HAMFORMER: A HAMILTONIAN FLOW-BASED SE(3)-TRANSFORMER FOR CONFORMATION OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Transformer architecture plays a central role in modern deep learning, and its SE(3)-equivariant variants have been widely applied to molecular conformation modeling and optimization. Given its importance, many methods attempt to interpret and improve the Transformer through the lens of dynamics. However, they mainly focus on a single attention head rather than the full Transformer architecture and seldom consider SE(3)-equivariance in their proposed models. In this study, we propose an improved SE(3)-Transformer architecture, called HamFormer, which revisits the design of SE(3)-Transformer through the lens of Hamiltonian flow. In particular, we define a parametrized Hamiltonian flow in the phase space of position and velocity, corresponding to optimizing an energy functional that combines an optimal transport-based potential with a damped momentum. The Lie-Trotter splitting framework of the Hamiltonian flow leads to the proposed HamFormer architecture, whose feedforward computation achieves the optimization step of the proposed energy functional. Furthermore, we combine HamFormer with Equivariant Graph Neural Network (EGNN), build a new model for molecular conformation optimization. Extensive experiments on molecular ground-state conformation prediction show that HamFormer provides a strong and physically interpretable surrogate for classic SE(3)-Transformer and consistently outperforms state-of-the-art competitors.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) have become the dominant architecture for modeling interacting elements, from tokens in language to atoms in molecules and materials. For the elements in three-dimensional space, whose underlying physics is invariant under rotations and translations, SE(3)-Transformers (Fuchs et al., 2020; Thomas et al., 2018) build this symmetry into attention itself, coupling invariant scalar features with equivariant coordinates so that predictions transform consistently with their inputs. This geometric inductive bias underpins a series of practical successes across the molecular sciences, including protein structure prediction in AlphaFold (Jumper et al., 2021; Abramson et al., 2024), molecular representation learning in Uni-Mol (Zhou et al., 2023), equivariant force fields and materials modeling (Satorras et al., 2021; Liao & Smidt, 2023), and ground-state conformation prediction (Wang et al., 2025). As the Transformer and its SE(3)-equivariant variants are entrusted with physically consequential predictions, understanding and improving what their layers compute has become a question of practical rather than merely theoretical significance.

Considerable efforts have gone into understanding the mechanism behind the Transformer, such as interpreting its self-attention layer through interacting particle dynamics (Geshkovski et al., 2023; Sander et al., 2022; Wang et al., 2025) and revisiting its residual architecture as ODE flows (Chen et al., 2018). Building on these interpretations, some recent methods improve the Transformer architecture through the lens of dynamics. For example, Sinkformer (Sander et al., 2021) applies doubly stochastic attention maps based on the Sinkhorn-scaling algorithm (Cuturi, 2013) and, accordingly, interprets the improved Transformer as Wasserstein gradient flows. Recent YurriFormer (Zimin et al., 2026) and SympFormer (Stein et al., 2026) introduce high-order dynamical information into the Transformer and design their models as the Nesterov-accelerated gradient descent algorithm. However, extending these dynamical interpretations to molecular refinement requires considering both feature evolution and molecular geometry. Beyond the attention operator, the architecture combines multiple attention heads, tokenwise FFNs, and pair relations that inform equivariant coordinate updates. The challenge

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

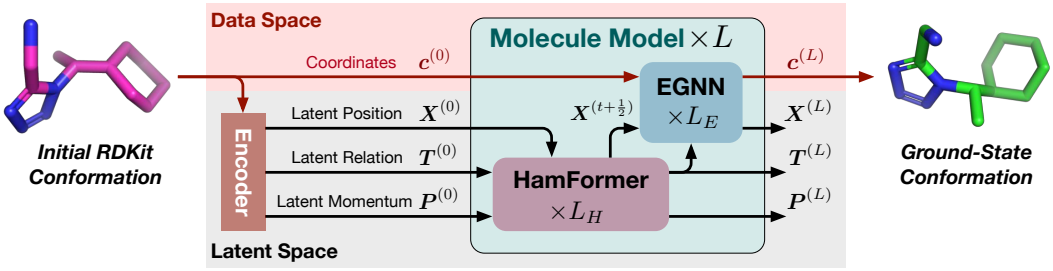

Figure 1: An illustration of the proposed molecule model based on HamFormer.

is to connect these components within a common dynamical framework that guides their interaction while preserving SE(3)-equivariance.

In this paper, we propose **HamFormer**, which revisits the SE(3)-Transformer from the viewpoint of parametric Hamiltonian flows and improves the architecture accordingly. In principle, HamFormer is built based on coupled interacting particle dynamics, which capture the evolution of a distribution in the phase space of position and momentum. To model the dynamics, we define a Hamiltonian flow in phase space generated by an energy functional consisting of an optimal transport-based potential and a learned kinetic term. We parameterize the energy functional and discretize the resulting dynamics in the Lie–Trotter splitting framework (Hairer et al., 2006), with momentum retention applied before the interaction update. Given a set of particles (e.g., atomic embeddings of a molecule) and their initial momentum and pairwise relations, the discrete implementation of splitting leads to the proposed HamFormer architecture, whose feedforward computation updates the positions, momentum, and relations of the particles following the reference kick–drift structure.

Notably, HamFormer augments the atom features and pair relations used in SE(3)-equivariant molecular architectures (Fuchs et al., 2020; Zhou et al., 2023; Wang et al., 2025) with an explicit latent momentum state, allowing interactions from earlier layers to influence subsequent feature updates. In addition, compared with existing accelerated and momentum-based Transformer architectures (Zimin et al., 2026; Stein et al., 2026; Gai et al., 2026), HamFormer combines an optimal transport-based potential with multi-head feature updates and SE(3)-equivariant coordinate refinement. As shown in Figure 1, combining HamFormer with Equivariant Graph Neural Network (EGNN) leads to a new model for molecular conformation optimization. We evaluate the proposed model on the task of ground-state conformation prediction. Experimental results demonstrate that HamFormer provides a strong and physically interpretable surrogate for classic SE(3)-Transformer, whose feature updates follow a Hamiltonian reference structure and coordinate refinement preserves SE(3)-equivariance. Accordingly, the HamFormer-based model outperforms state-of-the-art ground-state conformation predictors in both accuracy and efficiency.

2 Related Work and Preliminaries

Transformer and Its SE(3)-Equivariant Version. The Transformer uses attention to aggregate information across a sequence or a set of samples. Typically, given an input $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, where N indicates the sequence length or sample-set size and d is the input feature dimension, an attention head first obtains the value, query, and key matrices by linear maps to a head dimension d_h, i.e., $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$, $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q$, and $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K$, with $\boldsymbol{W}_V, \boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d\times d_h}$. It then aggregates $\boldsymbol{V}$ using weights determined by $\boldsymbol{Q}$ and $\boldsymbol{K}$ as $\mathrm{Att}(\boldsymbol{V}, \boldsymbol{Q}, \boldsymbol{K}) := \boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V}$, where $\boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{K}) \in \mathbb{R}^{N\times N}$ is the attention map parametrized by $\boldsymbol{Q}$ and $\boldsymbol{K}$. The vanilla Transformer (Vaswani et al., 2017) implements this map based on the softmax operation, i.e., $\boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{K}) = \mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{d_h})$, where $\mathrm{Softmax}(\cdot)$ is applied to each row, giving nonnegative attention weights that sum to one. The multi-head attention layer applies a group of linear maps, i.e., $\theta_{\mathrm{att}} = \{\boldsymbol{W}_{V,h}, \boldsymbol{W}_{Q,h}, \boldsymbol{W}_{K,h}\}_{h=1}^{H}$, to construct H attention heads with a common head dimension d_h and concatenates their outputs, with the output projection omitted, so that $\mathrm{MHA}_{\theta_{\mathrm{att}}}(\boldsymbol{X}) := \mathrm{Concat}_{\mathrm{col}}\big(\{\mathrm{Att}(\boldsymbol{V}_h, \boldsymbol{Q}_h, \boldsymbol{K}_h)\}_{h=1}^{H}\big)$, where $\mathrm{MHA}_{\theta_{\mathrm{att}}}(\boldsymbol{X}) \in \mathbb{R}^{N\times Hd_h}$, $\mathrm{Concat}_{\mathrm{col}}$ denotes column-wise concatenation, and $\boldsymbol{V}_h = \boldsymbol{X}\boldsymbol{W}_{V,h}$, $\boldsymbol{Q}_h = \boldsymbol{X}\boldsymbol{W}_{Q,h}$, and $\boldsymbol{K}_h = \boldsymbol{X}\boldsymbol{W}_{K,h}$ for $h = 1, \ldots, H$. A complete Transformer block combines multi-head attention with a token-wise feedforward network (FFN), residual connections, and layer normalization. For molecular coordinates $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N]^\top \in \mathbb{R}^{N\times 3}$, a coordinate predictor f

is SE(3)-equivariant if $f(\boldsymbol{C}\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{t}^\top) = f(\boldsymbol{C})\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{t}^\top$ for every rotation $\boldsymbol{R} \in \mathrm{SO}(3)$ and translation $\boldsymbol{t} \in \mathbb{R}^3$, with invariant atom and pair attributes held fixed.

Beyond sequence modeling, attention architectures support geometric prediction. SE(3)-Transformer and Equiformer combine attention with equivariant representations for three-dimensional data and atomistic learning (Fuchs et al., 2020; Liao & Smidt, 2023). EGNN constructs equivariant coordinate updates from invariant scalar weights and relative vectors (Satorras et al., 2021), providing a coordinate-refinement module for invariant feature processors.

Improved Transformer Architectures. Transformer layers can be interpreted as time discretizations of differential equations or iterations of optimization algorithms. The multi-particle formulation of Lu et al. (2019) associates attention and FFNs with interacting flows, while optimization-based analyses view layers as iterative updates (Yang & Wipf, 2022). Broader accounts connect Transformer dynamics to interacting-particle systems, gradient flows, and mean-field limits (Geshkovski et al., 2025).

Sinkformers replace softmax attention with Sinkhorn normalization, which approximately enforces both row and column sums (Sander et al., 2022). Entropy-regularized optimal transport provides the associated optimization problem (Cuturi, 2013). Under suitable assumptions, the resulting continuous-depth dynamics form a Wasserstein gradient flow. WGFormer applies transport-normalized attention to molecular conformation refinement and evolves pair relations across layers (Wang et al., 2025). Their gradient-flow analyses use a single attention head, although the implemented architectures can contain multiple heads.

MomentumNets introduce persistent momentum-like states in residual networks (Sander et al., 2021), and Hamiltonian neural networks learn energies governing coupled state–momentum dynamics (Greydanus et al., 2019). Energy Transformer derives attention and tokenwise updates from a designed energy (Hoover et al., 2023). More recently, YuriiFormer interprets attention and MLP updates through Lie–Trotter splitting of a composite objective (Zimin et al., 2026), while SympFormer derives accelerated attention from inertial dynamics on density spaces (Stein et al., 2026). Connecting these dynamical descriptions to molecular refinement requires accounting for both invariant feature processing and equivariant coordinate updates.

3 Proposed HamFormer Architecture

In HamFormer, a parametrized Hamiltonian reference flow couples invariant atom features with latent momentum. The potential force updates momentum, and the kinetic gradient advances the features. Its Lie–Trotter splitting determines the feedforward update structure for molecular conformation optimization. Evolving multi-head relations condition these interactions and provide invariant edge information to EGNN for equivariant coordinate refinement.

