# PHYSICS-GUIDED LEARNING OF KOHN-SHAM HAMILTONIANS FOR PERIODIC SOLIDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Deep learning prediction of Kohn-Sham Hamiltonians overcomes the computational bottleneck of density functional theory (DFT) in large-scale electronic structure calculations. Most existing methods, however, fit Hamiltonians numerically without fully using the rich quantum-mechanical priors in DFT. We propose two methods guided by physical priors: the HSE Hamiltonian model H2Net and the variational surrogate loss (VSLoss). Inspired by the separation of short- and long-range interactions in the HSE functional, H2Net uses a dual-branch architecture. Fewer layers in the long-range branch limit the receptive field and reduce training memory requirements. The model also incorporates the non-self-consistent part of the Hamiltonian and a PBE Hamiltonian from a pretrained model as input features, explicitly introducing physical priors to improve prediction accuracy. VSLoss guides the model toward minimizing the ground-state energy, bringing the Hohenberg–Kohn variational principle into training to improve the accuracy of downstream physical properties calculated from the Hamiltonian. By improving model representations and optimization objectives, respectively, the two methods make comprehensive use of DFT priors from basis sets, pseudopotentials, functionals, and the variational principle. Together, they offer a new approach to learning Kohn–Sham Hamiltonians for periodic solids.

## 1 INTRODUCTION

The electronic structure of periodic solids is a central subject in condensed matter physics, chemistry, and materials science, as it governs key physical properties such as energy bands, density of states, transport, and optical responses. Density functional theory (DFT) has become a widely adopted approach for electronic structure calculations. It determines the ground-state charge density through the self-consistent field (SCF) iteration of the single-particle Kohn–Sham equations, from which the Hamiltonian and other downstream physical quantities can subsequently be obtained (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). However, the high computational cost of SCF iterations has become a critical bottleneck limiting the application of DFT to large-scale calculations (Bowler & Miyazaki, 2012). For the PBE functional, the computational complexity of the iterative process is $O(N^3)$, while that of hybrid functionals can be as high as $O(N^4)$. Consequently, deep learning methods have been introduced to electronic structure calculations. By leveraging neural networks to learn the complex nonlinear mapping from atomic coordinates to the Hamiltonian, these methods can bypass time-consuming iterative solvers and directly predict the Kohn–Sham Hamiltonian at a fraction of the computational cost, thereby enabling the efficient computation of physical quantities of interest, such as energy bands and wavefunctions (Li et al., 2022; Gong et al., 2023; Zhong et al., 2023).

On the other hand, the introduction of equivariance represents a significant advance in the development of deep learning models for Hamiltonian prediction. Methods such as DeepH-E3 incorporate equivariance into graph neural networks (GNNs), enabling their outputs to transform consistently under group transformations applied to the inputs—including rotation, translation, and reflection—and thereby leveraging geometric symmetry as a physical prior and enforcing the correct transformation behavior of the outputs (Gong et al., 2023; Zhong et al., 2023; Yu et al., 2023). Notably, the incorporation of equivariance into neural networks did not originate from Hamiltonian prediction tasks (Thomas et al., 2018; Batzner et al., 2022), nor is spatial symmetry the only prior knowledge available for the Kohn–Sham Hamiltonian (Unke et al., 2021; Nigam et al., 2022).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

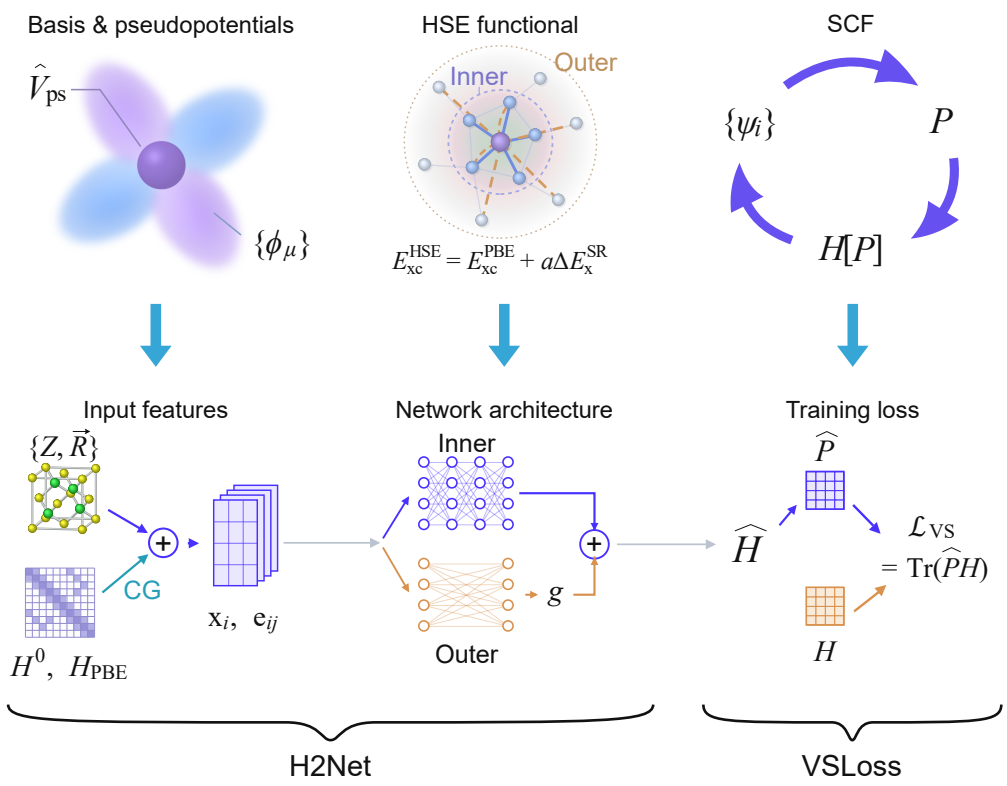


Figure 1: Physical priors and workflow of H2Net and the variational surrogate loss (VSLoss). The top row presents basis sets and pseudopotentials, the Heyd–Scuseria–Ernzerhof (HSE) functional, and self-consistent field (SCF) iterations. The bottom row maps these to input features, network architecture, and training loss.

This raises a deeper question: how can the specific prior knowledge rooted in quantum mechanics and solid-state physics and inherent to DFT be leveraged to guide Hamiltonian learning? DFT ingredients, such as basis sets, pseudopotentials, and functionals, contain rich information (Qiao et al., 2022). However, mainstream methods in the community tend to start from atomic coordinates and numerically regress the Hamiltonian, and therefore fail to fully exploit these sources of physical information. WANet has already shown that merely fitting the numerical values of the Hamiltonian does not guarantee accurate downstream physical quantities (Li et al., 2025). Several studies have begun to address this issue. Works such as DeepH-Zero (Li et al., 2024) and DEQHNet have explored integrating DFT computational paradigms into neural network architectures, achieving improvements in prediction accuracy (Zhang et al., 2024; Wang et al., 2024b). However, these approaches typically require substantial modifications to the model architecture, limiting their compatibility with mainstream Hamiltonian prediction frameworks and hindering their broader generalization and applicability. Compared with redesigning the entire network architecture, a more straightforward strategy is to inject DFT-provided physical priors into the model at different levels while preserving the mainstream Hamiltonian prediction framework. By translating these priors into mathematical forms compatible with existing learning paradigms, the model can leverage the physical information available from DFT calculations while retaining, as much as possible, the efficiency, scalability, and usability of existing methods.

Based on the above insights, this paper proposes two Hamiltonian prediction methods guided by physical priors: the HSE functional Hamiltonian model H2Net and the Variational Surrogate Loss (VSLoss). These two methods incorporate physical priors from three levels—basis sets and pseudopotentials, functionals, and fundamental principles of quantum mechanics—through input features, network architecture, and optimization objectives. This design enables the model to leverage physical information that previous methods do not fully exploit while preserving the mainstream architecture, leading to improvements in prediction accuracy and computational and storage effi-

ciency. Inspired by the formulation of the HSE functional (Heyd et al., 2003; Krukau et al., 2006), H2Net explicitly models the mixture of interactions using an equivariant neural network with a dual-branch architecture. Meanwhile, it extracts features through Clebsch-Gordan (CG) decomposition from the non-self-consistent partial Hamiltonian and the PBE Hamiltonian predicted by the universal Hamiltonian model Uni-HamGNN (Zhong et al., 2024). This enables the integration of multi-level physical priors spanning functionals, basis sets, pseudopotentials, and the PBE Hamiltonian. Compared with approaches such as DeepH-Hybrid that simply increase the cutoff radius, causing a substantial increase in the number of neighboring atoms and storage pressure (Tang et al., 2024), H2Net reduces the number of layers in the outer branch to control the model's receptive field, thereby significantly reducing GPU memory consumption during training while still improving prediction accuracy (Zhong et al., 2024). VSLoss incorporates the Hohenberg–Kohn variational principle (Hohenberg & Kohn, 1964) into deep learning training under certain approximations. As a result, the model no longer merely fits the numerical values of matrix elements during training; instead, guided by the surrogate loss function, it indirectly minimizes the ground-state energy and thereby enforces the variational principle, improving the prediction accuracy of downstream physical quantities (Hohenberg & Kohn, 1964; Li et al., 2024). These two methods exploit physical priors from complementary perspectives of model representation and optimization objectives, respectively: the former leverages existing physical information in DFT to enhance the model's expressive capacity for the Hamiltonian, while the latter draws on fundamental physical principles to provide a physically meaningful optimization objective. The two approaches complement each other, enabling deep learning models to more thoroughly leverage the rich physical prior information embedded in DFT without requiring substantial modifications to existing architectures.

## 2 METHOD

### 2.1 EQUIVARIANT GATED MATRIX FEATURES: BASIS-SET AND PSEUDOPOTENTIAL PRIORS

We introduce equivariant gated matrix features (EGMF) to explicitly incorporate physical priors from basis sets and pseudopotentials into deep learning models. Existing Hamiltonian prediction models usually start from atomic coordinates alone, leaving out the rich physical information associated with the choice of basis set and pseudopotential. Some studies use the non-self-consistent Hamiltonian $H^0$ for residual learning (Zhong et al., 2023), but this linear subtraction only allows the network to infer information indirectly from its prediction targets, rather than learn complex combinations of features. We instead supply precomputed matrices, such as the overlap matrix $S$ and the non-self-consistent part of the Hamiltonian $H^0$, directly as auxiliary features before message passing begins. This modular design is easy to integrate: $S$ and $H^0$ can be used for PBE Hamiltonian prediction, while a PBE Hamiltonian predicted by a universal model can serve as a strong physical prior for the more expensive HSE functional.

EGMF efficiently combines matrix features through Clebsch–Gordan (CG) decomposition and global gates while preserving the network's $E(3)$ equivariance. CG decomposition first converts blocks of an orbital matrix into irreducible tensors that transform according to angular momentum $L$. For orbital shells with angular momenta $l_a$ and $l_b$, the matrix block $M^s_{ij,ab}$, where $s$ denotes the matrix type, is decomposed into irreducible representations $z^s_{ij,ab;L\mu}$:

$$z^s_{ij,ab;L\mu} = \sum_{m=-l_a}^{l_a} \sum_{n=-l_b}^{l_b} \mathcal{C}^{L,l_a,l_b}_{\mu mn} M^s_{ij,am,bn}, \qquad |l_a - l_b| \le L \le l_a + l_b, \tag{1}$$

where $\mathcal{C}$ is a fixed coupling tensor in the real spherical harmonic basis. After collecting all components into node and edge features $\mathbf{z}^s_i$ and $\mathbf{z}^s_{ij}$, we project them into the original graph representation spaces using lightweight equivariant linear maps $P^n_s$ and $P^e_s$, without costly tensor-product operations. Independent learnable global scalar gates, $\alpha^n_s$ for nodes and $\alpha^e_s$ for edges, then add the matrix information to the initial node features $\mathbf{x}_i$ and edge features $\mathbf{e}_{ij}$ as residuals:

$$\widetilde{\mathbf{x}}_i = \mathbf{x}_i + \sum_{s\in\mathcal{S}} \alpha^n_s P^n_s \mathbf{z}^s_i, \quad \widetilde{\mathbf{e}}_{ij} = \mathbf{e}_{ij} + \sum_{s\in\mathcal{S}} \alpha^e_s P^e_s \mathbf{z}^s_{ij}. \tag{2}$$

The combined features $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{e}}_{ij}$ replace the less informative initial equivariant features as the starting point for message passing, without changing the dimensions of the backbone network.

EGMF offers three advantages over existing approaches to combining matrix features. The input matrices contain rich physical priors: $S$ provides basis-set information through two-orbital integrals, while $H^0$ adds kinetic-energy and pseudopotential information, giving the network physically meaningful initial inputs. The module is also easy to integrate and has very low computational overhead. Unlike methods that repeatedly invoke costly tensor products during message passing, EGMF introduces matrix information at initialization through a single equivariant linear layer. Its most important advantage is its inherent ability to accommodate basis-set mismatch. Projection through irreducible representations removes the requirement that the feature matrices and prediction targets use exactly the same basis set. When their maximum angular momentum $l$ is the same but their numbers of irreducible representations differ, the equivariant linear layer automatically aligns the channels. When the maximum $l$ differs, the method remains applicable by simply truncating the excess higher-order representations. This flexibility enables the transfer of priors across basis sets and explains how, in the experiments below, predictions from a model using an OpenMX basis can be used directly as priors for a model using an ABACUS basis.

