SPECTRALLY CONSTRAINED ELECTRONIC DENSITY MATRIX LEARNING WITH SUBSPACE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning electronic-structure matrices from density functional theory (DFT) has emerged as an active area of atomistic machine learning, providing electronic information beyond the energies and forces typically targeted by machine learning interatomic potentials (MLIPs). Machine-learned Hamiltonian models (MLHs) and density-matrix models (MLDMs) are two representative approaches. MLDMs avoid the numerically sensitive conversion from predicted Hamiltonians to density matrices, but often still underperform MLHs in energy evaluation. We posit that this performance gap arises in part from how density-matrix learning treats eigenvalues and eigenspaces. We introduce a spectrally constrained framework combining Cayley rotations with an occupied–virtual (OV) loss. Cayley rotations preserve the eigenvalues of a valid initial density in an orthonormal representation, leaving the eigenspaces as the remaining degrees of freedom. The OV loss targets the first-order density errors induced by small rotations away from the target eigenspaces. On QH9, our framework outperforms QHFlow2, a state-of-the-art Hamiltonian model, in energy and real-space density accuracy, achieving a mean one-shot energy error of 0.184 meV with no errors above 100 meV. The predicted densities also accelerate self-consistent field (SCF) calculations by reducing effective iteration cost. In our scaling comparison, a 65.9M-parameter Cayley model with OV loss achieves energy accuracy comparable to a 514M-parameter model without it. Our framework generalizes to larger QM40 molecules without fine-tuning, achieving a mean energy error of 0.98 meV/atom and improved SCF initialization.

1 INTRODUCTION

Machine learning in materials science and chemistry has largely focused on total energies and atomic forces computed within density functional theory (DFT) (Kohn & Sham, 1965), leading to the class of machine-learned interatomic potentials (MLIPs) (Unke et al., 2021b; Fu et al., 2025; Wood et al., 2025). More recently, many of these models have been adapted to learn intermediate DFT quantities containing rich electronic-level information in a compact atomic-orbital (AO) basis. Such models come in two main flavors. Machine learned Hamiltonian models (MLHs) predict Hamiltonian ($\mathbf{H}$) matrices, which contain all the electronic orbitals in the system of interest (Schütt et al., 2019; Unke et al., 2021a; Li et al., 2022; Yu et al., 2023b; Kim et al., 2025; Yu et al., 2025; Li et al., 2025b; Kaniselvan et al., 2026c; Kim et al., 2026; Kaniselvan et al., 2026b). Machine-learned density matrix models (MLDMs) predict one-electron density matrices ($\mathbf{D}$), which are built from the occupied electronic states (Shao et al., 2023; Hazra et al., 2024; Tang et al., 2024; Zhang et al., 2026; Yescas-Ramos et al., 2026). As summarized in Figure 1(a), predicted electronic-structure matrices can be used for real-space density reconstruction, one-shot DFT energy and force evaluation, and initialization of self-consistent field (SCF) calculations (Hazra et al., 2024; Rana et al., 2025; Kim et al., 2026; Liu et al., 2026).

For one-shot energy evaluations, MLDMs offer practical advantages over MLHs. Evaluating energies from a predicted $\mathbf{H}$ involves first building $\mathbf{D}$ analytically, which requires (1) solving a generalized eigenvalue problem and (2) combining the outer products of the lowest-energy eigenvectors, selected according to the electron count. During this conversion, small errors in the predicted $\mathbf{H}$ can be amplified by the eigensolver, altering the set of lowest eigenvalues and leading to substantial errors in $\mathbf{D}$ and downstream quantities (Li et al., 2025b; Kaniselvan et al., 2026c). By learning $\mathbf{D}$ directly, MLDMs naturally avoid error accumulation in both the eigendecomposition and orbital selection.

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.

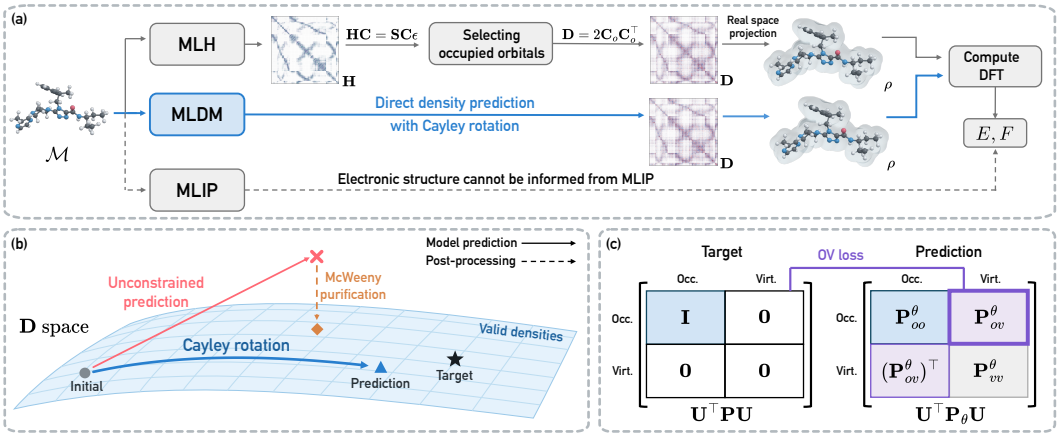


Figure 1: **Overview of spectrally constrained density-matrix learning with subspace loss.** (a) MLH and MLDM models predict $\mathbf{H}$ and $\mathbf{D}$, respectively. MLIPs typically predict energies and forces without explicitly predicting electronic structure. (b) Cayley rotations preserve the electron number, occupation spectrum, and idempotency of a valid initial density matrix. McWeeny purification reduces idempotency errors. (c) With these constraints enforced by the Cayley rotation, occupied–virtual (OV) loss additionally penalizes predicted mixing between target occupied and unoccupied states.

Despite this advantage, existing MLDMs have been reported to underperform MLHs in downstream calculations (Kim et al., 2026; Liu et al., 2026). We posit that this discrepancy arises from how density-matrix learning treats eigenvalues and eigenvectors. In $\mathbf{P}$, the normalized orthonormal representation of $\mathbf{D}$, eigenvectors describe electronic states, while eigenvalues specify their occupations. A valid $\mathbf{P}$ has $N_e/2$ eigenvalues equal to 1 (occupied) and all remaining eigenvalues equal to 0 (unoccupied), where N_e is the electron count. Therefore, fixing this spectrum leaves the eigenspaces as the remaining degrees of freedom. Standard entrywise losses, however, neither enforce this spectrum nor give explicit priority to aligning these eigenspaces with the target. This motivates preserving valid occupations by construction while learning the remaining eigenspace degrees of freedom.

Existing approaches partially address these challenges. dm-PhiSNet combines physically motivated training objectives with McWeeny purification (Yescas-Ramos et al., 2026; McWeeny, 1960). Purification promotes valid eigenvalues, but leaves eigenspace misalignment unresolved. Solver-aligned initialization learning (SAIL) addresses subspace supervision through energy gradients with respect to occupied–virtual rotations along SCF trajectories, improving SCF initialization (Eberhard et al., 2026). However, it does not enforce valid occupations on the direct density-matrix predictions and incurs additional cost from repeated solver evaluations and backpropagation through SCF iterations.

To this end, we introduce *a spectrally constrained density-matrix learning framework combining Cayley rotations with a target-basis occupied–virtual (OV) loss*. Cayley rotations transform a valid initial density while preserving its electron number and occupation spectrum by construction (Figure 1(b)). The OV loss supplements the matrix loss by targeting occupied-subspace misalignment, which appears to first order error in the target-basis when occupations are fixed (Figure 1(c)). The objective uses fixed target densities and requires no DFT evaluations during training. Our scaling experiments show comparable energy accuracy with $7.8\times$ fewer parameters when using OV loss.

We further investigate failure modes in converting predicted $\mathbf{H}$ into $\mathbf{D}$ for energy evaluation. We also revisit equivariance under spatial inversion (*i.e.*, reversing all coordinates about the molecular center) in MLDMs, identifying failures in models that do not enforce it. In our scaling comparisons, models enforcing this symmetry achieve comparable accuracy with roughly $2\times$ fewer parameters.

We evaluate density-matrix and real-space density accuracy, one-shot energies, and SCF acceleration on QH9 (Yu et al., 2023a) and QM40 (Madushanka et al., 2024). On QH9, our framework achieves a mean energy error of 0.184 meV, with lower mean errors and fewer large outliers than the evaluated MLH baselines. Its mean energy error remains lower than that of QHFlow2, a state-of-the-art MLH, even after excluding QHFlow2's errors above 1 eV. On larger QM40 molecules, it achieves 0.98 meV/atom without fine-tuning. Predicted densities also reduce effective SCF iteration cost by 46.6% on the QH9 evaluation subset and 30.0% on QM40 relative to conventional density initialization, demonstrating generalization in both energy accuracy and SCF acceleration.

2 BACKGROUND: THE DFT DENSITY MATRIX IN AN ATOMIC-ORBITAL BASIS

In Kohn-Sham DFT, molecular orbitals $\psi_i(\mathbf{r})$ are commonly represented in a finite atom-centered atomic-orbital (AO) basis $\{\chi_\mu\}_{\mu=1}^{n_{\mathrm{AO}}}$ (Slater, 1930; Boys, 1950; Szabo & Ostlund, 1996). The coefficients $C_{\mu i}$ in $\psi_i(\mathbf{r}) = \sum_\mu C_{\mu i}\chi_\mu(\mathbf{r})$ form the orbital coefficient matrix $\mathbf{C}$, whose ith column is denoted by $\mathbf{c}_i$. The AO basis is typically non-orthogonal, with overlap matrix $\mathbf{S}$ defined by $S_{\mu\nu} = \langle \chi_\mu \mid \chi_\nu \rangle$. Given a Hamiltonian $\mathbf{H}$ in this basis, the orbital coefficients $\mathbf{C}$ obtained by solving the generalized eigenvalue problem

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

where $\boldsymbol{\varepsilon}$ is the diagonal matrix of orbital energies. The eigenvectors are chosen to satisfy $\mathbf{C}^\top\mathbf{SC} = \mathbf{I}$, ensuring that the molecular orbitals are orthonormal. *i.e.* $\langle \psi_i \mid \psi_j \rangle = \mathbf{c}_i^\top \mathbf{S}\mathbf{c}_j = \delta_{ij}$.

The AO density matrix $\mathbf{D}$ is defined in terms of the orbital coefficient vectors and their occupation numbers as

$$\mathbf{D} = \sum_i f_i \mathbf{c}_i \mathbf{c}_i^\top = \mathbf{C}\,\mathrm{diag}(f_1, \ldots, f_{n_{\mathrm{AO}}})\mathbf{C}^\top, \tag{2}$$

where f_i is the occupation number of orbital ψ_i, with $\sum_i f_i = N_e$ for a system with N_e electrons. In the closed-shell ground states systems, the $n_{\mathrm{occ}} = N_e/2$ orbitals with the lowest-energy orbitals are occupied, with $f_i = 2$ for these orbitals and $f_i = 0$ otherwise. The corresponding real-space electron density is obtained as $\rho(\mathbf{r}) = \sum_i f_i |\psi_i(\mathbf{r})|^2 = \sum_{\mu\nu} \chi_\mu(\mathbf{r}) D_{\mu\nu} \chi_\nu(\mathbf{r})$.

Because $\mathbf{H}$ depends on $\mathbf{D}$, the density matrix and Hamiltonian must be determined self-consistently. The self-consistent field (SCF) method starts from an initial density matrix and repeats the cycle $\mathbf{D} \to \mathbf{H}[\mathbf{D}] \to \mathbf{C} \to \mathbf{D}$ until the $\mathbf{D}$ used to construct $\mathbf{H}$ agrees, within a chosen tolerance, with the $\mathbf{D}$ reconstructed from its occupied orbitals (Kohn & Sham, 1965; Szabo & Ostlund, 1996). Further background on DFT and the SCF procedure is provided in Sections B.1 to B.2.

Occupations and orbital subspaces in density matrix. Since the AO basis is nonorthogonal, the ordinary eigenvalues of $\mathbf{D}$ do not directly represent orbital occupations. We therefore transform $\mathbf{D}$ to orthonormal coordinates using Löwdin symmetric orthogonalization (Löwdin, 1950) and define $\mathbf{P} = \frac{1}{2}\mathbf{S}^{1/2}\mathbf{D}\mathbf{S}^{1/2}$. The eigenvalues of $\mathbf{P}$ are the normalized occupations $f_i/2$, equal to 1 for *occupied orbitals* and 0 for *virtual (unoccupied) orbitals*. Because its eigenvalues are zero or one, a valid $\mathbf{P}$ satisfies $\mathbf{P}^2 = \mathbf{P}$ (or $\mathbf{DSD} = 2\mathbf{D}$). This property, called *idempotency*, means that applying $\mathbf{P}$ twice has the same effect as applying it once. Its trace, $\mathrm{Tr}(\mathbf{P}) = n_{\mathrm{occ}}$, counts the occupied orbitals.

Let $\mathbf{U}_o$ and $\mathbf{U}_v$ contain orthonormal eigenvectors spanning the occupied and virtual subspaces, respectively. In the combined basis $\mathbf{U} = [\mathbf{U}_o\ \mathbf{U}_v]$, the projector takes the block-diagonal form

$$\mathbf{U}^\top\mathbf{P}\mathbf{U} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{3}$$

The complementary projector onto the virtual subspace is $\mathbf{Q} = \mathbf{I} - \mathbf{P} = \mathbf{U}_v\mathbf{U}_v^\top$, with $\mathbf{Q}^2 = \mathbf{Q}$ and $\mathbf{PQ} = \mathbf{0}$. For a fixed projector, the basis within either subspace is not unique. Let $\mathbf{R}_o$ and $\mathbf{R}_v$ be orthogonal matrices defining rotated bases $\mathbf{U}'_o = \mathbf{U}_o\mathbf{R}_o$ and $\mathbf{U}'_v = \mathbf{U}_v\mathbf{R}_v$. The corresponding projectors remain unchanged

$$\mathbf{P}' = \mathbf{U}'_o(\mathbf{U}'_o)^\top = \mathbf{U}_o\mathbf{R}_o\mathbf{R}_o^\top\mathbf{U}_o^\top = \mathbf{P},$$
$$\mathbf{Q}' = \mathbf{U}'_v(\mathbf{U}'_v)^\top = \mathbf{U}_v\mathbf{R}_v\mathbf{R}_v^\top\mathbf{U}_v^\top = \mathbf{Q}.$$

These rotations change only the choice of basis within each subspace, not the subspaces themselves, and therefore leave the $\mathbf{P}$ and $\mathbf{Q}$ unchanged. In contrast, mixing occupied and virtual states changes the occupied subspace and therefore changes $\mathbf{P}$.

3 METHODS: DENSITY-MATRIX LEARNING BY SUBSPACE ROTATION

Our model takes the molecular geometry $\mathcal{M} = \{Z_i, \mathbf{r}_i\}_i$, the AO overlap matrix $\mathbf{S}$, and a MINAO-derived initial density $\mathbf{D}_0$ as inputs.[1] Given these inputs, we predict the converged DFT target

[1] Details of the initial density are given in Section C.1.

AO density matrix $\mathbf{D}$ as $\mathbf{D}_\theta = f_\theta(\mathcal{M}, \mathbf{D}_0, \mathbf{S})$. We implement f_θ using an equivariant AO-matrix backbone based on QHFlow2, HELM, MALOQ, and eSEN (Kim et al., 2026; Kaniselvan et al., 2026c;b; Fu et al., 2025). The backbone produces a symmetric matrix $\mathbf{M}_\theta^{\mathrm{AO}}$. The model architectures are detailed in Section C.3.

Here, rather than treating $\mathbf{M}_\theta^{\mathrm{AO}}$ as a density prediction, we use it to construct a skew-symmetric generator. We use *Cayley rotations* to transform the occupied subspace of a valid initial density while preserving its electron number, occupation spectrum, and idempotency. An additional *OV loss* penalizes the predicted projector's occupied-virtual block in the target basis, complementing density-matrix supervision without solver evaluations during training.

Cayley rotation. A Cayley transform maps a skew-symmetric matrix to an orthogonal matrix (Li et al., 2020; Cartier & Giesbertz, 2025). We first represent the initial density as $\mathbf{P}_0 = \frac{1}{2}\mathbf{S}^{1/2}\mathbf{D}_0\mathbf{S}^{1/2}$ in orthonormal coordinates. For a skew-symmetric generator $\mathbf{K}_\theta$ and its rescaled form $\widetilde{\mathbf{K}}_\theta$ (Section C.2), the Cayley rotation and density prediction are

$$\mathbf{R}_\theta = \left(\mathbf{I} - \frac{1}{2}\widetilde{\mathbf{K}}_\theta\right)^{-1}\left(\mathbf{I} + \frac{1}{2}\widetilde{\mathbf{K}}_\theta\right), \qquad \mathbf{P}_\theta = \mathbf{R}_\theta\mathbf{P}_0\mathbf{R}_\theta^\top. \tag{4}$$

This map is the $[1/1]$ Padé approximant to $\exp(\widetilde{\mathbf{K}}_\theta)$ and preserves orthogonality exactly for skew-symmetric generators in exact arithmetic (Shepard et al., 2015). We construct $\mathbf{K}_\theta$ from the backbone output $\mathbf{M}_\theta^{\mathrm{AO}}$. We first express this output as $\mathbf{G}_\theta = \frac{1}{2}\mathbf{S}^{1/2}\mathbf{M}_\theta^{\mathrm{AO}}\mathbf{S}^{1/2}$, and define $\mathbf{K}_\theta = \mathbf{G}_\theta\mathbf{P}_0 - \mathbf{P}_0\mathbf{G}_\theta$. The resulting $\mathbf{K}_\theta$ is skew-symmetric, *i.e.*, $\mathbf{K}_\theta^\top = -\mathbf{K}_\theta$. This commutator also removes transformations that do not change $\mathbf{P}_0$. The rescaling preserves skew-symmetry.

Since $\mathbf{R}_\theta$ is orthogonal, $\mathbf{P}_\theta$ preserves the occupation spectrum of $\mathbf{P}_0$. Transforming back to the AO representation gives $\mathbf{D}_\theta = 2\mathbf{S}^{-1/2}\mathbf{P}_\theta\mathbf{S}^{-1/2}$, which preserves the electron number, $\mathrm{Tr}(\mathbf{D}_\theta\mathbf{S}) = N_e$, and satisfies idempotency, $\mathbf{D}_\theta\mathbf{S}\mathbf{D}_\theta = 2\mathbf{D}_\theta$. If its AO inputs and learned AO output are rotation equivariant, the Cayley head preserves equivariance (Section C.2.1).

Target-basis occupied–virtual loss. The Cayley rotation preserves the occupation spectrum, leaving the occupied subspace to be aligned with the target. Although the AO density-matrix loss supervises the full prediction, its entrywise reduction does not explicitly prioritize this alignment. We therefore add a target-basis occupied–virtual (OV) penalty to place explicit weight on the leading local error induced by rotations of the occupied subspace.

Let $\mathbf{U}_o$ and $\mathbf{U}_v$ contain orthonormal eigenvectors spanning the occupied and virtual subspaces of the target projector $\mathbf{P}$, respectively. In the basis $\mathbf{U} = [\mathbf{U}_o\ \mathbf{U}_v]$, the predicted projector has the block representation

$$\mathbf{U}^\top\mathbf{P}_\theta\mathbf{U} = \begin{bmatrix} \mathbf{P}_{oo}^\theta & \mathbf{P}_{ov}^\theta \\ (\mathbf{P}_{ov}^\theta)^\top & \mathbf{P}_{vv}^\theta \end{bmatrix}. \tag{5}$$

Since the target OV block is zero, $\mathbf{U}_o^\top\mathbf{P}\mathbf{U}_v = \mathbf{0}$, the predicted OV block directly measures the corresponding component of the prediction error, $\mathbf{P}_{ov}^\theta = \mathbf{U}_o^\top\mathbf{P}_\theta\mathbf{U}_v = \mathbf{U}_o^\top(\mathbf{P}_\theta - \mathbf{P})\mathbf{U}_v$. Near the target projector, small rotations of the occupied subspace produce first-order errors in the OV and VO blocks, whereas the OO and VV errors are second order (Section C.4.1).[2] The OV penalty therefore allows this leading local rotation error to be weighted explicitly alongside the AO entrywise error. However, a zero OV block does not guarantee identity to target: $\mathbf{P}_{ov}^\theta = \mathbf{0}$ may hold even when $\mathbf{D}_\theta \neq \mathbf{D}$. We therefore retain the full $\mathbf{D}$ loss and use the OV loss as an auxiliary term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{DM}}(\mathbf{D}_\theta, \mathbf{D}) + w_{\mathrm{OV}}\mathcal{L}_{\mathrm{OV}}(\mathbf{P}_{ov}^\theta). \tag{6}$$

Here, $\mathcal{L}_{\mathrm{DM}}$ supervises the full AO density matrix, $\mathcal{L}_{\mathrm{OV}}$ averages an elementwise penalty over the target-basis OV entries, and w_{OV} controls the relative weight of the OV loss.

The target bases are constructed from the fixed training densities and overlap matrices using only matrix operations. Evaluating the OV loss therefore requires no additional DFT solver evaluations. Exact loss definitions and implementation details are given in Section C.4.