3.1 Hamiltonian Flow

Hamiltonian dynamics are specified by an energy and a canonical state–momentum pairing (Evans, 2022, Sec. 3.3.1). On a periodic domain Ω of dimension d, let $\boldsymbol{x} \in \Omega$ denote position, ρ a smooth probability density, and S its conjugate potential, with velocity field ∇S. Let $\mathcal{F}$ be a potential-energy functional with first variation $\mathcal{F}'$. For N particles, collect positions $\boldsymbol{x}_i$ and momentum $\boldsymbol{p}_i$ as rows of $\boldsymbol{X}, \boldsymbol{P} \in \mathbb{R}^{N\times d}$, and let $\rho_N := N^{-1}\sum_{i=1}^N \delta_{\boldsymbol{x}_i}$ be the empirical approximation of ρ, where $\delta_{\boldsymbol{x}}$ denotes the Dirac measure at $\boldsymbol{x}$. The Wasserstein Hamiltonian (Chow et al., 2020) and its particle counterpart are

$$\mathcal{H}_W(\rho, S) = \frac{1}{2}\int_\Omega \|\nabla S(\boldsymbol{x})\|^2 \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \mathcal{F}(\rho) \quad \text{and} \quad H_N(\boldsymbol{X}, \boldsymbol{P}) = \frac{1}{2}\sum_{i=1}^N \|\boldsymbol{p}_i\|^2 + N\mathcal{F}(\boldsymbol{X}), \quad (1)$$

where the two terms of H_N define the kinetic energy $K_N(\boldsymbol{P})$ and potential energy $V_N(\boldsymbol{X})$, respectively. The particle expression uses total-energy normalization and assumes $\mathcal{F}(\boldsymbol{X})$ is well defined and smooth in $\boldsymbol{X}$.

Writing ρ_t, S_t for the state at time t, define the momentum along a characteristic trajectory $\boldsymbol{x}(t)$ by $\boldsymbol{p}(t) = \nabla S_t(\boldsymbol{x}(t))$ and the phase-space Hamiltonian by $h_t(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2 + \mathcal{F}'(\rho_t)(\boldsymbol{x})$. With the time-dependent additive constant in S_t chosen appropriately, the density and potential satisfy

the continuity and Hamilton–Jacobi equations, whose characteristic trajectories obey Hamilton's equations, i.e.,

$$\begin{cases} \partial_t \rho_t + \nabla \cdot (\rho_t \nabla S_t) = 0 & \text{(continuity)} \\ \partial_t S_t + \frac{1}{2}\|\nabla S_t\|^2 + \mathcal{F}'(\rho_t) = 0 & \text{(HJE)} \end{cases} \quad \begin{cases} \dot{\boldsymbol{x}} = \nabla_{\boldsymbol{p}} h_t = \boldsymbol{p} \\ \dot{\boldsymbol{p}} = -\nabla_{\boldsymbol{x}} h_t = -\nabla \mathcal{F}'(\rho_t)(\boldsymbol{x}) \end{cases} \text{(Hamilton)},$$

where the right-hand system is evaluated along $(\boldsymbol{x}(t), \boldsymbol{p}(t))$, with probability mass preserved for smooth periodic solutions and $\mathcal{H}_W$ conserved when $\mathcal{F}$ is time-independent (see also, Appendix A.1 for more details.).

More generally, for smooth, time-independent kinetic and potential energies K_N, V_N, the potential subflow changes momentum at fixed positions, while the kinetic subflow advances positions at fixed momentum. Their Lie–Trotter composition gives first-order *symplectic Euler* (see also, Wu et al., 2025, Sec. 4). For a step size $\eta > 0$, let $\boldsymbol{X}^k, \boldsymbol{P}^k$ approximate the states at $t_k = k\eta$. The particle dynamics and their discretization are

$$\begin{cases} \dot{\boldsymbol{P}} = -\nabla_{\boldsymbol{X}} V_N(\boldsymbol{X}) \\ \dot{\boldsymbol{X}} = \nabla_{\boldsymbol{P}} K_N(\boldsymbol{P}) \end{cases} \text{(continuous)} \quad \begin{cases} \boldsymbol{P}^{k+1} = \boldsymbol{P}^k - \eta \nabla_{\boldsymbol{X}} V_N(\boldsymbol{X}^k) \\ \boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \eta \nabla_{\boldsymbol{P}} K_N(\boldsymbol{P}^{k+1}) \end{cases} \text{(discretized)} \quad (2)$$

with the second-order *Störmer–Verlet* scheme obtained by symmetrizing this kick–drift splitting (Hairer et al., 2006).

3.2 Generalizing SE(3)-Transformer to Parametrized Hamiltonian Flows

A Parametric Instantiation of Hamiltonian. With θ denoting the model parameters, we retain the canonical structure of (1), replacing the quadratic kinetic energy by a learned density κ_θ and the potential by a transport value. The rows of $\boldsymbol{X}, \boldsymbol{P}$ now represent invariant atom features and latent momentum. Let μ be a probability measure over $(\xi, \boldsymbol{x}, \boldsymbol{p}) \in [0,1] \times \mathbb{R}^d \times \mathbb{R}^d$ with finite second moments, and let ν be its marginal on $(\xi, \boldsymbol{x})$. The fixed auxiliary label ξ preserves pair identity as features evolve and is used only in this distributional representation. For each head h, a bounded nonnegative measurable kernel $t_h(\xi, \xi')$ specifies the relations, with $t = (t_h)_{h=1}^H$ fixed during the reference dynamics.

The potential is the optimal value of a coupling problem, whose optimizer supplies the interaction weights. Let $\boldsymbol{z}_h$ project features to head dimension d_h. With $\Pi(\nu, \nu)$ denoting couplings with both marginals ν and KL denoting relative entropy, define

$$\begin{aligned} a_{\theta,h}(\xi, \boldsymbol{x}, \xi', \boldsymbol{x}' \mid t_h) &= \big[\boldsymbol{z}_h(\boldsymbol{x})^\top \boldsymbol{z}_h(\boldsymbol{x}')\big]/\sqrt{d_h} - \log\big(t_h(\xi, \xi') + \varepsilon_{\text{num}}\big), \\ \mathcal{V}_{\varepsilon_{\text{OT}},h}[\nu \mid t_h] &= \sup_{\pi_h \in \Pi(\nu,\nu)} \Big\{ \int a_{\theta,h} \, \mathrm{d}\pi_h - \varepsilon_{\text{OT}} \, \mathrm{KL}(\pi_h \| \nu \otimes \nu) \Big\}, \end{aligned} \quad (3)$$

where $\boldsymbol{z}_h(\boldsymbol{x}) = \boldsymbol{W}_h^\top \boldsymbol{x} + \boldsymbol{b}_h$ for $\boldsymbol{W}_h \in \mathbb{R}^{d \times d_h}$ and $\boldsymbol{b}_h \in \mathbb{R}^{d_h}$. The constants $\varepsilon_{\text{num}} > 0$ and $\varepsilon_{\text{OT}} > 0$ stabilize the logarithm and control entropic regularization, respectively. This is entropy-regularized optimal transport with cost $-a_{\theta,h}$, see also (Cuturi, 2013).

Extending the empirical representation above, let $\mu_N = N^{-1} \sum_{i=1}^N \delta_{(\xi_i, \boldsymbol{x}_i, \boldsymbol{p}_i)}$ for distinct labels ξ_i, with labeled feature marginal ν_N. Set $(\boldsymbol{T}_h)_{ij} = t_h(\xi_i, \xi_j)$ and $\mathbf{T} = (\boldsymbol{T}_h)_{h=1}^H$. Couplings in $\Pi(\nu_N, \nu_N)$ have masses π_{ij}/N, where the transport matrix $\boldsymbol{\pi} = (\pi_{ij})$ belongs to $\boldsymbol{\Pi} := \{\boldsymbol{\pi} \in \mathbb{R}_+^{N \times N} : \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\pi}^\top \mathbf{1} = \mathbf{1}\}$. For a score matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$, the finite transport value is

$$\Phi_{\varepsilon_{\text{OT}}}(\boldsymbol{S}) = \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \sum_{i,j} \pi_{ij} \big[S_{ij} - \varepsilon_{\text{OT}}(\log \pi_{ij} - 1)\big]. \quad (4)$$

Writing $\boldsymbol{Z}_h$ for the projected feature matrix, the distributional and particle Hamiltonians are

$$\mathcal{H}_\theta[\mu | t] = \int \kappa_\theta(\boldsymbol{p}) \, \mathrm{d}\mu + \frac{1}{2} \sum_{h=1}^H \sqrt{d_h} \mathcal{V}_{\varepsilon_{\text{OT}},h}[\nu \mid t_h] =: \mathcal{K}_\theta[\mu] + \mathcal{V}_\theta[\nu \mid t], \quad (5)$$

$$H_{\theta,N}(\boldsymbol{X}, \boldsymbol{P} | \mathbf{T}) = \sum_{i=1}^N \kappa_\theta(\boldsymbol{p}_i) + \frac{1}{2} \sum_{h=1}^H \sqrt{d_h} \Phi_{\varepsilon_{\text{OT}}}([\boldsymbol{Z}_h \boldsymbol{Z}_h^\top]/\sqrt{d_h} - \log(\boldsymbol{T}_h + \varepsilon_{\text{num}})), \quad (6)$$

where $\mathcal{K}_\theta, \mathcal{V}_\theta$ denote the kinetic and potential energy functionals, and the two terms of the particle Hamiltonian define $K_\theta(\boldsymbol{P})$ and $V_\theta(\boldsymbol{X} \mid \mathbf{T})$. The particle Hamiltonian equals $N\mathcal{H}_\theta[\mu_N \mid t]$ up to a

constant independent of $\boldsymbol{X}, \boldsymbol{P}$, as in the total-energy normalization of (1), with the exact relation derived in Appendix A.2. The projected features are $\boldsymbol{Z}_h = \boldsymbol{X}\boldsymbol{W}_h + \mathbf{1}\boldsymbol{b}_h^\top$, with matrix logarithms and scalar offsets applied entrywise. The kinetic density is $\kappa_\theta(\boldsymbol{p}) = \frac{1}{2}\|\sigma_{\sqrt{0.1}}(\boldsymbol{B}^\top\boldsymbol{p} + \boldsymbol{b})\|_2^2$ for learned $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$, where σ_a denotes entrywise leaky-ReLU with negative slope a. The slope $\sqrt{0.1}$ ensures that $\nabla_{\boldsymbol{p}}\kappa_\theta(\boldsymbol{p}) = \boldsymbol{B}\,\sigma_{0.1}(\boldsymbol{B}^\top\boldsymbol{p} + \boldsymbol{b})$, matching the kinetic drift. We take $\phi_{0.1}(u) = \int_0^u \sigma_{0.1}(s)\,\mathrm{d}s$, where $\sigma_{0.1}(u) = \max\{u, 0.1u\}$ is leaky-ReLU, applied entrywise.

Lie–Trotter realization. Fix θ and $\mathbf{T}$, and let $\boldsymbol{\pi}_h^\star$ be the unique optimizer of (4) for head h. The envelope identity $\nabla_{\boldsymbol{S}}\Phi_{\varepsilon_{\mathrm{OT}}}(\boldsymbol{S}) = \boldsymbol{\pi}^\star$, obtained from Danskin's theorem (Danskin, 1966), together with $\phi_{0.1}' = \sigma_{0.1}$, gives

$$\nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}) = \sigma_{0.1}(\boldsymbol{P}\boldsymbol{B} + \mathbf{1}\boldsymbol{b}^\top)\boldsymbol{B}^\top \quad \text{and} \quad \nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T}) = \frac{1}{2}\sum\nolimits_{h=1}^{H}\big(\boldsymbol{\pi}_h^\star + (\boldsymbol{\pi}_h^\star)^\top\big)\boldsymbol{Z}_h\boldsymbol{W}_h^\top,$$

where the prefactor $\sqrt{d_h}/2$ in the potential accounts for score normalization and the two feature arguments. Symmetric relations imply symmetric optimal plans, reducing the force to $-\sum_h \boldsymbol{\pi}_h^\star \boldsymbol{Z}_h \boldsymbol{W}_h^\top$. Substitution into (2) gives the reference kick–drift step, with momentum retention incorporated by applying $\boldsymbol{P} \leftarrow e^{-\gamma\eta}\boldsymbol{P}$ before the kick, where $\gamma \geq 0$ and $\gamma = 0$ recovers the conservative splitting. Appendices A.3 and A.4 derive the gradients and splitting, respectively.

Implementation. HamFormer uses finite Sinkhorn iterations for transport, normalized learned maps for the kick and drift, and a token-wise residual feed-forward network (FFN). Its stages alternate with EGNN stages, which use the final relations as invariant edge information for SE(3)-equivariant coordinate refinement, as shown in Figure 2.

