Autoregressive neural quantum states of Fermi Hubbard models

Eduardo Ibarra-García-Padilla,[1,2,*] Hannah Lange,[3,4,5] Roger G. Melko,[6,7] Richard T. Scalettar,[1] Juan Carrasquilla,[8] Annabelle Bohrdt,[5,9] and Ehsan Khatami[2,†]

[1]*Department of Physics and Astronomy, University of California, Davis, California 95616, USA*
[2]*Department of Physics and Astronomy, San José State University, San José, California 95192, USA*
[3]*Ludwig-Maximilians-University Munich, Theresienstr. 37, Munich D-80333, Germany*
[4]*Max-Planck-Institute for Quantum Optics, Hans-Kopfermann-Str.1, Garching D-85748, Germany*
[5]*Munich Center for Quantum Science and Technology, Schellingstr. 4, Munich D-80799, Germany*
[6]*Department of Physics and Astronomy, University of Waterloo, 200 University Ave. West, Waterloo, Ontario N2L 3G1, Canada*
[7]*Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[8]*Institut für Theoretische Physik, Eidgenössische Technische Hochschule Zürich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland*
[9]*University of Regensburg, Universitätsstr. 31, Regensburg D-93053, Germany*

(Received 9 November 2024; accepted 13 January 2025; published 3 February 2025)

Neural quantum states (NQSs) have emerged as a powerful ansatz for variational quantum Monte Carlo studies of strongly correlated systems. Here, we apply recurrent neural networks (RNNs) and autoregressive transformer neural networks to the Fermi-Hubbard and the (non-Hermitian) Hatano-Nelson-Hubbard models in one and two dimensions. In both cases, we observe that the convergence of the RNN ansatz is challenged when increasing the interaction strength. We present a physically motivated and easy-to-implement strategy for improving the optimization, namely, by ramping of the model parameters. Furthermore, we investigate the advantages and disadvantages of the autoregressive sampling property of both network architectures. For the Hatano-Nelson-Hubbard model, we identify convergence issues that stem from the autoregressive sampling scheme in combination with the non-Hermitian nature of the model. Our findings provide insights into the challenges of the NQS approach and make the first step towards exploring strongly correlated electrons using this ansatz.

DOI: 10.1103/PhysRevResearch.7.013122

I. INTRODUCTION

The Fermi-Hubbard model (FHM) describes itinerant, interacting spin-1/2 electrons hopping on a set of spatially localized orbitals. Despite its simplicity, it is paradigmatic for our understanding of electronic correlations in quantum materials in which strong Coulomb interactions play an essential role. Its importance lies in the fact that it accurately captures some of the key characteristics of strongly correlated materials. Numerous phases this model can display have striking similarities to the behaviors observed in a wide range of complex materials [1–5].

In recent years, neural network quantum states (NQSs) [6–16] have emerged as promising ansatz for variational wave functions. There are several remarkable aspects that make NQSs attractive for use in the field of quantum many-body physics. First is the ability of the NQS to capture a wide array of quantum states important to the study of quantum many-body systems [13,14]. It has been shown that they can encode volume law entangled states [17–21].

The second feature is related to the scaling of computational resources with the system size. The optimization and evaluation of observables in NQS relies on sampling, typically using a Metropolis sampling algorithm, which can become computationally very demanding. Here, we show that *autoregressive* networks, with normalized wave-function amplitudes that allow direct sampling instead of Metropolis sampling, can be very efficient for correlated electron systems. The autoregressive property refers to the use of a chain rule of probabilities to generate a sequence of data elements in which the probability of every element in the sequence depends only on the configuration of the set of elements that came before it. The scaling of the method with the system size can lead to a potentially huge advantage over conventional numerical treatment, such as the exact diagonalization (ED), density matrix renormalization group (DMRG), or quantum Monte Carlo (QMC) [22–25], which suffer from either exponential scaling or other technical issues arising due to the fermion "sign problem" [26–28].

Two prominent examples for autoregressive networks are recurrent neural networks (RNNs) [29] and transformer neural networks [30–35] (although nonautoregressive versions of transformers are possible, see, e.g., Refs. [36,37]). RNNs

*Contact author: edibarra@ucdavis.edu
†Contact author: ehsan.khatami@sjsu.edu

were initially developed for natural language processing and are *generative*, i.e., they can be trained to infer the probability distribution of unlabeled data, which can then be used to generate more data. The final joint distribution of configurations in an RNN $P(\sigma)$, where σ represents a configuration in a complete set, is normalized. This is a powerful property; not only can one sample from a trained RNN but also the trained RNN can, given a new configuration, return the associated normalized probability. The same property can be enforced for a transformer neural network by masking out future inputs to the attention mechanism, see, e.g., Refs. [30–35].

In a pioneering work [29], Hibat-Allah *et al.* showed that ground states of spin models can accurately be represented using a wave function based on an RNN with two output layers for the amplitude and phase of the wave function. Namely, samples are drawn from the generative model and the energy is computed and averaged over many samples and minimized with respect to the parameters of the RNN using the stochastic gradient descent. Weights and biases inside a shared RNN unit are among the quantities that form the optimization parameters. Since then, there have been several notable studies exploring aspects of the use of RNNs in quantum many-body physics, spanning representability, accuracy, performance [38–42], including through the use of symmetries of the Hamiltonian [43] or more advanced autoregressive architectures such as *transformers* [32–34,44].

Neural network wave functions have also been used recently to study itinerant electron models at half filling and beyond [41,45–57]. In Ref. [41], the authors use RNNs to represent the ground state of the t-J, t-XXZ, and t-J_z models in one and two spatial dimensions. They present a novel technique for mapping out the dispersion relation of the single hole, complementary to other schemes [58–60], and show that their results compare well with DMRG. This is accomplished through computing the expectation values of the translation operator for periodic systems. However, they establish that the ground-state uncertainties for these models are generally much larger than those for spin models studied previously.

Here, we utilize RNNs as a variational ansatz to access the ground state of the FHM. We find that, as expected, a naive application of the technique results in accuracies that are less accurate than in the case of the t-J model. We then introduce a ramping mechanism in which certain model parameters are gradually tuned from the initial values to the desired final values during the training process and show that the scheme leads to orders-of-magnitude improvements in the ground-state energy and other observables (this is also known as variational neural annealing [61]). We benchmark our results for the method applied to the one-dimensional (1D) and two-dimensional (2D) FHM at half filling and find that the accuracy of ground-state properties is generally independent of the system size for the same number of training steps, confirming that the computational resources will indeed grow linearly with system size. We then use the method to study a non-Hermitian Hamiltonian in which the tunneling rates to the left and right are unequal [the Hatano-Nelson-Hubbard model (HNHM)] and discuss the limitations of current RNN architectures for such problems.

The remainder of this paper is organized as follows: In Sec. II we present the Fermi-Hubbard model, the observables computed, the details of the RNNs, and the training schemes used. In Sec. III we present our main findings, first for the 1D FHM, second for the 2D FHM, and third for the 1D HNHM. Section IV summarizes our findings and presents an outlook for future studies.

II. MODEL AND METHODS

A. Fermi-Hubbard model

We study the FHM, whose Hamiltonian is expressed as

$$\hat{H}_{\mathrm{FH}} = -t \sum_{\langle i,j\rangle,\sigma} (\hat{c}^{\dagger}_{i\sigma}\hat{c}_{j\sigma} + \mathrm{H.c.}) + U \sum_{i} \hat{n}_{i\uparrow}\hat{n}_{i\downarrow} - \mu \sum_{i,\sigma} \hat{n}_{i\sigma}, \tag{1}$$

where $\hat{c}^{\dagger}_{i\sigma}$ ($\hat{c}_{i\sigma}$) is the creation (annihilation) operator for a fermion with spin σ on site i, $\hat{n}_{i\sigma}$ $\hat{c}^{\dagger}_{i\sigma}\hat{c}_{i\sigma}$ is the number operator for spin σ on site i, $\langle i, j\rangle$ denotes the sum over nearest neighbors, t is the nearest-neighbor hopping amplitude, U is the interaction strength, and μ is the chemical potential that controls the fermion density in the grand canonical ensemble. We consider the repulsive case $U > 0$. We set the energy scale to be $t = 1$. We consider 1D chains with $N = L$ sites and 2D square lattices with $N = L_x \times L_y$ sites. In all cases, open boundary conditions (OBCs) are considered. We work in the grand canonical ensemble and set the chemical potential to $\mu = U/2$ to achieve half filling on average in all cases involving this model.

The HNHM is described by a non-Hermitian Hamiltonian with unequal tunneling rates for left and right directions,

$$\hat{H}_{\mathrm{HNH}} = -t \sum_{i,\sigma} [(1 + g/t)\hat{c}^{\dagger}_{i+1\sigma}\hat{c}_{i\sigma} + (1 - g/t)\hat{c}^{\dagger}_{i\sigma}\hat{c}_{i+1\sigma}] + U \sum_{i} \hat{n}_{i\uparrow}\hat{n}_{i\downarrow} - \mu \sum_{i,\sigma} \hat{n}_{i\sigma}. \tag{2}$$

An important consequence of the anisotropic tunneling rates is that the HNHM is particle-hole symmetric (PHS) under the transformation $\hat{c}^{\dagger}_{i\sigma} \to (-1)^i \hat{c}_{i\sigma}$ and $g \to -g$. The regular particle hole transformation [only $\hat{c}^{\dagger}_{i\sigma} \to (-1)^i \hat{c}_{i\sigma}$] maps the left kinetic-energy term into the right one, and vice versa, therefore the change in sign of g is crucial to achieve PHS.

