Machine Learning Hamiltonians are Accurate Energy-Force Predictors

Seongsu Kim[1] **Chanhui Lee**[2] **Yoonho Kim**[1] **Seongjun Yun**[2] **Honghui Kim**[1] **Nayoung Kim**[1]
Changyoung Park[3] **Sehui Han**[3] **Sungbin Lim**[†3 4] **Sungsoo Ahn**[†1]

Abstract

Recently, machine learning Hamiltonian (MLH) models have gained traction as fast approximations of electronic structures such as orbitals and electron densities, while also enabling direct evaluation of energies and forces from their predictions. However, despite their physical grounding, existing Hamiltonian models are evaluated mainly by reconstruction metrics, leaving it unclear how well they perform as energy–force predictors. We address this gap with a benchmark that computes energies and forces directly from predicted Hamiltonians. Within this framework, we propose QHFlow2, a state-of-the-art Hamiltonian model with an SO(2)-equivariant backbone and a two-stage edge update. QHFlow2 achieves 40% lower Hamiltonian error than the previous best model with fewer parameters. Under direct evaluation on MD17/rMD17, it is the first Hamiltonian model to reach NequIP-level force accuracy while achieving up to $20\times$ lower energy MAE. On QH9, QHFlow2 reduces energy error by up to $20\times$ compared to MACE. Finally, we demonstrate that QHFlow2 exhibits consistent scaling behavior with respect to model capacity and data, and that improvements in Hamiltonian accuracy effectively translate into more accurate energy and force computations.

1. Introduction

Recently developed machine learning Hamiltonians (MLH; Schütt et al., 2019; Unke et al., 2021a; Li et al., 2022; Gong et al., 2023; Yu et al., 2023b;a; Li et al., 2025; Luo et al., 2025; Xia et al., 2025) predict the Kohn-Sham Hamiltonian from molecular geometry. Unlike machine-learning interatomic potentials (MLIPs; Unke et al., 2021b), they

predict an electronic-structure object that provides access to quantities such as electron density while also enabling downstream evaluation of energies and forces.

However, prior work has focused primarily on using predicted Hamiltonians to accelerate self-consistent field (SCF) convergence (Kim et al., 2025; Liu et al., 2025), leaving it unclear how accurately MLH models perform when energies and forces are computed directly from the predicted Hamiltonian. This question is particularly pressing given a recent study reporting that even strong Hamiltonian models yield far less accurate energy predictions than MLIP approaches (Kaniselvan et al., 2025).

Motivated by this gap, we re-examine the capabilities of MLH under direct evaluation and ask:

> *"How far have MLH models progressed, and can they achieve the energy–force accuracy required for practical atomistic simulations?"*

We answer this question by first establishing a benchmark that directly evaluates energies and forces from predicted Hamiltonians. Our analysis reveals that existing Hamiltonian models do not meet the MLIP level of accuracy, which serves as a practical reference for downstream applications.

To close this gap, we propose QHFlow2, which improves both scalability and robustness. First, we redesign the equivariant architecture for scalability by adapting an SO(2)-equivariant backbone based on eSEN (Fu et al., 2025), whose efficient edge updates are well suited for modeling orbital interactions. Second, we introduce a two-stage pair update that improves the robustness of the off-diagonal Hamiltonian blocks to cutoff and radial-basis choices. In addition, we extend a standard Hamiltonian benchmark to directly evaluate energies and forces from predicted Hamiltonians, facilitating controlled comparisons with MLIP baselines. Finally, we analyze the scaling behavior of Hamiltonian error and downstream energy and force accuracy with respect to model size and training-set size.

Under the extended benchmark, we show that sufficiently accurate Hamiltonian prediction yields accurate energies

Preprint. February 11, 2026.

Source code: https://github.com/seongsukim-ml/QHFlow2.

and forces under direct evaluation. On MD17 and rMD17, QHFlow2 achieves up to $20\times$ lower energy MAE than NequIP (Batzner et al., 2022), while reaching force accuracy comparable to MLIPs for the first time among Hamiltonian predictors. On QH9, QHFlow2 reduces energy error by up to $20\times$ relative to MACE (Batatia et al., 2022) and improves upon EquiformerV2 (Liao et al., 2024). Notably, QHFlow2 further reduces Hamiltonian prediction error by 40–50% relative to the prior state of the art while using roughly half the parameters and achieving $2.8\times$ faster inference.

Finally, we demonstrate that QHFlow2 exhibits consistent scaling behavior with respect to model capacity and data, and that improvements in Hamiltonian accuracy effectively translate to downstream energy and force predictions. Together, these results establish Hamiltonian models as a viable approach for atomistic modeling, combining accurate energies, competitive forces, and access to electronic-structure objects within a single framework.

Overall, our contributions are as follows:

- We propose QHFlow2, a scalable MLH that modernizes an eSEN-based SO(2)-equivariant backbone for Hamiltonian prediction and introduces a two-stage edge update, improving the robustness of cutoff and radial-basis choices while achieving the best accuracy with fewer parameters.

- We establish a unified benchmark that directly computes total energies and analytic forces from predicted Hamiltonians, and we construct an rMD17 benchmark with recomputed Hamiltonians, energies, and forces to facilitate controlled comparisons with MLIP baselines.

- We study model and data scaling in Hamiltonian prediction under direct evaluation, showing that increased scale consistently reduces Hamiltonian error and improves downstream energy and force accuracy.

2. Related Work

Machine learning Hamiltonians (MLHs). Deep learning approaches predict Kohn-Sham Hamiltonians from molecular geometries using equivariant message passing with orbital-based matrix representations (Schütt et al., 2019; Unke et al., 2021a; Li et al., 2022; Gong et al., 2023). Later work improves scalability through more efficient equivariant operations and training objectives (Yu et al., 2023a;b; Li et al., 2025; Luo et al., 2025; Xia et al., 2025). Beyond regression, Kim et al. (2025) introduces a generative formulation that treats Hamiltonians as structured objects.

Existing studies mainly evaluate Hamiltonian models using matrix and orbital-energy metrics, while some evaluate their utility within DFT workflows, such as improving SCF convergence (Liu et al., 2025). More recently, Kaniselvan et al.

(2025) used Hamiltonian models as a pre-training objective to transfer representations to downstream energy and force models via multi-head predictors. In contrast, our work evaluates Hamiltonian predictors through direct downstream accuracy, reporting energies and analytic forces computed from predicted Hamiltonians under a unified benchmark.

Machine learning interatomic potentials (MLIPs). MLIPs regress energies and forces directly from molecular configurations, typically using message passing over local atomic graphs and permutation-invariant aggregation (Deringer et al., 2019; Unke et al., 2021b). Directional and angular message passing improve the modeling of anisotropic interactions (Gasteiger et al., 2020; 2021), while group theory-based equivariant updates further enhance accuracy across molecular benchmarks (Batzner et al., 2022; Liao & Smidt, 2023; Batatia et al., 2022). Overall, most MLIPs focus on predicting and evaluating energy and force. In contrast, Hamiltonian models predict an electronic-structure object, thereby enabling the computation of energies, forces, and other observables simultaneously.

Equivariant architectures and efficiency. Rotational equivariance is a central inductive bias in molecular modeling and is widely used in both MLIPs and Hamiltonian prediction. State-of-the-art approaches often stem from SO(3)-equivariant updates with group-theoretic features and symmetry-preserving tensor operations (Thomas et al., 2018; Brandstetter et al., 2021; Geiger & Smidt, 2022), but the training cost can grow rapidly with higher-order angular features. Recent architectures mitigate this overhead by performing equivariant updates in local frames and exploiting SO(2) structure (Passaro & Zitnick, 2023; Liao et al., 2024; Fu et al., 2025). Building on this line of work, we adopt an SO(2) backbone to improve scalability, in addition to enhancing accuracy in Hamiltonian prediction.

3. Method

3.1. Overview: Machine Learning Hamiltonians

In this section, we overview machine learning Hamiltonian (MLH) models and how predicted Hamiltonians can be used to evaluate electronic-structure quantities, energies, and forces. We refer to Section A for a tutorial on Kohn-Sham (KS) density functional theory (DFT).

In KS DFT (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), the KS Hamiltonian $\mathbf{H} \in \mathbb{R}^{B \times B}$ is the central quantity that determines the ground-state electronic structure, where B is the number of basis functions under a fixed DFT setup. We propose QHFlow2, which predicts $\mathbf{H}$ directly from molecular geometry $\mathcal{M}$ and then evaluates the density matrix $\mathbf{D} \in \mathbb{R}^{B \times B}$, which uniquely determines the electron density ρ, as well as the KS total energy E_{KS} and atomic

forces $\mathbf{F}$ from the prediction. Our pipeline follows:

$$\mathcal{M} \to \hat{\mathbf{H}} \to \hat{\mathbf{D}} \to (\hat{E}_{\mathrm{KS}}, \hat{\mathbf{F}}), \tag{1}$$

where $\hat{\mathbf{H}}$, $\hat{\mathbf{D}}$, $\hat{E}_{\mathrm{KS}}$, and $\hat{\mathbf{F}}$ denote the predicted Hamiltonian, density matrix, energy, and forces, respectively.

Electronic structure computation. The Kohn-Sham Hamiltonian provides a direct interface to downstream electronic structure quantities through the Roothaan–Hall equation (Roothaan, 1951; Hall, 1951):

$$\hat{\mathbf{H}}\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\epsilon}. \tag{2}$$

where $\mathbf{S} \in \mathbb{R}^{B \times B}$ is the overlap matrix computed for the molecule $\mathcal{M}$ under the fixed DFT setup, $\mathbf{C} \in \mathbb{R}^{B \times n}$ contains the molecular-orbital coefficients for n orbitals, and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times n}$ is the diagonal matrix of the orbital energies.

Energies and forces evaluation from $\hat{\mathbf{H}}$. Given a predicted Hamiltonian $\hat{\mathbf{H}}$ and the overlap matrix $\mathbf{S}$, we solve for the orbital coefficients $\mathbf{C}$ satisfying Equation (2) under the same DFT setup with the reference dataset. Assuming a restricted closed-shell Kohn-Sham (RKS; Szabo & Ostlund, 1996) setting, we assign two electrons to every occupied orbital following eigenvalue ordering. With N_e electrons, we define the occupation vector $\mathbf{o} \in \mathbb{R}^B$ with

$$o_p = \begin{cases} 2, & p \leq \lceil N_e/2 \rceil, \\ 0, & p > \lceil N_e/2 \rceil, \end{cases} \tag{3}$$

where o_p is the p-th element of the occupation vector $\mathbf{o}$. This computes the approximation of the density matrix $\mathbf{D} = \mathbf{C}\operatorname{diag}(\mathbf{o})\mathbf{C}^\top \in \mathbb{R}^{B \times B}$. The density matrix $\mathbf{D}$ uniquely determines the electron density ρ. We evaluating the KS energy functional $E_{\mathrm{KS}}[\rho]$ and the corresponding forces as analytic gradients $\mathbf{F}_i = -\nabla_{\mathbf{r}_i} E_{\mathrm{KS}}[\rho]$. Further calculation details are provided in Section A.1.

3.2. Learning Hamiltonians with Flow Matching

We train our model using equivariant flow matching (Lipman et al., 2023; Song et al., 2023; Kim et al., 2025). Our model learns to map the molecule $\mathcal{M}$ to the target Hamiltonian matrix $\mathbf{H}$ via the time-dependent vector field v_t^θ, which satisfies rotational equivariance. To this end, we draw an initial noisy $\mathbf{H}_0$ from a rotation invariant distribution and sample intermediate $\mathbf{H}_t$ where $t \sim \mathcal{U}(0,1)$.

$$\mathbf{H}_t = (1-t)\mathbf{H}_0 + t\mathbf{H}. \tag{4}$$

Then the model predicts a terminal Hamiltonian $\mathbf{H}_1^\theta$ as

$$\mathbf{H}_1^\theta = f_\theta\left(\mathcal{M}, t, \mathbf{H}_t; \{\mathbf{M}_k\}\right), \tag{5}$$

where $\{\mathbf{M}_k\}$ is an auxiliary matrix feature, such as the overlap matrix or the initial Hamiltonian, which can be

easily computed from $\mathcal{M}$. This induces the vector field

$$v_t^\theta(\mathbf{H}_t, \mathcal{M}) = \frac{\mathbf{H}_1^\theta - \mathbf{H}_t}{1-t}, \tag{6}$$

which enables distributional probability path modeling of Hamiltonians. Corresponding training minimizes the conditional flow-matching objective

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}\left[\left\| v_t^\theta(\mathbf{H}_t, \mathcal{M}) - \frac{\mathbf{H} - \mathbf{H}_0}{1-t} \right\|^2\right], \tag{7}$$

encouraging v_t^θ to match the oracle velocity along the interpolation path. Further details about group and representation theory, invariant prior, and the training and inference algorithms, are in Sections B, C.2 and C.4 respectively.

3.3. QHFlow2 Architecture

We introduce **QHFlow2**, an MLH model that improves scalability by combining an SO(2)-equivariant message-passing backbone with an explicit two-stage update of pairwise features for Hamiltonian construction while preserving an SO(3)-equivariant feature structure. Figure 1 provides an overview of the method. Background on the rotational equivariance of Hamiltonian matrices and the associated representation theory is given in Section B. Model inputs and parameterization follow Equation (5).

Embeddings. We initialize atom-wise node features from the molecule $\mathcal{M}$, the flow time t, and the intermediate Hamiltonian state $\mathbf{H}_t$. The atomic number Z_i is embedded in an SO(3)-invariant feature $\mathbf{a}_i = g_{\mathrm{atom}}(Z_i)$, and the t is embedded using a sinusoidal encoding $\boldsymbol{\tau}_t = g_{\mathrm{time}}(t)$.

The intermediate Hamiltonian $\mathbf{H}_t$ is embedded in an SO(3)-equivariant irreducible representation (irrep) feature using a matrix encoder:

$$\mathbf{f}_H = g_{\mathrm{matrix}}^{(H)}(\mathbf{H}_t) = \{\mathbf{f}_H[\ell]\}_{\ell=0}^{\ell_{\max}}, \tag{8}$$

where $\mathbf{f}_H[\ell] \in \mathbb{R}^{2\ell+1}$ denotes the irreps component of degree ℓ. The additional conditioning matrices $\{\mathbf{M}_k\}$ are embedded in the same manner, i.e., $\mathbf{f}_k = g_{\mathrm{matrix}}^{(k)}(\mathbf{M}_k) = \{\mathbf{f}_k[\ell]\}_{\ell=0}^{\ell_{\max}}$.

The initial node feature is then constructed by combining only the invariant ($\ell = 0$) components,

$$\begin{aligned} \mathbf{x}_i^{(0)}[0] &= g_{\mathrm{mix}}\left(\mathbf{a}_i, \boldsymbol{\tau}_t, \mathbf{f}_H[0], \{\mathbf{f}_k[0]\}\right), \\ \mathbf{x}_i^{(0)}[\ell] &= \mathbf{f}_H[\ell] + \sum_k \mathbf{f}_k[\ell], \qquad \ell > 0. \end{aligned} \tag{9}$$

Here, g_{mix} produces atom-dependent scalar features via an MLP. All ($\ell > 0$) components are initialized by summing the matrix embeddings while preserving equivariance.
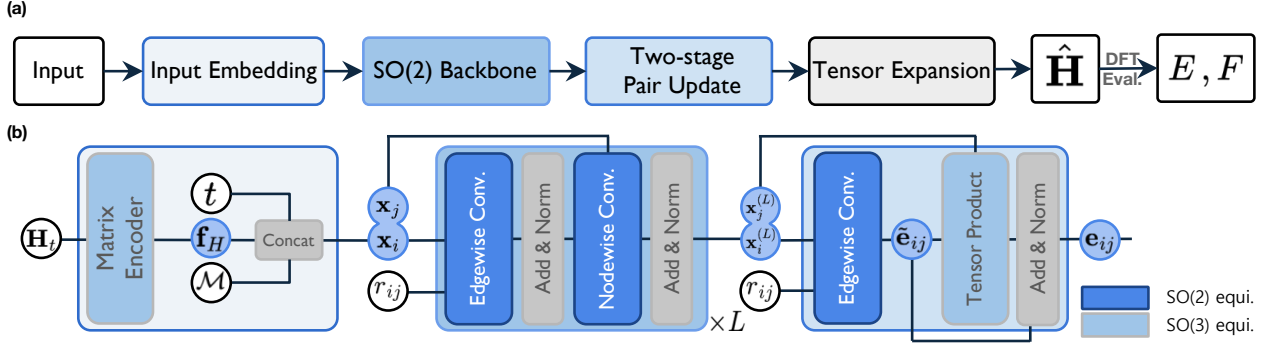
Figure 1. **QHFlow2 overall workflow.** Given the molecular structure $\mathcal{M}$, flow time t, and an intermediate Hamiltonian state $\mathbf{H}_t$, QHFlow2 applies an SO(2) backbone, a two-stage pairwise update, and construct Hamiltonian via tensor expansion for energy and force evaluation. *(a)* Pipeline overview. *(b)* Detailed architecture of the matrix encode, the SO(2) backbone, and the two-stage pair update.