### 2.2 A DUAL-BRANCH GRAPH NEURAL NETWORK: FUNCTIONAL PRIORS

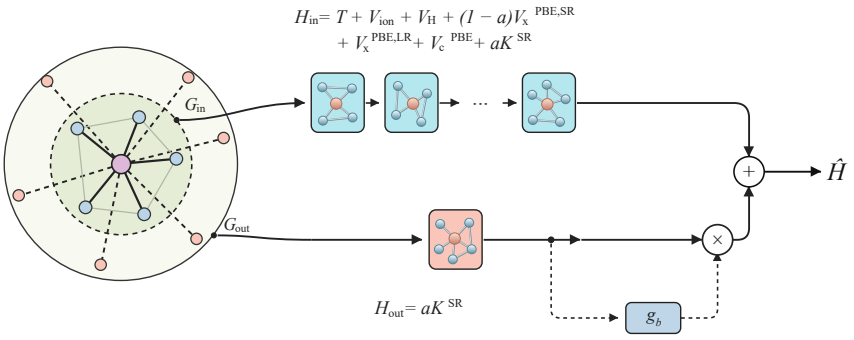


Figure 2: H2Net architecture. The inner branch uses a deeper network to learn the complex mixture of interactions, while the outer branch uses a shallower network to learn only Hartree–Fock (HF) interactions.

In addition to the basis-set and pseudopotential information supplied by the matrix inputs, we use the interaction structure of the HSE functional to guide the network design. The HSE exchange-correlation energy combines short-range Hartree–Fock (HF) exchange with short- and long-range PBE terms:

$$E_{\mathrm{xc}}^{\mathrm{HSE}} = aE_{\mathrm{x}}^{\mathrm{HF,SR}}(\omega) + (1-a)E_{\mathrm{x}}^{\mathrm{PBE,SR}}(\omega) + E_{\mathrm{x}}^{\mathrm{PBE,LR}}(\omega) + E_{\mathrm{c}}^{\mathrm{PBE}}, \tag{3}$$

where $a$ is the fraction of exact exchange, $\omega$ controls the screening length, and SR and LR denote short range and long range, respectively. The error function separates the Coulomb kernel into these two parts as $1/r = \mathrm{erfc}(\omega r)/r + \mathrm{erf}(\omega r)/r$, where $r$ is the electron separation; the first term on the right is the short-range kernel, and the second is the long-range kernel. HSE06 uses $a = 0.25$ and $\omega = 0.2\,\text{Å}^{-1}$ (approximately $0.11\,\mathrm{bohr}^{-1}$), mixing $25\%$ HF exchange with $75\%$ PBE exchange at short range and using PBE for long-range exchange and correlation.

The total Kohn–Sham Hamiltonian contains kinetic-energy, pseudopotential, and exchange-correlation terms, whose spatial ranges differ in a localized atomic-orbital basis. Apart from HF exchange and the nonlocal pseudopotential, these terms can be written as

$$O_{\mu\nu} = \int \phi_\mu^*(\mathbf{r})\,\hat{O}\,\phi_\nu(\mathbf{r})\,d\mathbf{r}, \tag{4}$$

where $\hat{O}$ denotes the operator for the corresponding term, and $\phi_\mu$ and $\phi_\nu$ are localized orbitals on atoms $i$ and $j$, respectively. Regardless of the specific operator, this integral can be nonzero only if orbitals $\mu$ and $\nu$ overlap in real space. Orbital basis sets for periodic solids usually impose a hard real-space cutoff, beyond which the orbital wavefunctions vanish. These terms therefore contribute no hopping between atoms separated by more than the sum of their orbital cutoffs. Although the

nonlocal pseudopotential takes a different form, its range extends only slightly beyond the basis-set cutoff and does not change this central conclusion.

By contrast, the short-range part of the four-center integral for HF exact exchange is

$$K_{\mu\nu}^{\mathrm{SR}} = -\sum_{\lambda\sigma} P_{\lambda\sigma} \iint \phi_\mu^*(\mathbf{r})\phi_\lambda(\mathbf{r}) \frac{\mathrm{erfc}(\omega|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|} \phi_\sigma^*(\mathbf{r}')\phi_\nu(\mathbf{r}')\, d\mathbf{r}\, d\mathbf{r}', \tag{5}$$

where $P_{\lambda\sigma}$ is the density matrix, and sums over periodic images are implicit in the orbital indices. In this integral, orbitals $\mu$ and $\lambda$ overlap at $\mathbf{r}$, while $\sigma$ and $\nu$ overlap at $\mathbf{r}'$; $\mu$ and $\nu$ need not overlap directly. The two orbital products are connected by the exchange kernel screened through the error function, so two charge-density distributions separated by a distance within the screening range can produce nonzero hopping. The inner region defined by the basis-set cutoff thus contains a complex mixture of interactions, whereas the extended coupling beyond it comes only from HF exchange. This physical distinction motivates separate network branches with different spatial receptive fields.

To translate these physical ranges into a network architecture, H2Net constructs two graphs with shared atomic nodes but different edge sets. For an edge $e = (i, j, \mathbf{T})$ in a periodic graph, let $d_e = \|\mathbf{r}_j + \mathbf{T} - \mathbf{r}_i\|$. The edge sets are

$$\begin{aligned} \mathcal{E}_{\mathrm{in}} &= \{e : d_e \le 1.2[r_c(Z_i) + r_c(Z_j)]\}, \\ \mathcal{E}_{\mathrm{out}} &= \{e : d_e \le 2[r_c(Z_i) + r_c(Z_j)]\}. \end{aligned} \tag{6}$$

Here, $r_c(Z)$ is the orbital cutoff radius for the basis set of element $Z$. The inner and outer distance thresholds are 1.2 and 2 times the sum of the orbital cutoffs at the two endpoints, respectively. The factor of 2 follows DeepH-hybrid (Tang et al., 2024); the cutoff selection procedure is described in Appendix G. Since $\mathcal{E}_{\mathrm{in}} \subseteq \mathcal{E}_{\mathrm{out}}$, the outer branch contributes to both near-neighbor corrections and coupling in the outer region. H2Net also incorporates the equivariant gated matrix features described above. Because the inner graph includes some PBE terms, it uses PBE predictions from Uni-HamGNN as auxiliary features, while the outer graph uses $H^0$. The two graphs then perform message passing independently from these input features.

The two $E(3)$-equivariant networks have independent parameters and extract features on their respective graphs. Equivariant readout then produces matrix blocks in the same target orbital basis. Using $b$ to denote either an onsite or an offsite block, the final Hamiltonian prediction is

$$\widehat{H}_b = F_{\theta_{\mathrm{in}}}(\mathcal{G}_{\mathrm{in}})_b + g_b F_{\theta_{\mathrm{out}}}(\mathcal{G}_{\mathrm{out}})_b. \tag{7}$$

The inner-branch output is set to zero outside the inner graph. Onsite and offsite predictions use separate scalar gates $g_b$, generated by an MLP from the invariant scalar features of the outer branch. These scalar gates preserve equivariance, and forward and reverse edges share gate weights to maintain Hermiticity. The gates adjust the correction magnitude to the chemical environment, while both branches are trained jointly end to end against the total HSE Hamiltonian.

This asymmetric design, with a deep inner branch and a shallow outer branch, directly reflects the physics of the functional and substantially reduces memory costs for large-scale calculations. The inner region contains a complex mixture of kinetic energy, pseudopotentials, PBE exchange-correlation, and HF exchange. Its complex local many-body environment requires a deeper network for adequate representation. The outer region contains only HF exchange, with a single physical contribution and many-body effects that weaken with distance, so one or a few layers suffice. In three dimensions, the number of neighbors grows cubically with the cutoff radius ($N_{\mathrm{nbr}} \propto R^3$), making a direct increase in the network cutoff costly in memory for intermediate features. H2Net retains deep message passing on the small inner graph and uses only a shallow network on the large outer graph, avoiding this memory bottleneck while improving prediction accuracy.

In this dual-graph architecture, H2Net uses HamGNN as the backbone for each branch and the EGMF method described in the previous subsection to provide features for both branches. Since the inner branch includes contributions from both the non-self-consistent part and PBE interactions, it uses the PBE Hamiltonian predicted by Uni-HamGNN as input features, while the outer branch uses $H^0$.

### 2.3 Variational surrogate loss: a variational-principle prior

Having incorporated basis-set and functional information into the input features and network architecture, we also introduce a physical prior into training. The similarity between neural network

training and the variational search for the minimum ground-state energy suggests optimizing a differentiable energy function $E(\widehat{H})$ of the predicted Hamiltonian. Constructing an explicitly differentiable energy calculation, however, is complex. DeepH-Zero implements energy optimization through its differentiable DFT framework, AI2DFT, improving accuracy at a high cost. We instead bypass the explicit construction of $E(\widehat{H})$ and build a surrogate loss whose gradient with respect to the predicted Hamiltonian equals the energy gradient $\nabla_{\widehat{H}} E$, thereby incorporating the ground-state energy into training indirectly.

Throughout this section, $\widehat{H}$ denotes the predicted Hamiltonian and $\widehat{P}$ the density matrix derived from it; $H$ and $P$ without hats denote the reference Hamiltonian and density matrix from self-consistent DFT calculations. The predicted Hamiltonian and its density matrix satisfy the generalized eigenvalue problem and normalization conditions

$$\widehat{H}C = SC\varepsilon, \qquad C^{\dagger}SC = I, \qquad \widehat{P} = C_{\mathrm{occ}}C_{\mathrm{occ}}^{\dagger}, \tag{8}$$

where $S$ is the overlap matrix, $C$ contains the eigenstate coefficients of the predicted Hamiltonian, $\varepsilon$ contains the corresponding eigenvalues, and $C_{\mathrm{occ}}$ contains the occupied-state coefficients. Since the density matrix is fully determined by the predicted Hamiltonian, the energy gradient follows from the chain rule:

$$\nabla_{\widehat{H}} E = \nabla_{\widehat{P}} E : \frac{\partial \widehat{P}}{\partial \widehat{H}}, \tag{9}$$

where the colon denotes contraction over the two density-matrix indices. In density functional theory, the first variation of the energy gives $\nabla_{\widehat{P}} E = H^{\mathrm{KS}}(\widehat{P})$ (see Appendix B.4 for a proof). Here, $H^{\mathrm{KS}}(\widehat{P})$ is the Hamiltonian obtained by evaluating the functional at the density matrix $\widehat{P}$; it is not equal to $\widehat{H}$.

Substituting this variational relation into the chain rule gives a surrogate objective in trace form. The matrix $H^{\mathrm{KS}}(\widehat{P})$ is computed by DFT from the current density matrix and is treated as a constant during differentiation, with no gradient propagated through it. The gradient relation is therefore

$$\nabla_{\widehat{H}} E = H^{\mathrm{KS}}(\widehat{P}) : \frac{\partial \widehat{P}}{\partial \widehat{H}} = \nabla_{\widehat{H}} \operatorname{Tr}\Big(\widehat{P} H^{\mathrm{KS}}(\widehat{P})\Big). \tag{10}$$

Thus, $\operatorname{Tr}\Big(\widehat{P} H^{\mathrm{KS}}(\widehat{P})\Big)$ is the desired surrogate loss. During backpropagation, gradients reach the predicted Hamiltonian $\widehat{H}$ only through the predicted density matrix $\widehat{P}$. This objective still requires assembling the effective Hamiltonian from the current predicted density at every iteration, however. Performing a non-self-consistent DFT single-point calculation for every sample in every training epoch remains prohibitively time-consuming. We therefore use a two-stage training strategy. The model is first pretrained with a matrix-element loss so that the predicted Hamiltonian approaches the reference. After pretraining, we assume that the prediction is sufficiently close to convergence for the approximation $H^{\mathrm{KS}}(\widehat{P}) \approx H^{\mathrm{KS}}(P) = H$ to hold. This substitution gives the final variational surrogate loss:

$$\mathcal{L}_{\mathrm{VS}} = \operatorname{Tr}\Big(\widehat{P} H\Big). \tag{11}$$

This expression requires only one generalized eigendecomposition to obtain gradient information about the ground-state energy, without changing the training framework or performing DFT single-point calculations. The value of VSLoss itself is neither the ground-state energy nor a conventional difference between a prediction and a reference; it is added to the other losses for optimization. The derivation applies in both real and reciprocal space. In our implementation, the predicted real-space Hamiltonian is first Fourier-transformed into reciprocal space, then diagonalized to obtain the density matrix and evaluate the loss.

The Hohenberg–Kohn theorem guarantees only that the minimum ground-state energy corresponds to the unique correct charge density, so VSLoss is particularly effective at optimizing wavefunctions. In DFT, the functional establishes the correspondence between charge density and the Hamiltonian, whereas deep learning predictions have no prescribed functional form. An additional band-structure loss is therefore needed during training to ensure that other properties calculated from the Hamiltonian, such as energy bands, are correct, as demonstrated by the experiments below.

# 3 DATASETS

Most publicly available Hamiltonian datasets focus on molecules, with relatively few covering periodic solids. Public datasets of HSE Hamiltonians are particularly scarce. Although NextHam and a study at KDD 2026 released universal solid-state Hamiltonian datasets spanning the periodic table, there is still no widely accepted benchmark for specialized solid-state Hamiltonian prediction, and many studies continue to use datasets from earlier work.

Dataset release formats and splitting procedures also lack a common standard. Most studies release data processed by their own software rather than raw DFT outputs, making comparisons between models difficult. When generating perturbed structures, many also draw the training, validation, and test sets from a single molecular dynamics (MD) trajectory. Temporal correlations between structures in different splits then create a substantial risk of data leakage.