We compute the energy $E = \langle \hat{H} \rangle / N$, the density

$$n = \frac{1}{N} \sum_{i,\sigma} \langle \hat{n}_{i\sigma} \rangle, \tag{3}$$

the kinetic energy

$$K = \frac{1}{N} \left\langle -t \sum_{\langle i,j\rangle,\sigma} (\hat{c}^{\dagger}_{i\sigma}\hat{c}_{j\sigma} + \mathrm{H.c.}) \right\rangle, \tag{4}$$

the double occupancy

$$\mathcal{D} = \frac{1}{N} \sum_{i} \langle \hat{n}_{i\uparrow}\hat{n}_{i\downarrow} \rangle, \tag{5}$$

and the nearest-neighbor (nn) spin-spin correlation function

$$\langle S_z S_z \rangle_{\rm nn} = \frac{1}{N_b} \sum_i \sum_{\delta \in \mathcal{S}(i)} \langle \hat{S}_z^i \hat{S}_z^{i+\delta} \rangle,$$

$$\text{where} \quad \langle \hat{S}_z^i \hat{S}_z^{i+\delta} \rangle = \frac{1}{4} \sum_\sigma (\langle \hat{n}_{i\sigma} \hat{n}_{i+\delta\sigma} \rangle - \langle \hat{n}_{i\sigma} \hat{n}_{i+\delta\bar{\sigma}} \rangle), \quad (6)$$

N_b is the number of bonds, $\mathcal{S}(i)$ denotes the set of nearest-neighbor vectors consistent with the OBC for site i, and $\bar{\sigma}$ denotes the opposite spin to σ.

B. Recurrent neural network wave functions

We use a recurrent neural network (RNN) to represent a pure quantum state in each of our case studies. The wave function ansatz is given by

$$|\psi_\lambda\rangle = \sum_\sigma \sqrt{p_\lambda(\sigma)} e^{i\phi_\lambda(\sigma)} |\sigma\rangle, \quad (7)$$

where λ denotes the variational parameters of the ansatz wave function $|\psi_\lambda\rangle$, and $|\sigma\rangle$ are the elements of the computational basis, i.e., $|\sigma\rangle = |\sigma_1, \sigma_2, \ldots, \sigma_N\rangle$, where σ_i can take any value of the local Hilbert space $\{0, \uparrow, \downarrow, \uparrow\downarrow\}$. λ includes *hidden variables* h_i used to pass information from site i to site $i+1$ during a given sampling step. The size of h_i is given by the number of hidden units, n_h.

In this work, we use one RNN cell and a Softmax layer to model the probability, together with a Softsign layer to model the phase. Specifically, we implement the gated recurrent unit (GRU) as the elementary cell in our RNNs. For further details on the parametrization of the GRU and RNN in this scheme, we refer the reader to Ref. [29].

To find the ground state of the system, we perform variational Monte Carlo (VMC) to minimize the energy. In that case, we minimize the expectation value of the energy of our ansatz wave function

$$\langle H_\lambda \rangle = \sum_\sigma |p_\lambda(\sigma)| \epsilon_\lambda(\sigma), \quad (8)$$

where we have defined

$$\epsilon_\lambda(\sigma) = \sum_\tau H_{\tau\sigma} \sqrt{\frac{p_\lambda(\tau)}{p_\lambda(\sigma)}} e^{i[\phi_\lambda(\tau) - \phi_\lambda(\sigma)]}, \quad (9)$$

and $H_{\tau\sigma} = \langle \tau | H | \sigma \rangle$.

In practice, we consider the cost function

$$C = \sum_\sigma |p_\lambda(\sigma)| [\epsilon_\lambda(\sigma) - \langle \epsilon_\lambda(\sigma) \rangle] \quad (10)$$

to minimize both the energy and its variance to stabilize the training, and we use the adaptive moment estimation (adam) optimizer to implement the gradient updates.

C. Pretraining with projective measurements

References [34,38,40] demonstrated the potential of hybrid quantum-classical methods for large-scale quantum many-body system simulation by merging experimental data from current quantum devices with autoregressive language models.

More specifically, they observed that if the RNN wave function is first trained using experimental projective measurements (which we name pretraining) and then VMC is performed, the convergence of the procedure is significantly improved. The only difference between these two settings is the loss function governing the optimization. In the VMC stage the loss function is the energy while in the pretraining stage the loss function is the Kullback-Leibler (KL) divergence [40],

$$L_{\rm KL} = \sum_\sigma p_d(\sigma) \ln \left[\frac{p_d(\sigma)}{p_\lambda(\sigma)} \right], \quad (11)$$

where p_d is the empirical distribution of the dataset. The minimum of this loss function occurs when $p_d(\sigma) = p_\lambda(\sigma)$. Therefore, minimizing the KL divergence drives the matching of the distributions. Note that in practice, it suffices to minimize $-\langle \ln[p_\lambda(\sigma)] \rangle_{p_d}$ in this stage.

In this work, we also experiment with pretraining our RNN models using projective measurements obtained from sampling of the ground-state function in the computational basis using exact diagonalization. For each case study, we generate and use 10 000 samples.

D. Ramping of Hubbard parameters

Additionally, we propose an alternative way of enhancing VMC simulators in which Hubbard parameters are ramped to their desired value. This idea is inspired by the experimental protocol for preparing low-entropy samples in optical lattices and optical tweezer arrays in which it is more efficient to load a band insulator and then modify the lattice (increase the number of sites) to get on average one particle per site or other target fillings [62,63]. In our proposed training scheme, instead of modifying the lattice geometry (as is often done in experiments), we modify the hopping amplitude during the training.

As we discuss in Sec. III, we find that ramping the tunneling rate from a larger t to the desired final value provides a scheme that either alone or in conjunction with pretraining with projective measurements, provides better results than VMC with pretraining using projective measurements alone.

The reason behind starting from large t rather than any of the other Hubbard parameters is threefold: (1) While the μ and U terms in the Fermi-Hubbard Hamiltonian are diagonal in the $n_{i\sigma}$ basis, the t term is not. For that reason, the kinetic-energy term mixes elements of the computational basis and its calculation involves knowledge of both the amplitudes and the phases (in contrast with the other two terms that only require the amplitudes). Therefore, starting from large t biases the RNN to learn amplitudes and phases simultaneously. (2) Starting from a larger t, the system promotes double occupancies, which are essential for reaching the true ground state [64] (3) Half-filling always occurs at $\mu = U/2$ independently of t, so the target density remains constant during all stages of the ramp.

Specific details of the ramps used in the study are presented in Appendix A.

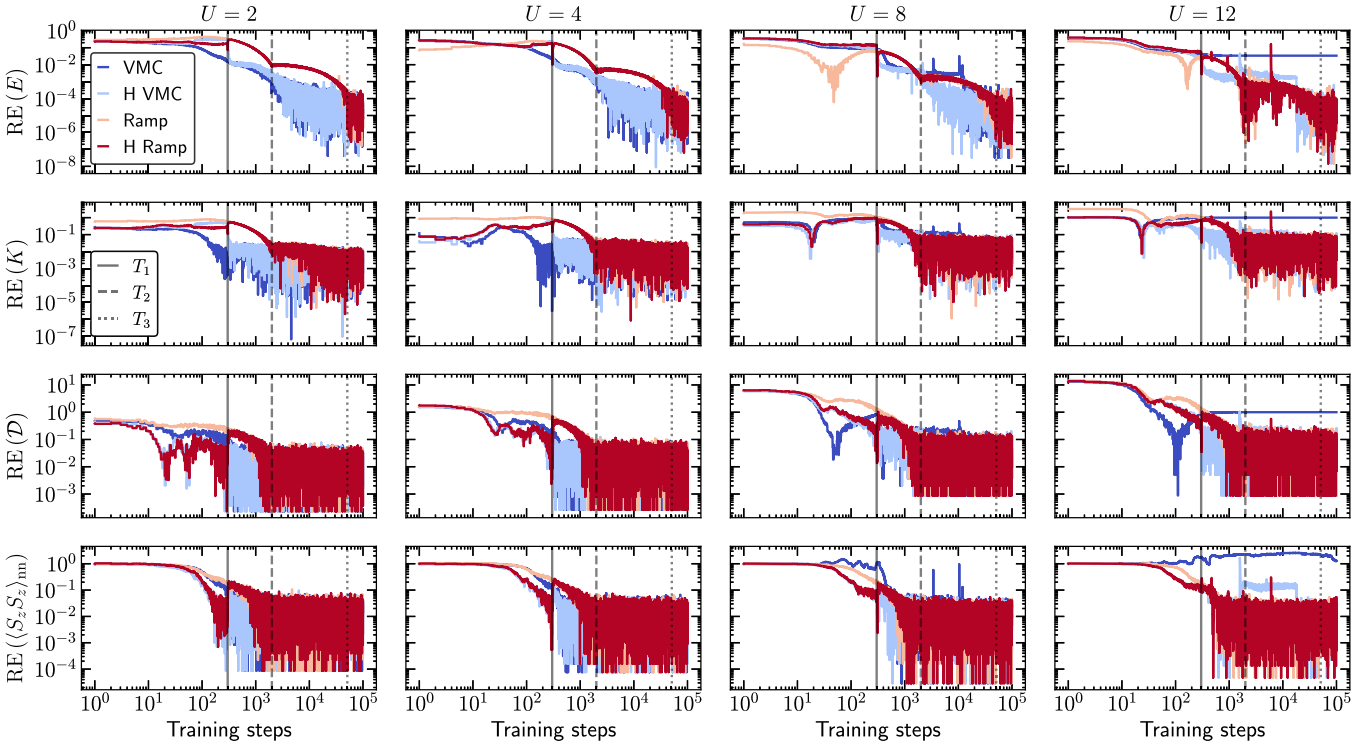

FIG. 1. Relative errors for the half filled FHM as a function of training steps for different methods on a 10-site chain. Columns correspond to $U = 2, 4, 8, 12$ and rows to the relative errors in energy, kinetic energy, double occupancy, and nearest-neighbor spin-spin correlation function. Results are presented for the best random seed for all methods. Vertical lines at $T_1 = 300$ (solid), $T_2 = 2000$ (dashed), and $T_3 = 51\,000$ (dotted) training steps corresponds to pretraining and ramping stages. Sharp jumps in the H Ramp curves (red) at T_1 are because the tunneling rate of the Hamiltonian is changed here from t to approximately $1.55t$.

E. Exact diagonalization and density matrix renormalization group

To benchmark our results we compare against numerically exact methods: ED, and the DMRG. We perform ED on $N = 6, 8, 10$ chains and DMRG on $N = 20$ and 100 chains as well as the $N = 4 \times 4$ lattice using the iTensor package [65] with an adaptable bond dimension and 30 sweeps, keeping the error for the energy below 10^{-8}.

III. RESULTS

For the results presented in this section, unless otherwise specified, we used $n_h = 100$ hidden units, a learning rate of $\ell_r = 0.001$, and generate $N_s = 1000$ samples for open-boundary conditions per training step. When pretraining is performed, we do it over the first 300 training steps, using 10 000 samples.

