

Linear Foundation Model for Quantum Embedding: Data-Driven Compression of the Ghost Gutzwiller Variational Space

Samuele Giuli ¹, Hasanat Hasan ^{2,3}, Benedikt Kloss ¹, Marius S. Frank ⁴, Tsung-Han Lee ⁵,
Olivier Gingras ^{1,6}, Yong-Xin Yao ^{7,8,*} and Nicola Lanatà ^{2,1,†}

¹Center for Computational Quantum Physics, *Flatiron Institute*, New York, New York 10010, USA

²School of Physics and Astronomy, *Rochester Institute of Technology*, Rochester, New York 14623, USA

³The Graduate Center, *City University of New York*, New York, New York 10016, USA

⁴Department of Chemistry, *Aarhus University*, 8000 Aarhus C, Denmark

⁵Department of Physics, *National Chung Cheng University*, Chiayi 62102, Taiwan

⁶Université Paris-Saclay, CNRS, CEA, *Institut de Physique Théorique*, 91191 Gif-sur-Yvette, France

⁷Ames National Laboratory, Ames, Iowa 50011, USA

⁸Department of Physics and Astronomy, *Iowa State University*, Ames, Iowa 50011, USA



(Received 5 January 2026; accepted 24 April 2026; published 28 July 2026)

Simulations of quantum matter rely mainly on Kohn-Sham density functional theory (DFT), which often fails for strongly correlated systems. Quantum embedding (QE) theories address this limitation by mapping the system onto an auxiliary embedding Hamiltonian (EH) describing fragment-environment interactions, but the EH is typically large and its iterative solution is the primary computational bottleneck. We introduce a linear foundation model for QE that utilizes principal component analysis to compress the space of quantum states needed to solve the EH within a small variational subspace. Using a data-driven active-learning scheme, we learn this subspace from EH ground states and reduce each embedding solution to a deterministic ground-state eigenvalue problem in the reduced space. Within the ghost-Gutzwiller approximation, we show for a three-orbital Hubbard model that a variational space learned on a Bethe lattice is transferable to square and cubic lattices without additional training, while substantially reducing the cost of the EH step. We further validate the approach on plutonium, where a single variational space reproduces the energetics of all six crystalline phases while reducing the cost of the EH solution by orders of magnitude. This provides a practical route to overcome the main computational bottleneck of QE frameworks, paving the way for high-throughput *ab initio* simulations of strongly correlated materials at a near-DFT cost.

DOI: [10.1103/phk3-pvbm](https://doi.org/10.1103/phk3-pvbm)

I. INTRODUCTION

Theoretical and computational tools capable of simulating solids and molecules “*ab initio*”—based solely on the quantum-mechanical interactions between their elementary constituents—are deeply integrated across scientific disciplines, from condensed matter physics and materials science to chemistry and biochemistry. By connecting emergent behaviors to microscopic origins, these tools enable interdisciplinary research and the development of new materials [1,2]. Kohn-Sham (KS) density functional theory (DFT) serves as the standard workhorse for many such calculations [3,4]. However, strong electron correlations [5] can drastically influence the electronic and structural properties of

materials, particularly in systems containing 3d transition metals, lanthanides, and actinides [6–15]. These strong electron correlations, related to the “multiconfigurational problem” in chemistry, underlie critical phenomena including magnetism and high-temperature superconductivity [16–23].

Quantum embedding (QE) theories provide a framework to calculate the electronic structure of extended strongly correlated systems. In QE methods, the interacting lattice or molecular problem is mapped onto a local fragment (impurity) coupled to an environment described by an embedding Hamiltonian (EH), which encodes how the fragment exchanges particles and correlations with the rest of the system. Examples include DFT combined with dynamical mean-field theory (DMFT) [24–29], density matrix embedding theory (DMET) [30,31], rotationally invariant slave-boson theory (RISB) [32–34], the multiorbital Gutzwiller approximation (GA) [35–43], and the ghost-Gutzwiller approximation (ghost-GA) [44–47]. These methods have demonstrated predictive power for correlated materials, but their widespread application in fields such as materials science and computational catalysis is hindered by a severe computational bottleneck: the solution of the EH. In iterative workflows such as charge self-consistency loops or structural optimizations, the EH

*Contact author: ykent@iastate.edu

†Contact author: nxlsps@rit.edu

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

parameters are updated recursively, which requires solving a quantum many-body impurity problem at each step. This makes QE calculations orders of magnitude more demanding than standard DFT.

Machine learning (ML) has emerged as a promising route to eliminate this bottleneck. In the context of ML-assisted DMFT, significant progress has been made in learning the mapping involving the infinite-dimensional hybridization function and the dynamical Green's function [48–55] or the impurity self-energy [56,57]. These approaches address the difficult problem of learning a mapping between continuous functions of frequency, which formally constitute infinite-dimensional input and output spaces. Within the class of QE methods such as DMET, RISB, and the ghost-GA, in contrast, the physical system is mapped to an EH coupled to a *finite* number of bath sites, and the output consists of static single-particle expectation values, specifically the density matrix. This implies that the solution to the embedding problem is entirely encoded in the ground-state energy function of the EH, as the required physical outputs can be obtained directly as gradients of this energy. Leveraging this structure, previous works have attempted to learn the ground-state energy of the EH directly [58,59]. However, such direct learning approaches do not automatically enforce fundamental physical constraints, such as the requirement that density matrix eigenvalues lie between zero and one (Pauli principle). Consequently, achieving high accuracy and stability requires extensive training data and complex sampling strategies to ensure that the learned landscape remains physical.

In this work, we present a framework that overcomes these limitations by shifting the learning target from the EH energy function to the low-dimensional structure of its ground states. We hypothesize that the physically relevant ground states of the EH reside on a low-dimensional manifold embedded within the exponentially large Hilbert space, and we show that this manifold can be efficiently learned from data using the principal component analysis (PCA). This procedure enables us to determine a linear variational space, which constitutes a “foundation model” for QE frameworks, offering two decisive advantages. First, by projecting the EH onto a learned subspace, we compress the EH many-body problem, reducing it to the diagonalization of a small matrix and eliminating the computational bottleneck. Second, because the method retains the structure of a variational ansatz, physical constraints such as the Pauli principle are satisfied by construction.

We validate this approach within the ghost-GA framework, by benchmarking it against the three-orbital Hubbard model, demonstrating remarkable transferability where a solver trained on a Bethe lattice accurately solves square and cubic lattices without retraining. Furthermore, we apply the method to the complex phase diagram of plutonium (Pu), reproducing the energetics of all six crystalline phases using a single variational space. Finally, we propose a pathway for extending our data-driven framework to larger EHs, using ground-state data generated by solvers such as matrix product states (MPS) [60–65], neural quantum states (NQS), variational impurity solvers based on superpositions of Gaussian states [66,67], or quantum-assisted methods [68–71].

II. THE GHOST-GA ALGORITHMIC STRUCTURE

The ghost-GA theoretical framework is rooted in the variational principle and, similarly to DMFT, in the idea of the infinite-dimension limit [38]. Specifically, it is based on a variational wavefunction represented as $|\Psi_G\rangle = \hat{P}_G|\Psi_0\rangle$, where $\hat{P}_G$ is an operator acting on a reference single-particle wavefunction (Slater determinant) $|\Psi_0\rangle$. The wavefunction $|\Psi_0\rangle$ is defined in an auxiliary Hilbert space, where the number of fermionic modes is larger than that in the physical Hilbert space by a tunable factor B . This factor B governs the size of the variational space and the accuracy of the method. The operator $\hat{P}_G$ maps states from the auxiliary space into the physical space, modifying the weights of the local electronic configurations of all fragments. Both $|\Psi_0\rangle$ and $\hat{P}_G$ are optimized to minimize the variational energy, evaluated neglecting all contributions that vanish in the limit of infinite-dimensional systems (as in DMFT). Therefore, like DMFT, the ghost-GA can accurately capture dynamical correlations, but not spatial correlations.

In the context of real-material calculations, the union of DFT with a many-body method is required to capture strong electronic correlations. The algorithmic structure of DFT + ghost-GA, resulting from a convenient reformulation of the energy minimization problem for the multiconfigurational wavefunction $|\Psi_G\rangle$ [36,44,45], closely mirrors that of DFT + DMFT: It features a charge self-consistent iterative procedure nested within the solution of an EH—which serves as the impurity model. Physically, the EH describes how each atomic fragment interacts with its environment, ensuring that the feedback between local electronic correlations and the global electron density is consistently accounted for throughout the simulation.

It is important to note that this algorithmic structure applies to both real materials (in DFT + ghost-GA) and theoretical models (such as the Hubbard model). In fact, like DFT + DMFT, the DFT + ghost-GA theory is ultimately encoded in a modified DFT functional, transforming the standard one-body KS reference system into a multiorbital Hubbard model with local interactions, termed the “KS-Hubbard Hamiltonian” (KSH). Similar to conventional DFT, the core of DFT + QE is a charge self-consistency loop, which continuously updates the KSH system. The KSH model is then solved with the QE method of choice. Thus, real-material calculations require a recursive QE solution of the EH, with an additional charge self-consistency loop nested within the process.

The EH of each fragment i is defined in second quantization as follows (omitting the label i for simplicity):

$$\begin{aligned} \hat{H}^{\text{EH}} = & \hat{H}^{\text{int}} + \sum_{\alpha,\beta=1}^v E_{\alpha\beta} c_{\alpha}^{\dagger} c_{\beta} \\ & + \sum_{a=1}^{Bv} \sum_{\alpha=1}^v (D_{a\alpha} c_{\alpha}^{\dagger} b_a + \text{H.c.}) + \sum_{a,b=1}^{Bv} \Lambda_{ab} b_a^{\dagger} b_b, \end{aligned} \quad (1)$$

where $c_{\alpha}^{\dagger}$ (c_{α}) is the creation (annihilation) operator for an embedded electron with spin-orbit index α , $b_a^{\dagger}$ (b_a) is the creation (annihilation) operator for an electron with index a in the bath, and v denotes the number of spin orbitals in the fragment (2 for s , 6 for p , 10 for d , and 14 for f shells).

The first term, $\hat{H}^{\text{int}}$, captures the local electronic interactions within the fragment:

$$\hat{H}^{\text{int}} = \frac{1}{2} \sum_{p,q,r,s=1}^{2l+1} \sum_{\sigma,\sigma' \in \uparrow, \downarrow} \mathcal{V}_{pqrs} c_{p\sigma}^\dagger c_{r\sigma'}^\dagger c_{s\sigma'} c_{q\sigma}, \quad (2)$$

where $l = 0, 1, 2$, and 3 correspond, respectively, to s , p , d , and f electronic shells, and $\mathcal{V}_{pqrs}$ are the interaction coefficients. The second term corresponds to the one-body part of the fragment (impurity), with E specifying its one-body energies. The subsequent terms in the Hamiltonian describe the coupling of the fragment to an auxiliary “bath” of $B\nu$ orbitals, governed by hybridization parameters D and bath energies Λ . The parameter B controls the balance between computational complexity and accuracy: $B = 1$ recovers the standard multi-orbital GA [35–43], while $B = 3$ has been systematically shown to achieve DMFT-level accuracy for ground-state properties [44,45,72–74], at a substantially lower computational cost.

From now on, we assume that the Coulomb integrals $\mathcal{V}_{pqrs}$ are fully encoded in the Hubbard-repulsion parameters U and the Hund’s coupling constant J [75], setting the Hubbard U as energy unit. For later convenience, we introduce the following compact notation:

$$\hat{H}^{\text{EH}}(\mathbf{X}) = \hat{H}_0 + \sum_{r=1}^{\dim(\mathbf{X})} X_r \hat{H}_r, \quad (3)$$

where $\hat{H}_0$ is the term corresponding to $U = 1$, while J/U and the entries of E/U , D/U , and Λ/U are stored in a real vector $\mathbf{X}$, which is updated self-consistently throughout the ghost-GA loop. We denote by $|\Phi(\mathbf{X})\rangle$ the ground state of $\hat{H}^{\text{EH}}(\mathbf{X})$.

