Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bridging the Training--Application Gap in Kohn--Sham Hamiltonian Learning through Dual-Space Supervision

Yifei Wang
Xi'an Jiaotong University
He Zhang
Zhongguancun Academy
Xinran Wei
Zhongguancun Academy
Jingwen Fu
Zhongguancun Academy
Yongqiang Ma
Xi'an Jiaotong University
Miao Kang
Xi'an Jiaotong University
Jianji Wang
Xi'an Jiaotong University
Chang Liu
liuchang@bza.edu.cn
Zhongguancun Academy
Nanning Zheng
Xi'an Jiaotong University

Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11240155/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bridging the Training–Application Gap in Kohn–Sham Hamiltonian Learning through Dual-Space Supervision

Abstract

Predicting the Kohn–Sham Hamiltonian in density functional theory (DFT) via machine learning offers a promising pathway toward efficient electronic structure modeling. Existing models are typically trained by directly minimizing element-wise Hamiltonian matrix errors, yet a significant training–application gap has been observed between Hamiltonian-space loss and errors in derived electronic properties, limiting their utility in practical applications. Noting that these properties are functions on the state space obtained by solving the Hamiltonian eigenvalue problem, this gap can be traced to a misalignment of the error metrics on the Hamiltonian space and its induced state space. The operator-to-state mapping is highly anisotropic, such that Hamiltonian errors of comparable numerical magnitude can induce markedly different downstream property errors. Motivated by the Hamiltonian–state variational duality rooted in Lieb's convex formulation of DFT, we introduce Dual-Space Loss (DSLoss) to bridge this training–application gap. By leveraging error metrics in both spaces, DSLoss allows Hamiltonian optimization to account for errors in both the predicted operator and its induced state. Across molecular benchmarks, DSLoss consistently improves a broad range of derived properties over conventional element-wise supervision. On the challenging ∇^2DFT benchmark, electron density and total energy errors are reduced by factors of 40 and 500, respectively, relative to the strongest competing method. The resulting model also extrapolates robustly to molecules containing up to twice as many heavy atoms as those encountered during training, achieving more than an order-of-magnitude lower errors for multiple properties and substantially outperforming direct density prediction models and end-to-end property predictors. Moreover, the model enables stable molecular dynamics simulations with accurate energetics, forces, and vibrational dynamics. These results show that DSLoss addresses an unrecognized bottleneck limiting operator-prediction models across diverse downstream tasks, thereby advancing scalable and reliable electronic structure modeling.

1 Introduction

Accurate prediction of molecular properties is central to applications ranging from drug discovery to materials design [1, 2]. These properties are governed by the underlying electronic structure, making electronic structure theory a fundamental tool for molecular modeling. Among existing approaches, density functional theory (DFT) [3–5] has become a computational workhorse because it provides a favorable balance between accuracy and cost, enabling broad scientific advances [6–8]. In the prevailing Kohn–Sham formulation (KSDFT) [4], ground-state properties are obtained by solving the Kohn–Sham equations through self-consistent field iterations. Despite its widespread success, this iterative procedure typically scales as $O(N^3)$ to $O(N^4)$ with the number of electrons N, creating a major computational bottleneck for large-scale molecular simulations.

Machine learning has therefore emerged as a promising means of bypassing these expensive electronic structure calculations. Substantial progress has been made in directly predicting task-specific properties, such as energies, forces and band gaps [9–19]. However, such models are usually trained separately for individual observables and do not explicitly preserve the underlying electronic structure, limiting their transferability across downstream tasks. A more general alternative is to predict the converged Kohn—Sham Hamiltonian (Fig. 1a), which in principle provides access to a broad range of ground-state observables without redefining the learning target for each property. This promise has stimulated growing interest in Hamiltonian learning, from early kernel-based models for low-dimensional systems [20] to deep neural networks for molecules or materials [21–23]. Because Hamiltonian matrix elements form geometric tensors that transform equivariantly under rotations, subsequent studies have developed symmetry-preserving architectures based on equivariant local frames [24–28] or high-order spherical tensors with Clebsch–Gordan tensor products [29–35].

Despite these advances in model architecture, prevailing Hamiltonian prediction methods are still typically trained by minimizing the mean absolute error (MAE) or mean squared error (MSE) between predicted and reference matrix elements [29, 30] ($\mathcal{L}_{\mathbf{H}}$ in Fig. 1a). This choice is natural because the converged Hamiltonian provides a direct and comparatively unambiguous training target. Its simplicity, however, confines supervision only to Hamiltonian space, whereas downstream properties depend on the

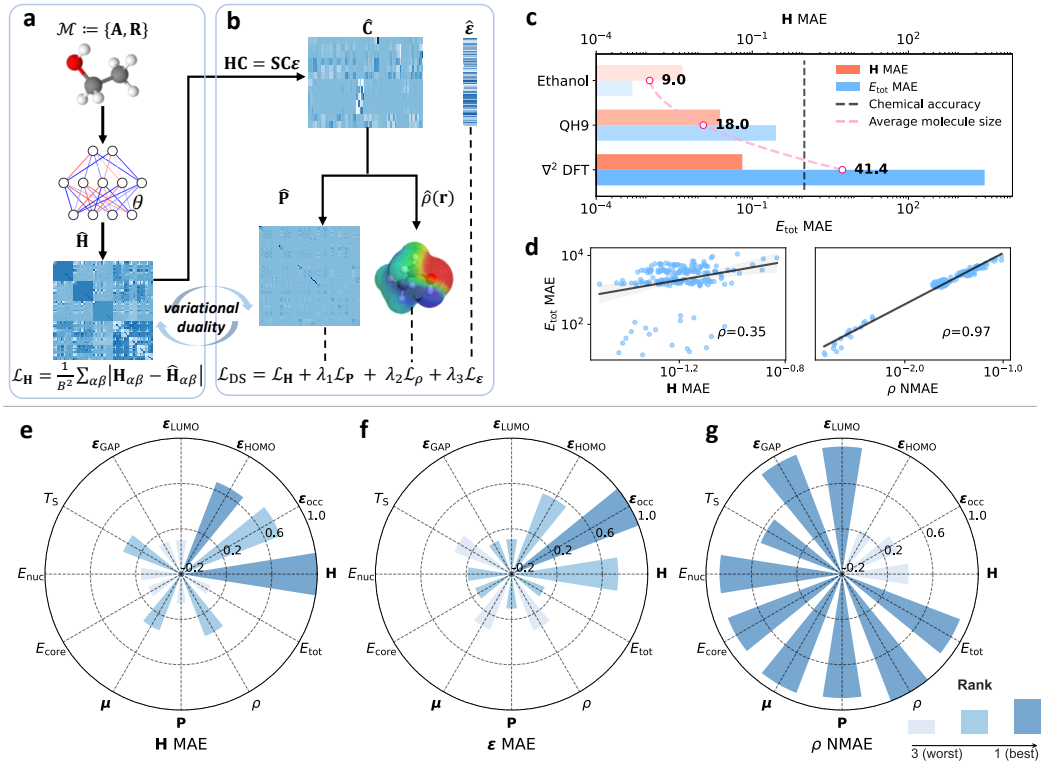


Figure 1: **Overview of Dual-Space Loss for Hamiltonian learning.** **(a)** Hamiltonian learning pipeline. Given a molecular system $\mathcal{M}$, an SE(3)-equivariant neural network predicts the ground-state Hamiltonian matrix $\hat{\mathbf{H}}$. Conventional Hamiltonian learning minimizes element-wise Hamiltonian errors using $\mathcal{L}_{\mathbf{H}}$. **(b)** Schematic of DSLoss. Inspired by Hamiltonian-state variational duality, DSLoss further constrains the electronic state induced by the predicted Hamiltonian by matching the reference density in the atomic orbital (AO) basis through $\mathcal{L}_{\mathbf{P}}$ and in real space through $\mathcal{L}_{\rho}$, together with an orbital energy regularization term $\mathcal{L}_{\boldsymbol{\varepsilon}}$. **(c)** Hamiltonian and total energy errors of QHNet across three benchmarks. Hamiltonian errors are evaluated on 4,500, 13,084 and 9,345 test structures, respectively, whereas total energy errors are evaluated on 50 test structures from each benchmark, except for ∇^2DFT, which uses 200. **(d)** Pearson correlations of Hamiltonian and electron density errors with total energy error, evaluated on 200 randomly sampled molecules from ∇^2DFT. **(e)-(g)** Pearson correlations of Hamiltonian, orbital energy and electron density errors with errors in representative derived properties. All quantities are evaluated using QHNet on 200 randomly sampled QH9 molecules. For each derived property, color intensity indicates the relative ranking of its correlations with errors in $\mathbf{H}$, $\boldsymbol{\varepsilon}$ and ρ.

derived electronic state, which reside in a space nonlinearly related to the Hamiltonian through its eigenvalue problem (Eq. (1)). The resulting separation creates a fundamental training–application gap: the operator-to-state mapping is highly anisotropic, such that Hamiltonian errors of comparable numerical magnitude can induce markedly different state errors, and reducing matrix-level errors does not necessarily yield commensurate improvements in downstream properties. This gap can become particularly severe as molecular size increases. As shown in Fig. 1c, across three datasets spanning distinct molecular scales, models trained with $\mathcal{L}_{\mathbf{H}}$ achieve comparable Hamiltonian MAEs yet exhibit markedly different total energy errors. While the resulting energies remain accurate for small molecules, errors in larger systems exceed chemical accuracy by more than three orders of magnitude, severely limiting the practical utility of the predicted Hamiltonians. This mismatch is further evidenced by the weak correlation between Hamiltonian MAE and total energy MAE (Fig. 1d, left), with similarly inconsistent behavior observed for other ground-state properties (Fig. 1e) and in independent studies [36]. Further analysis of the origin of this mismatch is provided in Results 2.1 and Supplementary Sec. B.1.

Several recent studies have therefore explored supervision signals beyond element-wise Hamiltonian errors. Ref. [37] proposed optimizing the total energy derived from the predicted Hamiltonian, whereas ref. [38] introduced a self-consistency objective that minimizes the discrepancy between the predicted Hamiltonian and the Hamiltonian reconstructed from its induced electronic state. These approaches leverage physical conditions satisfied by converged DFT solutions and provide informative supervision without requiring explicit Hamiltonian labels. Nevertheless, evaluating total energies or reconstructing Hamiltoni-

ans during training introduces substantial computational cost, limiting their scalability to larger molecular systems. More recently, WALoss was proposed to projects the predicted Hamiltonian onto the reference molecular orbital basis and supervises the resulting diagonal elements as orbital energies [36]. While effective for improving orbital energy prediction, these gains do not consistently transfer to downstream properties (Supplementary Table S17), because WALoss primarily constrains the eigenvalue spectrum rather than the electronic state that governs these observables.

In this work, we introduce a Dual-Space loss (***DSLoss***) to bridge this training–application gap by supervising the Hamiltonian prediction model from both the Hamiltonian space and the electronic state space. This can be seen as supervising the model from dual spaces under the view of Lieb's variational duality between an effective potential and electron density [39], where the former constructs the Hamiltonian and the latter is a representative of the state space. See Supplementary Sec. A.3 for detailed discussions. Specifically, we complete Hamiltonian supervision with electronic-state supervision, where the latter is implemented as the combination of losses on real-space density and on density matrix as the density representation on atomic orbital basis. By differentiating through the eigensolution, these signals directly align Hamiltonian optimization with the ground-state electronic structure used at evaluation while retaining computational efficiency. The effectiveness of this design is supported empirically by the strong correlations between density errors and downstream property errors (Fig. 1d; Fig. 1g), and theoretically by first-order perturbation analysis for the electronic state space (Supplementary Sec. B.1). A systematic comparison with alternative computationally efficient physical surrogates further demonstrates that electron density provides the most consistent improvements across downstream properties (Fig. 2a). To additionally constrain the eigenvalue spectrum, we incorporate supervision on differences between predicted and reference orbital energies, yielding the final DSLoss ($\mathcal{L}_{\mathrm{DS}}$ in Fig. 1b). The resulting objective is architecture-agnostic and can be integrated with general Hamiltonian predictors. Here, we instantiate it with the widely used QHNet architecture [30] and refer to the resulting framework as ***DSHam***.

We highlight the practical advantages of DSLoss across four aspects. **(1)** Across multiple benchmarks, DSLoss consistently improves prediction accuracy of electron density, total energy and other derived properties, with increasingly pronounced gains on more challenging datasets. On the ∇^2DFT benchmark, electron density and total energy errors are reduced by factors of 40 and 500, respectively, relative to the strongest competing method. **(2)** DSHam extrapolates robustly to molecules containing up to twice as many heavy atoms as those encountered during training. It also achieves substantially lower errors than both Hamiltonian prediction baselines and end-to-end property predictors, including improvements of more than an order of magnitude for several properties. **(3)** DSHam outperforms models that directly predict electron densities or density matrices, underscoring the advantage of learning the Hamiltonian while constraining its induced electronic state. **(4)** DSHam enables stable, DFT-consistent molecular dynamics simulations and closely reproduces the DFT vibrational density of states, with a spectral overlap of 95.3%. Together, these results establish dual-space supervision for Hamiltonian learning as a scalable and physically grounded framework for molecular electronic structure modeling.

2 Results

2.1 Methodology of DSLoss

For a given molecular structure $\mathcal{M} := \{\mathcal{Z}, \mathbf{X}\}$, where $\mathcal{Z} := \{Z_a\}_{a=1}^{A}$ and $\mathbf{X} := \{\mathbf{x}_a\}_{a=1}^{A}$ denote the atomic numbers and Cartesian coordinates of A atoms, KSDFT determines the ground-state electronic structure by minimizing the electronic energy with respect to a set of orthonormal one-electron orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N}$ ($\mathbf{r}$ is the electron coordinate). In practical calculations, each orbital is expanded in an atomic orbital (AO) basis $\{\eta_\alpha(\mathbf{r})\}_{\alpha=1}^{B}$ as $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} \mathbf{C}_{\alpha i}\eta_\alpha(\mathbf{r})$, where $\mathbf{C}$ is the molecular orbital (MO) coefficient matrix. The corresponding electron density is written as $\rho(\mathbf{r}) = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})$, where $\mathbf{P}_{\alpha\beta} = \sum_i \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i}$ is the density matrix. The stationary condition of the Kohn–Sham energy leads to the generalized eigenvalue problem

$$\mathbf{HC} = \mathbf{SC}\boldsymbol{\varepsilon}, \tag{1}$$

commonly referred to as the Kohn–Sham equation. Here, $\mathbf{H}_{\alpha\beta} = \int \eta_\alpha(\mathbf{r})\hat{H}\eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r}$ is the Hamiltonian matrix associated with the effective Kohn–Sham operator $\hat{H}$, $\mathbf{S}_{\alpha\beta} = \int \eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r}$ is the AO overlap matrix, and $\boldsymbol{\varepsilon}$ contains the orbital energies. In conventional DFT calculations, Eq. (1) is solved through self-consistent field (SCF) iterations because the Hamiltonian depends on the electronic density. The converged Hamiltonian $\mathbf{H}^\star$ and its induced ground-state electronic state then provide access to a broad range of ground-state molecular properties. Further details are provided in Supplementary Sec. A.1.

Hamiltonian prediction aims to bypass costly SCF iterations by learning a direct map from $\mathcal{M}$ to $\mathbf{H}^\star$ using a machine-learning model $\mathbf{H}_{\theta,\mathcal{M}}$ (Fig. 1a). Prevailing models are trained by minimizing element-

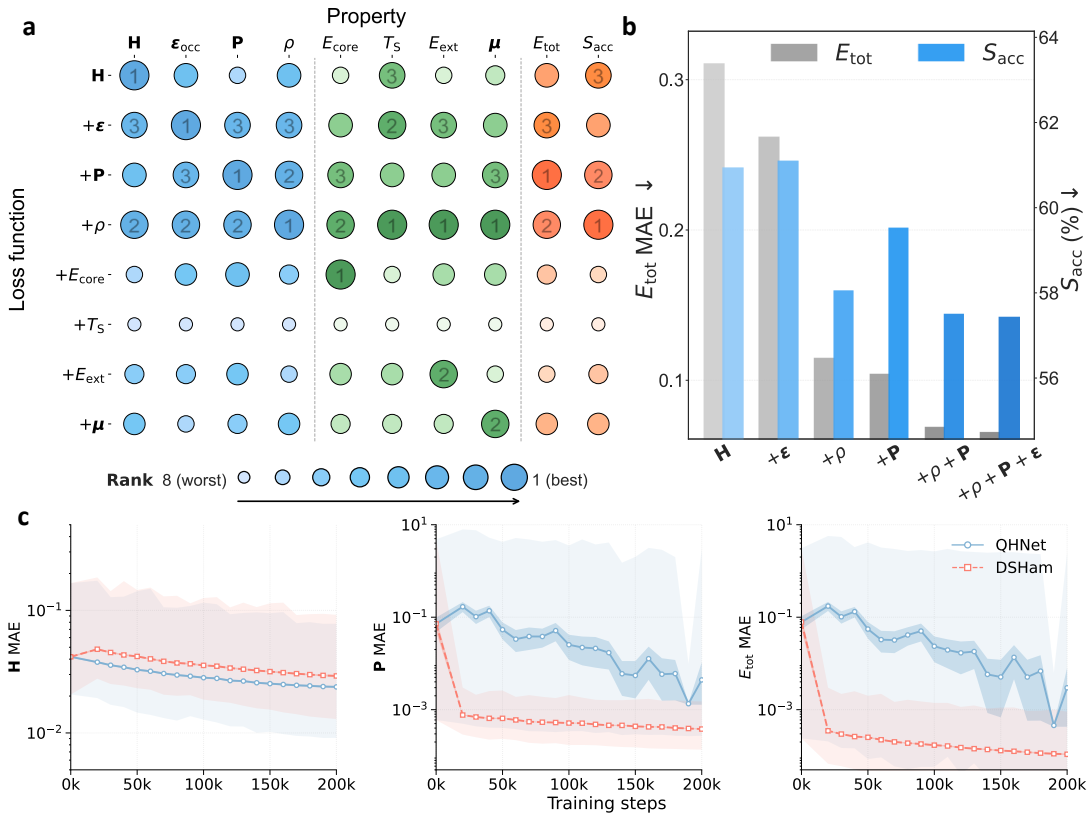


Figure 2: **Comparison of dual-space supervision with other physical objectives.** The units of $\mathbf{H}$ and E_{tot} are kcal/mol. **(a)** Performance obtained using different physical observables as auxiliary losses. Larger and darker bubbles indicate better performance. Blue, green and orange bubbles denote Hamiltonian-derived, density matrix-derived and molecule-level properties, respectively. **(b)** Performance of different loss combinations on QH9. E_{tot} and S_{acc} are evaluated on 50 test molecules, whereas all other quantities are evaluated on 13,084 test molecules. **(c)** Training curves of various Hamiltonian predictors on QH9, evaluated on 1,000 validation molecules. The outer shaded regions span the minimum and maximum values, while the inner shaded regions indicate 95% confidence intervals.

wise errors between the predicted Hamiltonian $\hat{\mathbf{H}}$ and the reference Hamiltonian $\mathbf{H}^\star$. While effective for reproducing Hamiltonian matrices, supervision restricted to the operator space does not directly constrain the electronic solution induced by the predicted Hamiltonian, and therefore may not provide sufficient guidance for optimizing downstream observables.

To address this training–application gap, we leverage the variational duality between the Kohn–Sham Hamiltonian and its ground-state density to extend supervision from the operator space to the induced electronic state through two complementary density representations. **(1)** In the AO space, we minimize the density matrix MAE, $\mathcal{L}_{\mathbf{P}} = \frac{1}{B^2}\sum_{\alpha\beta}^{B} |\hat{\mathbf{P}}_{\alpha\beta} - \mathbf{P}^\star_{\alpha\beta}|$, where $\hat{\mathbf{P}}$ and $\mathbf{P}^\star$ are the density matrices induced by the predicted and reference Hamiltonians, respectively. For fixed orbital occupations, the density matrix is invariant to unitary rotations within the occupied orbitals, thereby providing a gauge-invariant representation of the occupied subspace. Minimizing $\mathcal{L}_{\mathbf{P}}$ therefore directly penalizes distortions of the occupied subspace (Supplementary Sec. B.1.1). **(2)** In real space, we minimize the normalized MAE (NMAE) of density evaluated on irregular grids, $\mathcal{L}_\rho = \frac{\sum_g w_g |\hat{\rho}(\mathbf{r}_g) - \rho^\star(\mathbf{r}_g)|}{\sum_g w_g |\rho^\star(\mathbf{r}_g)|}$ where $\mathcal{G} = \{\mathbf{r}_g\}_{g=1}^{N_G}$ denotes sampled grid points around the molecule and w_g denotes grid weights. This term is invariant under $\text{SE}(3)$ transformations and provides spatially resolved supervision of the electron distribution, complementing the AO density matrix representation. The strong correlations between density errors in both representations and downstream property errors further support electron density as an effective supervisory signal (Fig. 1d; Fig. 1g). To further characterize eigenspectrum of the eigensolution, we supervise differences between the occupied orbital energies induced by the predicted and reference Hamiltonians. Unlike WALoss [36] that constructs orbital-energy surrogates by projecting the predicted Hamiltonian onto reference orbitals, we obtain $\hat{\varepsilon}$ by directly solving the Kohn–Sham equation in Eq. (1). Empirically, this implementation consistently outperforms WALoss across evaluated benchmarks (Supplementary Table S17). The final

Dual-Space Loss is defined as:

$$\mathcal{L}_{\mathrm{DS}} = \mathcal{L}_{\mathbf{H}} + \lambda_{\mathbf{P}}\mathcal{L}_{\mathbf{P}} + \lambda_{\rho}\mathcal{L}_{\rho} + \lambda_{\boldsymbol{\varepsilon}}\mathcal{L}_{\boldsymbol{\varepsilon}}, \tag{2}$$

$\lambda_{\mathbf{P}}$, λ_{ρ} and $\lambda_{\boldsymbol{\varepsilon}}$ are loss weights. Fig. 1b shows a schematic illustration. As shown in Fig. 2b, combining these complementary terms yields the best empirical performance.

We would like to mention that the dual-space supervision gives a universal improvement over derived properties than directly using auxiliary losses on specific targets. For comparison, we systematically compare it with several computationally efficient auxiliary objectives, including orbital energies $\boldsymbol{\varepsilon}$, one-electron energy E_{core}, external-potential energy E_{ext}, non-interacting kinetic energy T_{S}, and dipole moment $\boldsymbol{\mu}$. Fig. 2a summarizes the performance rankings of these objectives across downstream properties, with quantitative results reported in Supplementary Table S17. As expected, directly supervising a given quantity generally provides the largest improvement for that specific quantity. However, such gains are often localized to the supervised target and do not consistently transfer across other observables. Notably, density-based objectives yield the most consistent gains across nearly all evaluated properties, including total energy E_{tot} and the SCF acceleration ratio S_{acc}. This broad transferability supports electron density as a general state-level supervisory signal rather than a surrogate tailored to any individual property.

Beyond empirical comparisons, we perform a gradient analysis to understand why density supervision provides such broad benefits (Supplementary Sec. B.1), and find that: **(1)** Density-based losses propagate errors through the eigensolution to directly constrain the induced electronic state while distributing gradients across a broader range of Hamiltonian elements, thereby reducing gradient concentration and improving optimization stability (Supplementary Fig. S1). **(2)** This broader coverage remains physically structured, preferentially emphasizing valence and bonding interactions that strongly affect spatial electron distribution, whereas property-specific losses tend to concentrate gradients on limited AO interactions, leading to larger gradient variance, less stable optimization, and weaker transfer across downstream properties (Supplementary Fig. S2). We use the QHNet architecture [30] as the Hamiltonian prediction backbone. Additional results with SPHNet [40] are reported in Supplementary Table S21, where the advantage of DSLoss is consistently observed. Further implementation details are provided in Methods 4.

2.2 Performance of DSHam on Molecular Systems

We first evaluate DSHam on test molecules at comparable scales to those used for training. Three benchmarks are considered to probe complementary aspects of generalization. MD17 contains molecular dynamics trajectories of water, ethanol, malondialdehyde and uracil [41, 42] and is used to assess generalization across conformational space. QH9, evaluated using the `stable-random` split, contains molecules with 3–29 atoms and tests generalization across chemical space [30, 43]. The more challenging ∇^2DFT dataset spans molecules with 8–62 atoms and evaluates Hamiltonian learning at larger molecular scales [44, 45]. For each benchmark, all methods use identical training, validation and test splits.

We evaluate occupied orbital energies $\boldsymbol{\varepsilon}_{\mathrm{occ}}$, frontier orbital energies $\boldsymbol{\varepsilon}_{\mathrm{HOMO}}$ and $\boldsymbol{\varepsilon}_{\mathrm{LUMO}}$, the HOMO–LUMO gap $\boldsymbol{\varepsilon}_{\Delta}$, the density matrix $\mathbf{P}$, the real-space electron density ρ, the dipole moment $\boldsymbol{\mu}$ and the total energy E_{tot}. Errors are measured relative to KSDFT references using MAE, except for ρ, which is evaluated using normalized MAE. We also assess the utility of the predictions for accelerating DFT calculations. The SCF acceleration ratio S_{acc} is defined as the number of SCF iterations required when initialized with the model prediction relative to that required with the standard `MINAO` initialization. Further details on datasets, metrics and implementation are provided in Methods 4 and Supplementary Secs. B.3-B.5.

We begin by examining the training behavior of DSHam and QHNet, with QHNet trained using conventional element-wise Hamiltonian supervision. As expected, Fig. 2c shows that QHNet maintains a slight advantage in Hamiltonian MAE because its training objective directly targets element-wise matrix accuracy. This advantage, however, does not translate into superior downstream performance. During the early stages of training, QHNet exhibits large errors and substantial variance in downstream properties whereas DSHam converges more smoothly, yields more stable validation performance, and maintains lower density matrix and total energy errors throughout training. These results demonstrate that dual-space supervision provides more effective optimization toward physically consistent Hamiltonians.

