

Simulating correlated electrons with symmetry-enforced normalizing flows

Dominic Schuh^{1,2,*}, Janik Kreit^{1,2,†}, Evan Berkowitz^{3,4,5,6}, Lena Funcke^{1,2}, Thomas Luu^{2,4},
Kim A. Nicoli^{1,2,7} and Marcel Rodekamp⁸

¹Transdisciplinary Research Area (TRA) Matter, *University of Bonn*, Bonn, Germany

²Helmholtz Institute for Radiation and Nuclear Physics (HISKP), *University of Bonn*, Bonn, Germany

³Jülich Supercomputing Center (JSC), *Forschungszentrum Jülich*, Jülich, Germany

⁴Institute for Advanced Simulation 4 (IAS-4), *Forschungszentrum Jülich*, Jülich, Germany

⁵Center for Advanced Simulation and Analytics (CASA), *Forschungszentrum Jülich*, Jülich, Germany

⁶Department of Chemical and Physical Sciences, *College of Science and Mathematics*,

University of the Virgin Islands, U.S. Virgin Islands, Kingshill, St. Croix, USA

⁷Oldendorff Carriers GmbH & Co. KG, *Lübeck*, Regensburg, Germany

⁸Institut für Theoretische Physik, *Universität Regensburg*, Regensburg, Germany



(Received 22 July 2025; accepted 28 January 2026; published 25 March 2026)

We present the first proof of principle that normalizing flows can accurately learn the Boltzmann distribution of the fermionic Hubbard model—a key framework for describing the electronic structure of graphene and related materials. State-of-the-art methods like hybrid Monte Carlo often suffer from ergodicity issues near the time-continuum limit, leading to biased estimates. Leveraging symmetry-aware architectures as well as independent and identically distributed sampling, our approach resolves these issues and achieves significant speedups over traditional methods.

DOI: [10.1103/4741-cvvs](https://doi.org/10.1103/4741-cvvs)

Introduction. Understanding systems of strongly correlated electrons is fundamental to both condensed matter systems and quantum mechanics [1]. One of the most widely used theoretical frameworks for studying such systems is the Hubbard model [2–4], which captures the essential competition between electron kinetic energy and on-site interactions.

Over the years, a variety of methods have been developed to analyze the Hubbard model. In the weak interaction regime, perturbative approaches provide valuable insights [5]. However, outside this regime, nonperturbative effects become significant, rendering perturbative techniques insufficient. Nonperturbative methods are formulated either within a Hamiltonian framework, e.g., as in approaches based on neural quantum states [6], or within an action-based framework focusing on the finite-temperature auxiliary-field formulation of the Hubbard model. In the latter case, Monte Carlo simulations become an indispensable tool (see, e.g., Refs. [7–14] and references therein). Monte Carlo techniques, such as hybrid (or Hamiltonian) Monte Carlo (HMC), can suffer from (practical) ergodicity problems when potential barriers appear between modes of the sampled Boltzmann distribution [15]. These barriers can trap the sampler, leading to biased estimates of physical observables. A recent study addressed this issue by enhancing HMC with radial updates [16].

In this work, we develop the first flow-based deep generative machine learning approach to the finite-temperature auxiliary-field formulation of the Hubbard model on a half-filled bipartite lattice, designed to overcome ergodicity issues and enable efficient sampling from the Boltzmann distribution. We construct and train a symmetry-aware generative model and systematically compare its accuracy and efficiency to those of traditional HMC methods.

Generative neural samplers (GNSs), also known as Boltzmann generators [17], have demonstrated strong capabilities for modeling Boltzmann distributions of physical and chemical systems. In recent years, GNSs have been applied across a range of fields, including lattice quantum field theory [18–21], statistical mechanics [22,23], string theory [24,25], and quantum chemistry [17,26,27]. Prominent examples for GNSs include normalizing flows (NFs) [28–30] and autoregressive neural networks [31,32]. These models not only enable efficient sampling, but also facilitate the direct estimation of thermodynamic observables [19], entanglement entropies [33,34], and Feynman propagators [35]. Additionally, GNSs have been employed to guide advanced sampling techniques, such as Feynman-Hall resampling [36].

Recent studies have shown that GNSs excel at sampling from probability distributions with complex topologies, including bimodal structures [37,38] and gauge theories subject to topological freezing [39]. These findings establish GNSs as a promising alternative to traditional methods such as HMC, particularly in regimes where maintaining ergodicity is challenging. In this work, we leverage a GNS architecture, called symmetry-enforcing stochastic modulation (SESaMo) [40], to further enhance sampling performance.

The remainder of this Letter is structured as follows: We first review the Hubbard model, the framework of NFs, and the

*Contact author: schuh@hiskp.uni-bonn.de

†Contact author: jkreit@uni-bonn.de

SESaMo architecture; we then present our numerical results and conclude with a discussion of our findings.

The Hubbard model. Condensed matter systems can often be modeled as a fixed spatial lattice of atoms—ignoring vibrations and other degrees of freedom—on which electrons can hop between valence orbitals and interact via their electric charge. In this work, we employ the Hubbard model [2,2–4,41] on a bipartite lattice at half-filling to describe the dynamics and interactions of the electrons,

$$H = - \sum_{\langle x,y \rangle, \sigma} c_{x,\sigma}^\dagger c_{y,\sigma} - \frac{U}{2} \sum_x (n_{x,\uparrow} - n_{x,\downarrow})^2. \quad (1)$$

Here, $c_{x,\sigma}^\dagger$ ($c_{x,\sigma}$) denotes the electron creation (annihilation) operator at site x with spin σ ($=\uparrow\downarrow$) and $n_{x,\sigma} = c_{x,\sigma}^\dagger c_{x,\sigma}$ is the corresponding number operator. The first term on the right-hand side of Eq. (1) describes the *tight-binding* dynamics, while the second term captures the on-site repulsive interaction between electrons via the coupling U . Equation (1) is defined such that the electrically neutral half-filled state corresponds to the global ground state of the system. For stochastic simulations of this system, it is well known that this model can suffer from ergodicity problems [11,12,15,16].