SO(2)-equivariant message-passing backbone. We adopt the SO(2)-equivariant edgewise–nodewise update scheme of eSEN (Fu et al., 2025). The backbone takes the initial node embeddings $\mathbf{x}_i^{(0)}$ as input and stacks L layers, each alternating an SO(2)-equivariant edgewise convolution and a nodewise update.

At layer l, we construct an SO(2)-equivariant edgewise message on the fly for each directed pair (i, j) using relative geometry. With i-th atomic position $\mathbf{r}_i$ and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ and $r_{ij} = \|\mathbf{r}_{ij}\|$, define the neighborhood $\mathcal{N}(i) = \{\, j \neq i \mid r_{ij} < r_c \,\}$ with cutoff r_c. The distance r_{ij} is encoded by a radial basis expansion

$$\phi_{\mathrm{rbf}}(r_{ij}) = \big[\, b_1(r_{ij}), \ldots, b_M(r_{ij}) \,\big], \qquad (10)$$

where $\{b_m\}_{m=1}^{M}$ are smooth radial basis functions (e.g., Gaussian or Bessel bases) modulated by a smooth cutoff envelope (Gasteiger et al., 2020).

An edgewise message is computed as

$$\mathbf{m}_{ij}^{(l)} = \mathrm{SO2Conv}_{\mathrm{pair}}^{(l)}\Big([\mathbf{x}_i^{(l)}\|\mathbf{x}_j^{(l)}], \, \phi_{\mathrm{rbf}}(r_{ij})\Big), \qquad (11)$$

where $[\cdot\|\cdot]$ denotes concatenation. This is the SO(2)-equivariant edgewise convolution of eSEN (Fu et al., 2025); it operates in a local SO(2) frame while maintaining the node features in an SO(3)-irrep decomposition.

The edgewise messages are then aggregated over neighbors and used to update node features via a nodewise equivariant feed-forward block.

$$\mathbf{x}_i^{(l+1)} = \mathbf{x}_i^{(l)} + \mathrm{SO2Conv}_{\mathrm{self}}^{(l)}\left(\mathbf{x}_i^{(l)}, \, \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}^{(l)}\right). \quad (12)$$

Here, $\mathrm{SO2Conv}_{\mathrm{self}}^{(l)}$ is the nodewise feed-forward block of eSEN, applied independently to each SO(3) irrep component, thereby preserving the SO(3)-equivariant feature structure. Stacking L yields the final node features $\{\mathbf{x}_i^{(L)}\}$.

Two-stage pair update. The Hamiltonian decomposes into on-atom terms (diagonal blocks) and inter-atomic couplings (off-diagonal blocks). To model these couplings explicitly, we construct pairwise features $\mathbf{e}_{ij}$ for atom pairs (i, j) from the final node features $\{\mathbf{x}_i^{(L)}\}$ produced by the backbone. Unlike the backbone, which performs local message passing within a cutoff, this module forms pair features for all atom pairs to parameterize off-diagonal blocks.

We first initialize an intermediate pair feature using an SO(2)-equivariant pair encoder conditioned on a radial distance embedding,

$$\tilde{\mathbf{e}}_{ij} = \mathrm{SO2Conv}_{\mathrm{pair}}\Big([\mathbf{x}_i^{(L)}\|\mathbf{x}_j^{(L)}], \, \phi_{\mathrm{rbf}}(r_{ij})\Big), \qquad (13)$$

where $\phi_{\mathrm{rbf}}(r_{ij})$ denotes the radial basis embedding. For consistency, we use the same cutoff and radial basis configuration as in the backbone for this initialization.

In practice, we find that relying solely on this radial initialization can make Hamiltonian prediction sensitive to the choice of radial-basis hyperparameters. To reduce this dependence, we apply an additional refinement that injects pairwise interactions through an SO(3)-equivariant tensor-product coupling:

$$\mathbf{e}_{ij} = \tilde{\mathbf{e}}_{ij} + g_{\mathrm{ref}}\Big(\tilde{\mathbf{e}}_{ij}, \, \mathbf{x}_i^{(L)}, \, \mathbf{x}_j^{(L)}, \, r_{ij}\Big), \qquad (14)$$

where g_{ref} is an SO(3)-equivariant tensor-product layer that refines $\tilde{\mathbf{e}}_{ij}$ using node irreps and pair geometry. Implementation details are provided in Section C, and ablations are reported in Table 10.

Equivariant vector-to-matrix Hamiltonian construction. The backbone produces final node features $\{\mathbf{x}_i^{(L)}\}$ and explicit pairwise features $\{\mathbf{e}_{ij}\}$, which we convert into the predicted Hamiltonian $\hat{\mathbf{H}}$ in an atom-centered orbital basis. Let B_i denote the number of basis functions (orbitals) associated with atom i; in our setting, B_i is determined by

the atomic species Z_i. We construct $\hat{\mathbf{H}}$ as a block matrix $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_{ij}]_{i,j=1}^N$ with $\hat{\mathbf{H}}_{ij} \in \mathbb{R}^{B_i \times B_j}$.

For each atom pair (i, j), we first select an equivariant feature to represent their interaction,

$$\mathbf{q}_{ij} = \begin{cases} \mathbf{x}_i^{(L)}, & i = j, \\ \mathbf{e}_{ij}, & i \neq j, \end{cases}$$

and map it to the orbital-pair block via a learnable tensor expansion module adapted from QHNet (Yu et al., 2023b),

$$\tilde{\mathbf{H}}_{ij} = \mathcal{E}_{Z_i,Z_j}(\mathbf{q}_{ij}), \quad \mathcal{E}_{Z_i,Z_j} : \{\mathbf{q}[\ell]\}_{\ell=0}^{\ell_{\max}} \to \mathbb{R}^{B_i \times B_j}.$$

The expansion $\mathcal{E}_{Z_i,Z_j}$ is an SO(3)-equivariant linear readout that converts irreps components into the matrix block of the corresponding atom-type pair; details are provided in Section C. Finally, we enforce Hermiticity by symmetrization:

$$\hat{\mathbf{H}} = \frac{1}{2}(\tilde{\mathbf{H}} + \tilde{\mathbf{H}}^\top). \tag{15}$$

4. Experiments

In this section, we evaluate QHFlow2 under a direct downstream benchmark for Hamiltonian prediction. We also construct an rMD17 benchmark by re-computing reference Hamiltonians that are consistent with the provided energy and force labels. Our evaluation computes total energies and analytic forces from predicted Hamiltonians using PySCF (Sun et al., 2018), enabling controlled comparisons with MLIP baselines on molecular dynamics benchmarks. We report downstream energy and force accuracy (Figures 2 and 3), Hamiltonian prediction accuracy (Tables 1 and 2 and Figure 4), scaling behavior of MLH models (Figures 5 and 6), and runtime analysis (Figure 7).

4.1. Experimental Setup

Benchmarks and labels. We evaluate two benchmarks: an *MD benchmark* that combines MD17 and rMD17, and *QH9* (Chmiela et al., 2017; Schütt et al., 2019; Christensen & Von Lilienfeld, 2020; Ruddigkeit et al., 2012; Ramakrishnan et al., 2015; Yu et al., 2023a). QH9 is derived from QM9 and provides DFT labels, including Hamiltonians and orbital quantities. For QH9, we use an in-distribution split aligned with the QM9 molecule set (*QH9-stable-id*) for MLIP reference comparisons and additional out-of-distribution and dynamic splits to assess generalization in Hamiltonian prediction. MD benchmark labels are computed with PBE/def2-SVP and QH9 labels with B3LYP/def2-SVP. For rMD17, we compute reference Hamiltonians as well as downstream energies and forces, and we will release the benchmark upon publication. Full dataset construction and split definitions are provided in Section E.1.

Training and comparison. On the MD benchmark, we retrain both MLIP baselines and Hamiltonian predictors using the same train/validation/test splits and the same downstream evaluation pipeline. On QH9, all Hamiltonian prediction methods are trained and evaluated on identical splits and reference Hamiltonians. For MLIP comparisons on QH9, we report published QM9 results from the corresponding papers as reference points. Training configurations are provided in Sections E and F.

Models and baselines. For MLIP baselines, we compare against DimeNet, GemNet-T, and NequIP on the MD benchmark (Gasteiger et al., 2020; 2021; Batzner et al., 2022), and against MACE and Equiformer/EquiformerV2 on QH9 (Batatia et al., 2022; Liao & Smidt, 2023; Liao et al., 2024). For Hamiltonian baselines, we compare against QHNet, WANet, SPHNet, and QHFlow (Yu et al., 2023b; Li et al., 2025; Luo et al., 2025; Kim et al., 2025).

Evaluation metrics. We report Hamiltonian MAE (H), occupied orbital energy mean absolute error (MAE) (ϵ_{occ}), coefficient similarity (S_c), energy and forces MAE (E,F). For QH9, we additionally report HOMO, LUMO, and the HOMO–LUMO gap. For downstream evaluation, MLIPs predict energies and forces directly, whereas Hamiltonian predictors output $\hat{\mathbf{H}}$ from which energies and forces are computed. Detailed metric definitions and implementation details are provided in Section E.2.

4.2. Energy and Force Accuracy of MLH

MD benchmark. Figure 2 reports energy and force MAE under direct evaluation from predicted Hamiltonians on the MD benchmark. Overall, Hamiltonian-based evaluation provides competitive energy accuracy on MD trajectories, while force accuracy has been more challenging for prior Hamiltonian predictors. Within this setting, QHFlow2 substantially reduces energy error across all six systems and, for the first time among Hamiltonian predictors, reaches force accuracy comparable to NequIP under this benchmark. We connect these downstream trends to Hamiltonian reconstruction and orbital-level errors in Section 4.3.

QH9 benchmark. Figure 3 reports total-energy and orbital-energy errors on QH9. We train and evaluate all Hamiltonian predictors on the QH9-Stable-id split, which shares the same set of molecules as QM9 but uses fewer training molecules. We compare against MLIP reference results reported on QM9, including MACE, Equiformer, and EquiformerV2, as well as Hamiltonian prediction baselines. QHFlow2 achieves strong accuracy across all reported quantities, including total energy, HOMO, LUMO, and the HOMO-LUMO gap. We report exact values, including force errors, for Hamiltonian predictors in Table 2. MLIP force
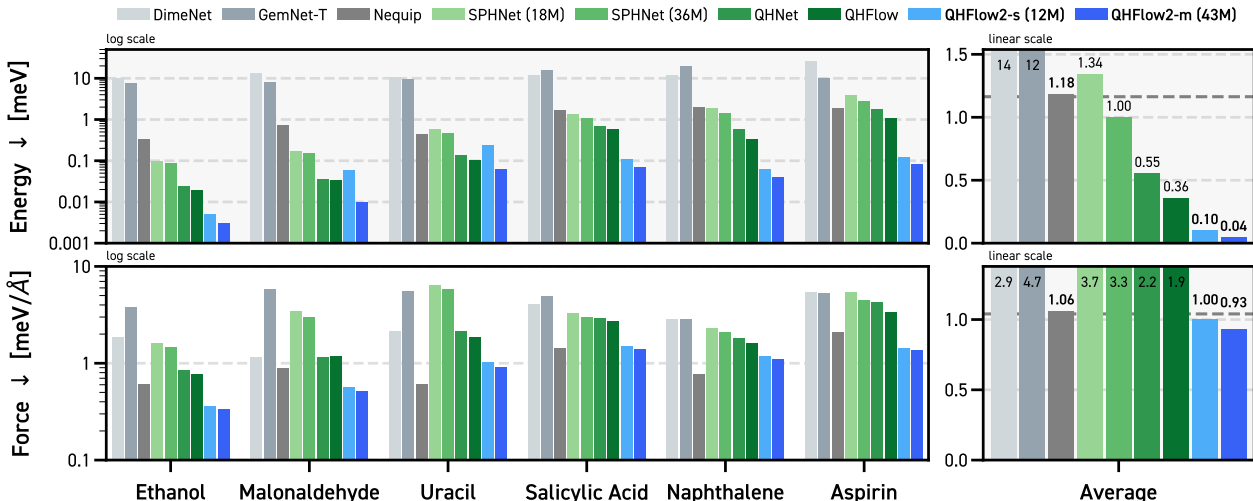
Figure 2. **Energy and force accuracy under direct evaluation on the MD benchmark.** We report mean absolute errors (MAE) of total energy (top) and forces (bottom) computed from predicted Hamiltonians on six molecular systems. gray bars denote MLIP baselines, green bars denote prior Hamiltonian predictors, and blue bars denote QHFlow2. All methods use the same data splits and evaluation setup. Overall, the results show that Hamiltonian-based direct evaluation yields competitive energy and force accuracy on MD trajectories, with prior Hamiltonian models narrowing the gap to MLIPs and QHFlow2 providing the strongest improvements. The right-hand panels summarize mean MAE across the six systems, and the dashed line marks the NequIP reference. QHFlow2 attains low energy error and, for the first time among Hamiltonian predictors, reaches NequIP-level force accuracy. Each number is reported in Table 14.
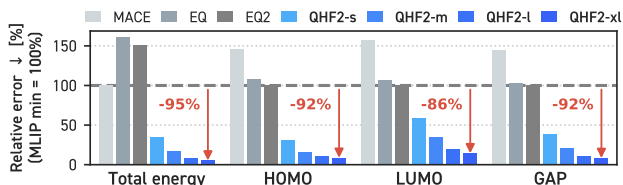

Figure 3. **Relative performance of MLIP baselines and QH-Flow2 on QH9.** Relative errors for total energy and orbital quantities are normalized by the best MLIP result of baselines for each quantity. EQ and EQ2 denote Equiformer and EquiformerV2, respectively. MLIP numbers are taken from published results on QM9 as reference points, whereas QHFlow2 is trained and evaluated on QH9-stable-id. All exact values are reported in Section G.
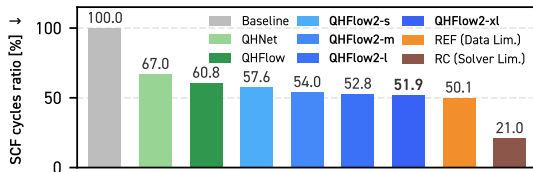

Figure 4. **Relative reduction of SCF cycles on QH9.** We report the SCF cycle ratio (lower is better): SCF iterations to converge when initializing with a predicted Hamiltonian $\hat{\mathbf{H}}$, normalized by the MinAO baseline (100%). All SCF/DFT settings are fixed; only the initialization is changed. Stronger predictors reduce SCF cycles and approach REF, which initializes with the dataset Hamiltonian $\mathbf{H}$ (data limit), while RC initializes with a converged Hamiltonian recomputed under our evaluation settings (solver limit).

comparisons are omitted because the corresponding QM9 references do not report force metrics. We further analyze scaling with model and data size in Figure 6.

4.3. Hamiltonian Prediction Accuracy

MD benchmark. Table 1 compares QHFlow2 with prior Hamiltonian models on the MD benchmark. We report Hamiltonian MAE (H), occupied orbital energy MAE (ϵ_{occ}), and coefficient similarity (S_c), together with downstream energy and force MAE (E, F) computed from each predicted Hamiltonian. QHFlow2 improves both Hamiltonian accuracy and orbital-level metrics, and these improvements coincide with lower energy errors while preserving competitive force accuracy. This trend suggests that reducing Hamiltonian errors and improving the accuracy of the

occupied orbital quantities are both associated with better downstream accuracy on MD trajectories.

QH9 benchmark. Table 2 compares Hamiltonian and orbital-energy metrics on QH9, along with energy and force errors. We report both the QH9-stable and QH9-dynamic subsets, with detailed splits depicted in Section E.1. We observe that QHFlow2 performs strongly across Hamiltonian and orbital metrics, and that these gains are accompanied by improved downstream accuracy.