We next compare DSHam with QHNet and the orbital-energy-supervised QHNet-WALoss. On QH9, DSHam substantially reduces errors in properties that are not directly supervised during training, including the dipole moment $\boldsymbol{\mu}$, total energy E_{tot} and SCF acceleration ratio S_{acc} (Fig. 3a). For several properties, the errors are more than an order of magnitude lower than those of both baselines, indicating that DSLoss aligns Hamiltonian prediction more effectively with downstream physical observables. This advantage becomes increasingly pronounced as molecular size and complexity increase (Fig. 3b). For

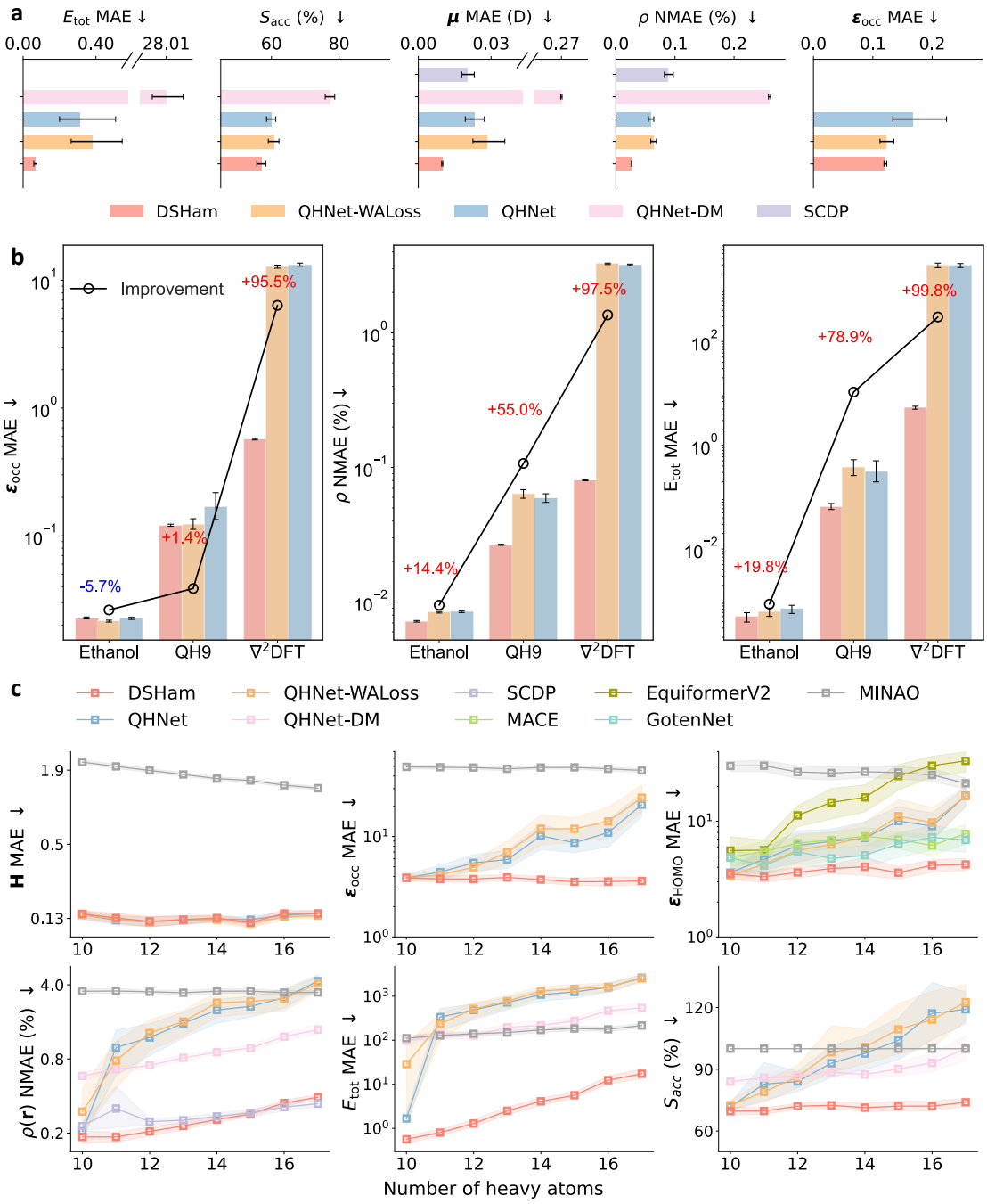


Figure 3: **Performance of DSHam against baseline methods.** Shaded regions and error bars indicate 95% confidence intervals. The units of $\mathbf{H}$, ε_{occ} and E_{tot} are kcal/mol. **(a)** Properties derived from predicted Hamiltonians on QH9. Results for E_{tot} and S_{acc} are evaluated on 50 test structures, whereas all other quantities are evaluated on 13,084 test structures. **(b)** Performance across datasets spanning different molecular scales. E_{tot} is evaluated on 50 structures per dataset, except for ∇^2DFT, which uses 200. All other quantities are evaluated on 4,500, 13,084 and 9,345 test structures from ethanol, QH9 and ∇^2DFT, respectively. **(c)** Extrapolation to larger PubChemQC molecules. All models are trained on QH9 molecules containing no more than nine heavy atoms and evaluated on PubChemQC molecules of increasing size. Each value is computed from 50 PubChemQC molecules.

total energy, DSHam achieves MAEs of 5×10^{-4}, 0.06 and 5.29 kcal/mol on ethanol, QH9 and ∇^2DFT, respectively. Relative to the strongest competing method, these results correspond to error reductions of 19.8%, 78.9% and 99.8% on ethanol, QH9 and ∇^2DFT, respectively. The result on ∇^2DFT is equivalent to a 500-fold error reduction, whereas the total energy MAE of QHNet reaches 2,920 kcal/mol. For

the dipole moment, DSHam similarly reduces the error by approximately 300-fold on ∇^2DFT (Supplementary Table S13). Although WALoss improves orbital energy and total energy prediction for smaller molecules, its advantage largely disappears on the more complex ∇^2DFT systems. Because most of these properties are not directly supervised during training, the broad performance gains indicate that DSLoss addresses a critical bottleneck shared across diverse downstream tasks. Additional results across properties, datasets and split settings are provided in Supplementary Sec. C.

Finally, we compare DSHam with two representative density learning models: QHNet-DM, which directly predicts the density matrix using the QHNet architecture, and SCDP, a leading model for real-space electron density prediction [46]. All methods are evaluated under matched experimental settings. As shown in Fig. 3a, DSHam outperforms both density learning baselines in real-space density prediction and density-derived quantities. Although QHNet-DM achieves competitive density matrix accuracy, this advantage does not translate into more accurate density-related observables such as $\boldsymbol{\mu}$ and E_{tot}. This result indicates that a low element-wise density matrix error alone is insufficient to ensure reliable downstream predictions. Further discussion are provided in Results 2.4 and Supplementary Sec. C.5.

2.3 Extrapolation of DSHam to Larger-Scale Molecules

To wield the advantage of DSHam for a more meaningful impact, we evaluate its extrapolation performance on molecules larger than those seen during training. For this purpose, we train all models on QH9 molecules containing no more than 9 heavy atoms and test them on PubChemQC molecules containing 10–17 heavy atoms. The test molecules are grouped by the number of heavy atoms N_{heavy}, with 50 structures sampled from each group to mitigate statistical bias and enable a reliable comparison across molecular sizes. To isolate the effect of molecular size, we retain only molecules composed of `H/C/N/O` elements and exclude radicals and uncommon isotopes, thereby aligning the elemental composition of the test set with that of QH9 (Methods 4.3, Supplementary Sec. B.3). Reference labels are computed at the same B3LYP/def2-SVP level used for QH9.

The extrapolation results are shown in Fig. 3c, with additional properties reported in Supplementary Sec. C.3. Across all molecular sizes, DSHam consistently achieves substantially lower property errors than QHNet, and the gap grows to several orders of magnitude as N_{heavy} increases, despite the two methods retain similar Hamiltonian MAEs. Notably, although QHNet predicts the residual between the `MINAO` initial guess and the reference Hamiltonian, its predicted Hamiltonians become less effective than the `MINAO` baseline for SCF initialization when $N_{\mathrm{heavy}} > 14$. By contrast, DSHam maintains more robust total energy accuracy and SCF acceleration across larger molecules. These results demonstrate that dual-space supervision facilitates the scalability of Hamiltonian learning and enables more reliable size extrapolation.

The conventional paradigm for molecular property prediction is to train a dedicated end-to-end (`e2e`) model for each target quantity [47, 11, 14, 17, 15, 18, 19]. Hamiltonian learning instead provides a shared electronic structure representation from which many properties can be derived using a single model. To examine whether this advantage persists under size extrapolation, we compare $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$ and ε_{Δ} derived from the predicted Hamiltonians with direct predictions from dedicated `e2e` models. We consider three representative architectures, EquiformerV2, MACE and GotenNet [15, 18, 19]. Each `e2e` model is trained separately for one target property using the same training data as the Hamiltonian predictors (Supplementary Sec. B.2). As shown in Fig. 3c, direct `e2e` predictors exhibit a pronounced generalization gap as molecular size exceeds the training range. With dual-space supervision, DSHam achieves the best performance for all three orbital energy quantities, reducing MAEs by one to two orders of magnitude relative to the `e2e` models. Additional results are provided in Supplementary Table S14. These results highlight Hamiltonian learning, particularly when trained with DSLoss, as a promising route to more reliable out-of-distribution molecular property prediction.

2.4 DSHam Unifies Hamiltonian and Electron Density Learning

As discussed in Results 2.1, dual-space supervision jointly constrains the predicted Hamiltonian and its induced electronic solution with two complementary density-based representations. An alternative way to exploit density information is to predict either representation directly from molecular structure, an approach increasingly explored in recent studies [46, 48–54]. To compare these two learning paradigms, we benchmark DSHam against two representative density-prediction methods, QHNet-DM and SCDP. As discussed in Results 2.2, DSHam outperforms both methods in downstream property prediction on molecules within the training scale. When extrapolated to larger molecules, QHNet-DM and SCDP retain substantially better density accuracy than QHNet and QHNet-WALoss (Fig. 3c), indicating that density labels may be easier to extrapolate than Hamiltonian matrices. This behavior may partly arise from the more bounded numerical range of density targets, which is more amenable to generalization by machine

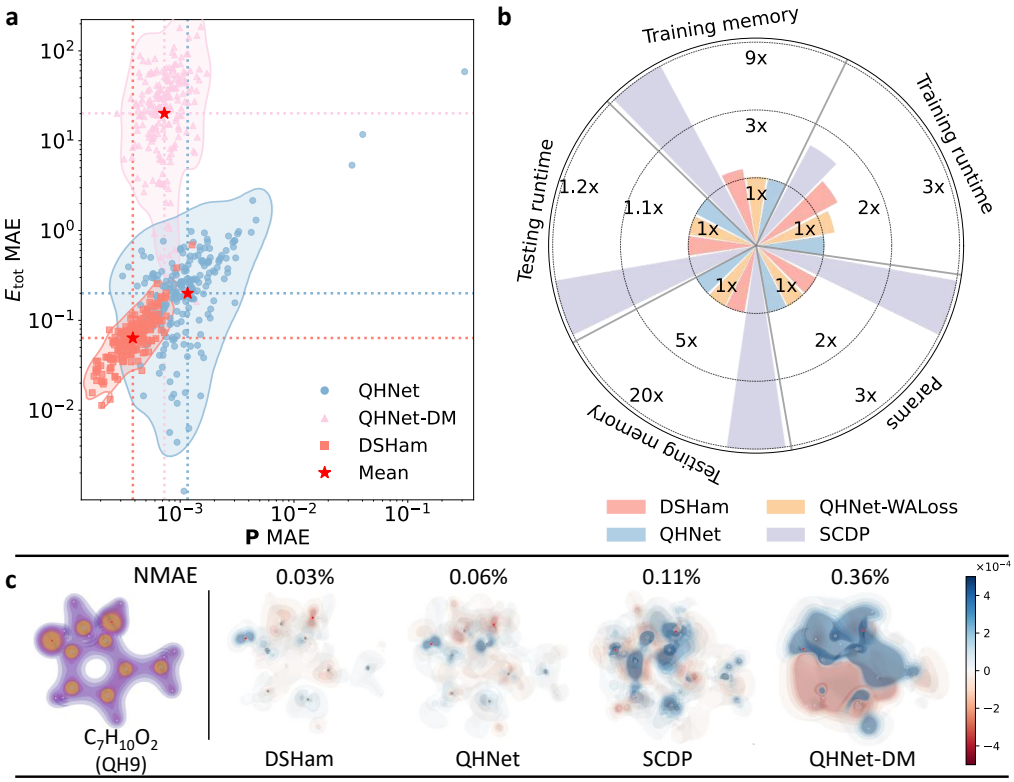


Figure 4: **Comparison of Hamiltonian and electron density learning methods.** **(a)** Scatter plot of density-matrix MAE versus total energy error, evaluated on 200 randomly selected molecules from the QH9 test set. E_{tot} MAE is shown in kcal/mol. **(b)** Computational cost of various methods. Runtime statistics are averaged over 500 training iterations on the QH9 dataset for each model. SCDP is evaluated with a batch size of 1, while the other models use a batch size of 8. **(c)** Visualization of predicted real-space densities for molecules randomly sampled from QH9. The left panels show the ground-truth density, and the right panels show the prediction error, with the NMAE indicated on each prediction.

learning models [55, 56]. However, this advantage in density extrapolation does not translate into better downstream properties, such as total energy and dipole moment (Supplementary Fig. S8).

We next qualitatively compare the predicted electron densities for representative molecules. As shown in Fig. 4c, for the $C_7H_{10}O_2$ molecule from QH9, the virtual nodes introduced by SCDP improve overall density accuracy but produce localized artifacts near the node positions. QHNet-DM exhibits a pronounced spatial bias, with systematic overestimation and underestimation in different molecular regions. By contrast, both Hamiltonian predictors recover densities that more closely resemble the KSDFT reference. Results on larger molecules from PubChemQC and ∇^2DFT are provided in Supplementary Sec. C.5. These results highlight the value of retaining the Hamiltonian as a structured intermediate representation while constraining its induced density through dual-space supervision, thereby linking transferable state information to reliable downstream observables.

To further understand the limitation of direct density prediction, we compare sample-wise density matrix and total energy errors on QH9 test molecules. As shown in Fig. 4a, QHNet-DM achieves lower density matrix errors than QHNet but substantially larger total energy errors, indicating that reducing the density matrix error does not necessarily ensure physical fidelity. One possible explanation is that direct regression does not explicitly enforce essential density matrix constraints, including electron number conservation and generalized idempotency in a non-orthogonal AO basis [57]. These constraints are naturally satisfied when the density matrix is constructed from the occupied eigenspace of a Hamiltonian. By constraining the Hamiltonian through dual-space supervision, DSHam further leverages this advantage thereby achieving superior performance in both $\mathbf{P}$ and E_{tot}. Developing direct density matrix predictors that rigorously preserve these constraints remains an important direction for future work. Direct real-space density regression faces related limitations because unconstrained models do not guarantee nonnegative densities throughout space. Moreover, deriving energies or initializing SCF calculations from a grid density generally requires an additional and potentially ill-posed projection onto an AO basis, limiting the practical utility of purely real-space density predictors such as SCDP.

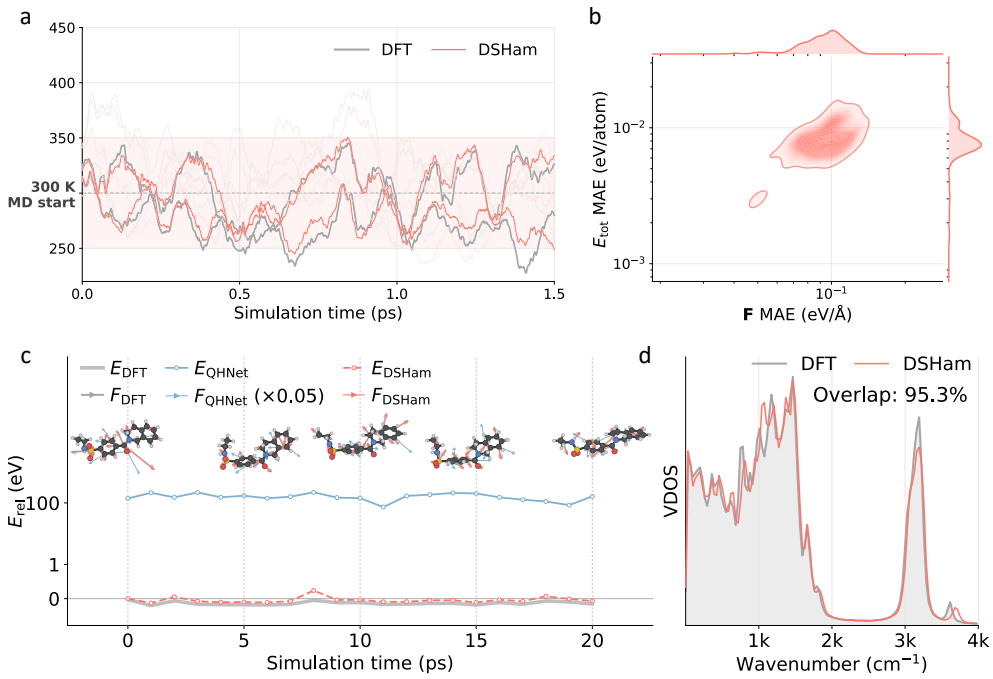


Figure 5: **Molecular dynamics simulations results.** Five randomly selected flexible molecules from the ∇^2DFT test set are used as case studies. **(a)** Temperature profiles of trajectories generated by DFT and DSHam. The shaded region denotes the 250–350 K interval. All five DFT–DSHam trajectory pairs are shown, with two randomly selected pairs highlighted and the remaining three displayed with reduced opacity. **(b)** Energy and force errors of DSHam relative to DFT, evaluated on the same 100 configurations sampled from DFT-generated trajectories. Marginal error distributions are shown along the top and right axes. **(c)** Evaluation results along the DSHam-generated trajectory of $C_{18}H_{20}N_2O_3S$. Representative structures are shown with force vectors evaluated by DSHam, QHNet and DFT on the same geometries. QHNet force vectors are scaled by a factor of 0.05 for visualization due to their substantially larger magnitudes. Relative energies E_{rel} from the three methods are also shown along the trajectory. **(d)** VDOS obtained from trajectories generated by DFT and DSHam for $C_{18}H_{20}N_2O_3S$.

For a comprehensive comparison of the two learning paradigms, we further benchmark computational cost and model size by reporting their relative ratios over QHNet in Fig. 4b. Built on the same backbone, QHNet-WALoss and DSHam introduce only marginal additional training memory consumption. Their per-iteration training runtimes increase by 17% and 35% over QHNet, respectively, representing moderate training-only overheads given the substantial improvements in density, energy and SCF acceleration performance. The additional runtime primarily arises from Hamiltonian diagonalization, with a detailed breakdown provided in Supplementary Sec. B.2.4. Importantly, DSHam retains the same Hamiltonian prediction cost as standard QHNet at inference time because the additional physical objectives are evaluated only during training. On large molecules in ∇^2DFT, Hamiltonian prediction can provide more than a 100-fold computational speedup over DFT under the same hardware configuration. SCDP requires substantially more training memory and runtime even with a smaller batch size, limiting its applicability to larger molecular systems.

2.5 DFT-Consistent Molecular Dynamics with DSHam

Having established the accuracy and extrapolation capability of DSHam, we next examine whether these advantages translate to molecular dynamics (MD) simulations. We use DSHam and QHNet models trained on ∇^2DFT to simulate five flexible test molecules containing 32–44 atoms. All simulations follow an NVT protocol at 300 K with a time step of 0.5 fs (Supplementary Sec. B.4). This setting provides a stringent test of conformational generalization because the ∇^2DFT training set offers only sparse conformational coverage, typically no more than five geometries per molecule, whereas MD continuously explores a much broader region of conformational space. Errors in the derived forces can therefore accumulate during propagation and rapidly destabilize the simulation.

We first assess simulation stability by monitoring temperature fluctuations and structural validity across all simulated trajectories. Structural validity is determined by whether any simulated bond length exceeds

a predefined threshold [58]. The DSHam trajectories remain largely within 250–350 K, comparable to DFT, and fluctuate around the target temperature of 300 K (Fig. 5a), with no invalid structures observed. By contrast, QHNet trajectories rapidly heat to above 10,000 K and collapse into nonphysical configurations. We further evaluate DSHam on configurations sampled from DFT-generated trajectories, thereby testing its accuracy on unseen geometries without confounding effects arising from the divergence of independently propagated trajectories. As shown in Fig. 5b, both energy and force errors remain concentrated in the low-error regime, with MAEs of 0.008 eV/atom and 0.09 eV/Å, respectively. These results demonstrate that DSHam generalizes robustly across a substantially broader conformational space than that sparsely sampled during training.

We next assess whether DSHam maintains DFT-consistent energetics and forces along its own generated trajectory using a 20-ps simulation of $C_{18}H_{20}N_2O_3S$. At sampled configurations, energies and forces are recomputed with DSHam, QHNet and DFT at the same geometries. As shown in Fig. 5c, DSHam closely follows the DFT references throughout the trajectory, yielding approximately 500-fold and 200-fold lower energy and force errors than QHNet, respectively. We also compare the vibrational dynamics of independently generated DFT and DSHam trajectories through their vibrational density of states (VDOS), which provides a spectral characterization of intramolecular vibrational motion [59]. The DSHam VDOS closely reproduces the DFT reference, with a spectral overlap of 95.3% across the frequency range (Fig. 5d). Notably, this trajectory-level fidelity is achieved without force supervision during training, showing that the electronic-structure accuracy of DSHam translates into accurate energy, forces and vibrational dynamics, bridging Hamiltonian prediction with practical molecular simulation.

3 Conclusions

We have developed DSHam, a Hamiltonian learning framework that uses dual-space supervision to bridge the training–application gap between operator prediction and eigensolution-based evaluation while retaining favorable computational efficiency. By leveraging the Hamiltonian-state variational duality, DSHam combines Hamiltonian prediction with direct supervision of its induced electronic state, consistently improving downstream physical properties across molecular benchmarks. These advantages become increasingly pronounced for larger and more challenging systems and persist when extrapolating well beyond the molecular scales encountered during training. DSHam also enables stable, DFT-consistent molecular dynamics and accurately reproduces vibrational dynamics. These results establish dual-space supervision as a physically grounded strategy for improving the reliability and scalability of electronic structure modeling. The general principle may inform other AI-for-science problems in which models predict physical operators but are ultimately evaluated through their induced solutions.

An important limitation of current Hamiltonian learning methods is their dependence on a predefined atomic orbital basis, which restricts direct transfer across basis sets. Using a sufficiently expressive common basis could provide a unified representation across different levels of theory, but would increase both reference calculation costs and learning complexity. Developing representations that can be transferred or systematically adapted across basis sets therefore remains an important direction for future research.

Although DSHam improves size extrapolation, reliable generalization across broader chemical and conformational spaces remains challenging. Extending the framework to more diverse molecular and materials systems, together with longer and more complex dynamical simulations, will help establish its practical scope. Hamiltonian predictors also involve higher-order tensor operations and can therefore be more computationally demanding than machine learning interatomic potentials. Improving model and eigensolver efficiency will be essential for scaling physically consistent Hamiltonian learning to large-scale molecular simulations.

4 Methods

This section provides methodological and experimental details for DSHam. We first describe the design of the Dual-Space loss function (Methods 4.1) and the construction of Hamiltonian prediction models (Methods 4.2). We then present the dataset preparation protocol (Methods 4.3) and evaluation metrics (Methods 4.4) used in the empirical studies. Additional implementation details are provided in Supplementary Sec. B. Atomic units are used throughout the paper unless otherwise specified.

We briefly summarize the main technical components here. Element-wise Hamiltonian supervision is insufficient to constrain the induce ground-state electronic state used for evaluation hence leading to a training–deployment gap. To address this mismatch, we propose the DSLoss and systematically compare it with alternative computationally efficient physical surrogates (Methods 4.1). Because Hamiltonian matrix in an AO basis transforms equivariantly under molecular rotations, we employ an $\mathrm{SE}(3)$-equivariant

architecture for Hamiltonian prediction and introduce several technical treatments for stable and efficient optimization (Methods 4.2).

4.1 Loss Function Design

As discussed in Results 2.1, for a molecular structure $\mathcal{M} := \{\mathcal{Z}, \mathbf{X}\}$, DSHam predicts the reference mean-field Hamiltonian $\mathbf{H}^\star$ through the model output $\hat{\mathbf{H}}$. Given $\hat{\mathbf{H}}$, the molecular orbital coefficient matrix $\hat{\mathbf{C}}$ and orbital energies $\hat{\varepsilon}$ are obtained by solving the generalized Kohn–Sham eigenvalue problem in Eq. (1). The Hamiltonian is represented in an AO basis containing B basis functions, and solving the corresponding B-dimensional eigenvalue problem requires a complete set of B molecular orbitals. For a closed-shell system with N electrons, $K = N/2$ orbitals are occupied, while the remaining $B - K$ unoccupied orbitals complete the single-particle space spanned by the AO basis. The density matrix is then constructed from the occupied orbitals as

$$\hat{\mathbf{P}}_{\alpha\beta} = \sum_{i=1} f_i \hat{\mathbf{C}}_{\alpha i} \hat{\mathbf{C}}_{\beta i},$$

where f_i denotes the occupation number of orbital ϕ_i. In the closed-shell KSDFT setting considered here, $f_i = 2$ for occupied orbitals and $f_i = 0$ for unoccupied orbitals.

Density matrix loss. To align the occupied eigenspace induced by $\hat{\mathbf{H}}$ with that induced by $\mathbf{H}^\star$, we introduce a density-matrix loss defined as the MAE between $\hat{\mathbf{P}}$ and $\mathbf{P}^\star$:

$$\mathcal{L}_{\mathbf{P}} = \frac{1}{B^2} \sum_{\alpha,\beta=1}^{B} \left| \hat{\mathbf{P}}_{\alpha\beta} - \mathbf{P}^\star_{\alpha\beta} \right|.$$

Evaluating $\mathcal{L}_{\mathbf{P}}$ requires differentiating through the eigensolver used to obtain $\hat{\mathbf{C}}$. We implement this step using the differentiable eigensolver provided by PyTorch [60]. However, backpropagation through the eigensolution can lead to numerically unstable gradients when orbital energies are nearly degenerate. We discuss this issue and the stabilization strategy in Methods 4.2 and Supplementary Sec. B.1.3.

Real-space density loss. Although the density-matrix loss directly constrains the electronic state, supervision in the AO basis may not fully capture fine-grained spatial features of the electron density. We therefore introduce a real-space density loss evaluated on the Lebedev-type irregular grid $\mathcal{G} := \{\mathbf{r}_g\}_{g=1}^{N_G}$ around each molecule, providing spatially resolved supervision over the electron distribution [46]. Specifically, we use the NMAE

$$\mathcal{L}_\rho = \frac{\sum_g w_g |\hat{\rho}(\mathbf{r}_g) - \rho^\star(\mathbf{r}_g)|}{\sum_g w_g |\rho^\star(\mathbf{r}_g)|},$$

where w_g denotes the weight of g-th grid point. Compared with $\mathcal{L}_{\mathbf{P}}$, which is defined in a basis-dependent representation, $\mathcal{L}_\rho$ is invariant to global rotations and translations of the molecule when evaluated on consistently transformed real-space grids. This invariant spatial supervision helps reduce unnecessary gradient variance and provides a complementary constraint to the AO-basis density-matrix loss.

