PHYSICS-ALIGNED ELECTRONIC GROUND-STATE LEARNING IMPROVES GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learned interatomic potentials (MLIPs) excel at in-distribution tasks, accelerating drug and material development, yet they struggle to generalize out-of-distribution. We propose to push the cost-accuracy Pareto frontier by designing observable-agnostic electronic ground-state descriptor models (GSMs) with computational costs situated between MLIPs and Kohn-Sham density functional theory (KS-DFT). We align the learning objectives and architectures of GSMs with the governing equations of KS-DFT by enforcing physical constraints and removing optimization pressure on unphysical or irrelevant degrees of freedom. In our size-extrapolation experiments from QM9 to QM40, our combined contributions OrthoNormal-Loss (ON-Loss) and Grassmann Restricted Occupied-Orbital Training (GROOT) reach a 79.1% energy and 83.4% force mean absolute error (MAE) reduction over previous state-of-the-art density GSMs. For Hamiltonian GSMs, ON-Loss and Residual Optimal-gauge Conditioning-aware KS-Eq. Training (ROCKET) together reduce the energy and force MAEs of the strongest baseline by 99.8% and 95.9%, respectively. Using a self-consistency rejection criterion, we filter out extrapolation errors on QMugs, rejecting fewer than 0.4% of predictions while reaching an energy MAE of 0.07 mHa. Finally, we demonstrate the efficiency of label-free self-consistency fine-tuning, and transfer GSMs to reactive chemistry in Transition1x, reaching energy errors below chemical accuracy.



Figure 1: **Size extrapolation of QM9-trained GSMs evaluated on up to $4\times$ larger QM40 molecules.** The GROOT models generalize robustly, **reaching errors below half chemical accuracy** (~ 0.8 mHa) **on 98% of the tested molecules.** All observables are computed at a budget of exactly one Fock build averaged over 3 training seeds each. **The GROOT variants are the best performing models for almost all observables** except for the Kohn–Sham (self-consistency) energy where ROCKET is best. MLIPs trained on forces still outperform even the best GSMs, but physically aligned GSMs achieve an order of magnitude better energies compared to all trained MLIPs.

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.

1 INTRODUCTION

Many computational chemistry applications aim to accurately approximate the electronic ground state of molecules and materials. The rapid developments in this field are driven by the abundance of applications enabling the *in silico* prescreening of potential materials and drugs before moving into resource- and time-intensive laboratory experiments or clinical trials. Since the exact computation of the ground state is intractable for almost all systems of interest, the art of computational chemistry is to find the correct trade-off between accuracy and computational cost for a given application (Szabo & Ostlund, 1996). Contemporary method development pushes this Pareto frontier as far as possible. For many applications, KS-DFT (Kohn & Sham, 1965) is the method of choice (Antypas et al., 2014; Haunschild et al., 2019). At a high level, its computational efficiency can be attributed to its lower-dimensional representation of the ground state based on the electron density $\rho : \mathbb{R}^3 \to [0, \infty)$, compared to wavefunction methods $\psi : \mathbb{R}^{N_{\text{elec}} \times 3} \to \mathbb{C}$ (Cramer, 2014).

Ground-state descriptor models (GSMs) have been investigated as a means to accelerate (Liu et al., 2026; Eberhard et al., 2026a) or completely circumvent (Schütt et al., 2019) the computationally demanding iterative solvers which drive KS-DFT calculations (Lehtola et al., 2020a). These solvers use finite basis sets to discretize the governing equations of KS-DFT, which yields a generalized eigenvalue problem (gEVP) (Roothaan, 1951). Present **state-of-the-art GSMs are not well-aligned with the goal of learning physical ground states**, motivating the following contributions:

1. Current literature has converged on supervising GSMs on the matrix representations of the gEVP (Sec. 3) measuring matrix errors in a non-orthonormal (NON) basis, whose spectral weighting is ill-aligned with physical observables. We show mathematically and empirically how **OrthoNormal-Loss (ON-Loss)** cures this by measuring errors in an orthonormal (ON) basis.

2. On the problem side of the gEVP, entire families of Hamiltonian matrices yield the same ground state. The established loss formulation views these physically equivalent Hamiltonians as different targets. We systematically treat this invariance w.r.t. changes in the Hamiltonian by introducing **Residual Optimal-gauge Conditioning-aware KS-Eq. Training (ROCKET)**.

3. On the solution side, physically valid density matrices are constrained to a manifold (the *Grassmannian*), but existing readout heads predict in the entire ambient space of symmetric matrices. We introduce **Grassmann Restricted Occupied-Orbital Training (GROOT)**, a matrix readout for $SO(3)$-equivariant graph neural networks (GNNs) that restricts predictions to valid densities, removing the leading order energy error term compared to unconstrained baselines.

4. We develop a **practical framework** under a **strict budget of one Fock build**, the cost-dominant operation of KS-DFT. For this, we adopt a computationally cheap **first-order energy correction**, identify a **GSM-native uncertainty measure** for rejecting unreliable predictions, and generalize self-consistency training (Zhang et al., 2024) for **label-free fine-tuning of foundational GSMs to new chemical domains** such as transition states. These constitute key advantages of GSMs over MLIPs that might justify their increased evaluation costs for some applications.

At this single Fock-build budget, the evaluation costs of GSMs are in between MLIPs and machine-learned (ML) approaches that require multiple Fock builds (Dick & Fernandez-Serra, 2021; Gao et al., 2024; Eberhard et al., 2026b; Luise et al., 2026; Kim et al., 2025; 2026). In our size extrapolation experiments on QMugs, we train on molecules with at most 150 electrons and evaluate on up to 660 electrons (Fig. 3). Here, ON-Loss alone reduces energy, force, dipole and density errors by 87–99% for density GSMs and by 93–99.5% for Hamiltonian GSMs. $\text{GROOT}_{\text{proj}}$ and ROCKET **compound** on top of that, reaching 99.6% and 99.9% lower energy errors than the NON baselines. At 0.21 and 0.56 mHa (without rejections), their *mean* energy errors lie below chemical accuracy and are 14 and 5 times lower than the best energy-trained MLIP, while they reject fewer than 0.4% of predictions. Using the predicted ground states as self-consistent field (SCF) initial guesses reduces the number of solver iterations needed to converge. For the QM9-trained models evaluated on QM40, ON-Loss lowers the density GSM's Effective Relative Iteration Count (ERIC) (Eberhard et al., 2026a) from 0.82 to 0.70, and GROOT lowers it further to 0.67, a third fewer iterations than the traditional MINAO initial guess. After label-free self-consistency fine-tuning on Transition1x (Schreiner et al., 2022), the QMugs-trained $\text{GROOT}_{\text{proj}}$ and ROCKET reach energy errors of 0.84 and 0.65 mHa, below chemical accuracy, while rejecting fewer than 1% of predictions.

2 BACKGROUND

The governing equation of KS-DFT in finite basis sets $\{\chi_\mu(\mathbf{r})\}_{\mu=1}^{B}$ is a gEVP of the Hamiltonian $\mathbf{H} \in \mathbb{R}^{B\times B}$ and *overlap* matrix $S_{\mu\nu} = \int_{\mathbb{R}^3} \chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\,\mathrm{d}\mathbf{r}$, with solutions (eigenvectors) $\mathbf{C} \in \mathbb{R}^{B\times B}$

$$\mathbf{HC} = \mathbf{SC}\,\mathrm{diag}(\epsilon_i)\,. \tag{1}$$

In this context, the solutions are also known as *molecular orbitals* $\mathbf{C}$. For further details and the connection of Eq. (1) to the underlying continuous theory, we refer to App. A and Lehtola et al. (2020a). By convention, the eigenvalues ϵ_i (*orbital energies*) are enumerated ascendingly. The $O = N_\mathrm{elec}/2$ lowest eigenvectors (columns of $\mathbf{C}$) form the *occupied block* $\mathbf{C}_O \in \mathbb{R}^{B\times O}$, while the remaining $V = B - O$ are *virtual*. The density matrix

$$P_{\mu\nu} = (\mathbf{C}_O\mathbf{C}_O^\mathsf{T})_{\mu\nu} = \textstyle\sum_{i=1}^{O} C_{\mu i}C_{\nu i} \tag{2}$$

encodes the electron density in the finite basis. The Hamiltonian $\mathbf{H}$ depends on the density matrix $\mathbf{P}$, which in turn depends on the coefficients $\mathbf{C}$, so Eq. (1) is non-linear. It can be linearized by iteratively solving for $\mathbf{C}_t$ given $\mathbf{H}_{t-1} := \mathbf{H}(\mathbf{P}(\mathbf{C}_{t-1}))$, which reduces the original problem to the **iterative solution of the gEVP** (Eq. 1) (Roothaan, 1951). This approach **requires an initial guess** $\mathbf{P}^{(0)}$ to start the iteration. The Superposition of Atomic *Densities* (SAD) method and its minimal-basis variant MINAO construct such a guess in density space by combining spherically averaged atomic density matrices (Almlöf et al., 1982; Van Lenthe et al., 2006). The ground state $\mathbf{C}^*$ is a fixed point of the iterative solver, and all three matrices $\mathbf{C}^*$, $\mathbf{P}^*$, and $\mathbf{H}^*$ are complete descriptors allowing access to all observables of the ground state. These matrices are commonly represented in atom-centered basis sets, which are **non-orthonormal (NON)** $\mathbf{S} \neq \mathbf{I}$. We can perform a change of basis $\widetilde{\chi}_\mu(\mathbf{r}) = \sum_{\nu=1}^{B} \chi_\nu(\mathbf{r})(\mathbf{S}^{-1/2})_{\nu\mu}$, to obtain **orthonormal (ON)** basis representations

$$\widetilde{\mathbf{C}} = \mathbf{S}^{1/2}\mathbf{C}, \qquad \widetilde{\mathbf{P}} = \mathbf{S}^{1/2}\mathbf{P}\mathbf{S}^{1/2}, \qquad \widetilde{\mathbf{H}} = \mathbf{S}^{-1/2}\mathbf{H}\mathbf{S}^{-1/2}, \qquad \widetilde{\mathbf{S}} = \mathbf{I}, \tag{3}$$

which reduce the gEVP to the regular eigenvalue problem $\widetilde{\mathbf{H}}\widetilde{\mathbf{C}} = \widetilde{\mathbf{C}}\,\mathrm{diag}(\epsilon_i)$ (Löwdin, 1950).

Infinitely many representations $(\mathbf{H}, \mathbf{C})$ encode the same physical ground state. Only the density matrix $\mathbf{P}$ is in one-to-one correspondence with this state. On the problem side of Eq. (1), for Hamiltonians $\mathbf{H}$, one can identify constant shifts

$$\widetilde{\mathbf{H}} \mapsto \widetilde{\mathbf{H}} + c\mathbf{I}, \qquad c \in \mathbb{R} \tag{4}$$

as a degree of freedom, leaving the eigenvectors $\mathbf{C}$ invariant. Similarly, so-called *virtual level shifts* $\widetilde{\mathbf{H}}_\xi := \widetilde{\mathbf{H}}^* + \xi(\mathbf{I} - \widetilde{\mathbf{P}}^*)$, for an arbitrary $\xi \geq 0$ only increase the eigenvalues of the virtual block $\widetilde{\mathbf{C}}_V$ by ξ. Both of these operations leave $\mathbf{C}$ and hence $\mathbf{P}$ unchanged. On the solution side of Eq. (1), we note that the solution space $\mathbf{C}$ carries considerably more degrees of freedom than the density matrix $\mathbf{P}$ it yields. In the mapping $\mathbf{C} \mapsto \mathbf{P}$ in Eq. (2), the virtual block $\mathbf{C}_V$ never enters and is discarded right away. Additionally, the occupied columns $\widetilde{\mathbf{C}}_O$ themselves are not unique and can be multiplied with an arbitrary orthogonal matrix $\mathbf{U}_{OO} \in \mathrm{O}(O)$ without changing the density

$$\widetilde{\mathbf{P}}' = \widetilde{\mathbf{C}}'_O(\widetilde{\mathbf{C}}'_O)^\mathsf{T} \equiv \big(\widetilde{\mathbf{C}}_O\mathbf{U}_{OO}\big)\big(\widetilde{\mathbf{C}}_O\mathbf{U}_{OO}\big)^\mathsf{T} = \widetilde{\mathbf{C}}_O\mathbf{U}_{OO}\mathbf{U}_{OO}^\mathsf{T}\widetilde{\mathbf{C}}_O^\mathsf{T} = \widetilde{\mathbf{P}}, \tag{5}$$

and the same holds for the virtual slice $\widetilde{\mathbf{C}}_\mathrm{V}$. **Whole families of Hamiltonian $\mathbf{H}$ and coefficient matrices $\mathbf{C}$ collapse onto the same physical density $\mathbf{P}$.** What distinguishes one ground state from another is the O-dimensional subspace spanned by their occupied columns $\mathrm{span}\{\mathbf{c}_i | i = 1, 2, \ldots, N_\mathrm{elec}/2\}$. The smallest problem instance $B = 3$ and $O = 2$ is intuitively accessible. In this case, the density is a plane through the origin of $\mathbb{R}^3$. Any two eigenvectors c_1, c_2 lying in that plane span it equally well. Predicting a density means predicting this subspace.

To constrain a model, we need to transfer this subspace criterion to the space of density matrices $\mathbf{P}$. The orthonormality of the occupied orbitals, $\widetilde{\mathbf{C}}_O^\mathsf{T}\widetilde{\mathbf{C}}_O = \mathbf{I}_O$, inserted into Eq. (2), gives $\widetilde{\mathbf{P}}^2 = \widetilde{\mathbf{C}}_O(\widetilde{\mathbf{C}}_O^\mathsf{T}\widetilde{\mathbf{C}}_O)\widetilde{\mathbf{C}}_O^\mathsf{T} = \widetilde{\mathbf{P}}$, so $\widetilde{\mathbf{P}}$ is idempotent with its eigenvalues either 0 or 1. The trace condition $\mathrm{Tr}(\widetilde{\mathbf{P}}) = O$ fixes their multiplicities to O ones and V zeros. Idempotency is not only a consequence of the construction but a complete test of it, because any **symmetric idempotent matrix** of the right trace factorizes back into an orthonormal occupied block (McWeeny, 1960). Thus, the set of valid density matrices is exactly

$$\mathcal{M}_{(N_\mathrm{elec},B)} \equiv \left\{ \widetilde{\mathbf{P}} \in \mathbb{R}^{B\times B} \;\middle|\; \widetilde{\mathbf{P}}^\mathsf{T} = \widetilde{\mathbf{P}},\;\; \widetilde{\mathbf{P}}^2 = \widetilde{\mathbf{P}},\;\; \mathrm{Tr}(\widetilde{\mathbf{P}}) = N_\mathrm{elec}/2 \right\}, \tag{6}$$

This set is the *Grassmann manifold*. Its dimension is the BO entries of $\widetilde{\mathbf{C}}_O$, minus $O(O+1)/2$ orthonormality conditions, minus $O(O-1)/2$ mixing directions (Edelman et al., 1998; Absil et al., 2008). Hence, learning valid density matrices $\mathbf{P} \in \mathbb{R}^{B\times B}$ is an OV-dimensional prediction task.

3 Related Work

ML-surrogates for ground-state observables prediction based on equivariant GNNs have been a key ingredient for recent advances in computational chemistry (Kulichenko et al., 2024; Wood et al., 2025). MLIPs are the most widely adopted of these surrogates. They predict the energy as a function of the molecular geometry $G = \{(Z_a, r_a)\}_{a=1}^{N_{\text{atoms}}}$, $G \mapsto E(G) \in \mathbb{R}$ in linear time with respect to system size (Schütt et al., 2018). These methods have demonstrated high accuracy and generalizability for a wide range of molecular systems (Kovács et al., 2023; Jacobs et al., 2025; Burger et al., 2025a;b; Koker et al., 2025) and are typically trained on KS-DFT reference data (Kulichenko et al., 2024). Another modeling ansatz focuses on **predicting the fundamental descriptors of KS-DFT**, aiming either at the Hamiltonian $\mathbf{H} \in \mathbb{R}^{B\times B}$ or the density matrix $\mathbf{P} \in \mathbb{R}^{B\times B}$, i.e., $G \mapsto \mathbf{H}(G), \mathbf{P}(G)$ (Schütt et al., 2019; Yu et al., 2023b). **From these descriptors, the computation of almost all observables, such as energy, forces, and dipole moments, requires at most one Fock build.** In particular, these models can be used to accelerate SCF calculations used to solve Eq. (1) (Liu et al., 2026; Eberhard et al., 2026a). A substantial body of work has built on QH9 (Yu et al., 2023a), a QM9-based (Ramakrishnan et al., 2014) benchmark for molecular Hamiltonian prediction. However, with recent models reporting energy errors more than two orders of magnitude below chemical accuracy (Kim et al., 2026) the benchmark has reached saturation. QH9's restriction to very small molecules limits its ability to assess size extrapolation to larger molecules, a shortcoming highlighted by recent studies (Li et al., 2025b; Eberhard et al., 2026a). Kaniselvan et al. (2026) have connected GSMs to MLIPs, using Hamiltonians for pretraining, and finetuning with an energy readout head. Zhang et al. (2026) propose to evaluate observables directly from the predicted density. Kim et al. (2026) do the same for the Hamiltonian, matching the force accuracy of MLIPs on in-distribution MD-trajectories albeit at higher computational cost, requiring two Fock builds, one for $\mathbf{H}(\mathbf{P}_{\text{MINAO}})$ Δ-learning (Eberhard et al., 2026a) and a second for the observable evaluation. **We compare the observables of all GSMs at a one Fock-build budget** starting from the molecular geometry.

Predicting and supervising the electronic ground state in the NON basis representation is the de-facto standard approach of the field (Schütt et al., 2019; Unke et al., 2021; Li et al., 2022; 2025b; Shmilovich et al., 2022; Gong et al., 2023; Yu et al., 2023a;b; Zhong et al., 2023; Hazra et al., 2024; Wang et al., 2024; Zhang et al., 2024; Febrer et al., 2025; Kim et al., 2025; 2026; Luo et al., 2025; Zhouyin et al., 2025; Eberhard et al., 2026a; Kaniselvan et al., 2026; Liu et al., 2026; Qian et al., 2026). Predicting in the NON basis is well-motivated by its spatial localization relative to the ON alternative in Eq. (3) (Li et al., 2022; Kaniselvan et al., 2026), although this argument is mainly valid for absolute Hamiltonian predictions, less so for Δ-learning, and not at all for densities (See App. G). Supervising model outputs in the NON basis lacks a similar justification, and Kaniselvan et al. (2026) report that matrix-element errors correlate only weakly with resulting eigenvalue errors in the NON basis, and we observe the same for downstream observables (App. M.1). In the few works that *do* supervise in the ON basis (Nigam et al., 2022; Cignoni et al., 2024; Zhang et al., 2026), the implications of ON supervision for physical observables remain poorly understood (Nigam et al., 2022), which might explain the lack of adoption. More recently, Li et al. (2025b) introduce an ON Hamiltonian loss-term, but perturb it (App. F), and combine it with conventional NON-supervision. We aim to close this gap by directly comparing ON-Loss and the NON-baseline **analytically via worst-case observable error bounds** (Sec. 4) and **empirically via correlations between loss residuals and observable errors** (Sec. 5), providing conclusive evidence in favor of ON-basis supervision.

Unrestricted density prediction in the space of symmetric matrices provides no guarantee for the physicality of $\hat{\mathbf{P}} \notin \mathcal{M}_{(N_{\text{elec}},B)}$. One way to restore the validity is to purify $\hat{\mathbf{P}}$ post-hoc (Lehtola, 2019), either by projecting onto $\mathcal{M}_{(N_{\text{elec}},B)}$ (Zhang et al., 2026) or by charge rescaling (charge conservation) and McWeeny purification (Yescas-Ramos et al., 2026), which only works within a basin of attraction (McWeeny, 1960). **Neither incorporates the constraint into the model**, correcting only at evaluation time (Zhang et al., 2026) or adding trace and idempotency loss terms (Yescas-Ramos et al., 2026). In contrast, **GROOT$_{\text{rot}}$ learns to transport an initial guess to the ground state $\hat{\mathbf{P}}^* \in \mathcal{M}_{(N_{\text{elec}},B)}$ end-to-end** (Sec. 4), while **GROOT$_{\text{proj}}$ includes the manifold projection in its parameter gradient computations**. The Hamiltonian-side degree of freedom of the gEVP (1) has been addressed for periodic structures (Wang et al., 2024; Yin et al., 2025). The molecular Hamiltonian losses in previous work leave this degree of freedom untreated, such that the model allocates expressivity to fit a component that no derived observable constrains. **ROCKET combines Δ-learning with a per-sample shift-invariant loss for molecular Hamiltonians** (Sec. 4).

4 METHODS

ON-Loss. When learning ground-state descriptors $\mathbf{X} \in \{\mathbf{H}, \mathbf{P}\}$, we aim to train on a loss that closely tracks the observable errors in the vicinity of the ground state. The tightest target-specific supervision is achieved by directly training on the observables. Although suitable for task-specific readouts used in MLIPs (Sec. 3), this ansatz deviates from the goal of learning general descriptors that support observables that are unspecified at training time. Instead, we propose to **minimize the Frobenius norm of matrix errors $\Delta\mathbf{X} = \hat{\mathbf{X}} - \mathbf{X}^*$ in their ON-basis representation** (Eq. 3)

$$\mathcal{L}_{\mathrm{ON}}(\Delta\mathbf{X}, \mathbf{S}) \equiv \mathcal{L}_{\mathrm{ON}}(\Delta\widetilde{\mathbf{X}}) := \frac{\|\Delta\widetilde{\mathbf{X}}\|_F}{B}\,. \tag{7}$$

This contrasts the standard approach of prior work, which directly supervises $\Delta\mathbf{X}$ in the non-orthonormal **NON**-basis (Sec. 3). The overlap matrix $\mathbf{S}$ mediates between matrix-errors ΔX and observable-errors (Table 1). Its eigenvalues satisfy $0 < \lambda_{\min} \leq 1 \leq \lambda_{\max}$ (Eq. 19) and act as a proxy for the relative importance of each direction w.r.t. physical observables, such that the **spectrum $\lambda(\mathbf{S})$ emphasizes some matrix-error directions more than others**. These spectral factors only vanish in ON-bases $\mathbf{S} = \mathbf{I}$ where the loss→observable error-propagation becomes approximately isotropic, which we also confirm empirically in section 5.

Table 1: **Worst-case leading-order observable errors in terms of matrix errors.** $\lambda_{\min}$ and $\lambda_{\max}$ are the extreme eigenvalues of the overlap matrix $\mathbf{S}$ for a given molecule, and $\Delta\epsilon_{\mathrm{HL}}$ its HOMO-LUMO gap. The overlap factors are the worst-case conversion factors from non-orthonormal to orthonormal matrix errors, attained along the extreme overlap eigenvectors. Derivations are in Appendix C.

Observable		Leading order in $\|\cdot\|_F$
Unconstrained $\Delta\mathbf{P}$		
E_{KS}	$\propto$	$\lambda_{\max}\|\Delta\mathbf{P}\|_F$
$\mu_\rho/L^1_\rho/L^2_\rho$	$\propto$	$\lambda_{\max}\|\Delta\mathbf{P}\|_F$
Manifold Constrained $\Delta\mathbf{P}$		
$E_{\mathrm{KS}}[\rho]$	$\propto$	$\lambda^2_{\max}\|\Delta\mathbf{P}\|^2_F$
$\mu_\rho/L^1_\rho/L^2_\rho$	$\propto$	$\lambda_{\max}\|\Delta\mathbf{P}\|_F$
Unconstrained $\Delta\mathbf{H}$		
E_{KS}	$\propto$	$\Delta\epsilon^{-2}_{\mathrm{HL}}\lambda^{-2}_{\min}\|\Delta\mathbf{H}\|^2_F$
$\mu_\rho/L^1_\rho/L^2_\rho$	$\propto$	$\Delta\epsilon^{-1}_{\mathrm{HL}}\lambda^{-1}_{\min}\|\Delta\mathbf{H}\|_F$

Constraining density GSMs to the manifold $\mathcal{M}_{(N_{\mathrm{elec}},B)}$ guarantees the physical validity of their predictions $\hat{\mathbf{P}}$ and further tightens the observable error bounds. While the dipole, L_1 and L_2 real-space density errors remain linear, the total-energy error becomes quadratic in $\|\Delta\widetilde{\mathbf{P}}\|'_F$ (Table 1). A priori, neither the initial guess nor the learned predictions are restricted to the Grassmannian. We therefore introduce Grassmann Restricted Occupied-Orbital Training (GROOT) to predict densities that lie on the Grassmannian manifold by construction. Following Lehtola (2019), we define an idempotent projection map $\mathcal{R} : \mathbf{P} \mapsto \mathbf{C}^{(\mathbf{P})}_O \mathbf{C}^{(\mathbf{P})\mathsf{T}}_O \in \mathcal{M}_{(N_{\mathrm{elec}},B)}$ that purifies any candidate density matrix. Given an arbitrary trial density $\mathbf{P}$, we diagonalize it in the orthonormal basis to extract its eigenvectors $\mathbf{C}^{(\mathbf{P})}$, retaining the O eigenvectors $\mathbf{C}^{(\mathbf{P})}_O$ corresponding to the largest eigenvalues. This yields the valid density matrix closest to $\mathbf{P}$ in the Frobenius norm (Eckart & Young, 1936). **We evaluate two approaches for predicting valid density matrices ($\mathbf{GROOT}_{\mathbf{rot}}$, $\mathbf{GROOT}_{\mathbf{proj}}$)**, using $\mathcal{R}$ either as a pre- or postprocessing step. $\mathbf{GROOT}_{\mathbf{rot}}$ is a one-step flow map (Boffi et al., 2025) that first projects the initial guess $\mathbf{P}_0 \equiv \mathcal{R}(\mathbf{P}_{\mathrm{MINAO}})$ and then learns the constant velocity $\hat{\mathbf{X}}$, Eq. (34), of the flow

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}(t) = \left[\hat{\mathbf{X}}, \mathbf{P}(t)\right] = \hat{\mathbf{X}}\mathbf{P}(t) - \mathbf{P}(t)\hat{\mathbf{X}}, \quad \mathbf{P}(0) = \mathbf{P}_0. \tag{8}$$

This flow equation has a closed-form solution $\mathbf{P}(t) = \exp(t\hat{\mathbf{X}})\,\mathbf{P}_0\exp(-t\hat{\mathbf{X}})$ that transports the initial projection $\mathbf{P}_0$ to the predicted density $\mathbf{P}(1) = \hat{\mathbf{P}}$ (Sakurai & Napolitano, 2020). The underlying transformation of the initial guess corresponds to a **rotation** $\hat{\mathbf{U}}(t) = \exp(t\hat{\mathbf{X}}) \in SO(B)$ and preserves the Grassmannian constraints of Eq. 6. These transformations $\mathbf{U}$ are also referred to as *orbital rotations* in this context. Wong (1967) showed that any two densities on $\mathcal{M}_{(N_{\mathrm{elec}},B)}$ are connected by pure mixing rotations. At the initial guess $\mathbf{P}_0$ these are generated by skew-symmetric matrices $\hat{\mathbf{X}}^\mathsf{T} = -\hat{\mathbf{X}}$ that *anti*commute with the reflection $\mathbf{R}_0 = 2\mathbf{P}_0 - \mathbf{I}$, i.e. $\hat{\mathbf{X}}\mathbf{R}_0 = -\mathbf{R}_0\hat{\mathbf{X}}$. We map any equivariant readout-head output $\hat{\mathbf{M}}$ to satisfy these conditions by applying $\hat{\mathbf{X}}(\hat{\mathbf{M}}) = \mathbf{P}_0\hat{\mathbf{M}}\mathbf{\Pi}_0 - \mathbf{\Pi}_0\hat{\mathbf{M}}^\mathsf{T}\mathbf{P}_0$, where $\mathbf{\Pi}_0 \equiv \mathbf{I} - \mathbf{P}_0$ is the projector onto the virtual orbitals. Since the unit-time map $\mathbf{P}(1)$ is available in closed form, we can bypass numerical ODE integration by directly supervising the prediction $\hat{\mathbf{P}}$ against the target density. (For details, see App. B.) $\mathbf{GROOT}_{\mathbf{proj}}$ enforces validity by projecting the symmetrized matrix readout $\hat{\mathbf{M}}$ to the Grassmannian manifold $\hat{\mathbf{P}} = \mathcal{R}(\mathbf{P}_{\mathrm{MINAO}} + \frac{1}{2}(\hat{\mathbf{M}} + \hat{\mathbf{M}}^\mathsf{T}))$ with the **parameter gradients flowing through** $\mathcal{R}$.

ROCKET is a loss function for Hamiltonian learning. Unlike GROOT which *enforces* additional *physical* constraints, ROCKET *lifts* optimization pressure from *unphysical or irrelevant* degrees of

freedom. **ROCKET systematically treats *gauge freedoms* of $\hat{\mathbf{H}}$**, meaning transformations of $\hat{\mathbf{H}}$ which leave the coefficients $\hat{\mathbf{C}} := \mathbf{C}(\hat{\mathbf{H}})$ unchanged. In the ON-basis, one such transformation is $\tilde{\mathbf{H}} \mapsto \tilde{\mathbf{H}} + c\mathbf{I}$ which only shifts the *eigenvalues* ϵ_i but does not affect the eigenvectors $\tilde{\mathbf{C}}$. We transfer this invariance to ON-Loss (Eq. 7) using a closed-form minimizer $c^* = \sum_\mu^B \Delta\widetilde{H}_{\mu\mu}/B$ and obtain

$$\mathcal{L}^{(\text{1st stage})}_{\text{ROCKET}}(\Delta\widetilde{\mathbf{H}}) := \min_c \mathcal{L}_{\text{ON}}(\Delta\widetilde{\mathbf{H}} - c\mathbf{I}) \,. \tag{9}$$

This leaves a flat $c\,\mathbf{S}$ loss-direction for every molecule. A pure data-normalization ansatz with a training-set-optimal gauge c_{t} and a standard matrix loss would strictly penalize gauge variation between molecules, while Eq. (9) permits arbitrary gauge variation. In between these extremes

$$\mathcal{L}^{(\text{2nd stage})}_{\text{ROCKET}}(\Delta\widetilde{\mathbf{H}}) := \mathcal{L}^{(\text{1st stage})}_{\text{ROCKET}}(\Delta\widetilde{\mathbf{H}}) + \alpha_{\text{ROCKET}}\ (c^* - \bar{c})^2 \tag{10}$$

softly restrains cross-molecule variation along the gauge direction, which performs best on our validation ablation (App. L.3). The reference gauge $\bar{c}$ in Eq. (10) is an exponential moving average (EMA) of the per-molecule optimal shifts c^*. Additionally, **we remove supervision from Hamiltonian directions that do not affect the lowest $O+k$ eigenvectors $\mathbf{C}_{O+k}(\hat{\mathbf{H}})$, leaving $\mathbf{P}(\mathbf{C}(\hat{\mathbf{H}}))$ unchanged.** We define the matrix $\widetilde{\mathbf{Q}}_{O+k}$ that projects onto the lowest $O+k$ reference eigenvectors as $\widetilde{\mathbf{Q}}_{O+k} := \widetilde{\mathbf{C}}^*_{O+k}\widetilde{\mathbf{C}}^{*\top}_{O+k}$, and mask out errors entirely within its orthogonal complement

$$\Delta\widetilde{\mathbf{H}}_{O+k} := \Delta\widetilde{\mathbf{H}} - (\mathbf{I} - \widetilde{\mathbf{Q}}_{O+k})\Delta\widetilde{\mathbf{H}}(\mathbf{I} - \widetilde{\mathbf{Q}}_{O+k}). \tag{11}$$

Plugging this into Eq. (10) yields the 3rd and final stage of ROCKET

$$\mathcal{L}^{(\text{3rd stage})}_{\text{ROCKET}}(\Delta\widetilde{\mathbf{H}}) := \mathcal{L}^{(\text{2nd stage})}_{\text{ROCKET}}(\Delta\widetilde{\mathbf{H}}_{O+k}) \,. \tag{12}$$

The lowest O eigenvectors determine the density matrix, while **retaining an additional k eigenvectors increases the robustness to prediction errors by separating the occupied states from the first unsupervised eigenstate.** This allows for a controlled tradeoff between learning a smooth but higher-dimensional Hamiltonian target vs learning a discontinuous but lower-dimensional density target. We use $k = 1$ by default and ablate larger k in Appendix E.

5 Experiments

We assess all models on the same equivariant GNN backbone (App. J), varying only the matrix readout. For the Hamiltonian ansatz, we obtain the density by solving the gEVP (Eq. 1), which is computationally cheap compared to the Fock-build (Eberhard et al., 2026a). **To evaluate size extrapolation, we benchmark our models across two complementary dataset regimes.** In the first regime, we train on equilibrium geometries from **QM9** (Ramakrishnan et al., 2014), restricted to molecules with at most 20 atoms and elements H, C, N, O, and F (Tab. 4). We then evaluate these models on the CHNOF subset of **QM40** (Madushanka et al., 2024) with up to 37 heavy atoms (Tab. 5), with molecules up to four times larger than those of the training distribution. In the second regime, we train on QMugs (Isert et al., 2021), testing extrapolation on conformer diversity searched at 300 °K across a wider chemical space (Tab. 6). Here, we **train on all QMugs molecules with up to 150 electrons and evaluate on systems with up to 660 electrons** (Tab. 6). Across all datasets, we generate reference density and Hamiltonian matrices with restricted KS-DFT at the B3LYP/def2-SVPStephens et al. (1994); Weigend & Ahlrichs (2005) level of theory using GPU-accelerated `GPU4PySCF` (Li et al., 2025a; Sun et al., 2020) (App. H). **To obtain the strongest possible baselines** and accurately characterize relative performances, we perform extensive ablations of both absolute and Δ-learning approaches (App. M) in our initial training on QM9, and select learning rates and weight decay for each method separately based on a 75-molecule QM40 validation set (App. L). All other hyperparameters are kept constant, and we observe no difference in training cost across all ablated ground-state predictors and loss combinations. We avoid the hidden Fock-build $\mathbf{H}(\mathbf{P}_{\text{MINAO}})$ for Hamiltonian Δ-learning (Eberhard et al., 2026a) by using the superposition of atomic potentials (SAP) ansatz $\mathbf{H}_{\text{SAP}}$ (Lehtola, 2019), and additionally ablate training-set-optimal gauge shifts ($c_t\mathbf{S}$) for the Hamiltonian baseline. For Δ-learning in density space, we use $\mathbf{P}_{\text{MINAO}}$ (Lehtola, 2019). When evaluating the energies from the predicted $\hat{\mathbf{P}}$ or $\mathbf{P}(\hat{\mathbf{H}})$ density matrix, we get the **first order energy correction in the self-consistency error** $\Delta^{\text{SC}}\mathbf{P} := \mathbf{P}(\mathbf{H}(\hat{\mathbf{P}})) - \hat{\mathbf{P}}$ to the ground-state energy expression of KS-DFT E_{KS} essentially for free, such that our predictions are given by $\hat{E} = E_{\text{KS}}(\hat{P}) + \sum_{\mu\nu} H_{\mu\nu}(\hat{P})\Delta^{\text{SC}}P_{\mu\nu}$ (Harris, 1985; Foulkes & Haydock, 1989).