The output parameters to be computed at each iteration of the ghost-GA algorithmic structure are the following ground-state single-particle density matrix elements:

$$h_{aa}(\mathbf{X}) = \langle \Phi(\mathbf{X}) | c_a^\dagger c_a | \Phi(\mathbf{X}) \rangle, \quad (4)$$

$$d_{mn}(\mathbf{X}) = \langle \Phi(\mathbf{X}) | b_m^\dagger b_n | \Phi(\mathbf{X}) \rangle. \quad (5)$$

A. Computational complexity of ghost-GA

The ghost-GA algorithmic structure outlined above has two key computational advantages:

(1) In ghost-GA, it is sufficient to solve an EH with a relatively small bath ($B = 3$) to obtain the total energy of the lattice model with an accuracy comparable with DMFT, which in principle requires a bath with infinite degrees of freedom.

(2) While in DMFT the output $\mathbf{Y}$ is the local Green’s function of the fragment i , the ghost-GA requires only the iterative calculation of the entries of the ground-state single-particle density matrix.

Because of these reasons, each ghost-GA iteration can be tackled with a wealth of efficient state-of-the-art methods that do not suffer from the so-called continuous-time quantum Monte Carlo fermionic “sign problem” [76,77]. Typical choices are the following:

(1) For s and p orbitals, the $B = 3$ EH can be solved using exact diagonalization (ED) methods (Arnoldi iterations).

(2) For d -orbital systems, the EH can be solved employing MPS techniques [60–65].

(3) Recently, quantum computing methods have also shown promise [68–71].

The accuracy and algorithmic structure of the ghost-GA framework have proven useful within a broad range of challenging strongly correlated problems [78–83] and provide us with the opportunity to tackle simulations otherwise impossible. This notably includes simulations of low-symmetry systems where the self-energy has a nondiagonal structure, such as multi-orbital systems with disorder or coexistence of crystal/ligand-field splittings and spin-orbit coupling (SOC). Another exciting example is computational catalysis, where strongly correlated materials have been shown to induce new mechanisms, potentially paving the way for more efficient ammonia production under mild conditions [84,85]. Meanwhile, even the ghost-GA would be too computationally expensive for extensive high-throughput calculations.

B. Challenge of designing an EH solver taking full advantage of the ghost-GA algorithmic structure

Despite the favorable algorithmic advantages discussed above, the computational cost of solving the EH remains the primary bottleneck within the self-consistency ghost-GA cycle, limiting the scale and throughput of simulations with respect to mean-field approaches such as Kohn-Sham DFT or DFT + U and, therefore, their applicability for large-scale materials discovery. This motivates us to address the following key question: Can this computational bottleneck be eliminated by developing a more efficient solver directly leveraging the ghost-GA algorithmic structure, thereby achieving DMFT-like accuracy at a near-DFT cost?

In this work, we introduce a pathway to achieve this goal by leveraging on the dimensionality reduction for learning an optimized variational space from data, systematically reducing the computational overhead of EH computations by orders of magnitude.

III. OVERVIEW OF PREVIOUS APPLICATIONS OF ML IN QE METHODS

The foundation of this work lies in a key observation made in Refs. [48–51,58,59]: within all QE methods, the form of the EH is universal. That is, its mathematical structure depends solely on the impurity’s structure (typically representing d -electron or f -electron shells) and remains independent of the specific material or molecule containing it. Based on this observation, the goal is to train a machine to learn once and for all the universal map between the embedding parameters $\mathbf{X}$ and the corresponding outputs $\mathbf{Y}$, for each impurity type. Note that, for a given bath size B , this requires to train only *one* machine for each type of impurity (e.g., one for d electrons and one for f electrons). This approach aims to bypass the computationally expensive quantum-mechanical solution of the EH, allowing DFT + QE simulations of any solid and molecule at a DFT-like cost.

The perspective of bypassing the computational bottleneck of QE simulations of strongly correlated methods motivated extensive work along the lines depicted above, particularly focused on DMFT. In that case, $\mathbf{X}$ includes the hybridization function of a fragment, its on-site energies, the Hubbard

parameter U , and Hund’s constant J , while $\mathbf{Y}$ represents the corresponding fragment’s Green’s function. This input-output structure demands learning from a function of infinite variables (the hybridization function), which involves aspects such as bath parametrization and fitting. Similarly, it involves the problem of learning a continuous output (the local Green’s function of the impurity). So far, research in this field has focused on identifying optimal parametrizations for these infinite-dimensional inputs and outputs. However, to date, benchmarks have focused only on the single-band impurity models with a *fixed* mathematical form, such as a featureless (flat) hybridization function. The scalability of ML-based impurity solvers to full DMFT self-consistent calculations—where the hybridization function is iteratively updated and does not retain any predetermined mathematical form—is yet to be proven, even for single-orbital lattice models.

As explained above in Sec. II, the ghost-GA introduces a notably simpler input-output structure with respect to DMFT. Here, the EH has a finite bath, defining the parameters with a finite set, and the output parameters are encoded in the gradient of the EH ground-state energy function:

$$\mathcal{E}(\mathbf{X}) = \langle \Phi(\mathbf{X}) | \hat{H}^{\text{EH}}(\mathbf{X}) | \Phi(\mathbf{X}) \rangle, \quad (6)$$

$$\mathbf{Y}(\mathbf{X}) = \nabla \mathcal{E}(\mathbf{X}), \quad (7)$$

where $\Phi(\mathbf{X})$ is the ground state of $\hat{H}^{\text{EH}}(\mathbf{X})$.

For example, the partial derivative of the EH ground-state energy with respect to $[\Lambda]_{aa}$ is $\langle \Phi(\mathbf{X}) | b_a^\dagger b_a | \Phi(\mathbf{X}) \rangle$. Therefore, if we could learn once and for all the universal energy function $\mathcal{E}(\mathbf{X})$, along with its gradient for all types of fragments i [particularly s , p , d , or f electronic shells, corresponding to $v_i = 2, 6, 10$, and 14 of Eq. (2), respectively], we would be able to perform ghost-GA calculations of any material and molecule, bypassing altogether the computational bottleneck of calculating the ground state of the EH. This mathematical structure motivated previous attempts to directly learn the function $\mathcal{E}(\mathbf{X})$ using ML frameworks [59]. However, these approaches do not automatically enforce all qualitative physical constraints that the energy landscape must satisfy. One such example is that the density matrix elements in Eqs. (4) and (5), encoded in the gradient of the energy, must satisfy N -representability constraints (e.g., the Pauli principle) [86]. The absence of this physical information in the model required generating extensive training data and complex sampling strategies to achieve sufficient accuracy [59], limiting scalability to complex materials.

Here, we propose a different approach that resolves these problems. Instead of directly learning the energy landscape, we introduce a method to construct a linear variational space for the EH ground state directly from data. This strategy effectively compresses the relevant Hilbert space, similarly to how image compression identifies and retains only the essential features of a picture while discarding redundant information. By solving the problem within this optimized subspace, the framework enforces N -representability constraints by construction, yielding a drastically more compact, accurate, and scalable approach.

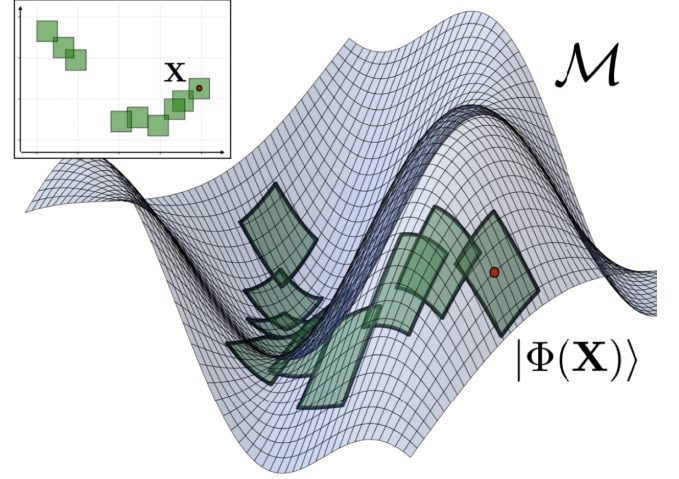


FIG. 1. Geometric representation of the ground-state manifold $\mathcal{M}$. The curved surface represents the manifold of ground states $|\Phi(\mathbf{X})\rangle$. The manifold is partitioned into local regions (green rectangles), each corresponding to a domain in the Hamiltonian parameter space $\mathbf{X}$, which is schematically illustrated in the inset.

IV. LEARNING AN EFFICIENT VARIATIONAL SPACE FOR THE EH

The set of all relevant ground states $|\Phi(\mathbf{X})\rangle$ forms a $\dim(\mathbf{X})$ -dimensional manifold $\mathcal{M}$ within the full Hilbert space $\mathcal{H}$, as represented in Fig. 1. Note that this is a direct consequence of the ghost-GA’s finite-bath structure. The fact that $\mathcal{M}$ lives in such a low-dimensional corner of the Hilbert space presents a unique opportunity: if we could learn an efficient variational space capturing the states in $\mathcal{M}$, we could construct a new type of EH solver.

The key difficulty to address is how to explicitly construct this latent space, given that $\mathcal{M}$ is not known explicitly. This makes our goal distinct from general-purpose variational frameworks where the structure of the ansatz is fixed *a priori* by physical considerations (e.g., MPS).

Our solution is based on the idea of learning the variational space directly from data, a concept also explored in other quantum many-body physics methods aimed at dimension reduction, such as eigenvector continuation [87–92] and foundation models based on neural networks [93–95]. Specifically, we realize a data-driven program constructing a piecewise-linear variational space, as follows:

- (1) We partition the space of $\mathbf{X}$ parameters into local domains, each representing a chart of the ground-state manifold $\mathcal{M}$ (see Fig. 1).
- (2) For each local domain, we construct from data a linear variational space with tunable dimension K (controlling the resulting accuracy).

We structure our explanation as follows: In Sec. IV A, we show that, once a suitable low-dimensional linear subspace has been specified within a given region of $\mathcal{M}$, the EH problem can be reformulated and solved within that reduced space. In Sec. IV B, we then show how this variational subspace can be constructed from data by PCA compression of EH ground states, and we place this construction on a principled footing

by interpreting it as a linear foundation model for the family of EHs.

A. Variational-space construction of the EH

Let $\mathcal{V}_K$ be a K -dimensional linear variational space for a given region of the EH ground-state manifold $\mathcal{M}$, and let $\{u_1, \dots, u_K\}$ be an orthonormal basis of $\mathcal{V}_K$. We define Π_K as the orthogonal projector onto $\mathcal{V}_K$. The problem of extremizing the corresponding variational energy is equivalent to calculating the ground state of the following $K \times K$ matrix:

$$[\tilde{H}^K(\mathbf{X})]_{nm} = [\tilde{H}_0^K]_{nm} + \sum_{r=1}^{\dim(\mathbf{X})} X_r [\tilde{H}_r^K]_{nm}, \quad (8)$$

where the matrix elements

$$[\tilde{H}_r^K]_{nm} = \langle u_n | \hat{H}_r | u_m \rangle \quad (9)$$

can be conveniently precalculated and stored for all $r = 0, \dots, \dim(\mathbf{X})$. In particular, the resulting variational approximation to the universal energy function previously introduced in Eq. (6) is the following:

$$\begin{aligned} \mathcal{E}(\mathbf{X}) &\approx \mathcal{E}_K(\mathbf{X}) = \sum_{n,m=1}^K c_n^*(\mathbf{X}) \langle u_n | \hat{H}^{\text{EH}}(\mathbf{X}) | u_m \rangle c_m(\mathbf{X}) \\ &= \lambda_{\min} \left(\tilde{H}_0^K + \sum_{r=1}^{\dim(\mathbf{X})} X_r \tilde{H}_r^K \right), \end{aligned} \quad (10)$$

where $\lambda_{\min}$ denotes the lowest eigenvalue of $\tilde{H}^K(\mathbf{X})$ and $c_m(\mathbf{X})$ are the components of the corresponding eigenvector:

$$|\Phi_K(\mathbf{X})\rangle = \sum_{k=1}^K c_k(\mathbf{X}) |u_k\rangle. \quad (11)$$

The gradient of $\mathcal{E}_K(\mathbf{X})$ is available through the Hellmann-Feynman theorem. For example,

$$\frac{\partial \mathcal{E}_K(\mathbf{X})}{\partial \Lambda_{aa}} = \sum_{n,m=1}^K c_n^*(\mathbf{X}) \langle u_n | b_a^\dagger b_a | u_m \rangle c_m(\mathbf{X}). \quad (12)$$

This framework has three fundamental properties:

(1) Since this representation is based on the variational principle, the resulting density matrices satisfy all N -representability conditions by construction [86]. For example, the occupation numbers are guaranteed to lie within physical bounds, $0 \leq \langle \Phi_K(\mathbf{X}) | b_a^\dagger b_a | \Phi_K(\mathbf{X}) \rangle \leq 1$, strictly enforcing the Pauli principle.