Orbital-energy loss. To further constrain the predicted Hamiltonian at the spectral level, we introduce an orbital-energy loss defined as the MAE between K occupied orbital energies derived from $\hat{\mathbf{H}}$ and those derived from $\mathbf{H}^\star$:

$$\mathcal{L}_{\boldsymbol{\varepsilon}} = \frac{1}{K} \sum_{i=1}^{K} |\hat{\varepsilon}_i - \varepsilon_i^\star|. \tag{3}$$

This implementation differs from WALoss [36], which evaluates orbital-energy discrepancies after projecting the predicted Hamiltonian onto the ground-truth eigenvectors $\mathbf{C}^\star$:

$$\mathcal{L}_{\mathrm{WA}} = \lambda_1 \sum_{i \in \mathcal{I}} \|(\mathbf{C}_i^\star)^\top \hat{\mathbf{H}} \mathbf{C}_i^\star - \varepsilon_i^\star\|_{\mathrm{F}}^2 + \lambda_2 \sum_{i \notin \mathcal{I}} \|(\mathbf{C}_i^\star)^\top \hat{\mathbf{H}} \mathbf{C}_i^\star - \varepsilon_i^\star\|_{\mathrm{F}}^2,$$

where $\mathcal{I}$ denotes the set of orbital indices below the lowest unoccupied molecular orbital (LUMO), and λ_1 and λ_2 are tunable weights with $\lambda_1 \gg \lambda_2$. In preliminary experiments, $\mathcal{L}_{\boldsymbol{\varepsilon}}$ consistently outperformed $\mathcal{L}_{\mathrm{WA}}$ across several benchmarks (Supplementary Table S17). We therefore adopt Eq. (3) for eigenvalue-level supervision in DSLoss.

One-electron energy losses. Ref. [37] showed that the total energy E_{tot} can provide informative supervision for Hamiltonian learning. However, evaluating E_{tot} from the predicted Hamiltonian requires computationally expensive integrals and matrix operations, which introduces substantial training overhead. We therefore consider several computationally efficient one-electron energy terms as alternative physical surrogates, including the non-interacting kinetic energy T_{S}, the external-potential energy E_{ext}, and their sum, the one-electron energy $E_{\text{core}} := T_{\text{S}} + E_{\text{ext}}$. These quantities can be computed directly from the density matrix and the corresponding AO integral matrices.

The non-interacting kinetic energy can be expressed as $T_{\text{S}} = \sum_i \langle \phi_i | \hat{T} | \phi_i \rangle$, where $\hat{T} := -\frac{1}{2}\nabla^2$ is the kinetic energy operator. In the AO basis, the kinetic-energy integral matrix is

$$\mathbf{T}_{\alpha\beta} = \int \eta_\alpha(\mathbf{r})(-\frac{1}{2}\nabla^2)\eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r}.$$

Thus, the kinetic energy can be evaluated as

$$T_{\text{S}} = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\mathbf{T}_{\alpha\beta}. \tag{4}$$

Similarly, the external-potential integral matrix is

$$(\mathbf{V}_{\text{ext}})_{\alpha\beta} = -\int \eta_\alpha(\mathbf{r})\Big(\sum_a \frac{Z^{(a)}}{\|\mathbf{r} - \mathbf{x}^{(a)}\|}\Big)\eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r},$$

where $Z^{(a)}$ and $\mathbf{x}^{(a)}$ denote the atomic number and Cartesian coordinates of atom a, respectively. The external-potential energy is then computed as

$$E_{\text{ext}} = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}(\mathbf{V}_{\text{ext}})_{\alpha\beta}. \tag{5}$$

The one-electron energy is defined as

$$E_{\text{core}} = T_{\text{S}} + E_{\text{ext}} = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}(\mathbf{T}_{\alpha\beta} + (\mathbf{V}_{\text{ext}})_{\alpha\beta}). \tag{6}$$

Dipole-moment loss. The dipole moment $\boldsymbol{\mu}$ is a fundamental molecular property that characterizes the response of an electron distribution to an external electric field. Using the electronic density, the electronic contribution to the dipole moment can be written as

$$\boldsymbol{\mu}_{\text{elec}} = -\int \mathbf{r}\rho(\mathbf{r})d\mathbf{r} = -\sum_{\alpha\beta} \mathbf{P}_{\alpha\beta} \int \eta_\alpha(\mathbf{r})\, \mathbf{r}\, \eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r}.$$

The total molecular dipole moment additionally includes the nuclear contribution,

$$\boldsymbol{\mu} = \sum_a^A Z^{(a)}\mathbf{x}^{(a)} + \boldsymbol{\mu}_{\text{elec}}. \tag{7}$$

We define the dipole-moment loss as the component-wise MAE between the predicted and reference dipole moments, $\mathcal{L}_{\boldsymbol{\mu}} = \sum_{y=1}^{3} |\hat{\boldsymbol{\mu}}_y - \boldsymbol{\mu}_y^\star|$.

As discussed in Results 2.1, we systematically evaluate different physical surrogates $\mathbf{O}$ as auxiliary training losses:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\mathbf{H}} + \lambda_{\mathbf{O}}\mathcal{L}_{\mathbf{O}},$$

where $\lambda_{\mathbf{O}}$ is tuned separately for each observable according to overall downstream performance. All models in this comparison are trained on QH9. The downstream-property rankings are summarized in Fig. 2a, with additional quantitative results provided in Supplementary Table S17. More technical details and discussions are provided in Supplementary Sec. A and Supplementary Sec. B.1.

4.2 Model Specification

Geometric equivariant neural networks. The Hamiltonian matrix in an AO basis transforms systematically under geometric transformations of the molecular structure. Translations leave the Hamiltonian matrix unchanged, whereas rotations induce a corresponding change of basis determined by the angular components of the atomic orbitals. This transformation property motivates the use of $\mathrm{SE}(3)$-equivariant

neural networks for Hamiltonian prediction, allowing the model to respect the geometric symmetry of the electronic-structure problem and thereby improve data efficiency and training stability. Formally, a function $h(\cdot)$ is SE(3)-equivariant if, for any group element $g \in \mathrm{SE}(3)$,

$$h(g \cdot \mathbf{x}) = \mathbf{D}(g)h(\mathbf{x}),$$

where $\mathbf{x}$ denotes the input geometric features and $\mathbf{D}(g)$ is the representation of the group action on the output. For the Hamiltonian matrix, rotational equivariance can be expressed as

$$\mathbf{H}_{\mathcal{M}'} = \mathbf{D}^{\top}(\mathcal{R})\mathbf{H}_{\mathcal{M}}\mathbf{D}(\mathcal{R}),$$

where $\mathcal{M}' = \{\mathcal{Z}, \mathcal{R}\mathbf{X}\}$ ($\mathcal{R}$ is arbitrary rotation action) and $\mathbf{D}(\mathcal{R})$ denotes the rotation representation acting on the AO space. Many SE(3)-equivariant graph neural networks have been developed for Hamiltonian learning by explicitly preserving rotation, translation and permutation symmetries [29–31, 33, 40]. A common building block in these models is the Clebsch–Gordan tensor product, which combines two irreducible representations u^{ℓ_1} and v^{ℓ_2} into an output representation of order ℓ_3:

$$\left(u^{\ell_1} \otimes v^{\ell_2}\right)^{\ell_3}_{m_3} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} (\mathbf{M}_{\mathrm{CG}})^{(\ell_3,m_3)}_{(\ell_1,m_1),(\ell_2,m_2)} u^{\ell_1}_{m_1} v^{\ell_2}_{m_2},$$

where $|\ell_1 - \ell_2| \le \ell_3 \le \ell_1 + \ell_2$, and m indexes the components of an irreducible representation with $-\ell \le m \le \ell$. Using these tensor products, atomic features $\{\mathbf{h}_a\}_{a=1}^{A}$ can be updated through equivariant message passing:

$$\tilde{\mathbf{h}}^{\ell_o}_a = \sum_{b \in \mathcal{N}_a} \sum_{\ell_i,\ell_f} \mathbf{W}_{\ell_i,\ell_f,\ell_o} \left(\mathbf{h}^{\ell_i}_b \otimes Y^{\ell_f}(\hat{\mathbf{r}}_{ab})\right)^{l_o},$$

where $\tilde{\mathbf{h}}^{\ell_o}_a$ is the updated feature of atom a at rotation order ℓ_o, $\mathbf{W}_{\ell_i,\ell_f,\ell_o}$ denotes learnable weights, and $Y^{\ell_f}(\hat{\mathbf{r}}_{ab})$ is the spherical-harmonic filter constructed from the normalized relative position $\hat{\mathbf{r}}_{ab} := \mathbf{r}_{ab}/\|\mathbf{r}_{ab}\|_2$. Here, ℓ_i, ℓ_f and ℓ_o denote the rotation orders of the input feature, geometric filter and output feature, respectively, and $\mathcal{N}_a$ is the neighbor set of atom a. Further details on the construction of SE(3)-equivariant neural networks can be found in refs. [29, 30].

Note that DSLoss is architecture-agnostic by design. We therefore adopt the widely used QHNet [30] architecture as the main backbone for empirical evaluation. Additional results using the SPHNet [40] architecture are reported in Supplementary Table S21.

Given the learned atomic representations, QHNet first constructs pairwise irreducible representations $\mathbf{f}_{ab}$ and then predicts the Hamiltonian matrix block by block through tensor expansion and filtering operations. The tensor expansion is defined as

$$\left(\bar{\otimes} w^{\ell_3}\right)^{(\ell_1,\ell_2)}_{(m_1,m_2)} = \sum_{m_3=-\ell_3}^{\ell_3} (\mathbf{M}_{\mathrm{CG}})^{(\ell_1,m_1),(\ell_2,m_2)}_{(\ell_3,m_3)} w^{\ell_3}_{m_3},$$

where $\bar{\otimes}$ denotes the tensor expansion operation. Each pair block $\mathbf{H}_{ab}$ consists of tensor components with different angular (rotation) orders and channels. For example, in the considered MD17 dataset, both C and O atoms contain 1s, 2s, 3s, 2p, 3p and 3d basis functions (Supplementary Sec. B.2.2). As a result, a pair block C−O can contain multiple channels for the same angular-order pair, such as nine channels with $(\ell_{o_1}, \ell_{o_2}) = (0, 0)$ and four channels with $(\ell_{o_1}, \ell_{o_2}) = (1, 1)$. In practice, QHNet introduces an atomic-number-dependent filter to generate weights for each valid path $(\ell_{o_1}, \ell_{o_2}, \ell_i)$, denoted as $\mathbf{u}^{(\ell_{o_1},\ell_{o_2},\ell_i)}_{ab} = \mathcal{U}(\mathcal{Z}_a, \mathcal{Z}_b, l_{o_1}, l_{o_2}, l_i))$. The corresponding Hamiltonian block is then constructed as

$$\hat{\mathbf{H}}^{(\ell_{o_1},\ell_{o_2})}_{ab,\zeta} = \sum_{\ell_i,\zeta'} \mathbf{u}^{(\ell_{o_1},\ell_{o_2},\ell_i)}_{ab,\zeta\zeta'} \bar{\otimes} \mathbf{f}^{\ell_i}_{ab,\zeta'},$$

where ζ' and ℓ_i denote the channel index and the angular-momentum order of the input features, respectively; ζ denotes the channel index of the output features. Additional model configuration details are provided in Supplementary Sec. B.2.1.

Gradient stabilization. As discussed in Results 2.1, DSLoss contains terms derived from the orbital coefficient matrix $\mathbf{C}$, and therefore requires backpropagation through the generalized eigenvalue problem in Eq. (1). Consequently, the gradients of $L_{\mathbf{P}}$ and L_ρ with respect to the Hamiltonian matrix involve the Jacobian $\nabla_{\mathbf{H}}\mathbf{P}$, which can be written as

$$\frac{\partial \mathbf{P}_{\alpha\beta}}{\partial \mathbf{H}_{\gamma\delta}} = \sum_{ij} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j} \mathbf{C}_{\alpha i}\mathbf{C}_{\beta j}\mathbf{C}_{\gamma i}\mathbf{C}_{\delta j} \tag{8}$$

where ε_i is the orbital energy of the i-th orbital and $f(\varepsilon_i)$ denotes its occupation number. For the derivation of the gradients, see Supplementary Sec. B.1.3. The denominator in Eq. (8) indicates that the gradient depends on inverse orbital-energy gaps. Equivalently, this dependence can be summarized by the matrix

$$\mathbf{G}_{ij} = \begin{cases} 1/(\varepsilon_i - \varepsilon_j), & \text{if } i \neq j \text{ and } f_i \neq f_j, \\ 0, & \text{else,} \end{cases}$$

so near-degenerate orbital energies with $|\varepsilon_i - \varepsilon_j| \approx 0$ can produce numerically unstable gradients. To mitigate this instability, we apply gradient truncation to $\nabla_{\mathbf{H}}\mathbf{P}$ during backpropagation [38]. Specifically, we replace $\mathbf{G}$ with a truncated matrix $\tilde{\mathbf{G}}$:

$$\tilde{\mathbf{G}}_{ij} = \begin{cases} \tau \cdot \mathrm{sgn}(\varepsilon_\mathrm{i} - \varepsilon_\mathrm{j}), & \text{if } \left|\frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\right| > \tau, \\ \mathbf{G}_{ij}, & \text{if } \left|\frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\right| \leq \tau. \end{cases}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function. The threshold τ is set to the 60th percentile of the absolute values of all entries in $\mathbf{G}$. In addition, if exact eigenvalue degeneracy occurs within a batch, we skip the parameter update for that batch to avoid ill-defined eigensolver gradients.

Warm-up initialization. The evaluation of DSLoss requires solving the generalized eigenvalue problem in Eq. (1). At the early stage of training, a randomly initialized Hamiltonian may yield unphysical eigenvalues or unstable eigenvectors, making DSLoss difficult to optimize. To improve training stability, we follow previous work [40, 38] and use the `MINAO` initialization [61] as a baseline electronic-structure guess, with the model learning a residual correction on top of it. The `MINAO` density matrix $\mathbf{P}^{\mathrm{minao}}$ is obtained directly from PySCF [61] and converted into an initial Hamiltonian guess for prediction. This residual learning strategy is applied for all models considered in this paper.

Numerical precision. When downstream properties are derived from the predicted Hamiltonian, numerical errors from the eigensolver and subsequent matrix operations can propagate into the target quantities and affect the quality of supervision. This effect is particularly pronounced for small molecules, where numerical errors can become comparable in magnitude to the model prediction errors; see Supplementary Table S19 in Supplementary Sec. C.6 for further discussion. To reduce precision-induced errors, we use double precision (`FP64`) for experiments on the MD17 dataset and single precision (`FP32`) for the other datasets. For all Hamiltonian models, downstream properties and the SCF acceleration ratio are computed in `FP64` during evaluation.

4.3 Dataset Preparation

To demonstrate the effectiveness of DSHam, we conduct experiments on MD17 and QH9, two representative benchmarks widely used to assess the generalization of Hamiltonian learning methods in conformational and chemical spaces, respectively. To evaluate scalability and out-of-scale extrapolation, we adopt the ∇^2DFT and PubChemQC datasets. For comparisons with electron-density learning methods, we use the MD and QM9 datasets, where QH9 is constructed on molecular structures from QM9. For each dataset, unless otherwise noted, we follow the default split used in the original work to divide molecular structures into training, validation, and test sets; detailed split statistics are provided in Supplementary Table S8. Supplementary information and data generation details are provided in Supplementary Sec. B.3.

MD17. To evaluate conformational-space generalization, we adopt the MD17 dataset [41, 21] and focus on four molecules: water (H_2O), ethanol (C_2H_5OH), malondialdehyde ($CH_2(CHO)_2$) and uracil ($C_4H_4N_2O_2$). Hamiltonian matrices are computed using the PBE [5] exchange–correlation (XC) functional with the def2-SVP basis set [62].

MD. The MD dataset is used to benchmark electron-density prediction. It contains molecular conformations of six small molecules: ethanol (C_2H_5OH), benzene (C_6H_6), phenol (C_6H_5OH), resorcinol ($C_6H_4(OH)_2$), ethane (C_2H_6) and malondialdehyde ($CH_2(CHO)_2$). The first four molecules are taken from ref. [63], with 1,000 sampled geometries per molecule, while ethane and malondialdehyde are taken from ref. [48], with 2,000 sampled geometries per molecule. Models are trained separately for each molecule. Because the original datasets provide only density labels computed at the PBE level with a plane-wave basis set, we regenerate Hamiltonian labels using PySCF with the PBE functional and the def2-SVP basis set for training Hamiltonian models. The electron density is evaluated on the Lebedev-type numerical integration grid used in PySCF, with the grid level set to 3. We further investigate the sensitivity of model performance to the grid level in Supplementary Sec. C.6.

QH9. QM9 is a widely used benchmark for predicting quantum-chemical properties with deep learning methods [64]. It contains 130,831 equilibrium geometries of small organic molecules with up to nine heavy atoms, composed of `H/C/N/O/F` elements. For benchmarking Hamiltonian learning methods, ref. [43] regenerated Hamiltonian matrices for these molecules at the B3LYP/def2-SVP level [65]. The resulting dataset, `QH9-stable`, is used to evaluate generalization across chemical space. We use the `stable-random` split for the main results reported in Results 2, and provide additional results on other split settings in Supplementary Sec. C.2. Notably, the `stable-random` split is by default adopted for experimental results on QH9 unless specified. To further evaluate conformational space generalization, QH9 provides Hamiltonian matrices for 2,998 molecular dynamics trajectories generated from QM9 molecules, forming the `QH9-dynamic` dataset. For experiments involving electron density prediction, real-space densities are evaluated on the Lebedev-type numerical integration grid used in PySCF, with the grid level set to 3. For the largest SCDP model, the grid level is set to 2 to avoid GPU memory overflow.

∇^2DFT. To evaluate the scalability of Hamiltonian learning on larger and more diverse molecular systems, we use the ∇^2DFT dataset [44, 45]. This dataset contains 12,676,264 conformations of drug-like molecules with heavy-atom counts ranging from 8 to 27 and includes multiple chemical elements, including `H/C/N/O/F/S/Cl/Br`. Hamiltonian matrices and other atomic and molecular properties were computed at the ωB97X-D/def2-SVP level [66]. The dataset is divided into a training split and three test splits. Both the training split and the conformational test split contain five progressively larger subsets. In this work, we train models on the medium-sized training subset, which includes 9,689 unique molecules with 49,725 conformations. Evaluation is performed on the corresponding medium-sized conformational test split, which contains unseen geometries of molecules present in the training subset.

PubChemQC. To evaluate extrapolation beyond the molecular scale observed during training, we construct an out-of-distribution benchmark based on PubChemQC [67]. We select molecules composed of `H/C/N/O` elements with heavy-atom counts ranging from 10 to 17. For each heavy-atom count, 50 molecules are randomly sampled to mitigate statistical bias and enable a reliable comparison across molecular sizes. Statistical information is provided in Supplementary Fig. S6. Open-shell systems, radicals and molecules containing uncommon isotopes are excluded. Reference Hamiltonian matrices are computed using PySCF at the B3LYP/def2-SVP level, consistent with the setting used for QH9.

4.4 Evaluation Metrics

To compare DSHam with baseline methods for Hamiltonian and electron density prediction, we evaluate three categories of physical quantities in addition to the MAE over all Hamiltonian matrix elements. For conciseness, Results 2 reports only representative metrics, while complete results are provided in Supplementary Sec. C. Additional technical background is provided in Supplementary Sec. A, and a comprehensive description of all metrics is given in Supplementary Table S10.

Orbital-energy quantities. To assess spectral accuracy, we report MAEs relative to KSDFT reference for four orbital-energy quantities: **(1)** occupied orbital energies $\boldsymbol{\varepsilon}_{\mathrm{occ}}$, which measure the overall quality of the occupied electronic spectrum; **(2)** the highest occupied molecular orbital (HOMO) energy $\boldsymbol{\varepsilon}_{\mathrm{HOMO}}$, which is closely related to chemical reactivity and molecular stability; **(3)** the lowest unoccupied molecular orbital (LUMO) energy $\boldsymbol{\varepsilon}_{\mathrm{LUMO}}$, which is relevant to electron affinity and chemical response; and **(4)** the HOMO–LUMO gap $\boldsymbol{\varepsilon}_{\Delta}$, which characterizes frontier-orbital separation and is associated with electronic excitation and transport-related properties.

Electron-density quantities. To assess the accuracy of the predicted electronic state, we report errors relative to KSDFT references for the density matrix $\mathbf{P}$ and real-space electron density ρ. The density matrix $\mathbf{P}$ is evaluated using MAE, while the real-space density ρ is evaluated using NMAE. We additionally report the cosine similarity between predicted and reference molecular-orbital coefficient matrices $\mathbf{C}$ to quantify agreement at the eigenvector level. Because molecular orbitals are non-unique in degenerate subspaces and highly sensitive to perturbations in near-degenerate subspaces, cosine similarity of $\mathbf{C}$ is reported only as a supplementary diagnostic; density-based metrics are used as the primary measures of electronic-state accuracy.

Molecular observables. To evaluate how well the predicted Hamiltonian supports downstream physical prediction, we report MAEs relative to KSDFT references for several molecular observables, including total energy E_{tot} (Eq. (S8)), kinetic energy T_{S} (Eq. (4)), external-potential energy E_{ext} (Eq. (5)), one-electron energy E_{core} (Eq. (6)), the LDA-based density functional quantity E_{LDA}, and dipole moment $\boldsymbol{\mu}$. We also assess the practical utility of the predicted Hamiltonian for accelerating DFT calculations by measuring the relative SCF iteration cost required for convergence. Specifically, we define

$$S_{\mathrm{acc}} := \frac{\mathcal{T}_{\hat{\mathbf{P}}}}{\mathcal{T}_{\mathrm{minao}}} \times 100\%,$$

where $\mathcal{T}_{\hat{\mathbf{P}}}$ and $\mathcal{T}_{\mathrm{minao}}$ denote the number of SCF iterations required when using the prediction-derived density matrix $\hat{\mathbf{P}}$ and the standard `MINAO` initialization, respectively. A smaller S_{acc} indicates stronger SCF acceleration. The LDA-based density functional quantity is computed as

$$E_{\mathrm{LDA}} = \int C_{\mathrm{F}} \rho(\mathbf{r})^{5/3} \mathrm{d}\mathbf{r},$$

where C_{F} is a constant. This metric is evaluated only on PubChemQC when comparing with real-space density predictors.

Code Availability

The source code for implementing the proposed method, trained neural network model checkpoints, and evaluation datasets are available in an anonymous Zenodo repository. The source-code archive also includes usage instructions, train/test splits, and a sample PySCF input file to facilitate reproducibility.

References

[1] Robert O Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of modern physics*, 87(3):897–923, 2015. 1

[2] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013. 1

[3] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 1964. 1, A.1

[4] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965. 1, A.2

[5] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996. 1, 4.3, A.1

[6] Pragya Verma and Donald G Truhlar. Status and challenges of density functional theory. *Trends in Chemistry*, 2(4):302–318, 2020. 1

[7] Kieron Burke. Perspective on density functional theory. *The Journal of chemical physics*, 136(15), 2012.

[8] Andrew M Teale, Trygve Helgaker, Andreas Savin, Carlo Adamo, Bálint Aradi, Alexei V Arbuznikov, Paul W Ayers, Evert Jan Baerends, Vincenzo Barone, Patrizia Calaminici, et al. DFT exchange: sharing perspectives on the workhorse of quantum chemistry and materials science. *Physical chemistry chemical physics*, 24(47):28700–28781, 2022. 1

[9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017. 1

[10] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet-a deep learning architecture for molecules and materials. *The Journal of chemical physics*, 148(24), 2018.

[11] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021. 2.3

[12] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in neural information processing systems*, 34:6790–6802, 2021.

[13] 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.

[14] Yilun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. In *International Conference on Learning Representations*, 2023. 2.3

[15] Yilun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, pages 39282–39309, 2024. 2.3, B.2.1

[16] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

[17] P Thölke and GD Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. 2.3

[18] Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022. 2.3, B.2.1

[19] Sarp Aykent and Tian Xia. GotenNet: Rethinking efficient 3D equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2.3, B.2.1

[20] Ganesh Hegde and R Chris Bowen. Machine-learned approximations to density functional theory Hamiltonians. *Scientific reports*, 7(1):42669, 2017. 1

[21] Kristof T Schütt, Michael Gastegger, Alexandre Tkatchenko, K-R Müller, and Reinhard J Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):5024, 2019. 1, 4.3

[22] Michael Gastegger, Adam McSloy, Michael Luya, Kristof T Schütt, and Reinhard J Maurer. A deep neural network for molecular wave functions in quasi-atomic minimal basis representation. *The Journal of Chemical Physics*, 153(4), 2020.

[23] 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. 1

[24] 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. 1

[25] Yuxiang Wang, He Li, Zechen Tang, Honggeng Tao, Yanzhen Wang, Zilong Yuan, Zezhou Chen, Wenhui Duan, and Yong Xu. DeepH-2: Enhancing deep-learning electronic structure via an equivariant local-coordinate transformer. *arXiv preprint arXiv:2401.17015*, 2024.

[26] Haiyang Yu, Yuchao Lin, Xuan Zhang, Xiaofeng Qian, and Shuiwang Ji. Efficient prediction of SO(3)-equivariant Hamiltonian matrices via SO(2) local frames. *arXiv preprint arXiv:2506.09398*, 2025.

[27] Seongsu Kim, Nayoung Kim, Dongwoo Kim, and Sungsoo Ahn. High-order equivariant flow matching for density functional theory Hamiltonian prediction. *Advances in Neural Information Processing Systems*, 38:13265–13307, 2026.

[28] YinZhangHao Zhou, Zixi Gan, Shishir Pandey, Linfeng Zhang, and Qiangqiang Gu. Learning local equivariant representations for quantum operators. In *International Conference on Learning Representations*, pages 50307–50326, 2025. 1

[29] Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *Advances in Neural Information Processing Systems*, 34:14434–14447, 2021. 1, 4.2

[30] Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum Hamiltonian. In *International Conference on Machine Learning*, pages 40412–40424. PMLR, 2023. 1, 2.1, 2.2, 4.2, B.2.1

[31] Shi Yin, Xinyang Pan, Xudong Zhu, Tianyu Gao, Haochong Zhang, Feng Wu, and Lixin He. Towards harmonization of SO(3)-equivariance and expressiveness: a hybrid deep learning framework for electronic-structure Hamiltonian prediction. *Machine Learning: Science and Technology*, 5(4): 045038, 2024. 4.2

[32] 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.

[33] 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. 4.2

[34] 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.