Graph cutoff radii are likewise not standardized, partly because DFT packages truncate Hamiltonian matrix elements differently. OpenMX, for example, sets Hamiltonian matrix elements beyond the basis-set cutoff to zero, allowing models to construct graphs using that cutoff. ABACUS does not apply the same truncation, so nonlocal pseudopotentials can produce matrix elements beyond the basis-set cutoff. As a result, studies that simply use the basis-set cutoff omit these contributions, while studies that connect nodes according to nonzero Hamiltonian elements introduce information from the prediction target into the graph structure, risking data leakage. The cutoff also affects error comparisons: a larger radius includes more small hopping terms between distant atoms, lowering the mean absolute error of the Hamiltonian matrix elements. Hamiltonian errors obtained with different cutoffs therefore cannot be used directly to compare model accuracy.

To address these issues, we built datasets for bilayer $\mathrm{MoS_2}$ and $\mathrm{MAPbI_3}$ from scratch. Structures for both materials were generated by MD simulations, with independent trajectories for the training, validation, and test sets to avoid leakage from splitting a single trajectory. We then calculated both PBE and HSE Hamiltonians for these structures using ABACUS. The graph cutoff was fixed at $1.2$ times the basis-set cutoff for PBE and twice the basis-set cutoff for HSE. We removed the f orbitals from the basis sets for reasons discussed below. We supplemented these datasets with HamGNN's public SACADA carbon-allotrope dataset and $\mathrm{SiO_2}$ isomer dataset. Both contain structurally diverse systems that provide a stronger test of generalization to complex structures, and both were calculated with the PBE functional in OpenMX. We also recalculated a subset of the $\mathrm{SiO_2}$ structures in ABACUS to provide an additional HSE dataset.

# 4 RESULTS

## 4.1 EVALUATION OF H2NET

We evaluate H2Net through a direct comparison with the HamGNN baseline. A comparison with DeepH-hybrid is provided in Appendix E.4. HamGNN has already achieved good results in HSE Hamiltonian prediction (Zhang et al., 2025). For basis sets containing f orbitals, however, its equivariant features must include irreducible representations up to at least $l = 6$. The resulting increase in feature dimension further raises memory requirements on large graphs. Table 1 reports the results for systems with f orbitals.

Under the same memory budget on an NVIDIA A800-SXM4-80GB GPU, HamGNN can train only a single layer on either system. H2Net supports three inner-branch layers and one outer-branch layer while achieving lower prediction errors. The Hamiltonian MAE decreases from $0.188$ to $0.046\,\mathrm{meV}$ for $\mathrm{MoS_2}$ and from $0.132$ to $0.038\,\mathrm{meV}$ for $\mathrm{MAPbI_3}$, reductions of approximately $75.5\%$ and $71.2\%$, respectively. H2Net uses $76.70$ and $78.98\,\mathrm{GB}$ of training memory on these systems, compared with $66.64$ and $46.23\,\mathrm{GB}$ for single-layer HamGNN. Both H2Net configurations remain within the same memory budget.

For a fair comparison of accuracy, we removed f orbitals from the basis sets of these two systems so that HamGNN could train with three layers and be compared with H2Net in the 3+1-layer configuration. These f orbitals do not participate in bonding, so removing them reduces the training burden without substantially changing the physical picture. The $\mathrm{SiO_2}$ basis set already contains no f orbitals and is unaffected. Table 2 reports the results. H2Net reduces both prediction error

Table 1: HSE Hamiltonian prediction with f orbitals retained. On an NVIDIA A800-SXM4-80GB GPU, H2Net supports training a 3+1-layer network and achieves lower MAE than a single-layer HamGNN. MAE is reported in units of $0.1\,\mathrm{meV}$, and Mem denotes training GPU memory usage in GB. The notation 3+1 denotes three inner-branch layers and one outer-branch layer; OOM denotes an out-of-memory error, and — indicates an unavailable result.

| Model | Layers | $MoS_2$ | | $MAPbI_3$ | |
|---|---|---|---|---|---|
| | | MAE ↓ | Mem ↓ | MAE ↓ | Mem ↓ |
| HamGNN | 1 | 1.88 | 66.64 | 1.32 | 46.23 |
| | 2 | — | OOM | — | OOM |
| | 3 | — | OOM | — | OOM |
| H2Net | 3+1 | **0.46** | 76.70 | **0.38** | 78.98 |

and training memory usage on all three systems. The Hamiltonian MAE decreases from $0.082$ to $0.064\,\mathrm{meV}$ for $MoS_2$, from $0.077$ to $0.050\,\mathrm{meV}$ for $MAPbI_3$, and from $0.509$ to $0.228\,\mathrm{meV}$ for $SiO_2$, with training memory usage decreasing from $46.17$ to $34.96\,\mathrm{GB}$, from $48.49$ to $28.83\,\mathrm{GB}$, and from $78.81$ to $47.25\,\mathrm{GB}$, respectively. These correspond to reductions of approximately $24.3\%$, $40.5\%$, and $40.0\%$. H2Net thus retains its advantages in both accuracy and memory efficiency even when HamGNN can also use three layers.

Table 2: HSE Hamiltonian prediction without f orbitals. HamGNN uses three layers, while H2Net uses the 3+1-layer configuration. MAE is reported in units of $0.1\,\mathrm{meV}$, and Mem denotes training GPU memory usage in GB; — indicates an unavailable result.

| Model | $MoS_2$ | | $MAPbI_3$ | | $SiO_2$ | |
|---|---|---|---|---|---|---|
| | MAE ↓ | Mem ↓ | MAE ↓ | Mem ↓ | MAE ↓ | Mem ↓ |
| HamGNN | 0.82 | 46.17 | 0.77 | 48.49 | 5.09 | 78.81 |
| H2Net | **0.64** | **34.96** | **0.50** | **28.83** | **2.28** | **47.25** |

### 4.2 Evaluation of EGMF

We examine EGMF separately by integrating it into other Hamiltonian prediction models. We supply the non-self-consistent Hamiltonian $H^0$ to DeepH-E3 and HamGNN through EGMF, with results in Table 3. Both backbones improve in accuracy across the different types of datasets when matrix features are added, with particularly pronounced gains on the carbon-allotrope and silica-isomer datasets. These results show that EGMF serves as a plug-and-play input module that supplies useful physical priors to different backbones. By contrast, residual learning improves HamGNN on the carbon-allotrope and silica-isomer datasets but worsens its performance on the other two systems. Residual learning changes only the prediction target by subtracting $H^0$, whereas EGMF feeds its basis-set and pseudopotential information directly into the network for use in message passing and nonlinear feature combinations. EGMF therefore offers both flexibility in integration and fuller use of matrix priors.

Table 3: EGMF performance across backbones and comparison with residual learning. Values are Hamiltonian MAEs in units of $0.1\,\mathrm{meV}$.

| Model | $MoS_2$ | $MAPbI_3$ | C | $SiO_2$ |
|---|---|---|---|---|
| DeepH-E3 | 9.13 | 5.72 | 9.92 | 19.64 |
| DeepH-E3 + EGMF ($H^0$) | **6.97** | **5.23** | **7.21** | **7.79** |
| HamGNN | 2.94 | 3.51 | 12.43 | 35.10 |
| HamGNN + $H^0$ residual learning | 5.25 | 4.19 | 6.18 | 15.24 |
| HamGNN + EGMF ($H^0$) | **2.22** | **2.75** | **5.06** | **9.69** |

### 4.3 EVALUATION OF VSLOSS

Table 4 compares the original HamGNN with variants trained using BandLoss, VSLoss, or both. We report Hamiltonian MAE, band MAE, and wavefunction cosine dissimilarity to show how these quantities respond to different training objectives. Definitions of the three evaluation metrics are provided in Appendix F.

Table 4: Hamiltonian, band, and wavefunction predictions under different training losses. $\Delta H$ and $\Delta\varepsilon$ denote Hamiltonian MAE and band MAE, in units of $0.1\,\mathrm{meV}$ and $\mathrm{meV}$, respectively. $\Delta\psi$ is wavefunction cosine dissimilarity, expressed as a percentage.

| Model | $MoS_2$ | | | $MAPbI_3$ | | | C | | | $SiO_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\Delta H\downarrow$ | $\Delta\varepsilon\downarrow$ | $\Delta\psi\downarrow$ | $\Delta H\downarrow$ | $\Delta\varepsilon\downarrow$ | $\Delta\psi\downarrow$ | $\Delta H\downarrow$ | $\Delta\varepsilon\downarrow$ | $\Delta\psi\downarrow$ | $\Delta H\downarrow$ | $\Delta\varepsilon\downarrow$ | $\Delta\psi\downarrow$ |
| HamGNN | 2.94 | 14.59 | 6.94% | **3.51** | 13.53 | 1.44% | 12.43 | 1786.15 | 40.73% | 35.10 | 1807.89 | 38.34% |
| + BandLoss | **2.93** | 10.72 | 5.24% | 3.54 | **9.61** | 0.92% | 46.27 | 988.16 | 49.59% | **24.22** | 26.02 | 2.13% |
| + VSLoss | 6.84 | 35.95 | 1.70% | 4.28 | 46.56 | 0.64% | **10.93** | 27.89 | 11.67% | 30.54 | 113.60 | 4.22% |
| + BandLoss & VSLoss | 6.81 | **2.45** | **1.37%** | 4.31 | 12.32 | **0.52%** | 15.58 | **15.23** | **10.66%** | 30.80 | **20.45** | **1.91%** |

VSLoss markedly improves wavefunction predictions on all four systems. Hamiltonian matrix-element errors do not follow the same pattern: for $MoS_2$, for example, the Hamiltonian MAE increases from $0.294\,\mathrm{meV}$ to $0.684\,\mathrm{meV}$ even as the wavefunctions become more accurate. This supports the earlier analysis that numerical accuracy in the Hamiltonian does not necessarily imply physical accuracy.

With VSLoss alone, however, band accuracy is often unsatisfactory. As discussed above, the HK theorem constrains only the charge density; perturbation theory provides a complementary explanation (see the appendix). Joint fine-tuning with BandLoss and VSLoss gives the best results for most downstream metrics. Although fine-tuning sacrifices Hamiltonian accuracy, it substantially improves predictions of downstream physical properties. EGMF and H2Net use physical priors to improve the model's ability to learn, while VSLoss uses them to provide the correct optimization objective.

## 5 CONCLUSION

We introduce H2Net and the variational surrogate loss VSLoss to make fuller use of physical priors from DFT in Hamiltonian learning for periodic solids. The experiments demonstrate the effectiveness of this complementary approach. H2Net improves prediction accuracy on all three HSE test systems and reduces training memory usage by approximately $24.3\%$ to $40.5\%$ in comparisons without f orbitals. EGMF results on DeepH-E3 and HamGNN further show that matrix priors benefit different backbones. VSLoss improves wavefunction predictions on all four test systems, and combining it with the band loss further balances band and wavefunction accuracy. These results show that physical priors help models learn Hamiltonians more effectively and connect training objectives to the physical quantities of interest in electronic structure calculations. Translating existing DFT information into network features, architectures, and optimization objectives therefore offers a practical way to balance prediction accuracy, computational efficiency, and downstream physical accuracy within existing equivariant learning frameworks. Several questions remain for future work, including whether the H2Net approach extends to hybrid functionals for molecular systems and how VSLoss should handle metallic occupations or Kramers degeneracy when computing density matrices.

### AI USE STATEMENT

We used generative AI tools to help draft and edit the paper, find related work, generate synthetic datasets, and develop proofs of mathematical claims. These uses are described in the paper and disclosed as required. We did not use generative AI for any other tasks requiring disclosure; the remaining disclosure categories do not apply to this work.

We reviewed all AI-assisted work. We checked and revised the writing, manually reviewed the literature to verify the relevance and accuracy of references and research ideas, and validated the synthetic datasets and their use in the study. We also independently checked the mathematical arguments and proofs for correctness. We take full responsibility for the final paper and all accompanying materials, including AI-assisted text, claims, mathematical arguments, datasets, and other artifacts.

## REFERENCES

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13 (1):2453, 2022. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

D R Bowler and T Miyazaki. O(N) methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, 2012. doi: 10.1088/0034-4885/75/3/036503. URL https://doi.org/10.1088/0034-4885/75/3/036503.

Xiaoxun Gong, He Li, Nianlong Zou, Runzhang Xu, Wenhui Duan, and Yong Xu. General framework for E(3)-equivariant neural network representation of density functional theory Hamiltonian. *Nature Communications*, 14(1):2848, 2023. doi: 10.1038/s41467-023-38468-8. URL https://doi.org/10.1038/s41467-023-38468-8.

Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, 2003. doi: 10.1063/1.1564060. URL https://doi.org/10.1063/1.1564060.

P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, 1964. doi: 10.1103/physrev.136.b864. URL https://doi.org/10.1103/physrev.136.b864.

W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/physrev.140.a1133. URL https://doi.org/10.1103/physrev.140.a1133.

Aliaksandr V. Krukau, Oleg A. Vydrov, Artur F. Izmaylov, and Gustavo E. Scuseria. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *The Journal of Chemical Physics*, 125(22):224106, 2006. doi: 10.1063/1.2404663. URL https://doi.org/10.1063/1.2404663.

He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, 2022. doi: 10.1038/s43588-022-00265-6. URL https://doi.org/10.1038/s43588-022-00265-6.