This quantum mechanical formulation may be rather intuitive to understand, but it is highly unwieldy to handle in practice. After all, a direct calculation involves a Hamiltonian matrix whose size grows exponentially with the number of lattice sites N_x . It is therefore common to cast the theory into an auxiliary-field formulation [2–4,41–43]. This reformulation begins by expressing observables in terms of statistical quantum mechanical expectation values:

$$\langle \mathcal{O} \rangle = \frac{1}{\mathcal{Z}} \text{tr}\{e^{-\beta H} \mathcal{O}\}, \quad (2)$$

where the partition function is given by $\mathcal{Z} = \text{tr}\{e^{-\beta H}\}$, $\beta = \frac{1}{k_B T}$ is the inverse temperature, and the trace runs over the entire Fock space.

This expectation value can be expressed in terms of an auxiliary bosonic field ϕ , introduced by a Hubbard-Stratonovich transformation [44],

$$\langle \mathcal{O} \rangle = \frac{1}{\mathcal{Z}} \int \mathcal{D}[\phi] e^{-S[\phi]} \mathcal{O}[\phi] \approx \frac{1}{N_{\text{cfg}}} \sum_{\{\phi\}} \mathcal{O}[\phi], \quad (3)$$

where N_{cfg} is the number of configurations $\{\phi\}$ sampled from e^{-S} [45], and S is the Hubbard action given by

$$S[\phi] = \frac{1}{2\tilde{U}} \sum_{x,t \in \Lambda} \phi_{xt}^2 - \ln \det (M[\phi] \cdot M[-\phi]), \quad (4)$$

where $\tilde{U} = U\beta/N_t$, $\Lambda = N_x \times N_t$ is the space-time volume of the lattice, xt labels a site in this volume, $\phi_{xt} \in \mathbb{R}$ is the corresponding field component, and M is the fermion matrix in exponential discretization [46] [see Eq. (A4) and the Appendix].

The Hubbard action consists of two main contributions: a Gaussian term, encoding interactions, and a fermionic term, describing the fermion dynamics. Importantly, the fermion determinant is manifestly positive [15], which ensures that the Boltzmann weight can be interpreted as a proper probability distribution.

The goal of this work is to provide the first proof of principle that NFs can be applied to the Hubbard model. In doing so, we demonstrate how ergodicity and autocorrelation issues can be essentially eliminated. We focus on a spatial lattice extent fixed to two sites, i.e., $N_x = 2$, where these effects are particularly pronounced, while we vary the temporal extent N_t .

An additional benefit of studying the two-site system is that the exact configuration distributions are analytically known in the strong-coupling limit, enabling direct comparisons with our GNS flows in this regime. We also benchmark our results against analytically known correlation functions [15]

$$C_{x,y}(t) \equiv \langle c_x(t) c_y^\dagger(0) \rangle = \langle M_{(t,x);(0,y)}^{-1}[\phi] \rangle, \quad (5)$$

where $M_{(t,x);(0,y)}^{-1}$ is an element of the inverse fermion matrix [see Eq. (A4) and the Appendix]. The correlator is an *observable* that provides information about the interacting spectrum of the system. The correlation functions are computed for all site pairs (x, y) , and subsequently diagonalized to obtain eigenfunctions $C_\pm(t)$ [15].

Symmetries of the Hubbard action. The symmetries of this system have been well documented (see, e.g., Ref. [15]). Of particular interest here are the symmetries of the fermion matrix M and its determinant. Reference [15] showed that, as a consequence of sublattice (also referred to as *chiral*) symmetry and particle-hole symmetry, one obtains

$$\det(M[\phi]M[-\phi]) = f[\phi]f[-\phi],$$

$$f[\phi] = e^{-\Phi/2} \det M[\phi] = e^{\Phi/2} \det M[-\phi] = f[-\phi],$$

where $\Phi = \phi_1 + \phi_2$, with $\phi_i = \sum_t \phi_{i,t}$. The function f is therefore an even function of the field variables. For the two-site model, this has important consequences in the strong-coupling limit ($U \gg 1$), where the determinant becomes nearly invariant under $\phi_i \rightarrow -\phi_i$. As a result, the system exhibits an approximate $\mathbb{Z}_4$ symmetry in the variables ϕ_1 and ϕ_2 . As we will show later, explicitly incorporating this symmetry into our neural network architecture is crucial for correctly capturing the probability distribution and observables, and for preventing mode collapse.

Normalizing flows. Normalizing flows [29,30] are generative models that enable efficient sampling from high-dimensional distributions by learning an invertible map between a simple base distribution q_z (e.g., a Gaussian) and a complex target distribution q_y defined by the physical system. Here, z denotes latent variables, and y corresponds to the physical degrees of freedom—in our case, the auxiliary Hubbard field configurations ϕ . This map is realized through a bijective transformation

$$f_\theta : z \sim q_z \mapsto y \sim q_y, \quad (6)$$

where f_θ is a bijective function parametrized by neural network weights θ . The primary objective is to construct a bijective map f_θ that is sufficiently expressive to model the target distribution. In recent years, numerous types of NFs have been proposed, including coupling-based NFs [47–49], autoregressive NFs [50], and continuous NFs [51]. In this work, we focus on coupling-based NFs with a real-valued non-volume-preserving (RealNVP) architecture [48]. These consist of L invertible and differentiable transformations $f_{\theta_i}^l : \mathbb{R}^{|\Lambda|} \rightarrow \mathbb{R}^{|\Lambda|}$, where each transformation is parametrized by

a neural network with parameters θ_i . The bijectivity of each individual transformation ensures that the overall map is also bijective:

$$y \equiv f_\theta(z) = (f_{\theta_L}^L \circ f_{\theta_{L-1}}^{L-1} \circ \dots \circ f_{\theta_1}^1)(z), \quad (7)$$

where $\theta = \{\theta_1, \dots, \theta_L\}$ collectively parametrizes the map f_θ . In order to enable density estimation of q_y , it is essential to efficiently compute the Jacobian determinant of f_θ during training.