[35] Chen Qian, Valdas Vitartas, James R Kermode, and Reinhard J Maurer. Equivariant electronic Hamiltonian prediction with many-body message passing. *npj Computational Materials*, 2026. 1

[36] Yunyang Li, Zaishuo Xia, Lin Huang, Xinran Wei, Samuel Harshe, Han Yang, Erpai Luo, Zun Wang, Jia Zhang, Chang Liu, et al. Enhancing the scalability and applicability of Kohn-Sham Hamiltonians for molecular systems. In *International Conference on Learning Representations*, pages 918–962, 2025. 1, 2.1, 4.1

[37] 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. 1, 4.1

[38] 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 *International Conference on Machine Learning*, pages 59329–59357. PMLR, 2024. 1, 4.2, 4.2

[39] Elliott H Lieb. Density functionals for coulomb systems. *International journal of quantum chemistry*, 24(3):243–277, 1983. 1, A.3

[40] Erpai Luo, Xinran Wei, Lin Huang, Yunyang Li, Han Yang, Zaishuo Xia, Zun Wang, Chang Liu, Bin Shao, and Jia Zhang. Efficient and scalable density functional theory Hamiltonian prediction through adaptive sparsity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 41368–41390. PMLR, 2025. 2.1, 4.2, 4.2, B.2.1

[41] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 2.2, 4.3, B.3

[42] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications*, 9(1): 3887, 2018. 2.2

[43] Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. QH9: A quantum Hamiltonian prediction benchmark for QM9 molecules. *Advances in Neural Information Processing Systems*, 36:40487–40503, 2023. 2.2, 4.3, B.1.2, B.3

[44] Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, et al. nabladft: Large-scale conformational energy and Hamiltonian prediction benchmark and dataset. *Physical Chemistry Chemical Physics*, 24(42):25853–25863, 2022. 2.2, 4.3, B.3

[45] Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, et al. ∇^2DFT: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. *Advances in Neural Information Processing Systems*, 37:36869–36889, 2024. 2.2, 4.3

[46] Xiang Fu, Andrew Rosen, Kyle Bystrom, Rui Wang, Albert Musaelian, Boris Kozinsky, Tess Smidt, and Tommi Jaakkola. A recipe for charge density prediction. *Advances in Neural Information Processing Systems*, 37:9727–9752, 2024. 2.2, 2.4, 4.1, B.2.1, C.5

[47] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical Review Letters*, 120 (14):143001, 2018. 2.3

[48] Felix Brockherde, Leslie Vogt, Li Li, Mark E Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the Kohn-Sham equations with machine learning. *Nature Communications*, 8(1):872, 2017. 2.4, 4.3

[49] Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8(1):183, 2022.

[50] Sheng Gong, Tian Xie, Taishan Zhu, Shuo Wang, Eric R Fadel, Yawei Li, and Jeffrey C Grossman. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Physical Review B*, 100(18):184103, 2019.

[51] Teddy Koker, Keegan Quigley, Eric Taw, Kevin Tibbetts, and Lin Li. Higher-order equivariant neural networks for charge density prediction in materials. *npj Computational Materials*, 10(1):161, 2024.

[52] Alberto Fabrizio, Andrea Grisafi, Benjamin Meyer, Michele Ceriotti, and Clemence Corminboeuf. Electron density learning of non-covalent systems. *Chemical science*, 10(41):9424–9432, 2019.

[53] Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller III. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31):e2205221119, 2022.

[54] Chaoran Cheng and Jian Peng. Equivariant neural operator learning with graphon convolution. *Advances in Neural Information Processing Systems*, 36:61960–61984, 2023. 2.4

[55] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002. 2.4

[56] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016. 2.4

[57] Trygve Helgaker, Helena Larsen, Jeppe Olsen, and Poul Jørgensen. Direct optimization of the ao density matrix in Hartree-Fock and Kohn-Sham theories. *Chemical Physics Letters*, 327(5-6):397–403, 2000. 2.4

[58] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gómez-Bombarelli, and T. Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023, 2022. 2.5, B.4

[59] Christoph Schran, Fabian L Thiemann, Patrick Rowe, Erich A Müller, Ondrej Marsalek, and Angelos Michaelides. Machine learning potentials for complex aqueous systems made simple. *Proceedings of the National Academy of Sciences*, 118(38):e2110077118, 2021. 2.5

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 4.1, B.2.1

[61] Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: the python-based simulations of chemistry framework. *WIREs Computational Molecular Science*, 8(1):e1340, 2018. 4.2, A.2, B.5

[62] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. 4.3

[63] Mihail Bogojeski, Leslie Vogt-Maranto, Mark E Tuckerman, Klaus-Robert Müller, and Kieron Burke. Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications*, 11(1):5223, 2020. 4.3

[64] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014. 4.3

[65] Axel D Becke. Density-functional thermochemistry. i. the effect of the exchange-only gradient correction. *The Journal of chemical physics*, 96(3):2155–2160, 1992. 4.3, A.1

[66] Jeng-Da Chai and Martin Head-Gordon. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of chemical physics*, 128(8), 2008. 4.3, A.1

[67] Maho Nakata and Toshiyuki Maeda. PubChemQC B3LYP/6-31G*//PM6 data set: The electronic structures of 86 million molecules using B3LYP/6-31G* calculations. *Journal of Chemical Information and Modeling*, 63(18):5734–5754, 2023. 4.3

[68] Llewellyn H Thomas. The calculation of atomic fields. In *Mathematical proceedings of the Cambridge philosophical society*, volume 23, pages 542–548. Cambridge University Press, 1927. A.1

[69] Yan Alexander Wang and Emily A Carter. Orbital-free kinetic-energy density functional theory. *Theoretical methods in condensed phase chemistry*, 5:117–184, 2000. A.1

[70] Sason Shaik, David Danovich, and Philippe C Hiberty. *A chemist's guide to Valence Bond Theory: Insights into chemical bonding, reactivity, and excited states*. John Wiley & Sons, 2026. B.1.2

[71] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. B.1.3

[72] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. B.2.1

[73] Saro Passaro and C Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant gnns. In *International Conference on Machine Learning*, pages 27420–27438. PMLR, 2023. B.2.1

[74] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. B.4

[75] John C Duda, Timothy S English, Edward S Piekos, William A Soffa, Leonid V Zhigilei, and Patrick E Hopkins. Implications of cross-species interactions on the temperature dependence of kapitza conductance. *Physical Review B—Condensed Matter and Materials Physics*, 84(19): 193301, 2011. B.4

[76] Ashutosh Giri, Connor J Dionne, and Patrick E Hopkins. Atomic coordination dictates vibrational characteristics and thermal conductivity in amorphous carbon. *npj Computational Materials*, 8(1): 55, 2022. B.4

[77] Peter Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. B.4

Appendix

A Mechanism of Density Functional Theory

A.1 Formulation of Density Functional Theory

Quantum chemistry provides a fundamental framework for computing molecular properties by describing the interactions among electrons and nuclei. Because nuclei are much heavier than electrons, molecular electronic-structure calculations commonly adopt the Born–Oppenheimer approximation, in which the nuclei are treated as fixed classical particles and the electronic problem is solved for a given nuclear configuration. For a molecular structure with A nuclei and N electrons, the nuclear configuration is specified by the atomic numbers $\mathcal{Z} := \{Z_a\}_{a=1}^{A}$ and nuclear coordinates $\mathbf{X} := \{\mathbf{x}_a\}_{a=1}^{A}$, while the electronic state is described by an antisymmetric many-electron wavefunction $\psi(\mathbf{r}_1, \cdots, \mathbf{r}_N)$. Under the Born–Oppenheimer approximation, the stationary electronic states are determined by the time-independent Schrödinger equation,

$$\hat{H}\psi(\mathbf{r}_1, \cdots, \mathbf{r}_N) = E\psi(\mathbf{r}_1, \cdots, \mathbf{r}_N),$$

where $\hat{H}$ is the electronic Hamiltonian operator, and E and ψ are the corresponding electronic energy and eigenstate. In atomic units, the electronic Hamiltonian is

$$\hat{H} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{a=1}^{A} \frac{Z_a}{\|\mathbf{r}_i - \mathbf{x}_a\|} + \sum_{i<j}^{N} \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|} = \hat{T} + \hat{V}_{\text{ext}} + \hat{V}_{\text{ee}},$$

where $\hat{T}$ is the electronic kinetic-energy operator, $\hat{V}_{\text{ext}}$ is the electron–nuclear attraction operator, and $\hat{V}_{\text{ee}}$ is the electron–electron repulsion operator. The total molecular energy additionally includes the nuclear–nuclear repulsion energy $E_{\text{nuc}} = \sum_{a<b}^{A} \frac{Z_a Z_b}{\|\mathbf{x}_a - \mathbf{x}_b\|}$, so that

$$E_{\text{tot}} = E + E_{\text{nuc}}. \tag{S1}$$

Many ground-state molecular properties are determined by the electronic ground state, namely the eigenstate with the lowest electronic energy. The ground-state energy can therefore be obtained through the variational principle,

$$E^\star = \min_{\psi:\text{antisym}, \langle\psi|\psi\rangle=1} \langle\psi|\hat{H}|\psi\rangle, \tag{S2}$$

where $\langle\psi|\hat{H}|\phi\rangle = \int \psi^\star(\mathbf{r})\hat{H}\phi(\mathbf{r})\mathrm{d}\mathbf{r}$ denotes the many-electron inner product with operator $\hat{H}$.

Although many approaches have been developed to parameterize the many-electron wavefunction ψ and minimize the electronic energy, directly optimizing a function on $\mathbb{R}^{3N}$ suffers from the exponential growth of the many-electron configuration space, making wavefunction-based approaches computationally demanding for large systems. Density functional theory (DFT) addresses this challenge by reformulating the ground-state problem in terms of the electron density $\rho(\mathbf{r})$, a function defined on $\mathbb{R}^3$. For an N-electron wavefunction ψ, the electron density is the one-electron marginal

$$\rho(\mathbf{r}) = N \int \cdots \int |\psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N)|^2 \mathrm{d}\mathbf{r}_2 \ldots \mathrm{d}\mathbf{r}_N.$$

The central foundation of DFT is provided by the Hohenberg–Kohn (HK) theorems [3]. The first HK theorem states that, for a non-degenerate ground state, the external potential V_{ext} is uniquely determined by the ground-state density $\rho(\mathbf{r})$ up to an additive constant. Since V_{ext} specifies the electronic Hamiltonian, the ground-state density determines ground-state observables in principle. The second HK theorem establishes a variational principle over densities: there exists a universal energy functional such that the exact ground-state energy is obtained by minimizing this functional over physically admissible densities, and the minimizing density is the exact ground-state density. The external-potential contribution is already an explicit density functional because $\hat{V}_{\text{ext}}$ is a one-body operator:

$$\langle\psi|\hat{V}_{\text{ext}}|\psi\rangle = \int \rho_{[\psi]}(\mathbf{r})V_{\text{ext}}(\mathbf{r})\,\mathrm{d}\mathbf{r} = E_{\text{ext}}[\rho_{[\psi]}].$$

Accordingly, the wavefunction variational problem in Eq. (S2) can be written in the constrained-search form,

$$\begin{aligned} E^\star &= \min_{\rho:\geqslant 0, \int \rho(\mathbf{r})\,\mathrm{d}\mathbf{r}=N} \left(\min_{\psi:\,\text{antisym}, \rho_{[\psi]}=\rho} \langle\psi|\hat{T} + \hat{V}_{\text{ee}}|\psi\rangle \right) + E_{\text{ext}}[\rho] \\ &= \min_{\rho:\geqslant 0, \int \rho(\mathbf{r})\,\mathrm{d}\mathbf{r}=N} \Big\{ E[\rho] := U[\rho] + E_{\text{ext}}[\rho] \Big\}. \end{aligned} \tag{S3}$$

Supplementary Table S1: **Main notations**. These notations are used consistently throughout the paper.

Notation	Description
$\mathbf{r} \in \mathbb{R}^3$	Electron coordinate
N	Number of electrons in the molecular system
$\langle f \vert g \rangle = \int f(\mathbf{r}) g(\mathbf{r}) \mathrm{d}\mathbf{r}$	Standard inner product in function space
$\left\langle f \middle\vert \hat{O} \middle\vert g \right\rangle = \int f(\mathbf{r}) (\hat{O} g)(\mathbf{r}) \mathrm{d}\mathbf{r}$	Function-space inner product associated with operator $\hat{O}$
$(f \vert g) = \int \frac{f(\mathbf{r}) g(\mathbf{r}')}{\Vert \mathbf{r} - \mathbf{r}' \Vert} \mathrm{d}\mathbf{r} \mathrm{d}\mathbf{r}'$	Coulomb integral between two charge distributions
$\mathbf{x} \in \mathbb{R}^3$	Nucleus coordinate
$\mathbf{X} = \{\mathbf{x}^{(a)}\}_{a=1}^{A}$	Molecular geometry specified by nuclear coordinates
$\mathcal{Z} = \{Z^{(a)}\}_{a=1}^{A}$	Molecular composition specified by atomic numbers
$\mathcal{M} = \{\mathcal{Z}, \mathbf{X}\}$	Molecular structure
$\psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$	N-electron wavefunction.
$\{\phi_i(\mathbf{r})\}_{i=1}^{N}$	Molecular orbitals
$i, j \in \{1, 2, \ldots, N\}$	Indices for molecular orbitals
$\{\eta_\alpha(\mathbf{r})\}_{\alpha=1}^{B}$	Atomic orbital basis functions
$\alpha, \beta, \delta, \gamma \in \{1, 2, \ldots, B\}$	Indices for atomic orbital basis functions
$\mathbf{H}_{\alpha\beta} = \left\langle \eta_\alpha \middle\vert \hat{H} \middle\vert \eta_\beta \right\rangle$	Hamiltonian matrix in the atomic orbital basis
$\mathbf{S}_{\alpha\beta} = \langle \eta_\alpha \vert \eta_\beta \rangle$	Overlap matrix of the atomic orbital basis
$\mathbf{C}_{\alpha i}$	Coefficient of molecular orbital i on atomic orbital α
$\mathbf{P}$	Density matrix
$\boldsymbol{\varepsilon}$	Orbital energies
$\rho(\mathbf{r})$	Electron density (function)
$\mathcal{G} := \{\mathbf{r}_g\}_{g=1}^{N_G}$	Spatial grid points for evaluating real-space quantities
s	SCF iteration index

Here, the inner constrained-search problem defines the universal density functional

$$U[\rho] := \min_{\psi:\, \text{antisym}, \rho_{[\psi]} = \rho} \langle \psi | \hat{T} + \hat{V}_{\text{ee}} | \psi \rangle,$$

which is universal in the sense that it does not depend on the specific molecule $\mathcal{M}$ or external potential V_{ext}. It contains the electronic kinetic energy and the electron-electron interaction energy associated with all wavefunctions yielding the density ρ. The ground-state problem is therefore formally transformed from a variational optimization over many-electron wavefunctions to a variational optimization over electron densities, as shown in Eq. (S3).

Because the exact analytical form of the universal functional is unknown, practical DFT calculations rely on approximations. A useful formal decomposition is

$$U[\rho] = T_{\text{S}}[\rho] + E_{\text{H}}[\rho] + E_{\text{XC}}[\rho],$$

where $T_{\text{S}}[\rho]$ denotes the kinetic-energy density functional, $E_{\text{H}}[\rho]$ denotes the classical electron–electron repulsion energy, also known as the Hartree energy, and $E_{\text{XC}}[\rho]$ collects the remaining exchange–correlation effects. The Hartree energy is given by

$$E_{\text{H}}[\rho] := \frac{1}{2} \int \frac{\rho(\mathbf{r}) \rho(\mathbf{r}')}{\Vert \mathbf{r} - \mathbf{r}' \Vert} \mathrm{d}\mathbf{r} \mathrm{d}\mathbf{r}',$$

which captures the dominant classical contribution to the electron–electron interaction but neglects exchange, correlation and self-interaction corrections. The kinetic energy density functional (KEDF) can be formally defined through a constrained search over antisymmetric wavefunctions that yield the prescribed density:

$$T_{\text{S}}[\rho] := \min_{\psi:\, \text{antisym}, \rho_{[\psi]} = \rho} \langle \psi | \hat{T} | \psi \rangle.$$

The exchange–correlation (XC) functional $E_{\text{XC}}[\rho]$ is then introduced to account for the residual part of the universal functional. Over the past decades, many successful XC approximations have been developed [5, 65, 66], while constructing accurate explicit kinetic-energy density functionals remains a grand challenge [68, 69].

A.2 Kohn–Sham DFT

To avoid directly approximating the KEDF, the Kohn–Sham (KS) formulation maps the interacting N-electron system onto a virtual non-interacting reference system composed of N single-particle auxiliary orbitals $\Phi := \{\phi_i\}_{i=1}^N$ [4], where the ground-state density of the original interacting system is equal to that of the chosen non-interacting system, $\rho(\mathbf{r}) = \sum_{i=1}^N |\phi_i(\mathbf{r})|^2$. Under this formulation, the dominant kinetic energy contribution is represented by the non-interacting kinetic-energy functional

$$T_{\mathrm{S}}[\rho] = \min_{\substack{\{\phi_i\}_{i=1}^N:\text{orthonormal},\\ \rho_{[\Phi]}=\rho}} \langle\psi_{[\Phi]}|\hat{T}|\psi_{[\Phi]}\rangle = \min_{\substack{\{\phi_i\}_{i=1}^N:\text{orthonormal},\\ \rho_{[\Phi]}=\rho}} \sum_{i=1}^N \langle\phi_i|\hat{T}|\phi_i\rangle.$$

The original density minimization problem in Eq. (S3) is then converted into an orbital optimization problem,

$$E^\star = \min_{\{\phi_i\}_{i=1}^N:\text{orthonormal}} \left\{ E[\Phi] := \sum_{i=1}^N \langle\phi_i|\hat{T}|\phi_i\rangle + E_{\mathrm{H}}[\rho_{[\Phi]}] + E_{\mathrm{XC}}[\rho_{[\Phi]}] + E_{\mathrm{ext}}[\rho_{[\Phi]}] \right\}.$$

Using the variational principle, the electronic energy is minimized with respect to the KS orbitals under the orthonormality constraint $\langle\phi_i|\phi_j\rangle = \delta_{ij}$. This can be written through the Lagrangian

$$\frac{\delta}{\delta\phi_i(\mathbf{r})} \left[E[\rho] - \sum_{j=1}^N \varepsilon_{ij} \left(\langle\phi_i|\phi_j\rangle - \delta_{ij} \right) \right] (\mathbf{r}) = 0.$$

where ε_{ij} are Lagrange multipliers enforcing orbital orthonormality. Taking the functional derivative with respect to each orbital and diagonalizing the multiplier matrix yields the canonical KS equations,

$$\left(\hat{T} + \underbrace{\int \frac{\rho(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|} \mathrm{d}\mathbf{r}'}_{:=V_{\mathrm{H}}(\mathbf{r})} + \underbrace{\frac{\delta E_{\mathrm{XC}}}{\delta\rho(\mathbf{r})}}_{:=V_{\mathrm{XC}}(\mathbf{r})} + \hat{V}_{\mathrm{ext}} \right) \phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}). \tag{S4}$$

Introducing the effective KS Hamiltonian operator (also called KS Fock operator) $\hat{H} := \hat{T} + \hat{V}_{\mathrm{H}} + \hat{V}_{\mathrm{XC}} + \hat{V}_{\mathrm{ext}}$, the stationarity condition can be written compactly as

$$\hat{H}\phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}). \tag{S5}$$

Because the effective Hamiltonian $\hat{H}$ depends on the electron density constructed from its own eigenfunctions $\{\phi_i\}$, the Kohn–Sham equations must be solved self-consistently. In practice, self-consistent field (SCF) iterations are used to update the orbitals $\{\phi_i\}$ and the corresponding density until the Kohn–Sham stationarity condition is satisfied.

Although the Kohn–Sham equation in Eq. (S5) provides an iterative route to the molecular ground state, optimizing continuous single-particle orbitals remains nontrivial because their explicit forms are unknown a priori. To make the problem computationally tractable, each orbital can be expanded using a finite basis set $\{\eta_\alpha(\mathbf{r})\}_{\alpha=1}^B$: $\phi_i(\mathbf{r}) = \sum_{\alpha=1}^B C_{\alpha i}\eta_\alpha(\mathbf{r})$, which converts the original functional optimization problem into the optimization of a finite coefficient matrix $\mathbf{C}$. The electron density specified by $\mathbf{C}$ can then be written as:

$$\rho_{\mathbf{C}}(\mathbf{r}) := \sum_{i=1}^N \sum_{\alpha\beta} \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i}\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r}) = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r}), \tag{S6}$$

where $\mathbf{P}_{\alpha\beta} := \sum_i \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i}$ is the density matrix in the chosen basis. Substituting the basis expansion into the Kohn–Sham equation gives

$$\hat{H} \sum_{\beta=1}^B \mathbf{C}_{\beta i}\eta_\beta(\mathbf{r}) = \varepsilon_i \sum_{\beta=1}^B \mathbf{C}_{\beta i}\eta_\beta(\mathbf{r}).$$

Projecting both sides onto each basis function $\eta_\alpha(\mathbf{r})$ yields

$$\sum_{\beta=1}^B \mathbf{C}_{\beta i} \left\langle \eta_\alpha \middle| \hat{H} \middle| \eta_\beta \right\rangle = \varepsilon_i \sum_{\beta=1}^B \langle\eta_\alpha|\eta_\beta\rangle .$$

By denoting $\left\langle \eta_\alpha \middle| \hat{H} \middle| \eta_\beta \right\rangle = \mathbf{H}_{\alpha\beta}$ and $\langle \eta_\alpha | \eta_\beta \rangle = \mathbf{S}_{\alpha\beta}$, the differential Kohn–Sham equation is transformed into the generalized eigenvalue problem: $\sum_{\beta=1}^{B} \mathbf{H}_{\alpha\beta}\mathbf{C}_{\beta i} = \varepsilon_i \sum_{\beta=1}^{B} \mathbf{S}_{\alpha\beta}\mathbf{C}_{\beta i}$, or, in compact matrix form: $\mathbf{HC} = \mathbf{SC}\varepsilon$, corresponding to Eq. (1). The Hamiltonian is represented in an AO basis containing B basis functions, and solving the corresponding eigenvalue problem therefore requires a complete set of B molecular orbitals (MOs). For a closed-shell system with N electrons, $K = N/2$ orbitals are occupied, while the remaining $B - K$ unoccupied orbitals complete the single-particle space spanned by the AO basis. The density matrix can be reformulated as $\mathbf{P}_{\alpha\beta} := \sum_i^B f_i \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i}$, where f_i is the occupation number for the i-th orbital. According to Eq. (S4), we can construct the Hamiltonian matrix by:

$$
\begin{aligned}
&\mathbf{H} = \mathbf{T} + \mathbf{V}_{\mathrm{H}} + \mathbf{V}_{\mathrm{XC}} + \mathbf{V}_{\mathrm{ext}}, \\
\text{where}\quad &\mathbf{T}_{\alpha\beta} = \int \eta_\alpha(\mathbf{r})(-\frac{1}{2}\nabla^2)\eta_\beta(\mathbf{r})\mathrm{d}\mathbf{r}, \\
&(\mathbf{V}_{\mathrm{H}})_{\alpha\beta} := \langle \eta_\alpha | V_{\mathrm{H}} | \eta_\beta \rangle = \iint \eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r}) \frac{\sum_{\gamma\delta} \mathbf{P}_{\gamma\delta}\eta_\gamma(\mathbf{r}')\eta_\delta(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\,\mathrm{d}\mathbf{r}'\mathrm{d}\mathbf{r} \\
&\qquad\qquad = \sum_{\gamma\delta} \mathbf{D}_{\alpha\beta,\gamma\delta}\mathbf{P}_{\gamma\delta}, \\
&\qquad\qquad \text{where } \mathbf{D}_{\alpha\beta,\gamma\delta} := \iint \frac{\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})\eta_\gamma(\mathbf{r}')\eta_\delta(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\,\mathrm{d}\mathbf{r}'\mathrm{d}\mathbf{r}, \\
&(\mathbf{V}_{\mathrm{XC}})_{\alpha\beta} := \langle \eta_\alpha | V_{\mathrm{XC}} | \eta_\beta \rangle = \int V_{\mathrm{XC}}(\mathbf{r})\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})\,\mathrm{d}\mathbf{r}, \\
&(\mathbf{V}_{\mathrm{ext}})_{\alpha\beta} := \langle \eta_\alpha | V_{\mathrm{ext}} | \eta_\beta \rangle = -\sum_{a=1}^{A} Z^{(a)} \int \frac{\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})}{\left\|\mathbf{r}-\mathbf{x}^{(a)}\right\|}\,\mathrm{d}\mathbf{r}.
\end{aligned}
\tag{S7}
$$

The total energy of the system corresponding to Eq. (S1) can be written as:

$$
E_{\mathrm{tot}}(\mathbf{C}) := E_{\mathrm{tot}}[\rho_{\mathbf{C}}] = T_{\mathrm{S}}(\mathbf{C}) + E_{\mathrm{H}}(\mathbf{C}) + E_{\mathrm{XC}}(\mathbf{C}) + E_{\mathrm{ext}}(\mathbf{C}) + E_{\mathrm{nuc}}, \tag{S8}
$$

$$
\begin{aligned}
\text{where}\quad &T_{\mathrm{S}}(\mathbf{C}) := \sum_{i=1}^{N} \langle \phi_{i,\mathbf{C}} | \hat{T} | \phi_{i,\mathbf{C}} \rangle = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\mathbf{T}_{\alpha\beta}, \\
&E_{\mathrm{H}}(\mathbf{C}) := E_{\mathrm{H}}[\rho_{\mathbf{C}}] = \frac{1}{2}\sum_{\alpha\beta\gamma\delta} \mathbf{P}_{\alpha\beta}\mathbf{D}_{\alpha\beta\gamma\delta}\mathbf{P}_{\gamma\delta}, \\
&E_{\mathrm{XC}}(\mathbf{C}) := E_{\mathrm{XC}}[\rho_{\mathbf{C}}] = E_{\mathrm{XC}}\Big[\sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}\eta_\alpha\eta_\beta\Big], \\
&E_{\mathrm{ext}}(\mathbf{C}) := E_{\mathrm{ext}}[\rho_{\mathbf{C}}] = \sum_{\alpha\beta} \mathbf{P}_{\alpha\beta}(\mathbf{V}_{\mathrm{ext}})_{\alpha\beta},
\end{aligned}
\tag{S9}
$$

A typical SCF procedure, as implemented in quantum chemistry packages such as PySCF [61], proceeds as follows:

1. **Initial guess.** An initial density matrix $\mathbf{P}^{(0)}$ is constructed using an approximate scheme, such as the minimal atomic orbital basis (`MINAO`) guess.
2. **Hamiltonian construction.** At iteration s, the KS Hamiltonian matrix $\mathbf{H}^{(s)} = \mathbf{H}[\mathbf{P}^{(s)}]$ is constructed from the current density matrix $\mathbf{P}^{(s)}$. This step evaluates the one-electron terms, the Hartree potential through electron-repulsion integrals, and the exchange–correlation potential V_{XC} on numerical integration grids.
3. **Solving the KS equations.** The generalized eigenvalue problem $\mathbf{H}^{(s)}\mathbf{C}^{(s+1)} = \mathbf{S}\mathbf{C}^{(s+1)}\varepsilon^{(s+1)}$ is solved to obtain updated MO coefficients $\mathbf{C}^{(s+1)}$ and orbital energies $\varepsilon^{(s+1)}$.
4. **Density update.** A new density matrix is constructed from the occupied orbitals: $\mathbf{P}_{\alpha\beta}^{(s+1)} = \sum_{i=1} f_i \mathbf{C}_{\alpha i}^{(s+1)}\mathbf{C}_{\beta i}^{(s+1)}$. In practical implementations, density mixing or damping may be applied to improve convergence.
5. **Convergence Check.** The procedure is repeated until the change in the density matrix, total energy, or Hamiltonian matrix falls below a predefined threshold.