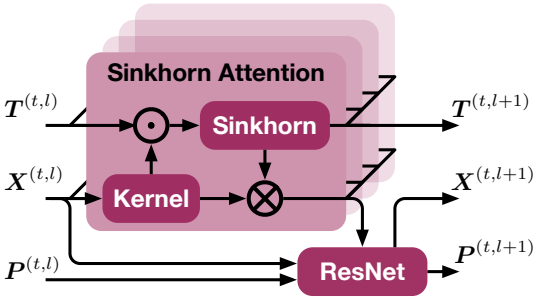

Figure 2: HamFormer architecture.

Let $\boldsymbol{c}_i^{\mathrm{in}} \in \mathbb{R}^3$ be the input coordinates and $D_{ij}^{\mathrm{in}} = \|\boldsymbol{c}_i^{\mathrm{in}} - \boldsymbol{c}_j^{\mathrm{in}}\|$. For each head, let $g_h(e_{ij})$ embed the unordered atom-pair type e_{ij} into a scalar. Write $\mathrm{Sinkhorn}_{\varepsilon_{\mathrm{OT}},I}(\boldsymbol{S})$ for I alternating row and column scaling iterations on $\exp(\boldsymbol{S}/\varepsilon_{\mathrm{OT}})$, targeting marginals $\mathbf{1}/N$. Assuming $\max_{u,v} D_{uv}^{\mathrm{in}} > 0$, initialize

$$\boldsymbol{T}_h^{(0)} = N\ \mathrm{Sinkhorn}_{\varepsilon_{\mathrm{OT}},I}\Big(\frac{[g_h(e_{ij})D_{ij}^{\mathrm{in}}]_{ij}}{\max_{u,v} D_{uv}^{\mathrm{in}}}\Big) \tag{7}$$

that N rescales the target row and column sums to one, and we use $I = 3$ iterations to approximate these constraints. At layer ℓ, use layer-indexed parameters and let $\mathrm{LN}_{\mathrm{in},\ell}$, $\mathrm{LN}_{P,\ell}$, and $\mathrm{LN}_{X,\ell}$ denote LayerNorm on the input features, interaction message, and drift. The relation update is

$$\begin{aligned}\boldsymbol{Z}_h^{(\ell)} &= \mathrm{LN}_{\mathrm{in},\ell}(\boldsymbol{X}^{(\ell)})\boldsymbol{W}_h^{(\ell)} + \mathbf{1}(\boldsymbol{b}_h^{(\ell)})^\top,\\ \boldsymbol{T}_h^{(\ell+1)} &= N\ \mathrm{Sinkhorn}_{\varepsilon_{\mathrm{OT}},I}\Big(\boldsymbol{Z}_h^{(\ell)}(\boldsymbol{Z}_h^{(\ell)})^\top/\sqrt{d_h} - \log(\boldsymbol{T}_h^{(\ell)} + \varepsilon_{\mathrm{num}})\Big),\end{aligned}$$

with the updated relations supplying the current interaction weights and conditioning the next layer. At $\varepsilon_{\mathrm{OT}} = 1$, the kernel divides the exponentiated affinity entrywise by $\boldsymbol{T}_h^{(\ell)} + \varepsilon_{\mathrm{num}}$. The initialization and updates depend only on invariant attributes, distances, and features, so the relations remain invariant under rigid motions.

With learned retention $\alpha_\ell \in \mathbb{R}$ and post-drift features $\boldsymbol{U}$, the state update is

$$\boldsymbol{P}^{(\ell+1)} = -\tanh(\alpha_\ell)\boldsymbol{P}^{(\ell)} + \mathrm{LN}_{P,\ell}\Big(-\sum\nolimits_{h=1}^{H}\boldsymbol{T}_h^{(\ell+1)}\boldsymbol{Z}_h^{(\ell)}(\boldsymbol{W}_h^{(\ell)})^\top\Big),\ \ \boldsymbol{X}^{(\ell+1)} = \boldsymbol{U} + \mathrm{FFN}_\ell(\boldsymbol{U}), \tag{8}$$

where $\boldsymbol{U} = \boldsymbol{X}^{(\ell)} + \mathrm{LN}_{X,\ell}\big[\sigma_{0.1}(\boldsymbol{P}^{(\ell+1)}\boldsymbol{B}_\ell + \mathbf{1}\boldsymbol{b}_\ell^\top)\boldsymbol{B}_\ell^\top\big]$. Only the interaction branch of the momentum update is normalized. The coefficient $-\tanh(\alpha_\ell) \in (-1, 1)$ is shared within each layer and allows signed retention, with $\alpha_\ell = 0$ at initialization. Also, it has no effect in the first layer because the initial momentum is $\boldsymbol{P}^{(0)} = \mathbf{0}$. Learned maps and normalizations replace explicit step sizes.

For hidden width d_{FF}, let $\boldsymbol{W}_{1,\ell} \in \mathbb{R}^{d\times d_{\mathrm{FF}}}$ and $\boldsymbol{W}_{2,\ell} \in \mathbb{R}^{d_{\mathrm{FF}}\times d}$ have compatible biases $\boldsymbol{b}_{1,\ell}, \boldsymbol{b}_{2,\ell}$. With input LayerNorm $\mathrm{LN}_{F,\ell}$, activation and output dropout $\mathrm{Drop}_{a,\ell}, \mathrm{Drop}_{o,\ell}$, and hidden activation

$\boldsymbol{G}$, define $\mathrm{FFN}_\ell(\boldsymbol{U}) = \mathrm{Drop}_{o,\ell}\big(\boldsymbol{G}\boldsymbol{W}_{2,\ell} + \mathbf{1}\boldsymbol{b}_{2,\ell}^\top\big)$, where $\boldsymbol{G} = \mathrm{Drop}_{a,\ell}\big[\mathrm{GELU}(\mathrm{LN}_{F,\ell}(\boldsymbol{U})\boldsymbol{W}_{1,\ell} + \mathbf{1}\boldsymbol{b}_{1,\ell}^\top)\big]$. Activation and output dropout rates are 0 and 0.1 during training, with both disabled during evaluation.

Before output normalization, the tied drift equals the reference kinetic gradient. With fixed relations and exact symmetric transport plans, the interaction message equals the reference force with respect to normalized features. The implementation omits the input LayerNorm Jacobian, and finite Sinkhorn iterations need not preserve plan symmetry. These changes, together with signed retention, varying parameters and relations, and the FFN, preserve the update order without guaranteeing symplecticity, energy conservation, or energy decrease. Fixed relations are assumed only in the reference derivation. Training differentiates through the relation updates, normalization, and Sinkhorn iterations using the supervised loss (see also, Appendix A.6 for more details).

The combined gradient of the multi-head OT potential plays the interaction role of multi-head attention in an SE(3)-Transformer: it supplies the force that updates latent momentum. The tied tokenwise feedforward map realizes the kinetic gradient and advances features using the updated momentum. Together, these operations form the reference kick–drift structure. For fixed parameters and relations, quadratic kinetic energy gives $\ddot{\boldsymbol{X}} = -\nabla_{\boldsymbol{X}} V_\theta$, introducing second-order feature dynamics in contrast to the first-order flows in the single-head analyses of Sinkformers and WGFormer (Sander et al., 2022; Wang et al., 2025). The learned kinetic energy generalizes the momentum-to-velocity relation. The residual FFN and EGNN extend this reference through learned feature corrections and equivariant coordinate updates. This provides a physics-aware organization of the complete architecture, connecting attention, momentum, feedforward processing, and molecular geometry.

Equivariant coordinate refinement. The initial relations depend only on atom-pair types and distances, and HamFormer preserves invariance of the features, momentum, and relations under rigid motions. Each EGNN stage (Satorras et al., 2021) uses these invariant quantities to construct coordinate displacements. At EGNN layer q, learned pair messages produce a residual feature update using atom features, current and stage-entry squared distances, and final HamFormer relations. A separate MLP uses the updated features and the same distances and relations to produce invariant scalar weights $w_{ij}^{(q)}$. The stage-entry distances and relations remain fixed throughout the stage. Coordinates are updated by $\boldsymbol{c}_i^{(q+1)} = \boldsymbol{c}_i^{(q)} + \sum_j w_{ij}^{(q)}(\boldsymbol{c}_i^{(q)} - \boldsymbol{c}_j^{(q)})/(\|\boldsymbol{c}_i^{(q)} - \boldsymbol{c}_j^{(q)}\| + \varepsilon_{\mathrm{dist}})$, where $\varepsilon_{\mathrm{dist}} > 0$ stabilizes the denominator and scalar normalization factors are absorbed into $w_{ij}^{(q)}$. The scalar coefficients are invariant, while relative vectors rotate and are unchanged by translation, giving an SE(3)-equivariant update. Appendix A.5 proves equivariance of the complete pipeline. EGNN leaves momentum unchanged and passes its feature output directly to the next HamFormer stage. Figure 1 shows the complete architecture.

Physical interpretation. Compared with the first-order transport-flow interpretations of Sinkformers and WGFormer (Sander et al., 2022; Wang et al., 2025), HamFormer introduces an independent momentum state between attention and feature motion. The combined gradient of the multi-head transport potential supplies the momentum kick, while the tied feedforward map supplies the kinetic drift. For fixed parameters and relations, quadratic kinetic energy gives $\ddot{\boldsymbol{X}} = -\nabla_{\boldsymbol{X}} V_\theta$, so the interaction force determines acceleration. The learned kinetic energy generalizes the momentum-to-velocity relation. For ground-state conformation prediction, momentum carries earlier interaction effects across refinement steps, while evolving relations retain atom-pair information. Transport couples the atom features across the molecule, and EGNN uses the resulting features and relations to produce equivariant coordinate displacements (Satorras et al., 2021). Table 1 summarizes these roles. The mechanical interpretation concerns latent features and momentum, with molecular coordinates updated by EGNN.

Table 1: Physical roles of HamFormer components.

Component	Form	Physics	State	Molecular Role
OT Attention	$-\nabla_{\boldsymbol{X}} V_\theta$	Force	$\boldsymbol{P}$	Atom Interactions
Kinetic Map	$\nabla_{\boldsymbol{P}} K_\theta$	Velocity	$\boldsymbol{X}$	Feature Motion
Retention	$-\tanh(\alpha_\ell)\boldsymbol{P}$	Signed Carry	$\boldsymbol{P}$	Refinement Memory
Relations	$\boldsymbol{T}_h^{(\ell)} \mapsto \boldsymbol{T}_h^{(\ell+1)}$	Coupling	$\mathbf{T}$	Pair Information
EGNN	$\boldsymbol{c}_i - \boldsymbol{c}_j$	Direction	$\boldsymbol{C}$	Coordinate Refinement

Table 2 compares the reference dynamics and architectural features. YuriiFormer introduces momentum through a Nesterov template, while SympFormer uses an attention-dependent kinetic metric.

Table 2: Reference dynamics under the respective theoretical assumptions and features of the reported architectures. OT denotes entropic optimal transport, and pair memory denotes relations retained across layers. The SympFormer row reports its softmax variant.

Model	Reference dynamics	Interaction	Momentum parametrization	Pair memory	SE(3)
Sinkformers (2022)	Wasserstein PDE: $\partial_t \rho = \nabla \cdot (\rho \nabla \mathcal{F}'(\rho))$	OT	–	×	×
WGFormer (2025)	Wasserstein PDE: $\partial_t \rho = \nabla \cdot (\rho \nabla \mathcal{F}'(\rho))$	OT	–	✓	✓
YuriiFormer (2026)	Nesterov: $\Delta \boldsymbol{X}^{k+1} = \beta_k \Delta \boldsymbol{X}^k - \eta_k \nabla E(\boldsymbol{X}^k + \alpha_k \Delta \boldsymbol{X}^k)$	Softmax	Learned coefficients	×	×
SympFormer (2026)	Hamiltonian PDE: $\partial_t \rho = \mathcal{H}'_S, \quad \partial_t S = -\mathcal{H}'_\rho - \gamma(t) S$	Softmax	Metric, damping	×	×
HamFormer **(ours)**	Hamiltonian PDE: $\partial_t \mu + \{\mu, h_\mu\} = 0$	OT	Kinetic map, signed retention	✓	✓

HamFormer combines a separable reference Hamiltonian at fixed relations with entropic transport interactions, persistent pair relations, and SE(3)-equivariant coordinate refinement.

