Supplementary Information for: Learning Canonical Representations for Unified 3D Molecular Modeling

Seungyeon Choi[1], **Hwanhee Kim**[1], **Yoonju Kim**[1], **Seungyong Lee**[1], **Inpyo Hong**[1], **and Sanghyun Park**[1,*]

[1]Yonsei University, Department of Computer Science, Seoul, South Korea
[*]e-mail: sanghyun@yonsei.ac.kr

ABSTRACT

This document contains the Supplementary Notes supporting "Learning Canonical Representations for Unified 3D Molecular Modeling".

Supplementary Note Contents

A Canonicalization for Invariant Prediction: Quotient Factorization and Stability 3
A.1 Setup
A.2 Exact factorization through a canonical representative
A.3 Approximate invariance and rotation-augmented risk stability
B Canonicalization for Equivariant Prediction 4
B.1 ρ-equivariant targets
B.2 Exact Canonicalize-and-Lift factorization
B.3 Approximate equivariance bounds
C Generative Modeling in Canonical Coordinates 5
C.1 Score matching and equivariant projection
C.2 Quotient-space evaluation beyond the ideal section
C.3 Equivariant scores without equivariant layers
D Task-Driven Inductive Bias: Canonicalization vs. Equivariance vs. Augmentation 7
D.1 Limitations of invariant architectures
D.2 Limitations of equivariant architectures
D.3 Limitations of rotation augmentation
D.4 Canonicalization as a unified solution
E How Should We Evaluate Molecular Representations? 9
E.1 Baseline Models for Representation Evaluation
E.2 Molecular Identity Preservation Analysis
E.3 Chemical Semantic Consistency
E.4 Embedding Space Geometry Analysis
E.5 Pretrain-Finetune Alignment Analysis
E.6 Analysis of Representation Quality Experiment
F Experimental Details for Canonical Latent-Space Diagnostics 15
F.1 Rotation Sensitivity
F.2 Latent-Distance Hierarchy
F.3 Latent Perturbation Decodability
F.4 Gauge-Waste Ratio and Search Efficiency
F.5 Interpolation Decodability
G Experimental Details for Figure 2 17
H Implementation Details 17
I Additional Experiments (Ablation Study) 21
I.1 In-Depth Analysis of Representation Paradigms

J Datasets 23
J.1 Pretraining Dataset
J.2 MoleculeNet
J.3 TDC
J.4 CrossDocked2020
J.5 GEOM-DRUG
J.6 Additional Benchmark
J.7 Evaluation Metrics
J.8 Baseline Models
J.9 MoleculeNet Analysis
K Flow Matching for Molecular Generation 28
K.1 Background: Conditional Flow Matching
K.2 Flow Matching in Proposed Framework
L Optimization via Latent Space Search 32
L.1 Problem Formulation
L.2 LRA-CMA-ES Overview
L.2 Application to Flow-Based Molecular Optimization

A Canonicalization for Invariant Prediction: Quotient Factorization and Stability

A.1 Setup

Let $\mathbf{x} \in \mathbb{R}^{N\times 3}$ be 3D coordinates and $\mathbf{h}$ be per-node features (e.g., atom types). We remove translations by centering

$$\mathbf{x}^{\mathrm{c}} := \mathbf{x} - \mathbf{1}\bar{\mathbf{x}}^{\top}, \qquad \bar{\mathbf{x}} := \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \in \mathbb{R}^3.$$

For $R \in \mathrm{SO}(3)$, the action on coordinates is $x \mapsto Rx$ (row-wise). We use a learned *rotation selector* $\widehat{R}_\theta(\mathbf{x},\mathbf{h}) \in \mathrm{SO}(3)$ and define the *canonicalization map*

$$A_\theta(\mathbf{x},\mathbf{h}) \;:=\; \widehat{R}_\theta(\mathbf{x},\mathbf{h})\,\mathbf{x}^{\mathrm{c}} \in \mathbb{R}^{N\times 3}. \tag{1}$$

A generic non-equivariant predictor in canonical space is

$$\widehat{y}(\mathbf{x},\mathbf{h}) \;:=\; \widetilde{f}_\eta\big(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}\big). \tag{2}$$

Permutation symmetry is not assumed: we construct the dataset using RDKit's canonical atom ordering, which fixes a deterministic node indexing.

A target $y(\mathbf{x},\mathbf{h}) \in \mathbb{R}$ is *rotation-invariant* if $y(R\mathbf{x},\mathbf{h}) = y(\mathbf{x},\mathbf{h})$ for all $R \in \mathrm{SO}(3)$.

A.2 Exact factorization through a canonical representative

Assumption 1 (Well-defined canonicalization on the data support)**.** There exists a subset $\mathscr{S} \subseteq \mathbb{R}^{N\times 3} \times \mathscr{H}$ containing the data support such that:

$$\text{(gauge consistency)} \quad \widehat{\mathbf{R}}_\theta(Rx,h) = \widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h})R^{\top},\ \ \forall R \in \mathrm{SO}(3),\ (\mathbf{x},\mathbf{h}) \in \mathscr{S},$$

$$\text{(orbit-separation)} \quad A_\theta(\mathbf{x},\mathbf{h}) = A_\theta(\mathbf{x}',\mathbf{h}') \;\Rightarrow\; \mathbf{h} = \mathbf{h}' \ \text{ and } \ \exists R \in \mathrm{SO}(3) \text{ s.t. } \mathbf{x}' = R\mathbf{x},\ \ \forall (\mathbf{x},\mathbf{h}),(\mathbf{x}',\mathbf{h}') \in \mathscr{S}.$$

Equivalently, $A_\theta(R\mathbf{x},\mathbf{h}) = A_\theta(\mathbf{x},\mathbf{h})$ and A_θ is injective on rotation orbits over $\mathscr{S}$.

Global orbit-separating sections may fail at symmetric/degenerate configurations (non-trivial stabilizers), leading to ambiguity or discontinuity. In practice this is mitigated by focusing on generic inputs (symmetries are measure-zero under perturbations) and by using deterministic node indexing (RDKit canonical ordering) so that $\widehat{R}_\theta(\mathbf{x},\mathbf{h})$ can implement consistent tie-breaking on the data support.

Proposition 2 (Invariant targets factor through canonicalization)**.** *Under Assumption 1, for any rotation-invariant target y there exists a function $\widetilde{y}$ such that*

$$y(\mathbf{x},\mathbf{h}) = \widetilde{y}\big(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}\big) \qquad \textit{for all } (\mathbf{x},\mathbf{h}) \in \mathscr{S}. \tag{3}$$

Proof. Fix $(\mathbf{x},\mathbf{h}) \in \mathscr{S}$ and set $u = A_\theta(\mathbf{x},\mathbf{h})$. Gauge consistency implies u is constant on the orbit $\{(R\mathbf{x},\mathbf{h}) : R \in \mathrm{SO}(3)\}$, and invariance implies y is constant on the same orbit. Orbit-separation ensures $(\mathbf{u},\mathbf{h})$ identifies the orbit within $\mathscr{S}$, so defining $\widetilde{y}(\mathbf{u},\mathbf{h}) := y(\mathbf{x},\mathbf{h})$ is well-defined. □

Proposition 3 (Universal approximation in canonical space)**.** *Assume Assumption 1. If y is continuous and rotation-invariant on $\mathscr{S}$ and $\widetilde{f}_\eta(\cdot,\mathbf{h})$ is a universal approximator for continuous functions on the domain of $\big(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}\big)$ (restricted to $(\mathbf{x},\mathbf{h}) \in \mathscr{S}$), then for any $\varepsilon > 0$ there exist parameters η such that*

$$\sup_{(\mathbf{x},\mathbf{h})\in\mathscr{S}} \big|\widetilde{f}_\eta(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}) - y(\mathbf{x},\mathbf{h})\big| \;\le\; \varepsilon. \tag{4}$$

By Proposition 2, $y(\mathbf{x},\mathbf{h}) = \widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})$ for some continuous $\widetilde{y}$ on the induced domain. Universality yields η approximating $\widetilde{y}$ uniformly to error ε.

A.3 Approximate invariance and rotation-augmented risk stability

Assumption 4 (Alignment consistency and Lipschitz stability)**.** Assume there exist $\delta \geq 0$, $L_f \geq 0$, and $L_\ell \geq 0$ such that for all $(x,h) \in \mathscr{S}$:

$$\sup_{R\in \mathrm{SO}(3)} \left\| A_\theta(R\mathbf{x},\mathbf{h}) - A_\theta(\mathbf{x},\mathbf{h}) \right\|_F \;\leq\; \delta,$$

$\widetilde{f}_\eta(\cdot,\mathbf{h})$ is L_f-Lipschitz in its first argument, and $\ell(\hat{y},y)$ is L_ℓ-Lipschitz in $\hat{y}$:

$$\left|\widetilde{f}_\eta(u_1,\mathbf{h}) - \widetilde{f}_\eta(u_2,\mathbf{h})\right| \leq L_f \|u_1-u_2\|_F, \qquad |\ell(\hat{y}_1,y)-\ell(\hat{y}_2,y)| \leq L_\ell |\hat{y}_1-\hat{y}_2|.$$

Proposition 5 (Approximate invariance of canonical predictors)**.** *Under Assumption 4, the predictor* (2) *satisfies*

$$\sup_{R\in \mathrm{SO}(3)} |\widehat{y}(R\mathbf{x},\mathbf{h}) - \widehat{y}(\mathbf{x},\mathbf{h})| \;\leq\; L_f\,\delta \qquad \textit{for all } (\mathbf{x},\mathbf{h}) \in \mathscr{S}.$$

If moreover y is rotation-invariant, then for each fixed $(\mathbf{x},\mathbf{h}) \in \mathscr{S}$,

$$\left| \mathbb{E}_{R\sim\mu}\left[\ell(\widehat{y}(R\mathbf{x},\mathbf{h}),y(\mathbf{x},\mathbf{h}))\right] - \ell(\widehat{y}(\mathbf{x},\mathbf{h}),y(\mathbf{x},\mathbf{h}))\right| \;\leq\; L_\ell L_f\,\delta.$$

Proof. Lipschitzness gives $|\widehat{y}(R\mathbf{x},\mathbf{h}) - \widehat{y}(\mathbf{x},\mathbf{h})| \leq L_f \|A_\theta(R\mathbf{x},\mathbf{h}) - A_\theta(\mathbf{x},\mathbf{h})\|_F \leq L_f\delta$. For invariant y, apply L_ℓ-Lipschitzness to the loss difference and average over R. □

B Canonicalization for Equivariant Prediction

B.1 ρ-equivariant targets

Let $\mathscr{Y}$ be an output vector space (e.g., $\mathbb{R}^3$ or $\mathbb{R}^{N\times 3}$) and $\rho : \mathrm{SO}(3) \to GL(\mathscr{Y})$ be a linear representation. A target $y(\mathbf{x},\mathbf{h}) \in \mathscr{Y}$ is *ρ-equivariant* if

$$y(R\mathbf{x},\mathbf{h}) = \rho(R)\,y(\mathbf{x},\mathbf{h}) \qquad \forall R \in \mathrm{SO}(3). \tag{5}$$

B.2 Exact Canonicalize-and-Lift factorization

Proposition 6 (Canonical factorization of ρ-equivariant maps)**.** *Assume Assumption 1. A function $y : \mathbb{R}^{N\times 3} \times \mathscr{H} \to \mathscr{Y}$ is ρ-equivariant on $\mathscr{S}$ if and only if there exists $\widetilde{y}$ such that*

$$y(\mathbf{x},\mathbf{h}) = \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))^{-1}\,\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}), \qquad \forall (\mathbf{x},\mathbf{h}) \in \mathscr{S}. \tag{6}$$

Proof. Assume y is ρ-equivariant on $\mathscr{S}$. For any $(\mathbf{x},\mathbf{h}) \in \mathscr{S}$, set $\mathbf{u} := A_\theta(\mathbf{x},\mathbf{h})$ and define

$$\widetilde{y}(\mathbf{u},\mathbf{h}) \;:=\; \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))\,y(\mathbf{x},\mathbf{h}). \tag{7}$$

To see that $\widetilde{y}$ is well-defined, let $(\mathbf{x}_1,\mathbf{h}),(\mathbf{x}_2,\mathbf{h}) \in \mathscr{S}$ satisfy $A_\theta(\mathbf{x}_1,\mathbf{h}) = A_\theta(\mathbf{x}_2,\mathbf{h}) = \mathbf{u}$. By orbit-separation there exists $R \in SO(3)$ such that $\mathbf{x}_2 = R\mathbf{x}_1$. Using ρ-equivariance, gauge consistency, and that ρ is a representation, we have

$$\begin{aligned}
y(\mathbf{x}_2,\mathbf{h}) &= y(R\mathbf{x}_1,\mathbf{h}) = \rho(R)\,y(\mathbf{x}_1,\mathbf{h}),\\
\widehat{\mathbf{R}}_\theta(\mathbf{x}_2,\mathbf{h}) &= \widehat{\mathbf{R}}_\theta(R\mathbf{x}_1,\mathbf{h}) = \widehat{\mathbf{R}}_\theta(x_1,h)R^\top,\\
\rho(\widehat{\mathbf{R}}_\theta(\mathbf{x}_2,\mathbf{h}))\,y(\mathbf{x}_2,\mathbf{h}) &= \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x}_1,\mathbf{h})R^\top)\,\rho(R)\,y(\mathbf{x}_1,\mathbf{h})\\
&= \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x}_1,\mathbf{h}))\,\rho(R^\top)\rho(R)\,y(\mathbf{x}_1,\mathbf{h}) = \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x}_1,\mathbf{h}))\,y(\mathbf{x}_1,\mathbf{h}),
\end{aligned} \tag{8}$$

where $\rho(R^\top) = \rho(R)^{-1}$ for $R \in SO(3)$. Hence $\widetilde{y}(\mathbf{u},\mathbf{h})$ does not depend on the choice of $(\mathbf{x},\mathbf{h})$ with $\mathbf{u} = A_\theta(\mathbf{x},\mathbf{h})$, and rearranging the definition yields $y(\mathbf{x},\mathbf{h}) = \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))^{-1}\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})$ for all $(\mathbf{x},\mathbf{h}) \in \mathscr{S}$.

Conversely, suppose there exists $\widetilde{y}$ such that $y(\mathbf{x},\mathbf{h}) = \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))^{-1}\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})$ on $\mathscr{S}$. Then for any $R \in SO(3)$, gauge consistency and the homomorphism property of ρ give

$$\begin{aligned}
y(Rx,h) &= \rho(\widehat{\mathbf{R}}_\theta(R\mathbf{x},\mathbf{h}))^{-1}\,\widetilde{y}(A_\theta(R\mathbf{x},\mathbf{h}),\mathbf{h})\\
&= \rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h})R^\top)^{-1}\,\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})\\
&= (\rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))\rho(R^\top))^{-1}\,\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})\\
&= \rho(R)\,\rho(\widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h}))^{-1}\,\widetilde{y}(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}) = \rho(R)\,y(\mathbf{x},\mathbf{h}),
\end{aligned}$$

which shows that y is ρ-equivariant on $\mathscr{S}$. □

B.3 Approximate equivariance bounds

Define the Canonicalize-and-Lift predictor

$$\widehat{y}(\mathbf{x},\mathbf{h}) := \rho(\widehat{R}_\theta(\mathbf{x},\mathbf{h}))^{-1}\, \widetilde{f}_\eta(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}).$$

Assumption 7 (Approximate canonicalization and Lipschitz stability)**.** Assume there exist $\delta_A, \delta_R \geq 0$ such that for all $(x,h) \in \mathscr{S}$ and all $R \in \mathrm{SO}(3)$,

$$\|A_\theta(R\mathbf{x},\mathbf{h}) - A_\theta(x,h)\|_F \leq \delta_A, \qquad \|\widehat{R}_\theta(R\mathbf{x},\mathbf{h}) - \widehat{R}_\theta(\mathbf{x},\mathbf{h})R^\top\|_F \leq \delta_R.$$

Assume $\widetilde{f}_\eta(\cdot,h)$ is $L_{\widetilde{f}}$-Lipschitz and bounded by $B_{\widetilde{f}}$:

$$\|\widetilde{f}_\eta(u_1,\mathbf{h}) - \widetilde{f}_\eta(u_2,\mathbf{h})\| \leq L_{\widetilde{f}}\|u_1 - u_2\|_F, \qquad \|\widetilde{f}_\eta(u,\mathbf{h})\| \leq B_{\widetilde{f}}.$$

Assume $\rho(Q)^{-1}$ is L_ρ-Lipschitz in operator norm w.r.t. Q:

$$\|\rho(Q_1)^{-1} - \rho(Q_2)^{-1}\|_{\mathrm{op}} \leq L_\rho\|Q_1 - Q_2\|_F.$$

Proposition 8 (Approximate equivariance of Canonicalize-and-Lift predictors)**.** *Under Assumption 7, for all $(x,h) \in \mathscr{S}$ and $R \in \mathrm{SO}(3)$,*

$$\left\|\widehat{y}(R\mathbf{x},\mathbf{h}) - \rho(R)\widehat{y}(\mathbf{x},\mathbf{h})\right\| \;\leq\; L_{\widetilde{f}}\,\delta_A \;+\; L_\rho\, B_{\widetilde{f}}\,\delta_R.$$

Proof. Add/subtract $\rho(\widehat{R}_\theta(R\mathbf{x},\mathbf{h}))^{-1}\widetilde{f}_\eta(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})$. Bound (i) the $\widetilde{f}_\eta$ change by $L_{\widetilde{f}}\delta_A$ and (ii) the $\rho(\widehat{R}_\theta)^{-1}$ change by $L_\rho\,\delta_R$ times $\|\widetilde{f}_\eta\| \leq B_{\widetilde{f}}$, then use $\rho(\widehat{R}_\theta(\mathbf{x},\mathbf{h})R^\top)^{-1} = \rho(R)\rho(\widehat{R}_\theta(\mathbf{x},\mathbf{h}))^{-1}$. □

C Generative Modeling in Canonical Coordinates

C.1 Score matching and equivariant projection

Let p_t be a time-marginal of an isotropic noising process on $\mathbb{R}^{N\times 3}$ and

$$\mathbf{s}^*(\mathbf{x},\mathbf{t}) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

For a score model $s(x,t)$, define the population MSE

$$\mathscr{J}(s) := \mathbb{E}_t\, \mathbb{E}_{\mathbf{x}\sim p_t}\left[\|s(x,t) - s^*(\mathbf{x},t)\|_F^2\right].$$

Assumption 9 (Rotation-invariant marginals)**.** Assume p_t is $\mathrm{SO}(3)$-invariant: $p_t(R\mathbf{x}) = p_t(\mathbf{x})$ for all $R \in \mathrm{SO}(3)$. Then $\mathbf{s}^*$ is $\mathrm{SO}(3)$-equivariant: $\mathbf{s}^*(R\mathbf{x},t) = R\,s^*(\mathbf{x},t)$ (1)

Definition 1 (Equivariant projection)**.** *For a vector field $\mathbf{s}(\cdot,t)$ define*

$$(\mathscr{P}_{\mathrm{eq}} \cdot \mathbf{s})(\mathbf{x},t) := \mathbb{E}_{R\sim\mu}\left[R\mathbf{s}(R^\top\mathbf{x},t)\right], \tag{9}$$

where μ is the Haar measure on $\mathrm{SO}(3)$ and $\mathscr{P}_{\mathrm{eq}}$ denotes the equivariant projection operator that maps an arbitrary vector field to an $SO(3)$-equivariant one by Haar-averaging over rotations.