The incorporation of physics priors, such as symmetries, has been shown to significantly improve the performance and training stability of generative models [52–54]. In Ref. [40], the SESaMo framework was introduced, which enhances NFs by incorporating stochastic modulation. This approach enforces symmetries on the output density by randomly applying symmetry transformations during sampling. In the context of this work, we employ a stochastic modulation

$$S_{\mathbb{Z}_4} \equiv S_2 \circ S_1 : y \sim q_y \mapsto \phi \sim q_\theta, \quad (8)$$

which imposes an approximate $\mathbb{Z}_4$ symmetry to q_θ by sequentially applying two stochastic transformations: a $\mathbb{Z}_2$ -symmetric modulation S_1 and an *approximate* $\mathbb{Z}_2$ -symmetric modulation S_2 .

The first transformation, S_1 , flips the sign of the entire field y with a probability of 50 %:

$$S_1 : y \mapsto \begin{cases} y, & \text{if } u_1 = 0, \\ -y, & \text{if } u_1 = 1, \end{cases} \quad \text{where } u_1 \sim \mathcal{B}(0.5). \quad (9)$$

The second transformation, S_2 , selectively flips the components $y_1 = (y_{11}, y_{12}, \dots, y_{1N_t})^T$ with a probability controlled by a *learnable* breaking parameter $b \in \mathbb{R}^-$:

$$S_2 : y_1 \mapsto \begin{cases} y_1, & \text{if } u_2 = 0, \\ -y_1, & \text{if } u_2 = 1, \end{cases} \quad \text{where } u_2 \sim \mathcal{B}(e^b), \quad (10)$$

with the Bernoulli distribution $\mathcal{B}(\cdot)$. Thus, S_1 enforces exact $\mathbb{Z}_2$ symmetry, while S_2 introduces a controlled symmetry *breaking* in the components y_{1t} , with the degree of breaking learned during training via the parameter b .

The complete transformation used to generate the final field configuration ϕ is thus given by

$$\phi \equiv g_\theta(z) = (S_2 \circ S_1 \circ f_\theta)(z), \quad (11)$$

where $\theta = \{\theta, b\}$ denotes the complete set of trainable parameters of the model. Here, the parameters θ are initialized randomly and b is initialized as $b = \ln(0.5)$ [55]. The resulting output probability density is expressed by

$$q_\theta(g_\theta(z)) = q_z(z) \cdot \left| \frac{\partial f_\theta}{\partial z} \right|^{-1} p_{S_1} \cdot p_{S_2}(u_2), \quad (12)$$

where p_{S_1} and p_{S_2} are the probabilities associated with the stochastic modulation steps S_1 and S_2 , respectively, which are given by

$$p_{S_1} = \frac{1}{2} \quad \text{and} \quad p_{S_2}(u_2) = \begin{cases} 1 - e^b, & \text{if } u_2 = 0, \\ e^b, & \text{if } u_2 = 1. \end{cases} \quad (13)$$

The objective of the training process is to minimize the *self-regularized* Kullback-Leibler (KL) divergence, an extension of the reverse KL divergence [56], which is used to train the NF with SESaMo. This divergence measures the distance

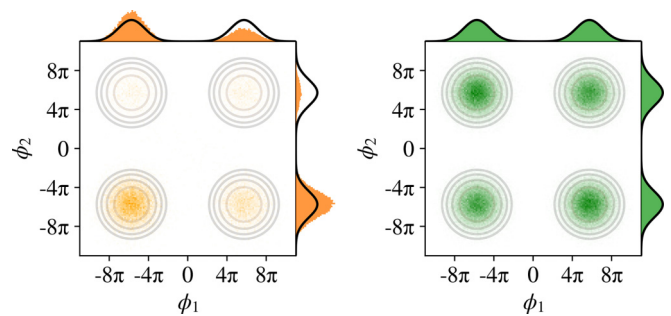


FIG. 1. Two-site Hubbard model ($U = 18$, $\beta = 1$): Sampled field configurations from 25k decorrelated samples generated by HMC (orange) and NF (green), summed over the time direction, i.e., $\phi_x = \sum_{t=1}^{N_t} \phi_{xt}$. Gray contours indicate the reference distribution in the strong-coupling limit. The outer histograms show the marginal distributions in ϕ_1 and ϕ_2 , with the black lines indicating the exact analytical results obtained from Eq. (70) in Ref. [15].

between the model output density q_θ and the target Boltzmann density $p = e^{-S}/Z$, and is given by

$$\widetilde{\text{KL}}(q_\theta || p) = \mathbb{E}_{z \sim q_z} [\ln q_\theta[g_\theta(z)] + S[g_\theta(z)] + \gamma \ln \hat{Z} + \Lambda(f_\theta(z))], \quad (14)$$

where $S[\cdot]$ is the physical action and $\hat{Z}$ is an estimator of the partition function scaled by a hyperparameter γ . The penalty term Λ acts as a regularization term that numerically enforces bijectivity of g_θ . Further details are given in the Appendix. Crucially, to ensure exactness, we apply an accept/reject step to the generated configurations [18].

Results. In the following, we evaluate the performance of our NF-based approach for the Hubbard model on a two-site lattice with parameters $U = 18$ and $\beta = 1$. We employ a Real-NVP network with an approximate $\mathbb{Z}_4$ symmetry enforced via the SESaMo approach. Specifically, we compare the resulting probability densities, physical observables, scaling behavior, and sampling efficiency.

In Fig. 1, we show the sampled distribution obtained from HMC with decorrelation measures (orange, left), i.e., taking only every tenth sample from the Markov chain [15], and from the NF (green, right). HMC struggles to distribute the relative weight correctly among all modes, as it predominantly collapses onto the lower-left mode and consequently undersamples the remaining ones, which can result in biased estimates of physical observables. In contrast, the NF successfully learns the correct multimodal distribution and samples with a high acceptance rate of $a = 98.3\%$. To further validate the superior performance of the NF, we compute the even one-body correlation function $C_+(t)$ as defined in Eq. (5), with results shown in Fig. 2. The figure compares outcomes from HMC with decorrelation measures (orange), NF (green), and exact diagonalization (ED, gray), where ED serves as the ground truth in this scenario. We observe that the strong ergodicity issues in the HMC samples lead to systematically biased results that deviate from the ground truth. In contrast, the NF approach accurately reproduces the observables with excellent agreement and significantly smaller uncertainties.