[2] For a local rotation $\mathbf{P}_\theta = \mathbf{R}_\delta\mathbf{P}\mathbf{R}_\delta^\top$, let $\varepsilon = \|\mathbf{R}_\delta - \mathbf{I}\|_2 \ll 1$. $\mathbf{P}_{ov}^\theta = O(\varepsilon)$, while $\mathbf{P}_{oo}^\theta - \mathbf{I}$ and $\mathbf{P}_{vv}^\theta$ are $O(\varepsilon^2)$.

Inversion equivariance by group averaging. We enforce inversion equivariance by evaluating the same backbone on the original and inverted inputs (Puny et al., 2022). Let $\mathcal{X}$ denote the full input and $\mathcal{I}\mathcal{X}$ its inverted counterpart, with the geometry and all AO-matrix inputs transformed consistently. The inverted branch is mapped back to the original AO frame before averaging:

$$\mathbf{M}_\theta^{\mathrm{inv}}(\mathcal{X}) = \mathcal{I}_{\mathrm{AO}}^{-1}[\mathbf{M}_\theta^{\mathrm{raw}}(\mathcal{I}\mathcal{X})], \quad \mathbf{M}_\theta^{\mathrm{AO}}(\mathcal{X}) = \frac{1}{2}\left[\mathbf{M}_\theta^{\mathrm{raw}}(\mathcal{X}) + \mathbf{M}_\theta^{\mathrm{inv}}(\mathcal{X})\right]. \tag{7}$$

Here, $\mathcal{I}_{\mathrm{AO}}^{-1}$ transforms the AO-matrix output back to the original frame. The averaged matrix is passed to the Cayley head. Implementations are given in Section C.5.

4 Experiments

Datasets and tasks. We train our models on QH9 (Yu et al., 2023a) and use its full 13,084-molecule test split for method comparisons and ablation studies. To assess molecule size extrapolation, we evaluate the QH9-trained models without fine-tuning on 2,243 QM40 (Madushanka et al., 2024) molecules with 10-37 heavy atoms and elements drawn from the QH9 (H,C,N,O,F). We also test their use in one-shot DFT evaluation and as initial guesses for SCF. The SCF evaluation uses a fixed 200 QH9 subset and all 2,243 QM40 (Sections D.1 and D.2). Both one-shot and SCF evaluations use B3LYP5/def2-SVP, matching the level of theory used to generate QH9 (Section D.1).

Models. Unless otherwise specified, the QH9 comparisons use 65.9M-parameter models. The $\Delta\mathbf{D}$ baseline predicts a correction to the initial density $\mathbf{D}_0$, giving $\mathbf{D}_\theta = \mathbf{D}_0 + \Delta\mathbf{D}$. Both density models share an $\mathrm{O}(3)$-equivariant backbone and a minimal atomic-orbital (MINAO)-derived initial density (Almlöf et al., 1982; Van Lenthe et al., 2006; Sun et al., 2018) (Section C.1). The controlled MLH uses this backbone with an initial Hamiltonian. We also evaluate the 20.4M-parameter QH-Net (Yu et al., 2023b) and the pretrained 43M QHFlow2-m model (Kim et al., 2026). Training configurations are detailed in Section D.1. All models are trained in FP32. Cayley evaluation uses FP64 outputs in Sections 4.1 and 4.2, and FP32 predictions with McW2 for Section 4.3. See Section E.1 for precision effects.

Evaluation metrics. On QH9 and QM40, we report entrywise density-matrix MAE and electron-count error $|\,\mathrm{Tr}(\widehat{\mathbf{D}}\mathbf{S}) - N_e|$ (Trace), where $\widehat{\mathbf{D}}$ is the evaluated density after any stated post-processing. Idempotency error (Idemp.) is the entrywise MAE of $\widehat{\mathbf{D}}\mathbf{S}\widehat{\mathbf{D}} - 2\widehat{\mathbf{D}}$. We also report target-basis OO/OV Frobenius errors ($\|\Delta\mathbf{P}_{oo}\|_F$ and $\|\Delta\mathbf{P}_{ov}\|_F$), the normalized integrated real-space density error (ρ_{NMAE}), and one-shot energy and force errors. On the QH9 SCF subset, we report the effective relative iteration count (ERIC) (Eberhard et al., 2026), normalized to MINAO initialization. Definitions are in Section D.3.

4.1 Occupation and Subspace Errors

We first probe the two error components in $\Delta\mathbf{D}$ predictions. 'McWeeny purification' can be applied as an analytical post-processing step to reduce occupation (eigenvalue) errors (McWeeny, 1960; Yescas-Ramos et al., 2026). Essentially, for predictions whose eigenvalues already lie near 0 or 1, it polarizes the spectrum toward the corresponding integer occupations while preserving the eigenvectors. This effect is visible for a representative molecule in Figure 2. Further details of the purification procedure are given in Section D.4.

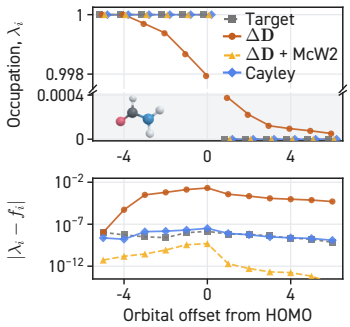
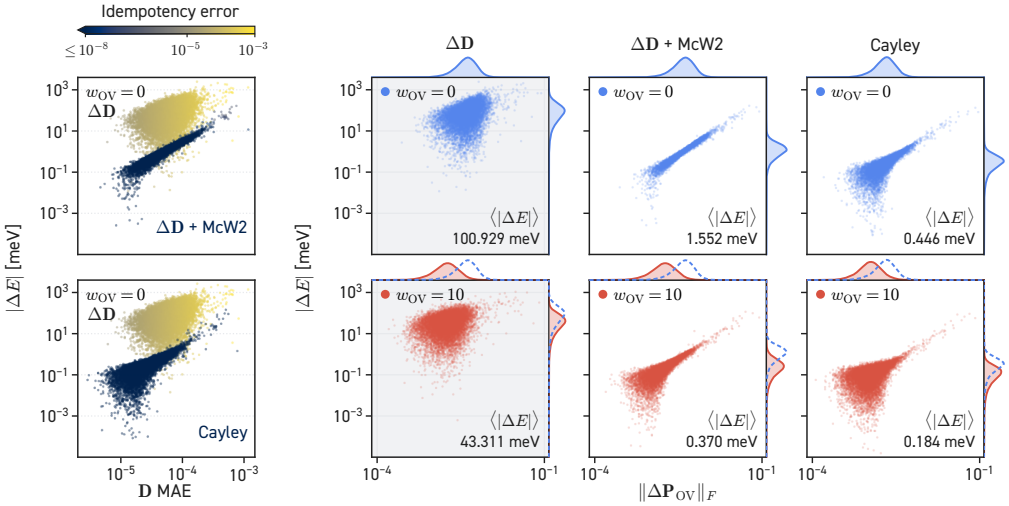


Figure 2: **Occupation correction.** Occupations (top) and deviations from 0 or 1 (bottom) for CH_3NO.

We apply two steps of McWeeny purification (*McW2*) to the same predictions to examine the effect of occupation correction. Across the full test set, McW2 reduces idempotency and absolute energy errors. At $w_{\mathrm{OV}} = 0$, the upper plot in Figure 3(a) shows a much larger reduction in energy error than in $\mathbf{D}$ MAE, with color indicating the idempotency residual. Absolute energy errors decrease for 99.08% and 99.62% of structures at $w_{\mathrm{OV}} = 0$ and 10, respectively. At $w_{\mathrm{OV}} = 0$, the OO error distribution shifts toward smaller values while the OV distribution remains nearly unchanged. Further OO/OV diagnostics are given in Section E.2. Consistently, Table 1 shows substantial reductions in trace, idempotency, energy, and force errors without a comparable reduction in $\mathbf{P}_{\mathrm{OV}}$.

Table 1: **MLDM comparison on QH9.** All models have 65.9M parameters. McW2 denotes two McWeeny steps. ERIC uses 200 test molecules (MINAO=1). Other metrics use the full test set. Best values outside Trace and Idemp. are **bold**.

Method	w_{OV}	Matrix	Validity		Subspace		Density	Downstream		SCF
		$\mathbf{D}\downarrow$ 10^{-6}	Trace↓ $10^{-6}e$	Idemp.↓ 10^{-6}	$\mathbf{P}_{\mathrm{OO}}\downarrow$ 10^{-3}	$\mathbf{P}_{\mathrm{OV}}\downarrow$ 10^{-3}	$\rho_{\mathrm{NMAE}}\downarrow$ 10^{-6}	Energy↓ meV	Force↓ meV/Å	ERIC↓
$\Delta\mathbf{D}$	0	65.58	3491.10	109.73	8.6187	4.3164	195.02	100.93	9.22	0.667
	1	67.10	2259.38	107.60	6.8664	3.2318	159.03	62.51	6.77	0.645
	10	93.98	1609.05	156.18	7.2369	2.0088	133.65	43.31	5.35	0.630
$\Delta\mathbf{D}$ + McW2	0	58.18	0.39	0.00	0.0097	4.3175	125.95	1.55	2.99	0.605
	1	52.73	0.21	0.00	0.0060	3.2331	101.59	0.87	2.66	0.579
	10	51.31	0.75	0.00	0.0037	2.0106	70.95	0.37	2.18	0.552
Cayley	0	30.74	0.20	0.00	0.0042	2.2594	60.99	0.45	2.11	0.551
	1	**28.45**	0.20	0.00	0.0033	1.7572	56.69	0.29	2.05	0.548
	10	29.64	0.20	0.00	**0.0024**	**1.2726**	**50.48**	**0.18**	**1.87**	**0.534**



(a) Idemp. constraints (b) OV supervision and energy accuracy

Figure 3: **Idempotency constraints and OV supervision on QH9.** (a) Absolute energy error versus $\mathbf{D}$ MAE at $w_{\mathrm{OV}} = 0$. The same raw $\Delta\mathbf{D}$ predictions are compared with their McW2 outputs (top) and Cayley predictions (bottom). Color indicates idemp/ error. (b) Absolute energy error versus target-basis OV error for $w_{\mathrm{OV}} = 0$ (blue) and 10 (red). Gray panels mark invalid projector $\Delta\mathbf{D}$. Annotations give mean absolute energy errors in meV. Marginals show distributions in logarithmic error coordinates, with dashed blue curves repeating the $w_{\mathrm{OV}} = 0$ baseline.

Even with the correct occupation spectrum, predicted occupied states may contain components along the target 'virtual' (unoccupied) subspace, contributing to the remnant error post-purification. We now target this subspace error by comparing models trained with different w_{OV}, evaluating each after the same McW2 protocol. Increasing w_{OV} shifts the OV and absolute energy-error distributions toward smaller values for McW2-processed $\Delta\mathbf{D}$ (Figure 3(b)). The corresponding rows in Table 1 also show improved density and force accuracy. The energy relationship becomes more dispersed at small OV errors, where the scalar OV error does not encode the direction-dependent energy response, and other residual errors become comparable. These results show that *occupation correction and subspace supervision address distinct limitations of elementwise density-matrix learning*.

Cayley subspace learning. We next test whether learning the subspace rotation directly improves on the strongest McW2-processed $\Delta\mathbf{D}$ baseline, using our Cayley-parametrized model (Cayley rows in Table 1). Both use the same initial density, equivariant backbone, dataset, and one-shot evaluation. While the baseline relies on McW2 to recover spectral validity after prediction, the Cayley model learns a subspace rotation while preserving the occupation structure by construction. The Cayley

Table 2: **MLH and MLDM comparison on QH9.** WA-FT denotes finetuning with WALoss. McW2 denotes two McWeeny purification steps. Both MLDMs use w_{OV}=10. '**All**' reports full-test means. For each row, the $E < 1\,\mathrm{eV}$ columns average only over molecules with absolute energy errors below $1\,\mathrm{eV}$. '**S filter**' denotes the overlap-eigenvalue threshold. $\mathbf{H}$ errors are in μE_h.

			$\mathbf{D}\downarrow$ (10^{-6})		$\rho_{\mathrm{NMAE}}\downarrow$ (10^{-6})		$E\downarrow$ (meV)			
Model	**S filter**	$\mathbf{H}\downarrow$	**All**	E<1 eV	**All**	E<1 eV	**All**	E<1 eV	Max.	$N_{>1\,\mathrm{eV}}\downarrow$
Hamiltonian models (MLHs)										
QHNet	10^{-6}	46.03	29342.64	1817.53	896.14	599.92	352.52	20.84	148670.43	85
	10^{-3}		2357.88	2182.12	730.72	714.41	52.22	35.28	63617.64	4
	10^{-2}		3780.58	3780.58	1973.67	1973.67	187.88	187.88	447.43	**0**
QHF2	10^{-6}	9.35	1706.42	282.88	118.66	105.58	14.69	0.71	72009.70	3
	10^{-3}		1785.32	1785.32	408.28	408.28	24.88	24.88	297.95	**0**
	10^{-2}		3765.64	3765.64	1899.40	1899.40	184.68	184.68	363.09	**0**
MLH	10^{-6}	**5.34**	2306.90	206.30	105.36	75.63	36.81	0.35	246661.40	5
	10^{-3}		1865.27	1768.42	396.95	392.11	30.39	24.66	75115.13	1
	10^{-2}		3765.18	3765.18	1892.98	1892.98	184.56	184.56	362.94	**0**
MLH + WA-FT	10^{-6}	11.64	5024.54	473.46	201.63	175.92	34.48	2.00	325736.05	2
	10^{-3}		1865.79	1821.35	444.45	439.77	28.67	25.51	39457.75	2
	10^{-2}		3766.42	3765.64	1902.42	1901.19	184.91	184.83	1264.25	1
Density-matrix models (MLDMs)										
$\Delta\mathbf{D}$ + McW2	–	–	51.31	51.31	70.95	70.95	0.37	0.37	113.68	**0**
Cayley	–	–	**29.64**	**29.64**	**50.48**	**50.48**	**0.18**	**0.18**	**85.38**	**0**

model further reduces subspace and downstream errors relative to the strongest purified baseline (Figure 3(b)), improving accuracy while preserving the occupation constraints. On the SCF on 200 subset, McW2 and stronger OV supervision also reduce the effective iteration count. Cayley yields the lowest ERIC at each OV weight (Table 1). Detailed SCF results are given in Figure 13.

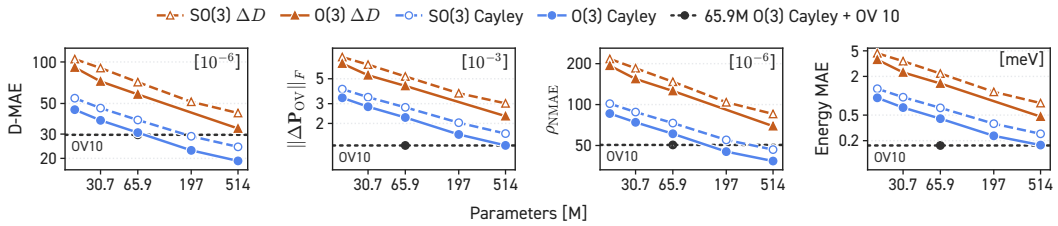
Together, the results support *treating the occupied subspace as the learned degree of freedom while enforcing electron number, occupation spectrum, and idempotency through the parameterization.*

4.2 Comparison with machine-learned Hamiltonians

We compare against a matched 65.9M MLH with the same O(3) backbone, 20.4M QHNet (Yu et al., 2023b), and 43M QHFlow2-m (Kim et al., 2026) (Table 2). The matched MLH achieves a Hamiltonian MAE of $5.34\,\mu E_h$, below the state-of-the-art QHFlow2-m's $9.35\,\mu E_h$, providing a strong controlled baseline. We also fine-tune the matched MLH with wavefunction alignment loss (MLH + WA-FT) (Li et al., 2025b) to assess spectral supervision. Each MLH yields $\mathbf{D}$ by generalized eigendecomposition of its predicted $\mathbf{H}$ with $\mathbf{S}$, whose small eigenvalues can amplify errors in orbital energies and the occupied subspace (Li et al., 2025b). We hold $\mathbf{H}$ fixed and vary the overlap-filter threshold across 10^{-6} (PySCF default), 10^{-3}, and 10^{-2}.[3] The MLDMs are the strongest $\Delta\mathbf{D}$ with McW2 and Cayley models from Table 1, both with $w_{\mathrm{OV}} = 10$. Section E.3.1 gives further details and a matched continuation with the original Hamiltonian loss.

At the default filter, QHFlow2-m and the un-finetuned matched MLH achieve sub-meV mean energy errors on their respective $|\Delta E| < 1$ eV subsets, yet rare large errors dominate their full-test means. For these two models, eliminating all errors above 1 eV by filtering raises the full-test mean above its default-filter value, revealing a tradeoff between suppressing failures and restricting the AO subspace (Figure 15). QHNet has a higher sub-eV mean of 20.84 meV and 85 errors above 1 eV at the default filter. WA-FT reduces the matched MLH's default-filter count from five to two, but increases its maximum error and mean over its own sub-eV subset. In comparison, $\Delta\mathbf{D}$ with McW2 and Cayley achieve full-test means of 0.37 and 0.18 meV, respectively, with no errors above 1 eV. These results support *learning* $\mathbf{D}$ *to avoid overlap-sensitive Hamiltonian-to-density conversion while retaining accurate one-shot energies*.

[3] Target-density reference energies are calculated with default threshold (τ=10^{-6}).



(a) Density matrix (b) OV error (c) Real-space density (d) Energy

Figure 4: **Parameter-size scaling on QH9.** (a) Density-matrix MAE (D-MAE), (b) target-basis OV error ($\|\Delta\mathbf{P}_{\mathrm{OV}}\|_F$), (c) real-space density error (ρ_{NMAE}), and (d) one-shot energy MAE for models trained with $w_{\mathrm{OV}} = 0$. Both $\Delta\mathbf{D}$ and Cayley predictions use McW2. Black dotted lines mark the corresponding errors of the 65.9M O(3) Cayley model trained with $w_{\mathrm{OV}} = 10$

4.3 Model ablations and scaling

Inversion symmetry. Preserving the occupation spectrum does not by itself ensure inversion equivariance. Recent Hamiltonian models are formulated with SO(3) equivariance (Yu et al., 2025; Kim et al., 2026), which alone does not guarantee inversion equivariance. We compare SO(3) and O(3) models to assess the empirical importance of this symmetry in density-matrix learning.

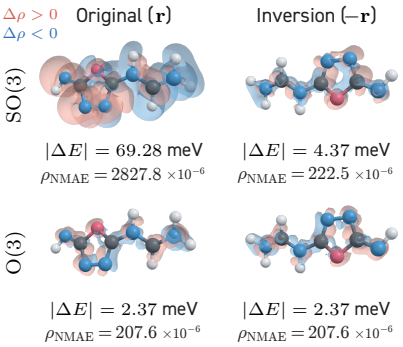


Figure 5: **Inversion sensitivity.** Signed real-space density errors of SO(3) and O(3) Cayley at original and inverted geometries.

We observe lower mean one-shot energy errors for O(3) in both $\Delta\mathbf{D}$ and Cayley on the full QH9 test set under the common McW2 protocol. In a separate symmetry audit, we identify two SO(3) Cayley cases with pronounced inversion sensitivity despite preserved occupations. In Figure 5, inversion substantially changes the SO(3) model's density and energy errors, whereas the O(3) model retains consistent predictions. For the illustrated Cayley comparison, the lower mean energy error persists after excluding these two molecules. Detailed comparisons and symmetry diagnostics are provided in Section E.4.

We find the same accuracy advantage in parameter-size scaling (Figure 4): for both parameterizations, the models achieve accuracy comparable to their counterparts across all four metrics, using approximately half as many parameters. Together, the symmetry diagnostics and accuracy comparisons provide empirical support for *the importance of inversion symmetry in density-matrix learning, reflected in both energy accuracy and parameter efficiency.*

Parameter-size scaling. We next examine how prediction errors change with model capacity (Figure 4). We compare SO(3) and O(3) variants of $\Delta\mathbf{D}$ and Cayley from 18.1M to 514M parameters, all trained with $w_{\mathrm{OV}} = 0$ and evaluated after the same McW2 post-processing. Increasing capacity reduces matrix, subspace, density, and energy errors in both families. At every shared model size, O(3) models achieve lower errors than their SO(3)counterparts across all four metrics. Our results show *consistent scaling behavior in spectral constrained models including Cayley*, with errors decreasing across all four metrics as model capacity increases.

