Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Learning Canonical Representations for Unified 3D Molecular Modeling

Sanghyun Park

sanghyun@yonsei.ac.kr

Yonsei University https://orcid.org/0000-0002-5196-6193

Seungyeon Choi

Yonsei University

Hwanhee Kim

Yonsei University https://orcid.org/0009-0004-9227-5550

yoonju kim

yonsei university

seungyong lee

yonsei university

inpyo hong

yonsei university

Article

Keywords:

Posted Date: July 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10292206/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

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

Three-dimensional molecular foundation models are increasingly expected to support scalar prediction, coordinate generation and structure-based optimization, yet these tasks impose different geometric-symmetry requirements on the learned representation. Existing models typically encode geometry through invariant, equivariant or rotation-augmented representations, which can discard task-relevant geometric information, entangle orientation with chemical variation or require multi-view inference. Here we introduce THEMOL, a 3D molecular representation framework that learns a canonical molecular frame before encoding and applies a non-equivariant backbone in this canonical space. This design separates global pose handling from representation learning, enabling a single pretrained latent space to be reused across prediction, generation and optimization. We provide a theoretical analysis of symmetry-induced trade-offs in reusable 3D molecular representations, covering invariant, equivariant and augmentation-based designs, and evaluate representation quality beyond downstream scores. Across molecular property prediction, 3D generation and structure-based optimization benchmarks, THEMOL achieves competitive or superior performance while improving rotational stability and latent-space structure.

Three-dimensional molecular structures provide the geometric information needed for many drug-discovery tasks, including molecular property prediction, coordinate generation and structure-based optimization. Recent molecular models increasingly aim to reuse representations pretrained on large-scale unlabelled data (1; 2; 3), rather than training separate models for each task. However, building such a reusable 3D representation is not straightforward, because different downstream tasks impose different requirements on how global rotations and translations should be handled.

Existing 3D molecular embeddings typically address euclidean symmetry through one of three strategies. Invariant models remove global pose and are effective for scalar prediction (4; 5), equivariant models preserve transformation structure and are well suited to coordinate- or force-like outputs (6; 7), and rotation-augmented non-equivariant models learn robustness statistically from randomly oriented inputs (8; 9). These designs have been highly successful in task-specific settings. Yet, when the same pretrained latent space must support prediction, generation and latent-space optimization, each strategy introduces a different compromise.

Invariant representations can suppress geometric information needed for coordinate-level reconstruction or chirality-sensitive reasoning (10; 11; 12). Equivariant latents preserve pose information, but may mix orientation degrees of freedom with chemically meaningful variation when used for latent generative modelling or optimization (13; 14; 15). Rotation augmentation reduces average sensitivity to orientation, but does not by itself guarantee single-view stability and may require multi-view inference to obtain robust predictions (16). These observations suggest that the central challenge is not simply choosing between invariance and equivariance, but separating global pose handling from molecular representation learning. Understanding these trade-offs calls for a principled account of how each geometric representation regime — invariant, equivariant and rotation-augmented — behaves when a single latent space is reused across property prediction, 3D molecule generation and structure-based optimization.

A second challenge concerns evaluation. Downstream benchmark performance alone does not directly measure the quality of a pretrained representation, because it can be confounded by task-specific heads, finetuning protocols and dataset effects (17). A reusable 3D molecular representation therefore requires direct evaluation of the latent space itself, including its geometric stability, chemical organization and transferability before downstream adaptation.

Here we introduce THEMOL, a 3D molecular representation framework that learns a canonical molecular frame before encoding. By mapping rotated copies of a molecule to a shared canonical pose, THEMOL allows a non-equivariant backbone to operate in a consistent coordinate system while retaining access to coordinate-level information. This design treats global pose as an interface variable rather than as a burden carried throughout the representation, enabling a single pretrained latent space to be reused across scalar prediction, coordinate generation and structure-based optimization.

We analyse the symmetry-induced trade-offs of invariant, equivariant and augmentation-based representations (Section.1.1), and show how learned canonicalization provides a principled alternative for reusable 3D molecular representations (Section.1.2). Beyond standard downstream benchmarks, we evaluate representation quality directly through molecular identity preservation, chemical semantic coherence, embedding geometry and pretrain–finetune alignment (Section.1.3,1.4). Across molecular property prediction, 3D molecule generation and structure-based optimization benchmarks, THEMOL achieves competitive or superior performance while improving rotational stability and latent-space structure (Section.1.5).

1 Results

1.1 Symmetry trade-offs in reusable 3D molecular representations

We first formalize the representation regimes that underlie reusable 3D molecular modelling. For notational clarity, we refer to invariant, equivariant and rotation-augmented representations as Repr.A, Repr.B and Repr.C, respectively, and denote our canonicalization-based representation as Repr.Z. Let $\Phi(\mathbf{x},\mathbf{h})$ denote a molecular representation map, where $\mathbf{x} \in \mathbb{R}^{N\times 3}$ are atomic coordinates and $\mathbf{h}$ denotes atom-level features. For a global Euclidean transformation $g \in \mathrm{SE}(3)$, Repr.A removes global pose at the latent level, satisfying

$$\Phi(g\cdot\mathbf{x},\mathbf{h}) = \Phi(\mathbf{x},\mathbf{h}),$$

typically through distance-, angle- or internal-coordinate-based invariant features. Repr.B preserves transformation structure through equivariant features that co-transform with molecular rotations,

$$\Phi(g\cdot\mathbf{x},\mathbf{h}) = \rho(g)\Phi(\mathbf{x},\mathbf{h}),$$

for a representation ρ acting on the feature space. Repr.C uses a non-equivariant backbone but relies on rotation augmentation to learn robustness to input orientation.

These strategies are well suited to task-specific settings when the symmetry requirement of the target is known in advance. However, when a single pretrained latent space must be reused across scalar prediction, coordinate generation and structure-based optimization, each regime introduces a different trade-off. We next analyse these trade-offs in the setting of pretrained molecular embeddings, focusing on how each representation regime behaves when a single latent space is reused across heterogeneous downstream tasks, as shown in Figure.1a.

Invariant representations. Because Repr.A removes global pose at the latent level, it provides an efficient and physically natural choice for scalar property prediction. However, the same removal introduces two trade-offs when the representation is reused for foundation-level modelling. First, distance-only geometric descriptors are invariant not only to global rotations and translations but also to reflections, effectively making them $O(3)$-invariant. They can therefore remove reflection-odd signals (10; 11) and fail to distinguish certain chiral configurations, as illustrated in Figure.1a. Second, a purely invariant latent specifies molecular geometry only up to a gauge, and therefore requires an additional frame-selection or lifting mechanism to reconstruct or generate coordinates in a chosen frame (12). Thus, Repr.A is well matched to invariant scalar prediction, but can be restrictive when a single pretrained latent space is expected to support coordinate-level generation or optimization. Further discussion and representative model examples are provided in Supplementary Note D.1.