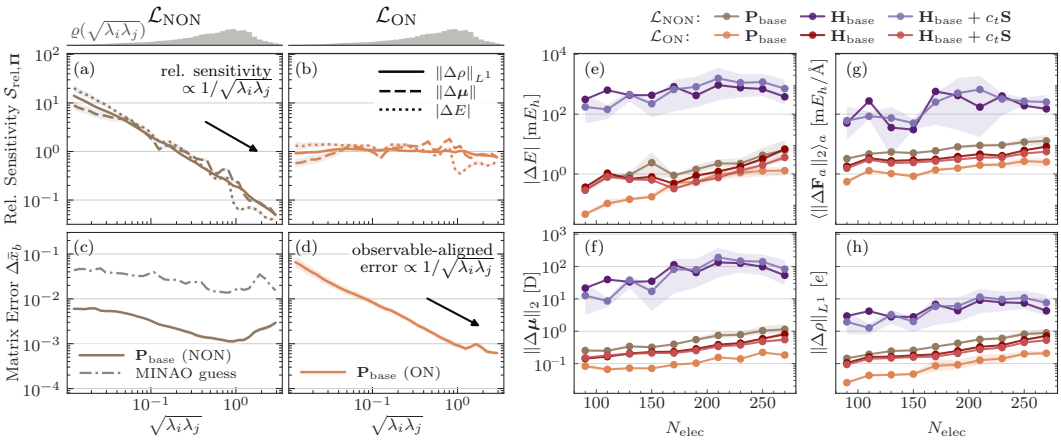


Figure 2: **ON-Loss is "observable-isotropic" and improves size extrapolation. (a,b,c,d)** Values log-binned by $\sqrt{\lambda_i\lambda_j}$, matching the directions $\mathbf{\Pi}_{ij}$ of QM40 molecules (Table 5); shading indicates the standard deviation, histograms show the direction-density ϱ. **(a,b)** Mean-normalized relative sensitivity for the L^1 density, dipole μ, and energy E errors (line styles): **(a)** $\mathcal{L}_{\text{NON}}$ falls as $1/\sqrt{\lambda_i\lambda_j}$, whereas **(b)** $\mathcal{L}_{\text{ON}}$ stays approximately constant for all three observables. **(c,d)** Matrix error $\Delta\bar{x}_b \equiv \sqrt{\langle \text{tr}(\mathbf{\Pi}_{ij}\Delta\mathbf{X})^2\rangle_b}$ of the trained network, with P_{MINAO} for reference: **(c)** $\mathcal{L}_{\text{NON}}$ leaves error that follows ϱ rather than observable impact, while **(d)** $\mathcal{L}_{\text{ON}}$ concentrates it in the high-impact directions. **(e–h)** Size extrapolation experiments from **QM9 to QM40**, featuring the same observables plus the force F acting on the atoms plotted against the electron count N_{elec}. We compare GSMs under both $\mathcal{L}_{\text{NON}}$ and $\mathcal{L}_{\text{ON}}$ loss functions, as well as Hamiltonian variants with a training-set-optimal constant gauge shift c_t. Markers are means and bands the minimum and maximum across identical runs; **every $\mathcal{L}_{\text{ON}}$ arm beats its $\mathcal{L}_{\text{NON}}$ counterpart.**

Loss Observable Alignment. We compare the alignment of $\mathcal{L}_{\text{ON}}$ and $\mathcal{L}_{\text{NON}}$ by investigating their agreement with key observables in the vicinity of the target density $\mathbf{P}^*$. To that end, we define the *sensitivity* $\mathcal{S}_{\mathbf{\Pi}}(\mathcal{O}) \in \mathbb{R}$ of an observable $\mathcal{O}$ (or Loss $\mathcal{L}$) w.r.t. a perturbation direction $\mathbf{\Pi} \in \mathbb{R}^{B\times B}$ as the absolute directional derivative (App. D.1). The ***relative directional sensitivity*** (App. D) $\mathcal{S}_{\text{rel},\mathbf{\Pi}}(\mathcal{L},\mathcal{O}) := \mathcal{S}_{\mathbf{\Pi}}(\mathcal{L}) \;/\; \mathcal{S}_{\mathbf{\Pi}}(\mathcal{O})$ captures if the loss changes more (or less) rapidly than the actual observable along a given direction. We analyze the sensitivities in a basis of symmetric matrices built from the eigenvectors $\mathbf{u}_k$ of the overlap $\mathbf{S} = \sum_k \mathbf{u}_k\mathbf{u}^\mathsf{T}\lambda_k$, with normalized elements $\mathbf{\Pi}_{ij} \propto \mathbf{u}_i\mathbf{u}_j^\mathsf{T} + \mathbf{u}_j\mathbf{u}_i^\mathsf{T}$. Our experiments show that the optimization process prioritizes directions according to their loss sensitivity $\mathcal{S}_{\mathbf{\Pi}}(\mathcal{L}_{\text{NON}}) = 1/B$ and $\mathcal{S}_{\mathbf{\Pi}}(\mathcal{L}_{\text{ON}}) = \sqrt{\lambda_i\lambda_j}/B$ for density matrices. We demonstrate this by comparing the matrix error components in the overlap basis, $\Delta\bar{x}_b := \sqrt{\langle \text{tr}(\mathbf{\Pi}_{ij}\Delta\mathbf{X})^2\rangle_b}$, binned by $\sqrt{\lambda_i\lambda_j}$ (Fig. 2 (b,d), and App. D). An ideal loss function should have direction-independent (isotropic) relative sensitivities $\mathcal{S}_{\text{rel},\mathbf{\Pi}}(\mathcal{L},\mathcal{O}) \approx c^{(\mathcal{O})}\;\forall\,\mathbf{\Pi}\,\forall\,\mathcal{O}$ across a diverse set of observables. Figure 2 (a,c) shows that this is approximately achieved by ON-Loss and violated by the NON baseline. Training on ON-Loss strictly outperforms this baseline on QM40 across model configurations on all tested observables (see Fig. 1 and Fig. 2 (e-h)). Across all evaluated models, we observe higher mutual information between the ON-Loss residual and downstream observable errors (Tab. 24), independent of the training loss that produced the model weights.

Size-Extrapolation. Reference calculations for KS-DFT ground states scale with the 3rd or 4th power with system size, depending on the level of theory. This restricts training data for ML surrogates to small to mid-sized molecules. Their practical utility, therefore, depends in part on their ability to accurately extrapolate to larger systems. In our extrapolation experiments on equilibrium structures from QM9 and QM40 (Fig. 1, Fig. 2 (e-h), App. N.1), the $\mathcal{L}_{\text{NON}}$ baselines generalize poorly to larger molecules, whereas ON-Loss consistently reduces prediction errors across both density and Hamiltonian representations (Fig. 3, Tab. 28). ROCKET and GROOT compound on top of these improvements: Although ROCKET already outperforms *all* tested Hamiltonian GSMs including the new ON-Loss on *all* tested observables, it is significantly improved upon by GROOT on almost all observables, with $\text{GROOT}_{\text{rot}}$ leading by a small margin over $\text{GROOT}_{\text{proj}}$ (see Fig. 1). Excluding the manifold projection from the training gradients and applying it only post hoc to the ON-Loss baseline rejects 66% of the molecules as unreliable, compared to 0.6% for $\text{GROOT}_{\text{proj}}$ (table 26).

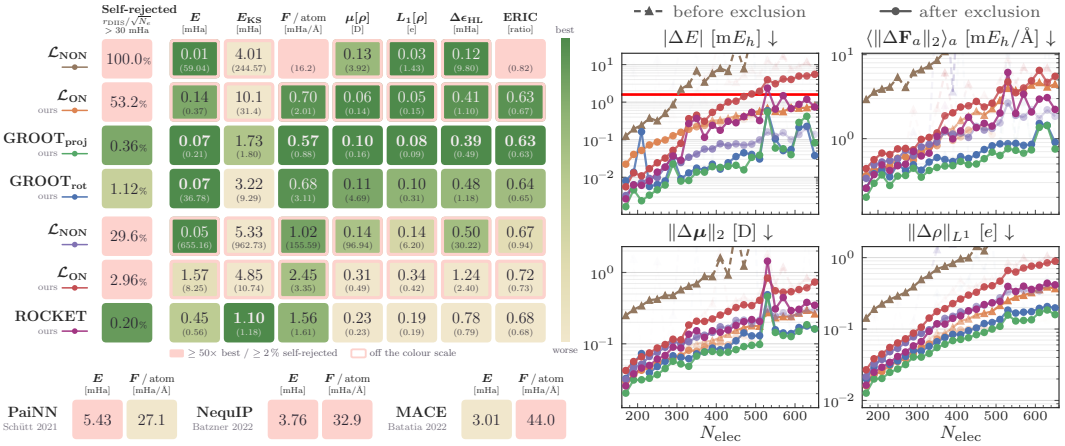


Figure 3: **GROOT and ROCKET show robust extrapolation, and can be equipped with GSMs-specific error indicators** (Tab. 28). GSMs and MLIPs trained on QMugs molecules with at most 150 electrons and evaluated on systems with up to 660 electrons. **The table** shows the fraction of molecules rejected by the self-consistency criterion and the observable errors of the remaining molecules, with the MAE including the rejected predictions in brackets below. Compared to linear scaling MLIPs like PaiNN (Schütt et al., 2021), NequIP (Batzner et al., 2022), and MACE (Batatia et al., 2022) trained on energies only (meaning at equal reference compute budget), $\text{GROOT}_{\text{proj}}$ achieves an order of magnitude better energies, rejecting less than 1% of its own predictions, characterizing the achievable cost-accuracy trade-off of GSMs compared to MLIPs. The ERIC (Eberhard et al., 2026a) relative to MINAO shows that our density models yield additional SCF acceleration over the baselines. **Observable MAEs vs. electron count.** GROOT and ROCKET stay orders of magnitude below the $\mathcal{L}_{\text{NON}}$ baselines with the margin widening as the system size increases.

Both versions of GROOT achieve energy errors of less than half *chemical accuracy* (1 kcal/mol) for 98% of the QM40 molecules in the test set with the vast majority of errors passing the even stricter threshold of $1/16\,\text{kcal/mol}$. This motivates the search for an **uncertainty measure that quantifies prediction confidence**. We identify the residual term $r = \sqrt{8}\,\|\tilde{H}\tilde{P} - \tilde{P}\tilde{H}\|_F$, commonly used in convergence acceleration methods for KS-DFT calculations (Pulay, 1980; 1982), as a suitable measure to equip GSMs with the approximately **size-intensive rejection criterion** $\|r\|_F/\sqrt{N_{\text{elec}}} > \varepsilon_{\text{SC}}$. Ideally, the self-consistency threshold ε_{SC} could be calibrated on a large validation set per model. For our experiments, we settled on uniform thresholds for all models, which we estimated based on the error distribution of a small large-molecule validation set to $\varepsilon_{\text{SC}} = 30\,\text{mE}_\text{h}$. **Besides size extrapolation** we also isolated the conformational generalization capabilities of GSMs by training on QMugs and evaluating on the 3BPA (Kovács et al., 2021) torsional surface (Fig. 4, table 29). We observe excellent performance for all our GSMs: compared to PaiNN, NequIP, and MACE, ROCKET reduces energy errors by roughly an order of magnitude (8–12×) and force errors by 5–7×, while the ON-Loss-trained density GSM $\mathbf{P}_{\text{base}}$ (ON) reduces them by 148–234× and 20–29×, respectively.

Towards foundational GSMs. Our evaluations suggest that *well-designed* GSMs can achieve significantly improved energies compared to MLIPs, both trained on energies only and trained on energies and forces, while force-trained MLIPs still lead in force accuracy (Fig. 1, Fig. 3). Despite their significantly higher inference time costs compared to MLIPs, viewed from an **equal reference compute perspective**, GSMs offer a highly interesting trade-off making them great foundation models. Not only do GSMs outperform energy-only trained MLIPs in terms of force accuracy by a large margin ($\sim \times 30 - 50$, Fig. 1), but they also offer **access to ground-state observables undefined at training time**. Moreover, **pretrained GSMs are conveniently adaptable to new chemistry by generalizing the idea of *label-free* Self-Consistency Training (SCT)** introduced by Zhang et al. (2024) to density GSMs. Depending on the application the resulting amortization cost analysis may be highly beneficial especially for research applications with low total evaluation counts since this incurs a much lower up front fine-tuning cost. We apply SCT to our QMugs-trained models and finetune them on the Transition1x dataset (Schreiner et al., 2022), containing molecular geometries along reaction pathways. Except for $\text{GROOT}_{\text{rot}}$ and $\mathbf{P}_{\text{base}}$ (ON), which have uncomfortably high rejection rates of 9%, all tested GSMs perform excellently on this task reaching far below chemical accuracy on the predictions accepted by the self-consistency rejection rule (Fig. 4, table 30).

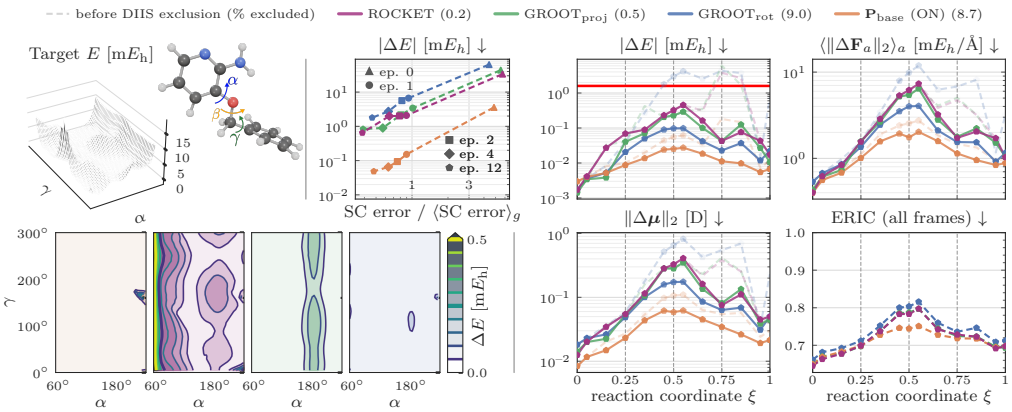


Figure 4: **Beyond size extrapolation: generalization across conformations and label-free adaptation to new chemistry.** QMugs-trained models generalize to the $\beta = 120°$ cut of the 3BPA backbone-dihedral energy surface (left panel, Tab. 29): the ground truth (α, γ) energy-surface, and the energy deviations of GSMs, with median (MAE-optimal) constant offset removed per model. **Label-free self-consistency training (SCT) on Transition1x.** The upper-right panel traces the validation error against the self-consistency residual over 0, 1, 2, 4, and 12 epochs of cosine-annealed SCT, suggesting that shorter training budgets might be feasible. The observable MAEs are resolved along the reaction coordinate ξ, with the transition state at $\xi = 0.5$. The energies of the finetuned GSMs consistently achieve chemical accuracy (dashed line), with ROCKET triggering the least amount of rejections.

6 Discussion

We have aligned GSMs with the governing equations of Kohn-Sham Density Functional Theory (KS-DFT) through ON-Loss, GROOT, and ROCKET. To this end, we performed a theoretical and empirical analysis of representational choices of $\mathbf{P}, \mathbf{H}$ for GSM training and their effects on downstream observables, finding strong evidence in favor of orthonormal (ON) basis representations (ON-Loss). Unconstrained density readouts can leave the manifold of valid density matrices, which can violate charge conservation and lead to fractional occupations, a regime where KS-DFT is known to misbehave (Perdew et al., 1982; Mori-Sánchez et al., 2008). GROOT enforces these constraints by construction and thus predicts physical density matrices. The densities obtained from Hamiltonian GSMs are inherently physical, but current loss functions overconstrain them. ROCKET reduces this excess supervision. In our QM9 to QM40 **size extrapolation** experiments, these contributions compound relative to state-of-the-art GSMs (Fig. 1). Alongside improvements across other observables, ON-Loss and GROOT reduce the energy error of the previous state-of-the-art density GSM by 79.1%, and ON-Loss and ROCKET reduce that of the strongest Hamiltonian baseline by 99.8%. Under GROOT, 98% of molecules achieve energy errors below $0.5\times$ chemical accuracy. On QMugs systems up to 660 electrons (table 28), GROOT$_{\text{proj}}$ and ROCKET reduce energy errors by 99.6% and 99.9%, respectively, relative to the corresponding NON baselines. Evaluating GSMs on energy observables requires a Fock build incurring significantly higher inference cost than MLIPs, but it yields an **inherent uncertainty measure** through the self-consistency residual. This self-rejection can filter large extrapolation errors while maintaining high yield. Our GSM framework adopts a generalization of **label-free fine-tuning** via Self-Consistency Training (Zhang et al., 2024) to density GSMs. When transferring QMugs-pretrained models to the reactive chemistry of Transition1x (Schreiner et al., 2022), this reaches energy errors well below chemical accuracy across the entire reaction coordinate (Fig. 4, table 30).

Limitations & Future Work. Our evaluations are restricted to closed-shell Kohn-Sham. Although a generalization to unrestricted systems is straightforward, this has not been covered by our evaluations. All experiments have been performed at the B3LYP/def2-SVP level of theory. Transfer to other functionals and basis sets remains untested. The unique capabilities of GSMs could enable active-learning pipelines for MLIP reference data generation on large molecules (Kulichenko et al., 2024). For each molecule, GROOT can predict the ground state at the cost of a single Fock build. If the self-consistency residual falls below the rejection threshold, the pipeline accepts the one-iteration observables. When the residual exceeds the threshold, the calculation falls back to iterative SCF, where the prediction can serve as an initial guess that accelerates convergence. The resulting solutions can continually update the GSM. The combination of this ansatz with ML exchange-correlation functionals (Gao et al., 2024; Luise et al., 2026; Eberhard et al., 2026b) could provide exceptional cost-accuracy trade-offs.

AI USE STATEMENT

We used large language models (LLMs) to produce the leading-order Taylor expansions in Tab. 1, improve the accessibility to a wider audience (for example by reframing an equivalence class argument in terms of data normalization and identifying unnecessary complexity and physics jargon in our first human-written draft), support code development and Slurm experiment management, and challenge author-generated interpretations of results to identify possible gaps in our own explanations. The authors evaluated this feedback and determined the final interpretation. Beyond the research uses, LLMs have been used for figure preparation, code editing and review, **appendix-only** captions, sentence-level readability revisions, literature discovery, and citation-key completion for author-selected sources.

We **did not** use LLMs to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, translate research material, or clean or reformat datasets. We also did not use LLMs to discover the research topic or identify the research gap, or suggest the structure of the paper. Qualitative and thematic data analysis and transcription of recordings are not applicable to this work.

The authors independently checked all AI-assisted mathematical claims, derivations, and proofs. We reviewed and tested all AI-assisted code used for the reported results. We inspected all AI-assisted figures against the underlying results. The authors independently read and assessed every AI-suggested paper that we cited. We reviewed all AI-assisted manuscript text for correctness and clarity. LLMs also assisted with proof verification, but the authors performed the final checks. We take responsibility for the final content of this work, including all text, claims, code, and figures produced with the aid of LLMs.

REPRODUCIBILITY STATEMENT

We provide proofs for all theoretical claims in the appendix, including the reachability, geometric validity, and expressivity of GROOT on the Grassmann manifold (section B) as well as the observable error bounds and quadratic energy scaling for valid density projectors (section C). Across all benchmarks, we document dataset sources, reference calculation protocols, and evaluation splits in section H, and specify model architectures and training hyperparameters in sections I and L. In the supplementary material, we upload a self-contained data analysis notebook that demonstrates our evaluation and the effectiveness of the first-order energy correction.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. doi: 10.1515/9781400830244. URL https://doi.org/10.1515/9781400830244.

J. Almlöf, K. Faegri, and K. 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.

K. Antypas, B. A. Austin, T. L. Butler, R. A. Gerber, C. L. Whitney, N. J. Wright, W. Yang, and Z. Zhao. NERSC workload analysis on hopper. Technical report, National Energy Research Scientific Computing (NERSC) Center, 2014.

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/google-deepmind.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022. doi: 10.52202/068431-0830. URL https://doi.org/10.52202/068431-0830.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

A. 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.

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.

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=cqDH0e6ak2.

Andreas Burger, Luca Thiede, Alan Aspuru-Guzik, and Nandita Vijaykumar. DEQuify your force field: More efficient simulations using deep equilibrium models. In *ICLR 2025 Workshops: AI4MAT*, 2025a. URL https://mlanthology.org/iclrw/2025/burger2025iclrw-dequify/.

Andreas Burger, Luca Thiede, Nikolaj Rønne, Varinia Bernales, Nandita Vijaykumar, Tejs Vegge, Arghya Bhowmik, and Alan Aspuru-Guzik. Shoot from the HIP: Hessian Interatomic Potentials without derivatives, 2025b. URL https://arxiv.org/abs/2509.21624.

B. C. Carlson and J. M. Keller. Orthogonalization procedures and the localization of Wannier functions. *Physical Review*, 105(1):102–103, 1957. doi: 10.1103/PhysRev.105.102. URL https://doi.org/10.1103/PhysRev.105.102.

Edoardo Cignoni, Divya Suman, Jigyasa Nigam, Lorenzo Cupellini, Benedetta Mennucci, and Michele Ceriotti. Electronic Excited States from Physically Constrained Machine Learning. *ACS Central Science*, 10(3):637–648, 2024. doi: 10.1021/acscentsci.3c01480. URL https://doi.org/10.1021/acscentsci.3c01480.

Christopher J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. Wiley, Chichester, 2 edition, 2014. ISBN 978-0-470-09181-4.

Sebastian Dick and Marivi Fernandez-Serra. Highly accurate and constrained density functional obtained with differentiable programming. *Physical Review B*, 104(16):L161109, October 2021. doi: 10.1103/PhysRevB.104.L161109. URL https://doi.org/10.1103/PhysRevB.104.L161109.

Eike S. Eberhard, Viktor Kotsev, Timm Güthle, and Stephan Günnemann. Transferable SCF-acceleration through solver-aligned initialization learning, 2026a. URL https://arxiv.org/abs/2604.21657v2.

Eike S. Eberhard, Luca A. Thiede, Abdulrahman Aldossary, Andreas Burger, Nicholas Gao, Vignesh Bhethanabotla, Alán Aspuru-Guzik, and Stephan Günnemann. Derivative informed learning of exchange-correlation functionals. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2606.04279.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

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.

Pol Febrer, Peter Bjørn Jørgensen, Miguel Pruneda, Alberto García, Pablo Ordejón, and Arghya Bhowmik. Graph2mat: Universal graph to matrix conversion for electron density prediction. *Machine Learning: Science and Technology*, 6(2):025013, 2025. doi: 10.1088/2632-2153/adc871. URL https://doi.org/10.1088/2632-2153/adc871.

W. Matthew C. Foulkes and Roger Haydock. Tight-binding models and density-functional theory. *Physical Review B*, 39(17):12520–12536, 1989.

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, 2025. URL https://arxiv.org/abs/2502.12147.

Nicholas Gao, Eike Eberhard, and Stephan Günnemann. Learning equivariant non-local electron density functionals. In *The Thirteenth International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2410.07972. URL https://doi.org/10.48550/arXiv.2410.07972.

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

Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(1):137, 2020. doi: 10.1038/s41597-020-0460-4.

J. Harris. Simplified method for calculating the energy of weakly interacting fragments. *Physical Review B*, 31(4):1770–1779, 1985.

Robin Haunschild, Andreas Barth, and Bernie French. A comprehensive analysis of the history of DFT based on the bibliometric method RPYS. *Journal of Cheminformatics*, 11(1):72, 2019. doi: 10.1186/s13321-019-0395-y. URL https://doi.org/10.1186/s13321-019-0395-y.

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.

Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. QMugs: Quantum mechanical properties of drug-like molecules, 2021. URL https://arxiv.org/abs/2107.00367.

Ryan Jacobs, Dane Morgan, Siamak Attarian, Jun Meng, Chen Shen, Zhenghao Wu, Clare Yijia Xie, Julia H. Yang, Nongnuch Artrith, Ben Blaiszik, Gerbrand Ceder, Kamal Choudhary, Gabor Csanyi, Ekin Dogus Cubuk, Bowen Deng, Ralf Drautz, Xiang Fu, Jonathan Godwin, Vasant Honavar, Olexandr Isayev, Anders Johansson, Boris Kozinsky, Stefano Martiniani, Shyue Ping Ong, Igor Poltavsky, Kj Schmidt, So Takamoto, Aidan P. Thompson, Julia Westermayr, and Brandon M. Wood. A practical guide to machine learning interatomic potentials – Status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, March 2025. ISSN 13590286. doi: 10.1016/j.cossms.2025.101214. URL https://linkinghub.elsevier.com/retrieve/pii/S1359028625000014.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Manasa Kaniselvan, Benjamin Kurt Miller, Meng Gao, Juno Nam, and Daniel S. Levine. Learning from the electronic structure of molecules across the periodic table. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d81cd83e7f6748af351485d73f305483-Abstract-Conference.html.

Seongsu Kim, Nayoung Kim, Dongwoo Kim, and Sungsoo Ahn. High-order equivariant flow matching for density functional theory hamiltonian prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0449. URL https://doi.org/10.52202/085713-0449.

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, 2026. URL https://arxiv.org/abs/2602.16897.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133. URL https://doi.org/10.1103/PhysRev.140.A1133.

Teddy Koker, Mit Kotak, and Tess Smidt. Training a Foundation Model for Materials on a Budget, 2025. URL https://doi.org/10.48550/arXiv.2508.16067.

Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice E. A. Allen, Daniel J. Cole, Christoph Ortner, and Gábor Csányi. Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE. *Journal of Chemical Theory and Computation*, 17(12):7696–7711, 2021. doi: 10.1021/acs.jctc.1c00647. URL https://doi.org/10.1021/acs.jctc.1c00647.

Dávid Péter Kovács, Ilyes Batatia, Eszter Sára Arany, and Gábor Csányi. Evaluation of the MACE force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics*, 159(4):044118, July 2023. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0155322. URL https://pubs.aip.org/jcp/article/159/4/044118/2904837/Evaluation-of-the-MACE-force-field-architecture.

Maksim Kulichenko, Benjamin Nebgen, Nicholas Lubbers, Justin S. Smith, Kipton Barros, Alice E. A. Allen, Adela Habib, Emily Shinkle, Nikita Fedik, Ying Wai Li, Richard A. Messerly, and Sergei Tretiak. Data generation for machine learning interatomic potentials and beyond. *Chemical Reviews*, 124(24):13681–13714, 2024. doi: 10.1021/acs.chemrev.4c00572. URL https://doi.org/10.1021/acs.chemrev.4c00572.

Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988. doi: 10.1103/PhysRevB.37.785. URL https://doi.org/10.1103/PhysRevB.37.785.

Susi Lehtola. Assessment of initial guesses for self-consistent field calculations. superposition of atomic potentials: Simple yet efficient. *Journal of Chemical Theory and Computation*, 15(3):1593–1604, 2019. doi: 10.1021/acs.jctc.8b01089. URL https://doi.org/10.1021/acs.jctc.8b01089.

Susi Lehtola, Frank Blockhuys, and Christian Van Alsenoy. An overview of self-consistent field calculations within finite basis sets. *Molecules*, 25(5):1218, 2020a. doi: 10.3390/molecules25051218. URL https://doi.org/10.3390/molecules25051218.

Susi Lehtola, Lucas Visscher, and Eberhard Engel. Efficient implementation of the superposition of atomic potentials initial guess for electronic structure calculations in Gaussian basis sets. *The Journal of Chemical Physics*, 152(14):144105, 2020b. doi: 10.1063/5.0004046. URL https://doi.org/10.1063/5.0004046.

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

Rui Li, Qiming Sun, Xing Zhang, and Garnet Kin-Lic Chan. Introducing GPU acceleration into the python-based simulations of chemistry framework. *The Journal of Physical Chemistry A*, 129(5):1459–1468, 2025a. doi: 10.1021/acs.jpca.4c05876. URL https://doi.org/10.1021/acs.jpca.4c05876.

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

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Zhe Liu, Yuyan Ni, Zhichen Pu, Qiming Sun, Siyuan Liu, and Wen Yan. Towards a transferable acceleration method for density functional theory. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/84fb08ae30ab266953d1ff72eb0611b1-Abstract-Conference.html.

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. URL https://doi.org/10.1063/1.1747632.

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, April 2026. URL https://doi.org/10.48550/arXiv.2506.14665. arXiv:2506.14665 [physics].

Erpai Luo, Xinran Wei, Lin Huang, Yunyang Li, Han Yang, Zaishuo Xia, Zun Wang, Chang Liu, Bin Shao, and Jia Zhang. Efficient and scalable density functional theory Hamiltonian prediction through adaptive sparsity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41368–41390. PMLR, 2025. URL https://proceedings.mlr.press/v267/luo25l.html.

Ayesh Madushanka, Renaldo T. Moura, and Elfi Kraka. QM40, realistic quantum mechanical dataset for machine learning in molecular science. *Scientific Data*, 11(1):1376, 2024. doi: 10.1038/s41597-024-04206-y. URL https://doi.org/10.1038/s41597-024-04206-y.

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.

Paula Mori-Sánchez, Aron J. Cohen, and Weitao Yang. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Physical Review Letters*, 100(14):146401, 2008. doi: 10.1103/PhysRevLett.100.146401. URL https://doi.org/10.1103/PhysRevLett.100.146401.

Jigyasa Nigam, Michael J. Willatt, and Michele Ceriotti. Equivariant Representations for Molecular Hamiltonians and N-Center Atomic-Scale Properties. *The Journal of Chemical Physics*, 156(1):014115, 2022. doi: 10.1063/5.0072784. URL https://doi.org/10.1063/5.0072784.

Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 27420–27438, 2023. URL https://proceedings.mlr.press/v202/passaro23a.html.

John P. Perdew, Robert G. Parr, Mel Levy, and Jose L. Balduz. Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Physical Review Letters*, 49(23):1691–1694, 1982. doi: 10.1103/PhysRevLett.49.1691. URL https://doi.org/10.1103/PhysRevLett.49.1691.

P. Pulay. Improved SCF convergence acceleration. *Journal of Computational Chemistry*, 3(4): 556–560, 1982. doi: 10.1002/jcc.540030413. URL https://doi.org/10.1002/jcc.540030413.

Péter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980. doi: 10.1016/0009-2614(80)80396-4. URL https://doi.org/10.1016/0009-2614(80)80396-4.

Chen Qian, Valdas Vitartas, James R. Kermode, and Reinhard J. Maurer. Equivariant electronic Hamiltonian prediction with many-body message passing. *npj Computational Materials*, 12 (1):169, 2026. doi: 10.1038/s41524-026-02020-1. URL https://doi.org/10.1038/s41524-026-02020-1.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

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. URL https://doi.org/10.1103/RevModPhys.23.69.

J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. Cambridge University Press, 3rd edition, 2020. doi: 10.1017/9781108587280. URL https://doi.org/10.1017/9781108587280.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1):779, 2022.

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

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9377–9388, 2021. URL https://mlanthology.org/icml/2021/schutt2021icml-equivariant/.

Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet - a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Kirill Shmilovich, Devin Willmott, Ivan Batalov, Mordechai Kornbluth, Jonathan Mailoa, and J. Zico Kolter. Orbital mixer: Using atomic orbital features for basis-dependent prediction of molecular wavefunctions. *Journal of Chemical Theory and Computation*, 18(10):6021–6030, 2022. doi: 10.1021/acs.jctc.2c00555. URL https://doi.org/10.1021/acs.jctc.2c00555.

P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry*, 98(45):11623–11627, 1994. doi: 10.1021/j100096a001. URL https://doi.org/10.1021/j100096a001.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S. Blunt, Nikolay A. Bogdanov, George H. Booth, Jia Chen, Zhi-Hao Cui, Janus J. Eriksen, Yang Gao, Sheng Guo, Jan Hermann, Matthew R. Hermes, Kevin Koh, Peter Koval, Susi Lehtola, Zhendong Li, Junzi Liu, Narbe Mardirossian, James D. McClain, Mario Motta, Bastien Mussard, Hung Q. Pham, Artem Pulkin, Wirawan Purwanto, Paul J. Robinson, Enrico Ronca, Elvira R. Sayfutyarova, Maximilian Scheurer, Henry F. Schurkus, James E. T. Smith, Chong Sun, Shi-Ning Sun, Shiv Upadhyay, Lucas K. Wagner, Xiao Wang, Alec White, James Daniel Whitfield, Mark J. Williamson, Sebastian Wouters, Jun Yang, Jason M. Yu, Tianyu Zhu, Timothy C. Berkelbach, Sandeep Sharma, Alexander Yu. Sokolov, and Garnet Kin-Lic Chan. Recent developments in the PySCF program package. *The Journal of Chemical Physics*, 153(2):024109, 2020. doi: 10.1063/5.0006074. URL https://doi.org/10.1063/5.0006074.

Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1996. ISBN 978-0-486-69186-2.

Oliver T. Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14434–14447, 2021. doi: 10.5555/3540261.3541367. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/78f1893678afbeaa90b1fa01b9cfb860-Abstract.html.

O. Vahtras, J. Almlöf, and M. W. Feyereisen. Integral approximations for LCAO-SCF calculations. *Chemical Physics Letters*, 213(5):514–518, 1993. doi: 10.1016/0009-2614(93)89151-7. URL https://doi.org/10.1016/0009-2614(93)89151-7.

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.

S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8):1200–1211, 1980. doi: 10.1139/p80-159. URL https://doi.org/10.1139/p80-159.

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

Florian Weigend. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Physical Chemistry Chemical Physics*, 4(18):4285–4291, 2002. doi: 10.1039/b204199p. URL https://doi.org/10.1039/b204199p.

Florian Weigend. Hartree–fock exchange fitting basis sets for H to Rn. *Journal of Computational Chemistry*, 29(2):167–175, 2008. doi: 10.1002/jcc.20702. URL https://doi.org/10.1002/jcc.20702.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. doi: 10.1039/B508541A. URL https://doi.org/10.1039/B508541A.

Yung-Chow Wong. Differential geometry of grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 57(3):589–594, March 1967. doi: 10.1073/pnas.57.3.589. URL https://doi.org/10.1073/pnas.57.3.589.

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, 2025.

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

Shi Yin, Zujian Dai, Xinyang Pan, and Lixin He. Advancing universal deep learning for electronic-structure Hamiltonian prediction of materials, 2025.

Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. QH9: A quantum hamiltonian prediction benchmark for QM9 molecules. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40487–40503, 2023a. doi: 10.52202/075280-1761. URL https://doi.org/10.52202/075280-1761.

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*, pp. 40412–40424. PMLR, 2023b. URL https://proceedings.mlr.press/v202/yu23i.html.

He Zhang, Chang Liu, Zun Wang, Xinran Wei, Siyuan Liu, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Self-consistency training for density-functional-theory Hamiltonian prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59329–59357. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ak.html.

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(4):1868–1880, 2026. doi: 10.1039/d5dd00230c. URL https://doi.org/10.1039/d5dd00230c.

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

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

APPENDIX CONTENTS

A Kohn–Sham Theory in a Finite Basis **19**

B Rotational GROOT: transport on the Grassmannian **20**

C Observable Bounds for the Orthonormal Loss **23**

C.1 AO and orthonormal Frobenius errors 23

C.2 Choice of orthonormal basis 23

C.3 General density errors 24

C.4 Grassmann-constrained density errors 24

C.5 Hamiltonian-induced density errors 25

C.6 Rotation covariance and the entrywise term 26

D Directional Sensitivity in the Overlap Spectrum **26**

D.1 Sensitivity as One-Sided Directional Derivatives 26

D.2 The overlap-spectrum basis of directions 27

D.3 Loss Sensitivities 28

D.4 Observable Sensitivities 28

E A Unifying Perspective on Hamiltonian and Density Learning **29**

F Relation of the Orthonormal Loss to the Wavefunction Alignment Loss **29**

G AO versus Löwdin Spatial Decay **31**

H Datasets **32**

I Model Hyperparameters **40**

I.1 Ground-State Descriptor Models 40

I.2 Machine-learned Interatomic Potentials MLIP reference models 40

J Model Implementation Details **41**

K Training Costs **41**

L Training and Optimization **42**

L.1 Density Models 43

L.2 Hamiltonian Models 45

L.3 ROCKET-Loss ablations 47

L.4 Self-Consistency-Training Learning-Rate Selection 48

M Loss vs Observables **49**

M.1 Frobenius Error versus Observable Error 50

N Model Ablation Data **54**

N.1 Training on QM9 54

N.2 Training on QMugs 58

O Runtime **63**

A KOHN–SHAM THEORY IN A FINITE BASIS ↑toc

We provide a brief overview of the mathematical fundamentals underlying KS-DFT, aimed at a broader audience at the intersection of quantum chemistry and machine learning. Hence, we refrain from any explicit references to electronic spin degrees of freedom. This spin-free treatment is sufficient for so-called closed-shell systems where $\rho^{(\text{up})} = \rho^{(\text{down})}$, requiring an even number of electrons. For further details and generalization to *open-shell* systems, we refer to the works of Szabo and Lehtola Szabo & Ostlund (1996); Lehtola et al. (2020a).

In their seminal work, Kohn & Sham (1965) showed that the ground state electron density $\rho(\mathbf{r})$ can be constructed from eigenfunctions $\varphi_i(\mathbf{r}) : \mathbb{R}^3 \to \mathbb{R}$ which are solutions of the Kohn-Sham equations

$$\left(-\frac{1}{2}\nabla^2 - \sum_{A=1}^{N_{\text{nuc}}} \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|} + \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}' + v_{\text{xc}}[\rho](\mathbf{r})\right)\varphi_i(\mathbf{r}) = \epsilon_i\varphi_i(\mathbf{r}), \tag{13}$$

which are the governing equations of KS-DFT. The first term is the kinetic energy operator, the second term the potential due to the atomic nuclei of the system, the third term the electron-electron potential, the fourth term the so-called *exchange-correlation* potential, and ϵ_i is the eigenvalue of $\varphi_i(\mathbf{r})$. The eigenfunctions $\varphi_i(\mathbf{r})$ are commonly referred to as molecular orbitals (MOs) and relate to the density via

$$\rho(\mathbf{r}) = 2\sum_{i}^{O} |\varphi_i(\mathbf{r})|^2\,, \tag{14}$$

where $O = N_{\text{elec}}/2$ and $V = B - O$ are the numbers of *occupied* and *virtual orbitals*, respectively, and the eigenfunctions are indexed lowest to highest with respect to their eigenvalues. By construction, the $\varphi_i(\mathbf{r})$ are orthonormal $\int_{\mathbb{R}^3} \varphi_i(\mathbf{r})\varphi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{ij}$ and hence $\int_{\mathbb{R}^3} \rho(\mathbf{r})\,\mathrm{d}\mathbf{r} = N_{\text{elec}}$. Eq. (13) can be discretized by representing the $\varphi_i(\mathbf{r})$ in a **finite basis**

$$\left\{\chi_\mu(\mathbf{r})\middle|\chi_\mu : \mathbb{R}^3 \mapsto \mathbb{R} \text{ and } \int_{\mathbb{R}^3} \chi_\mu(\mathbf{r})^2\,\mathrm{d}\mathbf{r} = 1\right\}_{\mu=1}^{B} \tag{15}$$

with coefficients $\mathbf{C} \in \mathbb{R}^{B\times B}$ such that $\varphi_i(\mathbf{r}) = \sum_{\mu=1}^{B} C_{\mu i}\chi_\mu(\mathbf{r})$ (Lehtola et al., 2020a). Inserting this into Eq. (14) yields

$$\rho(\mathbf{r}) = 2\sum_{i}^{O}\sum_{\mu\nu}^{B} C_{\mu i}C_{\nu i}\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r}) = 2\sum_{\mu\nu}^{B}\sum_{i}^{O} C_{\mu i}C_{\nu i}\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r}) =: 2\sum_{\mu\nu}^{B} P_{\mu\nu}\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r}), \tag{16}$$

which defines the **density matrix** $\mathbf{P} \in \mathbb{R}^{B\times B}$. For molecules, basis functions are commonly constructed from atom-centered Gaussians aligned with the Cartesian axes, whereas periodic chemistry codes often use plane waves. This allows for an efficient computation of many integrals involved in the computation, for example, the integral over the density in Eq. (14) becomes

$$N_{\text{elec}} = \int_{\mathbb{R}^3} \rho(\mathbf{r})\,\mathrm{d}\mathbf{r} = 2\sum_{\mu\nu}^{B} P_{\mu\nu}\int_{\mathbb{R}^3} \chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\,\mathrm{d}\mathbf{r} =: 2\sum_{\mu\nu}^{B} P_{\mu\nu}S_{\mu\nu} \tag{17}$$

with the overlap matrix $\mathbf{S} \in \mathbb{R}^{B\times B}$. Notably, this integral tensor $S_{\mu\nu}$ is a constant for a given basis set such that it only needs to be computed once per molecule. This is especially convenient since it can be shown that in finite basis sets Eq. (13) is equivalent to solving Eq. (1). Notably, the overlap matrix $\mathbf{S}$ is a **positive-definite Gram matrix** of the atomic orbital (AO) basis functions, defined as

$$S_{\mu\nu} := \int_{\mathbb{R}^3} \chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{18}$$

The inner product is symmetric, so $S_{\mu\nu} = S_{\nu\mu}$ and $\mathbf{S} = \mathbf{S}^\mathsf{T}$. For any nonzero $\mathbf{x} \in \mathbb{R}^B$, $\mathbf{x}^\mathsf{T}\mathbf{S}\mathbf{x} = \sum_{\mu,\nu} x_\mu x_\nu S_{\mu\nu} = \left\|\sum_\mu x_\mu\chi_\mu\right\|^2$. Linear independence of the AO basis implies $\sum_\mu x_\mu\chi_\mu \neq 0$ for all $\mathbf{x} \neq 0$, and therefore $\mathbf{x}^\mathsf{T}\mathbf{S}\mathbf{x} > 0$. The overlap matrix is thus symmetric positive definite, $\mathbf{S} \succ 0$. Hence $\mathbf{S}$ is invertible. Normalization of the individual AO basis functions implies $S_{\mu\mu} = 1$, and therefore $\mathrm{tr}(\mathbf{S}) = \sum_{\mu=1}^{B} S_{\mu\mu} = B$. The B eigenvalues of $\mathbf{S}$ are strictly positive and sum to B s.t.

$$0 < \lambda_{\min}(\mathbf{S}) \leq 1 \leq \lambda_{\max}(\mathbf{S}) \leq B, \tag{19}$$

such that **if any eigenvalue differs from 1, the remaining eigenvalues compensate with the opposite total deviation.**

B ROTATIONAL GROOT: TRANSPORT ON THE GRASSMANNIAN ↑toc

This appendix gives the full construction of $\text{GROOT}_{\text{rot}}$: why an occupied–virtual rotation suffices to reach any target density, how the generators of such rotations are characterized, and how an unconstrained decoder output is mapped into them.

This section proves that GROOT can connect any two physical densities with the same number of occupied orbitals. The argument has two logically separate parts. First, we study the geometry: does there exist an occupied–virtual rotation that transports the initial density to the target, does its generator follow a Grassmann geodesic, and how does it relate to nonminimal rotations that produce the same density? Second, we study the readout: does every model output produce such a valid generator, and can the readout reach every generator required by the geometry?

We answer these questions in four propositions. The first two concern only rotations on the Grassmannian and are independent of the neural network. The third proves that the GROOT head is valid for every input matrix. The fourth proves surjectivity, which is the statement that the head itself introduces no expressivity bottleneck.

$\text{GROOT}_{\text{rot}}$ is a one-step flow map (Boffi et al., 2025), which transports a computationally cheap initial density $\mathbf{P}_0 \in \mathcal{M}_{(N_{\text{elec}},B)}$ to a final density $\widehat{\mathbf{P}}$.[1] Rotating with an orthogonal matrix $\mathbf{U} \in SO(B)$ preserves these Grassmannian constraints and is therefore the transformation of choice

$$\widehat{\mathbf{P}} := \mathbf{U}\mathbf{P}_0\mathbf{U}^\mathsf{T}, \qquad \mathrm{tr}(\widehat{\mathbf{P}}) = \mathrm{tr}(\mathbf{P}_0\mathbf{U}^\mathsf{T}\mathbf{U}) = \mathrm{tr}(\mathbf{P}_0), \qquad \widehat{\mathbf{P}}^2 = \mathbf{U}\mathbf{P}_0^2\mathbf{U}^\mathsf{T} = \widehat{\mathbf{P}}. \tag{20}$$

After dropping all orbital rotations that leave $\mathbf{P}$ unchanged, cf. Fig. 5 and Eq. (5), any two valid densities are connected by a matrix $\mathbf{U} \in SO(B)$ that mixes the occupied and virtual subspaces. We formalize this by defining the projector into the virtual subspace $\mathbf{\Pi} := \mathbf{I} - \mathbf{P}$ and the reflection matrix $\mathbf{R} := \mathbf{P} - \mathbf{\Pi}$. When evaluated at a specific density matrix, such as $\mathbf{P}_0$, we denote the corresponding operators with a subscript ($\mathbf{\Pi}_0, \mathbf{R}_0$). The reflection $\mathbf{R}$ flips the virtual orbitals around the origin while not affecting the occupied orbitals, i.e. $\mathbf{R}\mathbf{C}_\text{V} = -\mathbf{C}_\text{V}$, $\mathbf{R}\mathbf{C}_\text{O} = \mathbf{C}_\text{O}$. Rotations that commute with the reflection, $\mathbf{R}\mathbf{U} = \mathbf{U}\mathbf{R}$, are those that are block-diagonal in the molecular orbitals, mixing occupied with occupied and virtual with virtual, and thus span the subspace of redundant orbital-rotation directions that leave $\mathbf{P}$ unchanged, as shown in Eq. (5). Complementarily, there also exists a canonical purely mixing rotation between the occupied and the virtual orbitals $\mathbf{U} \equiv \exp(\mathbf{X}) \in SO(B)$ generated by a skew-symmetric matrix $\mathbf{X} \in \mathfrak{p}$ that *anti*commutes with the reflection,

$$\mathfrak{p} \equiv \left\{ \mathbf{X} \in \mathbb{R}^{B\times B} \;\middle|\; \mathbf{X}^\mathsf{T} = -\mathbf{X},\ \ \mathbf{X}\mathbf{R} = -\mathbf{R}\mathbf{X} \right\}, \quad \text{with} \quad \mathbf{R} = 2\mathbf{P} - \mathbf{I}. \tag{21}$$

Throughout this section all matrices are represented in the orthonormal basis (Eq. 3). Let $\widetilde{\mathbf{P}}_0$ and $\widetilde{\mathbf{P}}_1$ be physical densities with the same rank, and define

$$\widetilde{\mathbf{\Pi}}_a = \mathbf{I} - \widetilde{\mathbf{P}}_a, \qquad \widetilde{\mathbf{R}}_a = \widetilde{\mathbf{P}}_a - \widetilde{\mathbf{\Pi}}_a, \qquad a \in \{0, 1\}. \tag{22}$$

Each reflection is symmetric and involutory,

$$\left(\widetilde{\mathbf{R}}_a\right)^\mathsf{T} = \widetilde{\mathbf{R}}_a, \qquad \left(\widetilde{\mathbf{R}}_a\right)^2 = \mathbf{I}, \tag{23}$$

and acts as $+1$ on the occupied subspace and as -1 on the virtual subspace. Wong (1967) showed that these pure mixing rotations suffice to connect any two densities on $\mathcal{M}_{(N_{\text{elec}},B)}$.

Proposition 1 (Reachability by an occupied–virtual rotation)**.** *Let* $\mathbf{P}_0$ *and* $\mathbf{P}_1$ *be physical densities with the same rank in the orthonormal basis,* $\mathbf{P}_0, \mathbf{P}_1 \in \mathcal{M}_{(N_{\text{elec}},B)}$*. Then there exists a generator* $\mathbf{X} \in \mathfrak{p}$ *such that* $\mathbf{P}_1 = \exp(\mathbf{X})\,\mathbf{P}_0 \exp(-\mathbf{X})$*. Consequently,* $\mathbf{U} := \exp(\mathbf{X}) \in \mathrm{SO}(B)$*, and* $\mathbf{U}$ *transports the occupied subspace of* $\mathbf{P}_0$ *to that of* $\mathbf{P}_1$*.*

Proof. If $O > V$, consider instead the complementary projectors $\widetilde{\mathbf{\Pi}}_a = \mathbf{I} - \widetilde{\mathbf{P}}_a$, which have rank V and codimension O. Their reflections are $-\widetilde{\mathbf{R}}_a$, so they define the same anticommuting generator space $\mathfrak{p}$. Moreover, a rotation transports $\widetilde{\mathbf{\Pi}}_0$ to $\widetilde{\mathbf{\Pi}}_1$ if and only if it transports $\widetilde{\mathbf{P}}_0$ to $\widetilde{\mathbf{P}}_1$. It therefore suffices to prove the case $O \leq V$.

[1] All matrices in this appendix are in the orthonormal basis introduced in Eq. (3). We omit tildes to avoid cluttering the equations.

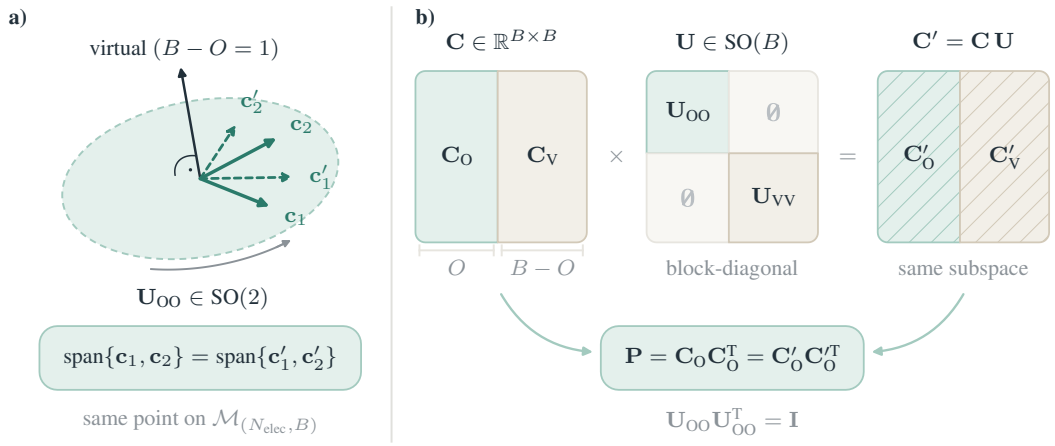


Figure 5: **Many coefficient matrices, one density. (a)** The smallest case, $B = 3$ and $O = 2$: the occupied orbitals c_1, c_2 span a plane which determines the density. Rotating the arrows within the plane by $\mathbf{U}_{OO}$ changes the orbitals, but not the plane. **(b)** For arbitrary $O, B \in \mathbb{N}$ with $O < B$. The coefficients $\mathbf{C} \in \mathbb{R}^{B \times B}$ split into an occupied block $\mathbf{C}_O \in \mathbb{R}^{B \times O}$ and a virtual block $\mathbf{C}_V \in \mathbb{R}^{B \times V}$. Only the occupied block determines the density matrix as seen in Eq. (2). Multiplication from the right by a block-diagonal orthogonal matrix $\mathbf{U}$ does not change the density matrix, see Eq. (5).

Since the physical densities in the orthonormal basis are themselves orthogonal projectors, we can always find $\widetilde{\mathbf{O}}_0, \widetilde{\mathbf{O}}_1 \in O(B)$ such that $\widetilde{\mathbf{P}}_a = \widetilde{\mathbf{O}}_a \widetilde{\mathbf{P}}_{\mathrm{ref}} \widetilde{\mathbf{O}}_a^{\mathsf{T}}$, with $a \in \{0, 1\}$ and $\widetilde{\mathbf{P}}_{\mathrm{ref}} = \mathrm{diag}(\mathbf{I}_O, \mathbf{0}_V)$. Express the second MO frame in the first through

$$\widetilde{\mathbf{Q}} = \widetilde{\mathbf{O}}_0^{\mathsf{T}} \widetilde{\mathbf{O}}_1, \qquad \widetilde{\mathbf{O}}_1 = \widetilde{\mathbf{O}}_0 \widetilde{\mathbf{Q}}. \tag{24}$$

Consequently, in the $\widetilde{\mathbf{O}}_0$-MO basis,

$$\widetilde{\mathbf{O}}_0^{\mathsf{T}} \widetilde{\mathbf{P}}_0 \widetilde{\mathbf{O}}_0 = \widetilde{\mathbf{P}}_{\mathrm{ref}}, \qquad \widetilde{\mathbf{O}}_0^{\mathsf{T}} \widetilde{\mathbf{P}}_1 \widetilde{\mathbf{O}}_0 = \widetilde{\mathbf{Q}} \widetilde{\mathbf{P}}_{\mathrm{ref}} \widetilde{\mathbf{Q}}^{\mathsf{T}}. \tag{25}$$

The general cosine-sine (CS) decomposition states that for every orthogonal $\widetilde{\mathbf{D}} \in \mathrm{SO}(B)$, there exist block-diagonal orthogonal matrices $\widetilde{\mathbf{L}} = \mathrm{diag}(\widetilde{\mathbf{L}}_{\mathrm{o}}, \widetilde{\mathbf{L}}_{\mathrm{v}})$ and $\widetilde{\mathbf{S}} = \mathrm{diag}(\widetilde{\mathbf{S}}_{\mathrm{o}}, \widetilde{\mathbf{S}}_{\mathrm{v}})$, as well as principal angles $0 \le \theta_i \le \frac{\pi}{2}$, such that $\widetilde{\mathbf{D}} = \widetilde{\mathbf{L}} \widetilde{\mathbf{G}} \widetilde{\mathbf{S}}^{\mathsf{T}}$, where, in the $O + O + (V - O)$ partition,

$$\widetilde{\mathbf{G}} = \begin{bmatrix} \cos \mathbf{\Theta} & \sin \mathbf{\Theta} & \mathbf{0} \\ -\sin \mathbf{\Theta} & \cos \mathbf{\Theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{V-O}, \end{bmatrix} \qquad \text{with} \qquad \mathbf{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_O). \tag{26}$$

We proceed to apply the general CS decomposition to $\widetilde{\mathbf{Q}}$. Because $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{S}}$ preserve the occupied–virtual splitting, they are block-diagonal orbital rotations that leave $\widetilde{\mathbf{P}}_{\mathrm{ref}}$ unchanged and can be freely exchanged for one another in

$$\widetilde{\mathbf{L}}^{\mathsf{T}} \widetilde{\mathbf{P}}_{\mathrm{ref}} \widetilde{\mathbf{L}} = \widetilde{\mathbf{P}}_{\mathrm{ref}} = \widetilde{\mathbf{S}}^{\mathsf{T}} \widetilde{\mathbf{P}}_{\mathrm{ref}} \widetilde{\mathbf{S}}. \tag{27}$$

Still working in the $\widetilde{\mathbf{O}}_0$-MO basis, we obtain

$$\widetilde{\mathbf{O}}_0^{\mathsf{T}} \widetilde{\mathbf{P}}_1 \widetilde{\mathbf{O}}_0 = (\widetilde{\mathbf{L}} \widetilde{\mathbf{G}} \widetilde{\mathbf{L}}^{\mathsf{T}}) \widetilde{\mathbf{P}}_{\mathrm{ref}} (\widetilde{\mathbf{L}} \widetilde{\mathbf{G}}^{\mathsf{T}} \widetilde{\mathbf{L}}^{\mathsf{T}}). \tag{28}$$

Returning to the orthonormal basis, we absorb the block-diagonal factor into the MO frame by defining $\widetilde{\mathbf{O}}_0' \equiv \widetilde{\mathbf{O}}_0 \widetilde{\mathbf{L}}$. This gives

$$\widetilde{\mathbf{P}}_0 = \widetilde{\mathbf{O}}_0' \widetilde{\mathbf{P}}_{\mathrm{ref}} (\widetilde{\mathbf{O}}_0')^{\mathsf{T}}, \quad \widetilde{\mathbf{P}}_1 = \big(\widetilde{\mathbf{O}}_0' \widetilde{\mathbf{G}} (\widetilde{\mathbf{O}}_0')^{\mathsf{T}}\big) \widetilde{\mathbf{P}}_0 \big(\widetilde{\mathbf{O}}_0' \widetilde{\mathbf{G}}^{\mathsf{T}} (\widetilde{\mathbf{O}}_0')^{\mathsf{T}}\big). \tag{29}$$

As with 2D rotations, we express the rotation in the MO basis as the matrix exponential $\widetilde{\mathbf{G}} = \exp(\widetilde{\mathbf{K}})$ of the skew-symmetric matrix

$$\widetilde{\mathbf{K}} = \begin{bmatrix} \mathbf{0} & \mathbf{\Theta} & \mathbf{0} \\ -\mathbf{\Theta} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{30}$$

which, in the orthonormal basis, yields the generator $\mathbf{X} = \widetilde{\mathbf{O}}_0' \widetilde{\mathbf{K}} (\widetilde{\mathbf{O}}_0')^\mathsf{T}$ such that the principal rotation is

$$\exp(\mathbf{X}) \widetilde{\mathbf{P}}_0 \exp(-\mathbf{X}) = \widetilde{\mathbf{P}}_1. \tag{31}$$

It remains to verify the second defining anticommutation condition of $\mathfrak{p}$. In the MO basis, the reflection operator similarly becomes diagonal and takes the form $\widetilde{\mathbf{R}}_{\text{ref}} = \text{diag}(\mathbf{I}_O, -\mathbf{I}_O, -\mathbf{I}_{V-O})$. Hence, the block forms of $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{R}}_{\text{ref}}$ imply $\widetilde{\mathbf{K}} \widetilde{\mathbf{R}}_{\text{ref}} + \widetilde{\mathbf{R}}_{\text{ref}} \widetilde{\mathbf{K}} = \mathbf{0}$. Since commutation relations are basis-independent, this statement carries over into the orthonormal basis such that

$$\mathbf{X} \widetilde{\mathbf{R}}_0 + \widetilde{\mathbf{R}}_0 \mathbf{X} = \mathbf{0}. \tag{32}$$

Finally, since $\mathbf{X}$ is real and skew-symmetric, $\mathbf{U} = \exp(\mathbf{X}) \in \text{SO}(B)$, which proves the claim. □

For the same density pair, any other transporting rotation differs from a pure mixing rotation only by a block-diagonal orbital rotation that leaves the density unchanged (Edelman et al., 1998).

Proposition 2 (Decomposition of a general transporting rotation)**.** *Let* $\mathbf{W} \in \text{SO}(B)$ *satisfy* $\mathbf{P}_1 = \mathbf{W} \mathbf{P}_0 \mathbf{W}^\mathsf{T}$. *Then* $\mathbf{W} = \exp(\mathbf{X})\, \mathbf{V}$ *with* $\mathbf{X} \in \mathfrak{p}$ *and a block-diagonal orbital rotation* $\mathbf{V} \in \text{SO}(B)$ *that commutes with the reflection,* $\mathbf{V} \mathbf{R}_0 = \mathbf{R}_0 \mathbf{V}$.

Proof. Define $\mathbf{V} = \mathbf{U}^\mathsf{T} \widetilde{\mathbf{W}}$. It lies in $\text{SO}(B)$ and gives $\widetilde{\mathbf{W}} = \mathbf{U}\mathbf{V}$. Both rotations transport the initial reflection to the target reflection, so

$$\mathbf{V} \widetilde{\mathbf{R}}_0 \mathbf{V}^\mathsf{T} = \mathbf{U}^\mathsf{T} \widetilde{\mathbf{R}}_1 \mathbf{U} = \widetilde{\mathbf{R}}_0. \tag{33}$$

Since $\mathbf{V}$ is orthogonal, this is equivalent to $\mathbf{V} \widetilde{\mathbf{R}}_0 = \widetilde{\mathbf{R}}_0 \mathbf{V}$. Thus $\mathbf{V}$ is a block-diagonal orbital rotation that leaves the density unchanged. □

Rather than direct density prediction, GROOT$_{\text{rot}}$ builds on equivariant matrix decoders (Yu et al., 2023b; Febrer et al., 2025) to construct only an occupied-virtual rotation that moves a computationally cheap density toward the target density. For every decoder output $\hat{\mathbf{M}} \in \mathbb{R}^{B \times B}$, we construct

$$\hat{\mathbf{X}}_{\text{GROOT}} \left(\hat{\mathbf{M}} \right) = \mathbf{P}_0 \hat{\mathbf{M}} \mathbf{\Pi}_0 - \mathbf{\Pi}_0 \hat{\mathbf{M}}^\mathsf{T} \mathbf{P}_0 \tag{34}$$

which lies in $\mathfrak{p}$, as shown in Appendix B, Proposition 3. The GROOT$_{\text{rot}}$ map $\hat{\mathbf{M}} \mapsto \hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}})$ is surjective onto $\mathfrak{p}$ because every valid $\mathbf{X} \in \mathfrak{p}$ is a fixed point, i.e. choosing $\hat{\mathbf{M}} = \mathbf{X}$ yields $\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}}) = \mathbf{X}$, as shown in Appendix B, Proposition 4.

Proposition 3 (Validity of the GROOT map)**.**
For an arbitrary matrix $\hat{\mathbf{M}} \in \mathbb{R}^{B \times B}$, *the generator* $\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}})$ *of Eq. (34) lies in* $\mathfrak{p}$ *for every input.*

Proof. Let $\widetilde{\mathbf{A}} = \widetilde{\mathbf{P}}_0 \hat{\mathbf{M}} \widetilde{\mathbf{\Pi}}_0$. Since the projectors are symmetric, $\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}}) = \widetilde{\mathbf{A}} - \widetilde{\mathbf{A}}^\mathsf{T}$ is skew-symmetric. Moreover, since $\widetilde{\mathbf{R}}_0 \widetilde{\mathbf{P}}_0 = \widetilde{\mathbf{P}}_0$ and $\widetilde{\mathbf{R}}_0 \widetilde{\mathbf{\Pi}}_0 = -\widetilde{\mathbf{\Pi}}_0$, together with the corresponding right-multiplication identities, we have $\widetilde{\mathbf{R}}_0 \widetilde{\mathbf{A}} \widetilde{\mathbf{R}}_0 = -\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{R}}_0 \widetilde{\mathbf{A}}^\mathsf{T} \widetilde{\mathbf{R}}_0 = -\widetilde{\mathbf{A}}^\mathsf{T}$. Therefore,

$$\widetilde{\mathbf{R}}_0 \hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}}) \widetilde{\mathbf{R}}_0 = -\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}}). \tag{35}$$

Since $\widetilde{\mathbf{R}}_0^2 = \mathbf{I}$, this is equivalent to the required anticommutation relation. Thus $\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}})$ lies in $\mathfrak{p}$. □

Proposition 4 (Surjectivity of the GROOT map)**.**
Every generator $\mathbf{X} \in \mathfrak{p}$ *can be reached by the GROOT map, since* $\hat{\mathbf{X}}_{\text{GROOT}}(\mathbf{X}) = \mathbf{X}$.

Proof. Let $\mathbf{X} \in \mathfrak{p}$. Choose $\hat{\mathbf{M}} = \mathbf{X}$. Since $\mathbf{X}^\mathsf{T} = -\mathbf{X}$, the readout gives

$$\hat{\mathbf{X}}_{\text{GROOT}}(\mathbf{X}) = \widetilde{\mathbf{P}}_0 \mathbf{X} \widetilde{\mathbf{\Pi}}_0 + \widetilde{\mathbf{\Pi}}_0 \mathbf{X} \widetilde{\mathbf{P}}_0 = \mathbf{X}. \tag{36}$$

Thus every valid generator has a preimage, and the map is surjective. □

The first two propositions show that any target occupied subspace is geometrically reachable and that non-minimal transporting rotations differ only by block-diagonal orbital rotations that leave the endpoint densities unchanged. Propositions 3 and 4 show that the GROOT map reaches the entire space of minimal geodesic generators while producing a valid generator for every possible neural-network output.

The GROOT$_{\text{rot}}$ readout must also preserve the $SO(3)$-equivariance of the underlying matrix decoder. A spatial rotation induces a corresponding $\mathbf{U} \in SO(B)$ in the orthonormal basis, transforming both $\mathbf{P}_0$ and $\hat{\mathbf{M}}$ as $\mathbf{Y} \to \mathbf{U}\mathbf{Y}\mathbf{U}^\mathsf{T}$, so Eq. (34) gives $\hat{\mathbf{X}}_{\text{GROOT}} \to \mathbf{U}\hat{\mathbf{X}}_{\text{GROOT}}\mathbf{U}^\mathsf{T}$. Hence, GROOT$_{\text{rot}}$'s covariant generator can learn to produce a rotation to any valid target density $\hat{\mathbf{P}} \approx \mathbf{P}^*$ from $\mathbf{P}_0$ (Proposition 1)

$$\hat{\mathbf{P}} = \exp\left(\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}})\right) \mathbf{P}_0 \exp\left(-\hat{\mathbf{X}}_{\text{GROOT}}(\hat{\mathbf{M}})\right). \tag{37}$$

C Observable Bounds for the Orthonormal Loss

We derive the observable-error bounds summarized in table 1 at a fixed molecular geometry. Density errors give linear bounds for one-body observables, including the dipole and real-space density. For valid density projectors, stationarity makes the energy error quadratic. Hamiltonian predictions inherit these bounds through their occupied eigenspaces, with an additional dependence on the HOMO–LUMO gap.

We use real matrices and the closed-shell, spin-free convention $\widetilde{\mathbf{P}}^* = \widetilde{\mathbf{C}}_\text{o}^* \widetilde{\mathbf{C}}_\text{o}^{*\mathsf{T}}$, where $\text{tr}(\widetilde{\mathbf{P}}^*) = O$ and O is the number of occupied orbitals. The physical electron density includes an occupation factor of two. We write $\mathbf{D} = \hat{\widetilde{\mathbf{P}}} - \widetilde{\mathbf{P}}^*$ for the density error and assume that the overlap matrix $\mathbf{S}$ is positive definite. For the energy bounds, we assume that E_m is three times continuously differentiable near the reference, with locally bounded third derivative, and stationary there under rotations between occupied and virtual orbitals. All asymptotic statements concern small errors for a fixed sample m. The bounds use unnormalized Frobenius norms, so substitution of $\|\Delta\widetilde{\mathbf{X}}\|_F = B\mathcal{L}_{\text{ON}}$ gives the corresponding bounds for the loss in Eq. (7).

C.1 AO and orthonormal Frobenius errors

The overlap eigenvalues determine how the AO and orthonormal norms weight each error direction. Write

$$\mathbf{S} = \mathbf{U}\,\text{diag}(\lambda_1, \ldots, \lambda_B)\mathbf{U}^\mathsf{T}, \qquad 0 < \lambda_{\min} \le \lambda_i \le \lambda_{\max}. \tag{38}$$

In this eigenbasis, let $\Delta\bar{\mathbf{X}} = \mathbf{U}^\mathsf{T}\Delta\mathbf{X}\mathbf{U}$. The density transformation in Eq. (3) gives $\|\Delta\widetilde{\mathbf{P}}\|_F^2 = \sum_{i,j} \lambda_i\lambda_j |\Delta\bar{P}_{ij}|^2$, whereas the Hamiltonian transformation gives $\|\Delta\widetilde{\mathbf{H}}\|_F^2 = \sum_{i,j} |\Delta\bar{H}_{ij}|^2/(\lambda_i\lambda_j)$. Bounding each weight by its extreme values yields

$$\lambda_{\min}\|\Delta\mathbf{P}\|_F \le \|\Delta\widetilde{\mathbf{P}}\|_F \le \lambda_{\max}\|\Delta\mathbf{P}\|_F, \tag{39}$$

$$\lambda_{\max}^{-1}\|\Delta\mathbf{H}\|_F \le \|\Delta\widetilde{\mathbf{H}}\|_F \le \lambda_{\min}^{-1}\|\Delta\mathbf{H}\|_F. \tag{40}$$

Errors of the form $\mathbf{u}_i\mathbf{u}_i^\mathsf{T}$ attain these endpoints when $\mathbf{u}_i$ is an extreme overlap eigenvector. The constants are therefore sharp for unrestricted symmetric errors. Restrictions to projector tangent directions or occupied–virtual Hamiltonian blocks can lower the constants for a particular sample.

Relative to the orthonormal norm, the plain AO density norm gives more weight to directions with small overlap eigenvalues and less weight to directions with large eigenvalues. For Hamiltonian errors, the weighting reverses. Equations (39) and (40) supply the worst-case AO conversion factors in table 1. They do not establish that the orthonormal bound is tighter for every observable and error direction.

