WHERE THE ERROR LIVES: SPECTRAL REPRESENTATIONS OF LEARNED KOHN–SHAM HAMILTONIANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Kohn–Sham density functional theory (DFT) is the standard method for computing the electronic structure of molecules and materials, but its self-consistent-field (SCF) iterations make it expensive for large systems and high-throughput screening. Machine-learned Hamiltonian models offer a shortcut: an equivariant network predicts the Kohn–Sham Hamiltonian matrix from the atomic structure in one forward pass, and diagonalizing the prediction yields orbital energies, the HOMO–LUMO gap, molecular orbitals and the electron density, or an initial guess that accelerates DFT. These models are trained to match the reference matrix entry by entry, yet a small entry-wise error does not guarantee accurate eigenvalues and eigenvectors, which are what every downstream use requires. We represent the prediction error in the eigenbasis of the reference Hamiltonian, where it separates into sectors that each property, and each training signal, reads selectively. This representation explains why existing physics-informed losses help only partially: the wavefunction-alignment loss either weights all sectors uniformly or sees orbital energies only to first order, and the electron density is unchanged along about two thirds of the directions in which the prediction can be wrong, including every orbital-energy error. It also points to a representation of the Hamiltonian that sees every sector while staying in the language of the density: the *cumulative local density of states* (LDOS), the number of electrons at each point in space below each energy. We prove that it contains the density as a special case, responds to every component of the error, determines the Hamiltonian uniquely, and yields gradients that do not vanish for orbital energies far from their targets. Supervising a network through this representation requires no change to its architecture. Fine-tuning three Hamiltonian networks (QHNet, SPHNet and WANet) on QH9 in this way reduces the HOMO–LUMO gap error by 49–79%, 2.0–2.6 times more than the wavefunction-alignment loss, without increasing the element-wise error by more than 0.3%, and the gains carry over to out-of-distribution and molecular-dynamics splits (36–77%), where the wavefunction-alignment loss helps less and, on QHNet, makes the gap worse. These results show that how a predicted Hamiltonian is represented during training, not only the architecture that produces it, decides whether it is useful in practice.

1 INTRODUCTION

Kohn–Sham density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), recognized by the 1998 Nobel Prize in Chemistry (Kohn, 1999), is the most widely used method for computing the electronic structure of molecules and materials. It summarizes the electrons of a molecule in a single matrix, the Kohn–Sham Hamiltonian, whose generalized eigendecomposition gives the orbital energies, the energies of the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) and their gap, the molecular orbitals, and the electron density. These quantities determine chemical reactivity, optical and electronic behaviour and, through the total energy, molecular stability. Obtaining the Hamiltonian, however, requires self-consistent-field (SCF) iterations. Each iteration builds the Hamiltonian from the current electron density, formally $O(n^4)$

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.

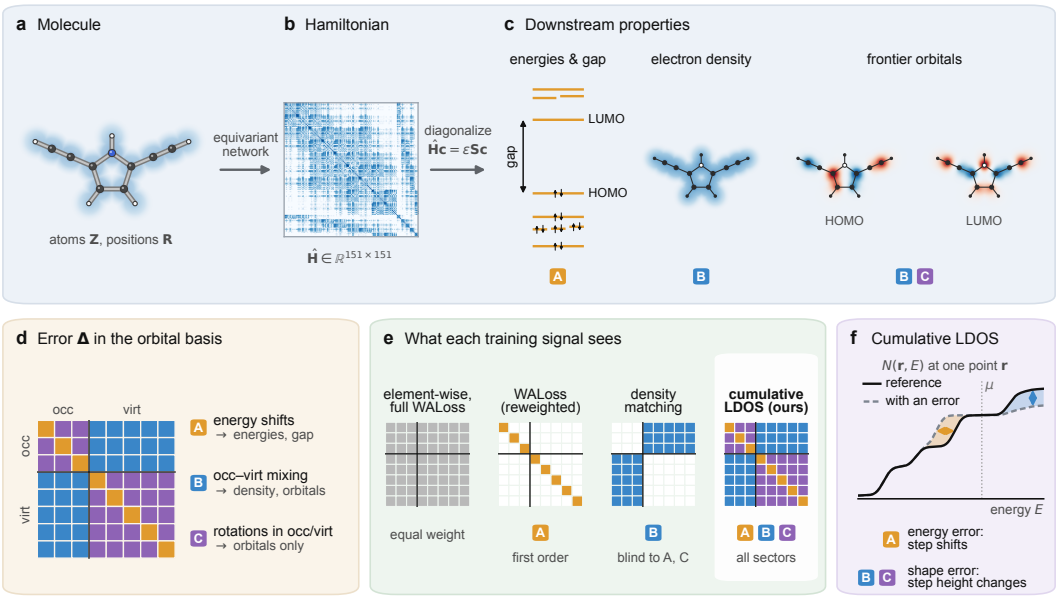


Figure 1: **Overview. (a–c)** A Hamiltonian network maps a molecule, QH9 #23470 (C_8H_5N), to its Kohn–Sham Hamiltonian $\hat{\mathbf{H}}$, whose eigendecomposition gives the orbital energies and HOMO–LUMO gap, the electron density and the molecular orbitals. All panels show reference DFT data. **(d)** In the reference orbital basis, the prediction error splits into three sectors: orbital-energy shifts (A), occupied–virtual mixing (B) and rotations within the occupied or the virtual space (C). The tags in (c) mark the sectors each property depends on. **(e)** The sectors each training signal sees. **(f)** At a point $\mathbf{r}$, the cumulative LDOS $N(\mathbf{r}, E)$ is a staircase with one step per orbital. An energy error shifts a step (A), and a shape error changes its height (B, C).

operations in the number of basis functions n for the hybrid functionals used in practice, and diagonalizes it at $O(n^3)$ cost; a calculation repeats this T times until the Hamiltonian and the density agree ($T \approx 9$ for the QH9 molecules studied here; Appendix A). This loop limits DFT on large systems and in high-throughput screening.

Machine-learned Hamiltonian models remove this bottleneck. An equivariant neural network predicts the Hamiltonian directly from the atomic numbers and positions (Schütt et al., 2019; Unke et al., 2021; Li et al., 2022; Yu et al., 2023b; Li et al., 2025; Luo et al., 2025), replacing all T iterations by one forward pass. The prediction is used in two ways: it is diagonalized once to obtain the electronic properties above, or it initializes a conventional DFT calculation so that fewer iterations are needed (Yu et al., 2023a; Li et al., 2025). In the first use, no SCF iteration is left to correct the prediction: the eigenpairs of the predicted matrix are the result. Progress has been rapid: on the QH9 benchmark (Yu et al., 2023a), the entry-wise error of the predicted matrix has more than halved since the benchmark was introduced (Table 1).

The entry-wise error, however, does not translate directly into accurate eigenvalues and eigenvectors, which are what the predicted Hamiltonian is used for. For example, continuing to train QHNet with its own element-wise loss lowers its matrix error but raises its HOMO–LUMO gap error, consistently on all four QH9 splits and by 53% on the standard split (Appendix G.2). Li et al. (2025) observed the same decoupling from the other side: on QH9, their alignment loss increased the matrix error while improving the orbital energies, and on larger molecules, matrices with tiny relative errors produced unusable total energies. The reason is that these models are trained with an *element-wise* loss, the mean absolute or squared error over matrix entries, which treats every entry alike, whereas eigenvalues and eigenvectors respond very unevenly to an error. Classical matrix perturbation theory makes this precise (Kato, 1995; Davis & Kahan, 1970; Stewart & Sun, 1990): to first order, an eigenvalue moves only with the corresponding diagonal entry of the error expressed in the eigenbasis, and the eigenvectors rotate by the off-diagonal entries divided by eigenvalue differences. On QH9, an error that couples the HOMO and the LUMO, about 0.25 Ha apart, therefore changes the LUMO about 75 times more than an equally large error that couples it to a core orbital some 19 Ha lower.

To make this precise, we represent the prediction error in the eigenbasis of the reference Hamiltonian, where it separates into three interpretable blocks, which we call *sectors* (Figure 1d): diagonal entries, which shift orbital energies; entries corresponding to an occupied and an unoccupied orbital, which mix the two and change the electron density; and entries corresponding to two occupied or two unoccupied orbitals, which only rotate orbitals within those spaces. This view explains why existing physics-informed losses help only partly. The wavefunction-alignment loss (WALoss) of Li et al. (2025) evaluates the predicted Hamiltonian in the reference orbital basis. In its full form it weights every entry of the error uniformly, like the element-wise loss, and therefore does not concentrate on the errors that matter; in the diagonal-only form proposed by Li et al. (2025), it penalizes only the orbital energies, and only to first order. A natural alternative is to diagonalize the prediction and match its electron density, which is exact and automatically focuses on the frontier orbitals, those closest to the HOMO–LUMO gap. We prove, however, that energy shifts and rotations within the occupied or the unoccupied space leave the density exactly unchanged, so a loss that depends only on the density does not respond to them at all; we say that it is *blind* to them. These errors span about 69% of the degrees of freedom of the error on QH9 and include every orbital-energy error.

The same representation tells us what a complete training signal must do: see every sector, weight each by its effect on the properties of interest, and use the exact eigenpairs of the prediction. We find such a signal in a physical observable, the *local density of states* (LDOS), which resolves the electron density by energy and is what a scanning tunnelling microscope measures (Tersoff & Hamann, 1985). Its cumulative form, $N(\mathbf{r}, E)$, the number of electrons at point $\mathbf{r}$ with energy below E, represents the Hamiltonian in the language of the density without discarding the orbital energies (Figure 1f). We prove four properties. The density is its slice at the Fermi level, the energy midway between the HOMO and the LUMO. It responds to every sector. It determines the Hamiltonian uniquely. And unlike a peak-shaped LDOS, it keeps pulling an orbital energy toward its target however far inside the energy window it starts. We turn N into a training loss, the *cumulative LDOS loss* $\mathcal{L}_{\mathrm{cum}}$: the mean squared difference between N computed from the predicted and from the reference Hamiltonian, over sampled points $\mathbf{r}$ and energies E. A network is fine-tuned on $\mathcal{L}_{\mathrm{elem}} + w\,\mathcal{L}_{\mathrm{cum}}$, where $\mathcal{L}_{\mathrm{elem}}$ is its usual element-wise loss and w a weight, with no change to its architecture.

Across three architectures and four QH9 splits, this loss improves every eigen-derived metric in all twelve settings and is the best method in each (Tables 1 and 6). On the standard split, it lowers the HOMO–LUMO gap error by 49–79%, 2.0–2.6 times more than WALoss, while the element-wise error rises by at most 0.3% and on QHNet even falls by 3.5%. Under distribution shift, it still lowers the gap error by 36–77%, whereas WALoss makes it worse in three of the nine shifted settings, all on QHNet.

Our contributions are as follows.

- **Problem.** We show that the mismatch between how Hamiltonian networks are trained, entry by entry, and how they are used, through eigenpairs, is a question of *where* the error lies. In the reference orbital basis, the error separates into sectors that each property and each training signal read selectively, and the electron density, the most natural physical target, is blind to every orbital-energy error (Sec. 3.1).
- **Solution.** We identify the cumulative LDOS as a representation of the Hamiltonian that contains the density, sees every sector and determines the Hamiltonian uniquely, and we turn it into a training loss that any Hamiltonian network can use unchanged (Secs. 3.2 and 3.3).
- **Results and implications.** The loss improves every eigen-derived metric in all twelve backbone–split settings without raising the element-wise error by more than 0.3% (Sec. 4). How a predicted Hamiltonian is represented during training, and not only the architecture that produces it, decides whether it is useful.

2 Background

Notation. We follow standard quantum-chemistry notation (Szabo & Ostlund, 1996); Appendix A explains the terms we use and illustrates the SCF procedure. Bold upright symbols denote matrices and vectors. A star marks quantities of the reference DFT calculation and a hat quantities computed from a network's prediction; unadorned symbols (ε, $\mathbf{C}$, ψ, ρ, N) denote the same maps applied to a generic Hamiltonian. A hat thus means *predicted*, as in statistics.

Kohn–Sham Hamiltonians. Consider a molecule of N_{at} atoms with atomic numbers $\mathbf{Z} \in \mathbb{Z}_+^{N_{\mathrm{at}}}$, which identify the element of each atom, and positions $\mathbf{R} \in \mathbb{R}^{3N_{\mathrm{at}}}$. Its electrons are described in a basis of n real-valued, atom-centred basis functions $\{\phi_\mu : \mathbb{R}^3 \to \mathbb{R}\}_{\mu=1}^n$, also called atomic orbitals; for the def2-SVP basis used here, they are Gaussian-type functions. The basis functions are linearly independent but not orthogonal. Their inner products form the overlap matrix $\mathbf{S}$, with entries $S_{\mu\nu} = \int \phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r})\, d\mathbf{r}$; it depends only on the geometry and is positive definite, written $\mathbf{S} \succ 0$. Kohn–Sham DFT (Hohenberg & Kohn, 1964; Kohn & Sham, 1965) represents the electronic structure by a second symmetric $n \times n$ matrix, the Hamiltonian $\mathbf{H}^\star$, with entries $H^\star_{\mu\nu} = \langle\phi_\mu|\mathcal{F}[\rho^\star]|\phi_\nu\rangle$, where $\mathcal{F}[\rho]$ is the Kohn–Sham operator built from an electron density ρ and $\rho^\star$ is the converged, self-consistent density. The molecular orbitals of the reference are linear combinations of the basis functions, $\psi_i(\mathbf{r}) = \sum_\mu C^\star_{\mu i}\phi_\mu(\mathbf{r})$ (we omit their star), whose coefficients and energies solve the generalized eigenvalue problem

$$\mathbf{H}^\star\mathbf{C}^\star = \mathbf{S}\mathbf{C}^\star\,\mathrm{diag}(\boldsymbol{\varepsilon}^\star), \qquad \mathbf{C}^{\star\top}\mathbf{S}\mathbf{C}^\star = \mathbf{I}. \tag{1}$$

The second condition states that, unlike the basis functions, the molecular orbitals are orthonormal, $\int \psi_i(\mathbf{r})\psi_j(\mathbf{r})\, d\mathbf{r} = \delta_{ij}$. Column $\mathbf{c}_i^\star$ of $\mathbf{C}^\star$ holds the coefficients of orbital ψ_i, and $\varepsilon_i^\star$ is its energy. A closed-shell molecule with N_e electrons occupies the $n_o = N_e/2$ lowest orbitals (the set $\mathcal{O}$); the remaining $n_v = n - n_o$ orbitals are unoccupied, or virtual (the set $\mathcal{V}$). The HOMO and LUMO energies are $\varepsilon^\star_{n_o}$ and $\varepsilon^\star_{n_o+1}$, and the electron density is

$$\rho^\star(\mathbf{r}) = 2\sum_{i\in\mathcal{O}} |\psi_i(\mathbf{r})|^2. \tag{2}$$

Because $\mathcal{F}[\rho]$ depends on the density, DFT finds $\mathbf{H}^\star$ by SCF iteration, alternating between building the Hamiltonian from the current density and solving equation 1; these T iterations dominate the cost of the calculation (Figure 3 in Appendix A).

Learning Hamiltonians. A Hamiltonian network predicts $\hat{\mathbf{H}} = f_\theta(\mathbf{Z}, \mathbf{R})$ and is built to transform correctly under rotations of the molecule; $\mathbf{S}$ is computed exactly from the geometry. Existing models are trained with the element-wise loss $\mathcal{L}_{\mathrm{elem}} = \|\hat{\mathbf{H}} - \mathbf{H}^\star\|$, the mean absolute or squared error over matrix entries (Yu et al., 2023b; Li et al., 2025; Luo et al., 2025). At inference, $\hat{\mathbf{H}}$ replaces $\mathbf{H}^\star$ in equation 1, and every downstream quantity follows from its eigenpairs $(\hat{\varepsilon}, \hat{\mathbf{C}})$. Following QH9 (Yu et al., 2023a), models are evaluated by the element-wise error of $\hat{\mathbf{H}}$ together with the errors of the orbital energies and the similarity of the orbital coefficients (Sec. 4.1).

Physics-informed losses. The wavefunction-alignment loss (WALoss) of Li et al. (2025) adds a spectral signal without diagonalizing the prediction. It precomputes the reference orbitals $\mathbf{C}^\star$ and evaluates the prediction in their basis: since $\mathbf{C}^\star$ diagonalizes $\mathbf{H}^\star$, the matrix $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star$ should equal $\mathrm{diag}(\boldsymbol{\varepsilon}^\star)$. The full form penalizes every entry of the deviation alike, $\mathcal{L}_{\mathrm{WA}} = \ell\big(\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star - \mathrm{diag}(\boldsymbol{\varepsilon}^\star)\big)$, where ℓ is the mean absolute or squared error over the entries (Appendix G.3); we use this form for all backbones. Li et al. (2025) also propose a diagonal-only form that keeps only the diagonal, $\sum_i w_i\big(\mathbf{c}_i^{\star\top}\hat{\mathbf{H}}\,\mathbf{c}_i^\star - \varepsilon_i^\star\big)^2$, with large weights on the occupied orbitals and the LUMO. A second option is to diagonalize $\hat{\mathbf{H}}$, form the predicted density $\rho[\hat{\mathbf{H}}]$ by equation 2, and penalize its deviation from $\rho^\star$ at a set $\mathcal{R}$ of sample points,

$$\mathcal{L}_\rho(\hat{\mathbf{H}}) = \Big\langle \Big(\frac{\rho[\hat{\mathbf{H}}](\mathbf{r}) - \rho^\star(\mathbf{r})}{\mathrm{div}(\mathbf{r})}\Big)^2 \Big\rangle_{\mathbf{r}\in\mathcal{R}}, \tag{3}$$

where $\mathrm{div}(\mathbf{r}) \geq 1$ is a geometry-only factor that down-weights the region near the nuclei, where the density is orders of magnitude larger (Ushenin et al., 2025). We call equation 3 density matching; Sec. 3.2 shows that it is a special case of our loss.

3 Spectral Representations of Hamiltonian Errors

We first represent the prediction error in the basis of reference orbitals and determine which part of it each eigen-property depends on (Sec. 3.1). This identifies what a complete training signal must see and leads to the cumulative LDOS (Sec. 3.2), whose guarantees we establish in Sec. 3.3.