For every case study, we run the training using at least 25 random initial parameters of the neural network. We call these *realizations*. For all the observables $\mathcal{O}$ considered, we compute the relative error as $\mathrm{RE} = |1 - \langle \mathcal{O} \rangle / \mathcal{O}_{GS}|$, where $\langle \mathcal{O} \rangle$ is either the value of the observable at each training step or the average of the observable over the last 100 training steps as specified, and $\mathcal{O}_{GS}$ is the exact value of the observable obtained using ED (or DMRG where explicitly indicated). We also define the *best realization* as the realization with the smallest relative error in energy.

In the following, we present results for different training schemes which we refer to as *methods*. These are labeled as follows:

(1) VMC: Minimize the energy only.

(2) H VMC: (Hybrid VMC) Perform pretraining, followed by energy minimization.

(3) Ramp: Perform the training while ramping t. The tunneling rate is initialized from a large value $t_i > t$ and decreases exponentially. The tunneling is set to its final value t at 51 000 training steps.

(4) H Ramp: (Hybrid Ramp) Perform pretraining, followed by training while ramping t, which is set to its final value at 51 000 training steps. We use the final value of t also in the pretraining stage to evaluate E. After pretraining is finished, the tunneling rate is set to the value obtained with the Ramp after 300 training steps (which is approximately $1.55t$). For the remainder of the realization, the tunneling rate is updated using the Ramp.

A. One-dimensional Hubbard chain

In Fig. 1, we present the relative errors of E, K, $\mathcal{D}$, and $\langle S_z S_z \rangle_{\rm nn}$ as functions of training steps for $U = 2, 4, 8, 12$ on a 10-site chain at half filling for the best realization of the RNN. In this figure, the relative errors are calculated for the expectation values of the observables at each training step and no averaging over training steps is performed.

For $U \leqslant 8$, all methods are able to get converged results, and their relative errors are consistent with each other.

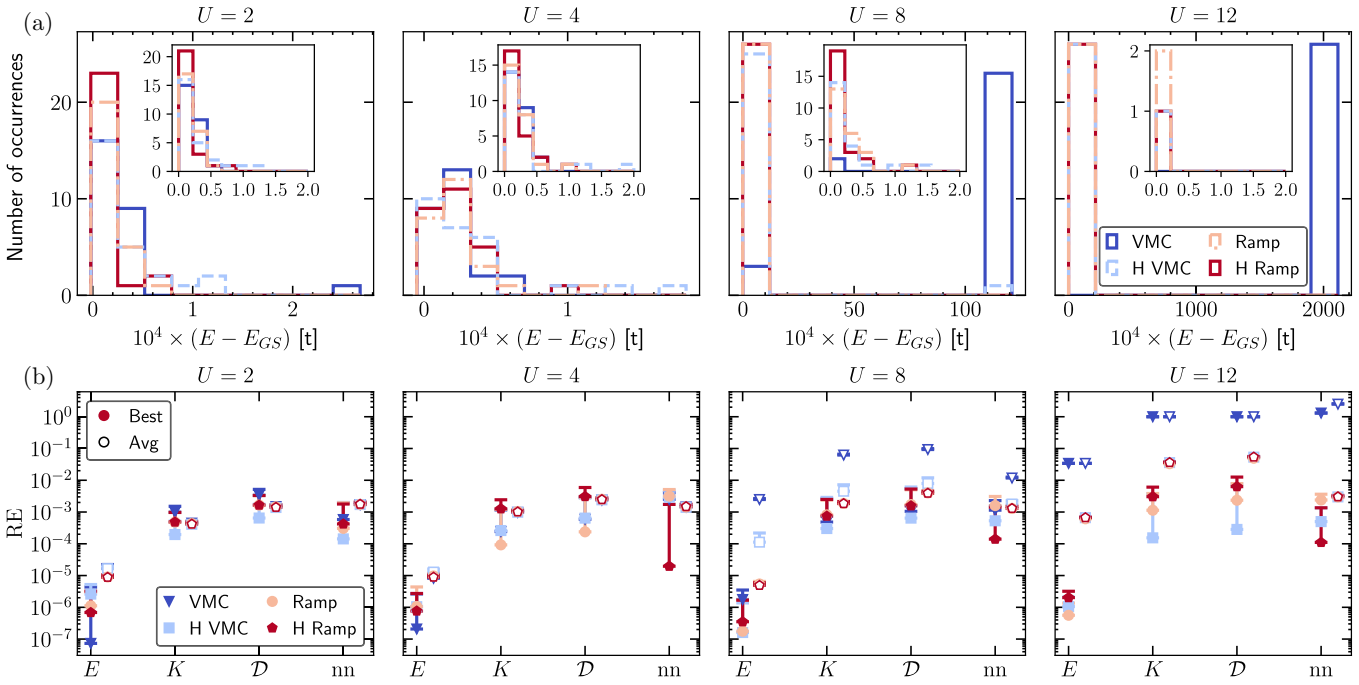

FIG. 2. (a) Energy histograms for the half filled FHM for different methods on a 10-site chain. Columns correspond to $U = 2, 4, 8, 12$. Results are obtained for 26 different initial random seeds, and the energies are reported as the average of the last 100 training steps for each random initial condition. The insets replot the data in the same zoomed-in range for all U to aid the comparison. (b) Relative errors for the half filled FHM for different methods on a 10-site chain. Columns correspond to $U = 2, 4, 8, 12$. Solid markers correspond to the best realization. Error bars are only presented for the upper bound and correspond to the mean of the relative errors obtained using $\langle \mathcal{O} \rangle \pm \sigma_{\mathcal{O}}$, where $\sigma_{\mathcal{O}}$ is the standard error of the mean (s.e.m.). Open markers correspond to the average of the relative errors obtained using 26 different random seeds and their error bars are the s.e.m. of these different realizations.

However, for $U = 12$, VMC alone is unable to do so and quickly gets stuck in the atomic limit solution ($U/t \to \infty$) where double occupancies are strongly suppressed.

It is important to note that, for the converged results, the relative error in E is approximately constant for all values of U and falls within the range $(10^{-7}, 10^{-3})$. In contrast, the relative errors in K and $\mathcal{D}$ worsen as the interaction strength is increased. This is evinced by the rise in the lower and upper bounds of the ranges containing these relative errors [for example, for K at $U = 2$ these are $(10^{-6}, 10^{-2})$, but increase to $(10^{-5}, 10^{-1})$ at $U = 12$]. This is indicative of the fact that the errors in K and $\mathcal{D}$ are correlated, similarly to what is observed in quantum Monte Carlo simulations. In contrast, the relative error of $\langle S_z S_z \rangle_{\mathrm{nn}}$ rapidly converges and also remains mostly constant for all U considered. It is worth noting that the lower bounds in the relative errors for $\mathcal{D}$ and $\langle S_z S_z \rangle_{\mathrm{nn}}$ are flat, whereas those for K and E continue to decrease and exhibit oscillations as a function of the training step for longer. We speculate that such difference in behavior may be traced back to the requirement of the knowledge of phases in calculating the latter quantities and the fact that $\phi_\lambda(\sigma)$ may continue to be learned after $p_\lambda(\sigma)$ are converged.

Although almost all methods (except for VMC alone for $U = 12$) are able to yield a converged RNN wave function with comparable relative errors [$\lesssim 10^{-6}$ in E and $\lesssim 10^{-3}$ in $\langle S_z S_z \rangle_{\mathrm{nn}}$], an extensive search over initial random configurations may be needed to find such lowest-energy configurations for some of these methods. To reveal that, in Fig. 2(a) we present histograms of the lowest energies obtained with the different methods, while in Fig. 2(b) we compare the relative errors in the observables between the best realization (solid markers) and those obtained by averaging 26 different realizations (open markers) for each method.

At $U = 2$, all methods have a large weight in states with energies very close to the ground state [see Fig. 2(a)]. However, performing VMC alone produces the least number of realizations close to the exact ground-state energy. While pretraining followed by VMC (H VMC) produces better realizations, the two ramps yield a larger number of realizations closer to the exact ground state and with shorter tails. When comparing the relative errors of the different methods for $U = 2$ in Fig. 2(b), we observe that, except for the energy, the best realization (as judged by the relative error in energy) and the averaged data for all methods are consistent with each other within error bars.

As the interaction strength is increased to $U = 4$, the histograms exhibit a similar profile as the one observed for $U = 2$, but for all methods the tails have grown, which increases the relative error of the averaged results, as illustrated in Fig. 2(b). Furthermore, the results obtained using the Ramp, H VMC, and the hybrid ramp (H Ramp) yield the lowest relative errors for the kinetic energy, double occupancy, and spin-spin correlation functions.

These findings are more evident when analyzing the $U = 8$ results. At this interaction strength, most of the realizations using VMC alone get stuck in higher-energy configurations,

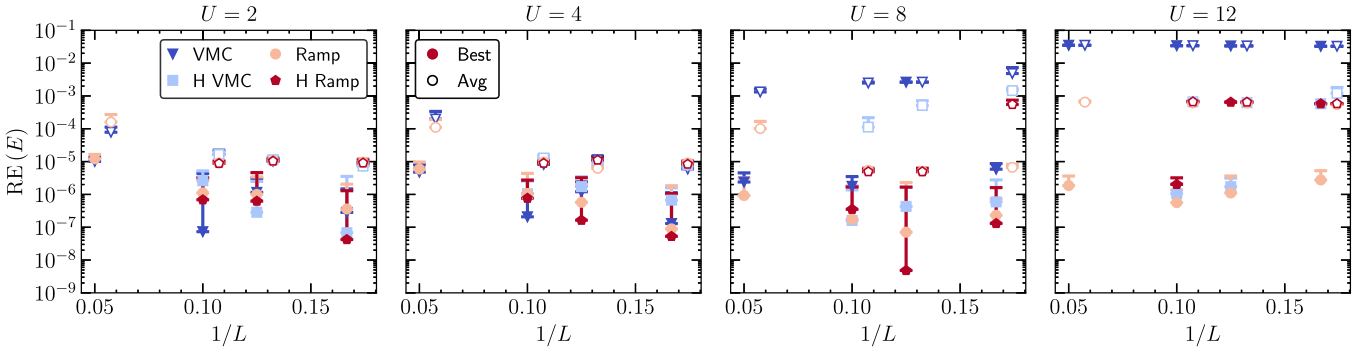

FIG. 3. Relative errors of the energy for the half filled FHM for different methods as a function of system size. Columns correspond to $U = 2, 4, 8, 12$. Solid and open markers are the same as in Fig. 2(b).

and only two realizations are able to get close to the ground state [see Fig. 2(a)]. On the other hand, the rest of the methods yield realizations with energies very close to the ground state, where the H Ramp method yields the largest number of realizations close to the ground state and a shorter tail for the histogram. Their relative errors in Fig. 2(b) highlight that, for the best realization, hybrid VMC and the hybrid ramp yield the best results, but the ramps surpass all other methods if the number of realizations is limited.