Effect of OV-loss weight. Here, we vary $w_{\mathrm{OV}} \in \{0, 1, 10, 20, 50\}$ for the 65.9M O(3) Cayley model under fixed training settings (Table 9) and the same McW2 protocol. Increasing the weight from 0 to 50 reduces energy MAE from 0.440 to 0.154 meV, whereas $\mathbf{D}$ MAE is lowest at $w_{\mathrm{OV}} = 1$ (Figure 6). Within this sweep, $w_{\mathrm{OV}} = 10$ offers a practical trade-off, keeping density-matrix MAE near its minimum while achieving an energy MAE of 0.168 meV. This energy error is comparable to 0.171 meV for the 514M O(3) Cayley model with $w_{\mathrm{OV}} = 0$ and 7.8× as many parameters (Figure 4). These results show that, in this comparison, *explicit subspace supervision improves both the accuracy and parameter efficiency of one-shot energy prediction.*

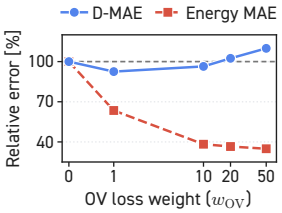
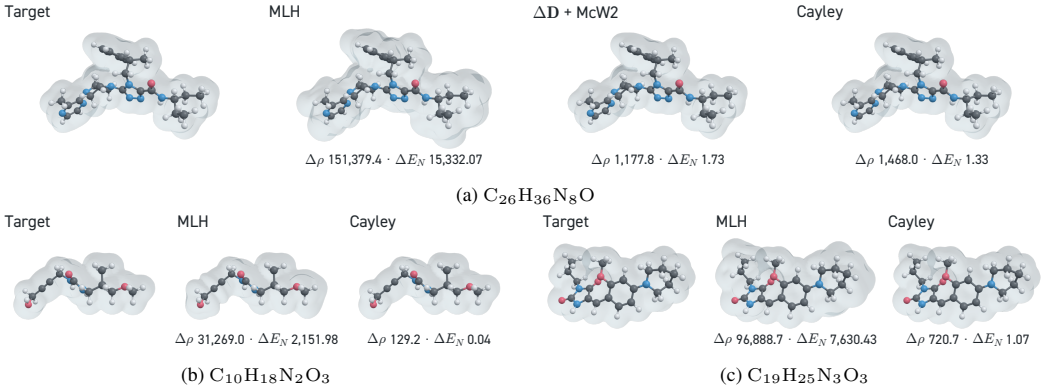


Figure 6: **OV-weight sweep.** Errors relative to the w_{OV}=0.

Table 3: **QM40 size extrapolation.** QH9-trained models are evaluated on 2,243 molecules without fine-tuning. McW2 denotes two McWeeny steps. Conv. denotes the SCF convergence rate, and ERIC is normalized to MINAO. N is the total atom count. Best MLDM errors (excluding Trace and Idemp.) and ERIC values are **bold**.

Method	w_{OV}	Matrix MAE		Validity		Subspace	Density	Energy MAE		SCF	
		H ↓	**D** ↓	Trace↓	Idemp.↓	$\mathbf{P}_{\mathrm{OV}}$ ↓	ρ_{NMAE} ↓	E ↓	E/N ↓	Conv.↑	ERIC↓
		μE_h	10^{-6}	$10^{-6}e$	10^{-6}	10^{-3}	10^{-6}	meV	meV/atom	%	
MLH	–	91.91	250296.12	18857.47	1268.55	262.85	73006.36	399646.96	6920.06	95.5	1.68
ΔD	0	–	193.87	63325.90	324.25	44.63	1039.39	848.83	15.68	100.0	0.77
	10	–	197.08	71953.88	290.90	26.49	801.44	677.41	12.02	100.0	0.74
ΔD + McW2	0	–	190.40	964.51	0.20	44.63	800.88	192.08	3.40	100.0	0.73
	10	–	152.42	654.39	0.08	26.48	**561.17**	79.03	1.38	100.0	**0.70**
Cayley	0	–	121.61	0.31	0.00	31.21	660.08	84.31	1.51	100.0	0.71
	10	–	**118.45**	0.31	0.00	**24.19**	801.61	**55.51**	**0.98**	100.0	**0.70**



(a) $C_{26}H_{36}N_8O$

(b) $C_{10}H_{18}N_2O_3$

(c) $C_{19}H_{25}N_3O_3$

Figure 7: **QM40 electron densities.** Total (heavy) atoms: (a) 71 (35), (b) 33 (15), (c) 50 (25). Values are density errors $\Delta\rho$ (10^{-6}) and energy MAE per atom ΔE_N (meV/atom). MLDMs use $w_{\mathrm{OV}} = 10$.

4.4 Molecular-size extrapolation on QM40

We evaluate QH9-trained models on 2,243 larger QM40 molecules without fine-tuning (Figure 8 and Table 3). All evaluated MLDMs achieve lower density-matrix, real-space density, and energy errors than the MLH baseline, supporting improved extrapolation to larger molecules. At $w_{\mathrm{OV}} = 10$, Cayley achieves lower matrix, OV, and energy errors than **ΔD** with McW2, with an energy error of 0.98 versus 1.38 meV/atom. These energies are evaluated directly from predicted densities. The predicted densities also improve SCF initialization. All evaluated MLDM conditions converge on every molecule, compared with 95.5% for MLH. Cayley and **ΔD** with McW2 at $w_{\mathrm{OV}} = 10$ both achieve 0.70 ERIC. Evaluation details are provided in Section D.2.

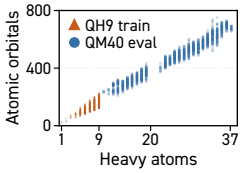


Figure 8: **Size coverage.**

We nevertheless find that lower density error does not necessarily imply lower energy error. At w_{OV}=10, **ΔD** with McW2 achieves a lower ρ_{NMAE} than Cayley (561.17×10^{-6} versus 801.61×10^{-6}), despite its higher energy error. ρ_{NMAE} measures the magnitude of spatial deviations, whereas their energetic impact depends on where electrons are redistributed and how the orbital states change. Figure 7 illustrates both errors alongside the predicted densities. These results support *effective molecular-size extrapolation of the spectral constrained MLDMs with subspace alignment*.

5 Conclusion

Our results show that occupation-preserving Cayley rotations and occupied-subspace supervision provide complementary improvements in downstream prediction beyond training with an entrywise density-matrix loss alone. The training objective uses fixed target densities without electronic-structure solver evaluations. Our framework improves density and energy accuracy and SCF initialization on QH9, with energy gains extending to QM40 without fine-tuning. We further identify overlap-filtering tradeoffs in MLHs and inversion sensitivity in MLDMs, highlighting the importance of both numerical stability and spatial symmetry.

AI USE STATEMENT

We developed the proposed method, designed the experiments, derived the mathematical formulations, and wrote the manuscript. OpenAI's ChatGPT and Codex were used under our supervision to assist with drafting and revising manuscript text, refining mathematical proofs, and writing plotting code. All final mathematical derivations, proofs, and scientific decisions were verified and made by us.

REPRODUCIBILITY STATEMENT

We provide the code as supplementary material, and will publicly release the full set of model checkpoints and the processed QH9 and QM40 datasets used in this work upon acceptance. The mathematical conventions and derivations underlying the proposed method are provided in Section B.3, Section C.4.1, and Section C. Model and training details are reported in Sections C.3 and D.1, and the dataset split, electronic-structure settings, evaluation protocol, and metric definitions are specified in Sections D.1 to D.3.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3. 16

Jan Almlöf, Knut Faegri, Jr., and Karl Korsell. Principles for a direct SCF approach to LCAO–MO ab initio calculations. *Journal of Computational Chemistry*, 3(3):385–399, 1982. doi: 10.1002/jcc.540030314. URL https://doi.org/10.1002/jcc.540030314. 5, 19

Axel D. Becke. A multicenter numerical integration scheme for polyatomic molecules. *The Journal of Chemical Physics*, 88(4):2547–2553, 1988. doi: 10.1063/1.454033. URL https://doi.org/10.1063/1.454033. 18

Axel D. Becke. Density-functional thermochemistry. III. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi: 10.1063/1.464913. URL https://doi.org/10.1063/1.464913. 18

S. F. Boys. Electronic wave functions. i. a general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 200(1063):542–554, 1950. doi: 10.1098/rspa.1950.0036. URL https://doi.org/10.1098/rspa.1950.0036. 3

Felix Brockherde, Leslie Vogt, Li Li, Mark E. Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the Kohn–Sham equations with machine learning. *Nature Communications*, 8:872, 2017. doi: 10.1038/s41467-017-00839-3. URL https://doi.org/10.1038/s41467-017-00839-3. 16

Éric Cancès, Gaspard Kemlin, and Antoine Levitt. Convergence analysis of direct minimization and self-consistent iterations. *SIAM Journal on Matrix Analysis and Applications*, 42(1):243–274, 2021. doi: 10.1137/20M1332864. URL https://doi.org/10.1137/20M1332864. 18, 19

Nicolas G. Cartier and Klaas J. H. Giesbertz. Impact of parametrizations of the one-body reduced density matrix on the energy landscape. *The Journal of Physical Chemistry Letters*, 16(15):3822–3831, 2025. doi: 10.1021/acs.jpclett.5c00308. URL https://doi.org/10.1021/acs.jpclett.5c00308. 4, 16

Chaoran Cheng and Jian Peng. Equivariant neural operator learning with graphon convolution. *Advances in Neural Information Processing Systems*, 36:61960–61984, 2023. 16

Eike S Eberhard, Viktor Kotsev, Timm Güthle, and Stephan Günnemann. Transferable scf-acceleration through solver-aligned initialization learning. *arXiv preprint arXiv:2604.21657*, 2026. 2, 5, 16, 17, 24, 25

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954. 16

Jonas Elsborg, Luca Thiede, Alán Aspuru-Guzik, Tejs Vegge, and Arghya Bhowmik. ELECTRA: A cartesian network for 3D charge density prediction with floating orbitals. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28092–28121, 2025. doi: 10.52202/085713-0947. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/288b63aa98084366c4536ba0574a0f22-Abstract-Conference.html. 16, 17

Jonas Elsborg, Felix Ærtebjerg, Luca Thiede, Alán Aspuru-Guzik, Tejs Vegge, and Arghya Bhowmik. Global plane waves from local gaussians: Periodic charge densities in a blink. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.19966v2. 16

Vladimir Fanaskov, Vladislav Trifonov, Alexander Rudikov, Ekaterina Muravleva, and Ivan Oseledets. Deep learning for subspace regression. In *International Conference on Learning Representations*, 2026. 16

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, and C Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *International Conference on Machine Learning*, 2025. 1, 4

Xiaoxun Gong, He Li, Nianlong Zou, Runzhang Xu, Wenhui Duan, and Yong Xu. General framework for E(3)-equivariant neural network representation of density functional theory hamiltonian. *Nature Communications*, 14:2848, 2023. doi: 10.1038/s41467-023-38468-8. URL https://doi.org/10.1038/s41467-023-38468-8. 15

S. Hazra, U. Patil, and S. Sanvito. Predicting the one-particle density matrix with machine learning. *Journal of Chemical Theory and Computation*, 20(11):4569–4578, 2024. doi: 10.1021/acs.jctc.4c00042. URL https://doi.org/10.1021/acs.jctc.4c00042. 1, 16

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1969–1978. PMLR, 2018. URL https://proceedings.mlr.press/v80/helfrich18a.html. 16

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864. URL https://doi.org/10.1103/PhysRev.136.B864. 17

Peter Bjørn Jørgensen and Arghya Bhowmik. Deepdft: Neural message passing network for accurate charge density prediction. *arXiv preprint arXiv:2011.03346*, 2020. 16

Manasa Kaniselvan, Alexander Maeder, Chen Hao Xia, Alexandros Nikolaos Ziogas, and Mathieu Luisier. Distributed equivariant graph neural networks for large-scale electronic structure prediction. *arXiv preprint arXiv:2507.03840*, 2025. URL https://arxiv.org/abs/2507.03840. 15

Manasa Kaniselvan, Mauro Dossena, Denghui Lu, Alexander Maeder, Nicolas Vetsch, Alexandros Nikolaos Ziogas, and Mathieu Luisier. Ab initio modeling of MoS_2/oxide device interfaces with machine learned electronic structures. *arXiv preprint arXiv:2608.27533*, 2026a. URL https://arxiv.org/abs/2608.27533. 15

Manasa Kaniselvan, Alexander Maeder, Denghui Lu, Alexandros Nikolaos Ziogas, and Mathieu Luisier. MALOQ: Massively accelerated learning of operators for quantum transport. *arXiv preprint arXiv:2606.28911*, 2026b. URL https://arxiv.org/abs/2606.28911. 1, 4, 15, 21

Manasa Kaniselvan, Benjamin Kurt Miller, Meng Gao, Juno Nam, and Daniel S. Levine. Learning from the electronic structure of molecules across the periodic table. *International Conference on Learning Representations*, 2026c. URL https://arxiv.org/abs/2510.00224. 1, 4, 15

Seongsu Kim and Sungsoo Ahn. Gaussian plane-wave neural operator for electron density estimation. In *International Conference on Machine Learning*, 2024. 16

Seongsu Kim, Nayoung Kim, Dongwoo Kim, and Sungsoo Ahn. High-order equivariant flow matching for density functional theory hamiltonian prediction. *Advances in Neural Information Processing Systems*, 38:13265–13307, 2025. 1, 15

Seongsu Kim, Chanhui Lee, Yoonho Kim, Seongjun Yun, Honghui Kim, Nayoung Kim, Changyoung Park, Sehui Han, Sungbin Lim, and Sungsoo Ahn. Machine learning hamiltonians are accurate energy-force predictors. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.16897. 1, 2, 4, 5, 7, 8, 15, 19, 21, 36

Manuel V. Klockow, Marc K. Ickler, Peter Lippmann, and Fred A. Hamprecht. A function-centric graph neural network approach for predicting electron densities. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HDdkFjFEZd. 16, 17

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133. URL https://doi.org/10.1103/PhysRev.140.A1133. 1, 3, 18

Teddy Koker, Keegan Quigley, Eric Taw, Kevin Tibbetts, and Lin Li. Higher-order equivariant neural networks for charge density prediction in materials. *npj Computational Materials*, 10(1):161, 2024. 16

Chenghan Li, Or Sharir, Shunyue Yuan, and Garnet Kin-Lic Chan. Image super-resolution inspired electron density prediction. *Nature Communications*, 16:4811, 2025a. doi: 10.1038/s41467-025-60095-8. URL https://www.nature.com/articles/s41467-025-60095-8. 17

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. doi: 10.1038/s43588-022-00265-6. URL https://doi.org/10.1038/s43588-022-00265-6. 1, 15

Jun Li, Fuxin Li, and Sinisa Todorovic. Efficient riemannian optimization on the Stiefel manifold via the Cayley transform. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.01113. 4, 16

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*, 2025b. URL https://openreview.net/forum?id=twEvvkQqPS. 1, 7, 15, 17, 37

Zhe Liu, Yuyan Ni, Zhichen Pu, Qiming Sun, Siyuan Liu, and Wen Yan. Towards a transferable acceleration method for density functional theory. In *International Conference on Learning Representations*, volume 2026, pp. 82169–82189, 2026. 1, 2

Per-Olov Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *The Journal of Chemical Physics*, 18(3):365–375, 1950. doi: 10.1063/1.1747632. 3

Giulia Luise, Chin-Wei Huang, Thijs Vogels, Derk P. Kooi, Sebastian Ehlert, Stephanie Lanius, Klaas J. H. Giesbertz, Amir Karton, Deniz Gunceler, Stefano Battaglia, Gregor N. C. Simm, P. Bernát Szabó, Megan Stanley, Wessel P. Bruinsma, Lin Huang, Xinran Wei, José Garrido Torres, Abylay Katbashev, Rodrigo Chavez Zavaleta, Bálint Máté, Sékou-Oumar Kaba, Roberto Sordillo, Yingrong Chen, David B. Williams-Young, Christopher M. Bishop, Jan Hermann, Rianne van den Berg, and Paola Gori-Giorgi. Accurate and scalable exchange-correlation with deep learning. *arXiv preprint arXiv:2506.14665*, 2025. doi: 10.48550/arXiv.2506.14665. URL https://arxiv.org/abs/2506.14665. 18

Erpai Luo, Xinran Wei, Lin Huang, Yunyang Li, Han Yang, Zaishuo Xia, Zun Wang, Chang Liu, Bin Shao, and Jia Zhang. Efficient and scalable density functional theory hamiltonian prediction through adaptive sparsity. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K3lykWhXON. 15

Ayesh Madushanka, Renaldo T. Moura, Jr., and Elfi Kraka. QM40, realistic quantum mechanical dataset for machine learning in molecular science. *Scientific Data*, 11:1376, 2024. doi: 10.1038/s41597-024-04206-y. URL https://www.nature.com/articles/s41597-024-04206-y. 2, 5, 27

R. McWeeny. Some recent advances in density matrix theory. *Reviews of Modern Physics*, 32(2):335–369, 1960. doi: 10.1103/RevModPhys.32.335. URL https://doi.org/10.1103/RevModPhys.32.335. 2, 5, 30

Jigyasa Nigam, Tess Smidt, and Geneviève Dusson. Machine learning of electronic structure and atomistic properties from the external potential. *The Journal of Chemical Physics*, 165(3):034119, 2026. doi: 10.1063/5.0332678. URL https://doi.org/10.1063/5.0332678. 16

Robert G. Parr and Weitao Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, New York, 1989. 18

John P. Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*, volume 577, pp. 1–20, 2001. doi: 10.1063/1.1390175. URL https://doi.org/10.1063/1.1390175. 18

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zIUyj55nXR. 5

Bhaskar Rana, Nicolas Viot, Jessica A. Martinez B., Xuecheng Shao, Pablo Ramos, and Michele Pavanello. Learning the one-electron reduced density matrix at SCF convergence thresholds. *Journal of Chemical Theory and Computation*, 21(24):12652–12663, 2025. doi: 10.1021/acs.jctc.5c01564. URL https://doi.org/10.1021/acs.jctc.5c01564. 1, 16

C. C. J. Roothaan. New developments in molecular orbital theory. *Reviews of Modern Physics*, 23(2):69–89, 1951. doi: 10.1103/RevModPhys.23.69. 18

Reinhold Schneider, Thorsten Rohwedder, Alexej Neelov, and Johannes Blauert. Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure. *Journal of Computational Mathematics*, 27(2–3):360–387, 2009. URL https://arxiv.org/abs/0805.1190. 18

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(1):5024, 2019. doi: 10.1038/s41467-019-12875-2. URL https://doi.org/10.1038/s41467-019-12875-2. 1, 15

Xuecheng Shao, Lukas Paetow, Mark E. Tuckerman, and Michele Pavanello. Machine learning electronic structure methods based on the one-electron reduced density matrix. *Nature Communications*, 14:6281, 2023. doi: 10.1038/s41467-023-41953-9. URL https://doi.org/10.1038/s41467-023-41953-9. 1, 16

Ron Shepard, Scott R. Brozell, and Gergely Gidofalvi. The representation and parametrization of orthogonal matrices. *The Journal of Physical Chemistry A*, 119(28):7924–7939, 2015. doi: 10.1021/acs.jpca.5b02015. URL https://doi.org/10.1021/acs.jpca.5b02015. 4, 20

John C. Slater. Atomic shielding constants. *Physical Review*, 36(1):57–64, 1930. doi: 10.1103/PhysRev.36.57. URL https://doi.org/10.1103/PhysRev.36.57. 3

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. doi: 10.1002/wcms.1340. URL https://doi.org/10.1002/wcms.1340. 5, 18, 19, 27

Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1996. 3, 18

Zechen Tang, Nianlong Zou, He Li, Yuxiang Wang, Zilong Yuan, Honggeng Tao, Yang Li, Zezhou Chen, Boheng Zhao, Minghui Sun, Hong Jiang, Wenhui Duan, and Yong Xu. Improving density matrix electronic structure method by deep learning. *arXiv preprint arXiv:2406.17561*, 2024. URL https://arxiv.org/abs/2406.17561. 1, 16

Oliver T. Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess E. Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems*, volume 34, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/78f1893678afbeaa90b1fa01b9cfb860-Abstract.html. 1, 15

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical reviews*, 121(16):10142–10186, 2021b. 1

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(1):65, 2025. 16

J. H. Van Lenthe, R. Zwaans, H. J. J. Van Dam, and M. F. Guest. Starting SCF calculations by superposition of atomic densities. *Journal of Computational Chemistry*, 27(8):926–932, 2006. doi: 10.1002/jcc.20393. URL https://doi.org/10.1002/jcc.20393. 5, 19

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7:3297–3305, 2005. doi: 10.1039/B508541A. 27

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4310. URL https://papers.nips.cc/paper_files/paper/2025/hash/bbf23e81b0ad7637fe9a731d0b676ca6-Abstract-Conference.html. 1

Chen Hao Xia, Manasa Kaniselvan, Alexandros Nikolaos Ziogas, Marko Mladenović, Rayen Mahjoub, Alexander Maeder, and Mathieu Luisier. Learning the electronic Hamiltonian of large atomic structures. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68236–68257. PMLR, 2025. URL https://proceedings.mlr.press/v267/xia25b.html. 15

Zuriel Y. Yescas-Ramos, Andrés Álvarez-García, and Huziel E. Sauceda. Towards accelerated SCF workflows with equivariant density-matrix learning and analytic refinement. *arXiv preprint arXiv:2604.27256*, 2026. URL https://arxiv.org/abs/2604.27256. 1, 2, 5, 16

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, Datasets and Benchmarks Track*, 2023a. URL https://openreview.net/forum?id=71uRr9N39A. 2, 5, 19, 27

Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum hamiltonian. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023b. URL https://arxiv.org/abs/2306.04922. 1, 5, 7, 15

Haiyang Yu, Yuchao Lin, Xuan Zhang, Xiaofeng Qian, and Shuiwang Ji. Efficient prediction of SO(3)-equivariant hamiltonian matrices via SO(2) local frames. *Proceedings of the 43rd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.09398. 1, 8, 15

Liwei Zhang, Patrizia Mazzeo, Michele Nottoli, Edoardo Cignoni, Lorenzo Cupellini, and Benjamin Stamm. A symmetry-preserving and transferable representation for learning the Kohn–Sham density matrix. *Digital Discovery*, 5:1868, 2026. doi: 10.1039/D5DD00230C. URL https://doi.org/10.1039/D5DD00230C. 1, 16

Zhanghao Zhouyin, Zixi Gan, Shishir Kumar Pandey, Linfeng Zhang, and Qiangqiang Gu. Learning local equivariant representations for quantum operators. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kpq3IIjUD3. 16

APPENDIX

A Related work **15**
A.1 Occupation constraints and subspace supervision 16
A.2 Reported real-space density results 17

B DFT background and density-matrix representation **17**
B.1 From the electronic Schrödinger equation to Kohn–Sham DFT 17
B.2 Self-consistent field methods 18
B.3 Electron count, occupations, and idempotency in an AO basis 19

C Method details **19**
C.1 Construction of the initial density 19
C.2 Cayley parameterization 20
C.2.1 Rotation equivariance of the Cayley head 20
C.3 Model architecture and output heads 21
C.3.1 Operator details 23
C.4 Density-matrix and occupied–virtual objectives 23
C.4.1 Occupied-subspace error and principal angles 24
C.4.2 Relation to solver-aligned initialization learning 24
C.5 Inversion equivariance by group averaging 25
C.6 Computational complexity 26

D Experimental setup and evaluation **27**
D.1 Common experimental setup 27
D.2 QM40 extrapolation protocols 27
D.3 Evaluation metrics and aggregation 30
D.4 McWeeny purification 30
D.5 Computational resources and runtime 31

E Additional experiments and analyses **32**
E.1 Effects of numerical precision on the Cayley model 32
E.2 Occupation and subspace errors 32
E.2.1 OO versus OV supervision in Cayley fine-tuning 36
E.3 Comparison with machine-learned Hamiltonians 36
E.3.1 Orthonormal-basis fine-tuning (WA-FT) 37
E.4 Inversion-symmetry diagnostics 38
E.5 Molecular-size extrapolation on QM40 38

A RELATED WORK

Learning DFT Hamiltonian matrices Machine-learned Hamiltonian models predict effective one-electron operators in an AO basis, from which orbitals, orbital energies, and density matrices can be obtained. SchNOrb, PhiSNet, DeepH, DeepH-E3, QHNet, SPHNet and QHNetV2 established equivariant architectures for this task (Schütt et al., 2019; Unke et al., 2021a; Li et al., 2022; Gong et al., 2023; Yu et al., 2023b; 2025; Luo et al., 2025). QHFlow models Hamiltonians with flow matching, whereas QHFlow2 evaluates predicted Hamiltonians through downstream energies and forces (Kim et al., 2025; 2026). WANet introduces wavefunction-alignment supervision which targets the orbital energy of the matrix (Li et al., 2025b). HELM extends chemical and basis-set coverage (Kaniselvan et al., 2026c). Augmented graph partitioning supports Hamiltonian learning for extended materials (Xia et al., 2025), while distributed equivariant GNNs and MALOQ scale prediction across GPUs (Kaniselvan et al., 2025; 2026b;a).