3.1 WHERE THE PREDICTION ERROR LIVES

Write the prediction as $\hat{\mathbf{H}} = \mathbf{H}^\star + \delta\mathbf{H}$ and express the error in the basis of reference orbitals,

$$\mathbf{\Delta} := \mathbf{C}^{\star\top}\delta\mathbf{H}\,\mathbf{C}^\star, \qquad \text{so that} \qquad \mathbf{C}^{\star\top}\hat{\mathbf{H}}\,\mathbf{C}^\star = \mathrm{diag}(\boldsymbol{\varepsilon}^\star) + \mathbf{\Delta}. \tag{4}$$

Entry Δ_{ij} measures how strongly the error couples reference orbitals i and j. We group the entries of the symmetric matrix $\mathbf{\Delta}$ into three *sectors* (Figure 1d): (A) the diagonal entries $d_i := \Delta_{ii}$; (B) the occupied–virtual block $\mathbf{\Delta}_{\mathcal{OV}}$; and (C) the off-diagonal parts of the occupied–occupied and virtual–virtual blocks, $\mathbf{\Delta}_{\mathcal{OO}}$ and $\mathbf{\Delta}_{\mathcal{VV}}$. The full WALoss averages the entries of $\mathbf{\Delta}$ without weights (Lemma B.1), so it treats all sectors alike. Classical matrix perturbation theory (Kato, 1995; Stewart & Sun, 1990) shows what each sector does.

Proposition 3.1 (What each property depends on)**.** *Let the reference orbital energies be non-degenerate. To first order in* $\mathbf{\Delta}$*:*
(a) orbital energies depend only on the diagonal: $\hat{\varepsilon}_i - \varepsilon_i^\star = d_i$*;*
(b) orbital shapes depend on the off-diagonal entries, weighted by inverse energy spacings: $\hat{\mathbf{c}}_i - \mathbf{c}_i^\star = \sum_{j\neq i} \frac{\Delta_{ji}}{\varepsilon_i^\star - \varepsilon_j^\star}\,\mathbf{c}_j^\star$*;*
(c) the density depends only on the occupied–virtual block, with the largest weights on orbital pairs near the HOMO–LUMO gap:

$$\delta\rho(\mathbf{r}) = 4 \sum_{i\in\mathcal{O},\, a\in\mathcal{V}} \frac{\Delta_{ai}\,\psi_i(\mathbf{r})\psi_a(\mathbf{r})}{\varepsilon_i^\star - \varepsilon_a^\star}. \tag{5}$$

Appendix B gives the proof and the second-order corrections. Item (c) holds because the occupied–occupied contributions cancel in pairs: rotating occupied orbitals among themselves leaves their total density unchanged. This cancellation is exact, not only first-order. Items (a) and (b) need non-degenerate orbital energies, but (c) needs only a nonzero HOMO–LUMO gap, because the density matrix is the spectral projector onto the occupied subspace, which depends analytically on the Hamiltonian while the gap is open (Kato, 1995). This matters on QH9, where 12% of the molecules have two orbital energies closer than 0.1 mHa. Remark B.3 in Appendix B relates the sectors to the reported metrics.

Proposition 3.2 (The density is blind to sectors A and C)**.** *If* $\mathbf{\Delta}_{\mathcal{OV}} = 0$ *and no eigenvalue of* $\hat{\mathbf{H}}$ *crosses the HOMO–LUMO gap, then* $\rho[\hat{\mathbf{H}}] = \rho^\star$ *exactly. Any loss that depends on* $\hat{\mathbf{H}}$ *only through its density therefore vanishes on a subspace of dimension* $\frac{n_o(n_o+1)}{2} + \frac{n_v(n_v+1)}{2}$ *out of* $\frac{n(n+1)}{2}$*.*

On a uniform sample of 4,000 QH9-stable molecules, with median $n = 171$ and $n_o = 33$, this blind subspace holds a median of 69.0% of the degrees of freedom of $\mathbf{\Delta}$ and 66–71% between the 5th and 95th percentiles (Table 5). It contains every orbital-energy error, including the HOMO–LUMO gap error. On a real molecule, shifting the HOMO or LUMO energy by 2 mHa changes the density loss by less than 10^{-30}, whereas rotating the HOMO and LUMO into each other changes it by 3×10^{-7} (Figure 11a in Appendix H).

Remark 3.3 (Why density matching still improves the gap). Proposition 3.2 is a statement about directions, not about what a trained network does. A network produces all sectors from the same atomic-orbital blocks, so correcting the frontier couplings of sector B, which the density weights by $1/(\varepsilon_a^\star - \varepsilon_i^\star)$, also changes the frontier orbital energies. In practice it improves them, and density matching alone lowers the gap error substantially (Sec. 4.4). The density loss is, however, exactly flat along any change of the predicted orbital energies that leaves the predicted orbitals fixed and moves no level across the gap, so it cannot correct the energy error that remains. That error lies in sector A, which is where energy resolution adds most (Sec. 4.4).

What existing losses see. Figure 1e applies these results to the losses of Sec. 2, and Table 3 in Appendix B lists how each loss weights the sectors. The element-wise loss and the full WALoss penalize every sector, but uniformly: a small coupling between two frontier orbitals, which shifts the gap at second order with weight $1/\text{gap}$ (Appendix B), costs as much as an equally large coupling between two deep core orbitals, which barely affects the properties of interest. The diagonal-only WALoss responds only to the diagonal and only to first order. Because off-diagonal entries also shift the orbital energies at second order, it can be lowered while the true orbital energies stay wrong, and increasing its weight indeed makes the gap worse (Sec. 4.4). Density matching uses the exact eigenpairs and focuses on the frontier, but is blind to the sectors of Proposition 3.2.

Desiderata. A loss that reliably improves eigen-properties should therefore (i) respond to every sector, so that no error goes unpenalized; (ii) weight the sectors by their effect on the quantities of interest, so that model capacity is spent where it matters; and (iii) use the exact eigenpairs of $\hat{\mathbf{H}}$ rather than a first-order proxy. No existing loss meets all three requirements. We now construct one that does.

3.2 From the density to the cumulative LDOS

The density is blind to orbital energies because it sums the occupied orbitals and discards their energies, as equation 2 shows. Keeping the energy of each orbital gives the local density of states, $\rho(\mathbf{r}, E) = \sum_i |\psi_i(\mathbf{r})|^2 \, \delta(E - \varepsilon_i)$, which describes how the electrons at point $\mathbf{r}$ are distributed in energy and is measured directly by scanning tunnelling microscopy (Tersoff & Hamann, 1985). We use its cumulative form, with each peak smoothed by a Gaussian of width σ, so that each step becomes an error-function step:

$$N(\mathbf{r}, E) = \sum_{i=1}^{n} |\psi_i(\mathbf{r})|^2 \, \Phi\Big(\frac{E - \varepsilon_i}{\sigma}\Big), \tag{6}$$

where Φ is the standard normal cumulative distribution function and σ is a smoothing width. $N(\mathbf{r}, E)$ is the number of electrons per spin at $\mathbf{r}$ with energy below E. As a function of E, it is a smoothed staircase with one step per orbital: the step sits at the orbital energy ε_i, and its height is the orbital density $|\psi_i(\mathbf{r})|^2$ at that point (Figure 8b in the appendix). An error in an orbital energy moves a step and an error in an orbital shape changes its height, so both kinds of error are visible.

Definition 3.4 (Cumulative LDOS loss)**.** Let $\mathcal{R}$ be a set of sample points, $\{E_k\}_{k=1}^{K}$ an energy grid on the window $[\mu - E_\downarrow, \mu + E_\uparrow]$ around the reference Fermi level $\mu = \frac{1}{2}(\varepsilon^\star_{n_o} + \varepsilon^\star_{n_o+1})$, and div the factor in equation 3. The cumulative LDOS loss is

$$\mathcal{L}_{\mathrm{cum}}(\hat{\mathbf{H}}) = \Big\langle \Big(\frac{N[\hat{\mathbf{H}}](\mathbf{r}, E_k) - N[\mathbf{H}^\star](\mathbf{r}, E_k)}{\mathrm{div}(\mathbf{r})} \Big)^2 \Big\rangle_{\mathbf{r} \in \mathcal{R},\, k}, \tag{7}$$

where $N[\hat{\mathbf{H}}]$ is computed from the exact eigenpairs of $(\hat{\mathbf{H}}, \mathbf{S})$ and gradients are propagated through the eigensolver.

We train with $\mathcal{L}_{\mathrm{elem}} + w\,\mathcal{L}_{\mathrm{cum}}$, where the weight $w > 0$ balances the scales of the two terms and is selected on the validation split (Table 15); on QHNet, the results change little over a tenfold range of w (Sec. 4.4). Two variants serve as ablations: replacing the step Φ by a Gaussian peak $e^{-x^2/2}$ gives a *peak-shaped* LDOS loss $\mathcal{L}_{\mathrm{LDOS}}$, and a single grid point at $E_1 = \mu$ recovers density matching (Theorem 3.5(i)).

3.3 Theoretical properties

Theorem 3.5 (Properties of the cumulative LDOS)**.** *Let the reference orbital energies be non-degenerate, as in Proposition 3.1, let $g = \varepsilon^\star_{n_o+1} - \varepsilon^\star_{n_o}$ be the reference HOMO–LUMO gap, and let I be any open energy interval, for example the loss window. Then:*
(i) Density matching is a special case. *If no orbital energy lies within $g/2$ of μ, then $N(\mathbf{r}, \mu) = \rho(\mathbf{r})/2$ up to $\frac{1}{2}\,\mathrm{erfc}\,\big(g/(2\sqrt{2}\sigma)\big)$ per orbital, so $\mathcal{L}_\rho$ is the $E = \mu$ slice of $\mathcal{L}_{\mathrm{cum}}$.*
(ii) No blind sector. *The first-order response of N on $\mathbb{R}^3 \times I$ to $\boldsymbol{\Delta}$ is injective: every energy error d_i (sector A) and every off-diagonal entry Δ_{ij} (sectors B and C) changes N.*
(iii) Identifiability. *$N[\hat{\mathbf{H}}] = N[\mathbf{H}^\star]$ on $\mathbb{R}^3 \times I$ if and only if $\hat{\mathbf{H}} = \mathbf{H}^\star$.*
(iv) Non-saturating gradient. *If one orbital energy is displaced by δ with its shape fixed, the gradient of the cumulative mismatch with respect to δ is proportional to $\mathrm{erf}(\delta/2\sigma)$, which tends to $\mathrm{sign}\,\delta$ as $|\delta|$ grows, whereas that of the peak-shaped mismatch is proportional to $(\delta/\sigma)\, e^{-\delta^2/4\sigma^2}$, which vanishes.*

Appendix B proves each part (Propositions B.4 and B.6, Theorem B.5) for any linearly independent basis of Gaussian-type orbitals, and Appendix H checks every part numerically on QH9 molecules.

Implications. By (i), which holds at every energy inside the gap and hence at the two grid energies next to μ, the cumulative loss keeps everything that density matching sees, including its automatic focus on the frontier orbitals (Proposition 3.1c); the other energies add the orbital-energy information that the density lacks. By (ii) and (iii), all three sectors are visible, and on the loss window N determines the Hamiltonian, so it cannot be matched by trading one kind of error for another, as the diagonal-only WALoss can. Pairs of orbitals that both lie far outside the window, such as two core orbitals or two high virtual orbitals, enter only through the Gaussian tails of their energy profiles; they barely affect ground-state properties and remain supervised by the element-wise term. The loss therefore concentrates on the valence and low virtual orbitals, the part of the spectrum that governs chemistry. Together, these properties meet the three desiderata of Sec. 3.1, and (iv) makes training robust: unlike a peak-shaped loss, the cumulative loss keeps pulling an orbital energy toward its target once the error exceeds a few σ (Proposition B.6 in Appendix B).

Finally, because most degrees of freedom of $\mathbf{\Delta}$ lie in sectors that barely affect the target properties, a model can lower $\mathcal{L}_{\mathrm{cum}}$ by moving error out of the sensitive sectors rather than by enlarging it, so the element-wise error need rise only slightly (Proposition B.7 in Appendix B).

4 Experiments

4.1 Setup

Datasets. QH9 (Yu et al., 2023a) provides B3LYP/def2-SVP Hamiltonians for QM9 molecules (Ramakrishnan et al., 2014) with up to nine heavy atoms. We use its four official splits: *stable-id* is a random split, *stable-ood* tests on larger molecules, *dynamic-geo* on unseen geometries of training molecules and *dynamic-mol* on unseen molecules. Appendix C gives sizes and statistics.

Backbones. We fine-tune three pretrained Hamiltonian networks that span the development of the field: QHNet (Yu et al., 2023b), the original QH9 baseline, and WANet (Li et al., 2025) and SPHNet (Luo et al., 2025), which make equivariant message passing cheaper through SO(2) convolutions and adaptive sparsity.

Baselines and protocol. We compare with WALoss (Li et al., 2025), the state-of-the-art physics-informed loss for Hamiltonian prediction, in its full form. For each backbone, every method fine-tunes the same pretrained checkpoint for 30k steps with the same optimizer, learning-rate schedule, batch size and data. Only the auxiliary loss and its weight differ, and all methods are evaluated with the same code. Appendix J lists all hyper-parameters.

Training objective and cost. We apply $\mathcal{L}_{\mathrm{cum}}$ when fine-tuning a pretrained model. At each step it is evaluated on eight molecules of the batch, at 256 sample points per molecule drawn once from the PySCF quadrature grid (Sun et al., 2018), whose atomic-orbital values are cached, and on $K = 64$ energies on $[\mu - 1.5, \mu + 0.3]$ Ha with $\sigma = 0.02$ Ha, covering the valence occupied orbitals, the LUMO and the low virtual orbitals. The eigenpairs of $(\hat{\mathbf{H}}, \mathbf{S})$ are computed in double precision, with gradients through the eigensolver (Appendix F). On QHNet the loss adds about 5% to the cost of pretraining, and it costs nothing at inference (Appendix G.6).

Metrics. Following QH9, we report the MAE of $\hat{\mathbf{H}}$ over all entries (H), over same-atom blocks (diag) and over the remaining blocks (non-diag). We also report the MAE of the HOMO and LUMO energies, of the gap and of all occupied orbital energies ($\varepsilon_{\mathcal{O}}$), and the cosine similarity ψ of the occupied orbital coefficients. Energies are in μHa and ψ in % (Appendix E).

4.2 Main results

Table 1 reports results on the QH9-stable-id test set. On QHNet, SPHNet and WANet, the cumulative LDOS loss gives the best orbital energies and orbital coefficients. It reduces the HOMO–LUMO gap error by 79% on QHNet, 49% on SPHNet and 53% on WANet, and the LUMO error by 56–84%; its gap error is 1.6–2.9 times lower than that of WALoss. These gains come without a cost in element-wise accuracy: the Hamiltonian MAE stays within 0.3% of the pretrained value on SPHNet and WANet, and on QHNet it falls by 3.5%, making our loss the best in every downstream column.

Table 1: Results on the QH9-stable-id test set (13,084 molecules) after 30k fine-tuning steps. Energies in μHa, ψ in %. For each backbone, the gray rows are the two baselines, the pretrained model and fine-tuning with WALoss (Li et al., 2025). The violet row, *density only*, is the ablation at weight $w = 100$ that keeps only the $E = \mu$ slice of our loss, which is density matching (Proposition B.4). The blue row is our cumulative LDOS loss, which adds the energy resolution. All fine-tuning runs share every other setting, and the peak-shaped variant is in Appendix G.4. Bold marks the best downstream metric per backbone.

backbone	role	method	H	diag	non-diag	HOMO	LUMO	gap	$\varepsilon_{\mathcal{O}}$	ψ
QHNet	baseline	pretrained	77.56	129.25	73.68	1041	9445	9632	1206	96.28
		+ WALoss	78.04	119.03	74.95	847	5458	5835	570	96.79
	ablation	+ density only (w=100)	76.82	118.70	73.65	719	1575	2081	356	97.87
	ours	+ cumulative LDOS (ours)	74.81	109.54	72.19	**678**	**1482**	**2027**	**341**	**97.91**
SPHNet	baseline	pretrained	36.60	44.79	35.38	345	1309	1554	380	99.06
		+ WALoss	37.28	46.03	35.98	341	957	1259	288	99.17
	ablation	+ density only (w=100)	37.63	47.14	36.23	363	684	970	143	99.08
	ours	+ cumulative LDOS (ours)	36.64	44.83	35.42	**292**	**572**	**789**	**137**	**99.25**
WANet	baseline	pretrained	40.07	51.28	38.42	335	1454	1684	288	98.98
		+ WALoss	40.99	53.26	39.19	328	1025	1306	157	99.13
	ablation	+ density only (w=100)	40.21	51.63	38.53	324	823	1101	140	**99.22**
	ours	+ cumulative LDOS (ours)	40.19	51.59	38.52	**287**	**573**	**791**	**137**	**99.22**

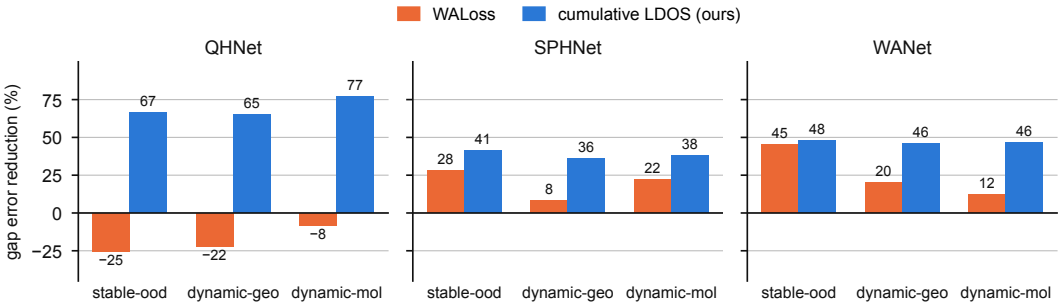


Figure 2: Reduction of the HOMO–LUMO gap error relative to each pretrained model on the three shifted QH9 splits (higher is better; negative values mean that the error grew). The cumulative LDOS loss lowers the gap error on every backbone and split, and it also improves every other downstream metric; Table 6 in Appendix G.1 lists all metrics.

WALoss improves the gap by 19–39% but raises the Hamiltonian MAE on all three backbones (by 0.6–2.3%), consistent with the trade-off that Li et al. (2025) reported on QH9.

This pattern follows from the analysis of Sec. 3.1. WALoss weights all entries of $\mathbf{\Delta}$ equally, so much of its effect goes to the large sectors that barely influence the spectrum, and its spectral gains come at the expense of element-wise accuracy. The cumulative LDOS loss concentrates on the sectors that control the frontier spectrum, and the model can satisfy it by moving error out of these sectors at little cost in total error (Proposition B.7 in Appendix B).

