

Learning minimal representations of fermionic ground states

Felix Frohnert^{✉,*}, Emiel Koridon[✉], and Stefano Polla^{✉,†}
 (*aQa^L*) *Applied Quantum Algorithms, Universiteit Leiden, Leiden, The Netherlands*



(Received 4 February 2026; accepted 27 May 2026; published 6 July 2026)

We introduce an unsupervised machine-learning framework that discovers optimally compressed representations of quantum many-body ground states. Using an autoencoder neural network architecture on reduced density matrix data from L -site Fermi-Hubbard models, we identify minimal latent spaces with a sharp reconstruction quality threshold at $L - 1$ latent dimensions, matching the system's intrinsic degrees of freedom. We demonstrate the use of the trained decoder as a differentiable variational ansatz to minimize energy directly within the latent space. Crucially, this approach circumvents the N -representability problem, as the learned manifold implicitly restricts the optimization to physically valid quantities.

DOI: [10.1103/qs15-cyq2](https://doi.org/10.1103/qs15-cyq2)

I. INTRODUCTION

The exponential growth of the Hilbert space with system size constitutes a fundamental challenge in quantum many-body physics, rendering direct simulation intractable beyond modest particle numbers [1]. Despite this barrier, approximate methods enable accurate simulations of systems containing tens to hundreds of particles across quantum chemistry and condensed matter physics [2–6]. These successes rely on exploiting the structure of the underlying physical models through tailored representations of the relevant quantum states, which compress the information of the exponentially large Hilbert space into a manageable form while remaining useful for specific tasks.

Modern electronic-structure methods exemplify this principle through diverse approaches [7–11]. Tensor networks leverage entanglement area laws to identify a relevant state manifold within the Hilbert space, enabling algorithms such as the density matrix renormalization group (DMRG) [12–15]. Density functional theory (DFT) reformulates the many-body problem in terms of the electron density, whose dimensionality scales linearly with the system size, at the cost of approximate exchange-correlation functionals [16–18]. Variational two-electron reduced density matrix (v2RDM) methods [19–21] provide an alternative formulation based directly on reduced density matrices, where the energy is minimized over the space of two-body RDMs subject to (approximate) N -representability constraints that ensure consistency with a quantum state [22,23]. From this perspective, DMRG provides an explicit compressed representation of the

wavefunction, while DFT and v2RDM approaches define implicit representations in terms of reduced quantities. In v2RDM, consistency with the underlying state is enforced through representability constraints, whereas in DFT it follows from the one-to-one mapping between density and Hamiltonian (within a fixed class). Each approach leverages specific physical features, locality, symmetries, or correlation structure, to define either explicit ansätze or implicit constraints that make compression possible. Indeed, for arbitrary Hamiltonians where any state can be a ground state, no universal compression scheme can exist.

These handcrafted representations, while powerful, are inherently limited by their design principles. Tensor networks excel for states with mostly short-range entanglement, DFT on weakly correlated systems, and coupled cluster methods for capturing dynamic correlation. Strongly correlated systems, characterized by competing interactions that preclude simple effective descriptions, remain particularly challenging [24]. This limitation motivates a fundamental question: Can compressed representations be learned directly from data, adapting to the specific system and task of interest rather than relying on predetermined structures?

Recent advances in machine learning offer a natural route toward this goal [25–33]. Representation learning techniques aim to automatically extract informative descriptors from data, transforming information into formats that are optimal for downstream objectives, such as prediction, interpretability, or transferability [34–37]. Autoencoding frameworks are an example of this, as they discover compact encodings by mapping their input through a bottleneck and then reconstructing it, revealing underlying factors of variation while discarding redundancy in the data [38–41].

In this article, we introduce an unsupervised machine learning framework that learns minimal compressed representations of quantum many-body ground states, optimized for variational energy minimization in the latent space. The model automatically discovers the essential degrees of freedom of the family of Hamiltonians on which it is trained, without prior knowledge of the underlying physics. We apply

*Contact author: f.frohnert@liacs.leidenuniv.nl

†Contact author: polla@lorentz.leidenuniv.nl

this framework to the Fermi-Hubbard model in the strongly correlated regime and investigate the natural compression limit and its relation to the system's intrinsic degrees of freedom [42–44]. The representations are learned from reduced density matrices computed via exact diagonalization, thereby isolating the analysis from errors introduced by approximate data. Our approach enables determining the ground-state energies of unseen Hamiltonian instances through variational energy optimization directly in the latent space. In contrast to v2RDM, we achieve this without the need to explicitly enforce N -representability constraints, as valid physical structure is learned implicitly from data.

This work bridges two complementary perspectives: the physics-driven identification of efficient quantum state descriptors and the data-driven discovery of task-specific representations through machine learning. The resulting framework demonstrates that learned compressions can serve both as practical variational ansätze for many-body calculations and as diagnostic tools for complexity in ground state- physics. While the Fermi-Hubbard model provides a controlled test bed in which the relevant degrees of freedom are well defined, the broader goal is to develop a methodology applicable to complex quantum systems, such as molecular, material, or experimental many-body settings, where the relevant effective variables are not known *a priori*.

II. METHODS

The overarching goal of our machine learning framework is to learn compressed representations of ground states for a given family of Hamiltonians that can serve as ansätze for variational energy optimization, as illustrated in Fig. 1.

Dataset and task. We consider the family of one-dimensional (1D) Fermi-Hubbard models on L sites with N electrons, parametrized by the site-dependent chemical potential $\mu = (\mu_1, \dots, \mu_L)$:

$$\begin{aligned} \hat{H}(\mu) = & -t \sum_i^L \sum_{\sigma} (\hat{c}_{i\sigma}^{\dagger} \hat{c}_{i+1\sigma} + \text{H.c.}) \\ & + U \sum_{i=1}^L \hat{n}_{i\uparrow} \hat{n}_{i\downarrow} + \sum_{i=1}^L \mu_i \sum_{\sigma} \hat{n}_{i\sigma}, \end{aligned} \quad (1)$$

with fixed hopping $t = 1$ and interaction strength $U = 4$. This matches the strongly correlated regime, which is especially challenging to solve on lattices of dimensionality 2 or larger [45]. We restrict the analysis to the block with fixed particle number $\sum_{i,\sigma} \hat{n}_{i\sigma} = N$ (choosing N at or close to half filling, $N = L/2$), zero spin projection $S_z = 0$, and total spin singlet $S^2 = 0$. We limit ourselves to system sizes $L \leq 14$ that can be solved by exact diagonalization.

The dataset for our representation learning task is generated by sampling random potentials μ from the distribution described in Ref. [46] (see Appendix A). For each instance, we compute the ground state $|\psi(\mu)\rangle$ via exact diagonalization and extract the vector of Hamiltonian expectation values:

$$\omega = \left\{ \left(\left\langle \sum_{\sigma} \hat{c}_{i\sigma}^{\dagger} \hat{c}_{i+1\sigma} \right\rangle, \left\langle \sum_{\sigma} \hat{n}_{i\sigma} \right\rangle, \langle \hat{n}_{i\uparrow} \hat{n}_{i\downarrow} \rangle \right) \right\}_{i=1, \dots, L}. \quad (2)$$

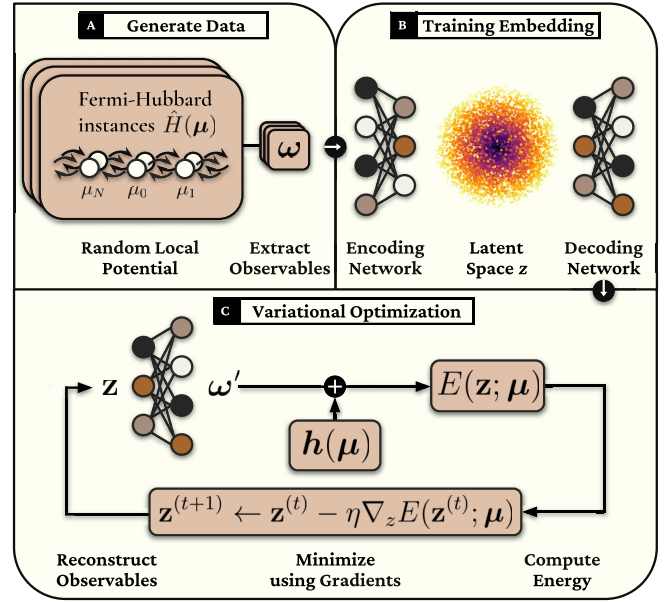


FIG. 1. Overview: Workflow for learning minimal representations of fermionic ground states. (A) The process starts by generating instances of the Hubbard Hamiltonian $\hat{H}(\mu)$ from random local potentials μ . For each instance, the expectation values of the Hamiltonian terms ω are computed via exact diagonalization. (B) The resulting dataset is used to train a neural-network-based autoencoder that compresses ω into a low-dimensional latent representation $\mathbf{z}$. (C) The decoder defines a differentiable mapping from latent space to the expectation values of Hamiltonian terms, and consequently to physical observables such as the energy. This property enables energy optimization (using any gradient-based optimizer) directly in the latent space for new, unseen potentials.

This representation provides a compact description of the underlying states in which the total energy is recovered exactly via the contraction of ω with the Hamiltonian coefficients $h(\mu) := \{(t, U, \mu_i)\}_{i=1, \dots, L}$:

$$\langle \hat{H}(\mu) \rangle = h(\mu) \cdot \omega. \quad (3)$$

Crucially, unlike the exponentially large wavefunction, the dimensionality of ω scales linearly with L .

The motivation for using this representation as the starting point for our machine-learning compression framework is twofold. First, while ω constitutes the minimal representation required to exactly extract the energy, the dimensionality of the underlying parameters generating the data is strictly smaller than the dimensionality of the Hamiltonian term description ($\dim(\mu) < \dim(\omega)$), hinting toward additional compressibility. Second, the target application for our models is variational energy optimization; yet we cannot optimize within the space of ω directly. Direct variational minimization is fundamentally precluded by the N -representability problem: Arbitrary vectors in ω space do not necessarily map to physical quantum states. Our compression scheme aims to automatically identify a manifold in ω space that implicitly enforces these physical constraints.