Proposition 10 (WLOG canonical coordinates for score matching)**.** *Assume (i) p_t is $\mathrm{SO}(3)$-invariant and hence s^* is $\mathrm{SO}(3)$-equivariant, and (ii) Assumption 1 holds. Then for any score field s, there exists a canonical-space field $\widetilde{s}(\cdot,t)$ such that*

$$s_{\mathrm{canonical}}(\mathbf{x},t) := \widehat{R}_\theta(\mathbf{x},\mathbf{h})^\top\, \widetilde{s}(A_\theta(\mathbf{x},\mathbf{h}),t) \tag{10}$$

satisfies

$$\mathscr{J}(s_{\mathrm{canonical}}) \;\leq\; \mathscr{J}(s). \tag{11}$$

Proof. We construct $s_{\mathrm{canonical}}$ via the equivariant projection. Let $\Delta := s - s^*$. Since s^* is $\mathrm{SO}(3)$-equivariant, $\mathscr{P}_{\mathrm{eq}}s^* = s^*$, and thus $\mathscr{P}_{\mathrm{eq}}s - s^* = \mathscr{P}_{\mathrm{eq}}\Delta$. By Jensen's inequality and orthogonality invariance of $\|\cdot\|_F$,

$$\|\mathscr{P}_{\mathrm{eq}}\Delta(\mathbf{x},t)\|_F^2 = \left\|\mathbb{E}_R[R\,\Delta(R^\top\mathbf{x},t)]\right\|_F^2 \leq \mathbb{E}_R\|\Delta(R^\top\mathbf{x},t)\|_F^2.$$

Taking $\mathbb{E}_t\mathbb{E}_{\mathbf{x}\sim p_t}$ and using $\mathrm{SO}(3)$-invariance of p_t yields $\mathscr{J}(\mathscr{P}_{\mathrm{eq}}s) \leq \mathscr{J}(s)$.

Now, $\mathscr{P}_{\mathrm{eq}}s$ is $\mathrm{SO}(3)$-equivariant by construction. Under Assumption 1, any $\mathrm{SO}(3)$-equivariant score admits the canonicalize-and-lift form (cf. Proposition 6):

$$(\mathscr{P}_{\mathrm{eq}}s)(\mathbf{x},t) = \widehat{R}_\theta(\mathbf{x},\mathbf{h})^\top\widetilde{s}(A_\theta(\mathbf{x},\mathbf{h}),t)$$

for some $\widetilde{s}$. Setting $s_{\mathrm{canonical}} := \mathscr{P}_{\mathrm{eq}}s$ completes the proof. □

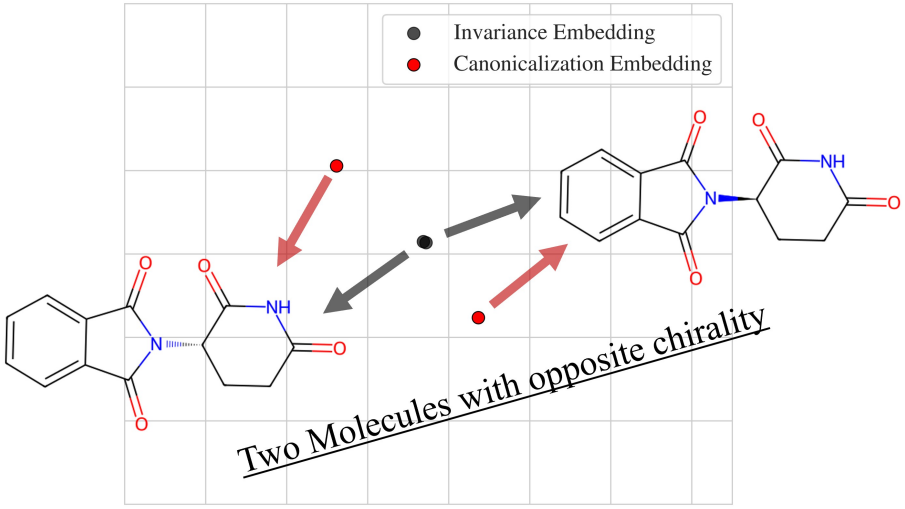


Figure 1. Scatter plots of Repr.**A** (Uni-Mol2) and Repr.**Z** (THEMOL) embeddings for a pair of molecules with opposite chirality.

C.2 Quotient-space evaluation beyond the ideal section

Let $G = \mathrm{SO}(3)$, $\varphi(\mathbf{x}) := [\mathbf{x}]_G$, and $d_{\mathrm{orb}}(\varphi(\mathbf{x}), \varphi(y)) := \min_{R \in \mathrm{SO}(3)} \|R\mathbf{x} - y\|_F$. Let W_1^{orb} be the 1-Wasserstein distance induced by d_{orb}.

Assumption 11 (Approximate orbit consistency)**.** On $\mathscr{S}$, for every $(\mathbf{x}, \mathbf{h})$ there exists $R \in \mathrm{SO}(3)$ such that $\|A_\theta(\mathbf{x}, \mathbf{h}) - R\mathbf{x}\|_F \leq \delta_A$.

Proposition 12 (Orbit-space guarantee under approximate canonicalization)**.** *Let q be a target distribution on $\mathscr{S}$, with $\widetilde{q} := \varphi_\# q$ and $q_A := (A_\theta)_\# q$. Let p be obtained by sampling $u \sim p_A$ and lifting $x := R^\top u$ with $R \sim \mu$ (Haar), and set $\widetilde{p} := \varphi_\# p$. Under Assumption 11,*

$$W_1^{\mathrm{orb}}(\widetilde{p}, \widetilde{q}) \leq W_1(p_A, q_A) + \delta_A, \qquad \textit{hence if } p_A = q_A \textit{ then } W_1^{\mathrm{orb}}(\widetilde{p}, \widetilde{q}) \leq \delta_A.$$

Proof (sketch). Rotation-invariance gives $\widetilde{p} = \varphi_\# p_A$. Since φ is 1-Lipschitz from $(\mathbb{R}^{N\times 3}, \|\cdot\|_F)$ to $(\varphi(\mathbb{R}^{N\times 3}), d_{\mathrm{orb}})$, $W_1^{\mathrm{orb}}(\varphi_\# p_A, \varphi_\# q_A) \leq W_1(p_A, q_A)$. Couple $(x,h) \sim q$ with $u = A_\theta(x,h)$; by Assumption 11, $d_{\mathrm{orb}}(\varphi(u), \varphi(x)) \leq \delta_A$, hence $W_1^{\mathrm{orb}}(\varphi_\# q_A, \varphi_\# q) \leq \delta_A$. Apply triangle inequality. □

C.3 Equivariant scores without equivariant layers

Assume p_t is $\mathrm{SO}(3)$-invariant so that the true score $s^*(x,t) = \nabla_x \log p_t(x)$ is $\mathrm{SO}(3)$-equivariant. Under Assumption 1, any $\mathrm{SO}(3)$-equivariant score field admits the canonicalize-and-lift parameterization (cf. Proposition 6):

$$s(x,t) = \widehat{R}_\theta(x,h)^\top \widetilde{s}(u,t), \tag{12}$$

for some canonical-space field $\widetilde{s}(\cdot, t)$.

Remark 1 (Comparison to SE(3)-equivariant backbones)**.** *Equation 12 implies that restricting to canonical-space modeling does not shrink the class of equivariant scores under exact canonicalization assumptions. Consequently, SE(3)-equivariant layers are not required for expressivity: equivariance can be enforced at the interface by a learned gauge selector and a lift, while the backbone operates in canonical coordinates.*

Remark 2 (Computational implication in iterative denoising)**.** *In diffusion or score-based generation with T denoising steps, an SE(3)-equivariant backbone enforces equivariance at every layer and at every step. The canonicalize-and-lift construction enforces symmetry through (i) canonicalization and (ii) a lightweight lift, while the main backbone can be non-equivariant. Hence, when equivariant layers are substantially more expensive than canonicalization plus a standard backbone, the per-step overhead compounds over T steps. This separation is most beneficial in foundation settings where the same backbone is reused across heterogeneous tasks, including invariant objectives and equivariant outputs.*

D Task-Driven Inductive Bias: Canonicalization vs. Equivariance vs. Augmentation

D.1 Limitations of invariant architectures

Purely invariant representations face two fundamental limitations for foundation settings.

Lemma 13 (Invariant latent cannot identify pose)**.** *Let* Enc *be* $SE(3)$*-invariant in* $\mathbf{x}$*:* $\mathrm{Enc}(R\mathbf{x}+\mathbf{1}t^\top,\mathbf{h})=\mathrm{Enc}(\mathbf{x},\mathbf{h})$ *for all* $(R,t)\in SE(3)$*. Assume there exist* $(\mathbf{x},\mathbf{h})$ *and a nontrivial* $(R,t)\in SE(3)$ *such that* $(R\mathbf{x}+\mathbf{1}t^\top,\mathbf{h})\neq(\mathbf{x},\mathbf{h})$*. Then no deterministic decoder* Dec *that takes only* $z=\mathrm{Enc}(\mathbf{x},\mathbf{h})$ *can satisfy* $\mathrm{Dec}(\mathrm{Enc}(\mathbf{x},\mathbf{h}))=\mathbf{x}$ *simultaneously for both* $\mathbf{x}$ *and* $R\mathbf{x}+\mathbf{1}t^\top$*.*

Proof. Let $z=\mathrm{Enc}(\mathbf{x},\mathbf{h})=\mathrm{Enc}(R\mathbf{x}+\mathbf{1}t^\top,\mathbf{h})$ by invariance. If Dec depends only on z, then $\mathrm{Dec}(z)$ must equal both $\mathbf{x}$ and $R\mathbf{x}+\mathbf{1}t^\top$, a contradiction. □

Proposition 14 (Chirality loss in distance-only invariant latents)**.** *Assume* $\mathrm{Enc}_{\mathbf{h}}(\mathbf{x},\mathbf{h})=\Phi(\mathbf{D}(\mathbf{x}),\mathbf{h})$ *where* $\mathbf{D}(\mathbf{x})_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|_2$*. Then* $\mathrm{Enc}_{\mathbf{h}}(Q\mathbf{x},\mathbf{h})=\mathrm{Enc}_{\mathbf{h}}(\mathbf{x},\mathbf{h})$ *for all* $Q\in O(3)$*. Consequently, there exists an* $\mathrm{SO}(3)$*-invariant target* y *that cannot be written as a function of* $\mathrm{Enc}_{\mathbf{h}}(\mathbf{x},\mathbf{h})$*, as shown Figure 1.*

Proof sketch. Distance matrices satisfy $\mathbf{D}(Q\mathbf{x})=\mathbf{D}(\mathbf{x})$ for all $Q\in O(3)$, hence $\mathrm{Enc}_{\mathbf{h}}(Q\mathbf{x},\mathbf{h})=\mathrm{Enc}_{\mathbf{h}}(\mathbf{x},\mathbf{h})$. Let

$$y(\mathbf{x}):=\det\big[(\mathbf{x}_i-\mathbf{x}_\ell),(\mathbf{x}_j-\mathbf{x}_\ell),(\mathbf{x}_k-\mathbf{x}_\ell)\big]$$

for some indices i,j,k,ℓ. Then $y(R\mathbf{x})=y(\mathbf{x})$ for all $R\in\mathrm{SO}(3)$ (since $\det(R)=1$), while $y(Q\mathbf{x})=-y(\mathbf{x})$ for any reflection $Q\in O(3)\setminus\mathrm{SO}(3)$ (since $\det(Q)=-1$). Choose $\mathbf{x}$ with $y(\mathbf{x})\neq 0$ and set $\mathbf{x}':=Q\mathbf{x}$; then $\mathrm{Enc}_{\mathbf{h}}(\mathbf{x}',\mathbf{h})=\mathrm{Enc}_{\mathbf{h}}(\mathbf{x},\mathbf{h})$ but $y(\mathbf{x}')\neq y(\mathbf{x})$, so y cannot be a function of $\mathrm{Enc}_{\mathbf{h}}$. □

D.2 Limitations of equivariant architectures

Equivariant representations provide a principled and often highly effective inductive bias for force prediction, vector-field estimation, coordinate-level generation, and other tasks whose outputs transform under rotations. Modern equivariant networks can also propagate scalar invariant channels together with higher-order geometric features, making them capable of handling both invariant and equivariant prediction targets. Our claim is instead specific to the unified autoencoder setting considered in this work: **the same pretrained latent is not only consumed by prediction heads, but also used as the state space for latent generative modeling and black-box molecular optimization. In this setting, the geometry of the latent space itself becomes a central design issue.**

Equivariant latent orbits. Let Enc be an equivariant encoder and let $\mathbf{z}=\mathrm{Enc}(\mathbf{x},\mathbf{h})$. Then, for any $R\in\mathrm{SO}(3)$,

$$\mathrm{Enc}(\mathbf{R}\mathbf{x},\mathbf{h})=\rho(R)\mathrm{Enc}(\mathbf{x},\mathbf{h}),$$

where ρ denotes the induced representation on the latent space. Hence, a single molecular structure corresponds not to one latent point, but to an orbit

$$\mathscr{O}_z=\{\rho(R)z:R\in\mathrm{SO}(3)\}.$$

This is the correct behavior for equivariant prediction, but it also means that **the latent represents both intrinsic molecular variation and global orientation**. Thus, when the latent is reused for generation or optimization, the model operates on an orbit-lifted space rather than directly on the quotient space of molecular structures modulo global rotations.

Implication for latent generative modeling. For latent diffusion or flow matching, orbit-lifted equivariant latents require the generator to model variations that are chemically redundant. If a clean latent can be written as $Z_1=\rho(R)Z_{\mathbf{c}}$, where $Z_{\mathbf{c}}$ denotes the intrinsic molecular latent and $R\sim\mu_{\mathrm{Haar}}$ is a global rotation, then the target distribution contains orientation-induced variability in addition to molecular variability. In high-noise regimes, the corrupted state carries limited information about the clean orientation, so the denoising or flow target can include uncertainty over R that is irrelevant to molecular identity(2). This does not make equivariant generation invalid; rather, it indicates that the generator must spend capacity modeling global orbit variation that a quotient- or canonical-space formulation can avoid.

Implication for latent-space optimization. The same orbit structure can reduce the efficiency of black-box latent optimization. In our setting, optimization searches a pretrained latent space while keeping the generator and decoder fixed. If rotated versions of the same molecule occupy different latent points, then an optimizer may explore multiple regions corresponding to the same underlying molecular candidate. Moreover, an unconstrained search direction in an equivariant latent can mix chemically meaningful molecular changes with changes in global orientation or gauge. Restricting the search only to invariant scalar components may reduce this redundancy, but it can also remove pose information needed for faithful 3D reconstruction and coordinate-level generation. Thus, the issue is not the predictive expressivity of equivariant modules, but the suitability of the induced latent geometry as a shared search and generation space.

Gauge handling as interface-level bookkeeping. Canonicalization does not eliminate symmetry handling; it relocates it. Any 3D coordinate-based model must account for global orientation in some form. In a fully equivariant latent, this orientation dependence remains inside the reusable latent through the orbit $\mathscr{O}_z$. In our formulation, a learned selector $\widehat{R}_\theta(\mathbf{x},\mathbf{h})$ first maps the input to a canonical representative,

$$A_\theta(\mathbf{x},\mathbf{h}) = \widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h})\mathbf{x}^{\mathbf{c}},$$

with the desired consistency condition

$$\widehat{\mathbf{R}}_\theta(\mathbf{R}\mathbf{x},\mathbf{h}) = \widehat{\mathbf{R}}_\theta(\mathbf{x},\mathbf{h})\mathbf{R}^\top \quad \Longleftrightarrow \quad A_\theta(\mathbf{R}\mathbf{x},\mathbf{h}) = A_\theta(\mathbf{x},\mathbf{h}).$$

The reusable latent is then learned in this canonical frame, while the selected rotation is retained only for lifting coordinate-level or equivariant outputs back to the original frame. Therefore, gauge bookkeeping is preserved at the input/output interface, but removed from the latent space on which generation and optimization operate.

D.3 Limitations of rotation augmentation

Rotation augmentation is a common alternative to architectural equivariance. We show that it provides weaker guarantees than canonicalization.

Setup. Let $G = \mathrm{SO}(3)$ with Haar measure μ, and let y be G-invariant: $y(R\mathbf{x},\mathbf{h}) = y(\mathbf{x},\mathbf{h})$. For a predictor f, define the rotation-augmentation risk

$$\mathscr{R}_{\mathrm{aug}}(f) := \mathbb{E}_{(\mathbf{x},\mathbf{h})}\,\mathbb{E}_{R\sim\mu}\big[\ell(f(R\mathbf{x},\mathbf{h}), y(\mathbf{x},\mathbf{h}))\big], \qquad (Pf)(\mathbf{x},\mathbf{h}) := \mathbb{E}_{R\sim\mu}\big[f(R\mathbf{x},\mathbf{h})\big].$$

We distinguish *single-view inference* (one forward pass) from *group-averaged inference* that explicitly computes Pf at test time. *Interpretation:* group-averaged inference is a Monte-Carlo approximation of a group integral, i.e., invariance is obtained by paying inference-time compute. For a non-equivariant backbone, achieving small variation of $f(R\mathbf{x},\mathbf{h})$ over $R \in \mathrm{SO}(3)$ is itself a hard learning problem; when such variation persists, the required number of views K becomes large.

