IMPROVING KOHN–SHAM HAMILTONIAN LEARNING WITH GRASSMANNIAN SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the Kohn–Sham Hamiltonian in density functional theory has become an important task in AI for science, as it can accelerate electronic structure calculations and enable the prediction of multiple molecular properties from a single model. However, element-wise Hamiltonian regression does not explicitly enforce the accuracy of the occupied orbital subspace, which determines the density matrix from which ground-state observables are evaluated. This subspace therefore provides a natural auxiliary target for improving downstream property predictions. Yet, directly supervising orbital coefficients in an atomic orbital basis presents two challenges: (1) *Gauge redundancy.* Rotations among occupied orbitals preserve their subspace and density matrix, making the coefficient representation non-unique. (2) *Gradient instability.* Orbital-based losses require differentiation through an eigensolver, whose eigenvector derivatives can become unstable near degeneracies. To address these challenges, we propose **GrassDist**, a principled framework that aligns predicted and reference occupied subspaces using a Grassmannian geodesic loss on the Grassmann manifold. GrassDist is invariant to gauge transformations among occupied orbitals by construction, hence resolving the ambiguity in coefficient supervision. We additionally incorporate two gradient-stabilization treatments to ensure stable optimization. Across multiple benchmarks, GrassDist achieves substantial improvements in various downstream properties, while enabling stable training and faster convergence. On QH9, it reduces density matrix and total energy errors by one to two orders of magnitude compared with the Hamiltonian-only baseline and WALoss. GrassDist also exhibits strong data efficiency: on MD17 uracil, using only 10% to 20% of the training data already matches or surpasses the full-data baseline. The resulting model further generalizes to molecules larger than those seen during training. These results establish GrassDist as a significant methodological advance toward reliable and scalable molecular modeling.

1 INTRODUCTION

Density functional theory (DFT) (Hohenberg & Kohn, 1964) is widely used in quantum chemistry and materials science to calculate electronic structures and molecular properties. In its Kohn–Sham (KS) formulation (Kohn & Sham, 1965), molecular orbitals are obtained by solving an eigenvalue problem for an effective one-electron Hamiltonian. This Hamiltonian, however, depends on the electron density constructed from the occupied orbitals. Conventional calculations therefore rely on a self-consistent field procedure that repeatedly updates the Hamiltonian and solves for the orbitals until convergence. The cost of these repeated calculations grows rapidly with system size (Goedecker, 1999), limiting the scale and throughput of molecular simulations.

Machine learning offers a way to reduce this cost by predicting the converged KS Hamiltonian directly from molecular geometry (Hegde & Bowen, 2017; Schütt et al., 2019). Once predicted, the Hamiltonian can be solved to obtain molecular orbitals and orbital energies, allowing a single model to support the evaluation of multiple molecular properties without property-specific prediction heads. The predicted Hamiltonian can also serve as an accurate initialization for subsequent SCF calculations. This formulation has benefited substantially from advances in equivariant neural networks, which ensure that the predicted Hamiltonian transforms consistently under molecular

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

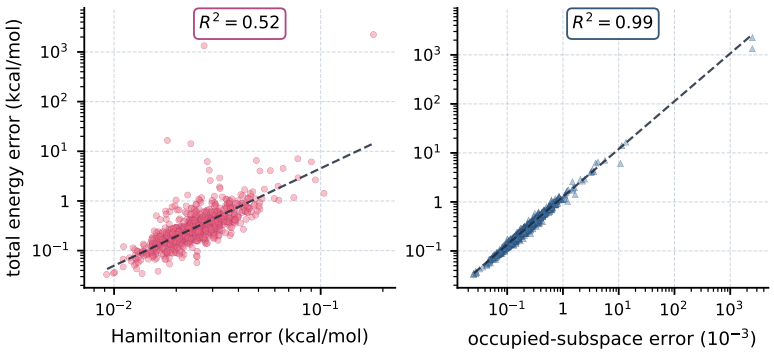


Figure 1: Correlation of total energy error with **(a)** Hamiltonian error and **(b)** occupied-subspace error. The coefficients of determination (R^2) are computed over the first 800 configurations of the QH9 `stable-iid` test set, using an SPHNet model trained with element-wise Hamiltonian supervision. The occupied-subspace error is defined by Eq. 7.

rotations (Unke et al., 2021; Yu et al., 2023b; Li et al., 2022; Gong et al., 2023; Luo et al., 2025; Qian et al., 2026; Tang et al., 2025). A common training strategy is to minimize element-wise errors between predicted and reference Hamiltonian matrices represented in an atomic orbital (AO) basis. While simple and effective for Hamiltonian reconstruction, this objective does not account for how errors in different directions of the Hamiltonian affect its eigensolutions. Consequently, a more accurate Hamiltonian in terms of element-wise error does not necessarily yield more accurate physical predictions (Li et al., 2025). Figure 1 illustrates this mismatch: the total energy error shows only a moderate correlation with the Hamiltonian error ($R^2 = 0.52$), whereas its correlation with the occupied-subspace error is substantially stronger ($R^2 = 0.99$).

This observation motivates us to examine the electronic quantity that is more directly connected to ground-state physics than element-wise Hamiltonian error. For a closed-shell system, the occupied orbitals span an occupied subspace that uniquely determines the density matrix and thereby governs many ground-state observables. Recent work has also recognized the limitation of element-wise Hamiltonian regression. For example, WALoss (Li et al., 2025) transforms the predicted Hamiltonian into the reference orbital basis and places greater emphasis on its diagonal elements associated with occupied orbitals and the LUMO. While effective for improving spectral accuracy, this formulation does not directly measure whether the predicted Hamiltonian recovers the correct occupied subspace. The occupied subspace therefore provides a natural auxiliary target for improving downstream physical predictions.

However, directly supervising orbital coefficients in an AO basis presents two challenges. (1) *Gauge redundancy.* For a system with B basis functions and N_{occ} occupied orbitals, the occupied coefficient matrix $\mathbf{C}_{\text{occ}} \in \mathbb{R}^{B \times N_{\text{occ}}}$ is obtained from the generalized Hamiltonian eigenvalue problem, with its columns corresponding to eigenvectors. As illustrated in Figure 2a, an orthogonal rotation $\mathbf{Q} \in \mathrm{O}(N_{\text{occ}})$ changes the orbital coefficients but preserves the occupied subspace, and therefore all density-derived ground-state observables. Consequently, the same physical state can be represented by many equally valid coefficient matrices. A coefficient-level loss nevertheless treats them as distinct labels, rendering the supervision target non-unique. (2) *Gradient instability.* Backpropagating coefficient-based losses requires differentiation through an eigensolver, whose eigenvector derivatives contain inverse eigenvalue-gap terms and can become numerically unstable near degeneracies. This issue can be particularly pronounced during early training, when inaccurate predicted spectra may contain spurious near-degeneracies.

To address these challenges, we introduce **GrassDist**, a principled framework that exploits the gauge symmetry of occupied orbitals to remove redundant supervision while maintaining stable optimization. We first express the occupied orbitals in an orthonormalized AO basis as $\mathbf{X} = \sqrt{\mathbf{S}}\mathbf{C}_{\text{occ}}$, which resides on the Stiefel manifold $\mathrm{St}(N_{\text{occ}}, B)$. Since any orthogonal rotation $\mathbf{XQ}$ (with $\mathbf{Q} \in \mathrm{O}(N_{\text{occ}})$) spans the identical occupied subspace and thus corresponds to the same physical state, we introduce the Grassmann manifold $\mathrm{Gr}(N_{\text{occ}}, B)$ that retains only the intrinsic degrees of freedom of the occupied subspace. We then measure the discrepancy between predicted and reference subspaces using

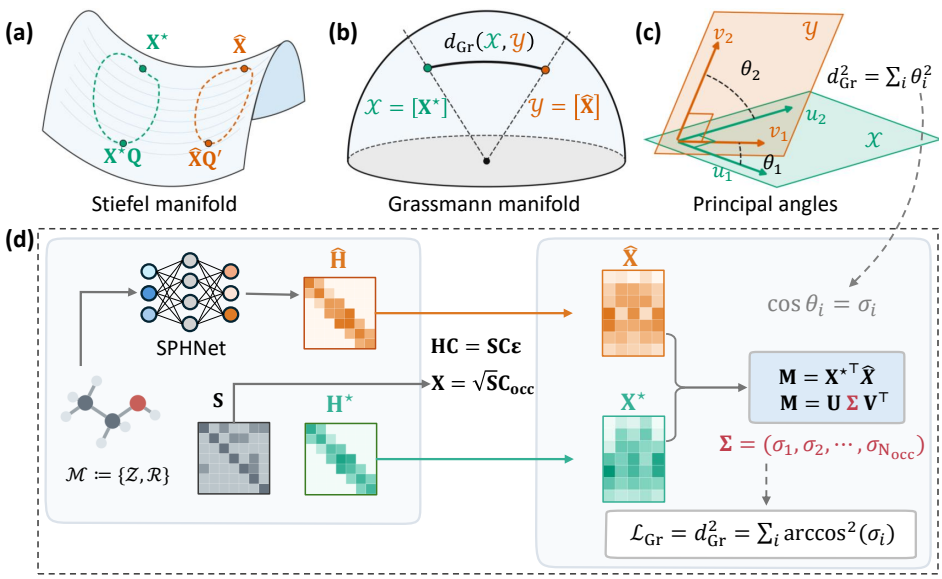


Figure 2: Overview of GrassDist. **(a)** Occupied-orbital representations on the Stiefel manifold. Representations related by $\mathbf{Q} \in \mathrm{O}(N_{\mathrm{occ}})$ span the same occupied subspace. **(b)** The Grassmann manifold identifies such gauge-equivalent representations as the same element, and the geodesic distance $d_{\mathrm{Gr}}(\mathcal{X}, \mathcal{Y})$ measures the discrepancy between the reference subspace $\mathcal{X}$ and predicted subspace $\mathcal{Y}$. **(c)** The geodesic distance is characterized by the principal angles $\{\theta_i\}_{i=1}^{N_{\mathrm{occ}}}$ between $\mathcal{X}$ and $\mathcal{Y}$. **(d)** Given the molecular geometry $\mathcal{M}$, an equivariant backbone (SPHNet) predicts the Hamiltonian $\widehat{\mathbf{H}}$. Solving the generalized eigenvalue problem yields orbital energies and occupied orbitals, represented in the orthonormalized AO basis as $\mathbf{X} = \sqrt{\mathbf{S}}\mathbf{C}_{\mathrm{occ}}$. The singular values $\sigma_i = \cos\theta_i$ of the cross-projection matrix $\mathbf{X}^{\star\top}\widehat{\mathbf{X}}$ yield the Grassmannian geodesic loss $\mathcal{L}_{\mathrm{Gr}}$.

a Grassmannian geodesic loss $\mathcal{L}_{\mathrm{Gr}}$ on the Grassmann manifold (Figure 2d). This loss is invariant to rotations among occupied orbitals, resolving the ambiguity of coefficient-level supervision. To ensure stable optimization, we evaluate $\mathcal{L}_{\mathrm{Gr}}$ using a local Taylor approximation of $\arccos^2(\sigma_i)$ in the small-angle regime, avoiding the numerical instability of direct differentiation through $\arccos(\sigma_i)$ as $\sigma_i \to 1$.

Across multiple benchmarks, GrassDist substantially improves the accuracy of various downstream properties. On QH9, it reduces density matrix and total energy errors by one to two orders of magnitude over the Hamiltonian-only baseline and WALoss. GrassDist also exhibits strong data efficiency: on MD17 uracil, using only 10% to 20% of the training data already matches or surpasses the full-data baseline. With these advantages, the resulting model further generalizes to molecules larger than those seen during training without fine-tuning. These results establish GrassDist as a significant methodological advance toward reliable and scalable molecular modeling.

2 Method

2.1 Preliminaries

We first review the basic formulation of KS-DFT and the conventional Hamiltonian learning paradigm before introducing GrassDist. Additional details are provided in Appendix A.1.

KS-DFT. For a molecular geometry $\mathcal{M} := \{\mathcal{Z}, \mathcal{R}\}$, where $\mathcal{Z} := \{Z^{(a)}\}_{a=1}^{A}$ and $\mathcal{R} := \{\mathbf{R}^{(a)}\}_{a=1}^{A}$ denote the atomic numbers and coordinates of the A nuclei, KS-DFT (Kohn & Sham, 1965) represents the electronic state using a set of one-electron orbitals $\{\phi_i(\mathbf{r})\}_{i=1}^{N_{\mathrm{occ}}}$ ($\mathbf{r}$ denotes coordinates of an electron). For practical calculations, each orbital is expanded in an AO basis $\{\chi_\mu(\mathbf{r})\}_{\mu=1}^{B}$ with the form $\phi_i(\mathbf{r}) = \sum_{\mu=1}^{B}(\mathbf{C}_{\mathrm{occ}})_{\mu i}\chi_\mu(\mathbf{r})$, where $\mathbf{C}_{\mathrm{occ}} \in \mathbb{R}^{B\times N_{\mathrm{occ}}}$ contains the coefficients of the

N_{occ} occupied orbitals. The KS orbitals are obtained by solving the generalized eigenvalue problem

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

where $\mathbf{H} \in \mathbb{R}^{B\times B}$ is the KS Hamiltonian, $\mathbf{S}_{\mu\nu} := \int \chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\, d\mathbf{r}$ is the AO overlap matrix, and $\boldsymbol{\varepsilon} = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_B)$ contains the orbital energies. Solving the KS equation in the B-dimensional AO space yields B generalized eigenvectors, comprising the occupied and virtual orbitals. We collect their coefficients as $\mathbf{C} = [\mathbf{C}_{\mathrm{occ}}, \mathbf{C}_{\mathrm{virt}}]$. Because $\mathbf{H}$ itself depends on the electron density constructed from the occupied orbitals, Eq. 1 is conventionally solved through an iterative self-consistent field (SCF) procedure (Szabo & Ostlund, 2012). For a closed-shell system, the N_{occ} lowest-energy orbitals are doubly occupied, with coefficients $\mathbf{C}_{\mathrm{occ}} \in \mathbb{R}^{B\times N_{\mathrm{occ}}}$. The corresponding density matrix is $\mathbf{P} = 2\mathbf{C}_{\mathrm{occ}}\mathbf{C}_{\mathrm{occ}}^\top$, from which ground-state observables are evaluated.

Limitations of element-wise Hamiltonian regression. Existing Hamiltonian prediction models (Unke et al., 2021; Qian et al., 2026) commonly minimize element-wise discrepancies between the predicted Hamiltonian $\widehat{\mathbf{H}}$ and the reference Hamiltonian $\mathbf{H}^\star$. Following Yu et al. (2023b), we use a combination of mean absolute error (MAE) and mean squared error (MSE) for Hamiltonian supervision. While effective at producing numerically accurate Hamiltonians, such element-wise objectives do not directly assess how prediction errors affect the resulting electronic state. To illustrate this, consider a perturbation $\delta\mathbf{H} = \widehat{\mathbf{H}} - \mathbf{H}^\star$ to the reference Hamiltonian. The first-order response of the density matrix can be written as (McWeeny, 1962; Baroni et al., 2001)

$$\delta\mathbf{P} = \sum_{i<j} \frac{f_i - f_j}{\varepsilon_i - \varepsilon_j} \left(\mathbf{C}_j^\top \delta\mathbf{H}\, \mathbf{C}_i\right) \left(\mathbf{C}_j\mathbf{C}_i^\top + \mathbf{C}_i\mathbf{C}_j^\top\right), \tag{2}$$

where f_i denotes the occupation of orbital i. A detailed derivation is provided in Appendix A.2. Eq. 2 makes explicit that the effect of a Hamiltonian error on the electronic state is mediated by the orbital coefficients. Therefore, element-wise accuracy of $\widehat{\mathbf{H}}$ alone does not directly measure the accuracy of the electronic state it induces. This motivates complementing Hamiltonian supervision with direct supervision of the occupied subspace that determines the electronic state.