Our methodology is not limited to Hamiltonian term expectation values; we explore the compression of another meaningful set of observables, the spin-adapted two-particle

reduced density matrix (2-RDM), in Appendix I. The 2-RDM encodes all two-body correlation functions and serves as the fundamental variable in v2RDM [22,23,47] methods. The 2-RDM provides a richer, though higher-dimensional, representation of the ground states from which all two-particle observables (including the energy) can be computed.

Architecture. We train neural-network autoencoders [38,39] to learn compressed representations of Hubbard model ground states. Below, we outline the structure and training of the best-performing models; the motivation behind each design choice is discussed in the ablation study (Appendix E). The models comprise an encoder network, $E_\theta : \omega \mapsto \mathbf{z}$, and a decoder network, $D_\phi : \mathbf{z} \mapsto \omega'$, where ω denotes the input observables and $\mathbf{z} \in \mathbb{R}^d$ the latent representation, in which the latent space dimensionality d is a hyperparameter of the model. The encoder compresses the input through four nonlinear fully connected layers, reducing its initial dimension (set by the input size) to a d -dimensional latent bottleneck. The decoder mirrors this structure, reconstructing the observables from the compressed representation.

The model is trained to minimize the mean-squared reconstruction loss

$$\mathcal{L}_{\text{rec}} = \|\omega - D_\phi(E_\theta(\omega))\|_2^2. \quad (4)$$

The data are standardized to ensure numerical stability during training. We employ Optuna [48] to guide hyperparameter optimization, tuning both architectural choices and training parameters. All architectural and training choices are free of system-specific or physics-inspired priors. Full details on the machine learning model are provided in Appendix B.

Regularization. To ensure stable training and learned latent representations that are geometrically well behaved, we augment the reconstruction loss with four regularization terms.

First, we employ a radial well loss that bounds the latent norm to a finite radius, preventing unbounded encodings [38]. This constraint helps maintain stable representations by limiting the magnitude of latent vectors:

$$\mathcal{L}_{\text{well}} = \max(0, \|\mathbf{z}\|_2 - r)^2, \quad (5)$$

where r denotes the radius parameter that defines the maximum allowed norm in latent space.

Second, we incorporate a contrastive repulsion term to preserve the relative distances between distinct ground states in latent space, preventing degeneracy or collapse of the representation. This regularization follows principles inspired by contrastive learning methods [36], ensuring that distinct inputs maintain meaningful separation in the learned representation. For a batch $\{(\omega_i, \mathbf{z}_i)\}$, the repulsion loss is formulated as

$$\mathcal{L}_{\text{repel}} = \sum_{i < j} \frac{\|\omega_i - \omega_j\|_2^2}{\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \varepsilon} \bigg/ \sum_{i < j} \|\omega_i - \omega_j\|_2^2, \quad (6)$$

where ε is a small constant for numerical stability.

Finally, we apply Lipschitz regularization to the encoder and decoder weights to promote smoothness and improve training stability. This approach constrains the Lipschitz constant of each layer, limiting how rapidly the network output can change with respect to its inputs. Conceptually, this has an effect similar to weight decay, since both penalize large

weight magnitudes; however, Lipschitz regularization more directly controls the global smoothness of the learned mapping. Empirically, this encourages smoother latent-to-output interpolations and stabilizes gradients during optimization. Following a variant of the formulation in Ref. [49], we implement per-layer trainable Lipschitz log bounds, which are softly enforced via a differentiable constraint:

$$\mathcal{L}_{\text{lip}} = \sum_i \log_{10} \text{softplus}(c_i), \quad (7)$$

where c_i denotes a trainable log bound on the ∞ norm of the weight matrix $\hat{W}_i$ for layer i . Each layer's weight matrix $\hat{W}_i$ is defined by rescaling the trainable parameter matrix W_i according to this bound,

$$\hat{W}_i = \text{diag}\left(\min\left[1, \frac{\text{softplus}(c_i)}{\sum_k |(W_i)_{rk}|}\right]\right) W_i, \quad (8)$$

so that $\text{softplus}(c_i)$ defines an adaptive upper limit on the layer's Lipschitz constant. The softplus function ensures positivity while maintaining smooth gradients, allowing each layer to adaptively learn its Lipschitz limit. This form of regularization yields a decoder that is both smoother and more robust than standard weight-decayed networks [49].

The total training objective combines and weights all terms,

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{well}} + \beta \mathcal{L}_{\text{repel}} + \gamma \mathcal{L}_{\text{lip}}^{\text{enc}} + \delta \mathcal{L}_{\text{lip}}^{\text{dec}}, \quad (9)$$

where $\mathcal{L}_{\text{lip}}^{\text{enc}}$ and $\mathcal{L}_{\text{lip}}^{\text{dec}}$ are the Lipschitz losses [Eq. (7)] separately computed on the encoder and decoder layers, respectively. This aims to learn a bounded, uniformly covered, and smooth latent manifold that encodes meaningful physical properties. The ablation study in Appendix E illustrates how each regularization term affects this goal.

Energy optimization. After training, the decoder D_ϕ defines a differentiable mapping from the latent representation $\mathbf{z}$ to the corresponding Hamiltonian terms or observables. For a given potential μ , the total energy can then be expressed as

$$E_\phi(\mathbf{z}; \mu) = \mathbf{h}(\mu) \cdot D_\phi(\mathbf{z}), \quad (10)$$

where $\mathbf{h}(\mu)$ denotes the potential-dependent coefficients of the Hamiltonian and $D_\phi(\mathbf{z})$ outputs the corresponding expectation values. This formulation yields a fully differentiable map from latent space to energy, enabling the direct optimization of the latent variable to identify the ground-state representation for a given potential:

$$\mathbf{z}^* = \arg \min_{\mathbf{z}} E_\phi(\mathbf{z}; \mu). \quad (11)$$

This optimization can be interpreted as a variational procedure over the manifold of states implicitly defined by the decoder D_ϕ . The optimization over $\mathbf{z}$ can be performed efficiently by exploiting gradients $\nabla_{\mathbf{z}} E_\phi(\mathbf{z}^{(t)}; \mu)$ obtained through automatic differentiation, while keeping the decoder parameters fixed. This allows the latent space itself to serve as a search domain for new ground-state configurations under varying external conditions.

We emphasize that this is not a trivial task. One might, in principle, attempt to minimize the energy directly with respect to the Hamiltonian expectation values ω , but this is fundamentally obstructed by the N -representability problem:

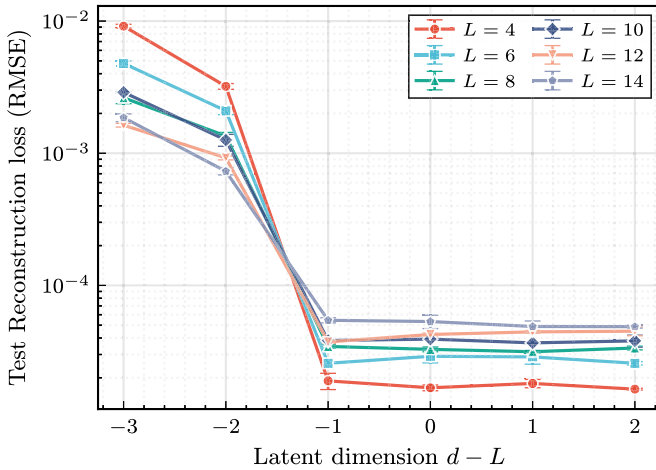


FIG. 2. Compression threshold: Test reconstruction loss (RMSE) of Hamiltonian term vectors as a function of latent dimension d . Data points represent the mean and standard deviation across three independent models. A sharp decrease in error is observed at $d = L - 1$, identifying a compression threshold that coincides with the intrinsic number of independent degrees of freedom in the half-filled Hubbard ground state. This transition marks the point at which the autoencoder fully captures the physically relevant subspace of Hamiltonian observables.

Not every collection of expectation values corresponds to a physical N -electron wavefunction. Enforcing the required N -representability constraints would entail imposing a hierarchy of semidefinite conditions, the computational cost of which grows prohibitively with the system size [19,23,47,50,51]. Our approach of optimizing in the learned latent space operates independently of these explicit conditions. Although the decoder is not explicitly constrained to generate N -representable states, a high quality of reconstruction and a sufficiently regular latent space can result in the optimization naturally remaining within or close to the physical ground-state manifold.

III. RESULTS

Learning the Intrinsic Structure of the Ground-State Manifold. We first train and test the specified autoencoders on a set $2L \times 10^5$ of Hamiltonian term expectation values ω for even system sizes $L = [4, \dots, 14]$ at half filling $N = L/2$. We train the models for latent dimensions $d \in [L - 3, \dots, L + 2]$, while keeping all other architectural hyperparameters fixed. This controlled scaling isolates the effect of the information bottleneck size on the representational capacity of the model.

Figure 2 visualizes the test reconstruction loss as a function of latent dimension d . We plot the mean and standard deviation over three independent models with random initialization. Across all system sizes, the loss decreases with increasing d , exhibiting a sharp drop at $d = L - 1$, followed by saturation for $d \geq L - 1$. This sudden improvement marks a compression threshold: For $d < L - 1$, the latent space is insufficient to capture all relevant physical variations, while additional dimensions beyond $L - 1$ provide no further benefit.

This behavior mirrors the physical structure of the underlying Hamiltonian family: The number of degrees of freedom in

the definition of $\hat{H}(\mu)$ corresponds to L , i.e., the dimension of μ . However, as the particle number is fixed, a constant shift in the potential $\mu_i \mapsto \mu_i + c$ will shift the Hamiltonian by a constant $c \times N$ without affecting the ground state. Thus, the number of independent degrees of freedom for the ground states of this Hamiltonian family is $L - 1$. Empirically, the autoencoder autonomously identifies the true number of independent degrees of freedom, discovering through unsupervised compression that the ground-state manifold is intrinsically $(L - 1)$ dimensional. The transition at $d = L - 1$ reveals a direct correspondence between empirically optimal compression and the physical structure of the Hubbard model, demonstrating that at this point the learned latent representation effectively captures the topology of the ground-state family covered by the training distribution with minimal redundancy.