Direct density-matrix learning Direct density-matrix models avoid constructing the density through a predicted Hamiltonian eigensystem. Early approaches predicted molecule-specific one-

electron density matrices (Shao et al., 2023; Hazra et al., 2024; Rana et al., 2025; Tang et al., 2024; Zhouyin et al., 2025; Zhang et al., 2026). dm-PhiSNet combines physically motivated training objectives with analytic trace and idempotency corrections (Yescas-Ramos et al., 2026). Op2Op reports that restoring projector conditions can improve derived properties (Nigam et al., 2026).

Real-space electron-density learning Real-space density learning is another approach for modeling electronic structures. Early work learned the external-potential-to-density map to bypass self-consistent KS calculations (Brockherde et al., 2017). Atom-to-query message-passing-based architectures are introduced (Jørgensen & Bhowmik, 2020; Cheng & Peng, 2023; Koker et al., 2024), while LAGNet adapts equivariant DeepDFT to LCAO-based densities (Ushenin et al., 2025). GPWNO combines Gaussian and plane-wave bases (Kim & Ahn, 2024). ELECTRA learns adaptive floating Gaussians (Elsborg et al., 2025), and ELECTRAFI extends this approach to periodic densities (Elsborg et al., 2026). BOA combines function-based message passing with quadratic products of atom-centered basis functions (Klockow et al., 2026).

Predicting the AO density matrix $\mathbf{D}$, rather than only the real-space density $\rho(\mathbf{r})$, provides direct access to the occupied subspace and orbital-dependent energy terms, which can be evaluated through matrix contractions without an additional KS diagonalization. The matrix representation also allows us to preserve electron number, the occupation spectrum, and idempotency through Cayley rotations, and to supervise occupied-subspace alignment with the OV loss. These physical constraints and subspace information can also benefit density-modeling accuracy, while $\rho(\mathbf{r})$ remains accessible by AO contraction and the same $\mathbf{D}$ can be used for SCF initialization. We compare the reported ELECTRA, BOA, and Cayley density errors under their respective evaluation protocols in Section A.2.

Occupied-subspace parameterization and alignment For fixed-electron-number closed-shell systems, the normalized orthonormal density $\mathbf{P}$ is a fixed-rank projector, connecting density prediction to optimization on the Grassmann manifold (Edelman et al., 1998; Absil et al., 2008). More generally, input-dependent subspace prediction has been studied as subspace regression (Fanaskov et al., 2026). Cayley transforms provide a parameterization of orthogonal rotations used in optimization on Stiefel and Grassmann manifolds (Li et al., 2020; Helfrich et al., 2018), and have also been applied to orbital parameterizations of one-particle density matrices (Cartier & Giesbertz, 2025). For SCF initialization, SAIL supervises orbital stationarity along differentiable SCF trajectories, without requiring ground-state labels during fine-tuning (Eberhard et al., 2026). We combine geometry-dependent Cayley rotations with target-basis OV supervision against fixed target densities, without Fock construction or SCF unrolling during loss evaluation.

A.1 Occupation constraints and subspace supervision

Table 4: **Comparison between occupation and subspace-related methods.**

Method	Direct $\mathbf{D}$	Exact occupations	OV-focused loss	No Fock/SCF in loss
dm-PhiSNet	✓	▲	✗	✓
SAIL (density)	✓	✗	✓	✗
WA-FT (MLH)	✗	✓	✗	✓
Cayley + OV (ours)	✓	✓	✓	✓

We distinguish our approach from prior methods by how they enforce occupations, supervise the occupied subspace, and depend on SCF differentiation (Table 4).

Our framework combines direct density-matrix prediction with exact occupation preservation through Cayley rotations and target-subspace alignment through an auxiliary OV loss, without Fock construction or SCF unrolling during loss evaluation.

dm-PhiSNet enforces electron number by trace rescaling and improves idempotency through iterative McWeeny purification (Yescas-Ramos et al., 2026). Because finite purification steps do not generally yield exact binary occupations, we mark this support as partial (Section D.4).

SAIL supervises OV orbital gradients along SCF trajectories (Eberhard et al., 2026). It targets orbital stationarity and requires solver differentiation, but does not constrain the initial predicted density to be a projector (Section C.4.2).

WALoss penalizes the eigenvalues of the orbitals directly (Li et al., 2025b). Since this methods can be used for Hamiltonian prediction, its valid occupations follow from diagonalization and selecting the lowest eigenvalues according to the number of electrons. Therefore, it retains MLH's overlap-sensitive density construction, which can yield large density and energy errors despite valid occupations.

A.2 Reported real-space density results

Here, we report the recent real-space density model and compare it with our approach in Table 5. The published QM9 errors are taken from Table 1 of Elsborg et al. (2025) for ELECTRA and Table 1 of Klockow et al. (2026) for BOA. Our 65.9M Cayley results use $w_{\mathrm{OV}} = 0$ and 10 on the full QH9 test set from Table 1.

Table 5: **Real-space density errors on QM9 and QH9.** Published ELECTRA (Elsborg et al., 2025) and BOA (Klockow et al., 2026) results are shown alongside our Cayley results from Table 1.

Model	Dataset	**Target calculation**			Density type	**Density error**
		Program	Functional	Basis / core treatment		ρ_{NMAE} (%) ↓
ELECTRA	QM9	VASP	PBE	Plane waves PAW	Valence	0.176
BOA large	QM9	VASP	PBE	Plane waves PAW	Valence	0.1339
BOA large	QM9	PySCF	PBE	GTH-TZV2P GTH	Valence	0.116
Cayley OV0	QH9	PySCF	B3LYP5	def2-SVP	All-electron	0.006099
Cayley OV10	QH9	PySCF	B3LYP5	def2-SVP	All-electron	0.005048

The QM9 density dataset used for ELECTRA and BOA consists of valence densities on uniform grids (Elsborg et al., 2025; Klockow et al., 2026; Li et al., 2025a), whereas our QH9 is based on all-electron densities (Section D.3).

B DFT background and density-matrix representation

B.1 From the electronic Schrödinger equation to Kohn–Sham DFT

For a molecular geometry $\mathcal{M} = \{Z_i, \mathbf{r}_i\}_i$, with nuclear charges Z_i and fixed nuclear positions $\mathbf{r}_i$, the electronic Schrödinger equation is

$$\widehat{\mathcal{H}}_e(\mathcal{M})\Psi = E\Psi. \tag{8}$$

Here $\widehat{\mathcal{H}}_e$ is the many-electron Hamiltonian, containing electronic kinetic energy, electron–nuclear attraction, and electron–electron repulsion. Its eigenfunction Ψ is the wavefunction of all N_e electrons, including their spatial and spin coordinates, and E is the corresponding electronic energy. Representing Ψ becomes difficult as N_e grows because it depends jointly on all electrons. Density functional theory (DFT) instead minimizes an energy functional over admissible electron densities $\rho(\mathbf{r})$ (Hohenberg & Kohn, 1964):

$$E_0(\mathcal{M}) = \min_{\rho:\, \int \rho(\mathbf{r})d\mathbf{r} = N_e} E_{\mathrm{DFT}}[\rho; \mathcal{M}]. \tag{9}$$

Here E_0 is the electronic ground-state energy and $E_{\mathrm{DFT}}[\rho; \mathcal{M}]$ is the electronic energy functional at the fixed geometry. The density $\rho(\mathbf{r})$ gives the number of electrons per unit volume at position $\mathbf{r}$, so the integral constraint fixes the electron number. With the exact functional, this reformulation is exact for the ground state.

Kohn–Sham (KS) DFT introduces an auxiliary system of noninteracting electrons whose ground-state density matches that of the interacting system in the exact formulation (Kohn & Sham, 1965). Its orthonormal one-electron orbitals $\psi_i(\mathbf{r})$ construct the density as

$$\rho(\mathbf{r}) = \sum_i f_i |\psi_i(\mathbf{r})|^2, \qquad \sum_i f_i = N_e,$$

where f_i is the occupation of spatial orbital i, equal to two for occupied orbitals and zero for virtual orbitals in the restricted closed-shell setting. The orbitals describe one-electron states of the auxiliary system, rather than distinguishable individual electrons.

This independent-particle construction provides the noninteracting kinetic energy explicitly. The Hartree term accounts for classical electrostatic repulsion, while the exchange–correlation functional contains the remaining interaction and kinetic-energy contributions. A practical calculation approximates this functional. The orbitals share an effective one-electron operator that depends on the density being constructed, so they must be determined self-consistently (Section B.2).

Energy evaluation from the density matrix We express the Kohn–Sham energy in terms of the density matrix at a fixed geometry $\mathcal{M}$. In the AO basis, the matrix specifies the one-electron density kernel and its real-space diagonal (Szabo & Ostlund, 1996):

$$\gamma_{\mathbf{D}}(\mathbf{r}, \mathbf{r}') = \sum_{\mu\nu} \chi_\mu(\mathbf{r}) D_{\mu\nu} \chi_\nu(\mathbf{r}'), \qquad \rho_{\mathbf{D}}(\mathbf{r}) = \gamma_{\mathbf{D}}(\mathbf{r}, \mathbf{r}). \tag{10}$$

The total energy separates into the standard Kohn–Sham contributions (Kohn & Sham, 1965; Parr & Yang, 1989):

$$E_{\mathrm{KS}}[\mathbf{D}; \mathcal{M}] = T_s[\mathbf{D}] + E_{\mathrm{ne}}[\rho_{\mathbf{D}}; \mathcal{M}] + E_{\mathrm{H}}[\rho_{\mathbf{D}}] + E_{\mathrm{xc}}[\mathbf{D}] + E_{\mathrm{nn}}(\mathcal{M}). \tag{11}$$

Here T_s is the noninteracting kinetic energy, E_{ne} is electron–nuclear attraction, E_{H} is Hartree electrostatic repulsion, and E_{nn} is nuclear repulsion. The notation $E_{\mathrm{xc}}[\mathbf{D}]$ accommodates density and orbital dependence in the chosen exchange–correlation approximation. In an AO implementation, $T_s = \mathrm{Tr}(\mathbf{DT})$ and $E_{\mathrm{ne}} = \mathrm{Tr}(\mathbf{DV}_{\mathrm{ne}})$, where $\mathbf{T}$ and $\mathbf{V}_{\mathrm{ne}}$ are the corresponding one-electron integral matrices. The Hartree term is computed by contracting AO electron-repulsion integrals with $\mathbf{D}$ (Sun et al., 2018).

The semilocal XC contribution is evaluated by numerical quadrature on an integration grid (Becke, 1988; Sun et al., 2018):

$$E_{\mathrm{xc}}^{\mathrm{sl}}[\mathbf{D}] = \int e_{\mathrm{xc}}(\mathbf{x}_{\mathbf{D}}(\mathbf{r}))\, d\mathbf{r} \approx \sum_g w_g\, e_{\mathrm{xc}}(\mathbf{x}_{\mathbf{D}}(\mathbf{r}_g)). \tag{12}$$

Here e_{xc} is an energy per unit volume, $\mathbf{r}_g$ and w_g are quadrature points and weights, and $\mathbf{x}_{\mathbf{D}}$ contains features derived from $\mathbf{D}$ and the AO functions. The local-density approximation (LDA) uses the local density, the generalized-gradient approximation (GGA) also uses its gradient, and meta-GGA may additionally use the kinetic-energy density or density Laplacian (Perdew & Schmidt, 2001). Hybrid functionals also include nonlocal exact exchange computed from $\mathbf{D}$ (Becke, 1993; Luise et al., 2025). Functional evaluation therefore combines matrix contractions with integration of density-derived features.

B.2 Self-consistent field methods

The Kohn–Sham DFT problem can be viewed as the constrained optimization (Schneider et al., 2009; Cancès et al., 2021):

$$\min_{\mathbf{D}=\mathbf{D}^\top} E_{\mathrm{KS}}[\mathbf{D}; \mathcal{M}] \quad \text{subject to} \quad \mathbf{DSD} = 2\mathbf{D}, \qquad \mathrm{Tr}(\mathbf{DS}) = N_e. \tag{13}$$

The constraints restrict the search to density matrices with the prescribed electron number and integer occupations.

Self-consistent field (SCF) solves the resulting nonlinear equations by alternating between constructing a Hamiltonian from the current density and constructing a new density from its occupied orbitals (Kohn & Sham, 1965; Roothaan, 1951). For a fixed geometry and AO basis, the overlap

$\mathbf{S}$ and core Hamiltonian $\mathbf{H}_{\text{core}} = \mathbf{T} + \mathbf{V}_{\text{ne}}$ remain fixed. The Hartree and exchange–correlation contributions depend on $\mathbf{D}$ and must be updated. Starting from an initial guess $\mathbf{D}^{(0)}$, the basic, unaccelerated iteration is

$$\begin{aligned} \mathbf{H}^{(k)} &= \mathbf{H}_{\text{core}} + \mathbf{V}_{\text{Hxc}}[\mathbf{D}^{(k)}], \\ \mathbf{H}^{(k)}\mathbf{C}^{(k)} &= \mathbf{S}\mathbf{C}^{(k)}\boldsymbol{\varepsilon}^{(k)}, \qquad \mathbf{C}^{(k)\top}\mathbf{S}\mathbf{C}^{(k)} = \mathbf{I}, \\ \mathbf{D}^{(k+1)} &= 2\mathbf{C}_o^{(k)}\mathbf{C}_o^{(k)\top}. \end{aligned} \tag{14}$$

Here k indexes the SCF iteration, $\boldsymbol{\varepsilon}^{(k)}$ contains the orbital energies, and $\mathbf{C}_o^{(k)}$ contains the $N_e/2$ lowest-energy orbitals. The matrix $\mathbf{V}_{\text{Hxc}}$ includes Hartree and XC contributions, including nonlocal exact exchange for a hybrid functional. The updated $\mathbf{D}^{(k+1)}$ is then used to rebuild $\mathbf{H}^{(k+1)}$, repeating the sequence $\mathbf{D} \to \mathbf{H} \to \mathbf{C}_o \to \mathbf{D}$. Self-consistency is reached when the occupied orbitals of $\mathbf{H}[\mathbf{D}]$ reproduce the density used to construct it The iterations stop when the energy change and orbital-gradient residual satisfy the prescribed tolerances (Cancès et al., 2021).

B.3 Electron count, occupations, and idempotency in an AO basis

For a restricted closed-shell KS-DFT, each occupied spatial orbital contains two electrons. Substituting the spin-summed AO density $\mathbf{D} = 2\mathbf{C}_o\mathbf{C}_o^\top$ and $\mathbf{C}_o^\top\mathbf{S}\mathbf{C}_o = \mathbf{I}$ gives

$$\mathbf{D}\mathbf{S}\mathbf{D} = 4\mathbf{C}_o(\mathbf{C}_o^\top\mathbf{S}\mathbf{C}_o)\mathbf{C}_o^\top = 2\mathbf{D}, \tag{15}$$

$$\operatorname{Tr}(\mathbf{D}\mathbf{S}) = 2\operatorname{Tr}(\mathbf{C}_o^\top\mathbf{S}\mathbf{C}_o) = N_e. \tag{16}$$

Defining

$$\mathbf{P} = \frac{1}{2}\mathbf{S}^{1/2}\mathbf{D}\mathbf{S}^{1/2} = \mathbf{U}_o\mathbf{U}_o^\top, \qquad \mathbf{U}_o = \mathbf{S}^{1/2}\mathbf{C}_o, \tag{17}$$

gives $\mathbf{U}_o^\top\mathbf{U}_o = \mathbf{I}$ and hence

$$\mathbf{P}^2 = \mathbf{P}, \qquad \operatorname{Tr}(\mathbf{P}) = n_{\text{occ}} = N_e/2. \tag{18}$$

The eigenvalues of $\mathbf{P}$ are the normalized orbital occupations. The identities in Equations (16) and (18) are the same constraints expressed in nonorthogonal and orthonormal coordinates, respectively. Idempotency and the correct electron number do not select the target occupied subspace, so a valid projector can still yield density and energy errors.

C Method details

C.1 Construction of the initial density

We follow the MINAO initialization used in the QH9 preprocessing (Yu et al., 2023a; Kim et al., 2026). PySCF first assigns element-specific atomic occupations in a compact ANO-derived minimal basis, projects the corresponding atomic orbitals into the molecular AO basis, and forms the raw guess $\mathbf{D}_{\text{MINAO}}$ (Almlöf et al., 1982; Van Lenthe et al., 2006; Sun et al., 2018). One Kohn–Sham Hamiltonian is then constructed from this guess, $\mathbf{H}_0 = \mathbf{H}_{\text{core}} + \mathbf{V}_{\text{Hxc}}[\mathbf{D}_{\text{MINAO}}]$, where $\mathbf{H}_{\text{core}}$ contains the one-electron kinetic and nuclear-attraction terms and $\mathbf{V}_{\text{Hxc}}$ is the density-dependent Hartree–exchange–correlation contribution. We obtain the initial density by solving

$$\mathbf{H}_0\mathbf{C}_0 = \mathbf{S}\mathbf{C}_0\boldsymbol{\varepsilon}_0, \qquad \mathbf{D}_0 = 2\mathbf{C}_{0,o}\mathbf{C}_{0,o}^\top, \tag{19}$$

where $\boldsymbol{\varepsilon}_0$ contains the generalized eigenvalues and $\mathbf{C}_{0,o}$ contains the $n_{\text{occ}} = N_e/2$ lowest-energy eigenvectors. Thus $\mathbf{D}_0$ is obtained after one Hamiltonian construction and the assignment of two electrons to each selected orbital. It is not the raw MINAO guess. This construction gives $\mathbf{D}_0$ the required electron count and generalized idempotency up to numerical precision (Section B.3). Equivalently, its orthonormal representation has n_{occ} unit occupation eigenvalues and all remaining eigenvalues equal to zero. Its occupied subspace nevertheless generally differs from that of the converged target.

C.2 CAYLEY PARAMETERIZATION

We implement the Cayley rotation in orthonormal coordinates obtained from the Cholesky factorization $\mathbf{S} = \mathbf{L}\mathbf{L}^\top$. These coordinates differ from the symmetric $\mathbf{S}^{1/2}$ coordinates in the main text only by an orthogonal rotation. The initial density and the symmetric AO-head output are mapped as

$$\mathbf{P}_0 = \frac{1}{2}\mathbf{L}^\top \mathbf{D}_0 \mathbf{L}, \qquad \mathbf{G}_\theta = \frac{1}{2}\mathbf{L}^\top \mathbf{M}_\theta^{\mathrm{AO}} \mathbf{L}. \tag{20}$$

Here $\mathbf{G}_\theta$ parameterizes the rotation rather than a density. We construct a skew-symmetric generator and apply its Cayley transform:

$$\begin{aligned} \mathbf{K}_\theta &= \mathbf{G}_\theta \mathbf{P}_0 - \mathbf{P}_0 \mathbf{G}_\theta, \\ \widetilde{\mathbf{K}}_\theta &= \frac{0.25\,\mathbf{K}_\theta}{\max(1, \|\mathbf{K}_\theta\|_F / \sqrt{n_{\mathrm{AO}}})}, \\ \mathbf{R}_\theta &= \left(\mathbf{I} - \frac{1}{2}\widetilde{\mathbf{K}}_\theta\right)^{-1} \left(\mathbf{I} + \frac{1}{2}\widetilde{\mathbf{K}}_\theta\right). \end{aligned} \tag{21}$$

The rescaling ensures $\|\widetilde{\mathbf{K}}_\theta\|_F \le 0.25\sqrt{n_{\mathrm{AO}}}$. Only the occupied–virtual blocks of $\mathbf{G}_\theta$ relative to $\mathbf{P}_0$ contribute to the commutator; the implementation computes it directly without constructing the occupied and virtual eigenvectors (Section C.4.1).

