


## Global sampling of Feynman's diagrams through normalizing flow

Luca Leoni <sup>1,\*</sup> and Cesare Franchini <sup>1,2</sup>

<sup>1</sup>*Department of Physics and Astronomy "Augusto Righi", Alma Mater Studiorum–Università di Bologna, Bologna 40127, Italy*

<sup>2</sup>*University of Vienna, Faculty of Physics and Center for Computational Materials Science, Kolingasse 14-16, 1090 Vienna, Austria*



(Received 1 February 2024; accepted 13 June 2024; published 8 July 2024)

Normalizing flows (NF) are powerful generative models with increasing applications in augmenting Monte Carlo algorithms due to their high flexibility and expressiveness. In this work we explore the integration of NF in the diagrammatic Monte Carlo (DMC) method, presenting an architecture designed to sample the intricate multidimensional space of Feynman's diagrams through dimensionality reduction. By decoupling the sampling of diagram order and interaction times, the flow focuses on one interaction at a time. This enables one to construct a general diagram by employing the same unsupervised model iteratively, dressing a zero-order diagram with interactions determined by the previously sampled order. The resulting NF-augmented DMC method is tested on the widely used single-site Holstein polaron model in the entire electron-phonon coupling regime. The obtained data show that the model accurately reproduces the diagram distribution by reducing sample correlation and observables' statistical error, constituting the first example of global sampling strategy for connected Feynman's diagrams in the DMC method.

DOI: [10.1103/PhysRevResearch.6.033041](https://doi.org/10.1103/PhysRevResearch.6.033041)

### I. INTRODUCTION

The investigation of many-body effects in interacting systems has long presented a significant challenge in the physics community. A modern approach to address this challenge centers on the strategic use of Feynman's diagrams for perturbative assessments of key quantities like self-energies and correlation functions [1,2]. These evaluations provide insights into the intricate ways in which single-particle properties are impacted by interactions. Consequently, this approach establishes a comprehensive framework with versatile applications extending to thermodynamics and quasiparticle properties, thereby attracting interest from a variety of fields [3].

The diagrammatic quantum Monte Carlo (DMC) method is a powerful numerical method to obtain approximation-free estimates of diagrammatic perturbative expansions [4,5]. A target quantity, usually the interacting Matsubara Green's function  $G(\mathbf{k}, \tau)$ , is sampled stochastically through the Markov chain Monte Carlo (MCMC) exploration of the Feynman's diagrams contributing to its power series. This approach has been widely used in condensed-matter physics to accurately solve a wide range of physical systems described in terms of effective Hamiltonians. Representative examples are given by polarons [4–10] and excitons [11,12], thermodynamics of spin systems [13,14], and correlation of

electron gas [15]. While considered to be numerically exact, the DMC method becomes progressively less efficient with the increasing complexity of the diagram's phase space, which rapidly scales when approaching real materials. For example, the incorporation of *ab initio* material-specific band dispersions and electron-phonon coupling introduces distinctive features in the phase space, leading to a sluggish MCMC exploration with increased autocorrelation between samples, inducing critical slowing down effects [16]. Other Monte Carlo algorithms tackle this issue by employing global sampling strategies, which enhance the exploration of phase space within the MCMC procedure [17]. These approaches, however, are not yet present in the DMC literature [18]. Nevertheless, recent advancements in generative machine learning (GML) have demonstrated the potential for such architectures to serve as effective global updates in various physics-related scenarios [19–25].

From a broader perspective, the integration of machine learning techniques in many-body physics has gathered substantial interest in recent years [26]. Noteworthy successes in performance enhancement have been achieved across various contexts, encompassing density functional theory [27–29], numerical renormalization groups [30,31], and interacting spin models [32]. Particularly relevant to the current context is the integration of GML models into Monte Carlo schemes to refine sampling from high-dimensional, intricate probability distributions [33,34]. Various GML algorithms have proven effective as global updates [17,35,36], enhancing MCMC procedures [23–25]. This enables the generation of uncorrelated samples within the chain and mitigates the critical slowing down problem.

Normalizing flows (NFs) represent an emerging family of generative models, standing out as one of the most promising techniques in GML. NFs provide a method to

\*Contact author: [luca.leoni12@unibo.it](mailto:luca.leoni12@unibo.it)

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

approximate invertible maps to desired probability distributions [37,38]. The NF architecture facilitates both rapid sampling and inference of target distributions, exhibiting significant versatility in representing complex objects. It has found successful applications across a wide spectrum, including image and video generation, reinforcement learning, and in the Monte Carlo integration of quantum lattice field models [19–22]. The prospect of designing a flow-based model to capture the features of the Feynman diagram’s phase space within a DMC framework is highly appealing. However, the added complexity in this case lies in the fact that the sample’s dimensionality is a variable of the distribution, rendering standard NF architectures unsuitable for addressing the problem. In this paper, we introduce an approach to circumvent such limitations by proposing a model that generates a diagram iteratively, focusing on individual interactions. To evaluate the feasibility and efficiency of this hypothesis, we employ the single-site polaron Holstein model as a test case, a paradigmatic electron-phonon (el-ph) effective Hamiltonian [39].

Our designed NF-augmented DMC approach yields a significant reduction in sample correlation within the Markov chain across a broad spectrum of electron-phonon coupling strengths. This reduction translates into a diminished number of samples required to achieve convergence. The proposed data-driven global updates enhance statistical convergence, leading to a roughly 50% reduction in both statistical errors associated with observed polaron properties and sample autocorrelation compared to the standard DMC local procedure. Since the code developed in this study was designed to enhance the exploration of Feynman’s diagram space rather than to optimize performance, the actual sampling cost is higher compared to the standard approach.

