Gaussian Process States: A Data-Driven Representation of Quantum Many-Body Physics

Aldo Glielmo[1,2,*] Yannic Rath[1,*] Gábor Csányi,[3] Alessandro De Vita,[1] and George H. Booth[1,†]

[1]*Department of Physics, King's College London, Strand, London WC2R 2LS, United Kingdom*
[2]*Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*
[3]*Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

We present a novel, nonparametric form for compactly representing entangled many-body quantum states, which we call a "Gaussian process state." In contrast to other approaches, we define this state explicitly in terms of a configurational data set, with the probability amplitudes statistically inferred from this data according to Bayesian statistics. In this way, the nonlocal physical correlated features of the state can be analytically resummed, allowing for exponential complexity to underpin the ansatz, but efficiently represented in a small data set. The state is found to be highly compact, systematically improvable, and efficient to sample, representing a large number of known variational states within its span. It is also proven to be a "universal approximator" for quantum states, able to capture any entangled many-body state with increasing data-set size. We develop two numerical approaches which can learn this form directly—a fragmentation approach and direct variational optimization—and apply these schemes to the fermionic Hubbard model. We find competitive or superior descriptions of correlated quantum problems compared to existing state-of-the-art variational ansatzes, as well as other numerical methods.

Subject Areas: Computational Physics,
Condensed Matter Physics,
Strongly Correlated Materials

I. INTRODUCTION

Representing entangled quantum many-body states efficiently and compactly is a major challenge, with the need for tractable approaches underpinning a diverse set of fields including quantum computation, electronic or nuclear structure, and quantum chemistry. Advances in these fields are often punctuated by the discovery and exploitation of efficient wave-function forms for capturing certain physics with polynomial-scaling resources. These nonlinear parametrizations avoid the famed "exponential wall" in the complexity of the general solution, including forms (and physical features) such as the Gutzwiller wave function [1] (suppression of local double occupancy), tensor network [2–4] (low entanglement), Laughlin [5] (fractionalized excitations), or coupled-cluster states [6] (low-rank reducible correlations). However, it is not just the quantum many-body problem that encounters a curse of dimensionality. Increasingly, computational fields are turning to statistical inference and machine learning approaches to

circumvent this bottleneck, and to allow a more general framework to capture arbitrary functional forms in high-dimensional spaces. Along this line, in this work, we espouse a new philosophy, whereby wave functions are not defined and constrained by a fixed parametrization, but rather, they are defined explicitly in terms of "data," with their parameters considered statistically as random variables. This philosophy leads to the description of entangled wave functions in a systematically improvable and nonparametric fashion. The expressiveness, accuracy, and compactness of the resulting state is found to be competitive or surpass that of state-of-the-art wave-function parametrizations, encoding many important correlated states in a simple form.

The data used to define this wave function consist of a small "training" sample of underlying many-body configurations and their (approximate) probability amplitudes. The question we wish to address first is how to optimally use this training set to infer the wave-function amplitudes on any other configuration in the exponentially large Hilbert space, in a statistically rigorous fashion. This inference allows us to define a complete wave function spanning all configurations, which can then be efficiently sampled to extract any desired property of the system. In this way, the state is explicitly parametrized by the data set, rather than this data being used to optimize a fixed functional form. It is clear that this data-driven approach will not be possible

*These two authors contributed equally to this work.
†george.booth@kcl.ac.uk

for a random state where no structure exists in the form of the amplitudes, but it is known that physical many-body states exhibit much structure that can be effectively learnt in this approach, emerging from the underlying principles encoded in the Hamiltonian. After this inferred wave function is defined and its limits discussed, we turn the idea into a practical and general method for quantum problems and demonstrate its accuracy and improvability in describing full N-body quantum correlations by application to the fermionic Hubbard model of the cuprates and strongly correlated solids. In this work, we consider two sources for the data: configurations from the solution to a small fragment of the full system, and a self-consistent approach to iterative selection and refinement of the data. Furthermore, we show that an automatic selection of the relevant data required to capture the complexity of the many-body correlations allows for a compact description of this state compared to other approaches, with significant potential for a wide range of applications.

II. GAUSSIAN PROCESS REGRESSION FOR WAVE FUNCTIONS

In order to describe the amplitudes of a quantum state, $\Psi(\mathbf{x})$, evaluated for any many-body configuration $\mathbf{x}$ in the underlying Hilbert space, we work within a Bayesian inference framework known as Gaussian process (GP) regression. This approach falls under the umbrella of "kernel" methods, which constitute the alternate paradigm to neural networks in machine learning. The great flexibility and ease of use of GPs and other kernel methods has brought them into prominence in a range of fields, including the interpolation of potential energy surfaces and other high-dimensional functions [7–10]. Advantages over neural networks include a rigorous underlying probablistic model and uncertainty bounds, the ease of incorporation of symmetries and conserved quantities, well-understood regularization procedures, and robust and theoretically justified algorithms [11]. In this nonparametric approach, an unknown target function is learnt from samples at a finite number of input points, allowing for the statistical inference of an estimator for the underlying function. The weights of the model are considered as random variables, which define a distribution over all linear models of the state in a high-dimensional space of "features," defined by the map $\Phi(\mathbf{x})$. The scalar product of two configurations in this feature space defines a kernel function, which can be evaluated between any two configurations, as $k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')$. Estimating the expectation of the Gaussian process then involves evaluating a linear combination of kernel functions $k(\mathbf{x}, \mathbf{x}'_b)$ between the target configuration $\mathbf{x}$ and each configuration $\mathbf{x}'_b$ of the training database.

We introduce a Gaussian Process State (GPS) as the exponential of a GP estimator, where a configurational probability amplitude $\Psi_g(\mathbf{x})$ is given as

$$\Psi_g(\mathbf{x}) = \exp\left(\sum_{b=1}^{N_b} k(\mathbf{x}, \mathbf{x}'_b) w_b\right), \qquad (1)$$

where w_b is the weight parameter of each data configuration defining the ansatz, which can, in general, be a complex number. Many existing parametrized states can be reexpressed in this form. For instance, the Gutzwiller ansatz can be considered via a feature map into a single-dimensional space, given by the number of doubly occupied sites in the configuration. However, the key to circumvent the complexity of a general mapping is to note that the only quantity required for predictions of the configurational amplitudes is the scalar product of two feature vectors, defined by the kernel function. By constructing the kernel function directly, the feature vectors are only ever implicitly defined, allowing for an arbitrary number of features to define the state, and a definition of the model explicitly and solely in terms of the underlying data configurations $\{\mathbf{x}'_b\}$.

This GPS is a "universal approximator," meaning that it can approximate a quantum state with arbitrary accuracy when provided with sufficient data. A GPS can also be recast as a feed-forward neural network with a infinite-dimensional single hidden layer [12,13], connecting the approach to the recently developed neural-network states [14–22]. These states have recently come into prominence, particularly in the form of a "restricted Boltzmann machine" (RBM), which also forms a nonparametric and universal approximator for quantum states, where "features" are effectively learnt variationally through the network architecture [23–25]. The exponential form in Eq. (1) is reminiscent of other wave-function forms, such as the Jastrow [26] or coupled-cluster ansatzes [6]. However, in contrast, the GPS is defined explicitly in terms of N-body configurations, rather than parametrized in terms of low-order features. The exponential form also ensures that the overall state for our additive kernel is appropriately product separable in terms of its local features, and it is size extensive, allowing its properties to appropriately scale to the thermodynamic limit. As a scalar product of the configurational features, the kernel function $k(\mathbf{x}, \mathbf{x}'_b)$ aims to quantify the similarity between any two configurations in the system, and it is clear that the specific choice of kernel function, as well as basis configurations, $\{\mathbf{x}'_b\}$, and weights w_b, is of paramount importance for the expressibility and success of the model.

A. Kernel function

The kernel function defines the set of correlation features for the state, which are optimally modeled from the data, without explicitly applying the corresponding feature map $\Phi(\mathbf{x})$ into this space. With a kernel including all features, we are able to reproduce any state with our ansatz, assuming we have a complete set of exact data points.

More importantly though, we can introduce a set of kernels that gives rise to a systematically improvable ansatz, which can accurately approximate the state using a far smaller set of training configurations. These kernels are obtained via the addition of hyperparameters, which weight the relative importance of fitting some features over others but which still leave an exact limit as the data set increases.

We can write a configuration $\mathbf{x}$ in terms of its local Hilbert space on each spatial degree of freedom, i. As a specific example of fermionic lattice models used later, we can consider four local Fock states for each lattice site, with $\mathbf{x} = x_i$, where $x_i \in \{\cdot, \uparrow, \downarrow, \uparrow\downarrow\}$ (for a spin or qubit model, there would just be two local states, $x_i \in \{\uparrow, \downarrow\}$). The kernel function compares these local occupations between two configurations, extracting multisite correlation features common to both. Thus, the kernel $k(\mathbf{x}, \mathbf{x}')$ is written as a linear combination of products of delta functions $\delta_{x_i x'_j}$, which we define to be 1 if the local occupation x_i is equal to the local occupation x'_j, and 0 otherwise. We do not assume that the dimensions of the two configurations are the same and can therefore compare configurations with different numbers of lattice sites.

The simplest kernel just compares the local occupancies, given by

$$k^{(1)}(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x'_j}, \qquad (2)$$

where L refers to the size of the lattice corresponding to $\mathbf{x}$ (denoting the test amplitude to infer) and x'_j denotes the local occupancy of the training point at a chosen reference site j. The reference site for the training lattice can be chosen arbitrarily for translationally symmetric systems since there always exists a potential training configuration that can describe all local environments of each site. Additional summation over the sites of the training lattice in Eq. (2) is possible, to ensure translational symmetry in the training data. However, this additional summation was not found to improve the results sufficiently, given the additional cost to evaluate, so only the sum over the test lattice is included, ensuring the wave function still retains overall translational symmetry. The superscript "(1)" emphasizes that this kernel extracts features corresponding to the Fock space of local, single-site plaquettes, therefore implicitly mapping each configuration to a four-dimensional space corresponding to its number of each local state. Therefore, this mapping extracts the features used in the Gutzwiller ansatz (number of doubly occupied sites), with the only difference being that the parameter of the feature is given implicitly by the data amplitudes, rather than explicitly in this feature space as is traditionally done.