In the Nesterov template, $\Delta \boldsymbol{X}^k = \boldsymbol{X}^k - \boldsymbol{X}^{k-1}$, and $\alpha_k, \beta_k, \eta_k$ control lookahead, retention, and step size for the reference objective E. In the Hamiltonian rows, primes denote first variations, and $h_\mu = \mathcal{H}'_\theta[\mu \mid t]$ is evaluated at fixed parameters and relations. The canonical Poisson bracket $\{f, g\} = \nabla_{\boldsymbol{x}} f \cdot \nabla_{\boldsymbol{p}} g - \nabla_{\boldsymbol{p}} f \cdot \nabla_{\boldsymbol{x}} g$ describes transport in feature–momentum space.

With feature dimensions, head count, and Sinkhorn iterations fixed, each dense HamFormer layer requires $O(N^2)$ time and storage, as detailed in Remark A.4. In practice, we use three Sinkhorn iterations and reuse each head's projected features for both score construction and transport aggregation.

Implementation. The practical layer retains the reference kick–drift order, with finite Sinkhorn iterations, normalized learned maps, signed retention, and a residual FFN. The relative magnitudes of the kinetic and potential energies do not directly measure their influence on the implemented dynamics, which depend on their gradients and the subsequent normalization. These extensions do not guarantee conservation or decrease of the reference Hamiltonian. Appendix A.6 gives the complete pipeline, both algorithms, and the supervised training objective.

4 Experiments

To validate the effectiveness and physical interpretability of HamFormer, we evaluate our model on molecular ground-state conformation prediction following the established experimental protocol of WGFormer Wang et al. (2025). In addition to quantitative comparisons with existing 2D graph models and 3D equivariant architectures, we provide an empirical analysis of the latent dynamics learned by HamFormer, focusing on the layerwise evolution of phase-space energy and the role of the latent momentum state.

4.1 Experimental Setup

Datasets and splitting strategy. We evaluate our model on two standard molecular conformation benchmarks: QM9 and the larger-scale Molecule3D (Xu et al., 2021). QM9 comprises approximately 134k small organic molecules with up to nine heavy atoms, with ground-state geometries calculated via density functional theory (DFT). Molecule3D contains substantially larger and more flexible drug-like molecules. Following the benchmark protocol in Wang et al. (2025), we adopt the standard training, validation, and test splits for QM9, and evaluate on the standard random split for Molecule3D.

Baselines. Following Wang et al. (2025), we compare HamFormer with two categories of baseline methods: (1) 2D molecular graph models, including GINE (Hu et al., 2019), GATv2 (Brody et al., 2022), GPS (Rampášek et al., 2022), and GTMGC (Xu et al., 2024), which generate 3D coordinates directly from 2D molecular graphs; and (2) 3D conformation refinement methods, including ConfOpt with TwoAtom and ThreeAtom energy functions (Guan et al., 2022), ConfOpt variants using SE(3)-Transformer (Fuchs et al., 2020) and EGNN (Satorras et al., 2021) backbones, and WGFormer (Wang et al., 2025).

Evaluation metrics. Following the established evaluation protocol in Xu et al. (2021); Wang et al. (2025), we evaluate predicted conformations using three complementary geometric metrics: (1)

D-MAE (Distance Mean Absolute Error), (2) **D-RMSE** (Distance Root Mean Square Error), which measure the accuracy of internal molecular geometry via pairwise interatomic distances, and (3) **C-RMSD** (Coordinate Root Mean Square Deviation), which assesses global 3D structural alignment by computing the coordinate error against the ground truth after optimal rotation and translation. All metrics are reported in Ångströms (Å), where lower values indicate better predictive quality.

4.2 Main Results

Table 3 reports the quantitative comparison on Molecule3D and QM9. Compared to WGFormer (Wang et al., 2025), HamFormer achieves lower errors across all metrics on QM9 (e.g., test C-RMSD of 0.198 Å vs. 0.206 Å), while delivering improved D-MAE and C-RMSD on Molecule3D. HamFormer attains these results with 16.3M parameters, an approximate 79% reduction compared to WGFormer (78.9M). Compared to the first-order Wasserstein gradient flow in WGFormer, incorporating second-order phase-space dynamics provides a more parameter-efficient inductive bias for molecular conformation refinement.

Table 3: Performance comparison on Molecule3D and QM9 datasets. ↓ indicates that lower values are better. **Bold** indicates the best performance, and underline indicates the second best. Baseline results follow Wang et al. (2025).

Dataset	Model	Validation			Test		
		D-MAE ↓	D-RMSE ↓	C-RMSD ↓	D-MAE ↓	D-RMSE ↓	C-RMSD ↓
Molecule3D (Random)	GINE(2D)	0.590	1.014	1.116	0.592	1.018	1.116
	GATv2(2D)	0.563	0.983	1.082	0.564	0.986	1.083
	GPS(2D)	0.528	0.909	1.036	0.529	0.911	1.038
	GTMGC(2D)	0.432	0.719	0.712	0.433	0.721	0.713
	SE(3)-Transformer	0.466	0.712	0.800	0.467	0.774	0.802
	EGNN	0.461	0.704	0.798	0.462	0.766	0.799
	ConfOpt-TwoAtom	0.438	0.668	0.748	0.438	0.670	0.749
	ConfOpt-ThreeAtom	0.429	0.659	0.734	0.430	0.661	0.736
	WGFormer	0.391	**0.649**	0.662	0.392	**0.652**	0.664
	HamFormer	**0.386**	0.653	**0.652**	**0.387**	0.656	**0.654**
QM9	GINE(2D)	0.357	0.673	0.685	0.357	0.669	0.693
	GATv2(2D)	0.339	0.663	0.661	0.339	0.659	0.666
	GPS(2D)	0.326	0.644	0.662	0.326	0.640	0.666
	GTMGC(2D)	0.262	0.468	0.362	0.264	0.470	0.367
	SE(3)-Transformer	0.254	0.451	0.296	0.256	0.455	0.303
	EGNN	0.248	0.442	0.257	0.251	0.449	0.265
	ConfOpt-TwoAtom	0.245	0.439	0.245	0.248	0.444	0.254
	ConfOpt-ThreeAtom	0.241	0.433	0.237	0.244	0.438	0.246
	WGFormer	0.223	0.416	0.198	0.227	0.422	0.206
	HamFormer	**0.215**	**0.414**	**0.191**	**0.218**	**0.419**	**0.198**

4.3 Ablation Studies and Further Analysis

Momentum Evolution over Layers. Figure 3 examines the six HamFormer layers separated into two alternating stages (H1–H3 for Stage 1 and H4–H6 for Stage 2). As expected in a dissipative physical system, the latent dynamics exhibit a clear two-stage relaxation behavior: 1) across the network as a whole, both the potential energy and total Hamiltonian follow an overall downward trajectory, achieving net energy dissipation toward a stable ground-state equilibrium; 2) the latent kinetic energy remains relatively suppressed in Stage 1, but experiences a prominent transient surge at the onset of Stage 2 (layer H4, marked by the dashed line) as stored potential energy is rapidly unleashed, followed by smooth dissipation across the remaining layers.

Architectural and Momentum Ablations. Figure 4 (left) systematically investigates the necessity of individual components on the Molecule3D random test set: (1) Inference-time momentum retention: Zeroing out the learned momentum at inference ($\rho_l = 0$) induces pronounced performance degradation across all metrics, with test C-RMSD rising from 0.654 Å to 0.714 Å and D-MAE increasing from 0.387 Å to 0.420 Å. This demonstrates that preserving momentum across layers acts as an indispensable dynamical memory, effectively preventing error accumulation during iterative refinement.

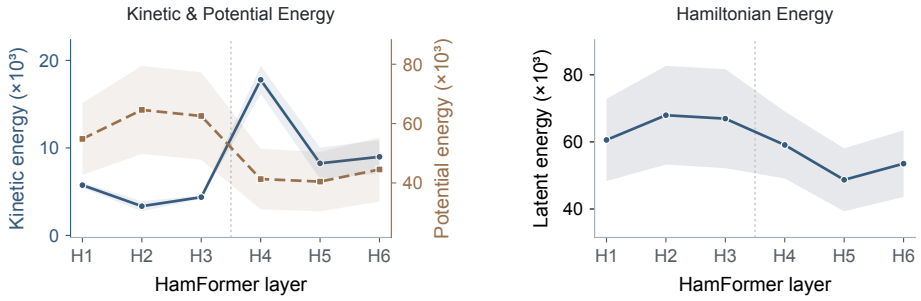

Figure 3: Layer-wise evolution of the latent kinetic energy, potential energy, and total Hamiltonian evaluated on 1,000 random molecules from the Molecule3D test set. Solid curves denote the mean, and shaded regions represent one standard deviation across molecules. The dashed line marks the transition between Stage 1 (layers H1–H3) and Stage 2 (layers H4–H6).

(2) Hamiltonian block ablation: Removing HamFormer and using pure EGNN, which eliminates phase-space feature updates and relies solely on local message-passing—causes notable performance drops (e.g., test C-RMSD rises to 0.697 Å). This confirms that our phase-space Hamiltonian flow provides an essential non-local inductive bias for capturing global molecular geometry.

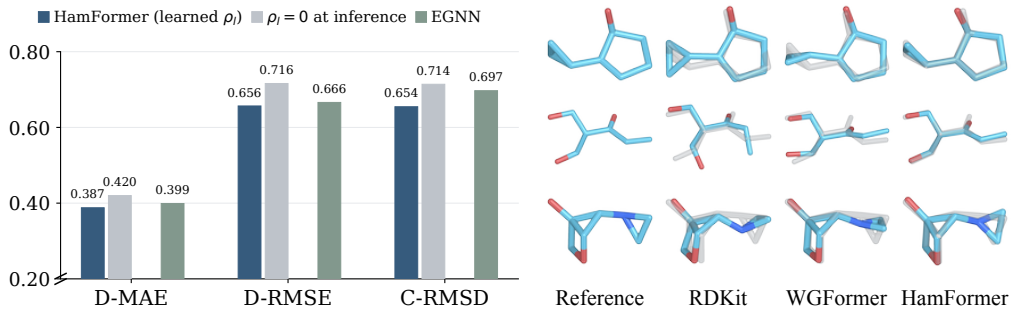

Figure 4: Qualitative conformation visualizations and quantitative ablation studies on Molecule3D. **Left:** Quantitative ablation on the test set comparing the full HamFormer with zeroed inference momentum ($\rho_l = 0$) and the EGNN baseline. **Right:** Visual comparison of predicted 3D molecular geometries against DFT references for RDKit, WGFormer, and HamFormer.

Conformation Visualizations. Figure 4 (right) presents visual comparisons of predicted 3D molecular conformations aligned against ground-truth reference structures for representative molecules. Compared WGFormer, the conformations generated by HamFormer consistently align more closely with the true reference geometries, exhibiting fewer steric clashes and more accurate bond lengths and angles.

5 Conclusion and Future Work

We presented HamFormer, which brings Hamiltonian-inspired latent dynamics into SE(3)-equivariant conformation refinement through the joint evolution of features, momentum, and transport-based relations. Its integration with EGNN yields improved refinement on QM9 and Molecule3D, suggesting that structured latent dynamics can complement geometric message passing. These findings motivate further study of the balance between representational capacity and computational efficiency, particularly as molecular size and structural complexity increase. Extending this framework beyond small molecules remains a promising direction for assessing its scalability and broader applicability.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, grammar polishing, and preliminary code structuring. We have not used generative AI tools for core mathematical formulation design, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for code implementation and debugging. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors, and all text was carefully checked for accuracy and academic integrity. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research focuses on computational molecular conformation prediction using publicly available datasets (QM9 and Molecule3D). Our work does not involve human subjects, sensitive personal data, or physical experiments that pose safety, environmental, or bioethical concerns. There are no known conflicts of interest or sponsorship issues associated with this research.