2.2 Learning challenges of coefficient-level orbital supervision

Gauge redundancy. A straightforward way to supervise the occupied subspace is to solve the generalized eigenvalue problem for the predicted Hamiltonian $\widehat{\mathbf{H}}$ and directly match its occupied orbital coefficients $\mathbf{C}_{\mathrm{occ}} \in \mathbb{R}^{B\times N_{\mathrm{occ}}}$ to the reference coefficients. However, the coefficient representation is inherently non-unique. To expose this redundancy in an orthonormal basis, we represent the occupied orbitals as the orthonormalized matrix $\mathbf{X} = \sqrt{\mathbf{S}}\,\mathbf{C}_{\mathrm{occ}}$, which satisfies $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_{N_{\mathrm{occ}}}$ (Löwdin, 1950). For any orthogonal transformation $\mathbf{Q} \in \mathrm{O}(N_{\mathrm{occ}})$, the rotated matrix $\mathbf{XQ}$ is column-orthonormal and spans exactly the same occupied subspace as $\mathbf{X}$. Equivalently, rotating the AO coefficients as $\widetilde{\mathbf{C}}_{\mathrm{occ}} = \mathbf{C}_{\mathrm{occ}}\mathbf{Q}$ leaves the density matrix unchanged:

$$\widetilde{\mathbf{P}} = 2\widetilde{\mathbf{C}}_{\mathrm{occ}}\widetilde{\mathbf{C}}_{\mathrm{occ}}^\top = 2\mathbf{C}_{\mathrm{occ}}\mathbf{Q}\mathbf{Q}^\top\mathbf{C}_{\mathrm{occ}}^\top = \mathbf{P}. \tag{3}$$

We define the rotation class $[\mathbf{X}] := \{\mathbf{XQ} : \mathbf{Q} \in \mathrm{O}(N_{\mathrm{occ}})\}$ as the set of all orthogonal rotations of $\mathbf{X}$; Eq. 3 shows that every matrix in this class represents the same physical state. An element-wise coefficient loss nevertheless treats these matrices as distinct labels, making the supervision target non-unique—the practical drawbacks of which are demonstrated in Section 3.3. The physically relevant target is therefore the occupied subspace, represented by the rotation class $[\mathbf{X}]$, rather than any particular matrix, as illustrated in Figure 2(a,b). This motivates a geometric formulation that compares occupied subspaces directly while remaining invariant to such rotations.

Gradient instability. Occupied-subspace supervision introduces two sources of numerical instability during backpropagation. First, the Grassmannian geodesic loss involves $\arccos(\sigma_i)$, whose derivative becomes singular as $\sigma_i \to 1^-$ and can destabilize automatic differentiation when the predicted and reference subspaces are closely aligned. We address this issue by replacing $\arccos^2(\sigma_i)$ with a local Taylor approximation when σ_i is close to one, as discussed in Section 2.3. Second, backpropagation through the eigensolver involves inverse orbital-energy differences, $(\widehat{\varepsilon}_i - \widehat{\varepsilon}_j)^{-1}$, which become ill-conditioned near spectral degeneracies. We stabilize this step by truncating the corresponding ill-conditioned gradient components. The detailed formulations and implementation are provided in Appendix B.2.

2.3 GEODESIC DISTANCE ON THE GRASSMANN MANIFOLD

Following Section 2.2, we formulate the occupied-subspace supervision on the Grassmann manifold (Edelman et al., 1998; Absil et al., 2008). Since the occupied coefficients satisfy the $\mathbf{S}$-orthonormality condition $\mathbf{C}_{\mathrm{occ}}^{\top}\mathbf{S}\mathbf{C}_{\mathrm{occ}} = \mathbf{I}_{N_{\mathrm{occ}}}$, the transformed coefficients $\mathbf{X} = \sqrt{\mathbf{S}}\mathbf{C}_{\mathrm{occ}}$ constitute an orthonormal frame, $\mathbf{X}^{\top}\mathbf{X} = \mathbf{I}_{N_{\mathrm{occ}}}$, and therefore lie on the Stiefel manifold

$$\mathrm{St}(N_{\mathrm{occ}}, B) = \left\{ \mathbf{X} \in \mathbb{R}^{B \times N_{\mathrm{occ}}} : \mathbf{X}^{\top}\mathbf{X} = \mathbf{I}_{N_{\mathrm{occ}}} \right\}. \tag{4}$$

Each frame $\mathbf{X}$ spans an N_{occ}-dimensional occupied subspace $\mathrm{col}(\mathbf{X}) \subset \mathbb{R}^{B}$. The collection of all such subspaces forms the Grassmann manifold $\mathrm{Gr}(N_{\mathrm{occ}}, B)$, which can also be expressed as

$$\mathrm{Gr}(N_{\mathrm{occ}}, B) \simeq \mathrm{St}(N_{\mathrm{occ}}, B)/\mathrm{O}(N_{\mathrm{occ}}). \tag{5}$$

Let $\mathbf{X}^{\star} = \sqrt{\mathbf{S}}\mathbf{C}_{\mathrm{occ}}^{\star}$ and $\widehat{\mathbf{X}} = \sqrt{\mathbf{S}}\widehat{\mathbf{C}}_{\mathrm{occ}}$ denote the reference and predicted orthonormalized frames, with $\mathcal{X} = [\mathbf{X}^{\star}]$ and $\mathcal{Y} = [\widehat{\mathbf{X}}]$ representing the corresponding occupied subspaces as points on $\mathrm{Gr}(N_{\mathrm{occ}}, B)$. Their relative geometry is characterized by the principal angles $\{\theta_i\}_{i=1}^{N_{\mathrm{occ}}}$, obtained via the singular value decomposition of the cross-projection matrix

$$\mathbf{M} := \mathbf{X}^{\star\top}\widehat{\mathbf{X}} = \mathbf{C}_{\mathrm{occ}}^{\star\top}\mathbf{S}\,\widehat{\mathbf{C}}_{\mathrm{occ}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \qquad \mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{N_{\mathrm{occ}}}), \tag{6}$$

where $\sigma_i \in [0, 1]$ and $\cos\theta_i = \sigma_i$ (Björck & Golub, 1973). The geodesic distance on the Grassmann manifold is therefore (Zhang et al., 2018)

$$d_{\mathrm{Gr}}(\mathcal{X}, \mathcal{Y}) = \left(\sum_{i=1}^{N_{\mathrm{occ}}} \theta_i^2 \right)^{1/2} = \left(\sum_{i=1}^{N_{\mathrm{occ}}} \arccos^2(\sigma_i) \right)^{1/2}. \tag{7}$$

By construction, this distance is invariant to the gauge freedom of occupied orbitals.

For optimization, however, its direct implementation introduces a numerical issue. Differentiating through $\arccos(\sigma_i)$ becomes unstable as $\sigma_i \to 1^{-}$. Although $\arccos^2(\sigma_i)$ has a finite overall derivative, the intermediate derivative

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\arccos(\sigma) = -\frac{1}{\sqrt{1-\sigma^2}} \xrightarrow{\sigma \to 1^{-}} -\infty \tag{8}$$

is singular and can destabilize automatic differentiation. To avoid this numerical instability, we introduce $u_i = 1 - \sigma_i$ and evaluate the Grassmannian geodesic loss using a local Taylor approximation in the small-angle regime:

$$\mathcal{L}_{\mathrm{Gr}} = \sum_{i=1}^{N_{\mathrm{occ}}} \begin{cases} 2u_i + \frac{1}{3}u_i^2 + \frac{4}{45}u_i^3 + \frac{1}{35}u_i^4, & u_i < 10^{-3}, \\ \arccos^2(\sigma_i), & \text{otherwise}, \end{cases} \qquad u_i = 1 - \sigma_i. \tag{9}$$

This formulation provides a numerically stable evaluation of the squared Grassmannian geodesic distance with bounded gradients near $\sigma_i \to 1$. Beyond numerical stability, the geodesic objective also provides stronger corrective signals for large subspace discrepancies, as analyzed in Appendix A.3 and empirically verified in Appendix C.8.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate GrassDist on MD17, QH9, and QMugs (Chmiela et al., 2017; Yu et al., 2023a; Isert et al., 2022) against established baselines, following standard evaluation protocols (Yu et al., 2023b; Luo et al., 2025). For MD17, we adopt trajectory geometries of four representative molecules (water, ethanol, malondialdehyde, and uracil) with reference Hamiltonians regenerated at the PBE/def2-SVP level (Perdew et al., 1996; Weigend & Ahlrichs, 2005). QH9 features small organic molecules at the B3LYP/def2-SVP (Becke, 1993; Lee et al., 1988) level across four standard benchmark splits: two equilibrium splits (`stable-iid`, `stable-ood`) and two dynamic splits (`dynamic-geo`, `dynamic-mol`). For out-of-distribution evaluation on larger drug-like molecules, we test a QMugs subset with 10–17 heavy atoms.

Table 1: Results on the four QH9 splits using the SPHNet backbone; the best-performing result in each split is boldfaced. For E_{tot}, values are reported as the mean (median) absolute error.

Split	Method	$\mathbf{H}\downarrow$	$d_{\text{Gr}}\downarrow$	$\mathbf{P}\downarrow$	$E_{\text{tot}}\downarrow$	$\varepsilon_{\Delta}\downarrow$	$\varepsilon_{\text{occ}}\downarrow$	$\boldsymbol{\mu}\downarrow$	$S_{\text{acc}}\downarrow$
stable-iid	Baseline	**41.03**	66.28	9.01	4.87 (0.28)	1607.37	488.08	21.36	67.2%
	WALoss	56.93	57.28	7.32	4.51 (0.66)	2006.25	299.30	26.22	70.0%
	GrassDist	182.26	**4.72**	**0.26**	**0.04** (**0.03**)	**889.23**	190.54	**6.04**	**61.7%**
stable-ood	Baseline	**33.52**	29.51	1.64	1.98 (0.25)	850.61	**118.33**	11.03	69.5%
	WALoss	47.97	55.32	3.10	2.42 (0.54)	1852.51	195.16	19.43	72.1%
	GrassDist	124.09	**4.96**	**0.22**	**0.04** (**0.04**)	**634.11**	150.18	**5.09**	**63.7%**
dynamic-geo	Baseline	**41.87**	32.29	2.74	0.27 (0.25)	1058.08	118.46	15.32	67.0%
	WALoss	42.70	15.62	1.50	0.28 (0.25)	746.13	**114.91**	11.70	66.8%
	GrassDist	173.43	**3.61**	**0.19**	**0.02** (**0.02**)	**381.99**	120.18	**4.01**	**60.8%**
dynamic-mol	Baseline	**98.72**	312.62	149.00	0.82 (0.74)	12364.72	3666.81	279.87	71.1%
	WALoss	99.21	230.00	56.56	0.90 (0.74)	8556.34	1050.80	191.55	71.8%
	GrassDist	219.80	**16.26**	**0.96**	**0.19** (**0.17**)	**2213.21**	**1021.92**	**43.95**	**67.1%**

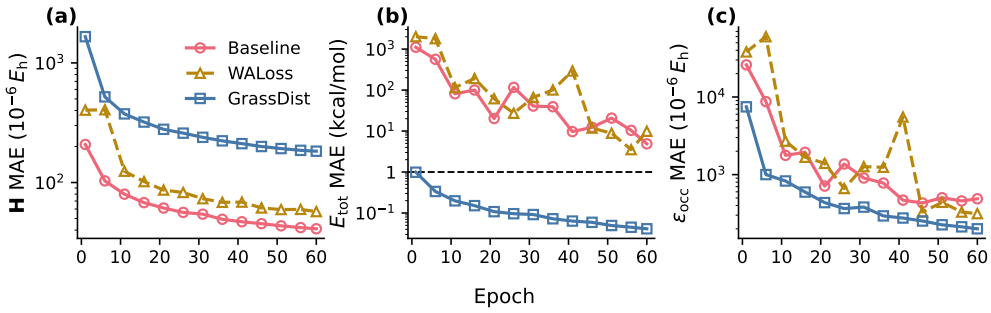


Figure 3: Convergence of GrassDist, WALoss, and the Hamiltonian-only baseline on `stable-iid` across training epochs. From left to right: (a) Hamiltonian matrix MAE ($\mathbf{H}$, in $10^{-6}\,E_{\text{h}}$), (b) total-energy MAE (E_{tot}, in kcal/mol), and (c) occupied orbital-energy MAE (ε_{occ}, in $10^{-6}\,E_{\text{h}}$).

Metrics and Units. We assess model performance across direct representations, the electronic spectrum, and derived physical observables: (i) **Matrix representations**: element-wise MAE on the Hamiltonian matrix $\mathbf{H}$ ($10^{-6}\,E_{\text{h}}$) and the density matrix $\mathbf{P}$ (10^{-3} for QH9, 10^{-6} for MD17); (ii) **Subspace and orbital fidelity**: the geodesic distance d_{Gr} (10^{-3}) between predicted and reference occupied subspaces, and, for the coefficient-supervision ablation, the $\mathbf{C}_{\text{occ}}$ MAE (10^{-3}) and column-wise cosine similarity $\mathbf{C}_{\text{occ}}$ sim (%); (iii) **Spectral properties**: MAE on the occupied-orbital energies ε_{occ} ($10^{-6}\,E_{\text{h}}$) and the HOMO–LUMO gap ε_{Δ} ($10^{-6}\,E_{\text{h}}$); (iv) **Physical observables**: MAE on the total energy E_{tot} (kcal/mol) and the vector error on the dipole moment $\boldsymbol{\mu}$ (10^{-3} a.u. for QH9, 10^{-6} a.u. for MD17); and (v) **Downstream utility**: the SCF iteration ratio $S_{\text{acc}} = N_{\text{pred}}/N_{\text{MINAO}}$ (%), where N_{pred} and N_{MINAO} denote the number of SCF iterations required to achieve convergence when initialized from the predicted Hamiltonian and the MINAO guess (Almlöf et al., 1982; Van Lenthe et al., 2006), respectively. Complete dataset partition statistics and computational setups are detailed in Appendix B.1 and Appendix B.2.

Table 2: Coefficient-level versus gauge-invariant subspace supervision on `QH9-stable-iid`; the best-performing result is boldfaced.