In Appendix H, we learn and interpret optimal compressed representations of systems with uneven system size L and N a multiple of 4. These systems have a degenerate ground state for $\mu = 0$, which imposes a topologically nontrivial structure on the ground state manifold. We show that this $U(1)$ ring structure is captured in the latent representation, indicating that properties of the ground-state manifold can be inferred directly from the latent space.

Latent Geometry of the Ground-State Representations. Having identified the compression threshold at $d = L - 1$, we now examine the geometry of the learned latent representation at this critical dimension. Figure 3 illustrates feature-wise projections of the latent coordinates for test data (for a representative system size $L = N = 6$), color coded by the strength of the external potential μ . The potential μ serves as the generative parameter used to sample distinct ground states; visualizing the latent encodings thus reveals how states at different potentials are represented. As the mean potential $\bar{\mu} = \sum_j \mu_j$ shifts Eq. (10) only by a constant and does not influence optimization, we quantify the strength of the potential by its standard deviation, $\sigma(\mu) = \|\mu - \bar{\mu}\|_2$. Color coding by this quantity highlights how ground states with varying potential strengths are organized within the latent space. The set of encoded latent vectors forms a nearly uniform sphere of radius $r \approx 2$. This geometry arises from the well loss in Eq. (5), which confines embeddings to a bounded spherical region, while the contrastive repulsion in Eq. (6) encourages uniform coverage. Notably, the color map reveals a smooth radial gradient: Points near the center correspond to weaker potentials, whereas those near the outer boundary represent stronger potentials. The emergence of such structure implies that the encoding extracts physically meaningful information from the Hamiltonian terms ω . Hence, the latent geometry demonstrates that the autoencoder learns a compact and physically organized representation of the ground-state manifold.

While this visualization exposes correlations with the potential strength, it does not, by itself, reveal which physical quantities fundamentally shape the latent space. In Appendix G, we further analyze how the learned representation relates to the potential and the density, showing that it cannot be reduced to either descriptor alone.

Variational Energy Optimization in Learned Latent space. Having characterized the geometry of the latent space, we

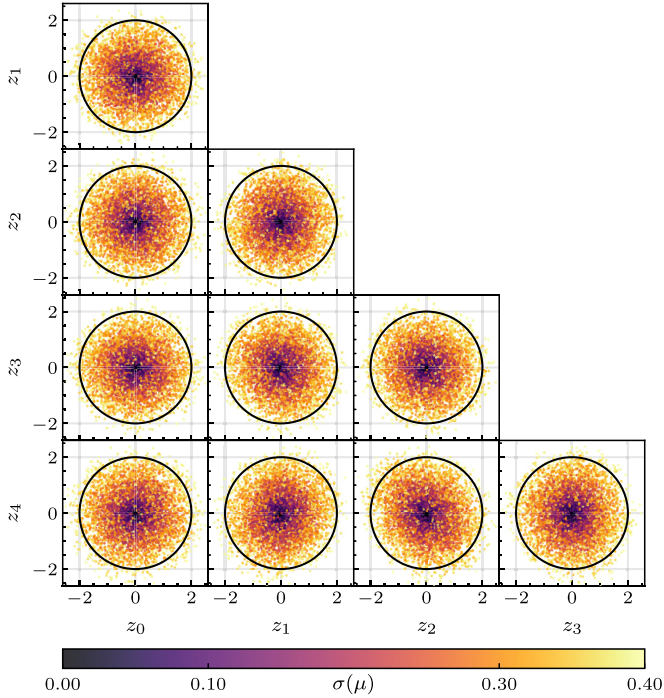


FIG. 3. Visualizing the learned representation: Latent space of the autoencoder for $L = N = 6$, color coded by the strength of the potential μ . The circular structure in the feature-wise projections results from the well constraint and uniform coverage enforced by contrastive repulsion, while the smooth radial color gradient shows that states with similar potential strengths cluster together, indicating that the latent geometry captures physically meaningful structure in the ground-state manifold.

now test the learned representation as a variational ansatz for ground-state energy search. For a given test-set potential μ , we perform the minimization in Eq. (11) using limited-memory Broyden–Fletcher–Goldfarb–Shanno [52] with a learning rate of 8×10^{-2} , starting from $\mathbf{z}_0 = \mathbf{0}$ and iterating until the convergence condition $\|\partial_{\mathbf{z}} E\|_2 \leq 3 \times 10^{-4}$ is met. We repeat this procedure on 10^4 test potentials.

To monitor potential divergence of the optimization procedure, we introduce a simple flag based on the latent-space regularization in Eq. (5). During training, latent vectors are softly constrained to $\|\mathbf{z}\|_2 \leq r = 2$, as illustrated in Fig. 3. If optimization pushes a point *significantly* beyond this range (we choose $\|\mathbf{z}^*\|_2 > 3$), it is marked as diverged, indicating that the energy might be much lower than physical and the result cannot be trusted.

Figure 4 illustrates the root-mean-squared energy error of the unflagged points, together with the fraction of flagged points, as a function of latent dimension d . We observe that the error drops sharply as d increases and reaches its minimum at $d = L - 1$, matching the compression threshold identified in Sec. III. At this dimension, the latent variables span the physically sufficient manifold supporting the Hubbard ground states. For $d > L - 1$, the error increases again: The additional latent directions introduce redundancy and weakly constrained, potentially unphysical degrees of freedom that deteriorate the conditioning of the variational landscape, even though the reconstruction quality remains

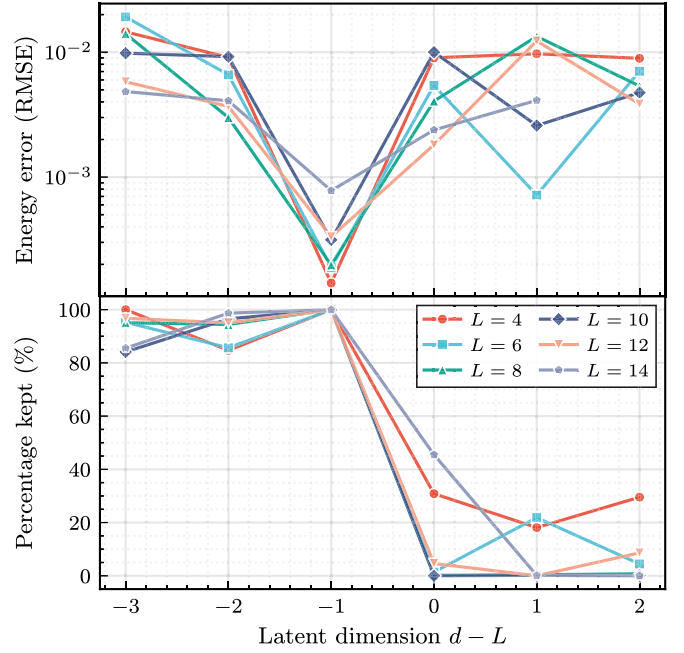


FIG. 4. Energy optimization threshold: Error (RMSE) of non-diverged optimized energies vs latent dimension d . Energies are obtained via gradient-based minimization in the learned latent space, using the decoder as a differentiable variational ansatz. A distinct error minimum occurs at $d = L - 1$, matching the intrinsic degrees of freedom of the half-filled Hubbard ground state. For $d \geq L$, performance degrades as excess latent dimensions introduce unconstrained, unphysical directions. Lower panel: Percentage of optimization trajectories retained after discarding diverged solutions (note: for $L = 14$ at the largest d , optimization diverged on all samples). Results shown for even system sizes $L \in [4, \dots, 14]$ at $U/t = 4$.

high. In these regimes, the optimizer often diverges, pushing $\|\mathbf{z}\|_2 \gg r$ far beyond the valid region identified by regularization, in an attempt to minimize the energy. Importantly, for the model at $d = L - 1$ (i.e., the minimal- d model that accurately represents the ground state manifold), optimization converges on 100% of the tested cases. This demonstrates that this model also represents the ground-state manifold *tightly*; i.e., it locally only represents valid states.

These results demonstrate that the autoencoder learns a latent representation that can serve as a practical variational manifold. At $d = L - 1$, the optimization landscape is empirically well conditioned and yields accurate ground-state energies without imposing explicit physical constraints. For $d \geq L$, weakly constrained latent directions reduce stability, indicating that higher-dimensional latent spaces require additional regularization.

IV. DISCUSSION AND OUTLOOK

In this article, we have introduced an unsupervised machine-learning framework that learns compressed representations of fermionic quantum many-body ground states. For the considered family of Hubbard Hamiltonians, we identify the minimal latent dimension $d = L - 1$ required to learn a faithful representation. This dimensionality matches the number of intrinsic degrees of freedom of the model. For

this choice of latent dimension, the representations of valid ground states form a smooth and bounded manifold reproducing the topology of the system's degrees of freedom (an $L - 1$ -dimensional ball). The decoder can then serve as a stable and differentiable ansatz for valid ground states, enabling variational energy optimization. Further increasing the latent dimension does not improve the quality of the learned representation. This, instead, introduces irrelevant directions in the latent space, orthogonal to the manifold of valid ground-state representations, degrading the performance of the decoder as a variational ansatz. Together, these results demonstrate that learned latent representations can serve both as probes of intrinsic dimensionality and as practical variational ansätze.

The analysis of the encoder's feature sensitivity (Appendix G) indicates a strong reliance on local density and on-site terms, effectively rediscovering the importance of key variables central to Density Functional Theory. This establishes a natural connection to ML-assisted DFT approaches [53–56] that leverage neural networks to refine exchange-correlation approximations.