Predicted Hamiltonians as SCF initial guesses on QH9. Although SCF acceleration is not a primary goal, we use it to assess the utility of predicted Hamiltonians as initial guesses in iterative KS-DFT. We run SCF on 300 molecules from the QH9-stable-iid test split using a fixed solver and

Table 1. **Hamiltonian prediction model performance on the MD benchmark.** Comparison of Hamiltonian prediction methods on six molecular systems. Reported metrics include Hamiltonian MAE H (μHa), occupied orbital energy MAE ϵ_{occ} (μHa), coefficient similarity S_c (%), and derived total-energy and force MAE (E in meV; F in meV/Å) computed from each predicted Hamiltonian. QHFlow2 rows are highlighted with skyblue , and the best result in each column is shown in **bold**. Size suffixes -s/-m denote increasing model capacity.

		Ethanol (9 atoms)					Malonaldehyde (9 atoms)					Uracil (12 atoms)				
Model	**Param.**	$H\downarrow$	$\epsilon_{occ}\downarrow$	$S_c\uparrow$	Energy$\downarrow$	Force$\downarrow$	$H\downarrow$	$\epsilon_{occ}\downarrow$	$S_c\uparrow$	Energy$\downarrow$	Force$\downarrow$	$H\downarrow$	$\epsilon_{occ}\downarrow$	$S_c\uparrow$	Energy$\downarrow$	Force$\downarrow$
Regression																
SPHNet	18M	11.39	30.29	99.99	0.097	1.609	8.81	25.47	99.99	0.166	3.430	11.42	34.63	99.99	0.576	6.399
SPHNet	36M	9.97	26.74	99.99	0.086	1.469	7.55	22.14	99.99	0.155	2.989	9.66	29.52	99.98	0.454	5.846
QHNet	27M	6.57	36.53	**100.00**	0.024	0.854	4.52	23.90	99.99	0.035	1.145	4.71	39.85	99.97	0.139	2.139
Flow-based																
QHFlow	30M	5.60	31.16	**100.00**	0.019	0.775	4.26	23.61	99.98	0.034	1.166	3.99	34.31	99.99	0.106	1.869
QHFlow2-s	14M	3.21	14.51	**100.00**	0.005	0.355	2.66	11.32	**100.00**	0.058	0.568	2.52	16.91	99.99	0.234	1.012
QHFlow2-m	43M	**2.63**	**13.56**	**100.00**	**0.003**	**0.338**	**2.03**	**10.33**	**100.00**	**0.010**	**0.514**	**1.85**	**15.88**	**100.00**	**0.062**	**0.901**

Model	**Param.**	Salicylic Acid (16 atoms)					Naphthalene (19 atoms)					Aspirin (21 atoms)				
Regression																
SPHNet	18M	8.17	54.72	99.82	1.337	3.264	6.09	37.98	98.83	1.905	2.315	10.74	98.05	99.51	3.972	5.347
SPHNet	36M	7.25	49.39	99.84	1.106	2.960	5.36	36.30	98.90	1.418	2.076	8.87	80.98	99.61	2.773	4.456
QHNet	27M	5.45	58.66	99.81	0.701	2.909	3.79	40.78	98.95	0.589	1.816	6.42	86.92	99.61	1.828	4.239
Flow-based																
QHFlow	30M	4.76	54.10	99.83	0.589	2.701	3.22	36.07	99.04	0.335	1.610	4.99	71.50	99.70	1.071	3.357
QHFlow2-s	14M	2.60	31.61	**99.91**	0.110	1.482	2.42	**28.30**	**99.28**	0.063	1.171	2.33	32.98	99.88	0.124	1.423
QHFlow2-m	43M	**2.05**	**30.31**	**99.91**	**0.070**	**1.381**	**2.11**	28.74	99.27	**0.040**	**1.106**	**1.68**	**31.14**	**99.90**	**0.084**	**1.350**

convergence criteria in Section E, and report the SCF-cycle ratio relative to the default MinAO initialization. As shown in Figure 4, stronger predictors consistently reduce SCF cycles. We include two references: *REF*, which initializes SCF with the dataset Hamiltonian **H**, and *RC*, which initializes SCF with a Hamiltonian that has converged under our evaluation settings. We use *REF* as a practical upper bound in our evaluation stack and report how closely each model approaches it under identical SCF settings.

4.4. Scaling Behavior

Data scaling on the MD benchmark. Figure 5 reports energy and force MAE (E, F) as a function of training set size, together with Hamiltonian MAE (H) and occupied orbital energy MAE (ϵ_{occ}). As the training set increases, QHFlow2 exhibits consistent reductions in Hamiltonian and orbital errors, accompanied by corresponding improvements in energy and force accuracy under direct evaluation. Compared to representative MLIP baselines, QHFlow2 improves energy accuracy rapidly with additional data while maintaining competitive force accuracy at larger training sizes. These trends support the view that increased data can be effectively translated into improved electronic-structure fidelity and downstream physical accuracy on MD trajectories.

Model parameter scaling on the QH9 benchmark. We next vary the parameter count of QHFlow2 using size variants (-s/-m/-l/-xl) while keeping the training and evaluation pipeline fixed. Figure 6 summarizes how Hamiltonian MAE and occupied orbital energy MAE change with model ca-
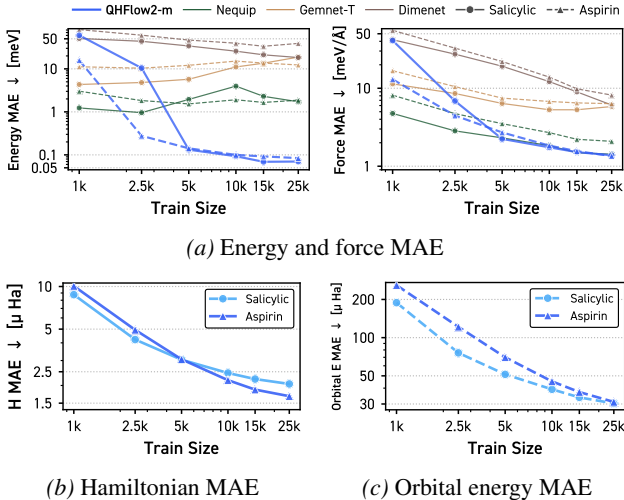

(a) Energy and force MAE

(b) Hamiltonian MAE

(c) Orbital energy MAE

Figure 5. **Data scaling on the MD benchmark.** *(a)* Energy and force MAE versus training set size for QHFlow2-m and MLIP baselines on salicylic acid and aspirin. *(b)* Hamiltonian MAE versus training set size for QHFlow2-m. *(c)* Occupied orbital energy MAE versus training set size for QHFlow2-m. Improved Hamiltonian accuracy consistently accompanies improved orbital, energy, and force accuracy.

pacity on QH9 and includes a direct comparison to QHFlow under the same protocol. Increasing capacity consistently improves both metrics, with QHFlow2 achieving lower errors than QHFlow at comparable parameter budgets. These results show that increased model capacity yields consistent gains in both Hamiltonian and orbital accuracy, while QHFlow2 remains more parameter-efficient than QHFlow across the explored regime.

Table 2. **Hamiltonian prediction model performance on QH9 benchmark.** Comparison of Hamiltonian prediction methods across QH9-Stable (id/ood) and QH9-Dynamic (geo/mol) splits. Reported metrics include Hamiltonian MAE H (μHa), occupied orbital energy MAE ϵ_{occ} (μHa), coefficient similarity S_c (%), and orbital errors for HOMO, LUMO, and the HOMO–LUMO gap (all in μHa), as well as total-energy and force MAE (E in meV; F in meV/Å) computed from each predicted Hamiltonian. QHFlow2 rows are highlighted with skyblue , and the best result in each column is shown in **bold**. Size suffixes -s/-m/-l/-xl denote increasing model capacity.

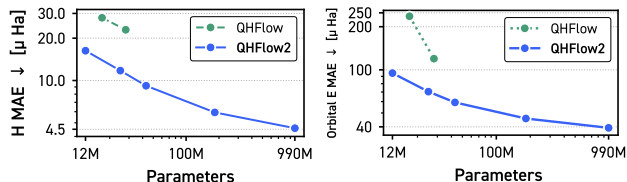|		QH9-stable-id								QH9-stable-ood							
Model	**Param.**	$H(\downarrow)$	$\epsilon_{\text{occ}}(\downarrow)$	$S_c(\uparrow)$	HOMO$(\downarrow)$	LUMO$(\downarrow)$	GAP$(\downarrow)$	E$(\downarrow)$	F$(\downarrow)$	$H(\downarrow)$	$\epsilon_{\text{occ}}(\downarrow)$	$S_c(\uparrow)$	HOMO$(\downarrow)$	LUMO$(\downarrow)$	GAP$(\downarrow)$	E$(\downarrow)$	F$(\downarrow)$
Regression																	
QHNet (w/o init)	20M	77.7	963.5	94.8	1546.3	18257.3	17822.6	434.63	2868.3	69.7	885.0	93.0	1046.0	25848.8	25370.1	427.22	66.1
QHNet	20M	46.0	459.6	98.3	470.6	2781.2	2967.9	36.92	16.3	38.1	170.5	97.8	255.6	1326.8	1359.9	13.18	6.3
WANet	29M	80.0	833.6	96.9	–	–	–	–	–	–	–	–	–	–	–	–	–
SPHNet	30M	45.5	334.3	97.8	–	–	–	–	–	43.3	186.4	98.2	–	–	–	–	–
Flow-based																	
QHFlow	28M	23.0	119.7	99.5	179.5	438.0	553.9	2.37	4.7	20.0	84.5	99.0	130.7	321.2	395.8	8.78	3.3
QHFlow2-s	12M	16.3	94.9	99.6	148.4	304.7	411.4	1.42	3.1	12.5	54.6	99.4	72.8	182.9	200.2	0.97	1.6
QHFlow2-m	43M	9.2	59.2	99.7	77.7	175.7	216.0	0.69	2.1	7.8	42.1	99.5	49.5	134.8	147.3	0.51	1.7
QHFlow2-l	183M	5.9	45.7	99.8	50.5	103.3	119.6	0.32	1.4	5.2	31.9	99.7	33.7	99.0	105.9	0.30	0.8
QHFlow2-xl	990M	**4.6**	**39.4**	**99.8**	**40.3**	**73.2**	**87.7**	**0.21**	**1.2**	**4.1**	**26.8**	**99.7**	**28.8**	**97.8**	**105.7**	**0.20**	**0.7**
		QH9-dynamic-geo								QH9-dynamic-mol							
Regression																	
QHNet (w/o init)	20M	88.4	1170.5	93.6	2040.1	23269.4	22408.0	448.63	119.3	121.4	5554.4	86.0	4352.8	53505.1	50424.9	841.74	206.5
WANet	29M	74.7	416.6	99.7	–	–	–	–	–	–	–	–	–	–	–	–	–
SPHNet	30M	52.2	100.9	99.1	–	–	–	–	–	108.2	1724.1	91.5	–	–	–	–	–
Flow-based																	
QHFlow	28M	25.9	103.1	99.6	175.2	425.2	547.3	4.28	5.2	45.9	442.6	98.7	479.7	1344.7	1605.0	8.93	11.4
QHFlow2-s	12M	15.4	53.8	99.7	67.3	430.1	447.5	1.34	2.9	29.7	261.3	99.2	263.2	644.0	772.9	4.29	9.0
QHFlow2-m	43M	8.7	42.5	99.9	51.8	113.5	144.0	0.51	1.7	22.0	240.1	99.4	207.2	378.8	468.6	2.51	8.1
QHFlow2-l	183M	5.2	28.2	99.9	30.6	88.8	101.7	0.22	1.1	17.5	214.5	99.5	179.2	261.7	325.0	1.85	7.0
QHFlow2-xl	990M	**3.7**	**22.8**	**100.0**	**21.0**	**44.4**	**50.6**	**0.15**	**0.9**	**15.9**	**202.3**	**99.6**	**167.4**	**219.5**	**278.1**	**1.66**	**6.5**

Figure 6. **Scaling behavior with model size.** (right) Hamiltonian MAE and (left) Occupied orbital MAE as a function of the number of parameters (log scale).

Figure 7. **Inference speed and memory scaling on QH9.** (left) inference speed (samples/s) versus parameter count. (right) peak inference GPU memory (GB) versus parameter count.

4.5. Runtime Analysis

We measure inference speed and peak GPU memory on QH9 under a shared evaluation setup. Inference times refer to the forward pass that outputs the predicted Hamiltonian $\hat{H}$ on an NVIDIA RTX A6000 (48GB) with a batch size of 32, excluding data loading and I/O. Figure 7 summarizes how throughput and memory scale with the parameter count. For flow-based models, we report one-step inference. Across model sizes, QHFlow2 achieves a consistently better speed-memory trade-off than prior Hamiltonian models, delivering higher throughput at comparable memory.

5. Conclusion and Limitations

We evaluated machine-learning Hamiltonians (MLH) via direct downstream energy and force calculations from predicted Hamiltonians and introduced a unified pipeline with recomputed rMD17 Hamiltonians for controlled comparisons to MLIP baselines. Within this framework, we pro-
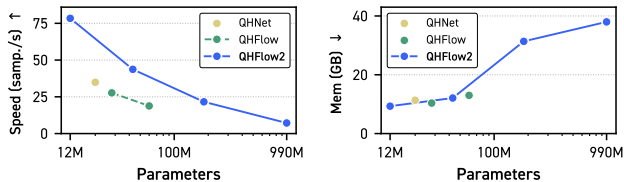
posed QHFlow2, a scalable MLH model that combines an SO(2) backbone with a two-stage pair update that reduces cutoff sensitivity. Across datasets, improved Hamiltonian accuracy consistently translated into lower downstream energy and force errors, and we characterized scaling trends with respect to model capacity and training data.

Our study has several limitations and concrete paths to address them. First, QHFlow2 is trained under a fixed functional basis set, and we do not explicitly evaluate transferability across settings; this could be improved by multimodal training, explicit conditioning on DFT choices, or lightweight adaptation. Second, we focus on restricted closed-shell systems; extending to open-shell cases is a natural next step by predicting spin-resolved Hamiltonians and validating against unrestricted references. Finally, end-to-end cost can still be dominated by diagonalization and DFT evaluation for large basis sizes; complementary advances in faster Hamiltonian-to-energy/force pipelines would improve throughput beyond predictor speedups.

Acknowledgements

Impact Statement

This work advances machine learning methods for quantum chemistry and molecular modeling. It does not involve human participants, personal data, or other sensitive information, and we therefore do not anticipate direct ethical risks arising from the methodology or experiments. Nevertheless, improved computational tools for molecular simulation may enable downstream applications in areas such as drug discovery and materials design, which can carry broader societal implications. We encourage responsible use of our methods in line with applicable laws, safety considerations, and established scientific and professional guidelines.