Method	$\mathbf{H}\downarrow$	$d_{\text{Gr}}\downarrow$	$\mathbf{P}\downarrow$	$E_{\text{tot}}\downarrow$	$\boldsymbol{\mu}\downarrow$	$\mathbf{C}_{\text{occ}}$ MAE$\downarrow$	$\mathbf{C}_{\text{occ}}$ sim$\uparrow$
$\mathbf{C}_{\text{occ}}$ MAE	51.42	12.64	0.79	0.19	17.37	22.14	99.2%
$\mathbf{C}_{\text{occ}}$ sim	52.77	12.30	0.61	0.18	13.91	20.89	**99.3%**
GrassDist-lite	**50.78**	9.21	0.61	0.12	9.16	**20.07**	99.2%
GrassDist	182.26	**4.72**	**0.26**	**0.04**	**6.04**	23.93	98.8%

3.2 RESULTS ON QH9

We first study whether GrassDist improves ground-state physical property predictions on QH9, using the SPHNet architecture (Luo et al., 2025) across all four benchmark splits. Using the same backbone, GrassDist is compared against a Hamiltonian-only baseline and WALoss (Li et al., 2025) (*i.e.*, surrogate orbital-energy supervision via a basis transformation with ground-truth eigenvectors).

As shown in Table 1, GrassDist consistently reduces errors across ground-state observables despite exhibiting a larger $\mathbf{H}$ MAE. On `stable-iid`, the reduction of d_{Gr} from 66.28 to 4.72 leads to a sharp decline in the density matrix MAE ($\mathbf{P}$, from 9.01 to 0.26). Remarkably, the E_{tot} MAE decreases from 4.87 to 0.04 kcal/mol, representing an over 100-fold improvement. Notably, although the baseline mean error (4.87 kcal/mol) is heavily inflated by extreme outliers relative to its median (0.28 kcal/mol), GrassDist still cuts the median error nearly tenfold, from 0.28 to 0.03 kcal/mol. We provide a structural and spectral analysis of these outliers in Appendix C.7. Across all four splits, GrassDist consistently achieves the best d_{Gr}, $\mathbf{P}$, E_{tot}, $\boldsymbol{\varepsilon}_{\Delta}$, and S_{acc}. Meanwhile, WALoss performs broadly on par with the Hamiltonian-only baseline.

Figure 3 reports the test-set errors of the three supervision strategies at matched checkpoints on `stable-iid`. GrassDist exhibits near-monotonic decreases in the errors of $\mathbf{H}$, E_{tot}, and $\boldsymbol{\varepsilon}_{\mathrm{occ}}$, whereas the baseline and WALoss show pronounced oscillations in downstream observables. GrassDist also converges substantially faster, reaching chemical accuracy (Pople, 1999) in E_{tot} (0.99 kcal/mol) within a single epoch, while the other two methods do not reach this threshold within 60 epochs. The backbone ablation in Appendix C.5 further shows that these gains are not specific to a particular network architecture. Overall, Grassmannian subspace supervision ensures both optimization stability and rapid convergence toward physically reliable electronic structures.

3.3 GAUGE REDUNDANCY IN PRACTICE

To empirically demonstrate the negative effect of gauge redundancy and the advantage of GrassDist's gauge-invariant formulation, we compare GrassDist against coefficient-level supervision on `stable-iid`, as summarized in Table 2. For the coefficient-level baselines, we report their best-performing auxiliary weights (0.05 for $\mathbf{C}_{\mathrm{occ}}$ MAE and 0.005 for $\mathbf{C}_{\mathrm{occ}}$ sim). GrassDist-lite adopts a minimal Grassmannian weight of $w_{\mathrm{Gr}} = 0.05$, whereas standard GrassDist uses the same setup as in Section 3.2.

As shown in Table 2, subspace supervision consistently outperforms coefficient-level supervision. Beyond improving physical predictions (d_{Gr}, E_{tot}, $\boldsymbol{\mu}$), GrassDist-lite strikingly achieves a lower $\mathbf{C}_{\mathrm{occ}}$ MAE (20.07) than direct supervision with $\mathbf{C}_{\mathrm{occ}}$ MAE itself (22.14) or $\mathbf{C}_{\mathrm{occ}}$ sim (20.89) while matching $\mathbf{C}_{\mathrm{occ}}$ sim at 99.2%, evidencing that orbital gauge redundancy severely hinders direct coefficient regression. Standard GrassDist delivers further improvements, achieving the best d_{Gr} (4.72), $\mathbf{P}$ (0.26), E_{tot} (0.04), and $\boldsymbol{\mu}$ (6.04). Overall, eliminating gauge ambiguity via GrassDist is essential for learning physically reliable electronic structures.

3.4 RESULTS ON MD17

Table 3: Results on four molecules from MD17 using the SPHNet backbone; the best-performing result for each molecule is boldfaced.

Molecule	Method	$\mathbf{H}\downarrow$	$d_{\mathrm{Gr}}\downarrow$	$\mathbf{P}\downarrow$	$E_{\mathrm{tot}}\downarrow$	$\boldsymbol{\varepsilon}_{\Delta}\downarrow$	$\boldsymbol{\varepsilon}_{\mathrm{occ}}\downarrow$	$\boldsymbol{\mu}\downarrow$	$S_{\mathrm{acc}}\downarrow$
Water	Baseline	**1.31**	0.02	**1.70**	$\mathbf{4.08\times10^{-7}}$	4.85	1.66	**19.88**	55.7%
	GrassDist	1.58	0.02	2.06	7.68×10^{-7}	**3.74**	1.66	22.95	55.7%
Ethanol	Baseline	**12.68**	1.36	64.27	2.11×10^{-3}	90.26	**97.47**	3289.23	**54.6%**
	GrassDist	17.02	**1.13**	**37.70**	$\mathbf{1.60\times10^{-3}}$	**81.45**	97.97	**3173.87**	55.8%
Malondialdehyde	Baseline	**3.97**	0.74	61.79	5.46×10^{-4}	**21.10**	20.95	1156.95	**48.4%**
	GrassDist	8.07	**0.48**	**29.80**	$\mathbf{2.86\times10^{-4}}$	21.94	**20.76**	**895.15**	48.5%
Uracil	Baseline	**3.13**	1.12	90.39	1.73×10^{-3}	17.68	15.23	1190.75	**42.6%**
	GrassDist	8.65	**0.59**	**36.49**	$\mathbf{6.25\times10^{-4}}$	**15.45**	**14.47**	**958.89**	42.8%

We next evaluate GrassDist against the Hamiltonian-only baseline on MD17 molecular-dynamics trajectories, reporting results on four representative molecules of increasing size.

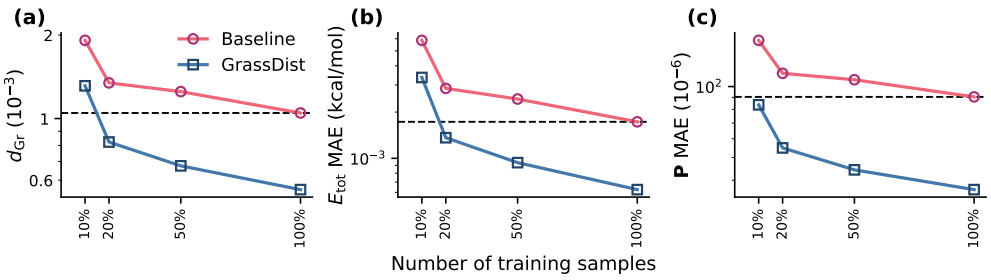


Figure 4: Data efficiency of GrassDist compared to the Hamiltonian-only baseline across different training-set sizes (10%, 20%, 50%, and 100%) on uracil. From left to right: (a) d_{Gr}, in 10^{-3}; (b) E_{tot}, in kcal/mol; and (c) $\mathbf{P}$, in 10^{-6}. Dashed horizontal lines mark the baseline performance trained on 100% data.

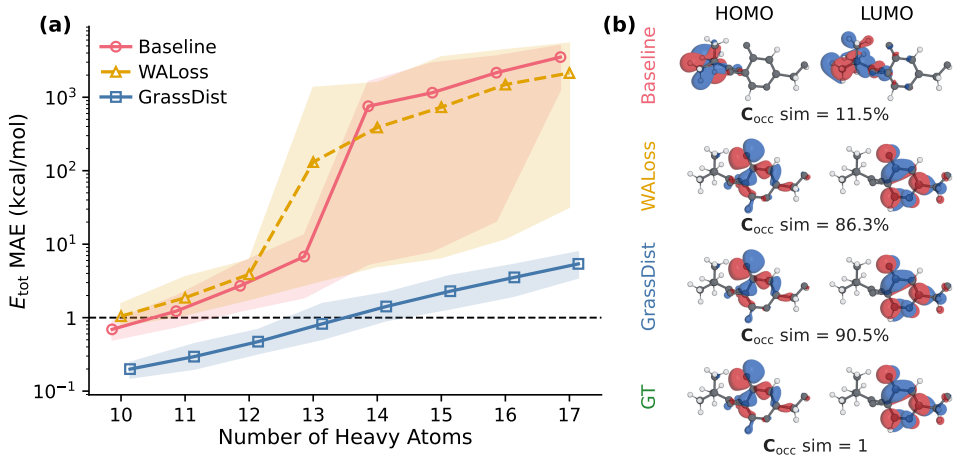


Figure 5: **(a)** Total-energy MAE versus heavy-atom count ($n = 50$ per size) for the Hamiltonian-only baseline, WALoss, and GrassDist; curves and shaded bands denote the mean and 5th–95th percentile range, and the dotted line marks chemical accuracy (1 kcal/mol). **(b)** HOMO and LUMO isosurfaces of a representative 16-heavy-atom QMugs molecule for the baseline, WALoss, GrassDist, and the ground truth (GT); each row is annotated with the cosine similarity of orbital coefficients ($\mathbf{C}_{\mathrm{occ}}$ sim) to the GT.

Water is a limiting case (only 10 electrons) where the baseline already recovers the occupied subspace well ($d_{\mathrm{Gr}} = 0.02 \times 10^{-3}$); here GrassDist performs slightly worse than the baseline, yielding improvements only on ε_{Δ}. In contrast, for larger molecules—ethanol, malondialdehyde, and uracil (26, 38, and 58 electrons, respectively)—the advantage of GrassDist scales with molecular size, monotonically increasing the E_{tot} MAE reduction from 24% to 48% and 64%, while consistently improving $\mathbf{P}$ and $\boldsymbol{\mu}$ predictions across all molecules.

To examine whether the benefits of occupied-subspace supervision persist in the low-data regime, we further evaluate GrassDist on uracil while varying the training-set size from 2,400 to 24,000 configurations. As shown in Figure 4, GrassDist consistently outperforms the baseline on all three metrics (d_{Gr}, E_{tot}, and $\mathbf{P}$) across every training-set size. Specifically, with only 10% of the training data, GrassDist already surpasses the full-data baseline on the $\mathbf{P}$ MAE (83.66×10^{-6} vs. 90.39×10^{-6}), and with 20% of the data it surpasses the full-data baseline on the E_{tot} MAE (1.36×10^{-3} vs. 1.73×10^{-3} kcal/mol). The E_{tot} trend closely mirrors that of d_{Gr}, consistent with their strong correlation shown in Figure 1. Overall, these results demonstrate that GrassDist substantially improves data efficiency on downstream ground-state observables while matching or surpassing the full-data baseline.

3.5 Extrapolation Results on QMugs

To test whether GrassDist remains reliable beyond the training size regime, we directly transfer models trained on `QH9-stable-iid` to larger, unseen drug-like molecules from QMugs (a subset containing 10 to 17 heavy atoms) without fine-tuning.

As shown in Figure 5(a), extrapolation across molecular size reveals a stark contrast. Beyond 13 heavy atoms, the baseline breaks down, with total-energy errors escalating into the 10^2–10^3 kcal/mol regime and exhibiting a very broad 5th–95th percentile band; WALoss fails earlier, at 13 heavy atoms. In contrast, GrassDist exhibits monotonic error scaling: its mean (median) total-energy error grows from 0.20 (0.20) kcal/mol at 10 heavy atoms to 5.39 (5.20) kcal/mol at 17, remaining within chemical accuracy up to 13 heavy atoms (Table 8). Furthermore, GrassDist maintains a consistently narrow error distribution across the entire size range; the same suppression of large errors is observed on `QH9-stable-iid` (Table 1). In Figure 5(b), the baseline predicts distorted frontier orbitals under size transfer, whereas GrassDist faithfully reproduces the reference HOMO and LUMO distributions. A per-size statistical breakdown of this benchmark is reported in Appendix C.4.

4 Related Work

Notable progress has been achieved in predicting KS-DFT Hamiltonians. Early efforts pioneered this direction using kernel ridge regression for one-dimensional semi-empirical Hamiltonians (Hegde & Bowen, 2017), followed by SchNorb (Schütt et al., 2019), which adopted neural networks for molecular systems. Yu et al. (2023b) designed QHNet to intrinsically guarantee equivariance by streamlining tensor products for Hamiltonian block prediction, after which Qian et al. (2026) introduced MACE-H to capture complex many-body correlations up to high-angular-momentum blocks via node degree expansions. Focusing on efficiency and representation fidelity, Luo et al. (2025) proposed SPHNet with adaptive sparse gating to alleviate computational overheads for large basis sets, while Yin et al. (2026) developed NextHAM to leverage the zeroth-step Hamiltonian for residual correction and jointly supervise real- and reciprocal-space representations. Beyond architectural modifications, Li et al. (2025) proposed WALoss, which employs a basis transformation to penalize deviations from reference eigenstructures and thereby mitigates the bottlenecks of element-wise regression. Aiming to scale electronic structure learning across broad chemical space, Kaniselvan et al. (2025) developed HELM to scale to systems with more than 100 atoms, high elemental diversity, and large basis sets, while introducing "Hamiltonian pretraining" to extract transferable atomic environment descriptors for downstream energy prediction. Departing from standard deterministic schemes, Kim et al. (2026a) introduced an equivariant flow matching framework (QHFlow), which was subsequently extended in QHFlow2 via an $\mathrm{SO}(2)$-equivariant backbone to achieve MLIP-level force accuracy (Kim et al., 2026b). Nonetheless, these methods minimize element-wise Hamiltonian errors without strictly respecting the geometry of the Grassmannian manifold associated with the occupied subspace.

5 Conclusions

This work rests on a simple observation: the Hamiltonian is an intermediate representation, while its occupied eigensubspace directly determines the ground-state density matrix. Element-wise Hamiltonian regression does not explicitly account for the accuracy of this subspace. GrassDist therefore complements conventional Hamiltonian supervision with a geometric objective defined directly on the gauge-invariant occupied subspace. This objective directly preserves its intrinsic geometry. Extensive experiments demonstrate that GrassDist improves ground-state observables across benchmarks, backbones, and molecules outside the training distribution. These results support occupied-subspace fidelity as a principled and transferable target for machine-learned electronic structure. Looking ahead, the cost of differentiating through generalized eigendecompositions motivates cheaper or implicit subspace objectives, and extending the formulation to larger, more diverse basis sets and to periodic systems will test how far occupied-subspace fidelity generalizes as a design principle.

AI USE STATEMENT

We used generative AI tools only to assist with language polishing. No experimental data were generated by AI. We have reviewed and verified all AI-assisted text. The authors take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available molecular benchmark datasets and does not involve human subjects, personal data, or clinical decision-making. The method is intended for computational chemistry and materials-science research. We do not identify foreseeable ethical risks beyond those associated with the responsible use, reporting, and release of computational models and benchmark data.