Traditional SCF iterations typically scale as $O(N^3)$ to $O(N^4)$ with system size and constitute a major computational bottleneck in large-scale molecular simulations. Hamiltonian learning aims to bypass these expensive iterative updates by directly predicting the converged Hamiltonian $\mathbf{H}^\star$ from the molecular structure $\mathcal{M}$.

A.3 An Extension of Lieb's Variational Duality

Lieb's convex formulation of DFT establishes a variational duality between the local one-body potential $v(\mathbf{r})$ and the ground-state electron density $\rho(\mathbf{r})$. With the natural pairing

$$\langle v, \rho \rangle = \int v(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r},$$

the concave ground-state energy $E[v]$ and the convex density functional $F[\rho]$ satisfy

$$\begin{aligned} F[\rho] &= \sup_v \left\{ E[v] - \langle v, \rho \rangle \right\}, \qquad \text{[39, Eq. (3.17)]} \\ E[v] &= \inf_\rho \left\{ F[\rho] + \langle v, \rho \rangle \right\}. \qquad \text{[39, Thm. 3.5]} \end{aligned} \tag{S10}$$

For a ground-state density ρ_v attaining the minimum in Eq. (S10), the potential and density satisfy the corresponding equality condition of these two variational problems. In this sense, v and its induced ground-state density ρ_v form a variationally conjugate pair.

The Hamiltonian–state relation considered in this work follows by extending the variable local potential in Lieb's construction to a general finite-dimensional one-body operator. This is a finite-dimensional one-body extension of the original potential–density duality, rather than a relation stated explicitly in Lieb's formulation. Such an extension is natural because the eigenvalue problem of a one-body Hamiltonian determines its ground-state occupied subspace, which is directly characterized by the corresponding one-particle density matrix. Lieb also discusses admissible one-particle density matrices as an alternative basic variable [39, Secs. 2 and 4B].

To make this relation explicit, we choose a real orthonormal one-particle basis of dimension M; the variational conjugacy derived below is independent of this choice. A one-body Hamiltonian is then represented by a symmetric matrix $\mathbf{H} \in \mathbb{S}^M$, where

$$\mathbb{S}^M = \left\{ \mathbf{A} \in \mathbb{R}^{M\times M} \,\middle|\, \mathbf{A}^\top = \mathbf{A} \right\}$$

The natural operator–state pairing is the trace pairing $\langle \mathbf{H}, \mathbf{P} \rangle = \mathrm{Tr}(\mathbf{HP})$, which is the finite-dimensional analogue of $\langle v, \rho \rangle = \int v(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r}$. For N electrons in a spin-orbital representation, the admissible one-particle density matrices satisfy

$$\mathcal{D}_N = \left\{ \mathbf{P} \in \mathbb{S}^M \,\middle|\, 0 \preceq \mathbf{P} \preceq \mathbf{I}, \quad \mathrm{Tr}(\mathbf{P}) = N \right\}.$$

For a closed-shell spin-summed spatial-orbital density matrix, the corresponding conditions are $0 \preceq \mathbf{P} \preceq 2\mathbf{I}$ and $\mathrm{Tr}(\mathbf{P}) = N_\mathrm{e}$. We retain the spin-orbital convention in the following derivation. For a fixed kinetic-energy matrix $\mathbf{T} \in \mathbb{S}^M$, let $\mathbf{V} \in \mathbb{S}^M$ denote a variable one-body potential matrix. We define the non-interacting kinetic-energy density functional as

$$\mathcal{F}_s(\mathbf{P}) = \begin{cases} \mathrm{Tr}(\mathbf{TP}), & \mathbf{P} \in \mathcal{D}_N, \\ +\infty, & \mathbf{P} \notin \mathcal{D}_N, \end{cases}$$

and the corresponding ground-state energy as

$$\mathcal{E}_s(\mathbf{V}) = \inf_{\mathbf{P}\in\mathbb{S}^M} \left\{ \mathcal{F}_s(\mathbf{P}) + \mathrm{Tr}(\mathbf{VP}) \right\}. \tag{S11}$$

The reverse variational relation is

$$\mathcal{F}_s(\mathbf{P}) = \sup_{\mathbf{V}\in\mathbb{S}^M} \left\{ \mathcal{E}_s(\mathbf{V}) - \mathrm{Tr}(\mathbf{VP}) \right\}. \tag{S12}$$

Eq. (S11) and Eq. (S12) establish the finite-dimensional analogue of Lieb's potential–density duality between $\mathbf{V}$ and $\mathbf{P}$ under the trace pairing. Since $\mathbf{T}$ is fixed, the change of variables $\mathbf{H} = \mathbf{T} + \mathbf{V}$ is an affine reparameterization of the same one-body operator space. Defining

$$\begin{aligned} \mathcal{E}_N(\mathbf{H}) &:= \mathcal{E}_s(\mathbf{H} - \mathbf{T}) \\ &= \min_{\mathbf{P}\in\mathcal{D}_N} \mathrm{Tr}(\mathbf{HP}) = \sum_{i=1}^{N} \varepsilon_i(\mathbf{H}), \end{aligned} \tag{S13}$$

where $\varepsilon_1(\mathbf{H}) \le \cdots \le \varepsilon_M(\mathbf{H})$ are the eigenvalues of $\mathbf{H}$, the reverse relation becomes

$$\iota_{\mathcal{D}_N}(\mathbf{P}) = \sup_{\mathbf{H}\in\mathbb{S}^M} \left\{\mathcal{E}_N(\mathbf{H}) - \mathrm{Tr}(\mathbf{HP})\right\}.$$

Thus, $\mathcal{E}_N$ and $\iota_{\mathcal{D}_N}$ are related by Fenchel duality under the trace pairing between one-body Hamiltonians and one-particle density matrices. A minimizer $\mathbf{P}_\mathbf{H}$ in Eq. (S13) is the ground-state one-particle density matrix induced by $\mathbf{H}$, thereby identifying the dual density-matrix variable with the electronic state obtained from the Hamiltonian eigensolution. For a general finite-dimensional one-body Hamiltonian, the full one-particle density matrix $\mathbf{P}$ is therefore the natural conjugate state variable. Restricting the Hamiltonian variation to local potentials recovers Lieb's original potential–density pairing, since $\mathrm{Tr}(\mathbf{V}[v]\mathbf{P}) = \int v(\mathbf{r})\rho_\mathbf{P}(\mathbf{r})\,\mathrm{d}\mathbf{r}$. We accordingly refer to the finite-dimensional relation between $\mathbf{H}$ and $\mathbf{P}$ as a *Hamiltonian–density-matrix variational duality*, or more broadly a *Hamiltonian–state variational duality* (Introduction 1). This construction is a finite-dimensional one-body extension of Lieb's potential–density duality and applies specifically to one-body Hamiltonians.

B DSHam Technical Details

In this section, we provide additional technical details of the proposed DSHam framework. We first present an in-depth gradient analysis of the supervision signals discussed in the main text and elaborate on the design motivation of DSLoss in Supplementary Sec. B.1. We then describe the configurations of the employed $\mathrm{SE}(3)$-equivariant graph neural network (SEGNN) models for Hamiltonian prediction, electron density prediction and physical property prediction in Supplementary Sec. B.2. Finally, we provide additional implementation details for data preparation (Supplementary Sec. B.3), DFT calculations (Supplementary Sec. B.5), MD simulations (Supplementary Sec. B.4), hardware configuration (Supplementary Sec. B.6) and error analysis (Supplementary Sec. B.7).

B.1 Analysis for Dual-Space Loss

B.1.1 Perturbation Analysis

As discussed in Results 2.1, supervision restricted to the Hamiltonian space does not directly constrain the electronic solution induced by the predicted Hamiltonian, and therefore may not provide sufficient guidance for optimizing downstream observables. Here we first understand this mismatch by examining how a perturbation to the Hamiltonian propagates to its electronic solution. Unless otherwise stated, we consider only responses to first order in the Hamiltonian perturbation throughout this section.

We first transform the generalized Kohn–Sham eigenvalue problem into an orthonormal representation. Let the overlap matrix have the eigendecomposition $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, and define the symmetric orthogonalization matrix as $\widetilde{\mathbf{Q}} = \mathbf{Q}\mathbf{\Lambda}^{-1/2}\mathbf{Q}^\top$. With $\widetilde{\mathbf{H}} = \widetilde{\mathbf{Q}}^\top\mathbf{H}\widetilde{\mathbf{Q}}$ and $\mathbf{C} = \widetilde{\mathbf{Q}}\mathbf{U}$, Eq. (1) becomes

$$\widetilde{\mathbf{H}}\mathbf{U} = \mathbf{U}\boldsymbol{\varepsilon}, \qquad \mathbf{U}^\top\mathbf{U} = \mathbf{I}.$$

Thus, the original generalized eigenvalue problem can be analyzed using standard symmetric eigenvalue perturbation theory.

Consider a predicted Hamiltonian $\hat{\mathbf{H}} = \mathbf{H}^\star + \Delta\mathbf{H}$. Since the overlap matrix is fixed for a given molecular geometry and AO basis, $\widetilde{\mathbf{Q}}$ does not respond to $\Delta\mathbf{H}$, and the corresponding perturbation is $\Delta\widetilde{\mathbf{H}} = \widetilde{\mathbf{Q}}^\top\Delta\mathbf{H}\widetilde{\mathbf{Q}}$. Let $\mathbf{U}^\star$ denote the eigenvectors of $\widetilde{\mathbf{H}}^\star$ and $\mathbf{C}^\star = \widetilde{\mathbf{Q}}\mathbf{U}^\star$ the corresponding MO coefficients. Expressing the perturbation in the reference MO basis gives

$$\Delta\mathbf{H}_{\mathrm{MO}} = \mathbf{U}^{\star\top}\Delta\widetilde{\mathbf{H}}\mathbf{U}^\star = \mathbf{C}^{\star\top}\Delta\mathbf{H}\mathbf{C}^\star. \tag{S14}$$

Let $\mathbf{f} = \mathtt{diag}(f_1, \ldots, f_B)$ denote the orbital occupation matrix and $\widetilde{\mathbf{P}} = \mathbf{U}\mathbf{f}\mathbf{U}^\top$ the density matrix in the orthogonalized AO basis. We express its response in the reference MO basis as $\Delta\widetilde{\mathbf{P}}_{\mathrm{MO}} = \mathbf{U}^{\star\top}\Delta\widetilde{\mathbf{P}}\mathbf{U}^\star$. Assuming no degeneracies between orbitals of different occupations, first-order perturbation theory gives

$$\Delta\varepsilon_i = (\Delta\mathbf{H}_{\mathrm{MO}})_{ii}, \qquad (\Delta\widetilde{\mathbf{P}}_{\mathrm{MO}})_{ij} = \frac{f_i - f_j}{\varepsilon_i^\star - \varepsilon_j^\star}(\Delta\mathbf{H}_{\mathrm{MO}})_{ij} \quad (i \ne j). \tag{S15}$$

These relations separate two distinct responses of the electronic solution. The diagonal components of $\Delta\mathbf{H}_{\mathrm{MO}}$ shift the orbital energy spectrum directly. By contrast, changes in the density matrix arise from couplings between orbitals with different occupations. Under the usual fixed integer occupations, these are the couplings between occupied and virtual orbitals, which rotate the occupied subspace and thereby

Supplementary Table S2: **Gradients of various physical observables.**

Quantity	Gradient	Reference
$\mathbf{P}$	$\frac{\partial \mathbf{P}_{\alpha\beta}}{\partial \mathbf{H}_{\gamma\delta}} = \sum_{ij} \frac{f(\boldsymbol{\varepsilon}_i) - f(\boldsymbol{\varepsilon}_j)}{\boldsymbol{\varepsilon}_i - \boldsymbol{\varepsilon}_j} \mathbf{C}_{\alpha i} \mathbf{C}_{\beta j} \mathbf{C}_{\gamma i} \mathbf{C}_{\delta j}$	Supplementary Sec. B.1.3
ρ	$\frac{\partial \rho}{\partial \mathbf{P}_{\alpha\beta}} = \int \eta_\alpha(\mathbf{r}) \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r}$	Supplementary Sec. B.1.4
$\boldsymbol{\varepsilon}$	$\frac{\partial \boldsymbol{\varepsilon}_i}{\partial \mathbf{H}_{\gamma\delta}} = \mathbf{C}_{\gamma i} \mathbf{C}_{\delta i}$	Supplementary Sec. B.1.6
T_S	$\frac{\partial T_\mathrm{S}}{\partial \mathbf{P}_{\alpha\beta}} = -\frac{1}{2} \int \eta_\alpha(\mathbf{r}) \nabla^2 \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r} = \mathbf{T}_{\alpha\beta}$	Supplementary Sec. B.1.7
E_ext	$\frac{\partial E_\mathrm{ext}}{\partial \mathbf{P}_{\alpha\beta}} = \sum_a \int \eta_\alpha(\mathbf{r}) \frac{1}{\lVert \mathbf{r} - \mathbf{x}_a \rVert} \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r} = (\mathbf{V}_\mathrm{ext})_{\alpha\beta}$	Supplementary Sec. B.1.7
E_core	$\frac{\partial E_\mathrm{core}}{\partial \mathbf{P}_{\alpha\beta}} = \mathbf{T}_{\alpha\beta} + (\mathbf{V}_\mathrm{ext})_{\alpha\beta}$	Supplementary Sec. B.1.7
$\boldsymbol{\mu}$	$\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{P}_{\alpha\beta}} = \int \eta_\alpha(\mathbf{r}) \mathbf{r} \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r}$	Supplementary Sec. B.1.5

modify the ground state electron distribution. Importantly, the density response depends not only on the magnitude of $\Delta\mathbf{H}_\mathrm{MO}$ but also on the occupation differences and inverse orbital energy separations in Eq. (S15). Hamiltonian perturbations of comparable magnitude can therefore produce very different changes in the occupied subspace. In particular, couplings across smaller occupied–virtual energy gaps induce larger state responses, whereas perturbations along directions that preserve the occupied subspace can have little effect on the density, even though conventional element-wise supervision assigns equal weight to both types of perturbations. This anisotropic response provides a perturbative explanation for why numerical Hamiltonian accuracy alone does not determine physical fidelity.

The density matrix in the original AO representation satisfies $\mathbf{P} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}^\top$ and therefore $\Delta\mathbf{P} = \widetilde{\mathbf{Q}}\Delta\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}^\top$. Since this transformation is fixed, the state response identified above is carried directly from the orthogonalized representation to the AO density matrix used in our supervision. Accordingly, $\mathcal{L}_\mathbf{P}$ constrains this state response in the AO representation, while $\mathcal{L}_\rho$ further constrains its real-space electron distribution. Together with orbital energy supervision $\mathcal{L}_{\boldsymbol{\varepsilon}}$ of the spectral response, these objectives extend supervision from the Hamiltonian itself to complementary components of its induced eigensolution, realizing our dual-space supervision. In contrast, conventional element-wise Hamiltonian supervision weights matrix errors solely by their numerical magnitudes, without accounting for their strongly nonuniform effects on the eigensolution. Density supervision instead propagates electronic-state errors back to the Hamiltonian through a physically structured sensitivity map, providing the basis for the gradient reweighting analyzed below.

B.1.2 Gradient Analysis

Dual-space supervision through $\mathcal{L}_\mathbf{P}$ and $\mathcal{L}_\rho$ in Eq. (2) directly constrains the electronic solution induced by the predicted Hamiltonian. Empirical comparisons with several computationally efficient auxiliary objectives further show that density supervision provides the most consistent improvements across downstream properties (Results 2.1). To understand the origin of this broad transferability, we trace the gradient pathways of different supervision signals back to the Hamiltonian and analyze the resulting sensitivity maps from both optimization and chemical perspectives. These maps characterize how each objective distributes its corrective signal across Hamiltonian elements.

For an auxiliary quantity $\mathbf{O}$ that depends differentiably on the density matrix, its loss gradient with respect to the Hamiltonian follows the chain

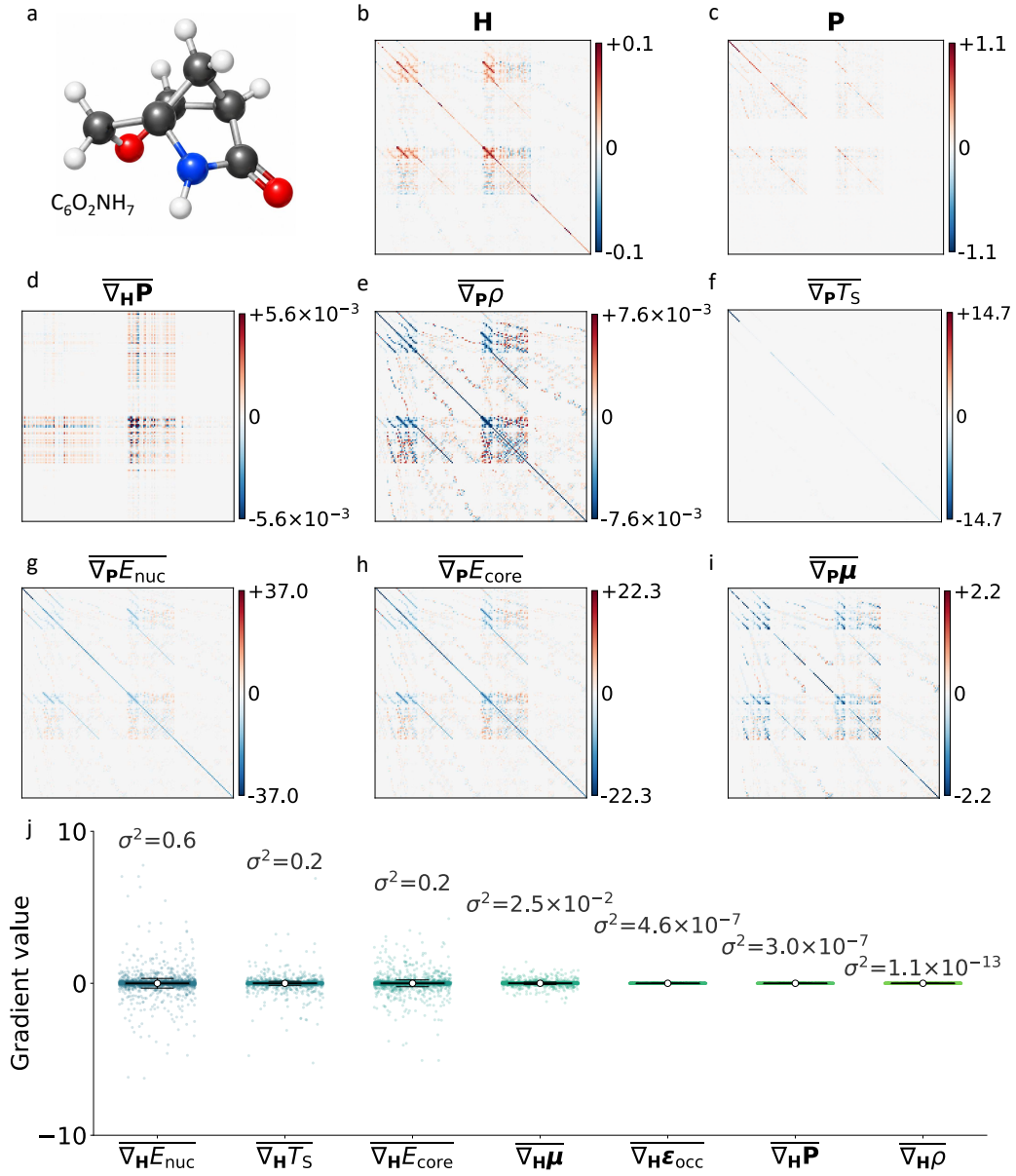
$$\frac{\partial \mathcal{L}_\mathbf{O}}{\partial \mathbf{H}} = \frac{\partial \mathcal{L}_\mathbf{O}}{\partial \mathbf{O}} \frac{\partial \mathbf{O}}{\partial \mathbf{P}} \frac{\partial \mathbf{P}}{\partial \mathbf{H}}.$$

This decomposition separates the Hamiltonian-to-state response, $\partial\mathbf{P}/\partial\mathbf{H}$, from the objective-specific reweighting, $\partial\mathbf{O}/\partial\mathbf{P}$. Different auxiliary objectives therefore propagate through the same underlying electronic response while assigning different weights to its Hamiltonian directions. We visualize these patterns using a randomly sampled molecule from QH9, $C_6O_2NH_7$, as an illustrative example (Supplementary Fig. S1a).

The Jacobian of the density matrix with respect to the Hamiltonian is

$$\nabla_\mathbf{H}\mathbf{P} = \left\{ \frac{\partial \mathbf{P}_{\alpha\beta}}{\partial \mathbf{H}_{\gamma\delta}} \right\}^B_{\alpha\beta\gamma\delta},$$

which is derived in Eq. (8) and summarized in Supplementary Table S2. This Jacobian is the AO-space differential counterpart of the occupied-subspace response established in Supplementary Sec. B.1.1. For



Supplementary Figure S1: **Gradient of different observables used as auxiliary losses.** **(a)** Structure of the example molecule $C_6O_2NH_7$ used for gradient visualization in **(b)-(i)**. White, gray, blue, and red spheres denote H, C, N, and O atoms, respectively. **(b)-(c)** Heat maps of the Hamiltonian matrix and density matrix. **(d)** Aggregated gradient map of the density matrix with respect to the Hamiltonian matrix. **(e)-(i)** Aggregated gradient maps of density-matrix-derived quantities with respect to the density matrix. **(j)** Distributions of the Jacobian entries of different observables with respect to Hamiltonian matrix elements. The corresponding gradient variance σ^2 is annotated above each distribution.

visualization, we reduce the four-dimensional Jacobian to a two-dimensional sensitivity map, $\overline{\nabla_{\mathbf{H}}\mathbf{P}}$, by aggregating over the density-matrix indices α and β while retaining the Hamiltonian indices γ and δ. The resulting map measures how strongly individual Hamiltonian elements can influence the density matrix and therefore identifies the directions through which $\mathcal{L}_{\mathbf{P}}$ provides corrective supervision.

As shown in Supplementary Fig. S1d, Hamiltonian elements associated with core-like AOs, such as deeply bound `1s` orbitals, exhibit substantially lower sensitivities than those involving valence orbitals. This behavior is consistent with the perturbative picture above: deeply bound core orbitals generally have larger energy separations from virtual states and undergo weaker mixing, whereas valence orbitals participate more strongly in bonding and in density variations relevant to molecular properties [70]. Our empirical results further show that core-related Hamiltonian elements can already be fitted accurately using element-wise Hamiltonian supervision alone, consistent with observations in [43]. Thus, $\mathcal{L}_{\mathbf{P}}$ complements $\mathcal{L}_{\mathbf{H}}$ by assigning greater corrective sensitivity to Hamiltonian directions that more strongly modify the occupied subspace, many of which involve valence and bonding interactions. Additional analysis is provided in Supplementary Sec. B.1.3.

To compare alternative physical surrogates derived from the density matrix, we further evaluate their Jacobians $\nabla_{\mathbf{P}}\mathbf{O}$ (Supplementary Table S2) and visualize the corresponding aggregated maps $\overline{\nabla_{\mathbf{P}}\mathbf{O}}$ in Supplementary Figs. S1e–i, where $\mathbf{O}$ includes quantities such as ρ and T_{S}. Several property-specific objectives exhibit sensitivities that are strongly concentrated to a limited subset of AO pairs. When propagated through $\nabla_{\mathbf{H}}\mathbf{P}$, this concentration produces larger variation across Hamiltonian gradients (Supplementary Fig. S1j), leaving other directions comparatively weakly constrained. This behavior is consistent with less stable optimization and with the limited improvements observed for properties beyond the directly supervised one in Fig. 2a.

Real-space density supervision instead produces a broader and smoother reweighting across AO pairs while retaining chemically meaningful selectivity. For a fixed AO basis, $\frac{\partial \rho(\mathbf{r})}{\partial \mathbf{P}_{\alpha\beta}} = \eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})$. The density Jacobian therefore weights each AO pair according to the spatial distribution of its corresponding pair density. Because $\mathcal{L}_\rho$ aggregates density deviations over spatial grid points, these contributions are integrated across the molecular electron distribution, producing a less sharply concentrated sensitivity pattern than objectives derived from individual scalar properties (Supplementary Fig. S1e). At the same time, the weighting remains nonuniform and preferentially responds to AO pairs with substantial spatial contributions to the density.

In the localized AO representation used here, these high-sensitivity components are enriched in valence and σ-bonding regions (Supplementary Fig. S2d). Real-space density supervision therefore complements the AO-space density-matrix objective $\mathcal{L}_{\mathbf{P}}$ by smoothly reweighting density matrix deviations according to their consequences for the spatial electron distribution. Together, $\mathcal{L}_{\mathbf{P}}$ and $\mathcal{L}_\rho$ provide broad gradient coverage without treating all Hamiltonian directions equivalently, thereby combining optimization stability with chemically structured supervision.

Orbital-energy supervision targets the complementary spectral component of the eigensolution. Under our formulation, the corresponding aggregated Hamiltonian sensitivity map of the occupied orbital energies, $\overline{\nabla_{\mathbf{H}}\boldsymbol{\varepsilon}_{\mathrm{occ}}}$, reduces to the density matrix $\mathbf{P}$; see Supplementary Sec. B.1.6. This complementary role is also reflected in the distinct first-order response of the orbital energies: unlike the density response, which is governed by occupied–virtual mixing and the associated inverse energy gaps, the orbital-energy response depends directly on diagonal Hamiltonian perturbations in the MO basis. $\mathcal{L}_{\boldsymbol{\varepsilon}}$ therefore provides a complementary signal for constraining the spectrum induced by the predicted Hamiltonian and empirically improves optimization stability in near-degenerate cases encountered when backpropagating through the density matrix loss (Methods 4.2; Supplementary Sec. B.1.6).

Together, these analyses explain the complementary optimization mechanisms underlying DSLoss. The AO-space density matrix objective propagates state errors over a broad set of Hamiltonian directions according to their influence on the occupied subspace. The real-space density objective further introduces a smooth and spatially informed reweighting that preferentially emphasizes chemically relevant valence and bonding interactions. Orbital energy supervision complements these state constraints by directly targeting the eigenvalue spectrum. These mechanisms connect the perturbative response of the eigensolution to the gradient structure encountered during training and explain why dual-space supervision yields more consistent improvements across downstream molecular properties. Further discussions of the individual auxiliary objectives are provided in Supplementary Secs. B.1.3–B.1.7.