Finally, for $U = 12$, VMC alone is unable to converge to the ground state as it gets stuck in the atomic limit solution where no doublons are present at half filling. While the rest of the methods are capable of producing one or two converged realizations [see inset in Fig. 2(a)], the majority of the realizations exhibit an energy difference with respect to the ground-state energy of $\approx 4 \times 10^{-3}$ (results lie outside the range plotted in the inset). This larger energy difference is reflected in larger relative errors in Fig. 2(b) for the open markers. Such a struggle to produce converged realizations for random initial parameters is evinced by the relative errors in energy: while for the best realization this number is $\approx 10^{-6}$, for the average over many realizations, it is roughly three orders of magnitude larger.

To summarize this section, we find that, for all the interaction strengths considered, the best realizations for each method after 10^5 training steps are consistent with each other within error bars. However, any of the ramping methods yields the largest number of runs of the lowest relative errors among the different realizations. This means that ramping the Hubbard parameters can achieve better converged results than VMC or H VMC alone.

An important question for numerical methods is how the amount of computational resources required to obtained converged results scales with system size. In Fig. 3 we demonstrate that, for a fixed number of hidden units and number of samples and training steps, results on $L = 6$-, 8-, 10-, and 20-site chains yield relative errors for the energy that are more or less flat within error bars (except for some slight upward trend that can be observed for $U \leqslant 4$ when increasing the system size). This is indicative that, if the number of realizations is limited, the computational cost to keep the relative error fixed as the system size is increased beyond $L = 20$ will likely scale at most linearly with system size. The observed behavior may be unique to 1D and not hold in higher dimensions, where one can in general expect a functional dependence of n_h on N to keep the error fixed. In Appendix B we present results for the 1D FHM also as functions of the system size and the interaction strength, where we demonstrate that they follow the expected trends as functions of these parameters.

So far, we have benchmarked the RNNs' performance in studying the 1D FHM using various training schemes. It is worth noting that while many one-dimensional problems can be studied using efficient techniques, such as matrix product states (MPSs) or the DMRG, RNNs present unique advantages over these approaches. For instance, dispersion relations are not straightforward to calculate with MPS but can be calculated using RNNs, as demonstrated in Ref. [41]. To accomplish this, the model is trained to represent the ground state initially and then a constraint in the loss function is activated, forcing the system to reach a higher-energy state with the corresponding target momentum.

B. Two-dimensional square lattice

Simulations of the FHM, to obtain even the basic properties, are the most challenging in dimensions higher than one and away from half filling. For these reasons, we also explore the capabilities of RNNs with different training schemes in two-dimensional systems, 3×2, and 4×4 for $U = 8$ at half filling. In Fig. 4 we present results for the 3×2 system using the same methods and architecture details as was done for the 1D system. In the 2D case, the benefits of using the ramps are even more evident. For the results presented in Fig. 4, the ramping methods achieve relative errors that are one or two orders of magnitude smaller than the hybrid VMC for all observables, thus highlighting the benefits of our physics-driven method.

Further inspection in Fig. 5 shows that, in 1D, the system is able to learn the balance between spin species ($\langle \frac{1}{N} \sum_i S_z^i \rangle = 0$) for all methods considered. Note that this symmetry is not enforced during the training. In particular, as can be seen in the left panels of Fig. 5 for $L = 10$, we observe that the pretraining and the ramping methods rapidly help the model to learn the correct spin populations. On the other hand, the VMC oscillates and struggles to figure that out, until eventually it converges to correct values. However, in 2D, the oscillations in the VMC method are uncontrolled and lead to a polarized sample, as can be seen in the right panels of Fig. 5 for an

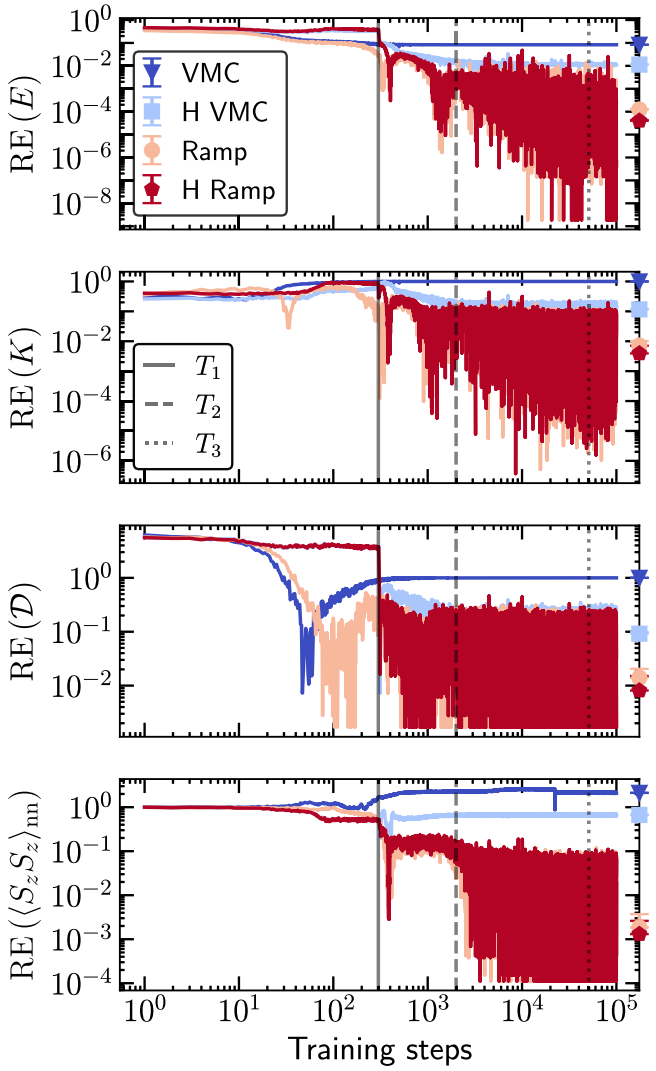

FIG. 4. Relative errors for the half filled FHM as a function of training steps for different methods on a 3×2 system at $U = 8$. Rows correspond to the relative errors in energy, kinetic energy, double occupancy, and nearest-neighbor spin-spin correlation function. Results are presented for the best random seed for all methods. Light vertical lines at $T_1 = 300$ (solid), $T_2 = 2000$ (dashed), and $T_3 = 51\,000$ (dotted) training steps corresponds to pretraining and ramping stages. Markers on the right axis correspond to averages of the observables for the last 100 training steps and error bars correspond to the s.e.m.

$N = 2 \times 3$ system. What is even more surprising is that, although the pretraining method favors spin balanced mixtures, when VMC is turned on after pretraining, the model immediately tends to spin polarization. Clearly the ramping method has a major advantage as it yields the best convergence towards exact results and ensures spin-balanced mixtures.

We further evaluate the performance of the RNN using the ramping training scheme for the 4×4 square lattice in Fig. 6. Specifically, we explore how the energy and the relative error in energy depend on the number of hidden units for the Ramp method. This is motivated by the fact that using $n_h = 100$, the ramping method achieves a relative error in the energy of 1.2×10^{-4} for the 3×2 system, but this error grows to 3.3×10^{-3} for the 4×4 system. This worsening in the relative error as the system size increases emphasizes the significance of establishing the trend with n_h, since we know the method's computational requirements scale linearly in n_h.

In Fig. 6 we observe that the energy decreases with $1/n_h$, and up to $n_h = 300$, its behavior seems to be well described by a linear fit. As n_h increases, the relative error in energy decreases and reaches 7.2×10^{-4} for the extrapolation in the limit $n_h \to \infty$. These extrapolated results are comparable to those presented in Ref. [48], where the authors benchmark their relative errors against auxiliary field QMC from Ref. [66], and illustrate the power and viability of the physics-based training scheme proposed here.

Finally, increasing the number of hidden units above 300 might change the scaling of E with n_h from linear to a power law, further reducing the relative error. However, these runs surpass our current computational capabilities. Lastly, it is worth mentioning that the functional dependence of n_h on N to keep the error fixed as a function of N for the FHM in 2D remains an open question, which will be explored in future studies.

C. Hatano-Nelson-Hubbard model

In addition to the FHM in 1D and 2D, we also explore the one-dimensional Hatano-Nelson-Hubbard model (HNHM). We study this model to evaluate the applicability, power, and versatility of the method in tackling a wide range of models that do not lend themselves to traditional numerical treatments. In particular, we focus on the HNHM because (1) non-Hermitian Hamiltonians are used in the study of open quantum systems [67], (2) display important connections to topological materials [68–70], and (3) pose difficulties for solving with established numerical methods, despite recent efforts that have been made to explore these type of models with DMRG [65,71].

Here, we focus on the HNHM with open boundary conditions. In this limit, after a gauge transformation, the model exhibits a real spectra [68], and therefore we expect the RNN architecture to be able to accurately obtain the ground state. This is because, in current VMC implementations, the energy (or loss function) is real. Nevertheless, we find that despite the spectra being real, for sufficiently large value of $|g/t|$, the RNN fails to converge to the ground state, as shown in Fig. 7 (more below).

In Fig. 7, we present results for an eight-site chain for $U/t = 2$ at $\mu = U/2$ as a function of the tunneling anisotropy $|g/t|$. RNN results are obtained by performing the autoregressive sampling from left to right for two cases: (i) for $g > 0$ (blue circles), which means the left-to-right tunneling is favored and we label it as $+g$, and (ii) for $g < 0$ (red squares), in which the right-to-left tunneling is now favored and we label it as $-g$. In Fig. 7 we also compare the RNN results against ED (green diamonds) for $\mathcal{D}$, n, and $\langle S_z S_z \rangle_{\rm nn}$.

