jiang *Annals Phys.* **391** (2018) 312–331, arXiv:1703.02369 [cond-mat.dis-nn].

PoS(LATTICE2022)036

[52] S. J. Wetzel and M. Scherzer *Phys. Rev. B* **96** no. 18, (2017) 184410, arXiv:1705.05582 [cond-mat.stat-mech].

[53] C. Alexandrou, A. Athenodorou, C. Chrysostomou, and S. Paul *Eur. Phys. J. B* **93** no. 12, (2020) 226, arXiv:1903.03506 [cond-mat.stat-mech].

[54] S. Blücher, L. Kades, J. M. Pawlowski, N. Strodthoff, and J. M. Urban *Phys. Rev. D* **101** no. 9, (2020) 094507, arXiv:2003.01504 [hep-lat].

[55] H. M. Yau and N. Su *SciPost Phys. Core* **5** (2022) 032, arXiv:2006.15021 [cond-mat.dis-nn].

[56] K. Zhou, G. Endrődi, L.-G. Pang, and H. Stöcker *Phys. Rev. D* **100** no. 1, (2019) 011501, arXiv:1810.12879 [hep-lat].

[57] W. Detmold, G. Kanwar, H. Lamm, M. L. Wagman, and N. C. Warrington *Phys. Rev. D* **103** no. 9, (2021) 094517, arXiv:2101.12668 [hep-lat].

[58] W. Hu, R. R. Singh, and R. T. Scalettar *Physical Review E* **95** no. 6, (2017) 062122, arXiv:1704.00080 [cond-mat.stat-mech].

[59] S. J. Wetzel, R. G. Melko, J. Scott, M. Panju, and V. Ganesh *Phys. Rev. Res.* **2** no. 3, (2020) 033499, arXiv:2003.04299 [physics.comp-ph].

[60] B. Yoon, T. Bhattacharya, and R. Gupta *Phys. Rev. D* **100** no. 1, (2019) 014504, arXiv:1807.05971 [hep-lat].

[61] R. Zhang, Z. Fan, R. Li, H.-W. Lin, and B. Yoon *Phys. Rev. D* **101** no. 3, (2020) 034516, arXiv:1909.10990 [hep-lat].

[62] P. E. Shanahan, D. Trewartha, and W. Detmold *Phys. Rev. D* **97** no. 9, (2018) 094506, arXiv:1801.05784 [hep-lat].

[63] R. J. Hudspith and D. Mohler arXiv:2112.01997 [hep-lat].

[64] L. Kades, J. M. Pawlowski, A. Rothkopf, M. Scherzer, J. M. Urban, S. J. Wetzel, N. Wink, and F. P. G. Ziegler *Phys. Rev. D* **102** no. 9, (2020) 096001, arXiv:1905.04305 [physics.comp-ph].

[65] S. Offler, G. Aarts, C. Allton, J. Glesaaen, B. Jäger, S. Kim, M. P. Lombardo, S. M. Ryan, and J.-I. Skullerud *PoS* **LATTICE2019** (2019) 076, arXiv:1912.12900 [hep-lat].

[66] J. Horak, J. M. Pawlowski, J. Rodríguez-Quintero, J. Turnwald, J. M. Urban, N. Wink, and S. Zafeiropoulos *Phys. Rev. D* **105** no. 3, (2022) 036014, arXiv:2107.13464 [hep-ph].

[67] S. Y. Chen, H. T. Ding, F. Y. Liu, G. Papp, and C. B. Yang arXiv:2110.13521 [hep-lat].

[68] L. Wang, S. Shi, and K. Zhou in *35th Conference on Neural Information Processing Systems*. 12, 2021. arXiv:2112.06206 [physics.comp-ph].

PoS(LATTICE2022)036

[69] S. Shi, L. Wang, and K. Zhou arXiv:2201.02564 [hep-ph].