Yang Li, Zechen Tang, Zezhou Chen, Minghui Sun, Boheng Zhao, He Li, Honggeng Tao, Zilong Yuan, Wenhui Duan, and Yong Xu. Neural-network Density Functional Theory Based on Variational Energy Minimization. *Physical Review Letters*, 133(7):076401, 2024. doi: 10.1103/physrevlett.133.076401. URL https://doi.org/10.1103/physrevlett.133.076401.

Yunyang Li, Zaishuo Xia, Lin Huang, Xinran Wei, Samuel Harshe, Han Yang, Erpai Luo, Zun Wang, Jia Zhang, Chang Liu, Bin Shao, and Mark Gerstein. Enhancing the scalability and applicability of kohn-sham hamiltonians for molecular systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=twEvvkQqPS.

Jigyasa Nigam, Michael J. Willatt, and Michele Ceriotti. Equivariant representations for molecular Hamiltonians and N-center atomic-scale properties. *The Journal of Chemical Physics*, 156(1): 014115, 2022. doi: 10.1063/5.0072784. URL https://doi.org/10.1063/5.0072784.

Zhuoran Qiao, Anders S. Christensen, Matthew Welborn, Frederick R. Manby, Anima Anandkumar, and Thomas F. Miller. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31): e2205221119, 2022. doi: 10.1073/pnas.2205221119. URL https://doi.org/10.1073/pnas.2205221119.

K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):5024, 2019. doi: 10.1038/s41467-019-12875-2. URL https://doi.org/10.1038/s41467-019-12875-2.

Kirill Shmilovich, Devin Willmott, Ivan Batalov, Mordechai Kornbluth, Jonathan Mailoa, and J. Zico Kolter. Orbital Mixer: Using Atomic Orbital Features for Basis-Dependent Prediction of Molecular Wavefunctions. *Journal of Chemical Theory and Computation*, 18(10):6021–6030, 2022. doi: 10.1021/acs.jctc.2c00555. URL https://doi.org/10.1021/acs.jctc.2c00555.

Zechen Tang, He Li, Peize Lin, Xiaoxun Gong, Gan Jin, Lixin He, Hong Jiang, Xinguo Ren, Wenhui Duan, and Yong Xu. A deep equivariant neural network approach for efficient hybrid density functional calculations. *Nature Communications*, 15(1):8815, 2024. doi: 10.1038/s41467-024-53028-4. URL https://doi.org/10.1038/s41467-024-53028-4.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, 2018. URL https://arxiv.org/abs/1802.08219.

Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. Se(3)-equivariant prediction of molecular wavefunctions and electronic densities. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14434–14447. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/78f1893678afbeaa90b1fa01b9cfb860-Paper.pdf.

Yuxiang Wang, Yang Li, Zechen Tang, He Li, Zilong Yuan, Honggeng Tao, Nianlong Zou, Ting Bao, Xinghao Liang, Zezhou Chen, Shanghua Xu, Ce Bian, Zhiming Xu, Chong Wang, Chen Si, Wenhui Duan, and Yong Xu. Universal materials model of deep-learning density functional theory Hamiltonian. *Science Bulletin*, 69(16):2514–2521, 2024a. doi: 10.1016/j.scib.2024.06.011. URL https://doi.org/10.1016/j.scib.2024.06.011.

Zun Wang, Chang Liu, Nianlong Zou, He Zhang, Xinran Wei, Lin Huang, Lijun Wu, and Bin Shao. Infusing self-consistency into density functional theory hamiltonian prediction via deep equilibrium models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=PSVkinBs4u.

Shi Yin, Zujian Dai, Xinyang Pan, and Lixin He. Advancing universal deep learning for electronic-structure hamiltonian prediction of materials. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YvmR4vNai2.

Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum Hamiltonian. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40412–40424. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yu23i.html.

Changwei Zhang, Yang Zhong, Zhi-Guo Tao, Xinming Qin, Honghui Shang, Zhenggang Lan, Oleg V. Prezhdo, Xin-Gao Gong, Weibin Chu, and Hongjun Xiang. Advancing nonadiabatic molecular dynamics simulations in solids with E(3) equivariant deep neural hamiltonians. *Nature Communications*, 16(1):2033, 2025. doi: 10.1038/s41467-025-57328-1. URL https://doi.org/10.1038/s41467-025-57328-1.

He Zhang, Chang Liu, Zun Wang, Xinran Wei, Siyuan Liu, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Self-consistency training for density-functional-theory hamiltonian prediction. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Vw4Yar2fmW.

Yang Zhong, Hongyu Yu, Mao Su, Xingao Gong, and Hongjun Xiang. Transferable equivariant graph neural networks for the Hamiltonians of molecules and solids. *npj Computational Materials*, 9(1):182, 2023. doi: 10.1038/s41524-023-01130-4. URL https://doi.org/10.1038/s41524-023-01130-4.

Yang Zhong, Hongyu Yu, Jihui Yang, Xingyu Guo, Hongjun Xiang, and Xingao Gong. Universal Machine Learning Kohn–Sham Hamiltonian for Materials. *Chinese Physics Letters*, 41(7): 077103, 2024. doi: 10.1088/0256-307x/41/7/077103. URL https://doi.org/10.1088/0256-307x/41/7/077103.

# A Related work

Hamiltonian learning has progressed from early conventional machine learning models to equivariant graph neural networks and, more recently, universal models that work across materials. Early studies, including SchNOrb (Schütt et al., 2019) and DeepH (Li et al., 2022), showed that neural networks and machine learning could predict Hamiltonians in an orbital basis from atomic structures and use them to calculate electronic properties. By incorporating spatial symmetry, PhiSNet (Unke et al., 2021), DeepH-E3 (Gong et al., 2023), and HamGNN (Zhong et al., 2023) encoded transformation rules directly in explicit equivariant representations within the network, while subsequent work continued to improve computational efficiency (Yu et al., 2023). Building on these advances, Uni-HamGNN (Zhong et al., 2024), NextHAM (Yin et al., 2026), and Universal DeepH (Wang et al., 2024a) extended Hamiltonian prediction across elements and structures. Representative universal models for solid-state Hamiltonians currently focus on the PBE functional because hybrid functionals require more costly training data and involve more complex coupling ranges. Although universal prediction for hybrid functionals remains challenging, universal PBE models can provide inexpensive and effective priors for training hybrid-functional models.

The spatial locality of LCAO basis functions naturally matches the inductive bias of graph neural networks, so most existing methods predict Hamiltonians in an LCAO basis. The overlap matrix contains information about the shapes and spatial overlap of the basis functions, which motivates its use as an additional network input in methods such as Orbital Mixer (Shmilovich et al., 2022) and OrbNet-Equi (Qiao et al., 2022). DEQHNet (Wang et al., 2024b) introduced diagonal reduction, which transforms the overlap matrix using Gaussian auxiliary basis functions to obtain an equivariant representation. However, it uses only onsite matrix blocks and does not directly use interatomic overlap blocks. Since the overlap matrix contains only real-space integrals between orbitals, some studies have turned to readily computed components of the full Kohn–Sham Hamiltonian to obtain additional priors, including pseudopotential information. HamGNN (Zhong et al., 2023) uses the non-self-consistent part of the Hamiltonian as a baseline for residual learning, with the network predicting the remaining correction, demonstrating the usefulness of these physical components. Although residual learning is straightforward to implement, it does not allow the network to learn complex nonlinear correlations from these components. NextHAM (Yin et al., 2026) combines residual learning with equivariant input features, converting a zero-step Hamiltonian that includes the initial electron interaction potential into irreducible representations as initial features. Its feature coupling relies on direct sums and tensor products, however, which add computation and memory costs. Our equivariant gated matrix features use both onsite and offsite blocks: after equivariant projection, these blocks are added to the initial node and edge representations through gates. This requires no complex operations and adds only one equivariant linear layer. The method also preserves the network's core tensor-product mechanism, making it compatible with existing architectures and easy to integrate.

Hybrid functionals are essential for accurate calculations of electronic structure in solids, but their complex interactions also make prediction challenging. Building on DeepH-E3 (Gong et al., 2023), DeepH-hybrid (Tang et al., 2024) enlarged the graph cutoff radius to cover HF exchange interactions, pioneering equivariant Hamiltonian prediction for hybrid functionals. N2AMD (Zhang et al., 2025) followed a similar approach, using HamGNN to perform nonadiabatic complex molecular dynamics simulations with the HSE functional in various periodic solids. Increasing the cutoff radius while keeping the backbone unchanged is simple to implement. However, applying multiple message-passing layers directly to a large graph with many edges and neighbors substantially increases the size of intermediate representations and training memory requirements, limiting scalability to complex systems. Guided by the form of the HSE functional, H2Net uses branches of unequal depth: a shallow branch operates on the outer graph spanning the full target coupling range, while a deep branch performs message passing on a smaller inner graph that captures complex interactions. The two branches jointly predict the total HSE Hamiltonian, reducing graph computation and memory costs.

Numerical accuracy in Hamiltonian learning does not necessarily imply physical accuracy. WANet (Li et al., 2025) showed that even small errors in Hamiltonian matrix elements can lead to large errors in derived properties such as system energies and wavefunctions. To address this problem, HamGNN (Zhong et al., 2023) fine-tunes a model trained on Hamiltonians with an additional band-structure loss. WALoss (Li et al., 2025) uses wavefunction projections to enforce

alignment with the reference eigensystem and gives greater weight to important orbitals, while NextHAM (Yin et al., 2026) extends this approach to solids through a k-space loss. These supervised methods explicitly align the downstream physical properties of the predicted Hamiltonian with the reference values. Another line of work incorporates physics into the supervision process itself to avoid unphysical numerical fitting. DeepH-Zero (Li et al., 2024) integrates neural networks with differentiable DFT and trains without labels by optimizing the ground-state energy. DEQHNet (Wang et al., 2024b) incorporates the self-consistent iterations of DFT into a neural network to improve downstream predictions. Although these methods have a clear physical basis, their complex procedures entail substantial computational costs and compatibility requirements. We therefore propose VSLoss, which indirectly minimizes the ground-state energy by optimizing a surrogate loss. Compatible with standard supervised training, VSLoss uses the HK variational principle to ensure the accuracy of downstream physical properties, balancing accuracy with compatibility.

# B Density functional theory in an LCAO basis

The main text uses the overlap matrix and the non-self-consistent part of the Hamiltonian as input features, and constructs VSLoss from the gradient of the energy with respect to the density matrix. These matrices share a common localized atomic-orbital representation. This appendix introduces that representation and the origin of each matrix, then derives the gradient relation $\nabla_P E = H^{\mathrm{KS}}(P)$ from the components of the Hamiltonian. We use atomic units throughout and hold atomic positions, basis functions, and pseudopotentials fixed.

## B.1 From the one-electron equation to orbital matrices

The stationary Schrödinger equation relates the state of a system to its energy: applying the Hamiltonian operator to an eigenstate gives that state multiplied by its energy. For a many-electron system, solving this equation directly requires treating interactions between electrons. Kohn–Sham DFT recasts the problem as a set of effective one-electron equations, incorporating the interactions into an effective potential determined by the electron density (Hohenberg & Kohn, 1964; Kohn & Sham, 1965):

$$\hat{h}^{\mathrm{KS}}[P]\psi_n = \varepsilon_n \psi_n, \tag{12}$$

where $\psi_n$ is a one-electron orbital, $\varepsilon_n$ its eigenvalue, and $P$ the density matrix describing the occupied states. Because the effective operator depends on the electronic state being sought, the state and operator must be updated iteratively until self-consistency is reached.

The linear combination of atomic orbitals (LCAO) representation expands one-electron orbitals in basis functions $\{\phi_\mu\}$ localized around atoms, turning the wavefunction problem into one of determining expansion coefficients:

$$\psi_n(\mathbf{r}) = \sum_\mu C_{\mu n}\phi_\mu(\mathbf{r}). \tag{13}$$

Here, $\mu$ labels an atomic orbital, and $C_{\mu n}$ is its coefficient in eigenstate $n$. Substituting this expansion into the one-electron equation, multiplying by $\phi_\mu^*$ from the left, and integrating over space gives a matrix eigenvalue problem. The Hamiltonian and overlap matrices are

$$\begin{aligned} H^{\mathrm{KS}}_{\mu\nu}(P) &= \int \phi_\mu^*(\mathbf{r})\hat{h}^{\mathrm{KS}}[P]\phi_\nu(\mathbf{r})\, d\mathbf{r}, \\ S_{\mu\nu} &= \int \phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r})\, d\mathbf{r}. \end{aligned} \tag{14}$$

Atomic orbitals are generally nonorthogonal, and $S$ records their inner products. The eigenvalue equation and normalization condition therefore become

$$H^{\mathrm{KS}}(P)C = SC\varepsilon, \qquad C^\dagger SC = I. \tag{15}$$

Once the eigenstates are obtained, the occupied-state coefficients determine the density matrix:

$$P = C_{\mathrm{occ}}C_{\mathrm{occ}}^\dagger, \qquad P_{\mu\nu} = \sum_{n\in\mathrm{occ}} C_{\mu n}C_{\nu n}^*. \tag{16}$$

Substituting these coefficients back into the orbital expansion gives the real-space electron density,

$$n(\mathbf{r}) = \sum_{\mu\nu} P_{\mu\nu}\phi_\mu(\mathbf{r})\phi^*_\nu(\mathbf{r}). \tag{17}$$

Thus, $P$ collects occupied-state information in matrix form, while $n(\mathbf{r})$ describes the spatial distribution of electrons. Integrating the density gives the electron count $N_e = \mathrm{Tr}(PS)$, which also shows how the overlap matrix enters electron-number normalization.