REPRODUCIBILITY STATEMENT

To ensure the complete reproducibility of our experimental results, we have taken the following measures: 1) **Model Architecture and Parameters:** Detailed mathematical derivations, architectural designs, and complete configuration parameters required for reproduction are clearly documented in the Appendix. 2) **Code and Checkpoints:** We provide the key implementation components and pre-trained model weights in the supplementary materials, facilitating the reproduction of the experimental results presented in this paper.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations*, 2022.

Ricky T.Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6572–6583, 2018.

Shui-Nee Chow, Wuchen Li, and Haomin Zhou. Wasserstein Hamiltonian flows. *Journal of Differential Equations*, 268(3):1205–1219, 2020.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2292–2300, 2013.

John M. Danskin. The Theory of Max-Min, with Applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Lawrence C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020.

Peyré Gabriel and Cuturi Marco. Computational optimal transport: With applications to data science. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Jingchu Gai, Nai-Chieh Huang, and Jiayun Wu. Momentum Streams for Optimizer-Inspired Transformers. *arXiv preprint arXiv:2605.24425*, 2026.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on Transformers. *Bulletin (New Series) of the American Mathematical Society*, 62(3):427, 2025.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiaqi Guan, Wesley Wei Qian, Qiang Liu, Wei-Ying Ma, Jianzhu Ma, and Jian Peng. Energy-inspired molecular conformation optimization. In *International Conference on Learning Representations*, 2022.

Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 3(1):805–882, 2006.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy Transformer. volume 36, pp. 27532–27559, 2023.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for Pre-training Graph Neural Networks. 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. In *International Conference on Learning Representations*, 2023.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 14501–14515, 2022.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Momentum residual neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9276–9287, 2021.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

Viktor Stein, Wuchen Li, and Gabriele Steidl. SympFormer: Accelerated attention blocks via inertial dynamics on density manifolds. *arXiv preprint arXiv:2603.16535*, 2026.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Fanmeng Wang, Minjie Cheng, and Hongteng Xu. WGFormer: An SE(3)-transformer driven by Wasserstein gradient flows for molecular ground-state conformation prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62652–62671. PMLR, 13–19 Jul 2025.

Hao Wu, Shu Liu, Xiaojing Ye, and Haomin Zhou. Parameterized Wasserstein Hamiltonian flow. *SIAM Journal on Numerical Analysis*, 63(1):360–395, 2025.

Guikun Xu, Yongquan Jiang, PengChuan Lei, Yan Yang, and Jim Chen. GTMGC: Using graph Transformer to predict molecule's ground-state conformation. In *International Conference on Learning Representations*, volume 2024, pp. 47712–47731, 2024.

Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, Maho Nakata, and Shuiwang Ji. Molecule3D: A Benchmark for Predicting 3D Geometries from Molecular Graphs. *arXiv preprint arXiv:2110.01717*, 2021.

Yongyi Yang and David P. Wipf. Transformers from an Optimization Perspective. volume 35, pp. 36958–36971, 2022.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*, 2023.

Aleksandr Zimin, Yury Polyanskiy, and Philippe Rigollet. YuriiFormer: A suite of Nesterov-Accelerated Transformers. *arXiv preprint arXiv:2601.23236*, 2026.

APPENDIX CONTENTS

A Theoretical Details **14**

A.1 Hamiltonian Dynamics on Density Space 14

A.2 Entropic Interaction and Particle Representation 15

A.3 Gradients of the Reference Hamiltonian 16

A.4 Lie–Trotter Splitting and the Implemented Layer 17

A.5 SE(3)-Equivariance of the Complete Pipeline 18

A.6 HamFormer for Molecular Conformation Optimization 18

A.7 Transport Computation, Training, and Complexity 19

B Experimental Details **20**

B.1 Datasets 20

B.2 Evaluation 20

B.3 Hyperparameters 21

B.4 Ablation Study on Model Components 21

B.5 Inference-Time Momentum-Retention Ablation 21

B.6 Layer-wise Energy Evolution on the QM9 Dataset 22

B.7 Additional Visualizations 22

A Theoretical Details

Section A.1 derives the density equations, their characteristics, and conservation properties. Sections A.2 and A.3 establish the particle representation of the learned Hamiltonian and compute its force and drift. Section A.4 derives the Lie–Trotter updates and examines momentum retention, and Section A.5 proves SE(3)-equivariance of the complete pipeline. Section A.6 presents the molecular refinement algorithms and supervised objective. Finally, Section A.7 details Sinkhorn scaling, rotational alignment, and computational complexity. Continuum calculations assume sufficient regularity to justify differentiation and integration by parts.

A.1 Hamiltonian Dynamics on Density Space

Consider the Wasserstein Hamiltonian in (1) on a periodic domain Ω,

$$\mathcal{H}_W(\rho, S) = \frac{1}{2}\int_\Omega \|\nabla S(\boldsymbol{x})\|^2 \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \mathcal{F}(\rho).$$

Let r be a smooth density perturbation with $\int_\Omega r\,\mathrm{d}\boldsymbol{x} = 0$, and let ψ be a smooth perturbation of S. Writing D for a directional derivative, the two variations are

$$D_\rho \mathcal{H}_W(\rho, S)[r] = \int_\Omega \left(\frac{1}{2}\|\nabla S\|^2 + \mathcal{F}'(\rho)\right) r\,\mathrm{d}\boldsymbol{x},$$
$$D_S \mathcal{H}_W(\rho, S)[\psi] = \int_\Omega \rho\,\nabla S\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla\cdot(\rho\nabla S)\,\psi\,\mathrm{d}\boldsymbol{x},$$

where periodicity removes the boundary term. Since density perturbations have zero integral, the first variation with respect to ρ is determined up to a spatial constant.

The canonical pairing identifies the density derivative with the variation in S, and the potential derivative with the negative variation in ρ (Chow et al., 2020). Choosing the time-dependent additive constant in S_t gives

$$\begin{cases} \partial_t \rho_t = -\nabla\cdot(\rho_t \nabla S_t), \\ \partial_t S_t = -\frac{1}{2}\|\nabla S_t\|^2 - \mathcal{F}'(\rho_t), \end{cases}$$

where the first equation transports probability with velocity ∇S_t, while the second determines the evolution of the conjugate potential.

Let $\boldsymbol{x}(t)$ satisfy $\dot{\boldsymbol{x}}(t) = \nabla S_t(\boldsymbol{x}(t))$, and define $\boldsymbol{p}(t) = \nabla S_t(\boldsymbol{x}(t))$. The chain rule along this trajectory yields

$$\begin{aligned} \dot{\boldsymbol{p}}(t) &= \nabla\partial_t S_t(\boldsymbol{x}(t)) + \nabla^2 S_t(\boldsymbol{x}(t))\dot{\boldsymbol{x}}(t) \\ &= \nabla\left(\partial_t S_t + \frac{1}{2}\|\nabla S_t\|^2\right)(\boldsymbol{x}(t)) \\ &= -\nabla\mathcal{F}'(\rho_t)(\boldsymbol{x}(t)). \end{aligned}$$

Thus, with $h_t(\boldsymbol{x}, \boldsymbol{p}) = \frac{1}{2}\|\boldsymbol{p}\|^2 + \mathcal{F}'(\rho_t)(\boldsymbol{x})$, the characteristic equations are

$$\dot{\boldsymbol{x}} = \nabla_{\boldsymbol{p}} h_t = \boldsymbol{p}, \qquad \dot{\boldsymbol{p}} = -\nabla_{\boldsymbol{x}} h_t = -\nabla\mathcal{F}'(\rho_t)(\boldsymbol{x}),$$

which recover Hamilton's equations through the classical Hamilton–Jacobi construction (Evans, 2022, Sec. 3.3.1). Integrating the continuity equation over Ω gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \rho_t\,\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla\cdot(\rho_t\nabla S_t)\,\mathrm{d}\boldsymbol{x} = 0.$$

If $\mathcal{F}$ has no explicit time dependence, differentiating the Hamiltonian and integrating by parts also gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_W(\rho_t, S_t) &= \int_\Omega \left(\frac{1}{2}\|\nabla S_t\|^2 + \mathcal{F}'(\rho_t)\right)\partial_t\rho_t\,\mathrm{d}\boldsymbol{x} + \int_\Omega \rho_t\nabla S_t\cdot\nabla\partial_t S_t\,\mathrm{d}\boldsymbol{x} \\ &= \int_\Omega \left(-\partial_t S_t\,\partial_t\rho_t + \partial_t\rho_t\,\partial_t S_t\right)\mathrm{d}\boldsymbol{x} = 0. \end{aligned}$$

The conserved quantity is the total functional $\mathcal{H}_W$. The characteristic Hamiltonian h_t can still depend on time through ρ_t.

A.2 Entropic Interaction and Particle Representation

The learned Hamiltonian in (5) combines a parametrized kinetic energy density κ_θ with an entropic transport potential. We first verify that this energy is finite and then derive its particle representation.

For the interaction term, the inequality

$$\left|\boldsymbol{z}_h(\boldsymbol{x})^\top \boldsymbol{z}_h(\boldsymbol{x}')\right| \le \frac{1}{2}\|\boldsymbol{z}_h(\boldsymbol{x})\|^2 + \frac{1}{2}\|\boldsymbol{z}_h(\boldsymbol{x}')\|^2$$

and the finite second moments of the fixed marginals bound the affinity integral uniformly over admissible couplings. The logarithmic relation term is bounded because t_h is bounded and $\varepsilon_{\mathrm{num}} > 0$. Relative entropy is nonnegative, and the product coupling $\nu \otimes \nu$ is admissible with zero relative entropy. These observations give finite upper and lower bounds for the interaction value. The kinetic density has at most quadratic growth, so its integral is finite as well.

For distinct fixed atom identifiers ξ_i, consider the empirical measures

$$\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{(\xi_i,\boldsymbol{x}_i,\boldsymbol{p}_i)} \quad \text{and} \quad \nu_N = \frac{1}{N}\sum_{i=1}^N \delta_{(\xi_i,\boldsymbol{x}_i)}.$$

The identifiers distinguish atoms with identical features. For a coupling of ν_N with itself, let π_{ij}/N denote the mass assigned to the ordered atom pair (i, j), where

$$\boldsymbol{\pi} \in \boldsymbol{\Pi} := \left\{\boldsymbol{\pi} \in \mathbb{R}_+^{N\times N} : \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\pi}^\top\mathbf{1} = \mathbf{1}\right\},$$

so that $\boldsymbol{\pi}$ has unit row and column sums, whereas $\boldsymbol{\pi}/N$ is a probability coupling. With $(\boldsymbol{T}_h)_{ij} = t_h(\xi_i, \xi_j)$ and $\boldsymbol{Z}_h = \boldsymbol{X}\boldsymbol{W}_h + \mathbf{1}\boldsymbol{b}_h^\top$, the sampled affinity is

$$\boldsymbol{S}_h = \frac{\boldsymbol{Z}_h\boldsymbol{Z}_h^\top}{\sqrt{d_h}} - \log(\boldsymbol{T}_h + \varepsilon_{\mathrm{num}}),$$

with finite transport value $\Phi_{\varepsilon_{\mathrm{OT}}}(\boldsymbol{S}_h)$ from (4). The following identity accounts for the difference between the distributional and matrix entropy conventions.

Lemma A.1 (Particle normalization). *For μ_N and its corresponding relation tensor,*

$$H_{\theta,N} = N\mathcal{H}_\theta[\mu_N \mid t] + \varepsilon_{\mathrm{OT}}N(1+\log N)\sum_{h=1}^H \frac{\sqrt{d_h}}{2},$$

where the additive correction is independent of $\boldsymbol{X}$ and $\boldsymbol{P}$.