Equivariant representations Equivariant representations provide a natural inductive bias for outputs that must co-transform with rotations, such as forces, vector fields and coordinate-level predictions. They are also effective for tasks that require both invariant and equivariant outputs within a single model, as demonstrated by modern equivariant architectures such as EquiformerV2 (7). However, when a single pretrained autoencoder latent is frozen and reused beyond prediction (for latent generative modelling and latent-space optimization) the same equivariant structure can introduce a distinct representational cost.

For latent generative modelling, an equivariant encoder maps rotated copies of the same molecule to an orbit rather than to a single latent point. After centering the coordinates, for $R \in \mathrm{SO}(3)$, $\mathrm{Enc}(R\mathbf{x},\mathbf{h}) = \rho(R)\mathrm{Enc}(\mathbf{x},\mathbf{h})$, where ρ acts on the latent feature space. Thus, non-scalar equivariant channels encode the global orientation of a molecule in addition to its intrinsic chemical and geometric variation. A standard Euclidean latent generative model must therefore model not only chemically meaningful variation, such as bonding patterns, conformational changes and substitutions, but also the $\mathrm{SO}(3)$-orbit structure induced by arbitrary global rotations. This can broaden the latent distribution, increase orientation-induced variance and make flow or diffusion trajectories mix chemical changes with gauge changes (13; 14; 15).

Latent-space optimization faces an analogous difficulty. Effective optimization assumes that local perturbations in the latent space correspond to meaningful changes in molecular properties, such as binding affinity, drug-likeness, synthesizability or structural stability. In an equivariant latent space, however, some directions can change orientation or gauge components

rather than molecular structure. This effect becomes more pronounced when the latent contains high-dimensional non-scalar equivariant channels, such as vector or higher-order irreducible representation features (7). As a result, the effective search space can include orientation degrees of freedom that are irrelevant to molecular improvement, making property-guided gradients or black-box search less aligned with chemically meaningful directions. Thus, while Repr.B is well suited to coordinate- and force-level modelling, it can make frozen latent generation and optimization unnecessarily sensitive to global orientation.

Rotation-augmented non-equivariant representations. Repr.C handles global rotations statistically rather than structurally: a non-equivariant backbone is trained on randomly rotated inputs so that its expected loss over $\mathrm{SO}(3)$ is small. This design is simple and broadly applicable, but it creates a mismatch between training and deployment. The augmentation objective improves average-case robustness over sampled orientations, whereas downstream use typically evaluates the model from a single input orientation. Thus, low augmented risk does not necessarily imply that the embedding or prediction is uniformly stable across all rotations of the same molecule. Different test-time orientations can still produce different latent codes or predictions, especially when the model has only learned rotational robustness from finite samples.

Stability can be improved by explicitly averaging predictions over multiple rotated views at inference time. However, this shifts invariance from the representation itself to test-time computation. Because $\mathrm{SO}(3)$ is a three-dimensional continuous group, uniform rotational control requires a rapidly growing number of sampled views; under standard covering-number arguments (16; 18), achieving ε-level control requires $K = \Omega(\varepsilon^{-3})$ views. Therefore, Repr.C trades architectural simplicity for either unguaranteed single-view stability or expensive multi-view inference. As illustrated in Figure.1b, rotation-augmented embeddings may remain dispersed across random orientations, whereas canonicalization maps rotated copies of a molecule to a consistent representation. Formal analysis is provided in Supplementary Note D.3.

1.2 Proposed representation for reusable 3D molecular modelling: Learned canonicalization

To address the trade-offs above, we introduce Repr.Z, a canonicalization-based representation framework. The core idea is to separate global pose handling from representation learning. Before encoding a molecule, Repr.Z maps its coordinates to a learned canonical frame; a standard non-equivariant backbone then operates in this canonical coordinate system. Ideally, rotated copies of the same molecule are mapped to the same canonical representative, so that the reusable latent space describes molecular variation rather than arbitrary global orientation. When coordinate-level outputs are required, predictions made in the canonical frame can be lifted back to the original coordinate system using the selected rotation.

Let $\mathbf{x} \in \mathbb{R}^{N\times 3}$ denote atomic coordinates and $\mathbf{h}$ denote atom-level features. We first remove translation by centering

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

For row-wise coordinate rotations, $R \in \mathrm{SO}(3)$ acts as $\mathbf{x} \mapsto \mathbf{x}R^{\top}$. We learn a rotation selector $\widehat{R}_\theta(\mathbf{x},\mathbf{h}) \in \mathrm{SO}(3)$ and define the canonicalized coordinates as

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

The desired gauge-consistency condition is

$$\widehat{R}_\theta(\mathbf{x}R^{\top},\mathbf{h}) = \widehat{R}_\theta(\mathbf{x},\mathbf{h})R^{\top}, \qquad \text{which implies} \qquad A_\theta(\mathbf{x}R^{\top},\mathbf{h}) = A_\theta(\mathbf{x},\mathbf{h}).$$

Thus, global rotations are removed before the backbone sees the input.

Proposition 1 (One canonical backbone for invariant and equivariant outputs)**.** *Assume that the canonicalization map satisfies*

$$A_\theta(\mathbf{x}R^{\top},\mathbf{h}) = A_\theta(\mathbf{x},\mathbf{h}) \qquad \forall R \in \mathrm{SO}(3).$$

Then any predictor of the form

$$\widehat{y}_{\mathrm{inv}}(\mathbf{x},\mathbf{h}) = \widetilde{f}_\eta(A_\theta(\mathbf{x},\mathbf{h}),\mathbf{h})$$

is rotation-invariant. Moreover, if the output space carries a representation ρ of $\mathrm{SO}(3)$, and the selector satisfies

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

then the lifted predictor

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

is ρ-equivariant.