C.2 Choice of orthonormal basis

Every full-rank orthonormalization gives the same Frobenius loss. To see this, take any square matrix $\mathbf{W}$ satisfying $\mathbf{W}^\mathsf{T}\mathbf{S}\mathbf{W} = \mathbf{I}$ and set $\mathbf{Q} = \mathbf{S}^{1/2}\mathbf{W}$. Then $\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}$ and $\mathbf{W} = \mathbf{S}^{-1/2}\mathbf{Q}$. The resulting

Hamiltonian and density representations are $\mathbf{Q}^\mathsf{T}\widetilde{\mathbf{H}}\mathbf{Q}$ and $\mathbf{Q}^\mathsf{T}\widetilde{\mathbf{P}}\mathbf{Q}$, respectively. Orthogonal invariance gives $\|\mathbf{Q}^\mathsf{T}\Delta\widetilde{\mathbf{X}}\mathbf{Q}\|_F = \|\Delta\widetilde{\mathbf{X}}\|_F$ for either matrix. The symmetric Löwdin choice additionally minimizes the sum of squared distances between the original basis functions and their orthonormal counterparts (Löwdin, 1950; Carlson & Keller, 1957).

Observable-specific bounds also depend on the observable operator. For example, an AO one-body operator $\mathbf{M} = \mathbf{S}^{1/2}\widetilde{\mathbf{M}}\mathbf{S}^{1/2}$ satisfies $|\operatorname{tr}(\mathbf{M}\Delta\mathbf{P})| \le \|\mathbf{M}\|_F\|\Delta\mathbf{P}\|_F$, with

$$\lambda_{\min}\|\widetilde{\mathbf{M}}\|_F \le \|\mathbf{M}\|_F \le \lambda_{\max}\|\widetilde{\mathbf{M}}\|_F.$$

Thus, a direct AO bound can have a smaller or larger constant than its orthonormal counterpart.

C.3 General density errors ↑toc

An unconstrained density prediction can incur a first-order energy error even at a stationary reference. Under our occupation convention, $\nabla E_m[\widetilde{\mathbf{P}}^*] = 2\widetilde{\mathbf{H}}^*$, so Taylor expansion gives

$$E_m[\widetilde{\mathbf{P}}^* + \mathbf{D}] - E_m[\widetilde{\mathbf{P}}^*] = 2\langle\widetilde{\mathbf{H}}^*, \mathbf{D}\rangle_F + \mathcal{O}(\|\mathbf{D}\|_F^2).$$

Cauchy–Schwarz bounds the magnitude of the linear term by $2\|\widetilde{\mathbf{H}}^*\|_F\|\mathbf{D}\|_F$. Stationarity makes $\widetilde{\mathbf{H}}^*$ block diagonal in the occupied–virtual basis, so only the occupied–occupied and virtual–virtual blocks of $\mathbf{D}$ contribute to this term. An unconstrained predictor can change these blocks at first order through occupation or charge errors.

One-body observables depend linearly on the density matrix. For an observable with orthonormal-basis matrix $\widetilde{\mathbf{M}}$, including its fixed charge and occupation factors, Cauchy–Schwarz gives

$$|\Delta M| = |\operatorname{tr}(\widetilde{\mathbf{M}}\mathbf{D})| \le \|\widetilde{\mathbf{M}}\|_F\|\mathbf{D}\|_F.$$

For the dipole, we apply this inequality to each Cartesian component, square the three inequalities, and add them. With $C_{\mu,m}^2 = \sum_{\alpha\in\{x,y,z\}}\|\widetilde{\boldsymbol{\mu}}_\alpha\|_F^2$, the result is $\|\Delta\boldsymbol{\mu}\|_2 \le C_{\mu,m}\|\mathbf{D}\|_F$. This proves the dipole row of table 1. The nuclear contribution cancels because we compare predictions at the same geometry.

The real-space electron density is another linear function of the density matrix. For real orthonormal basis functions $\widetilde{\chi}_\mu$, define $\Phi_{\mu\nu}(\mathbf{r}) = 2\widetilde{\chi}_\mu(\mathbf{r})\widetilde{\chi}_\nu(\mathbf{r})$ and $\Delta\rho(\mathbf{r}) = \sum_{\mu,\nu} D_{\mu\nu}\Phi_{\mu\nu}(\mathbf{r})$. Its L_1 norm measures the integrated absolute density error and satisfies

$$\|\Delta\rho\|_{L_1} \le 2\sum_{j=1}^{B}|d_j| \le 2\sqrt{B}\|\mathbf{D}\|_F,$$

where d_j are the eigenvalues of $\mathbf{D}$. To obtain the first inequality, diagonalize $\mathbf{D}$ and rotate the basis functions by the same orthogonal matrix. Then $\Delta\rho = 2\sum_j d_j\psi_j^2$ with $\int\psi_j^2 = 1$, so the triangle inequality applies term by term. The second inequality is Cauchy–Schwarz. Taking $\mathbf{D} = \mathbf{I}$ attains both inequalities, so the induced Frobenius-to-L_1 constant for unrestricted symmetric errors is $C_{1,m} = 2\sqrt{B}$.

The L_2 density error depends on the spatial overlap of the basis-function products. Assuming these products are square-integrable, define $\Gamma_{(\mu\nu),(\kappa\ell)} = \int\Phi_{\mu\nu}(\mathbf{r})\Phi_{\kappa\ell}(\mathbf{r})\,\mathrm{d}\mathbf{r}$. Expanding the squared density error gives

$$\|\Delta\rho\|_{L_2}^2 = \operatorname{vec}(\mathbf{D})^\mathsf{T}\boldsymbol{\Gamma}\operatorname{vec}(\mathbf{D}) \le \lambda_{\max}(\boldsymbol{\Gamma})\|\mathbf{D}\|_F^2.$$

Thus $\|\Delta\rho\|_{L_2} \le C_{2,m}\|\mathbf{D}\|_F$, where $C_{2,m} = \sqrt{\lambda_{\max}(\boldsymbol{\Gamma})}$. The L_2 row in table 1 refers to the norm itself, whose bound is linear in $\|\mathbf{D}\|_F$. Its square has a quadratic bound. Substituting (39) into these observable bounds supplies one factor of $\lambda_{\max}$ for each power of the density error.

C.4 Grassmann-constrained density errors ↑toc

For a valid rank-O projector, the leading density error only mixes occupied and virtual orbitals. This is the constraint that removes the first-order energy term. In the occupied–virtual basis, write

$\widetilde{\mathbf{P}}^* = \mathrm{diag}(\mathbf{I}_O, \mathbf{0})$ and parameterize a nearby projector by

$$\mathbf{K}(\boldsymbol{X}) = \begin{pmatrix} \mathbf{0} & \boldsymbol{X}^\mathsf{T} \\ -\boldsymbol{X} & \mathbf{0} \end{pmatrix}, \qquad \widetilde{\mathbf{P}}(\boldsymbol{X}) = e^{\mathbf{K}(\boldsymbol{X})} \widetilde{\mathbf{P}}^* e^{-\mathbf{K}(\boldsymbol{X})}.$$

Here $\boldsymbol{X} \in \mathbb{R}^{(B-O)\times O}$ contains the occupied–virtual rotation parameters. Because $\mathbf{K}$ is skew-symmetric, this construction preserves rank and idempotency. Expansion around $\boldsymbol{X} = \mathbf{0}$ gives

$$\mathbf{D} = \begin{pmatrix} \mathbf{0} & -\boldsymbol{X}^\mathsf{T} \\ -\boldsymbol{X} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}\boldsymbol{X}^\mathsf{T} \end{pmatrix} + \mathcal{O}(\|\boldsymbol{X}\|_F^3). \tag{41}$$

The diagonal blocks begin at second order. The squared error satisfies $\|\mathbf{D}\|_F^2 = 2\|\boldsymbol{X}\|_F^2 + \mathcal{O}(\|\boldsymbol{X}\|_F^4)$.

To make the energy constant precise, let $\mathbf{x} = \mathrm{vec}(\boldsymbol{X})$ and let $\mathbf{A}_m$ be the Hessian of $f_m(\mathbf{x}) = E_m[\widetilde{\mathbf{P}}(\boldsymbol{X})]$ at $\mathbf{x} = \mathbf{0}$. Stationarity gives $\nabla f_m(\mathbf{0}) = \mathbf{0}$, hence $f_m(\mathbf{x}) - f_m(\mathbf{0}) = \mathbf{x}^\mathsf{T}\mathbf{A}_m\mathbf{x}/2 + \mathcal{O}(\|\mathbf{x}\|_2^3)$. Define $h_m = \|\mathbf{A}_m\|_2/2$, where $\|\cdot\|_2$ denotes the spectral norm. The factor of two accounts for the two off-diagonal blocks in the Frobenius norm of a tangent perturbation. We obtain

$$\big|E_m[\hat{\widetilde{\mathbf{P}}}] - E_m[\widetilde{\mathbf{P}}^*]\big| \le \frac{h_m}{2}\|\mathbf{D}\|_F^2 + \mathcal{O}(\|\mathbf{D}\|_F^3). \tag{42}$$

This local bound holds for nearby rank-O projectors, including predictions from both GROOT constructions. The dipole and real-space density bounds remain linear because their maps are exactly linear in $\mathbf{D}$.

The projector error also has an exact geometric interpretation. For principal angles θ_i between the predicted and reference occupied subspaces,

$$\|\mathbf{D}\|_F^2 = 2\sum_{i=1}^{O} \sin^2\theta_i = 2\big(O - \mathrm{tr}(\hat{\widetilde{\mathbf{P}}}\widetilde{\mathbf{P}}^*)\big). \tag{43}$$

Equation (43) connects the Frobenius error to the orbital-projection score without choosing individual occupied orbitals.

C.5 Hamiltonian-induced density errors ↑toc

A Hamiltonian prediction determines a valid density projector through its lowest O eigenvectors. Write $\hat{\widetilde{\mathbf{H}}} = \widetilde{\mathbf{H}}^* + \mathbf{V}$ and let $g = \Delta\epsilon_{\mathrm{HL}} = \epsilon^*_{\mathrm{LUMO}} - \epsilon^*_{\mathrm{HOMO}} > 0$. We assume that $\widetilde{\mathbf{P}}^*$ projects onto the lowest O eigenvectors of $\widetilde{\mathbf{H}}^*$ and that $\|\mathbf{V}\|_2 < g/2$, which keeps the occupied and virtual eigenvalues separated. In the reference eigenbasis, first-order expansion of the projector and eigenvalue equations gives

$$\mathbf{D} = \begin{pmatrix} \mathbf{0} & \boldsymbol{T}^\mathsf{T} \\ \boldsymbol{T} & \mathbf{0} \end{pmatrix} + \mathcal{O}\big(\|\mathbf{V}\|_F^2/g^2\big), \qquad T_{ai} = -\frac{V_{ai}}{\epsilon^*_a - \epsilon^*_i}.$$

Indices i and a denote occupied and virtual orbitals, respectively, and the matrix remainder is in Frobenius norm. The first-order diagonal blocks vanish by idempotency, and the occupied–virtual eigenvalue equation is $(\epsilon^*_a - \epsilon^*_i)T_{ai} + V_{ai} = 0$. Thus

$$\begin{aligned}\|\mathbf{D}\|_F^2 &= 2\sum_{i\in\mathrm{o},\, a\in\mathrm{v}} \frac{|V_{ai}|^2}{(\epsilon^*_a - \epsilon^*_i)^2} + \mathcal{O}\big(\|\mathbf{V}\|_F^3/g^3\big) \\ &\le \frac{\|\mathbf{V}\|_F^2}{g^2} + \mathcal{O}\big(\|\mathbf{V}\|_F^3/g^3\big).\end{aligned}$$

The inequality uses $\epsilon^*_a - \epsilon^*_i \ge g$ and $2\sum_{i,a}|V_{ai}|^2 \le \|\mathbf{V}\|_F^2$ for symmetric $\mathbf{V}$. Equivalently, $\|\mathbf{D}\|_F \le \|\mathbf{V}\|_F/g + \mathcal{O}(\|\mathbf{V}\|_F^2/g^2)$, which gives the inverse-gap factor for the dipole and density bounds.

For the energy rebuilt from this induced density, substitution into (42) gives

$$\big|E_m[\hat{\widetilde{\mathbf{P}}}] - E_m[\widetilde{\mathbf{P}}^*]\big| \le \frac{h_m}{2g^2}\|\mathbf{V}\|_F^2 + \mathcal{O}\big(\|\mathbf{V}\|_F^3/g^3\big). \tag{44}$$

The curvature h_m and the remainder constants depend on the sample and energy functional. Equation (44) gives the inverse-gap-squared factor in table 1 without assuming how these constants vary across molecules. Using (40) adds a factor of $\lambda_{\min}^{-1}$ to linear bounds and $\lambda_{\min}^{-2}$ to quadratic bounds in the AO Hamiltonian error.

For the fixed independent-particle functional $E_{0,m}[\mathbf{P}] = \mathrm{tr}(\widetilde{\mathbf{H}}^*\mathbf{P})$, we can retain each occupied–virtual gap to obtain a sharper energy bound. Substituting $\boldsymbol{X} = -\boldsymbol{T}$ to first order into (41) gives

$$E_{0,m}[\hat{\widetilde{\mathbf{P}}}] - E_{0,m}[\widetilde{\mathbf{P}}^*] = \sum_{i\in\mathrm{o},\, a\in\mathrm{v}} \frac{|V_{ai}|^2}{\epsilon_a^* - \epsilon_i^*} + \mathcal{O}\big(\|\mathbf{V}\|_F^3/g^2\big) \leq \frac{\|\mathbf{V}\|_F^2}{2g} + \mathcal{O}\big(\|\mathbf{V}\|_F^3/g^2\big). \tag{45}$$

For this functional, $h_{0,m} = \max_{i,a}(\epsilon_a^* - \epsilon_i^*)$. The generic curvature bound would have a quadratic coefficient larger by $h_{0,m}/g$. The inverse-gap bound in (45) follows from keeping the individual gaps and does not require $h_{0,m}$ to scale with g. An explicit occupation factor of two doubles the energy and its bound. Evaluating energy directly from the predicted Hamiltonian eigenvalues can instead introduce a linear error through their first-order shifts.

C.6 Rotation covariance and the entrywise term ↑toc

Both AO and orthonormal Frobenius losses are invariant under physical rotations for the orthonormal spherical-harmonic shell convention used here. Such rotations act through an orthogonal matrix $\mathbf{R}$, so $\|\mathbf{R}^\mathsf{T}\Delta\mathbf{X}\mathbf{R}\|_F = \|\Delta\mathbf{X}\|_F$. The symmetric Löwdin transformation also preserves the same rotation law for the matrix representations because

$$(\mathbf{R}^\mathsf{T}\mathbf{S}\mathbf{R})^{\pm 1/2} = \mathbf{R}^\mathsf{T}\mathbf{S}^{\pm 1/2}\mathbf{R}. \tag{46}$$

Order-dependent orthogonalizers, such as Cholesky or Gram–Schmidt, need not satisfy the covariance law in (46), although they give the same invariant Frobenius loss. General nonorthogonal AO coordinate changes can alter the raw AO Frobenius norm.

An entrywise matrix L_1 term does not share this orthogonal invariance. Writing $\|\mathbf{A}\|_{1,1} = \sum_{i,j}|A_{ij}|$ and $\|\mathbf{A}\|_1 = \max_j \sum_i |A_{ij}|$, we obtain

$$\|\Delta\widetilde{\mathbf{P}}\|_{1,1} \leq \|\mathbf{S}^{1/2}\|_1^2\|\Delta\mathbf{P}\|_{1,1}, \qquad \|\Delta\widetilde{\mathbf{H}}\|_{1,1} \leq \|\mathbf{S}^{-1/2}\|_1^2\|\Delta\mathbf{H}\|_{1,1}.$$

These inequalities follow from vectorization and $\|\mathbf{A}\otimes\mathbf{A}\|_1 = \|\mathbf{A}\|_1^2$. The constants depend on the overlap eigenvectors as well as the eigenvalues. The overlap bounds and orthogonal invariance proved above therefore apply to the Frobenius loss in Eq. (7). Adding an entrywise term does not preserve those guarantees for the full loss.

D Directional Sensitivity in the Overlap Spectrum ↑toc

This appendix formalizes the relative sensitivity analysis of Section 5. We formally introduce the sensitivity of a variable as its absolute positive directional derivative. We construct the eigenbasis from the overlap matrix and analytically derive the derivatives of $\mathcal{L}_{\mathrm{NON}}$ and $\mathcal{L}_{\mathrm{ON}}$.

D.1 Sensitivity as One-Sided Directional Derivatives ↑toc

Let $\mathcal{O}$ be an observable evaluated at the converged reference ground state $\mathbf{X}^* \in \{\mathbf{H}^*, \mathbf{P}^*\}$, and let $\mathbf{\Pi} \in \mathbb{R}^{B\times B}$ be a symmetric perturbation direction with $\|\mathbf{\Pi}\|_F = 1$. We measure the sensitivity of $\mathcal{O}$ along $\mathbf{\Pi}$ by the one-sided directional derivative

$$D_{\mathbf{\Pi}}\mathcal{O}(\mathbf{X}^*) := \lim_{\epsilon\to 0^+} \frac{\mathcal{O}(\mathbf{X}^* + \epsilon\mathbf{\Pi}) - \mathcal{O}(\mathbf{X}^*)}{\epsilon}. \tag{47}$$

The one-sided limit is required because the quantities we compare are not all differentiable at the reference. The L^2 real-space density error and the Frobenius matrix norms entering both losses vanish at $\mathbf{X}^*$ and grow linearly in ϵ in either direction, so their two-sided derivatives do not exist

while Eq. (47) is well defined and positive. The *sensitivity* of $\mathcal{O}$ along $\mathbf{\Pi}$ and the *relative directional sensitivity* of a loss $\mathcal{L}$ with respect to $\mathcal{O}$ are then

$$\mathcal{S}_{\mathbf{\Pi}}(\mathcal{O}) := |D_{\mathbf{\Pi}}\mathcal{O}(\mathbf{X}^*)|\,, \qquad \mathcal{S}_{\mathrm{rel},\mathbf{\Pi}}(\mathcal{L},\mathcal{O}) := \frac{\mathcal{S}_{\mathbf{\Pi}}(\mathcal{L})}{\mathcal{S}_{\mathbf{\Pi}}(\mathcal{O})}. \tag{48}$$

D.2 The overlap-spectrum basis of directions ↑toc

Due to its importance in deriving the Lipschitz continuity relationships between the NON loss and the physical observables we probe the sensitivity in a basis derived from the overlap $\mathbf{S}$. We write the overlap matrix in its eigenbasis,

$$\mathbf{S} = \sum_{k=1}^{B} \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}} = \mathbf{U}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_B)\mathbf{U}^{\mathsf{T}}, \qquad 0 < \lambda_{\min} \le \lambda_k \le \lambda_{\max}, \tag{49}$$

consistent with Eq. (38). Because the matrix-error bounds of Appendix C are mediated by $\mathbf{S}$, we resolve the analysis in the directions built from its eigenvectors,

$$\mathbf{\Pi}_{ij} := \begin{cases} \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}, & i = j, \\ \dfrac{1}{\sqrt{2}}\left(\mathbf{u}_i \mathbf{u}_j^{\mathsf{T}} + \mathbf{u}_j \mathbf{u}_i^{\mathsf{T}}\right), & i < j, \end{cases} \tag{50}$$

where the prefactors normalize each direction to $\|\mathbf{\Pi}_{ij}\|_F = 1$. The $B(B+1)/2$ matrices $\{\mathbf{\Pi}_{ij}\}_{i\le j}$ are mutually orthogonal under the Frobenius trace product $\mathrm{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{B})$ and therefore form a complete orthonormal basis of the space of real symmetric $B \times B$ matrices, which is exactly the space in which a symmetric matrix error $\Delta\mathbf{X}$ lives.

In addition to probing orthonormal perturbations the chosen basis decomposes the total Frobenius matrix error in its components. Since the total Frobenius error is itself invariant under orthonormal basis transformations we write

$$||\Delta\mathbf{X}||_F^2 = \mathrm{tr}(\Delta\mathbf{X}^2) = \mathrm{tr}((\mathbf{U}^{\mathsf{T}}\Delta\mathbf{X}\mathbf{U})^2) = \mathrm{tr}(\Delta\bar{\mathbf{X}}^2) = ||\Delta\bar{\mathbf{X}}||_F^2. \tag{51}$$

The total Frobenius error is then the sum of the squared components in the overlap basis, which we project out using the previously defined projectors,

$$\begin{aligned} ||\Delta\bar{\mathbf{X}}||_F^2 &= \sum_i \left(\Delta\bar{\mathbf{X}}\right)_{ii}^2 + 2\sum_{i<j}\left(\Delta\bar{\mathbf{X}}\right)_{ij}^2 \\ &= \sum_i \left(\mathbf{u}_i^{\mathsf{T}}\Delta\mathbf{X}\,\mathbf{u}_i\right)^2 + 2\sum_{i<j}\left(\mathbf{u}_i^{\mathsf{T}}\Delta\mathbf{X}\,\mathbf{u}_j\right)^2 \\ &= \sum_{i\le j}\mathrm{tr}(\mathbf{\Pi}_{ij}\Delta\mathbf{X})^2 \equiv \sum_{i\le j}\Delta\bar{x}_{ij}^2. \end{aligned} \tag{52}$$

This decomposition leads us to define the binned matrix error along a direction as the root mean square of its elements. Indexing the logarithmically spaced bins in $\sqrt{\lambda_i\lambda_j}$ with b, the binned error is given by

$$\Delta\bar{x}_b \equiv \sqrt{\langle\Delta\bar{x}_{ij}^2\rangle_b} = \left(\frac{1}{N_b}\sum_{(m,i,j)\in b}\left(\Delta\bar{x}_{ij}^{(m)}\right)^2\right)^{1/2}, \quad N_b \equiv \left|\left\{(m,i,j)\,:\,\sqrt{\lambda_i^{(m)}\lambda_j^{(m)}} \in b\right\}\right|. \tag{53}$$

Here the superscript m indexes molecules. Squaring makes the bins additive and connects the binned errors back to total prediction error over the dataset

$$\sum_b N_b\,\Delta\bar{x}_b^2 = \sum_m \left\|\Delta\mathbf{X}^{(m)}\right\|_F^2. \tag{54}$$

After picking the basis $\mathbf{\Pi}_{ij}$ in Eqs. 50, we use the shorthands $D_{ij} \equiv D_{\mathbf{\Pi}_{ij}}$, $\mathcal{S}_{ij} \equiv \mathcal{S}_{\mathbf{\Pi}_{ij}}$ and $\mathcal{S}_{\mathrm{rel},ij} \equiv \mathcal{S}_{\mathrm{rel},\mathbf{\Pi}_{ij}}$.

D.3 Loss Sensitivities ↑toc

The reason Eq. (50) is the natural basis for this analysis is that it diagonalizes the overlap matrix $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ and therefore the transformation of both the non-orthonormal and the orthonormal loss functions $\mathcal{L}_{\text{NON}}$, $\mathcal{L}_{\text{ON}}$. Applying the orthonormalizing basis transformations on the perturbation basis yields

$$\mathbf{S}^{1/2}\mathbf{\Pi}_{ij}\mathbf{S}^{1/2} = \sqrt{\lambda_i\lambda_j}\,\mathbf{\Pi}_{ij}, \qquad \mathbf{S}^{-1/2}\mathbf{\Pi}_{ij}\mathbf{S}^{-1/2} = \frac{1}{\sqrt{\lambda_i\lambda_j}}\,\mathbf{\Pi}_{ij}. \tag{55}$$

Both losses are the normalized Frobenius error $\|\Delta\mathbf{X}\|_F/B$ of Eq. (7), evaluated in the non-orthonormal and the orthonormal basis respectively. Perturbing along a single direction leaves $\Delta\mathbf{X} = \epsilon\mathbf{\Pi}_{ij}$, so Eq. (47) together with Eq. (55) gives

$$\mathcal{S}_{ij}(\mathcal{L}_{\text{NON}}) = \frac{\|\mathbf{\Pi}_{ij}\|_F}{B} = \frac{1}{B}, \qquad \mathcal{S}_{ij}(\mathcal{L}_{\text{ON}}) = \frac{\left\|\mathbf{S}^{s/2}\mathbf{\Pi}_{ij}\mathbf{S}^{s/2}\right\|_F}{B} = \frac{(\lambda_i\lambda_j)^{s/2}}{B}, \tag{56}$$

with $s = +1$ for $\mathbf{X} = \mathbf{P}$ and $s = -1$ for $\mathbf{X} = \mathbf{H}$ (Eq. (3)). The non-orthonormal loss weights every direction equally, while the orthonormal loss carries the factor $(\lambda_i\lambda_j)^{s/2}$ with which the direction enters the orthonormal representation, the same factor that mediates the observable bounds of Appendix C.

D.4 Observable Sensitivities ↑toc

Energy The derivative of the energy E with respect to the density at the ground state is per definition the Hamiltonian matrix

$$\frac{\partial E}{\partial \mathbf{P}} = \mathbf{H}. \tag{57}$$

Hence the sensitivity, i.e. the directional derivative of energy with respect to a perturbation, is given by

$$\mathcal{S}_{ij}(E) = |\mathrm{tr}(\mathbf{H}^*\mathbf{\Pi}_{ij})| = \sqrt{2-\delta_{ij}}\left|\left(\mathbf{U}^\mathsf{T}\mathbf{H}^*\mathbf{U}\right)_{ij}\right| \tag{58}$$

so it reduces to rotating $\mathbf{H}^*$ into the overlap eigenbasis with $\mathbf{U}$ and reading off its (i,j) entry. Note that this is the unconstrained linear response with the Hamiltonian matrix fully pointing out of the Grassmannian manifold.

Dipole Since the dipole is linear in the density, $\mu_\alpha(\mathbf{P}) = \mathrm{tr}(\mathbf{P}\,\boldsymbol{\mu}_\alpha)$ with the dipole-integral matrices $\boldsymbol{\mu}_\alpha$, $\alpha \in \{x, y, z\}$, so is the dipole error with respect to a perturbation $\epsilon\mathbf{\Pi}_{ij}$

$$\Delta\mu_\alpha = \epsilon\mathrm{tr}(\boldsymbol{\mu}_\alpha\mathbf{\Pi}_{ij}) = \epsilon\sqrt{2-\delta_{ij}}\left(\mathbf{U}^\mathsf{T}\boldsymbol{\mu}_\alpha\mathbf{U}\right)_{ij} \tag{59}$$

The reported observable is the error magnitude $\|\Delta\boldsymbol{\mu}\|_2$. Making use of the componentwise error linearity we obtain

$$\mathcal{S}_{ij}(\|\Delta\boldsymbol{\mu}\|_2) = \sqrt{2-\delta_{ij}}\left(\sum_{\alpha\in\{x,y,z\}}\left(\mathbf{U}^\mathsf{T}\boldsymbol{\mu}_\alpha\mathbf{U}\right)_{ij}^2\right)^{1/2}. \tag{60}$$

L^1 density error The real-space density is linear in the density matrix, $\rho(\mathbf{r}) = \sum_{\mu\nu} P_{\mu\nu}\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})$, so with the eigenvector orbitals $\bar{u}_i(\mathbf{r}) := \sum_\mu (\mathbf{u}_i)_\mu\chi_\mu(\mathbf{r})$ the perturbation is again exact in ϵ, $\Delta\rho(\epsilon,\mathbf{r}) = \epsilon\sqrt{2-\delta_{ij}}\,\bar{u}_i(\mathbf{r})\bar{u}_j(\mathbf{r})$, and the one-sided limit gives

$$\mathcal{S}_{ij}(\|\Delta\rho\|_{L^1}) = \sqrt{2-\delta_{ij}}\,\|\bar{u}_i\bar{u}_j\|_{L^1} = \sqrt{2-\delta_{ij}}\int|\bar{u}_i(\mathbf{r})\bar{u}_j(\mathbf{r})|\,\mathrm{d}\mathbf{r}, \tag{61}$$

the absolute charge displaced along the direction. On the diagonal the prefactor is $\sqrt{2-\delta_{ii}} = 1$ and the integrand $\bar{u}_i^2$ is already non-negative, so the absolute value drops and the integral collapses onto the overlap itself,

$$\mathcal{S}_{ii}(\|\Delta\rho\|_{L^1}) = \int\bar{u}_i(\mathbf{r})^2\,\mathrm{d}\mathbf{r} = \mathbf{u}_i^\mathsf{T}\mathbf{S}\mathbf{u}_i = \lambda_i. \tag{62}$$

L^2 density error Measuring the same perturbation in the L^2 norm instead gives

$$\mathcal{S}_{ij}(\|\Delta\rho\|_{L^2}) = \sqrt{2-\delta_{ij}}\,\|\bar{u}_i\bar{u}_j\|_{L^2} = \sqrt{2-\delta_{ij}}\left(\int \bar{u}_i(\mathbf{r})^2\bar{u}_j(\mathbf{r})^2\,\mathrm{d}\mathbf{r}\right)^{1/2}, \tag{63}$$

the spatial overlap of the two orbital densities that span the direction.

E A Unifying Perspective on Hamiltonian and Density Learning ↑toc

Naively, examining the inverse HOMO-LUMO gap scaling in table 1 invites the question whether artificially increasing the gap could be utilized to reduce observable errors to zero. **Combining ROCKET with the level shifts formally unifies Hamiltonian and density matrix learning problems through a continuous interpolation parameter ξ.** Assume we learn a level-shifted Hamiltonian

$$\widetilde{\mathbf{H}}_\xi = \widetilde{\mathbf{H}} + \xi(\mathbf{I} - \widetilde{\mathbf{P}}) = \xi\mathbf{I} + (\widetilde{\mathbf{H}} - \xi\widetilde{\mathbf{P}})\,. \tag{64}$$

Since any rescaling of the matrix target is absorbed by data normalization and $\mathcal{L}_{\text{ROCKET}}$ is invariant under matrix shifts $\xi\mathbf{I}$ we formalize an equivalence relation for Hamiltonian learning targets

$$\left[\widetilde{\mathbf{H}}_1\right] = \left[\widetilde{\mathbf{H}}_2\right] \Leftrightarrow \left(\exists\, c\in\mathbb{R}\,,\, \sigma>0:\ \widetilde{\mathbf{H}}_1 = \sigma\big(\widetilde{\mathbf{H}}_2 + c\mathbf{I}\big)\right)\,. \tag{65}$$

Such that we can rewrite Eq. (64) to obtain $[\widetilde{\mathbf{H}}_\xi] = [\widetilde{\mathbf{H}} - \xi\widetilde{\mathbf{P}}]$. Dividing by ξ and taking the limit $\xi\to\infty$ gives $[\widetilde{\mathbf{H}}_\xi] = [-\widetilde{\mathbf{P}}]$, meaning that in the limit of large level shifts, Hamiltonian learning reduces to density learning. This says nothing about the optimality of the level shift, but it makes the tradeoff between learning a smooth but higher-dimensional Hamiltonian target vs learning a discontinuous but lower-dimensional density target explicit.

F Relation of the Orthonormal Loss to the Wavefunction Alignment Loss ↑toc

Li et al. (2025b) introduce Wavefunction Alignment Loss (WALoss) for Hamiltonian prediction. Their construction is closely related to our orthonormal Hamiltonian supervision, but differs in several important respects. We distinguish the mathematical loss presented in the paper from the loss in the released implementation.[2] The generalized eigenvectors of (1) satisfy

$$\mathbf{H}^*\mathbf{C}^* = \mathbf{S}\mathbf{C}^*\boldsymbol{\epsilon}^*, \qquad \mathbf{C}^{*\mathsf{T}}\mathbf{S}\mathbf{C}^* = \mathbf{I}. \tag{66}$$

We further define $\Delta\mathbf{H} = \widehat{\mathbf{H}} - \mathbf{H}^*$. **The initial WALoss formulation in Eq. (2) of Li et al. (2025b) is**

$$\mathcal{L}^{(2)}_{\text{WA}} := \left\|\mathbf{C}_i^{*\mathsf{T}}\widehat{\mathbf{H}}_i\mathbf{C}_i^* - \boldsymbol{\varepsilon}_i^*\right\|_F^2 \equiv \left\|\mathbf{C}_i^{*\mathsf{T}}\widehat{\mathbf{H}}_i\mathbf{C}_i^* - \mathbf{C}_i^{*\mathsf{T}}\mathbf{H}_i^*\mathbf{C}_i^*\right\|_F^2 = \left\|\mathbf{C}^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{C}^*\right\|_F^2\,. \tag{67}$$

This expression has a direct relation to our ON-Loss. With $\widetilde{\mathbf{H}} = \mathbf{S}^{-1/2}\mathbf{H}\mathbf{S}^{-1/2}$ and $\widetilde{\mathbf{C}}^* = \mathbf{S}^{1/2}\mathbf{C}^*$, the matrix $\widetilde{\mathbf{C}}^*$ is orthogonal and $\mathbf{C}^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{C}^* = \widetilde{\mathbf{C}}^{*\mathsf{T}}\Delta\widetilde{\mathbf{H}}\widetilde{\mathbf{C}}^*$, s.t. the invariance of the norm gives

$$\left\|\mathbf{C}^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{C}^*\right\|_F = \left\|\Delta\widetilde{\mathbf{H}}\right\|_F. \tag{68}$$

Hence, if all entries are weighted uniformly, Eq. (67) uses exactly the same matrix geometry as our Löwdin formulation, up to normalization and the use of a squared rather than unsquared norm. The reference MO basis is simply a different choice of an orthonormal basis. **However, the final WALoss equation presented in the paper is different.** Equation (3) assigns weights $\rho \gg \xi$ to the occupied orbitals and LUMO, but evaluates each state separately,

$$\mathcal{L}^{(3)}_{\text{WA}} = \rho\sum_{j\in\mathcal{I}}\left(\mathbf{c}_j^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{c}_j^*\right)^2 + \xi\sum_{j\notin\mathcal{I}}\left(\mathbf{c}_j^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{c}_j^*\right)^2. \tag{69}$$

[2] Paper: https://arxiv.org/pdf/2502.19227, Sec. 3, Eqs. (2)–(3). Code: https://github.com/microsoft/SPHNet/blob/main/src/training/module.py#L409-L421 and https://github.com/microsoft/SPHNet/blob/main/src/training/module.py#L441-L455.

Thus Eq. (69) contains only the diagonal of $\mathbf{C}^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{C}^{*}$. These terms are the first-order orbital-energy perturbations. In particular, the off-diagonal occupied–virtual couplings that drive first-order orbital rotations are absent. This contrasts with the preceding Eq. (2), whose full Frobenius norm does contain them.

The released implementation differs again. It explicitly forms both the diagonal and strict off-diagonal entries and, for its mean-squared-error (MSE) variant, uses [3]

$$\mathcal{L}_{\mathrm{WA}}^{\mathrm{code}} = \frac{1}{B}\sum_{i} T_{ii}^{2} + \frac{1}{B(B-1)}\sum_{i\neq j} T_{ij}^{2}, \qquad \mathbf{T} = \mathbf{C}^{*\mathsf{T}}\Delta\mathbf{H}\mathbf{C}^{*}. \tag{70}$$

Equation (70) is therefore not the uniformly weighted Frobenius norm in Eq. (68). The separate means give each diagonal entry $B-1$ times the weight of each off-diagonal entry. Moreover, the released code does not apply the occupied–LUMO ρ/ξ weighting of Eq. (69).