(2) Partitioning the parameter space in multiple charts reduces the complexity of the learning problem for each individual chart, and the cost of evaluating observables decreases as a function of the variational-space truncation K .

(3) The partitioning strategy also provides us with the option of focusing only on training our model within physically realizable regions of the parameters space, which, as we are going to show, is considerably more efficient and scalable.

It is important to point out that the possibility of calculating an efficient linear variational approximation to the EH problem is theoretically possible only thanks to the finite size of

the EH bath, as for an infinite bath Anderson's orthogonality catastrophe [96] would render ground states for different parameters mutually orthogonal, precluding the possibility of approximating them within any finite K .

B. Construction of variational space for each chart of $\mathcal{M}$

Let us consider any given chart of $\mathcal{M}$, represented in Fig. 1. We aim to identify the basis vectors of $\{u_1, \dots, u_K\}$ spanning the corresponding "optimal" variational space $\mathcal{V}_K$, of any dimension K smaller than the Hilbert space dimension M_B .

This optimization problem can be formalized mathematically as follows. We denote Π_K as the orthogonal projector over $\mathcal{V}_K$. Consider a probability distribution $p(\mathbf{X})$ over the parameter space for the Hamiltonians $\hat{H}(\mathbf{X})$. This induces a corresponding distribution $P(\Phi)$ over their respective ground states $\Phi(\mathbf{X})$. We aim to calculate, for any given integer K , the orthogonal projector Π_K that minimizes the following quantity:

$$C(\Pi_K) = E_P[\|(\Pi_K - I)\Phi\|^2], \quad (13)$$

where I is the identity and E_P denotes the expectation value with respect to $P(\Phi)$. In practice, we approximate $C(\Pi_K)$ using a finite number N of samples $\{\mathbf{X}^1, \dots, \mathbf{X}^N\}$, assuming a uniform (constant) probability distribution $p(\mathbf{X})$:

$$C(\Pi_K) \simeq \frac{1}{N} \sum_{s=1}^N \|(\Pi_K - I)\Phi(\mathbf{X}^s)\|^2, \quad (14)$$

where we are introducing a superscript s to indicate the s th sampled vectors $\mathbf{X}^s$, with components $[\mathbf{X}^s]_r$.

To solve this optimization problem, we first construct the data matrix containing the sampled ground states as columns: $\mathbb{M} = [\Phi(\mathbf{X}^1) | \Phi(\mathbf{X}^2) | \dots | \Phi(\mathbf{X}^N)]$. We then perform a singular-value decomposition (SVD) on this matrix:

$$\mathbb{M} = U \Sigma V^\dagger = \sum_{a=1}^N \sigma_a |u_a\rangle \langle v_a|, \quad (15)$$

where U and V are unitary matrices containing the left and right singular vectors, respectively, and Σ is the diagonal matrix of singular values σ_a , sorted in descending order. A standard result from PCA states that the leading singular vectors provide the optimal low-rank approximation to the data. Applying this to our case, we obtain that the optimization problem of minimizing $C(\Pi_K)$ is solved by the orthogonal projector:

$$\Pi_K = \sum_{a=1}^K |u_a\rangle \langle u_a|, \quad (16)$$

which minimizes the projection error in approximating the sampled ground states. In conclusion, our variational space is $\mathcal{V}_K = \text{Span}(\{u_1, \dots, u_K\})$ and the resulting impurity solver is constructed using it, as detailed in Sec. IV A.

PCA compression as a foundation model

The construction above shows that applying PCA compression to a dataset consisting of ground states $\{\Phi(\mathbf{X}^s)\}$, sampled according to any probability distribution $p(\mathbf{X})$ over

Hamiltonian parameters, is mathematically equivalent to determining an optimal variational subspace. Specifically, for a given $p(\mathbf{X})$, the resulting principal components span a shared variational subspace $\mathcal{V}_K$ that is optimal, in the sense of average fidelity, for all ground states generated by the family of embedding Hamiltonians $\hat{H}(\mathbf{X})$. In this sense, our PCA procedure provides a *linear foundation model* for this family of Hamiltonians, in direct connection with variational foundation models based on neural-network quantum states that are trained to approximate ground states across continuous families of Hamiltonians [93–95]. The main difference is that here the cost functional is fidelity based rather than energy based.

Our strategy described in Sec. IV A, which consists of projecting the entire EH landscape $\{\hat{H}(\mathbf{X})\}$ into the learned variational space $\mathcal{V}_K$, is directly enabled by this conceptual connection between compression and foundation-model perspectives. The resulting improvements in the efficiency of the overall QE procedure are discussed in Secs. V and VI.

V. BENCHMARK CALCULATIONS

In this section, we benchmark our PCA-based framework as follows. First, in Sec. V A, we benchmark it on the three-orbital Hubbard-Kanamori model using a pretrained library of local variational spaces described in Sec. V B, quantifying its accuracy and the resulting computational speedup. Second, in Sec. V C, we introduce an active-learning scheme that builds this library only on demand along the ghost-GA self-consistency procedure, showing that this substantially simplifies the learning process. Finally, in Sec. V D, we apply the same strategy to real materials, focusing on all allotropes of Pu under pressure, and show that the learned variational space captures the local $5f$ electronic environment and can be reused across all crystallographic phases.

A. Degenerate three-band Hubbard-Kanamori models

1. Lattice models

To illustrate our approach, we first focus on solving model systems with three degenerate orbitals and with a local electron-electron Kanamori interaction [97,98] and only nearest-neighbor hoppings. The Hamiltonian will generically read

$$\hat{H} = \hat{H}^0 + \sum_{\mathbf{R}} \hat{H}_{\mathbf{R}}^{\text{int}}(U, J), \quad (17)$$

$$\hat{H}^0 = -t \sum_{\langle \mathbf{R}\mathbf{R}' \rangle} \sum_{\alpha\sigma} c_{\mathbf{R}\alpha\sigma}^\dagger c_{\mathbf{R}'\alpha\sigma} - \mu \sum_{\mathbf{R}\alpha\sigma} c_{\mathbf{R}\alpha\sigma}^\dagger c_{\mathbf{R}\alpha\sigma}, \quad (18)$$

where $c_{\mathbf{R}\alpha\sigma}^\dagger$ ($c_{\mathbf{R}\alpha\sigma}$) is the electronic creation (annihilation) operator at site $\mathbf{R}$ for orbital α and spin σ , $\hat{H}^0$ is the non-interacting part of the Hamiltonian, and $\hat{H}_{\mathbf{R}}^{\text{int}}$ is the local Hubbard-Kanamori interaction. We consider $U' = U - 2J$, for which the interaction becomes rotationally invariant and will be parametrized by a Hubbard interaction U and a Hund's coupling J . For each site $\mathbf{R}$, the interaction is defined as follows:

$$\hat{H}_{\mathbf{R}}^{\text{int}}(U, J) = \frac{U - 3J}{2} (\hat{N}_{\mathbf{R}})^2 - 2J (\vec{S}_{\mathbf{R}})^2 - \frac{J}{2} (\vec{L}_{\mathbf{R}})^2, \quad (19)$$

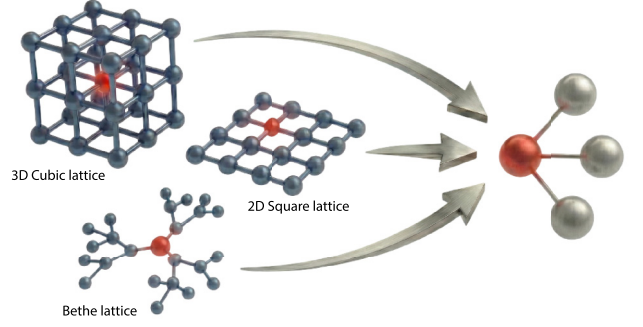


FIG. 2. Pictorial representation of how different lattice models (cubic, square, and Bethe) map onto an embedding Hamiltonian with the same structure.

where

$$\hat{N}_{\mathbf{R}} = \sum_{\alpha\sigma} c_{\mathbf{R}\alpha\sigma}^\dagger c_{\mathbf{R}\alpha\sigma}, \quad (20)$$

$$\vec{S}_{\mathbf{R}} = \frac{1}{2} \sum_{\alpha} \sum_{\sigma\sigma'} c_{\mathbf{R}\alpha\sigma}^\dagger \vec{\tau}_{\sigma\sigma'} c_{\mathbf{R}\alpha\sigma}, \quad (21)$$

$$\vec{L}_{\mathbf{R},\gamma} = i \sum_{\alpha\beta\sigma} \epsilon_{\gamma\alpha\beta} c_{\mathbf{R}\alpha\sigma}^\dagger c_{\mathbf{R}\beta\sigma}, \quad (22)$$

with $\vec{\tau} = (\tau_x, \tau_y, \tau_z)$ the vector of Pauli matrices and ϵ the Levi-Civita symbol. The noninteracting part of the Hamiltonian in the basis for which the hopping term is diagonal is

$$\hat{H}^0 = \sum_{\mathbf{k},\alpha\sigma} (\epsilon_{\mathbf{k}} - \mu) c_{\mathbf{k}\alpha\sigma}^\dagger c_{\mathbf{k}\alpha\sigma}, \quad (23)$$

where $\epsilon_{\mathbf{k}}$ is the Fourier transform of the hopping Hamiltonian and μ is the chemical potential. We will consider the Bethe lattice, featuring a semicircular density of states $\rho(\epsilon) = (2/\pi W^2) \sqrt{W^2 - \epsilon^2}$, the square lattice, for which $\epsilon_{\mathbf{k}} = -(W/2)(\cos k_x + \cos k_y)$, and the cubic lattice, for which $\epsilon_{\mathbf{k}} = -(W/3)(\cos k_x + \cos k_y + \cos k_z)$, where W is the half-bandwidth, and we will express most of our results in units of W . Once interactions are taken into account, the lattice model retains a charge- $U(1)$, spin- $SU(2)$, and orbital- $SO(3)$ symmetry.

2. Structure of the EH in the case of three degenerate orbitals

The global symmetries of the lattice Hamiltonian will be reflected as symmetries of the EH [34,44]. Different lattices with the same point-group symmetry will share the structure of the embedding Hamiltonian, as shown in Fig. 2, and imposing those symmetries will drastically reduce the number of free parameters in the Hamiltonian. In particular, we impose that the ghost-GA variational wavefunction $|\Psi_G\rangle$ respects charge $U(1)$, spin $SU(2)$, and orbital $SO(3)$ symmetries. These constraints translate into corresponding restrictions on the EH [34,36]. The $U(1)$ constraint implies that the EH is at half filling. The $SU(2)$ constraint implies that the EH can be chosen to be spin-rotation invariant, with a spin-singlet ground state and a one-body part that splits into two equivalent spin-up and spin-down terms. Finally, the $SO(3)$ constraint implies that the one-body part of the EH contains no terms

coupling different orbitals, and that all one-body parameters are identical for the three orbitals. The Hamiltonian then reads

$$\begin{aligned} \hat{H} = & \hat{H}^{\text{int}}(U, J) + E \sum_{\alpha=1}^3 \sum_{\sigma \in \uparrow, \downarrow} c_{\alpha\sigma}^\dagger c_{\alpha\sigma} \\ & + \sum_{a=1}^3 D_a \sum_{\alpha=1}^3 \sum_{\sigma \in \uparrow, \downarrow} (c_{\alpha\sigma}^\dagger b_{a\alpha\sigma} + \text{H.c.}) \\ & + \sum_{a,a'=1}^3 \Lambda_{aa'} \sum_{\alpha=1}^3 \sum_{\sigma \in \uparrow, \downarrow} b_{a\alpha\sigma}^\dagger b_{a'\alpha'\sigma}, \end{aligned} \quad (24)$$

where $c_{\alpha\sigma}^\dagger$ ($c_{\alpha\sigma}$) creates (annihilates) an electron in the impurity p orbital $\alpha \in \{1, 2, 3\}$ with spin σ , and $b_{a\alpha\sigma}^\dagger$ ($b_{a\alpha\sigma}$) creates (annihilates) an electron on the corresponding bath site in the bath channel $a \in \{1, 2, 3\}$. The term $\hat{H}^{\text{int}}(U, J)$ is the standard two-body Kanamori interaction, whose matrix elements are determined by the parameters U and J , as defined in Eq. (19). The parameters D_a control the hybridization strength between the impurity and bath channel a , while $\Lambda_{aa'}$ sets the on-site energies and hybridizations for all bath sites within that physical orbital channel.