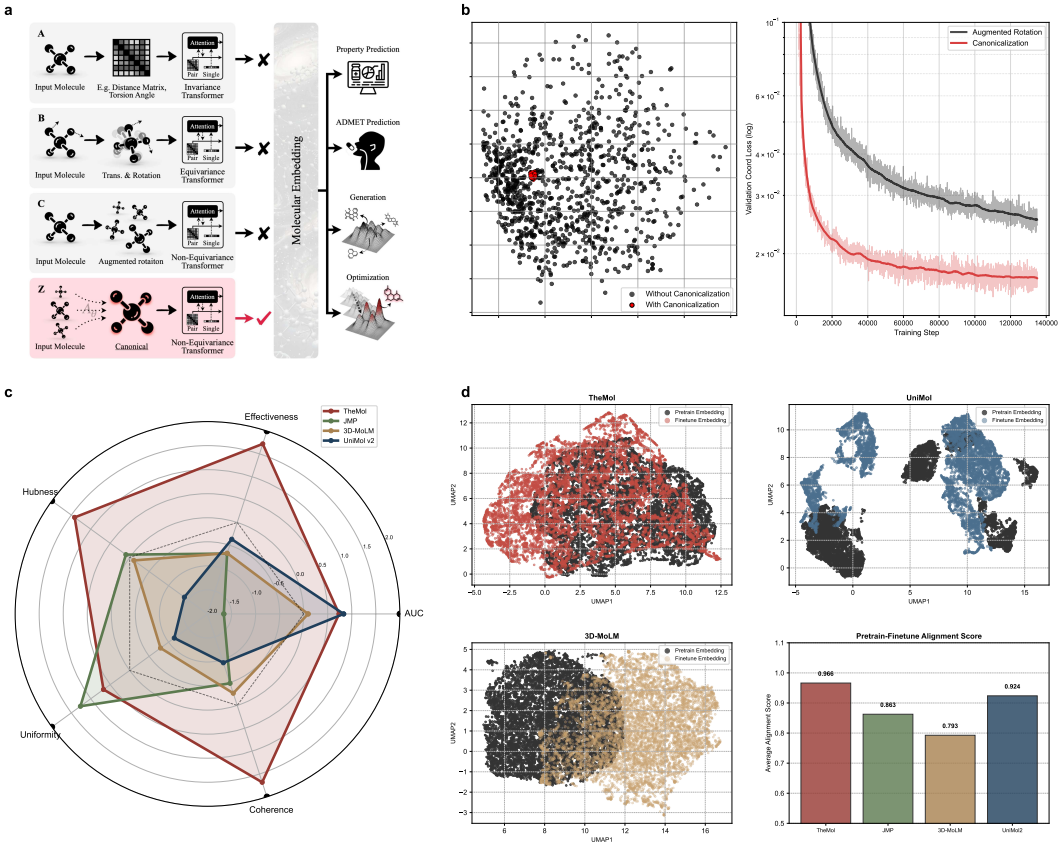


Figure 1. Overview of THEMOL and representation-level diagnostics. a, Overview of the THEMOL framework and representative 3D molecular representation regimes. Repr.A encodes molecules using geometric invariants, Repr.B preserves coordinate transformation structure throughout the network, and Repr.C trains a non-equivariant backbone with random rotations. Our Repr.Z first maps each molecule to a learned canonical pose and then processes the canonicalized coordinates with a non-equivariant transformer, enabling scalar and coordinate-level downstream tasks within a unified representation framework. **b**, Effect of canonicalization on rotation stability and pretraining convergence. Left, embeddings of aspirin under 1,000 random rotations without canonicalization (Repr.C) and with canonicalization (Repr.Z). Right, validation coordinate reconstruction loss during pretraining for canonicalization and rotation augmentation. Experimental details are provided in Supplementary Note G. **c**, Normalized radar chart of five intrinsic representation-quality diagnostics: molecular identity preservation (AUC), coherence, effective dimension, hubness and uniformity. Hubness and uniformity are sign-adjusted so that larger values consistently indicate better representation quality. **d**, UMAP visualization of pretrained and finetuned embedding distributions, together with average alignment scores computed by principal subspace similarity. Higher alignment indicates stronger overlap between pretraining and finetuning variation directions. Detailed definitions and protocols for representation-quality diagnostics and alignment analysis are provided in Supplementary Note E.

Proposition 1 clarifies why canonicalization can unify different symmetry requirements within a single backbone. The backbone operates only in the canonical coordinate system: invariant predictions are read out directly, whereas equivariant outputs are obtained by lifting canonical predictions back through the selected frame. Thus, unlike Repr.B, Repr.Z does not leave the SO(3)-orbit inside the reusable latent space. The selector $\widehat{R}_\theta$ absorbs this orbit before representation learning, allowing latent generation and optimization to act on canonical molecular variation rather than on a mixture of chemical and orientation degrees of freedom.

Canonicalization also avoids the main deployment issue of Repr.C. Rotation augmentation learns robustness only on average over sampled orientations, so single-view predictions can remain orientation-dependent. By contrast, Repr.Z explicitly selects a canonical frame before encoding; residual rotation sensitivity is therefore controlled by the learned alignment error rather than by the number of test-time views. Under approximate gauge consistency with error δ, the induced invariance violation is bounded by $O(\delta)$, without requiring multi-view averaging. Empirically, Figure.1b shows that canonicalization collapses randomly rotated copies of a molecule into a compact embedding cluster and improves reconstruction convergence over rotation augmentation. Formal proofs and extended generation-space analysis are provided in Supplementary Note A–C.

1.3 Evaluating representation quality beyond downstream benchmarks

Downstream benchmark performance is an indirect proxy for representation quality, as it can reflect task-specific heads, finetuning protocols and dataset effects as much as the pretrained embedding itself. We therefore evaluate the frozen molecular embedding space directly using five intrinsic diagnostics: molecular identity preservation, chemical semantic coherence, effective dimension, hubness and uniformity.

As shown in Figure 1c, the compared models exhibit different representation-level trade-offs. Uni-Mol2 (2) shows favourable geometry-related behaviour in hubness and uniformity, indicating a regular nearest-neighbour structure, but this does not translate into the strongest chemical coherence. JMP (3) and 3D-MoLM (19) provide competitive scores on individual axes, but do not show the same balanced profile across semantic organization, dimensional utilization and overall intrinsic quality. In contrast, THEMOL achieves the strongest coherence and the highest effective dimension while maintaining competitive identity preservation and aggregate representation quality. This pattern suggests that canonicalization preserves richer coordinate-dependent structural variation without allowing arbitrary global orientation to dominate the latent space.

We further assess transfer compatibility using pretrain–finetune alignment (Figure 1d). This diagnostic compares the principal subspaces of pretrained and finetuned embeddings, measuring whether downstream adaptation follows variation directions already present in the pretrained representation. THEMOL achieves the highest alignment score among the compared models, indicating stronger overlap between pretraining and finetuning variation directions. Together, the intrinsic diagnostics and alignment analysis suggest that THEMOL provides a latent space that is chemically organized, geometrically well utilized and compatible with downstream adaptation. These representation-level results complement downstream benchmark scores and support the use of THEMOL as a reusable 3D molecular representation for prediction, generation and optimization. We provide detailed definitions, implementation protocols and further analysis of these diagnostics in Supplementary Note E.