Figure 3 compares the sampling efficiency of the presented methods using the effective sample size (ESS) and

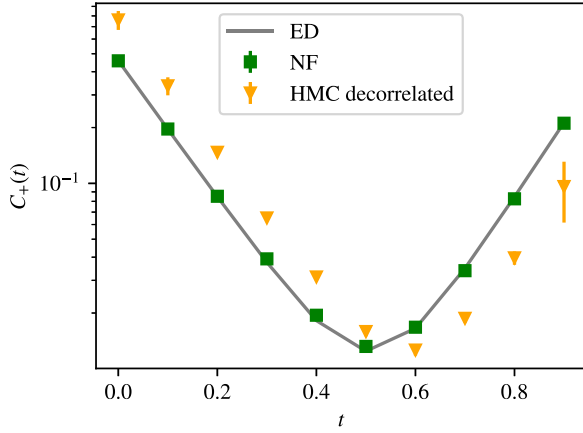


FIG. 2. One-body correlation function $C_+(t)$ of the two-site Hubbard model ($U = 18$, $\beta = 1$). Results from 25k decorrelated configurations generated by HMC (orange) and NF (green) are shown, alongside the ground truth obtained from exact diagonalization (gray).

the integrated autocorrelation time τ_{int} , evaluated on the correlators shown in Fig. 2. The NF (green) achieves both a higher ESS and a lower τ_{int} compared to HMC without decorrelation measures (blue), demonstrating superior sampling performance. Even when decorrelation measures are applied to HMC (orange), the NF remains significantly more efficient. These improvements underscore the advantages of (1) generating independent, parallel samples and (2) incorporating symmetry-aware training through SESaMo.

One of the main limitations of flow-based samplers is their unfavorable scaling with the lattice volume [57–60], as well as the relatively high computational cost of training. To directly assess the scaling performance of HMC and the NF, we compare the total wall-clock time—including training and the accept/reject step for the NF and decorrelation measures for HMC—required to generate 25k decorrelated configurations across a range of temporal extents N_t , as shown in Fig. 4.

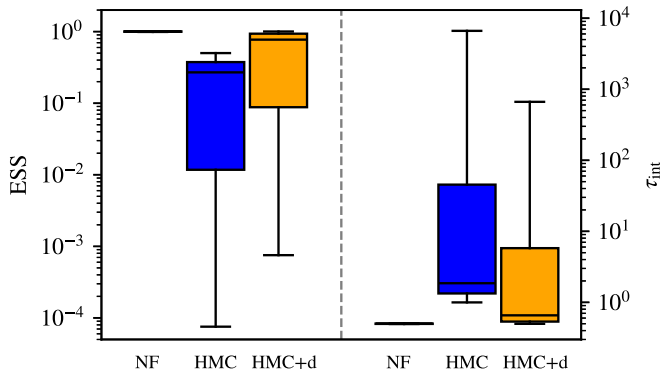


FIG. 3. Comparison of the effective sample size (ESS, left axis) and the integrated autocorrelation time τ_{int} (right axis) for three sampling methods: NF (green), HMC (blue), and HMC with decorrelation measures (HMC + d, orange). Both metrics are computed from the one-body correlation function shown in Fig. 2. A higher ESS and lower τ_{int} indicate more efficient sampling, underscoring the superior performance of the NF approach.

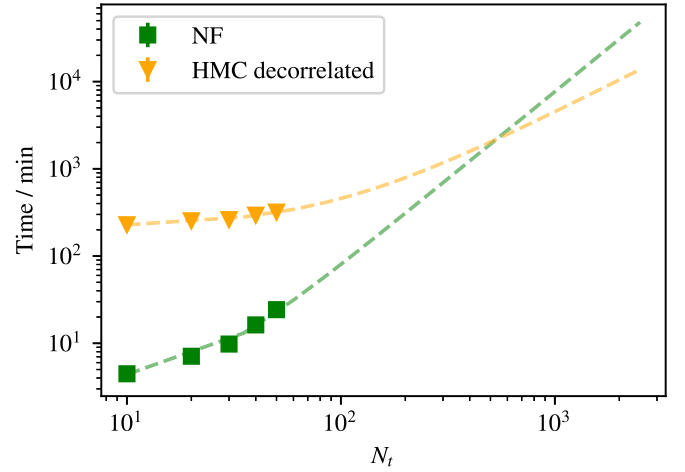


FIG. 4. Sampling time—including the training time required for the NF to reach an acceptance rate of at least 97%, as well as the accept/reject step, and including the decorrelation measures for HMC—as a function of the temporal lattice extent N_t , for generating 25k decorrelated configurations. HMC is considered thermalized after 10k steps. Results are shown for HMC (orange) and an NF (green). Error bars indicate the statistical uncertainty in the measured time across three independent runs. The dashed lines show fits to the observed scaling behavior: $N_t^{5/4}$ for HMC, consistent with theoretical expectations, and an empirical quadratic dependence for the NF.

The results show that incorporating symmetry constraints via SESaMo reduces the total sampling time for the NF by up to 98% compared to HMC.

Finally, in Fig. 5, we present a comparison of the total sampling time—again including training and the accept/reject step for the NF and decorrelation measures for HMC—for a system with $N_t = 10$ time slices, using HMC (blue), HMC with decorrelation measures (orange), and the NF (green). We

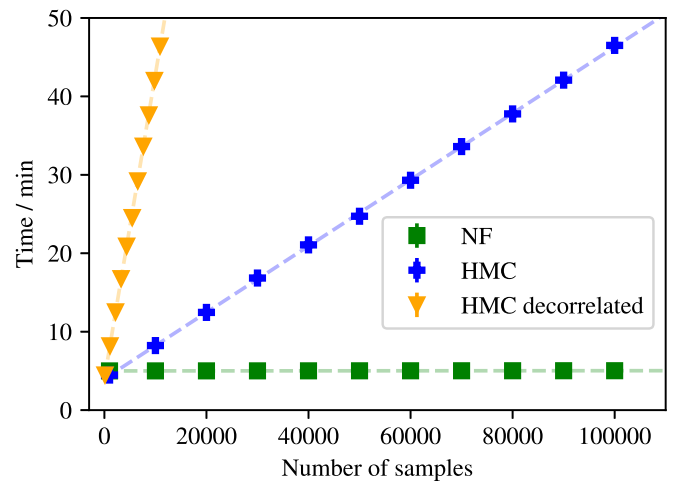


FIG. 5. Comparison of sampling costs for HMC (blue), HMC with decorrelation measures (orange), and the NF (green), including 5 min of training time and the accept/reject step for the NF. For HMC, thermalization is assumed after 10k steps. All results correspond to the Hubbard model with $U = 18$, $\beta = 1$, and $N_t = 10$. The trained flow model achieved an acceptance rate of 97.43%. The dashed lines indicate a linear fit to the number of samples.