To reduce the parameter space that must be covered by the ML model, we proceed as follows:

(1) Unitary canonical transformations acting among the bath orbitals do not affect the hybridization function and, in turn, the physical interaction between the impurity and the bath. Within the ghost-GA formalism, such transformations are referred to as gauge transformations. Following Ref. [59], we use gauge transformations to restrict our ML task to a bath representation with Λ diagonal. Furthermore, we sort the eigenvalues of Λ in ascending order ($\Lambda_a < \Lambda_{a+1}$) and choose the hybridization parameters D_a to be real and negative.

(2) Multiplying the EH by an overall positive factor does not change its ground-state eigenvector. Therefore, without loss of generality we fix the energy unit by setting $U = 1$ in the EH parametrization used for the ML task.

(3) As explained above, the charge $U(1)$ symmetry implies that the EH ground state has to be calculated within the half-filled sector. Since adding to the EH a chemical-potential term proportional to the total number of fermions (including bath and impurity) only shifts all eigenvalues by a constant, but does not change eigenvectors, we can set $E = 0$ with no loss of generality.

With these simplifications, the EH reads

$$\begin{aligned} \hat{H} = & \hat{H}^{\text{int}}(U, J) + \sum_{a=1}^3 \Lambda_a \sum_{\alpha=1}^3 \sum_{\sigma \in \uparrow, \downarrow} b_{a\alpha\sigma}^\dagger b_{a\alpha\sigma} \\ & + \sum_{a=1}^3 D_a \sum_{\alpha=1}^3 \sum_{\sigma \in \uparrow, \downarrow} (c_{\alpha\sigma}^\dagger b_{a\alpha\sigma} + \text{H.c.}), \end{aligned} \quad (25)$$

which can be rewritten as follows:

$$\hat{H}(\mathbf{X}) = \hat{H}^{\text{int}}(U, J) + \sum_{r=1}^6 X_r \hat{H}_r, \quad (26)$$

where $\hat{H}_r$ denote the hybridization and bath-energy terms in Eq. (25), and the parameter vector is

$$\mathbf{X} = (D_1, D_2, D_3, \Lambda_1, \Lambda_2, \Lambda_3). \quad (27)$$

As discussed above, we set $U = 1$.

Since J is not updated throughout the ghost-GA self-consistency procedure, from now on we set $J/U = 0.1$ for simplicity.

B. Pretraining approach

In this section, we construct an atlas of local charts in the parameter space defined in Eq. (27). The atlas is built over a prescribed parameter domain of embedding-Hamiltonian parameters $\mathbf{X}$. Each chart is assumed to be hyperrectangular, characterized by a center $\mathbf{X}_j$ and by fixed linear sizes L_D and L_Λ along the D and Λ directions, respectively.

To build the atlas, we repeatedly draw random samples $\mathbf{X}_j$ from a uniform distribution over the target domain. If a sampled point $\mathbf{X}_j$ lies inside at least one chart that has already been learned, we discard the sample. Otherwise, we define a new chart centered at $\mathbf{X}_j$ (with the sizes specified above) and associate with it a variational solver. We iterate this procedure until additional random samples no longer generate new charts, at which point the atlas provides a covering of the chosen domain.

1. Training procedure for a single chart

Throughout this section, we fix $L_D = 0.4$ and $L_\Lambda = 4.0$. For each chart, we follow the PCA construction described in Sec. IV using a finite local training set. Specifically, we draw N_{budget} parameter vectors $\{\mathbf{X}^s\}$ uniformly within the chart, compute the corresponding EH ground states, and perform the SVD of the resulting data matrix. In our calculations, we set $N_{\text{budget}} = 600$.

We then determine dynamically the dimension K of the local variational space from a prescribed accuracy threshold, which we set to $\varepsilon_{\text{tol}} = 10^{-6}$. Concretely, we set

$$K = \min\{k \mid \sigma_k/\sigma_1 < \varepsilon_{\text{tol}}\}, \quad (28)$$

where $\{\sigma_k\}$ are the singular values sorted in descending order. This criterion is motivated by the fact that the singular values quantify the weight of each principal direction in representing the training set: if σ_k decays rapidly with k , the sampled ground states lie close to a low-dimensional linear subspace. Requiring $\sigma_K/\sigma_1 < \varepsilon_{\text{tol}}$ therefore ensures that principal components beyond K carry a negligible weight compared to the leading one, so that the span of the first K components provides an accurate compressed representation of the training data within the chart.

Since the data matrix $\mathbb{M}$ has at most rank N_{budget} , the SVD spectrum is resolved only up to $k = N_{\text{budget}}$. For this reason, if the value of K determined by Eq. (28) approaches N_{budget} , it becomes sensitive to the finite sampling budget and cannot be interpreted as a robust estimate of the intrinsic truncation dimension. To guard against this undersampling effect, we impose the additional buffer condition $K/N_{\text{budget}} < 0.8$. If this condition is not satisfied, we keep N_{budget} fixed and replace the chart by a smaller one with the same center $\mathbf{X}_j$ but with L_D and L_Λ reduced by a factor of 2, and repeat the sampling and PCA procedure. This process is iterated until the singular-value decay is sufficiently well resolved at the target tolerance.

Once the condition $K/N_{\text{budget}} < 0.8$ is met, we retain the first K principal components $\{|u_n\rangle\}_{n=1}^K$, which form an orthonormal basis of the compressed variational space used to

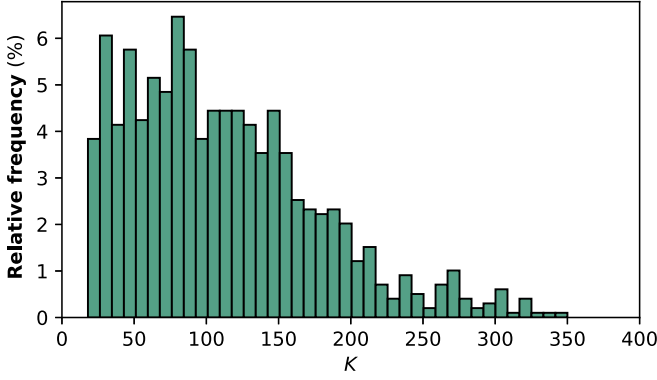


FIG. 3. Distribution of the number of principal components retained to define the local variational space across the 990 charts of the pretrained atlas described in Sec. VB 2.

represent EH ground states within the chart. We then compute and store the corresponding matrix elements in this basis for the operators needed within the ghost-GA loop. In the present symmetry-reduced setting, these can be taken as

$$[\tilde{O}_{\alpha\alpha}^h]_{nm} = \sum_{\sigma} \langle u_n | c_{\alpha\sigma}^{\dagger} b_{\alpha\sigma} | u_m \rangle, \quad (29)$$

$$[\tilde{O}_{\alpha\alpha}^d]_{nm} = \sum_{\sigma} \langle u_n | b_{\alpha\sigma}^{\dagger} b_{\alpha\sigma} | u_m \rangle, \quad (30)$$

$$[\tilde{H}^{\text{int}}]_{nm} = \langle u_n | \hat{H}^{\text{int}} | u_m \rangle. \quad (31)$$

This precomputation allows subsequent EH solutions with $\mathbf{X}$ lying within the trained chart to be carried out as a deterministic $K \times K$ eigenvalue problem within the corresponding K -dimensional variational space.

2. Pretraining construction and atlas statistics

For the degenerate three-orbital problem introduced in Sec. VA, we take the pretraining domain to be $D_i \in [-1.0, 0.0]$ and $\Lambda_i \in [-10.0, 10.0]$ for $i = 1, 2, 3$.

Applying the procedure above across this pretraining domain yields an atlas consisting of 990 charts. In practice, we find that the atlas construction saturates after $\sim 10^4$ random trial centers, in the sense that additional samples no longer generate new charts.

Across all charts, the resulting distribution of variational dimensions K is shown in Fig. 3. The largest value selected by the criteria above is $\max(K) = 350$, compared to a Fock-space dimension of 853 776 for the full EH. The resulting database of projected operators for all charts has size 1.75 GB.

3. Ghost-GA results with the pretrained solver

Having constructed the atlas and the corresponding database of projected operators, we next use the resulting compressed solver within full ghost-GA calculations for the three-orbital Hubbard-Kanamori model introduced in Sec. VA. We benchmark ghost-GA calculations in which the EH is solved using the pretrained PCA solver against reference calculations in which the EH is solved by exact diagonalization.

During the ghost-GA self-consistency cycle, the EH parameters $\mathbf{X}$ at a given iteration may fall within multiple charts

of the pretrained atlas. In such cases, we exploit the variational character of the projected solver: we evaluate the EH ground-state energy in each admissible chart and select the solution yielding the lowest energy.

Figure 4 compares ghost-GA results obtained with the ED solver (squares) and with the pretrained PCA solver (circles) for the degenerate three-orbital model as a function of U/W on the Bethe, square, and cubic lattices (left, center, and right columns, respectively), and for the fillings indicated in the figure. The quasiparticle weight Z [Figs. 4(a)–4(c)] and the total energy per site [Figs. 4(g)–4(i)] are reproduced with high accuracy by the PCA solver. The insets of Figs. 4(g)–4(i) report the corresponding energy differences $E_{\text{ED}} - E_{\text{PCA}}$, which remain small across the explored interaction range. At the same time, the total wall time of the ghost-GA cycle is reduced substantially when using the PCA solver, as shown in Figs. 4(d)–4(f), with speedups ranging from one order of magnitude to nearly two orders of magnitude in the vicinity of the Mott transition.

These benchmarks demonstrate that a solver pretrained once for a given EH structure can be reused across different lattice realizations of the same local problem, delivering both robust accuracy and significant computational acceleration. This reduction of the EH cost directly enhances the feasibility of systematic scans and high-throughput studies within ghost-GA.

C. Active-learning approach

In the previous subsection, we showed that constructing an atlas of local variational solvers can substantially accelerate quantum-embedding calculations by reducing the cost of the EH step. However, as shown in Ref. [59], the physically relevant EHs populate a sparse, low-dimensional manifold, rather than being spread throughout the full parameter domain. Consequently, globally pretraining a solver over a large parameter region can waste significant computational effort on regions that are never visited in QE calculations.

Motivated by this observation, we adopt an active-learning (on-demand) strategy in which the database is initialized as empty and updated only when needed. During the self-consistency cycle, if the current EH parameter $\mathbf{X}$ is contained in at least one previously learned chart, we solve the EH using the corresponding PCA-projected solver; otherwise, we solve the EH by ED and record the uncovered $\mathbf{X}$ values. After convergence at a given set of physical parameters, any newly recorded $\mathbf{X}$ values are used as training centers to construct additional charts following Sec. VB 1, and the resulting projected operators are appended to the database before proceeding.

We benchmark this approach by running ghost-GA calculations over the combined set of choices of lattice realization, electron filling n , and interaction strength U . We first perform the calculations on the Bethe lattice (left column of Fig. 5), starting from an empty database, and then apply the database accumulated on the Bethe lattice to the square and cubic lattices (center and right columns). In all cases, we traverse the (n, U) grid as follows: We loop over $n = 3.0, 2.9, \dots, 0.5$. For each value of n , we then sweep $U/W = 0.4, 0.5, \dots, 2.5$.