In order to include higher-order correlated features, we can include comparisons of the occupations of plaquettes over multiple sites for the two configurations (see Fig. 1). Here, the advantage of using a kernel function to extract
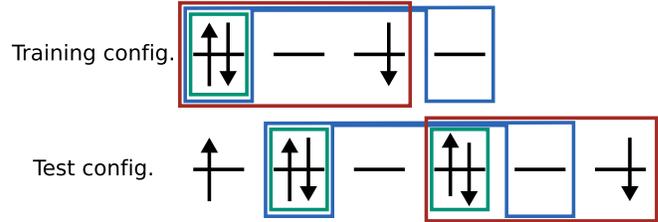

FIG. 1. Exemplification of the kernel function feature space. The kernel counts correlation features common to the test configuration (bottom) and training configuration (top), including, e.g., the single-site Gutzwiller feature (green), two-site features explicit in the Jastrow wave function (blue), and nearest-neighbor plaquette features used in EPS ansatzes (red).

descriptors implicitly becomes very prominent; while the number of possible multisite features grows exponentially with the number of sites in the plaquette considered, the kernel can still be efficiently evaluated. Multisite features can simply be included via products of delta functions evaluated for different displacements of sites i and j. Exploiting this property, we can analytically resum the contribution to the kernel function from all possible topologies and ranks of correlated descriptors defined up to a maximum number of sites (controlled by the hyperparameter p) and maximum length (controlled by a set of displacements D), giving a general kernel form

$$k(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x'_j} \left(\frac{\tilde{k}(\mathbf{x}_i, \mathbf{x}'_j)}{\sqrt{\tilde{k}(\mathbf{x}_i, \mathbf{x}_i)\tilde{k}(\mathbf{x}'_j, \mathbf{x}'_j)}} \right)^p, \qquad (3)$$

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}'_j) = \theta + \frac{1}{p}\sum_{d \in D} \delta_{x_{d(i)} x'_{d(j)}} f(d). \qquad (4)$$

In this kernel, the local occupancy over all sites i of the test configuration is compared to the occupancy of reference site j in the training configuration; if the local occupancy is the same ($\delta_{x_i x'_j} = 1$), then the local surroundings (of $\mathbf{x}_i$ and $\mathbf{x}'_j$) around sites i and j are also compared by means of the "local" kernel $\tilde{k}(\mathbf{x}_i, \mathbf{x}'_j)$. The local comparison includes all displacements $d \in D$ of i and j, running over the $|D|$ closest sites. In this work, we choose D to include all possible displacements over the lattice that the training configurations $\mathbf{x}'$ are associated with. Raising the local kernel to the power of p, we obtain a linear combination of all possible multisite correlation features in our implicit parametrization of the state, in the form of an "additive" kernel, which hierarchically includes all lower-rank features [27]. As $p \to (L-1)$, the implicit feature map of each configuration, $\Phi(\mathbf{x})$, is injective, fully characterizing the configuration. Further flexibility in this form is provided by the hyperparameter θ, controlling the relative importance of fitting high-order vs low-order features, as

well as the function $f(d)$, which in this work, unless otherwise stated, we choose to be the inverse of the displacement distance, $f(d) = 1/|d|$, to weight the importance of fitting short-range correlation features more than longer-range ones. This distance weighting will be further generalized to add more flexibility in Sec. III B. The complexity of the underlying parametrization of the GPS is therefore controlled by hyperparameters θ and p. These hyperparameters will optimally model the correlated features in the data, describing entanglement ranging up to $p + 1$ sites, with θ controlling the relative importance in the rank of these features.

The introduced kernel can be shown to model well-known ansatzes under specific limits of these hyperparameters. For $\theta \to \infty$ or $p \to 0$, we recover the $k^{(1)}$ kernel introduced in Eq. (2), which gives rise to a purely local Gutzwiller underlying form. For $\theta = 0$ and $p = 1$, we recover a generalized two-body Jastrow factor, and finally, $\theta = 0$ and $p \to \infty$ generates an entangled plaquette state (EPS) [28,29] with all plaquette sizes up to $|D| + 1$. As D is then allowed to increase in size to include all system displacements, the kernel is at its most expressive, allowing all features to be considered and an underlying (over) parametrization of the complete wave-function space.

Other important wave functions are also compactly expressed within a GPS formalism. The n-qubit W state [30] is a multipartite entangled state, important in quantum computing, which cannot be efficiently represented as an EPS [31]. However, this state can be represented as a GPS ($\theta = 0$, $p \to \infty$) with a single data configuration. Similarly, strongly interacting states such as the Laughlin wave function [5,32] describing the topologically ordered fractional quantum Hall effect can be exactly represented as a product of two GPS forms acting on a single data configuration. A full derivation of these limiting cases of the kernel is found in the Appendixes A and B.

Symmetries of the system can also be simply included in the kernel form. For instance, to incorporate the spin symmetry of the Hubbard model considered later into our ansatz, we can use the symmetrized kernel

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') + k(\mathbf{x}, \bar{\mathbf{x}}'), \qquad (5)$$

where $\bar{\mathbf{x}}'$ represents the spin reversal of configuration $\mathbf{x}'$. Key to the performance of the GPS is that, despite the fact that an exponential number of correlation features can underpin the state, the kernel presented in Eq. (3) can be computed efficiently in only $\mathcal{O}[L|D|]$ time (which can be reduced to $\mathcal{O}[L]$ when only a local change to $\mathbf{x}$ is made to a known kernel value, as is common when sampling the space). A full configurational amplitude evaluation will require N_b kernel computations, as given in Eq. (1). This construction of the state in the space of N-body data configurations therefore allows for its flexibility and expressibility. We now turn to the consideration of how compact the set of data configurations can be, and we study the Bayesian framework for computation of their selection and optimization of weights as key for the efficiency of the model.

B. Training and selection of data

Similar to the constructive algorithm for a matrix product state via successive SVD and decimation steps, we aim to devise a deterministic algorithm that allows for the compression of an arbitrary state into GPS form. This goal requires selection of a set of training configurations and associated trained weights to optimally represent the original wave function. Ideally, we want this training set to be *sparse* for a compact description of the state, by exploiting the underlying structure of the dominant correlation features of the target wave function. The required size of the training set will depend on the choice of kernel hyperparameters and desired accuracy in reproducing the state, as given by the variance of the likelihood estimator. In particular, the quality of the wave function can be improved if a more complex kernel is used with a larger set of correlation features (e.g., by increasing p). However, we expect this improvement will consequently require a larger training set of configurations to appropriately fit—a common tenet of machine learning.

We base our approach on the relevance vector machine (RVM) [33], a Bayesian algorithm for sparse learning. Within the RVM, the weight of each of a large set of candidate training configurations is considered as a separate random variable, with Gaussian prior distribution. The variances of such prior distributions can be efficiently and deterministically optimized through the maximization of the marginal likelihood of the model [34]. A large optimal variance for the configuration indicates its relative importance in reproducing the wave-function amplitudes to a desired fidelity (given by $\tilde{\sigma}^2$ in Appendix C) and is therefore selected in the data set. Concurrently, the optimal weights of the resulting configurations, w_b, are given by the mean of the posterior distribution for each data point. Complex Gaussian forms for this distribution allow for the optimization of complex weights, admitting signed amplitudes in the overall model. This method yields a rigorous Bayesian approach to automatically pick a highly compact training set of support configurations and optimized weights for the GPS, in order to faithfully reproduce a state to a desired accuracy. This is a systematically improvable approach to GPS construction, whereby increasing the complexity of the kernel and hence underlying features parametrizing the state, or reducing the variance of the likelihood, $\tilde{\sigma}^2$, automatically results in a corresponding increase in the data-set size selected by the RVM. Further details on this statistical approach for selecting and optimizing the data of the GPS are given in Appendix C, while a more in-depth exposition of this Bayesian approach can be found in Ref. [35].
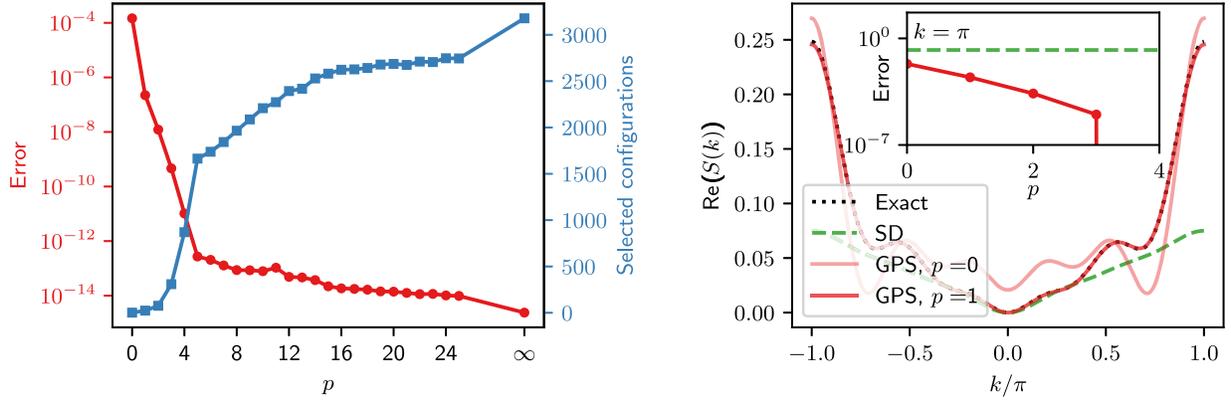
FIG. 2. Left panel: Mean-squared GPS amplitude error for the ground state of the 10-site, half-filled 1D Hubbard chain at $U = 8t$, as the kernel power hyperparameter is varied (p, with $\theta = 0$). Also shown is the size of the training set of configurations selected by the RVM from the exact data to construct the GPS with this underlying kernel complexity. Right panel: Real part of the spin structure factor $\text{Re}(S(k))$ corresponding to the exact wave function (black dotted line), the GPS with $p = 0$ and $p = 1$ (red solid line), and the single Slater determinant representation obtained with the Hartree-Fock method (green dashed line) for the same system. The inset shows the absolute error of $\text{Re}(S(k = \pi))$ for the GPS with respect to the kernel hyperparameter p, as well as the error obtained for a single Slater determinant representation.