find that, due to the embarrassingly parallel sampling with the NF and the low training overhead enabled by the efficient symmetry incorporation via SESaMo, the total sampling time for the NF increases only marginally beyond an initial training period of approximately 5 min. In contrast, the sampling time for HMC grows linearly with the number of configurations and, when applying decorrelation measures, becomes roughly two orders of magnitude larger than that of the NF.

Conclusion. In this work, we presented the first proof of principle that NFs can accurately learn the Boltzmann distribution of the *fermionic* Hubbard model and yield high-precision results for physical observables such as correlators, especially in regimes where HMC fails due to ergodicity issues. Our results establish that NFs can surpass HMC by a wide margin, both in accuracy and in computational efficiency. Furthermore, we demonstrated that incorporating symmetry constraints via SESaMo enables favorable (quadratic) scaling with the temporal lattice extent. Additional results for larger spatial lattice sizes are provided in the Appendix.

Finally, we note that extending the algorithm to larger spatial lattices may, in principle, lead to an exponential growth in the number of breaking parameters of the stochastic modulation. However, we expect the presence of additional, yet unexploited, symmetries in the system that substantially reduce the number of *independent* breaking parameters. A detailed investigation of this aspect is left for future work.

All in all, these findings establish a broadly applicable framework for efficient sampling in complex lattice systems, where traditional methods are hindered by ergodicity issues and long autocorrelations.

Acknowledgments. The authors thank Simran Singh, Luca Johannes Wagner, Gurtej Kanwar, and Daniel Hackett for inspiring discussions. The authors gratefully acknowledge the granted access to the Marvin cluster hosted by the University of Bonn. This project was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the CRC 1639 NuMerIQS—Project No. 511713970.

Data availability. The data that support the findings of this article are openly available [61].

Appendix:

Basis choice for the Hubbard model. In the main part of this Letter, we introduced the Hubbard model in the so-called *spin basis*. This choice, however, is not unique. More generally, the Hubbard Hamiltonian can be written in a continuous family of bases as

$$H = H_0 + \alpha \sum_x \frac{U}{2} (n_{x,\uparrow} - n_{x,\downarrow})^2 - (1 - \alpha) \frac{U}{2} \times \sum_x (n_{x,\uparrow} - n_{x,\downarrow})^2, \quad (\text{A1})$$

where the parameter $\alpha \in [0, 1]$ interpolates between the spin basis ($\alpha = 0$) and the particle/hole basis ($\alpha = 1$). The two limiting cases correspond to purely real ($\alpha = 0$) and purely imaginary ($\alpha = 1$) Hubbard-Stratonovich fields. Intermediate values of α , in contrast, typically lead to complex-valued fields. For this reason, the two limiting bases are especially

advantageous for numerical simulations, as they reduce the dimensionality of the field space.

The Hamiltonian in the particle/hole basis thus reads

$$H_{ph} = - \sum_{\langle x,y \rangle} (a_x^\dagger a_y - b_x^\dagger b_y) + \frac{U}{2} \sum_x (n_x - \tilde{n}_x), \quad (\text{A2})$$

where $n_x = a_x^\dagger a_x$ and $\tilde{n}_x = b_x^\dagger b_x$ count the number of particles and holes at position x , respectively. Following the same steps as in the spin basis, one obtains the Hubbard action in the particle/hole basis

$$S_{ph}[\phi] = \frac{1}{2\tilde{U}} \sum_{x,t \in \Lambda} \phi_{xt}^2 - \ln \det (M[i\phi] \cdot M[-i\phi]). \quad (\text{A3})$$

In contrast to the spin basis, the fermion matrix is now evaluated solely on imaginary fields $i\phi$.

Discretizations of the fermion matrix. In the discretized path integral formulation of the Hubbard model, the inverse temperature β is evenly divided into N_t time slices. The resulting fermion matrix $M[\phi]$, which emerges after integrating out the fermionic degrees of freedom, depends on the chosen discretization scheme. While all schemes converge to the same continuum limit $\delta = \beta/N_t \rightarrow 0$ [15], they differ at finite δ , resulting in distinct matrix structures and ergodicity properties.

In the main analysis of this Letter, we focused on the so-called *exponential* discretization,

$$M[\phi]_{x't',xt} = \delta_{x',x} \delta_{t',t} - [e^h]_{x',x} e^{\phi_{xt}} \mathcal{B}_{t'} \delta_{t',t+1}, \quad (\text{A4})$$

where the hopping matrix is defined as $h_{x',x} = \delta_{(x',x)}$, with $\delta_{(x',x)}$ restricting to allowed hoppings. The factor $\mathcal{B}$ explicitly encodes antiperiodic boundary conditions in the temporal direction t . To facilitate comparison with previous studies, we also consider the *diagonal* discretization, in which the exponential of the hopping matrix is expanded to first order. The resulting fermion matrix takes the form

$$M^d[\phi]_{x't',xt} = (\delta_{x',x} - h_{x',x}) \delta_{t',t} - e^{\phi_{xt}} \delta_{x',x} \mathcal{B}_{t'} \delta_{t',t+1}. \quad (\text{A5})$$

Although both discretizations formally agree in the continuum limit [15], they differ in their ergodic properties: The diagonal discretization does not suffer from a formal ergodicity problem, while the exponential discretization in the particle/hole basis encounters codimension-1 manifolds in configuration space (i.e., subspaces of dimension $N_x N_t - 1$) on which the probability weight vanishes. As a result, field configurations separated by such manifolds cannot be connected through finite-action paths, and standard Monte Carlo updates may become trapped within disconnected regions of configuration space. These manifolds are depicted as solid black lines in Fig. 6.