Due to the unequal tunneling rates in the HNHM, the trends displayed by the left (K_L) and right (K_R) kinetic energies in Fig. 7 as a function of $|g/t|$ are expected. At $g = 0$, $K_L = K_R$. As $|g/t|$ increases towards the $g = t$ limit, the magnitude of the unfavored kinetic energy decreases, in a linear fashion, toward zero. On the other hand, as $|g/t|$ increases, the magnitude

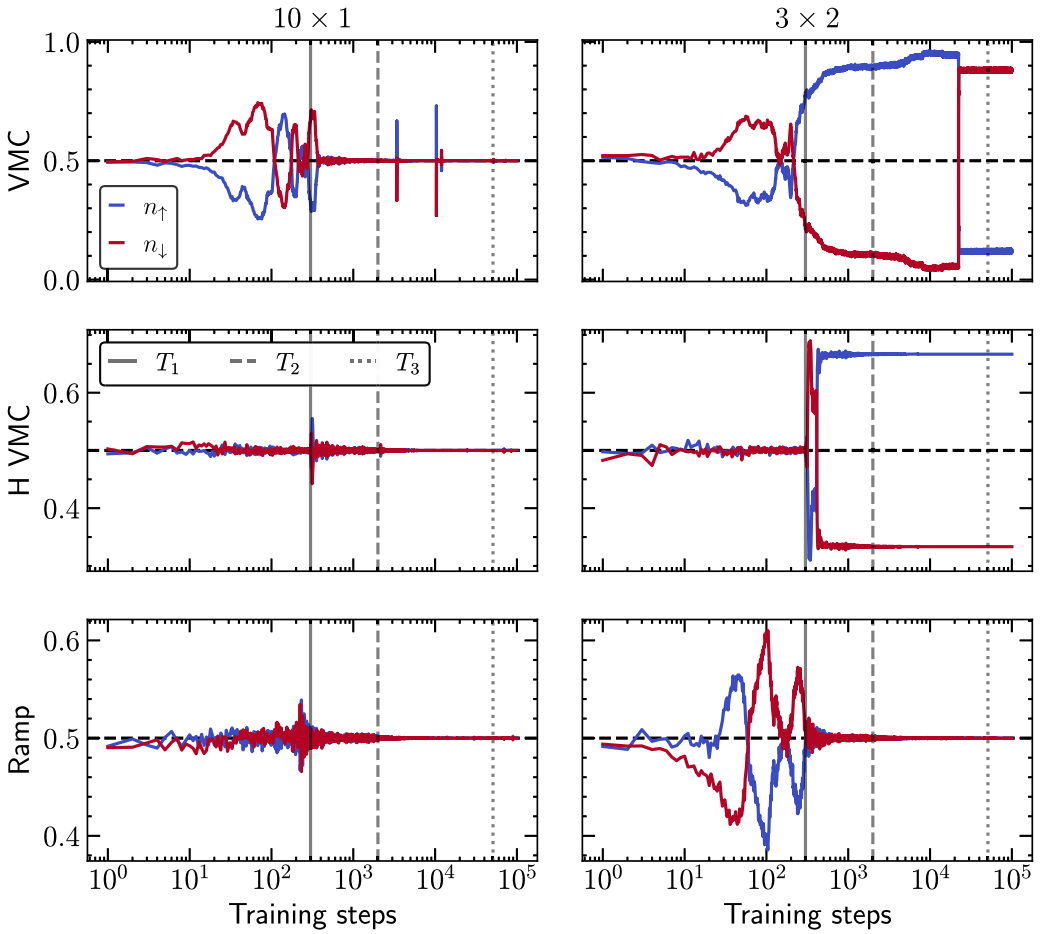

FIG. 5. Spin populations of the half filled FHM as a function of training steps for different methods on a 10-site chain and a 3 × 2 system at $U = 8$. Rows correspond to three different methods used. Results are presented for the best random seed for all methods and system sizes. Light vertical lines are the same as in Fig. 4.

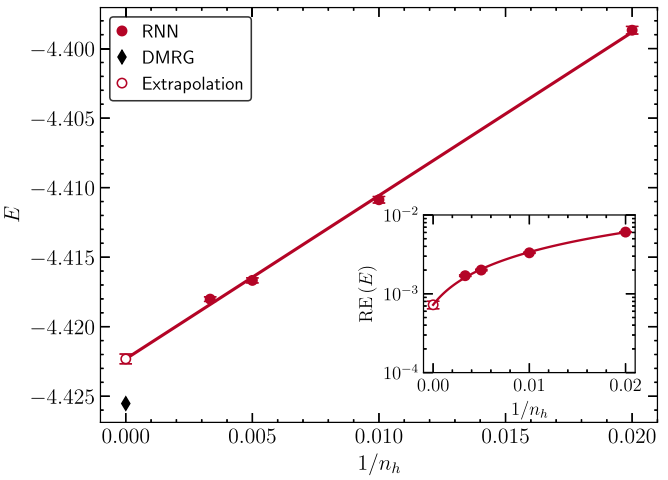

FIG. 6. Energy for the half filled FHM as a function of the inverse of the number of hidden units on a 4 × 4 system at $U = 8$. Red solid circles are RNN results, the black diamond is the DMRG result, and the solid line corresponds to a linear fit $E = E_0 + m/n_h$, where $E_0 = -4.4223 \pm 0.0004$, and $m = 1.17 \pm 0.03$. The open red circle corresponds to the extrapolation to $n_h \to \infty$. Inset presents the relative error of the energy, which for the extrapolated case corresponds to 7.2×10^{-4}.

of the favored kinetic energy increases. Although the results for the favored K_x suggest a linear increase in magnitude, followed by an upturn and a decrease around $|g/t| = 0.6$, results for $|g/t| > 0.5$ are likely unconverged, as we discuss below.

In Fig. 7 comparisons against ED (green diamonds) illustrate that convergence is only achieved for $|g/t| \leqslant 0.5$. For larger $|g/t|$, the two realizations differ significantly: In the case of $g > 0$ the system adds a particle into the chain and favors the formation of double occupancies, while in the case of $g < 0$ it prefers to remove a particle from the array and disfavors the formation of double occupancies. Furthermore, the densities for the $g > 0$ and $g < 0$ curves behave as if they are the particle-hole transformation of each other. This behavior exposes that the relative direction of the favored tunneling rate with respect to the direction of the autoregressive sampling in the RNN plays an important role in its convergence, and that the current sampling scheme alone is not capable of recovering the PHS completely.

Motivated by these results, we then averaged the $g > 0$ and $g < 0$ results (gray pentagons in Fig. 7). Although these are in good agreement with the exact results for the local observables ($\mathcal{D}$ and n) for most values of $|g/t|$, they do not agree for the spin-correlation function for $|g/t| > 0.5$, further reflecting the lack of convergence of the individual

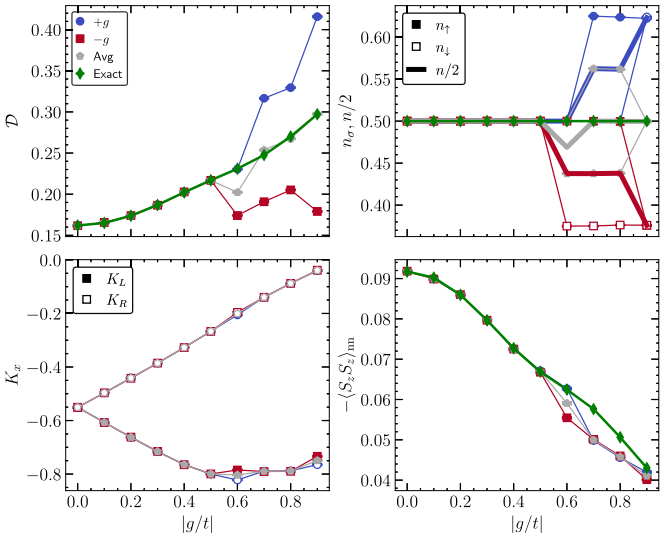

FIG. 7. Observables of the Hatano-Nelson-Hubbard model at $U/t = 2$ and $\mu = U/2$ for the Ramp method as a function of $|g/t|$ on an eight-site chain. Results are presented for RNNs where the autoregressive sampling is performed from left to right and the favored tunneling rate is to the right (blue circles) and left (red squares). Gray pentagons correspond to the average of these two independent runs, and green diamonds correspond to exact results with ED. Results are presented for the best realization, where results are obtained by averaging the observables for the last 100 training steps and error bars correspond to the s.e.m.

runs. These findings suggest a need for designing a different autoregressive sampling scheme. Therefore, we examined a new scheme in which we reverse the direction of the autoregressive sampling after each training step. However, we observed that the RNN convergence did not improve using this scheme (not shown). The behavior to add or remove a particle persists.

A similar calculation using an autoregressive version of a transformer quantum state (TQS) yields similar results. The architecture is similar to the one used in Ref. [34] for spin models, but adapted to the larger local Hilbert space of the HNHM. Figure 8 shows the relative errors obtained with the TQS using different numbers of layers n_l and embedding dimensions n_h (top) and the average filling (bottom) at the end of the training. As observed for the RNN, the accuracy decreases for $g/t \neq 0$. Furthermore, the average filling depends on the sign of g: It is overestimated for $g > 0$ and underestimated for $g < 0$, as can be seen in Fig. 8 (bottom) for $g/t = \pm 0.8$.

Finally, we find that the RNN's behavior of moving away from half filling occurs for other values of U/t and for different system sizes too (see Appendix C). In particular we observe that the "critical" g/t at which the system decides to add or remove a particle shifts to lower values as the system size increases and U/t decreases. Such behavior is reminiscent of nonergodicity issues in determinant QMC (DQMC), in which the method *sticks* at incorrect densities at large U, and the incorrect density corresponds to adding or subtracting integer number of particles [72]. Additionally, these issues aggravate for larger systems (signatures of *sticking* are present at higher temperatures for larger systems). For the FHM the relevant parameter is U/t. For the HNHM, we have two relevant ratios $U/(t \pm g)$. While in the isotropic limit $g = 0$, $U/t = 2$ corresponds to weak coupling, for $g = 0.6$, $U/(t - g) = 5$, and for $g = 0.9$, $U/(t - g) = 20$. These findings call for further investigation into the design of RNN architectures capable of addressing non-Hermitian Hamiltonians, which will be a subject of our future studies.