Average risk does not control worst-case invariance.

Proposition 15 (Augmentation risk does not certify worst-case invariance)**.** *Fix $(\mathbf{x},\mathbf{h})$ and consider the orbit $\mathscr{O}(\mathbf{x}) = \{(R\mathbf{x},\mathbf{h}) : R \in G\}$. For squared loss $\ell(a,b) = (a-b)^2$ and invariant label $y(\mathbf{x},\mathbf{h}) \equiv 0$, for any $\varepsilon > 0$ and $M > 0$, there exists a measurable f on $\mathscr{O}(\mathbf{x})$ such that*

$$\mathbb{E}_{R\sim\mu}\big[\ell(f(R\mathbf{x},\mathbf{h}),0)\big] \le \varepsilon \quad \textit{but} \quad \sup_{R\in G} |f(R\mathbf{x},\mathbf{h}) - f(\mathbf{x},\mathbf{h})| \ge M.$$

Proof sketch. Choose $A \subseteq G$ with $\mu(A) = \varepsilon/M^2$ and set $f(R\mathbf{x},\mathbf{h}) = M \cdot \mathbf{1}\{R \in A\}$, $f(\mathbf{x},\mathbf{h}) = 0$. □

This formalizes a structural gap: minimizing an *average* risk over random rotations does not control *worst-case* rotational variation under single-view inference. Canonicalization provides worst-case control directly via $\sup_R \|A_\theta(R\mathbf{x},\mathbf{h}) - A_\theta(\mathbf{x},\mathbf{h})\|_F \le \delta$ and Proposition 5.

Group averaging incurs multi-view cost. To obtain an explicitly invariant predictor, one must approximate Pf via Monte-Carlo averaging:

$$\widehat{f}_K(\mathbf{x},\mathbf{h}) := \frac{1}{K}\sum_{k=1}^{K} f(R_k\mathbf{x},\mathbf{h}), \qquad R_k \overset{\text{i.i.d.}}{\sim} \mu.$$

This estimator directly approximates the group integral defining Pf; unless $f(R\mathbf{x},\mathbf{h})$ is already nearly constant over R, its variance can be large, forcing large K. The mean-square approximation error satisfies $\mathbb{E}[(\widehat{f}_K(\mathbf{x},\mathbf{h}) - (Pf)(\mathbf{x},\mathbf{h}))^2] = \mathrm{Var}_{R\sim\mu}(f(R\mathbf{x},\mathbf{h}))/K$, so achieving MSE $\le \varepsilon^2$ requires $K = \Omega(\varepsilon^{-2})$.

Uniform control requires covering $\mathrm{SO}(3)$**.** For worst-case guarantees, we need uniform control over all rotations.

Proposition 16 (Covering number scaling of $\mathrm{SO}(3)$)**.** *Assume f is L_{rot}-Lipschitz on rotation orbits under a bi-invariant geodesic distance on* $\mathrm{SO}(3)$*. Then achieving ε-uniform control, i.e.,* $\sup_{R\in\mathrm{SO}(3)} |f(R\mathbf{x},\mathbf{h}) - \widehat{f}_K(\mathbf{x},\mathbf{h})| \le \varepsilon$*, requires*

$$K \;=\; \Omega\!\left(\left(\frac{L_{\mathrm{rot}}}{\varepsilon}\right)^3\right).$$

Remark 3 (Why the exponent is 3)**.** $\mathrm{SO}(3)$ *is a compact 3-dimensional Riemannian manifold. Metric balls have volume $\Theta(\eta^3)$, so covering* $\mathrm{SO}(3)$ *with η-balls requires $\Omega(\eta^{-3})$ balls.*

Summary: canonicalization vs. augmentation. Canonicalization achieves worst-case stability $\sup_R |\widehat{y}(R\mathbf{x},\mathbf{h}) - \widehat{y}(\mathbf{x},\mathbf{h})| \leq L_f\delta$ with *one* backbone pass (Proposition 5). In contrast, multi-view augmentation realizes invariance by approximating a group integral at inference time, shifting the burden to compute: it requires $K = \Omega(\varepsilon^{-2})$ views for MSE control and $K = \Omega(\varepsilon^{-3})$ views for uniform control, and it asks a non-equivariant backbone to implicitly learn small rotation-wise variation of $f(R\mathbf{x},\mathbf{h})$ over $\mathrm{SO}(3)$ compute tradeoff that canonicalization avoids.

D.4 Canonicalization as a unified solution

The preceding analysis highlights complementary limitations of common symmetry-handling strategies: (i) pose-level coordinate reconstruction is not identifiable from a purely invariant latent without a gauge/section, (ii) equivariant coordinate-state modeling can be computationally demanding in iterative generation, and (iii) augmentation alone does not provide worst-case guarantees without paying multi-view inference cost. Under mild section/stability assumptions, canonicalization offers a unified route that addresses these issues within a single framework.

Proposition 17 (Task-wise factorization via canonicalization)**.** *Assume Assumption 1. Then for any continuous invariant target y_{inv}, there exists a continuous head g_{inv} such that*

$$y_{\mathrm{inv}}(\mathbf{x},\mathbf{h}) = g_{\mathrm{inv}}\big(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}\big).$$

Moreover, for any continuous ρ-equivariant target y_{eq}, there exists a continuous head g_{eq} such that

$$y_{\mathrm{eq}}(\mathbf{x},\mathbf{h}) = \rho\big(\widehat{R}_\theta(\mathbf{x},\mathbf{h})\big)^{-1} g_{\mathrm{eq}}\big(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h}\big).$$

Under Assumptions 4 and 7, the same statements hold up to an additive stability term bounded by $C_{\mathrm{inv}}\delta$ and $C_{\mathrm{eq}}\delta$, respectively.

This proposition shows that canonicalization reduces both invariant prediction and equivariant outputs to learning in canonical coordinates, with task-specific heads and an explicit lifting by $\rho(\widehat{R}_\theta)^{-1}$. For generation tasks evaluated on rotation orbits, learning in canonical space and applying uniform lifting yields orbit-level fidelity controlled by the canonical modeling error plus an additive alignment term (Section C.2).

E How Should We Evaluate Molecular Representations?

E.1 Baseline Models for Representation Evaluation

To comprehensively evaluate the quality of learned 3D molecular representations, we compare against three representative foundation models that employ distinct pretraining strategies and architectural designs.

Uni-Mol2 (3) is a 3D molecular representation learning framework that adopts SE(3)-invariant representations (REPR.**A** in our taxonomy). The model employs a Transformer architecture operating on pairwise distance matrices and 3D spatial encodings derived from inter-atomic distances, pretrained via masked atom prediction and coordinate denoising on large-scale molecular conformations. While Uni-Mol incorporates equivariant layers in its output head for coordinate prediction, the core molecular embeddings remain inherently invariant.

JMP (4) extends molecular foundation modeling to materials science by pretraining on diverse atomic systems spanning small molecules, crystals, and bulk materials. JMP employs GemNet-OC as its backbone, which uses directed edge embeddings with geometric message passing based on distances, angles, and dihedral angles. Notably, while GemNet-OC can produce rotationally equivariant predictions (e.g., forces) through gradient computation or direct equivariant output blocks, the internal representations themselves remain SE(3)-invariant. This places JMP within the REPR.**A** category, similar to Uni-Mol, though with a more expressive geometric feature set incorporating higher-order invariants. The model is pretrained on approximately 120M systems from OC20, OC22, ANI-1x, and Transition-1x using multi-task learning objectives.

3D-MoLM (5) represents a multimodal approach that bridges 3D molecular structures with natural language understanding. The model integrates a 3D molecular encoder with a large language model through a cross-modal projector, enabling molecule-text interpretation tasks. The 3D encoder component, which we utilize for embedding extraction in our evaluation, processes molecular geometries through attention mechanisms with 3D positional encodings. This model provides an interesting comparison point as it learns molecular representations optimized for cross-modal alignment rather than purely geometric reconstruction or property prediction objectives, potentially capturing different aspects of molecular semantics.

These three baselines collectively represent prevalent approaches in 3D molecular foundation modeling: invariant representations with distance-based features (Uni-Mol), invariant representations with higher-order geometric invariants (JMP), and multimodal learning with cross-modal alignment objectives (3D-MoLM). By evaluating our canonicalization-based approach (**REPR.Z**) against these models, we systematically assess whether the proposed framework offers improved representation quality over existing 3D molecular foundation models across diverse evaluation criteria.

E.2 Molecular Identity Preservation Analysis

E.2.1 Dataset Construction

To evaluate whether learned representations appropriately capture molecular identity while accommodating conformational flexibility, we constructed a separate evaluation dataset from the PDBbind 2020 database (6).

We employed stratified sampling based on heavy atom count, partitioning molecules into five strata (10–20, 20–30, 30–40, 40–50, and 50–100 heavy atoms) and sampling 200 molecules per stratum to yield 1,000 unique molecules. For each molecule, five conformers were generated using the ETKDGv3 algorithm (7; 8) in RDKit (9), resulting in 5,000 conformer structures. Individual conformer embeddings $\mathbf{z}_{m,k}$ were retained to assess intra-molecular consistency versus inter-molecular separation.

E.2.2 Evaluation Evaluation

To quantify how well learned representations preserve molecular identity while distinguishing between different molecules, we evaluate the separability of same-molecule conformer pairs versus different-molecule pairs in the embedding space.

Given a dataset of M molecules, each with K conformers, we construct two types of embedding pairs:

- Positive pairs (intra-molecular): Pairs of embeddings from different conformers of the same molecule, i.e., $(\mathbf{z}_{m,k}, \mathbf{z}_{m,k'})$ where $k \neq k'$. These pairs should exhibit small distances if the representation correctly captures molecular identity invariant to conformational changes.
- Negative pairs (inter-molecular): Pairs of embeddings from different molecules, i.e., $(\mathbf{z}_{m,k}, \mathbf{z}_{m',k'})$ where $m \neq m'$. These pairs should exhibit large distances if the representation effectively discriminates between distinct molecules.

Separation Score via ROC-AUC. We assess the discriminability between positive and negative pairs using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC). For each pair, we compute the Euclidean distance between embeddings:

$$d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2 \tag{13}$$

We then frame the evaluation as a binary classification task where the objective is to distinguish positive pairs (label 1) from negative pairs (label 0) based on the pairwise distance. Since positive pairs should have smaller distances, we use the negative distance $-d_{ij}$ as the prediction score. The ROC-AUC is computed over all sampled pairs:

$$\text{ROC-AUC} = P\left(d_{ij}^{(-)} > d_{ij}^{(+)}\right) \tag{14}$$

where $d_{ij}^{(+)}$ denotes distances of positive pairs and $d_{ij}^{(-)}$ denotes distances of negative pairs.

An ROC-AUC of 1.0 indicates perfect separation, where all intra-molecular distances are smaller than all inter-molecular distances. An ROC-AUC of 0.5 indicates no discriminability, equivalent to random assignment. This metric provides a single scalar summary of how well the embedding space clusters conformers of the same molecule while separating different molecules, without requiring any downstream task-specific training.

E.3 Chemical Semantic Consistency

E.3.1 Dataset Construction

To evaluate whether learned representations capture chemically meaningful structure, we constructed two evaluation datasets from the PDBbind 2020 database (6): one based on functional groups and another based on molecular scaffolds.

Functional Group Dataset. We selected five functional groups (amide, ether, halogen, hydroxyl, and primary amine) identified via SMARTS pattern matching in RDKit (9). To ensure unambiguous assignment, only molecules containing exactly one target functional group were included. We sampled 100 molecules per group (500 total), generating 10 conformers each using ETKDGv3 (7; 8) for 5,000 structures. Molecule-level embeddings were obtained by averaging conformer embeddings.

Scaffold Dataset. Generic scaffolds were extracted using Murcko decomposition (10), retaining only core ring systems and linker topology. We identified 19 scaffolds with at least 100 molecules each in PDBbind 2020, sampling 100 molecules per scaffold (1,900 total). Following the same conformer protocol yielded 19,000 structures with averaged molecule-level embeddings.

E.3.2 Evaluation Metrics

To quantify how well the learned embedding space reflects chemical group structure, we compute a coherence score that measures the ratio of inter-group separation to intra-group compactness.

For each chemical group g (either functional group or scaffold), we compute the mean pairwise Euclidean distance among all molecules within that group:

$$D_{\text{intra}}(g) = \frac{1}{|g|(|g|-1)} \sum_{i,j \in g, i \neq j} \|\mathbf{z}_i - \mathbf{z}_j\|_2 \tag{15}$$

where $\mathbf{z}_i$ denotes the embedding of molecule i and $|g|$ is the number of molecules in group g. The overall intra-group distance is the average across all groups:

$$D_{\text{intra}} = \frac{1}{|G|} \sum_{g \in G} D_{\text{intra}}(g) \tag{16}$$

where G is the set of all groups. We first compute the centroid of each group:

$$\boldsymbol{\mu}_g = \frac{1}{|g|} \sum_{i \in g} \mathbf{z}_i \tag{17}$$

The inter-group distance is defined as the mean Euclidean distance between all pairs of group centroids:

$$D_{\text{inter}} = \frac{2}{|G|(|G|-1)} \sum_{g,h \in G, g \neq h} \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|_2 \tag{18}$$

Coherence Score. The coherence score is defined as the ratio of inter-group distance to intra-group distance:

$$\text{Coherence} = \frac{D_{\text{inter}}}{D_{\text{intra}}} \tag{19}$$

A higher coherence score indicates that molecules sharing the same chemical group are embedded closer together while molecules from different groups are well separated. This metric provides a quantitative assessment of how well the embedding space aligns with known chemical taxonomy, without requiring any task-specific supervision.

E.4 Embedding Space Geometry Analysis

E.4.1 Dataset Construction

We sampled 50,000 molecules from the PDBbind 2020 database (6) using stratified sampling across seven heavy atom count ranges (10–50) to ensure molecular size diversity. Duplicates were removed via canonical SMILES comparison. For each molecule, 10 conformers were generated using ETKDGv3 in RDKit, yielding 500,000 structures total. Molecule-level embeddings were computed by averaging across conformers:

$$\mathbf{z}_{\text{mol}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}^{(k)} \tag{20}$$

where $\mathbf{z}^{(k)}$ denotes the k-th conformer embedding and $K = 10$. All models were evaluated on identical molecule-conformer sets for fair comparison.

E.4.2 Evaluation Metrics

We evaluate the structural properties of learned embedding spaces using three complementary metrics. These metrics do not directly measure representation quality in terms of downstream performance; rather, they characterize whether the embedding space satisfies basic geometric prerequisites for reliable distance-based interpretation (11; 12).

Effective Dimension. The effective dimension quantifies how many dimensions are substantially utilized by the embedding distribution. Given a set of embeddings $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^d$, we first center the embeddings by subtracting the mean and then compute the covariance matrix:

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i - \frac{1}{n} \sum_{j=1}^{n} \mathbf{z}_j, \quad \boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^\top \tag{21}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ denote the eigenvalues of $\boldsymbol{\Sigma}$. The effective dimension is defined via the participation ratio:

$$D_{\text{eff}} = \frac{\left(\sum_{i=1}^{d} \lambda_i\right)^2}{\sum_{i=1}^{d} \lambda_i^2} \tag{22}$$

This metric ranges from 1 (all variance concentrated in a single direction) to d (variance uniformly distributed across all dimensions). To enable comparison across models with different embedding dimensions, we report the normalized effective dimension $D_{\text{eff}}/d \in (0,1]$. A low D_{eff}/d indicates that the embedding distribution is effectively confined to a low-dimensional subspace, suggesting limited representational capacity regardless of the nominal dimensionality.

Hubness. Hubness is a phenomenon in high-dimensional spaces where certain points appear disproportionately often as nearest neighbors of other points (11). Such points, called *hubs*, distort distance-based neighbor relationships and undermine the reliability of similarity-based retrieval and classification.

For each embedding $\mathbf{z}_i$, we compute its k-occurrence $N_k(i)$, defined as the number of times $\mathbf{z}_i$ appears among the k-nearest neighbors of all other points: $N_k(i) = \left|\left\{ j \neq i : \mathbf{z}_i \in \text{kNN}(\mathbf{z}_j) \right\}\right|$ We then measure the skewness of the k-occurrence distribution:

$$S_{N_k} = \frac{\mathbb{E}\left[(N_k - \mu_{N_k})^3\right]}{\sigma_{N_k}^3} \tag{23}$$

where μ_{N_k} and σ_{N_k} are the mean and standard deviation of $\{N_k(i)\}_{i=1}^n$.

A skewness close to zero indicates that all points serve as neighbors with similar frequency, implying stable and symmetric neighbor relationships. High positive skewness indicates the presence of hubs, which can compromise distance-based analyses. We use $k = 10$ throughout our experiments.

Uniformity. Uniformity measures how evenly the embeddings are distributed on the hypersphere, serving as an indicator of representation collapse (13). We first ℓ_2-normalize the embeddings to obtain unit vectors: $\mathbf{u}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}$

The uniformity loss is defined as:

$$L_{\text{uniform}} = \log \mathbb{E}_{i \neq j}\left[\exp\left(-t\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\right)\right] \tag{24}$$

where $t > 0$ is a temperature parameter, set to $t = 2$ following standard practice.

A lower (more negative) uniformity value indicates that embeddings are spread uniformly across the hypersphere, suggesting that the representation space is efficiently utilized without collapse. A higher value (closer to zero) indicates that embeddings are concentrated in a small region, which may signal representation collapse where the model fails to preserve discriminative information across different molecules.

E.5 Pretrain-Finetune Alignment Analysis

E.5.1 DATASET CONSTRUCTION

To evaluate how well pretrained representations align with downstream task distributions, we constructed evaluation datasets comprising both pretraining and finetuning molecular sets for each model.

For each model under evaluation, we sampled 50,000 molecules from its respective pretraining dataset. This model-specific sampling ensures that the alignment analysis reflects the actual distributional relationship between each model's learned representations and downstream tasks. The three-dimensional molecular structures provided in the original pretraining datasets were used directly without additional conformer generation.