The article is organized as follows. In the next section, we introduce the NF-augmented DMC architecture tailored for the Holstein polaron model in the atomic limit. Subsequently, we present and analyze the performance of the developed code in estimating specific polaron quantities.

## II. NORMALIZING FLOW DIAGRAMMATIC MONTE CARLO METHOD

### A. Diagram representation

The single-site Holstein model is expressed as

$$H = -\varepsilon \hat{c}^\dagger \hat{c} + \Omega \sum_{\mathbf{q}} \hat{b}_{\mathbf{q}}^\dagger \hat{b}_{\mathbf{q}} + \frac{g}{\sqrt{N}} \sum_{\mathbf{q}} \hat{c}^\dagger \hat{c} (\hat{b}_{-\mathbf{q}}^\dagger + \hat{b}_{\mathbf{q}}), \quad (1)$$

consisting of a single nondispersive electronic band with energy  $\varepsilon$ , interacting with an Einstein-like phononic branch of frequency  $\Omega$ . The electron-phonon coupling is characterized by a quadratic term with a constant interaction vertex  $g$ , and  $N$  represents the number of possible phonon states.

We are interested in the quasiparticle properties of the model, which are encoded in the interacting Green’s function. For polaronic systems the Green’s function can be expressed as a sum over diagrams having an increasing number of phonons, as sketched in Fig. 1(a). Using Feynman’s rules we can translate such an expansion in a mathematical expression

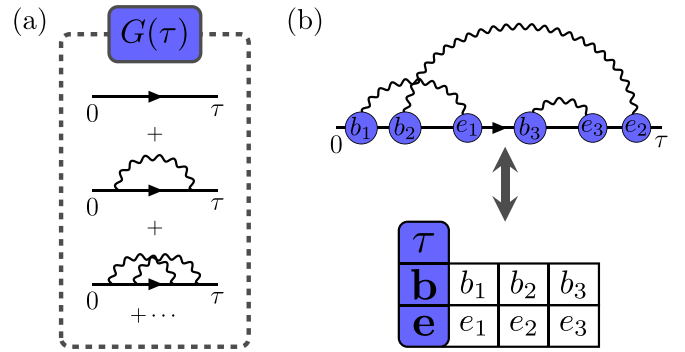


FIG. 1. (a) Graphical representation of the Green’s function expansion for the single-site Holstein model. (b) Vector description of a general diagram used in this work. The interaction times are stored in two separate vectors collecting the beginning and the end of the phonon lines while  $\tau$  encodes the length of the electron one, defining the diagram’s length.

that, in the Matsubara formalism, takes the form

$$G(\tau) = e^{-\varepsilon\tau} \sum_{n=0}^{\infty} \int_0^{\tau} dx_1 \int_{x_1}^{\tau} dx_2 \cdots \int_{x_{2n-1}}^{\tau} dx_{2n} W(\mathbf{x}). \quad (2)$$

The ordered integration accounts for every diagram having  $n$  phonons, each described by a vector of  $2n$  interaction times  $\mathbf{x}$ . The contribution of each diagram is expressed by the zero-temperature electronic and phononic propagators, resulting in the following diagram’s weight:

$$W(\mathbf{x}) = g^{2n} \prod_{i=1}^n e^{-\Omega(e_i - b_i)}, \quad (3)$$

where  $\mathbf{x}$  is divided in the phonon’s creation,  $\mathbf{b}$ , and annihilation,  $\mathbf{e}$ , times, allowing the diagram to be represented by a vector  $[\tau, \mathbf{b}, \mathbf{e}]$  [see Fig. 1(b)].

The DMC method directly integrates Eq. (2) by randomly sampling the expansion’s diagrams from a distribution proportional to their weight function. Detailed information on how the MCMC procedure is implemented in the diagram space can be found in Refs. [40,41], while the observable estimators for polaron quantities are introduced by Mishchenko *et al.* [5]. Here we focus our attention on  $G(\tau)$  and the polaron binding energy  $E_p$ . The Green’s function is evaluated by the histogram collecting the distribution of the variable  $\tau$ , while  $E_p$  is evaluated through the ground-state energy estimator [5]:

$$E_p = \lim_{\tau \rightarrow \infty} \left\langle \frac{\Omega \sum (e_i - b_i) - 2n}{\tau} \right\rangle. \quad (4)$$

Analytical expressions for both observables can be found in the literature [2], making them a suitable choice for testing the convergence properties of a new algorithm.

### B. Flow architecture

Normalizing flows transform a simple starting probability distribution,  $p_Z$ , into a more complex desired density,  $p_X$ , through a series of invertible and smooth transformations. This is achieved using invertible networks [42–44] to construct a variational map  $T_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^D$  that, when applied to