1.4 Canonicalization organizes latent space for molecular modelling

The intrinsic diagnostics above show that canonicalization improves the geometry and semantic organization of frozen molecular embeddings. We next ask whether this organization is operationally useful for molecular modelling. A reusable 3D latent space should support chemically meaningful comparison, decoding, interpolation and optimization without being dominated by arbitrary global pose. Thus, rotations of the same molecule should be suppressed as nuisance variation, conformational and chemical changes should remain distinguishable, nearby latent states should decode to plausible structures, and directed search should improve molecular objectives rather than spend displacement on frame changes.

Figure 2 evaluates these requirements across Repr.A, Repr.B, Repr.C and Repr.Z. Repr.Z strongly reduces rotation-induced latent drift, approaching the stability of invariant representations while retaining coordinate decodability. This is distinct from Repr.A: canonicalization removes global pose before encoding, but does not reduce geometry to invariant descriptors. The latent-distance hierarchy (Figure.2b) further shows that this removal is selective. In a well-organized molecular latent space, rotations should induce the smallest displacement, followed by conformational changes, matched-molecular-pair changes and scaffold-level differences. Repr.Z most consistently follows this ordering, indicating that canonicalization separates nuisance pose variation from chemically meaningful structural variation.

We then tested whether this latent organization remains compatible with generation, as shown in Figure 2c. Local perturbation and interpolation both probe whether the decoder sees a smooth, physically plausible neighbourhood around molecular embeddings. Repr.Z maintains higher PB-Valid values under latent perturbations and along interpolation paths, whereas equivariant or augmentation-based latents show greater validity loss away from observed training points. These results suggest that removing the rotation orbit before representation learning makes the latent manifold more decoder-compatible, rather than merely stabilizing embeddings.

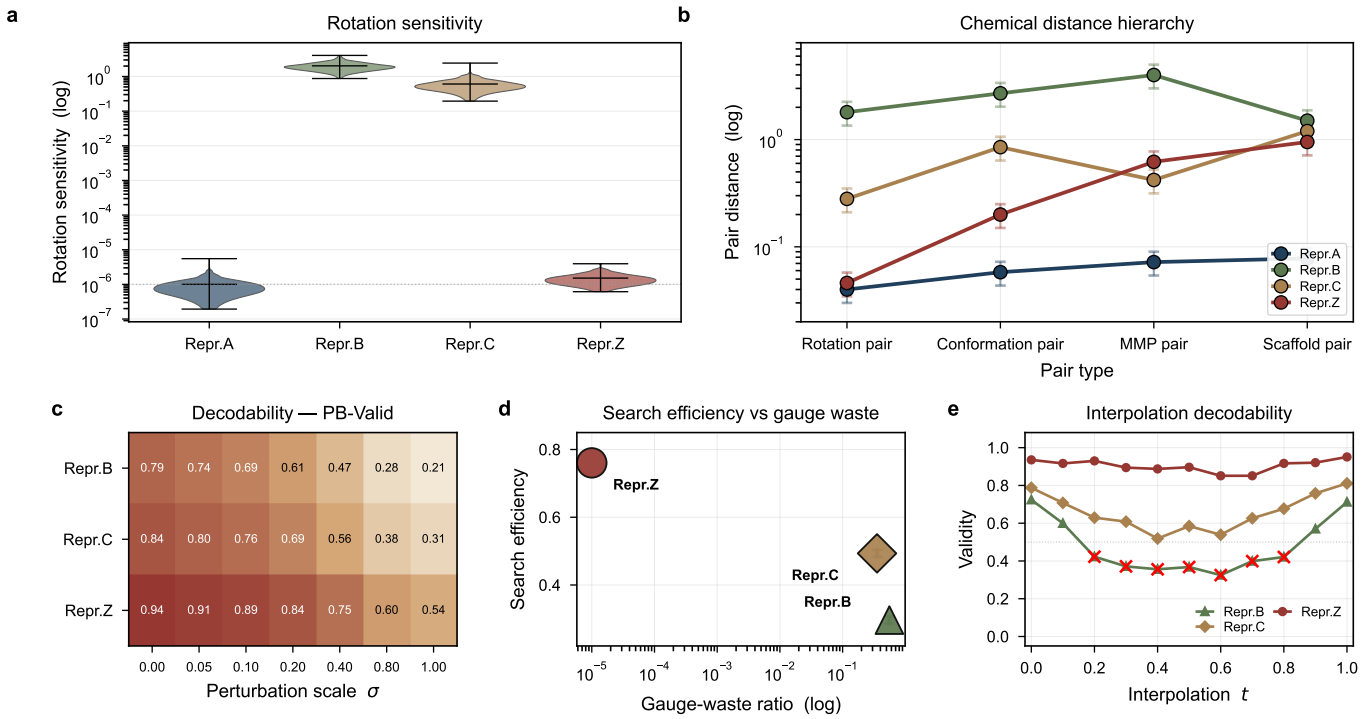


Figure 2. Task-readiness diagnostics of reusable 3D molecular representations. We compare Repr.A, Repr.B, Repr.C and Repr.Z using diagnostics designed to probe whether a frozen latent representation is suitable for prediction, generation and optimization. **a**, Rotation sensitivity, measured by latent drift under random global rotations. **b**, Latent-distance hierarchy across increasingly semantic molecular changes, from global rotation to conformer, matched molecular pair (MMP) and scaffold-level variation. **c**, Local latent decodability, measured by PB-Valid after perturbing latent codes and decoding the resulting molecules. **d**, Optimization-readiness analysis, comparing gauge waste, defined as the fraction of latent search displacement projected onto a rotation-induced latent subspace, and search efficiency, defined as objective improvement per unit latent path length. Point size denotes molecular diversity and colour denotes valid generation rate. **e**, Interpolation decodability, measured by PB-Valid along linear interpolations between molecular latent codes. Repr.A is omitted from coordinate-decoding diagnostics where an invariant-only latent does not define a canonical coordinate gauge without an additional frame-selection or lifting mechanism. Detailed metric definitions and experimental protocols are provided in Supplementary Note F.

Table 1. The overall results on 9 molecule classification datasets from the MoleculeNet (20). We report ROC-AUC score (higher is better) under scaffold splitting. The best result is highlighted in bold and the second-best is underlined. * denotes the mean calculated on reported tasks only.