Proof. The product measure $\nu_N \otimes \nu_N$ assigns mass $1/N^2$ to each ordered atom pair. Using $0\log 0 = 0$ and $\sum_{ij}\pi_{ij} = N$, a coupling with masses π_{ij}/N satisfies

$$\mathrm{KL}(\pi_h \| \nu_N \otimes \nu_N) = \sum_{i,j}\frac{\pi_{ij}}{N}\log\left(\frac{\pi_{ij}/N}{1/N^2}\right) = \frac{1}{N}\sum_{i,j}\pi_{ij}\log\pi_{ij} + \log N.$$

Consequently,

$$\begin{aligned} N\mathcal{V}_{\varepsilon_{\mathrm{OT}},h}[\nu_N \mid t_h] &= \max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big\{\sum_{i,j}\pi_{ij}(S_h)_{ij} - \varepsilon_{\mathrm{OT}}\sum_{i,j}\pi_{ij}\log\pi_{ij}\Big\} - \varepsilon_{\mathrm{OT}}N\log N \\ &= \Phi_{\varepsilon_{\mathrm{OT}}}(\boldsymbol{S}_h) - \varepsilon_{\mathrm{OT}}N(1+\log N). \end{aligned}$$

The kinetic term satisfies

$$N\int \kappa_\theta(\boldsymbol{p})\,\mathrm{d}\mu_N = \sum_{i=1}^N \kappa_\theta(\boldsymbol{p}_i) = K_\theta(\boldsymbol{P}),$$

and summing the interaction identity over heads with weights $\sqrt{d_h}/2$ proves the result. □

This is an exact evaluation on empirical measures and does not require a limit as $N \to \infty$. After total-energy normalization, the additive constant has no effect on the feature or momentum gradients.

A.3 Gradients of the Reference Hamiltonian

For matrices of the same size, write $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. The entropy term makes the finite coupling problem strictly concave in its optimization variable.

Lemma A.2 (Transport gradient). *For finite $\boldsymbol{S}$ and $\varepsilon_{\mathrm{OT}} > 0$, the problem in (4) has a unique positive optimizer $\boldsymbol{\pi}^\star$, and*

$$\nabla_{\boldsymbol{S}} \Phi_{\varepsilon_{\mathrm{OT}}}(\boldsymbol{S}) = \boldsymbol{\pi}^\star.$$

If $\boldsymbol{S}$ is symmetric, then $\boldsymbol{\pi}^\star$ is symmetric.

Proof. The admissible set $\boldsymbol{\Pi}$ is nonempty and compact, and the objective is continuous under $0 \log 0 = 0$. A maximizer therefore exists, and strict concavity of the entropy term gives uniqueness.

If an optimizer has a zero entry, interpolate it with $\mathbf{1}\mathbf{1}^\top / N$. For a small interpolation weight $\tau > 0$, the entropy gain at a zero entry has order $-\tau \log \tau$, while the remaining changes have order τ. The objective therefore increases for sufficiently small τ, contradicting optimality.

At a fixed coupling, the derivative of the objective with respect to $\boldsymbol{S}$ is $\boldsymbol{\pi}$. Danskin's theorem and uniqueness give the stated gradient (Danskin, 1966). For symmetric $\boldsymbol{S}$, transposition preserves feasibility and the objective. Uniqueness then implies $\boldsymbol{\pi}^\star = (\boldsymbol{\pi}^\star)^\top$. □

Fix θ and $\mathbf{T}$. For head h, the score differential is

$$\mathrm{d}\boldsymbol{S}_h = \frac{(\mathrm{d}\boldsymbol{Z}_h)\boldsymbol{Z}_h^\top + \boldsymbol{Z}_h(\mathrm{d}\boldsymbol{Z}_h)^\top}{\sqrt{d_h}},$$

where $\mathrm{d}\boldsymbol{Z}_h = (\mathrm{d}\boldsymbol{X})\boldsymbol{W}_h$. Applying Lemma A.2 gives

$$\begin{aligned}
\mathrm{d}\left[\frac{\sqrt{d_h}}{2}\Phi_{\varepsilon_{\mathrm{OT}}}(\boldsymbol{S}_h)\right] &= \frac{1}{2}\left\langle \boldsymbol{\pi}_h^\star, (\mathrm{d}\boldsymbol{Z}_h)\boldsymbol{Z}_h^\top + \boldsymbol{Z}_h(\mathrm{d}\boldsymbol{Z}_h)^\top \right\rangle = \frac{1}{2}\left\langle \left(\boldsymbol{\pi}_h^\star + (\boldsymbol{\pi}_h^\star)^\top\right)\boldsymbol{Z}_h, \mathrm{d}\boldsymbol{Z}_h \right\rangle \\
&= \frac{1}{2}\left\langle \left(\boldsymbol{\pi}_h^\star + (\boldsymbol{\pi}_h^\star)^\top\right)\boldsymbol{Z}_h\boldsymbol{W}_h^\top, \mathrm{d}\boldsymbol{X} \right\rangle.
\end{aligned}$$

Summing over heads yields

$$\nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T}) = \frac{1}{2}\sum_{h=1}^{H}\left(\boldsymbol{\pi}_h^\star + (\boldsymbol{\pi}_h^\star)^\top\right)\boldsymbol{Z}_h\boldsymbol{W}_h^\top.$$

The factor $\sqrt{d_h}/2$ cancels the score normalization and averages the contributions from the two feature arguments. If each $\boldsymbol{T}_h$ is symmetric, then $\boldsymbol{S}_h$ and its optimal plan are symmetric, so the force reduces to

$$-\nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T}) = -\sum_{h=1}^{H}\boldsymbol{\pi}_h^\star\boldsymbol{Z}_h\boldsymbol{W}_h^\top.$$

These derivatives use the exact optimizer and hold relations fixed.

For the kinetic term, the primitive of leaky-ReLU is

$$\phi_{0.1}(u) = \int_0^u \sigma_{0.1}(s)\,\mathrm{d}s = \begin{cases} \frac{1}{2}u^2, & u \geq 0, \\ \frac{0.1}{2}u^2, & u < 0, \end{cases}$$

so $\phi_{0.1}'(u) = \sigma_{0.1}(u)$. Under the row convention for $\boldsymbol{P}$,

$$K_\theta(\boldsymbol{P}) = \sum_{i,r}\phi_{0.1}\big((\boldsymbol{P}\boldsymbol{B} + \mathbf{1}\boldsymbol{b}^\top)_{ir}\big).$$

Its differential is

$$\mathrm{d}K_\theta = \left\langle \sigma_{0.1}(\boldsymbol{P}\boldsymbol{B} + \mathbf{1}\boldsymbol{b}^\top), (\mathrm{d}\boldsymbol{P})\boldsymbol{B} \right\rangle = \left\langle \sigma_{0.1}(\boldsymbol{P}\boldsymbol{B} + \mathbf{1}\boldsymbol{b}^\top)\boldsymbol{B}^\top, \mathrm{d}\boldsymbol{P} \right\rangle,$$

and hence

$$\nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}) = \sigma_{0.1}(\boldsymbol{P}\boldsymbol{B} + \mathbf{1}\boldsymbol{b}^\top)\boldsymbol{B}^\top.$$

The transpose appears because both linear maps arise from differentiating the same scalar energy. This gradient converts momentum into feature velocity. Momentum and velocity coincide for $K_\theta(\boldsymbol{P}) = \frac{1}{2}\|\boldsymbol{P}\|_F^2$, but need not coincide for the learned energy.

The primitive $\phi_{0.1}$ is continuously differentiable with a Lipschitz derivative, but is only piecewise twice differentiable. Calculations requiring a classical Hessian therefore apply within regions where no kinetic preactivation crosses zero.

A.4 Lie–Trotter Splitting and the Implemented Layer

For fixed parameters and relations, write $H_{\theta,N} = K_\theta + V_\theta$. Hamilton's equations are

$$\begin{cases} \dot{\boldsymbol{X}} = \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}), \\ \dot{\boldsymbol{P}} = -\nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T}). \end{cases}$$

Along a differentiable solution,

$$\frac{\mathrm{d}}{\mathrm{d}t} H_{\theta,N} = \langle \nabla_{\boldsymbol{X}} V_\theta, \nabla_{\boldsymbol{P}} K_\theta \rangle - \langle \nabla_{\boldsymbol{P}} K_\theta, \nabla_{\boldsymbol{X}} V_\theta \rangle = 0,$$

so the autonomous reference flow conserves its Hamiltonian.

The potential subflow changes momentum at fixed features, while the kinetic subflow advances features at fixed momentum. Their exact maps over a step $\eta > 0$ are

$$\begin{aligned} \varphi_V^\eta(\boldsymbol{X}, \boldsymbol{P}) &= \big(\boldsymbol{X}, \boldsymbol{P} - \eta \nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T})\big), \\ \varphi_K^\eta(\boldsymbol{X}, \boldsymbol{P}) &= \big(\boldsymbol{X} + \eta \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}), \boldsymbol{P}\big). \end{aligned}$$

Applying φ_V^η first and φ_K^η second gives $(\boldsymbol{X}^{k+1}, \boldsymbol{P}^{k+1}) = \varphi_K^\eta \circ \varphi_V^\eta(\boldsymbol{X}^k, \boldsymbol{P}^k)$, which expands to

$$\begin{cases} \boldsymbol{P}^{k+1} = \boldsymbol{P}^k - \eta \nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X}^k \mid \mathbf{T}), \\ \boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \eta \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}^{k+1}). \end{cases}$$

This is the momentum-first symplectic Euler step in (2). For sufficiently smooth energies, Lie–Trotter composition has first-order accuracy, while the symmetric composition $\varphi_V^{\eta/2} \circ \varphi_K^\eta \circ \varphi_V^{\eta/2}$ gives the second-order Störmer–Verlet method (Hairer et al., 2006).

Symplecticity follows from the subflow Jacobians. After vectorizing $\boldsymbol{X}$ and $\boldsymbol{P}$ and suppressing evaluation points, these have the block forms

$$D\varphi_V^\eta = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ -\eta \nabla^2 V_\theta & \boldsymbol{I} \end{pmatrix}, \qquad D\varphi_K^\eta = \begin{pmatrix} \boldsymbol{I} & \eta \nabla^2 K_\theta \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}.$$

Symmetry of the Hessians gives $(D\varphi)^\top \boldsymbol{J} D\varphi = \boldsymbol{J}$ for $\boldsymbol{J} = \left(\begin{smallmatrix} \mathbf{0} & \boldsymbol{I} \\ -\boldsymbol{I} & \mathbf{0} \end{smallmatrix}\right)$, and composition preserves this identity.[1] The resulting numerical maps need not conserve the original Hamiltonian exactly.

Momentum retention can be introduced through the linear decay subflow $\dot{\boldsymbol{P}} = -\gamma \boldsymbol{P}$, with $\gamma \geq 0$. Its exact solution is $\boldsymbol{P} \leftarrow e^{-\gamma\eta} \boldsymbol{P}$. Applying decay before the kick gives

$$\begin{cases} \boldsymbol{P}^{k+1} = e^{-\gamma\eta} \boldsymbol{P}^k - \eta \nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X}^k \mid \mathbf{T}), \\ \boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \eta \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}^{k+1}), \end{cases}$$

with $\gamma = 0$ recovering the conservative splitting. Applying decay after the kick would also scale the newly added force.

For the corresponding continuous damped system, the energy balance is

$$\frac{\mathrm{d}}{\mathrm{d}t} H_{\theta,N} = -\gamma \left\langle \boldsymbol{P}, \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P}) \right\rangle.$$

[1] This calculation assumes twice differentiable energies and applies to the learned kinetic map within its smooth activation regions.

For $K_\theta(\boldsymbol{P}) = \frac{1}{2}\|\boldsymbol{P}\|_F^2$, this becomes $-\gamma\|\boldsymbol{P}\|_F^2$, and the feature dynamics satisfy

$$\ddot{\boldsymbol{X}} + \gamma\dot{\boldsymbol{X}} + \nabla_{\boldsymbol{X}} V_\theta(\boldsymbol{X} \mid \mathbf{T}) = 0.$$

For a learned kinetic energy with a bias, the sign of $\langle \boldsymbol{P}, \nabla_{\boldsymbol{P}} K_\theta(\boldsymbol{P})\rangle$ is not automatic. Linear momentum decay therefore requires an additional condition to imply energy dissipation.

These calculations concern the reference system with fixed parameters and relations. The learned update in (8) follows the same momentum-first order, with the implementation differences described in the main text. Algorithm 1 gives the complete layer computation.

A.5 SE(3)-Equivariance of the Complete Pipeline

The Hamiltonian states $\boldsymbol{X}$ and $\boldsymbol{P}$ are invariant latent features, while molecular coordinates are updated separately. We hold atom attributes, pair types, and validity masks fixed under rigid motions. During training, we compare transformations using the same feature dropout masks.