Finally, we rotate the initial projector and recover the AO density:

$$\mathbf{P}_\theta = \mathbf{R}_\theta \mathbf{P}_0 \mathbf{R}_\theta^\top, \qquad \mathbf{D}_\theta = 2\mathbf{L}^{-\top} \mathbf{P}_\theta \mathbf{L}^{-1}. \tag{22}$$

The inverse map is evaluated using triangular solves. Because $\widetilde{\mathbf{K}}_\theta$ is skew-symmetric, $\mathbf{R}_\theta$ is orthogonal. For a valid initial projector and an invertible $\mathbf{L}$ on the retained AO subspace, this construction preserves the occupation spectrum, electron number, and idempotency exactly in exact arithmetic, up to numerical errors in implementation.

Relation to the matrix exponential. For a generic skew-symmetric argument $\mathbf{K}$, the map $\mathrm{cay}(\mathbf{K}) = (\mathbf{I} - \mathbf{K}/2)^{-1}(\mathbf{I} + \mathbf{K}/2)$ is the $[1/1]$ Padé approximant to $\exp(\mathbf{K})$ (Shepard et al., 2015). Expanding around $\mathbf{K} = \mathbf{0}$ gives

$$\begin{aligned} \mathrm{cay}(\mathbf{K}) &= \mathbf{I} + \mathbf{K} + \tfrac{1}{2}\mathbf{K}^2 + \tfrac{1}{4}\mathbf{K}^3 + O(\|\mathbf{K}\|^4), \\ \exp(\mathbf{K}) &= \mathbf{I} + \mathbf{K} + \tfrac{1}{2}\mathbf{K}^2 + \tfrac{1}{6}\mathbf{K}^3 + O(\|\mathbf{K}\|^4). \end{aligned} \tag{23}$$

The maps agree through second order, with $\mathrm{cay}(\mathbf{K}) - \exp(\mathbf{K}) = \frac{1}{12}\mathbf{K}^3 + O(\|\mathbf{K}\|^4)$ as $\mathbf{K} \to \mathbf{0}$. For every finite real skew-symmetric $\mathbf{K}$, however, $\mathbf{I} - \mathbf{K}/2$ is invertible and $\mathrm{cay}(\mathbf{K})^\top = \mathrm{cay}(\mathbf{K})^{-1}$. Thus orthogonality and the resulting preservation of a valid initial projector's spectrum, trace, and idempotency hold exactly in exact arithmetic, independently of the local approximation error. Our model constructs $\mathbf{K}_\theta$ from the learned AO output and uses $\mathbf{R}_\theta = \mathrm{cay}(\widetilde{\mathbf{K}}_\theta)$ as its rotation parameterization. Its constraint guarantees therefore require no small-generator assumption.

C.2.1 ROTATION EQUIVARIANCE OF THE CAYLEY HEAD

We show that the Cayley construction in Equations (20) and (21) preserves the rotation equivariance of its AO inputs and learned AO output. The proof covers both symmetric and Cholesky orthogonalization. Here a spatial rotation acts on the molecular geometry and AO basis. $\mathbf{R}_\theta$ remains the learned occupied-subspace update.

Assumptions and claim. Let $g \in \mathrm{SO}(3)$ rotate the molecular geometry $\mathcal{M}$, and let $\mathbf{U}_g \in \mathbb{R}^{n_{\mathrm{AO}} \times n_{\mathrm{AO}}}$ be its orthogonal representation in the real spherical AO basis, $\mathbf{U}_g^\top \mathbf{U}_g = \mathbf{I}$. Atom ordering is fixed, and each angular-momentum shell contains all of its components. A subscript g denotes a quantity evaluated on the rotated geometry with all AO inputs transformed consistently. Assume that $\mathbf{S}$ is positive definite and the symmetric AO inputs obey

$$\begin{aligned} \mathbf{S}_g &= \mathbf{U}_g \mathbf{S} \mathbf{U}_g^\top, \\ \mathbf{D}_{0,g} &= \mathbf{U}_g \mathbf{D}_0 \mathbf{U}_g^\top, \\ \mathbf{M}_{\theta,g}^{\mathrm{AO}} &= \mathbf{U}_g \mathbf{M}_\theta^{\mathrm{AO}} \mathbf{U}_g^\top. \end{aligned} \tag{24}$$

In particular, the last identity is an assumption on the AO predictor as a function of all its inputs. Then, in exact arithmetic, the Cayley density obeys

$$\boxed{\mathbf{D}_{\theta,g} = \mathbf{U}_g \mathbf{D}_\theta \mathbf{U}_g^\top.} \tag{25}$$

Proof. Choose any invertible factors $\mathbf{A}$ and $\mathbf{A}_g$ satisfying $\mathbf{A}^\top\mathbf{A} = \mathbf{S}$ and $\mathbf{A}_g^\top\mathbf{A}_g = \mathbf{S}_g$. The induced transformation between their orthonormal coordinates is $\mathbf{W}_g = \mathbf{A}_g\mathbf{U}_g\mathbf{A}^{-1}$. It is orthogonal because

$$\mathbf{W}_g^\top\mathbf{W}_g = \mathbf{A}^{-\top}\mathbf{U}_g^\top\mathbf{S}_g\mathbf{U}_g\mathbf{A}^{-1} = \mathbf{A}^{-\top}\mathbf{S}\mathbf{A}^{-1} = \mathbf{I}. \tag{26}$$

Using $\mathbf{A}_g\mathbf{U}_g = \mathbf{W}_g\mathbf{A}$, the two inputs to the commutator transform in the same orthonormal coordinates:

$$\begin{aligned} \mathbf{P}_{0,g} &= \tfrac{1}{2}\mathbf{A}_g\mathbf{D}_{0,g}\mathbf{A}_g^\top = \mathbf{W}_g\mathbf{P}_0\mathbf{W}_g^\top, \\ \mathbf{G}_{\theta,g} &= \tfrac{1}{2}\mathbf{A}_g\mathbf{M}^{\mathrm{AO}}_{\theta,g}\mathbf{A}_g^\top = \mathbf{W}_g\mathbf{G}_\theta\mathbf{W}_g^\top. \end{aligned} \tag{27}$$

Consequently, multiplication and subtraction give

$$\begin{aligned} \mathbf{K}_{\theta,g} &= \mathbf{G}_{\theta,g}\mathbf{P}_{0,g} - \mathbf{P}_{0,g}\mathbf{G}_{\theta,g} \\ &= \mathbf{W}_g(\mathbf{G}_\theta\mathbf{P}_0 - \mathbf{P}_0\mathbf{G}_\theta)\mathbf{W}_g^\top = \mathbf{W}_g\mathbf{K}_\theta\mathbf{W}_g^\top. \end{aligned} \tag{28}$$

The Frobenius norm is invariant under orthogonal conjugation, so $\|\mathbf{K}_{\theta,g}\|_F = \|\mathbf{K}_\theta\|_F$. The rescaling in Equation (21) therefore preserves the same law:

$$\widetilde{\mathbf{K}}_{\theta,g} = \mathbf{W}_g\widetilde{\mathbf{K}}_\theta\mathbf{W}_g^\top. \tag{29}$$

Since $\widetilde{\mathbf{K}}_\theta$ is real and skew-symmetric, $\mathbf{I} - \frac{1}{2}\widetilde{\mathbf{K}}_\theta$ is invertible: a real vector in its nullspace would satisfy $0 = \mathbf{v}^\top(\mathbf{I} - \frac{1}{2}\widetilde{\mathbf{K}}_\theta)\mathbf{v} = \|\mathbf{v}\|_2^2$. Orthogonal conjugation commutes with the inverse, so

$$\begin{aligned} \mathbf{R}_{\theta,g} &= \big[\mathbf{W}_g(\mathbf{I} - \tfrac{1}{2}\widetilde{\mathbf{K}}_\theta)\mathbf{W}_g^\top\big]^{-1}\big[\mathbf{W}_g(\mathbf{I} + \tfrac{1}{2}\widetilde{\mathbf{K}}_\theta)\mathbf{W}_g^\top\big] \\ &= \mathbf{W}_g\mathbf{R}_\theta\mathbf{W}_g^\top, \\ \mathbf{P}_{\theta,g} &= \mathbf{R}_{\theta,g}\mathbf{P}_{0,g}\mathbf{R}_{\theta,g}^\top = \mathbf{W}_g\mathbf{P}_\theta\mathbf{W}_g^\top. \end{aligned} \tag{30}$$

Finally, the inverse AO map cancels the intermediate coordinate change:

$$\begin{aligned} \mathbf{D}_{\theta,g} &= 2\mathbf{A}_g^{-1}\mathbf{W}_g\mathbf{P}_\theta\mathbf{W}_g^\top\mathbf{A}_g^{-\top} \\ &= \mathbf{U}_g(2\mathbf{A}^{-1}\mathbf{P}_\theta\mathbf{A}^{-\top})\mathbf{U}_g^\top = \mathbf{U}_g\mathbf{D}_\theta\mathbf{U}_g^\top. \end{aligned} \tag{31}$$

This establishes Equation (25).

Scope of the guarantee. Equivariance requires the initial density and learned AO output to satisfy Equation (24), and Cayley alone does not ensure this property. Projector validity is required for idempotency, but not for equivariance. The argument extends to $\mathrm{O}(3)$ when the inputs also obey the AO parity action.

C.3 Model architecture and output heads

The backbone is based on QHFlow2/MALOQ (Kim et al., 2026; Kaniselvan et al., 2026b). Hyperparameters are listed in Table 8. The $\Delta\mathbf{D}$, Cayley, and MLH models differ in readout conditioning and output maps.

1. Inputs and encoding. Let $\mathcal{M} = \{Z_i, \mathbf{r}_i\}$ denote the molecular geometry. All atom pairs $i \neq j$ are connected by directed edges. The edge descriptor $\mathbf{e}_{ij} \in \mathbb{R}^{768}$ concatenates radial features of $d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$ and source/destination element embeddings. The encoder combines geometry, atomic embeddings, and contractions of the input matrices:

$$\{\mathbf{h}_i^{(0)}\}_i = \mathrm{Enc}_\theta(\mathcal{M}, \mathbf{H}_0, \mathbf{S}), \qquad \mathbf{h}_i^{(k)} \in \mathcal{V}_{192}, \quad \mathcal{V}_C = \bigoplus_{\ell=0}^{4} \mathbb{R}^{(2\ell+1)\times C}.$$

Here k indexes node layers and (ℓ, m) angular components. The encoder uses regression time $t = 0$ and charge/spin embeddings. The AO readout uses $\mathbf{B}_0 = \mathbf{D}_0$ for density models and $\mathbf{B}_0 = \mathbf{H}_0$ for MLH.

2. Residual backbone and pair features. Each of the three node blocks applies an edge-message operator Φ_k and a grid feed-forward operator Ψ_k, each with a residual addition:

$$\begin{aligned} \mathbf{u}_j^{(k)} &= \mathbf{h}_j^{(k)} + \sum_{i \neq j} \Phi_k(\mathbf{h}_i^{(k)}, \mathbf{h}_j^{(k)}, \mathbf{e}_{ij}), \\ \mathbf{h}_j^{(k+1)} &= \mathbf{u}_j^{(k)} + \Psi_k(\mathbf{u}_j^{(k)}), \qquad k = 0, 1, 2. \end{aligned} \tag{32}$$

Φ_k applies pre-normalization, charge/spin injection into scalar channels, edge-frame $\mathrm{SO}(2)$ convolutions, gating, and a distance envelope. Ψ_k applies pre-normalization, a pointwise MLP on a spherical grid, and projection back to angular features. Each block has independent weights. Edge frames and molecular embeddings are implicit in the notation.

Two pair blocks $\mathcal{P}_p$ process the final node features in parallel. Their outputs are summed before the off-atom bottleneck:

$$\begin{aligned} \mathbf{f}_{ii} &= \mathcal{B}_{\mathrm{on}}(\mathbf{h}_i^{(3)}), \\ \mathbf{f}_{ij} &= \mathcal{B}_{\mathrm{off}}\left(\sum_{p=1}^{2} \mathcal{P}_p(\mathbf{h}_i^{(3)}, \mathbf{h}_j^{(3)}, \mathbf{e}_{ij}) \right), \qquad i \neq j. \end{aligned}$$

Each $\mathcal{P}_p$ normalizes the endpoint features, modulates the edge descriptor using scalar-channel products, and applies an edge message followed by a normalized grid FFN without a residual bypass. $\mathcal{B}_{\mathrm{on/off}}$ applies spherical RMS normalization and a $192 \to 96$ channel projection at each angular order, giving $\mathbf{f}_{ij} \in \mathcal{V}_{96}$.

3. Conditioned AO readout. Let $\mathbf{z}_{ij}$ collect scalar endpoint features and learned scalar contractions of $\mathbf{S}_{ij}$ and $(\mathbf{B}_0)_{ij}$. Separate on/off-atom MLPs map $\mathbf{z}_{ij}$ to path weights w and scalar biases β. For radial shells a, b with angular momenta ℓ_a, ℓ_b, the resulting coefficients are expanded into AO blocks:

$$\begin{aligned} \alpha_{ij,ab}^{\ell m} &= \frac{1}{96} \left[\sum_{c=1}^{96} w_{ab\ell c}(\mathbf{z}_{ij})(\mathbf{f}_{ij}^{\ell})_{mc} + \delta_{\ell 0} \beta_{ab}(\mathbf{z}_{ij}) \right], \\ [\widetilde{\mathbf{M}}_{ij}^{ab}]_{m_a m_b} &= \sum_{\ell=|\ell_a - \ell_b|}^{\ell_a + \ell_b} \sum_{m=-\ell}^{\ell} T_{\ell_a \ell_b \ell}^{m_a m_b m} \alpha_{ij,ab}^{\ell m}. \end{aligned} \tag{33}$$

Here c indexes feature channels and m_a, m_b index AO components. T is a fixed real Wigner-$3j$ tensor, and $1/96$ is the channel-reduction scale. The padded basis is $3s + 2p + 1d$ (14 AOs). H retains $2s + 1p$ (5 AOs), while C/N/O/F retain all 14. Thus $\ell \leq 4$ covers every shell pair. The blocks are symmetrized and assembled as

$$\mathbf{M}_\theta^{\mathrm{raw}} = \mathrm{Assemble}_Z \left\{ \frac{\widetilde{\mathbf{M}}_{ij} + \widetilde{\mathbf{M}}_{ji}^{\top}}{2} \right\}_{i,j} \in \mathbb{R}^{n_{\mathrm{AO}} \times n_{\mathrm{AO}}}.$$

Assembly applies element masks and physical AO ordering. Output blocks have no distance mask.

4. Symmetry and output maps. For $\mathcal{X} = (\mathcal{M}, \mathbf{H}_0, \mathbf{S}, \mathbf{B}_0)$, let $\mathcal{I}\mathcal{X}$ invert the geometry and transform each AO input by the parity matrix $\mathbf{J}_{\mathcal{I}}$. We apply inversion averaging to the matrix output:

$$\mathbf{M}_\theta^{\mathrm{AO}}(\mathcal{X}) = \frac{1}{2} \left[\mathbf{M}_\theta^{\mathrm{raw}}(\mathcal{X}) + \mathbf{J}_{\mathcal{I}}^{\top} \mathbf{M}_\theta^{\mathrm{raw}}(\mathcal{I}\mathcal{X}) \mathbf{J}_{\mathcal{I}} \right].$$

Both branches share weights. This adds inversion averaging to the backbone's proper-rotation convention. The internal all-even irrep tags do not specify physical AO parities (Section C.5).

The output maps are

$$\begin{aligned} \Delta\mathbf{D}: &\quad \mathbf{D}_\theta = \mathbf{D}_0 + \mathbf{M}_\theta^{\mathrm{AO}}, \\ \text{Cayley}: &\quad \mathbf{D}_\theta = 2\mathbf{L}^{-\top} \mathbf{R}_\theta \mathbf{P}_0 \mathbf{R}_\theta^{\top} \mathbf{L}^{-1}, \\ \text{MLH}: &\quad \mathbf{H}_\theta = \mathbf{H}_0 + \mathbf{M}_\theta^{\mathrm{AO}}, \quad \mathbf{D}_\theta = 2\mathbf{C}_o \mathbf{C}_o^{\top}. \end{aligned} \tag{34}$$

For MLH, $\mathbf{C}_o$ contains the lowest $N_e/2$ generalized eigenvectors satisfying $\mathbf{H}_\theta \mathbf{C}_o = \mathbf{S}\mathbf{C}_o \boldsymbol{\varepsilon}_o$ and $\mathbf{C}_o^{\top} \mathbf{S} \mathbf{C}_o = \mathbf{I}$. Overlap filtering is described in Section E.3. For Cayley, Inversion averaging is applied before the Cayley map. The resulting density is not averaged or subjected to spectral projection or McWeeny purification. Occupation, electron-number, and idempotency preservation hold in exact arithmetic under the conditions in Section C.2.

C.3.1 OPERATOR DETAILS

Radial encoding. For distance d, the Gaussian radial features and message envelope are

$$g_n(d) = \exp\left[-\frac{(d-\mu_n)^2}{2\sigma_r^2}\right], \qquad n = 0, \dots, 511,$$
$$f_{\text{cut}}(d) = \mathbf{1}_{u<1}(1 - 21u^5 + 35u^6 - 15u^7), \quad u = \frac{d}{15\,\text{Å}},$$

where $\mu_n = 15n/511$ Å and $\sigma_r = 30/511$ Å. The encoder places radial-MLP outputs in the edge-frame $m = 0$ components and sets the remaining components to zero. It then rotates the features back and sums the envelope-weighted contributions with scale $1/5$.

Edge messages and gating. Edge messages apply two $\mathrm{SO}(2)$ convolutions ($384 \to 192 \to 192$) to concatenated endpoint features in the edge frame. Gating between the convolutions uses SiLU for $\ell = 0$. Higher-order channels use sigmoid gates from the first convolution, shared over m. Messages rotate back before envelope scaling.

Grid feed-forward and pair modulation. Grid FFNs use three bias-free 192×192 channel maps with SiLU after the first two. Pair modulation scales $\mathbf{e}_{ij}$ using a $192 \to 192 \to 768$ SiLU MLP applied to the elementwise product of normalized endpoint scalars.

Readout conditioning. Let $\mathbf{s}_i$ denote the scalar channels of $\mathbf{f}_{ii}$. Let $\mathbf{c}_{ij}$ concatenate the contractions of $\mathbf{S}_{ij}$ and $(\mathbf{B}_0)_{ij}$. With brackets denoting concatenation, the coefficient MLP inputs are

$$\mathbf{z}_{ii} = [\mathbf{s}_i, \mathbf{c}_{ii}] \in \mathbb{R}^{288},$$
$$\mathbf{z}_{ij} = [\mathbf{s}_i, \mathbf{s}_j, \mathbf{c}_{ij}, \mathbf{c}_{ij}] \in \mathbb{R}^{576}, \qquad i \neq j.$$

Each weight/bias MLP has one width-96 hidden layer with SiLU and a zero-initialized final layer.

C.4 DENSITY-MATRIX AND OCCUPIED–VIRTUAL OBJECTIVES

DM loss. For each minibatch, let $\mathcal{I}_{\text{on}}$ and $\mathcal{I}_{\text{off}}$ collect the active entries of all on-atom and ordered off-atom AO blocks, respectively. Padded AO slots are removed using element-dependent orbital masks. With $e_k = (D_\theta)_k - D_k$, the density term is the sum of two batch-level masked means:

$$\mathcal{L}_{\text{DM}} = \frac{1}{|\mathcal{I}_{\text{on}}|} \sum_{k \in \mathcal{I}_{\text{on}}} |e_k| + \frac{1}{|\mathcal{I}_{\text{off}}|} \sum_{k \in \mathcal{I}_{\text{off}}} |e_k|. \tag{35}$$

Thus each active entry has equal weight within its block type, while the on-atom and off-atom terms receive equal outer weights.