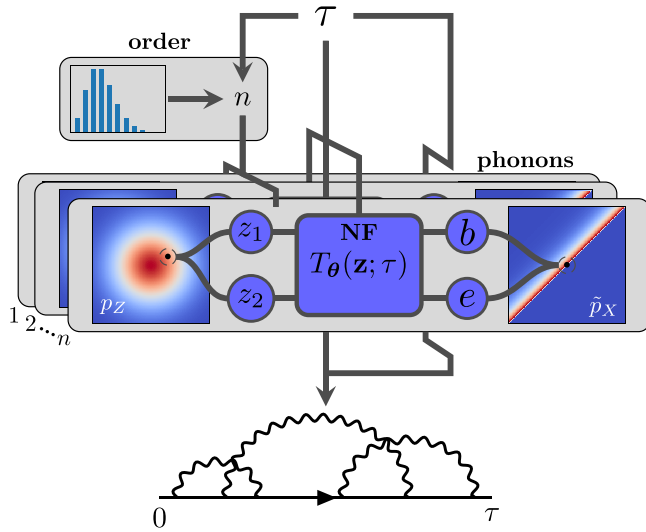


FIG. 2. Representation of the NF-DMC workflow. The model takes as input the diagram's length  $\tau$ , which is used as conditional parameter to sample both the order  $n$  and phonon interaction times  $\mathbf{b}$  and  $\mathbf{e}$ .  $n$  is initially sampled from a suited distribution (e.g., the Poisson distribution shown in histogram form in the inset). Then an increasing number of phonon lines are added to the initial zero-order diagram ( $n = 0$ ) defined by  $\tau$  by transforming Gaussian samples  $\mathbf{z}$  using the  $\tau$ -conditioned NF transform  $T_\theta$ . Finally, for each  $\tau$  an  $n$ -order diagram (i.e., with  $n$  phonon lines) is obtained, as sketched in the bottom.

samples  $\{\mathbf{z}_i\}_{i=0}^N$  drawn from  $p_Z$ , generates a set  $\mathbf{x}_i = T_\theta(\mathbf{z}_i)$  distributed as [37]

$$p_\theta(\mathbf{x}) = p_Z \circ T_\theta^{-1}(\mathbf{x}) \left| \det J_{T_\theta^{-1}}(\mathbf{x}) \right|, \quad (5)$$

with  $J_{T_\theta^{-1}}$  being the Jacobian of the invertible map. This provides fast sampling and density estimation of a parametrized distribution  $p_\theta(\mathbf{x})$ , which can be fitted to the target  $p_X(\mathbf{x})$  by minimizing the Kullback-Leibler (KL) divergence. For DMC the target  $p_X(\mathbf{x})$  is the diagram's weight distribution proportional to the weight  $W(\mathbf{x})$ , with  $\mathbf{x}$  being the diagram's vector representation introduced in Eq. (3) and Fig. 1(b). The complexity of approximating the full  $p_X(\mathbf{x})$  in a NF framework lies in the dimensionality of the vector space spanned by the different diagrams. In fact,  $\mathbf{x}$  changes dimensions based on the number of phonons, described by  $n$  [see Eq. (2)], typically referred to as the diagram's order. Approaching this problem with standard NF architectures would mean creating a different model for every possible dimension  $n$ , leading to the impossibility of covering the entire space. Here, we avoid such restriction by decoupling the sampling of the diagram's order and the interaction times with the approach schematized in Fig. 2.

The strategy relies on sampling single phonons using a conditional NF model  $T_\theta$  [20], with specifics reported in Appendix A, trained to fit a two-dimensional (2D) distribution ( $\tilde{p}_X$ ) describing  $b$  and  $e$  statistics based on the current diagram's length  $\tau$ , thus allowing the generation of a new diagram by multiple use of  $T_\theta$  to dress the zero-order diagram, defined by  $\tau$ , with a random number of phonons

independently sampled from a preselected integer distribution  $p_n(n|\tau)$ . In this way, the entire diagram space at different  $\tau$  can be explored using a single model possessing the following final density:

$$p_W(\mathbf{x}) = p_n(n|\tau) \left[ n! \prod_{i=1}^n p_\theta(b_i, e_i|\tau) \right], \quad (6)$$

where the  $n!$  is needed since every permutation of the phonons' indices leads to the same diagram. The combination of this density with Eq. (5) can be used to completely define an update in the DMC chain.

Finally, the model was completed by selecting  $p_n(n|\tau)$  as a Poisson distribution exhibiting a mean value  $\lambda$  dependent on both  $\tau$  and the el-ph coupling strength  $g$ . The form of  $\lambda$  was chosen to best fit the order statistic collected on pre-existing runs at different values of  $\tau$  with varying  $g$  (see Appendix A for details). This leads to a final  $p_n$  conditioned also by the coupling constant  $g$ . For instance, sampling from different coupling regimes comes at no additional increase in the NF complexity due to the  $g$  dependence being completely contained in the simpler order distribution.

### C. Loss function

In the considered Holstein model, comparing Eq. (6) with the target weight in Eq. (3) shows that the model distribution  $p_\theta$  should be trained to fit the un-normalized density  $p_X(b, e) = e^{-(e-b)}$ , where  $\Omega$  is set as the unit of measure for the energies. Also, to generate a physical phonon the flow sampling needs to take into account that the annihilation  $e$  must happen after the creation  $b$  and both of them need to be placed before the end of the diagram. We found that the most flexible way that allows  $e$  and  $b$  to satisfy such a constraint is to multiply the target  $p_X$  by steep sigmoid functions to create a new  $\tau$ -dependent target  $\tilde{p}_X$  with vanishing contributions in the unphysical regions of the domain:

$$\tilde{p}_X(b, e; \tau) = p_X(b, e) \sigma(b) \sigma(e - b) \sigma(\tau - e), \quad (7)$$

where the sigmoids used were defined as

$$\sigma(x) = \frac{1}{e^{-\alpha x} + 1}, \quad (8)$$

with the  $\alpha$  hyperparameter controlling the steepness of the function, recovering the exact form of the distribution for large values. Therefore, we selected an  $\alpha$  value of 50 in order to obtain a smooth enough  $\tilde{p}_X$  that still presents all the main features of the real distribution, as can be seen in Fig. 3. This allowed us to train an autoregressive neural spline model [45] to map a 2D diagonal Gaussian into  $\tilde{p}_X$  by minimizing the standard reverse KL divergence of the model,

$$\mathcal{L}(\theta) = \langle \ln |\det J_{T_\theta}| - \ln p_X \circ T_\theta + S \circ T_\theta \rangle_{p_Z}, \quad (9)$$

where  $S(x)$  is a sum of softplus functions rising from the sigmoid contributions. We obtained the final  $T_\theta(\mathbf{z}; \tau)$  by an unsupervised training where new samples were generated, at every step, from a set of  $\tau$ 's uniformly distributed in the interesting domain of  $G$  and used to estimate  $\mathcal{L}$  as an average.