REPRODUCIBILITY STATEMENT

Section 2 and the derivations specify the geometric objective and occupied-subspace formulation. Appendix B documents the training configuration, model architecture, and evaluation protocol used for all reported results.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Jan Almlöf, Knut Fægri Jr, and Knut Korsell. Principles for a direct scf approach to licao–moab-initio calculations. *Journal of Computational Chemistry*, 3(3):385–399, 1982.

Stefano Baroni, Stefano De Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Reviews of modern Physics*, 73(2): 515, 2001.

Axel D Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of chemical physics*, 98(7):5648–5652, 1993.

Åke Björck and Gene H Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

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.

John H Conway, Ronald H Hardin, and Neil JA Sloane. Packing lines, planes, etc.: Packings in grassmannian spaces. *Experimental mathematics*, 5(2):139–159, 1996.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Stefan Goedecker. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4): 1085, 1999.

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.

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

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

Manasa Kaniselvan, Benjamin Kurt Miller, Meng Gao, Juno Nam, and Daniel S Levine. Learning from the electronic structure of molecules across the periodic table. *arXiv preprint arXiv:2510.00224*, 2025.

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, 2026a.

Seongsu Kim, Chanhui Lee, Yoonho Kim, Seongjun Yun, Honghui Kim, Nayoung Kim, Changyoung Park, Sehui Han, Sungbin Lim, and Sungsoo Ahn. Machine learning hamiltonians are accurate energy-force predictors. *arXiv preprint arXiv:2602.16897*, 2026b.

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

Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2):785, 1988.

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.

Rui Li, Qiming Sun, Xing Zhang, and Garnet Kin-Lic Chan. Introducing gpu-acceleration into the python-based simulations of chemistry framework, 2024. URL `https://arxiv.org/abs/2407.09700`.

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*, volume 2025, pp. 918–962, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Per-Olov Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *The Journal of Chemical Physics*, 18(3):365–375, 1950.

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. *arXiv preprint arXiv:2502.01171*, 2025.

Rev McWeeny. Perturbation theory for the fock-dirac density matrix. *Physical Review*, 126(3):1028, 1962.

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.

John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

John A Pople. Nobel lecture: Quantum chemical models. *Reviews of Modern Physics*, 71(5):1267, 1999.

Chen Qian, Valdas Vitartas, James R Kermode, and Reinhard J Maurer. Equivariant electronic hamiltonian prediction with many-body message passing. *npj Computational Materials*, 12(1): 169, 2026.

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.

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, et al. Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S Blunt, Nikolay A Bogdanov, George H Booth, Jia Chen, Zhi-Hao Cui, et al. Recent developments in the pyscf program package. *The Journal of chemical physics*, 153(2), 2020.

Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

Zechen Tang, Haoxiang Chen, Yang Li, Yubing Qian, Yuxiang Wang, Weizhong Fu, Jialin Li, Chen Si, Wenhui Duan, Ji Chen, et al. Deep-learning electronic structure calculations. *Nature Computational Science*, 5(12):1133–1146, 2025.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

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.

JH Van Lenthe, Renate Zwaans, Huub JJ Van Dam, and MF Guest. Starting scf calculations by superposition of atomic densities. *Journal of computational chemistry*, 27(8):926–932, 2006.

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.

Xiaojie Wu, Qiming Sun, Zhichen Pu, Tianze Zheng, Wenzhi Ma, Wen Yan, Xia Yu, Zhengxiao Wu, Mian Huo, Xiang Li, Weiluo Ren, Sheng Gong, Yumin Zhang, and Weihao Gao. Enhancing gpu-acceleration in the python-based simulations of chemistry framework, 2024. URL `https://arxiv.org/abs/2404.09452`.

Shi Yin, Zujian Dai, Xinyang Pan, and Lixin He. Advancing universal deep learning for electronic-structure hamiltonian prediction of materials. In *International Conference on Learning Representations*, volume 2026, pp. 152342–152385, 2026.

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, 2023a.

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*, pp. 40412–40424. PMLR, 2023b.

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. *arXiv preprint arXiv:2403.09560*, 2024.

Jiayao Zhang, Guangxu Zhu, Robert W Heath Jr, and Kaibin Huang. Grassmannian learning: Embedding geometry awareness in shallow and deep learning. *arXiv preprint arXiv:1808.02229*, 2018.

A Additional Technical Background

A.1 Kohn–Sham Density Functional Theory

A molecular configuration is specified by $\mathcal{M} := \{\mathcal{Z}, \mathcal{R}\}$, where $\mathcal{Z} := \{Z^{(a)}\}_{a=1}^{A}$ and $\mathcal{R} := \{\mathbf{R}^{(a)}\}_{a=1}^{A}$ denote the atomic numbers and coordinates of the A nuclei, respectively. For a closed-shell system with N electrons, the $N_{\mathrm{occ}} = N/2$ occupied spatial orbitals determine the electron density as

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N_{\mathrm{occ}}} |\phi_i(\mathbf{r})|^2. \tag{10}$$

Expanding the orbitals in an AO basis $\{\chi_\mu(\mathbf{r})\}_{\mu=1}^{B}$,

$$\phi_i(\mathbf{r}) = \sum_{\mu=1}^{B} (\mathbf{C}_{\mathrm{occ}})_{\mu i}\chi_\mu(\mathbf{r}), \tag{11}$$

gives the occupied coefficient matrix $\mathbf{C}_{\mathrm{occ}} \in \mathbb{R}^{B\times N_{\mathrm{occ}}}$. The occupied orbitals satisfy the $\mathbf{S}$-orthonormality condition $\mathbf{C}_{\mathrm{occ}}^\top \mathbf{S}\mathbf{C}_{\mathrm{occ}} = \mathbf{I}_{N_{\mathrm{occ}}}$, where $\mathbf{S}$ is the AO overlap matrix. The corresponding one-particle density matrix is $\mathbf{P} = 2\mathbf{C}_{\mathrm{occ}}\mathbf{C}_{\mathrm{occ}}^\top$ and satisfies

$$\mathbf{P}\mathbf{S}\mathbf{P} = 2\mathbf{P}, \qquad \mathrm{Tr}(\mathbf{P}\mathbf{S}) = N, \tag{12}$$

These relations express $\mathbf{S}$-idempotency and electron-number conservation.

Ground-state observables can be evaluated from $\mathbf{P}$. In particular, the total energy is

$$E_{\mathrm{tot}} = \sum_{\mu\nu} P_{\mu\nu}h_{\mu\nu} + \frac{1}{2}\sum_{\mu\nu\lambda\sigma} P_{\mu\nu}P_{\lambda\sigma}\left[(\mu\nu|\lambda\sigma) - \frac{a_{\mathrm{x}}}{2}(\mu\sigma|\lambda\nu)\right] + E_{\mathrm{xc}}[\rho] + E_{\mathrm{nuc}}, \tag{13}$$

where $h_{\mu\nu}$ denotes the core-Hamiltonian matrix elements, $(\mu\nu|\lambda\sigma)$ the two-electron repulsion integrals in Mulliken notation, a_{x} the exact-exchange mixing coefficient, $E_{\mathrm{xc}}[\rho]$ the remaining exchange–correlation contribution excluding the explicit exact-exchange term, and E_{nuc} the nuclear-repulsion energy.

The molecular dipole moment $\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathrm{elec}} + \boldsymbol{\mu}^{\mathrm{nuc}}$ is given by

$$\boldsymbol{\mu}^{\mathrm{elec}} = -e\int \mathbf{r}\,\rho(\mathbf{r})\,d\mathbf{r} = -e\ \mathrm{Tr}\left(\mathbf{P}\boldsymbol{\mu}_{1\mathrm{e}}\right), \qquad \boldsymbol{\mu}^{\mathrm{nuc}} = e\sum_{a=1}^{A} Z^{(a)}\mathbf{R}^{(a)}, \tag{14}$$

where e is the elementary charge and $(\boldsymbol{\mu}_{1\mathrm{e}})_{\mu\nu} = \int \chi_\mu(\mathbf{r})\,\mathbf{r}\,\chi_\nu(\mathbf{r})\,d\mathbf{r}$ denotes the one-electron dipole integral.

A.2 First-Order Response of the Density Matrix

We derive Eq. 2, the first-order response of the density matrix to a Hamiltonian perturbation $\delta\mathbf{H}$, for a closed-shell system. We keep the molecular geometry and the AO overlap matrix $\mathbf{S}$ fixed and assume non-degenerate reference orbital energies.

Recall that the molecular-orbital coefficients satisfy the generalized eigenvalue problem

$$\mathbf{H}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\varepsilon}, \qquad \mathbf{C}^\top\mathbf{S}\mathbf{C} = \mathbf{I}, \tag{15}$$

where $\mathbf{C}_p$ denotes the p-th column of $\mathbf{C}$ and ε_p the corresponding orbital energy. Since $\mathbf{S}$ is symmetric positive definite, we introduce the symmetrically orthonormalized coefficients and Hamiltonian

$$\widetilde{\mathbf{X}} = \sqrt{\mathbf{S}}\mathbf{C}, \qquad \bar{\mathbf{H}} = \frac{1}{\sqrt{\mathbf{S}}}\mathbf{H}\frac{1}{\sqrt{\mathbf{S}}}, \tag{16}$$

which transform Eq. 15 into the ordinary symmetric eigenvalue problem $\bar{\mathbf{H}}\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}\boldsymbol{\varepsilon}$ with $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} = \mathbf{I}$. Consider $\mathbf{H} = \mathbf{H}^\star + \delta\mathbf{H}$, which gives $\bar{\mathbf{H}} = \bar{\mathbf{H}}^\star + \delta\bar{\mathbf{H}}$ with $\delta\bar{\mathbf{H}} = \frac{1}{\sqrt{\mathbf{S}}}\delta\mathbf{H}\frac{1}{\sqrt{\mathbf{S}}}$. Letting $\widetilde{\mathbf{x}}_p$ denote the p-th column of $\widetilde{\mathbf{X}}$, standard first-order perturbation theory gives

$$\delta\widetilde{\mathbf{x}}_p = \sum_{q\neq p} \frac{\widetilde{\mathbf{x}}_q^{\star\top}\delta\bar{\mathbf{H}}\,\widetilde{\mathbf{x}}_p^\star}{\varepsilon_p^\star - \varepsilon_q^\star}\widetilde{\mathbf{x}}_q^\star. \tag{17}$$

Using $\widetilde{\mathbf{x}}_p^\star = \sqrt{\mathbf{S}}\mathbf{C}_p^\star$, the matrix element in the numerator satisfies

$$\widetilde{\mathbf{x}}_q^{\star\top}\delta\bar{\mathbf{H}}\,\widetilde{\mathbf{x}}_p^\star = \mathbf{C}_q^{\star\top}\delta\mathbf{H}\,\mathbf{C}_p^\star. \tag{18}$$

Multiplying Eq. 17 by $\frac{1}{\sqrt{\mathbf{S}}}$ therefore gives the first-order variation of the AO coefficients,

$$\delta\mathbf{C}_p = \sum_{q\neq p} \frac{\mathbf{C}_q^{\star\top}\delta\mathbf{H}\,\mathbf{C}_p^\star}{\varepsilon_p^\star - \varepsilon_q^\star}\mathbf{C}_q^\star. \tag{19}$$

For a closed-shell system, the lowest $N_{\mathrm{occ}} = N/2$ orbitals are doubly occupied, and the reference density matrix is

$$\mathbf{P}^\star = 2\sum_{i=1}^{N_{\mathrm{occ}}} \mathbf{C}_i^\star\mathbf{C}_i^{\star\top}. \tag{20}$$

Its first-order variation is

$$\delta\mathbf{P} = 2\sum_{i=1}^{N_{\mathrm{occ}}} \left(\delta\mathbf{C}_i\,\mathbf{C}_i^{\star\top} + \mathbf{C}_i^\star\,\delta\mathbf{C}_i^\top\right). \tag{21}$$

Substituting Eq. 19 and separating the sum over $q \neq i$ into occupied and virtual orbitals, the occupied–occupied contribution is

$$\delta\mathbf{P}_{\text{occ-occ}} = 2\sum_{i=1}^{N_{\mathrm{occ}}}\sum_{\substack{j=1\\ j\neq i}}^{N_{\mathrm{occ}}} \frac{\mathbf{C}_j^{\star\top}\delta\mathbf{H}\,\mathbf{C}_i^\star}{\varepsilon_i^\star - \varepsilon_j^\star}\left(\mathbf{C}_j^\star\mathbf{C}_i^{\star\top} + \mathbf{C}_i^\star\mathbf{C}_j^{\star\top}\right). \tag{22}$$

Because $\delta\mathbf{H}$ is symmetric, $\mathbf{C}_j^{\star\top}\delta\mathbf{H}\,\mathbf{C}_i^\star = \mathbf{C}_i^{\star\top}\delta\mathbf{H}\,\mathbf{C}_j^\star$. The (i,j) and (j,i) terms in Eq. 22 therefore cancel pairwise due to the opposite energy denominators, so $\delta\mathbf{P}_{\text{occ-occ}} = 0$.

The remaining contribution is

$$\delta\mathbf{P} = 2\sum_{i=1}^{N_{\mathrm{occ}}}\sum_{v=N_{\mathrm{occ}}+1}^{B} \frac{\mathbf{C}_v^{\star\top}\delta\mathbf{H}\,\mathbf{C}_i^\star}{\varepsilon_i^\star - \varepsilon_v^\star}\left(\mathbf{C}_v^\star\mathbf{C}_i^{\star\top} + \mathbf{C}_i^\star\mathbf{C}_v^{\star\top}\right). \tag{23}$$

Equivalently, defining the orbital occupation as $f_p = 2$ for occupied orbitals and $f_p = 0$ for virtual orbitals, and ordering the orbitals by increasing energy, this can be written as

$$\delta\mathbf{P} = \sum_{p<q} \frac{f_p - f_q}{\varepsilon_p^\star - \varepsilon_q^\star}\left(\mathbf{C}_q^{\star\top}\delta\mathbf{H}\,\mathbf{C}_p^\star\right)\left(\mathbf{C}_q^\star\mathbf{C}_p^{\star\top} + \mathbf{C}_p^\star\mathbf{C}_q^{\star\top}\right). \tag{24}$$

Dropping the $\star$ superscripts on the reference eigensystem for notational simplicity recovers Eq. 2.