Symmetries of the Hubbard action in the particle/hole basis. The Hubbard model exhibits a rich symmetry structure that depends on the chosen basis. These symmetries can apply to the full action or, more restrictively, only to the fermion determinant. For a comprehensive classification and discussion of the symmetry structure across different formulations, we refer the reader to Ref. [15]. Here, we focus on the periodicity symmetry relevant to our analysis in the particle/hole basis.

In the particle/hole basis (i.e., $\alpha = 1$), the probability weight defined in Eq. (A3) depends only on $e^{i\phi}$. As a result, it

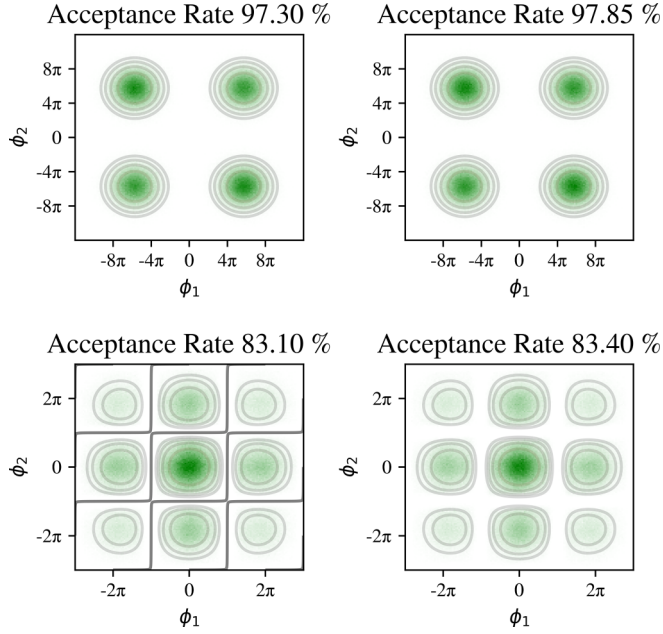


FIG. 6. Two-site distributions of the Hubbard model ($U = 18$, $\beta = 1$, $N_t = 10$). Results in the spin (top) and particle/hole (bottom) bases are shown for exponential (left) and diagonal (right) discretizations; gray contours indicate the strong-coupling reference, and all distributions are generated by an NF using $\mathbb{Z}_4$ stochastic modulation (top) or 2π canonicalization (bottom).

is invariant under discrete shifts of the auxiliary field

$$\phi_{xt} \rightarrow \phi_{xt} + 2\pi \cdot n_{xt}, \quad n_{xt} \in \mathbb{Z}. \quad (\text{A6})$$

This transformation leaves the fermion determinant unchanged but does not preserve the Gaussian part of the action. It is therefore only an *approximate* symmetry of the system, valid strictly at the level of the fermionic sector.

Details on the self-regularized KL divergence. The regularization term in the self-regularized KL divergence, introduced in Eq. (14), depends on the field configuration y after the NF f_θ and before the stochastic modulation S . It is defined as

$$\Lambda(y) = \sum_{x=1}^{N_x} A \sigma(B \lambda_x(y)) \cdot \theta(\lambda_x(y)), \quad (\text{A7})$$

where σ and θ denote the sigmoid and Heaviside functions, respectively. The hyperparameters A and B control the overall strength of the regularization and the steepness of its gradient. The penalty function $\lambda_x(y)$ is given by

$$\lambda_x(y) = - \sum_{t=1}^{N_t} y_{xt}. \quad (\text{A8})$$

Finally, the hyperparameter used to scale the the partition function estimate $\hat{Z}$ is fixed to $\gamma = 0.5$.

Additional results for the particle/hole basis and diagonal discretization. To enable direct comparison with the HMC results reported in Ref. [15], we train an NF using the

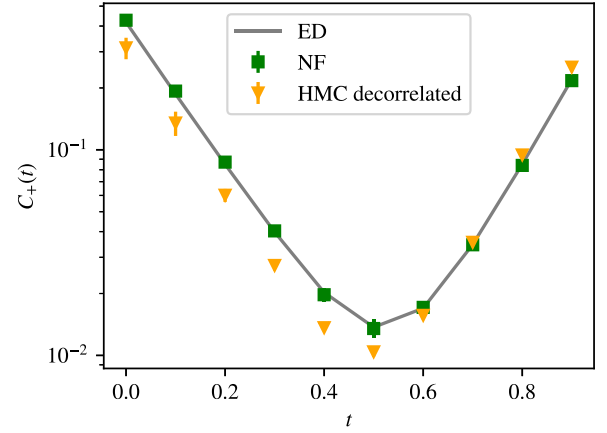


FIG. 7. One-body correlation function $C_+(t)$ of the four-site Hubbard model on a square lattice ($U = 18$, $\beta = 1$). Results obtained using HMC equipped with decorrelation measures (orange) and an NF (green) are shown, along with the ground truth from exact diagonalization (black). The NF accurately reproduces the ground truth with significantly smaller uncertainties, whereas HMC fails to capture the correct behavior.

canonicalization procedure [54] to account for the 2π -periodicity symmetry (A6), enabling it to learn the path integral distribution in the particle/hole basis. Additionally, we extend our analysis to include the diagonal discretization alongside the exponential discretization. As demonstrated in Fig. 6, the NF successfully captures the correct structure of the target distribution across all bases and discretizations.

At comparable acceptance rates, the NF significantly outperforms HMC by generating independent and identically distributed samples with full mode coverage [15,16]. This advantage is particularly pronounced in the spin basis, where the SESaMo framework enables a more expressive, symmetry-aware representation, further improving accuracy and acceptance rates.

Additional results for the four-site square lattice. In the main part of this Letter, we focused on the two-site model, where ergodicity issues are particularly pronounced and can be clearly visualized. To demonstrate the scalability and broader applicability of our method, we now present results for a two-dimensional four-site square lattice. At this system size, exact reference data from exact diagonalization are still available, allowing a direct quantitative comparison.