Datasets	BACE (↑)	BBBP (↑)	Tox21 (↑)	SIDER (↑)	HIV (↑)	MUV (↑)	PCBA (↑)	ClinTox (↑)	ToxCast (↑)	Mean (↑)
PretrainGNN(21)	84.5	72.6	78.1	62.7	79.9	81.3	86.0	72.6	65.7	75.93
GROVER(22)	82.6	70.0	74.3	64.8	62.5	62.5	76.5	81.2	65.4	71.09
MolCLR(23)	82.4	72.2	75.0	58.9	78.1	79.6	-	91.2	69.2	75.83*
MoleBLEND(24)	83.7	73.0	77.8	64.9	79.0	77.2	-	87.6	66.1	76.16*
Uni-Mol(25)	83.2	71.5	78.9	57.7	78.6	72.6	88.1	84.1	69.1	75.98
Mol-AE(26)	84.1	72.0	80.0	**67.0**	80.6	81.6	**88.9**	87.8	69.6	79.04
UniCorn(27)	85.8	74.2	79.3	64.0	-	**82.6**	-	92.1	69.4	78.0*
THEMOL	**89.9**$_{\pm 0.7}$	**91.2**$_{\pm 0.5}$	**81.3**$_{\pm 0.4}$	64.3$_{\pm 0.8}$	**85.0**$_{\pm 0.6}$	78.1$_{\pm 1.2}$	83.7$_{\pm 0.3}$	**98.9**$_{\pm 0.4}$	**69.7**$_{\pm 0.5}$	**81.65**

Table 2. 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 (↑)	
# Metrics	Atom Sta (%)	Valid (%)
Data	86.5	99.9
GeoLDM(28)	82.4	92.8
GeoBFN(29)	78.9	93.1
EquiFM(30)	84.1	98.9
GOAT(31)	84.8	96.2
THEMOL	**86.8**$_{\pm 0.11}$	**99.9**

	GEOM-DRUG v2			
# Metrics	PB-Valid (%)	OOD Rings (%)	Med. ΔE_{relax}	Med. ΔR_{relax}
Data	93.2$_{\pm 0.01}$	0.05	0.00	0.00
EQGAT-Diff(32)	77.6	0.28	6.51	0.60
Megalodon(33)	86.6	0.17	**3.17**	0.41
SemlaFlow(34)	88.5	0.0	31.92	0.24
FlowMol3(35)	91.9	**0.10**	3.83	0.39
THEMOL	**93.7**$_{\pm 0.93}$	0.18$_{\pm 0.01}$	4.97$_{\pm 0.02}$	**0.17**$_{\pm 0.03}$

Finally, we examined whether this organization improves property-guided search (Figure 2d). In equivariant latent spaces, part of the optimization trajectory can move along directions associated with global orientation rather than molecular change. We quantify this effect as gauge waste and compare it with search efficiency, defined as objective improvement per unit latent displacement. Repr.Z shows lower gauge waste and higher search efficiency while preserving validity and molecular diversity, consistent with the intended role of canonicalization: absorbing the global rotation orbit before encoding biases latent search toward chemically relevant directions.

Together, these diagnostics link representation quality to downstream utility. THEMOL produces a latent space in which orientation is controlled, chemical variation is structured, decoding remains stable and optimization is more efficiently aligned with molecular improvement. This task-readiness analysis motivates the downstream benchmarks in the next section, where the same canonical latent is evaluated for scalar prediction, 3D generation and structure-based optimization.

1.5 A canonical latent space transfers across prediction, generation and optimization

We next evaluate whether the canonical representation remains useful when the same pretrained latent space is placed under three distinct demands: scalar prediction, coordinate-level generation and structure-based optimization. These tasks stress different aspects of a 3D molecular representation. Property prediction requires chemically informative invariant features; generation requires coordinate-level geometric realism; and optimization requires a latent space in which property-guided search improves molecular objectives without collapsing chemical diversity. Dataset details, metrics and baseline models are provided in Appendices J, J.7 and J.8.

Scalar prediction tests whether canonicalization preserves biochemical information. On MoleculeNet, THEMOL achieves the highest mean ROC-AUC of 81.65 and ranks first on six of nine datasets, including BACE, BBBP and ClinTox. On the TDC ADMET benchmark, THEMOL attains a mean rank of 3.4 across 22 tasks and achieves first-place performance on 15 tasks spanning absorption, distribution, metabolism, excretion and toxicity. These results indicate that encoding molecules in a learned canonical frame does not reduce the representation to a purely geometric reconstruction code. Instead, the canonical latent remains strongly predictive of molecule-level biochemical and pharmacokinetic properties.

Generation tests whether the same latent retains coordinate-level realism. On GEOM-DRUG, THEMOL reaches 86.8% atom stability and 99.9% validity under the v1 protocol, matching the quality of the training distribution. Under the stricter v2 protocol, THEMOL achieves the highest PB-Valid score and the lowest median ΔR_{relax}, showing that generated geometries satisfy physical validity checks and require minimal structural displacement after relaxation. Although specialized 3D generators remain competitive on selected metrics, these results show that a canonical representation pretrained for reuse across tasks can still support realistic coordinate generation.

Optimization tests whether the latent space is searchable without chemical collapse. On CrossDocked2020 (36),

Table 3. Performance of THEMOL on the TDC Benchmark and its comparison with the TDC Leaderboard and MiniMol. The best results are **bolded**, underlined. The mean rank is computed using only the tasks in the TDC leaderboard.