We demonstrate the compression of a correlated state to GPS form by considering the paradigmatic fermionic Hubbard model, described by the Hamiltonian

$$\hat{H} = -t \sum_{\langle i,j \rangle, \sigma} \hat{c}^{\dagger}_{i,\sigma} \hat{c}_{j,\sigma} + U \sum_{i} \hat{n}_{i\uparrow} \hat{n}_{i\downarrow}. \tag{6}$$

The left panel of Fig. 2 shows the fidelity in compressing the ground state of this model to GPS form as the kernel complexity is increased via the p hyperparameter, for a 10-site system that can be exactly solved. The error is defined as the mean-squared error over all configurations, while the corresponding size of the automatically selected data set is also shown. With $p = 0$, only single-site local correlation features are represented, resulting in the selection of only two relevant configurations to specify the state with this level of description, as the entire (four-dimensional) feature space is rapidly saturated by the data. As the kernel complexity increases, larger training set sizes are chosen, with a corresponding rapid decrease in the error of the inferred GPS. For the most complex kernel containing all range and rank of correlations, up to about 3200 configurations are selected in the training set, resulting in the exact wave function reproduced as a GPS to numerical precision. This result represents a key feature of the GPS, that a desired kernel complexity can be used to automatically construct a sparse training set and weights required to optimally represent the complexity of correlations in the state. It is encouraging that the convergence of the GPS is so rapid with training set size, only requiring 76 training configurations to represent all amplitudes with a mean-squared error of about 10^{-8} ($p = 2$) for this system.

As the GPS directly models the wave function, all physical properties of the state can be extracted. This case

is exemplified in the right panel of Fig. 2, which shows the spin structure factor $S(k) = (1/L) \sum_j e^{-ikr_j} \langle \hat{S}_0 \cdot \hat{S}_j \rangle$ obtained for the GPS with $p = 0$ and $p = 1$, in comparison to the exact state, and the mean-field description. Although it only comprises 24 data configurations, the GPS with $p = 1$ already yields an accurate description of the structure factor, which is significantly better than that of a single Slater determinant. The inset shows the absolute error of the structure factor at the boundary of the first Brillouin zone as a function of p. Similar to the mean-squared error on the exact wave-function amplitudes, the structure factor error systematically decreases with increasing values of p. Already, the GPS with $p = 4$ matches the exact result up to numerical precision, which indicates the efficient description of even nonlocal correlations in this description.

The GPS model can also provide a unique insight into the underlying emergent physics of a state, through an analysis of the compact set of relevant configurations selected and their corresponding dominant features. A dimensionality reduction t-distributed stochastic neighbor embedding (t-SNE) algorithm can be used to project the data configurations onto a two-dimensional plane while optimally preserving a distance associated with the kernel function [36]. Figure 3 shows this process for all configurations under the $p \to \infty$ kernel given in Eq. (7). The full configurational space is shown as individual grey points, which naturally cluster together into four main groups based on the local Fock occupations of the reference site (up-spin, down-spin, unoccupied, and doubly occupied). Self-similar clusterings then also emerge in a clear hierarchical manner within each of these main clusters, based on their increasingly long-range correlation features that they have in common. However, the selected configurations chosen as basis configurations to support the GPS heavily
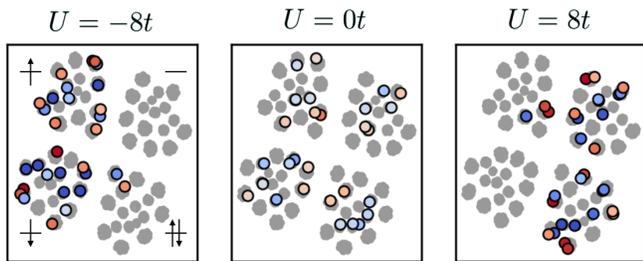
FIG. 3.   Two-dimensional t-SNE projection of the Hilbert space of a half-filled 1D Hubbard chain of eight sites. The four clusters are specified by the occupancy of the reference site of the configurations (clockwise from the top left: spin-up electron, holon, doublon, and spin-down electron). The colored points are the 30 configurations selected by the RVM as the most relevant, with red and blue colors indicating positive and negative weights.

depend on the dominant physics and correlations emerging from the model.

The optimal 30 configurations automatically selected by the RVM are colored from dark red (large positive weights) to dark blue (large negative weights). For negative U/t (the attractive Hubbard model), the RVM mostly selects configurations with single occupancies on the reference site, suppressing their amplitudes relative to the uniform state, resulting in an enhancement of short-range doublon formation and emergent pairing order [the symmetry of Eq. (5) ensures that the state is invariant to global spin flips of the data]. Conversely, at large positive U/t, doublon/hole occupancy is predominantly suppressed, resulting in the emergence of antiferromagnetic features from the chosen data points. At $U = 0$, data are selected uniformly through the Hilbert space in order to build up the long-range delocalization of the quasiparticles in the metallic state. Further insight into the subtle nature of the correlated physics emerging from these states can be found by consideration of the longer-range features within each chosen data configuration, giving a new perspective into emergent correlated physics within the GPS framework.

If the relevant physics in the wave function does not substantially change with lattice size, then we can reliably assert that the error per particle will remain constant as the system grows for the same fixed configurational data set and kernel complexity. While this case will only hold rigorously for systems beyond their bulk correlation length, we now use this observation within a practical construction of a GPS in the large system limit, by inferring from the truncated features of a smaller lattice.

III. PRACTICAL GAUSSIAN PROCESS STATES

A. Extrapolated GPS

While the compact representation of a GPS has been shown to have a number of promising features, we turn to practical algorithms for quantum many-body problems, where the desired state is not generally known in advance.

This case requires a method to choose a necessarily approximate initial state to compress to GPS form, before potential iterative refinement. We first consider an approach whereby we can exactly solve only a fragment of the system and use the information encoded about the correlations within that fragment to infer the wave-function amplitudes in GPS form on a larger simulated system—an approach that has recently also been considered for neural network states [37]. We denote this fragment the "training" lattice and the full system of interest the "simulation" lattice (e.g., in the thermodynamic limit). The approach is possible as the kernel function can compare correlation features on training and simulation configurations in different-sized Hilbert spaces, as well as the multiplicatively separable features of the GPS, ensuring a size-extensive and translationally invariant overall description. It is also important to ensure that single-particle physics beyond the length scale of the training lattice is still incorporated, which is done via a simple, fixed free-particle product state, which has the exact sign structure for this model. The full wave function on the simulation lattice is therefore the product of this free-fermion state and a GPS defined with a training data set on a small sublattice, which is designed just to infer the short-ranged correlated corrections to the underlying single-particle description. This construction mirrors the use of other successful fragmentation approaches for strong correlation, such as dynamical mean-field theory, where correlations in a small fragment are replicated across a larger system [38]. However, here, no symmetries are broken in the state, and a correlated wave function is defined over all configurations.

We use this approach to model the ground state of the Hubbard model for larger lattices, where the energy per site is well converged. In 1D, numerically exact results are accessible via the DMRG [2,39], allowing rigorous benchmarking of the accuracy of the resulting GPS state. For this work, we choose an $L' = 6$ site training lattice for the configurations $\mathbf{x}'$ and learn the model by training on the exact wave-function data of a 12-site system, which can be solved exactly. By considering training configurations with all possible particle number and spin fluctuations, it is ensured that a complete set of all features in the six-site range is spanned when fitting the 12-site data, with the other sites outside this range effectively acting as an "entanglement bath" to ensure a representation in the data of all the six-site features. There is no fundamental difficulty in the use of basis configurations to support the GPS defined on a different lattice than the one we used to define the training amplitudes, due to the sparse GPS approach taken to the fitting, as detailed in Appendix C. We then use the Bayesian fitting via the RVM to select the training set for different values of p by fitting the exact $L = 12$ solution, without any further optimization or adjustable parameters. This process results in an exceptionally computationally inexpensive approach. Since the
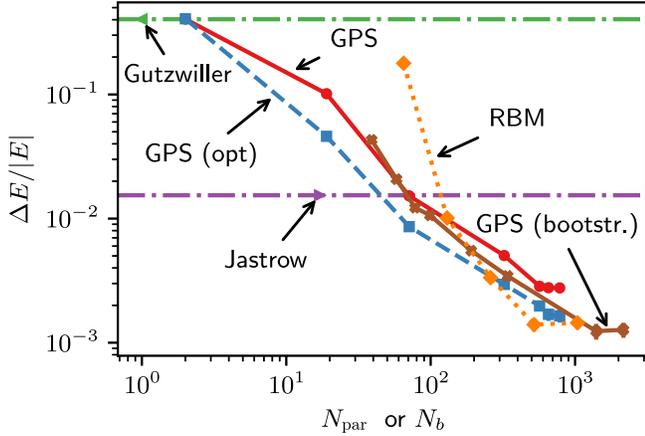
FIG. 4. Relative energy errors for the half-filled Hubbard chain (compared to DMRG) at $U = 8t, L = 32$, as a function of the size of the chosen set of configurations (N_b), as the kernel complexity p is increased from 0 to 6 ($\theta = 0, L' = 6$). For other methods, N_{par} denotes the number of variational parameters optimized in the state, as a measure of compactness and variational freedom of the state. The GPS is trained on the features of the ground state of a small 12-site lattice using configurations with all particle and spin fluctuations from an $L' = 6$ lattice. Subsequent variational optimization of the training weights of the GPS gives rise to the "GPS (opt)" values. For comparison, variationally optimized and translationally symmetric Gutzwiller, Jastrow, and RBM wavefunction results (with the same fixed free-fermion state) are included, with the RBM ratio of hidden to visible nodes (α) chosen as 1, 2, 4, 8, and 16 [23], taking the best result of 10 independent seeds. Also included are results from iterative bootstrapping of the GPS, as introduced in Sec. III B. The bootstrapped results are obtained by starting from a random initial set of data configurations. For these data points, the $p \to \infty$ limit of the kernel is used where the scaled distance weights $f(d)/\theta$ in the kernel function are optimized variationally.

resulting GPS on the simulation lattice ($L = 32$ sites in this example) is easily specified, a simple Markov chain Monte Carlo samples the energy of the state, without a sign problem, giving the results shown by the red line in Fig. 4.