A further difference is that none of these flavors of WALoss is used as a standalone orthonormal loss, marking a significant gap in the literature w.r.t. the understanding of the proposed objective. The final training objective in Appendix K, Eq. (19) of Li et al. (2025b) combines it with elementwise MSE and MAE of the Hamiltonian in the original non-orthogonal representation.[4] The AO Frobenius term retains the overlap-dependent anisotropy removed by Eq. (68), while the entrywise MAE is not invariant even under a general orthogonal change of basis. Finally, WALoss does not account for the Hamiltonian gauge freedom. For any $c \in \mathbb{R}$,

$$\mathbf{H} \mapsto \mathbf{H} + c\mathbf{S}, \qquad \widetilde{\mathbf{H}} \mapsto \widetilde{\mathbf{H}} + c\mathbf{I}, \tag{71}$$

leaves all generalized eigenvectors and therefore the density unchanged. Both the paper WALoss and its implementation nevertheless penalize this physically irrelevant direction through their diagonal terms. ROCKET instead quotients it out explicitly,

$$\mathcal{L}_{\mathrm{ROCKET}} = \min_{c\in\mathbb{R}} \mathcal{L}_{\mathrm{Löwdin}}\left(\Delta\widetilde{\mathbf{H}} - c\mathbf{I}\right), \tag{72}$$

before introducing any further conditioning or occupied-subspace structure. Thus, the occupied–LUMO emphasis of WALoss and our later ROCKET constructions should not be conflated: WALoss reweights a fixed reference Hamiltonian, whereas ROCKET first identifies Hamiltonians that represent the same physical ground state.

[3] The implementation constructs the transformed residual at https://github.com/microsoft/SPHNet/blob/main/src/training/module.py#L409-L421 and combines the two sectors at https://github.com/microsoft/SPHNet/blob/main/src/training/module.py#L441-L455. The default orbital-energy metric is MSE in https://github.com/microsoft/SPHNet/blob/main/src/utility/hydra_config.py.

[4] See https://arxiv.org/pdf/2502.19227, Appendix K, Eq. (19). The released implementation analogously adds `HamiltonianError` and `OrbitalEnergyError` as separate training losses: https://github.com/microsoft/SPHNet/blob/main/src/training/module.py#L555-L559.

G AO VERSUS LÖWDIN SPATIAL DECAY ↑toc

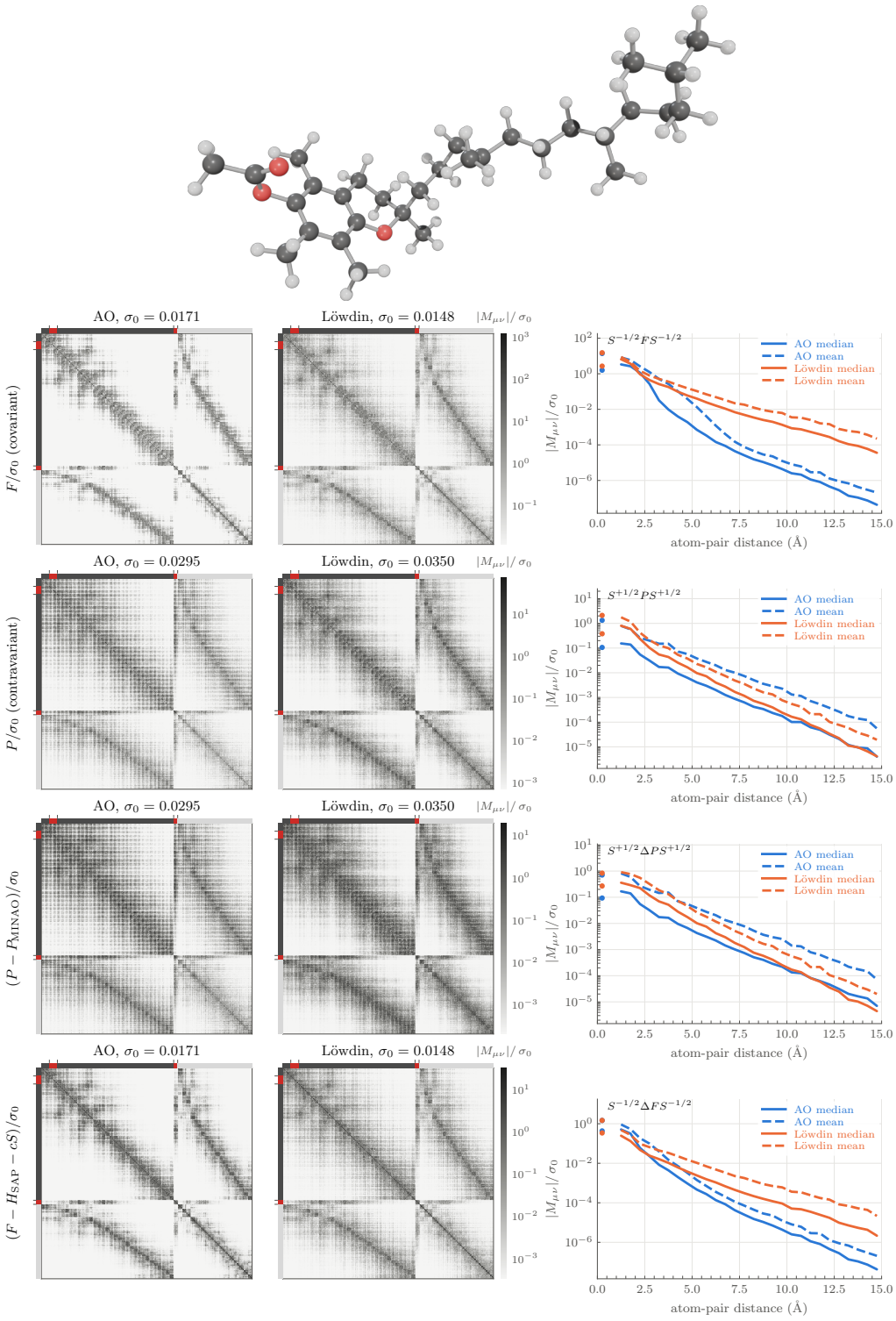


Figure 6: **AO versus Löwdin spatial decay for QM40 molecule 68311 at B3LYP/def2-SVP.** Columns are $\mathbf{H}$, $\mathbf{P}$, $\mathbf{P} - \mathbf{P}_{\mathrm{MINAO}}$, and $\mathbf{H} - \mathbf{H}_{\mathrm{SAP}} - c\mathbf{S}$, each in the AO (a.k.a. NON) basis and after Löwdin orthogonalization, scaled by the dataset root-mean-square (RMS) σ_0. AO entries decay with atom-pair distance. Löwdin entries remain large after the AO entries have dropped.

H DATASETS ↑ toc

We use five corpora. QM9 is the training set of the first regime, and QM40 tests the size extrapolation of the QM9-trained models. QMugs is the training set of the second regime and tests size extrapolation up to 660 electrons. 3BPA and Transition1x test the transfer of the QMugs-trained models to conformational and reactive chemistry. Tables 4 to 8 give one table per corpus. The upper block of each table describes the release and our copy of it. Each lower block covers one split, the rule that produced it, and the compute stages of table 3 we ran on it.

Conventions shared by all corpora. ↑ toc From every release we take only the atomic numbers and Cartesian coordinates. Released energies, forces and other properties do not enter any loss or metric. We recompute every electronic-structure quantity with the pipeline of table 2. All molecules are neutral and closed-shell. A checkpoint's element embedding is fixed by its training set. QM9-trained checkpoints know H, C, N, O and F, so we evaluate them only on corpora restricted to these elements. QMugs-trained checkpoints know H, C, N, O, F, P, S and Cl, so we can evaluate them on every corpus without Br or I. Each evaluation subset is a frozen list of sample identifiers, drawn with NumPy's `default_rng` under the seed given in the tables. Where a table stratifies by size, we assign each molecule to the electron bin $\lfloor N_{\mathrm{elec}}/20 \rfloor \cdot 20$ and draw the same number of molecules from every bin, so each size carries equal weight.

DFT reference. ↑ toc All reference labels use restricted Kohn–Sham DFT at B3LYP/def2-SVP. We run it with GPU4PySCF (Li et al., 2025a), the GPU backend of PySCF (Sun et al., 2020), with the settings in table 2.

Compute stages. ↑ toc We evaluate each prediction in up to three stages of increasing cost (table 3). The cheap stage holds every metric available at the cost of one Fock build, the budget at which we compare all models. The SCF and force stages cost considerably more on large molecules. On the larger test sets we therefore run them on a nested subset only, drawn equally from every electron bin so that it keeps the size distribution of the full set. The stage rows of each test set in the dataset tables give the number of molecules per bin on which each stage ran.

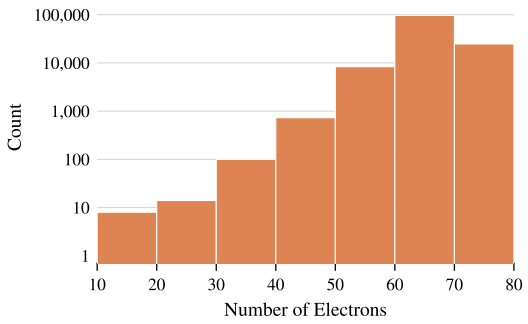


Figure 7: **Electron-count histogram of QM9.** The vertical axis is the molecule count on a logarithmic scale. QM9 contains only H, C, N, O, and F.

Table 2: **DFT reference pipeline (GPU4PySCF 1.8.1, B3LYP/def2-SVP, RI-JK).** This pipeline produces the evaluation references of all five corpora: all 130,831 QM9 molecules, 4,113 QM40 molecules, 191,591 QMugs conformers, 9,216 3BPA frames and 5,329 Transition1x frames. It also produces the training labels of the QMugs-trained checkpoints. The QM9 training labels come from a separate store computed with identical settings except for a looser convergence threshold (table 4). Transition1x self-consistency training uses no labels.

Parameter	Value	Notes
Basis set	def2-SVP	Ahlrichs split-valence double-zeta basis with polarization functions (Weigend & Ahlrichs, 2005).
XC functional	B3LYP	Three-parameter hybrid combining exact exchange (Becke, 1993) with LYP (Lee et al., 1988) and VWN local correlation (Vosko et al., 1980) in the parameterization of Stephens et al. (1994); we use the VWN-RPA form, as Gaussian does.
Density fitting	RI-J, RI-K	Density fitting of both the Coulomb (Vahtras et al., 1993) and the exact-exchange term (Weigend, 2002).
Auxiliary basis	def2-universal-JKFIT	One auxiliary basis serves both the RI-J and the RI-K contractions (Weigend, 2008).
Spin treatment	Restricted	Closed-shell restricted Kohn–Sham.
Quadrature grid	Level 3	PySCF grid level for the XC energy and potential (Becke, 1988; Sun et al., 2020).
Initial guess	MINAO	Superposition of atomic density matrices in a minimal ANO basis with spherically averaged occupations, projected onto def2-SVP (Almlöf et al., 1982; Van Lenthe et al., 2006; Lehtola, 2019). MINAO starts every reference SCF and is the base of the $\Delta\mathbf{P}$ readouts.
Hamiltonian base	SAP	$\mathbf{H}_{\mathrm{SAP}} = \mathbf{T} + \mathbf{V}_{\mathrm{SAP}}$: kinetic matrix plus a superposition of pretabulated screened atomic potentials, integrated on the level-3 grid (Lehtola, 2019; Lehtola et al., 2020b). It is the base of the $\Delta\mathbf{H}$ readouts and does not start the reference SCF.
SCF convergence	$10^{-10}\ E_h$	Converged when the energy changes by less than $10^{-10}\ E_h$ between cycles and the orbital-gradient norm is below 10^{-5} (GPU4PySCF default $\sqrt{10^{-10}}$); at most 100 cycles.

Table 3: **Compute stages.** Each split in tables 4 to 8 states on how many samples each stage ran. All errors are taken against the reference of table 2.

Stage	Cost	Computed quantities
Cheap	One Fock build	The predicted density matrix $\hat{\mathbf{P}}$ is the model output for density readouts and the aufbau density of the predicted $\hat{\mathbf{H}}$ (Eq. (1)) for Hamiltonian readouts. From $\hat{\mathbf{P}}$ we build $\mathbf{F}[\hat{\mathbf{P}}]$ once. Every quantity that needs no further Fock build and no nuclear gradient counts as cheap.
SCF	Up to 2× reference	A full SCF started from $\hat{\mathbf{P}}$ with the settings of table 2. A run is counted as not converged after twice the cycle count of the MINAO-started reference SCF.
Forces	One nuclear gradient	Nuclear forces evaluated at the aufbau density $\mathbf{P}_{\mathrm{aufbau}}(\mathbf{F}[\hat{\mathbf{P}}])$, compared with the reference forces $-\partial E/\partial \mathbf{R}$.

Table 4: **QM9** (Ramakrishnan et al., 2014). Training corpus of the first regime. The split is a deterministic size cut on the total atom count including hydrogens and involves no random draw. Compute stages are defined in table 3. The electron-count distribution is shown in Fig. 7.

Parameter	Value	Notes
Release	133,885 molecules	Small organic molecules with up to 9 heavy atoms (C, N, O, F), enumerated from GDB-17, with equilibrium geometries relaxed at B3LYP/6-31G(2df,p).
Our copy	130,831 molecules	We drop the 3,054 molecules that the authors flag as uncharacterized.
Geometries	0 K minima	Only the released thermochemistry refers to 298.15 K.
Elements	H, C, N, O, F	1,923 molecules contain F.
Size range		3–29 atoms, 1–9 heavy atoms, 10–74 electrons, 24–226 basis functions.
Training labels	$10^{-9}\ E_h$	The QM9 checkpoints are trained on labels converged to $10^{-9}\ E_h$ and otherwise identical to table 2. The total energies agree with the $10^{-10}\ E_h$ evaluation reference to a median of $2.5 \times 10^{-10}\ E_h$, with a maximum of $1.5 \times 10^{-5}\ E_h$.
Training split		
Rule	$N_{\text{atoms}} \leq 20$	Molecules with at most 20 atoms.
Size	103,790	All 1,923 F-containing molecules have 5–20 atoms and fall into this split, so the validation and test splits contain no F.
Validation split		
Rule	$N_{\text{atoms}} \in \{21, 22\}$	Molecules with 21 or 22 atoms.
Elements	H, C, N, O	No molecule contains F.
Size	17,594	The validation loss during training uses a fixed random subset of 5,000 molecules (seed 0), which also serves for checkpoint selection.
Test split		
Rule	$N_{\text{atoms}} \geq 23$	Uniform draw without replacement from the 9,447 molecules with 23–29 atoms (seed 0).
Elements	H, C, N, O	No molecule contains F.
Size	2,000	68% of the draw has 23 atoms; 1,960 molecules have 9 heavy atoms, 37 have 8 and 3 have 7. Used in table 25.
Cheap	All	
SCF	All	
Forces	All	

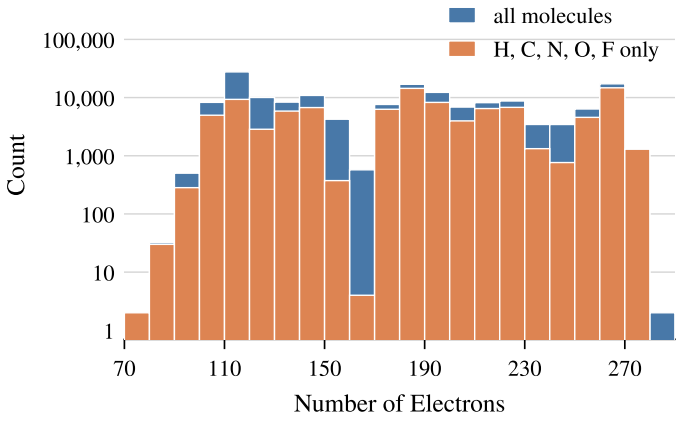


Figure 8: **Electron-count histogram of QM40.** The vertical axis is the molecule count on a logarithmic scale. The two series are the full dataset and its H, C, N, O, F subset.

Table 5: **QM40** (Madushanka et al., 2024). Size-extrapolation corpus for the QM9-trained checkpoints, which never see QM40 during training. To match their element vocabulary, we remove every molecule containing S or Cl. Compute stages are defined in table 3. The electron-count distribution is shown in Fig. 8.

Parameter	Value	Notes
Release	162,954 molecules	Drug-like molecules from ZINC14 with 10–40 heavy atoms (C, N, O, S, F, Cl), covering the size range of 88% of FDA-approved drugs. Geometries are pre-optimized with GFN2-xTB and relaxed at B3LYP/6-31G(2df,p), the level of theory of QM9; molecules with imaginary frequencies are removed.
Geometries	0 K minima	
Element filter	99,303 molecules	H, C, N, O, F only, 78–272 electrons. No molecule with 20 or 21 heavy atoms passes the filter, since all of them contain S or Cl.
Electron bins	$[80, 280)$	Ten bins of 20 electrons. The two molecules with 78 electrons fall below the first bin. The sparsest bin, 80–100 electrons, holds 311 molecules.
Validation (hyperparameter selection)		
Rule	50 regular, 25 hard	Used to select learning rate and weight decay (section L). We remove them from the pool before drawing the test sets.
Elements	H, C, N, O, F	Drawn from the element-filtered pool.
Size	75	
Cheap	All	Energy, dipole and L_1 error in table 23, alongside the validation loss.
Test split		
Rule	100 per bin	Nested in a parent draw of 200 molecules per bin from the remaining 99,228 molecules (seed 0, bins in ascending order): a first draw of 50 per bin, extended by 50 more per bin (seed 0).
Elements	H, C, N, O, F	Drawn from the element-filtered pool.
Size	1,000	Used in table 26 and all QM40 figures.
Cheap	100 per bin	All 1,000 molecules.
SCF	20 per bin	200 molecules: 10 per bin from each of the two draws of 50 (seed 0).
Forces	20 per bin	The same 200 molecules.

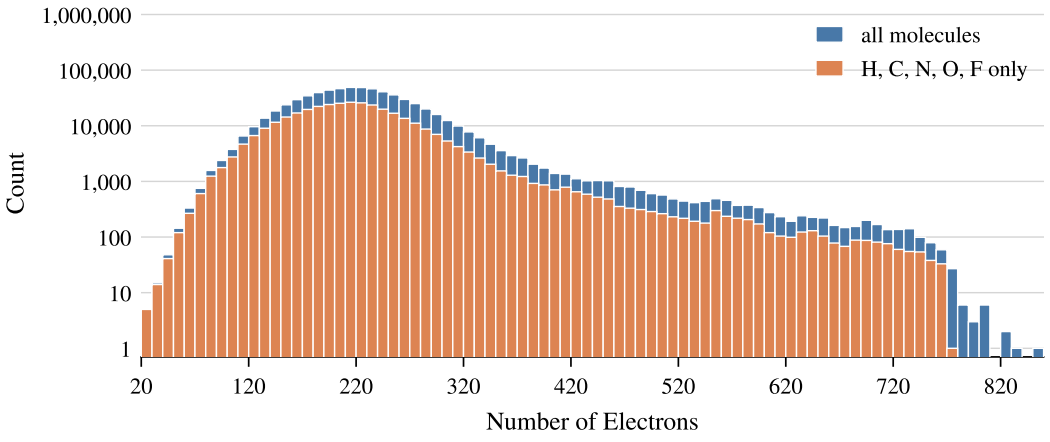


Figure 9: **Electron-count histogram of QMugs.** The vertical axis is the molecule count on a logarithmic scale. The two series are the full set and the H, C, N, O, F subset (Isert et al., 2021).

Table 6: **QMugs** (Isert et al., 2021). Training corpus of the second regime and size-extrapolation corpus for both checkpoint families. Every sample is a single conformer, and every split is disjoint by molecule. Sets with one conformer per molecule fix that choice in a single draw (seed 0). Compute stages are defined in table 3. The electron-count distribution is shown in Fig. 9.

Parameter	Value	Notes
Release	1,992,984 conformers of 665,911 molecules	Neutral, closed-shell drug-like molecules from ChEMBL 27 with 3–100 heavy atoms (H, C, N, O, F, P, S, Cl, Br, I). Each molecule has three conformers, selected by clustering GFN2-xTB metadynamics snapshots and then optimized with GFN2-xTB.
Geometries	0 K minima	GFN2-xTB local minima; 300 K is only the temperature of the conformer search.
Our copy	657,142 molecules	All conformers via OpenQDC, 22–850 electrons. Molecules are identified by SMILES, which merges some ChEMBL entries.
Training split		
Rule	$N_{\text{elec}} \leq 150$	Every conformer with at most 150 electrons (184,583 conformers of 60,367 molecules; 3–22 heavy atoms, 4–54 atoms) that is not in the validation split; all conformers kept.
Elements	H, C, N, O, F, P, S, Cl	Molecules with Br or I are excluded.
Size	175,256 conformers	57,349 molecules.
Validation split		
Rule	Largest 5%	The largest 5% of the molecules with at most 150 electrons, ranked by basis-set size, then electron count: 130–150 electrons, 16–22 heavy atoms, 383–436 basis functions.
Elements	H, C, N, O, F, P, S, Cl	Molecules with Br or I are excluded.
Size	9,327 conformers	3,018 molecules; used for checkpoint selection.
Test split, QMugs-trained checkpoints		
Rule	$N_{\text{elec}} \in [160, 660)$	25 bins of 20 electrons, one conformer per molecule; 100 molecules per bin, a first draw of 50 per bin extended by 50 more (seed 0). Disjoint from training because all conformers of a molecule share the electron count.
Elements	H, C, N, O, F, P, S, Cl	Molecules with Br or I are excluded.
Size	2,500	21–90 heavy atoms, 32–196 atoms. Used in table 28.
Cheap	100 per bin	All 2,500 molecules.
SCF	20 per bin	500 molecules: 10 per bin from each of the two draws of 50 (seed 0).
Forces	20 per bin	The same 500 molecules.
Test split, QM9-trained checkpoints		
Rule	$N_{\text{elec}} \in [280, 660)$	19 bins of 20 electrons, one conformer per molecule; 50 molecules per bin, nested in a parent draw of 200 per bin (seed 0).
Elements	H, C, N, O, F	The element vocabulary of the QM9-trained checkpoints.
Size	950	Used in table 27.
Cheap	50 per bin	All 950 molecules.
SCF	10 per bin	190 molecules (seed 0).
Forces	10 per bin	The same 190 molecules.

Table 7: **3BPA** (Kovács et al., 2021). Conformational transfer corpus for the QMugs-trained checkpoints, which never see 3BPA during training. We score each fixed-dihedral scan in full, without subsampling, because a subset of a potential-energy cut would be a different curve. Compute stages are defined in table 3.

Parameter	**Value**	**Notes**
Release	13,993 frames	A single flexible molecule, 3-(benzyloxy)pyridin-2-amine, with three consecutive rotatable bonds. The MD frames come from 25 ps Langevin trajectories at 300, 600 and 1200 K (1 fs time step) started in five conformational pockets; the dihedral scans are constrained DFT geometry optimizations. The released labels are ωB97X/6-31G(d), computed with ORCA.
Molecule	$C_{12}H_{12}N_2O$	27 atoms, 106 electrons, 270 basis functions.
Used subsets	Dihedral scans	7,047 frames. We do not use the two training sets (500 frames each) or the MD test sets (1,669 / 2,138 / 2,139 frames at 300 / 600 / 1200 K).
Dihedrals	α, β, γ	Three consecutive rotatable bonds of the benzyl-ether linker: α about the pyridine C–O bond, β about the O–CH_2 bond, and γ about the CH_2–phenyl bond. A fourth constraint fixes the amino group.
Grid	47×50	$\alpha = 55^\circ, \ldots, 239^\circ$ in steps of 4°; $\gamma = 5^\circ, \ldots, 299^\circ$ in steps of 6°.
Geometries	0 K	Constrained minima at each (α, γ) with β fixed.
Test split		
Rule	Three full scans	$\beta = 120^\circ$: 2,347 frames. The release lacks the three grid points $(\alpha, \gamma) = (239^\circ, 155^\circ/161^\circ/167^\circ)$. $\beta = 150^\circ$ and $\beta = 180^\circ$: complete grids of 2,350 frames. Each scan is scored as its own set. Used in table 29 and Fig. 4.
Size	7,047 frames	
Cheap	All	Since every frame is the same molecule, we remove a constant energy offset per scan and model, the median signed error.
Forces	All	

Table 8: **Transition1x** (Schreiner et al., 2022). Reactive corpus for label-free self-consistency training of the QMugs-trained checkpoints. We inherit the release split, which assigns each molecular formula to one split, so our training, validation and test sets share no formula and no reaction. Within each reaction we keep the reactant (R), transition state (TS) and product (P) and a window of path images around the TS. We sign each image's Kabsch root-mean-square deviation (RMSD) to the TS by the side of the TS it lies on. We divide $[-0.5, 0.5]$ Å of this signed distance into 20 bins of 0.05 Å and keep the image nearest each bin center, dropping empty bins and geometric duplicates. Compute stages are defined in table 3.

Parameter	**Value**	**Notes**
Release	9,654,813 frames	Configurations on and around the reaction paths of 10,073 reactions from Grambow et al. (2020), each with up to seven heavy atoms (C, N, O). The frames are images of nudged-elastic-band and climbing-image NEB runs, labeled at ωB97x/6-31G(d).
Elements	H, C, N, O	4–23 atoms.
Geometries	Reaction paths	Non-equilibrium configurations: many lie on or near reaction barriers.
Release split	9,561 / 225 / 287 reactions	Training / validation / test reactions of the release.
Reaction coordinate	$\xi \in [0, 1]$	$\xi = \frac{1}{2}(1 - d_{\mathrm{TS}}/d_{\mathrm{R,TS}})$ on the reactant side and $\xi = \frac{1}{2}(1 + d_{\mathrm{TS}}/d_{\mathrm{P,TS}})$ on the product side, where d is the Kabsch RMSD; R, TS and P sit at $\xi = 0$, 0.5 and 1.
Training split (label-free)		
Rule	All reactions	R, TS, P and the TS window for every training reaction.
Size	164,187 frames	8–23 frames per reaction (median 17), 56–178 basis functions.
Validation split		
Rule	100 reactions	One reaction per equal-size stratum of the 225 validation reactions, ranked by atom count (seed 0). Per reaction we keep R, TS, P and the two window images nearest ± 0.25 Å.
Size	500 frames	Used to select epoch and learning rate by the self-consistency error alone, which needs no reference.
Test split		
Rule	100 reactions	The validation draw rule applied to the 287 test reactions, with the full TS window of the training split.
Size	1,715 frames	5–14 atoms, 61–124 basis functions.
Cheap	All	
SCF	All	
Forces	All	

Table 9: **Output-scale normalization constants for different model types.** σ_0 is the element-wise root-mean-square of the supervision target in the representation the readout emits into, using the Löwdin frame for the Löwdin-basis heads and the AO basis for the AO-basis heads. We fit all constants on the full training split (103,790 molecules) and verify them end-to-end through the training data pipeline on 20 random training molecules with agreement within 2.7%. The readout applies the constant internally. The additive heads compose $\mathrm{init} + \sigma_0\,\mathrm{sym}(\hat{\mathbf{M}})$, while the rotation head scales the decoder output before generator construction ($\hat{\mathbf{X}}$ is linear in $\hat{\mathbf{M}}$), keeping the network output at unit scale while losses and gauge constants retain their physical units. $\bar{c} = 0.404129$ Ha is the training-set constant gauge. $\bar{c}_{\mathrm{AO}} = 0.382862$ Ha is the corresponding constant optimal in the AO Frobenius metric. The two metrics give different constants with spreads of 37.3 and 41.0 mHa. Because σ_0 anticorrelates with system size ($r = -0.86$ against the basis dimension), constants must be fitted on size-representative samples. For the rotation head, the decoder output $\hat{\mathbf{M}}$ enters through the projector sandwich $\hat{\mathbf{X}}(\hat{\mathbf{M}})$ (Eq. (34)), which pins only the occupied–virtual block of $\hat{\mathbf{M}}$ at unit weight ($\mathbf{P}_0\mathbf{S}$ is idempotent) and annihilates the rest. The value 0.0110 is the first-order pinned occupied–virtual block rms, measured against the purified aufbau-MINAO projector. The true Grassmann-log value runs about 7% higher (≈ 0.0117), and we use the first-order convention because the fit and the configuration must share one convention. Under the Grassmann interpretation, σ_0 is a rotation-angle scale rather than a matrix-element scale. The gap-shifted target admits a closed-form constant because the shift lies in the span of $\{\mathbf{I}, \mathbf{\Pi}^{\mathrm{ref}}\}$, giving $\sigma_0^2(\xi) = \sigma_{0,\Delta H}^2 + 2b_1\xi + b_2\xi^2$ with $b_1 = \mathbb{E}[\mathrm{tr}\,\Delta - \mathrm{tr}(\Delta\Pi)]\,/B^2 = 2.03\times10^{-3}$ and geometric $b_2 = \mathbb{E}[B-N]\,/B^2 = 4.70\times10^{-3}$. No re-measurement is needed for a new shift ($\sigma_0 = 0.0378/0.0409/0.0471$ at $\xi = 0.05/0.1/0.2$ Ha).

model type	supervision target	σ_0[e/Ha/−]	$1/\sigma_0$
$\Delta\mathbf{P}_{\mathrm{MINAO}}$	$\mathbf{S}^{1/2}\big(\mathbf{P}^{\mathrm{ref}} - \mathbf{P}^{\mathrm{MINAO}}\big)\mathbf{S}^{1/2}$	0.0350	28.6
$\Delta\mathbf{P}_{\mathrm{MINAO}}$ (AO)	$\mathbf{P}^{\mathrm{ref}} - \mathbf{P}^{\mathrm{MINAO}}$	0.0295	33.9
$\Delta\mathbf{H}_{\mathrm{SAP}}$	$\mathbf{S}^{-1/2}\big(\mathbf{H}^{\mathrm{ref}} - \mathbf{H}^{\mathrm{SAP}}\big)\mathbf{S}^{-1/2}$	0.0349	28.7
$\Delta\mathbf{H}_{\mathrm{SAP}}$ (AO)	$\mathbf{H}^{\mathrm{ref}} - \mathbf{H}^{\mathrm{SAP}}$	0.0471	21.2
$\Delta\mathbf{H}_{\mathrm{SAP}}$ fitted gauge	$\mathbf{S}^{-1/2}\big(\mathbf{H}^{\mathrm{ref}} - \mathbf{H}^{\mathrm{SAP}} - \bar{c}\,\mathbf{S}\big)\mathbf{S}^{-1/2}$	0.0148	67.6
$\Delta\mathbf{H}_{\mathrm{SAP}}$ fitted gauge (AO)	$\mathbf{H}^{\mathrm{ref}} - \mathbf{H}^{\mathrm{SAP}} - \bar{c}_{\mathrm{AO}}\,\mathbf{S}$	0.0171	58.5
GROOT	$\hat{\mathbf{M}}:\ \hat{\mathbf{X}}(\hat{\mathbf{M}}) = \mathbf{P}_0\hat{\mathbf{M}}\mathbf{\Pi}_0 - \mathbf{\Pi}_0\hat{\mathbf{M}}^{\intercal}\mathbf{P}_0$	0.0110	90.9

I MODEL HYPERPARAMETERS ↑toc

I.1 GROUND-STATE DESCRIPTOR MODELS ↑toc

Table 10: **QHFlow2 backbone (Kim et al., 2026).** SO(2)-equivariant edgewise–nodewise updates in the style of eSEN (Fu et al., 2025), with the generative flow replaced by the GROOT readout.

Hyperparameter	Value	Notes
Number of layers	4	
Sphere channels	192	Shared by backbone and decoder.
Maximum degree $\ell_{\max}$	4	def2-SVP off-diagonal $2 \otimes 2$ requires $\ell \leq 4$.
Edge channels	128	
Node hidden channels	128	
Radial basis	128	fixed Gaussians on $[0, 5]$ Å
Cutoff radius	5.0 Å	Envelope only.
Decoder degree $\ell_{\max}$	4	Same irreps as the backbone.
Decoder channels	$192 \to 64$	Self and pair layers at 192 per degree, then an equivariant linear to 64 per degree before the AO-block expansion.
Parameter count	81.8 M	Identical across all model types.
Pair-stream blocks	2	Off-diagonal head.
Decoder self/pair layers	2 each	
S^2 grid resolution	12×25	
Decoder pair radial basis	32	exp.-Bernstein on $[0, 15]$ Å
AO block irreps	$3 \times 0e + 2 \times 1e + 1 \times 2e$	def2-SVP on H, C, N, O, F.

I.2 MACHINE-LEARNED INTERATOMIC POTENTIALS MLIP REFERENCE MODELS ↑toc

Table 11: **Architecture of the PaiNN, NequIP and MACE baselines in the heat maps.** All three models sum per-atom energies over the molecule; the PaiNN and MACE MLPs use SiLU. MACE hidden features are $256\times(0e+1o+2e)$, and its edge spherical harmonics reach $l = 3$. The NequIP per-element energy scale is the per-element force RMS of the QM9 training split for the E and E+F models, and 1 for QMugs, which has no forces. The NequIP per-element shift is a least-squares fit to the training energies. MACE uses linear read-outs after interaction layers 1 and 2 and an MLP after layer 3, summed over layers.

	PaiNN	NequIP	MACE
Interaction layers	3	4	3
Features / channels	128	128 per l	256
Max. angular order	$l = 1$ (vectors)	$l_{\max} = 3$	$L_{\max} = 2$
Correlation order	–	–	3 (body order 4)
Parity	–	off	–
Radial basis	Gaussian, 20	Bessel, 8	Bessel, 8
Cutoff [Å], function	5.0, cosine	5.0, polynomial ($p = 6$)	5.0, polynomial ($p = 5$)
Read-out	MLP 128–64–1	linear, scale, shift	linear; MLP 256–128–1
Trainable parameters	0.6 M	33.5 M	18.5 M (E, E+F) 20.4 M (QMugs)

J Model Implementation Details ↑toc

Following Kim et al. (2026), we combine the eSCNMD GNN backbone by Wood et al. (2025), which builds on eSEN and eSCN (Fu et al., 2025; Passaro & Zitnick, 2023), with equivariant matrix encoders for the initial-guess and overlap matrices and a QHNet readout (Yu et al., 2023b). From the atomic numbers Z_a, coordinates $\mathbf{R}_a$, and initial-guess sub-matrices $\mathbf{X}_a$ centered at each atom a, the backbone constructs spherical-harmonic node features through $\mathrm{SO}(2)$-equivariant message-passing layers. The QHNet readout constructs the target matrix $\mathbf{X}$ blockwise. Each atom pair (i, j) contributes a block $\mathbf{X}_{ij} \in \mathbb{R}^{B_{Z_i} \times B_{Z_j}}$, formed by a self-tensor product of the node features for $i = j$ and a filtered tensor product across the two nodes for $i \neq j$. Depending on the ansatz, $\mathbf{X}$ is the Hamiltonian $\mathbf{H}$, the density $\mathbf{P}$, or the GROOT input $\hat{\mathbf{M}}$. For the Hamiltonian ansatz, we obtain the density by solving Eq. (1). For the Hamiltonian and density ansätze, we symmetrize the decoded block matrix as $\mathbf{X} \leftarrow (\mathbf{X} + \mathbf{X}^\mathsf{T})/2$. The GROOT readout instead uses the raw $\hat{\mathbf{M}}$. Since we parametrize the generator in the same orthonormal basis as the Hamiltonian and density matrices, these readout heads remain interchangeable.