Methods	TDC Dataset			TDC Leaderboard (January 2026)		MiniMol		THEMOL	
	Dataset	Metric	Size	SOTA Model	SOTA Result	Result	Rank	Result	Rank
Absorption	Caco2	MAE(↓)	906	CaliciBoost	0.256±0.006	0.350±0.018	16	0.291±0.009	8
	HIA	AUROC(↑)	578	MiniMol	0.993±0.005	0.993±0.005	2	**0.996**±0.003	**1**
	Pgp	AUROC(↑)	1212	MapLight + GNN	0.938±0.002	-	-	**0.939**±0.004	**1**
	Bioavailability	AUROC(↑)	640	MiniMol	0.942±0.002	**0.942**±0.002	**1**	0.706±0.015	-
	Lipophilicity	MAE(↓)	4200	MiniMol	0.456±0.008	**0.456**±0.008	**1**	**0.456**±0.007	**1**
	Solubility	MAE(↓)	9982	MiniMol	0.741±0.013	0.741±0.013	2	**0.722**±0.010	**1**
Distribution	BBB	AUROC(↑)	1975	MiniMol	0.924±0.003	0.924±0.003	2	**0.927**±0.004	**1**
	PPBR	MAE(↓)	1797	Gradient Boost	7.440±0.024	7.696±0.125	5	7.461±0.078	2
	VDss	Spearman(↑)	1130	MapLight + GNN	0.713±0.007	0.535±0.027	10	0.704±0.016	3
Metabolism	CYP2D6 inhibition	AUPRC(↑)	13130	MapLight + GNN	0.790±0.001	0.719±0.004	5	0.703±0.005	7
	CYP3A4 inhibition	AUPRC(↑)	12328	MapLight + GNN	0.916±0.000	0.877±0.001	7	0.880±0.004	6
	CYP2C9 inhibition	AUPRC(↑)	12092	MapLight + GNN	0.859±0.001	0.823±0.006	4	0.798±0.006	5
	CYP2D6 substrate	AUPRC(↑)	664	ContextPred	0.736±0.024	0.695±0.032	8	**0.775**±0.021	**1**
	CYP3A4 substrate	AUROC(↑)	667	CFA	0.667±0.019	0.663±0.008	3	**0.787**±0.014	**1**
	CYP2C9 substrate	AUPRC(↑)	666	MiniMol	0.474±0.025	0.474±0.025	2	**0.542**±0.019	**1**
Excretion	Half life	Spearman(↑)	667	CFA	0.576±0.025	0.495±0.042	7	0.559±0.026	3
	Clearance mircosome	Spearman(↑)	1102	MapLight + GNN	0.630±0.010	0.628±0.005	3	**0.692**±0.011	**1**
	Clearance hepatocyte	Spearman(↑)	1020	CFA	0.536±0.020	0.446±0.029	6	0.479±0.023	3
Toxicity	hERG	AUROC(↑)	648	MapLight + GNN	0.880±0.002	0.846±0.016	7	0.863±0.013	5
	Ames	AUROC(↑)	7255	ZairaChem	0.871±0.002	0.849±0.004	7	0.835±0.005	13
	DILI	AUROC(↑)	475	MiniMol	0.956±0.006	**0.956**±0.006	**1**	0.933±0.010	2
	LD50	MAE(↓)	7385	BaseBoosting	0.552±0.009	0.585±0.008	3	0.608±0.009	6
					TDC Leaderboard Mean Rank:		4.9		**3.4**

Table 4. Summary of binding affinity and molecular properties of reference molecules and molecules generated by THEMOL and baselines. Top 2 results are **bolded** and underlined, respectively.

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	-	-
Gen.	TargetDiff (37)	-8.51	-8.56	-7.80	-7.91	0.58	0.58	0.72	0.71
	DecompDiff (38)	-8.33	-8.27	-8.39	-8.43	0.61	0.60	0.68	0.68
	MolCRAFT (39)	-8.56	**-8.41**	-7.59	-7.31	0.68	0.67	0.70	0.73
Gen. + Opt.	DecompOpt (40)	-8.10	-8.35	-8.98	**-9.01**	0.65	0.65	0.60	0.61
	TacoGFN (41)	-7.46	-7.15	-7.74	-7.35	**0.79**	**0.80**	0.56	0.56
	ALIDiff (42)	-8.47	-8.21	-8.90	-8.81	0.57	0.56	0.73	0.71
	THEMOL	**-8.63**	-8.36	**-9.12**	-8.80	0.75	0.77	**0.85**	**0.84**

THEMOL achieves the strongest predicted binding affinity, with the best average SMINA and Vina Dock scores, while maintaining competitive synthetic accessibility and the highest molecular diversity. This balance is important for structure-based design: improving docking scores is less useful if the search collapses to a narrow family of similar molecules. Compared with baselines that trade affinity or synthetic accessibility against diversity, THEMOL improves binding-oriented objectives while preserving broad chemical exploration.

Together, these benchmarks show that the main advantage of THEMOL is not isolated task-specific superiority, but cross-task stability of the same canonical latent space. The representation remains predictive for scalar molecular properties, realistic enough for 3D generation and structured enough for pocket-conditioned optimization. This pattern supports the central premise of THEMOL: removing global pose before representation learning yields a reusable molecular latent space that can be deployed across tasks with different geometric requirements.

2 Discussion

This work presents THEMOL as a canonicalization-based framework for reusable 3D molecular representation learning. Rather than discarding global pose through invariance, preserving it throughout the latent space through equivariance, or suppressing it statistically through augmentation, THEMOL handles pose as an interface variable before representation learning. This separation allows a non-equivariant backbone to operate in a learned canonical frame while retaining the ability to recover

coordinate-level outputs through frame lifting. Our results suggest that this design reduces symmetry-induced trade-offs when a single latent representation is reused across scalar prediction, coordinate generation and structure-based optimization.

A key implication is that the quality of a molecular foundation representation cannot be judged solely by downstream benchmark scores. Reusable latents must also support modelling operations such as molecular comparison, decoding, interpolation and property-guided search. The representation-level diagnostics show that canonicalization suppresses rotation-induced nuisance variation, preserves chemically meaningful structure, improves local decodability and reduces gauge-like waste during optimization. Thus, canonicalization does not merely stabilize input coordinates; it reshapes the latent space into a more suitable substrate for molecular modelling.

THEMOL is not intended to replace all geometric inductive biases. Invariant representations remain effective for scalar prediction, equivariant models remain essential for force fields, vector fields and coordinate denoising, and rotation augmentation remains a useful empirical strategy. The advantage of canonicalization is most relevant in the foundation-representation setting, where the same frozen or reusable latent must support tasks with different symmetry requirements. In this setting, explicitly separating global pose from molecular variation provides a practical route to cross-task reuse.

Several limitations remain. Learned canonical frames can be ambiguous for highly symmetric, near-degenerate or highly flexible molecules, and the present study focuses mainly on small-molecule modelling. Extending this principle to peptides, proteins and protein–ligand complexes will require handling larger conformational variability and multiple interacting coordinate frames. In addition, structure-based optimization is evaluated using docking-based proxies, which do not guarantee experimental binding affinity or selectivity. Future work should therefore combine canonical representation learning with multimodal molecular data, physics-based refinement and prospective experimental validation. More broadly, our results support canonicalization as a design principle for molecular foundation models: global pose should be handled explicitly at the representation interface, rather than being discarded from, or entangled within, the reusable molecular latent space.

3 Methods

3.1 Datasets

We use separate datasets for representation pretraining, representation-quality analysis and downstream benchmark evaluation. For pretraining, we train the canonical autoencoder on a large-scale 3D molecular corpus containing approximately 19 million molecules and 209 million conformations. Following the original data construction protocol (25), each molecule is associated with 11 RDKit-generated conformations (43), including one flattened structure. This corpus is used only for self-supervised representation learning, and no downstream property labels or protein-pocket information are used during pretraining.