OV loss. For molecule m, let $n_{\text{AO},m}$ denote its number of active AOs. We retain the corresponding AO submatrices and compute the Cholesky factorization $\mathbf{S}_m = \mathbf{L}_m \mathbf{L}_m^\top$. With $\mathbf{A}_m = \mathbf{L}_m^\top$, the target and predicted densities in these coordinates are

$$\mathbf{P}_m = \frac{1}{2}\mathbf{A}_m \mathbf{D}_m \mathbf{A}_m^\top, \qquad \mathbf{P}_{\theta,m} = \frac{1}{2}\mathbf{A}_m \mathbf{D}_{\theta,m} \mathbf{A}_m^\top. \tag{36}$$

All $n_{\text{occ},m}$ occupied and $n_{\text{AO},m} - n_{\text{occ},m}$ virtual eigenvectors of the fixed target $\mathbf{P}_m$ are retained. No frontier-orbital window is used. The target eigenspaces are detached, and gradients pass only through $\mathbf{P}_{\theta,m}$. For a minibatch of N_{batch} molecules, we form $\mathbf{P}^\theta_{ov,m} = \mathbf{U}_{o,m}^\top \mathbf{P}_{\theta,m} \mathbf{U}_{v,m}$ and concatenate every entry of $\{\mathbf{P}^\theta_{ov,m}\}_{m=1}^{N_{\text{batch}}}$ into $\{b_k\}_{k=1}^{N_{\text{OV}}}$. The loss is

$$\mathcal{L}_{\text{OV}} = \frac{1}{N_{\text{OV}}} \sum_k \left(|b_k| + b_k^2\right), \qquad \mathcal{L} = \mathcal{L}_{\text{DM}} + w_{\text{OV}} \mathcal{L}_{\text{OV}}. \tag{37}$$

This entrywise concatenation weights molecule m in proportion to $n_{\text{occ},m}(n_{\text{AO},m} - n_{\text{occ},m})$. We combine absolute and squared OV errors to penalize both small residuals and larger deviations in the chosen target basis.

C.4.1 Occupied-subspace error and principal angles

Let $\Delta\mathbf{P} = \mathbf{P}_\theta - \mathbf{P}$ be the error in the normalized predicted density, and let $\mathbf{U} = [\mathbf{U}_o\ \mathbf{U}_v]$ be the target basis defined in Section 2. In this basis, the error decomposes as

$$\mathbf{U}^\top \Delta\mathbf{P}\mathbf{U} = \begin{pmatrix} \Delta\mathbf{P}_{oo} & \Delta\mathbf{P}_{ov} \\ \Delta\mathbf{P}_{vo} & \Delta\mathbf{P}_{vv} \end{pmatrix}. \tag{38}$$

Since the target has identity and zero blocks in Equation (3), $\Delta\mathbf{P}_{oo} = \mathbf{P}^\theta_{oo} - \mathbf{I}$, $\Delta\mathbf{P}_{ov} = \mathbf{P}^\theta_{ov}$, and $\Delta\mathbf{P}_{vv} = \mathbf{P}^\theta_{vv}$, using the prediction blocks of Equation (5). Rotating the basis only within the occupied or virtual subspace leaves $\mathbf{P}$ unchanged. For a nearby valid projector, the first-order change produced by a small skew-symmetric generator $\mathbf{K}_\delta$ is $\delta\mathbf{P} = [\mathbf{K}_\delta, \mathbf{P}]$. In the target basis, this change lies in the OV and VO blocks:

$$\mathbf{U}^\top \delta\mathbf{P}\mathbf{U} = \mathbf{U}^\top [\mathbf{K}_\delta, \mathbf{P}]\mathbf{U} = \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0} \end{pmatrix}, \qquad \mathbf{K}_\delta^\top = -\mathbf{K}_\delta. \tag{39}$$

Here $\mathbf{B}$ is the first-order OV block. The local rotation is relative to the target, not the learned rotation of the initial projector $\mathbf{P}_0$. The diagonal OO and VV blocks of two nearby valid projectors change only at higher order. For an unconstrained prediction, however, these blocks can also contain occupation-spectrum error.

Principal angles make the relation precise. Let φ_i be the principal angles between two valid occupied subspaces with the same electron count. *i.e.* $\cos\varphi_i = \sigma_i\big(\mathbf{U}_o^\top \mathbf{U}_o^\theta\big)$. In principal-vector coordinates,

$$\|\mathbf{P}_\theta - \mathbf{P}\|_F^2 = 2\sum_i \sin^2\varphi_i, \qquad \|\Delta\mathbf{P}_{ov}\|_F^2 = \sum_i \sin^2\varphi_i \cos^2\varphi_i, \qquad \|\Delta\mathbf{P}_{oo}\|_F^2 = \sum_i \sin^4\varphi_i.$$

Thus, OV error is the leading local measure of a small subspace rotation, whereas OO error is higher order once the occupation spectrum is valid.

Why OV supervision is used for auxiliary. The OV norm is not a global subspace distance, since its contribution also vanishes at a principal angle of $\pi/2$. For example, exchange one target occupied unit vector $\mathbf{u}_o$ with one virtual unit vector $\mathbf{u}_v$:

$$\mathbf{P}_{\text{swap}} = \mathbf{P} - \mathbf{u}_o\mathbf{u}_o^\top + \mathbf{u}_v\mathbf{u}_v^\top, \qquad \mathbf{U}_o^\top \mathbf{P}_{\text{swap}}\mathbf{U}_v = \mathbf{0}, \qquad \|\mathbf{P}_{\text{swap}} - \mathbf{P}\|_F^2 = 2. \tag{40}$$

This prediction preserves the electron number and idempotency but occupies the wrong state. Unlike a basis rotation within either subspace, the exchange changes the occupied subspace and hence the density. We therefore retain the full AO density-matrix loss and use the OV term to emphasize local alignment, as in Equation (6).

C.4.2 Relation to solver-aligned initialization learning

SAIL and our OV objective both use occupied–virtual directions that change the occupied subspace. They supervise different quantities: orbital stationarity along an SCF trajectory and density error in a fixed target basis, respectively.

Solver-trajectory supervision. SAIL learns an initial density $\mathbf{D}_\theta^{(0)}$ by differentiating through T SCF steps, including the convergence accelerator (Eberhard et al., 2026). Let $\mathbf{H}[\mathbf{D}^{(t)}]$ be the physical Hamiltonian (Fock matrix) at step t. Let $\mathbf{C}_o^{(t)}$ and $\mathbf{C}_v^{(t)}$ contain the current occupied and virtual AO orbital coefficients defining $\mathbf{D}^{(t)}$, with $\mathbf{C}^{(t)\top}\mathbf{S}\mathbf{C}^{(t)} = \mathbf{I}$. Using SAIL's orbital-gradient normalization, its objective is

$$\mathbf{g}_{\text{OV}}^{(t)} = \mathbf{C}_o^{(t)\top}\mathbf{H}[\mathbf{D}^{(t)}]\mathbf{C}_v^{(t)}, \qquad \mathcal{L}_{\text{SAIL}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\|\mathbf{g}_{\text{OV}}^{(t)}\|_F}{\sqrt{n_{\text{occ}} n_{\text{virt}}}}. \tag{41}$$

Here $n_{\text{virt}} = n_{\text{AO}} - n_{\text{occ}}$ and $\mathbf{D}^{(t)} = 2\mathbf{C}_o^{(t)}\mathbf{C}_o^{(t)\top}$. Before convergence, these orbitals need not diagonalize the Hamiltonian rebuilt from $\mathbf{D}^{(t)}$. Its OV block measures departure from orbital stationarity at the current iterate.

Comparison with our OV loss. SAIL evaluates Hamiltonian OV blocks in the current orbital basis along an SCF trajectory, whereas our auxiliary loss penalizes the predicted density block $\mathbf{P}^{\theta}_{ov}$ in the fixed target basis (Equation (5)). SAIL uses supervised pretraining followed by solver-aligned fine-tuning through Hamiltonian construction and SCF iterations, without ground-state labels. Our objective uses fixed target densities and matrix operations alongside the full density-matrix loss, with no Hamiltonian construction or SCF unrolling during loss evaluation.

Comparison of SCF acceleration with SAIL on QM40. We compare our QM40 results with the reported SAIL results (Eberhard et al., 2026, Table 1, v2) (Table 6). Here, we followed the QM40 sampling protocol of Eberhard et al. (2026). Cayley with $w_{\mathrm{OV}} = 10$ achieves ERIC 0.70, compared with 0.72 for SAIL's density-matrix model, while requiring no SCF solver evaluations during training. Both indicate roughly 30% lower effective iteration cost relative to MINAO when transferring from QM9-based training to QM40. Baseline denotes ground-state supervision, including our $w_{\mathrm{OV}} = 0$ models, and OV loss denotes $w_{\mathrm{OV}} = 10$. The architectures, training and evaluation protocols differ, and identical QM40 molecule sets have not been verified. We retain each study's ERIC convention, with ours counting an additional initialization Fock build (Equation (48)).

Table 6: **Reported QM40 SCF results.** ERIC uses each study's MINAO baseline and evaluation protocol.

Model	Target	ERIC↓		
		Baseline	SAIL	OV loss
Reported by Eberhard et al. (2026)				
Density coefficients	$\mathbf{c}$	0.90	0.75	—
Hamiltonian	$\mathbf{H}$	1.65	0.74	—
Density matrix	$\mathbf{D}$	1.05	0.72	—
Ours				
$\Delta\mathbf{D}$	$\mathbf{D}$	0.77	—	0.74
$\Delta\mathbf{D}$ + McW2	$\mathbf{D}$	0.73	—	0.70
Cayley	$\mathbf{D}$	0.71	—	0.70

C.5 Inversion equivariance by group averaging

Inversion and group averaging. Inversion maps $\mathbf{r}_i \mapsto -\mathbf{r}_i$ while preserving atomic numbers and atom ordering. Each atom-centered spherical AO acquires the parity $(-1)^{\ell_\mu}$, giving

$$\mathbf{J}_{\mathcal{I}} = \mathrm{diag}\big((-1)^{\ell_1}, \ldots, (-1)^{\ell_{n_{\mathrm{AO}}}}\big), \qquad \mathbf{J}_{\mathcal{I}}^{\top} = \mathbf{J}_{\mathcal{I}}^{-1} = \mathbf{J}_{\mathcal{I}}. \tag{42}$$

The inverted input $\mathcal{I}\mathcal{X}$ transforms the geometry and every AO-matrix input, including $\mathbf{S}$ and $\mathbf{D}_0$, by

$$\mathcal{I}_{\mathrm{AO}}[\mathbf{Y}] = \mathcal{I}_{\mathrm{AO}}^{-1}[\mathbf{Y}] = \mathbf{J}_{\mathcal{I}}\mathbf{Y}\mathbf{J}_{\mathcal{I}}^{\top}. \tag{43}$$

Using shared parameters, we map the inverted branch back to the original AO frame and average:

$$\begin{aligned} \mathbf{M}_{\theta}^{\mathrm{inv}}(\mathcal{X}) &= \mathbf{J}_{\mathcal{I}}^{\top}\mathbf{M}_{\theta}^{\mathrm{raw}}(\mathcal{I}\mathcal{X})\mathbf{J}_{\mathcal{I}}, \\ \mathbf{M}_{\theta}^{\mathrm{AO}}(\mathcal{X}) &= \tfrac{1}{2}\left[\mathbf{M}_{\theta}^{\mathrm{raw}}(\mathcal{X}) + \mathbf{M}_{\theta}^{\mathrm{inv}}(\mathcal{X})\right]. \end{aligned} \tag{44}$$

The averaged matrix is then supplied to the model-specific output map. Equivalently, the parity action can be applied before linear AO expansion by multiplying each shell-pair coefficient by $(-1)^{\ell_a+\ell_b}$, irrespective of its coupled angular degree.

Equivariance of the averaged output. Because $\mathcal{I}^2\mathcal{X} = \mathcal{X}$ and $\mathbf{J}_{\mathcal{I}}^2 = \mathbf{I}$, inversion exchanges the two branches:

$$\begin{aligned} \mathbf{M}_{\theta}^{\mathrm{AO}}(\mathcal{I}\mathcal{X}) &= \tfrac{1}{2}\left[\mathbf{M}_{\theta}^{\mathrm{raw}}(\mathcal{I}\mathcal{X}) + \mathbf{J}_{\mathcal{I}}\mathbf{M}_{\theta}^{\mathrm{raw}}(\mathcal{X})\mathbf{J}_{\mathcal{I}}^{\top}\right] \\ &= \mathbf{J}_{\mathcal{I}}\mathbf{M}_{\theta}^{\mathrm{AO}}(\mathcal{X})\mathbf{J}_{\mathcal{I}}^{\top}. \end{aligned} \tag{45}$$

Thus the average is inversion equivariant even when the raw predictor is not. Inversion commutes with proper rotations, and $\mathbf{J}_{\mathcal{I}}$ commutes with their AO representation $\mathbf{U}_g$ (Section C.2.1), so averaging also preserves the backbone's $\mathrm{SO}(3)$ equivariance. Since every improper orthogonal transformation is inversion composed with a proper rotation, the averaged output is $\mathrm{O}(3)$ equivariant.

C.6 COMPUTATIONAL COMPLEXITY

We summarize the asymptotic cost of the matrix operations beyond the shared backbone, excluding initial-density preparation and downstream DFT evaluation. Let $n = n_{\mathrm{AO}}$ be the number of active AOs in one molecule. The analysis assumes standard dense linear algebra and distinguishes density prediction, additional OV supervision, and optional post-processing (Table 7).

Table 7: Asymptotic matrix-operation costs per molecule, with $n = n_{\mathrm{AO}}$ and p McWeeny steps. Costs for the backbone, initial-density preparation, and downstream DFT are excluded.

Component	Training	Inference	Leading operation
Prediction and density construction			
Additive $\Delta\mathbf{D}$	$O(n^2)$	$O(n^2)$	Entrywise AO correction
Cayley	$O(n^3)$	$O(n^3)$	Matrix products and linear solve
MLH $\mathbf{H} \to \mathbf{D}$	$O(n^2)$	$O(n^3)$	Generalized eigendecomposition at inference
Additional training objective			
OV loss	$O(n^3)$	—	Overlap factorization, target eigendecomposition, and projection
Optional post-processing			
McWeeny (p steps)	—	$O(pn^3)$	Matrix polynomial evaluation

D EXPERIMENTAL SETUP AND EVALUATION

D.1 COMMON EXPERIMENTAL SETUP

Model architectures are described in Section C.3. The corresponding hyperparameters and training configurations are summarized in Tables 8 to 10.

QH9 data and density-matrix targets. We use the QH9Stable random split and retain it across all head and loss ablations (Yu et al., 2023a). Dataset statistics are summarized in Table 11. The original QH9 Hamiltonians were generated using PySCF at the B3LYP/def2-SVP level, with integration grid level 3, an SCF energy tolerance of 10^{-13} Ha, a gradient threshold of 3.16×10^{-5}, and an eight-vector DIIS history (Yu et al., 2023a; Sun et al., 2018; Weigend & Ahlrichs, 2005). Because the release does not include density matrices, we generate AO density targets $\mathbf{D}$ at the original geometries using the same basis and the source B3LYP/VWN5 convention, specified as `b3lyp5` in current PySCF versions.

Downstream DFT evaluation. We evaluate predicted and target densities at the B3LYP/VWN5 and def2-SVP level using GPU4PySCF 1.6.1, integration grid level 3, NWChem pruning, and no density fitting. Predicted densities are symmetrized in FP64 before evaluation. Energy and force errors compare calculations from predicted and target densities under identical settings, rather than against the stored dataset energy and force fields.

QH9 SCF initialization. We use 200 test molecules sampled with seed 42, with 50 each at 6, 7, 8, and 9 heavy atoms and no energy-error filtering. Each model uses the density construction and post-processing of its corresponding row in Tables 1 and 2. All initializations, including MINAO, use the same geometries and FP64 overlaps. We use PySCF 2.12.1 with the shared DFT settings, an energy tolerance of 10^{-9} Ha, an orbital-gradient threshold of $\sqrt{10^{-9}}$, and at most 50 cycles. CDIIS uses eight history vectors and starts at cycle 1. The solver's overlap threshold remains 10^{-6}, independently of the threshold used to construct an MLH initial density. ERIC averages the paired ratios in Equation (48) over the 200 molecules and three repeats, each with a fresh MINAO baseline.

D.2 QM40 EXTRAPOLATION PROTOCOLS

QM40 evaluation. QM40 contains 162,954 neutral, drug-like molecules drawn from ZINC with B3LYP/6-31G(2df,p)-optimized geometries (Madushanka et al., 2024). We restrict evaluation to H, C, N, O, and F and sample at most 100 molecules per heavy-atom count using seed 0. The resulting 2,243 closed-shell molecules span 10–37 heavy atoms, beyond QH9's 1–9 range (Table 12 and Figure 8). The element filter leaves no molecules with 20 or 21 heavy atoms. All models are trained on QH9 without QM40 fine-tuning, and reported averages weight molecules equally within this size-stratified subset.

At the supplied geometries, we generate targets using PySCF 2.10.0 and the shared DFT settings in Section D.1. All target calculations converge within 50 SCF cycles at an energy tolerance of 10^{-9} Ha. The MINAO-derived density $\mathbf{D}_0$ follows Section C.1, with the initial Hamiltonian using B3LYP/VWN3 to match the stored QH9 convention and targets using B3LYP/VWN5. For SCF acceleration, we initialize with the densities used for frozen-density evaluation and compare against MINAO under matched FP64 solver settings (Section D.1). The gradient threshold remains at the library default, and convergence is determined by the solver flag.

Table 8: Architecture of the 65.9M QH9 models used for QM40 transfer.

Hyperparameter	Value
Inputs and graph	
Elements / AO basis	H, C, N, O, F / def2-SVP
Physical AOs / padded slots per atom	5 (H), 14 (C/N/O/F) / 14
Directed graph / neighbor cap	All distinct-atom pairs / none
Gaussian radial functions / centers	512 / $[0, 15]$ Å
Message envelope / AO-output mask	Smooth cutoff at 15 Å / none
Backbone and AO readout	
Node-interaction / pair-feature blocks	3 / 2
Angular orders / edge-frame orders	$\ell_{\max} = 4$ / $m_{\max} = 4$
Hidden / bottleneck channels per order	192 / 96
Hidden / bottleneck components	4800 / 2400
Edge-embedding width	128
Normalization / nonlinearity	Spherical RMS / gated + grid FFN
Initial edge-degree aggregation	Sum scaled by $1/5$
Scalar / each matrix-conditioning width	96 / 96
On-atom / off-atom coefficient inputs	288 / 576
Spatial inversion handling	Two shared-weight evaluations
Registered parameters	65,856,150 (65.9M)
Cayley head and evaluation	
Rotation steps / scale	1 / 0.25
Clipping bound on $\|\mathbf{K}\|_F/\sqrt{n_{\mathrm{AO}}}$	1
Orthogonalization	Cholesky
Extra initial / final projection	None / none
Internal / optional post-hoc McWeeny steps	0 / 2
Neural / Cayley evaluation precision	FP32 / FP64

Table 9: Shared optimizer and checkpoint settings for 18 QH9 runs. Learning rates are peak values.

Hyperparameter	Value
Optimizer	Muon + AdamW
Muon learning rate	2×10^{-2}
Muon momentum	0.95 (Nesterov)
Newton–Schulz steps	5
AdamW learning rate	5×10^{-4}
AdamW moments	$(0.9, 0.95)$
AdamW stabilizer	10^{-10}
Weight decay (both)	10^{-4}
LR schedule	Linear warmup/decay
Warmup steps	1,000
Final LR ratio	0
Gradient clipping	1.0
Random seed	42
Validation interval	1 epoch
Reported checkpoint	Final epoch (80)

Table 10: QH9 settings by model size and head. "Both" denotes additive $\Delta\mathbf{D}$ and Cayley. B is molecules per GPU.

Size	Head	w_{OV}	GPU(s)	B
18.1M	Both	0	1×H100	32
30.7M	Both	0	1×H100	32
65.9M	Both	$0, 1, 10$	1×H100	32
197M	$\Delta\mathbf{D}$	0	1×H200	32
197M	Cayley	$0, 1, 10$	1×H200	32
514M	$\Delta\mathbf{D}$	0	1×H200	32
514M	Cayley	$0, 1, 10$	2×H200	16
Shared settings				
Gradient accumulation steps				1
Effective batch (molecules)				32
Training epochs				80
Optimizer steps per epoch				3,271
Total optimizer steps				261,680
Network training precision				FP32

Table 11: QH9Stable dataset statistics and fixed random split. Molecule counts denote geometries. Atom counts are per molecule, and heavy atoms exclude hydrogen.

Statistic	**Value**
Total molecules	130,831
Training molecules	104,664
Validation molecules	13,083
Test molecules	13,084
Chemical elements	H, C, N, O, F
Atoms (all data)	3–29
Heavy atoms (all data)	1–9
Atoms (test split)	4–29

Table 12: **Composition of the QM40 evaluation subset.** Per-molecule statistics cover all 2,243 selected molecules. SD denotes the population standard deviation.

Count	Mean	SD	Median	Min.	Max.
Total atoms	48.94	13.88	49	18	88
Heavy atoms	24.44	7.68	25	10	37
Atomic orbitals	464.64	136.00	478	207	746
Electrons	181.40	54.24	188	78	270

Heavy atoms	10–14	15–19	20–24	25–29	30–34	35–37
Molecules	312	500	211	500	500	220

D.3 Evaluation metrics and aggregation

Unless stated otherwise, metrics are computed per molecule and averaged arithmetically over the evaluation set.

Matrix and target-basis block metrics. Matrix MAEs and idempotency residuals average over all n_{AO}^2 unpadded entries. The target occupied basis contains the n_{occ} largest-eigenvalue eigenvectors of $\mathbf{P}$, with the remaining eigenvectors forming the virtual basis. We evaluate the error blocks in Equation (38) using post-processed predictions where applicable. The table labels $\mathbf{P}_{\mathrm{OO}}$ and $\mathbf{P}_{\mathrm{OV}}$ denote $\|\Delta\mathbf{P}_{oo}\|_F$ and $\|\Delta\mathbf{P}_{ov}\|_F$, without block-size normalization. The OV norm includes only the occupied-by-virtual block. These norms differ from the entrywise OV training loss (Section C.4).