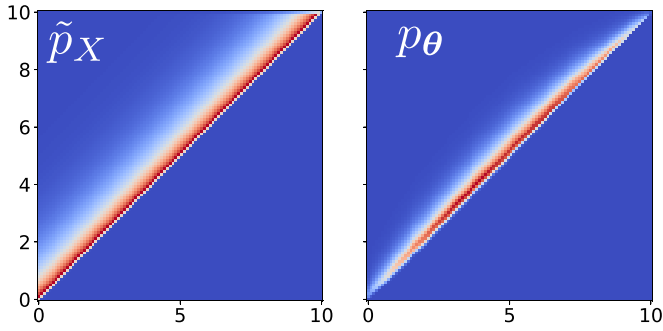


FIG. 3. Target density  $\tilde{p}_X$  and the obtained reconstruction using the NF model  $p_\theta$ . Both are plotted for a selected value of the conditional parameter  $\tau$  of 10, while the hyperparameter  $\alpha$  was set to 50.

Finally, the loss minimization was carried out using standard stochastic gradient descent methods using a learning rate of  $10^{-4}$ , reduced to  $10^{-5}$  after 30 000 steps.

### III. APPLICATION AND DISCUSSION

The developed NF-based architecture was integrated into the set of local updates in the Holstein model, establishing an NF-augmented Markov chain. Subsequently, the statistical properties of this augmented chain were systematically compared against those of the standard one. Extensive data on observables and correlation times were gathered for various coupling strengths ( $g$ ). Each run involved a set of 13 parallel chains, accumulating  $10^7$  steps. The utilization of separate chains facilitated the extraction of uncorrelated estimates, enabling the assessment of errors on individual observable values. This approach offered a direct means of comparing the performance of the two methodologies by analyzing the evolution of mean values and errors throughout the chain.

#### A. Statistical performances

The evolution of the polaron energy ( $E_p$ ) for a specific coupling strength ( $g$ ), as depicted in Fig. 4(a), clearly illustrates that the NF global update reduces the error in the estimate by approximately a factor of 2 compared to the local updates in the standard DMC method. Additionally, the Green's function presented in Fig. 4(b) exhibits nearly identical performance for both types of updates across the entire interaction spectrum. This outcome is anticipated since the diagram's length  $\tau$ , serving as the correlated variable in  $G(\tau)$ , is taken as input from  $p_W$  and is not sampled. Consequently, the correlation in the  $\tau$  channel remains consistent with the standard case, with improvements primarily observed in phonon-related quantities. To quantitatively evaluate the performance of the NF global updates, we present the effective sample size (ESS) as a function of the coupling strength  $g$  in Fig. 5. The NF-based updates exhibit a noticeable reduction in the overall sample correlation, leading to an increase in the ESS across all coupling regimes. This result underscores the effectiveness of the proposed model in accurately reconstructing the target density, thereby enhancing the overall performance of the NF-augmented Markov chain. However, a decrease

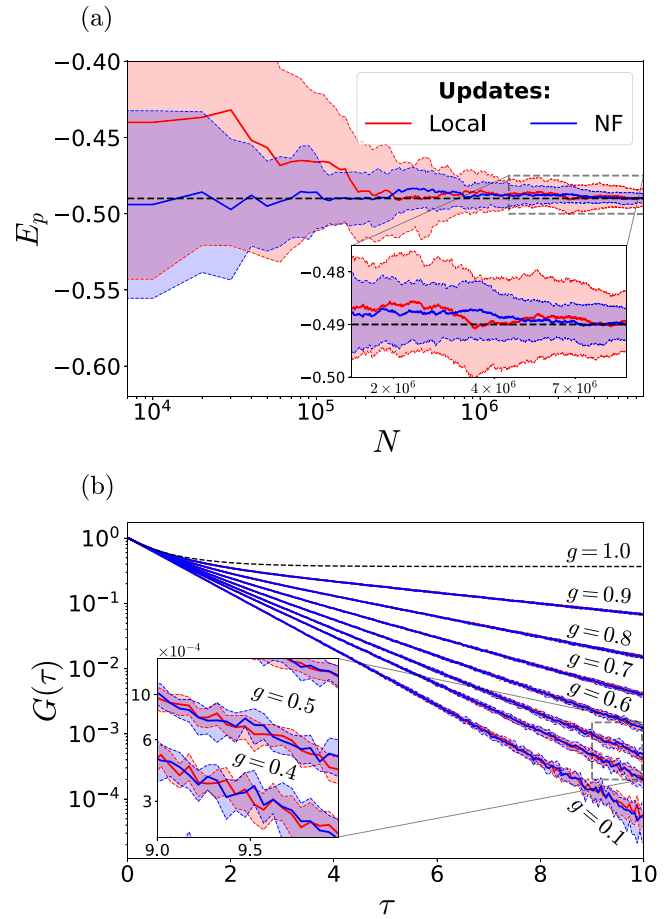


FIG. 4. Comparison between standard and NF-augmented DMC procedures for the estimate of  $E_p$  and  $G(\tau)$ . (a)  $E_p$  at  $g = 0.7$  as a function of the chain's steps. The bold line marks the mean value evolution, the colored area represents the error, and the black dashed line indicates the exact value. (b) Green's function estimate (in logarithmic scale) as a function of  $\tau$  for different coupling strengths  $g$ . Lines and colors have the same meanings as in panel (a).