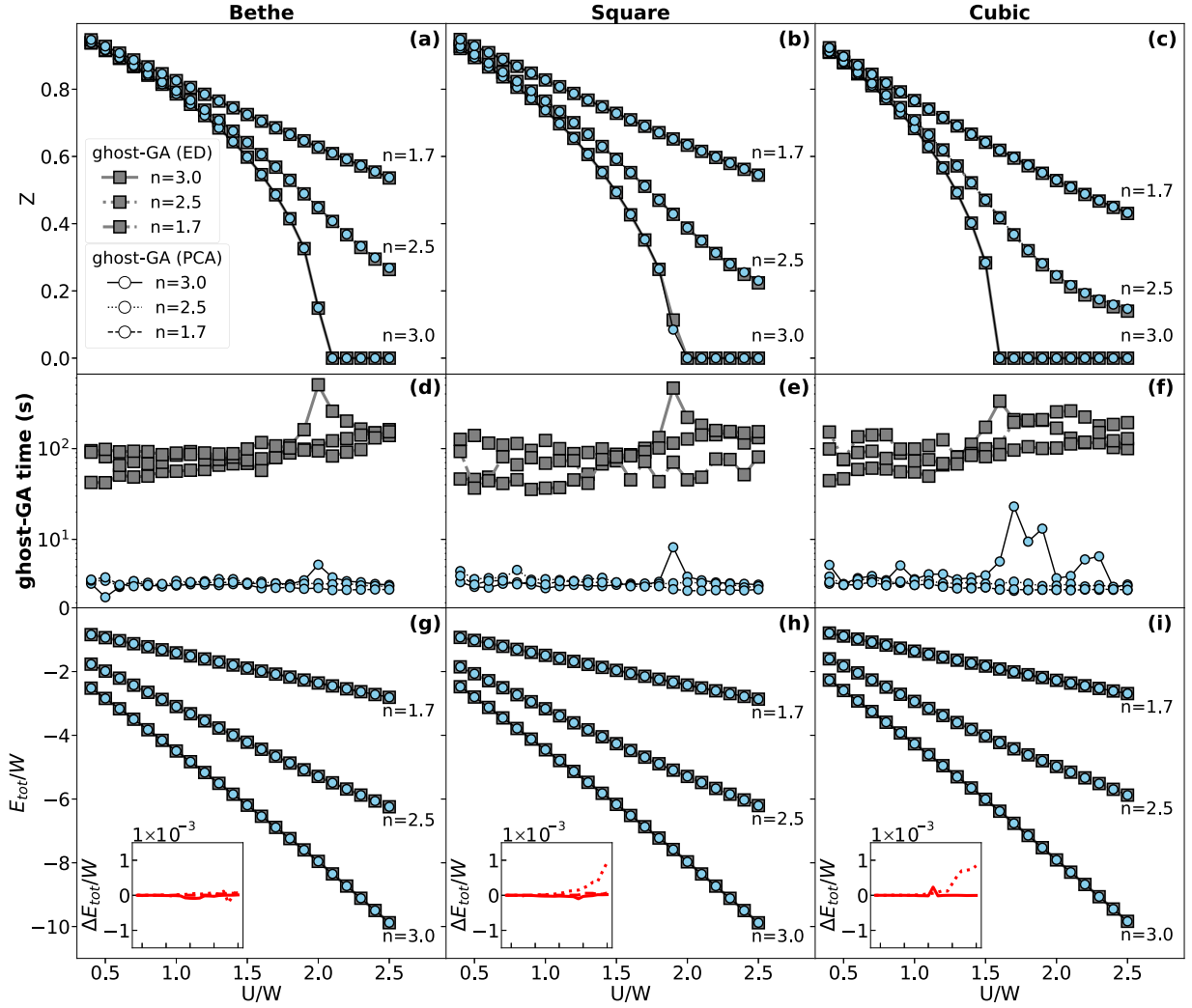


FIG. 4. Ghost-GA results for the three-degenerate-orbital Hubbard-Kanamori model with $J/U = 0.1$ on the Bethe (square, cubic) lattice for the left (center, right) column using pretrained PCA (circles) and ED (squares) as EH solvers. (a)–(c) Quasiparticle weight, (d)–(f) ghost-GA calculation time, and (g)–(i) total energy per site as a function of U/W at filling $n = 3.0, 2.5, 1.7$. The insets show the difference between the total ghost-GA energies per site computed with ED and PCA solvers for the EH.

On the Bethe lattice at half filling ($n = 3.0$), Figs. 5(a), 5(d), and 5(g) compare ghost-GA calculations using the ED solver for the EH (squares) and the active-learning PCA solver (circles). At $U/W = 0.4$, the database is empty, so EH calls are solved by ED (red markers), which triggers chart creation. Already at $U/W = 0.5$, the learned charts cover almost all EH calls (blue markers). Over the half-filled sweep, the EH solutions are learned with four charts created at $U/W = 0.4, 0.5, 0.9, 2.1$. The corresponding deviations in the total energy are shown in the inset of Fig. 5(g) and are 4 orders of magnitude smaller than the bandwidth W .

Continuing the Bethe-lattice calculations down to $n = 0.5$, in the first column of Fig. 5 we report representative results for Z [Fig. 5(a)], ghost-GA calculation time [Fig. 5(d)], and total energy [Fig. 5(g)] at $n = 3.0, 2.5, 1.0$. After the initial learning at half filling, only a small number of additional charts is required; in the present calculations, the atlas grows to a total of 13 charts. The resulting database of compressed solvers has size ~ 100 MB, corresponding to an order-of-magnitude

reduction in memory footprint with respect to the pretraining approach.

Using the database accumulated on the Bethe lattice, the center (square) and right (cubic) columns of Fig. 5 report the corresponding results for lattices with different geometries. For the data points considered here, no additional training steps are required. In both cases, the error in the total energy remains 4 orders of magnitude smaller than the half-bandwidth, while the overall ghost-GA calculation time is reduced by approximately a factor of 30 [Figs. 5(e) and 5(f)]. We note that for the square lattice a single point shows a slightly larger deviation, which can be attributed to the increased difficulty of accurately resolving the solution in proximity to the Mott transition within the present approximate solver. This deviation is consistent with the increased computational effort observed at that point and reflects the intrinsic difficulty of this regime.

Overall, active learning reduces the number of required charts from 990 (global pretraining) to 13 in this

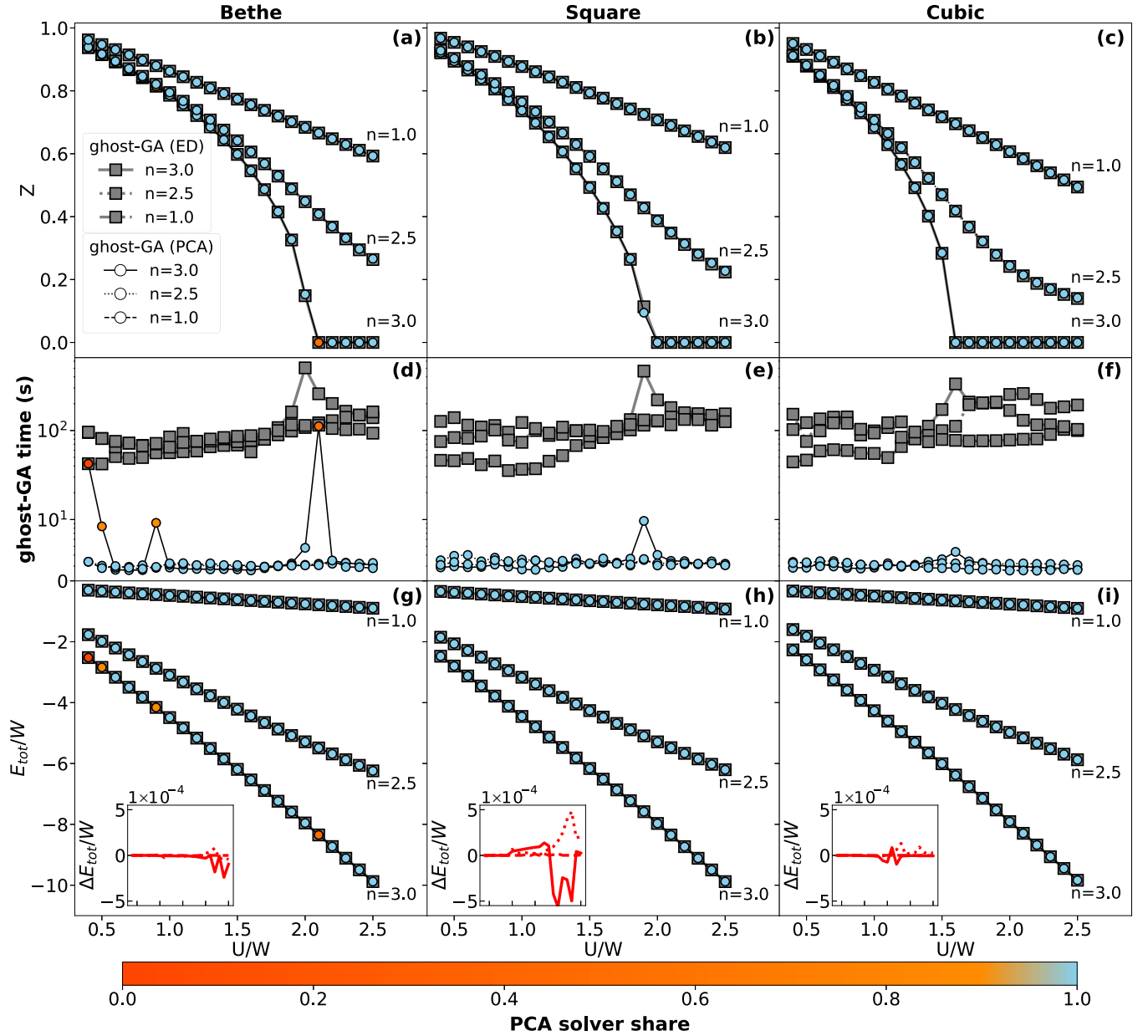


FIG. 5. Ghost-GA results for the three-degenerate-orbital Hubbard-Kanamori model with $J/U = 0.1$ on the Bethe (square, cubic) lattice in the left (center, right) column using the active-learning PCA solver (circles) and ED (squares) as EH solvers. (a)–(c) Quasiparticle weight, (d)–(f) ghost-GA calculation time, and (g)–(i) total energy per site as a function of U/W at fillings $n = 3.0, 2.5, 1.0$. Insets: differences between the total energies per site computed with ED and PCA solvers for the EH. Marker color indicates the fraction of EH solves within the ghost-GA cycle handled by the PCA solver (red: 0; blue: 1).

scan, yielding substantial savings in both computational cost and memory footprint, while retaining accuracy and transferability.

D. Active-learning application to all phases of Pu

In this section, our aim is to assess the performance benefits afforded by the PCA-based active-learning framework in a realistic materials context, where the environments of different structural fragments—here encoded in the embedding Hamiltonians—are constrained by physical factors such as crystal symmetry, structural stability, and the nature of

chemical bonding. To this end, we apply the PCA-compressed ghost-GA solver to a concrete real-material problem: computing the energy-volume $E(V)$ curves of six crystalline phases of Pu with a *single* trained solver reused for all volumes and phases.

In the Pu systems, the f -electron states are split by spin-orbit coupling and crystal-field splitting (CFS). Since SOC is dominant, the standard CFS-averaged approximation is adopted, where the f shell is treated as isotropic within the EH, and correlation effects distinguish only the $J = 5/2$ and $J = 7/2$ manifolds. We set $B = 1$, $U = 4$ eV, and $J/U = 0.09$ for the practical calculations of all the Pu phases [36].