We evaluate the learned representation across three drug-discovery task families. For molecular property prediction, we use 9 classification datasets from MoleculeNet (20) and 22 ADMET endpoints from the Therapeutics Data Commons benchmark (44). These tasks test whether the frozen canonical latent $\mathbf{X_z}$ retains molecule-level biochemical, pharmacokinetic and toxicity-relevant information.

For unconditional 3D molecule generation, we evaluate on GEOM-DRUG (45) under two established protocols. The first follows the evaluation protocol used by EDM-based 3D molecule generation studies (46), focusing on low-energy conformers and reporting atom stability and molecular validity. The second adopts the evaluation setting used by MiDi (47) and FlowMol3 (35), which uses GFN2-xTB local-minimum conformers with explicit hydrogens and topology-aware preprocessing, and evaluates physical and chemical validity using stricter geometry checks.

3.2 Overview of THEMOL

THEMOL is a canonicalization-based framework for reusable 3D molecular representation learning. Given atomic coordinates $\mathbf{x} \in \mathbb{R}^{N\times 3}$ and atom-level features $\mathbf{h} \in \mathbb{R}^{N\times d}$, the model first removes global translation, maps the molecule to a learned canonical frame, and then encodes the canonicalized geometry with a non-equivariant transformer autoencoder. The pretrained latent tensor is reused for three downstream modes: molecular property prediction, latent-space 3D molecule generation and structure-based molecular optimization.

We use row-wise coordinates. For $R \in \mathrm{SO}(3)$, a global rotation acts as $\mathbf{x} \mapsto \mathbf{x}R^\top$. Translation is removed by centering

$$\mathbf{x}^{\mathrm{c}} = \mathbf{x} - \mathbf{1}\bar{\mathbf{x}}^\top, \qquad \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i. \tag{1}$$

Variable-size molecules are handled by padding to $N_{\max}$. We denote by $\Omega \subseteq \{1,\dots,N_{\max}\}$ the valid atom mask, set padded rows to zero, and evaluate token-level losses only on Ω. Unless otherwise specified, the latent width is fixed to $H = 64$, so the reusable latent representation is $\mathbf{X_z} \in \mathbb{R}^{N_{\max}\times H}$. Dataset construction, preprocessing and hyperparameters are provided in Supplementary Note J and Supplementary Note H.1.

3.3 Learned canonicalization operator

The canonicalization module predicts a rotation selector $\widehat{R}_\theta(\mathbf{x},\mathbf{h}) \in \mathrm{SO}(3)$. The selector is obtained by predicting an unconstrained matrix $M_\theta(\mathbf{x},\mathbf{h}) \in \mathbb{R}^{3\times 3}$ and projecting it to the closest proper rotation:

$$\widehat{R}_\theta(\mathbf{x},\mathbf{h}) = \mathrm{SVD}^+(M_\theta(\mathbf{x},\mathbf{h})), \qquad \mathrm{SVD}^+(M) = U\,\mathrm{diag}(1,1,\det(UV^\top))V^\top, \tag{2}$$

where $M = U\Sigma V^\top$. This is the orthogonal Procrustes projection onto $\mathrm{SO}(3)$. The canonicalization map is

$$A_\theta(\mathbf{x},\mathbf{h}) = \mathbf{x}^{\mathrm{c}}\widehat{R}_\theta(\mathbf{x},\mathbf{h})^\top. \tag{3}$$

The intended gauge-consistency condition is

$$\widehat{R}_\theta(\mathbf{x}R^\top,\mathbf{h}) = \widehat{R}_\theta(\mathbf{x},\mathbf{h})R^\top \quad \Longrightarrow \quad A_\theta(\mathbf{x}R^\top,\mathbf{h}) = A_\theta(\mathbf{x},\mathbf{h}). \tag{4}$$

Thus, rotated copies of the same molecule are mapped to the same canonical representative before they are processed by the representation backbone.

The matrix predictor M_θ is implemented as a shallow pair-biased transformer. Initial atom tokens are built from centered coordinates and atom features, and pair biases are built from Gaussian basis features of pairwise distances and bond types. The rotation encoder ζ_{rot} processes these tokens, a mask-aware pooling operation produces a molecular summary vector, and an MLP head outputs nine scalars that are reshaped into M_θ:

$$\mathbf{X}_{\mathrm{rot}} = \zeta_{\mathrm{rot}}(\mathbf{X}^{(0)},\mathbf{B}^{(0)}), \qquad M_\theta(\mathbf{x},\mathbf{h}) = \mathrm{Reshape}_{3\times 3}\left(\mathrm{FFN}_{\mathrm{rot}}\left(\mathrm{Pool}_\Omega(\mathbf{X}_{\mathrm{rot}})\right)\right). \tag{5}$$

Because no unique ground-truth canonical orientation exists for a general molecule, $\widehat{R}_\theta$ is not supervised against a fixed reference frame. It is learned jointly with the autoencoder using reconstruction losses and the gauge-consistency regularizer

$$\mathscr{L}_{\mathrm{gc}} = \mathbb{E}_{R\sim\mu}\left[\left\|A_\theta(\mathbf{x}R^\top,\mathbf{h}) - A_\theta(\mathbf{x},\mathbf{h})\right\|_{F,\Omega}^2\right], \tag{6}$$

where μ is the Haar measure on $\mathrm{SO}(3)$. The canonical frame is therefore a learned gauge that co-adapts with the backbone, rather than a fixed PCA- or inertia-based alignment.

3.4 Canonical autoencoder pretraining

The pretrained model consists of the canonicalization encoder, a main encoder and a decoder. The canonicalization encoder predicts $\widehat{R}_\theta$. The main encoder and decoder operate in the canonical coordinate system.