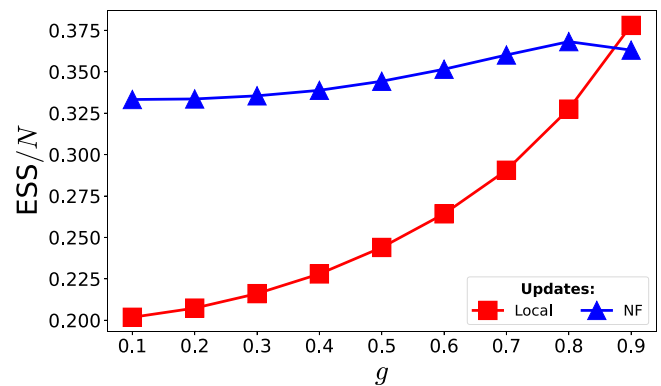


FIG. 5. Effective sample size of the two different Markov chains compared at different coupling constants. The algorithm using the NF global update remains superior along the whole line, showing a reduction in performances only upon reaching  $g$  close to the model instability given by  $G(\tau)$  becoming the identity.

in the performance of the NF model is observed in the strong-coupling regime ( $g = 0.9$ ). It is important to note that this anomalous behavior is attributed to a peculiarity of the physical model rather than the NF architecture itself. Specifically, the single-site Holstein Green's function tends to a constant as  $g$  approaches unity, leading to a flattening of the Green's function with increasing  $g$  in Fig. 4(b). This causes every value of  $\tau$  to have the same statistical weight, resulting in very large  $\tau$  values appearing in the chain. Consequently, the model enters an extrapolation regime, which reduces its performance. This is a system-specific issue that does not undermine the applicability of the NF model. It is always possible to adjust parameters, such as increasing  $\varepsilon$  (currently set to 1), to artificially make high  $\tau$  values less probable without altering the observable estimate.

### B. Computational aspects

The C++ package is constructed on top of the C++ API of PYTORCH [46], while the plots were produced using NUMPY [47] and MATPLOTLIB [48]. The training was carried out on a single Nvidia GeForce GTX 1650 GPU and took a little less than an hour of computer time. The code is openly available in GitHub [49]. Additional details on the NF-DMC architecture are given in Appendix B.

As mentioned in the Introduction, our code was not optimized for performance since the focus was placed on exploring the possible statistical advantages of the methodology. Therefore, no parallelization or batching was implemented, and GPUs were only utilized for training, leading to longer sampling times at higher orders. As a result, at present, the cost of our approach is approximately three times higher compared to the standard one at  $g = 0.1$ , where an  $\sim 40\%$  gain in effective sample size is observed, and becomes worse at larger  $g$ .

## IV. CONCLUSIONS

In conclusion, we presented a global sampling strategy for connected Feynman diagrams based on the use of normalizing flows, while integrating it into an NF-augmented diagrammatic Monte Carlo framework. Our devised procedure was tested on the Holstein polaron model, demonstrating superior statistical performance when compared to the conventional DMC method, albeit at an increased computational cost. Diagrams are constructed using a bottom-up approach that sequentially samples interactions from a single unsupervised NF model trained solely utilizing the weight function of the desired diagram type. This strategy represents an initial attempt to enhance DMC procedures using generative machine learning. It could potentially be extended to other model Hamiltonians by designing NF networks capable of representing the distinctive features of the targeted diagrams. For instance, potential future developments include the incorporation of phonon momenta to account for dispersion and the utilization of multiple NF models to describe systems with combined interactions [12,14]. Moreover, its high flexibility, stemming from the ever-growing variety of NF architectures, the freedom to choose the order distribution, and the option to train without the need for a database

may attract interest from the broader quantum Monte Carlo community, facilitating its application to other perturbative methods.

## ACKNOWLEDGMENTS

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [SFB-F81]. This work was partially supported by the NextGenerationEU-Piano Nazionale Resilienza e Resilienza (PNRR) CN-HPC Grant No. (CUP) J33C22001170001, SPOKE 10. The computational results were achieved using the Vienna Scientific Cluster (VSC). Fruitful discussions with Andrey Mishchenko and Stefano Ragni are greatly acknowledged.

## APPENDIX A: DESCRIPTION OF THE ORDER DISTRIBUTION

The form of the distribution used to sample the order  $p_n(n|\tau, g)$  was selected to best fit the order statistic collected from a set of DMC runs performed at different electron-phonon coupling strengths  $g$ . The data collected showed a form nearly identical to the one of a Poisson distribution for all the  $\tau$  and  $g$  values investigated, leading us to the following choice for the density used during the work:

$$p_n(n|\tau, g) = \frac{\lambda(\tau, g)^n}{n!} e^{-\lambda(\tau, g)}. \quad (\text{A1})$$

To complete the distribution we noticed how the normalization constant  $e^{\lambda(\tau, g)}$  must coincide with the final result of the integration in Eq. (2) reported here:

$$\begin{aligned} G(\tau) &= e^{-\varepsilon\tau} \sum_{n=0}^{\infty} \int_0^{\tau} dx_1 \int_{x_1}^{\tau} dx_2 \cdots \int_{x_{2n-1}}^{\tau} dx_{2n} W(\mathbf{x}) \\ &= e^{-\varepsilon\tau - \lambda(\tau, g)} \sum_{n=0}^{\infty} p_n(n|\tau, g). \end{aligned} \quad (\text{A2})$$

Therefore, by substituting the known result for  $G(\tau)$  [2] and using the fact that  $p_n$  is normalized, one would obtain

$$\exp[-\varepsilon\tau - g^2(\tau - 1 + e^{-\tau})] = \exp[-\varepsilon\tau - \lambda(\tau, g)], \quad (\text{A3})$$

where  $\Omega$  is set as the unit of measure for the energy. At the end, a form of the function  $\lambda(\tau, g)$  was obtained that respected perfectly the behavior of the distribution

$$\lambda(\tau, g) = g^2(\tau - 1 + e^{-\tau}), \quad (\text{A4})$$

which avoided the need to interpolate the means found by fit of the collected statistics to get an approximated version of the exact distribution. A representative test is displayed in Fig. 6.