The embedding Hamiltonian takes the following form:

$$\begin{aligned}
 \hat{H}_{\text{emb}}(\mathbf{X}) = & \hat{H}_{\text{int}}(U, J) + E_{5/2} \sum_{j_z=-5/2}^{5/2} c_{5/2,j_z}^\dagger c_{5/2,j_z} + E_{7/2} \sum_{j_z=-7/2}^{7/2} c_{7/2,j_z}^\dagger c_{7/2,j_z} \\
 & + \epsilon_{5/2} \sum_{j_z=-5/2}^{5/2} b_{5/2,j_z}^\dagger b_{5/2,j_z} + \epsilon_{7/2} \sum_{j_z=-7/2}^{7/2} b_{7/2,j_z}^\dagger b_{7/2,j_z} \\
 & + v_{5/2} \sum_{j_z=-5/2}^{5/2} (c_{5/2,j_z}^\dagger b_{5/2,j_z} + \text{H.c.}) + v_{7/2} \sum_{j_z=-7/2}^{7/2} (c_{7/2,j_z}^\dagger b_{7/2,j_z} + \text{H.c.}), \quad (32)
 \end{aligned}$$

with independent parameters $\mathbf{X} = (E_{5/2} - E_{7/2}, v_{5/2}, v_{7/2}, \epsilon_{5/2}, \epsilon_{7/2})$.

The EH structure above is identical for every Pu phase and volume considered. The training and inference therefore always occur in the same symmetry sector with fixed half filling at $N_e = 14$ and total angular momentum $J = 0$ for the impurity-bath system. In numerical ED calculations, we employ a Fock-state basis with $N_e = 14$ and total angular momentum projection $J_z = 0$, resulting in a basis set dimension of 2831 584.

We construct the training set within a hyperrectangular box in the five-dimensional (5D) parameter space of $\mathbf{X}$. The range of $\mathbf{X}$ is chosen to be about twice the parameter value range in self-consistent LDA + ghost-GA calculation for δ -Pu at a minimal volume ($14.5 \text{ \AA}^3/\text{atom}$) and maximal volume point ($33.5 \text{ \AA}^3/\text{atom}$). Specifically, the range is $[0.25, 0.29]$ for $(E_{5/2} - E_{7/2})/U$, $[-1, 0]$ for $v_{5/2}/U$ and $v_{7/2}/U$, $[3.38, 6.10]$ for $\epsilon_{5/2}/U$, and $[2.92, 5.48]$ for $\epsilon_{7/2}/U$. From this hyperrectangular box, we randomly draw $N = 500$ sets of parameters to construct EH instances, compute their exact ground states via ED, and apply PCA analysis to the resulting columnwise stacked state vectors. We retain $K = 250$ principal components to define the variational subspace with the singular value $\sigma_a \geq 10^{-4}$, as shown in Fig. 6(c). The projected operators $\{\hat{H}_r^{(K)}\}$ are precomputed once and cached, requiring only 14 MB of storage. The computational cost of constructing the PCA model is dominated by the exact diagonalization of the embedding Hamiltonian, each taking approximately 4 min on a single core, resulting in a total run-time of about 34 CPU hours. The variational subspace is fixed and reused for every EH solution in the LDA + ghost-GA calculations of all volumes and across all Pu phases without retraining, patching, or basis updates.

In Fig. 6(a), we compare LDA + ghost-GA $E(V)$ curves obtained with the ED solver and with the single PCA-compressed solver for the α , β , δ , δ' , ϵ , and γ phases. Across the entire sampled volume range, the two solvers are indistinguishable at the plot scale. The volumewise energy differences are shown in Fig. 6(b), which are mostly smaller than 5 meV per atom.

Because the same $K = 250$ basis is used for every EH call, the savings are uniform across phases and volumes. Replacing ED with the PCA solver reduces the cumulative time spent inside the EH routine by a factor of ~ 4000 (~ 200 min versus 0.05 min), while substantially lowering memory usage by diagonalizing a Hamiltonian in the $K \times K$ subspace instead of

the full Hilbert space. Thus, a single training on one 5D box yields ED-level accuracy for all considered Pu phases with orders-of-magnitude reductions in CPU time and RAM.

VI. SCALABILITY AND INTEGRATION WITH HIGH-LEVEL SOLVERS

The results presented thus far establish that the ground-state manifold of the embedding Hamiltonian admits a highly

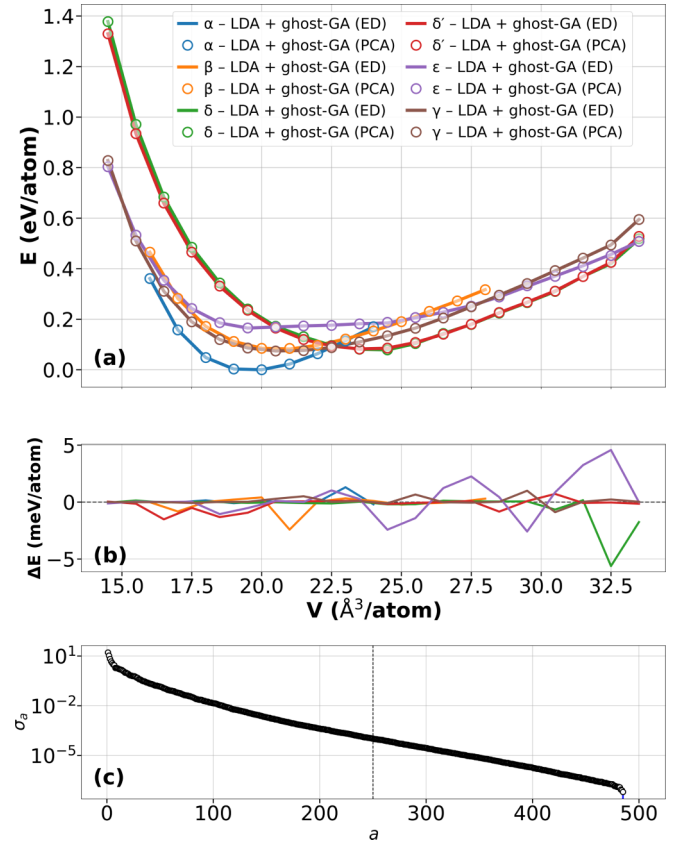


FIG. 6. (a) Energy-volume curves for the α , β , δ , δ' , ϵ , and γ phases of Pu. Lines: LDA + ghost-GA with exact ED solver. Circle markers: LDA + ghost-GA with the same PCA-compressed solver trained once on a fixed EH-parameter box and reused for all phases and volumes. (b) The energy difference between the calculations with ED and PCA solvers, $\Delta E = E_{\text{ED}} - E_{\text{PCA}}$, for each of the six phases. (c) Singular values σ_a of matrix $\mathbb{M}$ in descending order.

compact representation. This property has enabled us to drastically accelerate simulations of f -electron materials and multiorbital models using training data generated by exact diagonalization. However, extending this methodology to regimes governed by even larger Hilbert spaces requires overcoming the exponential scaling of the basis size. This challenge arises, for example, in ghost-GA calculations with larger bath sizes ($B \geq 3$), such as for transition metal oxides [74] or cluster impurities, where the dimension of the embedding Hamiltonian precludes the use of Arnoldi solvers. In these regimes, one must rely on high-level approaches like MPS [60–65], NQS, or variational impurity solvers based on superpositions of Gaussian states [66,67].

While these solvers can effectively handle larger systems, integrating them with the PCA framework described in Sec. IV poses a specific technical hurdle: standard PCA requires the explicit storage of full wavefunction vectors, which is often infeasible or computationally prohibitive for high-level representations. To unlock the capability of our framework to interface with these advanced solvers, it is necessary to implement a generalized formulation that remains agnostic to the underlying data structure. In this section, we introduce the “method of snapshots,” which constructs the variational space relying solely on state overlaps, providing a general pathway for scalable quantum embedding simulations.

A. Training PCA models from general impurity solvers using the “method of snapshots”

The core observation underlying this generalization is that the construction of the PCA variational space does not fundamentally require explicit state vectors, but only access to the overlaps between training states. This naturally leads to the formulation known as the “method of snapshots,” originally introduced in the context of reduced-order modeling in fluid dynamics, which recasts the diagonalization problem in terms of the overlap matrix in the space of samples rather than in the full Hilbert space. By lifting the restriction to solvers that explicitly return the wavefunction as a column vector, this reformulation makes the PCA-based solver compatible with any impurity solver capable of computing overlaps between ground states.

First, we diagonalize the overlap matrix of the sampled states:

$$\sum_{\alpha, \beta=1}^N V_{a\alpha}^* \langle \Phi(\mathbf{X}^\alpha) | \Phi(\mathbf{X}^\beta) \rangle V_{\beta b} = \sigma_a^2 \delta_{ab}, \quad (33)$$

where $a, b = 1, \dots, N$ and V is the matrix of eigenvectors. Note that this matrix is only $N \times N$, and so it is considerably smaller than the dimension of the Hilbert space M_B . Second, this diagonalization produces a set of orthonormal vectors:

$$|u_a\rangle = \sigma_a^{-1} \sum_{\alpha=1}^N V_{a\alpha} |\Phi(\mathbf{X}^\alpha)\rangle, \quad (34)$$

with σ_a being the singular values sorted in descending order. This approach provides an indirect but exact way of computing the singular values and principal components of the training data, bypassing the need to store wavefunctions as explicit M_B -dimensional arrays.

Significantly, Eqs. (33) and (34) require only the overlaps $\langle \Phi(\mathbf{X}^\alpha) | \Phi(\mathbf{X}^\beta) \rangle$ and linear combinations of the corresponding states. As a result, the same PCA-based EH reduction can be trained on data generated by virtually any advanced impurity solver, including MPS-based methods or quantum-assisted algorithms [68,70,71]. In this way, the PCA framework presented here becomes a general reduction layer that can sit on top of a wide class of solvers, amortizing their cost over many ghost-GA calculations.

The formulation of the PCA solver via the method of snapshots [Eqs. (33) and (34)] essentially decouples the generation of the variational basis from the solution of the embedding problem. This separation offers a strategic advantage when integrating with emerging NQS solvers. Recent benchmarks of NQS impurity solvers for quantum embedding have identified that, while the variational optimization of the wavefunction is computationally efficient, the subsequent “inference” step—sampling physical observables with the high precision required for self-consistency—remains a primary computational bottleneck [94]. Our PCA framework may provide a direct resolution to this inference problem. Instead of requiring the NQS to sample observables repeatedly during the self-consistency loop, one can use the NQS solely as a high-quality generator of basis snapshots. Once these states are projected into the compressed PCA subspace via the method of snapshots, all subsequent observable calculations become deterministic matrix operations within the low-dimensional subspace. This effectively shifts the heavy computational burden of Monte Carlo sampling entirely to the pretraining phase, leaving a fast, noise-free solver for the embedding loop.

B. Integration with neural-network quantum states

Looking beyond the generation of snapshots, the existence of a compact variational basis suggests the potential for more advanced hybrid architectures. Recent foundation models based on neural networks have demonstrated the capability to learn ground-state approximations across continuous families of Hamiltonians by taking the physical couplings as explicit inputs [93–95]. A promising future direction lies in combining these neural-network approaches with the linear subspace structure identified in this work. By explicitly exploiting problem-specific mathematical features such as the affine dependence of the embedding Hamiltonian on its parameters, it may be possible to construct hybrid variational ansätze that learn the optimal linear basis and the mixing coefficients simultaneously. Such strategies, combining the representational power of neural networks with the physical robustness of linear projections, may enable simulations of strongly correlated systems with even larger and more complex embedding problems.

VII. CONCLUSIONS

In this work, we introduced a linear foundation model for ghost-GA based on learning low-dimensional variational spaces for families of embedding Hamiltonians. Using PCA, we constructed local linear subspaces adapted to the EH ground-state manifold and projected the EH onto these subspaces. Once the projected operators are precomputed, each

EH problem is reduced to a small deterministic eigenproblem, enforcing by construction N -representability conditions [86] (e.g., the Pauli principle). Furthermore, we developed an active-learning strategy to generate the variational basis “on demand,” avoiding redundant training. We benchmarked the method on multiorbital Hubbard models and on all crystalline phases of Pu. In the Hubbard model, we showed that a database trained on the Bethe lattice could be transferred to square and cubic lattices without additional training, requiring only a small number of local domains (13) to cover the entire phase diagram. In the case of plutonium, the efficiency was even more pronounced: a single training box proved sufficient to describe the ground states across all six allotropes. This reflects that our method effectively captures the intrinsic similarity of the local $5f$ -shell environment across different structural phases. By extracting the universal features of the local chemical bonding from a single representative domain, the solver avoids redundant training and generalizes across the phase diagram, reproducing ghost-GA total energies obtained with full ED impurity solvers to within a few meV per atom, while reducing the cost of the EH step by orders of magnitude. These applications leverage the fact that the physically relevant embedding Hamiltonians, representing the interactions between fragments and their environments, span a much more limited range within the full parameter space—which is the same physical principle underlying scalability in foundation models for universal interatomic potentials [99–101]. Finally, we showed that this PCA construction can be reformulated via the method of snapshots, enabling the integration of data from high-level solvers such as MPS [60–65] and neural-network

quantum states [93–95]. Overall, our results establish that the essential information required for the embedding loop can be efficiently learned. By compressing the relevant Hilbert space, we can effectively eliminate the primary computational bottleneck of the QE frameworks, opening a path toward accurate high-throughput *ab initio* investigations of strongly correlated matter.