Real-space density metric. Predicted and target densities are evaluated in FP64 on the same PySCF level-3 molecular grid. For grid points $\mathbf{r}_g$ and weights w_g,

$$\rho_{\mathrm{NMAE}} = \frac{\sum_g w_g |\rho_{\widehat{\mathbf{D}}}(\mathbf{r}_g) - \rho_{\mathbf{D}}(\mathbf{r}_g)|}{\sum_g w_g |\rho_{\mathbf{D}}(\mathbf{r}_g)|}. \tag{46}$$

One-shot energy and force metrics. We evaluate predicted and target densities under identical DFT settings without SCF relaxation. Occupied-orbital MAE compares the lowest n_{occ} generalized eigenvalues of $\mathbf{H}[\widehat{\mathbf{D}}]$ and $\mathbf{H}[\mathbf{D}]$ with overlap $\mathbf{S}$. Energy MAE averages the absolute difference

$$\Delta E_{\mathrm{1shot}} = E_{\mathrm{KS}}[\widehat{\mathbf{D}}; \mathcal{M}] - E_{\mathrm{KS}}[\mathbf{D}; \mathcal{M}]. \tag{47}$$

Per-atom errors divide each molecule's absolute error by its total atom count, including hydrogen, before averaging. Force MAE averages absolute Cartesian-component errors over all atoms.

Effective relative iteration count. For recorded iteration counts $N_{i,m}$ and $N_{i,\mathrm{MINAO}}$, we define

$$\mathrm{ERIC}_m = \frac{1}{|\mathcal{C}_m|} \sum_{i \in \mathcal{C}_m} \frac{N_{i,m} + 1}{N_{i,\mathrm{MINAO}}}, \tag{48}$$

where $\mathcal{C}_m$ includes molecules for which both runs converge. The $+1$ counts an additional Fock build for model initialization. ERIC measures relative iteration cost and excludes inference and density-preparation wall times.

D.4 McWeeny purification

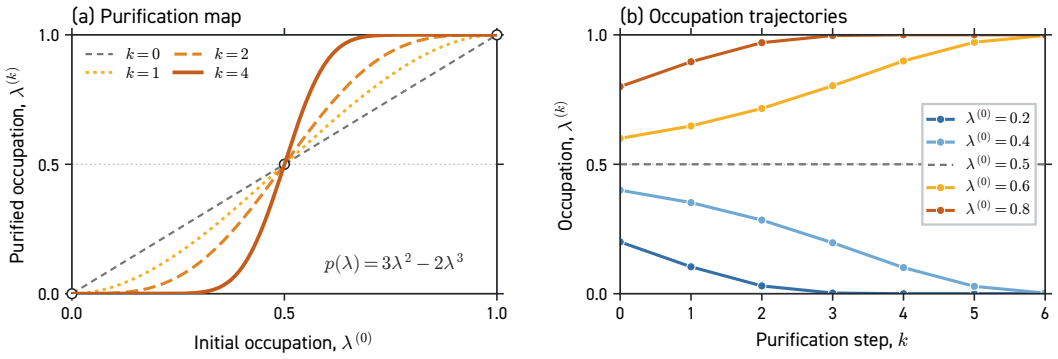


Figure 9: **McWeeny purification.** (a) Occupation maps after $k = 0, 1, 2$, and 4 steps. (b) Trajectories of representative initial occupations.

For the intervention in Section 4.1, one step applies the polynomial (McWeeny, 1960)

$$\widehat{\mathbf{P}}^{(k+1)} = p\big(\widehat{\mathbf{P}}^{(k)}\big) = 3\big(\widehat{\mathbf{P}}^{(k)}\big)^2 - 2\big(\widehat{\mathbf{P}}^{(k)}\big)^3, \qquad p(x) = 3x^2 - 2x^3. \tag{49}$$

This update is evaluated directly using matrix multiplications and requires no eigensolver. Its effect on the occupation spectrum can be understood by writing the current prediction in its eigenbasis:

$$\begin{aligned}\widehat{\mathbf{P}}^{(k)} &= \widehat{\mathbf{U}}\,\mathrm{diag}\big(\lambda_1^{(k)},\dots,\lambda_{n_{\mathrm{AO}}}^{(k)}\big)\widehat{\mathbf{U}}^\top,\\ \widehat{\mathbf{P}}^{(k+1)} &= \widehat{\mathbf{U}}\,\mathrm{diag}\big(p(\lambda_1^{(k)}),\dots,p(\lambda_{n_{\mathrm{AO}}}^{(k)})\big)\widehat{\mathbf{U}}^\top.\end{aligned} \tag{50}$$

This eigendecomposition is used only to interpret the transformation. Each eigenvalue follows $\lambda^{(k+1)} = p(\lambda^{(k)})$, while the eigenvectors $\widehat{\mathbf{U}}$ remain unchanged. For eigenvalues in $[0,1]$, repeated application moves values below $1/2$ toward zero and values above $1/2$ toward one. A value exactly at $1/2$ remains at this unstable fixed point. Figure 9 shows the purification maps and representative iteration trajectories. Purification can therefore reduce occupation errors without rotating a misaligned occupied subspace toward the target, which is the distinction probed in Section 4.1.

Using the AO-to-orthonormal relation in Section 2, the same update in the spin-summed AO density is

$$\begin{aligned}\widehat{\mathbf{D}}^{(k+1)} = &\frac{3}{2}\widehat{\mathbf{D}}^{(k)}\mathbf{S}\widehat{\mathbf{D}}^{(k)}\\ &-\frac{1}{2}\widehat{\mathbf{D}}^{(k)}\mathbf{S}\widehat{\mathbf{D}}^{(k)}\mathbf{S}\widehat{\mathbf{D}}^{(k)}.\end{aligned} \tag{51}$$

The reported McW2 results apply this update twice in FP64. A finite number of steps does not generally enforce an exact projector or electron count. For initial eigenvalues in $[0,1]$ with none equal to $1/2$, the limiting electron count is correct only when exactly n_{occ} eigenvalues lie above $1/2$. Comparisons with fixed-rank spectral retraction on QH9 and QM40 are given in Sections E.2 and E.5.

D.5 Computational resources and runtime

Each 65.9M main-comparison model was trained on one NVIDIA H100 GPU with 80 GB of device memory. Table 13 records the shared hardware and software environment, including the Intel Xeon Platinum 8468 CPU. GPU allocations for other model sizes are specified separately in Table 10. Data loading for the 65.9M runs used two persistent worker processes with a prefetch factor of two.

Table 14 compares total runtime and estimated training time. Total runtime includes validation and final testing. Training time is estimated from the recorded training intervals in each epoch, including data loading and optimization but excluding evaluation and checkpoint saving. The separate FP64 output evaluations and downstream DFT calculations are described in Sections D.1 and E.1.

Table 13: Hardware and software per 65.9M QH9 training run. Precision refers to the backbone. OV-loss linear algebra uses FP64.

Resource or setting	**Value**
GPU model	NVIDIA H100 80GB HBM3
GPUs allocated per run	1
Device memory per GPU	80 GB
CPU model	Intel Xeon Platinum 8468
Data-loader worker processes	2
Network training precision	FP32 (`32-true`)
Python	3.12.13
PyTorch	2.11.0+cu126
PyTorch CUDA build	12.6
PyTorch Lightning	2.6.1
e3nn	0.6.0

Table 14: Runtime for 80 epochs of QH9 training per 65.9M model on one H100. Total time includes validation and testing, which estimated training time excludes.

Method	w_{OV}	**Total time** (h)	**Training time** (h, est.)
$\Delta\mathbf{D}$	0	43.2	40.3
$\Delta\mathbf{D}$	1	51.9	48.1
$\Delta\mathbf{D}$	10	51.7	48.0
Cayley	0	70.0	64.6
Cayley	1	78.7	72.4
Cayley	10	79.0	72.7

Inference and energy evaluation. We time the 65.9M models on the 200 QH9 molecules used for SCF benchmarking, with batch size one and two repetitions per molecule on an H100. The density models use $w_{\mathrm{OV}} = 0$. The Cayley head takes 4.17 ms, accounting for 2.9% of inference and 0.35%

of the composed inference-to-energy time. Inference and energy are timed separately; their sum excludes initial-density preparation, full input construction, checkpoint loading, and disk I/O. The $H \to D$ and McW2 operations use one CPU thread; the Cayley head runs on the GPU.

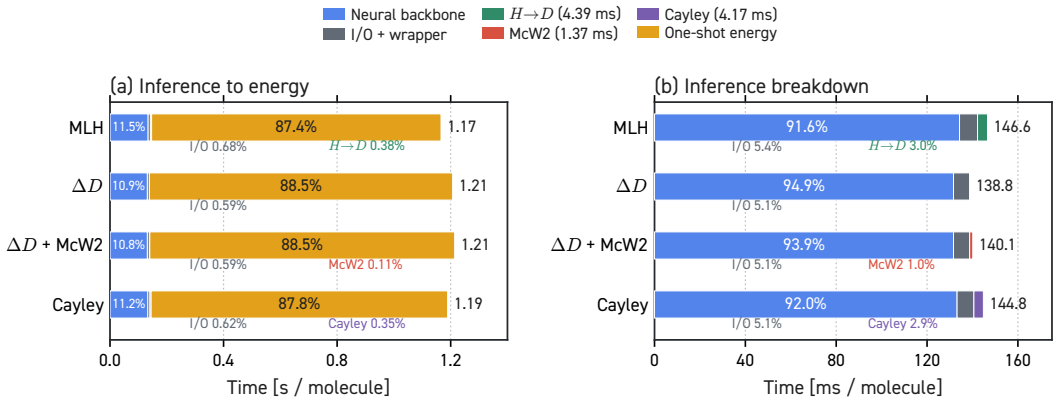


Figure 10: **Runtime breakdown on QH9.** (a) Composed inference-to-energy time and (b) inference components, averaged over 200 molecules. Percentages are relative to each panel's total.

E Additional experiments and analyses

Density visualization conventions. In the real-space comparisons below and their main-text counterparts, panels for the same molecule use matched cameras, physical scales, and framing. Signed errors are $\Delta\rho = \rho_{\text{pred}} - \rho_{\text{target}}$, with red and blue surfaces at equal positive and negative isovalues within each comparison. Smoothing is applied only for display.

E.1 Effects of numerical precision on the Cayley model

Cayley preserves occupations and idempotency in exact arithmetic, but FP32 evaluation introduces numerical residuals. We compare FP32 outputs, FP32 outputs followed by McW2, and FP64 Cayley rotation/output with an FP32 backbone, using the same 65.9M model at $w_{\text{OV}} = 0$ without retraining. Across all 13,084 QH9 test molecules, McW2 and FP64 rotation/output reduce occupation errors and lower mean one-shot energy error from 1.066 meV to 0.440 and 0.446 meV, respectively (Figure 11). Both yield similar energy-error distributions with essentially unchanged OV errors. Unlike Cayley, additive $\Delta\mathbf{D}$ predictions can violate idempotency even in exact arithmetic, so higher precision alone cannot enforce valid occupations. Neither numerical correction nor purification ensures occupied-subspace alignment.

E.2 Occupation and subspace errors

McWeeny purification and spectral retraction on QH9. We compare McW2 with fixed-rank spectral retraction (SR) on the same 65.9M additive predictions at $w_{\text{OV}} \in \{0, 10\}$. SR diagonalizes the normalized orthonormal density $\widehat{\mathbf{P}} = \frac{1}{2}\mathbf{S}^{1/2}\widehat{\mathbf{D}}\mathbf{S}^{1/2}$, sets its n_{occ} largest eigenvalues to one and the remainder to zero, and transforms the resulting projector back to the AO basis. This enforces the prescribed electron number and idempotency without using target orbitals. Both operations are applied in FP64 without retraining.

SR reduces idempotency errors from approximately 10^{-9} after McW2 to 10^{-15}, but lowers energy MAE by only 0.6% and 1.4% at the two OV weights, with force MAE changes below 0.02% (Table 15). Residual occupation errors therefore have little effect on QH9 downstream accuracy, unlike the larger effects on QM40 (Section E.5). OV supervision remains beneficial after SR, reducing energy MAE from 1.543 to 0.364 meV despite both predictions having the prescribed occupation spectrum.

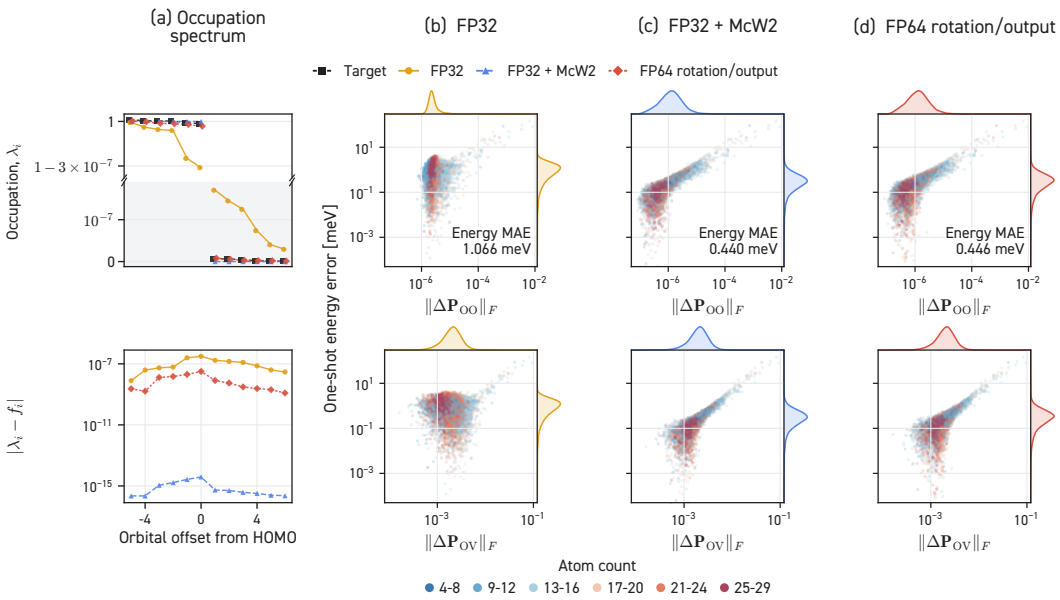


Figure 11: **Numerical precision of Cayley predictions.** The same 65.9M FP32-trained model ($w_{\mathrm{OV}} = 0$) is evaluated with FP32 output, FP32 + McW2, and FP64 rotation/output. (a) Occupations for CH_3NO. (b–d) Energy error versus OO (top) and OV (bottom) error on all 13,084 QH9 test molecules, colored by atom count.

Table 15: **McWeeny purification versus spectral retraction on QH9.** The same 65.9M $\Delta\mathbf{D}$ predictions undergo McW2 or fixed-rank spectral retraction (SR) in FP64. Means cover all 13,084 test molecules with the metrics and units of Table 1.

Method	w_{OV}	Matrix	Validity		Subspace		Density	Downstream	
		$\mathbf{D}\downarrow$ 10^{-6}	Trace$\downarrow$ $10^{-6}e$	Idemp.$\downarrow$ 10^{-6}	$\mathbf{P}_{\mathrm{OO}}\downarrow$ 10^{-3}	$\mathbf{P}_{\mathrm{OV}}\downarrow$ 10^{-3}	$\rho_{\mathrm{NMAE}}\downarrow$ 10^{-6}	Energy$\downarrow$ meV	Force$\downarrow$ meV/Å
$\Delta\mathbf{D}$	0	58.18	0.39	0.00	0.0097	4.3175	125.95	1.55	2.99
+ McW2	10	51.31	0.75	0.00	0.0037	2.0106	70.95	0.37	2.18
$\Delta\mathbf{D}$	0	58.18	0.00	0.00	0.0096	4.3175	125.95	1.54	2.99
+ SR	10	51.31	0.00	0.00	0.0034	2.0107	70.95	0.36	2.18

Matrix and subspace errors versus downstream accuracy. McW2 reduces OO and energy errors while leaving OV errors nearly unchanged. Increasing w_{OV} from 0 to 10 reduces OV and energy errors for both purified additive predictions and Cayley (Figure 12). Matrix MAE closely tracks OV and real-space density errors, but its relationship with energy depends on the parameterization and post-processing. Neither matrix MAE nor the OV norm alone fully characterizes energy accuracy.

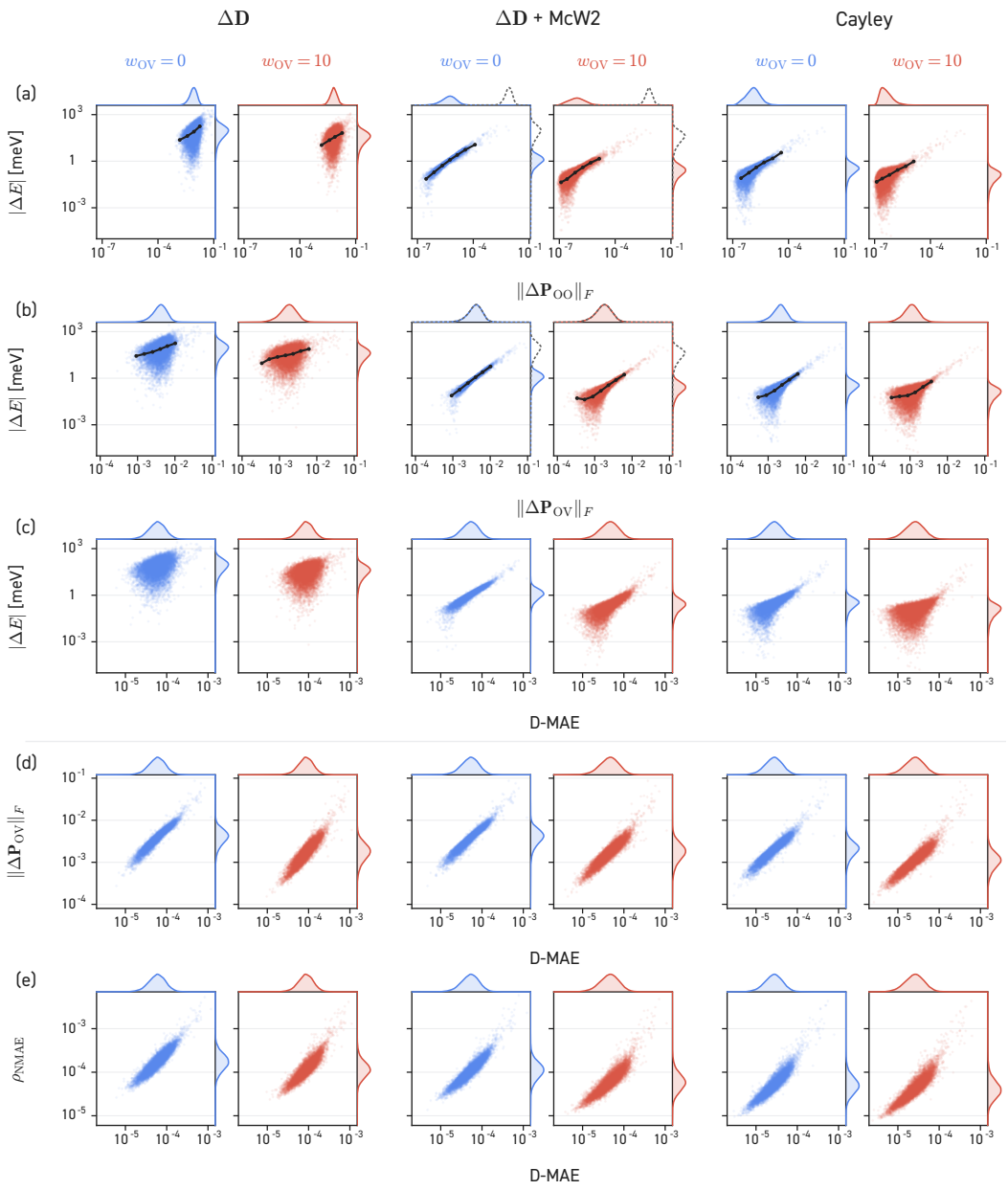


Figure 12: **Density and energy errors on QH9.** All 13,084 test molecules are shown for $\Delta\mathbf{D}$, $\Delta\mathbf{D}$ + McW2, and native Cayley with FP64 rotation/output at $w_{OV} = 0$ (blue) and 10 (red). Rows (a–c) compare energy error with OO error, OV error, and D-MAE. Rows (d,e) compare D-MAE with OV error and ρ_{NMAE}. In (a,b), black curves show binned median energy errors and gray dashed marginals in the McW2 columns show the raw predictions.

SCF initialization across molecular sizes. Using the size-balanced QH9 cohort in Section D.1, Cayley requires fewer effective SCF iterations than either additive variant at each OV weight across all four heavy-atom counts (Figure 13 and Table 16).

SCF-only timings use an H100 with DIIS starting at cycle one. We average three repetitions per molecule, except for raw ΔD, which uses one measurement. Time-ratio pairs each model with MINAO from the same run;the absolute MINAO curve uses the three-repeat run. MLH uses an overlap filter of 10^{-6}.

Table 16: **SCF initialization on QH9.** ERIC is relative to MINAO on 200 molecules over three paired repeats. MLDMs use $w_{\mathrm{OV}} = 10$ without overlap filtering. All runs converge. Settings follow Table 2. Best ERIC is **bold**.