## APPENDIX B: DESCRIPTION OF THE NORMALIZING FLOW ARCHITECTURE

The model employed in this study is a conditional normalizing flow architecture using rational quadratic splines

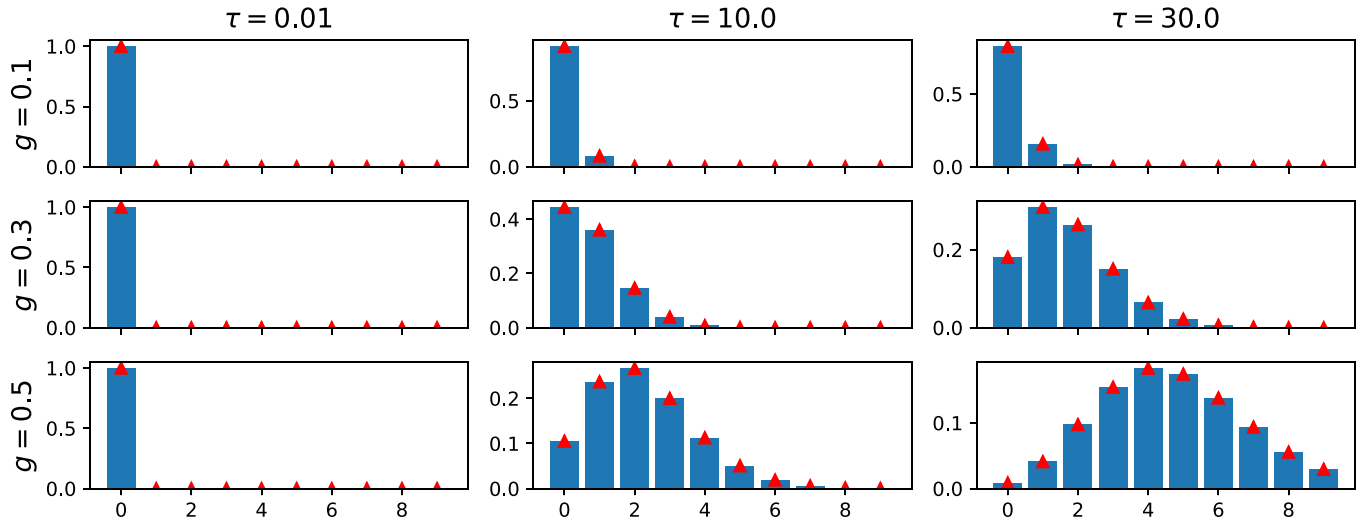


FIG. 6. Test of the obtained  $p_n(n|\tau, g)$  (red triangles) on the collected DMC distribution (blue histograms) at different  $\tau$  and  $g$  values. A perfect overlap of the analytic form used in this work and the raw data is clearly seen.

(RQS) [45] parametrized by the output of a masked multilayer perceptron used as an autoregressive layer [50]. To elucidate this architecture in greater detail, we present the mathematical formulation of  $T_\theta$  in an introductory way.

The parametric transformation constituting the flow is defined by its action on the components of the two-dimensional input vector  $\mathbf{z}$ , which is represented as follows:

$$\mathcal{R}_{\phi_j}(z_j) = \begin{cases} g_{\phi_j}(z_j), & z_j \in [-B, B], \\ z_j, & \text{otherwise,} \end{cases} \quad (\text{B1})$$

where the function  $g_{\phi_j}$  is an invertible RQS parametrized by  $\phi_j$  containing the coordinates and derivatives, strictly positive, of each node. Based on the number of nodes,  $\mathcal{R}_{\phi_j}$  is capable of representing more complex invertible transformations within the boundaries defined by  $B$ , which was fixed to 22

in order to operate within the Green's function's domain of interest [0, 20]. Generally, the transformation is completed by employing specialized neural networks to compute the appropriate  $\phi_j$ , based on the input  $\mathbf{z}$  and our conditional parameter  $\tau$ , leading to the specific transformation mapping the initial distribution  $p_Z$  to the target  $\tilde{p}_X$ . Our model employed a multilayer perceptron (MLP) to evaluate positions and derivatives of 5 nodes per spline, leading to two  $\phi_j$  composed of 15 entries each for the 2D model used in the work. In particular, a masked MLP (MMLP) was constructed based on the implementation of the Masked Autoencoder for Distribution Estimation (MADE) [42], shown in Fig. 7, to compute  $\phi_j$  as a function of  $\tau$  and  $\mathbf{z}_{<j}$  as follows:

$$[\phi_1(\tau), \phi_2(\tau, z_1)] = \text{MMLP}(\mathbf{x}, \theta), \quad \mathbf{x} = \text{concat}(\tau, \mathbf{z}). \quad (\text{B2})$$