Lemma A.3 (SE(3)-equivariance). *The complete coordinate predictor f, including centering and centroid restoration, satisfies*

$$f(\boldsymbol{C}\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{t}^\top) = f(\boldsymbol{C})\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{t}^\top$$

on valid atoms for every $\boldsymbol{R} \in \mathrm{SO}(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$.

Proof. Under $\boldsymbol{c}_i' = \boldsymbol{R}\boldsymbol{c}_i + \boldsymbol{t}$, pairwise distances satisfy

$$\|\boldsymbol{c}_i' - \boldsymbol{c}_j'\| = \|\boldsymbol{R}(\boldsymbol{c}_i - \boldsymbol{c}_j)\| = \|\boldsymbol{c}_i - \boldsymbol{c}_j\|.$$

Relation initialization uses only pair types and these distances, so the initialized relations are unchanged. Atom embeddings and zero initial momentum are unchanged as well. Every HamFormer operation acts on invariant quantities, so induction over its layers gives identical features, momentum, and relations for the original and transformed inputs. This argument holds for any fixed number of Sinkhorn iterations.

At EGNN layer q, the current and stage-entry distances remain invariant. The atom features and relation vectors are also invariant, so the pair messages, updated features, and scalar coordinate weights $w_{ij}^{(q)}$ are unchanged. Writing $\boldsymbol{c}_i'^{(q)} = \boldsymbol{R}\boldsymbol{c}_i^{(q)} + \boldsymbol{t}$, the coordinate update gives

$$\begin{aligned}\boldsymbol{c}_i'^{(q+1)} &= \boldsymbol{R}\boldsymbol{c}_i^{(q)} + \boldsymbol{t} + \sum_j w_{ij}^{(q)} \frac{\boldsymbol{R}(\boldsymbol{c}_i^{(q)} - \boldsymbol{c}_j^{(q)})}{\|\boldsymbol{c}_i^{(q)} - \boldsymbol{c}_j^{(q)}\| + \varepsilon_{\mathrm{dist}}} \\ &= \boldsymbol{R}\boldsymbol{c}_i^{(q+1)} + \boldsymbol{t}.\end{aligned}$$

Thus the coordinate update is equivariant, as in the EGNN construction (Satorras et al., 2021). Its feature output remains invariant and can enter the next HamFormer stage. Passing momentum through EGNN unchanged preserves its invariance.

For mask $\boldsymbol{m}$ and $N = \mathbf{1}^\top\boldsymbol{m}$, the input centroid is $\boldsymbol{c}_{\mathrm{ctr}} = (\boldsymbol{C}^{\mathrm{in}})^\top\boldsymbol{m}/N$. It transforms as $\boldsymbol{c}_{\mathrm{ctr}}' = \boldsymbol{R}\boldsymbol{c}_{\mathrm{ctr}} + \boldsymbol{t}$ and $\boldsymbol{C}_{\mathrm{c}}' = \boldsymbol{C}_{\mathrm{c}}\boldsymbol{R}^\top$. The centered network output therefore rotates by $\boldsymbol{R}$. Restoring the input centroid gives $\widehat{\boldsymbol{C}}' = \widehat{\boldsymbol{C}}\boldsymbol{R}^\top + \mathbf{1}\boldsymbol{t}^\top$, which proves the claim. □

Predicted pairwise distances are consequently invariant. This argument depends on the transformation properties of the network operations and does not require an exact Hamiltonian integrator or exact transport marginals.

A.6 HamFormer for Molecular Conformation Optimization

Given atom types $\boldsymbol{a}$, pair types $\boldsymbol{E} = (e_{ij})$, and input coordinates $\boldsymbol{C}^{\mathrm{in}}$, the atom embedding serves as the encoder and the final EGNN as the SE(3)-equivariant coordinate decoder. The network operates on centered coordinates and restores the input centroid to obtain $\widehat{\boldsymbol{C}}$, with pairwise distances $\widehat{\boldsymbol{D}}$. Algorithms 1 and 2 specify the layer update and pipeline.

Both algorithms update states in place. For a valid-atom mask $\boldsymbol{m} \in \{0,1\}^{N_{\mathrm{pad}}}$, let $N = \mathbf{1}^\top \boldsymbol{m}$. Masked Sinkhorn targets marginals $\boldsymbol{m}/N$, and geometric statistics use only valid atoms. Let L_E denote the number of HamFormer–EGNN stage pairs. For $l = 1, \ldots, L_E$, the stage maps $\mathrm{HamFormer}_l$ and EGNN_l apply their layers in order. Each EGNN stage records its incoming coordinates to compute the fixed stage-entry distances. The map $\mathrm{InitRelation}$ implements (7).

Algorithm 1 HamFormer layer

▷ **Require:** $\boldsymbol{X}, \boldsymbol{P}, \mathbf{T}, \boldsymbol{m}$, parameters θ_ℓ, iterations I
$N \leftarrow \mathbf{1}^\top \boldsymbol{m}$
for $h = 1, \ldots, H$ **do**
  $\boldsymbol{Z}_h \leftarrow \mathrm{LN}_{\mathrm{in},\ell}(\boldsymbol{X}) \boldsymbol{W}_h^{(\ell)} + \mathbf{1}(\boldsymbol{b}_h^{(\ell)})^\top$
  $\boldsymbol{S} \leftarrow \boldsymbol{Z}_h \boldsymbol{Z}_h^\top / \sqrt{d_h} - \log(\boldsymbol{T}_h + \varepsilon_{\mathrm{num}})$
  $\boldsymbol{T}_h \leftarrow N\, \mathrm{Sinkhorn}_{\varepsilon_{\mathrm{OT}}, I}(\boldsymbol{S}, \boldsymbol{m})$
end for
$\boldsymbol{P} \leftarrow -\tanh(\alpha_\ell)\boldsymbol{P} + \mathrm{LN}_{P,\ell}\big(-\sum_{h=1}^{H} \boldsymbol{T}_h \boldsymbol{Z}_h (\boldsymbol{W}_h^{(\ell)})^\top\big)$
$\boldsymbol{X} \leftarrow \boldsymbol{X} + \mathrm{LN}_{X,\ell}\big[\sigma_{0.1}(\boldsymbol{P}\boldsymbol{B}_\ell + \mathbf{1}\boldsymbol{b}_\ell^\top)\boldsymbol{B}_\ell^\top\big]$
$\boldsymbol{X} \leftarrow \boldsymbol{X} + \mathrm{FFN}_\ell(\boldsymbol{X})$
▷ **return** $\boldsymbol{X}, \boldsymbol{P}, \mathbf{T}$

Algorithm 2 Conformation refinement

▷ **Require:** $\boldsymbol{a}, \boldsymbol{E}, \boldsymbol{C}^{\mathrm{in}}, \boldsymbol{m}$
$N \leftarrow \mathbf{1}^\top \boldsymbol{m}$, $\boldsymbol{X} \leftarrow \mathrm{Embed}(\boldsymbol{a})$, $\boldsymbol{P} \leftarrow \mathbf{0}$
$\mathbf{T} \leftarrow \mathrm{InitRelation}(\boldsymbol{E}, \boldsymbol{C}^{\mathrm{in}}, \boldsymbol{m})$
$\boldsymbol{c}_{\mathrm{ctr}} \leftarrow (\boldsymbol{C}^{\mathrm{in}})^\top \boldsymbol{m}/N$
$\boldsymbol{C} \leftarrow \boldsymbol{C}^{\mathrm{in}} - \mathbf{1}\boldsymbol{c}_{\mathrm{ctr}}^\top$
for $l = 1, \ldots, L_E$ **do**
  $(\boldsymbol{X}, \boldsymbol{P}, \mathbf{T}) \leftarrow \mathrm{HamFormer}_l(\boldsymbol{X}, \boldsymbol{P}, \mathbf{T}, \boldsymbol{m})$
  $(\boldsymbol{X}, \boldsymbol{C}) \leftarrow \mathrm{EGNN}_l(\boldsymbol{X}, \boldsymbol{C}, \mathbf{T}, \boldsymbol{m})$
end for
$\widehat{\boldsymbol{C}} \leftarrow \boldsymbol{C} + \mathbf{1}\boldsymbol{c}_{\mathrm{ctr}}^\top$ and $\widehat{\boldsymbol{D}} \leftarrow [\|\widehat{\boldsymbol{c}}_i - \widehat{\boldsymbol{c}}_j\|]_{ij}$
▷ **return** $\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{D}}$

For ground-state conformation prediction (Wang et al., 2025), we minimize distance and aligned coordinate errors relative to target distances $\boldsymbol{D}^{\mathrm{gt}}$ and coordinates $\boldsymbol{C}^{\mathrm{gt}}$:

$$\mathcal{L}_{\mathrm{train}}(\theta) = \frac{1}{N^2} \sum_{(i,j)\in\mathcal{I}_{\mathrm{pair}}} \big|\widehat{D}_{ij} - D_{ij}^{\mathrm{gt}}\big| + \frac{\lambda}{N} \min_{\boldsymbol{R}\in \mathrm{SO}(3)} \big\|\widehat{\boldsymbol{C}}_{\mathrm{c}} - \boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}} \boldsymbol{R}^\top\big\|_F^2, \tag{9}$$

where the subscript c denotes centering over the N valid atoms, $\mathcal{I}_{\mathrm{pair}}$ contains all ordered valid atom pairs, and $\lambda \geq 0$ weights the coordinate loss. The first term measures absolute distance errors, while the second measures squared coordinate errors after optimal rotation. During training, this rotation is computed from detached predictions and held fixed during backpropagation. Section A.7 details rotational alignment and masked Sinkhorn scaling. All network parameters are trained jointly, with gradients propagated through the relation updates, normalization, and Sinkhorn iterations.

A.7 Transport Computation, Training, and Complexity

For the positive optimizer of (4), introduce row and column multipliers ζ_i and β_j. Stationarity gives

$$S_{ij} - \varepsilon_{\mathrm{OT}} \log \pi_{ij} - \zeta_i - \beta_j = 0,$$

so the optimizer has the form $\pi_{ij} = u_i \exp(S_{ij}/\varepsilon_{\mathrm{OT}}) v_j$, where $u_i = e^{-\zeta_i/\varepsilon_{\mathrm{OT}}}$ and $v_j = e^{-\beta_j/\varepsilon_{\mathrm{OT}}}$. Sinkhorn scaling alternately adjusts $\boldsymbol{u}$ and $\boldsymbol{v}$ to satisfy the row and column constraints (Cuturi, 2013; Gabriel & Marco, 2019).

The implementation targets probability marginals $\boldsymbol{m}/N$. Write $\boldsymbol{\pi}^{\mathrm{prob}}$ for the resulting matrix, with invalid rows and columns excluded from normalization. On the valid block, the scaling updates are

$$u_i \leftarrow \frac{1/N}{\sum_j \exp(S_{ij}/\varepsilon_{\mathrm{OT}}) v_j}, \qquad v_j \leftarrow \frac{1/N}{\sum_i u_i \exp(S_{ij}/\varepsilon_{\mathrm{OT}})}.$$

The relation update returns $N\boldsymbol{\pi}^{\mathrm{prob}}$, targeting unit row and column sums on valid atoms. With finitely many iterations, both sets of constraints need not hold exactly. Alternating row and column normalization also need not preserve symmetry at intermediate iterations.

For the relation-conditioned score in (6),

$$\exp(\boldsymbol{S}_h/\varepsilon_{\mathrm{OT}}) = \exp\left(\frac{\boldsymbol{Z}_h \boldsymbol{Z}_h^\top}{\varepsilon_{\mathrm{OT}}\sqrt{d_h}}\right) \odot (\boldsymbol{T}_h + \varepsilon_{\mathrm{num}})^{-1/\varepsilon_{\mathrm{OT}}},$$

where $\odot$ and the power act entrywise. At $\varepsilon_{\mathrm{OT}} = 1$, the exponentiated feature affinity is divided entrywise by $\boldsymbol{T}_h + \varepsilon_{\mathrm{num}}$ before normalization.