[70] L. Del Debbio, T. Giani, J. Karpie, K. Orginos, A. Radyushkin, and S. Zafeiropoulos *JHEP* **02** (2021) 138, arXiv:2010.03996 [hep-ph].

[71] J. Karpie, K. Orginos, A. Rothkopf, and S. Zafeiropoulos *JHEP* **04** (2019) 057, arXiv:1901.05408 [hep-lat].

[72] M. Lüscher *Commun. Math. Phys.* **293** (2010) 899–919, arXiv:0907.5491 [hep-lat].

[73] D. J. Rezende and S. Mohamed arXiv:1505.05770 [stat.ML].

[74] L. Dinh, J. Sohl-Dickstein, and S. Bengio arXiv:1605.08803 [cs.LG].

[75] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan *Journal of Machine Learning Research* **22** no. 57, (2021) 1–64.

[76] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati *PoS* **LATTICE2021** (2022) 338, arXiv:2111.11303 [hep-lat].

[77] S. Kullback and R. A. Leibler *The Annals of Mathematical Statistics* **22** no. 1, (1951) 79 – 86.

[78] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller *J. Chem. Phys.* **21** (1953) 1087–1092.

[79] W. K. Hastings *Biometrika* **57** (1970) 97–109.

[80] L. Tierney *the Annals of Statistics* (1994) 1701–1728.

[81] M. Creutz *Phys. Rev. Lett.* **43** (1979) 553–556. [Erratum: Phys.Rev.Lett. 43, 890 (1979)].

[82] N. Cabibbo and E. Marinari *Phys. Lett. B* **119** (1982) 387–390.

[83] A. D. Kennedy and B. J. Pendleton *Phys. Lett. B* **156** (1985) 393–399.

[84] Y. Yamauchi and S. Lawrence *PoS* **LATTICE2021** (2022) 621, arXiv:2112.15035 [hep-lat].

[85] S. Lawrence, H. Oh, and Y. Yamauchi arXiv:2205.12303 [hep-lat].

[86] R. Abbott *et al. Flow-based sampling for lattice gauge theories in higher dimensions (in preparation)* (2022) .

[87] K. He, X. Zhang, S. Ren, and J. Sun arXiv:1502.01852 [cs.CV].

[88] D. P. Kingma and J. Ba arXiv:1412.6980 [cs.LG].

[89] J. Zhang, T. He, S. Sra, and A. Jadbabaie arXiv:1905.11881 [math.OC].

[90] SciDAC, LHPC, UKQCD Collaboration, R. G. Edwards and B. Joo *Nucl. Phys. B Proc. Suppl.* **140** (2005) 832, arXiv:hep-lat/0409003.

[91] R. Abbott *et al. Aspects of scaling and scalability for flow-based sampling of lattice QCD (in preparation)* (2022) .

[92] A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, *et al. 2018 IEEE High Performance extreme Computing Conference (HPEC)* (Sep, 2018) 1–6, `arXiv:1807.07814 [cs.DC]`.

[93] A. Paszke *et al.* in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., pp. 8024–8035. Curran Associates, Inc., 2019. `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance\-deep-learning-library.pdf`.

[94] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, 2018. `http://github.com/google/jax`.

[95] T. Hennigan, T. Cai, T. Norman, and I. Babuschkin, 2020. `http://github.com/deepmind/dm-haiku`.

[96] A. Sergeev and M. Del Balso `arXiv:1802.05799 [cs.LG]`.

[97] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al. Nature* **585** no. 7825, (2020) 357–362.

[98] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al. Nature methods* **17** no. 3, (2020) 261–272.

[99] M. A. Clark, R. Babich, K. Barros, R. C. Brower, and C. Rebbi *Comput. Phys. Commun.* **181** (2010) 1517–1528, `arXiv:0911.3191 [hep-lat]`.

[100] J. D. Hunter *Computing in Science & Engineering* **9** no. 3, (2007) 90–95.

PoS(LATTICE2022)036