ACKNOWLEDGMENTS

N.L. gratefully acknowledges funding from the National Science Foundation under Award No. DMR-2532771 and from the Simons Foundation (Grant No. 00010910). The Flatiron Institute is a division of the Simons Foundation. Part of this work by Y.-X.Y. was supported by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences, Materials Science and Engineering Division, including the grant of computer time at the National Energy Research Scientific Computing Center (NERSC) in Berkeley, California. This part of research was performed at the Ames National Laboratory, which is operated for the U.S. DOE by Iowa State University under Contract No. DE-AC02-07CH11358. T.-H.L. gratefully acknowledges funding from the National Science and Technology Council (NSTC) of Taiwan under Grant No. NSTC 112-2112-M-194-007-MY3 and the National Center for Theoretical Sciences (NCTS) in Taiwan.

DATA AVAILABILITY

The data that support the findings of this article are openly available [102].

- [1] Materials genome initiative for global competitiveness, National Science and Technology Council (U.S.) (2011), available at: https://www.mgi.gov/sites/mgi/files/materials_genome_initiative-final.pdf.
- [2] L.-Q. Chen, L.-D. Chen, S. V. Kalinin, G. Klimeck, S. K. Kumar, J. Neugebauer, and I. Terasaki, Design and discovery of materials guided by theory and computation, *npj Comput. Mater.* **1**, 15007 (2015).
- [3] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964).
- [4] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* **140**, A1133 (1965).
- [5] P. R. C. Kent and G. Kotliar, Toward a predictive theory of correlated materials, *Science* **361**, 348 (2018).
- [6] N. F. Mott, *Metal-Insulator Transitions*, 2nd ed. (Taylor & Francis, London, 1990).
- [7] J. A. Schiller, L. K. Wagner, and E. Ertekin, Phase stability and properties of manganese oxide polymorphs: Assessment and insights from diffusion Monte Carlo, *Phys. Rev. B* **92**, 235209 (2015).
- [8] H. Peng and S. Lany, Polymorphic energy ordering of MgO, ZnO, GaN, and MnO within the random phase approximation, *Phys. Rev. B* **87**, 174113 (2013).
- [9] A. Schrön, C. Rödl, and F. Bechstedt, Energetic stability and magnetic properties of MnO in the rocksalt, wurtzite, and zinc-blende structures: Influence of exchange and correlation, *Phys. Rev. B* **82**, 165109 (2010).
- [10] J. Kolorenč and L. Mitás, Quantum Monte Carlo calculations of structural properties of FeO under pressure, *Phys. Rev. Lett.* **101**, 185502 (2008).
- [11] I. Leonov, L. Pourovskii, A. Georges, and I. A. Abrikosov, Magnetic collapse and the behavior of transition metal oxides at high pressure, *Phys. Rev. B* **94**, 155135 (2016).
- [12] K. Haule, Structural predictions for correlated electron materials using the functional dynamical mean field theory approach, *J. Phys. Soc. Jpn.* **87**, 041005 (2018).
- [13] N. Lanatà, T.-H. Lee, Y.-X. Yao, V. Stevanović, and V. Dobrosavljević, Connection between Mott physics and crystal structure in a series of transition metal binary compounds, *npj Comput. Mater.* **5**, 30 (2019).
- [14] J. N. Cross, T.-H. Lee, C.-J. Kang, Y.-X. Yao, S. K. Cary, J. T. Stritzinger, M. J. Polinski, C. D. McKinley, T. E. Albrecht-Schmitt, and N. Lanatà, Origins of the odd optical observables in plutonium and americium tungstates, *Chem. Sci.* **10**, 6508 (2019).
- [15] S. S. Galley, A. A. Arico, T.-H. Lee, X. Deng, Y.-X. Yao, J. M. Sperling, V. Proust, J. S. Storbeck, V. Dobrosavljevic, J. N. Neu, T. Siegrist, R. E. Baumbach, T. E. Albrecht-Schmitt, N. Kaltsoyannis, and N. Lanatà, Uncovering the origin of divergence in the $\text{CsM}(\text{CrO}_4)_2$ ($M = \text{La, Pr, Nd, Sm, Eu, Am}$) family through the chemical bonding in a molecular cluster

- and by band structure analysis, *J. Am. Chem. Soc.* **140**, 1674 (2018).
- [16] P. A. Lee, N. Nagaosa, and X.-G. Wen, Doping a Mott insulator: Physics of high-temperature superconductivity, *Rev. Mod. Phys.* **78**, 17 (2006).
- [17] P. Johnpierre and R. L. Greene, High-temperature superconductivity in iron-based materials, *Nat. Phys.* **6**, 645 (2010).
- [18] M. Yi, Y. Zhang, Z.-X. Shen, and D. Lu, Role of the orbital degree of freedom in iron-based superconductors, *npj Comput. Mater.* **2**, 57 (2017).
- [19] L. Miao, R. Basak, S. Ran, Y. Xu, E. Kotta, H. He, J. D. Denlinger, Y.-D. Chuang, Y. Zhao, Z. Xu, J. W. Lynn, J. R. Jeffries, S. R. Saha, I. Giannakis, P. Aynajian, C.-J. Kang, Y. Wang, G. Kotliar, N. P. Butch, and L. A. Wray, High temperature singlet-based magnetism from Hund's rule correlations, *Nat. Commun.* **10**, 644 (2019).
- [20] E. M. Nica, R. Yu, and Q. Si, Orbital-selective pairing and superconductivity in iron selenides, *npj Comput. Mater.* **2**, 24 (2017).
- [21] P. Dai, J. Hu, and E. Dagotto, Magnetism and its microscopic origin in iron-based high-temperature superconductors, *Nat. Phys.* **8**, 709 (2012).
- [22] J.-F. Ge, Z.-L. Liu, C. Liu, C.-L. Gao, D. Qian, Q.-K. Xue, Y. Liu, and J.-F. Jia, Superconductivity above 100 K in single-layer FeSe films on doped SrTiO₃, *Nat. Mater.* **14**, 285 (2015).
- [23] S. Manna, A. Kamlapure, L. Cornils, T. Hanke, E. M. J. Hedegaard, M. Bremholm, B. B. Iversen, P. Hofmann, J. Wiebe, and R. Wiesendanger, Interfacial superconductivity in a bi-collinear antiferromagnetically ordered FeTe monolayer on a topological insulator, *Nat. Commun.* **8**, 14074 (2017).
- [24] V. Anisimov and Y. Izyumov, *Electronic Structure of Strongly Correlated Materials* (Springer, Berlin, 2010).
- [25] V. I. Anisimov, A. I. Oteryaev, M. A. Korotin, A. O. Anokhin, and G. Kotliar, First-principles calculations of the electronic structure and spectra of strongly correlated systems: Dynamical mean-field theory, *J. Phys.: Condens. Matter* **9**, 7359 (1997).
- [26] K. Held, A. Nekrasov, G. Keller, V. Eyert, N. Blumer, A. K. McMahan, R. T. Scalettar, Th. Pruschke, V. I. Anisimov, and D. Vollhardt, Realistic investigations of correlated electron systems with LDA + DMFT, *Phys. Status Solidi (B)* **243**, 2599 (2006).
- [27] 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).
- [28] J.-Z. Zhao, J.-N. Zhuang, X.-Y. Deng, Y. Bi, L.-C. Cai, Z. Fang, and X. Dai, Implementation of LDA + DMFT with the pseudo-potential-plane-wave method, *Chin. Phys. B* **21**, 057106 (2012).
- [29] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti, Electronic structure calculations with dynamical mean-field theory, *Rev. Mod. Phys.* **78**, 865 (2006).
- [30] G. Knizia and G. K.-L. Chan, Density matrix embedding: A simple alternative to dynamical mean-field theory, *Phys. Rev. Lett.* **109**, 186404 (2012).
- [31] G. H. Booth and G. K.-L. Chan, Spectral functions of strongly correlated extended systems via an exact quantum embedding, *Phys. Rev. B* **91**, 155107 (2015).
- [32] F. Lechermann, A. Georges, G. Kotliar, and O. Parcollet, Rotationally invariant slave-boson formalism and momentum dependence of the quasiparticle weight, *Phys. Rev. B* **76**, 155102 (2007).
- [33] T. Li, P. Wolfle, and P. J. Hirschfeld, Spin-rotation-invariant slave-boson approach to the Hubbard model, *Phys. Rev. B* **40**, 6817 (1989).
- [34] N. Lanata, Y. Yao, X. Deng, V. Dobrosavljevic, and G. Kotliar, Slave boson theory of orbital differentiation with crystal field effects: Application to UO₂, *Phys. Rev. Lett.* **118**, 126401 (2017).
- [35] M. C. Gutzwiller, Correlation of electrons in a narrow *s* band, *Phys. Rev.* **137**, A1726 (1965).
- [36] N. Lanata, Y. X. Yao, C.-Z. Wang, K.-M. Ho, and G. Kotliar, Phase diagram and electronic structure of praseodymium and plutonium, *Phys. Rev. X* **5**, 011008 (2015).
- [37] N. Lanata, P. Barone, and M. Fabrizio, Fermi-surface evolution across the magnetic phase transition in the Kondo lattice model, *Phys. Rev. B* **78**, 155127 (2008).
- [38] W. Metzner and D. Vollhardt, Correlated lattice Fermions in $d = \infty$ dimensions, *Phys. Rev. Lett.* **62**, 324 (1989).
- [39] X.-Y. Deng, L. Wang, X. Dai, and Z. Fang, Local density approximation combined with Gutzwiller method for correlated electron systems: Formalism and applications, *Phys. Rev. B* **79**, 075114 (2009).
- [40] K. M. Ho, J. Schmalian, and C. Z. Wang, Gutzwiller density functional theory for correlated electron systems, *Phys. Rev. B* **77**, 073101 (2008).
- [41] N. Lanata, H. U. R. Strand, X. Dai, and B. Hellsing, Efficient implementation of the Gutzwiller variational method, *Phys. Rev. B* **85**, 035133 (2012).
- [42] J. Bunemann, F. Gebhard, T. Ohm, S. Weiser, and W. Weber, Gutzwiller-correlated wave functions: Application to ferromagnetic nickel, in *Frontiers in Magnetic Materials*, edited by A. V. Narlikar (Springer, Berlin, 2005), p. 117.
- [43] C. Attacalite and M. Fabrizio, Properties of Gutzwiller wave functions for multiband models, *Phys. Rev. B* **68**, 155117 (2003).
- [44] N. Lanata, T.-H. Lee, Y.-X. Yao, and V. Dobrosavljevic, Emergent Bloch excitations in Mott matter, *Phys. Rev. B* **96**, 195126 (2017).
- [45] M. S. Frank, T.-H. Lee, G. Bhattacharyya, P. K. H. Tsang, V. L. Quito, V. Dobrosavljevic, O. Christiansen, and N. Lanata, Quantum embedding description of the Anderson lattice model with the ghost Gutzwiller approximation, *Phys. Rev. B* **104**, L081103 (2021).
- [46] N. Lanata, Operatorial formulation of the ghost rotationally invariant slave-boson theory, *Phys. Rev. B* **105**, 045111 (2022).
- [47] N. Lanata, Derivation of the ghost Gutzwiller approximation from quantum embedding principles: Ghost density matrix embedding theory, *Phys. Rev. B* **108**, 235112 (2023).
- [48] L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, and A. J. Millis, Machine learning for many-body physics: The case of the Anderson impurity model, *Phys. Rev. B* **90**, 155136 (2014).
- [49] E. Sheridan, C. Rhodes, F. Jamet, I. Rungger, and C. Weber, Data-driven dynamical mean-field theory: An error-correction