References

Argaman, N. and Makov, G. Density functional theory: An introduction. *American Journal of Physics*, 68(1):69–79, 2000. 13

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35: 11423–11436, 2022. 2, 5

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022. 2, 5

Brandstetter, J., Hesselink, R., van der Pol, E., Bekkers, E. J., and Welling, M. Geometric and physical quantities improve e (3) equivariant message passing. In *International Conference on Learning Representations*, 2021. 2

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 5, 21

Choudhary, V., Bhatt, A., Dash, D., and Sharma, N. Dft calculations on molecular structures, homo–lumo study, reactivity descriptors and spectral analyses of newly synthesized diorganotin(iv) 2-chloridophenylacetohydroxamate

complexes. *Journal of Computational Chemistry*, 40 (27):2354–2363, 2019. doi: https://doi.org/10.1002/jcc. 26012. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.26012. 29

Christensen, A. S. and Von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4): 045018, 2020. 5, 21

Cornwell, J. F. *Group theory in physics: An introduction*, volume 1. Academic press, 1997. 15

Deringer, V. L., Caro, M. A., and Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019. 2

Fu, X., Wood, B. M., Barroso-Luque, L., Levine, D. S., Gao, M., Dzamba, M., and Zitnick, C. L. Learning smooth and expressive interatomic potentials for physical property prediction. 2025. 1, 2, 4

Fukui, K. Role of frontier orbitals in chemical reactions. *science*, 218(4574):747–754, 1982. 29

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. 2020. 2, 4, 5

Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021. 2, 5

Geerlings, P., De Proft, F., and Langenaeker, W. Conceptual density functional theory. *Chemical Reviews*, 103(5): 1793–1874, 2003. doi: 10.1021/cr990029p. 29

Geiger, M. and Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022. 2, 16, 23

Gill, P. M. Molecular integrals over gaussian basis functions. In *Advances in quantum chemistry*, volume 25, pp. 141–205. Elsevier, 1994. 16

Gong, X., Li, H., Zou, N., Xu, R., Duan, W., and Xu, Y. General framework for e (3)-equivariant neural network representation of density functional theory hamiltonian. *Nature Communications*, 14(1):2848, 2023. 1, 2

Hall, G. G. The molecular orbital theory of chemical valency viii. a method of calculating ionization potentials. *Proceedings of the Royal Society of London*, 205(1083): 541–552, Mar 1951. doi: https://doi.org/10.1098/rspa.1951.0048. 3

Hehre, W. J., Stewart, R. F., and Pople, J. A. Self-consistent molecular-orbital methods. i. use of gaussian expansions of slater-type atomic orbitals. *The Journal of Chemical Physics*, 51(6):2657–2664, 1969. 16

Hohenberg, P. and Kohn, W. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964. 2, 12

Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Reviews of modern physics*, 87 (3):897–923, 2015. 13

Kaniselvan, M., Miller, B. K., Gao, M., Nam, J., and Levine, D. S. Learning from the electronic structure of molecules across the periodic table. 2025. 1, 2

Kim, S., Kim, N., Kim, D., and Ahn, S. High-order equivariant flow matching for density functional theory hamiltonian prediction. *Advances in Neural Information Processing Systems*, 2025. 1, 2, 3, 5, 17

Koch, W. and Holthausen, M. C. *A chemist's guide to density functional theory*. John Wiley & Sons, 2015. 16

Kohn, W. and Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965. 2, 12

Kosmann-Schwarzbach, Y. et al. Groups and symmetries. *Universitext, Springer, New York*, 2010. 15

Kumar, R., Kar, R., and Maity, D. K. Accurate prediction of homo–lumo gap using dft functional and application to next-generation organic telluro[n]helicenes materials. *Journal of Computational Chemistry*, 46 (18):e70175, 2025. doi: https://doi.org/10.1002/jcc.70175. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.70175. e70175 JCC-25-0056. 29

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. 23

Li, H., Wang, Z., Zou, N., Ye, M., Xu, R., Gong, X., Duan, W., and Xu, Y. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, 2022. 1, 2

Li, Y., Xia, Z., Huang, L., Wei, X., Harshe, S., Yang, H., Luo, E., Wang, Z., Zhang, J., Liu, C., et al. Enhancing the scalability and applicability of kohn-sham hamiltonians for molecular systems. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 5

Liao, Y.-L. and Smidt, T. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. 2023. 2, 5

Liao, Y.-L., Wood, B., Das, A., and Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. 2024. 2, 5

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3

Liu, Z., Ni, Y., Pu, Z., Sun, Q., Liu, S., and Yan, W. Towards a universally transferable acceleration method for density functional theory. 2025. 1, 2

Luo, E., Wei, X., Huang, L., Li, Y., Yang, H., Wang, Z., Liu, C., Xia, Z., Zhang, J., and Shao, B. Efficient and scalable density functional theory hamiltonian prediction through adaptive sparsity. 2025. 1, 2, 5

Miar, M., Shiroudi, A., Pourshamsian, K., Oliaey, A. R., and Hatamjafari, F. Theoretical investigations on the homo–lumo gap and global reactivity descriptor studies, natural bond orbital, and nucleus-independent chemical shifts analyses of 3-phenylbenzo [d] thiazole-2 (3 h)-imine and its para-substituted derivatives: Solvent and substituent effects. *Journal of Chemical Research*, 45(1-2):147–158, 2021. 29

Mortazavi, B. Recent advances in machine learning-assisted multiscale design of energy materials. *Advanced Energy Materials*, 15(9):2403876, 2025. doi: https://doi.org/10.1002/aenm.202403876. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/aenm.202403876. 29

Mortazavi, B., Zhuang, X., Rabczuk, T., and Shapeev, A. V. Atomistic modeling of the mechanical properties: the rise of machine learning interatomic potentials. *Materials horizons*, 10(6):1956–1968, 2023. 29

Neese, F., Wennmohs, F., Becker, U., and Riplinger, C. The orca quantum chemistry program package. *The Journal of chemical physics*, 152(22), 2020. 21

Neugebauer, J. and Hickel, T. Density functional theory in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(5):438–448, 2013. 13

Passaro, S. and Zitnick, C. L. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. pp. 27420–27438. PMLR, 2023. 2

Perdew, J. P., Burke, K., and Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996. 21

Ramakrishnan, R., Hartmann, M., Tapavicza, E., and Von Lilienfeld, O. A. Electronic spectra from tddft and

machine learning in chemical space. *The Journal of chemical physics*, 143(8), 2015. 5, 21

Roothaan, C. C. J. New developments in molecular orbital theory. *Reviews of Modern Physics*, 23(2):69–89, Apr 1951. doi: https://doi.org/10.1103/revmodphys.23.69. 3

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012. 5, 21

Schrödinger, E. An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6):1049, 1926. 12

Schütt, K., Gastegger, M., Tkatchenko, A., Müller, K.-R., and Maurer, R. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1), 2019. 1, 2, 5, 21

Slater, J. C. Atomic shielding constants. *Physical review*, 36(1):57, 1930. 16

Song, Y., Gong, J., Xu, M., Cao, Z., Lan, Y., Ermon, S., Zhou, H., and Ma, W.-Y. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36: 549–568, 2023. 3

Stephens, P. J., Devlin, F. J., Chabalowski, C. F., and Frisch, M. J. 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. 21

Sun, Q., Berkelbach, T. C., Blunt, N. S., Booth, G. H., Guo, S., Li, Z., Liu, J., McClain, J. D., Sayfutyarova, E. R., Sharma, S., et al. Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018. 5, 17, 21, 23

Szabo, A. and Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, Inc., Mineola, first edition, 1996. 3

Thomas, N., Smidt, T., Kearnes, S., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 2, 16

Unke, O., Bogojeski, M., Gastegger, M., Geiger, M., Smidt, T., and Müller, K.-R. Se (3)-equivariant prediction of molecular wavefunctions and electronic densities. *Advances in Neural Information Processing Systems*, 34: 14434–14447, 2021a. 1, 2, 21

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schutt, K. T., Tkatchenko, A., and Muller, K.-R. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021b. 1, 2

Wang, G., Wang, C., Zhang, X., Li, Z., Zhou, J., and Sun, Z. Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. *iScience*, 27(5):109673, 2024. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2024.109673. URL https://www.sciencedirect.com/science/article/pii/S2589004224008952. 29

Weigend, F. and Ahlrichs, R. 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. 21

Wigner, E. *Group theory: and its application to the quantum mechanics of atomic spectra*, volume 5. Elsevier, 2012. 15

Wu, X., Sun, Q., Pu, Z., Zheng, T., Ma, W., Yan, W., Yu, X., Wu, Z., Huo, M., Li, X., Ren, W., Gong, S., Zhang, Y., and Gao, W. Enhancing gpu-acceleration in the python-based simulations of chemistry framework, 2024. URL https://arxiv.org/abs/2404.09452. 21

Xia, C. H., Kaniselvan, M., Ziogas, A. N., Mladenović, M., Mahjoub, R., Maeder, A., and Luisier, M. Learning the electronic hamiltonian of large atomic structures. 2025. 1, 2

Yu, H., Liu, M., Luo, Y., Strasser, A., Qian, X., Qian, X., and Ji, S. Qh9: A quantum hamiltonian prediction benchmark for qm9 molecules. *Advances in Neural Information Processing Systems*, 36:40487–40503, 2023a. 1, 2, 5, 21, 22

Yu, H., Xu, Z., Qian, X., Qian, X., and Ji, S. Efficient and equivariant graph networks for predicting quantum hamiltonian. pp. 40412–40424. PMLR, 2023b. 1, 2, 5, 19, 21, 23

A. Density Functional Theory

A.1. Kohn–Sham density functional theory

Density functional theory (DFT) is a central tool in quantum chemistry and materials science for approximating the electronic structure of many-body systems. Since solving the many-electron Schrödinger equation directly is computationally prohibitive for large numbers of particles (Schrödinger, 1926), DFT reformulates the ground-state problem in terms of the electron density $\rho(\mathbf{r})$. Because $\rho(\mathbf{r})$ depends only on three spatial coordinates, this representation greatly reduces the effective complexity of the problem compared to working with an N-electron wavefunction. Here, we generally follow the contents of the DFT overview books.

In practice, DFT is most commonly used through the Kohn–Sham (KS) formulation (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), which replaces the interacting electron system with an auxiliary system of non-interacting electrons constructed to reproduce the same ground-state density as the true system. The wavefunction of this reference system, the Kohn–Sham wavefunction Ψ_{KS}, must satisfy fermionic antisymmetry (Pauli exclusion), and is therefore written as a Slater determinant of N single-particle KS orbitals $\{\psi_i\}_{i=1}^{N}$:

$$\Psi_{\mathrm{KS}}(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}} \det[\psi_i(\mathbf{r}_j)]_{i,j=1}^{N} \tag{16}$$

In general, the electron density associated with an N-electron wavefunction Ψ is defined by marginalizing out $N - 1$ coordinates:

$$\rho(\mathbf{r}) = N \int \cdots \int \left| \Psi(\mathbf{r}, \mathbf{r}_2, \ldots, \mathbf{r}_N) \right|^2 d\mathbf{r}_2 \cdots d\mathbf{r}_N. \tag{17}$$

For a Slater determinant constructed from orthonormal orbitals, this reduces to a simple sum over the occupied KS orbitals:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} \left| \psi_i(\mathbf{r}) \right|^2. \tag{18}$$

Rather than explicitly manipulating the full KS wavefunction, one can compute $\rho(\mathbf{r})$ directly from the single-particle orbitals. The orbitals $\{\psi_i(\mathbf{r})\}$ and their orbital energies $\{\epsilon_i\}$ are obtained by solving the Kohn–Sham equations,

$$\mathcal{H}[\rho]\psi_i(\mathbf{r}) = \left[-\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{H}}[\rho](\mathbf{r}) + V_{\mathrm{xc}}[\rho](\mathbf{r}) \right] \psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \tag{19}$$

where the bracketed operator defines the effective KS Hamiltonian $\mathcal{H}[\rho]$. It consists of the kinetic-energy operator, the external potential V_{ext} from the nuclei, the Hartree potential $V_{\mathrm{H}}[\rho]$ describing classical electrostatic repulsion due to the electron density, and the exchange–correlation potential $V_{\mathrm{xc}}[\rho]$, which captures the remaining many-body quantum effects.

The ground-state density is then assembled from the N occupied KS orbitals as

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} \left| \psi_i(\mathbf{r}) \right|^2. \tag{20}$$

It is worth emphasizing that the KS orbitals ψ_i are auxiliary quantities: they are introduced to reproduce the correct ground-state density, rather than to represent the physical many-electron wavefunction. Moreover, because $\mathcal{H}[\rho]$ depends on ρ and ρ is constructed from the solutions $\{\psi_i\}$, the KS equations define a self-consistent problem.

Density functional theory (DFT) thus provides a tractable route to electronic structure by expressing ground-state properties in terms of $\rho(\mathbf{r})$ (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). Within KS-DFT, the interacting system is mapped to non-interacting orbitals $\{\psi_i\}_{i=1}^{N}$ with $\psi_i : \mathbb{R}^3 \to \mathbb{C}$, which reproduce the ground-state density $\rho(\mathbf{r}) : \mathbb{R}^3 \to \mathbb{R}_{>0}$. Since the ground-state energy can be expressed as a functional of $\rho(\mathbf{r})$, one can evaluate it without explicitly handling the full many-electron interaction in wavefunction form.

The KS orbitals and density are obtained by solving

$$\hat{H}_{\mathrm{KS}}[\rho]\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \qquad \rho(\mathbf{r}) = \sum_{i=1}^{N} \left| \psi_i(\mathbf{r}) \right|^2, \tag{21}$$

where ϵ_i denotes the orbital energies. The KS Hamiltonian operator $\hat{H}_{\mathrm{KS}}[\rho]$ depends explicitly on the electron density ρ, which is in turn constructed from the orbitals.

Energy and force evaluation in KS-DFT. Following standard DFT texts (Argaman & Makov, 2000; Neugebauer & Hickel, 2013; Jones, 2015), we summarize how total energies and atomic forces are evaluated in KS-DFT. Given Kohn–Sham (KS) orbitals $\{\psi_i\}$ with occupations $\{f_i\}$, the electron density is constructed as

$$\rho(\mathbf{r}) = \sum_i f_i \, |\psi_i(\mathbf{r})|^2, \tag{22}$$

where typically $f_i \in [0, 2]$ in spin-restricted settings (e.g., $f_i \in \{0, 2\}$ for closed-shell systems).

The KS total energy is written as the sum of kinetic, external, Hartree, exchange-correlation (XC), and nuclear-repulsion contributions:

$$E_{\mathrm{KS}}[\rho] = T_s[\rho] + \int v_{\mathrm{ext}}(\mathbf{r}) \, \rho(\mathbf{r}) \, d\mathbf{r} + E_{\mathrm{H}}[\rho] + E_{\mathrm{xc}}[\rho] + E_{\mathrm{nn}}, \tag{23}$$

where $T_s[\rho]$ is the non-interacting kinetic energy of the KS reference system, v_{ext} is the electron–nuclear potential, $E_{\mathrm{H}}[\rho]$ is the classical Coulomb (Hartree) energy, $E_{\mathrm{xc}}[\rho]$ accounts for exchange and correlation effects beyond T_s and E_{H}, and E_{nn} is the nuclear repulsion energy. The Hartree term is

$$E_{\mathrm{H}}[\rho] = \frac{1}{2} \iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r} \, d\mathbf{r}', \qquad v_{\mathrm{H}}[\rho](\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, d\mathbf{r}', \tag{24}$$

and the XC potential entering the KS equations is defined by the functional derivative

$$v_{\mathrm{xc}}[\rho](\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}. \tag{25}$$

In practice, $E_{\mathrm{xc}}[\rho]$ is approximated by a chosen XC functional (e.g., LDA, GGA, meta-GGA, or hybrid).

Atomic forces are obtained as derivatives of the total energy with respect to nuclear coordinates $\{\mathbf{R}_A\}$:

$$\mathbf{F}_A = -\nabla_{\mathbf{R}_A} E_{\mathrm{KS}}[\rho]. \tag{26}$$

When atom-centered basis functions are used, analytic forces typically include the Hellmann–Feynman contribution as well as Pulay terms arising from the nuclear dependence of the basis. Additional contributions may appear when parts of the energy (notably XC terms in GGA/meta-GGA) are evaluated by numerical quadrature on atom-centered grids, since quadrature weights and partition functions can depend on $\{\mathbf{R}_A\}$.

A.2. Variational energy minimization and self-consistency

Here, we describe KS self-consistency as a variational minimization problem. The KS energy functional $E_{\mathrm{KS}}[\rho]$ and its decomposition are given in Equation (23) (with the associated XC definitions in Equation (25)). The electron density is constructed from KS orbitals as in Equation (22), and the KS equations and Hamiltonian decomposition are summarized in Equation (23).

Stationarity and the nonlinear eigenproblem. At the ground state, $E_{\mathrm{KS}}[\rho]$ is stationary with respect to variations of the orbitals under the orthonormality and particle-number constraints. This stationarity yields the KS equations (Equation (19)), where the Hamiltonian depends on the density through the Hartree and XC potentials (Equations (24) and (25)). Consequently, solving KS-DFT amounts to a *nonlinear* eigenvalue problem: the density determines the Hamiltonian, and the Hamiltonian eigenvectors determine the density.

Self-consistency as a fixed-point iteration. Self-consistent field (SCF) methods solve the nonlinear KS problem by iterating a density-to-density (or density-matrix-to-density-matrix) map until consistency is reached. Let $\rho^{(k)}$ denote the density at iteration k, and define a map $\mathcal{F}$ that returns the output density obtained by (i) building $\hat{H}_{\mathrm{KS}}[\rho^{(k)}]$ (Equation (23)), (ii) solving the KS eigenproblem (Equation (19)), and (iii) forming the density from the resulting orbitals (Equation (22)). This yields the fixed-point form

$$\rho^{(k+1)} = \mathcal{F}(\rho^{(k)}). \tag{27}$$

In practice, direct iteration can be unstable, so one applies *mixing* to damp oscillations and accelerate convergence. A common choice is linear mixing,

$$\rho^{(k+1)} = (1 - \alpha)\rho^{(k)} + \alpha\,\tilde{\rho}^{(k+1)}, \qquad \alpha \in (0, 1], \tag{28}$$

where $\tilde{\rho}^{(k+1)} = \mathcal{F}(\rho^{(k)})$ is the "output" density computed from the KS solve. More advanced schemes (e.g., Pulay/DIIS) construct $\rho^{(k+1)}$ from a linear combination of several past iterates/residuals to improve robustness.