For comparison, using standard variational Monte Carlo methodology developed in the mVMC package [40,41], we also optimize translationally symmetric, real-valued Gutzwiller, Jastrow, and RBM states on top of the same (fixed) free-fermion state, where the variationally lowest result over multiple independent seeds is shown, to allow for faithful comparison of the accuracy and compactness of these states. For the GPS, $p = 0$ is shown to be essentially Gutzwiller accuracy, while $p = 1$ is slightly worse than the Jastrow description due to the neglect of two-site correlated features beyond a distance of six sites. Increasing p increases the quality of results up to $p = 4$, at which point all correlations required to describe the length scales of the training lattice are captured, resulting in a relative energy error of about 3×10^{-3}, significantly beyond the accuracy of traditional Slater-Jastrow approaches, with a chosen data
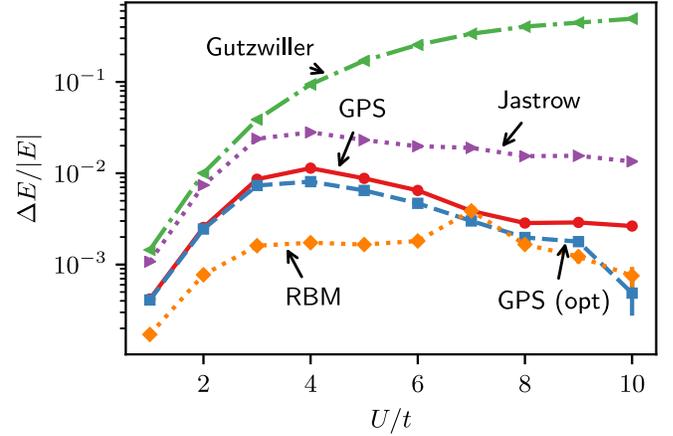

FIG. 5. Relative energy error (compared to DMRG) against U/t for GPS, Gutzwiller, Jastrow, and RBM states, for a half-filled 32-site Hubbard chain. The GPS ansatz is obtained from a fit on a 12-site chain solution with training configurations coming from an $L' = 6$ lattice with all possible particle and spin fluctuations using kernel hyperparameters ($p = 4$, $\theta = 0$). All ansatze are built on the same fixed free-fermion orbitals. Monte Carlo optimizations for the Gutzwiller, Jastrow, and RBM (with $\alpha = 5$) states are performed for 10 seeds, with the variationally lowest result shown. The red curve shows the GPS result, with the weights solely obtained from the fit on the fragment amplitudes; the blue curve represents additional variational optimization of the weights, with the $L' = 6$ data set fixed.

set of about 570 configurations. To mitigate the fact that the parameters of the model (the weights of the data) were deterministically trained by the RVM on the solution of a small sublattice, we can also variationally optimize these weights for the overall state, keeping the chosen data configurations fixed (blue squares in Fig. 4). This method removes the bias from the weights, having been optimized on the smaller lattice, but the improvement is rather marginal. The ultimate accuracy is comparable to the best RBM results, where increasing the number of hidden nodes allows for an, in principle, exact description of the state. The remaining error therefore primarily stems from the lack of correlated features in the kernel with a range beyond six sites, for which a larger training lattice for the data is required, and will be addressed in the next section.

Compact "extrapolated" GPS descriptions with competitive accuracy compared to other variational ansatzes can be found across a range of correlation strengths, as demonstrated in Fig. 5. We take the $p = 4$ GPS, with the same setup as Fig. 4, and consider the accuracy of the simple extrapolation procedure as the effective Coulomb strength U/t is varied. While the character of the ground state changes significantly in this range, the GPS results are consistent, outperforming Gutzwiller and Jastrow parametrizations for all U/t, even without any variational optimization of the weights. The $\alpha = 5$ RBM results are superior for $U \leq 6t$, where the correlation lengths exceed

the size of the training lattice features, but the increasing dominance of short-range correlations at larger U/t gives GPS results with similar accuracy to the RBM results shown. However, it is clear that to return to a systematically improvable description, it is necessary to go beyond the constraints imposed by a small training lattice and to include correlated features at all length scales and rank in an iterative approach.

B. Bootstrapped optimization of GPS

We propose a further numerical procedure for a series of optimization, compression, and data reselection steps to iteratively refine a GPS and its data set, illustrated for modeling a continuous function in Fig. 6. The approach is defined directly on the target system and therefore avoids the need to define a fragment system or restrict to short-range features. If we do not have access to any previous wave-function data, we initialize the GPS ansatz with a random set of data configurations with $w_b = 0$. Alternatively, we may simply pretrain the GPS by selecting the initial data configurations and weights from compression of a previously known approximate wave function, e.g., an optimized Jastrow ansatz, into GPS form. This step can give an improved starting point for the subsequent variational optimization of the ansatz. After the GPS ansatz is variationally optimized, the data set required for the optimized GPS is sparsified via the RVM procedure, which prunes redundant data from the original specification that are not required to describe the state to target accuracy (as given by the kernel complexity and the noise parameter $\tilde{\sigma}^2$ regularizing the fit of the data). This optimized GPS is then augmented with additional data points outside the original set, in order to acquire additional flexibility; then, all
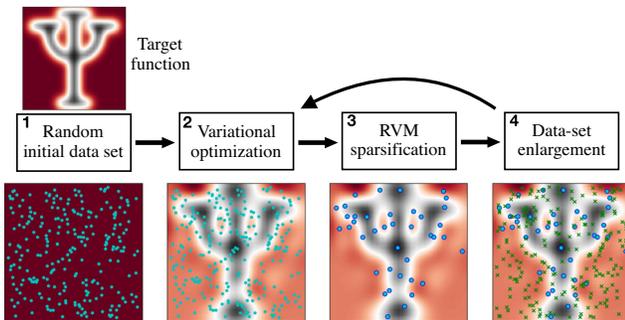

FIG. 6.   Illustration of the bootstrapping algorithm for the 2D target function shown in the top-left corner. (1) An initial set of configurations is randomly selected (blue points), with $w_t = 0$, ensuring no features are described. (2) The weights of each configuration are optimized variationally, reasonably describing the desired function. (3) The RVM is used to drastically sparsify the data set without changing the learned function. (4) Configurations with high local energy variance are added to the sparsified set (green crosses). Note that the final three panels model the function to essentially the same accuracy, allowing systematic improvement in refining the data set.

weights are reoptimized variationally, and the process is repeated until convergence. In the last step, we do not add more data configurations after the RVM has pruned non-relevant configurations from the model; we just run an additional variational optimization to obtain the final representation.

There are a number of possible criteria for the selection and addition of new trial data configurations to improve the GPS: the largest uncertainty in the GPS (obtained in a Bayesian framework), highest local energy, or largest contribution to the variance of the local energy. All these criteria indicate that the configuration is relatively poorly represented in the current GPS data set, which would benefit from explicit inclusion and optimization of these points. We choose the latter criterion, with the selection made from those configurations sampled during the previous Monte Carlo optimization of the GPS. We choose the number of configurations added to the data set to be a constant fraction (in this work 25%) of the size of the compressed data set. This choice ensures that a large fraction of the data at any point has already been deemed relevant by the RVM while allowing continued exploration of the relevant data and feature space. The complexity of the model (and therefore its accuracy) can be influenced by tuning the noise parameter $\tilde{\sigma}^2$ used to regularize the RVM fit of the data (see Appendix C). Choosing a smaller value of $\tilde{\sigma}^2$ enforces a higher fidelity in the reproduction of the wave function at each sparsification step, with a commensurate increase in required data set size.

The performance of the bootstrapped algorithm (without any pretraining) is compared to the extrapolated GPS method, shown as brown crosses in Fig. 4, for the $L = 32$ site system at $U = 8t$. To avoid restriction of the feature space, we use the $p \to \infty$ limit of the kernel (see Appendix A), which takes the form

$$k_{p\to\infty}(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x_j'} e^{-\sum_{d\in D}(1-\delta_{x_{d(i)} x_{d(j)}'})[f(d)/\theta]}, \quad (7)$$

and additionally variationally optimize the scaled distance weighting in the kernel of $f(d)/\theta$ to provide more flexibility in the distance weighting of the feature space. The total number of variational parameters of the ansatz reported in the figure is therefore $N_b + L - 1$, where N_b is the final number of data configurations reached in the bootstrapping and L is the lattice size. We note that the complete specification of a GPS requires more than just these parameters since the details of the discrete basis configurations to support the state are also required. However, the RVM procedure to select these configurations is highly efficient, such that picking these configurations is a negligible part of the simulation protocol. The bootstrapping procedure, along with the optimization of the distance weighting of the features in the kernel, allows us to reach representations that systematically surpass the

accuracy obtained from the fragmentation procedure used previously, as all long-range features are included in the description. This procedure allows us to reach a relative energy error of about 1.2×10^{-3} using $N_b = 1369$ data configurations.