### B.2 Real-space matrices and reciprocal-space representations

In a periodic solid, each atomic orbital has periodic images in different unit cells. For a lattice translation vector $\mathbf{T}$, the corresponding orbital is $\phi_{\mu\mathbf{T}}(\mathbf{r}) = \phi_{\mu\mathbf{0}}(\mathbf{r}-\mathbf{T})$. Matrix blocks between the reference cell and a translated cell are written as

$$H_{\mu\nu}(\mathbf{T}) = \langle\phi_{\mu\mathbf{0}}|\hat{h}|\phi_{\nu\mathbf{T}}\rangle, \qquad S_{\mu\nu}(\mathbf{T}) = \langle\phi_{\mu\mathbf{0}}|\phi_{\nu\mathbf{T}}\rangle. \tag{18}$$

When $\mu$ and $\nu$ belong to atoms $i$ and $j$, respectively, these matrix elements describe orbital coupling between atom $i$ and atom $j$ in the translated cell.

Lattice translation symmetry allows these real-space matrices to be transformed into reciprocal space through Bloch sums:

$$H(\mathbf{k}) = \sum_{\mathbf{T}} e^{i\mathbf{k}\cdot\mathbf{T}} H(\mathbf{T}), \qquad S(\mathbf{k}) = \sum_{\mathbf{T}} e^{i\mathbf{k}\cdot\mathbf{T}} S(\mathbf{T}). \tag{19}$$

Each wavevector $\mathbf{k}$ then defines a generalized eigenvalue problem,

$$H(\mathbf{k})C(\mathbf{k}) = S(\mathbf{k})C(\mathbf{k})\varepsilon(\mathbf{k}), \qquad C^\dagger(\mathbf{k})S(\mathbf{k})C(\mathbf{k}) = I. \tag{20}$$

The eigenvalues as a function of $\mathbf{k}$ form the energy bands, while the eigenstates determine the density matrix. This connects real-space orbital coupling to energy bands and electron density.

### B.3 Hamiltonian components and their integral forms

For semilocal functionals such as PBE, the Kohn–Sham Hamiltonian consists of a non-self-consistent part, the Hartree potential, and the exchange-correlation potential:

$$H^{\mathrm{KS}}(P) = H^0 + V_{\mathrm{H}}[n] + V_{\mathrm{xc}}[n], \qquad H^0 = T + V^{\mathrm{loc}}_{\mathrm{ion}} + V^{\mathrm{NL}}_{\mathrm{ion}}. \tag{21}$$

Here, $T$ is the kinetic-energy matrix, and $V^{\mathrm{loc}}_{\mathrm{ion}}$ and $V^{\mathrm{NL}}_{\mathrm{ion}}$ are the local and nonlocal parts of the ionic pseudopotential. These three terms can be calculated once the atomic positions, basis set, and pseudopotentials are specified, independently of the electron density to be determined, and are collectively denoted by $H^0$. A zero-step matrix obtained by evaluating the full Hamiltonian at an initial density guess also includes the Hartree and exchange-correlation contributions from that density.

The kinetic-energy and local-pseudopotential matrix elements are

$$\begin{aligned} T_{\mu\nu} &= -\frac{1}{2}\int \phi^*_\mu(\mathbf{r})\nabla^2\phi_\nu(\mathbf{r})\,d\mathbf{r}, \\ (V^{\mathrm{loc}}_{\mathrm{ion}})_{\mu\nu} &= \int \phi^*_\mu(\mathbf{r})v^{\mathrm{loc}}_{\mathrm{ion}}(\mathbf{r})\phi_\nu(\mathbf{r})\,d\mathbf{r}. \end{aligned} \tag{22}$$

The Hartree potential describes electrostatic interactions between electrons. Its matrix elements are likewise integrals of the two orbitals and the effective potential:

$$v_{\mathrm{H}}[n](\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}', \qquad (V_{\mathrm{H}})_{\mu\nu} = \int \phi^*_\mu(\mathbf{r})v_{\mathrm{H}}[n](\mathbf{r})\phi_\nu(\mathbf{r})\,d\mathbf{r}. \tag{23}$$

For periodic systems, this Coulomb potential is solved under periodic boundary conditions. The exchange-correlation potential accounts for exchange and correlation beyond the Hartree term. In PBE, the energy depends on the density and its gradient, while the corresponding potential still enters the matrix as a local multiplicative operator:

$$(V_{\mathrm{xc}})_{\mu\nu} = \int \phi^*_\mu(\mathbf{r})v_{\mathrm{xc}}[n](\mathbf{r})\phi_\nu(\mathbf{r})\,d\mathbf{r}. \tag{24}$$

For a given effective potential, all these terms take the form of two-orbital matrix elements. With basis functions that have hard real-space cutoffs, the two orbitals must overlap spatially for the corresponding matrix element to be nonzero.

The nonlocal pseudopotential couples orbitals through projector functions and has the separable form

$$\hat{v}_{\mathrm{ion}}^{\mathrm{NL}} = \sum_{Iab} |\beta_{Ia}\rangle D_{ab}^{I} \langle \beta_{Ib}|, \tag{25}$$

where $I$ labels an ion, $\beta_{Ia}$ is a projector on that ion, and $D_{ab}^{I}$ contains the pseudopotential coefficients. In the orbital basis, this gives

$$(V_{\mathrm{ion}}^{\mathrm{NL}})_{\mu\nu} = \sum_{Iab} \langle \phi_\mu | \beta_{Ia} \rangle D_{ab}^{I} \langle \beta_{Ib} | \phi_\nu \rangle. \tag{26}$$

The two orbitals couple separately to the projectors, so their matrix element can be nonzero even without direct overlap. Together with Equation (21), this shows that $S$ records overlap between basis functions, while $H^0$ additionally contains kinetic-energy and pseudopotential information.

HSE adds screened HF exchange to the semilocal terms (Heyd et al., 2003; Krukau et al., 2006), giving the Hamiltonian

$$\begin{aligned} H_{\mathrm{HSE}}^{\mathrm{KS}}(P) = H^0 + V_{\mathrm{H}}[n] + (1-a)V_{\mathrm{x}}^{\mathrm{PBE,SR}}[n] \\ + V_{\mathrm{x}}^{\mathrm{PBE,LR}}[n] + V_{\mathrm{c}}^{\mathrm{PBE}}[n] + aK^{\mathrm{SR}}[P]. \end{aligned} \tag{27}$$

Here, $a$ is the fraction of exact exchange, and SR and LR denote short range and long range. The short-range exchange matrix $K^{\mathrm{SR}}$ is given by a four-center integral, with periodic images included in the orbital indices:

$$\begin{aligned} K_{\mu\nu}^{\mathrm{SR}}[P] = -\sum_{\lambda\sigma} P_{\lambda\sigma} \iint \phi_\mu^*(\mathbf{r})\phi_\lambda(\mathbf{r}) \frac{\mathrm{erfc}(\omega|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|} \\ \times \phi_\sigma^*(\mathbf{r}')\phi_\nu(\mathbf{r}')\, d\mathbf{r}\, d\mathbf{r}'. \end{aligned} \tag{28}$$

The parameter $\omega$ controls the screening length. The two orbital products at $\mathbf{r}$ and $\mathbf{r}'$ are connected by the screened exchange kernel, so the outer orbitals $\mu$ and $\nu$ need not overlap directly. HF exchange can therefore produce coupling beyond their direct-overlap range. Equation (28) also shows that, for a fixed basis, $K^{\mathrm{SR}}$ is a linear function of the density matrix $P$, a relation used in the energy derivative below.

### B.4 Energy gradient with respect to the density matrix

Each Hamiltonian term corresponds to a contribution to the energy. Denoting the semilocal exchange-correlation part of HSE by

$$E_{\mathrm{xc}}^{\mathrm{sl}}[n] = (1-a)E_{\mathrm{x}}^{\mathrm{PBE,SR}}[n] + E_{\mathrm{x}}^{\mathrm{PBE,LR}}[n] + E_{\mathrm{c}}^{\mathrm{PBE}}[n], \tag{29}$$

the total energy can be written as

$$E[P] = \mathrm{Tr}(PH^0) + E_{\mathrm{H}}[n] + E_{\mathrm{xc}}^{\mathrm{sl}}[n] + aE_{\mathrm{x}}^{\mathrm{HF,SR}}[P] + E_{\mathrm{ion-ion}}, \tag{30}$$

where $E_{\mathrm{ion-ion}}$ is the ion–ion interaction energy. The PBE energy has the same form, with the HF exchange term removed and the semilocal part replaced by the full PBE exchange-correlation energy. Under the trace pairing, the matrix gradient is defined elementwise as $(\nabla_P E)_{\mu\nu} = \partial E / \partial P_{\nu\mu}$.

Since neither $H^0$ nor the ion–ion energy depends on $P$, their derivatives follow directly:

$$\frac{\partial\, \mathrm{Tr}(PH^0)}{\partial P_{\nu\mu}} = H_{\mu\nu}^0, \qquad \frac{\partial E_{\mathrm{ion-ion}}}{\partial P_{\nu\mu}} = 0. \tag{31}$$

The Hartree and semilocal exchange-correlation energies depend on $P$ through the electron density. From Equation (17),

$$\frac{\partial n(\mathbf{r})}{\partial P_{\nu\mu}} = \phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r}). \tag{32}$$

The Hartree energy is quadratic in the density,

$$E_{\mathrm{H}}[n] = \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'. \tag{33}$$

Differentiating the two density factors gives two equal terms, as seen by exchanging the integration coordinates $\mathbf{r}$ and $\mathbf{r}'$, which cancels the factor of $1/2$. Hence,

$$\begin{aligned}\frac{\partial E_{\mathrm{H}}}{\partial P_{\nu\mu}} &= \iint \frac{\phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}' \\ &= \int \phi_\mu^*(\mathbf{r})v_{\mathrm{H}}[n](\mathbf{r})\phi_\nu(\mathbf{r})\,d\mathbf{r} = (V_{\mathrm{H}})_{\mu\nu}.\end{aligned} \tag{34}$$

The semilocal exchange-correlation energy also depends on the spatial gradient of the density. Writing its energy density as $F$ gives $E_{\mathrm{xc}}^{\mathrm{sl}} = \int F(n, \nabla n)\,d\mathbf{r}$. Applying the chain rule to both arguments using Equation (32) yields

$$\begin{aligned}\frac{\partial E_{\mathrm{xc}}^{\mathrm{sl}}}{\partial P_{\nu\mu}} &= \int \left[\frac{\partial F}{\partial n}\phi_\mu^*\phi_\nu + \frac{\partial F}{\partial(\nabla n)}\cdot\nabla(\phi_\mu^*\phi_\nu)\right] d\mathbf{r} \\ &= \int \phi_\mu^*\left[\frac{\partial F}{\partial n} - \nabla\cdot\frac{\partial F}{\partial(\nabla n)}\right]\phi_\nu\,d\mathbf{r} = (V_{\mathrm{xc}}^{\mathrm{sl}})_{\mu\nu}.\end{aligned} \tag{35}$$

The second line follows by integration by parts of the gradient term; the boundary terms cancel under periodic boundary conditions. The expression in brackets is the semilocal exchange-correlation potential, so differentiation again gives the potential matrix listed above.

The HF exchange energy depends directly on the density matrix:

$$E_{\mathrm{x}}^{\mathrm{HF,SR}}[P] = \frac{1}{2}\,\mathrm{Tr}\big(PK^{\mathrm{SR}}[P]\big)\,. \tag{36}$$

Both factors depend on $P$, so the product rule gives

$$\begin{aligned}\frac{\partial E_{\mathrm{x}}^{\mathrm{HF,SR}}}{\partial P_{\nu\mu}} &= \frac{1}{2}K_{\mu\nu}^{\mathrm{SR}}[P] + \frac{1}{2}\sum_{\alpha\beta}P_{\alpha\beta}\frac{\partial K_{\beta\alpha}^{\mathrm{SR}}[P]}{\partial P_{\nu\mu}} \\ &= \frac{1}{2}K_{\mu\nu}^{\mathrm{SR}}[P] + \frac{1}{2}K_{\mu\nu}^{\mathrm{SR}}[P] = K_{\mu\nu}^{\mathrm{SR}}[P].\end{aligned} \tag{37}$$

Substituting Equation (28) into the second term and exchanging $\mathbf{r}$ and $\mathbf{r}'$ produces the same four-center integral as the first term. This holds because the screened exchange kernel depends only on the distance between the two points and is unchanged when their coordinates are exchanged.

Substituting these derivatives into Equation (30) recovers the components of the Kohn–Sham Hamiltonian:

$$\frac{\partial E}{\partial P_{\nu\mu}} = H_{\mu\nu}^0 + (V_{\mathrm{H}})_{\mu\nu} + (V_{\mathrm{xc}}^{\mathrm{sl}})_{\mu\nu} + aK_{\mu\nu}^{\mathrm{SR}}[P] = H_{\mu\nu}^{\mathrm{KS}}(P). \tag{38}$$

In matrix form,

$$\nabla_P E = H^{\mathrm{KS}}(P). \tag{39}$$

Here, $P$ is the density matrix formed from the orbital expansion coefficients, and the one-electron energy is already written as $\mathrm{Tr}(PH^0)$, so differentiation introduces no additional overlap matrix $S$. When differentiating the energy with respect to the predicted Hamiltonian through $P(H)$, the factor $\nabla_P E$ in the chain rule is therefore the Hamiltonian $H^{\mathrm{KS}}(P)$ assembled from the current density matrix.