Our JAX port (Babuschkin et al., 2020) corrects two bugs in the official reference implementation at commit `d82fa3a1` from 2026-08-09. The first bug occurs in `escn_backbone_v4_no_t.py`, lines 576 to 584, where the model adds contracted overlap and initial-guess blocks to the node representations. The contraction yields one scalar feature vector per atom, but the reference code reshapes this atom-indexed tensor as if its leading dimension counted graphs in `QHFlow_so2_v5_1_no_t.py`, lines 347 to 348 and then gathers from it with the graph assignment vector of the batch. Consequently, every atom of the g-th molecule in a batch receives the feature of the g-th atom of the batch instead of its own, so for a single molecule all atoms receive the feature of the first atom and the injected features depend on the atom ordering. Our port keeps the atom index and, going beyond the reference, contracts each block to features of degree $\ell = 0, 1, 2$ that it adds to the matching channels of its own node. In our tests, this correction reduces the relative rotation-equivariance error from 10^{-3} to 10^{-6}. The second bug appears where the backbone hands off features to the tensor-product decoder at `QHFlow_so2_v5_1_no_t.py`, line 356 for node features and line 365 for pair features. The backbone stores features as $(N, (\ell_{\max} + 1)^2, C)$ tensors, in which the magnetic index precedes the channel index within each degree ℓ, whereas the `e3nn` irreps of the decoder hold C copies of each degree with the channel index first. The reference code flattens the tensor with a raw reshape that omits the required blockwise transposition, so every $\ell \geq 1$ block reaches the decoder with its components misordered. Our port converts between the two memory layouts explicitly. With identical parameters and the same activation-normalization constant, our JAX decoder matches the PyTorch decoder to double-precision round-off.

K Training Costs ↑toc

L TRAINING AND OPTIMIZATION ↑ toc

Table 12: **Training and optimization hyperparameters.** Muon optimizer configuration. The optimizer configuration is shared by all reported runs. Family-specific deviations are stated in the notes. Model-init and training-PRNG seeds are drawn from OS entropy per run and logged.

Hyperparameter	Value	Notes
Batch size	16	
Optimizer	Muon	Muon on the 2D hidden weight matrices, AdamW on all remaining parameters. Both optimizer branches use Nesterov momentum, following the Optax (contrib) defaults (Babuschkin et al., 2020). We also tried the same momentum as used for standard Adam but observed no improvements.
Muon momentum	0.95	Default (Jordan et al., 2024).
Muon Newton–Schulz steps	5	Default (Jordan et al., 2024).
Adam (β_1, β_2)	$(0.9,\ 0.999)$	Default (Kingma & Ba, 2015).
Muon consistent RMS	0.2	Each orthogonalized $m \times n$ Muon update is scaled by $0.2\sqrt{\max(m,n)}$, giving approx. shape-independent update RMS (Liu et al., 2025).
Weight decay λ_{wd}		Tried {3e-4, 1e-3, 3e-3, 1e-2, 3e-2}. Same weight decay on the Muon and AdamW branches. Decay is masked and applied only to the 2D weight matrices and tensor-product path weights. Biases, normalizations and embeddings are undecayed.
Base learning rate (lr)	-	retuned for each model family
Warmup schedule	Linear	
Warmup steps	2,000	
Decay schedule	Cosine	With minimum learning rate of 1e-5 $\times$ base lr
Decay steps	101,776	Total 103,776 steps $=$ 16 epochs at 6,486 steps/epoch, batch size 16.
Training epochs	16	Tried $\{12,\ 16,\ 20\}$ tuned in favor of the baseline methods. Longer runs mainly improve in distribution, with no gains on the size extrapolation val sets.
Parameter EMA decay	0.995	Tried $\{0.99,\ 0.995\}$
Consecutive NaN skips	10	Maximum consecutive non-finite updates skipped before the run aborts.

Table 13: **ROCKET restraint hyperparameters.** The restraint weight α_{ROCKET} is the 3σ weight $\alpha/9$ with $\alpha = \mathbb{E}[\mathcal{L}_{\min}]/\sigma_c^2$ fitted on the full training split. A three-standard-deviation gauge offset then costs as much as the mean shift-minimized loss. See section 4.

Hyperparameter	Value	Notes
Optimal shift std σ_c	37.3 mHa	Standard deviation of the L_2-optimal gauge shifts over the training split.
Restraint weight α	1.161	Calibrated as $\alpha_{\text{ROCKET}} = \mathbb{E}[\mathcal{L}_{\min}(\Delta\widetilde{\mathbf{H}})]/\sigma_c^2$ over the full training split (103,790 molecules).
EMA decay	0.99	Decay of the EMA of the optimal overlap shift center. Initialized to training split optimal 0.404.

Table 14: **NON loss ablations.** Masked decay with σ_0 output norm. Target is Δ-MINAO except for the absolute-readout block. Validation and QM40 loss are the AO objective (masked Frobenius root-mean-square error, RMSE) on the QM9 validation split at the final epoch and on the fixed probe molecules (reg. 50 / hard 25). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch. ON-loss arms are in table 15.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	val loss	reg. 50	hard 25	Muon	Adam
abs P-learning (no-rescale)								
16	3e-4	1e-2	995756	5.73	16.3	29.6	0.16	0.55
16	1e-3	1e-2	995756	4.82	14.8	27.8	0.17	0.44
16	**3e-3**	**1e-2**	995756	4.48	**13.8**	**26.0**	0.26	0.40
"	"	"	684851	4.40	14.1	36.9	0.25	0.41
"	"	"	240706	**4.38**	18.0	384	0.26	0.41
16	3e-3	3e-3	995756	4.66	14.6	52.9	0.42	0.50
16	1e-2	1e-2	995756	5.31	15.0	**26.0**	0.32	1.03
abs P-learning								
16	**3e-3**	**1e-2**	995756	4.49	14.8	77.9	0.25	0.42
"	"	"	684851	**4.41**	14.1	27.2	0.25	0.42
"	"	"	240706	4.43	**13.9**	**26.5**	0.25	0.42
NON loss, Δ-Learning with P_{MINAO}								
16	1e-3	3e-3	982933	5.18	15.0	25.8	0.22	0.55
16	3e-3	3e-3	982933	4.84	14.5	25.6	0.41	0.50
16	**3e-3**	**1e-2**	982933	4.67	14.3	**24.6**	0.25	0.41
"	"	"	709626	4.51	13.8	34.8	0.26	0.42
"	"	"	744226	**4.50**	**13.0**	25.7	0.26	0.41
16	1e-2	3e-3	982933	7.80	16.9	28.6	0.79	1.36

L.1 Density Models ↑toc

L.1.1 NON Loss ↑toc

L.1.2 ON Loss ↑toc

Table 15: **ON loss ablation.** Masked decay with σ_0 output norm. Target is Δ-MINAO. Validation is the training objective (masked Frobenius RMSE, Löwdin picture) on the QM9 validation split at the final epoch. The QM40-loss columns are the *same objective* evaluated on the fixed probe molecules (reg. 50 / hard 25, same units). The density family ranks arms (Spearman $+0.87/+0.79$ vs E). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch. NON-loss arms are in table 14. GROOT arms are in table 16. Prediction-frame and auxiliary-supervision arms are in table 17.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	val loss	reg. 50	hard 25	Muon	Adam
ON loss								
12	1e-3	1e-2	0	1.46	7.54	9.06	0.17	0.48
"	"	"	749811	1.47	7.10	9.12	0.17	0.47
12	1e-3	3e-2	0	1.40	7.99	9.51	0.10	0.35
12	3e-3	1e-3	0	1.43	6.03	8.60	0.42	0.59
12	3e-3	3e-3	0	1.43	6.39	8.90	0.37	0.51
12	3e-3	1e-2	0	1.39	6.67	9.09	0.24	0.39
12	3e-3	3e-2	0	1.35	6.93	9.24	0.11	0.44
12	1e-2	1e-2	0	1.40	7.34	9.43	0.33	0.73
16	3e-3	3e-4	882901	1.33	6.13	9.47	0.50	0.62
16	3e-3	1e-3	118790	1.35	6.89	10.7	0.47	0.58
16	**3e-3**	**3e-3**	154091	1.30	5.93	8.97	0.40	0.48
"	"	"	406157	1.32	**5.75**	**8.28**	0.38	0.49
"	"	"	145343	1.35	6.73	9.25	0.40	0.48
16	3e-3	1e-2	0	**1.29**	6.36	9.17	0.24	0.40
"	"	"	261042	**1.29**	5.96	9.23	0.24	0.40

Table 16: **GROOT ON-loss ablations.** Masked decay with σ_0 output norm and a Δ-MINAO target. Validation is the training objective on the QM9 validation split at the final epoch. The QM40-loss columns evaluate that objective on the fixed probe molecules (reg. 50 / hard 25). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	val loss	reg. 50	hard 25	Muon	Adam
+ *GROOT*$_{\mathrm{proj}}$								
16	1e-3	3e-2	193360	**0.38**	1.24	2.48	0.09	0.32
16	3e-3	3e-3	193360	0.39	2.33	2.52	0.36	0.47
16	**3e-3**	**1e-2**	193360	0.39	1.17	2.97	0.22	0.36
"	"	"	141727	0.39	**1.06**	**2.44**	0.21	0.35
"	"	"	899526	0.39	1.12	2.72	0.22	0.35
16	3e-3	3e-2	193360	0.39	1.39	2.69	0.10	0.41
16	1e-2	3e-2	193360	0.44	1.18	2.60	0.11	0.83
+ *GROOT*$_{\mathrm{rot}}$								
16	1e-3	3e-3	628859	0.57	1.59	9.99	0.22	0.56
16	3e-3	3e-3	628859	0.56	1.54	11.8	0.39	0.48
16	3e-3	1e-2	628859	**0.54**	1.54	11.7	0.24	0.38
16	3e-3	3e-2	628859	0.55	1.63	10.1	0.10	0.46
16	**1e-2**	**1e-2**	628859	0.57	1.54	9.82	0.30	0.77
"	"	"	506493	0.60	1.53	9.90	0.28	0.74
"	"	"	900833	0.58	1.50	**9.74**	0.29	0.75
16	1e-2	3e-2	628859	0.60	**1.47**	9.89	0.11	1.02
"	"	"	798434	NaN	NaN	NaN	NaN	NaN
"	"	"	400920	NaN	NaN	NaN	NaN	NaN

L.1.3 Prediction Frame and Auxiliary Supervision ↑ *toc*

Table 17: **ON-loss prediction-frame and auxiliary-supervision ablations.** Density target is Δ-MINAO and Hamiltonian target is Δ-SAP with the training-set optimal constant gauge. Parenthesized validation is the arm's own training objective when that objective differs from the table metric. The QM40-loss columns use the Löwdin objective on the fixed probe molecules (reg. 50 / hard 25). The Hamiltonian AO-frame run was killed before the QM40 probe.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	val loss	reg. 50	hard 25	Muon	Adam
Density GROOT$_{\mathrm{proj}}$*, pre/post supervision*								
16	3e-3	1e-2	193360	(2.17)	3.75	5.77	0.23	0.38
Density GROOT$_{\mathrm{proj}}$*, AO-frame prediction*								
16	3e-3	1e-2	193360	0.58	1.39	3.03	0.21	0.34
Hamiltonian ROCKET, AO-frame prediction								
16	3e-4	3e-3	443905	(25.9)	–	–	–	–

L.2 HAMILTONIAN MODELS ↑toc

L.2.1 NON-ORTHONORMAL LOSS ↑toc

Table 18: **Hamiltonian NON-loss ablations.** Masked decay with σ_0 output norm. Target is Δ-SAP, with a training-set constant gauge in the second block only. Validation is the NON training objective on the QM9 validation split at the final epoch. The QM40-loss columns evaluate that objective on the fixed probe molecules (reg. 50 / hard 25). The covariant loss is gauge-dominated and nearly flat across arms, quoted for completeness, not for ranking. $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch. ON-loss arms are in table 20.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	QM9 val	reg. 50	hard 25	Muon	Adam
abs H-learning								
16	1e-4	1e-3	684464	0.78	3.95	12.2	0.18	0.62
ΔH_{SAP}*-learning*								
16	3e-6	1e-3	684464	2.12	8.7	18.4	0.18	0.63
16	1e-5	1e-3	684464	1.12	5.3	15.1	0.18	0.63
16	**3e-5**	**1e-3**	684464	0.66	**4.0**	**13.7**	0.18	0.63
16	1e-4	3e-4	684464	0.39	**4.0**	13.8	0.18	0.62
16	1e-4	1e-3	684464	0.39	**4.0**	13.8	0.18	0.62
16	1e-4	3e-3	684464	0.39	**4.0**	14.0	0.18	0.62
16	3e-4	1e-3	684464	0.27	50.2	119	0.18	0.62
16	1e-3	1e-3	684464	**0.26**	33.8	81.2	0.20	0.60
$\Delta(H_{\mathrm{SAP}} + c_{\textit{train-set-fitted}}S)$*-learning*								
16	1e-6	3e-3	684464	6.94	11.9	25.0	0.18	0.63
16	3e-6	1e-3	684464	1.65	4.8	15.3	0.18	0.63
16	3e-6	3e-3	684464	1.65	4.8	15.3	0.18	0.63
16	**1e-5**	**1e-3**	684464	0.78	**3.4**	**13.4**	0.18	0.63
"	"	"	896521	0.78	3.5	13.7	0.18	0.63
"	"	"	405602	0.92	5.0	15.2	0.18	0.63
"	"	"	748187	0.79	3.5	13.6	0.18	0.63
16	3e-5	1e-3	684464	0.45	3.6	13.9	0.18	0.63
16	1e-4	1e-3	684464	0.27	4.0	15.3	0.18	0.63
16	1e-4	3e-3	684464	0.27	4.0	15.4	0.18	0.62
16	3e-4	1e-3	684464	0.20	29.0	65.8	0.18	0.62
16	3e-4	3e-3	684464	0.20	28.1	64.4	0.17	0.60
16	1e-3	1e-3	684464	**0.19**	26.1	259	0.20	0.60
16	3e-3	1e-3	684464	0.21	2 608	35 112	0.36	0.58

L.2.2 WALOSS ↑toc

Table 19: **Hamiltonian WALoss ablation.** Target is Δ-SAP, with a training-set constant gauge in the second block only. Validation is the WALoss training objective on the QM9 validation split at the final epoch. The QM40-loss columns evaluate each run's own WALoss objective on the fixed probe molecules (reg. 50 / hard 25). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	QM9 val	reg. 50	hard 25	Muon	Adam
ΔH_{SAP}*-learning*								
16	**1e-4**	**1e-3**	684464	11.8	12.1	56	0.18	0.63
$\Delta(H_{\mathrm{SAP}} + c_{\textit{train-set-fitted}}S)$*-learning*								
16	3e-6	1e-3	684464	36.0	31.6	87	0.18	0.63
16	1e-5	1e-3	684464	9.17	12.1	110	0.18	0.63
16	3e-5	1e-3	684464	3.59	7.18	339	0.18	0.63
16	**1e-4**	**1e-3**	684464	2.37	**5.01**	**20**	0.18	0.63
"	"	"	485210	2.77	5.82	56.6	0.18	0.63
"	"	"	710469	3.10	5.55	24.3	0.18	0.63
16	3e-4	1e-3	684464	**1.91**	5.19	22	0.18	0.64
16	1e-3	1e-3	684464	3.93	10.0	62	0.21	0.73

L.2.3 ORTHONORMAL LOSS ↑ *toc*

Table 20: **Hamiltonian ON-loss ablations.** Target is Δ-SAP, with a training-set constant gauge in the $\Delta(H_{\mathrm{SAP}} + cS)$ block only. Validation is the training objective on the QM9 validation split at the final epoch. The QM40-loss columns evaluate that objective on the fixed probe molecules (reg. 50 / hard 25). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	QM9 val	reg. 50	hard 25	Muon	Adam
ΔH_{SAP}-learning								
16	**3e-5**	**3e-3**	670079	1.60	**3.6**	**9.7**	0.18	0.63
"	"	"	435126	1.61	3.7	9.9	0.18	0.62
"	"	"	301732	1.63	3.7	10.0	0.18	0.63
$\Delta(H_{\mathrm{SAP}} + c_{\text{train-set-fitted}}S)$-learning								
12	3e-5	1e-2	0	1.68	3.8	9.9	0.18	0.62
12	1e-4	1e-2	0	1.13	4.0	10.7	0.17	0.61
12	3e-4	1e-2	0	0.87	4.5	13.3	0.16	0.57
12	1e-3	1e-2	0	0.78	13.7	33.2	0.17	0.47
12	3e-3	1e-2	0	0.73	17.5	33.4	0.23	0.37
16	1e-5	3e-3	670079	2.20	4.2	9.9	0.18	0.63
16	**3e-5**	**3e-3**	670079	1.51	3.8	9.7	0.18	0.62
"	"	"	435126	1.52	3.7	9.7	0.18	0.62
"	"	"	301732	1.53	**3.6**	**9.6**	0.18	0.63
16	3e-5	1e-2	670079	1.51	3.8	9.8	0.18	0.62
16	1e-4	3e-3	670079	1.04	3.8	10.1	0.18	0.62
16	1e-4	1e-2	670079	1.03	3.8	10.3	0.17	0.60
16	1e-4	3e-2	670079	1.02	4.0	10.9	0.15	0.56
16	3e-4	1e-2	670079	0.80	4.5	13.4	0.16	0.56
16	1e-3	3e-3	940789	**0.71**	22.0	48.3	0.21	0.55
16	1e-3	1e-2	0	0.73	17.3	40.4	0.17	0.44

L.3 ROCKET-LOSS ABLATIONS ↑ toc

Table 21: **ROCKET ablations.** Target is Δ-SAP with a training-set constant gauge. Validation is the arm's training objective on the QM9 validation split at the final epoch. The QM40-loss columns evaluate that objective on the fixed probe molecules (reg. 50 / hard 25). $\|W\|$ columns are per-entry RMS of the final EMA weights by optimizer branch.

					QM40 loss		$\|W\|_{\mathrm{rms}}$	
ep	lr	wd	seed	QM9 val	reg. 50	hard 25	Muon	Adam
1st stage (unrestrained gauge)								
16	1e-4	1e-2	443905	0.96	41.0	32.5	0.17	0.60
16	**3e-4**	**1e-2**	443905	0.75	**14.2**	**24.5**	0.16	0.56
16	1e-3	1e-2	443905	**0.67**	24.6	26.8	0.17	0.44
2nd stage (restrained gauge)								
16	1e-4	1e-2	586405	1.38	3.6	7.8	0.17	0.60
16	1e-4	3e-3	586405	1.36	3.5	7.7	0.18	0.62
16	**3e-4**	**3e-3**	443905	0.76	2.4	7.5	0.18	0.60
"	"	"	452826	0.77	**2.3**	**7.1**	0.18	0.60
"	"	"	862718	0.77	2.5	8.2	0.18	0.61
16	3e-4	1e-2	443905	0.76	2.5	8.0	0.16	0.56
16	1e-3	1e-2	779430	0.73	2.5	8.9	0.17	0.44
16	3e-3	1e-2	958391	**0.70**	3.5	14.9	0.24	0.37
2nd stage (10 fold rel restrained gauge)								
16	**3e-4**	**3e-3**	443905	0.87	2.1	6.31	0.18	0.60
3rd stage (supervise occ+LUMO only; $k = 1$)								
16	3e-4	3e-3	443905	0.46	2.2	3.8	0.18	0.61
16	3e-4	1e-2	443905	0.48	2.3	4.2	0.16	0.56
16	**1e-3**	**3e-3**	443905	0.27	1.4	3.3	0.21	0.56
"	"	"	617814	0.43	2.2	3.9	0.21	0.56
"	"	"	307265	**0.25**	**1.3**	**3.2**	0.21	0.56
16	1e-3	1e-3	443905	0.33	1.7	3.5	0.23	0.60
16	3e-3	3e-3	443905	0.26	1.4	4.1	0.36	0.47
3rd stage (supervise occ+LUMO+1; $k = 2$)								
16	1e-3	3e-3	443905	0.25	1.1	3.3	0.21	0.56
3rd stage (supervise occ+LUMO+3; $k = 4$)								
16	1e-3	3e-3	443905	0.34	0.8	3.1	0.21	0.56
2nd stage + gap shifts (50 mHa, 100 mHa, 200 mHa, 800 mHa)								
16	**3e-4**	**3e-3**	443905	**0.83**	**2.4**	7.9	0.18	0.60
16	**3e-4**	**3e-3**	443905	0.92	2.6	**7.7**	0.18	0.60
16	**3e-4**	**3e-3**	443905	0.99	3.0	8.0	0.18	0.60
16	**3e-4**	**3e-3**	443905	3.08	7.1	9.6	0.18	0.60
3rd stage + gap shift 100 mHa								
16	**3e-4**	**3e-3**	443905	0.86	3.1	4.5	0.18	0.61

Table 22: Transition1x self-consistency training: the training objective's validation residual (mean over the Transition1x validation split of Table 8, in units of 10^{-4}) of every run, per configuration, learning rate and epoch; for ROCKET 3rd stage the same objective without its gauge-leash term. LR is the peak learning rate of the warmup–cosine schedule. Per epoch column and configuration, the lowest value is underlined: that rate is the checkpoint the ablation takes at that epoch. †: the epochs the Transition1x test ablation table (Table 30) scores. *: the run was stopped before its configured epochs; its later cells are empty and it competes only where it has a value. Every block is a residual of its own model – a density residual (dimensionless), or for ROCKET 3rd stage a Hamiltonian one in Ha – never compared across blocks.

LR	Validation self-consistency residual after epoch											
	1†	2†	3	4†	5	6	7	8	9	10	11	12†
GROOT$_{\text{rot}}$ – density residual, 10^{-4} (dimensionless)												
3×10^{-5}*	12.4	9.10	8.47	7.59	7.00	6.98						
10^{-4}	9.08	7.41	7.00	6.65	6.20	5.33	5.26	4.98	4.88	4.84	4.81	4.81
3×10^{-4}	8.13	7.42	6.71	5.97	5.86	5.25	5.01	4.66	4.34	4.28	4.23	4.19
10^{-3}	10.6	8.75	8.45	7.38	6.36	5.50	5.19	4.71	4.41	4.15	4.07	4.04
3×10^{-3}	13.1	10.4	8.87	8.01	7.02	6.85	5.78	5.43	5.07	4.75	4.58	4.52
GROOT$_{\text{proj}}$ – density residual, 10^{-4} (dimensionless)												
3×10^{-5}*	9.83	7.86	7.30	6.77								
10^{-4}	7.32	6.11	5.72	5.52	5.08	4.86	4.81	4.72	4.62	4.58	4.55	4.54
3×10^{-4}	7.61	6.46	6.06	5.42	5.18	4.77	4.66	4.27	4.12	4.11	4.01	4.02
10^{-3}	8.45	7.06	7.04	6.11	5.80	4.85	4.79	4.37	4.17	3.99	3.92	3.91
3×10^{-3}	9.93	8.06	7.57	6.93	6.58	5.85	5.58	4.82	4.61	4.41	4.33	4.26
P$_{\text{base}}$ (ON) – raw-readout density residual, 10^{-4} (dimensionless)												
10^{-4}*	8.52	6.99	6.27	5.84	5.58	5.40						
3×10^{-4}*	6.68	5.63	5.10	4.80	4.61	4.40	4.29	4.19	4.14	4.11		
10^{-3}	6.52	5.57	5.03	4.72	4.43	4.16	3.94	3.77	3.66	3.60	3.56	3.55
3×10^{-3}*	8.86	7.04	7.49									
10^{-2}*	1377	2037										
ROCKET 3rd stage – gauge-free occ+LUMO Hamiltonian residual, 10^{-4} Ha												
10^{-4}*	4.08	3.24	2.85	2.60								
3×10^{-4}	2.97	2.54	2.29	2.11	1.99	1.90	1.85	1.75	1.72	1.70	1.69	1.69
10^{-3}	2.97	2.61	2.38	2.14	2.00	1.80	1.68	1.60	1.52	1.49	1.46	1.45
3×10^{-3}*	3.68	2.96	2.78	2.48								
10^{-2}*	5.06	8.35										

L.4 Self-Consistency-Training Learning-Rate Selection ↑toc

Self-consistency training needs no reference labels, so we also select its learning rate without them. Table 22 reports the training objective on the Transition1x validation split for every learning rate and epoch. At each scored epoch we take the learning rate with the lowest residual; the resulting checkpoints are evaluated on the test split in table 30.

M LOSS VS OBSERVABLES ↑toc

Table 23: **QM40 probe observables of the production runs.** Every arm whose epoch, learning rate and weight decay are set to its production value, aggregated over its entropy-seed replicates as the mean over seeds. The loss column evaluates each family's own training objective on the fixed probe molecules (reg. 50, $\times 10^4$) and is comparable only within a block. The cata. column sums the reg. 50 and hard 25 catastrophe counts. E is the one-Fock-build electronic-energy MAE in mEh, and a catastrophe is $E > 1\,000$ mEh. Bold and underlined entries are the lowest and second-lowest values in each observable column, respectively. Red marks an error above 3/10 mEh for regular/hard E, 1/4 D for the dipole, 0.2/1 e for L_1, or cata. > 1.

					E [mEh]		dipole [D]		L_1 [e]		
	n	lr	wd	loss	reg.	hard	reg.	hard	reg.	hard	cata.
NON losses											
Density											
abs. P	3	3e-3	1e-2	14.3	**259**	**1 655**	0.76	31.9	0.50	4.13	**11.7**
$\Delta P_{\mathrm{MINAO}}$	3	3e-3	1e-2	13.7	**76**	**263**	0.51	4.50	0.38	1.4	1.3
Hamiltonian											
abs. H	1	1e-4	1e-3	4.0	**8 626**	**19 188**	**319**	**441**	**24.7**	**44.4**	**64**
ΔH_{SAP}	1	3e-5	1e-3	4.0	**712**	**2 989**	**50.8**	**120**	**3.57**	**11.1**	**24**
$\Delta(H_{\mathrm{SAP}} + cS)$	4	1e-5	1e-3	3.9	**1 396**	**3 532**	**69.5**	**125**	**5.44**	**11.2**	**35**
WALoss											
Hamiltonian											
ΔH_{SAP}	1	1e-4	1e-3	12.1	**2 032**	**4 338**	1.46	25.0	**4.64**	9.12	**68**
$\Delta(H_{\mathrm{SAP}} + cS)$	3	1e-4	1e-3	5.5	**614**	**3 333**	7.07	**51.7**	1.76	7.54	**19.7**
ON loss											
Density											
$\Delta P_{\mathrm{MINAO}}$	3	3e-3	3e-3	6.1	**35.0**	41.1	0.09	**0.60**	0.08	**0.24**	0
$\mathrm{GROOT}_{\mathrm{proj}}$	3	3e-3	1e-2	1.1	0.87	9.24	**0.07**	1.89	**0.05**	0.46	0
$\mathrm{GROOT}_{\mathrm{rot}}$	3	1e-2	1e-2	1.5	1.68	**121**	**0.07**	**51.4**	**0.05**	**3.57**	2
Hamiltonian											
ΔH_{SAP}	3	3e-5	3e-3	3.7	2.74	19.8	0.20	4.43	0.22	1.25	0
$\Delta(H_{\mathrm{SAP}} + cS)$	3	3e-5	3e-3	3.7	2.22	15.5	0.23	4.26	0.19	1.11	0
ROCKET (*Hamiltonian*)											
1st stage	1	3e-4	1e-2	14.2	3.83	69.7	0.46	1.87	0.38	**2.30**	0
2nd stage	3	3e-4	3e-3	2.4	1.22	8.38	0.12	3.63	0.11	0.83	0
+ gap shift 50 mHa	1	3e-4	3e-3	2.4	1.22	8.24	0.11	3.70	0.11	0.84	0
+ gap shift 100 mHa	1	3e-4	3e-3	2.6	1.25	8.60	0.11	3.49	0.11	0.83	0
+ gap shift 200 mHa	1	3e-4	3e-3	3.0	1.44	7.78	0.12	3.52	0.12	0.78	0
+ gap shift 800 mHa	1	3e-4	3e-3	7.1	2.83	9.42	0.13	2.08	0.13	0.56	0
3rd stage $k = 1$	3	1e-3	3e-3	1.6	**0.42**	7.69	0.09	1.58	0.09	0.72	0
+ gap shift 100 mHa	1	3e-4	3e-3	3.1	0.48	**6.34**	0.09	2.54	0.09	0.73	0
3rd stage $k = 2$	1	1e-3	3e-3	1.1	0.45	6.50	0.10	1.39	0.09	0.67	0
3rd stage $k = 4$	1	1e-3	3e-3	0.8	0.51	8.57	0.09	2.50	0.10	0.78	0

M.1 Frobenius Error versus Observable Error ↑ toc

Table 23 compares losses through the observables they produce. What it does not say is whether the quantity a loss actually minimizes – the Frobenius residual of the predicted matrix – is a faithful stand-in for those observables, and whether measuring that residual in the Löwdin basis rather than the raw AO one makes it a better stand-in. Figures 10 to 13 plot one observable error each against both Frobenius residuals, over the same arms, and table 24 scores every one of their panels with the Pearson correlation r of the two $\log_{10}$ errors and their mutual information I. The Löwdin congruence is the better predictor in 22 of the 24 panel pairs on r and in all 24 on I, and the gap is widest exactly where it matters: on the arms that were trained with $\mathcal{L}_{\mathrm{ON}}$, whose raw AO residual carries almost no information about the observable error it leaves behind.

Table 24: Every panel of the Frobenius–observable plates, scored both ways. r is the Pearson correlation of $\log_{10}$ of the Frobenius error with $\log_{10}$ of the observable error, and I their mutual information taken to base 2, k-nearest-neighbor (Kraskov) estimator at $k = 5$. Both are measured on the arm pooled over its seeds, the cloud the panel draws: NON is the raw congruence $\|\Delta\mathbf{X}\|_F$ and ON the Löwdin one, $\|\mathbf{S}^{+1/2}\Delta\mathbf{X}\mathbf{S}^{+1/2}\|_F$ for a density and $\|\mathbf{S}^{-1/2}\Delta\mathbf{X}\mathbf{S}^{-1/2}\|_F$ for a Hamiltonian. Higher is better for both statistics; the better of each pair is bold. Test set: QM40, 1000 molecules per seed, of which the 200 that carry forces form the force subset. The force rows are therefore measured on the force subset and the remaining observables on the whole test set, with no prediction excluded. Each arm is pooled over its seeds: 4000 points on the whole test set and 800 on the force subset for a 4-seed arm.

		r		I (base 2)	
Model	Seeds	NON	ON	NON	ON
$\lvert\Delta E\rvert$ [mE_h]					
$\mathbf{P}_{\mathrm{base}}$ (NON)	3	0.595	**0.669**	0.38	**0.56**
$\mathbf{P}_{\mathrm{base}}$ (ON)	3	0.778	**0.909**	0.87	**1.58**
$\mathbf{H}_{\mathrm{base}}$ (NON)	1	0.924	**0.932**	1.17	**1.32**
$\mathbf{H}_{\mathrm{base}}$ (ON)	3	0.555	**0.802**	0.30	**1.06**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	4	0.911	**0.927**	1.10	**1.24**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3	0.556	**0.760**	0.29	**0.84**
$\lVert\Delta\boldsymbol{\mu}\rVert_2$ [D]					
$\mathbf{P}_{\mathrm{base}}$ (NON)	3	0.695	**0.732**	0.51	**0.60**
$\mathbf{P}_{\mathrm{base}}$ (ON)	3	0.544	**0.616**	0.34	**0.41**
$\mathbf{H}_{\mathrm{base}}$ (NON)	1	**0.883**	0.880	1.09	**1.30**
$\mathbf{H}_{\mathrm{base}}$ (ON)	3	0.529	**0.602**	0.25	**0.46**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	4	**0.869**	0.867	1.04	**1.10**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3	0.524	**0.567**	0.29	**0.35**
$\lVert\Delta\rho\rVert_{L^1}$ [e]					
$\mathbf{P}_{\mathrm{base}}$ (NON)	3	0.784	**0.857**	0.78	**1.06**
$\mathbf{P}_{\mathrm{base}}$ (ON)	3	0.737	**0.837**	0.67	**0.97**
$\mathbf{H}_{\mathrm{base}}$ (NON)	1	0.907	**0.913**	1.18	**1.44**
$\mathbf{H}_{\mathrm{base}}$ (ON)	3	0.577	**0.727**	0.38	**0.88**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	4	0.905	**0.912**	1.14	**1.31**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3	0.580	**0.675**	0.37	**0.67**
$\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$ [mE_h/Å]					
$\mathbf{P}_{\mathrm{base}}$ (NON)	3	0.772	**0.800**	0.66	**0.72**
$\mathbf{P}_{\mathrm{base}}$ (ON)	3	0.618	**0.717**	0.51	**0.70**
$\mathbf{H}_{\mathrm{base}}$ (NON)	1	0.905	**0.911**	1.05	**1.36**
$\mathbf{H}_{\mathrm{base}}$ (ON)	3	0.599	**0.791**	0.45	**0.93**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	4	0.891	**0.910**	1.19	**1.38**
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3	[51]0.604	**0.757**	0.38	**0.73**

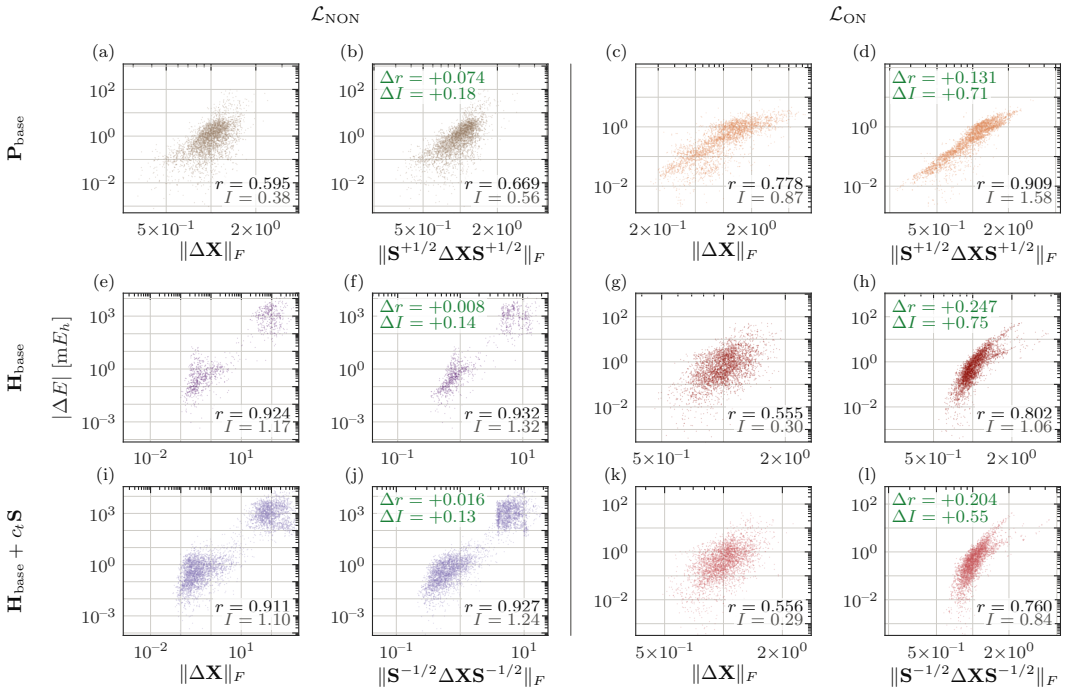


Figure 10: **Harris–Foulkes energy error against Frobenius residual.** One point per QM40 test molecule and seed, logarithmic on both axes. Rows are the model families P_{base}, H_{base}, $H_{\mathrm{base}} + c_t\mathbf{S}$; the column blocks are the training objective ($\mathcal{L}_{\mathrm{NON}}$, then $\mathcal{L}_{\mathrm{ON}}$), and within a block the left panel scores the raw residual $\|\Delta\mathbf{X}\|_F$ and the right the Löwdin one (Eq. (3)). Panels carry r and I, Löwdin panels also the gain over their raw neighbor; table 24 collects them for all four figures.