For the coordinate term in (9), both coordinate matrices are restricted to valid atoms and centered before alignment. Expanding the coordinate error gives

$$\big\|\widehat{\boldsymbol{C}}_{\mathrm{c}} - \boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}} \boldsymbol{R}^\top\big\|_F^2 = \|\widehat{\boldsymbol{C}}_{\mathrm{c}}\|_F^2 + \|\boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}}\|_F^2 - 2\,\mathrm{tr}\,\big(\boldsymbol{R}^\top \widehat{\boldsymbol{C}}_{\mathrm{c}}^\top \boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}}\big).$$

The first two terms are independent of $\boldsymbol{R}$. Let $\widehat{\boldsymbol{C}}_{\mathrm{c}}^{\top}\boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}} = \boldsymbol{U}_{\mathrm{svd}}\boldsymbol{\Sigma}\boldsymbol{V}_{\mathrm{svd}}^{\top}$ be a singular value decomposition with singular values in decreasing order. A minimizing proper rotation is

$$\boldsymbol{R}^{\star} = \boldsymbol{U}_{\mathrm{svd}}\,\mathrm{diag}\left(1, 1, \det(\boldsymbol{U}_{\mathrm{svd}}\boldsymbol{V}_{\mathrm{svd}}^{\top})\right)\boldsymbol{V}_{\mathrm{svd}}^{\top},$$

where the determinant correction enforces $\boldsymbol{R}^{\star} \in \mathrm{SO}(3)$. At points where the minimized coordinate loss is differentiable, for example when the minimizing rotation is unique, the envelope identity gives

$$\nabla_{\widehat{\boldsymbol{C}}_{\mathrm{c}}}\Big[\frac{\lambda}{N}\min_{\boldsymbol{R}\in\mathrm{SO}(3)}\big\|\widehat{\boldsymbol{C}}_{\mathrm{c}} - \boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}}\boldsymbol{R}^{\top}\big\|_F^2\Big] = \frac{2\lambda}{N}\Big(\widehat{\boldsymbol{C}}_{\mathrm{c}} - \boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}}(\boldsymbol{R}^{\star})^{\top}\Big).$$

Computing the exact optimal rotation from detached predictions and treating it as fixed therefore yields the gradient of the minimized value at such points. Degenerate configurations can have multiple minimizing rotations, so differentiability is not asserted everywhere.

Remark A.4 (**Computational complexity**). Consider one dense HamFormer layer with N valid atoms, feature width d, head widths d_h, FFN width d_{FF}, and I Sinkhorn iterations. Head projections cost $O(Nd\sum_h d_h)$. Forming scores and aggregating transport messages cost $O(N^2\sum_h d_h)$, and Sinkhorn scaling costs $O(IHN^2)$. The kinetic map and FFN cost $O(Nd^2)$ and $O(Nd\,d_{\mathrm{FF}})$, respectively. The total time complexity is therefore $O\big(Nd\sum_h d_h + N^2\sum_h d_h + IHN^2 + Nd^2 + Nd\,d_{\mathrm{FF}}\big)$, which reduces to $O\big(Nd^2 + N^2d + IHN^2 + Nd\,d_{\mathrm{FF}}\big)$ when $\sum_h d_h = d$.

The relation tensor requires $O(HN^2)$ storage. Including features, projected heads, and FFN activations gives forward activation storage $O\big(HN^2 + Nd + N\sum_h d_h + Nd_{\mathrm{FF}}\big)$. Training requires additional saved intermediates depending on the differentiation and recomputation strategy. For dense padded tensors, replace N by N_{pad} unless padded entries are excluded from computation. EGNN cost scales with the number of evaluated atom pairs at fixed MLP widths, and the complete network cost is obtained by summing over its layers.

B Experimental Details

B.1 Datasets

We use the QM9 splits and the Molecule3D random split datasets, following the benchmark protocol used by WGFormer. The processed QM9 validation and test sets contain 9,900 and 10,697 molecules, respectively. The corresponding Molecule3D sets contain 779,918 and 779,915 molecules.

B.2 Evaluation

Following WGFormer, we use the following metrics to evaluate the quality of the predicted molecular structures.

- **D-MAE** and **D-RMSE**: The mean absolute error and root mean square error between the predicted pairwise distances $\widehat{D}_{ij}$ and ground-truth distances D_{ij}^{gt}, both computed from the corresponding Cartesian coordinates:

$$\text{D-MAE} = \frac{1}{N^2}\sum\nolimits_{(i,j)\in\mathcal{I}_{\mathrm{pair}}}\big|\widehat{D}_{ij} - D_{ij}^{\mathrm{gt}}\big|,$$
$$\text{D-RMSE} = \sqrt{\frac{1}{N^2}\sum\nolimits_{(i,j)\in\mathcal{I}_{\mathrm{pair}}}\big(\widehat{D}_{ij} - D_{ij}^{\mathrm{gt}}\big)^2}.$$

- **C-RMSD**: The coordinate root mean square deviation between the predicted coordinates $\widehat{\boldsymbol{C}}$ and ground-truth coordinates $\boldsymbol{C}^{\mathrm{gt}}$ after optimal rigid alignment:

$$\text{C-RMSD} = \sqrt{\frac{1}{N}\min_{\boldsymbol{R}\in\mathrm{SO}(3),\,\boldsymbol{t}\in\mathbb{R}^3}\Big\|\widehat{\boldsymbol{C}} - (\boldsymbol{C}^{\mathrm{gt}}\boldsymbol{R}^{\top} + \mathbf{1}_N\boldsymbol{t}^{\top})\Big\|_F^2}.$$

  For this metric, hydrogen atoms are removed, N counts the retained atoms, and $\boldsymbol{C}_{\mathrm{c}}^{\mathrm{gt}}$ is understood to use the symmetry-equivalent atom ordering that minimizes the aligned RMSD.

B.3 HYPERPARAMETERS

Table 4 summarizes the full-model training settings used on the QM9 and Molecule3D random datasets, including the peak learning rate, scheduled epochs, warmup fraction, evaluation checkpoint epoch, number of training GPUs, batch size per GPU, global batch size, gradient accumulation steps, and random seed.

Table 4: Full-model training settings.

Setting	QM9	Molecule3D random
Peak learning rate	8×10^{-4}	4×10^{-4}
Scheduled epochs	500	40
Warmup fraction	0.06	0.05
Evaluated checkpoint epoch	294	40
Training GPUs	8	2
Batch size per GPU	32	64
Global batch size	256	128
Gradient accumulation	1	1
Random seed	1	1

B.4 ABLATION STUDY ON MODEL COMPONENTS

To further validate the effectiveness of our proposed design, we perform module ablation studies on the validation and test sets of both QM9 and Molecule3D random split. As shown in Table 5, removing the HamFormer feature block and the $P\&T$ branch leads to a prominent performance degradation across all metrics. This significant drop clearly demonstrates that the introduced components are essential for maintaining the model's high representation capacity and prediction accuracy.

Table 5: Module ablation on the test sets. "Feature block" denotes the module applied before each EGNN stage.

	Feature	$P\&T$	QM9			Molecule3D random		
Variant	block	branch	D-MAE	D-RMSE	C-RMSD	D-MAE	D-RMSE	C-RMSD
Full model	HamFormer	✓	**0.218**	**0.419**	**0.198**	**0.387**	**0.656**	**0.654**
EGNN-only	×	×	0.233	0.427	0.239	0.399	0.666	0.697

B.5 INFERENCE-TIME MOMENTUM-RETENTION ABLATION

To further validate the importance of the inference-time momentum-retention mechanism, we investigate the effect of the momentum factor on the Molecule3D (random split) dataset. As shown in Table 6, setting the retention factor to zero ($\rho_l = 0$) results in a substantial performance drop across both validation and test sets compared to using the learned momentum parameter (Learned ρ_l). This prominent degradation strongly confirms the crucial role and importance of the momentum-retention mechanism during inference.

Table 6: Inference-time momentum-retention ablation on Molecule3D random.

	Validation			Test		
Inference setting	D-MAE	D-RMSE	C-RMSD	D-MAE	D-RMSE	C-RMSD
Learned ρ_l	**0.386**	**0.653**	**0.652**	**0.387**	**0.656**	**0.654**
$\rho_l = 0$	0.418	0.684	0.712	0.420	0.716	0.714

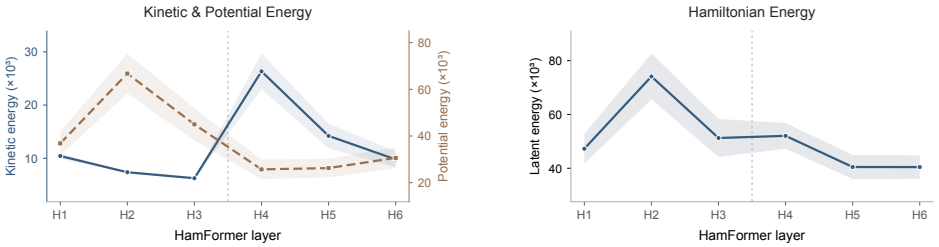

Figure 5: Layer-wise evolution of the latent kinetic energy, potential energy, and total Hamiltonian evaluated on 1,000 randomly selected molecules from the QM9 test set.

B.6 Layer-wise Energy Evolution on the QM9 Dataset

Figure 5 presents the layer-wise evolution of the latent kinetic energy, potential energy, and total Hamiltonian on the QM9 test set. The energy profiles exhibit clear changes across HamFormer layers, particularly around the transition from Stage 1 to Stage 2. This observation is consistent with the corresponding results on Molecule3D and further suggests that the Hamiltonian dynamics are progressively transformed throughout the network.

B.7 Additional Visualizations

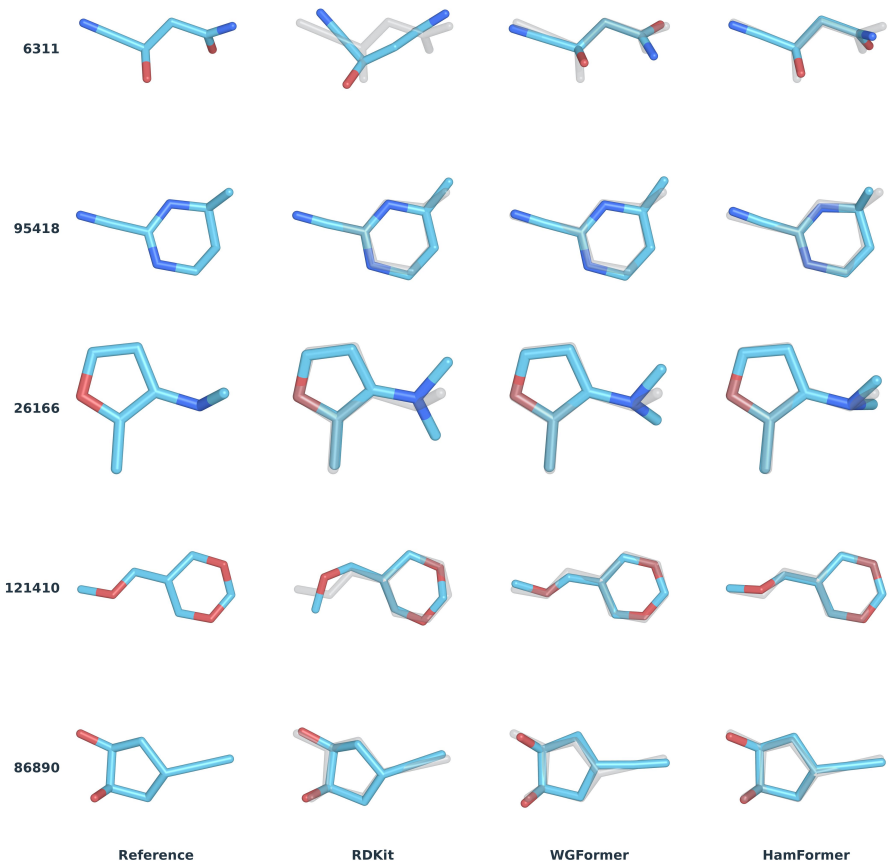

Figure 6: The visualizations for the ground-state conformations predicted by RDKit, WGFormer, and HamFormer.