4.3 Generalization across splits

Figure 2 applies the same protocol to the three shifted splits; Table 6 in Appendix G.1 lists all metrics. The gains persist under distribution shift. Across all four splits, the cumulative LDOS loss lowers the gap error of every model, by 36–77% on the shifted splits, and never raises the element-wise error by more than 0.3%, whereas WALoss moves both errors in either direction and increases the gap error on all three shifted splits of QHNet. On dynamic-mol, where the pretrained QHNet is least accurate, our loss lowers its gap error by 77% and raises its orbital similarity from 79.7% to 90.6%. Predicted Hamiltonians are also used to initialize DFT; starting SCF from them saves about a third of the iterations relative to the standard `minao` guess (Appendix G.5).

4.4 Ablation studies

Energy resolution. The ablation rows of Table 1 isolate the first design choice of the loss, the energy resolution; the second, step versus peak, is in Appendix G.4. Density matching alone already lowers the gap error by 35–78%, which confirms the value of its automatic focus on the frontier (Proposition 3.1c). Adding energy resolution helps most where the density is blind to the remaining error: on WANet the gap error falls from 1101 to 791 μHa, and on SPHNet from 970 to 789 μHa. On QHNet, where density matching already captures most of the gap improvement, energy resolution removes its element-wise cost (H from 76.82 to 74.81 μHa) and further lowers the LUMO and occupied-orbital errors.

Weight sensitivity. On QHNet, the gap error of the cumulative loss stays between 1811 and 2231 μHa as w varies tenfold, whereas the diagonal-only form of WALoss, which sees orbital energies only to first order, degrades as its weight grows (Appendix G.3).

5 Related Work

Hamiltonian prediction. SchNOrb (Schütt et al., 2019) and PhiSNet (Unke et al., 2021) predict molecular Hamiltonians and wavefunctions, and DeepH (Li et al., 2022) targets materials. QHNet (Yu et al., 2023b) introduced an efficient equivariant architecture, and QH9 (Yu et al., 2023a) established the benchmark used here. WANet (Li et al., 2025) and SPHNet (Luo et al., 2025) improve accuracy and scalability through SO(2) convolutions and adaptive sparsity. All of them are trained with element-wise losses; our loss applies to each of them without modification.

Physics-informed and spectral objectives. Gu et al. (2022) regress eigenvalues directly, which requires differentiating through an eigensolver at every step, supervises the energies but not the orbitals, and was found hard to optimize on large systems (Li et al., 2025). WALoss (Li et al., 2025) avoids the eigensolver by evaluating $\hat{\mathbf{H}}$ in the reference orbital basis. Lemma B.1 shows that its full form is the element-wise error measured in an orthonormal basis, and its perturbation theorem bounds eigenvalue errors by that error. Our analysis refines this bound into a map of which entries matter for which property, and explains the trade-off between matrix error and properties that Li et al. (2025) observed on QH9. Self-consistency training (Zhang et al., 2024) rebuilds a Hamiltonian from the predicted density through the exchange–correlation functional and matches it to the prediction on unlabeled molecules; it enforces the SCF fixed point rather than the spectrum and is complementary to our loss.

Electron density and LDOS in machine learning. DeepDFT (Jørgensen & Bhowmik, 2022), ChargE3Net (Koker et al., 2024) and LAGNet (Ushenin et al., 2025) predict the electron density as an output; LAGNet introduced the sampling grid and core suppression that we use. Ellis et al. (2021) predict the LDOS itself as the target of a finite-temperature DFT surrogate for materials. We use the LDOS as a training signal for a predicted Hamiltonian and show why its cumulative form is the one to optimize.

6 Conclusion

The element-wise error of a predicted Kohn–Sham Hamiltonian is a poor guide to its eigenproperties, because each property depends on a specific part of the error. Representing the error in the reference orbital basis makes this dependence explicit and shows what existing physics-informed losses miss. The cumulative LDOS represents the Hamiltonian in the language of the density without losing the orbital energies: it sees every part of the error, weights it by its effect on the spectrum, and determines the Hamiltonian uniquely. As a drop-in training signal, it reduces the HOMO–LUMO gap error by 49–79% on three architectures, maintains these gains under distribution shift, and leaves the element-wise accuracy intact. Appendix I discusses limitations and future work.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods: AI coding assistants helped write parts of the software implementation, including experiment orchestration scripts and figure-generation code. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat the dataset, support qualitative and thematic data analysis, or interpret results. All mathematical proofs were written by the authors; generative AI tools were subsequently used to polish the exposition, and numerical scripts were used to verify the correctness of every closed-form expression (see Appendix H). Additionally, we used generative AI tools to identify and summarize relevant literature during the related-work survey and to edit the paper to improve readability after the authors had written the full draft. We have reviewed all AI-assisted work. AI-generated code was tested for correctness and verified against quantum-chemical reference calculations by the authors. Literature identified by AI was verified through manual inspection of the original sources. All experiments were designed and executed by the authors, and all reported numbers were checked against the DFT reference data. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code for computing the proposed losses, the cached quadrature grids, and the evaluation scripts are provided in the supplementary material. Appendix J lists the fine-tuning settings and the evaluation protocol, and the README of the supplementary code gives the training commands with these settings. Complete proofs of the theoretical results appear in Appendix B, and Appendix H reports numerical checks of every analytical statement against reference data. The three backbone architectures we build on, QHNet, SPHNet and WANet, are publicly available together with the QH9 dataset and its published splits.

ETHICS STATEMENT

This work develops auxiliary training objectives for neural-network models of molecular Kohn–Sham Hamiltonians. The research uses only publicly available quantum-chemical datasets (QH9) and involves no human subjects, private data, or dual-use applications. We are not aware of any direct negative societal impact; by improving the physical accuracy of predicted Hamiltonians without requiring additional data or architectural changes, the method may help reduce the computational cost of molecular simulations. We have read and adhered to the ICLR Code of Ethics.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, 1997.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

J. Austin Ellis, Lenz Fiedler, Gabriel A. Popoola, Normand A. Modine, John A. Stephens, Aidan P. Thompson, Attila Cangi, and Sivasankaran Rajamanickam. Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks. *Physical Review B*, 104:035120, 2021.

Qiangqiang Gu, Linfeng Zhang, and Ji Feng. Neural network representation of electronic structure from ab initio molecular dynamics. *Science Bulletin*, 67(1):29–37, 2022.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

Peter Bjørn Jørgensen and Arghya Bhowmik. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Computational Materials*, 8:183, 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. In *International Conference on Learning Representations (ICLR)*, 2026.

Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer, Berlin, 1995.

Walter Kohn. Nobel lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253–1266, 1999.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

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

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*, 2:367–377, 2022.

Yunyang Li, Zaishuo Xia, Lin Huang, Xinran Wei, Samuel Harshe, Han Yang, Erpai Luo, Zun Wang, Jia Zhang, Chang Liu, Bin Shao, and Mark Gerstein. Enhancing the scalability and applicability of Kohn-Sham Hamiltonians for molecular systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Erpai Luo, Xinran Wei, Lin Huang, Yunyang Li, Han Yang, Zaishuo Xia, Zun Wang, Chang Liu, Bin Shao, and Jia Zhang. Efficient and scalable density functional theory Hamiltonian prediction through adaptive sparsity. In *International Conference on Machine Learning (ICML)*, 2025.

Péter Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Kristof T. Schütt, Michael Gastegger, Alexandre Tkatchenko, Klaus-Robert Müller, and Reinhard J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10:5024, 2019.

Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

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

Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover, Mineola, NY, 1996.

Jerry Tersoff and Donald R. Hamann. Theory of the scanning tunneling microscope. *Physical Review B*, 31(2):805–813, 1985.

Oliver T. Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Konstantin Ushenin, Kuzma Khrabrov, Artem Tsypin, Anton Ber, Egor Rumiantsev, and Artur Kadurin. LAGNet: better electron density prediction for LCAO-based data and drug-like substances. *Journal of Cheminformatics*, 17:65, 2025.

Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. QH9: A quantum Hamiltonian prediction benchmark for QM9 molecules. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 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 (ICML)*, 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. In *International Conference on Machine Learning (ICML)*, 2024.

A KOHN–SHAM DFT IN BRIEF

This section collects the quantum-chemistry background assumed in the main text: a glossary of terms (Table 2), and the self-consistent-field procedure and its cost (Figure 3).

Glossary. Table 2 defines the quantum-chemistry terms used in the paper.

Table 2: Quantum-chemistry terms used in this paper.

term	meaning
atomic number Z	Number of protons of a nucleus; it identifies the element (H 1, C 6, N 7, O 8, F 9).
basis set, basis function, atomic orbital	A fixed set of functions ϕ_μ centred on the atoms, in which the molecular orbitals are expanded; each function is called an atomic orbital.
def2-SVP	A standard Gaussian basis set of split-valence quality with polarization functions; it gives 5 functions per H atom and 14 per C, N, O or F atom.
molecular orbital ψ_i	A one-electron wavefunction of the molecule, a linear combination of basis functions; $\lvert\psi_i(\mathbf{r})\rvert^2$ is the probability density of an electron in that orbital.
Kohn–Sham DFT	Density functional theory in which the interacting electrons are replaced by non-interacting electrons in an effective potential that depends on the electron density.
exchange–correlation functional, B3LYP	The approximate part of the effective potential that accounts for many-electron effects; B3LYP is a widely used hybrid functional, which mixes in exact exchange.
level of theory	The functional and basis set of a calculation, here B3LYP/def2-SVP.
Kohn–Sham matrix $\mathbf{H}$	The matrix of the Kohn–Sham operator in the basis, also called the Hamiltonian or Fock matrix.
overlap matrix $\mathbf{S}$	The inner products of the basis functions; it differs from the identity because the basis functions are not orthogonal.
generalized eigenvalue problem	$\mathbf{Hc} = \varepsilon\mathbf{Sc}$; its eigenvalues are the orbital energies and its eigenvectors the orbital coefficients.
self-consistent field (SCF)	The iteration that makes the Hamiltonian and the density consistent with each other (Figure 3).
initial guess (`minao`, core)	The starting point of SCF: `minao` superposes atomic densities in a minimal basis; the core guess ignores the interaction between electrons.
DIIS	The standard extrapolation scheme that accelerates SCF convergence (Pulay, 1980).
closed shell	Every occupied orbital holds two electrons of opposite spin, so $n_o = N_e/2$.
occupied, virtual orbitals	The n_o lowest orbitals hold the electrons; the others, called virtual or unoccupied, are empty.
HOMO, LUMO, gap	The highest occupied and lowest unoccupied molecular orbitals, and the difference of their energies, a common indicator of reactivity and optical excitation.
frontier orbitals	The orbitals closest in energy to the HOMO–LUMO gap.
core orbitals	The deep $1s$ orbitals of the heavy atoms, far below the valence orbitals.
Fermi level μ	Here, the energy midway between the HOMO and the LUMO; the orbitals below it are occupied.
electron density $\rho(\mathbf{r})$	The expected number of electrons per unit volume at $\mathbf{r}$.
density matrix $\mathbf{P}$	$2\mathbf{C}_\mathcal{O}\mathbf{C}_\mathcal{O}^\top$, the matrix that gives the density through $\rho(\mathbf{r}) = \boldsymbol{\phi}(\mathbf{r})^\top \mathbf{P}\, \boldsymbol{\phi}(\mathbf{r})$.
local density of states (LDOS)	The density at $\mathbf{r}$ resolved by orbital energy; it is measured by scanning tunnelling microscopy.
units	Hartree (Ha), the atomic unit of energy: $1\,\text{Ha} = 27.211\,\text{eV}$, $1\,\mu\text{Ha} = 10^{-6}\,\text{Ha}$, $1\,\text{kcal/mol} = 1.594\,\text{mHa}$.

The SCF procedure and its cost. The Kohn–Sham operator $\mathcal{F}[\rho]$ depends on the electron density, and the density is computed from the eigenvectors of its matrix. DFT resolves this circularity by iteration (Figure 3a). Starting from an initial guess of the density, each iteration builds the Kohn–Sham matrix from the current density, solves the generalized eigenvalue problem equation 1, occupies the n_o lowest orbitals and forms a new density by equation 2, until the input and output densities agree. Building the matrix requires the Coulomb and, for hybrid functionals such as B3LYP, the exact-exchange contributions, formally $O(n^4)$ in the number of basis functions, together with a numerical quadrature of the exchange–correlation potential; the diagonalization costs $O(n^3)$. A calculation with T iterations costs T times one iteration. On 50 QH9-stable-ood test molecules, PySCF with its default settings (B3LYP/def2-SVP, DIIS (Pulay, 1980)) converges in a median of $T = 9$ iterations (range 8–10) from its default `minao` guess and in 17 from the core-Hamiltonian guess, taking a median of about 9 s per molecule on a CPU node. A Hamiltonian network replaces the whole loop (Figure 3b): one forward pass predicts the $O(n^2)$ entries of $\hat{\mathbf{H}}$, and one diagonalization gives every downstream quantity. On the same molecules, the QHNet forward pass on a GPU and the eigensolve take a median of 0.1 s together. Used instead as an initial guess, the prediction saves about a third of the SCF iterations (Appendix G.5). The loss proposed in this paper changes only training; inference is that of the backbone.

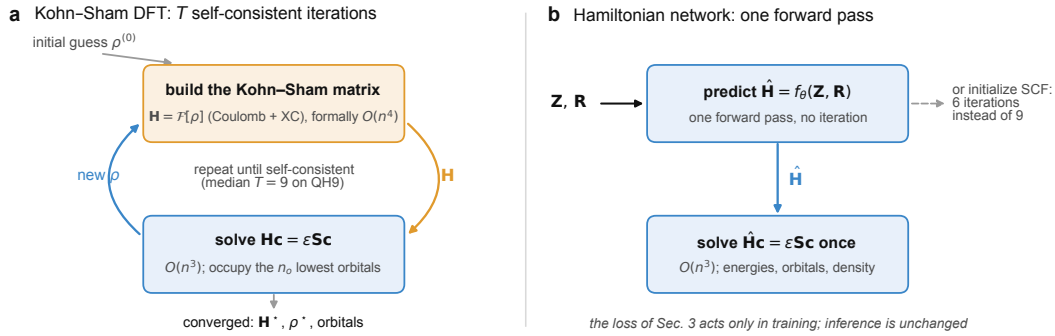


Figure 3: **(a)** The self-consistent-field loop of Kohn–Sham DFT: the Kohn–Sham matrix is built from the current density and diagonalized, and the occupied orbitals give the next density, until the two agree. **(b)** A Hamiltonian network predicts the converged matrix in one forward pass; one diagonalization gives the orbital energies, orbitals and density, and the prediction can also initialize SCF. Iteration counts are medians over 50 QH9-stable-ood test molecules (PySCF, B3LYP/def2-SVP).

B Proofs and Additional Theory

Throughout, $\mathbf{P} = 2\mathbf{C}^\star_{\mathcal{O}}\mathbf{C}^{\star\top}_{\mathcal{O}}$ denotes the density matrix, so that $\rho(\mathbf{r}) = \boldsymbol{\phi}(\mathbf{r})^\top \mathbf{P}\, \boldsymbol{\phi}(\mathbf{r})$ with $\boldsymbol{\phi}(\mathbf{r}) = (\phi_1(\mathbf{r}), \ldots, \phi_n(\mathbf{r}))^\top$; hats denote the same quantities computed from $\hat{\mathbf{H}}$. We write $g := \varepsilon^\star_{n_o+1} - \varepsilon^\star_{n_o}$ for the reference HOMO–LUMO gap and $\mathcal{U} := \operatorname{span}\{\phi_\mu\}_{\mu=1}^n$ for the functions representable in the basis, which contains every predicted and reference orbital. Gaussian-type basis functions are real-analytic on $\mathbb{R}^3$ and linearly independent, so every nonzero $u \in \mathcal{U}$ is a nonzero real-analytic function. Two consequences are used below: a nonzero $u \in \mathcal{U}$ does not vanish identically on any nonempty open set, and if $u, v \in \mathcal{U}$ satisfy $uv = 0$ on $\mathbb{R}^3$, then $u = 0$ or $v = 0$ (if $u \neq 0$, v vanishes on the dense open set where $u \neq 0$, hence everywhere).

B.1 The error in the orbital basis

Whitening a matrix $\mathbf{M}$ means expressing it in an orthonormal basis of the span of the atomic orbitals, $\mathbf{M} \mapsto \mathbf{S}^{-1/2}\mathbf{M}\,\mathbf{S}^{-1/2}$ (Löwdin orthogonalization). The error $\boldsymbol{\Delta}$ of equation 4 is the whitened error rotated into the reference eigenbasis, which gives the following identities.

Lemma B.1 (Whitening)**.** $\|\boldsymbol{\Delta}\|_F = \|\mathbf{S}^{-1/2}\delta\mathbf{H}\,\mathbf{S}^{-1/2}\|_F \le \|\mathbf{S}^{-1}\|_2\,\|\delta\mathbf{H}\|_F = \frac{\kappa(\mathbf{S})}{\|\mathbf{S}\|_2}\|\delta\mathbf{H}\|_F$. *Moreover* $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star - \operatorname{diag}(\boldsymbol{\varepsilon}^\star) = \boldsymbol{\Delta}$*, so the full WALoss of Li et al. (2025), an average of the absolute or squared entries of this matrix, weights every entry of* $\boldsymbol{\Delta}$ *alike (its mean-squared form is* $\|\boldsymbol{\Delta}\|_F^2/n^2$*), and the diagonal-only form equals* $\sum_i w_i d_i^2$*.*

Proof. Write $\mathbf{C}^\star = \mathbf{S}^{-1/2}\mathbf{Q}$ with $\mathbf{Q}$ orthogonal (from $\mathbf{C}^{\star\top}\mathbf{S}\mathbf{C}^\star = \mathbf{I}$). Then $\boldsymbol{\Delta} = \mathbf{Q}^\top\mathbf{S}^{-1/2}\delta\mathbf{H}\mathbf{S}^{-1/2}\mathbf{Q}$, and the Frobenius norm is orthogonally invariant. The bound follows from submultiplicativity and $\|\mathbf{S}^{-1/2}\|_2^2 = \|\mathbf{S}^{-1}\|_2$. The identity $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star - \operatorname{diag}(\boldsymbol{\varepsilon}^\star) = \boldsymbol{\Delta}$ is equation 4, because $\mathbf{C}^{\star\top}\mathbf{H}^\star\mathbf{C}^\star = \operatorname{diag}(\boldsymbol{\varepsilon}^\star)$ by equation 1, and the diagonal of $\boldsymbol{\Delta}$ is (d_i). □

The inequality is the core of the perturbation theorem of Li et al. (2025). On QH9 it is loose: Appendix D.4 measures a median amplification $\|\boldsymbol{\Delta}\|_F/\|\delta\mathbf{H}\|_F$ of 3.0 against a median bound of 1878, so the location of the error, rather than its size, is what determines its effect. Table 3 applies Propositions 3.1 and 3.2 to the losses of Section 2; Figure 1e shows the same comparison.