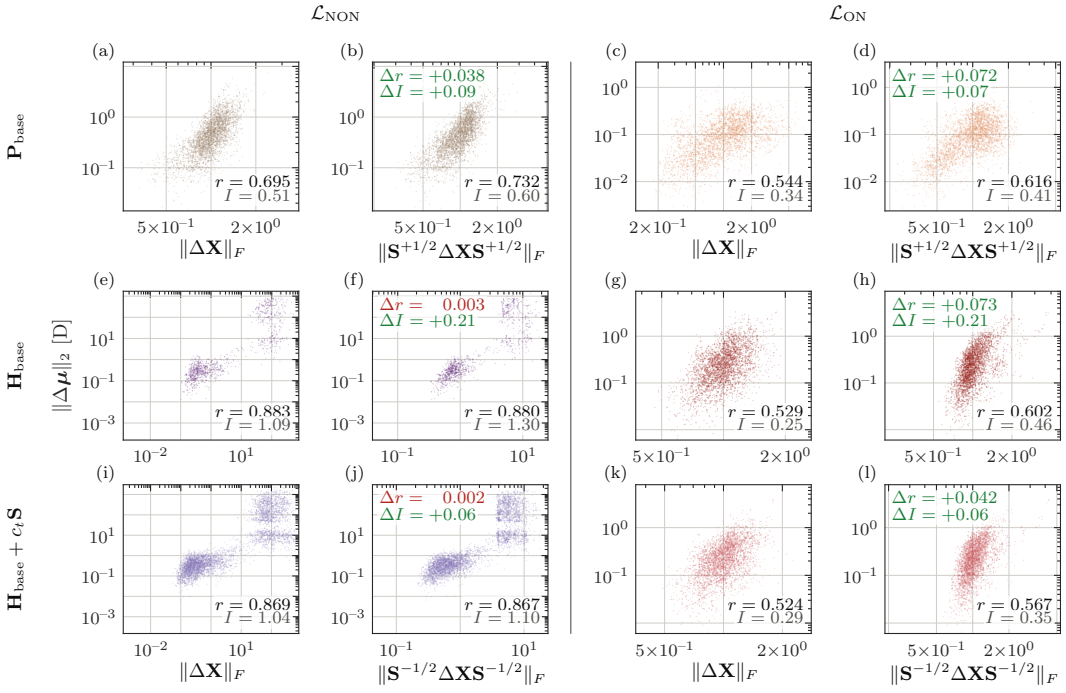


Figure 11: **Dipole error against Frobenius residual.** Same arms, panel layout, pooling and annotations as Fig. 10, with the dipole error $\|\Delta\boldsymbol{\mu}\|$ in Debye on the vertical axis.

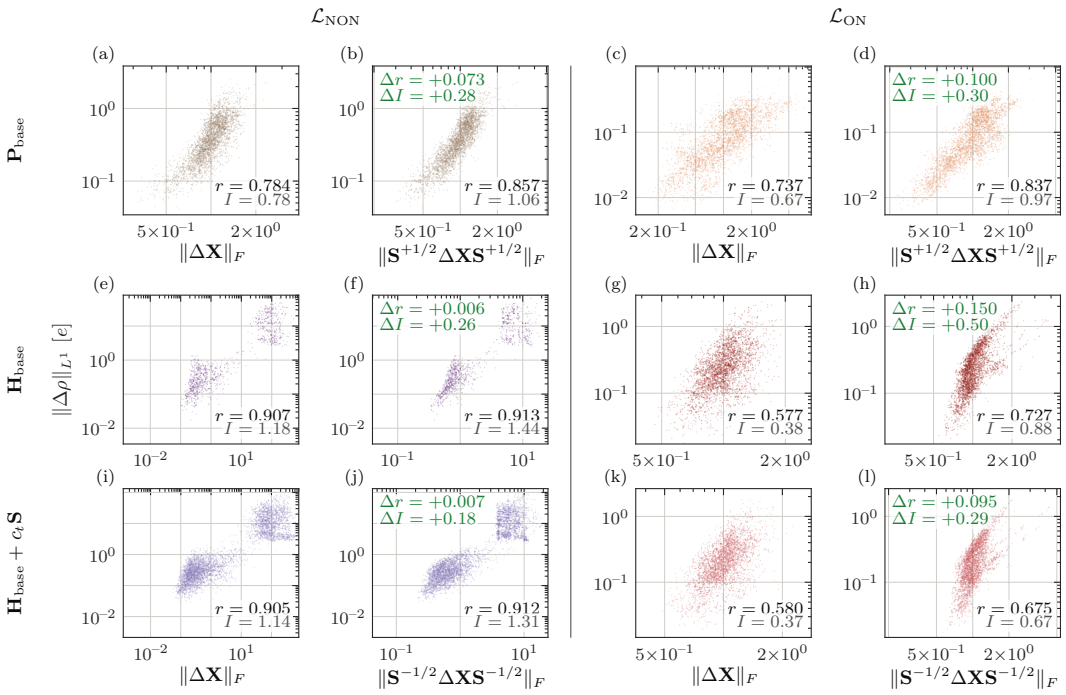


Figure 12: **Density L^1 error against Frobenius residual.** Same arms, panel layout, pooling and annotations as Fig. 10, with the real-space density residual $\|\Delta\rho\|_{L^1}$ in electrons on the vertical axis.

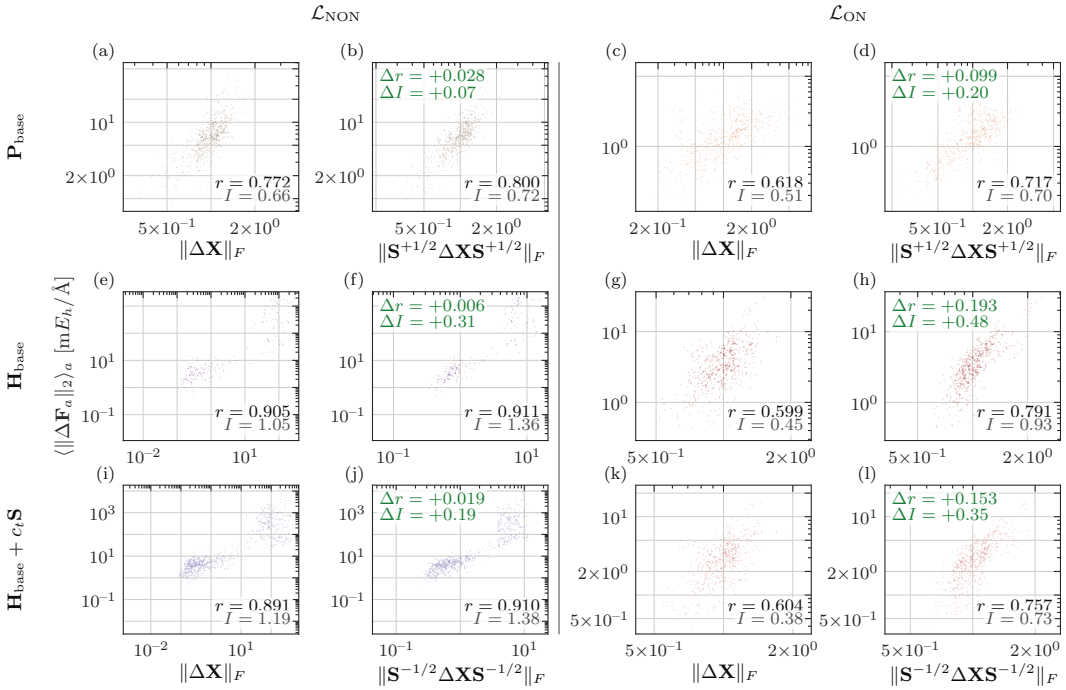


Figure 13: **Force error against Frobenius residual.** Same arms, panel layout, pooling and annotations as Fig. 10, with the per-atom mean force error $\langle\|\Delta\mathbf{F}_a\|\rangle$ on the vertical axis. Only the 200 molecules per seed that carry reference forces enter these panels, so the clouds are correspondingly sparser than in Figs. 10 to 12.

Table 25: Ablation on the QM9 test split (Table 4). Each cell is the mean over seeds of the per-molecule mean absolute error, with the range over seeds in square brackets. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed mean and the best bracketed single seed (the lower end of its range) are underlined. A prediction is excluded as an outlier when its direct inversion in the iterative subspace (DIIS) residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS}-\mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F} = \mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). An arm that lost every molecule keeps its row, with – before each bracketed value. Excluded, Energy (standard), Dipole, L_1, HOMO–LUMO: 2000 molecules before the outlier exclusion. Forces, ERIC: 2000 molecules before the outlier exclusion. ERIC is the number of SCF cycles from the predicted density relative to the MINAO guess, averaged over the full lane with unconverged runs pinned at the cap of 2. LR is the peak learning rate of the warmup–cosine schedule and WD the weight decay, read from each checkpoint's own training config; the seeds of a row share both. † one seed, no range.

Model	LR	WD	Excluded (%)		Standard energy (mHa) $\lvert\Delta E\rvert$	Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$	L_1 (e) $\lVert\Delta\rho\rVert_{L^1}$	Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$	HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$	ERIC (ratio)
Density-space models										
$\mathbf{P}_{\mathrm{abs.}}$ (NON)	3×10^{-3}	10^{-2}	62.9	9.65	(12.1)	0.0419 (0.0553)	0.0333 (0.0428)	0.974 (1.26)	0.578 (0.774)	0.745 (0.754)
			[62, 66]		[9.1, 10]	[0.040, 0.043]	[0.032, 0.034]	[0.94, 1.0]	[0.56, 0.62]	[0.74, 0.75]
					(12, 13)	(0.055, 0.056)	(0.043, 0.043)	(1.3, 1.3)	(0.77, 0.78)	(0.75, 0.75)
$\mathbf{P}_{\mathrm{abs.}}$ (NON), no rescale	3×10^{-3}	10^{-2}	61.8	10.4	(12.2)	0.0421 (0.0545)	0.0337 (0.0427)	1.00 (1.27)	0.570 (0.759)	0.745 (0.754)
			[60, 64]		[10, 11]	[0.042, 0.042]	[0.034, 0.034]	[1.00, 1.0]	[0.56, 0.59]	[0.74, 0.75]
					(12, 13)	(0.054, 0.055)	(0.042, 0.043)	(1.3, 1.3)	(0.75, 0.78)	(0.75, 0.75)
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	21.2	8.47	(9.32)	0.0512 (0.0563)	0.0399 (0.0439)	1.19 (1.31)	0.741 (0.813)	0.751 (0.755)
			[19, 25]		[8.4, 8.5]	[0.051, 0.052]	[0.039, 0.040]	[1.2, 1.2]	[0.73, 0.77]	[0.75, 0.75]
					(9.2, 9.6)	(0.055, 0.057)	(0.043, 0.045)	(1.3, 1.3)	(0.79, 0.86)	(0.75, 0.76)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	0	1.22	(1.22)	0.00634 (0.00634)	0.00460 (0.00460)	0.139 (0.139)	0.0798 (0.0798)	0.660 (0.660)
			[0, 0]		[1.2, 1.3]	[0.0062, 0.0064]	[0.0046, 0.0046]	[0.14, 0.14]	[0.079, 0.082]	[0.66, 0.66]
					(1.2, 1.3)	(0.0062, 0.0064)	(0.0046, 0.0046)	(0.14, 0.14)	(0.079, 0.082)	(0.66, 0.66)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-2}	10^{-2}	0	0.0499	(0.0499)	0.00385 (0.00385)	0.00368 (0.00368)	0.0798 (0.0798)	0.0559 (0.0559)	0.623 (0.623)
			[0, 0]		[0.048, 0.052]	[0.0037, 0.0041]	[0.0036, 0.0039]	[0.077, 0.084]	[0.055, 0.056]	[0.62, 0.63]
					(0.048, 0.052)	(0.0037, 0.0041)	(0.0036, 0.0039)	(0.077, 0.084)	(0.055, 0.056)	(0.62, 0.63)
$\mathrm{GROOT}_{\mathrm{proj}}$	3×10^{-3}	10^{-2}	0	0.0254	(0.0254)	0.00307 (0.00307)	0.00284 (0.00284)	0.0622 (0.0622)	0.0442 (0.0442)	0.605 (0.605)
			[0, 0]		[0.025, 0.026]	[0.0031, 0.0031]	[0.0028, 0.0028]	[0.062, 0.062]	[0.044, 0.044]	[0.60, 0.61]
					(0.025, 0.026)	(0.0031, 0.0031)	(0.0028, 0.0028)	(0.062, 0.062)	(0.044, 0.044)	(0.60, 0.61)
Hamiltonian-space models										
$\mathbf{H}_{\mathrm{abs.}}$ (NON)†	10^{-4}	10^{-3}	3.85	0.995	(30.3)	0.0208 (0.484)	0.0168 (0.0832)	0.381 (3.18)	0.263 (2.32)	0.703 (0.714)
$\mathbf{H}_{\mathrm{base}}$ (NON)†	3×10^{-5}	10^{-3}	1.15	0.635	(3.92)	0.0226 (0.0337)	0.0190 (0.0249)	0.427 (0.632)	0.331 (0.540)	0.709 (0.711)
$\mathbf{H}_{\mathrm{base}}$ (NON), QM9 best†	10^{-3}	10^{-3}	0.0500	0.107	(0.112)	0.00894 (0.00897)	0.00826 (0.00829)	0.179 (0.180)	0.127 (0.128)	0.667 (0.667)
$\mathbf{H}_{\mathrm{base}}$ (NON), WALoss†	10^{-4}	10^{-3}	100	–	(852)	– (0.461)	– (2.14)	– (44.8)	– (26.4)	– (1.06)
$\mathbf{H}_{\mathrm{base}}$ (ON)	3×10^{-5}	3×10^{-3}	0	0.146	(0.146)	0.0157 (0.0157)	0.0154 (0.0154)	0.326 (0.326)	0.238 (0.238)	0.690 (0.690)
			[0, 0]		[0.14, 0.15]	[0.015, 0.016]	[0.015, 0.016]	[0.32, 0.33]	[0.23, 0.24]	[0.69, 0.69]
					(0.14, 0.15)	(0.015, 0.016)	(0.015, 0.016)	(0.32, 0.33)	(0.23, 0.24)	(0.69, 0.69)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	10^{-5}	10^{-3}	5.76	1.05	(52.0)	0.0286 (0.213)	0.0242 (0.109)	0.543 (3.21)	0.385 (2.73)	0.725 (0.742)
			[2.4, 10]		[0.86, 1.4]	[0.028, 0.031]	[0.023, 0.026]	[0.52, 0.59]	[0.37, 0.40]	[0.72, 0.73]
					(22, 88)	(0.074, 0.46)	(0.057, 0.18)	(1.7, 5.6)	(1.2, 4.6)	(0.73, 0.76)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON), QM9 best†	10^{-3}	10^{-3}	0.350	0.0558	(5.45)	0.00666 (0.0258)	0.00599 (0.0162)	0.133 (0.376)	0.106 (0.448)	0.651 (0.653)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON), WALoss	10^{-4}	10^{-3}	100	–	(164)	– (0.152)	– (0.487)	– (8.54)	– (9.11)	– (0.930)
			[100, 100]		(130, 203)	(0.12, 0.19)	(0.39, 0.61)	(6.6, 11)	(7.4, 11)	(0.91, 0.95)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3×10^{-5}	3×10^{-3}	0	0.132	(0.132)	0.0145 (0.0145)	0.0136 (0.0136)	0.291 (0.291)	0.212 (0.212)	0.684 (0.684)
			[0, 0]		[0.13, 0.13]	[0.014, 0.015]	[0.013, 0.014]	[0.29, 0.29]	[0.20, 0.22]	[0.68, 0.68]
					(0.13, 0.13)	(0.014, 0.015)	(0.013, 0.014)	(0.29, 0.29)	(0.20, 0.22)	(0.68, 0.68)
ROCKET 2nd stage	3×10^{-4}	3×10^{-3}	0	0.0395	(0.0395)	0.00612 (0.00612)	0.00605 (0.00605)	0.128 (0.128)	0.0957 (0.0957)	0.653 (0.653)
			[0, 0]		[0.039, 0.040]	[0.0059, 0.0063]	[0.0060, 0.0061]	[0.13, 0.13]	[0.088, 0.10]	[0.65, 0.65]
					(0.039, 0.040)	(0.0059, 0.0063)	(0.0060, 0.0061)	(0.13, 0.13)	(0.088, 0.10)	(0.65, 0.65)
ROCKET 3rd stage	10^{-3}	3×10^{-3}	0	0.00958	(0.00958)	0.00402 (0.00402)	0.00428 (0.00428)	0.0898 (0.0898)	0.0690 (0.0690)	0.618 (0.618)
			[0, 0]		[0.0092, 0.0099]	[0.0040, 0.0041]	[0.0041, 0.0044]	[0.088, 0.091]	[0.067, 0.071]	[0.62, 0.62]
					(0.0092, 0.0099)	(0.0040, 0.0041)	(0.0041, 0.0044)	(0.088, 0.091)	(0.067, 0.071)	(0.62, 0.62)

N Model Ablation Data ↑toc

Table 25 reports the ablation arms of table 23 on the held-out QM9 test partition, the in-distribution reference point for the extrapolation results that follow. Figures 14 and 15 resolve the same density arms over the electron count N_{elec} instead of aggregating them into a single probe number, which separates errors that are already present in the training regime from errors that only grow under size extrapolation. Figures 16 and 17 do the same for the Hamiltonian arms. Each pair shows QM40 first and then QMugs, which extends the extrapolation range to roughly twice the QM40 electron counts. Tables 26 and 27 report the same arms aggregated over the full QM40 and QMugs test sets.

N.1 Training on QM9 ↑toc

N.1.1 DENSITY-PREDICTION PLOTS ↑toc

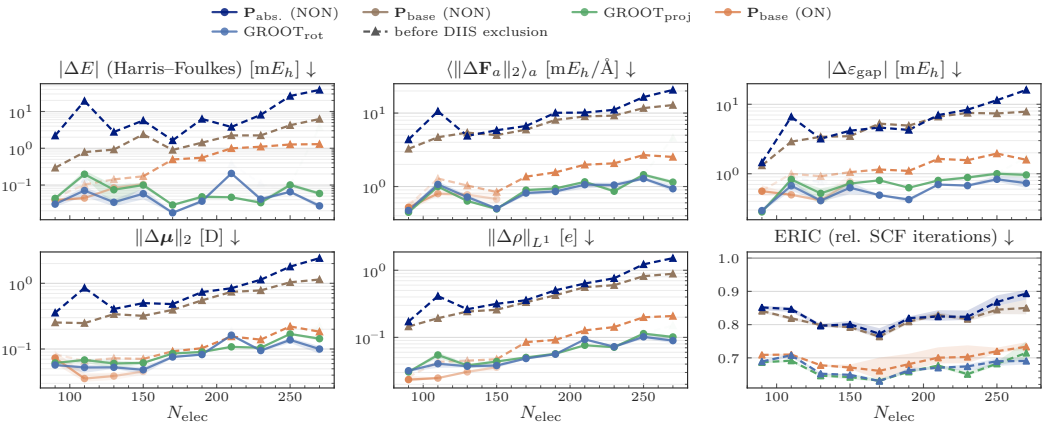


Figure 14: **Density-readout ablations on QM40, resolved over electron count.** All arms use the QM9-trained checkpoint of table 4 evaluated on the out-of-distribution QM40 set of table 5, binned by the molecular electron count N_{elec} and given a one-Fock-build budget. The arms are the absolute density readout, its unrescaled variant, the $\Delta\mathbf{P}_{\text{MINAO}}$ readout in the AO basis, the same Δ-readout in the Löwdin basis (ON loss), and the two manifold-restricted variants GROOT$_{\text{proj}}$ and GROOT$_{\text{rot}}$ (section 4). Lines are means over the entropy-seed replicates stated in the legend and the shaded band spans the seed spread, so the single-seed arms carry no band; the vertical axes of all error panels are logarithmic. ERIC is the effective relative iteration count (Eberhard et al., 2026a), where values below one mean fewer SCF iterations than the reference initial guess.

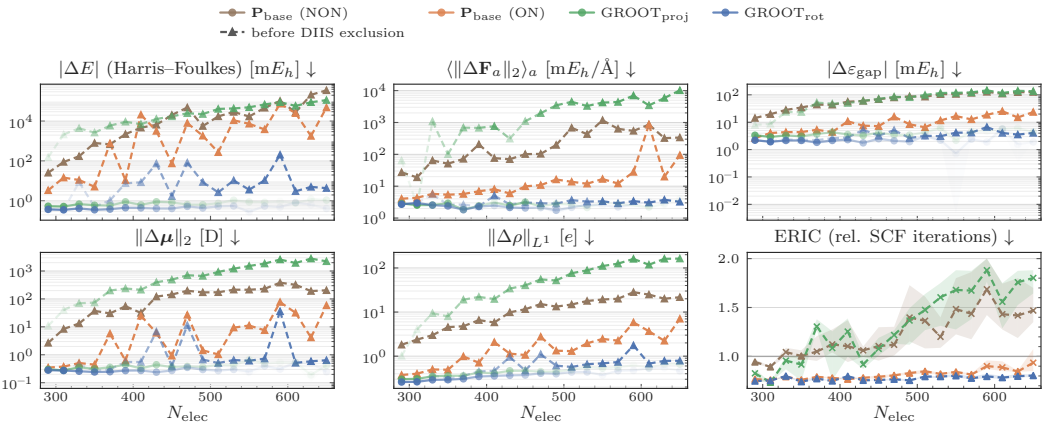


Figure 15: **Density-readout ablations on QMugs, resolved over electron count.** Same QM9-trained checkpoints, evaluation protocol, binning and aggregation as Fig. 14, on the larger QMugs molecules (Fig. 9), which reach $N_{\text{elec}} \approx 650$ against ≈ 270 on QM40. GROOT$_{\text{rot}}$ is the only arm whose error panels stay flat over the whole range and it leads every error panel, by up to three orders of magnitude on the energies; together with the Löwdin Δ-readout it is also the only arm that stays below the reference guess in ERIC, whereas GROOT$_{\text{proj}}$ rises above it from $N_{\text{elec}} \approx 320$ on and climbs towards the cap on the largest molecules. The corpus is the 950-molecule `qmugs_ebins200_cheap50` set of table 27.

N.1.2 HAMILTONIAN-PREDICTION PLOTS ↑toc

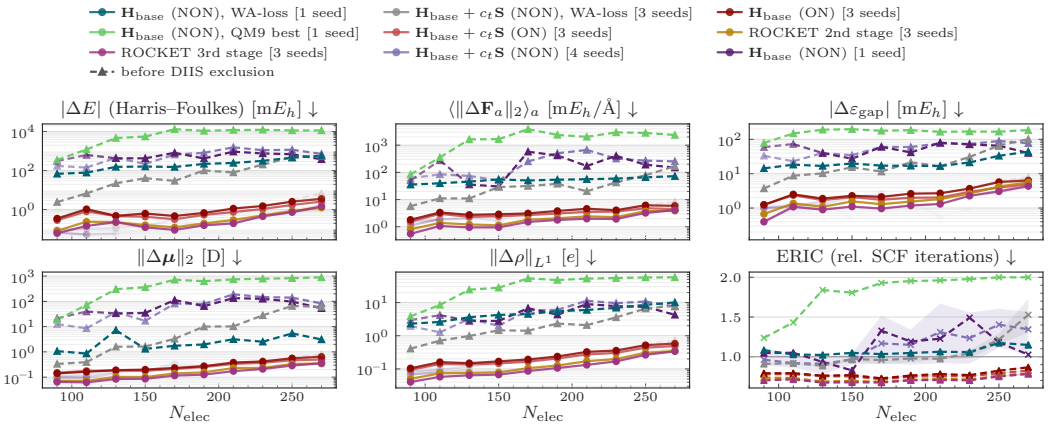


Figure 16: **Hamiltonian-readout ablations on QM40, resolved over electron count.** Same evaluation protocol, binning and aggregation as Fig. 14, for the $\Delta\mathbf{H}_{\text{SAP}}$ readouts. The arms cross the representation (AO versus Löwdin basis, Eq. (3)), the loss (WALoss versus the ROCKET objective of Eq. (9)) and the gauge treatment (none, a training-set fitted constant shift, and the EMA-centered restraint of Eq. (10)), with the occupied-plus-LUMO masked variant as the final arm. The Löwdin-basis ROCKET arms separate from the AO-basis arms by two to three orders of magnitude on every energy panel, and are the only arms that stay below the reference guess in ERIC across the whole range.

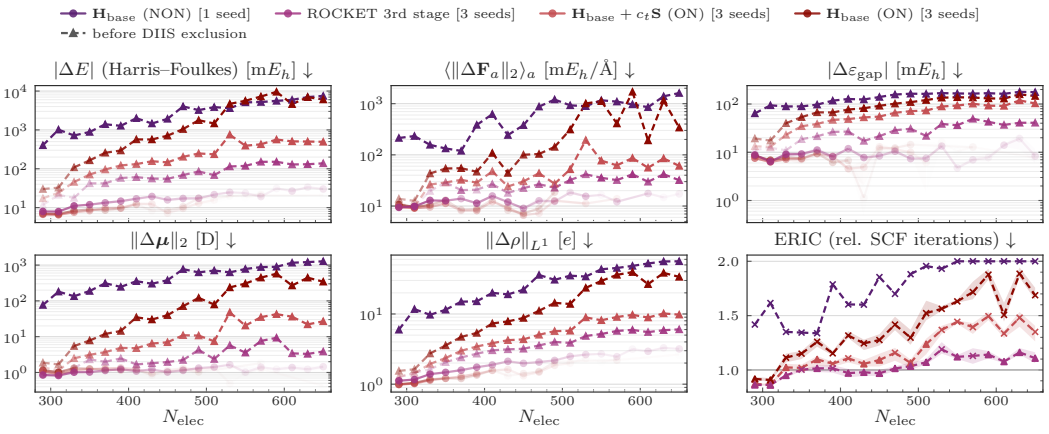


Figure 17: **Hamiltonian-readout ablations on QMugs, resolved over electron count.** Same protocol as Fig. 15, for the $\Delta\mathbf{H}_{\text{SAP}}$ readouts. The AO-basis arm sits one to two orders of magnitude above the Löwdin-basis ROCKET arms on every panel and is pinned at the ERIC cap across the range. Among the ROCKET arms the gauge-EMA, occupied-plus-LUMO masked variant leads every panel, but unlike on QM40 (Fig. 16) every Hamiltonian arm exceeds the reference guess in ERIC beyond $N_{\text{elec}} \approx 500$.

Table 26: Ablation on the QM40 test split (Table 5). Each cell is the mean over seeds of the per-molecule mean absolute error, with the range over seeds in square brackets. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed mean and the best bracketed single seed (the lower end of its range) are underlined. A prediction is excluded as an outlier when its DIIS residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS}-\mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F} = \mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). An arm that lost every molecule keeps its row, with – before each bracketed value. Excluded, Energy (standard), Energy (Harris–Foulkes), Dipole, L_1, HOMO–LUMO: 1000 molecules before the outlier exclusion. Forces, ERIC: 200 molecules before the outlier exclusion. ERIC is the number of SCF cycles from the predicted density relative to the MINAO guess, averaged over the full lane with unconverged runs pinned at the cap of 2. LR is the peak learning rate of the warmup–cosine schedule and WD the weight decay, read from each checkpoint's own training config; the seeds of a row share both. † one seed, no range.

Model	LR	WD	Excluded (%)	Standard energy (mHa) $\lvert\Delta E\rvert$	Harris–Foulkes energy (mHa) $\lvert\Delta E\rvert$	Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$	L_1 (e) $\lVert\Delta\rho\rVert_{L^1}$	Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$	HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$	ERIC (ratio)
Density-space models										
$\mathbf{P}_{\mathrm{abs.}}$ (NON)	3×10^{-3}	10^{-2}	100 [100, 100]	– (338) [245, 431]	– (11.5) [8.0, 16)	– (0.952) [0.76, 1.1)	– (0.616) [0.55, 0.67)	– (10.1) [8.5, 12)	– (6.72) [5.8, 7.7)	– (0.830) [0.82, 0.84)
$\mathbf{P}_{\mathrm{abs.}}$ (NON), no rescale†	3×10^{-3}	10^{-2}	100	– (175)	– (4.07)	– (0.594)	– (0.477)	– (8.45)	– (5.36)	– (0.821)
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100 [100, 100]	– (75.7) [61, 98)	– (2.17) [1.3, 3.8)	– (0.581) [0.50, 0.72)	– (0.446) [0.41, 0.52)	– (7.51) [6.7, 9.0)	– (5.06) [4.9, 5.3)	– (0.816) [0.81, 0.82)
$\mathbf{P}_{\mathrm{base}}$ (NON), Aufbau-purified	3×10^{-3}	10^{-2}	96.2 [96, 96]	2.90 (52.5) [2.9, 3.0] (47, 61)	0.0374 (0.981) [0.036, 0.040] (0.63, 1.5)	0.0305 (0.255) [0.030, 0.032] (0.19, 0.38)	0.0342 (0.217) [0.033, 0.037] (0.20, 0.24)	0.578 (2.53) [0.56, 0.59] (2.3, 2.8)	0.473 (1.89) [0.38, 0.56] (1.8, 1.9)	0.722 (0.758) [0.71, 0.73] (0.75, 0.76)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	75.3 [72, 80]	6.92 (45.5) [6.0, 7.6] (18, 92)	0.0581 (0.617) [0.050, 0.064] (0.42, 0.88)	0.0518 (0.118) [0.045, 0.061] (0.12, 0.12)	0.0267 (0.102) [0.025, 0.028] (0.083, 0.13)	0.663 (1.59) [0.61, 0.70] (1.3, 1.7)	0.524 (1.25) [0.51, 0.55] (1.1, 1.6)	0.694 (0.697) [0.69, 0.70] (0.69, 0.71)
$\mathbf{P}_{\mathrm{base}}$ (ON), Aufbau-purified	3×10^{-3}	3×10^{-3}	66.0 [63, 70]	2.87 (24.5) [2.8, 2.9] (19, 33)	0.0472 (0.214) [0.041, 0.057] (0.18, 0.24)	0.0559 (0.113) [0.053, 0.061] (0.11, 0.12)	0.0364 (0.113) [0.035, 0.039] (0.091, 0.14)	0.720 (1.36) [0.58, 0.81] (1.2, 1.6)	0.455 (1.09) [0.42, 0.50] (1.0, 1.2)	0.691 (0.702) [0.69, 0.69] (0.69, 0.72)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-2}	10^{-2}	1.97 [1.8, 2.1]	1.97 (2.09) [1.9, 2.1] (2.0, 2.2)	0.0586 (0.104) [0.054, 0.063] (0.098, 0.11)	0.0866 (0.0941) [0.083, 0.088] (0.091, 0.096)	0.0612 (0.0652) [0.059, 0.063] (0.063, 0.067)	0.875 (0.889) [0.87, 0.89] (0.87, 0.91)	0.585 (0.670) [0.58, 0.60] (0.66, 0.69)	0.671 (0.672) [0.67, 0.67] (0.67, 0.67)
$\mathrm{GROOT}_{\mathrm{proj}}$	3×10^{-3}	10^{-2}	0.633 [0.30, 1.0]	1.14 (2.88) [1.1, 1.2] (1.3, 5.6)	0.0730 (0.453) [0.063, 0.082] (0.075, 1.2)	0.0951 (0.104) [0.092, 0.097] (0.094, 0.12)	0.0636 (0.0692) [0.062, 0.065] (0.063, 0.078)	0.902 (1.25) [0.88, 0.92] (0.88, 1.9)	0.742 (0.831) [0.70, 0.78] (0.74, 1.0)	0.666 (0.668) [0.66, 0.67] (0.66, 0.67)
Hamiltonian-space models										
$\mathbf{H}_{\mathrm{base}}$ (NON)†	3×10^{-5}	10^{-3}	92.4	4.18 (1034)	0.0607 (584)	0.100 (71.9)	0.0748 (5.23)	1.50 (231)	1.07 (55.7)	0.779 (1.13)
$\mathbf{H}_{\mathrm{base}}$ (NON), QM9 best†	10^{-3}	10^{-3}	99.0	4.91 (9227)	1.04 (8390)	0.251 (537)	0.246 (38.8)	5.24 (2020)	3.56 (166)	0.881 (1.81)
$\mathbf{H}_{\mathrm{base}}$ (NON), WALoss†	10^{-4}	10^{-3}	100	– (2404)	– (246)	– (2.85)	– (5.43)	– (53.5)	– (21.8)	– (1.07)
$\mathbf{H}_{\mathrm{base}}$ (ON)	3×10^{-5}	3×10^{-3}	2.40 [1.9, 2.8]	3.35 (3.82) [3.2, 3.5] (3.6, 4.0)	1.22 (1.73) [1.1, 1.3] (1.5, 1.8)	0.316 (0.346) [0.30, 0.33] (0.33, 0.37)	0.275 (0.300) [0.27, 0.28] (0.29, 0.31)	3.82 (4.09) [3.7, 3.9] (3.9, 4.2)	3.06 (3.44) [2.9, 3.2] (3.3, 3.6)	0.781 (0.784) [0.78, 0.78] (0.78, 0.78)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	10^{-5}	10^{-3}	87.7 [79, 99]	5.21 (1513) [4.9, 5.7] (641, 2551)	0.123 (707) [0.072, 0.18] (257, 1141)	0.145 (80.0) [0.082, 0.20] (32, 128)	0.110 (5.96) [0.072, 0.15] (2.7, 9.4)	1.83 (254) [1.3, 2.4] (94, 515)	1.30 (56.7) [0.80, 1.5] (30, 84)	0.779 (1.14) [0.76, 0.81] (0.98, 1.3)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON), WALoss	10^{-4}	10^{-3}	100 [100, 100]	– (974) [829, 1108]	– (154) [102, 181)	– (18.2) [9.8, 23)	– (2.76) [2.4, 3.1)	– (43.7) [37, 52)	– (28.6) [23, 34)	– (1.04) [1.0, 1.0)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3×10^{-5}	3×10^{-3}	0.967 [0.60, 1.5]	2.91 (3.07) [2.9, 2.9] (3.0, 3.2)	0.900 (1.09) [0.84, 0.94] (0.93, 1.3)	0.283 (0.294) [0.27, 0.29] (0.28, 0.31)	0.238 (0.247) [0.23, 0.25] (0.24, 0.25)	3.13 (3.26) [3.0, 3.2] (3.1, 3.4)	2.49 (2.70) [2.4, 2.6] (2.5, 2.9)	0.763 (0.765) [0.76, 0.76] (0.76, 0.77)
ROCKET 2nd stage	3×10^{-4}	3×10^{-3}	0.0667 [0, 0.10]	1.70 (1.71) [1.6, 1.8] (1.6, 1.8)	0.395 (0.408) [0.31, 0.51] (0.33, 0.53)	0.179 (0.179) [0.18, 0.18] (0.18, 0.18)	0.153 (0.154) [0.14, 0.17] (0.14, 0.17)	2.07 (2.07) [1.9, 2.3] (1.9, 2.3)	2.17 (2.19) [2.0, 2.4] (2.0, 2.4)	0.722 (0.722) [0.72, 0.73] (0.72, 0.73)
ROCKET 3rd stage	10^{-3}	3×10^{-3}	0.0333 [0, 0.10]	0.665 (0.671) [0.62, 0.72] (0.62, 0.73)	0.369 (0.377) [0.32, 0.43] (0.32, 0.46)	0.157 (0.157) [0.15, 0.17] (0.15, 0.17)	0.133 (0.133) [0.13, 0.14] (0.13, 0.14)	1.81 (1.81) [1.7, 1.9] (1.7, 1.9)	1.66 (1.68) [1.6, 1.8] (1.6, 1.9)	0.709 (0.709) [0.71, 0.71] (0.71, 0.71)
ROCKET 3rd stage, k=2†	10^{-3}	3×10^{-3}	0	0.667 (0.667)	0.298 (0.298)	0.148 (0.148)	0.130 (0.130)	1.75 (1.75)	1.59 (1.59)	0.706 (0.706)

N.1.3 QM40 Extrapolation Data Tables ↑ *toc*

Table 27: Ablation on the QMugs test split of the QM9-trained checkpoints (Table 6). Each cell is the mean over seeds of the per-molecule mean absolute error, with the range over seeds in square brackets. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed mean and the best bracketed single seed (the lower end of its range) are underlined. A prediction is excluded as an outlier when its DIIS residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS}-\mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F}=\mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). An arm that lost every molecule keeps its row, with – before each bracketed value. Excluded, Energy (standard), Dipole, L_1, HOMO–LUMO: 950 molecules before the outlier exclusion. Forces, ERIC: 190 molecules before the outlier exclusion. ERIC is the number of SCF cycles from the predicted density relative to the MINAO guess, averaged over the full lane with unconverged runs pinned at the cap of 2. LR is the peak learning rate of the warmup–cosine schedule and WD the weight decay, read from each checkpoint's own training config; the seeds of a row share both. † one seed, no range.