## C Network backbone

Both H2Net branches use the HamGNN equivariant graph neural network as their backbone (Zhong et al., 2023), with independent parameters. We describe how the backbone encodes atomic and geometric information, learns environmental effects on orbital coupling through message passing, and reads out Hamiltonian matrix blocks.

### C.1 EQUIVARIANT REPRESENTATIONS OF ATOMS AND ATOM PAIRS

The network uses node features $\mathbf{x}_i$ and edge features $\mathbf{e}_{ij}$ to describe atomic environments and pairwise interactions, respectively, corresponding to onsite and offsite matrix blocks. To retain orbital directionality, both feature types consist of irreducible representations with different angular momenta and parities. Each channel with angular momentum $l$ contains $2l + 1$ components, and its parity $p = \pm 1$ specifies its behavior under spatial inversion.

Initial node features are obtained from atomic numbers $Z_i$ through one-hot encoding and a linear map, while edge features also incorporate interatomic distances and directions. For a periodic edge $(i, j, \mathbf{T})$, the relative displacement is $\mathbf{r}_{ij} = \mathbf{r}_j + \mathbf{T} - \mathbf{r}_i$, with length $d_{ij}$ and unit direction $\widehat{\mathbf{r}}_{ij}$. Distances are expanded in a radial basis, and directions in real spherical harmonics:

$$\boldsymbol{\rho}_{ij} = f_{\mathrm{cut}}(d_{ij})\,[b_1(d_{ij}), \ldots, b_{N_r}(d_{ij})], \qquad \mathbf{Y}_{ij} = \bigoplus_l Y^{(l)}(\widehat{\mathbf{r}}_{ij}), \tag{40}$$

Here, $b_n$ are radial basis functions, $N_r$ is their number, and $f_{\mathrm{cut}}$ is a smooth cutoff function. The element features at the two endpoints are separately mapped linearly and added, then combined with the spherical harmonics through a radially weighted equivariant tensor product to form the initial edge features. EGMF injects matrix information at this stage; the resulting features are denoted by $\mathbf{x}_i^{(0)}$ and $\mathbf{e}_{ij}^{(0)}$. Below, the superscript $t$ denotes the layer index, and the periodic translation $\mathbf{T}$ is omitted from edge subscripts.

### C.2 ALTERNATING NODE AND EDGE UPDATES

Coupling between an atom pair depends on the environments at both endpoints, so nodes and edges are updated alternately in each layer. A message along edge $i \to j$ uses both endpoint node features and the current edge features. These are separately coupled to the spherical harmonics, radially weighted, mixed across channels, and added:

$$\begin{aligned}\mathbf{m}_{ij}^{(t)} = \mathcal{A}_n^{(t)}(\boldsymbol{\rho}_{ij}) \left[(\mathbf{x}_i^{(t)} \oplus \mathbf{x}_j^{(t)}) \otimes_{\mathrm{CG}} \mathbf{Y}_{ij}\right] \\ + \mathcal{A}_e^{(t)}(\boldsymbol{\rho}_{ij}) \left[\mathbf{e}_{ij}^{(t)} \otimes_{\mathrm{CG}} \mathbf{Y}_{ij}\right].\end{aligned} \tag{41}$$

Here, $\oplus$ denotes concatenation by irreducible type, $\otimes_{\mathrm{CG}}$ is the CG tensor product, and $\mathcal{A}_n$ and $\mathcal{A}_e$ apply radial weighting and equivariant channel mixing, with distance-dependent weights generated by MLPs. The CG tensor product combines components according to angular-momentum and parity selection rules, preserving message equivariance.

At the receiving node $j$, messages are summed, processed by an equivariant residual block, and added to a linear skip connection from the original node features:

$$\mathbf{x}_j^{(t+1)} = W_{\mathrm{skip}}^{(t)} \mathbf{x}_j^{(t)} + \mathcal{R}_n^{(t)} \left( \sum_{(i,j,\mathbf{T}) \in \mathcal{E}} \mathbf{m}_{ij}^{(t)} \right), \tag{42}$$

Here, $\mathcal{E}$ is the edge set and $W_{\mathrm{skip}}$ is an equivariant linear map. The residual block $\mathcal{R}_n$ processes the aggregated information using equivariant linear layers and gated nonlinearities. Scalar channels gate the amplitudes of features at each order, while the skip connection retains the previous representation.

The updated nodes carry neighborhood information into a pair-interaction module that updates the edge representations:

$$\begin{aligned}\mathbf{e}_{ij}^{(t+1)} = W_e^{(t)} \mathbf{e}_{ij}^{(t)} \\ + \mathcal{P}^{(t)} \left(\mathbf{x}_i^{(t+1)}, \mathbf{x}_j^{(t+1)}, \mathbf{e}_{ij}^{(t)}, \mathbf{Y}_{ij}, \boldsymbol{\rho}_{ij}\right).\end{aligned} \tag{43}$$

Here, $W_e$ is an equivariant linear skip connection for edges. The module $\mathcal{P}$ includes linear maps of the endpoint nodes and node–direction and edge–direction couplings of the same form as Equation (41), with independent parameters. The updated edges contribute to node messages in the next layer, allowing multi-atom environmental information to propagate between nodes and edges across layers.

### C.3 Readout of Hamiltonian matrix blocks

After $T$ update layers, node and edge features pass through onsite and offsite readout heads, respectively, to produce irreducible coefficients for the target matrix blocks:

$$\mathbf{q}_i^{\mathrm{on}} = \mathcal{O}_n(\mathbf{x}_i^{(T)}), \qquad \mathbf{q}_{ij}^{\mathrm{off}} = \mathcal{O}_e(\mathbf{e}_{ij}^{(T)}). \tag{44}$$

Both readout heads, $\mathcal{O}_n$ and $\mathcal{O}_e$, consist of equivariant residual blocks and equivariant linear projections. For two orbital shells with angular momenta $l_a$ and $l_b$, the output coefficients span irreducible components with $|l_a - l_b| \le L \le l_a + l_b$ and parity $(-1)^{l_a+l_b}$. An inverse CG transform reconstructs the orbital-shell subblocks:

$$\widetilde{H}_{ij,am,bn} = \sum_{L=|l_a-l_b|}^{l_a+l_b} \sum_{\mu=-L}^{L} \mathcal{C}_{\mu mn}^{L,l_a,l_b}\, q_{ij,ab;L\mu}, \tag{45}$$

Here, $\mathcal{C}$ is the CG coupling tensor used in the main text, $a, b$ label orbital shells, $m, n$ are magnetic quantum numbers, $\mu$ indexes irreducible components, and $q_{ij,ab;L\mu}$ is the corresponding offsite readout coefficient. Onsite blocks are reconstructed in the same way from node readout coefficients. The subblocks are assembled according to the target basis, retaining the orbitals present for each element. Hermiticity is enforced by averaging an onsite block with its own conjugate transpose and an offsite block with the conjugate transpose of the corresponding reverse periodic-edge block:

$$\begin{aligned} H_i^{\mathrm{on}} &= \tfrac{1}{2}\left[\widetilde{H}_i^{\mathrm{on}} + (\widetilde{H}_i^{\mathrm{on}})^{\dagger}\right], \\ H_{ij}(\mathbf{T}) &= \tfrac{1}{2}\left[\widetilde{H}_{ij}(\mathbf{T}) + \widetilde{H}_{ji}(-\mathbf{T})^{\dagger}\right]. \end{aligned} \tag{46}$$

## D Further analysis of VSLoss

This appendix examines VSLoss in more detail. We show how the Ritz variational principle provides a lower bound on the loss and how a perturbation expansion around the reference relates it to coupling between occupied and unoccupied states. This analysis explains the origin of the energy-gap weighting and the role of joint band supervision.

### D.1 Variational lower bound and minima

VSLoss in the main text sums the energy expectations of the predicted occupied states under the reference Hamiltonian and typically takes a large negative value. To use it as a conventional non-negative loss in joint optimization, we subtract the energy expectation of the reference occupied states, giving zero loss at the reference solution. Following the main text, let $H, P$ denote the reference Hamiltonian and density matrix, and $\widehat{H}, \widehat{P}$ the predictions. The shifted loss is

$$\Delta\mathcal{L}_{\mathrm{VS}} = \mathcal{L}_{\mathrm{VS}} - \mathrm{Tr}(PH) = \mathrm{Tr}\big[(\widehat{P} - P)H\big]. \tag{47}$$

The reference baseline contributes no gradient, so it changes neither the gradient nor the optimum of the original objective. It shifts VSLoss toward zero to facilitate the operation of the optimizer and learning-rate scheduler. The Ritz variational principle shows that this shifted loss remains nonnegative during optimization.

Consider zero-temperature integer occupations, assigning unit occupation to each occupied state and omitting the $\mathbf{k}$ index. The prediction and reference share the same basis and positive-definite overlap matrix $S$, and each contains $N_{\mathrm{occ}}$ occupied states. Let $C, \varepsilon$ denote the reference eigenstates and eigenvalues, and $\widehat{C}, \widehat{\varepsilon}$ their predicted counterparts. The predicted occupied states satisfy

$$\widehat{P} = \widehat{C}_{\mathrm{occ}}\widehat{C}_{\mathrm{occ}}^{\dagger}, \qquad \widehat{C}_{\mathrm{occ}}^{\dagger} S \widehat{C}_{\mathrm{occ}} = I. \tag{48}$$

Substituting the density matrix into the loss and using the cyclic property of the trace gives

$$\mathrm{Tr}(\widehat{P}H) = \mathrm{Tr}(\widehat{C}_{\mathrm{occ}}^{\dagger} H \widehat{C}_{\mathrm{occ}}) = \sum_{i\in\mathrm{occ}} \widehat{C}_i^{\dagger} H \widehat{C}_i, \tag{49}$$

where $\widehat{C}_i$ is the $i$th predicted eigenstate. The loss is therefore the sum of energy expectations over these mutually $S$-orthogonal occupied orbitals.

The reference eigenstates satisfy

$$HC = SC\varepsilon, \qquad C^\dagger SC = I, \qquad P = C_{\mathrm{occ}} C_{\mathrm{occ}}^\dagger. \tag{50}$$

Order the reference eigenvalues from lowest to highest. The multi-state form of the Ritz principle states that, among all sets of $N_{\mathrm{occ}}$ mutually $S$-orthonormal orbitals, the lowest $N_{\mathrm{occ}}$ reference eigenstates minimize the sum of energy expectations. Thus,

$$\begin{aligned} \mathrm{Tr}(\widehat{P}H) &= \sum_{i \in \mathrm{occ}} \widehat{C}_i^\dagger H \widehat{C}_i \\ &\geq \sum_{i \in \mathrm{occ}} \varepsilon_i = \mathrm{Tr}(PH). \end{aligned} \tag{51}$$

Subtracting the reference baseline on the right gives $\Delta\mathcal{L}_{\mathrm{VS}} \geq 0$. This holds for every set of predicted occupied states satisfying the occupation and normalization conditions above.

If a gap separates the reference occupied and unoccupied states, equality in Equation (51) holds if and only if the predicted and reference occupied states span the same subspace:

$$\Delta\mathcal{L}_{\mathrm{VS}} = 0 \quad \Longleftrightarrow \quad \widehat{P} = P. \tag{52}$$

The global minimum of VSLoss over occupied subspaces is therefore attained exactly at the reference solution. Unitary rotations of occupied orbitals within this subspace leave the density matrix unchanged, so the equality condition concerns the occupied subspace as a whole. Having identified the optimum, we next examine how Hamiltonian errors shift occupied states away from the reference and how the loss responds.

### D.2 Perturbative response of the occupied subspace and local form of the loss

Hamiltonian errors mix other reference eigenstates into the predicted occupied orbitals. Let $\Delta H = \widehat{H} - H$; we consider the perturbative regime where $\Delta H$ is sufficiently small. Indices $i$ and $a$ denote occupied and unoccupied states, respectively, while $m, n$ run over all states. Since the reference eigenstates form a complete basis, a predicted occupied state can be expanded as

$$\widehat{C}_i = \sum_m C_m u_{mi}, \qquad u_{mi} = C_m^\dagger S \widehat{C}_i. \tag{53}$$

Using $C_m^\dagger S C_n = \delta_{mn}$, its normalization becomes

$$1 = \widehat{C}_i^\dagger S \widehat{C}_i = \sum_{mn} u_{mi}^* u_{ni} C_m^\dagger S C_n = \sum_m |u_{mi}|^2. \tag{54}$$

Likewise, $C_m^\dagger H C_n = \varepsilon_n \delta_{mn}$ gives the energy expectation

$$\widehat{C}_i^\dagger H \widehat{C}_i = \sum_{mn} u_{mi}^* u_{ni} C_m^\dagger H C_n = \sum_m \varepsilon_m |u_{mi}|^2. \tag{55}$$

Writing the reference energy $\varepsilon_i$ as $\varepsilon_i \sum_m |u_{mi}|^2$ using Equation (54) and subtracting yields

$$\widehat{C}_i^\dagger H \widehat{C}_i - \varepsilon_i = \sum_m (\varepsilon_m - \varepsilon_i) |u_{mi}|^2. \tag{56}$$

Each admixed state therefore contributes its mixing probability multiplied by its energy difference from the original occupied state.