We consider a system with interaction strength $U = 18$, inverse temperature $\beta = 1$, and $N_t = 10$ time slices. As an observable, we examine the one-body correlation function, shown in Fig. 7. Similarly to the two-site case, HMC (orange) exhibits clear deviations from the ground truth due to ergodicity issues. In contrast, the NF (green) accurately reproduces the exact result across all time separations.

This result highlights the effectiveness of the symmetry-aware NF beyond the two-site system and demonstrates its applicability to larger lattices. The accurate reproduction of exact observables at this scale suggests that the approach can be reliably extended to even larger system sizes.

- [1] D. P. Arovas, E. Berg, S. A. Kivelson, and S. Raghu, The Hubbard model, *Annu. Rev. Condens. Matter Phys.* **13**, 239 (2022).
- [2] J. Hubbard, Electron correlations in narrow energy bands, *Proc. R. Soc. London A* **276**, 238 (1963).
- [3] J. Hubbard, Electron correlations in narrow energy bands. II. The degenerate band case, *Proc. R. Soc. London A* **277**, 237 (1964).
- [4] J. Hubbard, Electron correlations in narrow energy bands. III. An improved solution, *Proc. R. Soc. London A* **281**, 401 (1964).
- [5] L. Razmadze and T. Luu, Hubbard interaction at finite T on a hexagonal lattice, *PoS LATTICE2024*, 071 (2024).
- [6] E. Ibarra-García-Padilla, H. Lange, R. G. Melko, R. T. Scalettar, J. Carrasquilla, A. Bohrdt, and E. Khatami, Autoregressive neural quantum states of Fermi Hubbard models, *Phys. Rev. Res.* **7**, 013122 (2025).
- [7] J. Ostmeyer, E. Berkowitz, S. Krieg, T. A. Lähde, T. Luu, and C. Urbach, Semimetal–Mott insulator quantum phase transition of the Hubbard model on the honeycomb lattice, *Phys. Rev. B* **102**, 245105 (2020).
- [8] J. Ostmeyer, L. Razmadze, E. Berkowitz, T. Luu, and U.-G. Meißner, Effective theory for graphene nanoribbons with junctions, *Phys. Rev. B* **109**, 195135 (2024).
- [9] P. Sinilkov, E. Berkowitz, T. Luu, and M. Rodekamp, Search for stable states in two-body excitations of the Hubbard model on the honeycomb lattice, *PoS LATTICE2024*, 075 (2025).
- [10] M. Rodekamp, E. Berkowitz, C. Gäntgen, S. Krieg, T. Luu, J. Ostmeyer, and G. Pederiva, Single-particle spectrum of doped $C_{20}H_{12}$ -perylene, *Eur. Phys. J. B* **98**, 36 (2025).
- [11] P. Buividovich, D. Smith, M. Ulybyshev, and L. von Smekal, Hybrid-Monte-Carlo study of competing order in the extended fermionic Hubbard model on the hexagonal lattice, *Phys. Rev. B* **98**, 235129 (2018).
- [12] S. Beyl, F. Goth, and F. F. Assaad, Revisiting the hybrid quantum Monte Carlo method for Hubbard and electron-phonon models, *Phys. Rev. B* **97**, 085144 (2018).
- [13] D. Smith and L. von Smekal, Monte Carlo simulation of the tight-binding model of graphene with partially screened Coulomb interactions, *Phys. Rev. B* **89**, 195429 (2014).
- [14] M. V. Ulybyshev, P. V. Buividovich, M. I. Katsnelson, and M. I. Polikarpov, Monte Carlo study of the semimetal-insulator phase transition in monolayer graphene with a realistic inter-electron interaction potential, *Phys. Rev. Lett.* **111**, 056801 (2013).
- [15] J.-L. Wynen, E. Berkowitz, C. Körber, T. A. Lähde, and T. Luu, Avoiding ergodicity problems in lattice discretizations of the Hubbard model, *Phys. Rev. B* **100**, 075141 (2019).
- [16] F. Temmen, E. Berkowitz, A. Kennedy, T. Luu, J. Ostmeyer, and X. Yu, Overcoming ergodicity problems of the hybrid Monte Carlo method using radial updates, *PoS LATTICE2024*, 068 (2024).
- [17] F. Noé, S. Olsson, J. Köhler, and H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, *Science* **365**, eaaw1147 (2019).
- [18] M. S. Albergio, G. Kanwar, and P. E. Shanahan, Flow-based generative models for Markov chain Monte Carlo in lattice field theory, *Phys. Rev. D* **100**, 034515 (2019).
- [19] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati, Estimation of thermodynamic observables in lattice field theories with deep generative models, *Phys. Rev. Lett.* **126**, 032001 (2021).
- [20] M. Caselle, E. Cellini, A. Nada, and M. Panero, Stochastic normalizing flows as non-equilibrium transformations, *J. High Energy Phys.* **07** (2022) 015.
- [21] K. Cranmer, G. Kanwar, S. Racanière, D. J. Rezende, and P. E. Shanahan, Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics, *Nat. Rev. Phys.* **5**, 526 (2023).
- [22] D. Wu, L. Wang, and P. Zhang, Solving statistical mechanics using variational autoregressive networks, *Phys. Rev. Lett.* **122**, 080602 (2019).
- [23] K. A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller, and P. Kessel, Asymptotically unbiased estimation of physical observables with neural samplers, *Phys. Rev. E* **101**, 023304 (2020).
- [24] M. Caselle, E. Cellini, and A. Nada, Sampling Nambu-Goto theory using normalizing flows, *PoS LATTICE2023*, 015 (2024).
- [25] M. Caselle, E. Cellini, and A. Nada, Sampling the lattice Nambu-Goto string using continuous normalizing flows, *J. High Energy Phys.* **02** (2024) 048.
- [26] N. W. A. Gebauer, M. Gastegger, and K. T. Schütt, Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates Inc., Red Hook, NY, 2019), Vol. 32, pp. 7566–7578.
- [27] N. W. A. Gebauer, M. Gastegger, S. S. P. Hessmann, K.-R. Müller, and K. T. Schütt, Inverse design of 3d molecular structures with conditional generative neural networks, *Nat. Commun.* **13**, 973 (2022).
- [28] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning* (PMLR, 2015), Vol. 37, p. 1530.
- [29] I. Kobyzev, S. J. Prince, and M. A. Brubaker, Normalizing flows: An introduction and review of current methods, *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3964 (2021).
- [30] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* **22**, 1 (2021).
- [31] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, Pixel recurrent neural networks, in *Proceedings of the 33rd International Conference on Machine Learning* (PMLR, 2016), Vol. 48, p. 1747.
- [32] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, Conditional image generation with PixelCNN decoders, in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2016), Vol. 29, p. 4797.
- [33] P. Białas, P. Korcyl, T. Stebel, and D. Zapolski, Rényi entanglement entropy of a spin chain with generative neural networks, *Phys. Rev. E* **110**, 044116 (2024).
- [34] A. Bulgarelli, E. Cellini, K. Jansen, S. Kühn, A. Nada, S. Nakajima, K. A. Nicoli, and M. Panero, Flow-based sampling for entanglement entropy and the machine learning of defects, *Phys. Rev. Lett.* **134**, 151601 (2025).