A typical stopping criterion checks either the density change $\|\rho^{(k+1)} - \rho^{(k)}\|$ or the energy change $|E_{\mathrm{KS}}[\rho^{(k+1)}] - E_{\mathrm{KS}}[\rho^{(k)}]|$ against a tolerance. At convergence, $\rho^\star$ satisfies the self-consistency condition $\rho^\star = \mathcal{F}(\rho^\star)$ and corresponds to a stationary point of the KS functional under the KS constraints.

Matrix form with finite-basis. Most implementations expand orbitals in a finite atom-centered basis $\{\phi_\mu\}_{\mu=1}^B$:

$$\psi_i(\mathbf{r}) = \sum_{\mu=1}^{B} C_{\mu i}\,\phi_\mu(\mathbf{r}), \tag{29}$$

which turns the KS equations (Equation (19)) into a generalized eigenvalue problem,

$$\mathbf{H}[\rho]\,\mathbf{C} = \mathbf{S}\,\mathbf{C}\,\boldsymbol{\epsilon}, \tag{30}$$

where $\mathbf{S}_{\mu\nu} = \langle \phi_\mu, \phi_\nu \rangle$ is the overlap matrix, $\mathbf{H}[\rho]$ is the KS (Fock) matrix assembled from the current density (via the Hartree and XC contributions; cf. Equations (24) and (25)), $\mathbf{C}$ is the coefficient matrix, and $\boldsymbol{\epsilon}$ is the diagonal matrix of orbital energies. The density matrix is formed as

$$\mathbf{P} = \mathbf{C}\,\operatorname{diag}\mathbf{o}\,\mathbf{C}^\top, \tag{31}$$

with an occupation vector $\mathbf{o}$, it induces the real-space density (equivalent to Equation (22)). In this representation, SCF can be viewed as iterating $\mathbf{P}^{(k)} \mapsto \mathbf{H}[\mathbf{P}^{(k)}] \mapsto \mathbf{P}^{(k+1)}$ (with mixing) until convergence.

Orbital occupations. Let $\varepsilon_1 \leq \cdots \leq \varepsilon_B$ be the generalized eigenvalues. For restricted (spin-unpolarized) calculations, the number of occupied spatial orbitals is

$$n_{\mathrm{occ}} = \left\lceil \frac{N_e}{2} \right\rceil, \tag{32}$$

and the occupied subspace corresponds to the n_{occ} lowest eigenpairs. The electron density (or density matrix) is then constructed from the occupied orbitals, while higher-energy orbitals form the virtual subspace.

B. Group theory and equivariance

B.1. Group theory

We review basic group-theoretic notions needed to motivate equivariant architectures for Hamiltonian prediction, following standard references (Kosmann-Schwarzbach et al., 2010; Cornwell, 1997; Wigner, 2012).

Groups. A *group* is a pair $(G, \circ)$ where G is a non-empty set and $\circ : G \times G \to G$ is a binary operation satisfying: (i) associativity $(a \circ b) \circ c = a \circ (b \circ c)$, (ii) identity $\exists e \in G$ s.t. $e \circ a = a \circ e = a$, (iii) inverse $\forall a \in G$, $\exists a^{-1} \in G$ s.t. $a \circ a^{-1} = a^{-1} \circ a = e$, and (iv) closure $a \circ b \in G$. We write ab for $a \circ b$ when unambiguous. Many symmetry groups in physics are *Lie groups* (continuous groups) such as $SO(3)$ and $SE(3)$ (Cornwell, 1997).

Group actions. To formalize symmetries of objects, we specify how a group acts on a set. A (left) *group action* of G on a set $\mathcal{X}$ is a map

$$\cdot_{\mathcal{X}} : G \times \mathcal{X} \to \mathcal{X}, \qquad (g, x) \mapsto g \cdot_{\mathcal{X}} x, \tag{33}$$

satisfying $e \cdot_{\mathcal{X}} x = x$ and $(gh) \cdot_{\mathcal{X}} x = g \cdot_{\mathcal{X}} (h \cdot_{\mathcal{X}} x)$. When clear, we omit the subscript and write $g \cdot x$. For example, $SO(3)$ acts on $\mathbb{R}^3$ by rotations: $R \cdot \mathbf{r} = R\mathbf{r}$.

Equivariance and invariance. Let G act on $\mathcal{X}$ and $\mathcal{Y}$ via $\cdot_{\mathcal{X}}$ and $\cdot_{\mathcal{Y}}$. A map $f : \mathcal{X} \to \mathcal{Y}$ is *G-equivariant* if

$$f(g \cdot_{\mathcal{X}} x) = g \cdot_{\mathcal{Y}} f(x), \qquad \forall g \in G, \ x \in \mathcal{X}. \tag{34}$$

If the output action is trivial ($g \cdot_{\mathcal{Y}} y = y$), this reduces to *invariance*: $f(g \cdot_{\mathcal{X}} x) = f(x)$. Equivariance is the appropriate symmetry notion for tensorial quantities (e.g., vectors, matrices) whose outputs transform under the group.

Representations. When $\mathcal{X}$ is a vector space, group actions are often linear. A *(linear) representation* of G on a vector space V is a homomorphism

$$\rho : G \to \mathrm{GL}(V), \qquad \rho(gh) = \rho(g)\rho(h), \ \rho(e) = I. \tag{35}$$

In finite dimensions, choosing a basis identifies $\rho(g)$ with an invertible matrix $D(g) \in \mathbb{C}^{d \times d}$. Two representations D and D' are *equivalent* if there exists an invertible Q such that $D'(g) = Q^{-1} D(g) Q$ for all g.

Equivariance Equation (34) is most commonly expressed using representations: if $x \in V_{\mathrm{in}}$ and $f(x) \in V_{\mathrm{out}}$ with actions $D_{\mathrm{in}}(g)$ and $D_{\mathrm{out}}(g)$, then

$$f\big(D_{\mathrm{in}}(g)x\big) = D_{\mathrm{out}}(g)f(x). \tag{36}$$

B.2. Irreducible representations and $SO(3)$

Irreducible representations. A subspace $W \subseteq V$ is *invariant* under ρ if $\rho(g)W \subseteq W$ for all $g \in G$. A representation ρ is *irreducible* if its only invariant subspaces are $\{0\}$ and V. For compact groups (including $SO(3)$), any finite-dimensional unitary representation decomposes into a direct sum of irreducible representations (irreps) (Wigner, 2012).

Irreps of $SO(3)$. The irreps of $SO(3)$ are indexed by $\ell \in \mathbb{N}_0$ and have dimension $2\ell + 1$. They can be realized via Wigner D-matrices $D^{(\ell)} : SO(3) \to \mathbb{C}^{(2\ell+1) \times (2\ell+1)}$. A convenient basis is provided by spherical harmonics Y_m^ℓ on $\mathbb{S}^2$ (with $m = -\ell, \dots, \ell$), which transform under rotations according to $D^{(\ell)}$ (Wigner, 2012; Cornwell, 1997).

Any square-integrable function on the sphere can be expanded as

$$f(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} f_{\ell m} Y_m^\ell(\theta, \phi), \tag{37}$$

and under a rotation $R \in SO(3)$, the coefficient vector $\mathbf{f}^{(\ell)} = (f_{\ell,-\ell}, \dots, f_{\ell,\ell})^\top \in \mathbb{C}^{2\ell+1}$ transforms as

$$\mathbf{f}^{(\ell)} \xmapsto{R} D^{(\ell)}(R) \mathbf{f}^{(\ell)}. \tag{38}$$

This viewpoint motivates representing geometric features as collections of ℓ-typed components (scalars $\ell = 0$, vectors $\ell = 1$, etc.), each of which transforms by the corresponding irrep matrix.

Clebsch–Gordan (CG) tensor product. Given irreps $D^{(\ell_1)}$ and $D^{(\ell_2)}$, their tensor product $D^{(\ell_1)} \otimes D^{(\ell_2)}$ is generally reducible and decomposes as

$$D^{(\ell_1)} \otimes D^{(\ell_2)} \cong \bigoplus_{\ell=|\ell_1-\ell_2|}^{\ell_1+\ell_2} D^{(\ell)}. \tag{39}$$

The CG coefficients implement the change of basis from the product space to the direct-sum irrep basis. Concretely, for $\mathbf{u}^{(\ell_1)} \in \mathbb{C}^{2\ell_1+1}$ and $\mathbf{v}^{(\ell_2)} \in \mathbb{C}^{2\ell_2+1}$, the ℓ-component $\mathbf{w}^{(\ell)} \in \mathbb{C}^{2\ell+1}$ is

$$w_m^{(\ell)} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} C_{(\ell_1,m_1),(\ell_2,m_2)}^{(\ell,m)} u_{m_1}^{(\ell_1)} v_{m_2}^{(\ell_2)}. \tag{40}$$

This construction is equivariant:

$$\left(D^{(\ell_1)}(R)\mathbf{u}^{(\ell_1)}\right) \otimes_{\mathrm{CG}} \left(D^{(\ell_2)}(R)\mathbf{v}^{(\ell_2)}\right) = D^{(\ell)}(R)\mathbf{w}^{(\ell)}, \qquad \forall R \in \mathrm{SO}(3). \tag{41}$$

Equivariant neural layers (TFN-style). Many SE(3)-equivariant GNNs represent node/edge features as a direct sum of SO(3) irreps and update them using CG tensor products with spherical-harmonic filters (Thomas et al., 2018; Geiger & Smidt, 2022). Let $\mathbf{V}_i^{(\ell)} \in \mathbb{C}^{(2\ell+1)\times C_\ell}$ denote the ℓ-typed feature at node i with C_ℓ channels. Given relative direction $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/\|\mathbf{r}_{ij}\|$ and distance $r_{ij} = \|\mathbf{r}_{ij}\|$, a typical TFN filter uses a radial network $R(r_{ij})$ and spherical harmonics $Y^{(\ell_f)}(\hat{\mathbf{r}}_{ij})$:

$$\mathbf{F}^{(\ell_f)}(\mathbf{r}_{ij}) = R^{(\ell_f)}(r_{ij})\, Y^{(\ell_f)}(\hat{\mathbf{r}}_{ij}), \tag{42}$$

and messages are formed via CG products so that the output transforms as a valid irrep (cf. Equations (39) and (41)).

B.3. Symmetry and equivariance of RH-DFT Hamiltonians

Atomic orbital basis and spherical harmonics. To solve the KS equations in practice, molecular orbitals are expanded in an atom-centered basis. A prototypical atomic orbital can be written as (Koch & Holthausen, 2015)

$$\phi_{n\ell m}(\mathbf{r}) = R_{n\ell}(\|\mathbf{r}\|)\, Y_m^\ell(\theta, \phi), \tag{43}$$

where $R_{n\ell}$ is a radial function (e.g., Gaussian or Slater-type) (Gill, 1994; Slater, 1930; Hehre et al., 1969), and Y_m^ℓ is a spherical harmonic. A molecular orbital is then expressed as a linear combination of shifted atomic orbitals,

$$\psi_i(\mathbf{r}) = \sum_{\alpha=1}^{B} C_{\alpha i}\, \phi_\alpha(\mathbf{r} - \mathbf{R}_\alpha), \tag{44}$$

with coefficient matrix $\mathbf{C} \in \mathbb{R}^{B\times O}$.

Hamiltonian blocks as intertwiners. In a spherical (angular-momentum) basis, basis functions are grouped by ℓ-type. Let $\mathbf{H}^{(\ell,\ell')}$ denote the submatrix coupling the ℓ-block on one center to the ℓ'-block on another center, with $\mathbf{H}^{(\ell,\ell')} \in \mathbb{C}^{(2\ell+1)\times(2\ell'+1)}$. Under a global rotation $R \in \mathrm{SO}(3)$, the spherical basis transforms by block-diagonal Wigner matrices $\mathcal{D}(R) = \mathrm{blkdiag}\big(D^{(\ell)}(R)\big)_\ell$. Equivariance of the KS operator in this basis implies that each block obeys the conjugation rule

$$\mathbf{H}^{(\ell,\ell')} \xmapsto{R} D^{(\ell)}(R)\, \mathbf{H}^{(\ell,\ell')}\, D^{(\ell')}(R)^{-1}. \tag{45}$$

Equivalently, $\mathbf{H}^{(\ell,\ell')}$ is an *intertwiner* between the ℓ' and ℓ representation spaces. This blockwise transformation law motivates parameterizing Hamiltonian predictors so that their outputs are collections of irrep-typed blocks that satisfy Equation (45) by construction.

C. Training, Priors, and Architecture

This section summarizes the learning objectives, rotationally invariant priors, and the network architecture used in QHFlow2 to predict Kohn–Sham Hamiltonians in an atom-centered orbital basis. Throughout, $\mathcal{M} = \{(Z_i, \mathbf{r}_i)\}_{i=1}^N$ denotes a molecular configuration, $\mathbf{H} \in \mathbb{R}^{B \times B}$ the (real-valued) Hamiltonian matrix in the chosen basis, and $\mathbf{S} \in \mathbb{R}^{B \times B}$ the overlap matrix.

C.1. Traininig objective as Residual parameterization

Rather than regressing the converged Hamiltonian directly, QHFlow2 learns the *residual* between a cheap initialization and the SCF solution, following residual-learning strategies used in prior Hamiltonian predictors (e.g., WANet and SHNet). For each molecule $\mathcal{M}$, let $\mathbf{H}_{\mathcal{M}}^{(0)}$ be an initial Hamiltonian guess and $\mathbf{H}_{\mathcal{M}}^{\star}$ be the converged SCF solution. We define the residual target

$$\mathbf{H}_{1,\mathcal{M}} := \mathbf{H}_{\mathcal{M}}^{\star} - \mathbf{H}_{\mathcal{M}}^{(0)}. \tag{46}$$

In our implementation, $\mathbf{H}_{\mathcal{M}}^{(0)}$ is computed using the `minao` initialization in PySCF (Sun et al., 2018), which does not involve any SCF iterations. At inference time, we reconstruct the Hamiltonian as

$$\widehat{\mathbf{H}}_{\mathcal{M}}^{\star} = \mathbf{H}_{\mathcal{M}}^{(0)} + \widehat{\mathbf{H}}_{1,\mathcal{M}}. \tag{47}$$

For completeness, we also evaluate a direct parameterization that does not rely on an explicit initialization (i.e., treating $\mathbf{H}_{\mathcal{M}}^{(0)} = \mathbf{0}$ and predicting $\mathbf{H}_{\mathcal{M}}^{\star}$ directly), and report the corresponding results in the Table 12.

C.2. Rotationally invariant priors for $\mathbf{H}_0$

We introduce two priors p_0 for sampling the initial residual $\mathbf{H}_0$: a Gaussian orthogonal ensemble (GOE) prior and a tensor expansion-based (TE) prior (Kim et al., 2025). In all flow-based Hamiltonian prediction experiments in this work, we use the TE prior as the default choice for p_0.

GOE prior. For the GOE prior, we sample a symmetric matrix by drawing independent entries $M_{ij} \sim \mathcal{N}(0, \sigma^2)$ for $i \leq j$ and setting $M_{ji} = M_{ij}$. We use $\sigma^2 = 1.0$ for MD17 and $\sigma^2 = 0.1$ for QH9.

Tensor expansion (TE) prior. The TE prior constructs an SO(3)-invariant distribution by sampling irrep-valued latent vectors and mapping them to matrix blocks via tensor expansion. Let $\mathbf{w}^{(\ell)} \in \mathbb{R}^{2\ell+1}$ be a type-ℓ irrep vector. We sample

$$\mathbf{w}^{(\ell)} = r \, D^{(\ell)}(\mathbf{R}) \, \mathbf{w}_0^{(\ell)}, \tag{48}$$

where $\mathbf{w}_0^{(\ell)}$ is a fixed unit-norm reference vector, $\mathbf{R} \sim \mathrm{Uniform}(\mathrm{SO}(3))$, and r is sampled independently from a radial distribution. This factorization makes $p(\mathbf{w}^{(\ell)})$ SO(3)-invariant. In practice, we use $r \sim \mathrm{LogNormal}(1, 0.1)$.

Given $\mathbf{w}^{(\ell)}$, tensor expansion maps it into a matrix block of type (ℓ_1, ℓ_2) via Clebsch–Gordan coefficients:

$$\left(\bar{\otimes} \, \mathbf{w}^{(\ell)} \right)_{m_1 m_2}^{(\ell_1, \ell_2)} = \sum_{m=-\ell}^{\ell} C_{(\ell_1, m_1),(\ell_2, m_2)}^{(\ell, m)} \, w_m^{(\ell)}, \tag{49}$$

where $m_1 \in \{-\ell_1, \ldots, \ell_1\}$ and $m_2 \in \{-\ell_2, \ldots, \ell_2\}$. The resulting block transforms equivariantly as

$$\left(\bar{\otimes} \, \mathbf{w}^{(\ell)} \right)^{(\ell_1, \ell_2)} \xrightarrow{\mathbf{R}} D^{(\ell_1)}(\mathbf{R}) \left(\bar{\otimes} \, \mathbf{w}^{(\ell)} \right)^{(\ell_1, \ell_2)} D^{(\ell_2)}(\mathbf{R})^{-1}. \tag{50}$$

We use this TE prior as a rotationally invariant noise distribution for sampling $\mathbf{H}_0$ in the flow matching path Equation (4).