We also apply this "bootstrapped" GPS approach to the 2D square Hubbard lattice at $U = 8t$, where the electronic structure problem is far more challenging. We consider the quality of the bootstrapped GPS for three different forms of the kernel: the restrictive $k^{(1)}$ kernel of Eq. (2) (obtained as $\theta \to \infty$), the $p \to \infty$ kernel of Eq. (7) where the hyperparameter controlling the relative importance of fitting high- vs low-rank features (θ) is variationally optimized but where the distance weighting is fixed at $f(d) = 1/|d|$, and finally, the kernel where this distance weighting is also entirely optimized. Further fine control of the accuracy vs compactness of the GPS can be achieved via the regularization parameter $\tilde{\sigma}^2$ within each of these kernel choices, defining the fidelity in reproducing the state upon the compression step and giving a clear route to improvability in the GPS results [35]. We improve the flexibility of the orbital part of the wave function by using a simultaneously optimized Pfaffian state, together with translational and spin-symmetry projections on this part of the wave function [41–43]. This method allows for more freedom to alter the nodal structure, and we restrict the weights of the GPS and RBM forms to real values.

Figure 7 shows the relative energy error of the bootstrapped GPS with respect to the number of variational parameters used, as the data-set size increases and kernel complexity changes for a 6×6 half-filled lattice. Comparison values are obtained with the Gutzwiller, Jastrow, and RBM states, all using the same form for the orbital and symmetry-projected reference, while auxiliary-field QMC is able to provide unbiased benchmarks at half-filling [44]. The bootstrapped GPS is pretrained via an initial RVM step on an optimized Jastrow wave function, and it demonstrates a systematic improvement with increasing data-set size, surpassing the Gutzwiller accuracy with two data points.

As the number of data points increases, it obtains the same level of accuracy as the Jastrow ansatz with similar numbers of parameters, and it can ultimately achieve a compact form with an energy error comparable to the RBM approach in a very systematic fashion. Because we choose representations with real variational parameters, the accuracy is ultimately limited by the nodal surface constraints of the Pfaffian state. This accuracy could be further improved with standard backflow or Lanczos step technology [21,45], while in the future, we will also try to describe the sign structure of these systems within the GPS model, which also represents a challenging problem for neural network quantum states [19,37,46]. However, even without these, the results demonstrate the convergence and improvability of the GPS relative to other approaches, and they
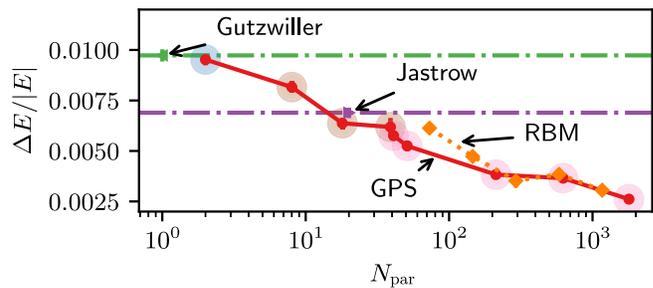

FIG. 7. Energy errors obtained for the bootstrapped GPS for the half-filled 6×6 2D square Hubbard lattice at $U = 8t$ (compared to benchmark AFQMC values [44]), plotted against the number of variational parameters (excluding those in the Pfaffian state). Results are systematically improved via use of a more complex kernel or a decrease of the $\tilde{\sigma}^2$ regularization parameter, which increases the chosen data-set size. The GPS with blue background shading was obtained with the $k^{(1)}$ kernel; values with brown shading correspond to the $p \to \infty$ kernel, where the distance weighting was fixed at $f(d) = 1/|d|$ and θ was variationally optimized; for the pink highlighted results, all scaled distance hyperparameters $f(d)/\theta$ were optimized variationally (giving $N_b + 35$ variational parameters). Also displayed for comparison are the energy errors obtained with the Gutzwiller, Jastrow, and RBM ansatzes (with α chosen as 1, 2, 4, 8, and 16). Each of these displayed comparison values corresponds to the lowest energy obtained for repeated calculations with five different random seeds.

underline the advantage of this data-driven approach, where these correlated features can be modeled with very few parameters, for a sparse and accurate representation.

Finally, we consider the tricky $n = 0.875$ hole-doped parameter regime of the model, where agreement between the most accurate numerical methods is substantially harder to achieve compared to the half-filled case [47]. Furthermore, no exact results are available for benchmarking, with many competing low-energy inhomogeneous phases claimed as the ground state [48–51]. We increase the lattice size to 8×8 and compare the convergence of the bootstrapped GPS to other variational approaches in Fig. 8. We again initialize the GPS by a fit on the Jastrow wave function and use the $p \to \infty$ kernel with variationally optimized distance weights $f(d)/\theta$ to obtain the maximal flexibility of the representation.

While the RBM at $\alpha = 16$ results in an appreciable improvement of the energy per site compared to the Jastrow ansatz, a further improvement by almost the same amount is achieved with the bootstrapped GPS, which comprises a similar number of variational parameters as the RBM (although this number fluctuates as the data set changes). Due in part to the efficient pretraining of the GPS on initial wave-function data (in this case sampled from the optimized Jastrow ansatz), the GPS ansatz exhibits rapid and robust convergence. However, we are cautious of firm conclusions as to their ultimate relative accuracy and compactness, especially with the rapid developments of
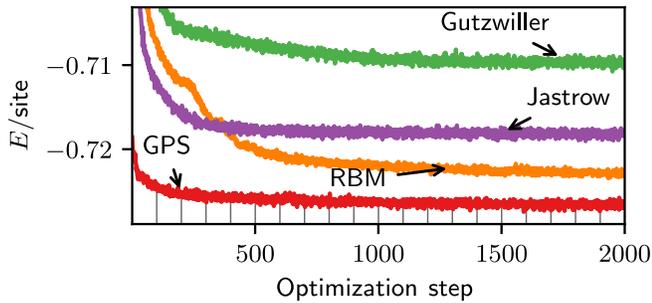
FIG. 8. Convergence of the ground-state energy for the 8×8 Hubbard model, at $U = 8t$ and with 0.875 electrons per site. The bootstrapped GPS representation is initialized at the optimized Jastrow wave function, resulting in rapid convergence. Vertical grey lines indicate the iterations at which the configurational data set of the GPS is updated. The convergence is shown for Gutzwiller (1 parameter), Jastrow (34 parameters), the RBM with $\alpha = 16$ (2064 parameters), and the bootstrapped GPS with optimized distance weights ($N_b + 63$ parameters). The number of data configurations specifying the GPS after each application of the RVM in the algorithm range from about 1000 to 2000 configurations.

more advanced network architectures [52–54] and associated optimization techniques [55] in the neural-network quantum state field.

IV. SUMMARY AND OUTLOOK

We present a compact and efficient representation of quantum many-body wave functions, the Gaussian process state. By allowing the parameters of the model to follow a statistical distribution, we derive a Bayesian inference framework for the probability amplitudes of the state in terms of an exponentially resummed space of physical correlation features. These features are expressed in a kernel function, which can compare any two configurations in this feature space, and allow for their amplitudes to be learnt from a data set, while also allowing insight into dominant features of this learnt state. The state can be deterministically and efficiently constructed with a manifestly compact data set from any initial state. As the data set is allowed to increase in size, the state is then shown to be able to universally approximate any entangled quantum state. In this way, these states complement the neural-network states, also derived from machine learning principles to approximate any quantum state, and we anticipate many synergies between these approaches in the future. We also see the importance of the Bayesian framework as set out in this work in this endeavor, and the compact nature of the data-driven GPS approach as a key tool to reliably and systematically move beyond existing ansatze-driven approaches in strongly correlated systems.

We also present two numerical approaches that exploit this wave-function form directly to approximate an unknown quantum state. In the first, the model is trained on amplitudes from a small subsystem, whose features are then implicitly extrapolated to define a state over the whole system. This process is numerically simple; however, the restriction to relatively short-ranged entanglement features inspires the development of the bootstrapped iterative optimization approach, which does not rely on any previous wave-function information. This model has broad applicability and is shown to be efficient, sparse, and accurate for both the half-filled and hole-doped strongly correlated regimes of the Hubbard model. We expect these schemes to work similarly well for other systems relevant for a broad range of scientific domains and to open a new route for the compact representation of quantum many-body systems directly in terms of data.

APPENDIX A: LIMITS OF THE KERNEL FUNCTION

The kernel function is one of the key ingredients of the GPS ansatz proposed in this work since it implicitly models the entanglement and correlations that can be represented. Its general form, repeated here for convenience,

$$k(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x_j'} \left(\frac{\tilde{k}(\mathbf{x}_i, \mathbf{x}_j')}{\sqrt{\tilde{k}(\mathbf{x}_i, \mathbf{x}_i)\tilde{k}(\mathbf{x}_j', \mathbf{x}_j')}} \right)^p, \quad \text{(A1)}$$

$$\tilde{k}(\mathbf{x}_i, \mathbf{x}_j') = \theta + \frac{1}{p} \sum_{d \in D} \delta_{x_{d(i)} x_{d(j)}'} f(d), \quad \text{(A2)}$$

depends on the two hyperparameters, p and θ, as well as the chosen set of displacements D and the distance weighting function $f(d)$. The "power" p controls the maximum rank of the correlations modeled (i.e., the maximum number of

sites for which an explicit feature is present); the "length scale" θ allows tuning of the importance of accurately fitting low-rank compared to high-rank features in the ansatz; the distance weighting function $f(d)$ controls the weighting of features depending on their range, and D determines the set of displacements between two sites, which are included in the correlations. In this work, this is generally chosen to be the set of all possible displacements. It is instructive to look at the behavior of the kernel for certain limiting cases of the hyperparameters just described since these give rise to well-known physical descriptions.