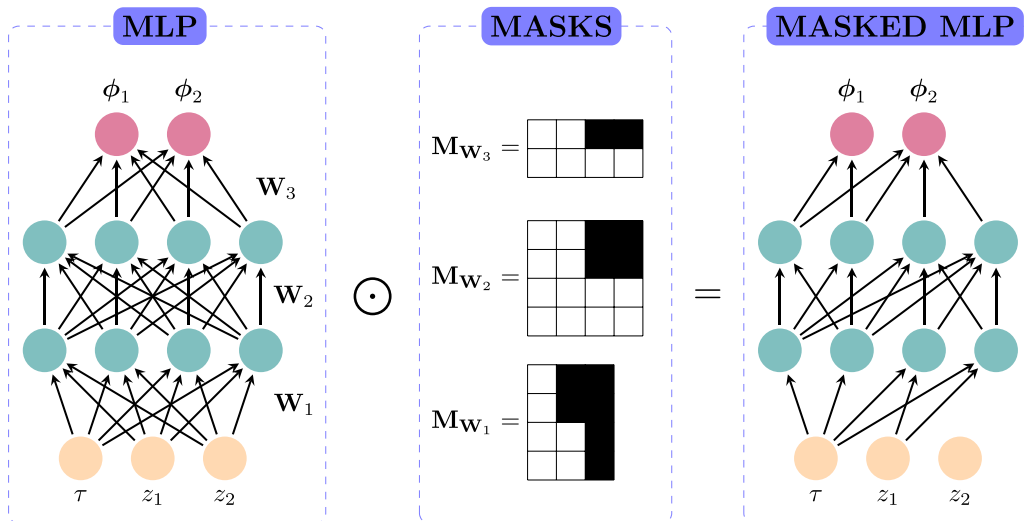


FIG. 7. Graphical description of the masked MLP used in this work adapted from the idea in the MADE implementation [42]. A fully connected network, on the left, gets truncated by multiplying the weight matrices  $\mathbf{W}$  with masks  $\mathbf{M}_W$ . The result is a simpler network, as can be seen on the right, where the  $j$ th output,  $\phi_j$ , clearly possesses the wanted dependence on the inputs  $\tau$  and  $\mathbf{z}_{<j}$ .

Once Eqs. (B1) and (B2) are coupled together they form a map acting on  $\mathbb{R}^2$  parametrized by the network weights  $\theta$  and conditioned by  $\tau$  that constitutes the following conditional flow transform:

$$R_\theta(\mathbf{z}, \tau) = [\mathcal{R}_{\phi_1(\tau)}(z_1), \mathcal{R}_{\phi_2(\tau, z_1)}(z_2)], \quad (\text{B3})$$

$$\det J_{R_\theta}(\mathbf{z}, \tau) = \mathcal{R}'_{\phi_1(\tau)}(z_1) \mathcal{R}'_{\phi_2(\tau, z_1)}(z_2), \quad (\text{B4})$$

where we can see how the determinant of the Jacobian,  $\det J_{R_\theta}(\mathbf{z}, \tau)$ , takes a simple form since the matrix is triangular. The final transformation  $T_\theta(\bullet, \tau)$ , presented as our model in Sec. II B, was given by a composition of two  $R_\theta$ , giving the following variational map:

$$T_\theta(\mathbf{z}, \tau) = R_{\theta_1}[R_{\theta_2}(\mathbf{z}, \tau), \tau], \quad \theta = [\theta_1, \theta_2]. \quad (\text{B5})$$

The Jacobian determinant can then be evaluated using the chain rule of differentiation.