The Hubbard model offers an ideal starting point for our framework: Its degrees of freedom are known *a priori*, providing a controlled test bed for benchmarking latent-space compression and interpretability. Additional complexity could be introduced by extending the considered family of Hubbard Hamiltonians to include varying interaction ratios U/t ; we expect this would require one additional latent dimension to capture the extra degree of freedom. However, because the parameter space is inherently anisotropic—i.e., changing the global U/t impacts the underlying ground state distinctively compared to local potential shifts μ_i —we anticipate that the resulting latent manifold will exhibit a more complex structure. A natural next step is to apply this framework to systems where the relevant degrees of freedom are not analytically known, such as models of molecules and correlated materials, or results from quantum simulators. In such cases, the learned latent representation could act as a data-driven probe that reveals how many and which degrees of freedom actually matter for describing a family of ground states.

A variational ansatz learned from data could become useful if it can be applied to systems larger than those on which it was trained. Future work may explore how such scalability can be achieved using geometric deep learning architectures, such as convolutional networks or transformers. This requires incorporating weight sharing and equivariance mechanisms that explicitly exploit the physical structure and symmetries of the underlying states. To extend our framework to more complex systems, the training pipeline could be adapted to rely on data generated by methods that are more efficient than exact diagonalization, such as tensor network techniques or quantum algorithms, and to incorporate datasets produced by a variety of computational approaches.

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank V. Zakharov, P. Emonts, A. Dawid, B. van Zwol, E. Costa, E. van Nieuwenburg, and V. Dunjko for fruitful discussions. This work was supported by the Dutch National Growth Fund (NGF), as part of the Quantum Delta NL pro-

gram. S.P. and E.K. acknowledge support from Shell Global Solutions BV. Parts of this work were performed using the computational resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

The views and opinions expressed here are solely those of the authors and do not necessarily reflect those of the funding institutions. Neither of the funding institutions can be held responsible for them.

DATA AVAILABILITY

The code, data, and model weights are available in the GitHub release [57].

APPENDIX A: DATASET DETAILS

To generate training data for the neural-network-based autoencoder, we sample Hubbard model ground states at varying random potentials μ using the procedure detailed in Ref. [46] (an adaptation of the procedure from Ref. [56]), which we recap here.

The first step in generating a data point is to sample a random potential μ . To do this, we first sample a strength parameter W uniformly at random from the interval $[0.005t, 2.5t]$. We then sample the on-site potential μ uniformly at random within the range $[-W, W]$ and calculate its standard deviation $\sigma(\mu) = \sqrt{\langle \mu_j^2 \rangle - \langle \mu_j \rangle^2}$, where the average is taken over all sites j . If $\sigma(\mu) < 0.4t$, we accept the potential and proceed; otherwise, we reject it and repeat the sampling procedure from the beginning. This acceptance criterion produces a representative distribution of potentials with approximately uniformly distributed standard deviations below $0.4t$, ensuring we sample both weak and strong potentials irrespective of the system size L .

Once a potential is accepted, we employ exact diagonalization to compute the ground state of the 1D Hubbard model at the given potential using the PySCF Full Configuration Interaction module [58]. From the converged ground state $|\psi(\mu)\rangle$, we compute the 2-RDM

$$\Gamma_{pqrs}(\mu) := \sum_{\sigma\tau} \langle \psi(\mu) | c_{p\sigma}^\dagger c_{r\tau}^\dagger c_{s\tau} c_{q\sigma} | \psi(\mu) \rangle. \quad (\text{A1})$$

We then extract the expectation values of Hamiltonian terms from the 2-RDM: the spin-summed on-site density $\langle \sum_\sigma \hat{n}_{i\sigma} \rangle$, the nearest-neighbor hopping correlators $\langle \sum_\sigma \hat{c}_{i\sigma}^\dagger \hat{c}_{i+1,\sigma} \rangle$, and the on-site interaction $\langle \hat{n}_{i\uparrow} \hat{n}_{i\downarrow} \rangle$. This procedure is repeated for $N_{\text{inst}} = 10^5$ independent potentials to generate the full dataset. Stacking these three observables for all realizations yields a dataset of shape $(N_{\text{inst}}, 3L)$. Finally, by exploiting translational and reflection symmetries, we increase the effective size of the training dataset by a factor of $2L$ (this symmetry augmentation procedure is not applied to validation and test set).

APPENDIX B: NEURAL NETWORK DETAILS

To learn unsupervised compressions of the ground-state data, we employ an autoencoder with a symmetric encoder-decoder design. The input $\omega \in \mathbb{R}^{n_{\text{in}}}$ contains the three local Hamiltonian terms across L sites (thus, $n_{\text{in}} = 3L$). The

encoder maps these observables to a latent vector $\mathbf{z}$ through four fully connected layers with softplus activations, progressively reducing the dimensionality to a d -dimensional bottleneck: $E_\theta : \omega \mapsto \mathbf{z}$. Here, d is a tunable hyperparameter controlling the size of the latent space. The decoder mirrors this structure, reconstructing the observables from the latent representation, $D_\phi : \mathbf{z} \mapsto \omega'$, using four hidden layers with softplus activations that expand from dimension d back to n_{in} . To keep the total parameter count largely independent of d , we impose a minimum hidden-layer width of $m_d = L^2$. The hidden-layer dimensions follow a geometric progression that interpolates smoothly between the narrow latent bottleneck and the wide reconstruction layers. Specifically, the first decoder layer starts at width m_d , and the final hidden layer terminates at λn_{in} , where the same scaling factor $\lambda = 20$ is also applied symmetrically to the encoder's first hidden layer relative to the input dimension. This ensures a balanced expansion and contraction around the latent space. This architecture provides a controlled and flexible framework for learning compact latent representations of the ground-state observables.

For the system sizes considered ($L = 4$ to $L = 14$), this architecture yields models with approximately 3×10^4 to 7×10^5 trainable parameters, scaling primarily with the input size. For the compression of the two-body reduced density matrix, which has significantly higher input dimensionality, the largest model contains approximately 5×10^7 parameters.

Training the largest models considered in this study empirically requires additional architectural modifications to prevent overfitting and ensure stable optimization. We implement this by soft constraining the Lipschitz constant of the layers, which has the additional conceptual benefit of encouraging the model to learn a smoother mapping between latent space and feature space [49]. This is achieved with a modification of the fully connected layers of the model and an additional term in the loss function. To encourage Lipschitz smoothness, each layer i applies row-wise weight normalization controlled by a learnable parameter c_i . For a weight matrix W_i , a bias b_i , and input x , the forward pass computes

$$\hat{W}_i = \text{diag}\left(\min\left[1, \frac{\text{softplus}(c_i)}{\sum_k |(W_i)_{ik}|}\right]\right) W_i, \quad (\text{B1})$$

$$y = \text{softplus}(\hat{W}_i x + b_i), \quad (\text{B2})$$

where the subscript r indexes the rows of W_i . This normalization constrains the ℓ^∞ operator norm of each layer, with $\text{softplus}(c_i)$ serving as an adaptive, learnable upper bound on the maximum row sum. The Lipschitz regularization loss in Eq. (7) penalizes large Lipschitz bounds, encouraging the network to use the minimal Lipschitz constant necessary for accurate reconstruction. This term is added to the reconstruction loss during training with a hyperparameter weight γ , δ , and the model is implemented using the normalization layers.

Importantly, the neural network architecture introduces no inductive bias beyond that inherent to a standard feedforward network and does not encode any explicit physics prior that would shape the learned representation.

APPENDIX C: REGULARIZATION DETAILS

We augment the reconstruction loss used to train neural network autoencoders with four regularization terms designed to yield a compact, smoothly structured latent manifold that meaningfully represents the ground-state physics while remaining amenable to downstream optimization tasks.

To prevent unbounded latent encodings and ensure the learned representation occupies a compact region, we apply a radial well potential as introduced in Eq. (5), which penalizes latent vectors $\mathbf{z}$ that exceed radius r from the origin. This soft confinement is essential for stable energy optimization in latent space, as it prevents the encoder from learning arbitrarily large embeddings that would complicate gradient-based search.

However, this confinement alone does not guarantee a well-structured latent space. Without additional constraints, the encoder could adopt a trivial solution in which all states collapse to a small region near the origin, satisfying the well loss while destroying any meaningful geometric structure. This phenomenon, known as latent collapse, causes distinct physical states to map to nearly identical latent codes, making the topology of the ground-state manifold less accessible and rendering the learned representation useless for downstream tasks. To prevent this collapse and preserve the intrinsic geometry of the input space, we employ a contrastive repulsion loss [36]. For a batch $\{(\omega_i, \mathbf{z}_i)\}_{i=1}^N$ of input states and their encodings, we penalize pairs that are well separated in input space yet close in latent space, as introduced in Eq. (6). This term encourages approximately isometric embeddings: Pairs of states separated by a distance d in input space should remain roughly a distance d apart in latent space. The normalization by the average input distance makes the loss scale invariant across different system sizes and potential strengths.

As introduced in Appendix B, models trained on larger system sizes benefit from weight regularization. Thus, to promote smooth latent-to-feature mappings and stabilize both training and subsequent latent-space optimization, we constrain the Lipschitz constant of the decoder network [49]. For each decoder layer ℓ with weight matrix W_ℓ , we introduce a learnable scalar parameter c_ℓ that bounds the layer's operator norm. Following the row-wise normalization scheme described in Appendix B, we minimize the loss in Eq. (7) that penalizes large Lipschitz bounds while allowing the network to adaptively learn the minimal smoothness required for accurate reconstruction. This regularization prevents exploding gradients during training and ensures that small perturbations in latent space correspond to small changes in the decoded output, a helpful property for gradient-based energy minimization.

The total training objective combines these terms:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{well}} + \beta \mathcal{L}_{\text{repe}} + \gamma \mathcal{L}_{\text{lip}}^{\text{enc}} + \delta \mathcal{L}_{\text{lip}}^{\text{dec}}, \quad (\text{C1})$$

where we set $\alpha = 10^{-7}$, $\beta = 10^{-7}$, $\gamma = 10^{-9}$, and $\delta = 10^{-8}$ across all experiments. These coefficients balance the regularization effects in relation to reconstruction accuracy.

Importantly, the mean squared error objective and additional regularization terms act solely as generic optimization criteria, without encoding any explicit physics prior.