B.1.3 Density Matrix

As described in Methods 4.2 and Supplementary Sec. B.1.1, evaluating $\mathcal{L}_{\mathbf{P}}$ requires the MO coefficients $\mathbf{C}$ obtained by solving the generalized Kohn–Sham eigenvalue problem in the nonorthogonal AO basis (Eq. (1)). Backpropagating through this eigensolver therefore involves the Jacobian $\nabla_{\mathbf{H}}\mathbf{P}$. Here, we provide a detailed derivation of $\nabla_{\mathbf{H}}\mathbf{P}$. We first introduce an infinitesimal perturbation $\delta\mathbf{H}$ to the Hamiltonian while keeping $\mathbf{S}$ fixed. In the orthogonalized representation, this perturbation becomes $\delta\widetilde{\mathbf{H}} = \widetilde{\mathbf{Q}}^\top \delta\mathbf{H}\widetilde{\mathbf{Q}}$. For non-degenerate eigenvalues and fixed occupations, the first-order response of the eigenvector $\mathbf{U}_i$ is

$$\delta\mathbf{U}_i = \sum_{j\neq i} \mathbf{U}_j \frac{\mathbf{U}_j^\top \delta\widetilde{\mathbf{H}}\mathbf{U}_i}{\varepsilon_i - \varepsilon_j}.$$

Transforming this response back to the AO basis gives

$$\delta\mathbf{C}_i = \widetilde{\mathbf{Q}}\delta\mathbf{U}_i = \sum_{j\neq i} \widetilde{\mathbf{Q}}\mathbf{U}_j \frac{\mathbf{U}_j^\top \widetilde{\mathbf{Q}}^\top \delta\mathbf{H}\widetilde{\mathbf{Q}}\mathbf{U}_i}{\varepsilon_i - \varepsilon_j} = \sum_{j\neq i} \mathbf{C}_j \frac{\mathbf{C}_j^\top \delta\mathbf{H}\mathbf{C}_i}{\varepsilon_i - \varepsilon_j},$$

where $\mathbf{C}_i$ denotes the i-th column of $\mathbf{C}$. The first-order variation of the density matrix is therefore

$$\delta\mathbf{P} = \sum_i f_i \left(\delta\mathbf{C}_i\mathbf{C}_i^\top + \mathbf{C}_i\delta\mathbf{C}_i^\top\right) = \sum_{ij} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\mathbf{C}_i\mathbf{C}_j^\top \left(\mathbf{C}_i^\top \delta\mathbf{H}\mathbf{C}_j\right), \tag{S16}$$

where the $i = j$ term is understood as zero for fixed occupations.

Writing the scalar perturbation term element-wise, $\mathbf{C}_i^\top \delta\mathbf{H}\mathbf{C}_j = \sum_{\gamma\delta} \mathbf{C}_{\gamma i}\delta\mathbf{H}_{\gamma\delta}\mathbf{C}_{\delta j}$, we obtain

$$\delta\mathbf{P}_{\alpha\beta} = \sum_{\gamma\delta} \left[\sum_{ij} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\mathbf{C}_{\alpha i}\mathbf{C}_{\beta j}\mathbf{C}_{\gamma i}\mathbf{C}_{\delta j}\right] \delta\mathbf{H}_{\gamma\delta}.$$

Thus, the Jacobian of the density matrix with respect to the Hamiltonian matrix is

$$\frac{\partial\mathbf{P}_{\alpha\beta}}{\partial\mathbf{H}_{\gamma\delta}} = \sum_{ij} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\mathbf{C}_{\alpha i}\mathbf{C}_{\beta j}\mathbf{C}_{\gamma i}\mathbf{C}_{\delta j},$$

which gives the expression in Eq. (8).

To connect the AO-space response derived above with the MO-space response in Supplementary Eq. S15, we identify $\mathbf{C}_i^\top \delta\mathbf{H}\mathbf{C}_j$ in Supplementary Eq. S16 as the Hamiltonian perturbation projected onto the MO basis according to Supplementary Eq. S14:

$$(\delta\mathbf{H}_{\mathrm{MO}})_{ij} = \mathbf{C}_i^\top \delta\mathbf{H}\mathbf{C}_j. \tag{S17}$$

Substituting Eq. (S17) into Supplementary Eq. S16 gives

$$\delta\mathbf{P} = \sum_{ij} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j}\mathbf{C}_i\mathbf{C}_j^\top \left(\delta\mathbf{H}_{\mathrm{MO}}\right)_{ij},$$

Using the equivalent representation $\delta\mathbf{P} = \mathbf{C}\delta\widetilde{\mathbf{P}}_{\mathrm{MO}}\mathbf{C}^\top$, we obtain, for $i \neq j$,

$$\left(\delta\widetilde{\mathbf{P}}_{\mathrm{MO}}\right)_{ij} = \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j} \left(\delta\mathbf{H}_{\mathrm{MO}}\right)_{ij},$$

which is aligned with Supplementary Eq. S16.

According to the aggregated gradient map $\overline{\nabla_{\mathbf{H}}\mathbf{P}}$, the gradient of the density matrix with respect to the Hamiltonian is strongly suppressed for Hamiltonian elements associated with core-like AO basis functions. Specifically, when the AO indices γ and δ correspond to heavy-atom `1s` basis functions, their coefficients are mainly concentrated in the lowest N_{heavy} MOs, where N_{heavy} denotes the number of heavy atoms in the molecule. These low-lying orbitals are typically core-like and are dominated by localized `1s` atomic orbitals, making the corresponding coefficient patterns highly sparse (Supplementary Fig. S2b). Moreover, nonzero contributions in Eq. (8) require the orbital indices i and j to have different occupations, namely one occupied and the other unoccupied. For core-like AO indices, however, $\mathbf{C}_{\gamma i}$ and $\mathbf{C}_{\delta j}$ are appreciable mainly within the occupied core subspace and are negligible for unoccupied orbitals. Consequently, the products $\mathbf{C}_{\gamma i}\mathbf{C}_{\delta j}$ that connect occupied and unoccupied orbitals are nearly zero,

leading to an approximately vanishing aggregated gradient $\sum_{\alpha\beta} \frac{\partial \mathbf{P}_{\alpha\beta}}{\partial \mathbf{H}_{\gamma\delta}}$, as illustrated in Supplementary Fig. S2c.

We also observe that density matrix supervision can introduce optimization instability through the factor $1/|\varepsilon_i - \varepsilon_j|$ in Eq. (8), which becomes large for nearly degenerate orbital energies. In practice, we tackle this issue by applying the gradient truncation technique described in Methods 4.2.

As discussed in Results 2.2 and Results 2.4, we compare Hamiltonian learning methods with approaches that directly predict the density matrix. Direct density matrix prediction exhibits stronger generalization in terms of density matrix MAE (Fig. 3). This behavior may be partly attributable to the more bounded numerical range of density labels, which can lead to better controlled optimization and generalization under the MAE objective [71]. However, improved density matrix MAE does not necessarily translate into more accurate downstream physical properties. As discussed in Results 2.4, direct density matrix regression does not explicitly enforce the structural constraints of physically valid density matrices, such as electron-number conservation and generalized idempotency in a nonorthogonal AO basis. Consequently, a density matrix with a low element-wise error may still fail to correspond to a valid occupied electronic subspace, leading to systematic errors in derived quantities such as total energy and dipole moment (Supplementary Fig. S8).

These observations suggest that the Hamiltonian serves as a crucial intermediate representation for preserving the physical structure of molecular electronic states. At the same time, density matrix supervision provides a complementary signal that emphasizes electronically important regions and helps align Hamiltonian learning with downstream physical observables.

B.1.4 Real-space Density

As discussed in Supplementary Sec. A.2, given an AO basis $\{\eta_\alpha(\mathbf{r})\}_{\alpha=1}^{B}$, the real-space electron density can be reconstructed from the density matrix through Eq. (S6). The gradient of $\mathcal{L}_\rho$ with respect to $\mathbf{P}$ is given by

$$\frac{\partial L_\rho}{\partial \mathbf{P}_{\alpha\beta}} = \int \frac{\partial L_\rho}{\partial \rho(\mathbf{r})} \eta_\alpha(\mathbf{r}) \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r},$$

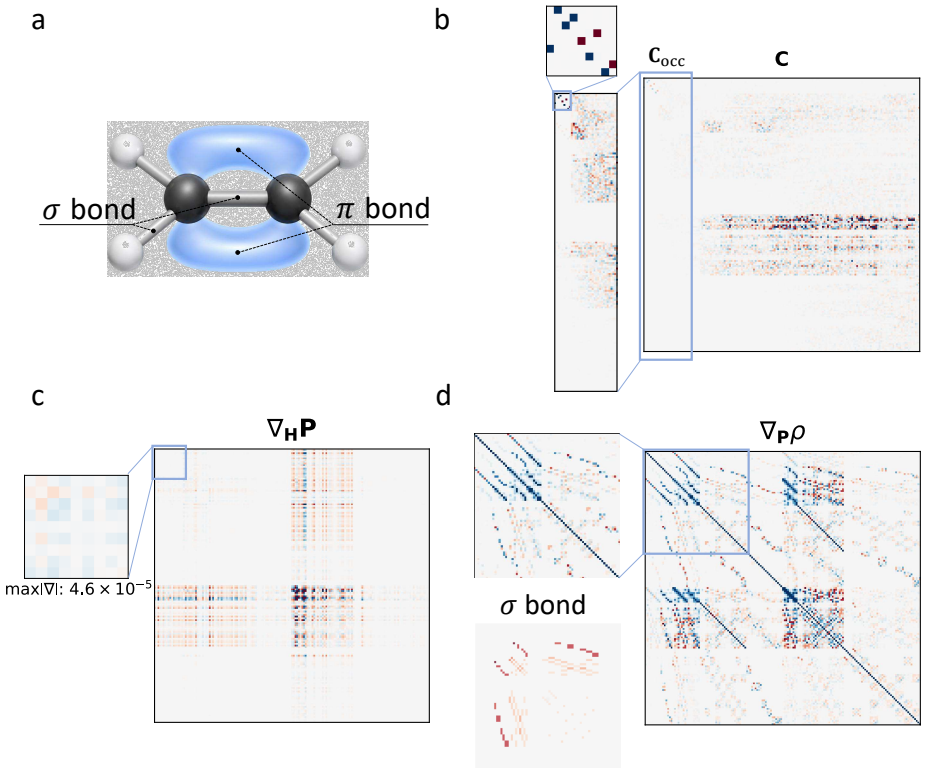
which shows that AO pairs with larger spatial overlap receive stronger supervision. Such AO pairs are often associated with chemically important bonding regions, particularly σ-bonding regions (Supplementary Fig. S2a,d). Because σ bonds constitute a major component of the molecular electron-density backbone, errors in these regions can substantially affect downstream molecular properties. Therefore, supervising the real-space electron density naturally emphasizes electronically important regions in downstream property prediction. A schematic comparison between σ and π bonds using ethene as an example is provided in Supplementary Fig. S2a.

To empirically validate the importance of the AO pairs emphasized by $\mathcal{L}_\rho$, we selected 500 molecules from the `stable-random` split of QH9 and applied small perturbations to their reference density matrices. Specifically, uniform noise was added to density matrix elements that receive large weights from the real-space density loss. To preserve Hermiticity, the perturbation matrix was constrained to be Hermitian. Because strictly maintaining both idempotency and electron-number conservation under perturbation is nontrivial, we used zero-mean noise with small variance and kept all diagonal elements unchanged to minimize global changes to the density matrix. For comparison, we generated a control perturbation by simultaneously permuting the rows and columns of the same noise matrix, thereby preserving Hermiticity and the noise magnitude distribution while removing its correlation with AO spatial overlap. As shown in Supplementary Fig. S4a, perturbations concentrated on AO pairs with large spatial overlap induce substantially larger total-energy errors than the permuted control, confirming the importance of these AO pairs in molecular property determination.

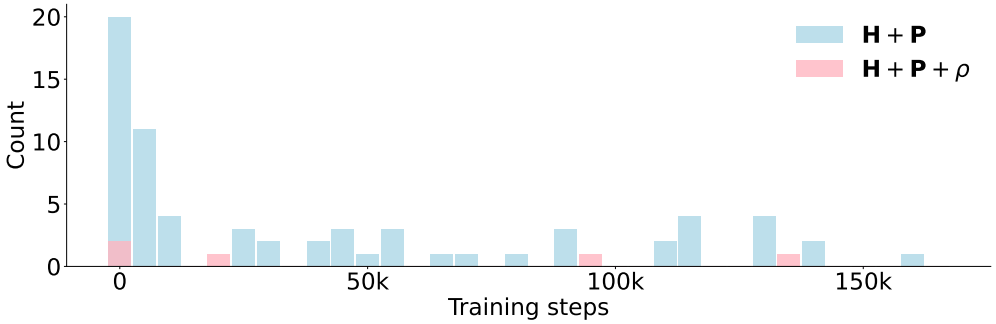
In addition, we observe that incorporating $\mathcal{L}_\rho$ improves training stability by reducing the frequency of near-degenerate orbital events. We define a near-degenerate orbital event during training as $|\hat{\varepsilon}_{\mathrm{HOMO}} - \hat{\varepsilon}_{\mathrm{LUMO}}| \leq 0.1\,|\varepsilon_{\mathrm{HOMO}} - \varepsilon_{\mathrm{LUMO}}|$. After introducing $\mathcal{L}_\rho$, the number of such events decreases markedly from 68 to 5 (Supplementary Fig. S3), suggesting that real-space density supervision helps regularize the induced electronic state during Hamiltonian optimization.

B.1.5 Dipole Moment

We also explore the dipole moment as an auxiliary supervision target because of its relevance to molecular polarity and its low computational cost. According to Eq. (7), differentiating the dipole moment with



Supplementary Figure S2: **Gradient visualization of density supervision.** The visualization is based on the same $C_6O_2NH_7$ molecule used in Supplementary Fig. S1. **(a)** Schematic illustration of σ- and π-bonds in ethene (C_2H_4). White and gray spheres denote H and C atoms. Sticks and the blue region indicate the σ- and π-bonding regions, respectively. **(b)** Heat map of the orbital coefficient matrix. AOs and MOs are reordered according to their angular momenta and orbital energies, respectively. The upper-left block corresponds to the nine core MOs associated with the nine heavy atoms, each of which is dominated by a `1s` AO. **(c)** Aggregated gradient map of the density matrix with respect to the Hamiltonian matrix. AOs are reordered according to their angular momentum orders. Elements associated with core-like AOs receive smaller gradient weights, reflecting the localized and `1s`-dominated character of core orbitals. **(d)** Aggregated gradient map of the electron density with respect to the density matrix. AOs are reordered according to their angular momentum orders. The highlighted upper-left region corresponds to AOs in the minimal-basis description, showing that density-based supervision primarily emphasizes AO pairs associated with σ-bonding regions. The σ-bond map in the lower-left region is obtained using valence bond theory.



Supplementary Figure S3: **Frequency of near-degeneracy events during training.** Incorporating density supervision markedly reduces the frequency of near-degenerate orbital events, indicating improved training stability.

respect to the density matrix yields the AO-basis representation of the position operator:

$$\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{P}_{\alpha\beta}} = \frac{\partial \boldsymbol{\mu}_{\text{elec}}}{\partial \mathbf{P}_{\alpha\beta}} = -\int \eta_\alpha(\mathbf{r}) \mathbf{r} \eta_\beta(\mathbf{r}) \mathrm{d}\mathbf{r},$$

where only the electronic contribution $\boldsymbol{\mu}_{\text{elec}}$ depends on $\mathbf{P}$, while the nuclear contribution is fixed for a given molecular geometry. This expression can be interpreted as a position-weighted aggregation of the density response $\frac{\partial \rho}{\partial \mathbf{P}_{\alpha\beta}}$. As shown in Supplementary Fig. S1i, the aggregated gradient map $\overline{\nabla_{\mathbf{P}} \boldsymbol{\mu}}$ is substantially sparser than $\overline{\nabla_{\mathbf{P}} \rho}$. A possible explanation is that when the pair function $\eta_\alpha(\mathbf{r})\eta_\beta(\mathbf{r})$ is nearly symmetric with respect to the origin or an inversion center, its contribution to the position-weighted integral can cancel out, leading to a small gradient for the corresponding AO pair. This selective sensitivity is also reflected in the significantly larger variance of $\overline{\nabla_{\mathbf{P}} \boldsymbol{\mu}}$ compared with $\overline{\nabla_{\mathbf{P}} \rho}$ (Supplementary Fig. S1j). As a result, dipole-moment supervision can introduce a biased and sparse optimization signal for the density matrix, leaving many density-matrix directions weakly constrained and potentially destabilizing training. As shown in Fig. 3a, a model can achieve low dipole-moment error while still failing to recover a physically consistent density matrix.

B.1.6 Orbital Energy

As discussed in Results 2.1, to further characterize the electronic structure at the eigenvalue level, we also introduce the occupied orbital-energy supervision. The first-order variation of the eigenvalue ε_i under a Hamiltonian perturbation $\delta\mathbf{H}$ is $\delta\varepsilon_i = \mathbf{C}_i^\top \delta\mathbf{H}\mathbf{C}_i$. Since $\delta\mathbf{H}_{\text{MO}} = \mathbf{C}^\top \delta\mathbf{H}\mathbf{C}$ is the perturbation projected onto the fixed MO basis, the expression above is equivalently $\delta\varepsilon_i = (\delta\mathbf{H}_{\text{MO}})_{ii}$, which recovers $\Delta\varepsilon_p = (\Delta\mathbf{H}_{\text{MO}})_{pp}$ in Supplementary Eq. S15. Therefore, when Hamiltonian matrix elements are treated as independent entries, we have

$$\frac{\partial \varepsilon_i}{\partial \mathbf{H}_{\alpha\beta}} = \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i}.$$

Summing over occupied orbitals with their occupation numbers gives

$$\sum_i f_i \frac{\partial \varepsilon_i}{\partial \mathbf{H}_{\alpha\beta}} = \sum_i f_i \mathbf{C}_{\alpha i}\mathbf{C}_{\beta i} = \mathbf{P}_{\alpha\beta}.$$

Thus, occupied orbital energy supervision naturally weights Hamiltonian elements according to the density matrix in the AO basis. To further understand this mechanism, we perform a perturbation analysis analogous to that in Supplementary Sec. B.1.4. We randomly select 500 molecules from QH9 and identify the Hamiltonian elements emphasized by occupied orbital energy supervision according to the magnitudes of the corresponding density matrix entries. Uniform noise is then added to the selected off-diagonal Hamiltonian elements. As a random baseline, we apply the same perturbation matrix after randomly permuting its element locations while preserving Hermiticity, thereby keeping the perturbation strength and sparsity unchanged while removing the density matrix guidance. The perturbed Hamiltonians are used to compute total energy. As shown in Supplementary Fig. S4b, perturbations concentrated on Hamiltonian elements associated with large density matrix entries induce substantially larger total energy errors than random perturbations. This confirms that occupied orbital energy supervision naturally emphasizes Hamiltonian elements that are more relevant to downstream molecular properties.

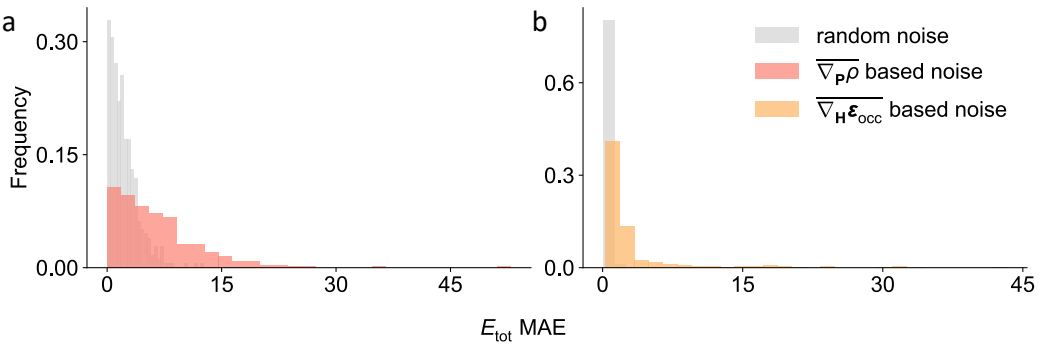
B.1.7 One-electron Energy

As discussed above, element-wise Hamiltonian errors are poorly aligned with the electronic state and downstream observables. Although total energy supervision provides a physically meaningful global objective, it introduces substantial computational overhead. We therefore investigate several one-electron energy components, including the kinetic energy T_{S}, the external potential energy E_{ext} and their sum E_{core}, as computationally efficient alternatives for global supervision.

To understand the behavior of these energy-based supervision signals, we analyze their gradients with respect to the density matrix. As discussed in Methods 4.1, each one-electron energy component is linear in the AO density matrix. Therefore, differentiating each component with respect to a density matrix element yields the corresponding one-electron operator in the AO basis:

$$\frac{\partial T_{\text{S}}}{\partial \mathbf{P}_{\alpha\beta}} = \mathbf{T}_{\alpha\beta}, \quad \frac{\partial E_{\text{ext}}}{\partial \mathbf{P}_{\alpha\beta}} = (\mathbf{V}_{\text{ext}})_{\alpha\beta}, \quad \frac{\partial E_{\text{core}}}{\partial \mathbf{P}_{\alpha\beta}} = (\mathbf{T} + \mathbf{V}_{\text{ext}})_{\alpha\beta}.$$

The gradient induced by E_{ext} is determined by the external potential matrix in Supplementary Eq. S7. Because the nuclear attraction potential has the largest magnitude near nuclei, this supervision tends to



Supplementary Figure S4: **Energy errors induced by Hamiltonian and density matrix perturbations**. Results are evaluated on 500 randomly selected molecules from the `stable-random` split of QH9. **(a)** Results of density matrix perturbations. AO pairs emphasized by the density gradient are selected according to $|g_{\alpha\beta} - \bar{g}| > 0.5\sigma_g$, where $g_{\alpha\beta}$ denotes the corresponding value in $\overline{\nabla_{\mathbf{P}}\rho}$, and $\bar{g}$ and σ_g denote its mean and standard deviation, respectively. Uniform noise scaled by $0.01\sigma_{\mathbf{P}}$ is then added to the selected off-diagonal elements, where $\sigma_{\mathbf{P}}$ denotes the standard deviation of density matrix entries. As a random baseline, we use the same noise matrix but randomly permute its element locations while preserving matrix symmetry, thereby keeping the perturbation strength and sparsity unchanged while removing the gradient-guided structure. **(b)** Results of Hamiltonian matrix perturbations. AO pairs emphasized by the orbital energy gradient $\overline{\nabla_{\mathbf{H}}\boldsymbol{\varepsilon}_{\text{occ}}}$ are selected analogously to panel **(a)**, and the random baseline is constructed using the same procedure.

assign large weights to density matrix elements associated with near-nuclear basis functions. These regions are already strongly emphasized by element-wise Hamiltonian supervision due to their large matrix magnitudes, making additional supervision on them largely redundant for learning valence-dominated chemical variations. This trend is also reflected in the qualitative gradient analysis in Supplementary Fig. S1b. Moreover, the gradients induced by one-electron energy losses exhibit excessively large numerical ranges and variances, leading to unbalanced gradient allocation (Supplementary Fig. S1f–h) and thereby limiting effective learning in some chemically important regions. The issue is particularly severe for the kinetic energy loss: many interatomic or off-diagonal kinetic energy matrix elements are close to zero (Supplementary Fig. S1f), resulting in sparse gradients that provide limited supervision for orbital interactions. Consequently, despite their clear physical interpretation and low computational cost, these one-electron energy losses fail to provide consistently effective learning signals and ultimately lead to inferior molecular property prediction performance (Fig. 2a).

B.2 Model Specification

In this section, we first provide the implementation details of the neural network models employed in the main paper (Supplementary Sec. B.2.1), then detail the configurations of training hyperparameters for the reproducibility of our empirical results (Supplementary Sec. B.2.3), and finally discuss the computational complexity of the DSHam framework in Supplementary Sec. B.2.4.

B.2.1 Model Configuration

Here we provide the configuration details of the neural network models used for Hamiltonian prediction, electron density prediction and physical property prediction. All models follow the settings specified in their official implementations, with task-specific adjustments when necessary. We summarize key hyperparameters below. All neural network models in this paper are implemented using PyTorch [60] and PyTorch Geometric [72].

Hamiltonian Prediction Models. To respect the geometric symmetry of the Hamiltonian, which is invariant to translations and equivariant to rotations in the AO basis, we build DSHam on the QHNet architecture, a widely used SEGNN for Hamiltonian prediction [30]. QHNet consists of four main modules: node-wise interaction, diagonal pair , non-diagonal pair and expansion. Given atom types $\mathcal{Z}$ and nuclear positions $\mathbf{X}$ as inputs, QHNet first applies five node-wise interaction layers to extract $\mathrm{SE}(3)$-equivariant atomic features. The features of diagonal and non-diagonal atom pairs are then processed by the corresponding pair modules to construct pairwise representations $\mathbf{h}_{aa}$ and $\mathbf{h}_{ab}$, respectively, where a and b index atoms. Finally, the expansion module maps these pairwise representations to blocks of the Hamiltonian matrix. Further details can be found in the original paper [30]. To evaluate the generality of

Supplementary Table S3: **Hyperparameters of Hamiltonian learning backbones for different datasets.** The top and bottom panels summarize the model configurations and training setups, respectively. $L_{\max}$ denotes the maximum degree of irreducible spherical harmonic representations used in the equivariant features. For SPHNet, both sparse node interactions and sparse tensor products are disabled because they degrade downstream property performance in our experiments.

Hyperparameter	MD17 (QHNet)	MD (QHNet)	QH9 (QHNet)	QH9 (SPHNet)	∇^2DFT (QHNet)
interaction layers	5	5	5	4	5
$L_{\max}$	4	4	4	4	4
hidden size	128	128	128	128	128
bottle hidden size	32	32	32	32	32
reference basis set	def2-SVP	def2-SVP	def2-SVP	def2-SVP	def2-SVP
batch size	16	16	8	32	4
gradient clipping	1.0	1.0	5.0	5.0	N/A
training steps	200k	200k	260k	200k	160k
precision	FP64	FP32	FP32	FP64	FP32
initial LR	5×10^{-4}	5×10^{-4}	5×10^{-4}	1×10^{-3}	5×10^{-4}
warmup steps	1,000	1,000	1,000	1,000	1,000
LR schedule	Polynomial	Polynomial	Polynomial	Polynomial	Polynomial

Supplementary Table S4: **Hyperparameters of the e2e predictors.** The top and bottom panels summarize the model configurations and training setups, respectively. $L_{\max}$ denotes the maximum degree of irreducible spherical harmonic representations used in the equivariant features.

Hyperparameters	EquiformerV2	MACE	GotenNet
interaction layers	12	2	6
hidden size	128	256	128
$L_{\max}$	6	3	2
batch size	8	8	8
training steps	260k	260k	260k
precision	FP32	FP32	FP32
gradient clipping	5.0	10.0	10.0
initial LR	5×10^{-4}	5×10^{-4}	5×10^{-4}
warm up steps	1,000	1,000	1,000
LR scheduler	Polynomial	Polynomial	Polynomial

DSLoss, we also use the recently proposed SPHNet as an alternative backbone architecture [40]. SPHNet introduces two sparse gates to selectively suppress noncritical interaction combinations, thereby improving computational efficiency. In our preliminary analysis, we found that sparse Hamiltonian predictors can fail to preserve global electronic consistency, leading to degraded accuracy in density matrices and total energies. Therefore, we disable sparsification in our implementation by setting the sparsity degree to zero. Both of these two architectures are implemented based on their official codebases[1], and the detailed configurations for different datasets are summarized in Supplementary Table S3.