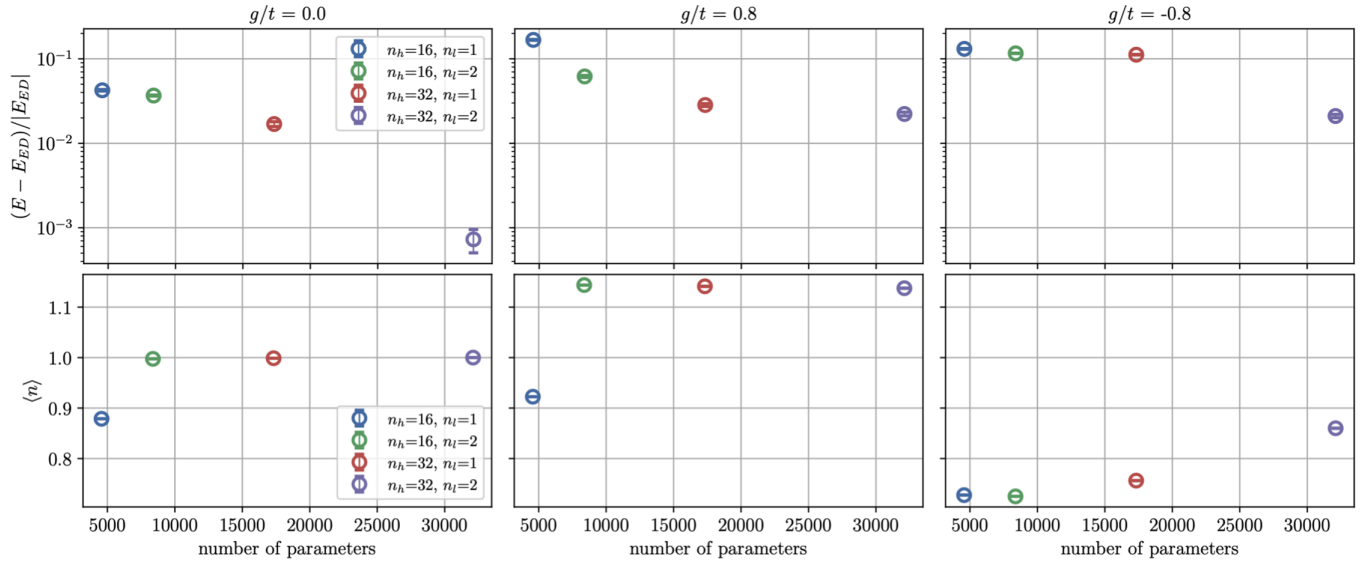

FIG. 8. Results obtained using a autoregressive transformer quantum state with different numbers of layers n_l and embedding dimensions n_h. We show the relative errors (top) and the average filling $\langle n \rangle$ (bottom) at the end of the training for $g/t = 0.0, \pm 0.8$ (left to right). In all calculations, eight attention heads are used.

IV. CONCLUSIONS

In this work we utilized RNNs as a variational ansatz to access the ground state of many-body Hamiltonians. We evaluated their applicability, power, and versatility using different training schemes for the FHM in 1D and 2D and in the 1D HNHM.

We introduced a physically motivated method for enhancing VMC simulations that is independent of having access to experimental or numerical projective measurements. Our method is based on ramping the tunneling rate from a larger t to the desired final value during the training of the RNN.

We first benchmarked our results for the FHM against ED and DMRG on 1D chains with open boundary conditions. In this regime, (1) we observed that as system size increases, the computational cost to keep the relative error fixed will likely scale at most linearly with system size, and (2) we demonstrated that our proposed training scheme, either alone or in conjunction with pretraining, produces results that are generally better than VMC with pretraining with projective measurements.

We then applied this training scheme to the FHM in the 2D square lattice. We found that our proposed method performs significantly better than the hybrid optimization technique, achieving relative errors that are one or two orders of magnitude smaller for all observables considered in this study. Moreover, for the largest 2D system examined (4×4) we obtained relative errors in the energy that are consistent with those obtained with NQS simulations with constrained hidden states [48].

Finally, our application of the method to the HNHM illustrated that further considerations need to be taken into account for the study of non-Hermitian physics with RNN. These point to interesting future studies, e.g., the use of stochastic reconfiguration for the optimization of the RNN's variational parameters [73] or the use of symmetries [43].

In addition to exploring non-Hermitian Hamiltonians, an immediate avenue for application of our method is to understand the effects of doping Mott insulators and magnetically ordered phases, which is one of the principal objectives in strongly correlated matter. In particular, the approach presented here provides a useful starting point for the exploration of the doped FHM in 1D and 2D with RNN- and TQS-based VMCs in which one may have to ramp multiple model parameters simultaneously, including μ, during the training to achieve a desired filling.

Finally, we also expect the method to perform well for the attractive FHM too since it is equivalent to the repulsive model at half filling due to particle-hole symmetry. Studies involving the former model may provide useful insight into the convergence and scaling properties of autoregressive neural networks on large lattices in two dimensions, where QMC methods do not exhibit a sign problem away from half filling and accurate calculations of very large system sizes at low temperature are available for comparison [74].

ACKNOWLEDGMENTS

E.I.-G.-P., R.T.S., and E.K. are supported by the grant DE-SC0022311, funded by the U.S. Department of Energy, Office of Science. Computing resources were supported by the Spartan high-performance computing facility at San José State University supported by the NSF under Grant No. OAC-1626645. A.B. and H.L. acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2111—390814868. H.L. acknowledges support by the International Max Planck Research School for Quantum Science and Technology (IMPRS-QST). R.G.M. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Perimeter Institute for Theoretical Physics. Research at the Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Economic Development, Job Creation and Trade.

APPENDIX A: DETAILS OF RAMPS

The tunneling ramps used in this study are defined as follows:

$$t_{\mathrm{Ramp}}(n) = \begin{cases} t + (t_i - t)e^{-Wn\Delta(n)} & n < 51\,000 \\ t & n \geqslant 51\,000, \end{cases} \tag{A1}$$

$$t_{\mathrm{H\,Ramp}}(n) = \begin{cases} t & n \leqslant 300 \\ t + (t_i - t)e^{-Wn\Delta(n)} & 300 < n < 51\,000 \\ t & n \geqslant 51\,000, \end{cases} \tag{A2}$$

where n is the training step number, and

$$\Delta(n) = \begin{cases} 5 \times 10^{-4}/t & n \leqslant 2000 \\ 2 \times 10^{-5}/t & n > 2000. \end{cases} \tag{A3}$$

We set $t_i = 2t$ and $W = 4t$. The ramps are illustrated in Fig. 9 for simplicity and visualization. We also observed that different ramp parameters do not affect convergence.

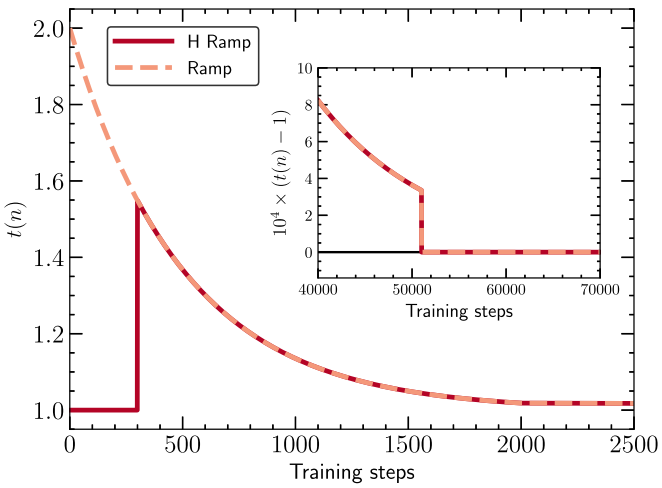

FIG. 9. Visualization of the tunneling ramps used in the study as a function of the number of training steps. The inset illustrates when the exponential ramp is quenched to lower the tunneling rate to the desired value of t.

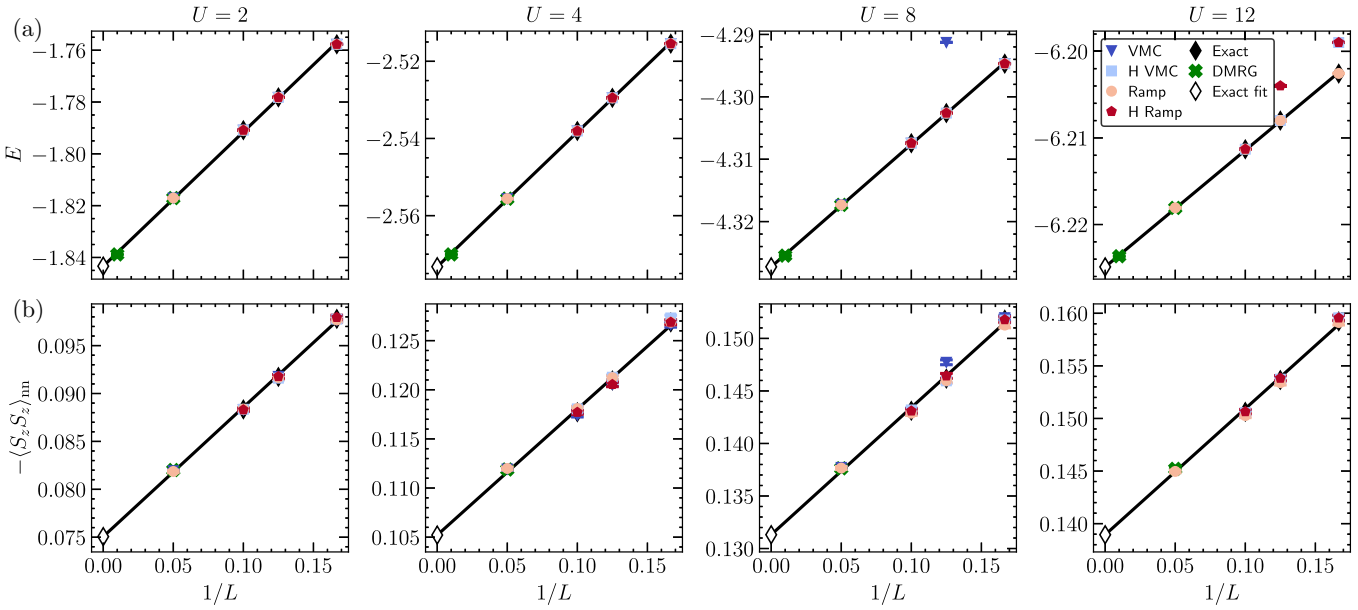

FIG. 10. (a) Energy and (b) nearest-neighbor spin-spin correlation for the halffilled FHM for different methods as a function of system size. Columns correspond to $U = 2, 4, 8, 12$. Results are presented for the best realization, where results are obtained by averaging the observables for the last 100 training steps and error bars correspond to the s.e.m. (except for $U = 12$ for the VMC method, where results lie outside the range plotted). For $L = 20$, results from DMRG are also presented for comparison (green crosses). DMRG results are also presented for $L = 100$ for the ground-state energy. Black solid diamonds correspond to results from ED. The black lines are linear fits to the ED and DMRG data, and the black open diamonds are the thermodynamic limit extrapolations of the linear fits.