APPENDIX D: TRAINING AND HYPERPARAMETER DETAILS

We implement all models in PyTorch [59], train them on graphics processing unit (GPU) using the Adam optimizer [60] with a batch size of 256, and use an 81/9/10 train-validation-test split. Training runs for up to 3×10^3 epochs with early stopping: If the validation loss fails to improve by at least $\Delta = 10^{-10}$ for 30 consecutive epochs, training terminates.

The learning rate schedule begins at $\eta_0 = 4 \times 10^{-3}$ and adapts using PyTorch's ReduceLROnPlateau scheduler, which reduces η by a factor of 0.5 whenever the validation loss plateaus (improvement $< 10^{-8}$) for more than ten epochs. This adaptive schedule allows the optimizer to make aggressive updates early in training while refining the solution with smaller steps as convergence approaches.

Hyperparameters, including the initial learning rate, batch size, number of hidden layers D , width scaling factor λ , minimum layer width m_d , and regularization coefficients $\{\alpha, \beta, \gamma, \delta\}$, are selected with the guidance of the Optuna framework [48]. The best-performing configuration from this search is then used to train the final model reported in all results.

All reported performance metrics are evaluated on a held-out test set of potential configurations $\{\mu\}$ that were not seen during training or hyperparameter tuning.

APPENDIX E: ABLATION STUDY

To understand the contribution of each architectural and regularization component to the model's performance, we conduct a systematic ablation study on the $L = N = 6$ system. We compare six model configurations, evaluating both reconstruction accuracy and optimization performance. For each configuration, we train five independent models with different random initializations to quantify variability, reporting mean and standard deviation across these runs. We evaluate the following model variants:

(1) *Full model*: Nonlinear encoder and decoder with $D = 4$ hidden layers, trained with all regularization terms ($\mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{well}} + \beta \mathcal{L}_{\text{repl}} + \gamma \mathcal{L}_{\text{lip}}^{\text{enc}} + \delta \mathcal{L}_{\text{lip}}^{\text{dec}}$).

(2) *Reconstruction*: Full architecture but trained with $\mathcal{L}_{\text{rec}}$ alone, removing all regularization ($\alpha = \beta = \gamma = \delta = 0$). This tests whether regularization provides benefits beyond accurate reconstruction.

(3) *No Lipschitz regularization*: Trained without smoothness constraints ($\gamma = \delta = 0$), while retaining well confinement and contrastive repulsion. This isolates the effect of enforcing smooth latent-to-output mappings.

(4) *No well loss*: Trained without confinement ($\alpha = 0$), allowing unbounded latent encodings. This tests whether compact latent regions are necessary for effective optimization.

(5) *No contrastive repulsion*: Trained without the anticollapse term ($\beta = 0$). This examines whether preventing latent

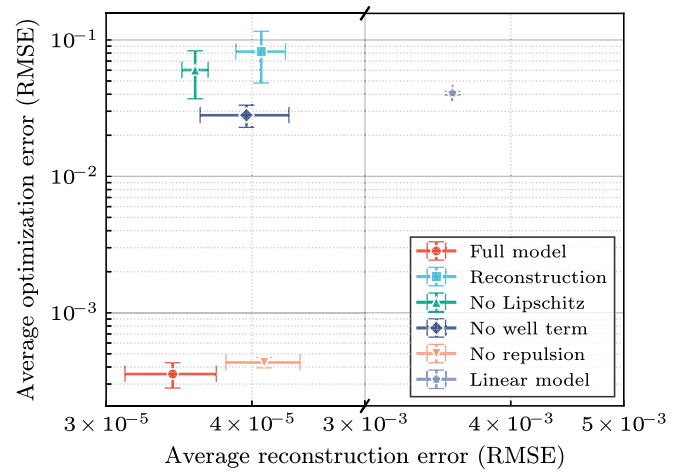


FIG. 5. Ablation study: Performance comparison of model variants on systems with $L = N = 6$ and latent dimensions $d = L - 1$. We evaluate six distinct model configurations in terms of reconstruction error and optimization performance. Error bars show the standard deviation over five independently trained models with different random initializations. The full model consistently outperforms all ablated variants, indicating that each regularization component contributes meaningfully to both accurate reconstruction and effective energy optimization.

collapse is critical for preserving the ground-state manifold geometry.

(6) *Linear model*: Replaces the nonlinear layers in encoder and decoder with a single linear layer. This tests whether nonlinearity in the model is essential.

All models use identical latent dimensions $d = L - 1$, training procedures, and hyperparameters (excluding the ablated components) to ensure a fair comparison. We evaluate each model configuration using two complementary metrics. First, we measure the reconstruction error on a held-out test set, which quantifies how faithfully the autoencoder can compress and reconstruct the Hamiltonian-term features. Second, to assess how well the learned latent space supports energy minimization, we perform gradient-based optimization in latent space to determine ground states for previously unseen potential configurations. Importantly, both metrics are purely data driven, reflecting reconstruction quality learned from the dataset without incorporating any explicit physics-based assumptions or inductive biases into the model.

Figure 5 summarizes reconstruction error versus optimization error for all model variants, with error bars showing the standard deviation across five random initializations. The results reveal several key observations.

Reconstruction Error. All nonlinear models, independent of which regularization terms were enabled, achieve reconstruction errors of comparable magnitude. This shows that the nonlinear encoder-decoder architecture is generally capable of faithfully representing the Hamiltonian-term features within the latent bottleneck of size $d = L - 1$. Among these, the full model achieves the best reconstruction performance, indicating that each regularization component contributes positively. In contrast, the linear model performs two orders of

magnitude worse, confirming that it lacks the expressivity required for the task.

Optimization Error. The full model provides the best overall trade-off, achieving the lowest optimization error. In contrast, the reconstruction-only model performs the worst. Although it achieves reasonable reconstruction accuracy, its unregularized and unbounded latent space leads to poor optimization behavior. The model without Lipschitz regularization maintains good reconstruction performance but also shows degraded optimization behavior. From this, we can infer that encouraging Lipschitz smoothness improves latent-to-output geometry: Small steps in latent space correspond to controlled, predictable changes in the reconstructed observables, which is essential for stable and reliable minimization. The model without the well constraint suffers from a similar issue. Without this constraint, the latent space can expand in an unbounded or highly nonuniform manner due to the repulsion term, making the gradient-based search for the minimum energy substantially harder. The model without the repulsion term shows solid optimization performance but slightly reduced reconstruction quality. This suggests that repulsion primarily helps structure the latent space in a way that improves reconstruction with only a mild influence on optimization when the other regularizers are still present. Interestingly, despite its poor reconstruction, the linear model achieves an optimization error comparable to the reconstruction-only baseline. We attribute this to its convex and well-behaved latent geometry, which facilitates gradient-based minimization even though the representation is less optimal for accurate reconstruction.

Overall, these ablations illustrate how each architectural component contributes to a latent space that is both expressive enough for accurate reconstruction and structured enough to support robust energy optimization.

APPENDIX F: TRAINING DATA EFFICIENCY

In the main experiments, the autoencoder was trained on $2L \times 10^5$ Hamiltonian term expectation values ω , generated by 10^5 unique potentials μ . A natural question is how the amount of training data required to learn optimal compression scales with the system size L . Figure 6 illustrates the relationship between the size of the training and validation set (already accounting for the $2L \times$ increase from the symmetry augmentation described in Appendix A) and the reconstruction loss for various system sizes, evaluated at the critical latent dimension $d = L - 1$. The point at which the reconstruction error drops sharply provides an approximate measure of the data required for efficient learning at this compression threshold. We observe a marginal increase in test reconstruction error for smaller system sizes when using the largest dataset sizes. We attribute this minor fluctuation to both the fixed hyperparameter configuration, which was optimized for the baseline dataset size of 10^5 points, and the training dynamics of the smaller model architecture (approximately 3×10^4 parameters for $L = 6$). Crucially, this discrepancy remains well below the convergence tolerance of 3×10^{-4} root mean squared error (RMSE) used for variational energy optimization.

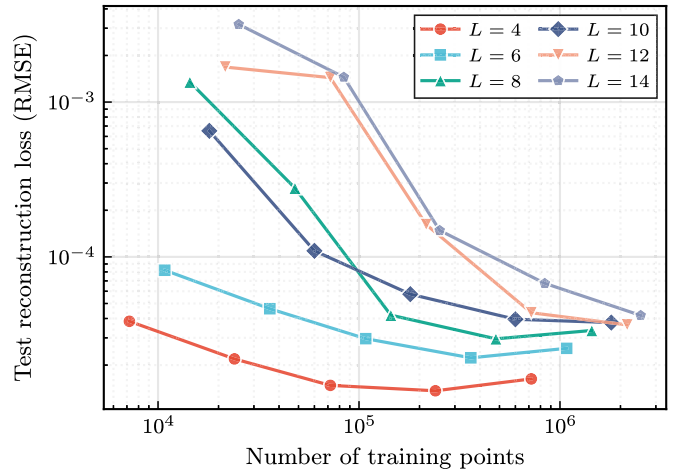


FIG. 6. Training data efficiency: Root-mean-squared error reconstruction loss as a function of the size of the training and validation dataset (accounting for symmetry augmentation) for different system sizes at the optimal latent dimension $d = L - 1$. The curves indicate how the data requirement grows with system size to achieve high-quality compression and display the best out of three training instances.

APPENDIX G: INTERPRETING THE ENCODER

While the visualization in Fig. 3 demonstrates that the learned representation is compact and varies smoothly with potential strength, it does not reveal which physical features fundamentally structure the latent space.

Interpreting deep nonlinear encoders is inherently challenging. Unlike in linear models, individual weights are not directly meaningful, and the effect of any input feature can vary substantially across different regions of the input space. In machine-learning research, this challenge is often addressed using gradient-based saliency maps [61–63] or SHapley Additive exPlanations-style local attribution methods [64,65], which quantify how sensitive a model’s output is to perturbations of each input feature.