Table 3: Which sectors of the error each loss responds to (first order), how it weights them, and whether it uses the eigenpairs of the prediction. The element-wise loss is uniform up to the overlap metric (Lemma B.1).

loss	energies d_i	$\boldsymbol{\Delta}_{\mathcal{OV}}$ (density)	$\boldsymbol{\Delta}_{\mathcal{OO}}, \boldsymbol{\Delta}_{\mathcal{VV}}$	weighting	eigenpairs of $\hat{\mathbf{H}}$
element-wise $\mathcal{L}_{\text{elem}}$	✓	✓	✓	uniform	–
WALoss, full (Li et al., 2025)	✓	✓	✓	uniform	–
WALoss, diagonal-only (Li et al., 2025)	first order	–	–	occupied orbitals and LUMO	–
density matching $\mathcal{L}_\rho$	–	✓	–	frontier, $1/(\varepsilon_a - \varepsilon_i)$	✓
cumulative LDOS $\mathcal{L}_{\text{cum}}$ (ours)	✓	✓	✓	frontier and energy window	✓

B.2 PERTURBATION THEORY FOR EACH PROPERTY

Besides Proposition 3.1, the main text uses two second-order facts.

Proposition B.2 (Second-order corrections)**.** *Under the assumptions of Proposition 3.1:*

(d) $\hat{\varepsilon}_i - \varepsilon_i^\star = d_i + \sum_{j\neq i} \frac{\Delta_{ij}^2}{\varepsilon_i^\star - \varepsilon_j^\star} + O(\|\mathbf{\Delta}\|^3)$*;*

(e) the Kohn–Sham total energy is stationary in $\mathbf{P}$ *at the ground state, so its error is second order in* $\mathbf{\Delta}_{\mathcal{OV}}$ *and has no first-order dependence on any sector.*

Item (d) is the channel through which off-diagonal errors reach the orbital energies: a coupling Δ_{ij} between two orbitals close in energy shifts both with weight $1/(\varepsilon_i^\star - \varepsilon_j^\star)$, which is largest across the HOMO–LUMO gap. It is also why a loss that sees only the diagonal, and only to first order, such as the diagonal-only WALoss, can be lowered while the true orbital energies stay wrong. Item (e) explains why total-energy errors are a late indicator of the problem (Li et al., 2025): the energy is the least sensitive probe of the occupied subspace and the density the most sensitive one.

Remark B.3 (What the sectors mean for the metrics). To first order, the gap error is $d_{\mathrm{LUMO}} - d_{\mathrm{HOMO}}$, two entries of sector A out of $n(n+1)/2$; the orbital coefficients, and with them the similarity ψ, depend on sectors B and C; and the density depends on sector B alone. The weights in (b) and (c) are inverse orbital-energy differences. On QH9, the HOMO and LUMO are about 0.25 Ha apart and the deepest core orbital lies about 19 Ha below the LUMO (Table 5), so an error coupling the two frontier orbitals changes the LUMO about 75 times more than an equally large error coupling it to the deepest core orbital (median over QH9; 5–95% range 51–109). Item (c) is the first-order form of the Davis–Kahan theorem (Davis & Kahan, 1970): the occupied subspace, and with it the density, is rotated only by sector B, with a strength set by the gap (Appendix B).

Proof of Propositions 3.1 and B.2. (a), (b) and (d) are Rayleigh–Schrödinger perturbation theory for the pencil $(\mathbf{H}^\star + \delta\mathbf{H}, \mathbf{S})$ after whitening (Kato, 1995; Stewart & Sun, 1990). For (c), $\delta\mathbf{P} = 2\sum_{i\in\mathcal{O}}(\delta\mathbf{c}_i\mathbf{c}_i^\top + \mathbf{c}_i\delta\mathbf{c}_i^\top)$; inserting (b) and pairing the terms (i,j) and (j,i) with $i,j \in \mathcal{O}$ gives $\Delta_{ji}/(\varepsilon_i - \varepsilon_j) + \Delta_{ij}/(\varepsilon_j - \varepsilon_i) = 0$ by symmetry of $\mathbf{\Delta}$, which leaves

$$\delta\mathbf{P} = 2\sum_{i\in\mathcal{O}}\sum_{a\in\mathcal{V}} \frac{\Delta_{ai}}{\varepsilon_i^\star - \varepsilon_a^\star}\left(\mathbf{c}_i^\star\mathbf{c}_a^{\star\top} + \mathbf{c}_a^\star\mathbf{c}_i^{\star\top}\right)$$

and equation 5. Part (c) does not need the non-degeneracy assumption. $\hat{\mathbf{P}}/2$ is the spectral projector of the pencil onto its n_o lowest eigenvalues, which is analytic in $\mathbf{\Delta}$ while $\|\mathbf{\Delta}\|_2 < g/2$ (Kato, 1995); expanding its resolvent integral to first order gives entries $\Delta_{ai}/(\varepsilon_i^\star - \varepsilon_a^\star)$ between occupied and virtual orbitals and zero elsewhere, whether or not levels within $\mathcal{O}$ or within $\mathcal{V}$ coincide. The pairwise cancellation above is the same computation for non-degenerate levels. Beyond first order, the $\sin\Theta$ theorem of Davis & Kahan (1970) gives the same picture. In the reference orbital basis, the predicted orbitals are the eigenvectors of $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star = \mathrm{diag}(\boldsymbol{\varepsilon}^\star) + \mathbf{\Delta}$, and the occupied reference subspace leaves a residual whose only nonzero block is $\mathbf{\Delta}_{\mathcal{VO}}$; if $\|\mathbf{\Delta}\|_2 < g/2$, the canonical angles Θ between the occupied subspaces of the reference and the prediction, which measure how far the occupied subspace, and with it $\hat{\mathbf{P}}$ and $\hat{\rho}$, has moved, therefore satisfy $\|\sin\Theta\|_F \le \|\mathbf{\Delta}_{\mathcal{OV}}\|_F/(g - 2\|\mathbf{\Delta}\|_2)$ (Stewart & Sun, 1990). (e) is the variational principle for the Kohn–Sham energy with respect to occupied–virtual rotations.

Proof of Proposition 3.2. If $\mathbf{\Delta}_{\mathcal{OV}} = 0$ then $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^\star$ is block diagonal in $(\mathcal{O},\mathcal{V})$, so $\mathrm{span}\{\mathbf{C}^\star_{\mathcal{O}}\}$ is an $\mathbf{S}$-invariant subspace of $\hat{\mathbf{H}}$; if no eigenvalue crosses the gap it is the occupied subspace of $\hat{\mathbf{H}}$, and $\hat{\mathbf{P}} = \mathbf{P}^\star$ because $\mathbf{P}$ depends only on the subspace. The dimension count is that of symmetric matrices on $\mathcal{O}$ and on $\mathcal{V}$.

Remark (directions versus norm). Proposition 3.2 counts directions, not norm: at the QH9 median, 69% of the parameters of $\mathbf{\Delta}$ lie in the two blind sectors (Table 5), and this is the room that density matching leaves unpenalized. How much of an actual residual sits in that room is a separate, empirical question. For the physically realized Hamiltonian changes of Appendix D.4, the blind sectors carry 15.6% of $\|\mathbf{\Delta}\|_F^2$ at a displacement of 0.02 Å and 30.3% at 0.30 Å, almost all of the growth in the virtual–virtual block that contains the LUMO energy error. The blind share is therefore smaller in norm than in dimension, but it contains every orbital-energy error and grows under distribution shift.

B.3 PROPERTIES OF THE CUMULATIVE LDOS LOSS

Theorem 3.5 collects the following results.

Proposition B.4 (Fermi-level identity; Theorem 3.5(i)). *Let (ε_i, ψ_i) be the orbitals of $\mathbf{H}^\star$ or of $\hat{\mathbf{H}}$, and let the energy E lie between their HOMO and LUMO energies at a distance of at least $t\sigma$ from both. Then, at every point $\mathbf{r}$, $|N(\mathbf{r}, E) - \rho(\mathbf{r})/2| \le \frac{1}{2}\operatorname{erfc}(t/\sqrt{2}) \sum_i |\psi_i(\mathbf{r})|^2$: each orbital deviates by at most a fraction $\frac{1}{2}\operatorname{erfc}(t/\sqrt{2})$ of its density. At $E = \mu$ and for the reference, $t = g/(2\sigma)$, and the fraction is 3×10^{-7} for $g = 0.2$ Ha and $\sigma = 0.02$ Ha.*

Proof. By equation 2, $N(\mathbf{r}, E) - \frac{1}{2}\rho(\mathbf{r}) = -\sum_{i\in\mathcal{O}} |\psi_i(\mathbf{r})|^2 \big[1 - \Phi\big((E - \varepsilon_i)/\sigma\big)\big] + \sum_{a\in\mathcal{V}} |\psi_a(\mathbf{r})|^2\, \Phi\big((E-\varepsilon_a)/\sigma\big)$. Every bracket and every Φ in this expression is at most $1 - \Phi(t) = \frac{1}{2}\operatorname{erfc}(t/\sqrt{2})$, and the two sums have opposite signs, so the total is bounded by the larger of them. □

The identity holds at every energy inside the gap, not only at μ. This matters for the training grid: with $K = 64$ points on $[\mu - 1.5, \mu + 0.3]$ Ha, μ is not a grid point but lies midway between the points $\mu \pm 0.014$ Ha. With $t = (g/2 - 0.014\,\text{Ha})/\sigma$, both are density slices up to 2×10^{-8} per orbital at the median QH9 gap and 1×10^{-4} at its 5th percentile, so the loss contains density matching through these two slices (Table 14). Orbitals far below the window contribute a constant step equal to their density, so nothing that density matching sees is lost; the other grid points add how that density is distributed in energy.

Theorem B.5 (No blind sector and identifiability; Theorem 3.5(ii, iii)). *Let the reference orbital energies be non-degenerate, and let $I \subseteq \mathbb{R}$ be an open energy interval, for example the loss window. Write φ for the standard normal density, $\Phi_i(E) := \Phi((E - \varepsilon_i^\star)/\sigma)$ and $\varphi_i(E) := \varphi((E - \varepsilon_i^\star)/\sigma)$. Then*

(i) the first-order response $\boldsymbol{\Delta} \mapsto \delta N|_{\mathbb{R}^3 \times I}$ is injective on symmetric $\boldsymbol{\Delta}$: energy errors d_i enter through the step positions with weight $|\psi_i(\mathbf{r})|^2 \varphi_i(E)/\sigma$, and every off-diagonal Δ_{ij}, including the occupied–occupied and virtual–virtual entries invisible to ρ, enters through the step heights with the E-profile $\Phi_i(E) - \Phi_j(E)$, a smoothed indicator of the interval between $\varepsilon_i^\star$ and $\varepsilon_j^\star$;

(ii) $N[\hat{\mathbf{H}}] = N[\mathbf{H}^\star]$ on $\mathbb{R}^3 \times I$ if and only if $\hat{\mathbf{H}} = \mathbf{H}^\star$, and then $\hat{\varepsilon}_i = \varepsilon_i^\star$ and $\hat{\mathbf{c}}_i = \pm\mathbf{c}_i^\star$ for every i. By contrast, $\mathcal{L}_\rho$ vanishes on the blind set of Proposition 3.2, and the diagonal-only WALoss vanishes whenever $d_i = 0$ for every i with $w_i > 0$.

Proof. We use one fact about the energy profiles. *(F) For distinct $\lambda_1, \dots, \lambda_m$, the $2m$ functions $\Phi((E - \lambda_k)/\sigma)$ and $\varphi((E - \lambda_k)/\sigma)$ are linearly independent on I.* All of them extend to entire functions of E, so a combination $\sum_k [\alpha_k \Phi((E - \lambda_k)/\sigma) + \beta_k \varphi((E - \lambda_k)/\sigma)]$ that vanishes on I vanishes on $\mathbb{R}$. Differentiating it and multiplying by $\sqrt{2\pi}\,\sigma\, e^{E^2/2\sigma^2}$ gives $\sum_k \big[\alpha_k - \beta_k(E - \lambda_k)/\sigma\big] e^{-\lambda_k^2/2\sigma^2} e^{\lambda_k E/\sigma^2} = 0$, and exponential polynomials with distinct exponents vanish only if every polynomial coefficient does, so $\alpha_k = \beta_k = 0$ for all k.

(i) By Proposition 3.1(a,b), the first-order change of equation 6 is

$$\delta N(\mathbf{r}, E) = \sum_{i=1}^{n} \Big[2\psi_i(\mathbf{r})\, \delta\psi_i(\mathbf{r})\, \Phi_i(E) - \frac{d_i}{\sigma}\, \psi_i(\mathbf{r})^2\, \varphi_i(E)\Big], \qquad \delta\psi_i = \sum_{j \ne i} \frac{\Delta_{ji}}{\varepsilon_i^\star - \varepsilon_j^\star}\, \psi_j \in \mathcal{U}.$$

Collecting the terms (i, j) and (j, i) gives $2\Delta_{ij}\, \psi_i \psi_j\, [\Phi_i(E) - \Phi_j(E)]/(\varepsilon_i^\star - \varepsilon_j^\star)$, the step-height profile of the statement. If $\delta N = 0$ on $\mathbb{R}^3 \times I$, then (F), applied at each $\mathbf{r}$, gives $\psi_i\, \delta\psi_i = 0$ and $d_i \psi_i^2 = 0$ on $\mathbb{R}^3$ for every i. Since $\psi_i \ne 0$, the product property of $\mathcal{U}$ gives $\delta\psi_i = 0$, and $d_i = 0$; linear independence of the ψ_j then gives $\Delta_{ji} = 0$ for all $j \ne i$. Hence $\boldsymbol{\Delta} = \mathbf{0}$.

(ii) For a Hamiltonian $\mathbf{H}$ with eigenpairs $(\varepsilon_i, \mathbf{c}_i)$, group equation 6 by distinct eigenvalues, $N[\mathbf{H}](\mathbf{r}, E) = \sum_\lambda w_\lambda(\mathbf{r})\, \Phi((E - \lambda)/\sigma)$ with $w_\lambda = \sum_{i:\, \varepsilon_i = \lambda} \psi_i^2$. If $N[\hat{\mathbf{H}}] = N[\mathbf{H}^\star]$ on $\mathbb{R}^3 \times I$, (F) applied to the union of the two spectra gives $\hat{w}_\lambda = w_\lambda^\star$ on $\mathbb{R}^3$ for every λ. For an eigenvalue λ of $\hat{\mathbf{H}}$, $\hat{w}_\lambda$ is a sum of squares of nonzero functions in $\mathcal{U}$, so it is not identically zero and λ is an eigenvalue of $\mathbf{H}^\star$; the converse holds likewise. $\mathbf{H}^\star$ has n distinct eigenvalues and $\hat{\mathbf{H}}$ has n eigenvalues counted with multiplicity, so every eigenvalue of $\hat{\mathbf{H}}$ is simple and $\hat{\varepsilon}_i = \varepsilon_i^\star$. Then $\hat{\psi}_i^2 = \psi_i^2$,

that is, $(\hat{\psi}_i - \psi_i)(\hat{\psi}_i + \psi_i) = 0$ on $\mathbb{R}^3$, so $\hat{\psi}_i = \pm\psi_i$ and $\hat{\mathbf{c}}_i = \pm\mathbf{c}_i^\star$. Finally, equation 1 gives $\mathbf{H} = \mathbf{S}\mathbf{C}\,\mathrm{diag}(\boldsymbol{\varepsilon})\mathbf{C}^\top\mathbf{S}$ for either Hamiltonian, and the signs cancel, so $\hat{\mathbf{H}} = \mathbf{H}^\star$; the converse is immediate. The statements about $\mathcal{L}_\rho$ and the diagonal-only WALoss are Proposition 3.2 and Lemma B.1. □

The theorem concerns the representation N over all of space; the loss is its sample estimate at 256 points and $K = 64$ energies per molecule (Section 4.1). It needs nothing of the basis beyond linear independence and the analyticity of Gaussian-type functions. We state it on all of space rather than on a finite point set because the natural finite-sample condition, linear independence of the products $\phi_\mu\phi_\nu$ on the points, fails in def2-SVP: on an atom with two p shells of radial parts R_1, R_2, $p_{1x}p_{2y} = xyR_1R_2 = p_{1y}p_{2x}$ identically. Because I may be any interval, the theorem applies to the loss window, but orbital pairs far outside the window enter only through the Gaussian tails of their profiles, which is the energy-window weighting of Table 3. Table 14 checks both statements on QH9 molecules. On the level-0 quadrature grid, the first-order response is injective over the full energy range, and on the loss window it resolves every direction that involves an orbital inside the window, although some only weakly. The theorem does not carry over to the 256 points sampled per molecule: there, some directions on the loss window remain unresolved, mostly couplings between occupied orbitals and virtual orbitals above the window, whose energy profiles are nearly constant over the window, and less than 1% of their weight lies on pairs of orbitals inside the window. The element-wise term supervises these directions, and a denser sample resolves them.