APPENDIX B: FURTHER FERMI-HUBBARD MODEL DETAILS IN ONE DIMENSION

For larger system sizes where ED is not feasible we compare against DMRG and the extrapolated results from ED. These are presented in Fig. 10, where we show E and $\langle S_z S_z \rangle_{\rm nn}$ as a function of $1/L$. The results for the energy and the correlation function from the RNN are consistent with ED and DMRG.

In addition, in Fig. 11 we present results as a function of U. The behaviors are as expected, as U increases, the ground-state energy grows in magnitude, and both the kinetic energy and the number of double occupancies decrease in magnitude. On the contrary, the antiferromagnetic nearest-neighbor spin-correlation function increases as U increases before the expected decline at larger U. In all cases, for converged results, there is good agreement with the exact results.

APPENDIX C: MORE RESULTS FOR THE HATANO-NELSON-HUBBARD MODEL

In Figs. 12 and 13 we present results for the HNHM as for different system sizes and interaction strengths, respectively. Figure 12 illustrates that, at fixed interaction strength, as the

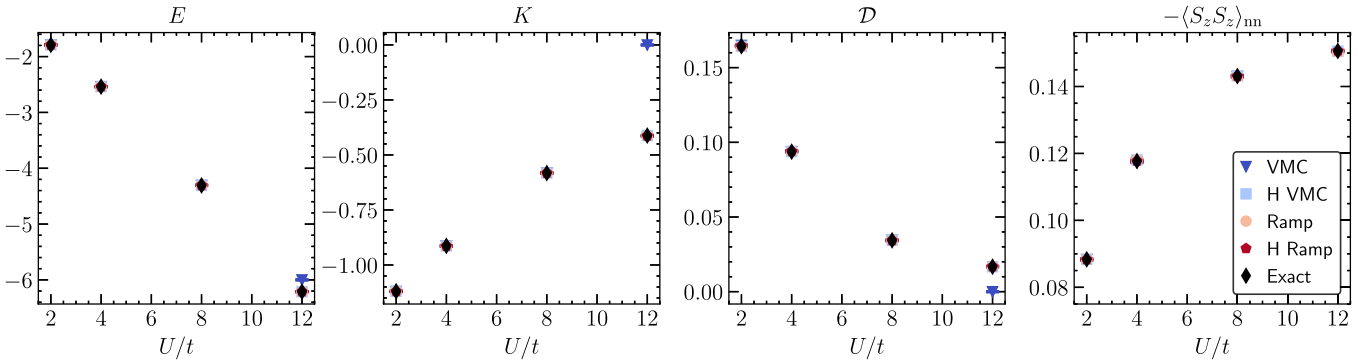

FIG. 11. E, K, $\mathcal{D}$, and $\langle S_z S_z \rangle_{\rm nn}$ for the half filled FHM for different methods as a function of the interaction strength U/t on a 10-site chain. Results are presented for the best realization, where results are obtained by averaging the observables for the last 100 training steps and error bars correspond to the s.e.m. For the spin-spin correlation function, the VMC method marker at $U/t = 12$ is not presented since results lie outside the range plotted. Black solid diamonds correspond to results from ED.

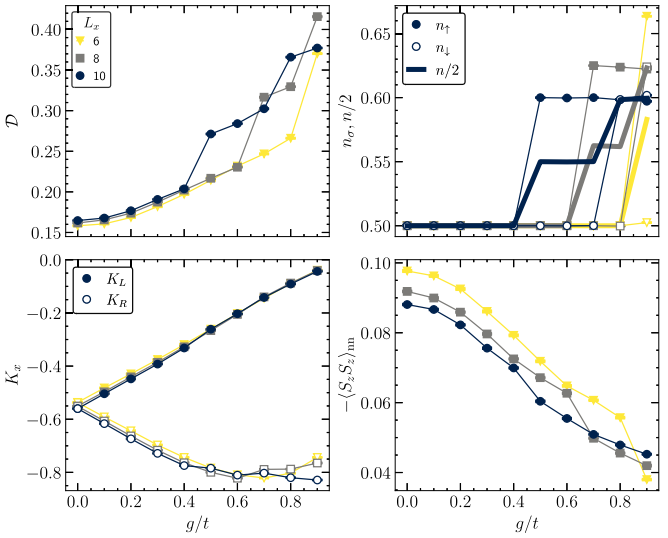

FIG. 12. Observables of the Hatano-Nelson-Hubbard model at $U = 2t$ and $\mu = U/2$ for the Ramp method as a function of g/t for different sizes L_x. Results are presented for the best realization, where results are obtained by averaging the observables for the last 100 training steps, and error bars correspond to the s.e.m.

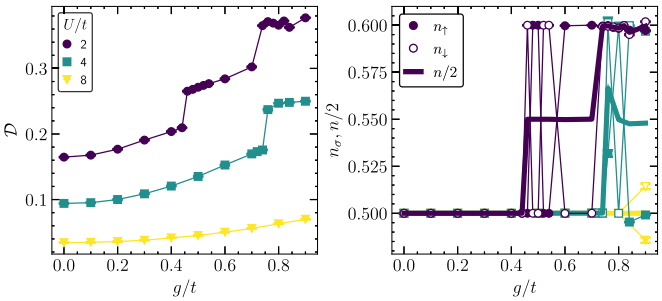

FIG. 13. Observables of the Hatano-Nelson-Hubbard model at $\mu = U/2$ for the Ramp method as a function of g/t and different values of U/t in a $L = 10$-site chain. Results are presented for the best realization, where results are obtained by averaging the observables for the last 100 training steps, and error bars correspond to the s.e.m.

system size increases, the "critical" g/t at which the system moves away from half filling shifts to lower values. Furthermore, Fig. 13 shows that the value of the "critical" g/t also shifts to lower values as U/t decreases.

[1] M. Rasetti, *The Hubbard Model: Recent Results* (World Scientific, Signapore, 1991), Vol. 7.

[2] F. Gebhard and F. Gebhard, *Metal-Insulator Transitions* (Springer, Berlin, 1997).

[3] P. Fazekas, *Lecture Notes on Electron Correlation and Magnetism* (World Scientific, Singapore, 1999), Vol. 5.

[4] D. P. Arovas, E. Berg, S. Kivelson, and S. Raghu, The Hubbard model, Annu. Rev. Condens. Matter Phys. **13**, 239 (2022).

[5] M. Qin, T. Schäfer, S. Andergassen, P. Corboz, and E. Gull, The Hubbard model: A computational perspective, Annu. Rev. Condens. Matter Phys. **13**, 275 (2022).

[6] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[7] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).

[8] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, Restricted Boltzmann machines in quantum physics, Nat. Phys. **15**, 887 (2019).

[9] Z.-A. Jia, B. Yi, R. Zhai, Y.-C. Wu, G.-C. Guo, and G.-P. Guo, Quantum neural network states: A brief review of methods and applications, Adv. Quantum Technol. **2**, 1800077 (2019).

[10] J. Carrasquilla, Machine learning for quantum matter, Adv. Phys.: X **5**, 1797528 (2020).

[11] J. Carrasquilla and G. Torlai, How to use neural networks to investigate quantum many-body physics, PRX Quantum **2**, 040201 (2021).

[12] M. Bukov, M. Schmitt, and M. Dupont, Learning the ground state of a non-stoquastic quantum Hamiltonian in a rugged neural network landscape, SciPost Phys. **10**, 147 (2021).

[13] M. Medvidović and J. R. Moreno, Neural-network quantum states for many-body physics, Eur. Phys. J. Plus **139**, 631 (2024).

[14] H. Lange, A. V. de Walle, A. Abedinnia, and A. Bohrdt, From architectures to applications: A review of neural quantum states, Quantum Sci. Technol. **9**, 040501 (2024).

[15] M. Reh, M. Schmitt, and M. Gärttner, Optimizing design choices for neural quantum states, Phys. Rev. B **107**, 195115 (2023).

[16] N. Yoshioka and R. Hamazaki, Constructing neural stationary states for open quantum many-body systems, Phys. Rev. B **99**, 214306 (2019).

[17] O. Sharir, A. Shashua, and G. Carleo, Neural tensor contractions and the expressive power of deep neural quantum states, Phys. Rev. B **106**, 205136 (2022).

[18] D.-L. Deng, X. Li, and S. Das Sarma, Quantum entanglement in neural network states, Phys. Rev. X **7**, 021021 (2017).

[19] X. Gao and L.-M. Duan, Efficient representation of quantum many-body states with deep neural networks, Nat. Commun. **8**, 662 (2017).

[20] Z. Denis, A. Sinibaldi, and G. Carleo, Comment on "Can neural quantum states learn volume-law ground states?", arXiv:2309.11534.

[21] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Quantum entanglement in deep learning architectures, Phys. Rev. Lett. **122**, 065301 (2019).

[22] S. R. White, Density matrix formulation for quantum renormalization groups, Phys. Rev. Lett. **69**, 2863 (1992).