For our encoder, the natural analog of such attribution tools is the Jacobian

$$J(\omega) = \partial E_\theta(\omega) / \partial \omega. \quad (G1)$$

This matrix measures how each latent coordinate responds to infinitesimal changes in each input observable at a specific sample, thus providing a precise notion of local feature relevance for nonlinear mappings. The main limitation of this approach is the same locality: If the encoder were strongly nonlinear, the Jacobian would vary substantially across input samples, making a single global interpretation unreliable. We emphasize that achieving more global interpretability would likely require architectural or regularization choices explicitly designed to constrain the encoder toward mechanistically transparent representations, which remains an important direction for future work.

To assess the feasibility of a global interpretation, we compute the Jacobian for a representative set of 10^4 ground-state data points for the $L = 6$, $d = L - 1$ model. We then analyze the mean and standard deviation of each Jacobian element [Fig. 7(a)]. In this visualization, the color map represents the

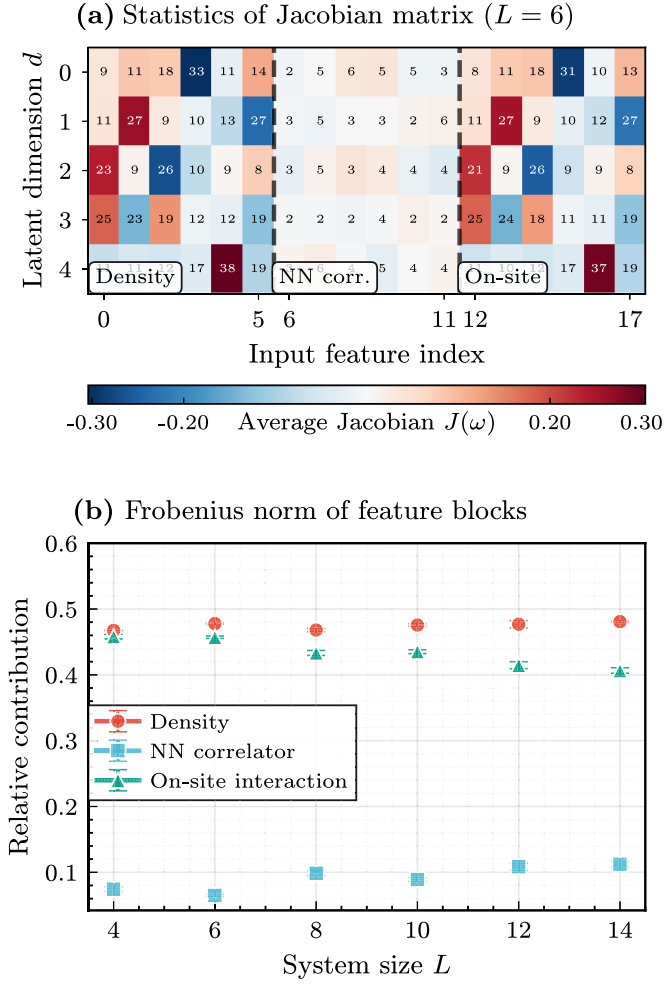


FIG. 7. Interpreting the trained encoder: (a) Average Jacobian matrix and variability. Visualization of the mean Jacobian matrix $J(\omega)$ of the encoder for the $L = N = 6$ model (latent dimension $d = L - 1$) computed over 10^4 input samples. The color map represents the mean magnitude (μ), while the overlaid numbers represent the sample standard deviation (σ), scaled by a factor of 10^3 . The consistently small variability ($\sum \sigma_{i,j} / \sum |\mu_{i,j}| \approx 16\%$) across the dataset suggests that the encoder has learned a stable, close to linear mapping from the input feature space to the latent representation. Distinct patterns are visible across the three input feature blocks: Spin-summed density and on-site interaction exhibit larger average gradients, while the nearest-neighbor pair correlator appears less influential. (b) Relative feature importance via Frobenius norm. Due to the inherent correlation between input features and the linear dependencies within the latent space, the element-wise Jacobian can only offer limited insight into feature importance. To provide a robust, rotationally invariant measure of effective sensitivity, we calculate the relative Frobenius norm contribution of each feature's Jacobian submatrix (G2) as a function of system size L (at fixed latent dimension $d = L - 1$). The error bars shown are averaged over three independently trained models. This heuristic confirms a consistent trend across system sizes: The spin-summed density and on-site interaction are the dominant features encoding information into the latent representation, while the nearest-neighbor correlator remains the least significant.

mean magnitude (μ), and the overlaid numbers denote the sample standard deviation (σ), which is scaled by a factor of 10^3 for improved readability.

Two key observations emerge from this analysis: First, the Jacobians vary only weakly across the dataset. The total relative variability $\sum \sigma_{i,j} / \sum |\mu_{i,j}| \approx 16\%$ across the dataset is small, indicating that the encoder applies nearly the same linearized transformation throughout the data manifold. This near-constancy implies that the encoder is close to a linear function, and the Jacobian can be interpreted almost like the weight matrix of a linear encoder. Second, the average Jacobian magnitude clearly shows three distinct blocks corresponding to spin-summed density, on-site interaction, and nearest-neighbor correlators. Density and on-site terms exhibit consistently larger magnitudes. This means that perturbations in these observables produce stronger change in the latent representation. In contrast, the nearest-neighbor correlator block has substantially smaller magnitudes, indicating that these features play a comparatively minor role in shaping the learned latent coordinates.

In summary, because the encoder operates in an effectively linear regime over the data manifold, the Jacobian magnitudes provide an interpretable measure of feature importance. The observed block structure highlights which physical observables dominate the latent representation and gives insight into how the encoder processes the input features.

To obtain a robust characterization of feature influence across different system sizes, we evaluate the relative Frobenius norm contribution of each feature's Jacobian submatrix. This metric, r_f , is defined as

$$r_f = \frac{\|J_f(\omega)\|_F}{\sum_{f'} \|J_{f'}(\omega)\|_F}, \quad (\text{G2})$$

where $J_f(\omega)$ denotes the block of the Jacobian matrix associated with feature f . The Frobenius norm $\|J_f(\omega)\|_F = \sqrt{\sum_{i,j} (J_f^{ij}(\omega))^2}$ serves as a natural measure of the total sensitivity of the latent representation to changes in feature f , aggregating information across all latent dimensions and spatial positions.

The Frobenius norm captures the overall strength of the feature's influence, preventing cancellation that occurs when averaging raw, signed gradients and is invariant to orthogonal transformations of the latent basis. The relative normalization ensures that contributions sum to unity, enabling direct comparison of feature importance across different system sizes and architectural choices.

Figure 7(b) shows these relative Frobenius norm contributions as a function of system size for $d = L - 1$, with the mean and standard deviation over three independently trained models. A consistent trend is observed: Density and on-site interaction terms dominate the learned representation across all sizes, while the contribution of the nearest-neighbor correlator remains low. This pattern suggests that the autoencoder primarily relies on local quantities that directly encode the potential and interaction structure of the Hamiltonian, with longer-range correlations only weakly influencing the

compact latent description. The narrow error bars confirm that this hierarchy of importance is robust.

APPENDIX H: DEGENERACIES IN GROUND-STATE MANIFOLD

In Appendix III, we studied system sizes $L \in \{4, \dots, 14\}$ with even particle numbers at half filling. To explore the limits of our compression approach, we also test our framework on systems with an odd number of sites L , close to half filling $N = L \pm 1$. In particular, we report the results for systems with $L = 5, N = 4$ and $L = 7, N = 8$. We identified these systems as exhibiting different behavior from others during training. In particular, they achieve a slightly worse reconstruction loss ($\approx 1e-4$) compared to other systems of a similar size ($\approx 3e-5$) for $d \geq L - 1$.

For these systems, Fig. 8(a) illustrates the feature-wise projections of the latent space at the critical dimension $d = L - 1$ and Fig. 8(b) the reconstruction loss as a function of the latent dimension. The results reveal markedly different behavior compared to the even- L systems studied previously. We observe that the encoded points extend well beyond the radial well boundary r , indicating that the confinement regularization is less effective. Moreover, the distribution is highly anisotropic. The feature projections show that the topology of the representation manifold is that of a ring, rather than the spherical one observed for even- L systems (see Fig. 3).

This qualitative difference reflects an underlying quasidegeneracy in the spectrum of the quantum system. In Fermi-Hubbard chains with periodic boundary conditions, with $N = 4k$ (where $k \in \mathbb{N}$) and odd L , the singlet ground state of the zero-potential system $\bar{\mu} = 0$ is doubly degenerate. (This can be observed numerically and explained by analyzing the model Hamiltonian in momentum space. The degenerate ground states can be labeled by the total momentum, which is a good quantum number when $\mu = 0$.) We define $|\phi_A\rangle$ and $|\phi_B\rangle$ as two orthonormal vectors forming a basis for the ground subspace of $H(\mu = 0)$. While a nonzero disorder potential μ lifts this degeneracy, the energy splitting is perturbatively small for potentials $\delta\mu$ with weak fluctuations $\sigma(\delta\mu) \ll t$. This symmetry breaking generates arbitrary combinations of the two degenerate zero-potential states for perturbatively small potentials; thus, the ground state of $H(\delta\mu)$ can be written as $a|\phi_A\rangle + b|\phi_B\rangle + o(\sigma(\delta\mu))$. As the considered model Hamiltonians can be represented as real symmetric matrices, the coefficients $a, b \in \mathbb{R}$; equivalently, $a = \sin \theta, b = \cos \theta$. Thus, the manifold of ground states for infinitesimal potentials has an $SO(2)$ structure (i.e., a circle). The latent representation faithfully captures this structure, with the ring topology visible in Fig. 3. The inner edge of the ring corresponds to the ground states of models with a weakly fluctuating potential. This is precisely the behavior we expect for a system exhibiting disorder-induced symmetry breaking of a doubly degenerate ground state.