Model	LR	WD	Excluded (%)	Standard energy (mHa) $\lvert\Delta E\rvert$	Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$	L_1 (e) $\lVert\Delta\rho\rVert_{L^1}$	Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$	HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$	ERIC (ratio)
Density-space models									
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100	– (46920)	– (145)	– (12.9)	– (314)	– (76.3)	– (1.23)
			[100, 100]	(1214, 137230)	(8.3, 369)	(5.5, 25)	(69, 756)	(59, 104)	(1.1, 1.4)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	100	– (13720)	– (14.2)	– (2.03)	– (60.1)	– (10.8)	– (0.817)
			[100, 100]	(887, 31790)	(1.0, 31)	(1.3, 3.2)	(8.7, 154)	(8.5, 14)	(0.81, 0.83)
GROOT$_{\mathrm{rot}}$	10^{-2}	10^{-2}	61.5	10.4 (36.9)	0.279 (3.29)	0.329 (0.647)	2.41 (3.27)	2.17 (3.48)	0.743 (0.776)
			[53, 66]	[10, 11]	[0.27, 0.30]	[0.32, 0.34]	[2.4, 2.4]	[2.1, 2.3]	[0.74, 0.74]
				(28, 44)	(3.3, 3.3)	(0.62, 0.69)	(3.2, 3.4)	(3.5, 3.5)	(0.77, 0.79)
GROOT$_{\mathrm{proj}}$	3×10^{-3}	10^{-2}	68.8	8.28 (26318)	0.360 (957)	0.392 (66.7)	2.65 (2794)	3.55 (79.7)	0.743 (1.32)
			[59, 74]	[8.0, 8.6]	[0.34, 0.37]	[0.39, 0.40]	[2.6, 2.7]	[3.4, 3.6]	[0.74, 0.75]
				(8755, 36267)	(422, 1317)	(31, 87)	(1006, 3982)	(57, 92)	(1.1, 1.4)
Hamiltonian-space models									
$\mathbf{H}_{\mathrm{base}}$ (NON)†	3×10^{-5}	10^{-3}	100	– (4417)	– (585)	– (29.9)	– (709)	– (137)	– (1.76)
$\mathbf{H}_{\mathrm{base}}$ (ON)	3×10^{-5}	3×10^{-3}	87.3	14.7 (730)	1.16 (160)	1.18 (16.3)	10.3 (366)	8.02 (95.9)	0.883 (1.39)
			[86, 88]	[14, 15]	[1.1, 1.2]	[1.1, 1.2]	[10, 10]	[7.8, 8.2]	[0.88, 0.89]
				(529, 1000)	(114, 228)	(13, 21)	(236, 551)	(91, 103)	(1.4, 1.4)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3×10^{-5}	3×10^{-3}	82.8	14.6 (164)	1.12 (15.5)	1.23 (5.63)	9.69 (51.8)	7.62 (64.5)	0.869 (1.18)
			[81, 84]	[14, 15]	[1.1, 1.1]	[1.2, 1.3]	[9.4, 10]	[7.5, 7.8]	[0.86, 0.88]
				(128, 193)	(9.9, 21)	(4.8, 6.4)	(39, 69)	(55, 72)	(1.1, 1.2)
ROCKET 2nd stage	3×10^{-4}	3×10^{-3}	69.5	16.9 (103)	1.04 (3.43)	1.49 (3.65)	10.9 (26.5)	9.30 (33.7)	0.879 (1.03)
			[61, 79]	[15, 18]	[0.99, 1.1]	[1.4, 1.6]	[11, 11]	[8.2, 10.0]	[0.87, 0.89]
				(58, 176)	(2.3, 4.5)	(3.0, 4.2)	(22, 31)	(26, 40)	(0.98, 1.1)
ROCKET 3rd stage	10^{-3}	3×10^{-3}	64.9	15.0 (152)	1.08 (3.22)	1.71 (3.76)	12.1 (27.0)	8.96 (28.1)	0.891 (1.04)
			[57, 69]	[14, 16]	[1.0, 1.1]	[1.6, 1.8]	[12, 12]	[8.8, 9.1]	[0.89, 0.89]
				(63, 207)	(2.4, 4.2)	(3.4, 4.0)	(25, 29)	(24, 33)	(1.0, 1.1)
ROCKET 3rd stage, k=2†	10^{-3}	3×10^{-3}	64.0	15.3 (85.9)	1.06 (3.28)	1.75 (3.76)	12.2 (27.1)	9.10 (28.4)	0.890 (1.02)

N.2 TRAINING ON QMUGS ↑ *toc*

N.2.1 QMUGS SIZE-EXTRAPOLATION DATA TABLES ↑ *toc*

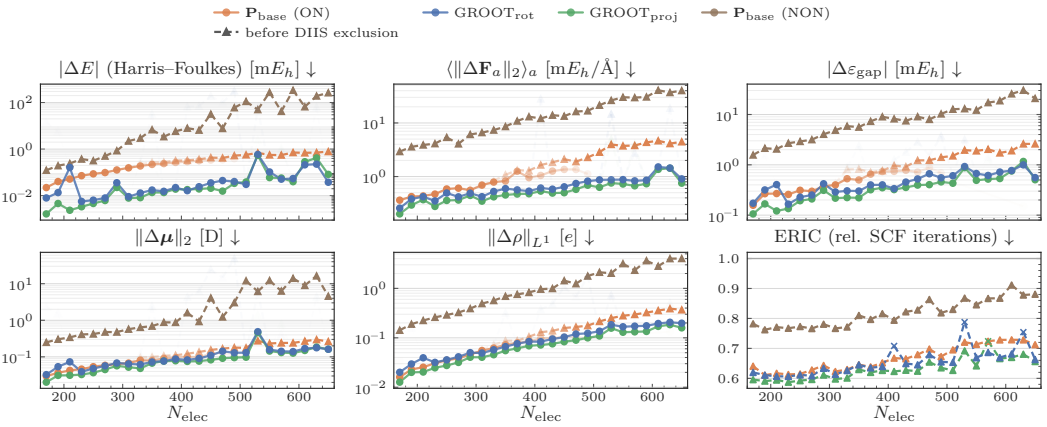


Figure 18: **Density-readout ablations on the QMugs out-of-distribution (OOD) corpus, resolved over electron count.** QMugs-trained 24-epoch checkpoints on the same corpus and protocol as table 28: 100 molecules per 20-electron bin over $N_{\mathrm{elec}} \in [160, 660)$, with forces and ERIC evaluated on the 500-molecule force subset. The arms are the $\Delta\mathbf{P}_{\mathrm{MINAO}}$ readout under $\mathcal{L}_{\mathrm{NON}}$ and under ON-Loss, and the two manifold-restricted variants $\mathrm{GROOT}_{\mathrm{rot}}$ and $\mathrm{GROOT}_{\mathrm{proj}}$ (section 4). Markers are the per-bin mean absolute error of each arm's single seed and every error panel uses a logarithmic vertical axis; table 28 reports the corresponding corpus-level means. ERIC is the effective relative iteration count (Eberhard et al., 2026a), where values below one mean fewer SCF iterations than the reference initial guess.

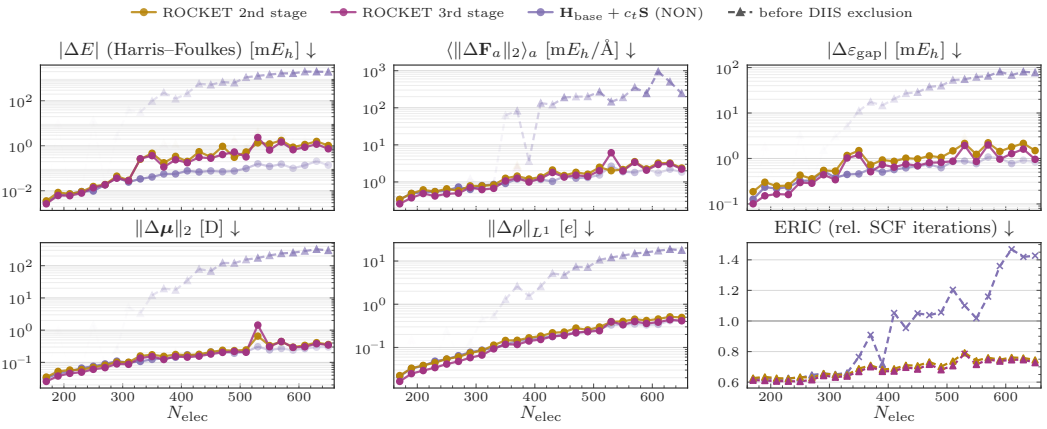


Figure 19: **Hamiltonian-readout ablations on the QMugs OOD corpus, resolved over electron count.** Same checkpoints, corpus, protocol, binning and aggregation as Fig. 18, for the $\Delta\mathbf{H}_{\mathrm{SAP}}$ readouts: the AO-basis readout with a training-set fitted constant gauge under $\mathcal{L}_{\mathrm{NON}}$, and the two ROCKET stages of Eqs. (9) and (10). Table 28 reports the corresponding corpus-level means.

Table 28: Ablation on the QMugs test split of the QMugs-trained checkpoints (Table 6). Each cell is the per-molecule mean absolute error of the arm's single seed. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed value within a training group is underlined. Rows are grouped by how the checkpoint was trained rather than by model family, and nothing is marked across groups. Within a group the rows are in this paper's conceptual arm order, which puts the density-space arms before the Hamiltonian-space ones. A prediction is excluded as an outlier when its DIIS residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS} - \mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F} = \mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). Excluded, Energy (standard), Energy (Harris–Foulkes), Dipole, L_1, HOMO–LUMO: 2500 molecules before the outlier exclusion. Forces, ERIC: 500 molecules before the outlier exclusion. ERIC is the number of SCF cycles from the predicted density relative to the MINAO guess, averaged over the full lane with unconverged runs pinned at the cap of 2. LR is the peak learning rate of the warmup–cosine schedule and WD the weight decay, read from each checkpoint's own training config; the seeds of a row share both. One seed per arm.

Model	LR	WD	Excluded (%)	Standard energy (mHa) $\lvert\Delta E\rvert$	Harris–Foulkes energy (mHa) $\lvert\Delta E\rvert$	Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$	L_1 (e) $\lVert\Delta\rho\rVert_{L1}$	Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$	HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$	ERIC (ratio)
Supervised, 24 epochs										
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100.0	4.01 (245)	0.0142 (59.0)	0.127 (3.92)	0.0343 (1.43)	– (16.2)	0.123 (9.80)	– (0.818)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	53.2	10.1 (31.4)	0.140 (0.365)	0.0638 (<u>0.143</u>)	0.0495 (0.155)	0.696 (2.01)	0.406 (1.10)	0.628 (0.668)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-3}	10^{-2}	1.12	3.22 (9.29)	0.0712 (36.8)	0.114 (4.69)	0.101 (0.313)	0.678 (3.11)	0.481 (1.18)	0.643 (0.654)
$\mathrm{GROOT}_{\mathrm{proj}}$	10^{-3}	10^{-2}	0.360	1.73 (1.80)	0.0699 (<u>0.209</u>)	0.0980 (0.165)	0.0818 (<u>0.0884</u>)	0.573 (<u>0.879</u>)	0.391 (<u>0.491</u>)	0.626 (<u>0.631</u>)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	3×10^{-4}	10^{-3}	29.6	5.33 (963)	0.0527 (655)	0.136 (96.9)	0.143 (6.20)	1.02 (156)	0.499 (30.2)	0.667 (0.937)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (ON)	3×10^{-3}	3×10^{-3}	2.96	4.85 (10.7)	1.57 (8.25)	0.307 (0.491)	0.340 (0.424)	2.45 (3.35)	1.24 (2.40)	0.719 (0.734)
ROCKET 2nd stage	3×10^{-3}	3×10^{-3}	0.600	3.42 (3.70)	0.529 (0.941)	0.212 (0.256)	0.223 (0.237)	1.57 (1.79)	1.04 (1.12)	0.692 (0.695)
ROCKET 3rd stage	3×10^{-3}	3×10^{-3}	0.200	1.10 (<u>1.18</u>)	0.445 (0.557)	0.226 (0.229)	0.189 (0.193)	1.56 (1.61)	0.782 (0.793)	0.678 (0.679)
Supervised, continued to 72 epochs										
$\mathrm{GROOT}_{\mathrm{rot}}$	3.3×10^{-4}	10^{-2}	0.960	2.77 (8.81)	0.0796 (36.7)	0.116 (4.67)	0.108 (0.317)	0.697 (3.18)	0.494 (1.20)	0.645 (0.658)
$\mathrm{GROOT}_{\mathrm{proj}}$	3.3×10^{-4}	10^{-2}	0.200	1.52 (<u>1.59</u>)	0.0715 (<u>0.217</u>)	0.0933 (<u>0.160</u>)	0.0783 (<u>0.0848</u>)	0.547 (<u>0.853</u>)	0.390 (<u>0.487</u>)	0.623 (<u>0.628</u>)

Table 29: Ablation on the three 3BPA fixed-dihedral scans (Table 7), one block per scan; the scans are never pooled. Each cell is the per-molecule mean absolute error of the arm's single seed. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed value is underlined. The marking is per block: each block is its own corpus and nothing is pooled across them. Rows are in this paper's conceptual arm order – base, then base plus a constant gauge shift, then GROOT, then ROCKET, each non-orthonormal before orthonormal – which interleaves the two model families rather than grouping them. The energy columns have a constant offset removed: the median of the per-frame signed error, subtracted per block and per model, as every frame is the same molecule. A prediction is excluded as an outlier when its DIIS residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS} - \mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F} = \mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). An arm that lost every molecule keeps its row, with – before each bracketed value. Every column is scored on all the frames of its block before the outlier exclusion: $\beta = 120°$ 2347, $\beta = 150°$ 2350, $\beta = 180°$ 2350. LR is the peak learning rate of the warmup–cosine schedule and WD the weight decay, read from each checkpoint's own training config; the seeds of a row share both. PaiNN, NequIP and MACE are machine-learned interatomic potentials trained on the QMugs energies alone (the reference data carry no forces), on the same training split as our QMugs-trained checkpoints. They predict the energy directly, which their Harris–Foulkes column reports, and their forces are its negative gradient; they predict no density, so their other cells are empty. Their LR is the initial learning rate of their own schedules (plateau for PaiNN, cosine for NequIP and MACE). One seed per arm.

Model	LR	WD	Excluded (%)	Harris–Foulkes energy (mHa) $\lvert\Delta E\rvert$		Standard energy (mHa) $\lvert\Delta E\rvert$		Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$		L_1 (e) $\lVert\Delta\rho\rVert_{L^1}$		Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$		HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$	
$\beta = 120°$															
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100	–	(0.204)	–	(13.5)	–	(0.187)	–	(0.148)	–	(4.12)	–	(3.32)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	0.0426	0.00634	(0.00643)	4.34	(4.35)	0.108	(0.108)	0.0572	(0.0572)	1.27	(1.27)	0.995	(0.995)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	3×10^{-4}	10^{-3}	0.213	0.104	(0.604)	0.157	(1.26)	0.235	(0.285)	0.183	(0.188)	4.35	(4.54)	1.19	(1.29)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-3}	10^{-2}	0	0.0159	(0.0159)	0.0278	(0.0278)	0.139	(0.139)	0.107	(0.107)	2.33	(2.33)	0.635	(0.635)
$\mathrm{GROOT}_{\mathrm{proj}}$	10^{-3}	10^{-2}	0	0.0316	(0.0316)	0.0209	(0.0209)	0.112	(0.112)	0.0968	(0.0968)	1.95	(1.95)	0.621	(0.621)
ROCKET 2nd stage	3×10^{-3}	3×10^{-3}	0	0.125	(0.125)	0.0732	(0.0732)	0.252	(0.252)	0.229	(0.229)	4.91	(4.91)	1.42	(1.42)
ROCKET 3rd stage	3×10^{-3}	3×10^{-3}	0	0.107	(0.107)	0.0595	(0.0595)	0.222	(0.222)	0.220	(0.220)	4.65	(4.65)	1.28	(1.28)
PaiNN	5×10^{-4}	10^{-2}			(1.14)								(26.7)		
NequIP	3×10^{-3}	5×10^{-7}			(1.16)								(25.6)		
MACE	3×10^{-3}	5×10^{-7}			(1.17)								(33.1)		
$\beta = 150°$															
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100	–	(0.225)	–	(7.60)	–	(0.193)	–	(0.131)	–	(3.83)	–	(2.92)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	0	0.00833	(0.00833)	3.55	(3.55)	0.0921	(0.0921)	0.0548	(0.0548)	1.18	(1.18)	0.970	(0.970)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	3×10^{-4}	10^{-3}	0	0.136	(0.136)	0.213	(0.213)	0.213	(0.213)	0.184	(0.184)	4.30	(4.30)	1.39	(1.39)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-3}	10^{-2}	0	0.0200	(0.0200)	0.0395	(0.0395)	0.158	(0.158)	0.105	(0.105)	2.43	(2.43)	1.03	(1.03)
$\mathrm{GROOT}_{\mathrm{proj}}$	10^{-3}	10^{-2}	0	0.0208	(0.0208)	0.0223	(0.0223)	0.107	(0.107)	0.0939	(0.0939)	1.89	(1.89)	0.855	(0.855)
ROCKET 2nd stage	3×10^{-3}	3×10^{-3}	0	0.162	(0.162)	0.0973	(0.0973)	0.236	(0.236)	0.236	(0.236)	5.06	(5.06)	1.64	(1.64)
ROCKET 3rd stage	3×10^{-3}	3×10^{-3}	0	0.140	(0.140)	0.0820	(0.0820)	0.215	(0.215)	0.227	(0.227)	4.79	(4.79)	1.46	(1.46)
PaiNN	5×10^{-4}	10^{-2}			(1.37)								(24.1)		
NequIP	3×10^{-3}	5×10^{-7}			(1.63)								(26.5)		
MACE	3×10^{-3}	5×10^{-7}			(1.62)								(33.5)		
$\beta = 180°$															
$\mathbf{P}_{\mathrm{base}}$ (NON)	3×10^{-3}	10^{-2}	100	–	(0.243)	–	(8.96)	–	(0.207)	–	(0.137)	–	(4.06)	–	(3.17)
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	3×10^{-3}	0	0.00802	(0.00802)	4.76	(4.76)	0.0868	(0.0868)	0.0538	(0.0538)	1.14	(1.14)	0.970	(0.970)
$\mathbf{H}_{\mathrm{base}} + c_t\mathbf{S}$ (NON)	3×10^{-4}	10^{-3}	0	0.151	(0.151)	0.196	(0.196)	0.188	(0.188)	0.181	(0.181)	4.18	(4.18)	1.37	(1.37)
$\mathrm{GROOT}_{\mathrm{rot}}$	10^{-3}	10^{-2}	0	0.0157	(0.0157)	0.0481	(0.0481)	0.128	(0.128)	0.0965	(0.0965)	2.27	(2.27)	1.05	(1.05)
$\mathrm{GROOT}_{\mathrm{proj}}$	10^{-3}	10^{-2}	0	0.0106	(0.0106)	0.0250	(0.0250)	0.0825	(0.0825)	0.0851	(0.0851)	1.66	(1.66)	0.838	(0.838)
ROCKET 2nd stage	3×10^{-3}	3×10^{-3}	0	0.154	(0.154)	0.0940	(0.0940)	0.195	(0.195)	0.230	(0.230)	4.86	(4.86)	1.59	(1.59)
ROCKET 3rd stage	3×10^{-3}	3×10^{-3}	0	0.153	(0.153)	0.0887	(0.0887)	0.189	(0.189)	0.224	(0.224)	4.70	(4.70)	1.45	(1.45)
PaiNN	5×10^{-4}	10^{-2}			(1.19)								(24.3)		
NequIP	3×10^{-3}	5×10^{-7}			(1.88)								(25.7)		
MACE	3×10^{-3}	5×10^{-7}			(1.66)								(33.2)		

N.2.2 3BPA Dihedral-Scan Data Tables ↑toc

Table 29 transfers the same QMugs-trained checkpoints to the 3BPA fixed-dihedral scans, which probe a single molecule over a dense conformational grid instead of a corpus of distinct molecules, and therefore separate conformational from chemical generalization. The three backbone-dihedral cuts are scored as three separate corpora and never pooled, so the bold marking is per block.

Table 30: Transition1x test split (Table 8); the learning rates were selected on the validation split. The blocks are the stages of label-free self-consistency training: the lv24 QMugs-trained parents before any, then $\mathbf{P}_{\mathrm{base}}$ (ON), GROOT$_{\mathrm{rot}}$, GROOT$_{\mathrm{proj}}$ and ROCKET 3rd stage after 1, 2, 4 and 12 epochs warm-started from their parents, each epoch at the learning rate with the lowest validation self-consistency residual at that epoch (for ROCKET 3rd stage the residual without its gauge-leash term); no child uses weight decay. Each cell is the per-molecule mean absolute error of the arm's single seed. Values in round brackets average all molecules of the lane before the outlier exclusion, the same molecules for every arm. Per column, the best bracketed value is underlined. The marking is per block, and nothing is pooled across blocks. A prediction is excluded as an outlier when its DIIS residual $r_{\mathrm{DIIS}} = \sqrt{8}\,\|\mathbf{S}^{-1/2}(\mathbf{FPS}-\mathbf{SPF})\mathbf{S}^{-1/2}\|_F$, with $\mathbf{P}$ the scored density and $\mathbf{F}=\mathbf{F}[\mathbf{P}]$, has $r_{\mathrm{DIIS}}/\sqrt{N_{\mathrm{elec}}} > 30$ mHa, judged per checkpoint and molecule; every unbracketed value averages that checkpoint's kept molecules only, and Excluded is the share of the molecules it lost (not underlined). Every column is scored on all 1715 frames, in every block, before the outlier exclusion. ERIC is the number of SCF cycles from the predicted density relative to the MINAO guess, averaged over the full lane with unconverged runs pinned at the cap of 2. LR is the peak learning rate of the warmup–cosine schedule, read from each checkpoint's own training config; the seeds of a row share it. One seed per arm.

Model	LR	Excluded (%)	Harris–Foulkes energy (mHa) $\lvert\Delta E\rvert$		Standard energy (mHa) $\lvert\Delta E\rvert$		Dipole (D) $\lVert\Delta\boldsymbol{\mu}\rVert_2$	L_1 (e) $\lVert\Delta\rho\rVert_{L^1}$	Forces (mHa/Å) $\langle\lVert\Delta\mathbf{F}_a\rVert_2\rangle_a$		HOMO–LUMO (mHa) $\lvert\Delta\varepsilon_{\mathrm{gap}}\rvert$		ERIC (ratio)
Before self-consistency training (the QMugs-trained parents)													
$\mathbf{P}_{\mathrm{base}}$ (ON)	3×10^{-3}	84.7	0.0386	(3.58)	44.1	(116)	0.0416	(0.805)	0.0323	(0.269)	2.64	(21.8)	0.455 (7.99)	0.741 (0.896)
GROOT$_{\mathrm{rot}}$	10^{-3}	72.4	0.202	(63.7)	0.574	(33.7)	0.146	(4.14)	0.0713	(1.01)	4.37	(66.4)	1.49 (23.9)	0.794 (0.999)
GROOT$_{\mathrm{proj}}$	10^{-3}	64.5	0.259	(43.9)	0.511	(21.3)	0.153	(3.29)	0.0716	(0.830)	4.87	(56.9)	1.75 (20.7)	0.791 (0.978)
ROCKET 3rd stage	3×10^{-3}	58.9	0.913	(33.5)	0.783	(16.3)	0.273	(2.72)	0.156	(0.770)	8.38	(45.8)	2.93 (20.3)	0.851 (0.985)
After 1 epoch of self-consistency training														
$\mathbf{P}_{\mathrm{base}}$ (ON)	10^{-3}	51.7	0.0382	(0.152)	8.35	(14.1)	0.0487	(0.167)	0.0264	(0.0551)	2.84	(5.23)	0.719 (1.45)	0.747 (0.786)
GROOT$_{\mathrm{rot}}$	3×10^{-4}	37.0	0.0835	(6.67)	0.603	(3.87)	0.107	(0.783)	0.0516	(0.188)	2.98	(12.7)	0.935 (3.32)	0.777 (0.833)
GROOT$_{\mathrm{proj}}$	10^{-4}	15.6	0.266	(3.40)	0.527	(1.43)	0.209	(0.540)	0.0749	(0.144)	4.90	(10.4)	1.60 (3.25)	0.788 (0.817)
ROCKET 3rd stage	10^{-3}	1.40	0.448	(2.08)	0.341	(0.628)	0.306	(0.404)	0.0971	(0.116)	6.50	(7.90)	2.06 (2.28)	0.813 (0.817)
After 2 epochs of self-consistency training														
$\mathbf{P}_{\mathrm{base}}$ (ON)	10^{-3}	36.2	0.0220	(0.0939)	7.80	(11.2)	0.0514	(0.114)	0.0228	(0.0394)	2.14	(3.46)	0.608 (1.09)	0.736 (0.765)
GROOT$_{\mathrm{rot}}$	10^{-4}	35.2	0.0530	(5.64)	0.550	(3.60)	0.0880	(0.668)	0.0395	(0.158)	2.36	(10.5)	0.789 (2.73)	0.754 (0.808)
GROOT$_{\mathrm{proj}}$	10^{-4}	6.94	0.270	(2.17)	0.455	(0.918)	0.223	(0.424)	0.0720	(0.114)	4.91	(7.82)	1.73 (2.56)	0.782 (0.797)
ROCKET 3rd stage	3×10^{-4}	1.63	0.343	(2.02)	0.287	(0.584)	0.242	(0.406)	0.0776	(0.108)	5.32	(7.26)	1.85 (2.09)	0.785 (0.791)
After 4 epochs of self-consistency training														
$\mathbf{P}_{\mathrm{base}}$ (ON)	10^{-3}	22.1	0.0158	(0.0652)	6.48	(8.23)	0.0516	(0.0972)	0.0209	(0.0321)	1.90	(2.80)	0.580 (0.902)	0.733 (0.752)
GROOT$_{\mathrm{rot}}$	3×10^{-4}	19.1	0.0648	(2.73)	0.517	(2.14)	0.103	(0.490)	0.0437	(0.117)	2.72	(7.83)	0.860 (1.91)	0.763 (0.794)
GROOT$_{\mathrm{proj}}$	3×10^{-4}	1.92	0.221	(0.896)	0.345	(0.487)	0.215	(0.283)	0.0651	(0.0799)	4.63	(5.53)	1.28 (1.46)	0.773 (0.777)
ROCKET 3rd stage	3×10^{-4}	1.17	0.289	(1.99)	0.207	(0.488)	0.228	(0.356)	0.0674	(0.0906)	4.79	(6.22)	1.45 (1.59)	0.770 (0.774)
After 12 epochs of self-consistency training														
$\mathbf{P}_{\mathrm{base}}$ (ON)	10^{-3}	8.75	0.0138	(0.0485)	5.00	(5.82)	0.0366	(0.0582)	0.0145	(0.0196)	1.29	(1.65)	0.368 (0.527)	0.707 (0.717)
GROOT$_{\mathrm{rot}}$	10^{-3}	8.98	0.0426	(1.79)	0.401	(1.27)	0.0859	(0.366)	0.0312	(0.0812)	2.06	(5.42)	0.706 (1.22)	0.728 (0.748)
GROOT$_{\mathrm{proj}}$	10^{-3}	0.525	0.104	(0.837)	0.202	(0.311)	0.141	(0.198)	0.0402	(0.0518)	2.87	(3.44)	0.889 (0.971)	0.730 (0.732)
ROCKET 3rd stage	10^{-3}	0.233	0.141	(0.649)	0.0951	(0.171)	0.153	(0.196)	0.0434	(0.0518)	3.09	(3.51)	1.01 (1.03)	0.730 (0.732)

N.2.3 Self-Consistency-Training Data Tables ↑ *toc*

Table 30 reports the self-consistency-trained arms on the Transition1x test split. Each arm is warm-started from the QMugs-trained parent of table 28, so the comparison isolates what the label-free objective adds on a chemistry the parents never saw. The arms are scored before training and after 1, 2, 4 and 12 epochs, each at the learning rate selected on the validation split by the self-consistency residual alone (table 22).

O RUNTIME ↑toc

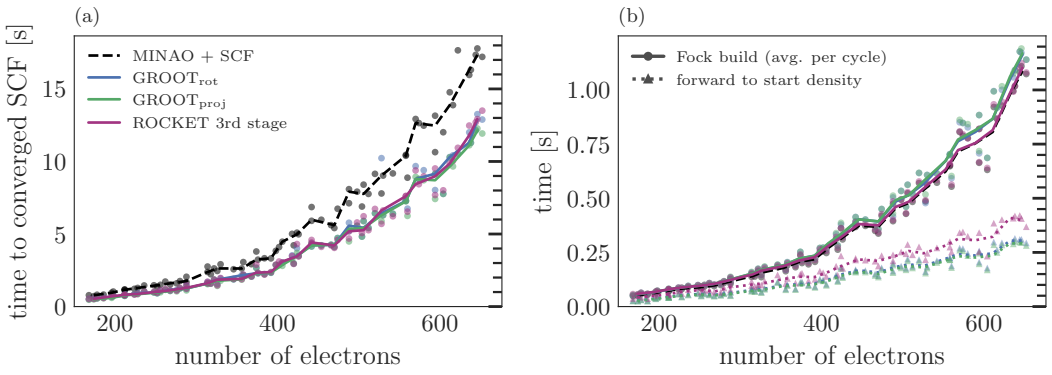


Figure 20: **Wall time to a converged SCF from ML and MINAO starts on out-of-distribution QMugs molecules. (a)** Time from the prepared molecule to a converged SCF against the electron count, for the MINAO baseline (dashed) and for SCFs started from the predictions of GROOT$_{rot}$, GROOT$_{proj}$ and ROCKET (third stage). The timed region covers the initial guess, the model input, the forward pass and the SCF; only the per-molecule system setup is excluded. **(b)** Solid lines: mean cost of one SCF cycle, i.e. SCF time divided by cycle count, dominated by the Fock build and including DIIS and diagonalization. Dotted lines: time from the forward pass to the starting density, including the copy to the CPU and, for ROCKET, the aufbau density of the predicted Hamiltonian; the GROOT arms start from their predicted density directly. Points are single molecules; lines join per-bin medians in (a) and per-bin means in (b) over 20-electron bins, without fitted curves. The set holds three random QMugs molecules per 20-electron bin in $[160, 660)$, 75 in total, all beyond the 150-electron training range, computed with the reference settings of table 2. Each molecule runs on one NVIDIA H200. One discarded warm-up round absorbs compilation and the grid and density-fitting setup; the four pipelines then alternate over 10 timed rounds, and each molecule contributes its minimum. All runs converge, and the ML-started SCFs reach the baseline total energy to within $3 \times 10^{-11}\ E_h$. Relative to MINAO, the median time to convergence is 0.70 for both GROOT arms and 0.72 for ROCKET.