We compiled finetuning datasets from two widely-used molecular property prediction benchmarks: **MoleculeNet** (14): We selected six tasks spanning classification and regression: BACE, BBBP, ClinTox, Lipophilicity, ESOL, and FreeSolv , **Therapeutics Data Commons (TDC)** (15): We included the ADMET benchmark tasks, which comprise pharmacokinetic and toxicity endpoints relevant to drug discovery.

For each molecule in the finetuning datasets, we generated three conformers using RDKit (9) to capture conformational diversity. Molecule-level embeddings were computed by averaging the embeddings across the three conformers. To ensure that the alignment analysis reflects genuine distributional differences rather than memorization effects, we removed molecules appearing in both the pretraining and finetuning sets. Duplicate detection was performed using canonical SMILES comparison.

E.5.2 Evaluation Metric

To quantify the alignment between pretraining and finetuning distributions in the embedding space, we measure the similarity of their principal subspaces using principal angles.

Given a set of pretraining embeddings $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$ and finetuning embeddings $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$, we first center each distribution by subtracting its mean:

$$\tilde{\mathbf{p}}_i = \mathbf{p}_i - \boldsymbol{\mu}_P, \quad \tilde{\mathbf{f}}_j = \mathbf{f}_j - \boldsymbol{\mu}_F \tag{25}$$

where $\boldsymbol{\mu}_P = \frac{1}{n}\sum_{i=1}^n \mathbf{p}_i$ and $\boldsymbol{\mu}_F = \frac{1}{m}\sum_{j=1}^m \mathbf{f}_j$. We then compute the covariance matrices:

$$\boldsymbol{\Sigma}_P = \frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{p}}_i \tilde{\mathbf{p}}_i^\top, \quad \boldsymbol{\Sigma}_F = \frac{1}{m}\sum_{j=1}^m \tilde{\mathbf{f}}_j \tilde{\mathbf{f}}_j^\top \tag{26}$$

Table 1. Quantitative results of the representation quality experiment. All values are normalized.

Datasets	AUC ↑	Uniformity ↑	Coherence ↑	Hubness ↑	Effectiveness ↑	Mean↑
JMP	-1.66	**1.266**	-0.48	0.098	-0.68	0.862
UniMol2	**0.831**	-1.14	-0.932	-1.406	-0.36	0.923
3D-MOLM	0.093	-0.79	-0.26	-0.107	-0.67	0.792
THEMOL	0.673	0.673	**1.681**	**1.415**	**1.718**	**0.966**

For each covariance matrix, we perform eigendecomposition and extract the top-k eigenvectors corresponding to the k largest eigenvalues, where k is set to 70% of the embedding dimension. This choice ensures that the principal subspace captures the majority of variance while excluding minor directions that may correspond to noise. Let $\mathbf{V}_P = [\mathbf{v}_P^{(1)}, \ldots, \mathbf{v}_P^{(k)}]$ and $\mathbf{V}_F = [\mathbf{v}_F^{(1)}, \ldots, \mathbf{v}_F^{(k)}]$ denote the matrices whose columns are the top-k eigenvectors of $\mathbf{\Sigma}_P$ and $\mathbf{\Sigma}_F$, respectively. These matrices define k-dimensional subspaces that capture the principal directions of variation in each distribution.

The principal angles $\theta_1, \theta_2, \ldots, \theta_k$ between the two subspaces provide a canonical measure of subspace similarity (16). To compute these angles, we first orthonormalize each subspace via QR decomposition:

$$\mathbf{V}_P = \mathbf{Q}_P \mathbf{R}_P, \quad \mathbf{V}_F = \mathbf{Q}_F \mathbf{R}_F \tag{27}$$

We then compute the singular value decomposition of the inner product matrix: $\mathbf{Q}_P^\top \mathbf{Q}_F = \mathbf{U}\mathbf{S}\mathbf{W}^\top$, where $\mathbf{S} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ contains the singular values. The principal angles are obtained as: $\theta_i = \arccos(\sigma_i), \quad i = 1, \ldots, k$ Principal angles of zero indicate perfectly aligned directions, while angles of $\frac{\pi}{2}$ indicate orthogonal directions.

Alignment Score. We summarize the principal angles into a single alignment score by computing the mean cosine of the principal angles:

$$\text{Alignment Score} = \frac{1}{k}\sum_{i=1}^{k} \cos(\theta_i) = \frac{1}{k}\sum_{i=1}^{k} \sigma_i \tag{28}$$

This score ranges from 0 to 1, where higher values indicate better alignment between the pretraining and finetuning subspaces. A high alignment score means that the principal directions of variation in the finetuning data are fully captured by the principal directions learned during pretraining.

Remark. A high alignment score suggests that the pretrained representation has learned generalizable features that transfer well to downstream tasks without requiring substantial modification. Conversely, a low alignment score raises concerns about the quality of the pretrained representation as a foundation model. When the pretraining and finetuning subspaces diverge substantially, strong downstream performance may result primarily from heavy adaptation during finetuning rather than from the inherent expressiveness of the pretrained representation. In such cases, it becomes difficult to attribute downstream success to the foundation model itself, as the finetuning process effectively learns a new representation rather than leveraging the pretrained one. Since the fundamental goal of foundation models is to learn universally useful representations that transfer across diverse downstream tasks, a low alignment score may indicate a limited capacity to embed diverse molecular structures into a coherent, shared manifold.

E.6 Analysis of Representation Quality Experiment

E.6.1 Overview

We evaluate representation quality through five complementary metrics that probe different aspects of learned embeddings: AUC for molecular identity preservation, Coherence for chemical semantic consistency, Effective Dimension, Hubness, and Uniformity for embedding space geometry. Our analysis reveals systematic patterns that validate the theoretical advantages of canonicalization while exposing the trade-offs inherent in alternative representation paradigms.

E.6.2 Embedding Space Geometry: Trade-offs and Correlations

Geometric Regularity vs. Chemical Semantics. A striking trade-off emerges between geometric regularity (measured by Hubness and Uniformity) and chemical semantic preservation (measured by Coherence). UniMol achieves the best geometric regularity with the lowest Hubness (-1.406) and Uniformity (-1.146), indicating a well-structured embedding space with uniform neighbor relationships and even distribution on the hypersphere. However, it simultaneously exhibits the lowest Coherence (-0.932), suggesting that its SE(3)-invariant representation, built solely on pairwise distances, fails to capture higher-level chemical semantics such as functional group relationships and scaffold similarities.

In contrast, THEMOL demonstrates the opposite pattern: while showing relatively higher Hubness (1.415), it achieves the highest Coherence (1.681) by a substantial margin. This suggests that the canonicalization-based approach, which processes

original 3D coordinates in a canonical frame rather than reducing them to distance matrices, preserves richer chemical structural information at the cost of geometric regularity in the embedding space.

This trade-off has important implications for downstream applications. Tasks requiring chemical similarity assessment, such as scaffold hopping or functional group-based clustering, would benefit from THEMOL's semantically rich representations. Conversely, tasks relying heavily on nearest neighbor retrieval may favor UniMol's geometrically uniform embedding space.

Effective Dimension and Hubness Correlation. We observe a positive correlation between Effective Dimension and Hubness across models. THEMOL exhibits both the highest Effective Dimension (1.718) and the highest Hubness (1.415), while UniMol shows lower values for both metrics (-0.367 and -1.406, respectively).

This correlation can be explained by the curse of dimensionality: when embeddings genuinely utilize high-dimensional space (high Effective Dimension), the concentration of measure phenomenon makes certain points more likely to appear as nearest neighbors of many other points, thereby increasing Hubness. THEMOL's high Effective Dimension indicates that its embedding space encodes rich information across many dimensions, but this inevitably introduces hub phenomena as a side effect.

Conversely, UniMol's low Effective Dimension suggests that its embeddings are effectively confined to a lower-dimensional manifold, which naturally mitigates hubness but may also indicate limited representational capacity. This interpretation aligns with the theoretical limitations of distance-only invariant features discussed in Section D.1.

E.6.3 Pretrain-Finetune Alignment Analysis

Generalizability and Credit Assignment. THEMOL achieves the highest Alignment Score (0.9663), indicating that 96.6% of the principal variation directions in finetuning data are already captured by the directions learned during pretraining. This suggests that the canonicalization-based approach learns representations that are inherently aligned with the molecular variations relevant to downstream tasks, requiring minimal adaptation during finetuning.

The Alignment Score provides a principled basis for attributing downstream performance to either pretrained representation quality or finetuning adaptation. For models with high alignment scores (THEMOL: 0.9663, UniMol: 0.9238), strong downstream performance can be confidently attributed to the quality of the pretrained representations themselves, as the finetuning process operates within a representation space that already captures task-relevant variations.

Conversely, for models with lower alignment scores (JMP: 0.8626, 3D-MoLM: 0.7927), strong downstream performance (if achieved) may result primarily from heavy adaptation during finetuning rather than from the inherent quality of the pretrained representations. This raises concerns about the true contribution of pretraining, as the foundation model's value lies in providing transferable representations rather than serving merely as an initialization for task-specific learning.

Relationship between Effective Dimension and Alignment Score. We observe a positive correlation between Effective Dimension and Alignment Score. THEMOL exhibits both the highest Effective Dimension (Z-score: 1.718) and the highest Alignment Score (0.9663), while models with lower Effective Dimension generally show lower alignment.

This correlation admits a natural interpretation: high Effective Dimension indicates that the embedding space encodes variations across many directions. Such rich representations have a higher probability of overlapping with the diverse molecular variations present in downstream tasks. THEMOL's canonicalization approach preserves coordinate-level information without the dimensionality reduction inherent in distance-based features, enabling the model to capture a broader range of molecular variations that prove relevant for downstream transfer.

Interestingly, UniMol achieves the second-highest Alignment Score (0.9238) despite a below-average Effective Dimension (Z-score: -0.367). This suggests that while UniMol's distance-based features are limited in expressiveness, they nonetheless capture the core molecular variations most relevant to downstream property prediction tasks.

Divergence between Coherence and Alignment Score. We observe no clear correlation between Coherence and Alignment Score. Most notably, UniMol exhibits the lowest Coherence (Z-score: -0.932) but the second-highest Alignment Score (0.9238), while 3D-MoLM shows moderate Coherence (Z-score: -0.266) but the lowest Alignment Score (0.7927).

This divergence reveals that the two metrics capture fundamentally different aspects of representation quality. Coherence measures alignment with chemical taxonomy (how well the embedding space reflects functional group and scaffold relationships) capturing semantic structure. Alignment Score measures consistency between pretraining and finetuning distributions (how well pretrained representations cover downstream task variations) capturing statistical structure.

UniMol's high Alignment Score despite low Coherence suggests that current downstream benchmarks may emphasize numerical property prediction over chemical semantic reasoning. THEMOL uniquely achieves top performance on both metrics, demonstrating that canonicalization learns representations capturing both chemical semantic structure and statistical variations relevant to downstream tasks.

E.6.4 Validation of Theoretical Framework

These empirical results provide evidence supporting the theoretical analysis presented in Section 3.1. First, the limitations of invariant representations (REPR.**A**) manifest clearly in UniMol's results: despite excellent geometric regularity, the lowest

Coherence confirms that distance-only features sacrifice chemical semantic information.

Second, the canonicalization approach (REPR.$\mathbf{Z}$) demonstrates its theoretical advantage in information preservation. THEMOL's substantially higher Effective Dimension validates that processing coordinates in canonical space retains richer information content, and the superior Coherence confirms that this preserved information is chemically meaningful.

Third, THEMOL's highest Alignment Score validates that canonicalization achieves the fundamental objective of foundation models: learning universally useful representations that transfer across diverse downstream tasks. The combination of leading performance on Coherence (semantic structure) and Alignment Score (statistical structure) demonstrates THEMOL's unique capability to serve as a comprehensive foundation model for 3D molecular representation learning.

F Experimental details for canonical latent-space diagnostics

This section describes the diagnostic experiments used to evaluate whether a frozen 3D molecular representation provides a latent space suitable for prediction, decoding, interpolation and black-box optimization. Unless otherwise stated, the diagnostics use the same PDBbind-derived molecular set used for the representation-quality analysis in Supplementary Note E. We compare four representation regimes: invariant representations Repr.A, equivariant representations Repr.B, rotation-augmented non-equivariant representations Repr.C, and the proposed canonicalization-based representation Repr.Z. All diagnostics are performed without additional downstream training. For Repr.B, Repr.C and Repr.Z, the encoder, decoder and latent flow model are kept frozen, and the compared models use the same latent dimensionality, decoder capacity and training budget.

For each molecule m, the frozen encoder produces a token-level latent tensor $\mathbf{Z}_m \in \mathbb{R}^{N_{\max} \times H}$. All latent-distance computations are performed after flattening the valid token-level latent tensors. Padded tokens are masked out and all models are evaluated under the same masking convention. No task-specific normalization is fitted for these diagnostics.

Exclusion of Repr.A **from coordinate-decoding diagnostics.** Repr.A is included in the rotation-sensitivity and latent-distance hierarchy analyses, but excluded from the perturbation-decoding, interpolation-decoding and optimization-readiness diagnostics. The reason is structural: an invariant-only latent specifies molecular geometry only up to a gauge. Therefore, coordinate-level decoding or latent-space optimization requires an additional frame-selection or lifting mechanism that is not part of Repr.A itself. Panels requiring direct coordinate decoding are therefore evaluated only for Repr.B, Repr.C and Repr.Z.

F.1 Rotation sensitivity

Rotation sensitivity measures how much the latent representation changes when the same molecule is globally rotated. For each molecule m, we sample $K = 1000$ independent rotations

$$R_1, \ldots, R_K \sim \mathrm{Haar}(\mathrm{SO}(3)).$$

Using the row-vector coordinate convention, the rotated coordinates are $\mathbf{x}_m R_k^\top$. Let

$$\mathbf{z}_{m,k} = \mathrm{vec}\left[\Phi(\mathbf{x}_m R_k^\top, \mathbf{h}_m)\right]$$

denote the flattened latent representation of the k-th rotated copy, where Φ denotes the frozen representation map (ex. ζ_{Enc}). We compute the molecule-wise rotation sensitivity as

$$\bar{\mathbf{z}}_m = \frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_{m,k}, \qquad S_{\mathrm{rot}}(m) = \frac{1}{K}\sum_{k=1}^{K} \left\|\mathbf{z}_{m,k} - \bar{\mathbf{z}}_m\right\|_2^2. \tag{29}$$

Lower values indicate that global rotation induces little change in the latent representation. The figure reports the distribution of $S_{\mathrm{rot}}(m)$ over the evaluated molecule set.

F.2 Latent-distance hierarchy

The latent-distance hierarchy diagnostic evaluates whether the representation assigns distances in a chemically sensible order. An ideal reusable 3D molecular representation should assign the smallest distance to global rotations of the same molecule, a larger distance to different conformers of the same molecule, a still larger distance to matched molecular pairs with local chemical substitutions, and the largest distance to scaffold-level changes. We construct four pair types: $\mathscr{P}_{\mathrm{rot}}$, $\mathscr{P}_{\mathrm{conf}}$, $\mathscr{P}_{\mathrm{MMP}}$, $\mathscr{P}_{\mathrm{scaf}}$.

Rotation pairs are generated by applying two independent random rotations to the same conformer. Conformer pairs are different conformers of the same molecule. Matched molecular pairs are constructed by pairing molecules that share the same core scaffold but differ by a local substituent transformation. Scaffold pairs are molecule pairs with different Bemis–Murcko scaffolds.

For a pair (i, j), the latent distance is computed as

$$d_\Phi(i,j) = \left\| \mathrm{vec}(\mathbf{Z}_i) - \mathrm{vec}(\mathbf{Z}_j) \right\|_2, \tag{30}$$

where padded tokens are masked consistently across models before vectorization. For each representation and each pair type, we report the mean latent distance over the corresponding pair set. The expected qualitative ordering is

$$d_{\mathrm{rot}} < d_{\mathrm{conf}} < d_{\mathrm{MMP}} < d_{\mathrm{scaf}}. \tag{31}$$

This diagnostic does not by itself define chemical correctness; rather, it tests whether nuisance variation from global pose is separated from increasingly semantic molecular variation.

F.3 Latent perturbation decodability

Latent perturbation decodability evaluates whether local neighborhoods of the latent space remain decodable into physically plausible 3D molecules. For each molecule m, we perturb its frozen latent representation by isotropic Gaussian noise:

$$\mathbf{Z}_{m,\sigma} = \mathbf{Z}_m + \sigma\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{32}$$

where noise is applied only to valid latent tokens and padded tokens remain masked. We evaluate

$$\sigma \in \{0, 0.05, 0.10, 0.20, 0.40, 0.80, 1.00\}.$$

Each perturbed latent is decoded by the frozen decoder,

$$\widehat{M}_{m,\sigma} = \zeta_{\mathrm{Dec}}(\mathbf{Z}_{m,\sigma}), \tag{33}$$

and the resulting molecule is evaluated using the same PB-Valid criterion used in the GEOM-DRUG v2 evaluation protocol. For each perturbation scale, we compute

$$\mathrm{PBValid}(\sigma) = \frac{\#\{\widehat{M}_{m,\sigma} \text{ passes PB-Valid}\}}{\#\{\widehat{M}_{m,\sigma}\}}. \tag{34}$$

Higher PB-Valid values at larger perturbation scales indicate a wider decoder-valid latent neighborhood.

F.4 Gauge-waste ratio and search efficiency

The optimization-readiness diagnostic measures whether latent-space search directions correspond to chemically meaningful improvement or are wasted along rotation-induced latent directions. For each representation, we analyse optimization trajectories generated by the same fixed-size latent search protocol used in the structure-based molecular optimization experiments.

Rotation-induced latent subspace. For a molecule m, we estimate a local rotation-induced latent subspace by applying $K = 100$ random rotations and encoding the rotated copies:

$$\mathscr{Z}_{\mathrm{rot}}(m) = \left\{ \mathrm{vec}\left[\Phi(\mathbf{x}_m R_k^\top, \mathbf{h}_m) \right] \right\}_{k=1}^{K}.$$

We apply PCA to this latent cloud and take the top three principal components as the rotation-induced subspace,

$$U_{\mathrm{rot}}(m) = \mathrm{span}\{u_1, u_2, u_3\},$$

reflecting the three-dimensional degree of freedom of $\mathrm{SO}(3)$. Let $P_{U_{\mathrm{rot}}}$ denote the orthogonal projection onto this subspace.