This analysis demonstrates that reconstruction quality, as well as latent space visualization, serves as a diagnostic for ground-state complexity beyond simple compressibility. Systems with accidental or near-degeneracies present a fundamentally harder learning problem: The autoencoder must either (1) commit to one branch of the degenerate manifold,

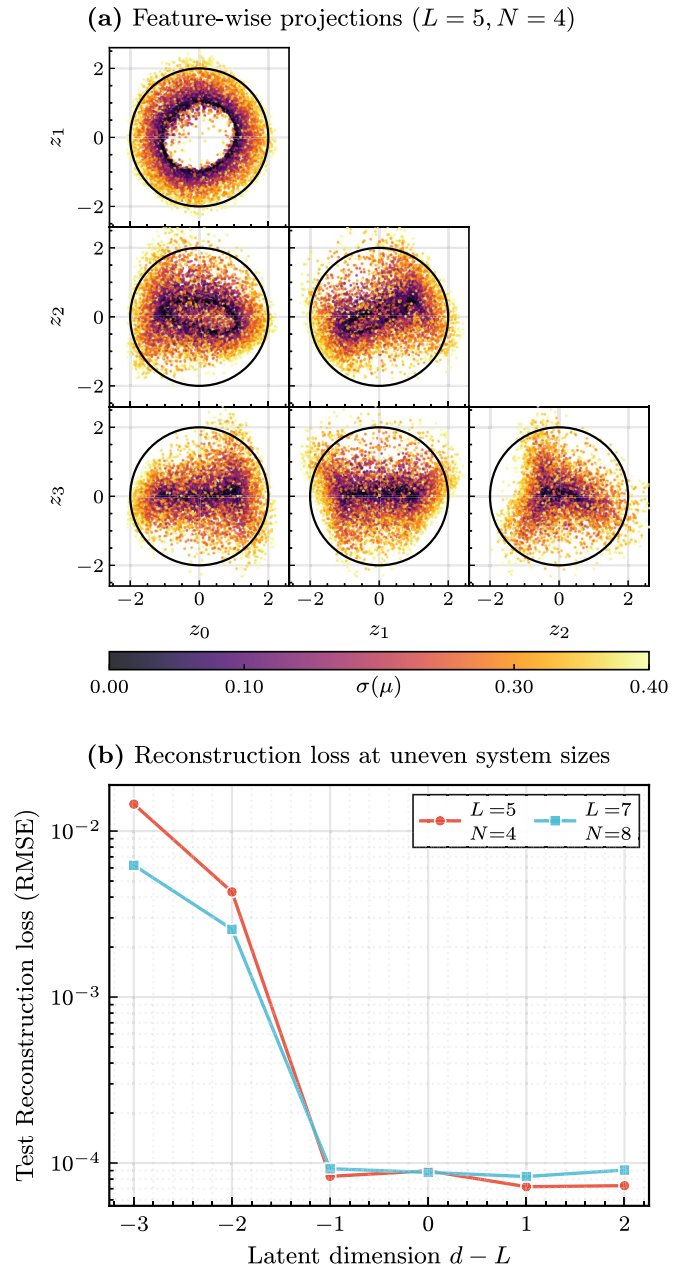


FIG. 8. Compression of degenerate systems: (a) Latent space of the autoencoder for $L = 5, N = 4$, color coded by the potential strength μ . Unlike the $L = N$ systems, the data do not exhibit the same circular structure in the feature-wise projections, indicating qualitatively different underlying physics. The ringlike structure illustrates an underlying degeneracy in the ground-state manifold. (b) Test reconstruction loss as a function of latent dimension d . A sharp decrease in reconstruction error is again observed at $d = L - 1$, but the overall reconstruction quality is comparatively worse than for the even systems.

breaking the underlying symmetry, or (2) represent the full degenerate subspace, which require additional latent dimensions or nonlinear manifold structures (such as the observed rings). The even- L systems studied in the main text avoid this issue due to the absence of such degeneracies at generic disorder strengths, enabling clean, low-dimensional compressed representations.

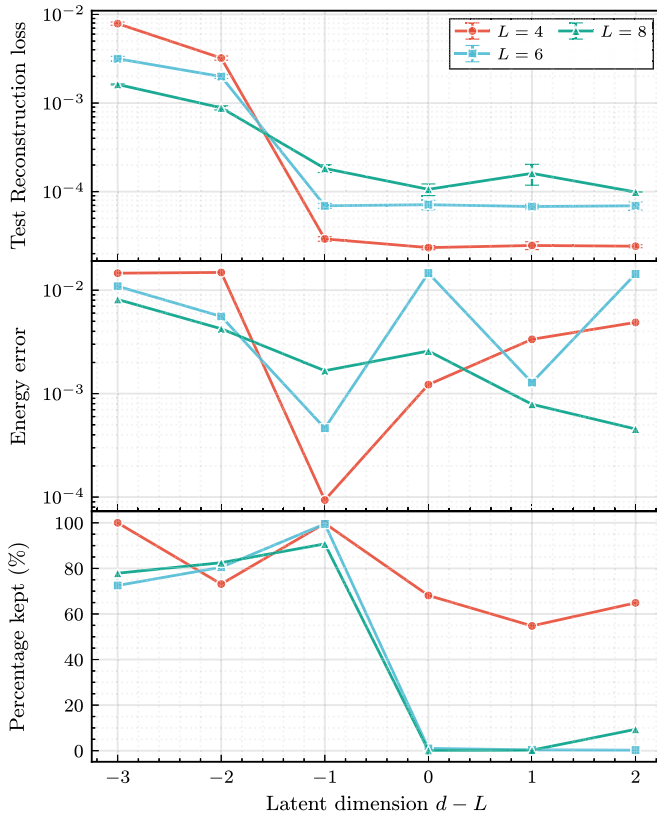


FIG. 9. Compression and energy optimization of two-body reduced density matrices: (Top) Test reconstruction RMSE (mean and standard deviation over three models) vs latent dimension d . A sharp drop at $d = L - 1$ confirms that the 2-RDM shares the same intrinsic degrees of freedom as the Hamiltonian terms, despite its larger dimensionality. (Middle) Energy optimization error using the decoder as a variational ansatz. The critical dimension $d = L - 1$ yields the optimal performance; for $d \geq L$, the unconstrained latent directions degrade the optimization stability. (Bottom) Fraction of optimization trajectories retained after rejecting solutions that violate the stability bound $\|\mathbf{z}^*\| < r_{\text{opt}}$. Data shown for system sizes $L \in [4, 6, 8]$ at $U/t = 4$.

APPENDIX I: COMPRESSION OF TWO-BODY REDUCED DENSITY MATRICES

While the Hamiltonian terms ω provide a compact representation sufficient for energy evaluation, they capture only limited information about ground-state correlations.

A strictly more informative representation is the two-particle reduced density matrix introduced in Eq. (A1), which encodes all two-body correlation functions of the ground state

and finds application in various quantum chemistry settings [22,23].

We now apply the same unsupervised autoencoding framework to learn compressed representations of 2-RDMs instead of Hamiltonian terms. Since the 2-RDM is a substantially larger object with dimension L^4 , we are limited to system sizes $L \leq 8$ due to memory and training constraints. We use 10^5 samples without symmetry augmentation to train, validate, and test the models. Figure 9 (top) shows the reconstruction error versus latent dimension for 2-RDM autoencoders on systems $L \in [4, 6, 8]$. We observe qualitatively similar compression behavior to Hamiltonian terms (Fig. 2): A critical dimension emerges at $d = L - 1$, below which reconstruction error does not decrease significantly. We observe a minor deviation for the $L = 8$ system, which we attribute to the limited training data available relative to the high dimensionality of the 2-RDM feature space. Consistent with the data efficiency analysis in Fig. 6, we anticipate that 2-RDM models would benefit from expanding the effective dataset size via symmetry augmentation. We leave the detailed exploration of this scaling regime for future study.

Despite the 2-RDM's larger input dimensionality, the optimal latent dimension remains $d = L - 1$, identical to that of the Hamiltonian terms expectation values. This again confirms that the intrinsic dimensionality of the ground-state manifold is determined by the underlying physical degrees of freedom and not by the representation's size. Since the Hamiltonian terms can be extracted as linear contractions of the 2-RDM and yield the same compression threshold, we conclude that the higher-order correlation information encoded in the 2-RDM does not add independent degrees of freedom for the Hubbard ground states studied here. The two-body correlations beyond those captured in local observables lie in a redundant subspace that does not require additional latent dimensions to represent.

We perform the same latent-space energy optimization introduced in Sec. III using a trained 2-RDM autoencoder. Figure 9 (middle and bottom panels) demonstrates that the optimization achieves qualitatively similar performance to Hamiltonian term compression, tested for 800 trial potentials. We attribute this to the high reconstruction fidelity achieved during training; the autoencoder learns to stay within the physically accessible subspace of the 2-RDM space, implicitly satisfying the N -representability conditions [66,67] without explicit enforcement. Figure 9 (bottom) again illustrates that for latent dimensions $d > L - 1$, the optimizer pushes solutions toward the boundary of the allowed region, with $\|\mathbf{z}\|_2 \approx r_{\text{opt}}$, in an attempt to minimize the energy, which results in an increasing number of rejected results.