A.3 Optimization Dynamics: Geodesic versus Chordal Loss

Using the principal angles $\{\theta_i\}_{i=1}^{N_{\mathrm{occ}}}$, we compare the squared chordal loss with the squared Grassmannian geodesic loss:

$$\mathcal{L}_{\mathrm{ch}} := \sum_{i=1}^{N_{\mathrm{occ}}} \sin^2\theta_i, \qquad \mathcal{L}_{\mathrm{Gr}} := d_{\mathrm{Gr}}^2(\mathcal{X},\mathcal{Y}) = \sum_{i=1}^{N_{\mathrm{occ}}} \theta_i^2. \tag{25}$$

Their gradients with respect to each principal angle are

$$\frac{\partial\mathcal{L}_{\mathrm{ch}}}{\partial\theta_i} = \sin(2\theta_i), \qquad \frac{\partial\mathcal{L}_{\mathrm{Gr}}}{\partial\theta_i} = 2\theta_i. \tag{26}$$

Hence,

$$\lim_{\theta_i\to\pi/2}\frac{\partial\mathcal{L}_{\mathrm{ch}}}{\partial\theta_i} = 0, \qquad \lim_{\theta_i\to\pi/2}\frac{\partial\mathcal{L}_{\mathrm{Gr}}}{\partial\theta_i} = \pi. \tag{27}$$

This difference is transparent in the idealized gradient flow defined on the principal-angle coordinates,

$$\dot{\theta}_i = -\frac{\partial\mathcal{L}}{\partial\theta_i}. \tag{28}$$

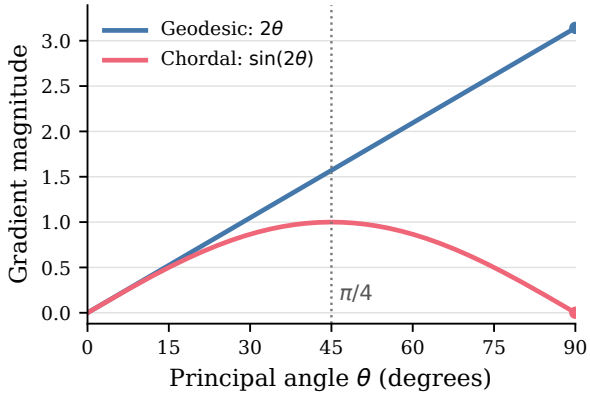


Figure 6: Gradient magnitude with respect to a principal angle for the squared geodesic loss, 2θ, and the squared chordal loss, $\sin(2\theta)$. The chordal gradient vanishes as the subspaces approach orthogonality ($\theta \to \pi/2$), whereas the geodesic gradient remains nonzero.

As $\theta_i \to \pi/2$, the chordal flow satisfies $|\dot{\theta}_i| \to 0$, whereas the geodesic flow retains $|\dot{\theta}_i| \to \pi$. This comparison is confined to the loss geometry in the principal-angle coordinates and isolates the subspace term; it does not by itself characterize the full parameter-space dynamics, which additionally depend on the network Jacobian. Within this reduced space, the chordal objective provides a vanishing correction signal for large principal-angle errors, whereas the geodesic objective remains non-vanishing, as illustrated in Figure 6.

B Experimental Settings and Model Configurations

B.1 Evaluation Protocol

All evaluations follow the standardized train/validation/test partitions listed in Table 4. Due to computational cost, total-energy and SCF acceleration metrics are evaluated on the first 800 test configurations of each split; all other properties are evaluated on the full test sets. The MD17 benchmark contains 5,000 water geometries and 30,000 configurations for each other molecule. For QMugs, we evaluate on a subset of 50 configurations per heavy-atom count (10–17 heavy atoms) using recomputed B3LYP/def2-SVP references.

For each configuration, the generalized eigenvalue problem $\mathbf{HC} = \mathbf{SC}\boldsymbol{\varepsilon}$ is solved in float64 via symmetric orthonormalization, diagonalizing $\bar{\mathbf{H}} = \frac{1}{\sqrt{\mathbf{S}}}\mathbf{H}\frac{1}{\sqrt{\mathbf{S}}}$ to yield orbital energies $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_B)$ and molecular orbital coefficients $\mathbf{C} \in \mathbb{R}^{B\times B}$ satisfying $\mathbf{C}^\top\mathbf{SC} = \mathbf{I}$. The occupied block corresponds to the first N_{occ} columns, $\mathbf{C}_{\text{occ}} \in \mathbb{R}^{B\times N_{\text{occ}}}$. Overlap matrices $\mathbf{S}$ and one-electron dipole integrals $\boldsymbol{\mu}_{1\text{e}}$ are computed with `gpu4pyscf` (Sun et al., 2018; 2020; Li et al., 2024; Wu et al., 2024). Every per-configuration error is averaged over the evaluated structures.

MAE for H Element-wise mean absolute error over all atomic-orbital basis pairs ($B \times B$).

MAE for d_{Gr} The geodesic distance $d_{\text{Gr}} = (\sum_{i=1}^{N_{\text{occ}}} \theta_i^2)^{1/2}$ measures the geometric distance between the predicted and reference occupied subspaces on the Grassmann manifold $\text{Gr}(N_{\text{occ}}, B)$, where θ_i are canonical principal angles from the singular values of the cross-projection matrix $\mathbf{M} = \mathbf{C}_{\text{occ}}^{\star\top}\mathbf{S}\widehat{\mathbf{C}}_{\text{occ}}$.

MAE for P Element-wise mean absolute error over all AO pairs of the one-particle density matrix $\mathbf{P} = 2\mathbf{C}_{\text{occ}}\mathbf{C}_{\text{occ}}^\top$.

MAE for $\boldsymbol{\varepsilon}_{\text{occ}}$ Average absolute error over the N_{occ} occupied orbital-energy levels.

MAE for ε_Δ Absolute error of the HOMO–LUMO gap $\varepsilon_\Delta = \varepsilon_{N_{\mathrm{occ}}+1} - \varepsilon_{N_{\mathrm{occ}}}$.

MAE for E_{tot} Absolute error of the DFT total energy, evaluated with `gpu4pyscf` using the benchmark functional and basis set and including nuclear repulsion.

Error for $\boldsymbol{\mu}$ Euclidean norm $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star\|_2$ between the predicted and reference molecular dipole moments, obtained from the one-electron dipole integrals $\boldsymbol{\mu}_{1\mathrm{e}}$.

SCF Iteration Ratio (S_{acc}) The SCF iteration ratio evaluates the convergence acceleration when initializing downstream SCF iterations with the predicted Hamiltonian, defined as $S_{\mathrm{acc}} = N_{\mathrm{pred}}/N_{\mathrm{MINAO}}$, where N_{pred} and N_{MINAO} count the monitored effective-potential calls (`get_veff`) required to converge from the predicted Hamiltonian and the MINAO guess, respectively.

$\mathbf{C}_{\mathrm{occ}}$ MAE In the coefficient-supervision ablation, the element-wise mean absolute error over all AO and occupied-orbital pairs ($B \times N_{\mathrm{occ}}$).

Cosine Similarity for $\mathbf{C}_{\mathrm{occ}}$ In the coefficient-supervision ablation, the cosine similarity evaluates the column-wise mean absolute cosine similarity $\frac{1}{N_{\mathrm{occ}}} \sum_{i=1}^{N_{\mathrm{occ}}} \frac{|\widehat{\mathbf{C}}_i^\top \mathbf{C}_i^\star|}{\|\widehat{\mathbf{C}}_i\|_2 \|\mathbf{C}_i^\star\|_2}$ between predicted and reference occupied orbital vectors, absorbing trivial column-wise sign flips.

Table 4: QH9 split definitions and sample counts.

Split	Definition	Train	Validation	Test
`QH9-stable-iid`	Stable geometries are randomly partitioned into training, validation, and test sets.	104,664	13,083	13,084
`QH9-stable-ood`	Molecules with 3–20 atoms are used for training, 21–22 atoms for validation, and 23–29 atoms for testing.	104,001	17,495	9,335
`QH9-dynamic-geo`	Each of 999 molecular-dynamics trajectories is partitioned geometry-wise, so test geometries belong to molecules observed during training.	79,920	9,990	9,990
`QH9-dynamic-mol`	The 999 trajectories are partitioned molecule-wise, assigning all geometries of a molecule to one subset.	79,900	9,900	10,100

B.2 Training Details

Training setup. The main experiments use the SPHNet backbone and are trained with AdamW (Loshchilov & Hutter, 2017) (learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.99, 0.999)$, zero weight decay), batch size 32, seed 123, and gradient clipping at 0.09 for 200,000 optimization steps. The learning rate is linearly warmed up over 1,000 steps and then follows a power-one polynomial schedule to 10^{-7}. Training is performed in 32-bit precision with PyTorch (Paszke et al., 2019). Since evaluating the orbital-based losses requires solving the generalized eigenvalue problem with sufficient numerical accuracy, we carry out the occupied-subspace eigendecompositions and the corresponding losses in float64.

The main QH9 results use loss weights $(w_{\mathbf{H}}, w_{\mathrm{Gr}}, w_{\boldsymbol{\varepsilon}}) = (1, 10, 0.25)$. The weight ablation fixes $w_{\boldsymbol{\varepsilon}} = 0$ and varies w_{Gr}; the coefficient-supervision ablation uses $(1, 10, 0)$; and the QHNet backbone ablation reuses $(1, 10, 0.25)$. For WALoss, we adopt the weight $w = 2.5$ reported as the best setting by the original work (Li et al., 2025). Models are selected by minimum validation loss.

Following prior work (Luo et al., 2025), the network predicts the Hamiltonian residual $\Delta\mathbf{H} = \mathbf{H}^\star - \mathbf{H}_{\mathrm{init}}$ relative to a MINAO initial guess (Almlöf et al., 1982; Van Lenthe et al., 2006) $\mathbf{H}_{\mathrm{init}}$, which is added back before eigensolves and evaluation. With the residual parameterization enabled, the shared initial guess cancels in the element-wise difference. All variants, including the baselines and ablations, share the same optimizer, learning-rate schedule, batch size, and optimization-step budget; only the supervision target and its weight are varied.

Gradient stabilization. As discussed in Section 2.2, orbital-based supervision requires differentiating through the eigensolver. For non-degenerate orbital energies, the differential of an occupied orbital coefficient vector is

$$\mathrm{d}\widehat{\mathbf{C}}_i = \sum_{j=1, j\neq i}^{B} \frac{\widehat{\mathbf{C}}_j^\top (\mathrm{d}\widehat{\mathbf{H}})\widehat{\mathbf{C}}_i}{\widehat{\varepsilon}_i - \widehat{\varepsilon}_j}\widehat{\mathbf{C}}_j, \qquad i \leq N_{\mathrm{occ}}. \tag{29}$$

Consequently, the eigensolver backward pass contains inverse orbital-energy differences $K_{ij} = (\widehat{\varepsilon}_i - \widehat{\varepsilon}_j)^{-1}$, which become ill-conditioned near spectral degeneracies.

To stabilize backpropagation, following Zhang et al. (2024), we truncate ill-conditioned components in the eigensolver backward pass:

$$\widetilde{K}_{ij} = \begin{cases} (\widehat{\varepsilon}_i - \widehat{\varepsilon}_j)^{-1}, & |\widehat{\varepsilon}_i - \widehat{\varepsilon}_j| > 10^{-3} \text{ and } \min(i,j) \leq N_{\mathrm{occ}}, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

This directly suppresses near-singular gradient components along the eigensolver backpropagation path while masking terms irrelevant to the occupied subspace.

As an additional safeguard, we supervise the occupied orbital energies and the LUMO,

$$\mathcal{L}_{\boldsymbol{\varepsilon}} = \frac{1}{N_{\mathrm{occ}}+1} \sum_{i=1}^{N_{\mathrm{occ}}+1} |\widehat{\varepsilon}_i - \varepsilon_i^\star|, \tag{31}$$

which helps reduce spurious near-degeneracies in the predicted spectrum before they enter the eigensolver backward pass. The element-wise Hamiltonian loss is

$$\mathcal{L}_{\mathbf{H}} = \frac{1}{|\Omega|} \sum_{(\mu,\nu)\in\Omega} \left(\left| \widehat{\mathbf{H}}_{\mu\nu} - \mathbf{H}^\star_{\mu\nu} \right| + \left(\widehat{\mathbf{H}}_{\mu\nu} - \mathbf{H}^\star_{\mu\nu} \right)^2 \right), \tag{32}$$

where Ω indexes the matrix elements in the molecular batch and $\mathbf{H}^\star$ denotes the reference Hamiltonian. This combination of MAE and MSE is used during training, whereas validation uses the MAE term alone to evaluate the Hamiltonian prediction error. The overall training objective is

$$\mathcal{L}_{\mathrm{total}} = w_{\mathbf{H}}\mathcal{L}_{\mathbf{H}} + w_{\mathrm{Gr}}\mathcal{L}_{\mathrm{Gr}} + w_{\boldsymbol{\varepsilon}}\mathcal{L}_{\boldsymbol{\varepsilon}}. \tag{33}$$

Hyperparameter sensitivity analyses for w_{Gr} and $w_{\boldsymbol{\varepsilon}}$ are reported in Appendix C (Table 5–Table 6).

B.3 Model Architecture

Following previous Hamiltonian prediction methods (Unke et al., 2021; Yu et al., 2023b; Gong et al., 2023; Qian et al., 2026; Luo et al., 2025), we employ $\mathrm{SE}(3)$-equivariant neural networks to predict the AO Hamiltonian. These models represent atomic features by irreducible representations (irreps) $\mathbf{h}^\ell$ of angular order ℓ. Given two irreps $\mathbf{u}^{\ell_1}$ and $\mathbf{v}^{\ell_2}$, their Clebsch–Gordan tensor product (Thomas et al., 2018; Geiger & Smidt, 2022) is

$$\left(\mathbf{u}^{\ell_1} \otimes \mathbf{v}^{\ell_2}\right)^{\ell_3}_{m_3} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} \mathcal{C}_{\mathrm{CG}}{}^{(\ell_3,m_3)}_{(\ell_1,m_1),(\ell_2,m_2)} u^{\ell_1}_{m_1} v^{\ell_2}_{m_2}, \tag{34}$$

where ℓ_1, ℓ_2, and ℓ_3 denote the angular orders of the two input irreps and the output irrep, respectively, with $|\ell_1 - \ell_2| \leq \ell_3 \leq \ell_1 + \ell_2$; $m_k \in \{-\ell_k, \ldots, \ell_k\}$ indexes the components of an order-ℓ_k irrep; and $\mathcal{C}_{\mathrm{CG}}$ denotes the Clebsch–Gordan coefficients.

For atoms a and b, let $\mathbf{r}_{ab} = \mathbf{R}^{(b)} - \mathbf{R}^{(a)}$ and $\widehat{\mathbf{r}}_{ab} = \mathbf{r}_{ab}/\|\mathbf{r}_{ab}\|_2$. Combining atomic irreps with spherical harmonics of the relative direction gives a generic equivariant message

$$\mathbf{c}^{\ell_o}_{ab} = \sum_{\ell_i,\ell_f} \mathbf{W}^{(\ell_i,\ell_f,\ell_o)}_{ab} \left(\mathbf{h}^{\ell_i}_b \otimes Y^{\ell_f}(\widehat{\mathbf{r}}_{ab}) \right)^{\ell_o}, \tag{35}$$

Here a and b index atoms, and ℓ_i, ℓ_f, and ℓ_o denote the angular orders of the input atomic feature, the spherical-harmonic filter, and the output message, respectively. Moreover, Y^{ℓ_f} is the spherical harmonic of order ℓ_f, and $\mathbf{W}^{(\ell_i,\ell_f,\ell_o)}_{ab}$ collects learnable pair-dependent weights.

The resulting atomic representations are further converted into pairwise irreps $\mathbf{f}_{ab}^{\ell_o}$ and expanded according to the angular momenta of the AO basis to construct the Hamiltonian blocks:

$$\widehat{\mathbf{H}}_{ab,\zeta}^{(\ell_1,\ell_2)} = \sum_{\ell_o,\zeta'} \mathbf{U}_{ab,\zeta\zeta'}^{(\ell_1,\ell_2,\ell_o)} \left(\bar{\otimes}_{\ell_o} \mathbf{f}_{ab,\zeta'}^{\ell_o}\right), \tag{36}$$

where ℓ_1 and ℓ_2 denote the angular momenta of the two AO shells defining the output block, ℓ_o is the angular order of the pairwise irrep, ζ' and ζ index the input and output channels, respectively, $\bar{\otimes}_{\ell_o}$ denotes the Clebsch–Gordan tensor expansion from an order-ℓ_o irrep to the corresponding (ℓ_1, ℓ_2) AO block, and $\mathbf{U}_{ab,\zeta\zeta'}^{(\ell_1,\ell_2,\ell_o)}$ denotes the learnable channel-mixing weights. The atomic-pair blocks are finally assembled into the full Hamiltonian $\widehat{\mathbf{H}}$.