Gauge-waste ratio. For each optimization run, we track the CMA-ES population mean in latent space, $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_T$, and define the step displacement as $\Delta\mathbf{z}_t = \mathbf{z}_{t+1} - \mathbf{z}_t$.

The per-step gauge-waste ratio is

$$W_t = \frac{\|P_{U_{\mathrm{rot}}} \Delta\mathbf{z}_t\|_2^2}{\|\Delta\mathbf{z}_t\|_2^2 + \varepsilon}, \qquad \varepsilon = 10^{-8}. \tag{35}$$

The trajectory-level gauge-waste ratio is the average over optimization steps:

$$\overline{W} = \frac{1}{T} \sum_{t=0}^{T-1} W_t. \tag{36}$$

Lower values indicate that the optimization trajectory spends less displacement along rotation-induced latent directions.

Search efficiency. Let $F(M;P)$ be the scalar optimization fitness for molecule M against pocket P, combining docking affinity, synthetic accessibility and invalid-molecule penalties as in the structure-based optimization protocol. Because docking scores are lower for stronger predicted binding, we use the sign convention

$$F(M;P) = -E_{\text{dock}}(M;P) + \lambda_{\text{SA}} S_{\text{SA}}(M) - \lambda_{\text{inv}} \mathbf{1}\{M \text{ invalid}\}, \tag{37}$$

with fixed coefficients across all compared representations. For an optimization trajectory, we define

$$\Delta F = F(M_T;P) - F(M_0;P), \qquad L_{\text{path}} = \sum_{t=0}^{T-1} \|\Delta \mathbf{z}_t\|_2, \tag{38}$$

and compute search efficiency as

$$\text{Eff} = \frac{\Delta F}{L_{\text{path}} + \varepsilon}. \tag{39}$$

Higher values indicate larger objective improvement per unit latent-space displacement.

In the gauge-waste/search-efficiency plot, the x-axis reports $\overline{W}$, the y-axis reports Eff, point size denotes the mean pairwise Tanimoto diversity among valid generated molecules, and colour denotes the valid generation rate. The desired region is low gauge waste, high search efficiency, high validity and high diversity.

F.5 Interpolation decodability

Interpolation decodability evaluates whether linear paths between molecular latents remain inside a decoder-valid region of the latent space. We sample molecule pairs with matched atom counts when possible and use their frozen latent tensors $\mathbf{Z}_i$ and $\mathbf{Z}_j$. For interpolation coefficient $t \in \{0, 0.1, \ldots, 1.0\}$, we construct

$$\mathbf{Z}_{ij}(t) = (1-t)\mathbf{Z}_i + t\mathbf{Z}_j. \tag{40}$$

Each interpolated latent is decoded with the frozen decoder, $\widehat{M}_{ij}(t) = \zeta_{\text{Dec}}(\mathbf{Z}_{ij}(t))$, and evaluated using PB-Valid. For each interpolation position, we compute

$$\text{PBValid}(t) = \frac{\#\{\widehat{M}_{ij}(t) \text{ passes PB-Valid}\}}{\#\{\widehat{M}_{ij}(t)\}}. \tag{41}$$

A flat curve with high PB-Valid across t indicates that the line segment between two molecular latents remains largely decodable.

F.6 Aggregation and reporting

For rotation sensitivity, we report molecule-level distributions. For latent-distance hierarchy, we report mean distances for each representation and pair type on a logarithmic scale. For latent perturbation and interpolation decodability, we report PB-Valid rates averaged over decoded samples at each perturbation scale or interpolation point. For optimization-readiness, we first compute gauge-waste ratio and search efficiency for each optimization trajectory and then average over runs. All comparisons use frozen representations and identical evaluation pipelines across the compared representation regimes.

G Experimental Details for Figure 1.b

Left Panel: Embedding Consistency Under Rotation. We selected a single aspirin molecule and generated 1,000 conformations by applying random rotations sampled uniformly from SO(3). Each rotated conformation was passed through the pretrained encoder to obtain molecular embeddings. The red points represent embeddings obtained after applying the learned canonical mapping A_θ, while the gray points represent embeddings obtained without canonical mapping. All embeddings were projected to 2D using PCA for visualization.

Right Panel: Pretraining Convergence Comparison. Using the same experimental setup, we compared the validation coordinate reconstruction loss during pretraining with and without the canonical mapping. Both models were trained under identical hyperparameters, differing only in whether the canonicalization operator $\widehat{R}_\theta$ was applied to input coordinates.

H Implementation Details

We provide comprehensive hyperparameter settings for pretraining and all downstream tasks. All experiments were conducted on 2 NVIDIA L40S GPUs.

Table 2. Training hyperparameters for autoencoder pretraining.

Hyperparameter	Value
Optimizer	Adam
Learning rate	1×10^{-3}
Adam (β_1, β_2)	$(0.9, 0.99)$
Weight decay	1×10^{-4}
LR scheduler	Polynomial decay
Warmup steps	10,000
Total steps	2,000,000
Batch size	512
Gradient clipping	1.0

Table 3. Loss weights for autoencoder pretraining.

Loss Term	Weight
Coordinate reconstruction ($\mathscr{L}_{\text{Coord}}$)	10.0
Atom type prediction ($\mathscr{L}_{\text{Type}}$)	0.5
Bond type prediction ($\mathscr{L}_{\text{Bond}}$)	1.0
Distance prediction ($\mathscr{L}_{\text{Dist}}$)	10.0
KL divergence ($\mathscr{L}_{\text{KLD}}$)	0.01
Gauge consistency ($\mathscr{L}_{\text{gc}}$)	1.0

Table 4. Architecture hyperparameters for pretraining.

Component	Hyperparameter	Value
Canonicalization Encoder	Layers	3
	Hidden dimension	128
	Attention heads	8
	FFN dimension	128
Main Encoder	Layers	10
	Hidden dimension	128
	Attention heads	8
	FFN dimension	128
Decoder	Layers	3
	Hidden dimension	128
	Attention heads	8
	FFN dimension	128
VAE Latent	Latent dimension	64
	Log variance clipping	$[-15, 15]$

Table 5. DiT architecture for flow matching.

Hyperparameter	Value
Layers	12
Hidden dimension	512
Attention heads	8
MLP ratio	4.0
Input dimension	8

Table 6. Training hyperparameters for flow matching.

Hyperparameter	Value
Optimizer	Adam
Learning rate	1×10^{-4}
Adam (β_1, β_2)	$(0.9, 0.99)$
Weight decay	1×10^{-4}
LR scheduler	Polynomial decay
Warmup steps	10,000
Total steps	2,000,000
Batch size	512
Gradient clipping	1.0
Self-conditioning probability	0.5

H.1 Pretraining

Table 4 summarizes the architectural configurations of THEMOL. And, Table 2 presents the training hyperparameters for autoencoder pretraining. The pretraining objective combines multiple reconstruction terms, as shown Table 3

H.1.1 Input Perturbation

During pretraining, we apply the following augmentations: **Random rotation**: Applied with probability 0.7 to enforce rotation robustness of the canonicalization operator. **Coordinate noise**: Uniform noise from $\mathscr{U}(-0.1, 0.1)$ Å added to atomic coordinates. **Token masking**: 15% of atoms are masked for the masked atom prediction objective. All hydrogen atoms are removed during preprocessing, and molecules are truncated to a maximum of 256 atoms.

H.2 Generation (Flow Matching)

Table 5 summarizes the architectural configurations of THEMOL. And, Table 6 presents the training hyperparameters for generative setting.

At inference, we integrate the learned flow using the Euler method with 100 timesteps. Self-conditioning is enabled during sampling, where the previous prediction is used as an additional input to the network.

H.3 Property Prediction

For property prediction tasks on MoleculeNet and TDC ADMET benchmarks, we finetune the pretrained encoder with a classification or regression head. We perform grid search, as shown: 7

Table 7. Grid search space for property prediction.

Hyperparameter	Search Space
Learning rate	$\{5 \times 10^{-5}, 8 \times 10^{-5}, 1 \times 10^{-4}\}$
Batch size	$\{32, 64\}$
Dropout	$\{0.0, 0.1, 0.2, 0.5\}$
Warmup ratio	$\{0.0, 0.06, 0.1\}$

Table 8. Training configuration for property prediction.

Hyperparameter	MoleculeNet	TDC ADMET
Optimizer	Adam	Adam
Adam (β_1, β_2)	$(0.9, 0.99)$	$(0.9, 0.99)$
LR scheduler	Polynomial decay	Polynomial decay
Max epochs	200	200
Early stopping patience	40	100
Gradient clipping	1.0	1.0
Conformers (train)	1	1

For classification tasks, we use cross-entropy loss for binary classification and binary cross-entropy for multi-label classification. For regression tasks, we use smooth L1 loss with target normalization. The best checkpoint is selected based on the validation metric: ROC-AUC for most classification tasks, AUPRC for CYP datasets, MAE for some regression tasks, and Spearman correlation for clearance and half-life prediction.

H.4 Structure-Based Optimization

For structure-based molecular optimization, we optimize the latent space using LRA-CMA-ES while keeping the flow model frozen.

H.4.1 CMA-ES Configuration

Table 9. CMA-ES hyperparameters for molecular optimization.

Hyperparameter	Value
Population size (λ)	50
Number of generations	15
Initial σ	1.0
Learning rate adaptation	Enabled

H.4.2 Scoring Function

The fitness function combines binding affinity and synthetic accessibility:

$$S(M;T) = E_{\text{dock}}(M;T) + \lambda_{\text{SA}} \cdot S_{\text{SA}}(M), \tag{42}$$

where E_{dock} is the Vina docking score (kcal/mol), S_{SA} is the normalized synthetic accessibility score scaled to $[-10, 0]$, and $\lambda_{\text{SA}} = 1.0$.

H.4.3 Docking Configuration

Table 10. UniDock configuration for molecular docking.

Parameter	Value
Scoring function	Vina
Search mode (optimization)	Fast
Search mode (evaluation)	Balance
Box size	$25 \times 25 \times 25$ Å
Number of poses	1

H.4.4 Final Sampling

After optimization, we sample 100 molecules from the optimized distribution $\mathcal{N}(\mathbf{m}^{(T)}, 0.1^2\mathbf{I})$, where $\mathbf{m}^{(T)}$ is the final mean from CMA-ES.

H.5 Coordinate Reconstruction Loss.

A key design choice in our framework is that the canonicalization operator $\hat{R}_\theta$ is learned *jointly* with the autoencoder, rather than being fixed a priori (e.g., via PCA or inertia-based alignment). This joint optimization has a crucial implication: the canonical frame and the representation co-adapt to each other.

Consider the alternative of using a fixed canonicalization. The backbone would be forced to learn representations compatible with an externally imposed frame, which may not align with the model's inductive biases. In contrast, our approach allows the canonicalization to discover orientations that the backbone can most effectively encode, while the backbone simultaneously adapts to the structure of canonicalized inputs.

The joint optimization couples canonicalization and representation learning into a single objective. While we do not claim that this yields a globally optimal frame among all consistent alternatives, we argue that it avoids a key failure mode: *mismatch between a fixed frame and the model's inductive bias*. By allowing mutual adaptation, the learned frame is at least a local optimum that the backbone can effectively exploit. The strong downstream performance across property prediction, generation, and optimization tasks suggests this local optimum is practically sufficient for general-purpose molecular modeling.

I Additional Experiments (Ablation Study)

Before presenting the ablation results, we provide implementation details for each representation paradigm evaluated in this study.

Repr.A (Invariant). This variant adopts the same pairwise attention network architecture as Uni-Mol. We construct an autoencoder by combining two such sub-networks as the encoder and decoder modules, respectively. The pairwise distance-based attention mechanism ensures SE(3)-invariant representations throughout the network.

Repr.B (Equivariant). This variant takes 3D atomic coordinates as input and processes them through stacked blocks, each consisting of SE(3)-equivariant layers and pairwise attention modules. The encoder-decoder architecture produces a composite latent representation $[\mathbf{z}_\mathbf{x}, \mathbf{z}_\mathbf{h}]$, where $\mathbf{z}_\mathbf{x} \in \mathbb{R}^{N\times 3}$ denotes the equivariant coordinate features from the final SE(3)-equivariant layer, and $\mathbf{z}_\mathbf{h} \in \mathbb{R}^{N\times H}$ denotes the invariant node features from the final pairwise attention layer.

Repr.C (Augmentation). This variant is identical to Repr.**Z** (our proposed method) except that the learned canonicalization operator is removed. Instead, random rotation augmentation is applied to input molecules during training to encourage rotation robustness.

Repr.Z (Canonicalization, Ours). Our proposed method first applies a learned canonicalization operator to align input molecules into a canonical pose, then processes them through a non-equivariant transformer backbone. This design enables both invariant and equivariant downstream tasks within a unified framework.

Table 11. 3D molecule generation results on GEOM-DRUG under two evaluation protocols. (Left) v1 setting evaluates Atom Stability and molecular Validity. (Right) v2 setting employs stricter metrics: PB-Valid, OOD Rings, and geometry relaxation measures.

	GEOM-DRUG v1 (↑)		**GEOM-DRUG v2** (↓)			
Method	Atom Sta (%)	Valid (%)	PB-Valid (%)	OOD Rings (%)	Med. ΔE_{relax}	Med. ΔR_{relax}
Data	86.5	99.9	93.2±0.01	0.05	0.00	0.00
THEMOL (Repr.**B**)	72.3±0.31	89.4±0.6	78.6±1.42	0.47±0.05	12.84±0.09	0.52±0.06
THEMOL (Repr.**C**)	78.6±0.24	94.2±0.4	84.1±1.18	0.35±0.03	9.21±0.06	0.38±0.04
THEMOL (Repr.**Z**)	**86.8**±0.11	**99.9**±0.1	**93.7**±0.93	**0.18**±0.01	**4.97**±0.02	**0.17**±0.03

Table 12. Performance of different type of THEMOL family on the TDC Benchmark (15).

Methods	TDC Dataset			THEMOL	THEMOL	THEMOL	THEMOL
	Dataset	Metric	Size	Repr.**A**	Repr.**B**	Repr.**C**	Repr.**Z**
Absorption	Caco2	MAE(↓)	906	0.324±0.014	0.487±0.032	0.361±0.018	0.291±0.009
	HIA	AUROC(↑)	578	0.974±0.008	0.847±0.024	0.948±0.011	0.996±0.003
	Pgp	AUROC(↑)	1212	0.917±0.007	0.782±0.019	0.891±0.009	0.939±0.004
	Bioavailability	AUROC(↑)	640	0.678±0.021	0.534±0.038	0.647±0.024	0.706±0.015
	Lipophilicity	MAE(↓)	4200	0.492±0.010	0.681±0.024	0.529±0.013	0.456±0.007
	Solubility	MAE(↓)	9982	0.769±0.013	1.042±0.028	0.815±0.016	0.722±0.009
Distribution	BBB	AUROC(↑)	1975	0.904±0.007	0.768±0.018	0.878±0.009	0.927±0.004
	PPBR	MAE(↓)	1797	7.842±0.105	10.234±0.187	8.193±0.121	7.461±0.076
	VDss	Spearman(↑)	1130	0.668±0.022	0.489±0.041	0.634±0.026	0.704±0.016
Metabolism	CYP2D6 inhibition	AUPRC(↑)	13130	0.674±0.007	0.521±0.016	0.647±0.009	0.703±0.004
	CYP3A4 inhibition	AUPRC(↑)	12328	0.852±0.006	0.694±0.015	0.824±0.008	0.880±0.003
	CYP2C9 inhibition	AUPRC(↑)	12092	0.766±0.008	0.598±0.018	0.736±0.010	0.798±0.005
	CYP2D6 substrate	AUPRC(↑)	664	0.738±0.026	0.561±0.045	0.704±0.030	0.775±0.019
	CYP3A4 substrate	AUROC(↑)	667	0.752±0.017	0.578±0.032	0.719±0.020	0.787±0.012
	CYP2C9 substrate	AUPRC(↑)	666	0.504±0.025	0.348±0.042	0.469±0.029	0.542±0.018
Excretion	Half life	Spearman(↑)	667	0.521±0.031	0.342±0.054	0.486±0.036	0.559±0.024
	Clearance mircosome	Spearman(↑)	1102	0.654±0.015	0.472±0.031	0.619±0.018	0.692±0.010
	Clearance hepatocyte	Spearman(↑)	1020	0.441±0.028	0.268±0.048	0.408±0.032	0.479±0.021
Toxicity	hERG	AUROC(↑)	648	0.828±0.017	0.654±0.034	0.796±0.020	0.863±0.012
	Ames	AUROC(↑)	7255	0.806±0.007	0.648±0.016	0.778±0.009	0.835±0.004
	DILI	AUROC(↑)	475	0.901±0.014	0.734±0.031	0.871±0.017	0.933±0.009
	LD50	MAE(↓)	7385	0.652±0.012	0.892±0.026	0.694±0.015	0.608±0.008

Table 13. Summary of binding affinity and molecular properties of reference molecules and molecules generated by THEMOL and baselines. (↑) / (↓) denotes whether a larger / smaller number is preferred.

Methods	SMINA (↓)		Vina Dock (↓)		SA (↑)		Diversity (↑)	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Reference	-6.37	-7.92	-6.36	-7.45	0.73	0.74	-	-
THEMOL (Repr.**B**)	-7.21	-6.94	-7.48	-7.21	0.61	0.62	0.58	0.57
THEMOL (Repr.**C**)	-8.27	-8.04	-8.71	-8.43	0.71	0.72	0.76	0.75
THEMOL (Repr.**Z**)	-8.63	-8.36	-9.12	-8.80	0.75	0.77	0.85	0.84

I.1 In-Depth Analysis of Representation Paradigms

I.1.1 Quantifying the Canonicalization Advantage

To understand *why* canonicalization consistently outperforms alternatives, we analyze the learned representations through the lens of our theoretical framework.