Proposition B.6 (Non-saturating restoring force; Theorem 3.5(iv))**.** *Displace one orbital energy by δ with its shape fixed. Per sample point and up to the factor $|\psi_i(\mathbf{r})|^4$, the mismatch integrated over E is, for a unit-height Gaussian peak, $g(\delta) = 2\sqrt{\pi}\,\sigma\big(1 - e^{-\delta^2/4\sigma^2}\big)$, which saturates at $2\sqrt{\pi}\sigma$ with $g'(\delta) = (\sqrt{\pi}\,\delta/\sigma)\,e^{-\delta^2/4\sigma^2} \to 0$; and, for the cumulative step, $c(\delta) = \delta\ \mathrm{erf}\big(\frac{\delta}{2\sigma}\big) + \frac{2\sigma}{\sqrt{\pi}}\big(e^{-\delta^2/4\sigma^2} - 1\big)$, with $c'(\delta) = \mathrm{erf}\big(\frac{\delta}{2\sigma}\big) \to \mathrm{sign}\,\delta$. As $\sigma \to 0$, $\mathcal{L}_{\mathrm{cum}}$ becomes the Cramér–von Mises distance between the local spectral measures $\sum_i |\psi_i(\mathbf{r})|^2\delta_{\varepsilon_i}$ of $\hat{\mathbf{H}}$ and $\mathbf{H}^\star$.*

Proof. For any square-integrable f, $\int (f(E-\delta) - f(E))^2 dE = 2\|f\|_2^2 - 2(f \star f)(\delta)$ with $f \star f$ the autocorrelation. For the unit-height Gaussian $f(E) = e^{-E^2/2\sigma^2}$, $\|f\|_2^2 = \sigma\sqrt{\pi}$ and $(f \star f)(\delta) = \sigma\sqrt{\pi}\,e^{-\delta^2/4\sigma^2}$, giving $g(\delta)$. For the step, differentiate $c(\delta) = \int(\Phi((E-\delta)/\sigma) - \Phi(E/\sigma))^2 dE$ under the integral: $c'(\delta) = \frac{2}{\sigma}\int(\Phi(E/\sigma) - \Phi((E-\delta)/\sigma))\varphi((E-\delta)/\sigma)\,dE = \mathrm{erf}(\delta/2\sigma)$, the last step being the Gaussian convolution identity $\int \Phi(x)\varphi(x-a)dx$ evaluated in closed form; integrating from $c(0) = 0$ gives the stated $c(\delta)$. Both closed forms agree with numerical quadrature to machine precision (Table 14). As $\sigma \to 0$, $c(\delta) \to |\delta|$, the W_1 distance between two unit point masses, and for general spectra $\int (N[\hat{\mathbf{H}}] - N[\mathbf{H}^\star])^2 dE$ is the Cramér–von Mises statistic between the two local spectral measures. □

A peak stops pulling once it no longer overlaps its target; a step does not, as long as the displaced level stays inside the energy window (Figure 11b,c). At $\sigma = 0.02$ Ha and an energy error of 0.1 Ha, the peak gradient has fallen to 1% of its maximum, while the cumulative gradient is at 99.96% of its limit. On a window $[a, b]$, the part of the integral for $c'(\delta)$ that the window omits is bounded by $\frac{2}{\sigma}\int \varphi((x-\delta)/\sigma)\,dx$ over $x \notin [a - \varepsilon_i^\star,\ b - \varepsilon_i^\star]$, so

$$\left| c'_{[a,b]}(\delta) - \mathrm{erf}\left(\frac{\delta}{2\sigma}\right)\right| \le 2\,\Phi\Big(\frac{a - \varepsilon_i^\star - \delta}{\sigma}\Big) + 2\,\Phi\Big(\frac{\varepsilon_i^\star + \delta - b}{\sigma}\Big).$$

The restoring force is therefore that of the proposition, up to an exponentially small error, while the displaced level is a few σ inside the window, and it vanishes, like the peak's, once the level has left the window. The step is also wide in E, so the coarse training grid does not alias it: on the $K = 64$ grid, whose spacing is 1.43σ, the gradient stays within 0.01 of $\mathrm{erf}(\delta/2\sigma)$ for $|\delta| \le 0.1$ Ha and every level at least 0.18 Ha inside the window (Table 14).

B.4 Error reallocation

Proposition B.7 (Error reallocation)**.** *Let $\mathcal{M} \not\ni 0$ be the set of errors a model can reach and W a sector weighting with large entries on d_i and $\mathbf{\Delta}_{\mathcal{OV}}$. If $\mathbf{\Delta}_{\mathrm{iso}}$ minimizes $\|\mathbf{\Delta}\|_F^2$ over $\mathcal{M}$ and $\mathbf{\Delta}_w$ minimizes $\|\mathbf{\Delta}\|_F^2 + w\|W \circ \mathbf{\Delta}\|_F^2$, then $\|\mathbf{\Delta}_w\|_F \geq \|\mathbf{\Delta}_{\mathrm{iso}}\|_F$ and $\|W \circ \mathbf{\Delta}_w\|_F \leq \|W \circ \mathbf{\Delta}_{\mathrm{iso}}\|_F$, both strictly unless the two objectives share a minimizer.*

Proof. $\mathbf{\Delta}_w$ is feasible for the first problem, so $\|\mathbf{\Delta}_{\mathrm{iso}}\|^2 \leq \|\mathbf{\Delta}_w\|^2$; $\mathbf{\Delta}_{\mathrm{iso}}$ is feasible for the second, so $\|\mathbf{\Delta}_w\|^2 + w\|W \circ \mathbf{\Delta}_w\|^2 \leq \|\mathbf{\Delta}_{\mathrm{iso}}\|^2 + w\|W \circ \mathbf{\Delta}_{\mathrm{iso}}\|^2$; combining the two gives both inequalities, and equality in either forces a common minimizer. □

Every eigen-derived metric is a function of the W-weighted sectors, while the complementary sectors $\mathbf{\Delta}_{\mathcal{OO}}, \mathbf{\Delta}_{\mathcal{VV}}$ hold most of the degrees of freedom. A spectral loss can therefore lower every eigen-derived metric at the price of at most a small increase in the element-wise error. The two losses also push on $\hat{\mathbf{H}}$ differently. With $f_E(x) := \Phi((E - x)/\sigma)$, the Daleckii–Krein formula for the derivative of a spectral function (Bhatia, 1997) gives $\nabla_{\hat{\mathbf{H}}} \mathcal{L}_{\mathrm{cum}} = \hat{\mathbf{C}}\, \mathbf{G}\, \hat{\mathbf{C}}^\top$ with

$$G_{ij} = \frac{2}{|\mathcal{R}|K} \sum_{\mathbf{r} \in \mathcal{R}} \sum_{k=1}^{K} \frac{N[\hat{\mathbf{H}}](\mathbf{r}, E_k) - N[\mathbf{H}^\star](\mathbf{r}, E_k)}{\mathrm{div}(\mathbf{r})^2}\, f_{ij}^{[1]}(E_k)\, \hat{\psi}_i(\mathbf{r}) \hat{\psi}_j(\mathbf{r}),$$

where $f_{ij}^{[1]}(E) = \big(f_E(\hat{\varepsilon}_i) - f_E(\hat{\varepsilon}_j)\big)/(\hat{\varepsilon}_i - \hat{\varepsilon}_j)$ for $i \neq j$ and $f_{ii}^{[1]}(E) = f_E'(\hat{\varepsilon}_i)$. The divided difference is bounded by $\min\{1/(\sqrt{2\pi}\sigma),\, 1/|\hat{\varepsilon}_i - \hat{\varepsilon}_j|\}$ and is exponentially small unless a grid energy lies between $\hat{\varepsilon}_i$ and $\hat{\varepsilon}_j$ or within a few σ of them. The gradient of $\mathcal{L}_{\mathrm{cum}}$ is thus concentrated on close and frontier pairs of predicted orbitals inside the window, whereas the gradient of the mean-absolute element-wise loss, $\mathrm{sign}(\delta\mathbf{H})/n^2$, weights every entry of $\delta\mathbf{H}$ alike. Read in reverse, the proposition also explains why continuing to minimize the element-wise loss alone on a converged model can move error *into* the sensitive sectors (Appendix G.2).

C THE QH9 DATASETS AND SPLITS

C.1 PROVENANCE AND QUANTUM-CHEMICAL SETTINGS

Every experiment in this paper uses QH9 (Yu et al., 2023a), which supplies converged Kohn–Sham Hamiltonians for QM9 (Ramakrishnan et al., 2014) molecules. The reference calculations are B3LYP with the def2-SVP basis, carried out in PySCF (Sun et al., 2018); each record stores the converged Fock matrix $\mathbf{H}^\star$ in the atomic-orbital basis together with the overlap matrix $\mathbf{S}$, the nuclear charges and the geometry. Because $\mathbf{H}^\star$ is the converged operator, the label is a fixed point of the self-consistency map and not a single SCF iterate, so a model that reproduces it reproduces the entire converged electronic structure of that geometry.

Two families are released. *QH9-stable* holds one equilibrium geometry for each of 130,831 molecules. *QH9-dynamic* holds short *ab initio* molecular-dynamics trajectories; we use the 100k release, which is 999 molecules sampled at 100 geometries each, 99,900 records in total. The dynamic family is the harder of the two: consecutive frames of a trajectory visit distorted, off-equilibrium geometries whose Hamiltonians are not close to any equilibrium structure in the stable set.

The basis is small enough to state completely. In def2-SVP a hydrogen carries five atomic orbitals ($2s$, $1p$) and each of C, N, O and F carries fourteen ($3s$, $2p$, $1d$), so the matrix dimension is determined by the composition alone: $n = 5n_{\mathrm{H}} + 14n_{\mathrm{heavy}}$. The two molecules used for the figures below check this exactly: molecule #12345 is $C_7H_{10}O$ with $n = 7 \cdot 14 + 1 \cdot 14 + 10 \cdot 5 = 162$, and molecule #23470 is C_8H_5N with $n = 9 \cdot 14 + 5 \cdot 5 = 151$.

C.2 SPLITS

The four splits used in this paper are given in Table 4. They probe different kinds of generalization, which matters for the interpretation of Table 6: *stable-id* is an interpolation test, *stable-ood* extrapolates in molecule size, *dynamic-geo* extrapolates to unseen geometries of molecules the model has seen, and *dynamic-mol* extrapolates to unseen molecules altogether. The size split is a substantial shift. Measured on a uniform 4,000-molecule sample of QH9-stable, the training side has a median of 17 atoms and $n = 166$, while the test side has a median of 23 atoms and $n = 196$, and the median gap moves from 6.43 to 8.10 eV: the test molecules are systematically larger, and have wider HOMO–LUMO gaps, than the training molecules.

Table 4: The four QH9 splits. Sizes are the exact record counts read from the released split masks. The size split thresholds are on the total atom count, hydrogens included; the random split uses a fixed permutation with seed 43.

split	selection rule	train	val	test	tests
stable-id	random $80/10/10$	104,664	13,083	13,084	interpolation
stable-ood	≤ 20 / 21–22 / ≥ 23 atoms	104,001	17,495	9,335	molecule size
dynamic-geo	geometries within a molecule	79,920	9,990	9,990	unseen geometry
dynamic-mol	disjoint molecules	79,900	9,900	10,100	unseen molecule

C.3 Composition, spectra, and the size of the blind subspace

Table 5 and Figure 4 summarize the sample. Two entries matter for the rest of the paper. First, the occupied fraction is small: the median molecule has $n_{\mathrm{occ}} = 33$ occupied orbitals out of $n = 171$ basis functions, so $f = n_{\mathrm{occ}}/n \approx 0.19$. Second, the dimension count of Proposition 3.2 is therefore severe. A symmetric residual $\mathbf{\Delta}$ has $n(n+1)/2$ free parameters, of which $n_{\mathcal{O}}(n_{\mathcal{O}}+1)/2+n_{\mathcal{V}}(n_{\mathcal{V}}+1)/2$ lie in the occupied–occupied and virtual–virtual blocks that the density cannot resolve. At the QH9 median that is a fraction 0.690, and the 5–95% range across the sample is $[0.661, 0.709]$: roughly seven of every ten directions in which a predicted Hamiltonian can be wrong are directions in which density matching is exactly blind, and this holds for every molecule in the sample.

The spectra are also wide. The median molecule has its lowest eigenvalue at -19.1 Ha and its highest at $+3.7$ Ha, while the quantities we care about live in a window of order 1 Ha around the Fermi level. Nine deep core levels of molecule #23470 sit below the loss window and thirty-two levels lie inside it; the cumulative form is defined so that the levels below the window contribute a constant step and drop out of any difference (Appendix J).

Table 5: QH9-stable composition and reference spectra, measured on a uniform 4,000-molecule sample. Ranges are 5–95% intervals unless a min–max is stated. The last row is the exact fraction of the parameters of a symmetric residual that lie in the two sectors the density loss cannot resolve.

quantity	median	range	min–max
atoms per molecule	18	–	5–29
heavy atoms (C, N, O, F)	9	–	1–9
basis functions n	171	141–196	34–226
occupied orbitals n_{occ}	33	29–35	5–37
occupied fraction $f = n_{\mathrm{occ}}/n$	0.193	0.178–0.219	0.147–0.256
HOMO–LUMO gap (Ha)	0.248	0.173–0.305	0.079–0.463
HOMO–LUMO gap (eV)	6.74	4.72–8.31	2.15–12.60
lowest eigenvalue ε_1 (Ha)	-19.14	–	-24.76–-10.17
highest eigenvalue ε_n (Ha)	3.69	–	2.91–4.14
density-blind fraction of $\mathbf{\Delta}$	0.690	0.661–0.709	–

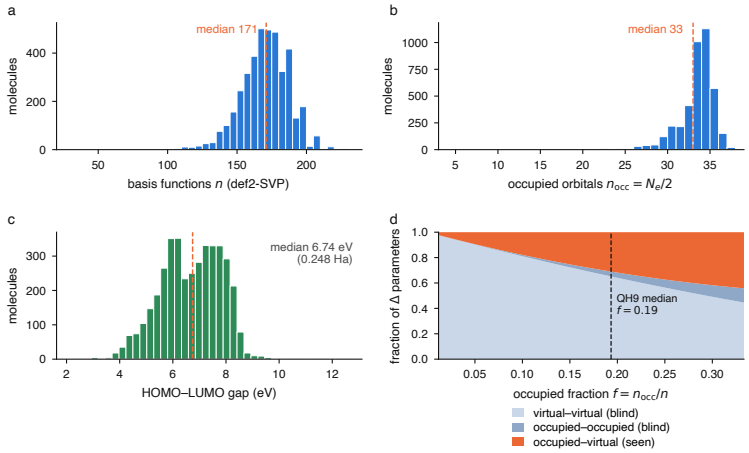


Figure 4: QH9-stable statistics from a uniform 4,000-molecule sample. **(a)** Basis size. **(b)** Number of occupied orbitals. **(c)** Reference HOMO–LUMO gap. **(d)** Exact partition of the $n(n+1)/2$ parameters of a symmetric residual $\mathbf{\Delta}$ into the three sectors of Proposition 3.1, at fixed $n = 171$, as a function of the occupied fraction f. Only the occupied–virtual slice changes the density; at the QH9 median occupied fraction it accounts for 31% of the parameters, and the remaining 69% is invisible to $\mathcal{L}_\rho$.

C.4 A GALLERY

Figure 5 shows the reference density of twelve molecules drawn at even spacing through the QH9-stable index range, on a common length scale. The sample is chemically diverse: saturated and conjugated skeletons, three- to six-membered rings, one to three heteroatoms, planar and strongly non-planar geometries, n from 100 to 191 and reference gaps from 0.14 to 0.30 Ha. A single auxiliary loss with a single weight has to be useful across all of it.

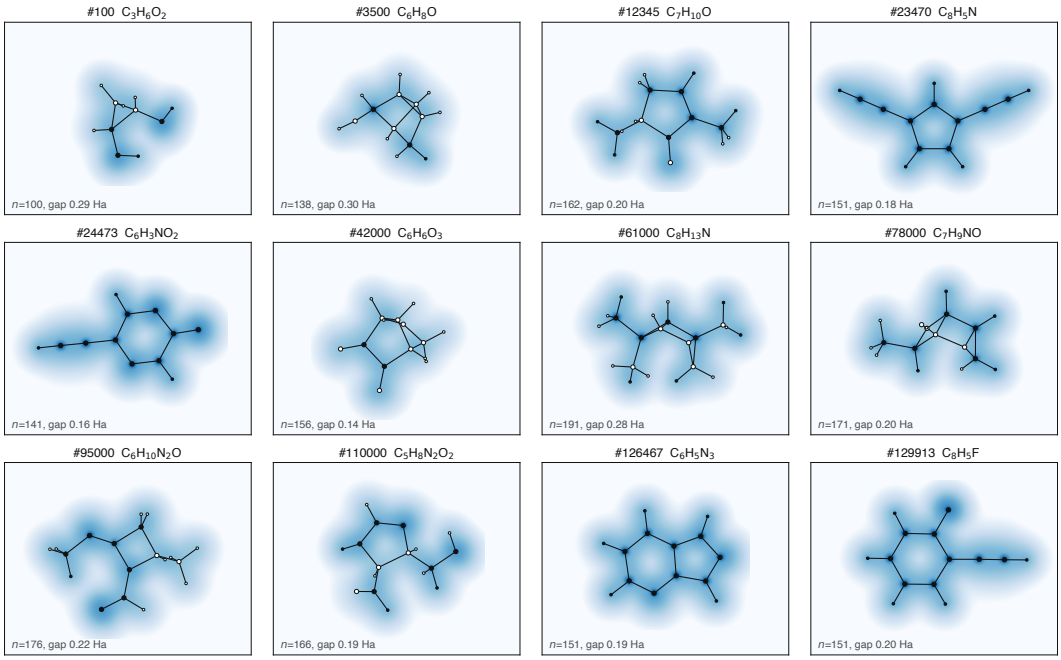


Figure 5: Reference electron density for twelve QH9-stable molecules, on the best-fit plane through the nuclei, logarithmic colour scale, all panels on the same length scale and the same colour scale. Bonds are drawn from covalent-radius criteria; nuclei more than $0.7\,a_0$ off the plane are drawn open, so a panel with many open markers is a non-planar molecule seen in projection. Molecular formulae are in Hill notation.

C.5 THE CACHED AUXILIARY QUANTITIES

The auxiliary losses need real-space and spectral quantities that the released dataset does not contain, so we precompute them once per geometry and cache them. Each cache record holds the level-0 atom-centred radial–Lebedev quadrature grid of PySCF (14,112 points for molecule #23470) with its weights, the reference density $\rho^\star$ on that grid, the geometry-only core-suppression factor div, the fixed random subset $\mathcal{R}$ of 256 sample points, and the atomic-orbital values at those points. Caching the orbital values is what makes the loss cheap: evaluating $N(\mathbf{r}, E)$ during training costs one eigendecomposition and two small matrix products, with no basis-function evaluation in the training loop.

Two checks confirm the cache is consistent with the released labels. Rebuilding the def2-SVP molecule from the stored geometry and recomputing the overlap integrals reproduces the released $\mathbf{S}$ to 2.2×10^{-7}, and integrating the cached density against the cached weights gives 59.978 electrons for a molecule with $N_e = 60$, the expected accuracy of a level-0 grid.

D VISUALIZING THE REFERENCE DATA

This section shows what the objects in Sections 2 and 3 actually look like. Everything plotted here is reference DFT data: no learned model enters any panel. The purpose is to make the structural claims of the sector analysis concrete, and to make the scale of the quantities the losses compare explicit.