Hamiltonian models (MLHs)				*Density-matrix models (MLDMs)*	
Model	ERIC↓ by S filter			**Model**	ERIC↓
	10^{-6}	10^{-3}	10^{-2}		
QHF2	0.590	0.669	0.817	**ΔD + McW2**	0.552
MLH	0.557	0.662	0.818	**Cayley**	**0.534**
MLH + WA-FT	0.645	0.694	0.818		

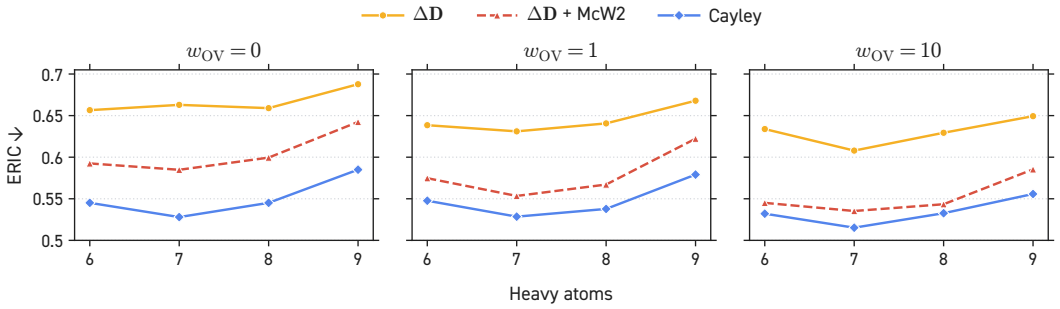


Figure 13: **SCF initialization by molecular size on QH9.** ERIC for 65.9M models at three OV weights, with 50 molecules per point. MINAO is the unit baseline. All runs converge.

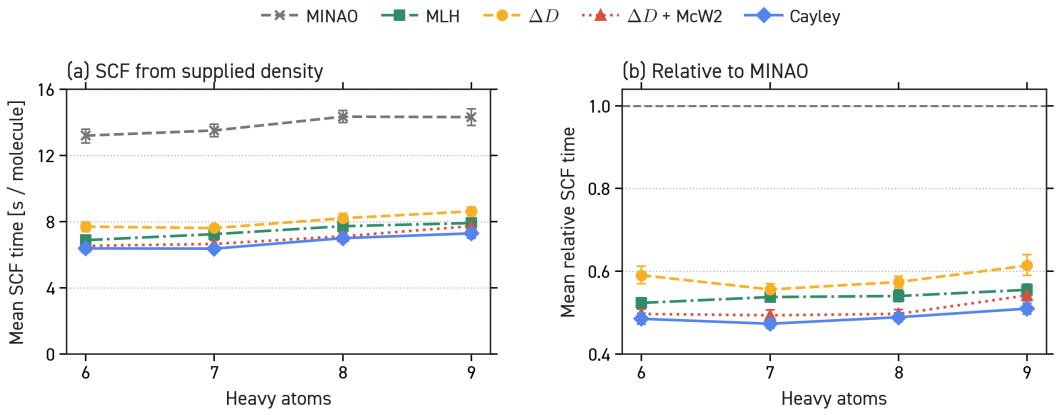


Figure 14: **SCF runtime from supplied initial densities on QH9.** (a) Mean time and (b) mean paired model/MINAO time ratio for 50 molecules per size. Density models use $w_{\mathrm{OV}} = 10$. Error bars show molecule-bootstrap 95% confidence intervals. Timings start from the supplied densities; their preparation, inference, and post-processing are excluded.

E.2.1 OO VERSUS OV SUPERVISION IN CAYLEY FINE-TUNING

We compare OO, OV, and combined supervision through 20-epoch fine-tuning of the same 65.9M O(3) Cayley model trained with matrix MAE, using seed 42. The matrix loss retains unit weight, while auxiliary losses use per-molecule target-basis block RMS errors.

Table 17: **OO and OV supervision in Cayley fine-tuning.** All rows use the 65.9M O(3) model and all 13,084 QH9Stable test molecules after McW2. $\mathbf{D}$ and E report MAEs, while OO and OV report mean per-molecule block RMS errors. Bold denotes minima at the displayed precision.

	Loss weights		**Matrix errors** (10^{-6})			**Energy** (meV)
Supervision	w_{OO}	w_{OV}	$\mathbf{D}\downarrow$	OO $\downarrow$	OV $\downarrow$	$E\downarrow$
Pretrained reference						
Matrix only	0	0	30.742	0.1466	33.858	0.4401
20-epoch fine-tuning						
+ OO	1	0	30.614	0.1459	33.778	0.4384
+ OO	10	0	30.613	0.1454	33.747	0.4376
+ OO	100	0	30.616	0.1431	33.537	0.4316
+ OV	0	1	**30.401**	**0.1383**	**31.428**	**0.3866**
+ OO + OV	1	1	**30.401**	0.1384	**31.428**	**0.3866**

OO-only supervision reduces energy MAE by less than 2% across $w_{\mathrm{OO}} \in \{1, 10, 100\}$, whereas OV supervision at $w_{\mathrm{OV}} = 1$ reduces it by 12.16% (Table 17). Adding OO supervision at $w_{\mathrm{OO}} = 1$ provides no further improvement at the reported precision.

E.3 COMPARISON WITH MACHINE-LEARNED HAMILTONIANS

MLH baselines. The controlled MLH uses the shared 65.9M backbone and residual output map in Equation (34), conditioned on geometry, overlap, and the MINAO-derived $\mathbf{H}_0$. It predicts B3LYP5/def2-SVP Hamiltonians with

$$\mathcal{L}_{\mathrm{H}} = \frac{10}{B}\sum_{m=1}^{B}\frac{1}{|\mathcal{I}_m|}\sum_{k\in\mathcal{I}_m}\left(|e_{mk}| + e_{mk}^2\right), \qquad e_{mk} = \widehat{H}_{mk} - H_{mk}, \tag{52}$$

where $\mathcal{I}_m$ contains all active AO entries of molecule m. We also evaluate the epoch-78 QHFlow2-m model (Kim et al., 2026) on the same 13,084 QH9 test molecules. Both baselines use FP64 density construction without McWeeny purification and matched target-density energy references.

The original QHFlow2 energy evaluation uses adaptive overlap filtering,[4] whereas our re-evaluation applies a fixed threshold across molecules for each condition. This protocol difference contributes to discrepancies between the originally reported and re-evaluated energy errors.

Density construction and overlap filtering. After symmetrizing $\widehat{\mathbf{H}}$ and $\mathbf{S}$ in FP64, we diagonalize $\mathbf{S} = \mathbf{V}\,\mathrm{diag}(s_i)\mathbf{V}^\top$ and retain modes above threshold τ:

$$\mathbf{X}_\tau = \left[\frac{\mathbf{v}_i}{\sqrt{s_i}}\right]_{s_i>\tau}, \qquad \mathbf{X}_\tau^\top \mathbf{S}\mathbf{X}_\tau = \mathbf{I}. \tag{53}$$

The predicted density follows from

$$\mathbf{X}_\tau^\top \widehat{\mathbf{H}}\mathbf{X}_\tau\mathbf{Z} = \mathbf{Z}\boldsymbol{\varepsilon}, \qquad \mathbf{C}_\tau = \mathbf{X}_\tau\mathbf{Z}, \qquad \widehat{\mathbf{D}}_\tau = 2\mathbf{C}_{\tau,o}\mathbf{C}_{\tau,o}^\top, \tag{54}$$

where $\mathbf{C}_{\tau,o}$ contains the $N_e/2$ lowest-energy orbitals. Retained ranks below $N_e/2$ are conversion failures.

[4] See the adaptive filtering implementation.

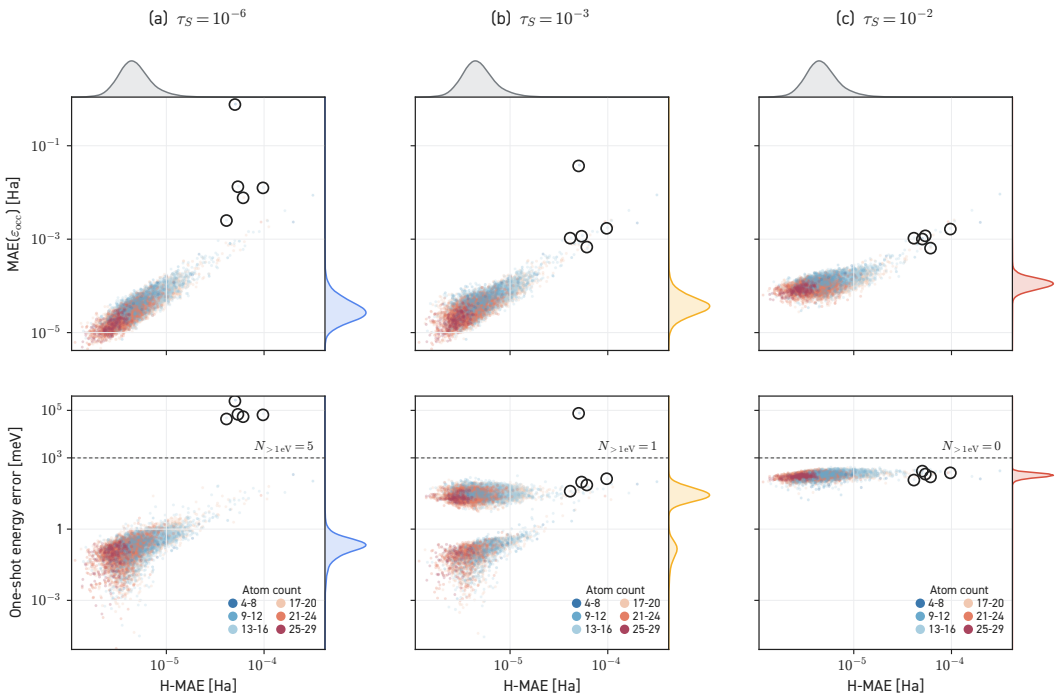


Figure 15: **Overlap-filter sensitivity on QH9.** Hamiltonian MAE versus occupied-orbital energy MAE (top) and one-shot energy error (bottom) at three thresholds, colored by atom count. Circles track the five molecules exceeding 1 eV at any threshold. Annotations count errors above 1 eV at each threshold.

Holding Hamiltonian predictions fixed, we vary $\tau \in \{10^{-6}, 10^{-3}, 10^{-2}\}$. The default threshold removes no modes, since the smallest test-set overlap eigenvalue is 1.89×10^{-5}. Increasing τ reduces the five energy errors above 1 eV to one at 10^{-3} and none at 10^{-2}, but the largest threshold increases errors for the bulk of the test set (Table 2 and Figure 15).

Spectral sensitivity. Valid occupations do not ensure that the selected occupied subspace is accurate. For positive-definite $\mathbf{S}$, define

$$\overline{\mathbf{H}} = \mathbf{S}^{-1/2}\mathbf{H}\mathbf{S}^{-1/2}. \tag{55}$$

The bound $\|\delta\overline{\mathbf{H}}\|_2 \leq \|\delta\mathbf{H}\|_2/\lambda_{\min}(\mathbf{S})$ shows how small overlap eigenvalues amplify AO-matrix errors. The first-order occupied–virtual response is

$$\kappa_{ai} \simeq \frac{\mathbf{u}_a^\top \delta\overline{\mathbf{H}}\mathbf{u}_i}{\varepsilon_i - \varepsilon_a}, \tag{56}$$

where i and a index occupied and virtual eigenvectors. Small orbital gaps provide a second source of sensitivity, also reflected in the perturbation analysis of Li et al. (2025b). Overlap filtering reduces metric amplification but changes the represented eigenproblem, producing the observed accuracy–stability tradeoff.

E.3.1 Orthonormal-basis fine-tuning (WA-FT)

We fine-tune the same 80-epoch MLH for 20 additional epochs, either retaining the original objective or adding

$$\mathcal{L}_{\mathrm{H}} + 25\mathcal{L}_{\mathrm{WA}}, \qquad \mathcal{L}_{\mathrm{WA}} = \frac{1}{B}\sum_m \frac{\|\mathbf{X}_m^\top(\widehat{\mathbf{H}}_m - \mathbf{H}_m)\mathbf{X}_m\|_F^2}{n_m^2}, \tag{57}$$

where n_m is the AO count and $\mathbf{X}_m$ is the FP64 overlap whitener with an eigenvalue floor of 10^{-8}. Both runs use batch size 32, seed 42, and peak learning rates of 0.01 for Muon and 2.5×10^{-4} for AdamW, with one epoch of linear warmup followed by cosine decay.

E.4 Inversion-symmetry diagnostics

Full-test comparison. For the 65.9M Cayley models with $w_{\mathrm{OV}} = 10$, we compare SO(3) and inversion-averaged O(3) predictions on all 13,084 QH9 test molecules at their original geometries, using FP32 inference followed by McW2. O(3) reduces mean energy error from 0.242 to 0.168 meV, a 30.3% reduction. After excluding the two inversion-sensitive examples in Figure 5, the reduction remains 26.7%.

Selected-case symmetry audit. We separately evaluate these two molecules at their original geometries, after a 90° rotation about the z axis, and after inversion about their geometric centroids. All AO inputs are transformed consistently with the geometry, and predicted densities are mapped back to the original AO frame before comparison (Section C.2.1 and Section C.5). Both 65.9M, $w_{\mathrm{OV}} = 10$ models use their epoch-80 weights, with all inference operations promoted to FP64 and no McWeeny post-processing. For an AO transformation $\mathbf{U}_g$, we measure

$$\epsilon_g = \frac{\|\mathbf{U}_g^\top \widehat{\mathbf{D}}(g\mathcal{M})\mathbf{U}_g - \widehat{\mathbf{D}}(\mathcal{M})\|_F}{\|\widehat{\mathbf{D}}(\mathcal{M})\|_F}. \tag{58}$$

Inversion changes SO(3) energy errors from 69.28 to 4.37 meV and from 92.62 to 6.07 meV for the two examples, with residuals 2.66×10^{-2} and 4.95×10^{-2}. O(3) retains energy errors of 2.37 and 4.08 meV, respectively, with inversion residuals below 4×10^{-15}. Errors use the shared DFT settings and metrics in Sections D.1 and D.3.

E.5 Molecular-size extrapolation on QM40

We test how QH9-trained density models transfer to the larger molecules in Section D.2. The comparison holds model capacity fixed and varies the output head, OV-loss weight, and post-processing under the protocol in Section D.2.

Table 18 reports mean density errors and energy-error distributions for the reference density, additive predictions with and without post-processing, and native Cayley predictions on the same 2,243 molecules. McW2 applies two McWeeny steps, while SR denotes fixed-rank spectral retraction of the raw additive prediction.

Table 18: **Supplementary QM40 results.** Mean density errors and energy-error distributions on the same 2,243 molecules. $\mathbf{D}_0$ is the reference density. SR denotes spectral retraction of raw additive predictions. For $e_N = |\Delta E|/N$, $n_{>\tau}$ counts cases above $\tau = 100$ meV/atom, and $\mathrm{Mean}_{\leq\tau}$ averages the remaining molecular energy errors. Best errors among learned models are **bold**, excluding the conditional means.

Method	w_{OV}	Mean density errors			Energy error [meV]				Per atom [meV/atom]		$n_{>\tau}\downarrow$
		$\mathbf{D}\downarrow$ 10^{-6}	$\mathbf{P}_{\mathrm{OV}}\downarrow$ 10^{-3}	$\rho_{\mathrm{NMAE}}\downarrow$ 10^{-6}	Mean↓	Median↓	Max.↓	$\mathrm{Mean}_{\leq\tau}\downarrow$	Mean↓	Median↓	
$\Delta\mathbf{D}$ + McW2	0	190.396	44.634	800.88	192.076	128.943	1612.975	192.076	3.402	2.608	0
	10	152.419	26.483	561.17	79.033	33.007	9703.650	74.740	1.376	0.677	1
$\Delta\mathbf{D}$ + SR	0	190.395	44.637	801.07	166.054	121.549	1148.726	166.054	2.961	2.457	0
	10	152.430	26.490	**560.22**	56.915	32.623	1117.967	56.915	1.009	0.670	0
Cayley	0	121.611	31.213	660.08	84.305	61.789	765.120	84.305	1.509	1.308	0
	1	**117.017**	27.334	754.99	64.731	44.631	**725.780**	64.731	1.152	0.927	0
	10	118.446	**24.190**	801.61	**55.511**	**31.153**	1700.365	55.511	**0.985**	**0.613**	0

Idempotency errors and spectral retraction. Two McWeeny steps can leave residual occupation errors in additive QM40 predictions. We therefore apply spectral retraction (SR, Section E.2) to the same raw predictions. SR reduces mean entrywise idempotency error to approximately 10^{-15} and lowers mean per-atom energy error relative to McW2 by 13.0% at $w_{\mathrm{OV}} = 0$ and 26.7% at $w_{\mathrm{OV}} = 10$. Matrix, OV, and real-space density errors change by less than 0.2%. Under SR, increasing w_{OV} from 0 to 10 further reduces energy error from 2.961 to 1.009 meV/atom.

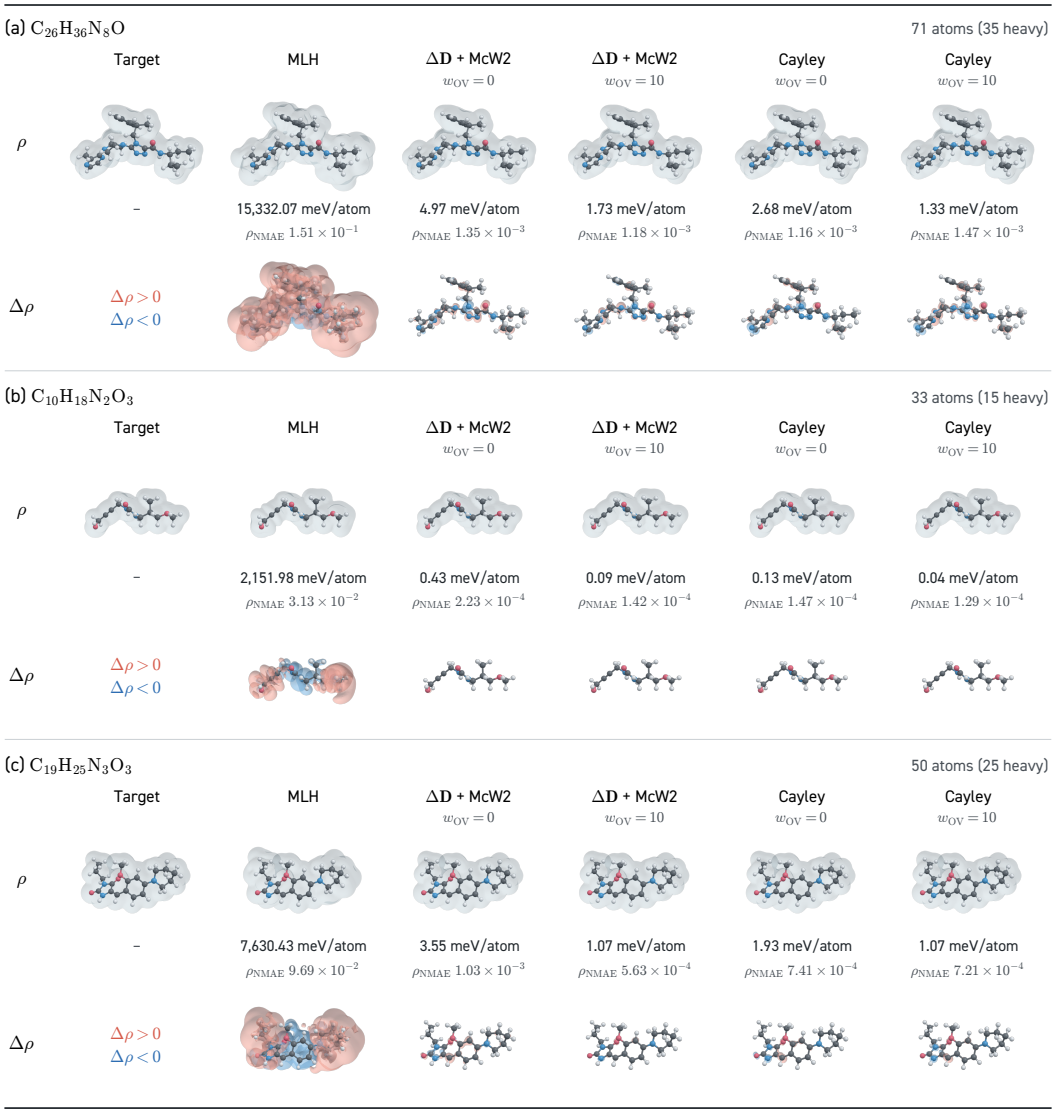


Figure 16: **QM40 electron densities and signed errors.** The molecules in Figure 7 are shown for the target, MLH, $\Delta\mathbf{D}$ + McW2, and native Cayley, with $w_{\mathrm{OV}} \in \{0, 10\}$ for both MLDMs. Red and blue denote positive and negative density errors. Labels give absolute energy error per atom (meV/atom) and ρ_{NMAE}.

Native Cayley has lower energy errors at both OV weights, reaching 1.509 and 0.985 meV/atom, respectively. It also has lower matrix and OV errors, whereas the retracted additive model at OV10 has lower ρ_{NMAE} (560.22×10^{-6} versus 801.61×10^{-6}).

Density visualizations. The examples in Figures 7 and 16 span 15, 25, and 35 heavy atoms. Size-selected examples are chosen near the median total atom count within each heavy-atom group, breaking ties by source index. The 15-heavy-atom example in Figure 7 is selected to illustrate lower Cayley errors than MLH in both density and energy.