For the main experiments, we instantiate this framework with SPHNet (Luo et al., 2025). Its node representations are constructed by four vectorial interaction blocks over short-range (5 Å) and long-range (15 Å) neighborhoods, followed by two spherical interaction blocks that introduce higher-order irreps. We use angular orders up to $L_{\max} = 4$ for the def2-SVP basis. Diagonal and non-diagonal pair modules construct intra-atomic ($a = b$) and inter-atomic ($a \neq b$) pair features, respectively, which are mapped to the corresponding AO Hamiltonian blocks by the expansion module. To verify that the proposed Grassmannian supervision is not tied to a particular network architecture, we additionally conduct backbone ablations with QHNet (Yu et al., 2023b); the results are reported in Appendix C.5.

C Additional Results

C.1 Grassmann-Weight Ablation

Figure 7 and Table 5 sweep the Grassmannian weight w_{Gr} with $w_{\mathbf{H}} = 1$ and $w_{\boldsymbol{\varepsilon}} = 0$ on `QH9-stable-iid`, thus isolating the contribution of occupied-subspace supervision on top of Hamiltonian regression. Under this setting, Hamiltonian regression augmented solely with occupied-subspace supervision already yields substantial gains: even a small weight $w_{\mathrm{Gr}} = 0.05$ lowers the total-energy MAE from 4.87 to 0.12 kcal/mol and reduces the density and subspace errors by up to an order of magnitude. The ground-state metrics then show diminishing returns beyond $w_{\mathrm{Gr}} = 10$ (e.g. $E_{\mathrm{tot}} = 0.03$ kcal/mol, more than 99% below the baseline), although d_{Gr} and $\mathbf{P}$ continue to decrease slowly at larger weights. The $\mathbf{H}$ MAE rises with geometric weighting, but the physical observables remain accurate. At larger weights (500 and ∞) the occupied-orbital-energy distribution and the gap error grow (Figure 7b; Table 5); since subspace supervision alone leaves a residual gap error of order $10^{-2}\,E_{\mathrm{h}}$, finer tuning of w_{Gr} yields little spectral benefit. We therefore fix $w_{\mathrm{Gr}} = 10$ and include a small auxiliary orbital-energy term $\mathcal{L}_{\boldsymbol{\varepsilon}}$ to improve spectral accuracy without compromising subspace fidelity (Table 6).

Table 5: Grassmann-weight ablation on `QH9-stable-iid` with $w_{\boldsymbol{\varepsilon}} = 0$ and $w_{\mathbf{H}} = 1$; $w_{\mathrm{Gr}} = 0$ is the Hamiltonian-only baseline and $w_{\mathrm{Gr}} = \infty$ the pure Grassmannian objective ($w_{\mathbf{H}} = 0$). For E_{tot}, values are reported as the mean (median) absolute error. The best result in each column is boldfaced.

w_{Gr}	$\mathbf{H}$ ($10^{-6}E_{\mathrm{h}}$) ↓	d_{Gr} (10^{-3}) ↓	$\mathbf{P}$ (10^{-3}) ↓	E_{tot} (kcal/mol) ↓	$\boldsymbol{\varepsilon}_{\Delta}$ ($10^{-6}E_{\mathrm{h}}$) ↓	$\boldsymbol{\varepsilon}_{\mathrm{occ}}$ ($10^{-6}E_{\mathrm{h}}$) ↓	$\boldsymbol{\mu}$ (10^{-3}a.u.) ↓	S_{acc} ↓
0	**41.03**	66.28	9.01	4.87 (0.28)	**1607.37**	488.08	21.36	67.2%
0.05	50.78	9.21	0.61	0.12 (0.11)	1938.70	**202.55**	9.16	64.9%
10	191.72	4.32	0.24	0.03 (0.03)	13983.00	1285.47	6.25	61.7%
100	1754.71	3.41	0.18	0.02 (0.02)	12694.79	20667.11	5.50	60.2%
500	3409.40	**2.80**	**0.13**	**0.02** (**0.01**)	17286.88	50452.46	**5.01**	**59.1**%
∞	4074.99	3.59	0.17	0.03 (0.02)	18665.26	66006.06	6.67	61.2%

C.2 Sensitivity to the Auxiliary Orbital-Energy Weight

Table 6 and Figure 8 fix $w_{\mathrm{Gr}} = 10$ and vary the orbital-energy weight $w_{\boldsymbol{\varepsilon}}$ on `QH9-stable-iid`. Increasing $w_{\boldsymbol{\varepsilon}}$ monotonically improves the HOMO–LUMO gap ($13983.00 \rightarrow 309.98$ in $10^{-6}E_{\mathrm{h}}$) while degrading the subspace, density, and total-energy metrics; the dipole error is minimized at the

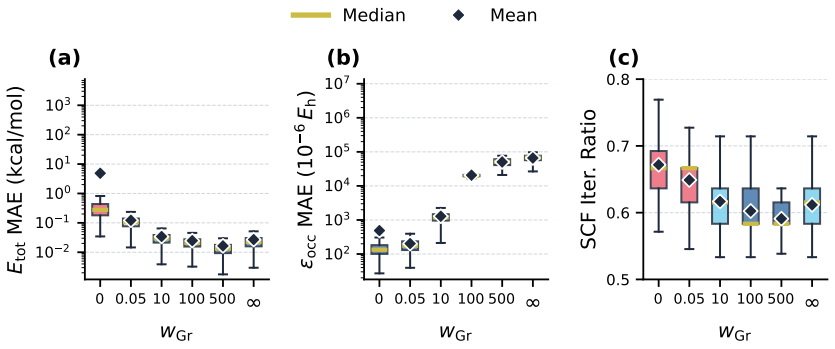


Figure 7: Grassmann-weight ablation on `QH9-stable-iid` with $w_{\varepsilon} = 0$ and $w_{\mathbf{H}} = 1$. Per-configuration distribution of (a) total-energy error, (b) occupied-orbital-energy error, and (c) SCF iteration ratio versus the Grassmannian weight w_{Gr}; the rightmost setting $w_{\text{Gr}} = \infty$ is the pure Grassmannian objective ($w_{\mathbf{H}} = 0$). Boxes span the interquartile range, yellow lines mark medians, and diamonds mark means; outlier markers are hidden.

default $w_{\varepsilon} = 0.25$ and the SCF iteration ratio at $w_{\varepsilon} = 0.05$. The default thus keeps the subspace metrics near their $w_{\varepsilon} = 0$ optimum while substantially reducing the gap error.

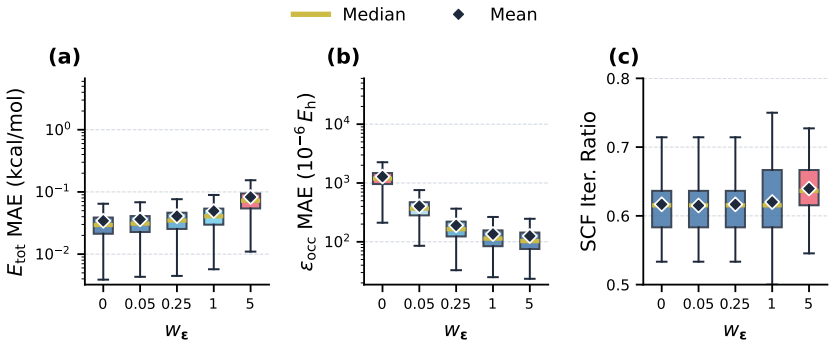


Figure 8: Orbital-energy weight ablation on top of GrassDist ($w_{\text{Gr}} = 10$) on `QH9-stable-iid`. Per-configuration distribution of (a) total-energy error, (b) occupied-orbital-energy error, and (c) SCF iteration ratio as the auxiliary orbital-energy weight w_{ε} increases. Boxes span the interquartile range, yellow lines mark medians, and diamonds mark means; outlier markers are hidden.

Table 6: Orbital-energy weight ablation on top of GrassDist ($w_{\text{Gr}} = 10$) on `QH9-stable-iid`; all runs with $w_{\varepsilon} > 0$ supervise the energy of occupied orbitals and the LUMO. For E_{tot}, values are reported as the mean (median) absolute error. The best result in each column is boldfaced.

w_{ε}	$\mathbf{H}$ ($10^{-6} E_{\text{h}}$) ↓	d_{Gr} (10^{-3}) ↓	$\mathbf{P}$ (10^{-3}) ↓	E_{tot} (kcal/mol) ↓	ε_{Δ} ($10^{-6} E_{\text{h}}$) ↓	ε_{occ} ($10^{-6} E_{\text{h}}$) ↓	$\boldsymbol{\mu}$ (10^{-3} a.u.) ↓	S_{acc} ↓
0	191.72	**4.32**	**0.24**	**0.03** (0.03)	13983.00	1285.47	6.25	61.7%
0.05	**178.66**	4.47	0.24	0.04 (0.03)	4748.63	404.10	6.21	**61.5%**
0.25	182.26	4.72	0.26	0.04 (0.03)	889.23	190.54	**6.04**	61.7%
1	194.57	5.12	0.28	0.05 (0.04)	399.00	135.67	6.23	62.0%
5	265.86	6.66	0.37	0.08 (0.07)	**309.98**	**125.48**	7.75	64.0%

C.3 Ablation of Auxiliary Orbital-Energy Supervision

To isolate the role of the auxiliary orbital-energy loss, Table 7 compares it with occupied-subspace supervision. On `QH9-stable-iid`, $\mathcal{L}_{\varepsilon}$ primarily improves the HOMO–LUMO gap, whereas

$\mathcal{L}_{\mathrm{Gr}}$ provides the dominant gains in ground-state quantities, reducing the density error from 2.25 to 0.24 and the total-energy error from 1.20 to 0.03 kcal/mol. Adding the auxiliary $\mathcal{L}_{\boldsymbol{\varepsilon}}$ term to $\mathcal{L}_{\mathrm{Gr}}$ substantially improves spectral accuracy while largely preserving these ground-state gains. The same trend holds on `QH9-stable-ood`, where adding $\mathcal{L}_{\mathrm{Gr}}$ at matched $w_{\boldsymbol{\varepsilon}} = 0.25$ reduces the density and total-energy errors from 1.42 to 0.22 and from 0.82 to 0.04 kcal/mol, respectively.

Table 7: Comparison of auxiliary supervision schemes on `QH9-stable-iid` and `QH9-stable-ood`. The columns $(w_{\mathbf{H}} : w_{\mathrm{Gr}} : w_{\boldsymbol{\varepsilon}})$ denote the weights of the Hamiltonian, occupied-subspace, and orbital-energy terms. For E_{tot}, values are reported as the mean (median) absolute error. The best result in each column within each split is boldfaced.

Split	$(w_{\mathbf{H}} : w_{\mathrm{Gr}} : w_{\boldsymbol{\varepsilon}})$	$\mathbf{H}$ $(10^{-6}E_{\mathrm{h}})\downarrow$	d_{Gr} $(10^{-3})\downarrow$	$\mathbf{P}$ $(10^{-3})\downarrow$	E_{tot} (kcal/mol) $\downarrow$	$\boldsymbol{\varepsilon}_{\Delta}$ $(10^{-6}E_{\mathrm{h}})\downarrow$	$\boldsymbol{\varepsilon}_{\mathrm{occ}}$ $(10^{-6}E_{\mathrm{h}})\downarrow$	$\boldsymbol{\mu}$ $(10^{-3}\mathrm{a.u.})\downarrow$	S_{acc} $\downarrow$
stable-iid	(1 : 0 : 0.25)	**78.08**	29.43	2.25	1.20 (1.03)	**289.78**	**139.23**	26.43	72.3%
	(1 : 10 : 0)	191.72	**4.32**	**0.24**	**0.03** (**0.03**)	13983.00	1285.47	6.25	61.7%
	(1 : 10 : 0.25)	182.26	4.72	0.26	0.04 (0.03)	889.23	190.54	**6.04**	**61.7**%
stable-ood	(1 : 0 : 0.25)	**61.66**	24.91	1.42	0.82 (0.81)	**571.85**	**118.48**	18.34	73.4%
	(1 : 10 : 0.25)	124.09	**4.96**	**0.22**	**0.04** (**0.04**)	634.11	150.18	**5.09**	**63.7**%

C.4 Per-Size Breakdown on QMugs Extrapolation

Table 8 reports the statistical breakdown of the size-extrapolation benchmark from Section 3.5. For each heavy-atom count from 10 to 17, we report the mean and median with the 5th–95th percentile range of the total-energy error over the $n = 50$ test molecules. The baseline develops failure cases beyond 13 heavy atoms—some configurations exceed 20 kcal/mol—which inflate its mean and widen its percentile range, and WALoss shows extreme outliers already at 13; in contrast, GrassDist stays sub-kcal/mol through 13 heavy atoms and then degrades smoothly.

C.5 Backbone Ablation with QHNet

To verify that the benefits of occupied-subspace supervision are agnostic to the choice of equivariant backbone, we evaluate GrassDist ($(w_{\mathbf{H}}, w_{\mathrm{Gr}}, w_{\boldsymbol{\varepsilon}}) = (1, 10, 0.25)$) against the element-wise Hamiltonian baseline using QHNet (Yu et al., 2023b) across both the chemical-space benchmark (`QH9-stable-iid`) and the molecular dynamics trajectory (MD17 uracil).

As shown in Table 9, GrassDist consistently delivers substantial improvements across both benchmarks when deployed on QHNet. On `QH9-stable-iid`, GrassDist reduces the geodesic distance, density-matrix MAE ($\mathbf{P}$), and total-energy MAE by 94.8%, 97.6%, and 99.3%, respectively, while also improving the occupied-orbital energy, HOMO–LUMO gap, dipole moment, and downstream SCF acceleration. On the MD17 uracil trajectory, GrassDist reduces d_{Gr} by 52.0%, the density error by 63.9%, and the total-energy MAE by 68.6%, while modestly improving the gap and dipole errors. In both cases, these physical improvements confirm that prioritizing occupied-subspace geometry provides consistent benefits independent of backbone architecture.