Density prediction models. To compare with density learning methods, we adopt SCDP[2], a state-of-the-art (SOTA) model for electron density prediction [46]. SCDP represents charge density using atomic and virtual orbitals and predicts the corresponding orbital coefficients with the equivariant eSCN architecture [73]. Due to its substantial GPU memory requirements, SCDP is restricted to a batch size of 4 on the MD dataset and 1 on the QH9 dataset under our experimental settings. All SCDP models are trained with virtual nodes and without the finetuning stage. The model configurations are reported in Supplementary Table S5.

End-to-end property prediction models. To demonstrate the advantage of DSHam as a general property predictor, we compare it with three representative end-to-end (e2e) equivariant property predictors: EquiformerV2 [15], MACE [18] and GotenNet [19]. We adopt their architectures following the origi-

[1]The codebases are released at https://github.com/divelab/AIRS/tree/main/OpenDFT/QHNet and https://github.com/microsoft/SPHNet, which are available under the terms of the GPL-3.0 and MIT licenses, respectively.

[2]https://github.com/kyonofx/scdp, the code is available under the terms of the MIT license.

Supplementary Table S5: **Hyperparameters of the SCDP model.** The top and bottom panels summarize the model configurations and training setups, respectively. $L_{\max}$ and $m_{\max}$ denote the maximum degree and maximum magnetic quantum number of irreducible spherical harmonic representations used in the equivariant features, respectively. β_{basis} denotes the exponent scaling factor that controls the size of the atomic orbital basis set.

Hyperparameter	QH9	MD
interaction layers	8	4
$L_{\max}$	6	3
$m_{\max}$	2	2
hidden size	128	128
β_{basis}	1.3	1.5
reference basis set	def2-QZVPPD	def2-QZVPPD
batch size	1	4
grid level	2	3
training steps	500k	250k
precision	FP32	FP32
gradient clipping	0.5	0.5
initial LR	1×10^{-3}	1×10^{-3}
warm up steps	1,000	1,000
LR scheduler	Exponential	Exponential

nal codebases[3], while aligning the training protocols with our experimental settings; see Supplementary Table S4 for configuration details. To adapt MACE for orbital energy prediction, we replace its default summation readout with an averaging readout, which produces size-normalized molecular-level predictions and improves performance.

B.2.2 Atomic Orbital Basis Specification

As discussed in Methods 4.2, quantum tensors expressed in an AO basis, such as the Hamiltonian matrix, inherit the angular momentum structure of the underlying basis functions. For example, in the def2-SVP basis set, the basis functions for carbon are specified as `3s2p1d`, meaning that the corresponding expansion contains three type-0, two type-1 and one type-2 tensor channels. The detailed orbital compositions used for different atom types and basis settings are summarized in Supplementary Table S6. To rigorously preserve geometric equivariance, SEGNN-based predictors need to construct equivariant atomic pair features that cover the maximum rotation order of the corresponding basis functions and their tensor product couplings in equivariant message passing; see $L_{\max}$ in Supplementary Table S3 and Supplementary Table S5. Notably, the orbital composition assigned to the same element may differ across methods, depending on the basis used for Hamiltonian representation or density fitting. Even when two elements share the same nominal orbital composition, their actual basis functions can still differ because of different exponents and contraction coefficients. In our experiments, methods built on the same backbone and target representation use identical orbital compositions to ensure fair comparison.

B.2.3 Training Hyperparameters

We first summarize the training protocols and loss configurations adopted in DSHam. The hyperparameters associated with the dual-space loss are listed in Supplementary Table S7. Compared with QH9 and ∇^2DFT, molecules in the MD and MD17 datasets generally contain fewer heavy atoms and exhibit less extensive electronic delocalization. As discussed in Supplementary Sec. B.1.4, the real-space density loss primarily emphasizes local σ-bonding regions, which constitute a larger fraction of the electronic structure in these smaller molecules. Accordingly, we reduce the relative weights of density matrix and orbital energy supervision in DSLoss and place greater emphasis on the real-space density term. For other molecular property-based auxiliary losses, the weighting coefficients are selected empirically according to training stability and validation performance (Supplementary Table S18). Other training hyperparameters are provided in Supplementary Tables S3-S5.

[3]The codebases of the three models are released at https://github.com/atomicarchitects/equiformer_v2, https://github.com/ACEsuit/mace and https://github.com/sarpaykent/GotenNet, respectively. All the code is available under the terms of the MIT license.

Supplementary Table S6: **Atomic orbital compositions for the def2-QZVPPD and def2-SVP basis sets.** For SCDP, virtual nodes are assigned the same atomic orbital basis as oxygen, following the original setting.

Atom type	def2-QZVPPD (SCDP)	def2-SVP (QHNet/SPHNet)
H	4s3p2d	2s1p
C	5s4p3d2f	3s2p1d
N	5s4p3d2f	3s2p1d
O	5s4p3d2f	3s2p1d
F	5s4p3d2f	3s2p1d
Virtual node	5s4p3d2f	N/A
S	N/A	4s3p1d
Cl	N/A	4s3p1d
Br	N/A	5s4p3d

Supplementary Table S7: **Loss combination weights of DSLoss for different datasets.**

Dataset	$\lambda_{\boldsymbol{\varepsilon}}$	$\lambda_{\mathbf{P}}$	λ_{ρ}
MD	0.001	0.001	0.02
MD17	0.001	0.001	0.02
QH9	0.02	0.03	0.02
∇^2DFT	0.02	0.03	0.02

B.2.4 Computational Complexity

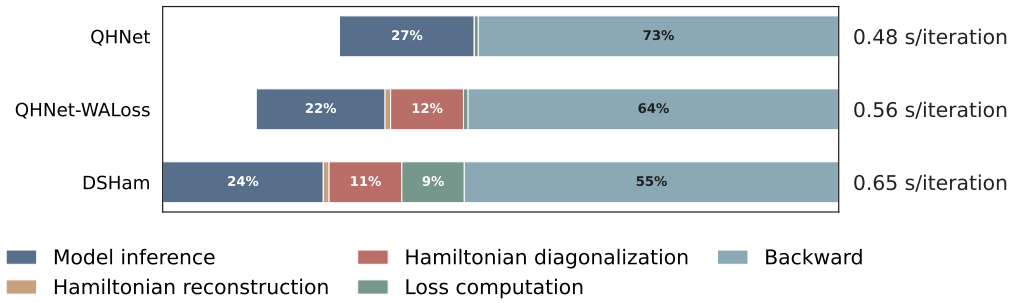
As described in Methods 4.1, evaluating DSLoss requires constructing several additional physical observables, which introduces extra computational overhead. All observables in our framework are computed from the orbital coefficients $\hat{\mathbf{C}}$ or orbital energies $\hat{\varepsilon}$ derived from the predicted Hamiltonian $\hat{\mathbf{H}}$, and the dominant cost comes from solving the generalized eigenvalue problem in Eq. (1). For a basis set containing B atomic orbitals, this eigendecomposition scales as $O(B^3)$. Once the orbital coefficients are obtained, the density matrix can be constructed from the occupied MOs with complexity $O(B^2K)$, where K is the number of occupied orbitals; since K scales linearly with B in typical molecular systems, this cost is $O(B^3)$. Evaluating the real-space electron density on N_G spatial grid points requires computing AO-pair contributions on the grid, leading to a complexity of $O(N_G B^2)$. In practice, both the basis size B and the number of grid points N_G scale approximately linearly with the system size N, so the overall asymptotic complexity of evaluating DSLoss is $O(N^3)$. Because these additional computations mainly consist of highly parallel matrix operations, they can be efficiently accelerated on modern GPUs.

As discussed in Results 2.4, we further analyze the practical training cost by comparing the per-iteration runtime breakdown of DSHam with two counterparts using the same architecture: QHNet and QHNet-WALoss. All models are evaluated using a density grid level of 2 under the same hardware configuration (Supplementary Sec. B.6). Runtime statistics are averaged over 500 training iterations on the QH9 dataset for each model. As shown in Supplementary Fig. S5, standard QHNet requires 0.48s per iteration, whereas QHNet-WALoss and DSHam require 0.56s and 0.65s per iteration, corresponding to training overheads of approximately 17% and 35%, respectively. For both QHNet-WALoss and DSHam, the additional cost mainly arises from Hamiltonian diagonalization, which accounts for around 11% of the total per-iteration training time. During Hamiltonian prediction at inference time, DSHam and QHNet-WALoss have the same computational cost as standard QHNet, since the additional loss terms are used only during training.

B.3 Data Preparation

This section provides supplementary information on the datasets used in our experiments, including dataset statistics and train/validation/test split settings. For MD17, QH9 and ∇^2DFT, we adopt the original data splits defined in their respective works [41, 43, 44]. Detailed statistics for all datasets and splits are summarized in Supplementary Table S8.

MD17 consists of molecular dynamics trajectories of small molecules, including water, ethanol, malondialdehyde and uracil. The MD dataset is used for comparison with the SCDP baseline and contains molecular dynamics trajectories for density prediction. For this dataset, we perform additional electronic-structure calculations to generate Hamiltonian matrices using the def2-SVP basis set, enabling direct



Supplementary Figure S5: **Per-iteration training runtime breakdown of different methods.** The total length of each right-aligned bar represents the overall training time per iteration, while the colored segments indicate the relative time spent on model inference, Hamiltonian reconstruction, Hamiltonian diagonalization, loss computation and backpropagation.

Supplementary Table S8: **Dataset statistics.**

Dataset	Train	Validation	Test
MD17-water	500	500	3,900
MD17-ethanol	25,000	500	4,500
MD17-malondialdehyde	25,000	500	1,478
MD17-uracil	25,000	500	4500
MD-ethanol	950	50	1,000
MD-benzene	950	50	1,000
MD-phenol	950	50	1,000
MD-resorcinol	950	50	1,000
MD-ethane	1,900	100	2,000
MD-malondialdehyde	1,900	100	2,000
QH9-stable-random	104,664	13,083	13,084
QH9-stable-ood	104,001	17,495	9,335
QH9-dynamic-geo	239,840	29,980	29,980
QH9-dynamic-mol	239,840	29,900	30,100
∇^2DFT	44,752	4,973	9,345

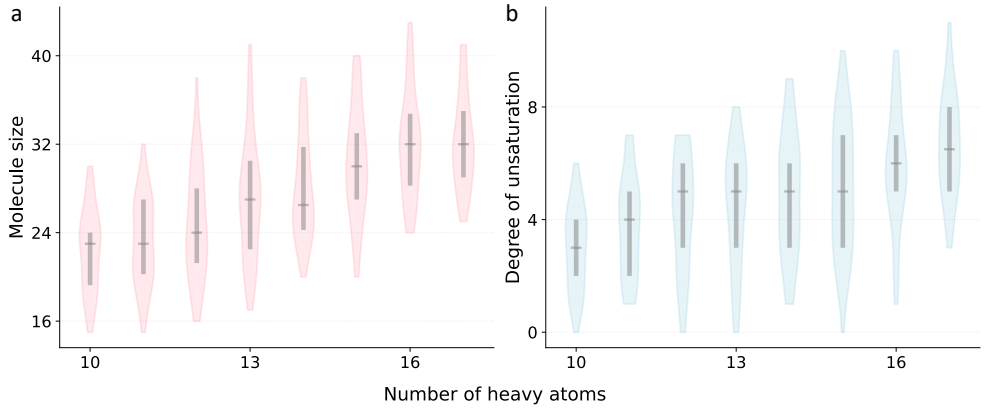
comparison between Hamiltonian learning and electron density prediction methods. QH9 is a quantum chemistry dataset constructed from molecules containing only `H/C/N/O/F` elements, with no more than nine heavy atoms and provides multiple evaluation settings, including `stable-random` and `stable-ood` for chemical space evaluation, as well as `dynamic-geo` and `dynamic-mol` for conformation space evaluation. For the extrapolation experiments described in Results 2.3, we construct a custom benchmark derived from PubChemQC, as described in Methods 4.3. The statistical distribution of this benchmark is shown in Supplementary Fig. S6.

B.4 MD Simulation Details

As described in Results 2.5, we evaluate the practical utility of DSHam through MD simulations. Here, we provide details of the test systems, simulation settings, and evaluation protocols.

Simulated molecular systems. We perform MD simulations on five molecules selected from the medium-sized ∇^2DFT test split. Their molecular formulas and dataset identifiers are summarized in Supplementary Table S9.

MD simulation setup. The initial geometries were taken directly from the dataset without prior geometry optimization. MD simulations were carried out using the Atomic Simulation Environment (ASE) [74], with either a Hamiltonian prediction model or DFT serving as the calculator that supplied energies and atomic forces at each integration step. For each molecule, the same set of initial velocities was sampled from a Maxwell–Boltzmann distribution at 300 K, rescaled to exactly 300 K, and used to initialize both the model- and DFT-driven simulations. The resulting trajectories were then propagated independently in the canonical ensemble using the Langevin integrator implemented in ASE, with a target temperature of 300 K, a time step of 0.5 fs, and a friction coefficient of 0.01 fs^{-1}.



Supplementary Figure S6: **Statistics of the selected PubChemQC molecules. (a)** Distribution of molecular sizes. **(b)** Distribution of molecular unsaturation degrees. Shaded violin envelopes represent kernel density estimates. Internal bars indicate medians, and thick vertical lines denote interquartile ranges.

Supplementary Table S9: **Molecular systems used in the MD experiments.**

Test index	Molecular formula	MOSES ID
10	$C_{14}H_{15}N_3O_2S$	1582087
100	$C_{18}H_{20}N_2O_3S$	360897
3000	$C_{20}H_{17}NO_4$	1647689
7000	$C_{13}H_{15}N_3O_2S$	619820
7500	$C_{10}H_{14}N_2O_4S_2$	137258

Single-point evaluation. To disentangle errors in the predicted Hamiltonian from the divergence of independently propagated trajectories, all energy and force comparisons were performed on the same molecular geometries. For Fig. 5c, conformations were sampled every 1 ps from the 20-ps trajectory generated by DSHam. Molecular structures and force vectors are visualized at 5-ps intervals, with force arrows shown only for heavy atoms. For each sampled conformation $\mathcal{M}^t$, the relative energy from calculator w ($w \in \{\mathrm{DSHam}, \mathrm{QHNet}, \mathrm{DFT}\}$) is defined using the DSHam energy of the initial conformation as a common reference:

$$E^w_{\mathrm{rel}}\left[\mathcal{M}^t\right] = E_w\left[\mathcal{M}^t\right] - E_{\mathrm{DSHam}}\left[\mathcal{M}^0\right].$$

For Fig. 5b, we performed a complementary evaluation using structures sampled from five 1.5-ps trajectories generated by DSHam. Conformations were extracted every 150 fs from each trajectory.

Structural validity. We also monitored the structural validity of every conformation along all simulation trajectories. Following Ref. [58], a conformation $\mathbf{x}(t)$ sampled at time t was classified as structurally unstable if the maximum deviation of any monitored bond length from its reference value exceeded a prescribed threshold:

$$\max_{(a,b)\in\mathcal{B}} \left| \|\mathbf{x}_a(t) - \mathbf{x}_b(t)\|_2 - e^{(0)}_{ab} \right| > \delta_{\mathrm{bond}},$$

where $\mathcal{B}$ denotes the set of bonds included in the evaluation, $\mathbf{x}_a(t)$ is the Cartesian position of atom a at simulation time t, and $\delta_{\mathrm{bond}} = 0.5\,\text{Å}$. $e^{(0)}_{ab} = \|\mathbf{x}_a(0) - \mathbf{x}_b(0)\|_2$ is the corresponding bond length in the initial equilibrium conformation. No conformation along any of the five DSHam-generated trajectories exceeded this threshold.

Vibrational density of states. Following previous work, we assess the quality of DSHam-generated trajectories by comparing their vibrational density of states (VDOS) with the DFT reference [75, 76]. The VDOS shown in Fig. 5d was computed from trajectories of $C_{18}H_{20}N_2O_3S$ generated independently by DSHam and DFT. At each time step, we remove the center-of-mass motion from the atomic velocities:

$$\widetilde{\mathbf{v}}_a(t) = \mathbf{v}_a(t) - \frac{\sum_b M_b \mathbf{v}_b(t)}{\sum_b M_b},$$

where M_a and $\mathbf{v}_a(t)$ denote the mass and instantaneous Cartesian velocity vector of atom a, respectively.

For each atom a and Cartesian component $\sigma \in \{x, y, z\}$, the velocity power spectral density was estimated using Welch's method with a Hann window [77]. The mass-weighted spectrum was then constructed as

$$I(f) = \sum_a M_a \sum_{\sigma \in \{x,y,z\}} P_{a\sigma}(f),$$

where f is the linear frequency and $P_{a\sigma}(f)$ denotes the Welch power spectral density of the center-of-mass-corrected velocity component $\widetilde{\mathrm{v}}_{a\sigma}(t)$. Frequencies were converted to wavenumbers according to $\widetilde{\nu} = \frac{f}{c}$, where $c = 2.99792458 \times 10^{10}$ cm/s. We retained the range $0 \le \widetilde{\nu} \le \widetilde{\nu}_{\max}$, with $\widetilde{\nu}_{\max} = 4000\ \mathrm{cm}^{-1}$. Any residual negative values arising from numerical roundoff were clipped to zero. For each trajectory $w \in \{\mathrm{DSHam}, \mathrm{DFT}\}$, the resulting spectrum was normalized to unit area over this interval:

$$\widehat{I}_w(\widetilde{\nu}) = \frac{I_w(\widetilde{\nu})}{\displaystyle\int_0^{\widetilde{\nu}_{\max}} I_w(\widetilde{\nu}')\mathrm{d}\widetilde{\nu}'},$$

where $\widehat{I}_w$ denotes the normalized VDOS obtained from the trajectory generated by method w. The reported VDOS overlap was calculated as

$$\mathcal{O}_{\mathrm{VDOS}} = \int_0^{\widetilde{\nu}_{\max}} \min\left[\widehat{I}_{\mathrm{DSHam}}(\widetilde{\nu}), \widehat{I}_{\mathrm{DFT}}(\widetilde{\nu})\right] \mathrm{d}\widetilde{\nu}.$$

Because both spectra are non-negative and normalized to unit area, $\mathcal{O}_{\mathrm{VDOS}} \in [0, 1]$ and is reported as a percentage.

B.5 DFT Implementation Details

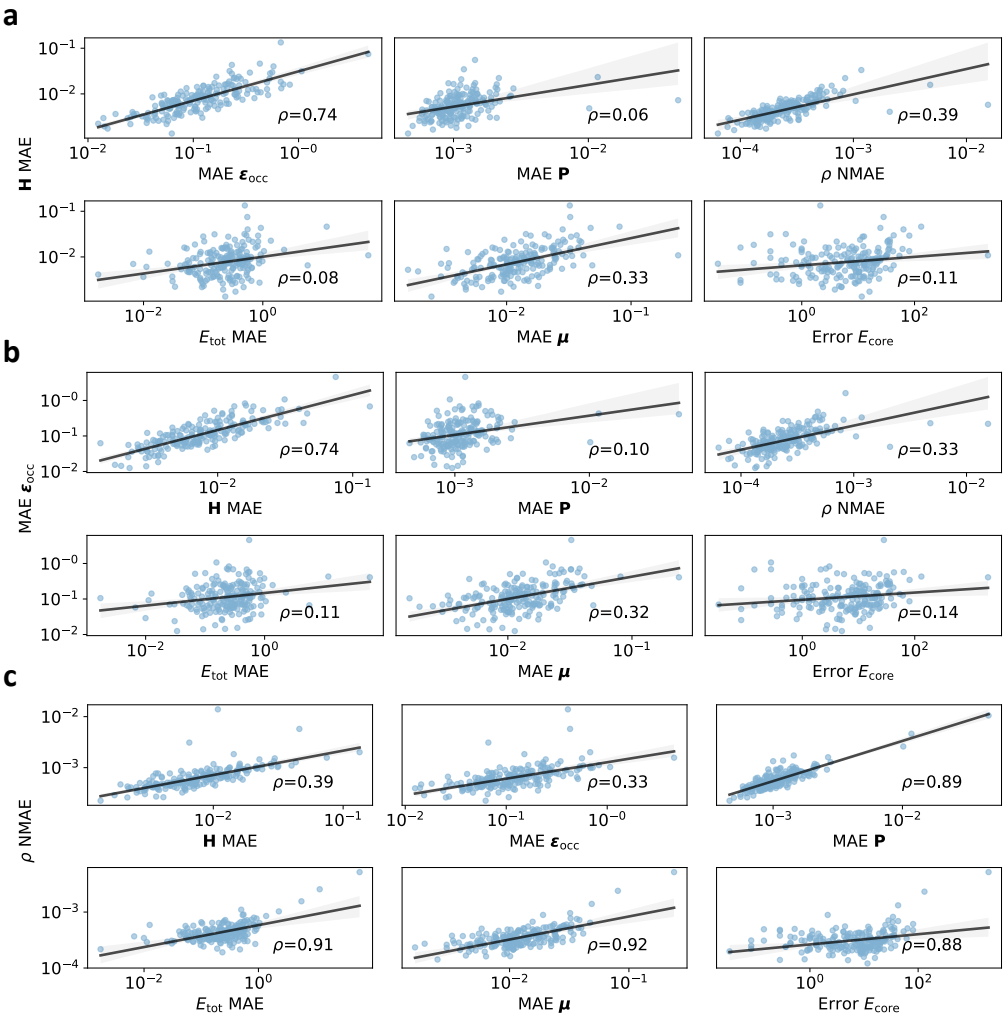
All DFT calculations, including those for dataset generation (Methods 4.3) and evaluation of the SCF acceleration ratio (Methods 4.4), are performed using PySCF [61]. Unless otherwise specified, all calculations use the default PySCF settings. The SCF energy convergence threshold is set to 10^{-8} Hartree.

B.6 Hardware Configurations

To facilitate reproducibility and provide context for runtime benchmarks, we summarize the hardware configurations used in this work. All neural network models are trained and evaluated on a workstation equipped with an NVIDIA A100 GPU with 80 GiB memory and a 24-core CPU. DFT calculations are conducted on the same platform. For consistency, all runtime and GPU memory measurements reported in Fig. 4b are benchmarked under this hardware configuration.

Supplementary Table S10: **Description of evaluation metrics.**

Symbol	Description
$\mathbf{H}$	Hamiltonian matrix.
$\varepsilon_{\mathrm{occ}}$	Occupied molecular orbital energy.
$\mathbf{C}$	Molecular orbital coefficient matrix.
$\varepsilon_{\mathrm{HOMO}}$	Energy of the highest occupied molecular orbital.
$\varepsilon_{\mathrm{LUMO}}$	Energy of the lowest unoccupied molecular orbital.
ε_{Δ}	HOMO–LUMO energy gap, defined as the difference between $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$.
$\mathbf{P}$	Density matrix.
$\rho(\mathbf{r})$	Real-space electron density evaluated on numerical spatial grids.
T_{S}	Non-interacting kinetic energy.
E_{ext}	External potential energy of electrons under the nuclear attraction field.
E_{core}	One-electron energy, including T_{S} and E_{ext}.
E_{tot}	Total molecular energy.
$\boldsymbol{\mu}$	Total molecular dipole moment.
S_{acc}	SCF acceleration ratio when using model outputs as the initial guess, computed relative to the semi-empirical initial guess.



Supplementary Figure S7: **Scatter plots of Hamiltonian, orbital energy and electron density errors against representative derived property errors. (a–c)** Correlations between representative derived property errors and errors in the Hamiltonian, orbital energies and electron density, respectively, extending the analysis in Fig. 1e–g. The analysis is performed on 200 randomly sampled QH9 molecules. Black lines show linear fits in log–log space, with Pearson correlation coefficients annotated. Shaded regions indicate 95% confidence intervals.

B.7 Correlation between Property Errors

In Fig. 1b, we analyze the Pearson correlation between prediction errors for different molecular properties. The Pearson correlation coefficient is defined as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_k (y_k - \bar{y})^2}},$$

where x_i and y_i denote the prediction errors of sample i for properties x and y, respectively, and $\bar{x}$ and $\bar{y}$ denote the corresponding mean errors across all samples. In addition to the correlation coefficients between different molecular property errors, we provide the detailed sample distributions in Supplementary Fig. S7.

C Additional Empirical Results

This section provides supplementary empirical results for the experiments reported in Results 2. We first summarize the evaluation metrics used throughout this work (Supplementary Sec. C.1), then provide comprehensive results across datasets and tasks (Supplementary Secs. C.2-C.3), and finally present ablation studies in Supplementary Sec. C.6.

C.1 Evaluation Metrics

Before presenting the detailed experimental results, we summarize the evaluation metrics used in this work. The complete list of molecular properties and electronic structure quantities is provided in Supplementary Table S10. Unless otherwise specified, all quantities are evaluated using the mean absolute error (MAE) with respect to the KSDFT reference. There are two exceptions: the orbital coefficient matrix $\mathbf{C}$, which is evaluated using cosine similarity $\mathbf{C}_{\text{sim}}$, and the real-space electron density ρ, which is evaluated using normalized MAE (NMAE). Cosine similarity $\mathbf{C}_{\text{sim}}$, density NMAE and the SCF acceleration ratio S_{acc} are reported as percentages (%). For the remaining quantities, energy-related errors are reported in units of kcal/mol and 10^{-6} Hartree (μE_h) in the main text and supplementary information, respectively; density matrix errors are reported in units of 10^{-6}; and dipole moment errors are reported in units of 10^{-6} Debye (μD).

C.2 Performance on Molecular Systems

As discussed in Results 2.2, DSHam achieves superior prediction accuracy for downstream physical observables across both conformational space (MD17) and chemical space (QH9 and ∇^2DFT). Here, we report quantitative results covering a broader set of evaluation metrics, molecular systems and data split settings. Results on additional MD17 molecules are listed in Supplementary Table S11, where DSHam outperforms QHNet across molecular benchmarks and almost all metrics. Moreover, as molecular complexity increases, the performance gap between DSHam and QHNet becomes increasingly pronounced, consistent with the trend observed in Fig. 3a.

As described in Supplementary Sec. B.3, QH9 provides four split settings for evaluating generalization across chemical space (`stable-random` and `stable-ood`) and conformational space (`dynamic-geo` and `dynamic-mol`). The results in Supplementary Table S12 show that DSHam maintains its advantage over QHNet across all split settings. When transferred to the more complex ∇^2DFT dataset, DSHam continues to provide accurate downstream physical quantities, whereas the two baseline methods exhibit substantially degraded performance. In particular, the total energy MAEs of both baselines are three orders of magnitude larger than that of DSHam. These results further demonstrate the advantage of DSLoss in aligning the predicted Hamiltonian with downstream physical fidelity.

C.3 Extrapolation Results

We provide supplementary results for the extrapolation experiments in Results 2.3. The complete extrapolation results for all physical quantities are shown in Supplementary Fig. S8. All models are trained on the `stable-random` split of QH9 and evaluated on larger PubChemQC molecules containing 10–17 heavy atoms. For each quantity, results are grouped by the number of heavy atoms N_{heavy}. Consistent with Results 2.3, DSHam exhibits strong extrapolation performance across all physical quantities, benefiting from dual-space supervision.