- [35] S. Chen, O. Savchuk, S. Zheng, B. Chen, H. Stoecker, L. Wang, and K. Zhou, Fourier-flow model generating Feynman paths, *Phys. Rev. D* **107**, 056001 (2023).
- [36] R. Abbott, D. Boyda, G. Kanwar, F. Romero-Lopez, D. Hackett, P. Shanahan, and J. Urban, Progress in normalizing flows for 4d gauge theories, *PoS LATTICE2024*, 066 (2025).
- [37] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati, Machine learning of thermodynamic observables in the presence of mode collapse, *PoS LATTICE2021*, 338 (2022).
- [38] K. A. Nicoli, C. J. Anders, T. Hartung, K. Jansen, P. Kessel, and S. Nakajima, Detecting and mitigating mode-collapse for flow-based sampling of lattice field theories, *Phys. Rev. D* **108**, 114501 (2023).
- [39] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, Equivariant flow-based sampling for lattice gauge theory, *Phys. Rev. Lett.* **125**, 121601 (2020).
- [40] J. Kreit, D. Schuh, K. A. Nicoli, and L. Funcke, SESaMo: Symmetry-enforcing stochastic modulation for normalizing flows, *arXiv:2505.19619*.
- [41] R. C. Brower, D. Schaich, and C. Rebbi, Hybrid Monte Carlo simulation on the graphene hexagonal lattice, *PoS Lattice* **2011**, 056 (2012).
- [42] T. Luu and T. A. Lähde, Quantum Monte Carlo calculations for carbon nanotubes, *Phys. Rev. B* **93**, 155106 (2016).
- [43] M. Rodekamp, Machine learning for path deformation and Bayesian data analysis in selected lattice field theories, Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 2024.
- [44] J. Hubbard, Calculation of partition functions, *Phys. Rev. Lett.* **3**, 77 (1959).
- [45] The approximation in Eq. (3) becomes exact in the limit $N_{\text{cfg}} \rightarrow \infty$, with the error scaling as $N_{\text{cfg}}^{-1/2}$.
- [46] Other choices for incorporating the hopping term into the fermion matrix include the *diagonal* discretization [41,42] and the *linear* discretization [13]. For more details, we refer to the Appendix.
- [47] L. Dinh, D. Krueger, and Y. Bengio, NICE: Non-linear independent components estimation, in *3rd International Conference on Learning Representations, ICLR 2015, Workshop Track Proceedings* (2015).
- [48] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using real NVP, in *International Conference on Learning Representations* (OpenReview.net, 2017).
- [49] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2019), Vol. 32, pp. 7509–7520.
- [50] D. P. Kingma and P. Dhariwal, Glow: Generative flow with invertible 1×1 convolutions, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY 2018), Vol. 31, pp. 10215–10224.
- [51] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, Neural ordinary differential equations, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2018), Vol. 31, pp. 6571–6583.
- [52] T. Cohen and M. Welling, Group equivariant convolutional networks, in *Proceedings of the 33rd International Conference on Machine Learning* (OpenReview.net, 2016), Vol. 48, p. 2990.
- [53] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, Spherical CNNs, in *International Conference on Learning Representations* (2018).
- [54] D. Boyda, G. Kanwar, S. Racanière, D. J. Rezende, M. S. Albergo, K. Cranmer, D. C. Hackett, and P. E. Shanahan, Sampling using $SU(N)$ gauge equivariant flows, *Phys. Rev. D* **103**, 074504 (2021).
- [55] The convergence of the breaking parameter during training is not affected by its initialization. We therefore initialize an exact $\mathbb{Z}_4$ symmetry with $b = \ln(0.5)$.
- [56] S. Kullback and R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.* **22**, 79 (1951).
- [57] R. Abbott, M. S. Albergo, A. Botev, D. Boyda, K. Cranmer, D. C. Hackett, A. G. Matthews, S. Racanière, A. Razavi, D. J. Rezende *et al.*, Aspects of scaling and scalability for flow-based sampling of lattice QCD, *Eur. Phys. J. A* **59**, 257 (2023).
- [58] L. Del Debbio, J. Marsh Rossney, and M. Wilson, Efficient modeling of trivializing maps for lattice ϕ^4 theory using normalizing flows: A first look at scalability, *Phys. Rev. D* **104**, 094507 (2021).
- [59] R. Abbott, M. S. Albergo, A. Botev, D. Boyda, K. Cranmer, D. C. Hackett, G. Kanwar, A. G. D. G. Matthews, S. Racanière, A. Razavi, D. J. Rezende, F. Romero-López, P. E. Shanahan, and J. M. Urban, Normalizing flows for lattice gauge theory in arbitrary space-time dimension, *arXiv:2305.02402*.
- [60] J. Komijani and M. K. Marinkovic, Generative models for scalar field theories: How to deal with poor scaling? *PoS LATTICE2022*, 019 (2023).
- [61] D. Schuh, J. Kreit, E. Berkowitz, L. Funcke, T. Luu, K. Nicoli, and M. Rodekamp, Dataset for “Simulating correlated electrons with symmetry-enforced normalizing flows”, Zenodo (2025), doi: <https://doi.org/10.5281/zenodo.15878888>.