C.3. Training and sampling algorithms

For completeness, we summarize the training and sampling procedures in Algorithms 1 and 2. **Unless stated otherwise, we set p_0 to the TE prior throughout this work.** The only difference between the GOE and TE settings is the choice of p_0 for sampling $\mathbf{H}_0$.

C.4. Training and Sampling Algorithms of QHFlow2

For completeness, we summarize the training and sampling procedures in Algorithms 1 and 2. The only difference between the GOE and TE settings is the choice of p_0 for sampling $\mathbf{H}_0$.

Algorithm 1 QHFlow2 training procedure (residual CFM)

Require: Dataset $\{(\mathcal{M}_i, \mathbf{S}_{\mathcal{M}_i}, \mathbf{H}^{\star}_{\mathcal{M}_i}, \mathbf{H}^{(0)}_{\mathcal{M}_i})\}$, model f_θ, prior p_0
 1: **for** each training step **do**
 2:     Sample minibatch $\mathcal{B} = \{\mathcal{M}_i\}_{i=1}^{|\mathcal{B}|}$
 3:     **for** each $\mathcal{M}$ in $\mathcal{B}$ **do**
 4:         Compute residual target $\mathbf{H}_{1,\mathcal{M}} \leftarrow \mathbf{H}^{\star}_{\mathcal{M}} - \mathbf{H}^{(0)}_{\mathcal{M}}$
 5:         Sample $\mathbf{H}_0 \sim p_0$ and $t \sim \mathcal{U}(0,1)$
 6:         Interpolate $\mathbf{H}_{t,\mathcal{M}} \leftarrow (1-t)\mathbf{H}_0 + t\,\mathbf{H}_{1,\mathcal{M}}$
 7:         Predict $\mathbf{H}^{(\theta)}_{1,\mathcal{M}} \leftarrow f_\theta(\mathbf{H}_{t,\mathcal{M}}, \mathbf{S}_{\mathcal{M}}, \mathcal{M}, t)$
 8:         Accumulate loss $\mathcal{L}_{\mathcal{M}} \leftarrow \|\mathbf{H}^{(\theta)}_{1,\mathcal{M}} - \mathbf{H}_{1,\mathcal{M}}\|_F^2$
 9:     **end for**
10:     Update θ by gradient descent on $\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{\mathcal{M}\in\mathcal{B}} \mathcal{L}_{\mathcal{M}}$
11: **end for**

Algorithm 2 QHFlow2 sampling procedure (ODE discretization)

Require: Molecular configuration $\mathcal{M}$, overlap matrix $\mathbf{S}_{\mathcal{M}}$, initialization $\mathbf{H}^{(0)}_{\mathcal{M}}$, model f_θ, prior p_0, time grid $\{t_k\}_{k=0}^{K}$ with $t_0 = 0, t_K = 1$
 1: Sample $\mathbf{H}_0 \sim p_0$ and set $\mathbf{H}_{t_0} \leftarrow \mathbf{H}_0$
 2: **for** $k = 0$ to $K-1$ **do**
 3:     $\mathbf{H}^{(\theta)}_{1,\mathcal{M}} \leftarrow f_\theta(\mathbf{H}_{t_k}, \mathbf{S}_{\mathcal{M}}, \mathcal{M}, t_k)$
 4:     $v_{t_k,\theta} \leftarrow (\mathbf{H}^{(\theta)}_{1,\mathcal{M}} - \mathbf{H}_{t_k})/(1-t_k)$
 5:     $\mathbf{H}_{t_{k+1}} \leftarrow \mathbf{H}_{t_k} + (t_{k+1} - t_k)\,v_{t_k,\theta}$
 6: **end for**
 7: Output residual $\widehat{\mathbf{H}}_{1,\mathcal{M}} \leftarrow \mathbf{H}_{t_K}$
 8: Reconstruct Hamiltonian $\widehat{\mathbf{H}}^{\star}_{\mathcal{M}} \leftarrow \mathbf{H}^{(0)}_{\mathcal{M}} + \widehat{\mathbf{H}}_{1,\mathcal{M}}$

D. Model architecture

D.1. Two-stage pair refinement

QHFlow2 maintains explicit pairwise representations $\mathbf{z}_{ij}$ to improve off-diagonal Hamiltonian blocks. The pair module follows the design of the *non-diagonal pair update* in QHNet (adapted to our setting), but is implemented with $SO(2)$-equivariant operations in edge-aligned local frames. Concretely, we perform a two-stage update: (i) *initialization* from geometric/chemical edge cues, and (ii) *refinement* using the final node context. We summarize the abstraction as

$$\mathbf{z}_{ij} \leftarrow \mathrm{InitPair}(\mathbf{e}_{ij}), \qquad \mathbf{z}_{ij} \leftarrow \mathrm{UpdPair}(\mathbf{z}_{ij}, \mathbf{z}_i, \mathbf{z}_j, \mathbf{e}_{ij}), \tag{51}$$

where both stages are type-preserving (irrep-wise) and equivariant by construction.

Notation and local frames. Let $\mathbf{z}_i = \bigoplus_{\ell=0}^{\ell_{\max}} \mathbf{z}_i^{(\ell)}$ be the node irreps and $\mathbf{z}_{ij} = \bigoplus_{\ell=0}^{\ell_{\max}} \mathbf{z}_{ij}^{(\ell)}$ the pair irreps. Edge attributes $\mathbf{e}_{ij}$ include distance encodings and chemical types (and optionally a reference-frame cue). For each edge (i,j), let $\mathbf{R}_{ij} \in SO(3)$ align $\hat{\mathbf{r}}_{ij}$ with the z-axis. We evaluate pair operations in the local frame $\tilde{\mathbf{z}}_i^{(\ell)} = D^{(\ell)}(\mathbf{R}_{ij})\,\mathbf{z}_i^{(\ell)}$ (and similarly for j), where the residual symmetry reduces to an $SO(2)$ action around the axis.

Stage I: pair initialization. The initialization stage injects purely edge-level information into $\mathbf{z}_{ij}$ (before any node-context refinement). We parameterize

$$\mathbf{z}_{ij}^{(\ell)} = \Phi_{\mathrm{init}}^{(\ell)}(\mathbf{e}_{ij}), \qquad \ell = 0, \ldots, \ell_{\max}, \tag{52}$$

where $\Phi_{\text{init}}^{(\ell)}$ outputs an ℓ-type feature (with channels) using $\mathrm{SO}(2)$-equivariant edgewise operations in the local frame (followed by rotation back to the global frame). Intuitively, Φ_{init} provides a geometry-aware seed representation for each pair, which is later refined by node features.

Stage II: non-diagonal pair refinement via tensor-product filtering. The refinement stage updates $\mathbf{z}_{ij}$ using the final node features, mirroring the non-diagonal pair update of QHNet. The core idea is to form an equivariant *pair message* by (a) gating and self-interacting the node irreps, (b) computing an attention-like scalar weight per channel path, and (c) applying a Clebsch–Gordan tensor product to couple the two node features into a pair irrep.

(a) Norm-gate and self-interaction. For each ℓ_{in}, we first apply a norm gate to stabilize magnitudes and then a type-preserving self-interaction (channel mixing):

$$\bar{\mathbf{z}}_i^{(\ell_{\text{in}})} = \mathrm{GateNorm}^{(\ell_{\text{in}})}(\tilde{\mathbf{z}}_i^{(\ell_{\text{in}})}), \qquad \hat{\mathbf{z}}_i^{(\ell_{\text{in}})} = \mathbf{W}_{\text{SI}}^{(\ell_{\text{in}})} \bar{\mathbf{z}}_i^{(\ell_{\text{in}})}, \tag{53}$$

and similarly for j. Here $\mathbf{W}_{\text{SI}}^{(\ell_{\text{in}})}$ acts only on channels, hence preserves equivariance.

(b) Edge-conditioned filter and attentive weight. For each path $(\ell_{\text{in}}, \ell_{\text{in}} \to \ell_{\text{out}})$, we produce a scalar filter from the edge attributes:

$$R^{(\ell_{\text{in}}, \ell_{\text{out}})}(\mathbf{e}_{ij}) = \mathrm{MLP}_R^{(\ell_{\text{in}}, \ell_{\text{out}})}(\mathbf{e}_{ij}), \qquad a_{ij}^{(\ell_{\text{in}}, \ell_{\text{out}})} = \mathrm{MLP}_a^{(\ell_{\text{in}}, \ell_{\text{out}})}\Big(\mathbf{z}_i^{(0)}, \mathbf{z}_j^{(0)}, \mathbf{e}_{ij}\Big), \tag{54}$$

and combine them as in QHNet (Yu et al., 2023b):

$$F_{ij}^{(\ell_{\text{in}}, \ell_{\text{out}})} = a_{ij}^{(\ell_{\text{in}}, \ell_{\text{out}})} R^{(\ell_{\text{in}}, \ell_{\text{out}})}(\mathbf{e}_{ij}). \tag{55}$$

The weight F_{ij} is a scalar over irrep indices, so it does not affect equivariance.

(c) Tensor-product pair message. We then form an equivariant pair message by coupling the two node irreps through a Clebsch–Gordan tensor product (computed in the local frame):

$$\tilde{\mathbf{f}}_{ij}^{(\ell_{\text{out}})} = \sum_{\ell_{\text{in}}} \Big(F_{ij}^{(\ell_{\text{in}}, \ell_{\text{out}})} \hat{\mathbf{z}}_i^{(\ell_{\text{in}})} \otimes_{\mathrm{CG}} \hat{\mathbf{z}}_j^{(\ell_{\text{in}})} \Big)^{(\ell_{\text{out}})}. \tag{56}$$

Finally, we rotate the message back to the global frame and update the stored pair feature by combining the previous pair state $\mathbf{z}_{ij}$ and the new message $\mathbf{f}_{ij}$ with an e3nn linear tensor product:

$$\mathbf{f}_{ij}^{(\ell)} = D^{(\ell)}(\mathbf{R}_{ij})^{-1} \tilde{\mathbf{f}}_{ij}^{(\ell)}, \qquad \Delta \mathbf{z}_{ij} = \mathrm{LTP}\Big(\mathbf{z}_{ij}, \mathbf{f}_{ij}, \mathbf{e}_{ij}\Big), \qquad \mathbf{z}_{ij} \leftarrow \mathbf{z}_{ij} + \Delta \mathbf{z}_{ij}, \tag{57}$$

where $\mathrm{LTP}(\cdot)$ denotes the e3nn *linear tensor product* (LinearTensorProduct), which maps two irrep-valued inputs to an irrep-valued output via Clebsch–Gordan coupling with learnable channel weights. The edge attributes $\mathbf{e}_{ij}$ are used to produce the tensor-product weights (e.g., through an MLP), while equivariance is preserved by construction.

D.2. Tensor expansion for Hamiltonian construction.

To represent on-site and inter-site interactions in a common equivariant form, we define a unified feature for each ordered pair (i, j):

$$\mathbf{w}_{ij} = \begin{cases} \mathbf{z}_i, & i = j, \\ \mathbf{z}_{ij}, & i \neq j, \end{cases} \qquad \mathbf{w}_{ij} = \bigoplus_{\ell=0}^{\ell_{\max}} \mathbf{w}_{ij}^{(\ell)}, \tag{58}$$

where $\mathbf{z}_i$ denotes the final node feature and $\mathbf{z}_{ij}$ the final pair feature. Each component $\mathbf{w}_{ij}^{(\ell)}$ is a type-ℓ irrep feature (channels suppressed) and transforms under a global rotation $\mathbf{R} \in \mathrm{SO}(3)$ as $\mathbf{w}_{ij}^{(\ell)} \xrightarrow{\mathbf{R}} D^{(\ell)}(\mathbf{R})\mathbf{w}_{ij}^{(\ell)}$. To map these irrep features into Hamiltonian blocks consistent with orbital angular-momentum coupling, we construct for each (ℓ_1, ℓ_2) an equivariant matrix block $\mathbf{H}_{ij}^{(\ell_1, \ell_2)} \in \mathbb{R}^{(2\ell_1+1) \times (2\ell_2+1)}$ by combining tensor expansion with a learnable channel projection:

$$\mathbf{H}_{ij}^{(\ell_1, \ell_2)} = \sum_{\ell=|\ell_1-\ell_2|}^{\ell_1+\ell_2} \mathbf{W}_{ij}^{(\ell \to \ell_1, \ell_2)} \Big(\bar{\otimes} \, \mathbf{w}_{ij}^{(\ell)} \Big)^{(\ell_1, \ell_2)}, \tag{59}$$

where $\left(\bar{\otimes}\, \mathbf{w}_{ij}^{(\ell)}\right)^{(\ell_1,\ell_2)}$ is the tensor expansion defined in Equation (49) and $\mathbf{W}_{ij}^{(\ell \to \ell_1,\ell_2)}$ acts only on channel dimensions (and may depend on edge attributes), thereby preserving equivariance. By construction, each block satisfies

$$\mathbf{H}_{ij}^{(\ell_1,\ell_2)} \xrightarrow{\mathbf{R}} D^{(\ell_1)}(\mathbf{R})\, \mathbf{H}_{ij}^{(\ell_1,\ell_2)}\, D^{(\ell_2)}(\mathbf{R})^{-1}, \tag{60}$$

and assembling all blocks across (i,j) yields a Hamiltonian that obeys the global similarity transform, guaranteeing exact $\mathrm{SO}(3)$ equivariance in the atom-centered orbital basis.

E. Experimental study settings

E.1. Dataset preparation

To demonstrate the effectiveness of flow-matching-based training, we conduct experiments on two molecular datasets: MD17 and QH9. The MD17 represents a relatively simple task compared to the QH9, focusing solely on small systems and their conformational space. The PubChemQH9 is not considered since their dataset and codebase are not publicly released.

MD17. The MD17 (Chmiela et al., 2017; Schütt et al., 2019) dataset consists of quantum chemical simulations for four small organic molecules: water (H_2O), ethanol (C_2H_5OH), malondialdehyde ($CH_2(CHO)_2$), and uracil ($C_4H_4N_2O_2$). It provides a comprehensive set of molecular properties, including geometries, total energies, forces, Kohn–Sham Hamiltonian matrices, and overlap matrices. All reference computations were implemented via the ORCA electronic structure package (Neese et al., 2020) using the PBE exchange–correlation functional (Perdew et al., 1996; Weigend & Ahlrichs, 2005) and the def2-SVP Gaussian-type orbital (GTO) basis set. We follow the standard data split protocol used in prior work (Schütt et al., 2019; Unke et al., 2021a; Yu et al., 2023b) to divide each molecule's conformational data into training, validation, and test sets. The detailed dataset statistics are summarized in Table 3 and MOs in the table imply molecular orbitals (*i.e.*, s, p, d, f)

Table 3. The statistics of MD17 dataset (Schütt et al., 2019).

Dataset	# of structures	Train	Val	Test	# of atoms	# of orbitals	# of occupied MOs
Ethanol	30,000	25,000	500	4,500	9	72	10
Malondialdehyde	26,978	25,000	500	1,478	9	90	19
Uracil	30,000	25,000	500	4,500	12	132	26

rMD17. The revised MD17 (rMD17) dataset (Christensen & Von Lilienfeld, 2020) was introduced to mitigate inconsistencies in the DFT settings of the original MD17 trajectories by recomputing energies and forces at the PBE/def2-SVP level of theory with very tight SCF convergence criteria and a dense numerical integration grid (using ORCA (Neese et al., 2020)). We build on rMD17 because its reference energies and forces are reproducible across standard DFT packages, which enables controlled comparisons between Hamiltonian-based models and machine-learning interatomic potentials.

In this work, we further consider a larger molecular system than in our MD17 Hamiltonian benchmark to study a more challenging regime. We recompute reference quantities with PySCF at the PBE/def2-SVP level, using density fitting (via GPU4PySCF (Wu et al., 2024)) and an SCF convergence tolerance of 10^{-7}. Unless otherwise noted, we use a numerical integration grid of level 3.

Table 4. The statistics of rMD17 dataset (Schütt et al., 2019).