1. Gutzwiller

Limit $\theta \to \infty$ or $p = 0$. A Gutzwiller representation is recovered by either letting θ go to infinity or, equivalently, by taking p to be zero. Indeed, in both cases, the kernel in Eq. (A1) takes the very simple form

$$k_G(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x_j'}. \qquad (A3)$$

This kernel scans over the sites in configuration $\mathbf{x}$ and counts the number of sites with occupancies identical to that of $\mathbf{x}'$ at reference site j. This reference site of the data in the training lattice can be chosen arbitrarily in translationally symmetric systems. It is straightforward to check that the above kernel can be written as the scalar product in a feature space, which is then symmetrized over the translations of the system lattice, as $k_G(\mathbf{x}, \mathbf{x}') = \sum_i^L \Phi_i^G(\mathbf{x}) \cdot \Phi_j^G(\mathbf{x}')$, where the feature vectors are defined as

$$(\mathbf{\Phi}_i^G)_s(\mathbf{x}) = \delta_{s, x_i}, \qquad (A4)$$

where s is a specific occupancy $s \in \{\cdot, \uparrow, \downarrow, \uparrow\downarrow\}$ and $\Phi_i^G(\mathbf{x})$ therefore simply checks whether site i in $\mathbf{x}$ has a given occupancy s. The choice to sum over the sites of the lattice (i) reflects the choice to symmetrize the features over the lattice, while the neglect of symmetrization over the training lattice (j) avoids additional cost in kernel evaluation, without substantial loss of accuracy, which means that the kernel k^G implicitly defines a linear model on the features $\Phi_i^G(\mathbf{x})$. This model can be considered a generalized Gutzwiller form that favors or suppresses a given configuration depending on the number of sites in any given local occupancy, s.

2. Jastrow. Limit $\theta = 0$ and $p = 1$

A generalization of the Jastrow representation is recovered by setting $\theta = 0$ and $p = 1$. Indeed, for such a choice of hyperparameters, the kernel takes the form

$$k_J(\mathbf{x}, \mathbf{x}') = \sum_i^L \sum_{d \in D} \delta_{x_i x_j'} \delta_{x_{d(i)} x_{d(j)}'} f(d). \qquad (A5)$$

As in the case of the Gutzwiller limit just discussed, it is also possible to write the above kernel as a (symmetrized) scalar product in some explicitly constructed feature space. In this case, the features with respect to site i and displacement vector d_{12} are defined by the following rank-3 tensor:

$$(\mathbf{\Phi}_i^J)_{s_1, s_2, d_{12}}(\mathbf{x}) = \sum_l^L \delta_{s_1 x_i} \delta_{s_2 x_l} \delta_{d_{12} d_{il}} \sqrt{f(d_{il})}. \qquad (A6)$$

The component s_1, s_2, d_{12} of Φ_i^J evaluates to $\sqrt{f(d_{12})}$ whenever the occupancies of the site with index i and the displaced site with index $d_{12}(i)$ are found with the local occupancies s_1 and s_2. Similarly to the Jastrow factor, this representation depends on two sites and their relative positions. However, Φ_i^J can be considered of generalized Jastrow form since it depends on the specific occupation of the two sites, rather than their charge densities. It should also be stressed that the $\sqrt{f(d_{il})}$ does not determine the overall amplitude of the feature but rather the importance in trying to fit these to the data. Its overall weight is instead determined by the weights over the training configurations that exhibit that correlation feature.

3. Squared exponential kernel. Limit $p \to \infty$

Setting the hyperparameter p to finite values is a physically meaningful way to restrict the space of correlations defined by the kernel function to a certain rank of correlations. However, in certain cases (e.g., in the bootstrapping calculations described in this work), it might be convenient to take the $p \to \infty$ limit, obtaining a very flexible kernel controlled by the single hyperparameter θ as well as the chosen distance weighting function $f(d)$. To evaluate this limit, we exploit the identity $e^x = \lim_{n \to \infty}[1 + (x/n)]^n$, obtaining the following kernel:

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x_j'} e^{-\gamma_S^2(\mathbf{x}_i, \mathbf{x}_j')/2\theta}, \qquad (A7)$$

$$\gamma_S^2(\mathbf{x}_i, \mathbf{x}_j') = S(\mathbf{x}_i, \mathbf{x}_i) + S(\mathbf{x}_j', \mathbf{x}_j') - 2S(\mathbf{x}_i, \mathbf{x}_j'), \qquad (A8)$$

$$S(\mathbf{x}_i, \mathbf{x}_j') = \sum_{d \in D} \delta_{x_{d(i)} x_{d(j)}'} f(d). \qquad (A9)$$

The above kernel can be seen as a squared exponential kernel on the distance induced by the scalar product $S(\mathbf{x}_i, \mathbf{x}_j')$. Such a distance metric is related to the concept of a Hamming distance [56] in discrete strings. This squared exponential kernel includes, by construction, all possible correlation features, depending on a single hyperparameter θ, which models the relative importance of fitting higher-order to lower-order features in the training data.

4. Entangled plaquette states.
Limit $p \to \infty$ and $\theta \to 0$

The entangled-plaquette-states ansatz (also known in the literature as correlator product states or complete graph tensor networks) decomposes the overall amplitude of a configuration into a product of contributions coming from a set of plaquettes over the degrees of freedom [28,29]. Each possible state local to each plaquette is given its own variational parameter, with the number of parameters therefore growing exponentially with plaquette size. Spanning the space of these features is recovered from the kernel in Eq. (A1) in the limit of $p \to \infty$ and $\theta \to 0$, with the range of effective plaquettes then being controlled by the allowed displacements in D. We have shown that the limit $p \to \infty$ evaluates to Eq. (A7). For $\theta \to 0$, the exponential $e^{-\gamma_S^2(\mathbf{x}_i,\mathbf{x}'_j)/2\theta}$ in Eq. (A7) tends towards a delta function $\delta^D_{\mathbf{x}_i \mathbf{x}'_j}$ evaluating to zero unless the local occupations of $\mathbf{x}$ around i and $\mathbf{x}'$ around j are identically equal, up to the set of displacements defined by D. We can hence write the resulting kernel as

$$k_C(\mathbf{x}, \mathbf{x}') = \sum_i^L \delta_{x_i x'_j} \delta^D_{\mathbf{x}_i \mathbf{x}'_j}. \tag{A10}$$

The combined Kronecker symbols $\delta_{x_i x'_j} \delta^D_{\mathbf{x}_i \mathbf{x}'_j}$ only evaluate to 1 if all the sites within a given cutoff in the two configurations are exactly the same, and we can hence write it as $\delta_{(x_i,\mathbf{x}_i),(x'_j,\mathbf{x}'_j)}$. Similarly to the Gutzwiller and Jastrow limits, the above kernel can also be rewritten as an explicit scalar product symmetrized across the lattice. In this case, the corresponding features are defined by the tensor

$$(\boldsymbol{\Phi}^C_i)_{\mathbf{s}}(\mathbf{x}) = \delta_{\mathbf{s},(x_i,\mathbf{x}_i)}, \tag{A11}$$

where now the index $\mathbf{s} = (s_1, s_2, \ldots)$ runs over all Fock states within the $|D| + 1$ site plaquette around site i. Linear regression on $\sum_{i=1}^L \Phi^C_i(\mathbf{x})$ gives rise to an entangled-plaquette-states ansatz, favoring or suppressing a given amplitude depending on each "plaquette" of $|D| + 1$ sites that is found within the corresponding configuration. It is important to note here that while the parametrization of a *direct* EPS ansatz grows exponentially with $|D|$ (roughly as $4^{|D|}$), the number of parameters when modeled *indirectly* through the above kernel only scales with the size of the data set. The amplitudes of these features are then directly inferred from the weights over the test configurations.

APPENDIX B: REPRESENTABILITY OF SPECIFIC STATES

1. W state

The GPS can also be used to represent certain many-qubit entangled states central to quantum information theory. As an example, here we show how to represent the entangled W state of n qubits,

$$|W\rangle = \frac{1}{\sqrt{n}}(|10\ldots0\rangle + |01\ldots0\rangle + \cdots + |00\ldots1\rangle), \tag{B1}$$

with the GPS. The W state can be represented using a single training configuration, namely, one of the n basis configurations that generate the state, e.g.,

$$\mathbf{x}' = |10\ldots0\rangle, \tag{B2}$$

and a kernel function with $p \to \infty$ and $\theta \to 0$, where the delta function only distinguishes between the two local occupancies $|0\rangle$ and $|1\rangle$. With this choice, the kernel $k(\mathbf{x}', \mathbf{x})$ always either evaluates to 1, if $\mathbf{x}$ is one of the n translations of the configurations $\mathbf{x}'$ that create the W state, or to 0 otherwise. Setting the weight associated with the data configuration to the normalization $w_{\mathbf{x}'} = \ln(1/\sqrt{n})$, we therefore obtain a GPS representation of the W state using only a single data configuration.

2. Laughlin wave function

Modifying the kernel function used also allows representation of other known ansatzes in very compact forms. As an example, we consider the Laughlin wave function discretized on a real-space lattice, an ansatz describing n spinless electrons on a two-dimensional lattice of L sites. Expressing the lattice position of site k as a complex number $Z_k = x_k + iy_k$, where $x_k(y_k)$ is the $x(y)$ coordinate of the site, this discretized Laughlin wave function can be written as [28,32]

$$\Psi_L(\mathbf{x}) \propto \prod_{i<j}^n (Z_i - Z_j)^q e^{-\alpha \sum_{k=1}^n |Z_k|^2}, \tag{B3}$$

where q and α are fixed by the external model and desired state, with indices i, j, and k referring to occupied sites in a sampled configuration $\mathbf{x}$.