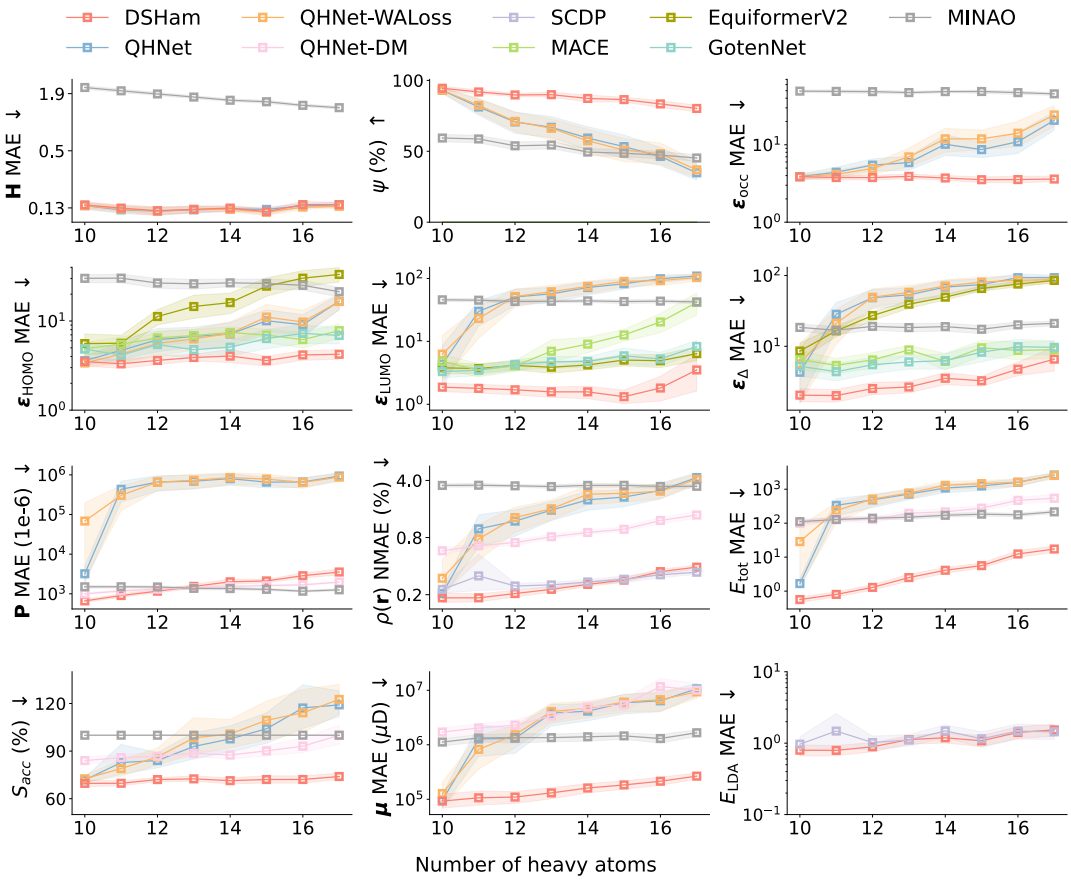
Although QHNet-DM shows reasonable robustness in density matrix prediction, its predicted density matrices do not explicitly satisfy physical constraints such as electron-number conservation and generalized idempotency, which limits the accuracy of derived energies. Similarly, SCDP directly constrains

Supplementary Table S11: **Results on the MD17 dataset, as an extension to the results narrated in Results 2.2**. Results are computed over 3,900, 4,500, 1,478 and 4,500 test structures of each MD17 molecule, respectively.

Dataset	Methods	$\mathbf{H}$ (μE_h)	$\mathbf{C}_{\text{sim}}$ (%)	$\boldsymbol{\varepsilon}_{\text{occ}}$ (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)
Water	QHNet	2.90	99.99	**12.92**	5.29	0.0014	0.02	**42.48**
	DSHam	**2.35**	99.99	13.37	**4.81**	**0.0013**	0.02	42.53
Ethanol	QHNet	7.38	99.99	**35.95**	64.47	0.0085	1.15	42.74
	DSHam	**7.25**	99.99	36.23	**51.03**	**0.0072**	**0.79**	**42.54**
Malondial-dehyde	QHNet	**5.41**	99.98	**25.03**	103.84	0.0068	1.70	43.96
	DSHam	5.72	99.98	25.93	**84.59**	**0.0057**	**1.36**	**43.04**
Uracil	QHNet	**5.54**	**99.98**	42.45	142.62	0.0095	6.84	39.80
	DSHam	5.67	99.97	**41.28**	**113.20**	**0.0074**	**4.50**	**38.43**
Dataset	Methods	$\boldsymbol{\varepsilon}_{\text{HOMO}}$ (μE_h)	$\boldsymbol{\varepsilon}_{\text{LUMO}}$ (μE_h)	$\boldsymbol{\varepsilon}_{\Delta}$ (μE_h)	$\boldsymbol{\mu}$ (μD)	E_{core} (μE_h)	T_{S} (μE_h)	E_{ext} (μE_h)
Water	QHNet	**9.92**	16.29	**15.15**	93.93	**211.85**	**202.55**	**412.15**
	DSHam	10.98	**15.75**	16.75	**81.64**	215.28	208.25	421.27
Ethanol	QHNet	**38.21**	95.64	96.84	1502.74	1245.05	798.39	1958.99
	DSHam	39.26	**86.43**	**87.38**	**1285.62**	**1070.09**	**732.94**	**1743.96**
Malondial-dehyde	QHNet	**19.84**	17.52	20.32	1911.92	1076.93	793.16	1800.31
	DSHam	20.82	**18.63**	**19.92**	**1676.66**	**987.64**	**723.99**	**1651.59**
Uracil	QHNet	**39.48**	**29.02**	41.59	3696.57	1735.87	1298.66	2774.46
	DSHam	39.89	29.31	**40.07**	**3063.92**	**1486.09**	**1079.71**	**2377.38**

Supplementary Table S12: **Results on the QH9 dataset, as an extension to the results narrated in Results 2.2.** Results are computed over 13,084, 9,335, 29,980 and 30,100 QH9 test molecules from the `stable-random`, `stable-ood`, `dynamic-geo` and `dynamic-mol` splits, respectively.

Dataset	Methods	$\mathbf{H}$ (μE_h)	$\mathbf{C}_{\text{sim}}$ (%)	$\boldsymbol{\varepsilon}_{\text{occ}}$ (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)
stable-random	QHNet	**38.44**	98.76	275.52	16678.08	0.0604	455.51	60.84
	DSHam	46.84	**99.24**	**192.75**	**384.37**	**0.0266**	**104.40**	**57.44**
stable-ood	QHNet	**33.50**	98.00	153.74	1306.77	0.0466	163.48	59.52
	DSHam	40.60	**98.31**	**148.04**	**336.92**	**0.0267**	**67.96**	**58.73**
dynamic-geo	QHNet	**51.05**	98.26	325.35	16673.62	0.0919	938.01	64.37
	DSHam	65.11	**99.11**	**271.93**	**428.51**	**0.0388**	**147.51**	**59.25**
dynamic-mol	QHNet	**68.94**	96.32	884.97	48886.73	0.1761	896.74	63.10
	DSHam	78.98	**98.22**	**536.44**	**685.74**	**0.0596**	**205.86**	**60.00**
Dataset	Methods	$\boldsymbol{\varepsilon}_{\text{HOMO}}$ (μE_h)	$\boldsymbol{\varepsilon}_{\text{LUMO}}$ (μE_h)	$\boldsymbol{\varepsilon}_{\Delta}$ (μE_h)	$\boldsymbol{\mu}$ (μD)	E_{core} (μE_h)	T_{S} (μE_h)	E_{ext} (μE_h)
stable-random	QHNet	362.97	1627.63	1824.48	24836.92	33489.75	8315.24	39140.70
	DSHam	**256.42**	**1189.70**	**1316.70**	**10595.87**	**6738.70**	**4074.39**	**9960.43**
stable-ood	QHNet	218.43	**968.71**	**1037.64**	13774.80	14197.74	5274.90	16125.89
	DSHam	**170.98**	2291.69	2259.71	**8437.57**	**7081.01**	**3587.67**	**9764.66**
dynamic-geo	QHNet	495.61	2536.39	2726.68	48884.23	78882.55	27486.70	105753.59
	DSHam	**359.56**	**1488.64**	**1689.33**	**15113.99**	**9977.72**	**5905.88**	**14854.85**
dynamic-mol	QHNet	855.81	6673.22	6695.07	133227.26	184003.34	39460.36	221944.56
	DSHam	**524.77**	**1688.68**	**1943.32**	**31927.62**	**17462.77**	**9038.69**	**24675.04**



Supplementary Figure S8: **Results on the selected PubChemQC molecules, as an extension to the results narrated in Results 2.3.** All models are trained on QH9 molecules containing no more than nine heavy atoms and evaluated on PubChemQC molecules with increasing molecular size. Each value is computed over 50 PubChemQC molecules. Shaded regions denote the 95% confidence intervals of the mean errors.

Supplementary Table S13: **Results on ∇^2DFT dataset, as an extension to the results narrated in Results 2.2.** Results are computed over 9,345 ∇^2DFT test molecules.

Model	$\mathbf{H}$ (μE_h)	ε_{occ} (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)	$\boldsymbol{\mu}$ (μD)
QHNet	102.40	21118.75	837873.34	3.21	4653194.90	122.77	25011278.15
QHNet-WALss	**102.12**	20297.54	845204.17	3.26	4687292.17	120.70	24071252.82
DSHam	132.39	**907.64**	**820.29**	**0.08**	**8431.23**	**70.70**	**77604.87**

Supplementary Table S14: **Results on orbital energy prediction, as an extension to the results narrated in Results 2.3**. All `e2e` models are trained separately for each target property. The MAE of each property is listed in μE_h. Results are computed over 13,084 and 30,100 QH9 test molecules from the `stable-random` and `dynamic-mol` splits, respectively.

Dataset	Methods	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	ε_{Δ}
stable-random	DSHam	**256.42**	1189.70	**1316.70**
	EquiformerV2	899.78	**627.88**	1540.17
	MACE	2859.80	1689.62	3285.74
	GotenNet	1051.96	738.48	1592.44
dynamic-mol	DSHam	**524.77**	**1688.68**	**1943.32**
	EquiformerV2	5632.91	3560.54	8197.24
	MACE	6333.38	6081.95	8403.72
	GotenNet	6470.40	3554.21	9984.05

the real-space electron density but does not guarantee a physically consistent underlying electronic state. Moreover, projecting a real-space density back to the AO representation is generally ill-posed, further limiting its applicability for deriving Hamiltonian-based properties. These results highlight the advantage of using the Hamiltonian as a structured intermediate representation while incorporating dual-space supervision to improve physical consistency.

C.4 Comparison with End-to-End Predictors

As discussed in Results 2.3, DSHam leverages the advantage of Hamiltonian learning as a general property prediction framework and achieves substantially better extrapolation accuracy for orbital energy prediction on molecules beyond the training scale (Fig. 3c). To further evaluate the effectiveness of DSHam, we provide additional in-scale results on the QH9 dataset in Supplementary Table S14. The `stable-random` and `dynamic-mol` splits are used to benchmark generalization across chemical and conformational spaces, respectively. Following the comparison in Fig. 3, we compare DSHam with three representative `e2e` property predictors: EquiformerV2, MACE and GotenNet. All `e2e` models are trained separately for each target quantity, and their model configurations are kept the same as those used in Fig. 3; see Supplementary Table S4.

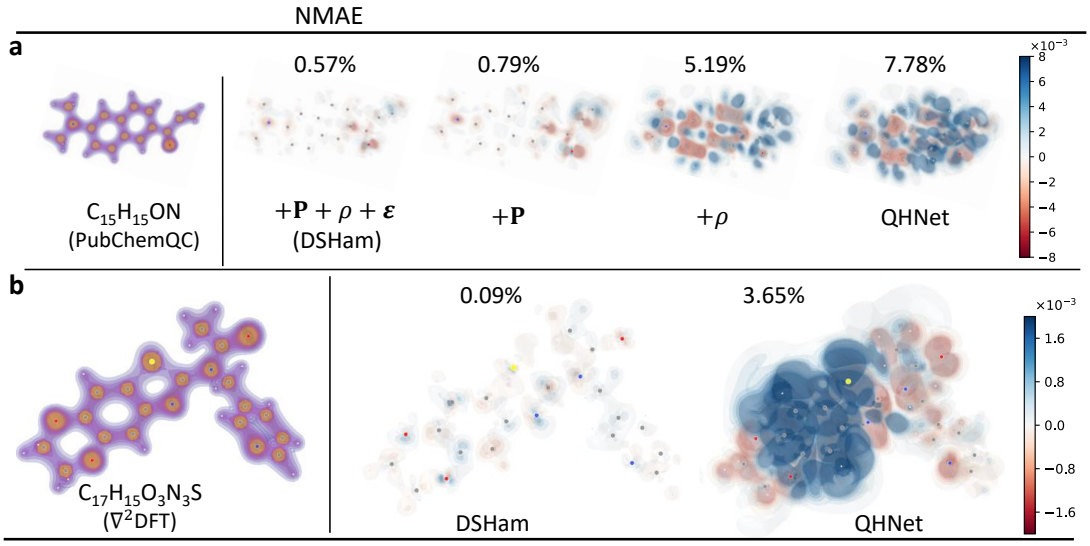
Consistent with the observations in Results 2.3, DSHam achieves substantially better accuracy across nearly all split settings and orbital energy properties. The only exception is $\varepsilon_{\text{LUMO}}$ on the `stable-random` split, where DSHam still attains performance competitive with `e2e` predictors. A possible explanation is that DSLoss directly supervises orbital energies only in the occupied space, while unoccupied orbital energies are constrained indirectly through the learned Hamiltonian. This interpretation is supported by the result that DSHam achieves an order-of-magnitude lower MAE for $\varepsilon_{\text{HOMO}}$ than the `e2e` baselines. Overall, these findings demonstrate the effectiveness of DSHam as a general Hamiltonian-based property prediction framework for generalization in both chemical and conformational spaces.

C.5 Comparison with Electron Density Learning Methods

As discussed in Results 2.4, DSHam outperforms direct electron density learning methods, highlighting the benefit of using the Hamiltonian as a structured intermediate representation. For a comprehensive evaluation, we report additional results on real-space density prediction in Supplementary Table S15 and density matrix prediction in Supplementary Table S16. Following the original implementation [46], we introduce virtual nodes in SCDP to enhance model expressivity. The fine-tuning stage is omitted because we observed severe training instability in preliminary experiments. To ensure fair comparison,

Supplementary Table S15: **Results for real-space density prediction, as an extension to the results narrated in Results 2.4.** Electron density errors are reported as NMAE percentages. The density is evaluated on Lebedev grids generated by PySCF with grid level 2. All methods are trained in single precision. For fair comparison, we additionally train SCDP to predict the residual between the reference density and the electron density obtained from the `MINAO` initial guess; this variant is denoted as SCDP*. By contrast, SCDP refers to the model directly supervised by the KSDFT density. Results are computed over 1,000, 1,000, 1,000, 1,000, 2,000, 2,000 and 13,084 test structures of each dataset, respectively.

Dataset	DSHam	QHNet	SCDP*	SCDP
MD-ethanol	**0.0127**	0.0168	0.0504	0.1030
MD-benzene	**0.0034**	0.0059	0.0122	0.0560
MD-phenol	**0.0064**	0.0101	0.1403	0.1504
MD-resorcinol	**0.0081**	0.0130	0.0297	0.0728
MD-ethane	**0.0049**	0.0153	0.0055	0.0759
MD-malonaldehyde	**0.0114**	0.0146	0.1697	0.4052
QH9	**0.0271**	0.0591	0.0890	0.2538



Supplementary Figure S9: **Visualization of predicted real-space densities for molecules sampled from PubChemQC and ∇^2DFT, as an extension to the results narrated in Results 2.4 and Fig. 4**. The left panels show the ground-truth density, and the right panels show the prediction error, with the NMAE indicated on each prediction.

all methods are trained using residual targets defined as the difference between the ground-truth labels and the corresponding quantities computed from the `MINAO` initial guess. As shown in Supplementary Table S15, DSHam achieves the best density NMAE across all datasets, consistent with the conclusion in Fig. 3.

The qualitative comparison in Fig. 4c shows that DSHam recovers real-space electron densities that more closely resemble the DFT references. To further examine this advantage, we provide visualizations for larger molecules in Supplementary Fig. S9. For the $C_{15}H_{15}ON$ molecule randomly selected from PubChemQC, we examine the contributions of individual components of DSLoss. QHNet exhibits substantial errors in regions dominated by AOs with higher principal quantum numbers. Adding $\mathcal{L}_\rho$ reduces the overall density error but is less effective in delocalized regions of the aromatic rings, consistent with its stronger sensitivity to AO pairs with substantial spatial overlap. In comparison, $\mathcal{L}_\mathbf{P}$ directly constrains the occupied subspace and complements $\mathcal{L}_\rho$ in capturing delocalized electronic contributions. Combining these two state-space signals substantially reduces the density across both localized and delocalized regions, including higher-shell contributions. For the more complex $C_{17}H_{15}O_3N_3S$ molecule randomly selected from ∇^2DFT, QHNet similarly exhibits substantially larger higher-shell errors than DSHam, consistent with its poor downstream property accuracy, whereas dual-space supervision markedly suppresses these errors.

Supplementary Table S16: **Results on density matrix prediction, as an extension to the results narrated in Results 2.4.** We additionally report molecular properties derived from the predicted density matrices. Results are computed over 13,084 QH9 test molecules. For fair comparison, QHNet-DM is trained to predict the residual between the reference density matrix and the density matrix obtained from the `MINAO` initial guess.

Methods	$\mathbf{P}$ (10^{-6})	$\rho(\mathbf{r})$ (%)	E_{tot} (μE_h)	S_{acc} (%)
DSHam	**384.37**	**0.0266**	**104.40**	**57.44**
QHNet	16678.08	0.0604	455.51	58.73
QHNet-DM	769.74	0.2601	44640.19	77.30

Supplementary Table S17: **Results of different auxiliary loss combinations, as an extension to the results shown in Fig. 2a.** $\lambda_{\mathbf{O}}$ denotes loss weight for observable $\mathbf{O}$. Bold and underlined values denote the best and second-best results, respectively. Results are computed over 13,084 QH9 test molecules. For each setting, $\lambda_{\mathbf{O}}$ is selected according to the overall validation performance across downstream properties. The detailed weighting coefficients for each auxiliary loss combination are provided in Supplementary Table S18.

No.	Loss	$\mathbf{H}$ (μE_h)	ε_{occ} (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)
1	$\mathcal{L}_{\mathbf{H}}$ (QHNet)	**38.44**	275.52	16678.08	0.0604	455.51	60.84
2	$\mathcal{L}_{\mathbf{H}} + \lambda_{\rho}\mathcal{L}_{\rho}$	<u>39.86</u>	194.16	698.77	<u>0.0268</u>	183.02	58.04
3	$\mathcal{L}_{\mathbf{H}} + \lambda_{\mathbf{P}}\mathcal{L}_{\mathbf{P}}$	44.30	230.40	555.21	0.0394	164.88	58.56
4	$\mathcal{L}_{\mathbf{H}} + \lambda_{\boldsymbol{\varepsilon}}\mathcal{L}_{\boldsymbol{\varepsilon}}$	40.57	**184.75**	1260.20	0.0491	437.94	60.24
5	$\mathcal{L}_{\mathbf{H}} + \lambda_{\text{WA}}\mathcal{L}_{\text{WA}}$	40.53	195.52	6938.54	0.0633	603.03	60.64
6	$\mathcal{L}_{\mathbf{H}} + \lambda_{E_{\text{core}}}\mathcal{L}_{E_{\text{core}}}$	53.97	306.64	1756.67	0.0799	813.54	64.65
7	$\mathcal{L}_{\mathbf{H}} + \lambda_{T_{\text{S}}}\mathcal{L}_{T_{\text{S}}}$	76.94	828455.57	2276967.79	28.80	19526279.65	164.35
8	$\mathcal{L}_{\mathbf{H}} + \lambda_{E_{\text{ext}}}\mathcal{L}_{E_{\text{ext}}}$	53.05	317.54	3077.05	0.0800	836.69	64.61
9	$\mathcal{L}_{\mathbf{H}} + \lambda_{\boldsymbol{\mu}}\mathcal{L}_{\boldsymbol{\mu}}$	51.36	388.71	3403.92	0.0796	730.52	62.87
10	$\mathcal{L}_{\mathbf{H}} + \lambda_{\rho}\mathcal{L}_{\rho} + \lambda_{\mathbf{P}}\mathcal{L}_{\mathbf{P}}$	46.89	227.90	<u>397.77</u>	0.0274	<u>109.03</u>	**57.26**
11	$\mathcal{L}_{\mathbf{H}} + \lambda_{\rho}\mathcal{L}_{\rho} + \lambda_{\mathbf{P}}\mathcal{L}_{\mathbf{P}} + \lambda_{\boldsymbol{\varepsilon}}\mathcal{L}_{\boldsymbol{\varepsilon}}$	46.84	<u>192.75</u>	**384.37**	**0.0266**	**104.40**	<u>57.44</u>

We further compare DSHam with QHNet-DM on the QH9 dataset, where QHNet-DM adopts QHNet as the backbone but directly predicts the density matrix. Although QHNet-DM achieves competitive density matrix accuracy, it fails to produce reliable real-space densities, total energies and SCF acceleration performance. As discussed in Results 2.4, a potential direction for improving QHNet-DM is to embed essential physical constraints, such as electron-number conservation and generalized idempotency, into the model, thereby encouraging more physically valid density matrix predictions.

C.6 Additional Ablation Results

Alternative training objectives. As discussed in Results 2.1, we systematically benchmark DSLoss against a range of computationally efficient physical surrogates used as alternative training objectives. As a quantitative extension of the ranking results shown in Fig. 2, we report the corresponding numerical results in Supplementary Table S17. All loss weights are selected according to validation performance, and the final weights used for different loss combinations are provided in Supplementary Table S18.

Precision. As mentioned in Methods 4.2, numerical precision can have a non-negligible effect on the training and evaluation of Hamiltonian models. Here, we report ablation results for models trained with `FP64` and `FP32` in Supplementary Table S19. The results show that although `FP64` brings a marginal improvement over `FP32` on ethanol conformations, the performance gain becomes negligible on the larger malondialdehyde conformations. Therefore, we adopt `FP64` only for MD17 experiments, while using `FP32` for all other datasets to balance accuracy and computational efficiency.

Grid resolution. We next investigate the sensitivity of DSHam to the grid resolution used for real-space density evaluation. Following the procedure described in Methods 4.3, we use PySCF to evaluate the electron density on numerical integration grids generated at different grid levels. We conduct experiments with grid levels ranging from 1 to 4, and summarize the results in Supplementary Table S20. The results show that grid resolution has only a minor impact on model performance. Therefore, to balance predictive

Supplementary Table S18: **Loss combination weights corresponding to the auxiliary loss combinations in Supplementary Table S17.**

No.	Loss combination weights
2	$\mathcal{L}_{\mathbf{H}} + 0.02\mathcal{L}_{\rho}$
3	$\mathcal{L}_{\mathbf{H}} + 0.03\mathcal{L}_{\mathbf{P}}$
4	$\mathcal{L}_{\mathbf{H}} + 0.02\mathcal{L}_{\boldsymbol{\varepsilon}}$
5	$\mathcal{L}_{\mathbf{H}} + 0.02\mathcal{L}_{\mathrm{WA}}$
6	$\mathcal{L}_{\mathbf{H}} + 0.0001\mathcal{L}_{E_{\mathrm{core}}}$
7	$\mathcal{L}_{\mathbf{H}} + 0.0002\mathcal{L}_{T_{\mathrm{S}}}$
8	$\mathcal{L}_{\mathbf{H}} + 0.0007\mathcal{L}_{E_{\mathrm{ext}}}$
9	$\mathcal{L}_{\mathbf{H}} + 0.0005\mathcal{L}_{\boldsymbol{\mu}}$
10	$\mathcal{L}_{\mathbf{H}} + 0.02\mathcal{L}_{\rho} + 0.03\mathcal{L}_{\mathbf{P}}$
11	$\mathcal{L}_{\mathbf{H}} + 0.02\mathcal{L}_{\rho} + 0.03\mathcal{L}_{\mathbf{P}} + 0.02\mathcal{L}_{\boldsymbol{\varepsilon}}$

Supplementary Table S19: **Sensitivity of DSHam to numerical precision.** Results are evaluated over 4,500 ethanol test molecules and 1,478 malondialdehyde test molecules.

Dataset	Precision	$\mathbf{H}$ (μE_h)	$\varepsilon_{\mathrm{occ}}$ (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)
Ethanol	FP32	8.41	40.72	57.93	0.0082	1.04	43.64
	FP64	**7.25**	**36.23**	**51.03**	**0.0072**	**0.79**	**42.54**
Malondialdehyde	FP32	6.18	30.33	91.04	0.0064	**1.33**	44.11
	FP64	**5.72**	**25.93**	**84.59**	**0.0057**	1.36	**43.04**

Supplementary Table S20: **Sensitivity of DSHam to grid resolution.** All models are trained with only the real-space density loss $\mathcal{L}_{\rho}$ as auxiliary supervision. Results are reported on 30,100 QH9 test molecules from the `dynamic-mol` split.

Grid level	$\mathbf{H}$ (μE_h)	$\varepsilon_{\mathrm{occ}}$ (μE_h)	$\mathbf{P}$ (10^{-6})	ρ (%)	E_{tot} (μE_h)	S_{acc} (%)
1	**70.34**	**529.41**	1161.11	**0.0620**	325.30	60.00
2	72.02	551.99	1153.34	0.0635	332.48	60.00
3	71.25	538.28	1255.03	0.0636	321.46	60.00
4	71.03	544.17	**1133.23**	0.0624	**320.52**	60.00

Supplementary Table S21: **Performance with the SPHNet backbone on QH9.** Both models are trained in double precision, and results are evaluated on 13,084 QH9 test molecules.

Model	$\mathbf{H}$ (μE_h)	$\mathbf{P}$ (10^{-6})	E_{tot} (μE_h)	S_{acc} (%)
SPHNet	**41.47**	22793.59	866.66	68.15
DSHam (SPHNet)	57.17	**464.70**	**183.26**	**64.48**

accuracy and computational efficiency, we use a grid level of 3 throughout Results 2, unless otherwise specified.

Backbone architecture. Finally, to assess the generality of DSHam across different Hamiltonian learning architectures, we replace the QHNet backbone with SPHNet. As shown in Supplementary Table S21, DSHam continues to deliver consistent and substantial performance improvements with the new backbone, demonstrating that the proposed supervision strategy is not tied to a specific network architecture.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryinformationDSHam.pdf