Dataset	# of structures	Train	Val	Test	# of atoms	# of orbitals	# of occupied MOs
Salicylic acid	30,000	25,000	500	4,500	16	170	36
Naphthalene	30,000	25,000	500	4,500	18	180	34
Aspirin	30,000	25,000	500	4,500	21	222	47

QH9. QH9 (Yu et al., 2023a) is a large-scale quantum chemistry benchmark for training and evaluating models that predict Hamiltonian matrices across diverse chemical structures. Built on QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2015), it contains 130,831 Hamiltonians from equilibrium geometries and 2,698 molecular dynamics (MD) trajectories. The dataset covers small organic molecules with up to nine heavy atoms (C, N, O, and F). All Hamiltonians are computed with PySCF (Sun et al., 2018) using the B3LYP (Stephens et al., 1994) exchange–correlation functional and the def2-SVP Gaussian-type orbital basis. We report detailed dataset statistics in Table 5.

QH9 consists of two subsets, `QH9-stable` and `QH9-dynamic-300k`, and defines four standard evaluation splits: stable-`id`, stable-`ood`, dynamic-300k-`geo`, and dynamic-300k-`mol`. For stable-`id`, `QH9-stable` is randomly split into training, validation, and test sets. For stable-`ood`, the split is defined by molecular size: molecules with 3–20 atoms are used for training, 21–22 atoms for validation, and 23–29 atoms for testing. This setup evaluates out-of-distribution generalization to larger and more complex molecules.

The dynamic-300k splits are constructed from MD trajectories, where each of the 2,698 molecules is associated with 100 geometry snapshots. In dynamic-300k-`geo`, snapshots are split within each molecule, with an 80/10/10 train/validation/test ratio. As a result, all molecular identities appear in every split, while the conformations are disjoint, isolating geometric generalization. In contrast, dynamic-300k-`mol` splits molecular identities into disjoint train/validation/test sets with the

same 80/10/10 ratio, and assigns all 100 snapshots of a molecule to the same split. This setting is more challenging than geo, as it requires generalization to unseen molecules rather than unseen conformations.

Table 5. The statistics of QH9 dataset (Yu et al., 2023a).

Dataset	# of structures	# of Molecules	Train	Val	Test
Stable-id	130,831	130,831	104,664	13,083	13,084
Stable-ood	130,831	130,831	104,001	17,495	9,335
Dynamic-300k-geo	269,800	2,698	215,840	26,980	26,980
Dynamic-300k-mol	269,800	2,698	215,800	26,900	27,100

E.2. Evaluation Metrics

We evaluate Hamiltonian prediction models at three levels: (i) element-wise reconstruction of the Hamiltonian matrix, (ii) spectral fidelity of the generalized eigenvalue problem, and (iii) accuracy of frontier-orbital quantities. Throughout this section, we use the following notation.

Notation. For a molecular configuration $\mathcal{M} = \{(Z_i, \mathbf{r}_i)\}_{i=1}^N$, let $\mathbf{H}^\star \in \mathbb{R}^{B \times B}$ denote the reference (SCF-converged) Kohn–Sham Hamiltonian matrix in an atom-centered orbital basis with B basis functions. The model outputs a prediction $\hat{\mathbf{H}} \in \mathbb{R}^{B \times B}$. We denote by $\mathbf{S} \in \mathbb{R}^{B \times B}$ the overlap matrix of the same basis, which is fixed given $\mathcal{M}$ and the basis choice.

Given $(\mathbf{H}, \mathbf{S})$, orbital energies and coefficients are obtained by solving the generalized eigenvalue problem

$$\mathbf{H}\,\mathbf{C} = \mathbf{S}\,\mathbf{C}\,\boldsymbol{\epsilon}, \tag{61}$$

where $\mathbf{C} \in \mathbb{R}^{B \times B}$ collects generalized eigenvectors (orbital coefficients) and $\boldsymbol{\epsilon} = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_B)$ contains the corresponding eigenvalues (orbital energies), ordered such that $\epsilon_1 \leq \cdots \leq \epsilon_B$. We write $(\hat{\mathbf{C}}, \hat{\boldsymbol{\epsilon}})$ for the solution obtained from $(\hat{\mathbf{H}}, \mathbf{S})$ and $(\mathbf{C}^\star, \boldsymbol{\epsilon}^\star)$ for that obtained from $(\mathbf{H}^\star, \mathbf{S})$ under the same numerical setup.

Hamiltonian MAE. This metric measures element-wise reconstruction accuracy of the Hamiltonian matrix:

$$\mathrm{MAE}(\mathbf{H}) = \frac{1}{B^2} \sum_{p=1}^{B} \sum_{q=1}^{B} \left| \hat{H}_{pq} - H_{pq}^\star \right|. \tag{62}$$

Here, $H_{pq}^\star$ and $\hat{H}_{pq}$ denote entries of the reference and predicted Hamiltonians in the same basis ordering. We report this quantity in μHa in the main tables, following standard Hamiltonian benchmarks.

Occupied orbital energy MAE ($\epsilon_{\mathbf{occ}}$). To assess spectral fidelity in the energy range relevant to the ground state, we compare orbital energies for occupied orbitals only:

$$\mathrm{MAE}(\epsilon_{\mathrm{occ}}) = \frac{1}{n_{\mathrm{occ}}} \sum_{p \in \mathcal{I}_{\mathrm{occ}}} \left| \hat{\epsilon}_p - \epsilon_p^\star \right|. \tag{63}$$

This metric focuses on the part of the spectrum that directly determines the ground-state density in RKS. Importantly, the energies $\hat{\epsilon}_p, \epsilon_p^\star$ are obtained by solving the generalized eigenproblem Equation (61) with the same overlap matrix $\mathbf{S}$.

Orbital coefficient similarity ($\mathcal{S}_c$). We measure alignment between predicted and reference orbital coefficient vectors. Let $\hat{\mathbf{c}}_p$ and $\mathbf{c}_p^\star$ denote the p-th columns of $\hat{\mathbf{C}}$ and $\mathbf{C}^\star$, respectively. Because each eigenvector is defined up to a sign (and may be unstable within degenerate or near-degenerate subspaces), we use an absolute cosine similarity and average over occupied orbitals:

$$\mathcal{S}_C(\hat{\mathbf{C}}, \mathbf{C}^\star) = \frac{1}{n_{\mathrm{occ}}} \sum_{p=1}^{n_{\mathrm{occ}}} \frac{\left| \left\langle \hat{\mathbf{c}}_p, \mathbf{c}_p^\star \right\rangle \right|}{\|\hat{\mathbf{c}}_p\|_2 \, \|\mathbf{c}_p^\star\|_2}. \tag{64}$$

Here $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product and $\| \cdot \|_2$ is the ℓ_2 norm. The absolute value makes the score invariant to sign flips ($\mathbf{c}_p \mapsto -\mathbf{c}_p$), which do not change the electron density.

HOMO, LUMO, and gap ($\epsilon_{\textbf{HOMO}}$, $\epsilon_{\textbf{LUMO}}$, $\epsilon_{\triangle}$). Frontier orbitals are key for chemical reactivity and are often sensitive to generalization. Under the RKS occupation rule, the HOMO index is $p_H = n_{occ}$ and the LUMO index is $p_L = n_{occ} + 1$. We report absolute errors for HOMO energy, LUMO energy, and the HOMO–LUMO gap:

$$\mathrm{MAE}(\epsilon_{\mathrm{HOMO}}) = \left| \hat{\epsilon}_{p_H} - \epsilon_{p_H}^{\star} \right|, \tag{65}$$

$$\mathrm{MAE}(\epsilon_{\mathrm{LUMO}}) = \left| \hat{\epsilon}_{p_L} - \epsilon_{p_L}^{\star} \right|, \tag{66}$$

$$\mathrm{MAE}(\Delta\epsilon) = \left| (\hat{\epsilon}_{p_L} - \hat{\epsilon}_{p_H}) - (\epsilon_{p_L}^{\star} - \epsilon_{p_H}^{\star}) \right|. \tag{67}$$

These quantities are computed from the eigenvalues of Equation (61) and therefore probe whether the predicted Hamiltonian reproduces the correct frontier spectrum relative to the reference.

E.3. Experimental setup

Environment. All experiments were run using a single GPU per model. Most runs were conducted on NVIDIA RTX 3090 and NVIDIA RTX A6000 GPUs. For larger model variants that required higher memory and throughput, we used NVIDIA H100 and NVIDIA B200 GPUs. For the extra-large model variant, we trained on an NVIDIA H100 with gradient accumulation over 4 steps to fit within memory constraints.

Our implementation is based on PyTorch 2.1.2 and PyG 2.3.0, compiled with CUDA 12.1. We additionally rely on standard scientific and atomistic ML libraries, including PySCF (Sun et al., 2018), e3nn (Geiger & Smidt, 2022), and ASE (Larsen et al., 2017). Full environment specifications (package versions and hardware details) will be released upon publication to support reproducibility.

To improve reproducibility, we fix random seeds wherever possible. Minor non-determinism may still remain due to GPU kernels and low-level library implementations.

Shared hyperparameters. For fair comparison, we follow the hyperparameter choices of the baseline QHNet (Yu et al., 2023b) whenever possible. QHFlow2 shares most architectural and training settings with QHNet so that performance differences primarily reflect our flow-matching formulation rather than extensive hyperparameter tuning. Key training and inference hyperparameters for each dataset are summarized in Table 9.

Table 6. Training and inference hyperparameters of QHFlow2 used across datasets.

Hyperparameter	Description	QH9	MD17
Learning Rate	Initial learning rate	1e-3	1e-3
Minimum Learning Rate	Minimum learning rate	1e-7	1e-9
Batch Size	Number of molecules per batch	32	10
Scheduler	Learning rate scheduler	Polynomial	Polynomial
LR Warmup Steps	Warmup steps for linear warmup	1,000	1,000
Max Steps	Maximum number of training steps	260,000	200,000
Fine-tuning LR	Initial learning rate for fine-tuning	1e-5	–
Fine-tuning Minimum LR	Minimum learning rate for fine-tuning	1e-7	–
Fine-tuning Steps	Maximum number of fine-tuning steps	60,000	–
Prior Distribution	Prior for flow matching	TE	TE
Sampling Steps	Number of ODE steps at inference	3	3

Architecture hyperparameters. Table 7 reports the architectural and input-feature settings for QHFlow2. Across all experiments, the model is conditioned on the time-dependent Hamiltonian state H_t. We additionally provide the overlap matrix S for QH9, while omitting it for MD17/rMD17. The remaining hyperparameters specify the equivariant backbone: the maximum irrep degree $\ell_{\max}$, embedding and bottleneck widths, the numbers of SO(2) backbone layers and two-stage update layers, the SO(2) neighbor cutoff radius, and the channel sizes used for spherical and edge features.

Hamiltonian model hyperparameters. In Table 8, we report the key architectural choices of baseline Hamiltonian predictors. Training strategies follow the original papers or are kept consistent with our main training setup; For model-scale comparisons, we vary only the capacity-related parameters while keeping the remaining architectural choices fixed.

Table 7. QHFlow2 architecture and input settings by model size.

Setting	QHFlow2-S	QHFlow2-S (md)	QHFlow2-M	QHFlow2-M (md)	QHFlow2-L	QHFlow2-XL
Parameters	14M	12M	43.3M	42.8M	183M	990M
Using H_t as embedding	True	True	True	True	True	True
Using H_{init} as embedding	False	False	False	False	True	True
Using S as embeding	True	False	True	False	True	True
Model order (ℓ_{max})	4	4	4	4	4	4
Embedding dimension	64	64	128	128	256	512
Bottleneck hidden size	32	32	64	32	64	128
Number of SO(2) backbone layers	3	2	3	3	4	5
Number of two-stage update layers	2	2	2	2	2	3
SO(2) max radius (Å)	5.0	5.0	5.0	5.0	5.0	5.0
Sphere channels	64	64	128	128	256	512
Edge channels	64	64	128	128	256	512

Table 8. Architecture and input hyperparameters of Hamiltonian predictors.

Hyperparameter	SPHNet(18M/36M)	QHNet	QHFlow
Using S as embedding	False / False	False	True
Model order (ℓ_{max})	4 / 4	4	4
Embedding dimension	128 / 184	128	128
Bottleneck hidden size	32 / 32	32	32
Number of backbone GNN layers	4 / 4	5	5
Cutoff radius (Å)	15 / 15	15	15
Sphere channels	128 / 184	128	128
Edge channels	32 / 32	32	32

E.4. SCF convergence test

Figure 4 reports SCF convergence behavior when replacing the default initialization with predicted Hamiltonians. All models in this experiment are trained on QH9-stable-iid. We randomly sample 300 molecules from the corresponding test split and, for each molecule, use the model-predicted Hamiltonian $\hat{H}$ to construct an initial electron density, from which we run SCF.

To ensure a controlled and fair comparison, we fix the DFT setup and SCF solver configuration across all runs and change only the initialization. Following the QH9 protocol, we use B3LYP/def2-SVP for all calculations. We match the convergence tolerance and numerical settings to the QH9 setup (e.g., `conv_tol`$= 10^{-13}$ and grid level 3), and keep the DIIS procedure at the PySCF default configuration. For computational efficiency, we use GPU4PySCF and density fitting, applied identically to the baseline and all Hamiltonian predictors. Under this setup, differences in SCF iterations primarily reflect the quality of the initial guess rather than solver-side changes.

SCF fixed points can be sensitive to subtle numerical details (e.g., grid, screening and hardware), so the converged solution under our evaluation stack may not exactly match the dataset configuration. Even so, the consistent trend toward the dataset-Hamiltonian initialization in Figure 4 suggests that improved Hamiltonian predictions provide better starting points for the intended electronic state, resulting in fewer SCF iterations.

F. MLFF Experimental Details

F.1. Dataset and Preprocessing

Dataset. We employed high-fidelity recalculated trajectories based on the MD17 benchmark. Specifically, we utilized the Revised MD17 (rMD17) dataset for four molecules: aspirin, naphthalene, salicylic acid, and ethanol. Additionally, we included uracil, malondialdehyde, and ethanol obtained from recalculated MD17 trajectories. For consistency across experiments, we employed a subset of 30,000 configurations for each molecule.

Data Splits and Representation. All molecular configurations are represented as atomistic graphs where nodes correspond to atoms with atomic number Z and 3D coordinates, and edges connect atom pairs within a model-specific cutoff radius (r_c). The graphs are treated as undirected, implemented via bidirectional directed edges (i.e., both $i \rightarrow j$ and $j \rightarrow i$ are included) without periodic boundary conditions.

For each molecule, we employed fixed train/validation/test splits of 25,000/500/4,500 configurations, totaling 30,000 samples. These splits were pre-generated and consistently applied across all models to ensure fair comparison. For low-data experiments, we keep the validation and test sets unchanged and subsample the training set to $N \in \{10,000, 5,000, 1,000\}$ by taking the first N configurations from the pre-shuffled 25,000-sample training pool. Since the training pool is already randomly ordered by construction, this procedure yields nested random subsets while remaining fully deterministic. Energies are reported in eV, and forces in eV/Å.

Target Normalization and Preprocessing. We adopted the standard hyperparameters and configurations provided by the MDsim framework for all reference models to ensure reproducibility.

- **GemNet-T & DimeNet:** We did not apply explicit label normalization (`normalize_labels: False`), consistent with their standard implementations.

- **NequIP:** We used the default configuration which automatically applies per-species rescaling based on dataset statistics (shifting by mean per-atom energy and scaling by force RMS).

F.2. Model Architectures and Hyperparameters

We evaluate three reference models: NequIP, GemNet-T, and DimeNet. Table 9 summarizes the key hyperparameters and architectural details adopted in our experiments. To ensure fair comparison, consistent loss weights were applied across models where applicable.

Table 9. Hyperparameters and architectural details for NequIP, GemNet-T, and DimeNet. Note that the loss weights (λ) denote the ratio between energy and force terms in the objective function.

Hyperparameter	NequIP	GemNet-T	DimeNet
Architecture			
Cutoff Radius (r_c)	4.0 Å	5.0 Å	5.0 Å
Interaction Blocks / Layers	5 layers	4 blocks	6 blocks
RBF Basis Size	8	6 (radial), 7 (spherical)	6 (radial), 7 (spherical)
Hidden Dimension	32	128 (atom), 64 (triplet)	128
Training Objective			
Loss Weights ($\lambda_E : \lambda_F$)	1 : 1000	1 : 1000	1 : 1000
Optimization			
Optimizer	Adam	AdamW	AdamW
Learning Rate	5×10^{-3}	1×10^{-3}	1×10^{-3}
Batch Size	5	1	32
Max Epochs	2000	10000	10000
Complexity			
# of Parameters	1,053,816	1,890,125	2,100,070

F.3. Training Infrastructure

Hardware and Software. All models were trained on a high-performance computing cluster utilizing NVIDIA Ampere and Hopper architecture GPUs (including A100, H100, and H200). Each training session was executed on a single GPU node. The software environment was configured using the MDsim framework[0]. Key dependencies include PyTorch 2.0 and CUDA 11.8, chosen to ensure compatibility across different GPU architectures.