- approach to solve the quantum many-body problem using machine learning, *Phys. Rev. B* **104**, 205120 (2021).
- [50] E. J. Sturm, M. R. Carbone, D. Lu, A. Weichselbaum, and R. M. Konik, Predicting impurity spectral functions using machine learning, *Phys. Rev. B* **103**, 245118 (2021).
- [51] X.-Y. Ren, R.-S. Han, and L. Chen, Learning impurity spectral functions from density of states, *J. Phys.: Condens. Matter* **33**, 495601 (2021).
- [52] C. Venturella, C. Hillenbrand, J. Li, and T. Zhu, Machine learning many-body Green's functions for molecular excitation spectra, *J. Chem. Theory Comput.* **20**, 143 (2024).
- [53] E. Agapov, O. Bertomeu, A. Carballo, C. B. Mendl, and A. Sander, Predicting interacting Green's functions with neural networks, [arXiv:2411.13644](https://arxiv.org/abs/2411.13644).
- [54] F. Deng, Y. Lu, X. Cao, and Z. Zhong, Neural network impurity solver for real-frequency dynamical mean-field theory, [arXiv:2511.14505](https://arxiv.org/abs/2511.14505).
- [55] H. Lee, Z. Zhao, G. H. Booth, W. Ge, and C. Weber, Language-inspired machine learning approach for solving strongly correlated problems with dynamical mean-field theory, *Phys. Rev. B* **112**, 035165 (2025).
- [56] X. Dong, E. Gull, and L. Wang, Equivariant neural network for Green's functions of molecules and materials, *Phys. Rev. B* **109**, 075112 (2024).
- [57] P. Mitra and H. Banerjee, Deep learning-based prediction of self-energies from *ab initio* dynamical mean-field theory for real materials with minimal data sets, *ChemRxiv*, <https://doi.org/10.26434/chemrxiv-2025-dp7rd>.
- [58] J. Rogers, T.-H. Lee, S. Pakdel, W. Xu, V. Dobrosavljević, Y.-X. Yao, O. Christiansen, and N. Lanatà, Bypassing the computational bottleneck of quantum-embedding theories for strong electron correlations with machine learning, *Phys. Rev. Res.* **3**, 013101 (2021).
- [59] M. S. Frank, D. G. Artiukhin, T.-H. Lee, Y. Yao, K. Barros, O. Christiansen, and N. Lanatà, Active learning approach to simulations of strongly correlated matter with the ghost Gutzwiller approximation, *Phys. Rev. Res.* **6**, 013242 (2024).
- [60] M. Fishman, S. R. White, and E. M. Stoudenmire, The ITensor software library for tensor network calculations, *SciPost Phys. Codebases* **4** (2022).
- [61] H. Zhai, H. R. Larsson, S. Lee, Z.-H. Cui, T. Zhu, C. Sung, L. Peng, R. Peng, K. Liao, J. Tölle, J. Yang, S. Li, and G. K.-L. Chan, Block2: A comprehensive open source framework to develop and apply state-of-the-art DMRG algorithms in electronic structure and beyond, *J. Chem. Phys.* **159**, 234801 (2023).
- [62] U. Schollwöck, The density-matrix renormalization group, *Rev. Mod. Phys.* **77**, 259 (2005).
- [63] S. R. White, Density matrix formulation for quantum renormalization groups, *Phys. Rev. Lett.* **69**, 2863 (1992).
- [64] S. R. White, Density-matrix algorithms for quantum renormalization groups, *Phys. Rev. B* **48**, 10345 (1993).
- [65] X. Cao, Y. Lu, P. Hansmann, and M. W. Haverkort, Tree tensor-network real-time multiorbital impurity solver: Spin-orbit coupling and correlation functions in Sr_2RuO_4 , *Phys. Rev. B* **104**, 115119 (2021).
- [66] S. Boutin and B. Bauer, Quantum impurity models using superpositions of fermionic Gaussian states: Practical methods and applications, *Phys. Rev. Res.* **3**, 033188 (2021).
- [67] A.-K. Wu, B. Kloss, W. Krinitsin, M. T. Fishman, J. H. Pixley, and E. M. Stoudenmire, Disentangling interacting systems with fermionic Gaussian circuits: Application to quantum impurity models, *Phys. Rev. B* **111**, 035119 (2025).
- [68] P. V. Sriluckshmy, F. Jamet, and F. Šimkovic IV, Quantum assisted ghost Gutzwiller ansatz, [arXiv:2506.21431](https://arxiv.org/abs/2506.21431).
- [69] N. Gomes, A. Mukherjee, F. Zhang, T. Iadecola, C.-Z. Wang, K.-M. Ho, P. P. Orth, and Y.-X. Yao, Adaptive variational quantum imaginary time evolution approach for ground state preparation, *Adv. Quantum Technol.* **4**, 2100114 (2021).
- [70] T. Jiang, J. Rogers, M. S. Frank, O. Christiansen, Y.-X. Yao, and N. Lanatà, Error mitigation in variational quantum eigen-solvers using tailored probabilistic machine learning, *Phys. Rev. Res.* **6**, 033069 (2024).
- [71] I.-C. Chen, A. Khindanov, C. M. Salazar, H. M. Barona, G. Harrabi, F. Zhang, C.-Z. Wang, T. Iadecola, N. Lanatà, and Y.-X. Yao, Quantum-classical embedding via ghost Gutzwiller approximation for enhanced simulations of correlated electron systems, *npj Quantum Inf.* (2026), doi:10.1038/s41534-026-01244-9.
- [72] T.-H. Lee, N. Lanatà, and G. Kotliar, Accuracy of ghost rotationally invariant slave-boson and dynamical mean field theory as a function of the impurity-model bath size, *Phys. Rev. B* **107**, L121104 (2023).
- [73] T.-H. Lee, C. Melnick, R. Adler, N. Lanatà, and G. Kotliar, Accuracy of ghost-rotationally-invariant slave-boson theory for multiorbital Hubbard models and realistic materials, *Phys. Rev. B* **108**, 245147 (2023).
- [74] T.-H. Lee, C. Melnick, R. Adler, X. Sun, Y. Yao, N. Lanatà, and G. Kotliar, Charge self-consistent density functional theory plus ghost rotationally invariant slave-boson theory for correlated materials, *Phys. Rev. B* **110**, 115126 (2024).
- [75] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, First-principles calculations of the electronic structure and spectra of strongly correlated systems: The LDA + U method, *J. Phys.: Condens. Matter* **9**, 767 (1997).
- [76] A. J. Kim, P. Werner, and R. Valentí, Alleviating the sign problem in quantum Monte Carlo simulations of spin-orbit-coupled multiorbital Hubbard models, *Phys. Rev. B* **101**, 045108 (2020).
- [77] R. Levy and B. K. Clark, Mitigating the sign problem through basis rotations, *Phys. Rev. Lett.* **126**, 216401 (2021).
- [78] S. Giuli, C. Mejuto-Zaera, and M. Capone, Altermagnetism from interaction-driven itinerant magnetism, *Phys. Rev. B* **111**, L020401 (2025).
- [79] D. Guerci, M. Capone, and M. Fabrizio, Exciton Mott transition revisited, *Phys. Rev. Mater.* **3**, 054605 (2019).
- [80] C. Mejuto-Zaera and M. Fabrizio, Efficient computational screening of strongly correlated materials: Multiorbital phenomenology within the ghost Gutzwiller approximation, *Phys. Rev. B* **107**, 235150 (2023).
- [81] A. M. Tagliente, C. Mejuto-Zaera, and M. Fabrizio, Revealing spinons by proximity effect, *Phys. Rev. B* **111**, 125110 (2025).
- [82] C. Mejuto-Zaera, Quantum embedding for molecules using auxiliary particles—The ghost Gutzwiller ansatz, *Faraday Discuss.* **254**, 653 (2024).
- [83] I. Pasqua, A. M. Tagliente, G. Bellomia, B. Monserrat, M. Fabrizio, and C. Mejuto-Zaera, Quasiparticle band

- picture bridging topology and strong correlations across energy scales, *Phys. Rev. B* **113**, 155133 (2026).
- [84] A. Cao, V. J. Bukas, V. Shadravan, Z. Wang, H. Li, J. Kibsgaard, I. Chorkendorff, and J. K. Nørskov, A spin promotion effect in catalytic ammonia synthesis, *Nat. Commun.* **13**, 2382 (2022).
- [85] K. Zhang, A. Cao, L. H. Wandall, J. Vernieres, J. Kibsgaard, J. K. Nørskov, and I. Chorkendorff, Spin-mediated promotion of Co catalysts for ammonia synthesis, *Science* **383**, 1357 (2024).
- [86] A. J. Coleman, Structure of fermion density matrices, *Rev. Mod. Phys.* **35**, 668 (1963).
- [87] D. Frame, R. He, I. Ipsen, D. Lee, D. Lee, and E. Rrapaj, Eigenvector continuation with subspace learning, *Phys. Rev. Lett.* **121**, 032501 (2018).
- [88] Y. Rath and G. H. Booth, Interpolating numerically exact many-body wave functions for accelerated molecular dynamics, *Nat. Commun.* **16**, 2005 (2025).
- [89] K. Atalar, Y. Rath, R. Crespo-Otero, and G. H. Booth, Fast and accurate nonadiabatic molecular dynamics enabled through variational interpolation of correlated electron wavefunctions, *Faraday Discuss.* **254**, 542 (2024).
- [90] N. Hogan, E. Kökcü, T. Steckmann, L. P. Doak, C. Mejuto-Zaera, D. Camps, R. V. Beeumen, W. A. de Jong, and A. F. Kemper, Efficient quantum implementation of dynamical mean field theory for correlated materials, *arXiv:2508.05738*.
- [91] C. Mejuto-Zaera and A. F. Kemper, Quantum eigenvector continuation for chemistry applications, *Electron. Struct.* **5**, 045007 (2023).
- [92] T. Duguet, A. Ekström, R. J. Furnstahl, S. König, and D. Lee, *Colloquium*: Eigenvector continuation and projection-based emulators, *Rev. Mod. Phys.* **96**, 031002 (2024).
- [93] R. Rende, L. L. Viteritti, F. Becca, A. Scardicchio, A. Laio, and G. Carleo, Foundation neural-networks quantum states as a unified ansatz for multiple Hamiltonians, *Nat. Commun.* **16**, 7213 (2025).
- [94] Z. Zhouyin, T.-H. Lee, A. Chen, N. Lanatà, and H. Guo, Neural-quantum-states impurity solver for quantum embedding problems, *Phys. Rev. B* **113**, 155123 (2026).
- [95] H. Ma, H. Shang, and J. Yang, Quantum embedding method with transformer neural network quantum states for strongly correlated materials, *npj Comput. Mater.* **10**, 220 (2024).
- [96] P. W. Anderson, Infrared catastrophe in Fermi gases with local scattering potentials, *Phys. Rev. Lett.* **18**, 1049 (1967).
- [97] J. Kanamori, Electron correlation and ferromagnetism of transition metals, *Prog. Theor. Phys.* **30**, 275 (1963).
- [98] A. Georges, L. de' Medici, and J. Mravlje, Strong correlations from Hund's coupling, *Annu. Rev. Condens. Matter Phys.* **4**, 137 (2013).
- [99] A. E. A. Allen, N. Lubbers, S. Matin, J. S. Smith, R. A. Messerly, S. Tretiak, and K. Barros, Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, *npj Comput. Mater.* **10**, 154 (2024).
- [100] J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang, and S. Han, Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials, *J. Am. Chem. Soc.* **147**, 1042 (2025).
- [101] S. Ju, J. You, G. Kim, Y. Park, H. An, and S. Han, Application of pretrained universal machine-learning interatomic potential for physicochemical simulation of liquid electrolytes in Li-ion batteries, *Digit. Discov.* **4**, 1544 (2025).
- [102] S. Giuli, PCA-gGA-data: Data and plotting scripts for "Linear foundation model for quantum embedding: Data-driven compression of the ghost Gutzwiller variational space" (v1.0), Zenodo (2026), doi: [10.5281/zenodo.19632781](https://doi.org/10.5281/zenodo.19632781).