D.1 THE HAMILTONIAN AND OVERLAP MATRICES

Figure 6 shows the two matrices for molecule #12345. Three features shape the metrics reported throughout the paper. The dynamic range is large: $\max|\mathbf{H}^{\star}_{\mu\nu}| = 19.1\,\text{Ha}$, carried by the core diagonal, against a median magnitude three orders of magnitude smaller. The matrices are strongly local: the median $|\mathbf{H}^{\star}_{\mu\nu}|$ is 3.0×10^{-2} Ha for basis functions on the same atom, 9.1×10^{-3} Ha for atoms 4–6 a_0 apart, and 1.9×10^{-5} Ha beyond 9 a_0. And the same-atom blocks dominate, which is why the diagonal-block MAE is roughly twice the off-diagonal-block MAE in every table in this paper; the split into "diag" and "non-diag" columns is a split by magnitude as much as by position.

This is also why element-wise accuracy is a weak proxy for spectral accuracy. An error budget dominated by a few large core entries can be reduced substantially while leaving the frontier of the spectrum untouched, and the element-wise-only control of Appendix G.2 does exactly that.

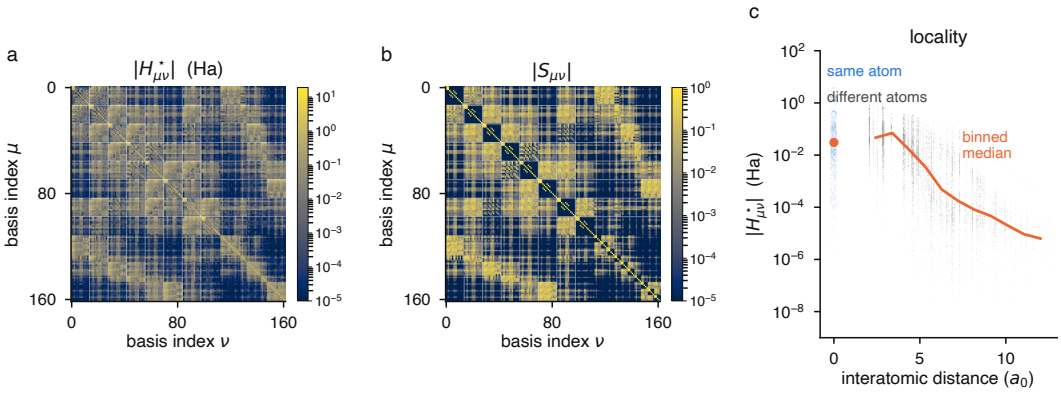


Figure 6: Anatomy of a reference Kohn–Sham Hamiltonian, molecule #12345 ($C_7H_{10}O$, $n = 162$, $n_{\text{occ}} = 30$). **(a)** $|\mathbf{H}^{\star}_{\mu\nu}|$ and **(b)** $|\mathbf{S}_{\mu\nu}|$ on a logarithmic colour scale; thin white rules mark heavy-atom block boundaries. Entries below 10^{-5} are shown at the floor of the scale. **(c)** Magnitude of the Hamiltonian elements against the distance between the atoms carrying the two basis functions, with same-atom pairs placed at zero and jittered. The orange curve is the binned median and the orange marker the same-atom median. The decay over six orders of magnitude is what local message-passing architectures exploit.

D.2 Density, frontier orbitals, and core suppression

Figure 7 shows the real-space side for molecule #23470, a planar molecule chosen so that the frontier orbitals can be drawn without projection artefacts. Both frontier orbitals are of π symmetry: their amplitude in the nuclear plane is 1.8×10^{-6}, numerically zero, so they are sampled $1\,a_0$ above it. The HOMO is the bonding π combination delocalized over the ring and both substituents; the LUMO adds a nodal plane across the ring. These are the objects that the occupied–virtual sector controls, and the reason the sector matters physically: an error in $\mathbf{\Delta}_{\mathcal{OV}}$ mixes the two pictures and moves both the gap and the frontier shapes.

Panel d explains the core-suppression factor. The reference density spans eighteen orders of magnitude, from 10^{-16} in the tails to $182\,e\,a_0^{-3}$ at the nuclei, and an unweighted squared residual would be a cusp-fitting objective. Dividing by the geometry-only factor div of Ushenin et al. (2025), which reaches 148 at the nuclei and decays to exactly one beyond about $2\,a_0$, compresses the peak by a factor of 139 and leaves the valence region untouched. Panel f records how far an uninformed start is from the converged density: the atomic-superposition guess has a whole-grid NMAE of 13%, with the error concentrated in the bonding region around 1.5–$2\,a_0$ where the atomic picture has nothing to say. Closing that gap is what buys the SCF savings in Table 12.

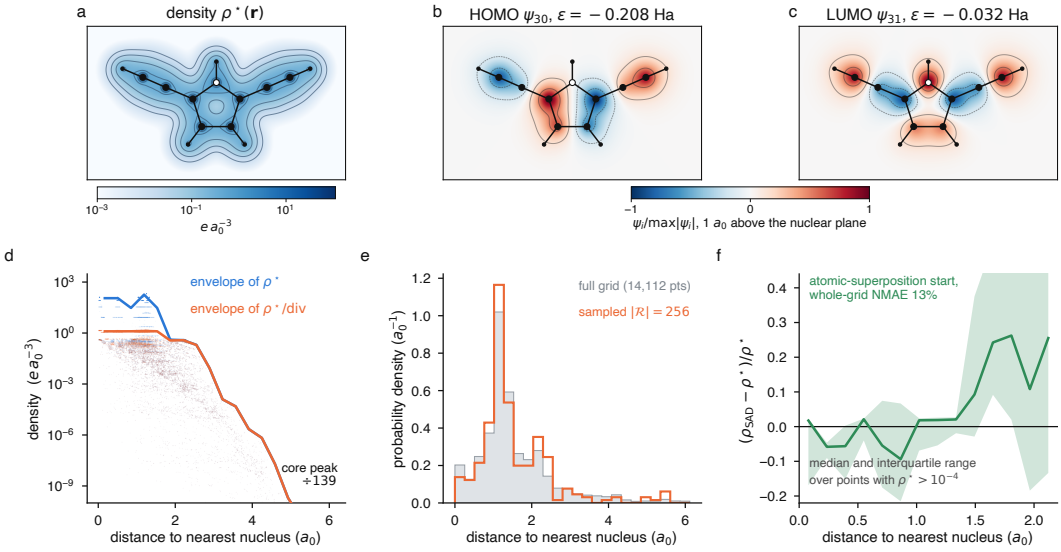


Figure 7: Real-space reference quantities for molecule #23470, 2,5-diethynyl-1H-pyrrole (C_8H_5N, $n = 151$, $n_{\mathrm{occ}} = 30$), which is exactly planar. **(a)** Electron density in the nuclear plane, logarithmic scale, with the molecular skeleton overlaid and the nitrogen drawn open. **(b,c)** HOMO and LUMO amplitudes sampled $1\,a_0$ above the nuclear plane and normalized by their own maxima; both vanish identically in the plane itself. **(d)** Upper envelope (99.5th percentile per radial bin) of $\rho^\star$ and of the core-suppressed target $\rho^\star/\mathrm{div}$ against distance to the nearest nucleus; the two coincide beyond $2\,a_0$, where $\mathrm{div} = 1$. **(e)** The 256 sampled points reproduce the radial distribution of the full 14,112-point quadrature grid. **(f)** Relative error of the atomic-superposition density against distance to the nearest nucleus.

D.3 THE SPECTRAL PICTURE AND THE FERMI-LEVEL IDENTITY

Figure 8 shows the spectral side. Panel a places the energy window in context: nine core levels sit far below it, the valence occupied levels and the low virtuals lie inside, and the window width of $1.8\,\mathrm{Ha}$ is an order of magnitude larger than the gap it straddles. Panel b shows the cumulative LDOS at five sample points. Each curve is a staircase whose risers are the orbital weights $|\psi_i(\mathbf{r})|^2$ at that point, and the curves differ by orders of magnitude between points, which is why the loss is informative: a single molecule supplies 256 independent spectral fingerprints rather than one.

Panel c is a pointwise test of Proposition B.4 on real data. Evaluating $N(\mathbf{r}, \mu)$ and $\rho^\star(\mathbf{r})/2$ independently at all 256 sample points agrees to a maximum relative deviation of 6×10^{-8}, which is the precision of the single-precision cached orbital values rather than of the identity. The cumulative loss therefore contains density matching as its $E = \mu$ slice, and the ablation between $\mathcal{L}_\rho$ and $\mathcal{L}_{\mathrm{cum}}$ in the main text compares one slice with the full profile.

Panel d shows why the extra slices help. At a fixed sample point the Gaussian-peak profile is a set of isolated bumps separated by intervals where it is numerically zero; a frontier orbital whose energy is wrong by more than a few σ falls into one of those intervals and receives no gradient, which is the saturation of Proposition B.6 and the mechanism behind the deep-orbital drift in Table 11. The staircase has no such intervals: it is monotone in E and its plateaus still record how many states lie below, so the mismatch integrated over the window grows with the displacement instead of saturating.

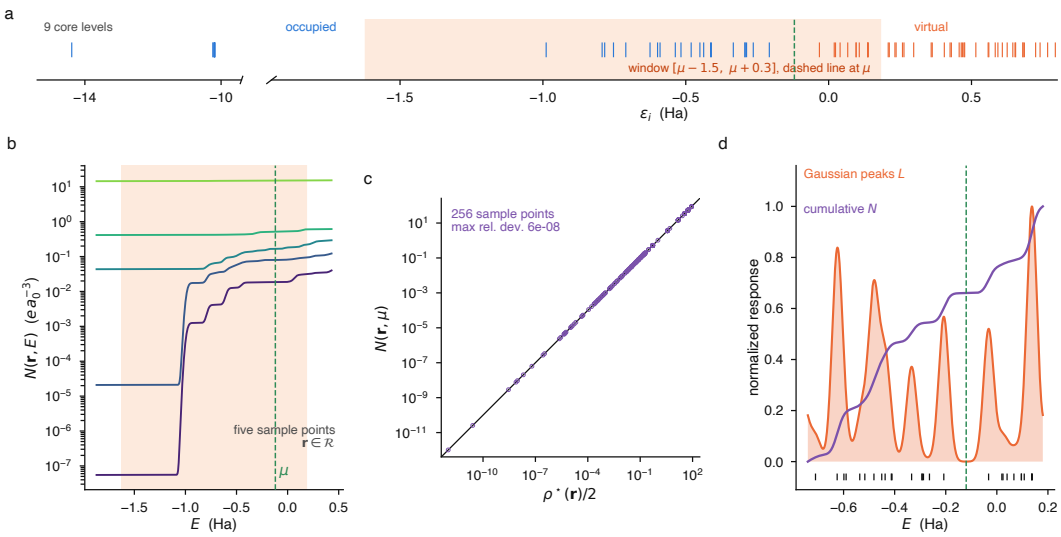


Figure 8: Spectral reference quantities for molecule #23470. **(a)** Generalized eigenvalues with the loss window shaded; the energy axis is broken so that the nine core levels are visible. **(b)** Cumulative LDOS $N(\mathbf{r}, E)$ of equation 6 at five of the 256 sample points, spanning three orders of magnitude in local density. **(c)** Proposition B.4 tested pointwise: $N(\mathbf{r}, \mu)$ against $\rho^\star(\mathbf{r})/2$ at all 256 sample points. **(d)** The two energy profiles at one sample point, with the constant contribution of the levels below the window removed and each curve normalized by its own maximum. Ticks mark the orbital energies.

D.4 WHERE A REAL RESIDUAL LIVES

The sector analysis of Section 3.1 is a statement about an arbitrary residual. It is worth asking where a *physically realized* residual actually sits, and the dynamic trajectories answer this without involving a model: taking two frames of the same molecule, $\delta\mathbf{H} = \mathbf{H}^\star(\text{frame } k) - \mathbf{H}^\star(\text{frame } 0)$ is a genuine change of the converged Kohn–Sham operator caused by a geometric displacement, and $\mathbf{\Delta} = \mathbf{C}_0^{\star\top}\delta\mathbf{H}\,\mathbf{C}_0^\star$ is its whitened form in the eigenbasis of the reference frame. It is the kind of error a model makes when it fails to track a change of geometry, which is what the dynamic splits test. Figure 9 reports the decomposition over 640 frame pairs drawn from 214 trajectories.

Three findings matter for the argument of this paper. First, the occupied–virtual sector dominates the magnitude: it carries a median of 70–84% of $\|\mathbf{\Delta}\|_F^2$ while occupying only 31% of the parameters. Density matching is therefore well aimed, which is why $\mathcal{L}_\rho$ alone already achieves a large share of the gap improvement in Table 1.

Second, the blind part is not negligible and it grows with the shift. At a displacement of 0.02 Å the two density-blind sectors hold 15.6% of $\|\mathbf{\Delta}\|_F^2$; at 0.30 Å they hold 30.3%, and essentially all of the growth is in the virtual–virtual block. That block is exactly where the frontier eigenvalue errors live: by Proposition 3.1(a) the LUMO energy error is $\Delta_{\mathrm{LUMO,LUMO}}$, a virtual–virtual entry, which Proposition 3.2 shows the density cannot resolve at any weight. The virtual–virtual share therefore grows exactly where a density-only signal becomes less informative, under geometric shift, which is the setting of the dynamic splits in Table 6.

Third, the whitening bound of Lemma B.1 is loose by a wide margin on real data. The observed amplification $\|\mathbf{\Delta}\|_F/\|\delta\mathbf{H}\|_F$ has a median of 3.0 with a 5–95% range of $[2.3, 5.2]$, while the bound $\|\mathbf{S}^{-1}\|_2$ has a median of 1878: the certificate overstates the true amplification by a median factor of 580. Such a loose bound cannot certify that a small element-wise error implies a small spectral error; the looseness reflects the conditioning of def2-SVP overlap matrices, whose condition number has a median of 1.0×10^4 across QH9-stable. This is the quantitative version of the paper's starting observation: element-wise accuracy and spectral accuracy are only weakly coupled, so the spectrum has to be supervised directly.

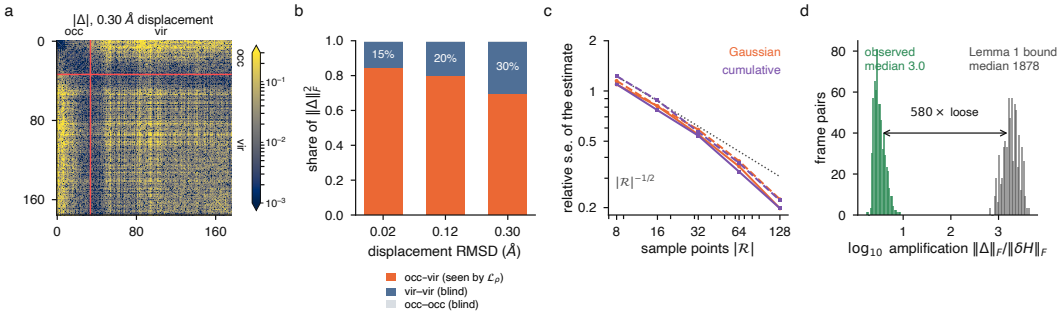


Figure 9: Sector content of a physically realized Hamiltonian change, and two numerical properties of the estimator. **(a)** $|\mathbf{\Delta}|$ for one pair of QH9-dynamic frames of the same molecule separated by 0.30 Å RMSD, in the eigenbasis of the reference frame; red rules mark the occupied/virtual boundary at $n_{\mathrm{occ}} = 34$, and values are clipped to the colour range at both ends. **(b)** Median share of $\|\mathbf{\Delta}\|_F^2$ per sector against displacement, over 640 frame pairs from 214 trajectories. The virtual–virtual share, which contains the LUMO energy error and is invisible to $\mathcal{L}_\rho$, doubles from 15% to 30% as the displacement grows. **(c)** Relative standard error of the per-molecule loss estimate when $|\mathcal{R}|$ of the 256 cached sample points are used, over 400 draws; solid lines are a random 75 μHa element-wise residual, dashed lines a 10 mHa LUMO shift, and the grey guide is the $|\mathcal{R}|^{-1/2}$ rate. **(d)** Observed amplification $\|\mathbf{\Delta}\|_F/\|\delta\mathbf{H}\|_F$ against the bound $\|\mathbf{S}^{-1}\|_2$ of Lemma B.1, same frame pairs.

E METRIC DEFINITIONS

All reported errors are test-set means over molecules, and all energies are converted to μHa. Let $\hat{\mathbf{H}}$ and $\mathbf{H}^\star$ be the predicted and reference Hamiltonians of a molecule with n basis functions and n_{occ} occupied orbitals, and let $\hat{\boldsymbol{\varepsilon}}$, $\hat{\mathbf{C}}$ and $\boldsymbol{\varepsilon}^\star$, $\mathbf{C}^\star$ solve the generalized eigenproblems $\hat{\mathbf{H}}\mathbf{c} = \varepsilon\mathbf{S}\mathbf{c}$ and $\mathbf{H}^\star\mathbf{c} = \varepsilon\mathbf{S}\mathbf{c}$ with the reference overlap $\mathbf{S}$ in both cases.

Element-wise errors. H is the mean absolute error over all matrix entries, $\frac{1}{n^2}\sum_{\mu\nu}|\hat{\mathbf{H}}_{\mu\nu} - \mathbf{H}^\star_{\mu\nu}|$. The *diag* column restricts the average to entries whose two basis functions sit on the same atom, and *non-diag* to the remaining entries; the two subsets partition the matrix, and by the locality shown in Figure 6c the diagonal subset carries the larger entries. All three are invariant to any permutation of the basis, so they are comparable across code bases that order the atomic orbitals differently.

Orbital energies. HOMO, LUMO and gap are the absolute errors of $\hat{\varepsilon}_{n_{\mathrm{occ}}}$, $\hat{\varepsilon}_{n_{\mathrm{occ}}+1}$ and their difference against the reference values. $\boldsymbol{\varepsilon}_{\mathcal{O}}$ is the mean absolute error over all occupied orbitals, $\frac{1}{n_{\mathrm{occ}}}\sum_{i\le n_{\mathrm{occ}}}|\hat{\varepsilon}_i - \varepsilon^\star_i|$, which is dominated by the deep levels and is therefore the column in which the deep-orbital drift of the peak-shaped loss (Table 11) is visible. The per-orbital table (Table 11) reports the same quantity resolved by orbital index relative to the frontier.

Orbital shapes. ψ is the mean cosine similarity between predicted and reference occupied orbitals, $\frac{1}{n_{\mathrm{occ}}}\sum_{i\le n_{\mathrm{occ}}}|\hat{\mathbf{c}}_i^\top\mathbf{S}\mathbf{c}^\star_i|$, reported as a percentage; the overlap metric is the natural inner product for the generalized problem, and the absolute value removes the sign ambiguity of an eigenvector. Values are high in absolute terms for every model, so the informative comparison is between methods within a backbone rather than against zero.