Computational Cost. The training duration varied significantly depending on the model architecture and convergence rate. On average, a full training run required approximately 24 to 48 hours of wall-clock time on a single A100 GPU.

F.4. Evaluation metric and protocol

We assessed the model performance on a held-out test set comprising 4,500 configurations. We report the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for both potential energy (E) and atomic forces (F).

- **Energy Metrics:** MAE and RMSE are computed on the predicted total potential energy per conformation (Unit: eV).

- **Force Metrics:** MAE and RMSE are calculated component-wise for all atomic forces (Unit: eV/Å), averaged over all atoms and spatial directions (x, y, z).

---

[0]The source code and environment details are available at: https://github.com/kyonofx/MDsim

G. Additional results

G.1. Ablation on QHFlow2 design

Ablation of two-stage edge update. Table 10 ablates the two-stage update by toggling the reference-frame cue g_{ref} and the SO(2) edge cutoff, while fixing the SO(2) backbone and parameter budget (43M) and training on QH9Stable-iid. Without g_{ref}, a larger cutoff is needed to reduce off-diagonal error. Enabling g_{ref} improves all metrics across cutoffs and narrows the gap between small and large cutoffs, giving strong performance already at the default 5 Å. The best result is obtained with g_{ref} and a 15 Å cutoff.

Table 10. **SO(2) ablations over two-stage update.** All variants use the same SO(2) backbone and are trained on QH9Stable-iid with the parameter budget fixed to 43M. We vary whether the reference-frame cue g_{ref} is enabled and the SO(2) edge cutoff (default: 5 Å). We report Hamiltonian MAE and its diagonal/off-diagonal components (lower is better).

g_{ref}	cutoff	H MAE ↓	Diag.↓	Off-diag. ↓
✓	–	27.91	31.37	14.35
✗	5	14.91	22.89	27.64
✗	8	9.58	22.06	8.63
✗	12	9.36	21.79	8.41
✗	15	9.86	21.11	9.00
✓	5	9.18	18.45	8.47
✓	12	8.55	17.77	7.84
✓	15	**8.43**	**17.30**	**7.75**

Ablation of flow matching. Table 11 compares flow matching with a regression-style variant of QHFlow2 that fixes $t = 0$ and $H_t = \mathbf{0}$ during training and evaluation, using the same architecture and trained on QH9Stable-iid. Flow matching improves Hamiltonian MAE on both diagonal and off-diagonal blocks and reduces orbital energy error. The gap between 1-step and 3-step is small, and we find that 1–3 steps are generally sufficient in our setting.

Table 11. **Regression vs. Flow-based Hamiltonian prediction.** Comparison between pointwise regression and flow-based modeling with different numbers of flow steps. Errors are reported separately for diagonal and off-diagonal Hamiltonian blocks. Lower is better.

Metric	Regression	Flow-based	
		1-step	3-step
All H MAE ↓	10.77	9.19	**9.18**
Diagonal H MAE ↓	21.01	18.49	**18.46**
Off-diagonal H MAE ↓	10.00	8.55	**8.47**

Ablation of initial Hamiltonian. Table 12 ablates the use of an initial Hamiltonian by evaluating models without providing an initial guess at inference. QHFlow2 retains strong accuracy under this setting and yields energy and force errors in the MLIP range on MD17. In contrast, prior Hamiltonian predictors show reasonable matrix-level scores but fail to produce accurate energies when the initial Hamiltonian is removed, highlighting the robustness of QHFlow2 for direct downstream evaluation.

Table 12. **MD17 benchmark w/o initial Hamiltonian.** Best results are shown in **bold**.

		Ethanol (9 atoms)					Malonaldehyde (9 atoms)					Uracil (12 atoms)				
Model	**Param.**	H ↓	ϵ_{occ} ↓	S_c ↑	Energy↓	Force↓	H ↓	ϵ_{occ} ↓	S_c ↑	Energy↓	Force↓	H ↓	ϵ_{occ} ↓	S_c ↑	Energy↓	Force↓
w/o initial Hamiltonian																
SchNOrb	–	187.4	334.4	**100.00**	–	–	191.1	400.6	99.00	–	–	227.8	1760.	90.00	–	–
PhiSNet	–	20.09	102.04	99.81	–	–	21.31	100.6	99.89	–	–	18.65	143.36	99.86	–	–
QHNet	20M	27.99	99.33	99.99	–	–	29.60	100.16	99.92	–	–	26.80	127.93	99.99	–	–
HELM	–	**5.79**	–	–	28.93	–	**4.86**	–	–	20.15	–	**3.61**	–	–	22.69	–
QHFlow2-m	43M	12.49	21.32	**100.00**	0.136	1.320	13.14	26.64	99.97	**1.145**	2.296	10.47	29.13	99.89	**0.638**	2.414

We additionally explore a simple *diagonal initialization* scheme. Since the on-atom (diagonal) Hamiltonian blocks are rotationally invariant for isolated atoms, we initialize each diagonal block using the corresponding atom's reference Hamiltonian computed from a one-time atomic DFT calculation. This precomputation is cheap and amortized across

Table 13. **QH9 benchmark init.** Results shown in **bold** denote the best result in each column, whereas those that are <u>underlined</u> indicate the second best.

Dataset	Model	Param.	$H\downarrow[\mu E_h]$	$\epsilon_{\text{occ}}\downarrow[\mu E_h]$	$\mathcal{S}_c\uparrow[\%]$	$\epsilon_{\text{LUMO}}\downarrow[\mu E_h]$	$\epsilon_{\text{HOMO}}\downarrow[\mu E_h]$	$\epsilon_\Delta\downarrow[\mu E_h]$
	QHFlow	28M	22.95	119.67	99.51	437.96	179.48	553.87
	QHFlow2-m	43M	9.18	59.23	99.73	175.68	77.67	215.99
Stable-iid	**QHFlow2**-l	183M	**5.93**	**45.73**	**99.81**	**103.32**	**50.50**	**119.62**
	QHFlow2-m (w/o init)	43M	25.50	279.84	98.31	834.31	255.53	1025.42
	QHFlow2-m (+ diagonal init.)	43M	22.81	130.88	99.39	965.63	218.24	1057.18
	QHFlow2-l (+ diagonal init.)	183M	16.22	150.04	99.52	794.71	109.16	825.32

molecules. Table 13 summarizes the results. Diagonal initialization improves over using no initialization, but still lags behind the full initialization setting, especially for orbital energies and the HOMO–LUMO gap. This suggests that accurate initialization must also capture inter-atomic couplings and environment-dependent on-site effects, which are not provided by isolated-atom diagonal blocks.

G.2. Exact values for plotted results

This section reports the exact numerical values used to generate the figures in the main text. Table 14 lists the values used in Figure 2, and Table 15 lists the values used in Figure 3.

Table 14. **MLIP vs. QHFlow2 comparison on MD17.** Energy and force errors are reported in meV and meV/Å. Molecules Eta., Malon., and Ura. belong to MD17, while SA, Naph., and Asp. belong to rMD17. Best results are shown in **bold**, and second best are <u>underlined</u>.

	MD17						rMD17					
	Eta.		Malon.		Ura.		SA		Naph.		Asp.	
Model	E↓	F↓	E↓	F↓	E↓	F↓	E↓	F↓	E↓	F↓	E↓	F↓
MLIP												
DimeNet	10.32	1.84	12.98	1.16	10.68	2.13	12.21	4.06	11.92	2.83	26.04	5.45
Gemnet-T	7.44	3.77	8.12	5.73	9.37	5.51	15.64	4.85	19.62	2.87	10.27	5.28
Nequip	0.33	0.61	0.73	0.88	0.45	0.60	1.71	1.43	2.01	0.76	1.87	2.08
QHFlow2												
s-14M	<u>0.005</u>	<u>0.36</u>	<u>0.058</u>	<u>0.57</u>	<u>0.234</u>	<u>1.01</u>	<u>0.110</u>	<u>1.48</u>	<u>0.063</u>	1.17	<u>0.124</u>	<u>1.42</u>
m-43M	**0.003**	**0.34**	**0.010**	**0.51**	**0.062**	**0.90**	**0.070**	**1.38**	**0.040**	<u>1.10</u>	**0.084**	**1.35**

Table 15. **MLIP vs. QHFlow2 comparion on QH9.** We report a subset of orbital- and energy-related metrics for representative MLFF baselines on QM9 and QHFlow-v2 variants on QH9-stable (iid).

Dataset	Model	$H(\downarrow)$	LUMO($\downarrow$)	HOMO($\downarrow$)	GAP($\downarrow$)	Energy($\downarrow$)
	MLFF baseline					
QM9[†]	Equiformer	–	551.24	514.49	1102.48	6.59
	MACE	–	808.48	698.24	1543.47	4.10
	EquiformerV2	–	514.49	477.74	1065.73	6.17
	QHFlow2					
QH9-stable	**QHFlow2-s**	16.29	304.69	148.42	411.41	1.42
(iid)	**QHFlow2-m**	9.18	175.68	77.67	215.99	0.69
	QHFlow2-l	5.93	103.32	50.50	119.62	0.32
	QHFlow2-xl	**4.58**	**73.19**	**40.30**	**87.74**	**0.21**

G.3. Energy and force calculation runtime on MD17

Table 16 reports a runtime breakdown for MLIPs and Hamiltonian predictors on MD17, averaged over 100 structures each from salicylic acid, naphthalene, and aspirin. For MLIPs, the reported time is end-to-end energy and force evaluation

(network energy prediction plus autodiff forces). For Hamiltonian predictors, we separate (i) Hamiltonian inference and (ii) a downstream PySCF step that maps $\hat{H}$ to energies and analytic forces. We use the same PySCF pipeline for all Hamiltonian models, so the downstream cost is comparable across predictors and provides a controlled runtime baseline. Under this decomposition, differences between Hamiltonian predictors are reflected in the Hamiltonian inference time, while the overall cost is currently dominated by the shared PySCF post-processing. A remaining limitation is that this evaluation depends on a DFT solver stack; while it ensures physically grounded energies and forces, the total runtime inherits solver-side overhead that is not optimized by the predictor itself.

Table 16. **Speed and runtime breakdown on the MD benchmark.** Runtime is reported in seconds (s), and speed is reported in samples/s, averaged over 100 structures each from salicylic acid, naphthalene, and aspirin. Inf. is the model forward-pass time (end-to-end energy/force evaluation for MLIPs; Hamiltonian prediction for Hamiltonian models). For Hamiltonian models, E and F are the downstream PySCF times to compute energy and forces from the predicted Hamiltonian $\hat{H}$, so Total = Inf. + E + F, while Total = Inf. + F for MLIPs. Best results are in **bold** within each model family.

Model	Inf.↓	E	F	Total ↓	Speed ↑
MLIPs			Inf./E: direct model output; F: autograd		
Dimenet	0.121	–	0.003	0.124	8.03
Gemnet-T	**0.086**	–	0.001	**0.087**	**11.44**
Nequip	0.164	–	0.002	0.167	5.97
Hamiltonian Models			Inf.: predict $\hat{H}$; E/F: PySCF from $\hat{H}$		
SPHNet-18M	0.089	0.504	0.883	1.483	0.694
SPHNet-36M	0.091	0.507	0.881	1.477	0.693
QHFlow2-s	**0.086**	0.506	0.882	**1.474**	**0.699**
QHFlow2-m	0.159	0.507	0.890	1.556	0.663

G.4. Extrapolation on Molecular Reactive Sites

For further investigation beyond the comparison with MLIPs, we conduct an electronic-structure-based reactivity analysis using Hamiltonian predictors, which is prohibitive for MLIPs. For this, we design molecular reactive bond-stretching experiments where their Frontier molecular orbital (FMO) (Fukui, 1982) energies (*i.e.* HOMO and LUMO) change, reflecting varying molecular reactivity. FMO energies are used as global reactivity descriptors (Fukui, 1982), following their widespread use in conceptual DFT (Geerlings et al., 2003) to characterize molecular stability, softness, and overall chemical reactivity (Choudhary et al., 2019; Miar et al., 2021; Kumar et al., 2025). This setup explicitly highlights a fundamental limitation of MLIPs: since MLIPs do not represent the electronic structure, they cannot be used for traditional chemistry applications relying on orbital- and density-based reactivity descriptors (Mortazavi et al., 2023; Wang et al., 2024; Mortazavi, 2025), whereas sufficiently accurate Hamiltonian predictors can reproduce DFT-level electronic reactivity trends and thereby enable such analyses. In detail, for the aspirin and naphthalene molecules included in the MD benchmark, target specific reactive sites: the ester carbonyl bond (C11=O12) in Aspirin (an electrophile) and the aromatic C1-C2 bond in Naphthalene (a nucleophile), we change the reactive bond length from 0.95 to 1.15, evaluate the FMO energy calculated from the predicted Hamiltonian, based on the MAE of the predicted values against DFT calculations. The results in Section G.4 show that the Hamiltonian predictors can be applied to orbital- and density-based reactivity descriptor analyses, with the proposed QHFlow2 demonstrating superior robustness compared to baseline. While both models exhibit low errors near the equilibrium bond length, SPHNet shows a significant divergence as the bond is stretched. SPHNet shows a rapid degradation in accuracy, with errors escalating sharply in the extrapolation regime (stretch ratio > 1.05). In contrast, QHFlow2 maintains a consistently low error profile across the entire stretching range. Specifically, for the HOMO energy, QHFlow2 keeps the MAE below 2.0 meV even at maximum distortion, whereas SPHNet's error exceeds 5.0 meV. A similar trend is observed for the LUMO energy, where QHFlow2 significantly outperforms the SPHNet. This demonstrates that the proposed QHFlow2 captures subtle electronic variations more robustly than SPHNet, validating its potential for reliable reactivity analysis beyond equilibrium geometries.
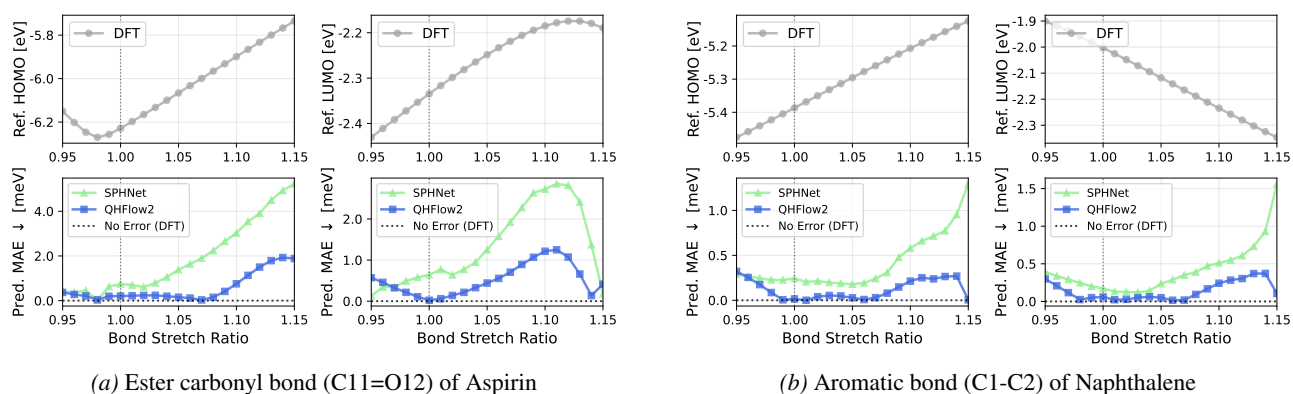
(a) Ester carbonyl bond (C11=O12) of Aspirin

(b) Aromatic bond (C1-C2) of Naphthalene

Figure 8. **Robustness of Frontier Molecular Orbital (FMO) Predictions under Reactive Bond Stretching.** The plots illustrate the Mean Absolute Error (MAE) of calculated HOMO and LUMO energies from Hamiltonian predictors (*i.e.* QHFlow2 and SPHNet), relative to DFT value as the reactive bond length is distorted from $0.95\times$ to $1.15\times$ the equilibrium length. (a) Performance on the reactive ester carbonyl bond (C11=O12) of Aspirin, which acts as an electrophilic site. (b) Performance on the aromatic C1-C2 bond of Naphthalene, representing a nucleophilic site.