We can express the Laughlin wave function as a GPS using a single data configuration with a modified kernel or, equivalently, as a product of two GPS forms. This representation may be of interest as a starting point to build more flexible ansatzes for this problem. We define the modified kernel as

$$k_L(\mathbf{x}, \mathbf{x}') = -\alpha k_{1B}(\mathbf{x}, \mathbf{x}') + k_{2B}(\mathbf{x}, \mathbf{x}'), \tag{B4}$$

which is a linear combination of the one- and two-body kernels k_{1B} and k_{2B} (with k_{1B} weighted by the parameter α). The one-body kernel is given as

$$k_{1B}(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^L \delta_{x_k x'_k} |Z_k|^2, \tag{B5}$$

where the sum runs over all sites of the lattice. This definition is equivalent to the local kernel $\tilde{k}$ of Eq. (A2) with $p = 1$ and $\theta = 0$, and with $f(d) = |d|^2$. To define the two-body kernel, we express the displacements in the local kernel as complex values $(d_x + id_y)$ and use the kernel of Eq. (A1) with $p = 1$, $\theta = 0$, and a displacement weighting function of $f(d) = \ln(d_x + id_y)^q$, where the site with index 0 is selected as a reference site in $\mathbf{x}'$. This construction gives a two-body kernel as

$$k_{2B}(\mathbf{x}, \mathbf{x}') = \sum_{i}^{L} \delta_{x_i x_0'} \sum_{d \in D} \delta_{x_{d(i)} x_{d(0)}'} \ln(d_x + id_y)^q. \quad \text{(B6)}$$

Here, the sum is taken over all possible displacements for which the displacement of site i gives a site with a larger index, i.e., $D = \{d : d(i) > i\}$. Using a single data configuration $\mathbf{x}'$ in which all sites are occupied, $\mathbf{x}' = |1, 1, \ldots, 1\rangle$, the factor $e^{k_{2B}(\mathbf{x},\mathbf{x}')}$ reproduces the two-body factor of the Laughlin wave function, $\prod_{i<j}^{n}(Z_i - Z_j)^q = e^{\sum_{i<j}^{n} \ln((Z_i - Z_j)^q)}$, and the factor $e^{-\alpha k_{1B}(\mathbf{x},\mathbf{x}')}$ gives the one-body factor $e^{-\alpha \sum_{k=1}^{n} |Z_k|^2}$. With the introduced definition of the kernel function k_L, the Laughlin wave function is therefore obtained as the GPS representation

$$\Psi_L(\mathbf{x}) \propto e^{k_L(\mathbf{x}, \mathbf{x}')}. \quad \text{(B7)}$$

APPENDIX C: RELEVANCE VECTOR MACHINE

In this Appendix, we describe the Bayesian algorithms used to compute the GPS model parameters to optimally fit a target state. This process includes the optimization of the weights of the data points and the selection of the sparse configurational data set itself, as well as, optionally, the optimization of kernel hyperparameters. An expanded discussion of this GPS model optimization can be found in Ref. [35].

The wave-function ansatz in Eq. (1) can be rewritten as

$$\ln \Psi_g(\mathbf{x}) = \sum_b w_b k(\mathbf{x}, \mathbf{x}_b), \quad \text{(C1)}$$

where the summation runs over all configurations in the data set $\{\mathbf{x}_b\}_{b=1}^{N_b}$. The goal of the RVM is to find a sparse set of weights $\mathbf{w}$ from the training data, which optimally represent the target state, as given by samples of its log wave-function amplitudes. As a standard modeling assumption, we model the likelihood of the logarithm of the amplitudes as a Gaussian distribution

$$p(\ln \mathbf{\Psi} | \mathbf{w}, \mathbf{\Lambda}) = \mathcal{N}(\mathbf{Kw}, \mathbf{\Lambda}), \quad \text{(C2)}$$

where we have defined the matrix $\mathbf{K}$ of dimensions $N \times N_b$ and entries $K_{ib} = k(\mathbf{x}_i, \mathbf{x}_b)$.

Using a constant variance $\mathbf{\Lambda} = \mathbf{I}\sigma^2$ would allow us to achieve a constant precision σ^2 on the logarithm of the amplitudes in Eq. (C1). This precision does not fully encompass the aim we are trying to achieve because a constant precision on the logarithm of the wave-function amplitudes does not correspond to a constant precision on the amplitudes themselves. As an example, a wave function with amplitudes ranging from 10^{-5} to 10^{-1} will have amplitude logarithms (in base 10) ranging from -5 to -1. Fitting all amplitudes to an accuracy of $\sigma^2 = 0.1$ means that, roughly speaking, the largest and the smallest amplitudes will be allowed to fluctuate between $10^{-1.1}$ and $10^{-0.9}$ and between $10^{-5.1}$ and $10^{-4.9}$, respectively. Clearly, these fluctuations are much larger for the larger amplitudes, meaning that they will be represented to a lower precision in the true amplitudes compared to the smaller amplitudes. These differences in the fluctuations also result in the selected data set being less sparse than it could be, as small weighted amplitudes are comparatively being very finely resolved.

To overcome this problem, we can let each log amplitude, $\ln \Psi_i$, be represented to a different precision σ_i^2 and define the variance Λ in Eq. (C2) as $\Lambda_{ij} = \delta_{ij}\sigma_i^2$. The variances σ_i^2 can be chosen to precisely compensate for the fact that we are fitting to the logarithm of the state, by ensuring that the likelihood distribution of a given log wave-function amplitude has a smaller variance if the magnitude of its expectation is larger. The variance of a given amplitude Ψ_i, under the assumption of Eq. (C2) that its logarithm is normally distributed, is

$$\text{Var}(\Psi_i) = \langle \Psi_i \rangle^2 (e^{\sigma_i^2} - 1). \quad \text{(C3)}$$

This relation demonstrates that the variance of Ψ_i increases for increasing $\langle \Psi_i \rangle$, which we aim to compensate. By assuming that the expectation value for the amplitude is the same as the one given by the data set, we can fix the variance of each amplitude as $\text{Var}(\Psi_i) = \tilde{\sigma}^2$ and solve for the amplitude-dependent precision in the likelihood, as

$$\sigma_i^2 = \ln\left(1 + \frac{\tilde{\sigma}^2}{\Psi_i^2}\right). \quad \text{(C4)}$$

This construction allows us to define a constant variance $\tilde{\sigma}^2$ to model all wave-function amplitudes and correspondingly use the configuration-dependent σ_i^2 above to model the variance in the likelihood distribution of specific values of $\ln \Psi_i$.

To enforce sparsity, RVM further models the prior distribution of each parameter w_b with a normal distribution centered around zero and an independent variance of α_b^{-1}. This distribution gives rise to the following sparse prior on the weights,

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \tag{C5}$$

$$\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha}). \tag{C6}$$

Having specified prior and likelihood distributions, the posterior distribution of the weights can be formally written using Bayes' theorem as

$$p(\mathbf{w}|\ln\boldsymbol{\Psi}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}) = \frac{p(\ln\boldsymbol{\Psi}|\mathbf{w}, \boldsymbol{\Lambda})p(\mathbf{w}|\boldsymbol{\alpha})}{p(\ln\boldsymbol{\Psi}|\boldsymbol{\alpha}, \boldsymbol{\Lambda})}. \tag{C7}$$

Since the prior and likelihood are Gaussian, the posterior is also Gaussian and can be computed in closed form [33] as

$$p(\mathbf{w}|\ln\boldsymbol{\Psi}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{C8}$$

with

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{K}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\ln\boldsymbol{\Psi}, \tag{C9}$$

$$\boldsymbol{\Sigma} = (\mathbf{K}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{K} + \mathbf{A})^{-1}. \tag{C10}$$

The mean $\boldsymbol{\mu}$ of the posterior distribution is the best estimate for the value of the parameters $\mathbf{w}$. The values of the variances $\boldsymbol{\alpha}$ are optimized by a procedure called type-II maximum likelihood. This procedure consists of the maximization of the denominator of Eq. (C7), the marginal likelihood, with respect to these variances on the prior of the weights and likelihood of reproducing the target state. The value of the marginal likelihood (or its logarithm) can also be given in closed form [34]. An efficient algorithm is used whereby we initialize with an empty data set, and data are "greedily" added by ensuring the update to a parameter α_b gives rise to the maximum increase in the log marginal likelihood. This algorithm also guarantees convergence to a local maxima within the RVM. During this optimization, the sparse prior dictates that many α_b will be optimized at values of infinity, indicating that the basis configuration $\mathbf{x}_b'$ is not relevant to describe the wave function within the feature space of the kernel. This configuration can be ignored within the ansatz in Eq. (C1), as its features are spanned effectively by the other data points.

We typically set $\tilde{\sigma}^2$ to a specific value to reach a desired accuracy and sparsity with the fit. Alternatively, we can determine a sensible choice for $\tilde{\sigma}^2$ for a given data set and kernel choice by running the RVM multiple times and maximizing the marginal likelihood with respect to the parameter $\tilde{\sigma}^2$, e.g., by the hyperparameter optimization technique presented in Ref. [57], implemented in the hyperopt python package [58].

Finally, we note that we typically want the likelihood variance $\tilde{\sigma}^2$ to be defined relative to the overall wave-function scale, which will depend on the normalization of the state we fit to. However, the optimized weights favor solutions distributed around zero due to the zero-mean prior

of Eq. (C5). We can therefore renormalize the training data by shifting the log amplitudes we fit on such that their mean vanishes. If we are interested in describing the true amplitudes (including their normalization) with the learned GPS, we can simply add the rescaling constant of the training data back to the exponential for GPS predictions, effectively shifting the predicted log amplitudes. Because the shift of the log amplitudes corresponds to a multiplication of the actual amplitudes, the normalization factor determined in the training can be ignored if the overall norm of the wave function is irrelevant, as is, e.g., the case for Monte Carlo evaluation of the energy. It was found that enforcing this zero-mean condition on the log amplitudes of the training data helps to numerically stabilize the described bootstrapping procedure since the magnitudes of the weights are well behaved and are then commensurate with the imposed zero-mean prior on these quantities.

[1] M. C. Gutzwiller, *Effect of Correlation on the Ferromagnetism of Transition Metals*, Phys. Rev. Lett. **10**, 159 (1963).