[23] U. Schollwöck, The density-matrix renormalization group, Rev. Mod. Phys. **77**, 259 (2005).

[24] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Monte Carlo calculations of coupled boson-fermion systems. I, Phys. Rev. D **24**, 2278 (1981).

[25] S. Sorella, S. Baroni, R. Car, and M. Parrinello, A novel technique for the simulation of interacting fermion systems, Europhys. Lett. **8**, 663 (1989).

[26] E. Y. Loh, J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar, Sign problem in the numerical simulation of many-electron systems, Phys. Rev. B **41**, 9301 (1990).

[27] M. Troyer and U.-J. Wiese, Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations, Phys. Rev. Lett. **94**, 170201 (2005).

[28] V. I. Iglovikov, E. Khatami, and R. T. Scalettar, Geometry dependence of the sign problem in quantum Monte Carlo simulations, Phys. Rev. B **92**, 045110 (2015).

[29] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, Recurrent neural network wave functions, Phys. Rev. Res. **2**, 023358 (2020).

[30] D. Luo, Z. Chen, J. Carrasquilla, and B. K. Clark, Autoregressive neural network for simulating open quantum systems via a probabilistic formulation, Phys. Rev. Lett. **128**, 090501 (2022).

[31] D. Luo, Z. Chen, K. Hu, Z. Zhao, V. M. Hur, and B. K. Clark, Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models, Phys. Rev. Res. **5**, 013216 (2023).

[32] Y.-H. Zhang and M. Di Ventra, Transformer quantum state: A multipurpose model for quantum many-body problems, Phys. Rev. B **107**, 075147 (2023).

[33] K. Sprague and S. Czischek, Variational Monte Carlo with large patched transformers, Commun. Phys. **7**, 90 (2024).

[34] H. Lange, G. Bornet, G. Emperauger, C. Chen, T. Lahaye, S. Kienle, A. Browaeys, and A. Bohrdt, Transformer neural networks and quantum simulators: A hybrid approach for simulating strongly correlated systems, arXiv:2406.00091.

[35] D. Fitzek, Y. H. Teoh, H. P. Fung, G. A. Dagnew, E. Merali, M. S. Moss, B. MacLellan, and R. G. Melko, RydbergGPT, arXiv:2405.21052.

[36] L. L. Viteritti, R. Rende, and F. Becca, Transformer variational wave functions for frustrated quantum spin systems, Phys. Rev. Lett. **130**, 236401 (2023).

[37] R. Rende, L. L. Viteritti, L. Bardone, F. Becca, and S. Goldt, A simple linear algebra identity to optimize large-scale neural network quantum states, Commun. Phys. **7**, 260 (2024).

[38] S. Czischek, M. S. Moss, M. Radzihovsky, E. Merali, and R. G. Melko, Data-enhanced variational Monte Carlo simulations for Rydberg atom arrays, Phys. Rev. B **105**, 205108 (2022).

[39] M. Hibat-Allah, R. G. Melko, and J. Carrasquilla, Investigating topological order using recurrent neural networks, Phys. Rev. B **108**, 075152 (2023).

[40] M. S. Moss, S. Ebadi, T. T. Wang, G. Semeghini, A. Bohrdt, M. D. Lukin, and R. G. Melko, Enhancing variational Monte Carlo simulations using a programmable quantum simulator, Phys. Rev. A **109**, 032410 (2024).

[41] H. Lange, F. Döschl, J. Carrasquilla, and A. Bohrdt, Neural network approach to quasiparticle dispersions in doped antiferromagnets, Commun. Phys. **7**, 187 (2024).

[42] F. Döschl, F. A. Palm, H. Lange, F. Grusdt, and A. Bohrdt, Neural network quantum states for the interacting Hofstadter model with higher local occupations and long-range interactions, Phys. Rev. B **111**, 045408 (2025).

[43] S. Morawetz, I. J. S. De Vlugt, J. Carrasquilla, and R. G. Melko, U(1)-symmetric recurrent neural networks for quantum state reconstruction, Phys. Rev. A **104**, 012401 (2021).

[44] E. R. Bennewitz, F. Hopfmueller, B. Kulchytskyy, J. Carrasquilla, and P. Ronagh, Neural error mitigation of near-term quantum simulations, Nat. Mach. Intell. **4**, 618 (2022).

[45] J. Stokes, J. R. Moreno, E. A. Pnevmatikakis, and G. Carleo, Phases of two-dimensional spinless lattice fermions with first-quantized deep neural-network quantum states, Phys. Rev. B **102**, 205122 (2020).

[46] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, Phys. Rev. B **96**, 205152 (2017).

[47] Z. Liu and B. K. Clark, Unifying view of fermionic neural network quantum states: From neural network backflow to hidden fermion determinant states, Phys. Rev. B **110**, 115124 (2024).

[48] J. R. Moreno, G. Carleo, A. Georges, and J. Stokes, Fermionic wave functions from neural-network constrained hidden states, Proc. Natl. Acad. Sci. USA **119**, e2122059119 (2022).

[49] D. Luo and B. K. Clark, Backflow transformations via neural networks for quantum many-body wave functions, Phys. Rev. Lett. **122**, 226401 (2019).

[50] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic Schrödinger equation, Nat. Chem. **12**, 891 (2020).

[51] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, *Ab initio* solution of the many-electron Schrödinger equation with deep neural networks, Phys. Rev. Res. **2**, 033429 (2020).

[52] J. Kim, G. Pescia, B. Fore, J. Nys, G. Carleo, S. Gandolfi, M. Hjorth-Jensen, and A. Lovato, Neural-network quantum states for ultra-cold Fermi gases, Commun. Phys. **7**, 148 (2024).

[53] I. Romero, J. Nys, and G. Carleo, Spectroscopy of two-dimensional interacting lattice electrons using symmetry-aware neural backflow transformations, arXiv:2406.09077.

[54] S. Humeniuk, Y. Wan, and L. Wang, Autoregressive neural Slater-Jastrow ansatz for variational Monte Carlo simulation, SciPost Phys. **14**, 171 (2023).

[55] K. Inui, Y. Kato, and Y. Motome, Determinant-free fermionic wave function using feed-forward neural networks, Phys. Rev. Res. **3**, 043126 (2021).

[56] N. Yoshioka, W. Mizukami, and F. Nori, Solving quasiparticle band spectra of real solids using neural-network quantum states, Commun. Phys. **4**, 106 (2021).

[57] M. Bortone, Y. Rath, and G. H. Booth, Impact of conditional modelling for a universal autoregressive quantum state, Quantum **8**, 1245 (2024).

[58] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Symmetries and many-body excitations with neural-network quantum states, Phys. Rev. Lett. **121**, 167204 (2018).

[59] Y. Nomura, Machine learning quantum states—extensions to fermion–boson coupled systems and excited-state calculations, J. Phys. Soc. Jpn. **89**, 054706 (2020).

[60] L. L. Viteritti, F. Ferrari, and F. Becca, Accuracy of restricted Boltzmann machines for the one-dimensional J_1-J_2 Heisenberg model, SciPost Phys. **12**, 166 (2022).

[61] M. Hibat-Allah, E. M. Inack, R. Wiersema, R. G. Melko, and J. Carrasquilla, Variational neural annealing, Nat. Mach. Intell. **3**, 952 (2021).

[62] C. S. Chiu, G. Ji, A. Mazurenko, D. Greif, and M. Greiner, Quantum state engineering of a Hubbard sys-

tem with ultracold fermions, Phys. Rev. Lett. **120**, 243201 (2018).

[63] Z. Z. Yan, B. M. Spar, M. L. Prichard, S. Chi, H.-T. Wei, E. Ibarra-García-Padilla, K. R. A. Hazzard, and W. S. Bakr, Two-dimensional programmable tweezer arrays of fermions, Phys. Rev. Lett. **129**, 123201 (2022).

[64] As an example, consider the two-particle two-site calculation, where one finds that the ground state of the model is the singlet state with a small admixture of doublons. In the $t/U \ll 1$ limit, the ground-state energy is $\approx -4t^2/U$, and the ground state is given by

$$|\psi\rangle \approx \mathcal{N}\left[\frac{2t}{U}(|0,\uparrow\downarrow\rangle + |\uparrow\downarrow,0\rangle) + (|\uparrow,\downarrow\rangle + |\downarrow,\uparrow\rangle)\right],$$

where $\mathcal{N}$ is a normalization factor. Note that although the coefficient associated with the states with double occupancies is small, it does not vanish.

[65] M. Fishman, S. R. White, and E. M. Stoudenmire, The ITensor software library for tensor network calculations, SciPost Phys. Codebases, 4 (2022).

[66] M. Qin, H. Shi, and S. Zhang, Benchmark study of the two-dimensional Hubbard model with auxiliary-field quantum Monte Carlo method, Phys. Rev. B **94**, 085103 (2016).

[67] I. Rotter, A non-Hermitian Hamilton operator and the physics of open quantum systems, J. Phys. A: Math. Theor. **42**, 153001 (2009).

[68] N. Okuma and M. Sato, Non-Hermitian topological phenomena: A review, Annu. Rev. Condens. Matter Phys. **14**, 83 (2023).

[69] A. Maddi, Y. Auregan, G. Penelet, V. Pagneux, and V. Achilleos, Exact analog of the Hatano-Nelson model in one-dimensional continuous nonreciprocal systems, Phys. Rev. Res. **6**, L012061 (2024).

[70] T. Orito and K.-I. Imura, Entanglement dynamics in the many-body Hatano-Nelson model, Phys. Rev. B **108**, 214308 (2023).

[71] P. Zhong, W. Pan, H. Lin, X. Wang, and S. Hu, Density-matrix renormalization group algorithm for non-Hermitian systems, arXiv:2401.15000.

[72] R. T. Scalettar, R. M. Noack, and R. R. P. Singh, Ergodicity at large couplings with the determinant Monte Carlo algorithm, Phys. Rev. B **44**, 10502 (1991).

[73] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, Cambridge, 2017).

[74] R. T. Scalettar, E. Y. Loh, J. E. Gubernatis, A. Moreo, S. R. White, D. J. Scalapino, R. L. Sugar, and E. Dagotto, Phase diagram of the two-dimensional negative-U Hubbard model, Phys. Rev. Lett. **62**, 1407 (1989).