Downstream cost. The SCF columns of Table 12 are ratios against a `minao` initialization on the same molecules and the same convergence threshold: iteration count, SCF wall time, and total wall time including the forward pass of the model. Ratios below one are savings.

Comparisons across backbones. Each backbone is evaluated with its own code base. All comparisons in this paper are made within a backbone, between methods that share a checkpoint, a data split, an evaluation code and a step budget.

F THE LOSSES AS IMPLEMENTED

This section states the auxiliary terms in the form the training code evaluates them.

Whitening and eigensolve. For each selected molecule the code assembles the full predicted Hamiltonian from the model's diagonal and off-diagonal blocks, permutes it from the backbone's basis ordering into the PySCF ordering, and casts to float64 on the CPU. With the Cholesky factorization $\mathbf{S} = \mathbf{L}\mathbf{L}^\top$ it forms $\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-\top}$, calls a symmetric eigensolver, and back-transforms with $\hat{\mathbf{C}} = \mathbf{L}^{-\top}\mathbf{C}_w$, so that $\hat{\mathbf{C}}^\top\mathbf{S}\hat{\mathbf{C}} = \mathbf{I}$ holds to roundoff. Gradients flow through the Cholesky factorization and the eigensolver. The reference spectrum $\boldsymbol{\varepsilon}^\star, \mathbf{C}^\star$ is obtained by the identical code path applied to the reference blocks, which guarantees that prediction and target are compared under the same conventions and the same precision.

Sample average and core suppression. Writing $\mathcal{R}$ for the cached set of $|\mathcal{R}| = 256$ sample points and div for the geometry-only suppression factor, the density term is

$$\mathcal{L}_\rho = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}\left(\frac{\hat{\rho}(\mathbf{r}) - \rho^\star(\mathbf{r})}{\mathrm{div}(\mathbf{r})}\right)^2, \qquad \hat{\rho}(\mathbf{r}) = 2\sum_{i\le n_{\mathrm{occ}}}|\hat{\psi}_i(\mathbf{r})|^2,$$

and the two energy-resolved terms replace the single residual by a mismatch averaged over the energy grid as well,

$$\mathcal{L} = \frac{1}{|\mathcal{R}|\,K}\sum_{\mathbf{r}\in\mathcal{R}}\sum_{k=1}^{K}\left(\frac{\hat{G}(\mathbf{r},E_k) - G^\star(\mathbf{r},E_k)}{\mathrm{div}(\mathbf{r})}\right)^2, \qquad G(\mathbf{r},E) = \sum_{i=1}^{n}|\psi_i(\mathbf{r})|^2\, s\!\left(\frac{E-\varepsilon_i}{\sigma}\right),$$

with $s = \Phi$, the standard normal distribution function, for $\mathcal{L}_{\mathrm{cum}}$ and $s(x) = e^{-x^2/2}$, the unit-height Gaussian, for $\mathcal{L}_{\mathrm{LDOS}}$. The same div appears in all three terms, so the core region is suppressed identically and no term is a cusp-fitting objective. The orbital sum runs over all n orbitals, not only those inside the window.

Window anchoring. The grid is $E_k = \mu^\star - 1.5 + (k-1)\cdot 1.8/(K-1)$ with $K = 64$ and $\sigma = 0.02$ Ha, where $\mu^\star = \frac{1}{2}(\varepsilon^\star_{n_{\mathrm{occ}}} + \varepsilon^\star_{n_{\mathrm{occ}}+1})$ is the *reference* Fermi level. Anchoring the window to the reference rather than to the prediction is what keeps the objective from drifting: a model cannot reduce the loss by moving its own Fermi level out of the window, and the identity of Proposition B.4 at $E = \mu^\star$ holds against a fixed target.

Batching and cost. Eight molecules per optimizer step contribute to the auxiliary term, drawn uniformly at random from the batch; the element-wise term uses the whole batch. This subsampling is the only randomness that the auxiliary term adds, and Figure 9c bounds the resulting per-molecule estimator noise. The auxiliary term has no clipping or rescaling of its own; the gradient-norm clipping of Table 15 applies to the gradient of the total loss.

SCF acceleration protocol. For Table 12 the predicted Hamiltonian of each test molecule is used to build an initial guess: we solve the generalized eigenproblem for $\hat{\mathbf{H}}$ with the reference overlap, form the density matrix from the resulting occupied orbitals, and hand it to PySCF as the starting density for an otherwise unmodified B3LYP/def2-SVP calculation with the default DIIS (Pulay, 1980) and the default convergence threshold. The baseline is the same calculation started from `minao`. Reported ratios are over the same 50 molecules, and the total-time column includes the model forward pass and the eigensolve, so a ratio below one is a net saving after paying for inference.

G Additional Results

G.1 All metrics on the shifted splits

Table 6 lists every metric for the shifted splits summarized in Figure 2.

Table 6: Results on the shifted QH9 splits (test sets; energies in μHa, ψ in %). Same protocol as Table 1. Bold marks the best downstream metric per block.

backbone	split	method	H	diag	non-diag	HOMO	LUMO	gap	$\varepsilon_{\mathcal{O}}$	ψ
QHNet	stable-ood	pretrained	73.53	122.32	70.78	561	4632	4522	740	95.54
		+ WALoss	72.70	115.00	70.30	475	5511	5671	393	95.04
		+ cumulative LDOS (ours)	71.10	110.81	68.87	**442**	**1420**	**1506**	**290**	**96.17**
	dynamic-geo	pretrained	98.4	161.4	93.6	960	5561	6090	840	96.66
		+ WALoss	100.0	162.2	95.3	976	6950	7441	597	96.06
		+ cumulative LDOS (ours)	95.6	148.5	91.6	**771**	**1517**	**2132**	**400**	**97.85**
	dynamic-mol	pretrained	173.5	420.1	153.7	7662	68568	62262	9696	79.74
		+ WALoss	174.5	419.4	154.8	8101	74057	67285	9280	79.10
		+ cumulative LDOS (ours)	172.3	415.9	152.8	**2494**	**15105**	**14356**	**4842**	**90.57**
SPHNet	stable-ood	pretrained	30.38	38.62	29.44	171	911	1004	110	98.62
		+ WALoss	30.95	39.60	29.95	187	605	720	107	98.67
		+ cumulative LDOS (ours)	30.47	38.87	29.51	**156**	**504**	**590**	**105**	**98.70**
	dynamic-geo	pretrained	37.77	41.27	37.23	225	758	928	104	99.39
		+ WALoss	38.55	42.73	37.90	241	642	854	104	99.49
		+ cumulative LDOS (ours)	37.74	41.21	37.20	**203**	**431**	**592**	**98**	**99.53**
	dynamic-mol	pretrained	102.32	221.59	83.38	1461	13758	13433	2862	93.54
		+ WALoss	102.09	221.60	83.12	1344	10632	10466	2346	94.11
		+ cumulative LDOS (ours)	102.28	221.43	83.36	**1306**	**8528**	**8362**	**2068**	**94.56**
WANet	stable-ood	pretrained	37.05	78.99	32.28	380	1259	1055	160	98.55
		+ WALoss	34.42	49.26	32.72	176	524	577	113	98.57
		+ cumulative LDOS (ours)	34.09	48.39	32.46	**162**	**486**	**549**	**110**	**98.59**
	dynamic-geo	pretrained	41.04	46.63	40.19	269	953	1126	106	99.32
		+ WALoss	41.93	48.54	40.93	252	675	898	103	99.45
		+ cumulative LDOS (ours)	41.06	46.64	40.20	**217**	**440**	**609**	**97**	**99.51**
	dynamic-mol	pretrained	127.51	290.69	101.52	3238	26236	24126	4112	89.17
		+ WALoss	127.15	290.41	101.15	3070	23111	21184	3571	89.88
		+ cumulative LDOS (ours)	127.46	290.44	101.51	**2020**	**13756**	**12936**	**2186**	**92.47**

G.2 ELEMENT-WISE-ONLY CONTROL

Setting. Every fine-tuning run in this paper starts from the same pretrained checkpoint and trains for 30k steps with the same optimizer, learning-rate schedule, batch size and data (Table 15); the runs differ only in the auxiliary loss. The *element-wise-only* run is the control that removes the auxiliary loss altogether: it continues training the pretrained model with its own element-wise loss $\mathcal{L}_{\mathrm{elem}}$ (the mean entry-wise error between $\hat{\mathbf{H}}$ and $\mathbf{H}^{\star}$, exactly the objective used in pretraining) for the same 30k steps, so it simply keeps minimizing the Hamiltonian MAE and nothing else. It isolates the effect of the additional training itself, so that any difference between the WALoss or cumulative-LDOS runs and this control is due to the auxiliary signal and not to the extra steps. Tables 7 and 8 repeat the rows of Tables 1 and 6 with this control added, for all three backbones on stable-id and for QHNet on all four splits.

Finding. Continued element-wise training does what it is asked to do and no more. The element-wise error goes down or stays put, while the eigen-derived quantities do not improve and, on QHNet, deteriorate sharply. On QHNet the Hamiltonian MAE falls on every split (by 4.0%, 3.6%, 3.4% and 0.8% on stable-id, stable-ood, dynamic-geo and dynamic-mol), while the gap error rises by 53%, 246%, 236% and 19%, ψ falls on every split, and on dynamic-mol the occupied-orbital error rises as well. On SPHNet every metric changes by less than 1.5% (the gap error by 0.1%); on WANet the Hamiltonian MAE is unchanged while the LUMO and gap errors rise by 10% and 8%. A matrix that is more accurate entry by entry therefore does not by itself give a better spectrum. The same decoupling has been reported independently: Li et al. (2025) found that on larger molecules matrices with tiny relative errors gave unusable total energies, and Kaniselvan et al. (2026) measured only a weak correlation between matrix-element errors and eigenvalue errors across the test splits of ∇^2DFT. This is Proposition B.7 read in reverse: the element-wise loss does not distinguish the sectors of the error, so without a spectral signal the optimizer is free to move error into the sectors that the properties are sensitive to.

Two directions of generalization. Table 7 holds the dataset fixed (stable-id) and varies the model: the control behaves the same way on three architectures of very different accuracy, so the effect is a property of the training signal rather than of one network. Table 8 holds the model fixed (QHNet) and varies the data: the same pattern appears on the random, size-extrapolation and molecular-dynamics splits, so it is not tied to one distribution either. In both tables the cumulative LDOS loss, which shares every setting with the control except the auxiliary term, improves every eigen-derived metric while keeping the Hamiltonian MAE at most 0.3% above the pretrained value.

Table 7: QH9-stable-id, all three backbones (Table 1 with the element-wise-only control added; test set, μHa, ψ in %). Bold marks the best downstream metric per backbone.

backbone	method	H	diag	non-diag	HOMO	LUMO	gap	$\varepsilon_{\mathcal{O}}$	ψ
QHNet	pretrained	77.56	129.25	73.68	1041	9445	9632	1206	96.28
	+ element-wise loss only (control)	74.44	108.11	71.90	1280	14946	14769	868	95.57
	+ WALoss	78.04	119.03	74.95	847	5458	5835	570	96.79
	+ cumulative LDOS (ours)	74.81	109.54	72.19	**678**	**1482**	**2027**	**341**	**97.91**
SPHNet	pretrained	36.60	44.79	35.38	345	1309	1554	380	99.06
	+ element-wise loss only (control)	36.55	44.67	35.35	349	1311	1555	381	99.05
	+ WALoss	37.28	46.03	35.98	341	957	1259	288	99.17
	+ cumulative LDOS (ours)	36.64	44.83	35.42	**292**	**572**	**789**	**137**	**99.25**
WANet	pretrained	40.07	51.28	38.42	335	1454	1684	288	98.98
	+ element-wise loss only (control)	40.08	51.25	38.43	335	1603	1827	292	98.95
	+ WALoss	40.99	53.26	39.19	328	1025	1306	157	99.13
	+ cumulative LDOS (ours)	40.19	51.59	38.52	**287**	**573**	**791**	**137**	**99.22**

Table 8: QHNet on all four QH9 splits (the QHNet rows of Tables 1 and 6 with the element-wise-only control added; test sets, μHa, ψ in %). Bold marks the best downstream metric per split.

split	method	H	diag	non-diag	HOMO	LUMO	gap	$\varepsilon_{\mathcal{O}}$	ψ
stable-id	pretrained	77.56	129.25	73.68	1041	9445	9632	1206	96.28
	+ element-wise loss only (control)	74.44	108.11	71.90	1280	14946	14769	868	95.57
	+ WALoss	78.04	119.03	74.95	847	5458	5835	570	96.79
	+ cumulative LDOS (ours)	74.81	109.54	72.19	**678**	**1482**	**2027**	**341**	**97.91**
stable-ood	pretrained	73.53	122.32	70.78	561	4632	4522	740	95.54
	+ element-wise loss only (control)	70.86	109.25	68.70	783	15801	15638	582	94.35
	+ WALoss	72.70	115.00	70.30	475	5511	5671	393	95.04
	+ cumulative LDOS (ours)	71.10	110.81	68.87	**442**	**1420**	**1506**	**290**	**96.17**
dynamic-geo	pretrained	98.4	161.4	93.6	960	5561	6090	840	96.66
	+ element-wise loss only (control)	95.1	146.4	91.3	1503	20655	20468	770	94.51
	+ WALoss	100.0	162.2	95.3	976	6950	7441	597	96.06
	+ cumulative LDOS (ours)	95.6	148.5	91.6	**771**	**1517**	**2132**	**400**	**97.85**
dynamic-mol	pretrained	173.5	420.1	153.7	7662	68568	62262	9696	79.74
	+ element-wise loss only (control)	172.1	415.8	152.5	9145	82180	74214	11343	76.80
	+ WALoss	174.5	419.4	154.8	8101	74057	67285	9280	79.10
	+ cumulative LDOS (ours)	172.3	415.9	152.8	**2494**	**15105**	**14356**	**4842**	**90.57**

G.3 WEIGHT SENSITIVITY ON QHNET

WALoss as used in this paper. Throughout the paper, WALoss means the full form of Li et al. (2025): the prediction is expressed in the basis of reference orbitals and every entry of $\mathbf{C}^{\star\top}\hat{\mathbf{H}}\mathbf{C}^{\star} - \mathrm{diag}(\boldsymbol{\varepsilon}^{\star})$ is penalized. On SPHNet and WANet, the loss is the mean absolute plus the mean squared entry of this matrix, with weight 0.05. On QHNet, it is the mean absolute entry, evaluated on eight molecules per batch like our own loss, and we report two weights: $w = 0.03$, at which the WALoss term is about as large as the element-wise term, and $w = 0.1$, at which it dominates; both are in Table 9 and give similar gap errors (6072 and 5835 μHa).

Table 9: QHNet on QH9-stable-id: all metrics against the auxiliary weight w (test set, μHa). WALoss is the full form used throughout the paper; the weight of the cumulative loss, $w = 30$, is selected on the validation split.

method	H	diag	non-diag	HOMO	LUMO	gap	$\boldsymbol{\varepsilon}_{\mathcal{O}}$	ψ
pretrained	77.56	129.25	73.68	1041	9445	9632	1206	96.28
element-wise loss only (control)	74.44	108.11	71.90	1280	14946	14769	868	95.57
WALoss w=0.03	76.01	112.51	73.26	823	5728	6072	593	96.81
WALoss w=0.1	78.04	119.03	74.95	847	5458	5835	570	96.79
density only w=1	77.72	126.93	74.02	780	1923	2376	571	97.44
density only w=10	79.01	138.44	74.55	802	2500	2896	373	97.43
density only w=100	76.82	118.70	73.65	719	1575	2081	356	97.87
cumulative LDOS w=10	74.88	109.66	72.26	675	1693	2231	348	97.80
cumulative LDOS w=30	74.81	109.54	72.19	678	1482	2027	341	97.91
cumulative LDOS w=100	74.85	109.90	72.21	655	1300	1811	336	97.98

Figure 10 plots the density-only and cumulative rows of Table 9 and the diagonal-only rows of Table 10 against w, with the pretrained model, the control and WALoss ($w = 0.1$) as horizontal references. Across all tested weights, the density-only and cumulative losses give 2–3 times lower gap error than WALoss, so the ranking does not depend on the choice of w. The cumulative loss responds only mildly to its weight (gap error 2231, 2027 and 1811 μHa at $w = 10$, 30 and 100) and its element-wise error stays essentially constant: at the selected weight, $w = 30$, its element-wise error is within 0.6% of the lowest value among all runs, that of the control, while its gap error is 7.3 times lower.

The diagonal-only form. Li et al. (2025) also propose a second form of WALoss that keeps only the diagonal of the projected matrix, $\sum_i w_i \left(\mathbf{c}_i^{\star\top} \hat{\mathbf{H}}\, \mathbf{c}_i^{\star} - \varepsilon_i^{\star}\right)^2$, with large weights w_i on the occupied orbitals and the LUMO. Each term is the first-order estimate of one orbital energy (Proposition 3.1a), so this form sees sector A only, and only to first order. We implemented it on QHNet ($w_i = 1$ on the occupied orbitals and the LUMO, 0.01 elsewhere) to test the prediction of Sec. 3.1 that such an objective can be lowered by rotating orbitals instead of correcting their energies. Table 10 confirms it: as the weight grows, the gap and occupied-orbital errors get monotonically worse, from 12,373 to 26,813 μHa and from 797 to 1573 μHa, while ψ falls.