- [1] R. P. Feynman, Simulating physics with computers, *Int. J. Theor. Phys.* **21**, 467 (1982).
- [2] J. Tindall, M. Fishman, E. M. Stoudenmire, and D. Sels, Efficient tensor network simulation of IBM's Eagle kicked Ising experiment, *PRX Quantum* **5**, 010308 (2024).
- [3] S. Patra, S. S. Jahromi, S. Singh, and R. Orús, Efficient tensor network simulation of IBM's largest quantum processors, *Phys. Rev. Res.* **6**, 013326 (2024).
- [4] A. Menczer, M. van Damme, A. Rask, L. Huntington, J. Hammond, S. S. Xantheas, M. Ganahl, and O. Legeza, Parallel implementation of the density matrix renormalization group method achieving a quarter petaFLOPS performance on a single DGX-H100 GPU node, *J. Chem. Theory Comput.* **20**, 8397 (2024).
- [5] J. Brabec, J. Brandeys, K. Kowalski, S. Xantheas, Ö. Legeza, and L. Veis, Massively parallel quantum chemical density

- matrix renormalization group method, *J. Comput. Chem.* **42**, 534 (2021).
- [6] P. R. Nagy and M. Kállay, Approaching the basis set limit of CCSD(T) energies for large molecules with local natural orbital coupled-cluster methods, *J. Chem. Theory Comput.* **15**, 5275 (2019).
 - [7] U. Schollwöck, The density-matrix renormalization group in the age of matrix product states, *Ann. Phys.* **326**, 96 (2011).
 - [8] G. K.-L. Chan and S. Sharma, The density matrix renormalization group in quantum chemistry, *Annu. Rev. Phys. Chem.* **62**, 465 (2011).
 - [9] R. Orús, Tensor networks for complex quantum systems, *Nat. Rev. Phys.* **1**, 538 (2019).
 - [10] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory*, 1st ed. (Wiley, Chichester, UK, 2000).
 - [11] C. D. Sherrill, Frontiers in electronic structure theory, *J. Chem. Phys.* **132**, 110902 (2010).
 - [12] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
 - [13] S. R. White, Density-matrix algorithms for quantum renormalization groups, *Phys. Rev. B* **48**, 10345 (1993).
 - [14] F. Verstraete, J. I. Cirac, and V. Murg, Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems, *Adv. Phys.* **57**, 143 (2008).
 - [15] J. Eisert, M. Cramer, and M. B. Plenio, Colloquium: Area laws for the entanglement entropy, *Rev. Mod. Phys.* **82**, 277 (2010).
 - [16] K. Burke and L. O. Wagner, DFT in a nutshell, *Int. J. Quantum Chem.* **113**, 96 (2013).
 - [17] K. Burke, Perspective on density functional theory, *J. Chem. Phys.* **136**, 150901 (2012).
 - [18] N. Mardirossian and M. Head-Gordon, Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals, *Mol. Phys.* **115**, 2315 (2017).
 - [19] M. Nakata, H. Nakatsuji, M. Ehara, M. Fukuda, K. Nakata, and K. Fujisawa, Variational calculations of fermion second-order reduced density matrices by semidefinite programming algorithm, *J. Chem. Phys.* **114**, 8282 (2001).
 - [20] D. A. Mazziotti, Variational two-electron reduced density matrix theory for many-electron atoms and molecules: Implementation of the spin- and symmetry-adapted T_2 condition through first-order semidefinite programming, *Phys. Rev. A* **72**, 032510 (2005).
 - [21] A. Eugene DePrince, III, Variational determination of the two-electron reduced density matrix: A tutorial review, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **14**, e1702 (2024).
 - [22] A. J. Coleman, Structure of fermion density matrices, *Rev. Mod. Phys.* **35**, 668 (1963).
 - [23] P.-O. Löwdin, Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction, *Phys. Rev.* **97**, 1474 (1955).
 - [24] A. Georges, Strongly correlated electron materials: Dynamical mean-field theory and electronic structure, *AIP Conf. Proc.* **715**, 3 (2004).
 - [25] A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, *et al.*, Modern applications of machine learning in quantum sciences, in *Machine Learning in Quantum Sciences* (Cambridge University Press, Cambridge, 2025).
 - [26] A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo, and S. Severini, Learning hard quantum distributions with variational autoencoders, *npj Quantum Inf.* **4**, 28 (2018).
 - [27] V. Dunjko and H. J. Briegel, Machine learning & artificial intelligence in the quantum domain: A review of recent progress, *Rep. Prog. Phys.* **81**, 074001 (2018).
 - [28] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
 - [29] J. Zang, M. Medvidović, D. Kiese, D. D. Sante, A. M. Sengupta, and A. J. Millis, Machine learning-based compression of quantum many body physics: PCA and autoencoder representation of the vertex function, *Mach. Learn.: Sci. Technol.* **5**, 045076 (2024).
 - [30] L. M. Sager-Smith and D. A. Mazziotti, Reducing the quantum many-electron problem to two electrons with machine learning, *J. Am. Chem. Soc.* **144**, 18959 (2022).
 - [31] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
 - [32] M. Sajjan, J. Li, R. Selvarajan, S. H. Sureshbabu, S. S. Kale, R. Gupta, V. Singh, and S. Kais, Quantum machine learning for chemistry and physics, *Chem. Soc. Rev.* **51**, 6475 (2022).
 - [33] J. R. Moreno, G. Carleo, and A. Georges, Deep learning the Hohenberg-Kohn maps of density functional theory, *Phys. Rev. Lett.* **125**, 076402 (2020).
 - [34] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Int.* **35**, 1798 (2013).
 - [35] S. N. Alshammari, M. Hamilton, A. Feldmann, J. R. Hershey, and W. T. Freeman, A unifying framework for representation learning, in *International Conference on Learning Representations* (ICLR, 2025).
 - [36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *International Conference on Machine Learning* (PMLR, 2020), pp. 1597–1607.
 - [37] E. Costa, G. Scriva, and S. Pilati, Solving deep-learning density functional theory via variational autoencoders, *Mach. Learn.: Sci. Technol.* **5**, 035015 (2024).
 - [38] S. J. D. Prince, *Understanding Deep Learning* (MIT Press, Cambridge, MA, 2023).
 - [39] S. Chen and W. Guo, Auto-Encoders in Deep Learning—A Review with New Perspectives, *Mathematics* **11**, 1777 (2023).
 - [40] F. Møller, G. Fernández-Fernández, T. Schweigler, P. d. Schoulepnikoff, J. Schmiedmayer, and G. Muñoz-Gil, Learning minimal representations of many-body physics from snapshots of a quantum simulator, *Phys. Rev. Res.* **8**, 023094 (2026).
 - [41] F. Frohnert and E. van Nieuwenburg, Explainable representation learning of small quantum states, *Mach. Learn.: Sci. Technol.* **5**, 015001 (2024).
 - [42] J. Hubbard, Electron correlations in narrow energy bands, *Proc. R. Soc. London A* **276**, 238 (1963).
 - [43] D. P. Arovas, E. Berg, S. A. Kivelson, and S. Raghu, The Hubbard model, *Annu. Rev. Condens. Matter Phys.* **13**, 239 (2022).
 - [44] L. W. Cheuk, M. A. Nichols, K. R. Lawrence, M. Okan, H. Zhang, E. Khatami, N. Trivedi, T. Paiva, M. Rigol, and M. W.

- Zwierlein, Observation of spatial charge and spin correlations in the 2D Fermi-Hubbard model, *Science* **353**, 1260 (2016).
- [45] J. P. F. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson, C. A. Jiménez-Hoyos, *et al.*, Solutions of the two-dimensional Hubbard model: Benchmarks and results from a wide range of numerical algorithms, *Phys. Rev. X* **5**, 041041 (2015).
- [46] E. Koridon, F. Frohnert, E. Prehn, E. van Nieuwenburg, J. Tura, and S. Polla, Learning density functionals from noisy quantum data, *Mach. Learn.: Sci. Technol.* **6**, 025020 (2025).
- [47] L. H. Delgado-Granados, L. M. Sager-Smith, K. Trifonova, and D. A. Mazziotti, Machine learning of two-electron reduced density matrices for many-body problems, *J. Phys. Chem. Lett.* **16**, 2231 (2025).
- [48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19* (ACM Press, New York, NY, 2019), pp. 2623–2631.
- [49] H.-T. D. Liu, F. Williams, A. Jacobson, S. Fidler, and O. Litany, Learning smooth neural functions via Lipschitz regularization, in *ACM SIGGRAPH 2022: Conference Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques, Vancouver, BC, Canada* (ACM Press, New York, NY, 2022), pp. 1–13.
- [50] M. Nakata, B. J. Braams, K. Fujisawa, M. Fukuda, J. K. Percus, M. Yamashita, and Z. Zhao, Variational calculation of second-order reduced density matrices by strong N -representability conditions and an accurate semidefinite programming solver, *J. Chem. Phys.* **128**, 164113 (2008).
- [51] D. A. Mazziotti, Variational minimization of atomic and molecular ground-state energies via the two-particle reduced density matrix, *Phys. Rev. A* **65**, 062511 (2002).
- [52] D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, *Math. Program.* **45**, 503 (1989).
- [53] G. Luise, C.-W. Huang, T. Vogels, D. P. Kooi, S. Ehlert, S. Lanius, K. J. H. Giesbertz, A. Karton, D. Gunceler, M. Stanley *et al.*, Accurate and scalable exchange-correlation with deep learning, [arXiv:2506.14665](https://arxiv.org/abs/2506.14665).
- [54] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, Pushing the frontiers of density functionals by solving the fractional electron problem, *Science* **374**, 1385 (2021).
- [55] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Finding density functionals with machine learning, *Phys. Rev. Lett.* **108**, 253002 (2012).
- [56] J. Nelson, R. Tiwari, and S. Sanvito, Machine learning density functional theory for the Hubbard model, *Phys. Rev. B* **99**, 075132 (2019).
- [57] S. Polla, F. Frohnert, and E. Koridon, StefanoPolla/DFTQML: Machine learning for quantum chemistry, GitHub, 2026, <https://github.com/StefanoPolla/DFTQML/releases/v2.0>.
- [58] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan, PySCF: The Python-based simulations of chemistry framework, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **8**, e1340 (2018).
- [59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, New York, 2019), Vol. 32.
- [60] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *International Conference on Learning Representations* (ICLR, 2015).
- [61] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, SmoothGrad: Removing noise by adding noise, [arXiv:1706.03825](https://arxiv.org/abs/1706.03825).
- [62] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in *International Conference on Learning Representations Workshop* (ICLR Workshop, 2014).
- [63] M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks, in *Proceedings of the 34th International Conference on Machine Learning (ICML)* (PMLR, 2017), pp. 3319–3328.
- [64] B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, The Shapley value in machine learning, in *Proceedings of the 31st International Joint Conference on Artificial Intelligence* (IJCAI, 2022), pp. 5572–5579.
- [65] S. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems* (NeurIPS, 2017), Vol. 30.
- [66] D. A. Mazziotti, Structure of fermionic density matrices: Complete n -representability conditions, *Phys. Rev. Lett.* **108**, 263002 (2012).
- [67] D. A. Mazziotti, Two-electron reduced density matrix as the basic variable in many-electron quantum chemistry and physics, *Chem. Rev.* **112**, 244 (2012).