C.6 Ablation of Direct Coefficient-level Supervision

To compare the geometric formulation on the Grassmannian manifold $\mathrm{Gr}(N_{\mathrm{occ}}, B)$ with direct coefficient-level supervision, we ablate different loss targets defined on the occupied molecular-orbital coefficient matrix $\mathbf{C}_{\mathrm{occ}} \in \mathbb{R}^{B \times N_{\mathrm{occ}}}$. We evaluate three strategies alongside the unconstrained Hamiltonian regression baseline on the `QH9-stable-iid` benchmark with the SPH-Net backbone: (1) $\mathbf{C}_{\mathrm{occ}}$ **MAE**, which imposes an element-wise mean absolute error on the coefficient matrices, $\mathcal{L}_{\mathrm{MAE}} = \sum_{\mu=1}^{B}\sum_{i=1}^{N_{\mathrm{occ}}} |\widehat{\mathbf{C}}_{\mu i} - \mathbf{C}^{\star}_{\mu i}|/(BN_{\mathrm{occ}})$; (2) $\mathbf{C}_{\mathrm{occ}}$ **sim**, which optimizes the orbital-wise cosine similarity directly on the raw coefficient vectors, $\mathcal{L}_{\mathrm{CosSim}} = 1 - \frac{1}{N_{\mathrm{occ}}}\sum_{i=1}^{N_{\mathrm{occ}}} |\widehat{\mathbf{C}}_i^{\top}\mathbf{C}^{\star}_i|/(\|\widehat{\mathbf{C}}_i\|_2\|\mathbf{C}^{\star}_i\|_2)$, taking the absolute value to absorb trivial column-wise sign flips; and (3) **GrassDist**, which supervises the occupied subspace via the squared gauge-invariant geodesic loss under the $\mathbf{S}$-metric. Here $w_{\mathrm{tar}}/w_{\mathbf{H}}$ is the auxiliary-to-Hamiltonian weight ratio and $w_{\boldsymbol{\varepsilon}} = 0$.

Table 8: QMugs extrapolation breakdown by molecular size ($n = 50$ per heavy-atom count). total-energy error (kcal/mol) for the Hamiltonian-only baseline, WALoss, and GrassDist, reported as the mean and median with the 5th–95th percentile range $[Q_5, Q_{95}]$.

Heavy Atoms	Method	Mean	Median	$[Q_5, Q_{95}]$
10	Baseline	0.69	0.70	[0.50, 0.90]
	WALoss	1.05	0.97	[0.77, 1.55]
	GrassDist	**0.20**	**0.20**	**[0.15, 0.25]**
11	Baseline	1.22	1.13	[0.75, 1.97]
	WALoss	1.87	1.51	[1.09, 3.60]
	GrassDist	**0.29**	**0.27**	**[0.20, 0.44]**
12	Baseline	2.71	2.36	[1.23, 5.55]
	WALoss	3.89	2.73	[1.76, 6.11]
	GrassDist	**0.47**	**0.44**	**[0.32, 0.69]**
13	Baseline	6.78	5.99	[1.87, 13.29]
	WALoss	131.07	5.59	[2.98, 1355.40]
	GrassDist	**0.82**	**0.73**	**[0.51, 1.55]**
14	Baseline	754.97	620.00	[5.34, 1629.08]
	WALoss	387.80	14.62	[4.97, 1578.82]
	GrassDist	**1.42**	**1.26**	**[0.91, 2.19]**
15	Baseline	1153.28	1425.97	[8.04, 2930.07]
	WALoss	732.35	22.79	[6.58, 3538.68]
	GrassDist	**2.29**	**2.21**	**[1.32, 3.74]**
16	Baseline	2158.26	2155.23	[20.70, 3643.43]
	WALoss	1487.16	1317.88	[11.93, 4424.69]
	GrassDist	**3.53**	**3.55**	**[1.96, 5.31]**
17	Baseline	3517.10	3617.29	[1262.28, 5070.87]
	WALoss	2135.57	1701.86	[31.89, 5436.30]
	GrassDist	**5.39**	**5.20**	**[3.47, 7.87]**

Table 9: Backbone ablation using QHNet on `QH9-stable-iid` and MD17 uracil. Metrics, units, and lower-is-better directions follow Table 1 and Table 3; Baseline denotes Hamiltonian-only QH-Net. For E_{tot}, values are the mean (median) absolute error. The best result in each column within each benchmark is boldfaced.

Benchmark	Method	$\mathbf{H}$ $(10^{-6}E_{\text{h}})\downarrow$	d_{Gr} $(10^{-3})\downarrow$	$\mathbf{P}\downarrow$	E_{tot} (kcal/mol)$\downarrow$	ε_{Δ} $(10^{-6}E_{\text{h}})\downarrow$	ε_{occ} $(10^{-6}E_{\text{h}})\downarrow$	$\boldsymbol{\mu}\downarrow$	$S_{\text{acc}}\downarrow$
`QH9-stable-iid`	Baseline	**40.89**	86.19	10.16	5.04 (0.25)	2020.29	202.50	30.45	66.8%
($\mathbf{P}: 10^{-3}$, $\boldsymbol{\mu}: 10^{-3}$ a.u.)	GrassDist	160.85	**4.45**	**0.24**	**0.04** (**0.03**)	**916.21**	**190.79**	**6.03**	**61.3%**
MD17 uracil	Baseline	**3.75**	1.27	103.28	2.179×10^{-3} (1.674×10^{-3})	20.10	19.60	1457.17	43.6%
($\mathbf{P}: 10^{-6}$, $\boldsymbol{\mu}: 10^{-6}$ a.u.)	GrassDist	10.08	**0.61**	**37.25**	$\mathbf{6.841 \times 10^{-4}}$ ($\mathbf{5.472 \times 10^{-4}}$)	**17.70**	**16.94**	**1080.14**	**43.0%**

Table 10 reports the full ratio sweep for both coefficient targets. At the smallest weight ($w_{\text{tar}}/w_{\mathbf{H}} = 0.005$) both coefficient losses improve on the baseline for most physical observables, with $E_{\text{tot}} \leq 0.26$ kcal/mol and dipole errors below the baseline. Raising the ratio, however, degrades the physical metrics: for $\mathbf{C}_{\text{occ}}$ MAE the subspace, density, and total-energy errors are lowest at 0.05, while the Hamiltonian error grows monotonically ($48.10 \rightarrow 489.20 \times 10^{-6}\,E_{\text{h}}$ from 0.005 to 10) and the total-energy MAE climbs from 0.19 to 4.15 kcal/mol; $\mathbf{C}_{\text{occ}}$ sim degrades faster, reaching $E_{\text{tot}} = 6.39$ kcal/mol at 10. At the common nominal weight ($w_{\text{tar}}/w_{\mathbf{H}} = 0.05$), GrassDist gives the best geodesic distance (9.21), total-energy error (0.12 kcal/mol), dipole error (9.16×10^{-3} a.u.), and intrinsic coefficient MAE (20.07×10^{-3}), while $\mathbf{C}_{\text{occ}}$ sim gives the lowest density-matrix error (0.60×10^{-3}).

The intrinsic $\mathbf{C}_{\text{occ}}$ columns show that neither coefficient loss minimizes its own target. The smallest coefficient MAE is obtained not by the MAE-trained model (22.14×10^{-3} at 0.05) but by Grass-

Table 10: Ablation of occupied-orbital supervision targets on `QH9-stable-iid`. Metrics follow Table 1, including the density MAE, mean dipole vector error, and mean geodesic distance. The weight column is the auxiliary-to-Hamiltonian ratio $w_{\mathrm{tar}}/w_{\mathbf{H}}$, where ∞ denotes the coefficient-only control with $w_{\mathbf{H}} = 0$. Best results per column across all configurations are boldfaced. The two $\mathbf{C}_{\mathrm{occ}}$ columns report the intrinsic coefficient metrics of the predicted occupied orbitals (lower MAE and higher cosine similarity are better).

Method	$w_{\mathrm{tar}}/w_{\mathbf{H}}$	$\mathbf{H}$ ($10^{-6}E_{\mathrm{h}}$) ↓	d_{Gr} (10^{-3}) ↓	$\mathbf{P}$ (10^{-3}) ↓	E_{tot} (kcal/mol) ↓	$\boldsymbol{\mu}$ (10^{-3}a.u.) ↓	$\mathbf{C}_{\mathrm{occ}}$ MAE (10^{-3}) ↓	$\mathbf{C}_{\mathrm{occ}}$ sim (%) ↑
Baseline	0	**41.03**	66.28	9.01	4.87 (0.28)	21.36	20.57	98.9%
$\mathbf{C}_{\mathrm{occ}}$ MAE	0.005	48.10	14.14	0.99	0.26 (0.21)	14.85	20.37	99.2%
$\mathbf{C}_{\mathrm{occ}}$ MAE	0.05	51.42	12.64	0.79	0.19 (0.16)	17.37	22.14	99.2%
$\mathbf{C}_{\mathrm{occ}}$ MAE	1	146.80	20.39	0.98	0.59 (0.54)	41.09	24.62	98.5%
$\mathbf{C}_{\mathrm{occ}}$ MAE	10	489.20	31.93	1.25	4.15 (4.01)	60.57	24.83	97.8%
$\mathbf{C}_{\mathrm{occ}}$ MAE (pure)	∞	4856.10	37.40	1.05	12.93 (12.49)	52.35	24.97	98.0%
$\mathbf{C}_{\mathrm{occ}}$ sim	0.005	52.77	12.30	0.61	0.18 (0.15)	13.91	20.89	**99.3**%
$\mathbf{C}_{\mathrm{occ}}$ sim	0.05	105.29	17.39	0.60	0.38 (0.33)	25.71	24.17	99.1%
$\mathbf{C}_{\mathrm{occ}}$ sim	10	3230.79	41.07	0.78	6.39 (5.57)	61.36	25.29	99.1%
GrassDist	0.05	50.78	9.21	0.61	0.12 (0.11)	9.16	**20.07**	99.2%
GrassDist	10	191.72	4.32	0.24	0.03 (0.03)	**6.25**	24.10	98.5%
GrassDist (pure)	∞	4074.99	**3.59**	**0.17**	**0.03** (**0.02**)	6.67	28.11	84.3%

Dist (20.07×10^{-3}), and the largest cosine similarity is obtained by $\mathbf{C}_{\mathrm{occ}}$ sim only at the smallest weight (99.3% at 0.005). The coefficient-only control ($w_{\mathrm{tar}}/w_{\mathbf{H}} = \infty$, Hamiltonian supervision removed) confirms that direct coefficient fitting cannot replace Hamiltonian regression: it still yields $\mathbf{H} = 4856.10 \times 10^{-6}\,E_{\mathrm{h}}$ and $E_{\mathrm{tot}} = 12.93$ kcal/mol, with a coefficient MAE of 24.97×10^{-3} that is not even the smallest in the table. The pure geometric objective at the same ratio behaves oppositely: with the Hamiltonian term removed entirely it still attains $E_{\mathrm{tot}} = 0.03$ kcal/mol and the best geodesic distance (3.59) and density-matrix error (0.17×10^{-3}) in the table, so occupied-subspace supervision alone suffices while direct coefficient fitting does not. Its coefficient cosine similarity is nevertheless low (84.3%) and its coefficient MAE (28.11×10^{-3}) is the largest in the table, consistent with a correct occupied subspace whose individual eigenvector columns are gauge-ambiguous.

C.7 Outlier Case Study: Quasi-One-Dimensional Chains

The dataset-level averages of Table 1 hide a small number of large baseline outliers. We inspect the three test configurations with the largest Hamiltonian-only total-energy error on `QH9-stable-iid` and report their observables for both models at their best validation checkpoints in Table 11.

Table 11: Case study of the three largest Hamiltonian-only total-energy outliers on `QH9-stable-iid` (out of 13,084 test configurations). Both models use their best validation checkpoints. Metrics and units follow Table 1: $\mathbf{H}$ and $\boldsymbol{\varepsilon}_{\Delta}$ in $10^{-6}\,E_{\mathrm{h}}$, d_{Gr} in 10^{-3}, the density matrix $\mathbf{P}$ in 10^{-3}, the dipole vector error $\boldsymbol{\mu}$ in 10^{-3} a.u., and the SCF iteration ratio S_{acc}. Lower is better in every column; the better of the two methods is boldfaced per molecule.

Molecule	Method	$\mathbf{H}$ ($10^{-6}\,E_{\mathrm{h}}$) ↓	d_{Gr} (10^{-3}) ↓	$\mathbf{P}$ (10^{-3}) ↓	E_{tot} (kcal/mol) ↓	$\boldsymbol{\varepsilon}_{\Delta}$ ($10^{-6}\,E_{\mathrm{h}}$) ↓	$\boldsymbol{\mu}$ (10^{-3} a.u.) ↓	S_{acc} ↓
C_2N_2	Baseline	**286.76**	1573.02	37259.32	2213.94	279547.42	190.95	122.2%
	GrassDist	543.65	**14.50**	**4.76**	**0.62**	**2967.05**	**33.32**	**100.0**%
C_8H_6O	Baseline	**43.61**	1570.95	2952.92	1321.25	171963.16	5108.86	116.7%
	GrassDist	146.92	**6.20**	**0.68**	**0.07**	**1524.14**	**2.76**	**75.0**%
C_9H_{12}	Baseline	**29.00**	116.31	53.01	16.74	174035.80	44.48	83.3%
	GrassDist	114.03	**3.92**	**0.29**	**0.04**	**352.21**	**2.65**	**66.7**%

Figure 9 shows the three molecules, all drawn to a common scale.

The three largest baseline failures are all acyclic, chain-like molecules, i.e., quasi-one-dimensional: cyanogen (C_2N_2) is essentially linear and spans only 3.7 Å, whereas C_8H_6O and C_9H_{12} are carbon chains spanning 11.4 and 11.6 Å(Figure 9). For the baseline, the element-wise Hamiltonian error on

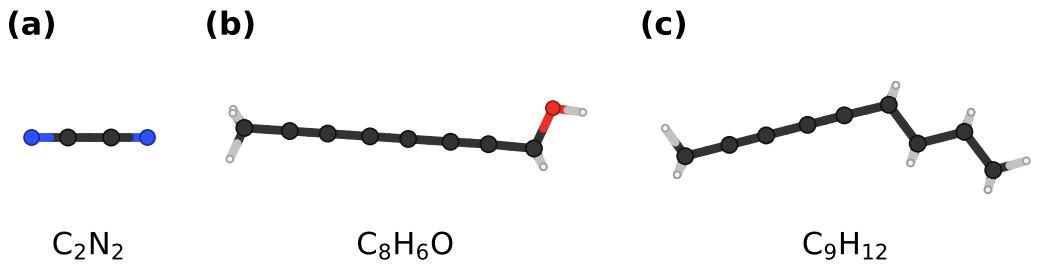


Figure 9: Structures of the three `QH9-stable-iid` configurations for which the Hamiltonian-only SPHNet baseline yields its largest total-energy errors: (a) C_2N_2 (total-energy error 2213.94 kcal/mol), (b) C_8H_6O (1321.25 kcal/mol), and (c) C_9H_{12} (16.74 kcal/mol). GrassDist reduces these errors to 0.62, 0.07, and 0.04 kcal/mol, respectively.

these configurations is small, yet the subspace, density, energy, and gap errors are large (Table 11); GrassDist instead matches the reference occupied subspace and the observables. To characterize the predicted spectra, we inspect the frontier orbital levels of C_8H_6O (Figure 10) and C_2N_2 (Figure 11).