- 
- [1] R. P. Feynman, Space-time approach to quantum electrodynamics, *Phys. Rev.* **76**, 769 (1949).
- [2] G. D. Mahan, *Many-Particle Physics* (Springer, Boston, 2000).
- [3] G. Stefanucci and R. van Leeuwen, *Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction* (Cambridge University, Cambridge, England, 2013).
- [4] N. V. Prokof'ev and B. V. Svistunov, Polaron problem by diagrammatic quantum Monte Carlo, *Phys. Rev. Lett.* **81**, 2514 (1998).
- [5] A. S. Mishchenko, N. V. Prokof'ev, A. Sakamoto, and B. V. Svistunov, Diagrammatic quantum Monte Carlo study of the Fröhlich polaron, *Phys. Rev. B* **62**, 6317 (2000).
- [6] C. Franchini, M. Reticciooli, M. Setvin, and U. Diebold, Polarons in materials, *Nat. Rev. Mater.* **6**, 560 (2021).
- [7] T. Hahn, S. Klimin, J. Tempere, J. T. Devreese, and C. Franchini, Diagrammatic Monte Carlo study of Fröhlich polaron dispersion in two and three dimensions, *Phys. Rev. B* **97**, 134305 (2018).
- [8] A. S. Mishchenko, N. Nagaosa, N. V. Prokof'ev, A. Sakamoto, and B. V. Svistunov, Optical conductivity of the Fröhlich polaron, *Phys. Rev. Lett.* **91**, 236401 (2003).
- [9] T. Hahn, N. Nagaosa, C. Franchini, and A. S. Mishchenko, Diagrammatic quantum Monte Carlo study of an acoustic lattice polaron, *Phys. Rev. B* **104**, L161111 (2021).
- [10] S. Ragni, T. Hahn, Z. Zhang, N. Prokof'ev, A. Kuklov, S. Klimin, M. Houtput, B. Svistunov, J. Tempere, N. Nagaosa, C. Franchini, and A. S. Mishchenko, Polaron with quadratic electron-phonon interaction, *Phys. Rev. B* **107**, L121109 (2023).
- [11] E. A. Burovski, A. S. Mishchenko, N. V. Prokof'ev, and B. V. Svistunov, Diagrammatic quantum Monte Carlo for two-body Problems: Applied to excitons, *Phys. Rev. Lett.* **87**, 186402 (2001).
- [12] E. Burovski, H. Fehske, and A. S. Mishchenko, Exact treatment of exciton-polaron formation by diagrammatic Monte Carlo simulations, *Phys. Rev. Lett.* **101**, 116403 (2008).
- [13] S. A. Kulagin, N. Prokof'ev, O. A. Starykh, B. Svistunov, and C. N. Varney, Bold diagrammatic Monte Carlo technique for frustrated spin systems, *Phys. Rev. B* **87**, 024407 (2013).
- [14] G. De Filippis, A. de Candia, A. S. Mishchenko, L. M. Cangemi, A. Nocera, P. A. Mishchenko, M. Sasseti, R. Fazio, N. Nagaosa, and V. Cataudella, Quantum phase transition of many interacting spins coupled to a bosonic bath: Static and dynamical properties, *Phys. Rev. B* **104**, L060410 (2021).
- [15] K. Chen and K. Haule, A combined variational and diagrammatic quantum Monte Carlo approach to the many-electron problem, *Nat. Commun.* **10**, 3725 (2019).
- [16] U. Wolff, Critical slowing down, *Nucl. Phys. B* **17**, 93 (1990).
- [17] U. Wolff, Collective Monte Carlo updating for spin systems, *Phys. Rev. Lett.* **62**, 361 (1989).
- [18] N. Prokof'ev, in *Many-Body Methods for Real Materials*, edited by E. Pavarini, E. Koch, and S. Zhang, Schriften des Forschungszentrums Jülich. Modeling and Simulation Vol. 9 (Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag, Jülich, 2019), Chap. 15, pp. 421–447.
- [19] M. S. Albergo, 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).
- [20] A. Singha, D. Chakrabarti, and V. Arora, Conditional normalizing flow for Markov chain Monte Carlo sampling in the critical region of lattice field theory, *Phys. Rev. D* **107**, 014512 (2023).
- [21] A. Singha, D. Chakrabarti, and V. Arora, Sampling  $U(1)$  gauge theory using a retrainable conditional flow-based model, *Phys. Rev. D* **108**, 074518 (2023).
- [22] 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).
- [23] L. Huang and L. Wang, Accelerated Monte Carlo simulations with restricted Boltzmann machines, *Phys. Rev. B* **95**, 035105 (2017).
- [24] M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden, Adaptive Monte Carlo augmented with normalizing flows, *Proc. Natl. Acad. Sci. USA* **119**, e2109420119 (2022).
- [25] 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).
- [26] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [27] J. R. Moreno, G. Carleo, and A. Georges, Deep learning the Hohenberg-Kohn maps of density functional theory, *Phys. Rev. Lett.* **125**, 076402 (2020).
- [28] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Finding density functionals with machine learning, *Phys. Rev. Lett.* **108**, 253002 (2012).
- [29] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Bypassing the Kohn-Sham equations with machine learning, *Nat. Commun.* **8**, 872 (2017).
- [30] S.-H. Li and L. Wang, Neural network renormalization group, *Phys. Rev. Lett.* **121**, 260601 (2018).

- [31] D. Di Sante, M. Medvidović, A. Toschi, G. Sangiovanni, C. Franchini, A. M. Sengupta, and A. J. Millis, Deep learning the functional renormalization group, *Phys. Rev. Lett.* **129**, 136402 (2022).
- [32] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
- [33] D. Wu, L. Wang, and P. Zhang, Solving statistical mechanics using variational autoregressive networks, *Phys. Rev. Lett.* **122**, 080602 (2019).
- [34] 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).
- [35] N. Prokof'ev and B. Svistunov, Worm algorithms for classical statistical models, *Phys. Rev. Lett.* **87**, 160601 (2001).
- [36] U. Wolff, Monte Carlo simulation of a lattice field theory as correlated percolation, *Nucl. Phys. B* **300**, 501 (1988).
- [37] 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).
- [38] 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).
- [39] T. Holstein, Studies of polaron motion: Part I. The molecular-crystal model, *Ann. Phys.* **8**, 325 (1959).
- [40] K. Van Houcke, E. Kozik, N. Prokof'ev, and B. Svistunov, Diagrammatic Monte Carlo, *Phys. Procedia* **6**, 95 (2010).
- [41] J. Greitemann and L. Pollet, Lecture notes on diagrammatic Monte Carlo for the Fröhlich polaron, *SciPost Phys. Lect. Notes* **2**, 1 (2018).
- [42] M. Germain, K. Gregor, I. Murray, and H. Larochelle, MADE: Masked autoencoder for distribution estimation, *J. Mach. Learn. Res.* **37**, 881 (2015).
- [43] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using Real NVP, [arXiv:1605.08803](https://arxiv.org/abs/1605.08803).
- [44] D. P. Kingma and P. Dhariwal, Glow: Generative flow with invertible  $1 \times 1$  convolutions, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Red Hook, NY, 2018), Vol. 31.
- [45] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, 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, Red Hook, NY, 2019), Vol. 32.
- [46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, PyTorch: An imperative style, high-performance deep learning library, 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, Red Hook, NY, 2019), Vol. 32, pp. 8024–8035.
- [47] 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, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant *et al.*, Array programming with NumPy, *Nature (London)* **585**, 357 (2020).
- [48] J. D. Hunter, Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [49] The code developed in this study is available at <https://github.com/QuantumMaterialsModelling/NF-DiagMC>.
- [50] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017), Vol. 30.