Consistency Error Analysis. We empirically measure the canonicalization consistency error δ from Assumption 4:

$$\delta = \mathbb{E}_{(\mathbf{x},\mathbf{h})\sim\mathscr{D}}\left[\sup_{R\in SO(3)} \|A_\theta(R\mathbf{x},\mathbf{h}) - A_\theta(\mathbf{x},\mathbf{h})\|_F\right] \tag{43}$$

Table 15 reports δ measured on held-out molecules using 100 random rotations per molecule:

The near-zero consistency errors validate that our learned canonicalization satisfies Assumption 4 and Proposition 1 of the main text in practice, with over 98% of molecules achieving effectively exact consistency ($\delta < 10^{-3}$).

I.1.2 Failure Case Analysis

To provide a balanced assessment, we analyze cases where Repr.**Z** shows smaller advantages or potential limitations.

Large Flexible Molecules. For molecules with many rotatable bonds, conformational flexibility can challenge canonicalization:

Table 14. The overall results on 9 molecule classification datasets from the MoleculeNet (14). We report ROC-AUC score (higher is better) under scaffold splitting.

Datasets	BACE (↑)	BBBP (↑)	Tox21 (↑)	SIDER (↑)	HIV (↑)	MUV (↑)	PCBA (↑)	ClinTox (↑)	ToxCast (↑)	Mean (↑)
THEMOL (Repr.**A**)	87.4±0.008	88.7±0.007	79.8±0.005	62.8±0.009	83.2±0.006	76.4±0.011	82.1±0.004	96.8±0.006	68.2±0.007	79.49
THEMOL (Repr.**B**)	78.6±0.018	79.4±0.016	72.1±0.012	56.2±0.019	74.8±0.014	67.3±0.022	74.5±0.011	87.2±0.015	61.4±0.014	72.39
THEMOL (Repr.**C**)	84.2±0.012	85.6±0.010	77.4±0.008	60.1±0.013	80.7±0.009	73.8±0.015	79.6±0.007	93.4±0.010	65.9±0.011	77.86
THEMOL (Repr.**Z**)	89.9±0.006	91.2±0.005	81.3±0.004	64.3±0.007	85.0±0.005	78.1±0.010	83.7±0.003	98.9±0.004	69.7±0.005	81.65

Table 15. Canonicalization consistency error across molecular datasets.

Dataset	δ (mean)	δ (99th pctl)	% Perfect ($\delta < 10^{-3}$)
GEOM-DRUG	2.3×10^{-3}	8.7×10^{-3}	99.2%
CrossDocked	3.1×10^{-3}	1.2×10^{-2}	98.7%
MoleculeNet	1.8×10^{-3}	6.4×10^{-3}	99.5%

Repr.**Z** maintains consistent improvements across flexibility ranges, though absolute performance decreases for all methods on highly flexible molecules—a known challenge in 3D molecular modeling independent of representation choice.

I.1.3 Computational Overhead Analysis

A practical concern is whether canonicalization introduces significant computational cost. Table 17 reports timing benchmarks. Notably, Repr.**Z** achieves only 11% overhead compared to the lightweight Repr.**C** baseline, while providing substantially better performance across all downstream tasks. In contrast, Repr.**A** incurs 54% overhead because its autoencoder architecture demands encoder and decoder modules each with UniMol-scale capacity; without sufficient depth in both components, training fails to converge properly.

J Datasets

J.1 Pretraing Dataset.

For pretraining, we utilize the dataset constructed by (17). This dataset comprises 19 million molecules, curated by merging a synthesizable database with collections from (18). Consistent with the settings in (17), it contains a total of 209 million conformations, where 11 conformations (including one flattened structure) were generated for each molecule using RDKit (9).

J.2 MoleculeNet

We conduct extensive experiments on the MoleculeNet (14) benchmark to evaluate the performance of the proposed model on molecular property prediction tasks and compare it with existing methodologies. These datasets encompass a diverse range of molecular properties, including biophysics, and physiology, ensuring a comprehensive evaluation of model generalization. The specifications of the primary datasets used in this study are summarized in Table 18.

J.3 TDC

To evaluate our model on ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) prediction tasks, we employed the Therapeutics Data Commons (TDC) (19) benchmark. TDC serves as a standardized ecosystem designed for machine learning in drug discovery and development. A detailed summary of the datasets utilized in our experiments is presented in Table 19.

J.4 CrossDocked2020

We utilized the CrossDocked2020 dataset (20) to assess the generation capability of our model. Constructed from Pocketome PDB structures, this dataset expanded the pose space via Smina cross-docking across 2,922 pockets. Notably, to ensure robustness against false positives and negatives, approximately 11.9 million counter-examples were incorporated. Consequently, the final dataset comprises 22,584,102 poses involving 13,780 ligands and 2,922 pockets.

Following the data processing protocol established by (21), we further refined the dataset to ensure high data quality. We filtered out data points with a binding pose RMSD greater than 1 Å, yielding a refined subset of 184,057 poses. To prevent data leakage, we utilized mmseqs2 to cluster the data based on a 30% sequence identity threshold. From these clusters, we randomly sampled 100,000 protein-ligand pairs for the training set and selected 100 proteins from the remaining non-overlapping clusters to serve as the test set.

Table 16. Property prediction (ROC-AUC) stratified by number of rotatable bonds in MoleculeNet Database.

Rotatable Bonds	Repr.**A**	Repr.**C**	Repr.**Z**	Δ (Z vs. A)
0–3	86.2	84.7	88.4	+2.2
4–7	79.8	78.1	82.3	+2.5
8+	74.3	72.6	76.1	+1.8

Table 17. Computational cost comparison. Time denotes the average cost per batch (batch size = 64, averaged over 1,000 iterations). Ratio denotes the slowdown ratio compared with Repr.**C** (augmentation baseline).

Method	Params (M)	Time (ms)			Ratio
		Input Proc.	Backbone	Total	
Uni-Mol	47.61	21.0	53.2	74.2	1.38
THEMOL (Repr.**A**)	52.46	24.8	57.8	82.6	1.54
THEMOL (Repr.**B**)	48.72	16.8	51.6	68.4	1.27
THEMOL (Repr.**C**)	42.15	15.2	38.5	53.7	1.00
THEMOL (Repr.**Z**)	44.38	22.3	37.5	59.8	1.11

J.5 GEOM-DRUG

To evaluate the performance of our proposed 3D molecular generation model, we utilized the GEOM-DRUG dataset (22). We excluded the QM9 dataset (23) from this study, as its limited scale and molecular simplicity are considered insufficient for the rigorous demands of our drug discovery. This dataset contains diverse 3D conformers of drug-like molecules along with their corresponding energy information, serving as a widely adopted benchmark for validating how precisely a generative model mimics molecular geometry and whether the generated structures are thermodynamically plausible. To strictly verify the robustness of our model, we adopted two distinct preprocessing and experimental configurations established in prior studies.

First, we adhered to the experimental setting defined in EDM (24). Under this configuration, we constructed the training data using approximately 430,000 molecules, selecting only the 30 conformers with the lowest energy for each molecule. This selection strategy aims to induce the model to learn 3D coordinates close to the physically stable ground state. To assess the physical validity of the generated molecules under this setting, we employed the Atom Stability metric and the Wasserstein distance between energy distributions.

Second, we adopted the data construction strategy from MiDi (25) and FlowMol3 (26). We utilized conformers located at local minima on the GFN2-xTB potential surface, resulting in a dataset comprising 243,480 unique molecules and 5,741,535 conformers. Notably, this setting treats all atoms, including explicit hydrogens, as independent nodes and applies a specific Kekulization process to distinguish aromatic bonds as single or double bonds, thereby enabling the model to explicitly learn chemical topology. Consistent with the data setup, we applied the evaluation protocol established in MiDi to rigorously verify the chemical validity of the complete molecular structures, including hydrogens.

The necessity of this dual experimental setup highlights the current absence of a consensus standard within the field of 3D molecular generation. The fragmentation of preprocessing methods and evaluation metrics not only hinders objective performance comparison between models but also introduces ambiguity in verifying utility for real-world drug design. Therefore,

Table 18. Summary information of the MoleculeNet benchmark datasets.

Dataset	Tasks	Task type	Molecules (train/test)	Description
BACE	1	Classification	1210/302	Binding results of human BACE-1 inhibitors
BBBP	1	Classification	1631/408	Blood-brain barrier penetration
ClinTox	2	Multi-label classification	1182/296	Clinical trial toxicity and FDA approval status
Tox21	12	Multi-label classification	6264/1566	Qualitative toxicity measurements
ToxCast	617	Multi-label classification	6860/1716	Toxicology data based on in vitro screening
SIDER	27	Multi-label classification	1141/286	Adverse drug reactions to the 27 systemic organs
HIV	1	Classification	32901/8226	The ability to suppress HIV replication
MUV	17	Multi-label classification	74469/18618	A subset of PubChem BioAssay
PCBA	128	Multi-label classification	350343/87586	Bioactivities data generated by high-throughput screening

Table 19. Summary information of the TDC benchmark datasets.

Dataset	Task type	Molecules (train/test)	Description
Caco2	Regression	728/182	Permeability values using the Caco-2 cell line
HIA	Classification	461/117	Human Intestinal Absorption (HIA) efficiency data
Pgp	Classification	973/245	P-glycoprotein (Pgp) inhibition status
Bioavailability	Classification	512/128	Oral Bioavailability fraction reaching systemic circulation
Lipophilicity	Regression	3360/840	Lipophilicity indicating membrane permeability
Solubility	Regression	7985/1997	Aqueous Solubility in physiological conditions
BBB	Classification	1624/406	Blood-brain barrier penetration
PPBR	Regression	2231/559	Human Plasma Protein Binding Rate
VDss	Regression	904/226	Volume of Distribution at Steady State
CYP2D6 inhibition	Classification	10504/2626	Inhibition of CYP2D6, responsible for metabolizing 25% of clinical drugs
CYP3A4 inhibition	Classification	9861/2467	Inhibition of CYP3A4, involved in metabolizing 50% of prescribed drugs
CYP2C9 inhibition	Classification	9673/2419	Inhibition of CYP2C9, an enzyme metabolizing 100 clinical drugs
CYP2D6 substrate	Classification	532/135	Substrate specificity data for the CYP2D6 enzyme
CYP3A4 substrate	Classification	535/135	Substrate specificity data for the CYP3A4 enzyme
CYP2C9 substrate	Classification	534/135	Substrate specificity data for the CYP2C9 enzyme
Half Life	Regression	532/135	Duration required for plasma drug concentration to decrease by 50%
Clearance microsome	Regression	881/221	Drug clearance rates in Microsome environments
Clearance hepatocyte	Regression	970/243	Drug clearance rates in Hepatocyte environments
hERG	Classification	523/132	Human ether-à-go-go related gene blockage indicating cardiotoxicity risk
Ames	Classification	2821/1457	Ames mutagenicity test results indicating potential DNA damage
DILI	Classification	379/96	Bioactivities data generated by high-throughput screening
LD50	Regression	5907/1478	Acute toxicity dosage causing death in 50% of test subjects

establishing a unified dataset standard that ensures both physicochemical realism and experimental consistency is essential for the robust advancement of future research and fair benchmarking.

J.6 Additional Benchmark

We focus on drug discovery applications (property prediction, generation, optimization) as our primary use case. Evaluation on quantum mechanical benchmarks (MD17 (27), MatBench (28), QMOF (29)), which are not directly related to drug discovery and target conformer-level physical quantities rather than molecule-level biological properties, is beyond the scope of this work but represents a valuable direction for future investigation.

J.7 Evaluation Metrics

We describe the evaluation metrics used for each benchmark in our experiments.

J.7.1 Molecular Property Prediction

MoleculeNet. We evaluate classification performance using **ROC-AUC** (Receiver Operating Characteristic Area Under the Curve), which measures the model's ability to distinguish between positive and negative classes across all classification thresholds. Higher values indicate better discriminative performance, with 1.0 representing perfect classification and 0.5 representing random guessing.

TDC ADMET Benchmark. The Therapeutics Data Commons benchmark employs task-specific metrics appropriate for each endpoint: **AUROC**, **AUPRC**, **MAE**, **Spearman Correlation**

J.7.2 3D Molecule Generation (GEOM-DRUG)

We adopt two evaluation protocols following prior work.

GEOM-DRUG v1 (EDM setting (24)):

- **Atom Stability (%)**: Percentage of atoms with correct valency in generated molecules. Higher is better.
- **Validity (%)**: Percentage of generated molecules that pass RDKit sanitization checks. Higher is better.

GEOM-DRUG v2 (MiDi setting (25), metrics from "GEOM-DRUG Revisited" paper (26; 30)):

- **PB-Valid (%)**: PoseBusters Validity (31). Percentage of molecules passing comprehensive physical validity checks including bond lengths, bond angles, planar aromatic rings, planar double bonds, internal steric clashes, and internal energy tests. Higher is better.
- **OOD Rings (%)**: Out-of-Distribution Ring Rate. Percentage of generated ring systems that are never observed in ChEMBL, a database of 2.4M bioactive compounds. Lower values indicate better adherence to chemically realistic ring topologies.
- **Med.** $\Delta E_{\mathbf{relax}}$: Median relaxation energy (kcal/mol). Energy change after geometry optimization with the GFN2-xTB semi-empirical potential. Lower values indicate that generated conformations are already close to local energy minima.
- **Med.** $\Delta R_{\mathbf{relax}}$: Median relaxation RMSD (Å). Root-mean-square deviation of atomic coordinates before and after energy minimization. Lower values indicate that generated 3D geometries require minimal adjustment to reach physically stable configurations.

J.7.3 3D Molecule Optimization (CrossDocked2020)

- **SMINA / Vina Dock**: Docking scores computed by SMINA and AutoDock Vina, respectively. These scores estimate binding affinity between the generated ligand and target protein pocket, measured in kcal/mol. Lower (more negative) values indicate stronger predicted binding affinity.
- **SA**: Synthetic Accessibility score (32), ranging from 1 (easy to synthesize) to 10 (difficult to synthesize). We report the normalized score where higher values indicate greater synthetic feasibility.
- **Diversity**: Molecular diversity among generated compounds, computed as the average pairwise Tanimoto distance based on Morgan fingerprints. Higher values indicate greater structural diversity in the generated molecule set, which is desirable for exploring broader chemical space.

J.8 Baseline Models

We provide descriptions of baseline models used in our experiments across different benchmarks.

J.8.1 MoleculeNet Baselines

- **PretrainGNN** (33) is a pioneering work on self-supervised pretraining for graph neural networks in molecular property prediction, employing strategies such as context prediction and attribute masking to learn transferable molecular representations.
- **GROVER** (34) is a self-supervised molecular representation learning framework based on Graph Transformers, which pretrains on large-scale unlabeled molecular data using motif-level and graph-level prediction tasks.
- **MolCLR** (35) is a contrastive learning framework for molecular representation that learns by maximizing agreement between differently augmented views of the same molecule through graph-level and node-level augmentations.
- **MoleBLEND** (36) is a multimodal molecular pretraining framework that learns unified representations by blending information across multiple molecular modalities including 2D graphs, 3D conformations, and molecular fingerprints.
- **Uni-Mol** (17) is an SE(3)-invariant transformer pretrained on 209 million molecular conformations, learning 3D molecular representations through pairwise distance matrices and coordinate denoising objectives.
- **Mol-AE** (37) is an autoencoder-based molecular representation learning model that employs a 3D cloze test objective, where the model learns to reconstruct masked 3D structural information to capture spatial molecular features.
- **UniCorn** (38) is a unified contrastive learning framework for multi-view molecular representation learning that aligns representations across different molecular views through contrastive objectives.
- **RadialFocus** (39) is a geometric graph transformer that introduces distance-modulated attention mechanisms, where attention weights are explicitly conditioned on interatomic distances to capture geometric relationships.

J.8.2 TDC ADMET Baselines

- **MiniMol** (40) is a parameter-efficient foundation model for molecular learning that achieves strong performance across diverse ADMET tasks while maintaining a compact model size suitable for resource-constrained settings.
- **MapLight + GNN** (41) combines molecular fingerprints with graph neural networks to predict ADMET properties, leveraging complementary information from handcrafted descriptors and learned representations.
- **ContextPred** (42) learns molecular representations through context-level language modeling by predicting contextual embeddings, transferring the masked prediction paradigm from NLP to molecular graphs.
- **CFA** (43) enhances ADMET property prediction through combinatorial fusion analysis, systematically combining predictions from multiple base models to improve overall predictive performance.
- **ZairaChem** (44) is a fully-automated AI/ML virtual screening platform that integrates multiple machine learning methods into a unified cascade, designed for practical deployment in drug discovery settings.
- **Gradient Boost** and **BaseBoosting** are gradient boosting ensemble methods applied to molecular property prediction, combining multiple weak learners to achieve robust predictions on ADMET endpoints.

J.8.3 GEOM-DRUG Generation Baselines

- **GeoLDM** (45) is an SE(3)-equivariant latent diffusion model that performs 3D molecule generation in a learned latent space, separating the learning of data structure from the generative modeling process.
- **GeoBFN** (46) applies Bayesian flow networks to unified generative modeling of 3D molecules, providing a probabilistic framework that jointly models discrete atom types and continuous 3D coordinates.
- **EquiFM** (47) combines equivariant flow matching with hybrid probability transport for 3D molecule generation, designing SE(3)-equivariant vector fields that respect molecular symmetries during generation.
- **GOAT** (48) accelerates 3D molecule generation through jointly geometric optimal transport, coupling atoms across source and target distributions to reduce variance and enable faster sampling.
- **EQGAT-Diff** (49) systematically explores the design space of equivariant diffusion-based generative models for de novo 3D molecule generation, investigating architectural choices including attention mechanisms and equivariant message passing.
- **Megalodon** (50) applies modular co-design principles to de novo 3D molecule generation, decomposing the generation process into modular components that can be independently optimized.
- **SemlaFlow** (51) combines latent attention mechanisms with equivariant flow matching for efficient 3D molecular generation, operating in a compressed latent space while maintaining SE(3)-equivariance.
- **FlowMol3** (26) is a flow matching approach for 3D de novo small-molecule generation that directly models the probability path between noise and molecular structures in Cartesian coordinates.