Table 10: QHNet on QH9-stable-id with the diagonal-only form of WALoss at three weights (test set, μHa).

method	H	diag	non-diag	HOMO	LUMO	gap	$\varepsilon_{\mathcal{O}}$	ψ
pretrained	77.56	129.25	73.68	1041	9445	9632	1206	96.28
diagonal-only WALoss w=1	74.48	108.29	71.93	1130	12504	12373	797	95.88
diagonal-only WALoss w=10	75.20	110.03	72.58	1090	14208	13701	977	95.59
diagonal-only WALoss w=100	77.84	117.44	74.86	1862	28282	26813	1573	93.66

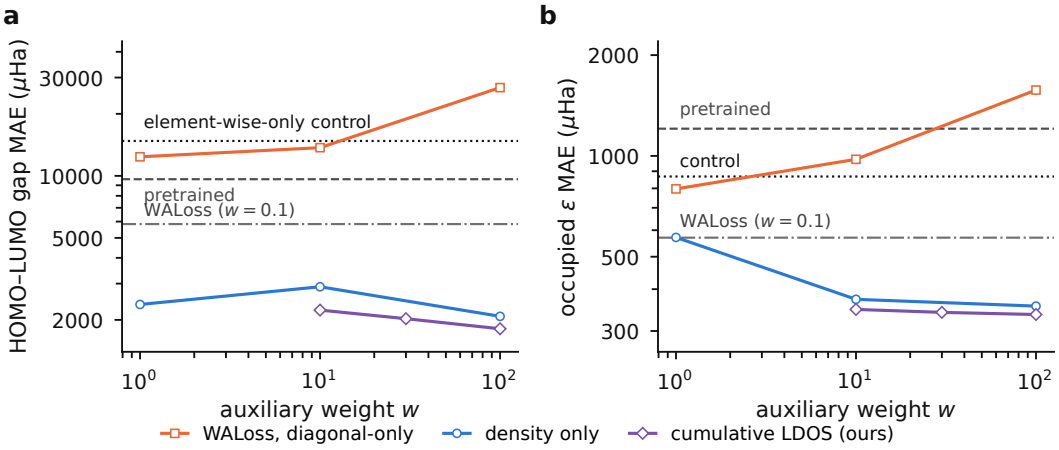


Figure 10: Weight sensitivity on QHNet (entries of Tables 9 and 10). **(a)** Gap error and **(b)** occupied-orbital energy error against the auxiliary weight w, with the pretrained checkpoint, the element-wise-only control and WALoss ($w = 0.1$) as horizontal references. The diagonal-only WALoss gets worse as its weight grows; the density-only and cumulative losses improve or stay flat, and the cumulative loss is the lowest at every weight.

G.4 PEAK-SHAPED LDOS VARIANTS

The peak-shaped loss $\mathcal{L}_{\mathrm{LDOS}}$ of Sec. 3.2 replaces the cumulative step by a unit-height Gaussian; everything else (window, σ, sample points) is shared with the cumulative loss. On SPHNet and WANet at $w = 150$ it reaches gap errors of 805 and 813 μHa (HOMO 281 and 270, LUMO 588 and 613, $\varepsilon_{\mathcal{O}}$ 235 and 140, ψ 99.22 and 99.19%), close to the cumulative loss at $w = 450$ (789 and 791 μHa), so on these two backbones the two forms perform similarly. On QHNet, whose initial orbital-energy errors are the largest, the peak-shaped loss is unreliable: two runs with identical settings end at gap errors of 6917 and 1994 μHa, and in the first run the occupied-orbital error doubles relative to the pretrained model. The three cumulative runs of Table 9, by contrast, vary smoothly with w (gap error 2231, 2027 and 1811 μHa). Per-orbital errors (Table 11) show why: the deep occupied orbitals, whose errors are largest, drift away from their targets (for HOMO$-$23, the 23rd orbital below the HOMO, the error grows from 1,049 to 20,979 μHa), whereas the cumulative loss, with the same window, σ and weight, shows no drift and gives the best $\varepsilon_{\mathcal{O}}$ and ψ of all variants at the same element-wise error.

The mechanism is Theorem 3.5(iv): once an orbital's energy error exceeds a few σ, the predicted and reference peaks no longer overlap, the mismatch saturates and its gradient vanishes (Figure 11b,c), so the orbital is no longer pulled back; the step of the cumulative form keeps a constant restoring force at any distance. This is the failure mode of overlap-based divergences between distributions with disjoint supports, whose gradients vanish while a transport (Wasserstein) distance keeps a usable one (Arjovsky et al., 2017); the cumulative form compares the local spectra through their distribution functions, as a transport distance does, and tends to their Cramér–von Mises distance as $\sigma \to 0$ (Proposition B.6). The deep occupied orbitals, whose initial errors on QHNet are the largest, are the first to be released.

Table 11: Per-orbital energy error on QHNet (1,000 test molecules, μHa) for the peak-shaped and the cumulative loss at the same weight ($w = 30$), window and σ. The peak-shaped loss corrects the virtual orbitals inside the window but lets the deep occupied orbitals drift; the cumulative loss improves every orbital.

loss	LUMO+2	LUMO+1	LUMO	HOMO	HOMO$-$5	HOMO$-$19	HOMO$-$21	HOMO$-$23
pretrained	3455	3838	9593	1005	403	635	1029	1049
peak LDOS	2287	1698	7685	1053	593	5955	11364	20979
cumulative LDOS	**1032**	**994**	**1567**	**690**	**235**	**195**	**251**	**367**

G.5 SCF ACCELERATION

A predicted Hamiltonian can also initialize a conventional DFT calculation (the protocol is in Appendix F). On 50 stable-id test molecules, every model in Table 12 saves about a third of the iterations and 13–15% of the total time relative to the `minao` guess, and the models differ little (0.671–0.674 in iterations, 0.80–0.82 in SCF time). The loss improves the eigen-properties without changing what the model offers as an SCF starting point.

Table 12: SCF acceleration on 50 QH9-stable-id test molecules: ratio of SCF iterations and of wall time relative to the `minao` initial guess (lower is better; total time includes model inference).

method	SCF iterations	SCF time	total time
pretrained	0.6732	0.8115	0.8619
element-wise loss only (control)	0.6735	0.8139	0.8655
density only ($w = 100$)	0.6710	0.8171	0.8682
cumulative LDOS ($w = 100$)	0.6713	0.8028	0.8505

G.6 TRAINING COST

With the cached orbital values, evaluating N costs one eigendecomposition and two small matrix products per molecule; QH9 matrices are at most 226×226. On QHNet the auxiliary term adds 0.72 s to each training step and no GPU memory, because the eigendecompositions run on the CPU (Table 13); over the 30k fine-tuning steps this amounts to about 6 hours, roughly 5% of the cost of pretraining QHNet for 300k steps. The loss is used only in training, so inference costs exactly what it costs for the backbone.

Table 13: Training cost of the auxiliary losses on QHNet (QH9-stable-id), measured on one NVIDIA L40S with the same code, data order and batch size (32). Seconds per step are averaged over steps 50–300 without validation; memory is the peak reserved CUDA memory. The eigendecompositions run on the CPU in double precision.

auxiliary loss	seconds/step	overhead	peak GPU memory (GB)
none (element-wise only)	1.47	–	31.2
density only	1.92	+0.45 s (+30%)	31.2
peak LDOS	2.28	+0.81 s (+55%)	31.2
cumulative LDOS	2.19	+0.72 s (+49%)	31.2

H NUMERICAL VERIFICATION

Every analytical statement in Sections 2 and 3 was checked numerically on reference data. Table 14 collects the checks and their outcomes. The Fermi-level identity holds to the single precision in which the orbital values are cached: the residual scales with the float32 unit round-off. The two perturbations used in Figure 11a are exact by construction: the energy-only shift moves the intended orbital energy by 2.000 mHa with an eigenvector overlap deviation at the level of double-precision round-off, and the shape rotation leaves every eigenvalue fixed to 2.4×10^{-13} Ha, so the sector test cleanly separates energy errors from shape errors.

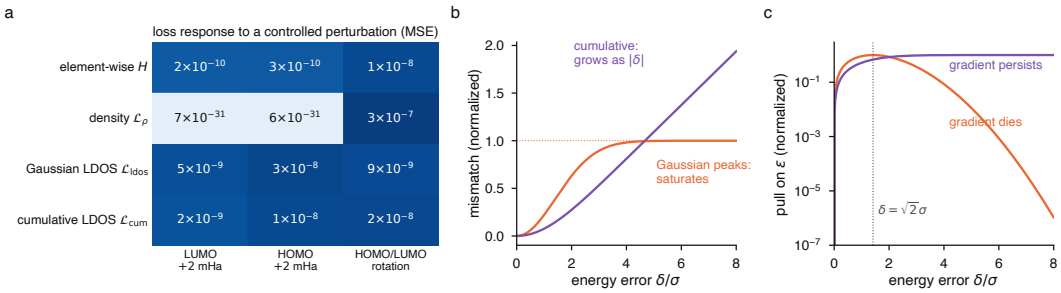


Figure 11: **(a)** Response of each loss to controlled perturbations of the reference Hamiltonian of one QH9 molecule (#12345): shifting the LUMO or HOMO energy by 2 mHa with all orbitals fixed, or rotating the HOMO and LUMO into each other with all energies fixed. Density matching does not respond to energy shifts at all (values at the level of floating-point round-off); both LDOS losses respond to both kinds of error. **(b)** Mismatch between a displaced and a reference orbital as a function of the energy error δ, for a Gaussian peak and for a cumulative step (closed forms of Theorem 3.5(iv)): the peak mismatch saturates, the step mismatch keeps growing. **(c)** The corresponding gradient with respect to δ, normalized by its maximum: the peak gradient vanishes beyond a few σ, whereas the step gradient approaches a constant.

The lower block of Table 14 checks Theorem B.5 and the statements around it on the reference Hamiltonians of three QH9-stable molecules, #18 ($n = 81$), #12345 ($n = 162$) and #23470 ($n = 151$), with orbital values evaluated on the level-0 quadrature grid. For injectivity, we form the first-order response of N to every entry of a symmetric $\mathbf{\Delta}$, one column per entry, normalize the columns, and report the smallest singular value relative to the largest. For recovery, we start from random symmetric errors with $\|\mathbf{\Delta}\|_2 = 50$ mHa on molecule #18 and minimize the mismatch in N over the full energy range, or in ρ, over $\mathbf{\Delta}$. Minimizing the mismatch in N recovers every orbital energy to 0.03 mHa and shrinks the error among orbitals inside the window about 30-fold; the residual, 20–30% of the initial $\|\mathbf{\Delta}\|_F$, lies mostly in couplings between core and virtual orbitals, the directions to which N is least sensitive. Minimizing the density mismatch instead leaves the orbital energies wrong by 10 mHa. The supplementary code contains the scripts for these checks (`tests/check.py` and `tests/verify_theory.py`).

Table 14: Numerical checks. Molecule #12345 and #23470 are the two QH9-stable geometries used in the appendix figures; the lower block also uses molecule #18. "Machine zero" means the response is at or below the double-precision noise floor of the quantity being differenced.

statement	what is checked	result	reference
generalized eigenbasis	$\|\hat{\mathbf{C}}^\top \mathbf{S}\hat{\mathbf{C}} - \mathbf{I}\|_\infty$	4.8×10^{-14}	–
cache consistency	recomputed vs. released $\mathbf{S}$	2.2×10^{-7}	App. C
cache consistency	$\int \rho^\star \, dV$ vs. $N_e = 60$	59.978	App. C
Prop. B.4	$N(\mathbf{r}, \mu)$ vs. $\rho^\star(\mathbf{r})/2$, all 256 points	6.4×10^{-8} rel. (float32 cache)	Fig. 8c
Prop. B.6	closed forms $g(\delta)$, $c(\delta)$ vs. quadrature	9.3×10^{-16} rel.	Fig. 11b
energy-only perturbation	induced $\Delta\varepsilon_i$	2.000 mHa (exact)	Fig. 11a
energy-only perturbation	eigenvector overlap deviation	$\leq 4.4 \times 10^{-16}$	Fig. 11a
shape-only perturbation	$\max_i \|\Delta\varepsilon_i\|$	2.4×10^{-13} Ha	Fig. 11a
Prop. 3.2	$\mathcal{L}_\rho$ response to a 2 mHa energy shift	6.2–6.6×10^{-31} (machine zero)	Fig. 11a
Prop. 3.1(c)	$\mathcal{L}_\rho$ response to a 0.05 rad rotation	2.7×10^{-7}	Fig. 11a
Gaussian saturation	$\|g'\|$ at $\delta = 4\sigma$, 6σ, rel. to its peak	8.6×10^{-2}, 8.7×10^{-4}	Fig. 11c
Prop. 3.1(a,b), first-order δN	exact minus first-order change, perturbation scale $t = 10^{-5}, 10^{-6}, 10^{-7}$	rel. error $\propto t$ ($3.8 \times 10^{-4} \to 3.8 \times 10^{-6}$)	Thm. B.5(i)
Prop. 3.1(c), degenerate levels	same for $\delta\rho$, two occupied and two virtual levels degenerate	rel. error $\propto t$ ($6.5 \times 10^{-5} \to 7.3 \times 10^{-7}$)	App. B
Prop. B.4 at the grid energies $\mu \pm 0.014$ Ha	$\max \|N - \rho/2\| / \max(\rho/2)$ on the grid; ratio to the bound	$\leq 8.5 \times 10^{-8}$; ≤ 0.22	App. B
Thm. B.5(i), full energy range	smallest relative singular value of the first-order response	$\geq 6.1 \times 10^{-6}$, no null direction	App. B
Thm. B.5(i), loss window ($K = 64$)	directions with an orbital inside the window / with both outside	all resolved, $\geq 2.5 \times 10^{-7}$ / below round-off	App. B
Thm. B.5(ii), recovery	max orbital-energy error after minimizing $\mathcal{L}_{\mathrm{cum}}$ (or $\mathcal{L}_\rho$), three random starts	0.03 mHa ($\mathcal{L}_\rho$: 10 mHa)	App. B
Davis–Kahan bound	$\|\sin\Theta\|_F$ over the bound, 600 random $\mathbf{\Delta}$ with $\|\mathbf{\Delta}\|_2 < g/2$	≤ 0.22	App. B
restoring force on the $K = 64$ grid	$\|c' - \mathrm{erf}(\delta/2\sigma)\|$, $\|\delta\| \leq 0.1$ Ha, levels ≥ 0.18 Ha inside	$\leq 7.2 \times 10^{-3}$	Prop. B.6
gradient of $\mathcal{L}_{\mathrm{cum}}$	Daleckii–Krein form vs. central differences	1.9×10^{-9} rel.	App. B
two p shells on one atom	$\max \|p_{1x}p_{2y} - p_{1y}p_{2x}\| / \max \|p_{1x}p_{2y}\|$	2.9×10^{-16} (identically zero)	App. B

I LIMITATIONS AND FUTURE WORK

The analysis is exact within its assumptions: first-order perturbation theory in the basis of reference orbitals, with non-degenerate reference energies for the statements about orbital energies and shapes and an open HOMO–LUMO gap for the statement about the density. Under these assumptions the sector picture is a change of basis rather than an approximation. What the analysis does not do is predict how a given network moves along the sectors during training, so the sector picture is best read as a diagnostic rather than a guarantee.

The loss has been tested where Hamiltonian prediction is currently benchmarked: on QH9 (B3LYP/def2-SVP), with three architectures that span the development of the field, on all four official splits, and against the strongest published physics-informed loss. Its ingredients, the predicted Hamiltonian, the overlap matrix and orbital values at a few hundred points, exist for any basis set, functional or system, so the same construction should carry over to larger molecules, other levels of theory and periodic systems, where the LDOS is resolved per k-point and is the quantity that scanning tunnelling microscopy measures; these are the settings in which we plan to test it next. Two design choices should be noted: the energy window covers the valence and low virtual orbitals, the right choice for ground-state properties, and would be widened for excited-state targets; and the smoothing width σ was fixed rather than tuned.

The cost is one eigendecomposition per sampled molecule and training step, the same operation that reads any property from a predicted Hamiltonian; on QH9 it adds 0.72 s per step on QHNet (Appendix G.6), and inference is untouched. At the scale of thousands of basis functions the dense solver would give way to the iterative and fragment-based solvers that DFT codes already use.

Two directions follow from the analysis itself. The cumulative LDOS can be used from the start of training rather than as a fine-tuning signal, where its non-vanishing gradients should matter most while the spectrum is still far from the reference. And the sector representation can serve as a diagnostic in its own right: computed on a validation set, it tells which part of the residual a model still carries and therefore which training signal will help, turning the choice of loss from trial and error into a measurement.

J Implementation Details and Protocol

Sample points and core suppression. Each molecule's sample set $\mathcal{R}$ is a fixed random subset of 256 points of the atom-centred radial–Lebedev quadrature grid of PySCF at level 0, cached once. The core-suppression factor $\mathrm{div}(\mathbf{r})$ is the element-wise fit of Ushenin et al. (2025), which depends only on the geometry.

Eigensolve. For each of the eight molecules per step we form $\mathbf{L}^{-1}\hat{\mathbf{H}}\mathbf{L}^{-\top}$ with $\mathbf{S} = \mathbf{L}\mathbf{L}^{\top}$, call a symmetric float64 eigensolver on the CPU, and back-transform; gradients flow through the eigensolver.

Energy grids and weights. $K = 64$ points on $[\mu - 1.5, \mu + 0.3]$ Ha with $\sigma = 0.02$ Ha for all backbones. The loss weights are listed in Table 15. The sum in equation 6 runs over all orbitals, so orbitals below the window contribute a constant step and $N(\mathbf{r}, E) = \rho/2$ holds at every energy inside the gap, up to the bound of Proposition B.4.

Evaluation. Element-wise MAE is basis-permutation invariant. Eigen-derived metrics are computed from the generalized eigenproblem of $\hat{\mathbf{H}}$ with the DFT overlap $\mathbf{S}$. All comparisons are made within a backbone using the backbone's own evaluator and matched settings. For SPHNet and WANet, whose evaluator runs on the GPU, TF32 matrix multiplication is disabled and the generalized eigenproblem is solved in float64, matching the CPU float64 evaluation of QHNet.

Optimization. Table 15 lists the settings. Every fine-tuning run of a given backbone starts from the same released checkpoint and differs only in the auxiliary term and its weight; the optimizer, learning rate, schedule, batch size, gradient clipping, step budget and data split are held fixed across runs.

Table 15: Fine-tuning settings. Every method of a given backbone shares all settings except the auxiliary loss and its weight; unlisted settings follow each backbone's released training recipe ("native"). The cumulative LDOS weight is selected on the validation split; Table 9 reports its test-set sensitivity on QHNet.

	QHNet	SPHNet / WANet
optimizer	AdamW, $\beta = (0.99, 0.999)$	native
learning rate	1×10^{-5}	1×10^{-5}
warm-up steps	1,000	native
fine-tuning steps	30,000	30,000
batch size	32	64
gradient clipping	norm 5.0	norm 5.0
molecules per auxiliary step	8	8
energy grid	$K = 64$, $\sigma = 0.02$ Ha	$K = 64$, $\sigma = 0.02$ Ha
cumulative LDOS weight w	30 (stable-id), 100 (other splits)	450
WALoss weight	0.1	0.05

Verification scripts. A sensitivity script perturbs the reference Hamiltonian of one molecule by an eigenvalue-only shift $\mathbf{H}^{\star} + d\,\mathbf{S}\mathbf{c}_i\mathbf{c}_i^{\top}\mathbf{S}$ (which moves ε_i by d and no eigenvector) and by a rotation in the HOMO/LUMO plane, and evaluates each loss (Figure 11a). A second script checks the identity of Prop. B.4, the closed forms of Prop. B.6 against numerical quadrature, and the sector sensitivities of both LDOS forms. Appendix H reports the outcome of every check.