[2] U. Schollwöck, *The Density-Matrix Renormalization Group in the Age of Matrix Product States*, Ann. Phys. (N.Y.) **326**, 96 (2011).

[3] R. Orús, *A Practical Introduction to Tensor Networks: Matrix Product States and Projected Entangled Pair States*, Ann. Phys. (N.Y.) **349**, 117 (2014).

[4] A. W. Sandvik and G. Vidal, *Variational Quantum Monte Carlo Simulations with Tensor-Network States*, Phys. Rev. Lett. **99**, 220602 (2007).

[5] R. B. Laughlin, *Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations*, Phys. Rev. Lett. **50**, 1395 (1983).

[6] R. J. Bartlett and M. Musiał, *Coupled-Cluster Theory in Quantum Chemistry*, Rev. Mod. Phys. **79**, 291 (2007).

[7] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning*, Phys. Rev. Lett. **108**, 058301 (2012).

[8] A. Glielmo, P. Sollich, and A. De Vita, *Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels*, Phys. Rev. B **95**, 214302 (2017).

[9] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons*, Phys. Rev. Lett. **104**, 136403 (2010).

[10] L. Cheng, S. Ramchandran, T. Vatanen, N. Lietzén, R. Lahesmaa, A. Vehtari, and H. Lhdesmki, *An Additive Gaussian Process Regression Model for Interpretable Non-parametric Analysis of Longitudinal Data*, Nat. Commun. **10**, 1798 (2019).

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2005).

[12] R. M. Neal, *Bayesian Learning for Neural Networks*, Ph.D. thesis, University of Toronto, Department of Computer Science (1994).

[13] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, *Deep Neural Networks as*

Gaussian Processes, International Conference on Learning Representations, 2018, https://openreview.net/forum?id= B1EA-M-0Z.

[14] Z.-A. Jia, B. Yi, R. Zhai, Yu.-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).

[15] X. Gao and L.-M. Duan, *Efficient Representation of Quantum Many-Body States with Deep Neural Networks,* Nat. Commun. **8,** 662 (2017).

[16] G. Carleo, Y. Nomura, and M. Imada, *Constructing Exact Representations of Quantum Many-Body Systems with Deep Neural Networks,* Nat. Commun. **9,** 5322 (2018).

[17] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, *Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions,* Nat. Commun. **10,** 5024 (2019).

[18] R. Zen, L. My, . Tan, F. Hébert, M. Gattobigio, C. Miniatura, D. Poletti, and S. Bressan, *Transfer Learning for Scalability of Neural-Network Quantum States,* Phys. Rev. E **101,** 053301 (2020).

[19] K. Choo, A. Mezzacapo, and G. Carleo, *Fermionic Neural-Network States for Ab-Initio Electronic Structure,* Nat. Commun. **11,** 2368 (2020).

[20] J. Hermann, Z. Schätzle, and F. Noé, *Deep-Neural-Network Solution of the Electronic Schrödinger Equation,* Nat. Chem. **12,** 891 (2020).

[21] 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. Research **2,** 033429 (2020).

[22] S. R. Clark, *Unifying Neural-Network Quantum States and Correlator Product States via Tensor Networks,* J. Phys. A **51,** 135301 (2018).

[23] G. Carleo and M. Troyer, *Solving the Quantum Many-Body Problem with Artificial Neural Networks,* Science **355,** 602 (2017).

[24] R. G. Melko, G. Carleo, J. Carrasquilla, and J. I. Cirac, *Restricted Boltzmann Machines in Quantum Physics,* Nat. Phys. **15,** 887 (2019).

[25] 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).

[26] R. Jastrow, *Many-Body Problem with Strong Forces,* Phys. Rev. **98,** 1479 (1955).

[27] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen, *Additive Gaussian Processes,* in *Advances in Neural Information Processing Systems 24,* edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2011), pp. 226–234.

[28] H. J. Changlani, J. M. Kinder, C. J. Umrigar, and G. K.-L. Chan, *Approximating Strongly Correlated Wave Functions with Correlator Product States,* Phys. Rev. B **80,** 245116 (2009).

[29] F. Mezzacapo, N. Schuch, M. Boninsegni, and J. I. Cirac, *Ground-State Properties of Quantum Many-Body Systems: Entangled-Plaquette States and Variational Monte Carlo,* New J. Phys. **11,** 083026 (2009).

[30] W. Dür, G. Vidal, and J. I. Cirac, *Three Qubits Can Be Entangled in Two Inequivalent Ways,* Phys. Rev. A **62,** 062314 (2000).

[31] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, *Matrix Product State Representations,* Quantum Inf. Comput. **7,** 401 (2007).

[32] I. Glasser, J. I. Cirac, G. Sierra, and A. E. B. Nielsen, *Lattice Effects on Laughlin Wave Functions and Parent Hamiltonians,* Phys. Rev. B **94,** 245104 (2016).

[33] M. E. Tipping, *The Relevance Vector Machine,* in *Advances in Neural Information Processing Systems 12* (MIT Press, 2000), pp. 652–658, http://papers.nips.cc/paper/1719-the-relevance-vector-machine.pdf.

[34] M. E. Tipping and A. C. Faul, *Fast Marginal Likelihood Maximisation for Sparse Bayesian Models,* in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003, Key West, Florida, USA, 2003,* edited by C. M. Bishop and B. J. Frey (Society for Artificial Intelligence and Statistics, 2003).

[35] Y. Rath, A. Glielmo, and G. H. Booth, *A Bayesian Inference Framework for Compression and Prediction of Quantum States,* J. Chem. Phys. **153,** 124108 (2020).

[36] L. van der Maaten and G. Hinton, *Visualizing Data Using t-SNE,* J. Mach. Learn. Res. **9,** 2579 (2008).

[37] T. Westerhout, N. Astrakhantsev, K. S. Tikhonov, M. I. Katsnelson, and A. A. Bagrov, *Generalization Properties of Neural Network Approximations to Frustrated Magnet Ground States,* Nat. Commun. **11,** 1593 (2020).

[38] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, *Dynamical Mean-Field Theory of Strongly Correlated Fermion Systems and the Limit of Infinite Dimensions,* Rev. Mod. Phys. **68,** 13 (1996).

[39] S. Sharma and G. K.-L. Chan, *Spin-Adapted Density Matrix Renormalization Group Algorithms for Quantum Chemistry,* J. Chem. Phys. **136,** 124121 (2012).

[40] S. Sorella, *Generalized Lanczos Algorithm for Variational Quantum Monte Carlo,* Phys. Rev. B **64,** 024512 (2001).

[41] T. Misawa, S. Morita, K. Yoshimi, M. Kawamura, Y. Motoyama, K. Ido, T. Ohgoe, M. Imada, and T. Kato, *mVMC—Open-Source Software for Many-Variable Variational Monte Carlo Method,* Comput. Phys. Commun. **235,** 447 (2019).

[42] D. Tahara and M. Imada, *Variational Monte Carlo Method Combined with Quantum-Number Projection and Multi-Variable Optimization,* J. Phys. Soc. Jpn. **77,** 114701 (2008).

[43] T. Mizusaki and M. Imada, *Quantum-Number Projection in the Path-Integral Renormalization Group Method,* Phys. Rev. B **69,** 125110 (2004).

[44] 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).

[45] D. Luo and B. K. Clark, *Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions,* Phys. Rev. Lett. **122,** 226401 (2019).

[46] A. Szabó and C. Castelnovo, *Neural Network Wave Functions and the Sign Problem,* Phys. Rev. Research **2,** 033075 (2020).

[47] J. P. F. LeBlanc *et al.* (Simons Collaboration on the Many-Electron Problem), *Solutions of the Two-Dimensional Hubbard Model: Benchmarks and Results from a Wide Range of Numerical Algorithms*, Phys. Rev. X **5,** 041041 (2015).

[48] B.-X. Zheng, C.-M. Chung, P. Corboz, G. Ehlers, M.-P. Qin, R. M. Noack, H. Shi, S. R. White, S. Zhang, and G. K.-L. Chan, *Stripe Order in the Underdoped Region of the Two-Dimensional Hubbard Model*, Science **358,** 1155 (2017).

[49] B.-X. Zheng and G. K.-L. Chan, *Ground-State Phase Diagram of the Square Lattice Hubbard Model from Density Matrix Embedding Theory*, Phys. Rev. B **93,** 035126 (2016).

[50] W.-J. Hu, F. Becca, and S. Sorella, *Absence of Static Stripes in the Two-Dimensional $t - J$ Model Determined Using an Accurate and Systematic Quantum Monte Carlo Approach*, Phys. Rev. B **85,** 081110(R) (2012).

[51] A. Macridin, M. Jarrell, and T. Maier, *Phase Separation in the Hubbard Model Using the Dynamical Cluster Approximation*, Phys. Rev. B **74,** 085104 (2006).

[52] X. Liang, W.-Y. Liu, P.-Z. Lin, G.-C. Guo, Y.-S. Zhang, and L. He, *Solving Frustrated Quantum Many-Particle Models with Convolutional Neural Networks*, Phys. Rev. B **98,** 104426 (2018).

[53] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, *Quantum Entanglement in Deep Learning Architectures*, Phys. Rev. Lett. **122,** 065301 (2019).

[54] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, *Recurrent Neural Network Wave Functions*, Phys. Rev. Research **2,** 023358 (2020).

[55] C. Roth, *Iterative Retraining of Quantum Spin Models Using Recurrent Neural Networks*, arXiv:2003.06228.

[56] R. W. Hamming, *Error Detecting and Error Correcting Codes*, Bell Syst. Tech. J. **29,** 147 (1950).

[57] J. Bergstra, D. Yamins, and D. Cox, *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*, in *Proceedings of the 30th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 28, edited by S. Dasgupta and D. McAllester (PMLR, Atlanta, Georgia, 2013), pp. 115–123.

[58] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, *Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization*, Comput. Sci. Discovery **8,** 014008 (2015).