J.8.4 CrossDocked2020 Optimization Baselines

- **TargetDiff** (52) is a 3D equivariant diffusion model for target-aware molecule generation that conditions the diffusion process on protein pocket structures to generate ligands with high binding affinity.
- **DecompDiff** (53) introduces decomposed priors for structure-based drug design, factorizing the molecular generation process into arms and scaffold components for more controllable ligand generation.
- **MolCRAFT** (54) performs structure-based drug design in continuous parameter space, optimizing molecular structures through gradient-based methods in a learned continuous latent representation.
- **DecompOpt** (55) combines controllable and decomposed diffusion models for structure-based molecular optimization, enabling targeted modifications of specific molecular substructures while preserving desired properties.
- **TacoGFN** (56) is a target-conditioned GFlowNet for structure-based drug design that learns to sample diverse molecules proportionally to a reward function combining binding affinity and drug-likeness.
- **ALIDiff** (57) aligns target-aware molecule diffusion models with exact energy optimization, incorporating physics-based binding energy calculations to guide the diffusion process toward high-affinity ligands.

J.9 MoleculeNet Analysis

To validate the practical utility of the proposed model in real-world drug discovery, we conducted a preliminary in-depth analysis of data integrity and validity for the MoleculeNet benchmark datasets. This analysis revealed that certain datasets exhibit discrepancies with actual drug discovery environments, possess data integrity issues, or suffer from incomplete problem definitions. Consequently, to ensure the reliability and practical validity of our evaluation, we excluded specific regression tasks—namely ESOL, FreeSolv, and the QM series (QM7, QM8, QM9)—from our experiments. The detailed analysis and rationale for the exclusion of each dataset are provided below.

J.9.1 ESOL

The ESOL dataset included in MoleculeNet represents aqueous solubility in log scale ($\log_{10}$ mol/L), spanning a remarkably broad dynamic range of approximately 13.18 log units, with values distributed from -11.60 to 1.58. This distribution encompasses compounds with extremely low or high solubility, a structural characteristic that allows even relatively simple regression models to easily achieve high correlation coefficients.

In contrast, the solubility of pharmaceutical compounds is evaluated within a much more restricted range during the actual drug development process. Experimental studies on pharmaceutical hydrates report that aqueous solubility typically falls between -3.19 and 1.96 on a log scale (58). Similarly, the standard coverage of the PASS assay, as proposed by the UNGAP consortium (0.001-100 mg/mL), corresponds to a log scale range of approximately -3 to 2 (59). This suggests that the solubility values encountered in real-world drug screening and optimization are concentrated within a relatively narrow range on the logarithmic scale.

Although achieving a high correlation between predicted and experimental values is intrinsically difficult within this realistic solubility range, the ESOL dataset entails a risk of overestimating model performance due to its excessively extended dynamic range. Consequently, prediction performance on the ESOL benchmark may not adequately reflect a model's generalization capability in a real-world drug discovery environment. Therefore, to ensure the realistic validity and interpretability of the evaluation, we did not conduct experiments on the ESOL regression task.

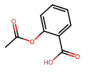
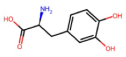
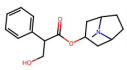
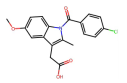
Molecule	Label	Dataset Input
	0	CC(=O)Oc1ccccc1C(O)=O
	1	C1=CC=CC(=C1C(O)=O)OC(C)=O
	0	[C@H](CC1=CC(=C(C=C1)O)O)(C(O)=O)N
	1	N[C@@H](Cc1ccc(O)c(O)c1)C(O)=O
	0	CN1C2CCC1CC(C2)OC(=O)C(CO)c3ccccc3
	1	C1=CC=CC=C1C(C(OC2CC3N(C)C(C2)CC3)=O)CO
	0	COc1ccc2n(c(C)c(CC(O)=O)c2c1)C(=O)c3ccc(Cl)cc3
	1	C1=CC(=CC2=C1[N](C(=C2CC(O)=O)C)C(C3=CC=C(Cl)C=C3)=O)OC

Figure 2. Examples of problem (Section J.9.4).

J.9.2 FreeSolv

The FreeSolv dataset was originally designed as a benchmark for validating how accurately Force Field parameters in Molecular Dynamics (MD) simulations reproduce hydration free energies.

However, from a drug discovery perspective, hydration free energy has limited capacity to independently represent drug efficacy or overall drug-likeness. Theoretically, it functions primarily as an intermediate component within the thermodynamic cycle used to derive binding free energy (60), and is rarely utilized as an independent optimization target in the actual lead optimization process. Accordingly, to verify the utility of the model in contributing to the actual drug discovery process, we excluded the FreeSolv task, which represents a purely thermodynamic parameter not directly linked to in vivo behavior or therapeutic efficacy.

J.9.3 QM7, QM8, QM9

The QM7, QM8, and QM9 datasets are benchmarks designed to predict quantum chemical properties from the 3D structure of molecules. The properties included in these datasets are derived from quantum mechanical calculations based on optimized 3D atomic coordinates and are intrinsically dependent on the geometric conformation of the molecule, such as bond lengths, bond angles, and spatial arrangement. Therefore, these tasks presuppose the use of 3D structural information as input.

Nevertheless, numerous preceding studies have reported prediction results for QM properties using only 1D representations, such as SMILES, without 3D coordinate information. While SMILES representations can capture molecular connectivity and certain chemical features, they have an inherent limitation in uniquely determining the quantum chemical properties corresponding to a specific structure among multiple possible 3D conformations. Due to this limitation, predicting QM properties in a SMILES-based setting may render the problem definition itself incomplete. Consequently, we excluded the QM7, QM8, and QM9 tasks from our experiments.

J.9.4 BBBP

Finally, regarding the binary classification dataset BBBP, we analyzed all compounds by converting them to canonical SMILES using RDKit to ensure data integrity. As a result, we identified 10 pairs of data errors where structurally identical molecules were assigned conflicting labels. This constitutes a logical contradiction where the same compound is labeled as both permeable and non-permeable, representing label noise that can negatively impact model training. The molecular structures of the 10 compound pairs where these errors were identified are shown in Figure 2.

Our analysis confirmed that these data conflicts were confined exclusively to the training set and did not affect the test set. Therefore, the test performance figures presented in this study are free from direct distortion caused by these data errors, and the reliability of the evaluation is considered valid. However, this analysis suggests that even widely accepted benchmark datasets may contain intrinsic defects, requiring caution in interpreting results. Furthermore, this underscores the necessity of conducting multifaceted supplementary experiments to prove model robustness, rather than relying solely on a specific benchmark for comprehensive performance verification.

K Flow Matching for Molecular Generation

In this section, we provide a detailed exposition of our flow-based generative model for molecular generation in the pretrained representation space. We adopt *optimal-transport conditional flow matching* (OT-CFM) (61) with a clean-target prediction parameterization inspired by FrameFlow (62).

K.1 Background: Conditional Flow Matching

Continuous Normalizing Flows. A continuous normalizing flow (CNF) (63) defines a generative model by learning a time-dependent vector field $v_\theta : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ that transports a simple prior distribution p_0 (e.g., Gaussian) to the data distribution p_1. The flow is characterized by the ordinary differential equation (ODE):

$$\frac{d\mathbf{x}_t}{dt} = v_\theta(t, \mathbf{x}_t), \qquad \mathbf{x}_0 \sim p_0, \tag{44}$$

where integrating from $t = 0$ to $t = 1$ transforms samples from p_0 to (approximately) p_1. The time-varying density p_t induced by the flow satisfies the continuity equation:

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v_\theta) = 0. \tag{45}$$

Flow Matching Objective. Given access to a target vector field $u_t(\mathbf{x})$ that generates the desired probability path p_t, we can train v_θ via simple regression:

$$\mathscr{L}_{\text{FM}}(\theta) = \mathbb{E}_{t \sim \mathscr{U}[0,1], \mathbf{x} \sim p_t} \left[\| v_\theta(t, \mathbf{x}) - u_t(\mathbf{x}) \|^2 \right]. \tag{46}$$

However, directly computing $u_t(\mathbf{x})$ and sampling from the marginal $p_t(\mathbf{x})$ is generally intractable. The key insight of *conditional flow matching* (CFM) is to decompose the problem using tractable conditional probability paths.

Conditional Flow Matching. CFM (61; 64) considers the marginal probability path as a mixture of simpler conditional paths:

$$p_t(\mathbf{x}) = \int p_t(\mathbf{x}|z)\, q(z)\, dz, \tag{47}$$

where z is a conditioning variable (e.g., a pair of source and target points) drawn from some distribution $q(z)$, and $p_t(\mathbf{x}|z)$ is a simple conditional probability path that we can sample from and whose generating vector field $u_t(\mathbf{x}|z)$ is known in closed form.

The marginal vector field that generates $p_t(\mathbf{x})$ can be expressed as:

$$u_t(\mathbf{x}) = \mathbb{E}_{q(z)} \left[\frac{u_t(\mathbf{x}|z)\, p_t(\mathbf{x}|z)}{p_t(\mathbf{x})} \right]. \tag{48}$$

The crucial result (61; 64) is that the *conditional flow matching* objective:

$$\mathscr{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, q(z), p_t(\mathbf{x}|z)} \left[\| v_\theta(t, \mathbf{x}) - u_t(\mathbf{x}|z) \|^2 \right] \tag{49}$$

has the same gradient as the intractable flow matching objective 46:

$$\nabla_\theta \mathscr{L}_{\text{FM}}(\theta) = \nabla_\theta \mathscr{L}_{\text{CFM}}(\theta). \tag{50}$$

This allows simulation-free training by sampling $(t, z, \mathbf{x})$ and regressing $v_\theta(t, \mathbf{x})$ to the conditional vector field $u_t(\mathbf{x}|z)$.

Gaussian Conditional Paths. A common choice is to let $z = (\mathbf{x}_0, \mathbf{x}_1)$ be a pair of source and target points, and define the conditional path as a Gaussian interpolation:

$$p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \mu_t(\mathbf{x}_0, \mathbf{x}_1), \sigma^2 I\right), \tag{51}$$

where $\mu_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ is the linear interpolation between the endpoints. The conditional flow $\mathbf{x}_t = \psi_t(\mathbf{x}_0|\mathbf{x}_1)$ satisfying $p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1)$ is:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1 + \sigma\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I), \tag{52}$$

and the conditional vector field generating this path is simply:

$$u_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0. \tag{53}$$

Note that the conditional velocity is constant in time and equals the displacement from source to target.

K.2 Flow Matching in Proposed Framework

Generation. We generate novel 3D molecules by learning a generative model directly on the pretrained representation space $\mathbf{X}_\mathbf{z} \in \mathbb{R}^{N\times H}$. After pre-training, we freeze the autoencoder and extract $\mathbf{X}_\mathbf{z}$ for each training molecule, which induces an empirical target distribution q_1 over $\mathbf{X}_\mathbf{z}$ (with padding masks for variable N). We then train a continuous-time flow model in this space using *optimal-transport conditional flow matching (OT-CFM)*, and decode generated representations back to molecular structures via the pretrained decoder heads.

Minibatch OT coupling. Let q_0 be a simple base distribution on $\mathbb{R}^{N\times H}$ (isotropic Gaussian on valid atoms; padded rows set to zero). Given a minibatch of target representations $\{\mathbf{X}_1^{(i)}\}_{i=1}^B \sim q_1$ and source samples $\{\mathbf{X}_0^{(i)}\}_{i=1}^B \sim q_0$, we compute a minibatch OT plan by solving

$$\Pi^\star \in \arg\min_{\Pi\in\mathscr{U}} \sum_{i=1}^{B}\sum_{j=1}^{B} \Pi_{ij} \left\|\mathbf{X}_0^{(i)} - \mathbf{X}_1^{(j)}\right\|_{F,\Omega}^2, \tag{54}$$

where $\mathscr{U}$ denotes the set of doubly-stochastic matrices (uniform marginals), Ω is the set of non-padding atom indices, and $\|\cdot\|_{F,\Omega}$ evaluates the Frobenius norm only on Ω. We then sample paired endpoints $(\mathbf{X}_0, \mathbf{X}_1)$ according to $\Pi^\star$. This OT coupling biases the endpoint pairing toward short displacements in representation space, which empirically stabilizes training and improves sample quality compared to independent pairing.

OT-CFM training with clean-target prediction. Given paired endpoints $(\mathbf{X}_0, \mathbf{X}_1)$, we define the conditional path by linear interpolation

$$\mathbf{X}_t := (1-t)\mathbf{X}_0 + t\mathbf{X}_1 \;+\; \sigma\boldsymbol{\varepsilon}, \qquad t \sim \mathrm{Unif}[0, 1-\varepsilon],\ \boldsymbol{\varepsilon} \sim \mathscr{N}(\mathbf{0}, \mathbf{I}), \tag{55}$$

where σ is an optional small noise level (we set $\sigma = 0$ unless stated otherwise) and $\varepsilon > 0$ avoids numerical singularities near $t \to 1$. Instead of directly regressing the conditional velocity, we train the network to predict the clean target $\mathbf{X}_1$:

$$\widehat{\mathbf{X}}_1 = f_\phi(t, \mathbf{X}_t). \tag{56}$$

We then reconstruct the time-dependent velocity field via the reparameterization

$$v_\phi(t, \mathbf{X}_t) \;:=\; \frac{\widehat{\mathbf{X}}_1 - \mathbf{X}_t}{1-t}. \tag{57}$$

Note that the ground-truth conditional velocity along (55) satisfies

$$\dot{\mathbf{X}}_t = \mathbf{X}_1 - \mathbf{X}_0 = \frac{\mathbf{X}_1 - \mathbf{X}_t}{1-t}. \tag{58}$$

This is consistent with the ground-truth velocity along the linear path, since $\dot{\mathbf{X}}_t = \mathbf{X}_1 - \mathbf{X}_0 = \frac{\mathbf{X}_1 - \mathbf{X}_t}{1-t}$ when $\sigma = 0$. Accordingly, we optimize the weighted regression loss

$$\mathscr{L}_{\mathrm{FM}} = \mathbb{E}_{(\mathbf{X}_0,\mathbf{X}_1)\sim\Pi^\star}\, \mathbb{E}_{t\sim\mathrm{Unif}[0,1-\varepsilon],\boldsymbol{\varepsilon}} \Big[w(t) \left\| f_\phi(t, \mathbf{X}_t) - \mathbf{X}_1 \right\|_{F,\Omega}^2 \Big], \tag{59}$$

with $w(t) = \min\{(1-t)^{-2}, w_{\max}\}$. In all experiments, we set $\varepsilon = 10^{-3}$ and clip the weight with $w_{\max} = 1/\varepsilon^2$ to prevent instabilities from the $(1-t)^{-2}$ factor near $t \approx 1$.

Sampling in representation space. To generate a new representation, we draw $\mathbf{X}_0 \sim q_0$ and integrate the ODE

$$\frac{d\mathbf{X}_t}{dt} \;=\; v_\phi(t, \mathbf{X}_t), \qquad t \in [0, 1-\varepsilon], \tag{60}$$

using a fixed-step solver with K steps. We default to Heun's method (predictor–corrector) for stability:

$$\tilde{\mathbf{X}}_{k+1} = \mathbf{X}_k + \Delta t\; v_\phi(t_k, \mathbf{X}_k), \tag{61}$$

$$\mathbf{X}_{k+1} = \mathbf{X}_k + \frac{\Delta t}{2}\Big(v_\phi(t_k, \mathbf{X}_k) + v_\phi(t_{k+1}, \tilde{\mathbf{X}}_{k+1}) \Big), \tag{62}$$

where $t_k = k\Delta t$ and $\Delta t = (1-\varepsilon)/K$. The final state $\mathbf{X}^{\mathrm{Final}} := \boldsymbol{\zeta}_{\mathrm{Dec}}(\mathbf{X}_K)$ is used as a generated sample in the representation space.

Algorithm 1 OT-CFM with Clean-Target Prediction: Sampling

Require: Trained network f_ϕ, source distribution q_0, number of steps K, Decoder $\boldsymbol{\zeta}_{\text{Dec}}$, decoder heads $\{\text{FFN}_{\text{coord}}, \text{FFN}_{\text{type}}\}$

Sample initial point: $\mathbf{X}_0 \sim q_0$
Initialize: $\mathbf{X} \leftarrow \mathbf{X}_0, \quad t \leftarrow 0, \quad \Delta t \leftarrow 1/K$
for $k = 0, 1, \ldots, K-1$ **do**
    Predict clean target: $\widehat{\mathbf{X}}_1 = f_\phi(t, \mathbf{X})$
    Compute velocity: $v = (\widehat{\mathbf{X}}_1 - \mathbf{X})/(1-t)$
    // Euler step:
    $\mathbf{X} \leftarrow \mathbf{X} + \Delta t \cdot v$
    $t \leftarrow t + \Delta t$
end for
$\mathbf{X}^{\text{Final}} \leftarrow \boldsymbol{\zeta}_{\text{Dec}}(\mathbf{X})$
// Decode to molecular structure:
$\widehat{\mathbf{x}}^\star \leftarrow \text{FFN}_{\text{coord}}(\mathbf{X}^{\text{Final}})$
$\widehat{\mathbf{h}} \leftarrow \text{FFN}_{\text{type}}(\mathbf{X}^{\text{Final}})$
// (Optional) Uniform lifting for arbitrary orientation:
// Sample $R \sim \mu$ (Haar measure on $SO(3)$)
// $\widehat{\mathbf{x}} \leftarrow \widehat{\mathbf{x}}^\star R$
Return Generated molecule $(\widehat{\mathbf{x}}^\star, \widehat{\mathbf{h}})$

Decoding to a 3D molecule (original scale). We map $\widetilde{\mathbf{X}}^{\text{Final}}$ back to molecular outputs using the pretrained decoder heads:

$$\widehat{\mathbf{x}}^\star = \text{FFN}_{\text{coord}}(\mathbf{X}^{\text{Final}}) \in \mathbb{R}^{N\times 3}, \quad \widehat{\mathbf{h}} = \text{FFN}_{\text{type}}(\mathbf{X}^{\text{Final}}), \tag{63}$$

which yields a centered 3D geometry $\widehat{\mathbf{x}}^\star$ in the same coordinate convention as pre-training. Our decoder outputs $\widehat{\mathbf{x}}^\star$ in a canonical coordinate system determined by the pretrained alignment, which fixes a representative on each $SO(3)$-orbit. This is sufficient when the evaluation (or downstream usage) is invariant to global rotations, since only the orbit $\{\widehat{\mathbf{x}}^\star R : R \in SO(3)\}$ matters rather than a particular pose. When samples must be expressed in arbitrary global orientations, we apply *uniform lifting* by drawing $R \sim \mu$ (the Haar measure on $SO(3)$) and returning $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^\star R$, which restores the correct distribution over orientations while preserving the canonical geometry. We summarize sampling procedures in Algorithm 1.