The predicted eigenvalue equation determines the mixing coefficients. Substituting $\widehat{H} = H + \Delta H$ and multiplying from the left by $C_m^\dagger$ gives

$$\begin{aligned} (H + \Delta H)\widehat{C}_i &= S \widehat{C}_i \widehat{\varepsilon}_i, \\ \varepsilon_m u_{mi} + C_m^\dagger \Delta H \widehat{C}_i &= \widehat{\varepsilon}_i u_{mi}. \end{aligned} \tag{57}$$

The second line uses $C_m^\dagger H = \varepsilon_m C_m^\dagger S$. Defining $\Delta H_{mi} = C_m^\dagger \Delta H C_i$ and rearranging for states with $\varepsilon_m \neq \varepsilon_i$ gives

$$u_{mi} = \frac{C_m^\dagger \Delta H \widehat{C}_i}{\widehat{\varepsilon}_i - \varepsilon_m} = \frac{\Delta H_{mi}}{\varepsilon_i - \varepsilon_m} + O(\Delta H^2). \tag{58}$$

The last equality uses a first-order approximation: near the reference, $\widehat{C}_i = C_i + O(\Delta H)$ and $\widehat{\varepsilon}_i = \varepsilon_i + O(\Delta H)$. Since the numerator already contains the first-order factor $\Delta H$, replacing the predicted orbitals and energies by their reference values introduces corrections only at second order. Here, $O(\Delta H^p)$ denotes terms of order $p$ as the entire error matrix is scaled down, with $S$ and the reference spectrum fixed. If the reference contains degeneracies, we choose eigenstates that diagonalize the perturbation within each degenerate subspace and track the predicted states continuously. Terms between equal reference energies are multiplied by zero in Equation (56), so subsequent sums include only terms between distinct energies.

Equation (58) gives the squared magnitude of the mixing coefficient:

$$|u_{mi}|^2 = \frac{|\Delta H_{mi}|^2}{(\varepsilon_i - \varepsilon_m)^2} + O(\Delta H^3), \qquad \varepsilon_m \neq \varepsilon_i. \tag{59}$$

Substituting into Equation (56) and summing over all occupied states yields

$$\begin{aligned} \Delta\mathcal{L}_{\mathrm{VS}} &= \sum_{i\in\mathrm{occ}} \left(\widehat{C}_i^\dagger H \widehat{C}_i - \varepsilon_i\right) \\ &= \sum_{i\in\mathrm{occ}} \sum_{m:\,\varepsilon_m\neq\varepsilon_i} (\varepsilon_m - \varepsilon_i) \frac{|\Delta H_{mi}|^2}{(\varepsilon_i - \varepsilon_m)^2} + O(\Delta H^3) \\ &= \sum_{i\in\mathrm{occ}} \sum_{m:\,\varepsilon_m\neq\varepsilon_i} \frac{|\Delta H_{mi}|^2}{\varepsilon_m - \varepsilon_i} + O(\Delta H^3). \end{aligned} \tag{60}$$

The inner sum contains both occupied and unoccupied states. For two occupied states $i, j$ with distinct energies, Hermiticity of $\Delta H$ gives $|\Delta H_{ji}|^2 = |\Delta H_{ij}|^2$, so their contributions cancel pairwise:

$$\frac{|\Delta H_{ji}|^2}{\varepsilon_j - \varepsilon_i} + \frac{|\Delta H_{ij}|^2}{\varepsilon_i - \varepsilon_j} = 0. \tag{61}$$

Terms between equal energies already vanish, leaving only coupling between occupied and unoccupied states:

$$\Delta\mathcal{L}_{\mathrm{VS}} = \sum_{i\in\mathrm{occ}} \sum_{a\in\mathrm{virt}} \frac{|\Delta H_{ai}|^2}{\varepsilon_a - \varepsilon_i} + O(\Delta H^3). \tag{62}$$

Equation 62 corresponds to a first-order perturbation expansion of the wavefunctions restricted to sums between occupied and unoccupied states. When the predicted Hamiltonian error is sufficiently small, VSLoss is therefore mathematically similar to first-order wavefunction perturbation theory. The difference is that VSLoss contains only coupling between the occupied and unoccupied subspaces, not errors within either subspace. This reflects the HK variational principle in training: because the minimum ground-state energy corresponds to a unique charge density, the difference between predicted and reference wavefunctions decreases as VSLoss lowers the ground-state energy. The numerator in Equation 62 is also the objective optimized by WALoss (Li et al., 2025) and the k-space loss in NextHAM (Yin et al., 2026). VSLoss adds energy-gap weighting in the denominator, improving wavefunction training. Its sum over occupied–unoccupied coupling terms is closely related to the high-energy–low-energy subspace coupling emphasized by NextHAM. Whereas NextHAM identifies this explicitly from empirical observations, VSLoss incorporates the same insight implicitly through the variational principle.

### D.3 Limitations

Despite its use of the variational principle, VSLoss has limitations. With the notation above, the first-order perturbation of an energy level is

$$\widehat{\varepsilon}_n - \varepsilon_n = C_n^\dagger \Delta H C_n + O(\Delta H^2). \tag{63}$$

The first-order energy shift depends on diagonal errors in the reference eigenbasis, whereas the VSLoss expansion in Equation (62) contains only coupling errors between occupied and unoccupied states. The diagonal terms therefore require a separate constraint. We add band supervision to VSLoss in the main text to constrain the predicted energy levels directly. The two losses act on

the occupied subspace and the energy spectrum, respectively, and together form the second-stage training objective.

VSLoss also requires first-stage pretraining on matrix elements. The true energy gradient depends on the Kohn–Sham Hamiltonian associated with the current predicted density. Replacing it with the fixed self-consistent reference avoids repeatedly assembling the effective Hamiltonian during training. When the prediction is far from the reference, however, the approximation $H^{\mathrm{KS}}(\widehat{P}) \approx H^{\mathrm{KS}}(P) = H$ fails, and VSLoss supplies incorrect gradients that disrupt training. VSLoss can therefore be used only for fine-tuning, not for training from scratch, which is a clear limitation relative to other approaches.

# E Additional ablation studies

## E.1 EGMF ablations: input matrices and injection locations

Section 4.2 demonstrates the applicability of EGMF across backbones and the accuracy gains from using matrix priors as input features. To better understand these gains, we ablate the input matrices and injection locations on the $SiO_2$ dataset using HamGNN as the backbone. We first compare the overlap matrix $S$ with the non-self-consistent Hamiltonian $H^0$ as inputs, then examine the individual roles of injecting $H^0$ features through nodes and edges. Table 5 reports the results.

Table 5: Ablation of EGMF input matrices and injection locations on $SiO_2$. Using $H^0$ gives lower errors than using $S$. Retaining only edge features from $H^0$ yields a lower error than either node-only or combined node-and-edge features. All configurations use HamGNN as the backbone. Values are Hamiltonian MAEs in units of $10^{-4}$ eV; lower is better. Nodes and edges receive features from onsite and offsite blocks, respectively. Bold indicates the lowest error.

| Configuration | MAE ↓ |
|---|---|
| HamGNN | 35.10 |
| HamGNN + $H^0$ residual learning | 15.24 |
| HamGNN + EGMF ($S$) | 28.30 |
| HamGNN + EGMF ($H^0$, nodes and edges) | 9.69 |
| HamGNN + EGMF ($H^0$, nodes only) | 14.45 |
| HamGNN + EGMF ($H^0$, edges only) | **8.95** |

The non-self-consistent Hamiltonian provides a more effective input prior than the overlap matrix. Adding $S$ reduces the Hamiltonian MAE from $35.10 \times 10^{-4}$ eV for the original HamGNN to $28.30 \times 10^{-4}$ eV, showing that basis-set and orbital-overlap information helps learning. Using $H^0$ as the EGMF input reduces the error further to $9.69 \times 10^{-4}$ eV. As described in Appendix B.3, $H^0$ contains kinetic-energy and local and nonlocal pseudopotential terms, providing physical information beyond orbital overlap. This result shows that incorporating this information directly into the input representation can further improve Hamiltonian prediction.

Having established the value of the physical information in $H^0$, we examine where it enters the network. EGMF converts onsite blocks into node features and offsite blocks into edge features. With node inputs alone, the MAE rises from $9.69 \times 10^{-4}$ eV for combined node-and-edge inputs to $14.45 \times 10^{-4}$ eV. Edge inputs alone give an error of $8.95 \times 10^{-4}$ eV, slightly below that obtained with both feature types. This comparison shows that interatomic matrix information plays a more prominent role in learning this system, and supplying it through edge features alone already yields substantial accuracy gains.

The role of edge features also supports the discussion of matrix inputs in the main text. DEQHNet uses onsite matrix inputs to capture intra-atomic orbital information (Wang et al., 2024b), whereas EGMF also incorporates interatomic matrix information into the initial edge representations for use in message passing. The ablations demonstrate the value of this extension for periodic $SiO_2$. Together, the two comparisons show that EGMF benefits both from the richer physical priors in $H^0$ and from its use of interatomic matrix information.

### E.2 H2Net ablations: dual-branch architecture and matrix inputs

To separate the effects of the dual-branch architecture and matrix inputs, we compare HamGNN, H2Net without EGMF, and the full H2Net on three HSE systems. These configurations use neither design, the dual-branch architecture alone, and both designs, respectively. Table 6 reports the results.

Table 6: Ablation of the dual-branch architecture and matrix inputs in H2Net. Retaining the dual-branch architecture without EGMF still gives lower Hamiltonian MAEs than HamGNN on all three systems. Adding EGMF reduces the errors further. MAE is reported in units of $10^{-4}$ eV; lower is better. Bold indicates the best result for each system.

| Configuration | $MoS_2$ | $MAPbI_3$ | $SiO_2$ |
|---|---|---|---|
| HamGNN | 0.82 | 0.77 | 5.09 |
| H2Net without EGMF | 0.72 | 0.54 | 2.56 |
| H2Net | **0.64** | **0.50** | **2.28** |

Even without matrix features, the dual-branch architecture improves accuracy on all three systems. Relative to HamGNN, H2Net without EGMF reduces the Hamiltonian MAE from $0.82$ to $0.72$ for $MoS_2$, from $0.77$ to $0.54$ for $MAPbI_3$, and from $5.09$ to $2.56$ for $SiO_2$, all in units of $10^{-4}$ eV. Assigning message-passing depth according to interaction range therefore improves Hamiltonian prediction as well as providing the memory savings shown in the main text.

Adding EGMF to the dual-branch architecture further reduces errors on all three systems: from $0.72$ to $0.64$ for $MoS_2$, from $0.54$ to $0.50$ for $MAPbI_3$, and from $2.56$ to $2.28$ for $SiO_2$. Matrix inputs thus provide an additional accuracy gain. Together, the comparisons show that H2Net benefits from both the dual-branch architecture, which adapts the modeling approach to different interaction ranges, and matrix priors, which supply directly usable physical information.

We also ablate network depth on $MoS_2$, with results in Figure 3. HamGNN's prediction accuracy drops markedly as the number of layers decreases. In contrast, when the inner branch is sufficiently deep, H2Net's accuracy changes little with the depth of the outer branch. This confirms the earlier analysis that the simpler interactions in the outer branch can be learned with a shallow network. A single outer-branch layer therefore reduces the receptive field and memory requirements while maintaining high prediction accuracy.

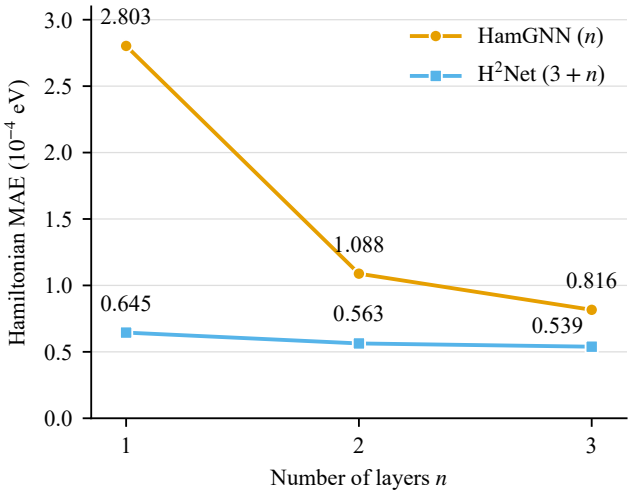


Figure 3: Network-depth ablation on $MoS_2$. Orange circles denote HamGNN with $n$ layers. Blue squares denote H2Net with three inner-branch layers and $n$ outer-branch layers. The vertical axis reports Hamiltonian MAE in units of $10^{-4}$ eV.

### E.3 JOINT FINE-TUNING WITH BAND LOSS AND VSLOSS FOR HSE HAMILTONIAN PREDICTION

To assess joint fine-tuning with band loss (BandLoss) and VSLoss for HSE Hamiltonian prediction, we compare H2Net before and after this fine-tuning on $\mathrm{MoS_2}$, $\mathrm{MAPbI_3}$, and $\mathrm{SiO_2}$. Table 7 reports the same three metrics as the main text: Hamiltonian MAE, band MAE, and wavefunction cosine dissimilarity.