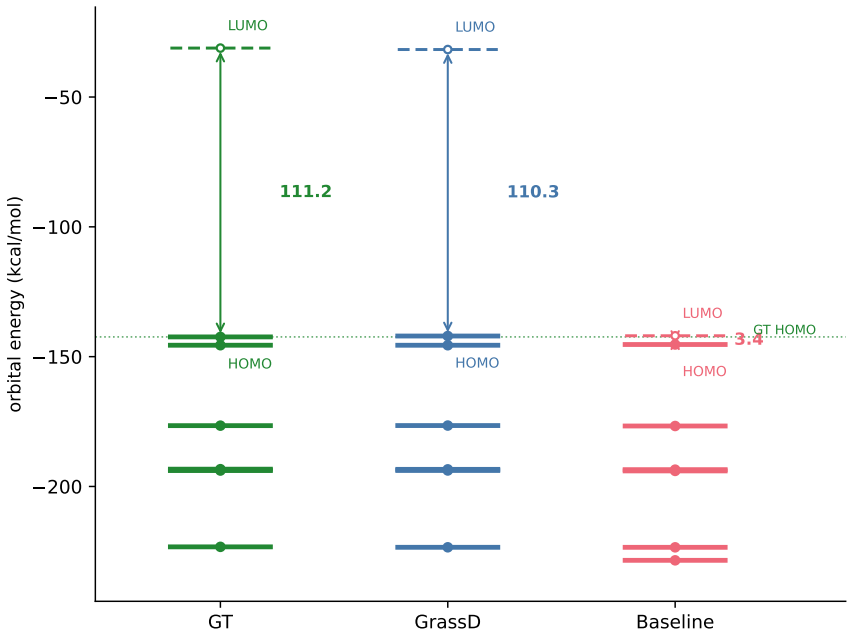


Figure 10: Frontier orbital levels of C_8H_6O for GT, GrassDist, and the Hamiltonian-only baseline. Solid filled bars are occupied levels and the dashed open bar is the LUMO; the arrow spans the HOMO–LUMO gap (in kcal/mol), and the dotted line marks the GT HOMO energy.

For C_2N_2, the reference HOMO and LUMO are each two-fold degenerate; the baseline predicts a closed 0.0 kcal/mol gap and places its lowest unoccupied level at the reference HOMO energy, whereas GrassDist recovers the gap (173.6 kcal/mol vs. the reference 175.4 kcal/mol).

For C_8H_6O, the reference HOMO–LUMO gap is 111.2 kcal/mol; the baseline predicts 3.4 kcal/mol and places its lowest unoccupied level at -142.0 kcal/mol, essentially the reference HOMO energy (-142.4 kcal/mol), whereas GrassDist recovers the gap (110.3 kcal/mol) and frontier positions. In both molecules the baseline's predicted occupied subspace differs from the reference, consistent with its large $\mathbf{P}$ and E_{tot} errors despite a small $\mathbf{H}$ error; GrassDist reproduces the reference occupied subspace.

Taken together, Table 11 and Figure 9, Figure 10, Figure 11 and Figure 12–Figure 14 show that a small element-wise Hamiltonian error does not imply accurate observables: for the baseline the $\mathbf{H}$

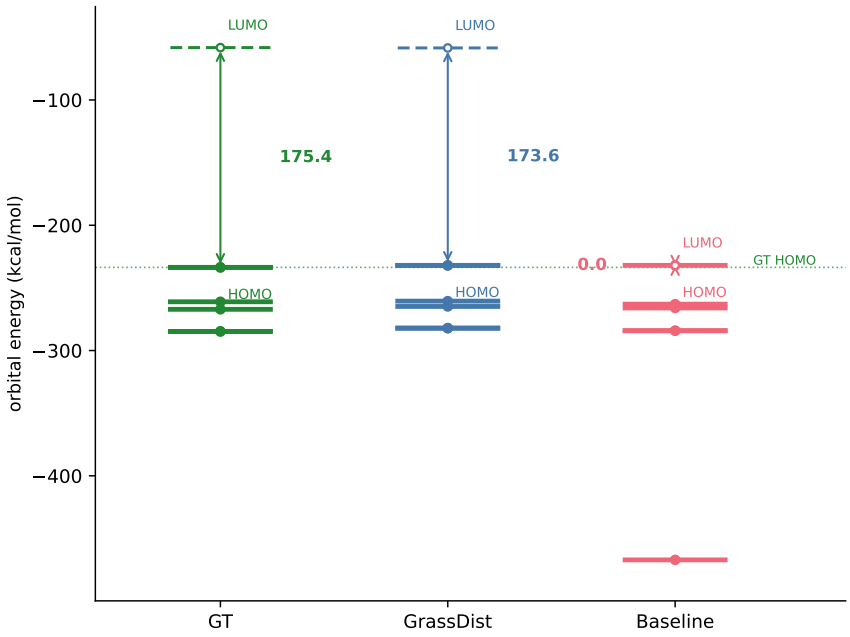


Figure 11: Frontier orbital levels of C_2N_2 for GT, GrassDist, and the Hamiltonian-only baseline. Solid filled bars are occupied levels and the dashed open bar is the LUMO; the arrow spans the HOMO–LUMO gap (in kcal/mol), and the dotted line marks the GT HOMO energy. The reference frontier is two-fold degenerate (HOMO and LUMO each form a degenerate pair).

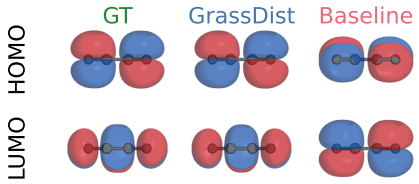


Figure 12: Frontier orbitals of C_2N_2: HOMO (top) and LUMO (bottom), for GT (left), GrassDist (middle), and the baseline (right).

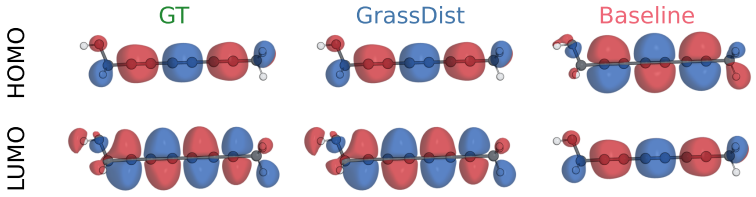


Figure 13: Frontier orbitals of C_8H_6O: HOMO (top) and LUMO (bottom), for GT (left), GrassDist (middle), and the baseline (right).

error on these configurations is 29–287 $\times 10^{-6}\,E_h$ (below the dataset mean for C_9H_{12}), yet the two worst give $d_{Gr} \approx 1.6 \times 10^3$, $\mathbf{P} \approx 3.0 \times 10^3$–$3.7 \times 10^4$ (in units of 10^{-3}), and $E_{tot} \approx 1321$–2214 kcal/mol. GrassDist reduces d_{Gr} to at most 14.5 and E_{tot} to 0.04–0.62 kcal/mol, at or near chemical accuracy, at the cost of a larger element-wise Hamiltonian error.

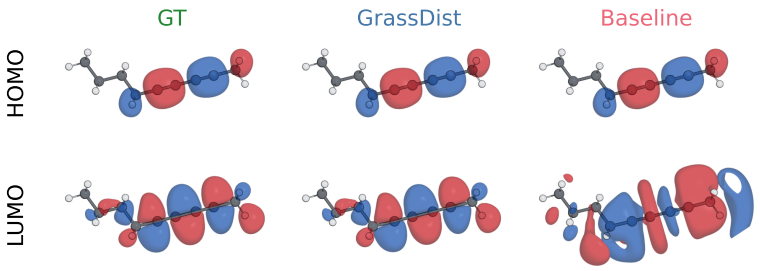


Figure 14: Frontier orbitals of C_9H_{12}: HOMO (top) and LUMO (bottom), for GT (left), GrassDist (middle), and the baseline (right).

C.8 Geodesic versus Chordal Subspace Loss: Empirical Verification

Appendix A.3 establishes that the chordal gradient vanishes as a principal angle approaches $\pi/2$ whereas the geodesic gradient grows linearly, which favors the squared geodesic loss $\mathcal{L}_{\mathrm{Gr}} = \sum_i \theta_i^2$ over the squared chordal loss (Conway et al., 1996) $\mathcal{L}_{\mathrm{ch}} = \sum_i \sin^2 \theta_i$. We verify this conclusion empirically. Two models are trained on `QH9-stable-iid` with identical backbones and optimization settings and differ only in the occupied-subspace term—the geodesic-Taylor metric used by GrassDist versus the chordal metric—with a Hamiltonian-to-subspace weight ratio of 1:10. To keep the comparison on electronically non-trivial systems and to avoid confounding by the initial-guess parameterization, both models regress the full Hamiltonian $\mathbf{H}^\star$ directly instead of a residual, and we evaluate on the large-molecule subset of the training split ($\geq$ 24 atoms, 2,449 configurations), recomputing all observables after each of the first ten epochs following Appendix B.1. All comparisons are paired at the configuration level.

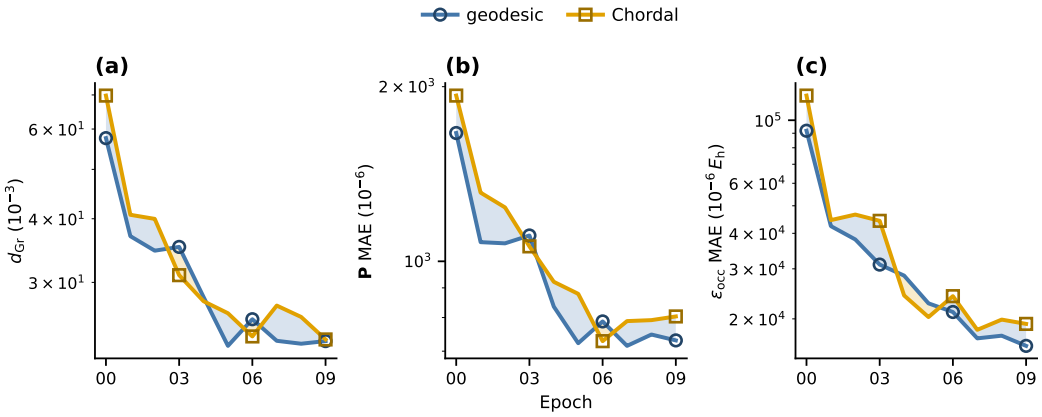


Figure 15: Per-epoch mean error of the geodesic and chordal subspace losses on the large-molecule `QH9-stable-iid` subset ($\geq$ 24 atoms, 2,449 configurations), for (a) d_{Gr} in 10^{-3}, (b) $\mathbf{P}$ MAE in 10^{-6}, and (c) $\varepsilon_{\mathrm{occ}}$ MAE in $10^{-6}\,E_{\mathrm{h}}$, following the conventions of Table 1. Shaded regions mark the gap between the two curves (blue where the geodesic error is lower).

Averaged over the ten epochs, the geodesic loss attains a lower error on 100% (2449/2449), 99.9% (2447/2449), and 100% (2449/2449) of the configurations for d_{Gr}, $\mathbf{P}$, and $\varepsilon_{\mathrm{occ}}$, with pooled error ratios (geodesic/chordal) of 0.92, 0.91, and 0.85 respectively (Figure 15–Figure 17). The advantage is thus consistent in sign and moderate but systematic in magnitude.

The chordal loss is not uniformly worse: on a small minority of epochs it is marginally better (for instance d_{Gr} at epochs 3, 4, and 6, and $\varepsilon_{\mathrm{occ}}$ at epochs 4 and 5; Figure 16). These reversals do not persist, so once the error is averaged over the training trajectory the geodesic loss wins on essentially every configuration (Figure 17). This is precisely the behavior expected when the chordal gradient vanishes for near-orthogonal subspaces: early in training, when the predicted subspace is

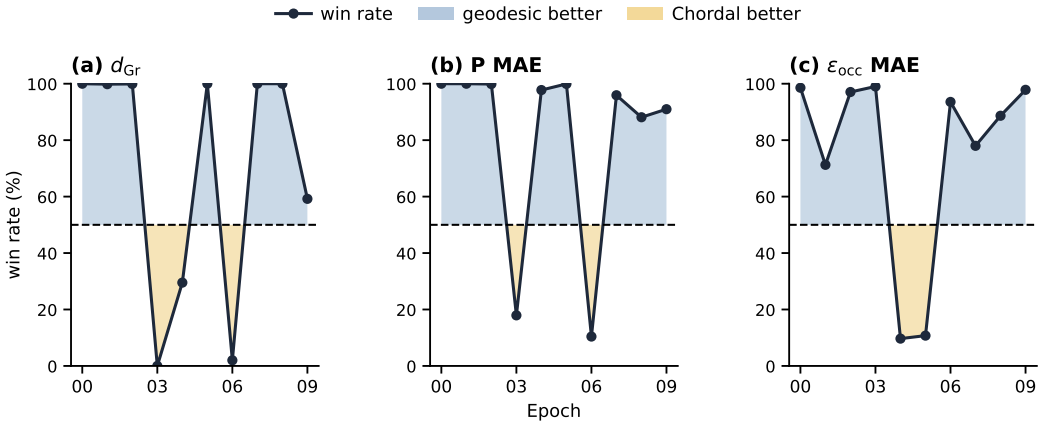


Figure 16: Per-epoch paired win rate of the geodesic versus the chordal subspace loss (fraction of configurations on which the geodesic error is lower), for (a) d_{Gr}, (b) $\mathbf{P}$, and (c) $\varepsilon_{\mathrm{occ}}$. The dashed line marks 50%; blue (orange) shading marks epochs at which the geodesic (chordal) loss wins on the majority of configurations.

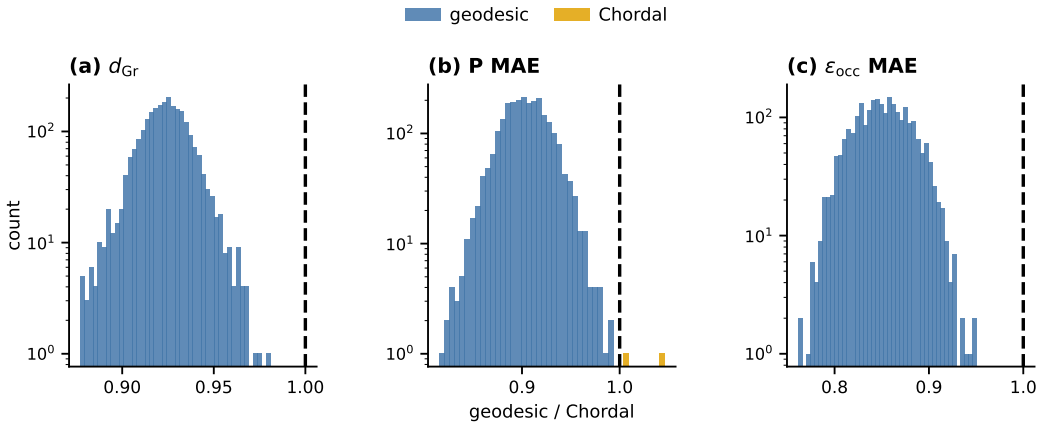


Figure 17: Distribution over the $2{,}449$ configurations of the per-configuration error ratio (geodesic/chordal) for (a) d_{Gr}, (b) $\mathbf{P}$, and (c) $\varepsilon_{\mathrm{occ}}$, where the per-configuration error is averaged over the first ten epochs. The dashed vertical line marks parity; mass to its left (blue) corresponds to a lower geodesic error.

still far from the reference, the geodesic loss supplies a stronger and more reliable signal whereas the chordal gradients saturate. Overall, the geodesic loss is consistently preferred to the chordal loss on this benchmark, which is why all main experiments of this work adopt the geodesic subspace loss.