Backbone Architecture (DiT). To parameterize the flow model that predicts the clean target $\mathbf{X}_1$ from noised representations, we adopt a Diffusion Transformer (DiT) (65) architecture. The key design choice is the use of *adaptive layer normalization with zero initialization* (adaLN-Zero) for time conditioning, which modulates the hidden representations based on the diffusion timestep.

Given the noised representation $\mathbf{X}_t \in \mathbb{R}^{N\times H}$ and timestep $t \in [0,1]$, we first compute a time embedding via sinusoidal encoding followed by a two-layer MLP:

$$\mathbf{c} = \text{MLP}(\text{SinusoidalEmb}(t)) \in \mathbb{R}^H. \tag{64}$$

Each DiT block applies self-attention and a feed-forward network, where both operations are modulated by time-dependent parameters. Specifically, for a hidden state $\mathbf{x}$, the time embedding $\mathbf{c}$ generates six modulation parameters via a linear projection:

$$(\boldsymbol{\gamma}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_2, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2) = \text{Linear}(\text{SiLU}(\mathbf{c})), \tag{65}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta}$ denote scale and shift parameters for layer normalization, and $\boldsymbol{\alpha}$ denotes gating parameters. The modulated layer normalization is defined as:

$$\text{modulate}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \text{LayerNorm}(\mathbf{x}) \odot (1+\boldsymbol{\gamma}) + \boldsymbol{\beta}. \tag{66}$$

A DiT block then computes:

$$\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\alpha}_1 \odot \text{Attention}(\text{modulate}(\mathbf{x}, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_1)), \tag{67}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\alpha}_2 \odot \text{MLP}(\text{modulate}(\mathbf{x}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_2)). \tag{68}$$

The modulation layers are initialized to zero, ensuring that at the start of training, each block acts as an identity function, which stabilizes early optimization. We additionally incorporate *self-conditioning*: with probability 0.5 during training, we condition on a previous prediction $\hat{\mathbf{X}}_1$ by concatenating it with the input $\mathbf{X}_t$ before embedding. This provides the model with an estimate of the target, improving sample quality without additional computational cost at inference.

L Optimization via Latent Space Search

We optimize the pretrained flow-based generator to produce molecules with high binding affinity to a target protein pocket. Rather than retraining the flow model, we keep the learned flow and decoder fixed and optimize only the *initial-state sampling distribution*. This section describes our optimization framework based on LRA-CMA-ES.

L.1 Problem Formulation

Let $\Phi_\phi : \mathbb{R}^{N\times H} \to \mathbb{R}^{N\times H}$ be the frozen flow map and $\boldsymbol{\zeta}_{\text{Dec}} : \mathbb{R}^{N\times H} \to M$ be the frozen decoder. Given a target protein pocket $\mathscr{P}$, we seek an initial-state distribution p_ψ that maximizes the expected score of generated molecules:

$$\max_{\psi} J(\psi) := \mathbb{E}_{\mathbf{X}_0\sim p_\psi}\Big[S\big(\boldsymbol{\zeta}_{\text{Dec}}(\Phi_\phi(\mathbf{X}_0)); \mathscr{P}\big)\Big], \tag{69}$$

where the scoring function combines docking affinity and synthetic accessibility:

$$S(\mathscr{L}; \mathscr{P}) = -E_{\text{bind}}(\mathscr{L}; \mathscr{P}) + \lambda_{\text{SA}} \cdot S_{\text{SA}}(\mathscr{L}). \tag{70}$$

Here E_{bind} is the binding energy from UniDock (66) and $S_{\text{SA}} \in [0,1]$ is the normalized synthetic accessibility score. Invalid molecules receive a large penalty $S_{\text{penalty}} \ll 0$.

Since S involves non-differentiable operations (docking, discrete decoding), we employ derivative-free optimization. We parameterize p_ψ as a Gaussian $\mathscr{N}(\mathbf{m}, \sigma^2\mathbf{C})$ and optimize $\psi = (\mathbf{m}, \sigma, \mathbf{C})$ using LRA-CMA-ES.

L.2 LRA-CMA-ES Overview

CMA-ES (67) is an evolution strategy that maintains a Gaussian search distribution and adapts its covariance matrix to the local landscape. At iteration t, it samples λ candidates, evaluates their fitness, selects the top μ candidates, and updates the distribution parameters. The mean update is:

$$\mathbf{m}^{(t+1)} = \mathbf{m}^{(t)} + c_m \sum_{i=1}^{\mu} w_i \big(\mathbf{x}^{(i:\lambda)} - \mathbf{m}^{(t)}\big), \tag{71}$$

where $\mathbf{x}^{(i:\lambda)}$ is the i-th best candidate, $w_i > 0$ are selection weights, and c_m is the learning rate. The covariance $\mathbf{C}$ and step size σ are updated via evolution paths that accumulate information across iterations (see (67) for details).

When fitness evaluations are noisy, LRA-CMA-ES (68) adapts the learning rates based on the signal-to-noise ratio (SNR) of consecutive updates. Let $\boldsymbol{\delta}^{(t)} = \mathbf{m}^{(t+1)} - \mathbf{m}^{(t)}$. The SNR is estimated as:

$$\widehat{\text{SNR}}^{(t)} \propto \frac{\langle \boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t-1)} \rangle_{(\mathbf{C}^{(t)})^{-1}}}{\|\boldsymbol{\delta}^{(t)}\|_{(\mathbf{C}^{(t)})^{-1}} \cdot \|\boldsymbol{\delta}^{(t-1)}\|_{(\mathbf{C}^{(t)})^{-1}}}, \tag{72}$$

where $\|\cdot\|_{\mathbf{A}}$ denotes the norm induced by $\mathbf{A}$. The learning rates are then adapted as: $c_m^{(t+1)} = \text{clip}\Big(c_m^{(t)} \cdot g(\widehat{\text{SNR}}^{(t)}),\, c_{\min},\, c_{\max}\Big)$, where $g(\cdot)$ is a monotonically increasing function. When SNR is high (consistent updates), learning rates increase to accelerate convergence; when SNR is low (noisy updates), learning rates decrease to stabilize the search.

L.3 Application to Flow-Based Molecular Optimization

We apply LRA-CMA-ES to optimize the initial-state distribution of our flow model. Let $\mathbf{X}_0$ denote the vectorized initial state (restricted to non-padding entries). The search distribution is:

$$p_\psi^{(t)}(\mathbf{X}_0) = \mathscr{N}\big(\mathbf{m}^{(t)}, (\sigma^{(t)})^2\mathbf{C}^{(t)}\big), \tag{73}$$

initialized with $\mathbf{m}^{(0)} = \mathbf{0}$, $\sigma^{(0)} = \sigma_{\text{init}}$, and $\mathbf{C}^{(0)} = \mathbf{I}$.

At each iteration t, we sample λ initial states $\{\mathbf{X}_0^{(t,i)}\}_{i=1}^{\lambda}$ from the current distribution $\mathscr{N}(\mathbf{m}^{(t)}, (\sigma^{(t)})^2\mathbf{C}^{(t)})$. Each sample is passed through the frozen flow and decoder to obtain a molecule $M^{(t,i)} = \boldsymbol{\zeta}_{\text{Dec}}(\Phi_\phi(\mathbf{X}_0^{(t,i)}))$, which is then scored as $f^{(t,i)} = -\mathscr{S}(M^{(t,i)}; T)$. Based on the fitness values $\{f^{(t,i)}\}_{i=1}^{\lambda}$, we apply the LRA-CMA-ES update to obtain the next distribution parameters $(\mathbf{m}^{(t+1)}, \sigma^{(t+1)}, \mathbf{C}^{(t+1)})$. After T iterations, we sample from the optimized distribution and return top-scoring molecules. The procedure is summarized in Algorithm 2.

Parallelization. Fitness evaluation (flow integration + decoding + docking) dominates the computational cost. Since evaluations are independent across population members, we parallelize across λ workers. Each iteration then takes approximately the time of a single docking call ($\sim$1–5 seconds depending on molecule size and target complexity).

Algorithm 2 Flow-Based Molecular Optimization with LRA-CMA-ES

Require: Frozen flow map Φ_ϕ, Decoder $\boldsymbol{\zeta}_{\text{Dec}}$, Scoring Function $\mathscr{S}$, Target Protein T
Require: Population size λ, max iterations T
Initialize $p_\psi = \mathscr{N}(\mathbf{m}^{(0)}, (\sigma^{(0)})^2\mathbf{C}^{(0)})$ with $\mathbf{m}^{(0)} = \mathbf{0}$, $\sigma^{(0)} = \sigma_{\text{init}}$, $\mathbf{C}^{(0)} = \mathbf{I}$
for $t = 0, 1, \ldots, T-1$ **do**
    // Sample population from current distribution
    Sample $\{\mathbf{X}_0^{(i)}\}_{i=1}^{\lambda} \sim \mathscr{N}(\mathbf{m}^{(t)}, (\sigma^{(t)})^2\mathbf{C}^{(t)})$
    // Generate molecules via frozen flow and decoder
    for $i = 1, \ldots, \lambda$ **do**
        $\mathbf{X}_1^{(i)} \leftarrow \Phi_\phi(\mathbf{X}_0^{(i)})$ *// Integrate flow ODE*
        $\mathscr{L}^{(i)} \leftarrow \boldsymbol{\zeta}_{\text{Dec}}(\mathbf{X}_1^{(i)})$ *// Decode to molecule*
    end for
    // Evaluate fitness
    for $i = 1, \ldots, \lambda$ **do**
        $f^{(i)} \leftarrow -\mathscr{S}(M^{(i)}; T)$ *// Negative score*
    end for
    // Update distribution parameters via LRA-CMA-ES
    $(\mathbf{m}^{(t+1)}, \sigma^{(t+1)}, \mathbf{C}^{(t+1)}) \leftarrow$ LRA-CMA-ES-UPDATE$(\mathbf{m}^{(t)}, \sigma^{(t)}, \mathbf{C}^{(t)}; \{(\mathbf{X}_0^{(i)}, f^{(i)})\}_{i=1}^{\lambda})$
end for
// Final sampling from optimized distribution
Sample $\{\mathbf{X}_{\text{Optim}}^{(i)}\}_{i=1}^{N_{\text{final}}} \sim \mathscr{N}(\mathbf{m}^{(T)}, (\sigma^{(T)})^2\mathbf{C}^{(T)})$
Generate and decode: $M^{(i)} = \boldsymbol{\zeta}_{\text{Dec}}(\Phi_\phi(\mathbf{X}_{\text{Optim}}^{(i)}))$
Return Top-scoring molecules $\{M^{(i)}\}$

References

1. Lu, H., Szabados, S. & Yu, Y. Diffusion models with group equivariance. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*.

2. Cao, Z. *et al.* Efficient molecular conformer generation with so (3)-averaged flow matching and reflow. *arXiv preprint arXiv:2507.09785* (2025).

3. Ji, X. *et al.* Uni-mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969* (2024).

4. Shoghi, N. *et al.* From molecules to materials: Pre-training large generalizable models for atomic property prediction. In *The Twelfth International Conference on Learning Representations* (2024).

5. Li, S. *et al.* Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923* (2024).

6. Liu, Z. *et al.* Forging the basis for developing protein–ligand interaction scoring functions. *Accounts chemical research* **50**, 302–309 (2017).

7. Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).

8. Wang, S., Witek, J., Landrum, G. A. & Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **60**, 2044–2058 (2020).

9. Landrum, G. *et al.* Rdkit: Open-source cheminformatics. https://www.rdkit.org (2023).

10. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *J. Medicinal Chem.* **39**, 2887–2893 (1996).

11. Radovanovic, M., Nanopoulos, A. & Ivanovic, M. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.* **11**, 2487–2531 (2010).

12. Malvar, H. S., He, L.-w. & Cutler, R. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 3, iii–485 (IEEE, 2004).

13. Wang, T. & Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939 (PMLR, 2020).

14. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).

15. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proc. Neural Inf. Process. Syst. NeurIPS Datasets Benchmarks* (2021).

16. Björck, Å. & Golub, G. H. Numerical methods for computing angles between linear subspaces. *Math. Comput.* **27**, 579–594 (1973).

17. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations* (2023).

18. Li, P. *et al.* An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings Bioinforma.* **22**, bbab109 (2021).

19. Huang, K. *et al.* Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* (2021).

20. Francoeur, P. G. *et al.* Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. chemical information modeling* **60**, 4200–4215 (2020).

21. Luo, S., Guan, J., Ma, J. & Peng, J. A 3d generative model for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **34**, 6229–6239 (2021).

22. Axelrod, S. & Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **9**, 185 (2022).

23. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. data* **1**, 1–7 (2014).

24. Hoogeboom, E., Satorras, V. G., Vignac, C. & Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, 8867–8887 (PMLR, 2022).

25. Vignac, C., Osman, N., Toni, L. & Frossard, P. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 560–576 (Springer, 2023).

26. Dunn, I. & Koes, D. R. Flowmol3: flow matching for 3d de novo small-molecule generation. *ArXiv* arXiv–2508 (2025).

27. Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. advances* **3**, e1603015 (2017).

28. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).

29. Rosen, A. S. *et al.* Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **4**, 1578–1597 (2021).

30. Nikitin, F., Dunn, I., Koes, D. R. & Isayev, O. Geom-drugs revisited: Toward more chemically accurate benchmarks for 3d molecule generation. *arXiv preprint arXiv:2505.00169* (2025).

31. Buttenschoen, M., Morris, G. M. & Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).

32. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. cheminformatics* **1**, 8 (2009).

33. Hu, W. *et al.* Strategies for pre-training graph neural networks. In *International Conference on Learning Representations* (2020).

34. Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, vol. 33, 12559–12571 (2020).

35. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. In *Nature Machine Intelligence*, vol. 4, 279–287 (2022).

36. Yu, Q. *et al.* Multimodal molecular pretraining via modality blending. *arXiv preprint arXiv:2307.06235* (2024).

37. Zhang, J., Liu, Z., Guo, Y., Zhang, Z. & Ke, G. Auto-encoder based molecular representation learning with 3d cloze test objective. *arXiv preprint* (2024).

38. Li, S. *et al.* Unicorn: A unified contrastive learning approach for multi-view molecular representation learning. *arXiv preprint* (2024).

39. Vonessen, C. *et al.* Radialfocus: Geometric graph transformers via distance-modulated attention. *arXiv preprint* (2025).

40. Klaser, K. *et al.* Minimol: A parameter-efficient foundation model for molecular learning. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

41. Notwell, J. H. & Wood, M. W. Admet property prediction through combinations of molecular fingerprints. *arXiv preprint arXiv:2310.00174* (2023).

42. Hu, W. *et al.* Strategies for pre-training graph neural networks. In *International Conference on Learning Representations* (2020). ContextPred is one of the pretraining strategies proposed in this work.

43. Quazi, M., Schweikert, C., Hsu, D. F., Oprea, T. *et al.* Enhancing admet property models performance through combinatorial fusion analysis. (2023).

44. Turon, G. *et al.* Zairachem: First fully-automated ai/ml virtual screening cascade implemented at a drug discovery centre in africa. *Nat. Commun.* (2023).

45. Xu, M., Powers, A. S., Dror, R. O., Ermon, S. & Leskovec, J. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, 38592–38610 (PMLR, 2023).

46. Song, Y. *et al.* Unified generative modeling of 3d molecules via bayesian flow networks. *arXiv preprint arXiv:2403.15441* (2024).

47. Song, Y. *et al.* Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *Advances in Neural Information Processing Systems* (2024).

48. Hong, H., Lin, W. & Tan, K. C. Accelerating 3d molecule generation via jointly geometric optimal transport. *arXiv preprint arXiv:2405.15252* (2024).

49. Le, T., Noe, F. & Clevert, D.-A. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *International Conference on Learning Representations* (2024).

50. Reidenbach, D., Nikitin, F., Isayev, O. & Paliwal, S. G. Applications of modular co-design for de novo 3d molecule generation. *Digit. Discov.* (2025).

51. Irwin, R., Tibo, A., Janet, J. P. & Olsson, S. Semlaflow–efficient 3d molecular generation with latent attention and equivariant flow matching. In *International Conference on Artificial Intelligence and Statistics*, 3772–3780 (PMLR, 2025).

52. Guan, J. *et al.* 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations* (2023).

53. Guan, J. *et al.* Decomposed diffusion models for structure-based drug design. In *International Conference on Machine Learning*, 11827–11846 (PMLR, 2023).

54. Qu, Y. *et al.* Molcraft: Structure-based drug design in continuous parameter space. In *International Conference on Machine Learning* (2024).

55. Controllable and decomposed diffusion models for structure-based molecular optimization. *arXiv preprint* (2024).

56. Shen, T., Qu, M., Gong, J. & Bengio, Y. Target-conditioned gflownet for structure-based drug design. In *International Conference on Learning Representations* (2024).

57. Huang, S. *et al.* Aligning target-aware molecule diffusion models with exact energy optimization. In *International Conference on Machine Learning* (2024).

58. Franklin, S. J., Younis, U. S. & Myrdal, P. B. Estimating the aqueous solubility of pharmaceutical hydrates. *J. pharmaceutical sciences* **105**, 1914–1919 (2016).

59. Vertzoni, M. *et al.* Ungap best practice for improving solubility data quality of orally administered drugs. *Eur. J. Pharm. Sci.* **168**, 106043 (2022).

60. Mobley, D. L. & Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J. computer-aided molecular design* **28**, 711–720 (2014).

61. Tong, A. *et al.* Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Mach. Learn. Res.* 1–34 (2024).

62. Yim, J. *et al.* Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297* (2023).

63. Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. *Adv. neural information processing systems* **31** (2018).

64. Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

65. Peebles, W. & Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205 (2023).

66. Yu, Y. *et al.* Uni-dock: Gpu-accelerated docking enables ultralarge virtual screening. *J. chemical theory computation* **19**, 3336–3345 (2023).

67. Hansen, N. & Ostermeier, A. Completely derandomized self-adaptation in evolution strategies. *Evol. computation* **9**, 159–195 (2001).

68. Nomura, M., Akimoto, Y. & Ono, I. Cma-es with learning rate adaptation. *ACM Transactions on Evol. Learn.* **5**, 1–28 (2025).