Table 7: Joint fine-tuning with band loss and VSLoss for HSE Hamiltonian prediction. Joint fine-tuning lowers both band MAE and wavefunction cosine dissimilarity on all three systems, with the largest improvement on $\mathrm{SiO_2}$. $\Delta H$ and $\Delta\varepsilon$ denote Hamiltonian MAE and band MAE, in units of $10^{-4}\,\mathrm{eV}$ and $\mathrm{meV}$, respectively. $\Delta\psi$ is wavefunction cosine dissimilarity, expressed as a percentage. Lower values are better for all three metrics; bold indicates the best result for each metric in each system.

| System | Configuration | $\Delta H \downarrow$ | $\Delta\varepsilon \downarrow$ | $\Delta\psi \downarrow$ |
|---|---|---|---|---|
| $\mathrm{MoS_2}$ | H2Net | **0.64** | 5.28 | 0.99% |
| | H2Net + BandLoss + VSLoss | 2.00 | **4.79** | **0.76%** |
| $\mathrm{MAPbI_3}$ | H2Net | **0.50** | 13.12 | 5.56% |
| | H2Net + BandLoss + VSLoss | 0.66 | **13.09** | **5.52%** |
| $\mathrm{SiO_2}$ | H2Net | **2.28** | 326.52 | 18.01% |
| | H2Net + BandLoss + VSLoss | 3.86 | **75.92** | **7.08%** |

Joint fine-tuning with band loss and VSLoss lowers both band error and wavefunction cosine dissimilarity on all three systems. The largest improvement occurs for $\mathrm{SiO_2}$, where band MAE falls from $326.52\,\mathrm{meV}$ to $75.92\,\mathrm{meV}$ and wavefunction cosine dissimilarity from $18.01\%$ to $7.08\%$. For $\mathrm{MoS_2}$, the two metrics decrease from $5.28\,\mathrm{meV}$ and $0.99\%$ to $4.79\,\mathrm{meV}$ and $0.76\%$, respectively. The gains for $\mathrm{MAPbI_3}$ are smaller, from $13.12\,\mathrm{meV}$ and $5.56\%$ to $13.09\,\mathrm{meV}$ and $5.52\%$. Compared with the broadly pronounced wavefunction improvements for PBE in the main text, the HSE gains are concentrated more strongly in $\mathrm{SiO_2}$, although downstream predictions improve consistently across all three systems.

At the same time, Hamiltonian MAE increases on all three systems. For example, the $\mathrm{SiO_2}$ Hamiltonian MAE rises from $2.28 \times 10^{-4}\,\mathrm{eV}$ to $3.86 \times 10^{-4}\,\mathrm{eV}$ even as its band and wavefunction predictions improve. This agrees with the observation in the main text that matrix-element fitting errors and downstream physical errors do not always change together. Jointly incorporating band loss and VSLoss into the fine-tuning objective thus also improves downstream physical predictions for HSE.

### E.4 COMPARISON WITH DEEPH-HYBRID

We compare HSE Hamiltonian prediction accuracy on the datasets used in the original DeepH-hybrid study (Tang et al., 2024). The two datasets were generated by randomly perturbing bilayer graphene and $\mathrm{MoS_2}$ structures, respectively. We train HamGNN and H2Net on these datasets and compare them with the results reported in the DeepH-hybrid paper, as shown in Table 8.

Table 8: Hamiltonian prediction on the original DeepH-hybrid datasets. H2Net achieves the lowest MAE on both bilayer graphene and $\mathrm{MoS_2}$. DeepH-hybrid values are taken from the original paper; HamGNN and H2Net results are from our training runs. MAE is reported in units of $10^{-4}\,\mathrm{eV}$. Lower values are better, and bold indicates the best result for each system.

| Method | Bilayer graphene | $\mathrm{MoS_2}$ |
|---|---|---|
| DeepH-hybrid | 1.47 | 2.65 |
| HamGNN | 0.77 | 0.95 |
| H2Net | **0.53** | **0.62** |

The two datasets show a consistent pattern. Relative to DeepH-hybrid, HamGNN reduces Hamiltonian MAE by approximately $47.6\%$ on bilayer graphene and $64.2\%$ on $MoS_2$. H2Net further reduces the errors relative to HamGNN by approximately $31.2\%$ and $34.7\%$, respectively, achieving the lowest Hamiltonian prediction error among the three methods on both datasets.

## F Evaluation metrics

The main text evaluates matrix-element and downstream prediction accuracy using Hamiltonian MAE, band MAE, and wavefunction cosine dissimilarity. The generalized eigenvalue equation $HC = SC\varepsilon$ shows that replacing $H$ with $H - \delta S$ shifts all eigenvalues by $-\delta$ without changing the eigenstates. We therefore remove this difference in the energy reference before computing Hamiltonian MAE:

$$\Delta H = \frac{1}{N_H} \sum_a \left| \widehat{H}_a - H_a - \delta_H S_a \right|, \qquad \delta_H = \frac{\sum_a (\widehat{H}_a - H_a) S_a}{\sum_a S_a^2}. \tag{64}$$

Here, $\widehat{H}$ and $H$ are the predicted and reference Hamiltonians, $a$ runs over the real-space matrix elements included in the evaluation, $N_H$ is their total number, and $S$ is the corresponding overlap matrix.

Band MAE is likewise computed after removing a uniform energy shift. For each test structure, we randomly select six $k$ points for evaluation, denoted by $\mathcal{K}$, with $N_k = |\mathcal{K}| = 6$. To focus on states near the Fermi level and exclude unrelated states, we select the five bands immediately below and the five immediately above the reference Fermi level at each $\mathbf{k} \in \mathcal{K}$, with index set $\mathcal{B}(\mathbf{k})$. Predicted and reference eigenvalues are sorted separately by energy and matched using the reference band indices. Band MAE is defined as

$$\begin{aligned} \Delta\varepsilon &= \frac{1}{N_k N_B} \sum_{\mathbf{k}\in\mathcal{K}} \sum_{n\in\mathcal{B}(\mathbf{k})} \left| \widehat{\varepsilon}_n(\mathbf{k}) - \varepsilon_n(\mathbf{k}) - \delta_\varepsilon \right|, \\ \delta_\varepsilon &= \operatorname{median}_{\substack{\mathbf{k}\in\mathcal{K} \\ n\in\mathcal{B}(\mathbf{k})}} \left( \widehat{\varepsilon}_n(\mathbf{k}) - \varepsilon_n(\mathbf{k}) \right). \end{aligned} \tag{65}$$

Here, $N_B = |\mathcal{B}(\mathbf{k})| = 10$ is the number of evaluated bands at each $k$ point. For a given structure, all evaluation points and selected bands share a single uniform shift $\delta_\varepsilon$.

At the same $k$ points and for the same selected bands, we measure eigenstate prediction errors using wavefunction cosine dissimilarity. Since wavefunctions have an arbitrary global phase, we take the magnitude of the inner product between normalized states to remove this phase ambiguity:

$$\begin{aligned} \Delta\psi &= \frac{1}{N_k N_B} \sum_{\mathbf{k}\in\mathcal{K}} \sum_{n\in\mathcal{B}(\mathbf{k})} \left( 1 - \left| \langle \psi_n(\mathbf{k}) \mid \widehat{\psi}_n(\mathbf{k}) \rangle \right| \right), \\ \langle \psi_n(\mathbf{k}) \mid \widehat{\psi}_n(\mathbf{k}) \rangle &= C_n^\dagger(\mathbf{k}) S(\mathbf{k}) \widehat{C}_n(\mathbf{k}). \end{aligned} \tag{66}$$

Here, $C_n(\mathbf{k})$ and $\widehat{C}_n(\mathbf{k})$ are the reference and predicted wavefunction coefficients in the LCAO basis, and $S(\mathbf{k})$ is the corresponding overlap matrix. The coefficients are normalized such that $C_n^\dagger(\mathbf{k}) S(\mathbf{k}) C_n(\mathbf{k}) = \widehat{C}_n^\dagger(\mathbf{k}) S(\mathbf{k}) \widehat{C}_n(\mathbf{k}) = 1$. Lower values indicate closer agreement between wavefunctions. The main text reports this metric as a percentage.

## G Choice of cutoff radii

H2Net uses inner- and outer-graph cutoffs of 1.2 and 2 times the sum of the orbital cutoff radii at the two endpoints, respectively. In ABACUS, matrix elements from kinetic energy and local potentials vanish when the two orbitals no longer overlap. The nonlocal pseudopotential, however, can connect them through projector functions, producing coupling beyond their direct-overlap range; Appendix B gives the integral form. The inner-graph cutoff therefore needs to extend slightly beyond the sum of the orbital cutoffs.

For $MoS_2$ under PBE, both Mo and S have orbital cutoff radii of $7\,\mathrm{bohr}$, giving a direct-overlap range of $14\,\mathrm{bohr}$. Figure 4 plots, for each distance interval, the 99th percentile of absolute nonzero

matrix elements, calculated separately for each element pair and then averaged. Matrix elements near $14\,\mathrm{bohr}$ remain on the order of $10^{-4}$ Ha. These extended couplings decay rapidly with distance, reaching approximately $10^{-7}$ Ha near $16.8\,\mathrm{bohr}$, close to machine precision, where they can be treated as noise. We therefore set the inner-graph cutoff to $1.2$ times the sum of the orbital cutoffs to include these weak but still contributing couplings.

For HSE, screened HF exchange introduces additional coupling beyond the direct-overlap range, requiring a larger graph cutoff. The outer graph uses twice the sum of the orbital cutoff radii to capture these extended exchange contributions, following DeepH-hybrid (Tang et al., 2024). The inner graph thus covers the main coupling range of the PBE terms, while the outer graph also captures the extended exchange coupling in HSE, consistent with the dual-branch design in the main text.

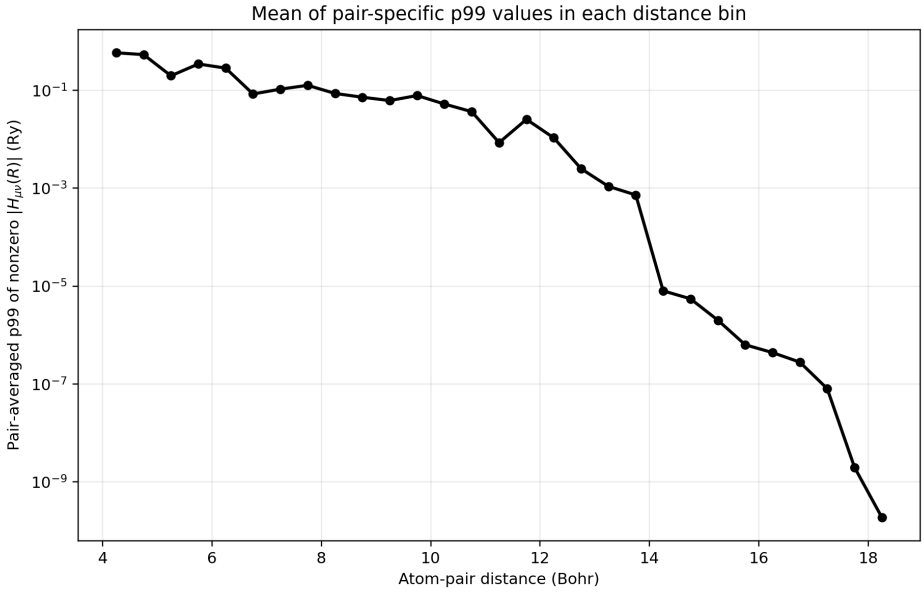


Figure 4: Decay of PBE Hamiltonian matrix elements with interatomic distance in $MoS_2$. Within each distance interval, the 99th percentile of absolute nonzero matrix elements is calculated separately for each element pair and then averaged over element pairs.

## H Dataset construction

We performed molecular dynamics simulations of bilayer $\mathrm{MoS_2}$ and $\mathrm{MAPbI_3}$ in ASE using the SevenNet neural network potential, a Bussi thermostat, and a time step of 1 fs. The simulation temperatures were 500 K and 300 K, and the supercells contained 54 and 48 atoms, respectively. To sample well-equilibrated configurations, we discarded the first 5 ps of each trajectory and then sampled structures every 50 fs. For each material, we ran three independent trajectories and collected 150, 50, and 50 structures to form the training, validation, and test sets, respectively. We then calculated PBE and HSE Hamiltonians for the sampled structures using ABACUS, with $k$-point grids of $4 \times 4 \times 1$ for $\mathrm{MoS_2}$ and $3 \times 3 \times 5$ for $\mathrm{MAPbI_3}$.

For the SACADA carbon-allotrope and $\mathrm{SiO_2}$ isomer datasets from the original HamGNN study, the PBE Hamiltonians were calculated with OpenMX. We retained the original $8 : 1 : 1$ training, validation, and test split. We also sampled 250 structures from the $\mathrm{SiO_2}$ dataset, assigning 150, 50, and 50 to the training, validation, and test sets, respectively. Their HSE Hamiltonians were calculated using ABACUS to supplement the HSE data.

# I TRAINING DETAILS

All training used the Adam optimizer. HamGNN and H2Net were trained from scratch with an MAE loss and an initial learning rate of $0.01$. We used a ReduceLROnPlateau scheduler with a decay factor of $0.5$ and a patience of $5$. Training stopped when the loss failed to decrease for $20$ consecutive epochs or after $300$ epochs.

DeepH-E3 also used an MAE loss and a ReduceLROnPlateau scheduler with a decay factor of $0.5$, but its initial learning rate and patience were set to $0.002$ and $60$, respectively. Training stopped when the learning rate reached $3 \times 10^{-5}$ or lower, or after $2000$ epochs.

After matrix-element pretraining, we fine-tuned the models following the two-stage strategy described in the main text. The initial learning rate was $10^{-4}$, and the Hamiltonian matrix-element loss, VSLoss, and band loss had relative weights of $1000 : 100 : 10$. Fine-tuning lasted at most $500$ epochs, with all other settings unchanged from training the corresponding model from scratch.