Given canonical coordinates $\mathbf{x}^\star = A_\theta(\mathbf{x},\mathbf{h})$, the main encoder constructs atom-token embeddings and pair biases as

$$\mathbf{X}^{(0)} = \mathrm{Linear}_x(\mathbf{x}^\star) + \mathrm{Linear}_h(\mathbf{h}), \tag{7}$$

$$\mathbf{B}_{ij}^{(0)} = \mathrm{Linear}_b\left(\mathrm{GBF}\left(\left\|\mathbf{x}_i^\star - \mathbf{x}_j^\star\right\|_2, e_{ij}\right)\right), \tag{8}$$

where e_{ij} denotes the bond or edge type. Pair-biased self-attention blocks produce contextual atom representations $\mathbf{X}_{\mathrm{Enc}} = \zeta_{\mathrm{Enc}}(\mathbf{X}^{(0)},\mathbf{B}^{(0)})$. The variational encoder outputs per-token Gaussian parameters,

$$\mu = \mathrm{Linear}_\mu(\mathbf{X}_{\mathrm{Enc}}), \qquad \log\sigma^2 = \mathrm{Linear}_\sigma(\mathbf{X}_{\mathrm{Enc}}), \qquad \mathbf{X}_{\mathbf{z}} = \mu + \sigma\odot\varepsilon, \quad \varepsilon\sim\mathscr{N}(\mathbf{0},\mathbf{I}). \tag{9}$$

The latent tensor $\mathbf{X}_{\mathbf{z}} \in \mathbb{R}^{N_{\max}\times 64}$ is the representation used in all downstream modules. During deterministic representation extraction, we use the posterior mean $\boldsymbol{\mu}$ and denote it by $\mathbf{X}_{\mathbf{z}}$ for simplicity. The decoder ζ_{Dec} maps $\mathbf{X}_{\mathbf{z}}$ to decoder tokens, from which coordinate, atom-type, bond-type and distance heads reconstruct the molecule, as follow: $\mathbf{X}_{\mathrm{Final}} = \zeta_{\mathrm{Dec}}(\mathbf{X}_{\mathbf{z}})$.

Pretraining is formulated as denoising reconstruction in the learned canonical frame. Given the perturbed canonical input $(\widetilde{\mathbf{x}}^\star,\widetilde{\mathbf{h}})$, the autoencoder produces

$$\widehat{\mathbf{X}}_{\mathbf{z}} = \zeta_{\mathrm{Enc}}(\widetilde{\mathbf{x}}^\star,\widetilde{\mathbf{h}}), \qquad \widehat{\mathbf{X}}_{\mathrm{Dec}} = \zeta_{\mathrm{Dec}}(\widehat{\mathbf{X}}_{\mathbf{z}}). \tag{10}$$

The reconstruction heads predict the clean molecule in the same canonical frame,

$$\widehat{\mathbf{x}}^\star = \pi_{\mathrm{coord}}(\widehat{\mathbf{X}}_{\mathrm{Dec}}), \qquad \widehat{\mathbf{h}} = \pi_{\mathrm{type}}(\widehat{\mathbf{X}}_{\mathrm{Dec}}), \qquad \widehat{\mathbf{e}} = \pi_{\mathrm{bond}}(\widehat{\mathbf{X}}_{\mathrm{Dec}}).$$

The reconstruction losses are computed against the clean ground-truth molecule after applying the same selected rotation:

$$\mathscr{L}_{\text{Coord}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \|\widehat{\mathbf{x}}_i^{\star} - \mathbf{x}_i^{\star}\|_2^2, \tag{11}$$

Thus, the model does not compare the corrupted coordinates with the clean coordinates directly. The corrupted, canonicalized molecule is used as the autoencoder input, and the loss supervises the decoder to reconstruct the clean coordinates in the same learned canonical frame. The full pretraining objective is

$$\mathscr{L}_{\text{pre}} = \lambda_{\text{coord}}\mathscr{L}_{\text{Coord}} + \lambda_{\text{type}}\mathscr{L}_{\text{Type}} + \lambda_{\text{bond}}\mathscr{L}_{\text{Bond}} + \lambda_{\text{dist}}\mathscr{L}_{\text{Dist}} + \lambda_{\text{kld}}\mathscr{L}_{\text{KLD}} + \lambda_{\text{gc}}\mathscr{L}_{\text{gc}}. \tag{12}$$

Here $\mathscr{L}_{\text{Type}}$ and $\mathscr{L}_{\text{Bond}}$ are cross-entropy losses for atom and bond reconstruction, $\mathscr{L}_{\text{Dist}}$ reconstructs pairwise distances, and $\mathscr{L}_{\text{KLD}}$ regularizes the variational posterior. Loss weights, perturbation schemes and architecture hyperparameters are reported in Supplementary Note H.1; additional discussion of the coordinate loss is provided in Supplementary Note H.5.

3.5 Property prediction from frozen encoder latents

For molecular property prediction, we use the frozen pretrained encoder as a representation extractor. For each molecule, the canonicalization module and encoder produce $\mathbf{X}_{\mathbf{z}}$, and only a lightweight prediction head is trained on top of this latent tensor. We do not introduce a special coordinate-free `[CLS]` atom, because such a token has no natural 3D position in a molecular coordinate representation.

The molecular readout is obtained by attention pooling over valid atom tokens. The pooled vector $\mathbf{g}$ is passed to task-specific classification or regression heads. Classification tasks use binary cross-entropy or multi-label binary cross-entropy, and regression tasks use smooth L1 loss with target normalization. The backbone remains frozen in these experiments, so benchmark performance reflects the quality of the learned representation rather than end-to-end adaptation. Hyperparameter search spaces and training protocols for MoleculeNet and TDC ADMET are reported in Supplementary Note H.3.

3.6 Latent flow matching for 3D molecule generation

For molecule generation, the flow model generates latent tensors $\mathbf{X}_{\mathbf{z}}$, not final coordinates directly. After pretraining, we freeze the autoencoder and encode training molecules into $\mathbf{X}_{\mathbf{z}} \in \mathbb{R}^{N_{\max} \times 64}$. This defines the empirical target distribution q_1. The base distribution q_0 is an isotropic Gaussian over valid latent entries, with padded rows fixed to zero. For variable-size generation, we first sample an atom count N from the empirical size prior, construct the corresponding mask Ω_N, and run the flow on the valid rows of the padded tensor.

We use optimal-transport conditional flow matching (OT-CFM) (48). Given data latents $\{\mathbf{X}_{\mathbf{z},1}^{(j)}\}_{j=1}^{B} \sim q_1$ and Gaussian sources $\{\mathbf{X}_{\mathbf{z},0}^{(i)}\}_{i=1}^{B} \sim q_0$, we compute a minibatch OT plan

$$\Pi^{\star} \in \arg\min_{\Pi \in \mathscr{U}} \sum_{i,j} \Pi_{ij} \left\| \mathbf{X}_{\mathbf{z},0}^{(i)} - \mathbf{X}_{\mathbf{z},1}^{(j)} \right\|_{F,\Omega}^2, \tag{13}$$

where $\mathscr{U}$ is the set of doubly stochastic matrices. Paired endpoints $(\mathbf{X}_{\mathbf{z},0}, \mathbf{X}_{\mathbf{z},1})$ are sampled according to $\Pi^{\star}$. For $t \sim \text{Unif}[0, 1-\varepsilon]$, the conditional path is

$$\mathbf{X}_{\mathbf{z},t} = (1-t)\mathbf{X}_{\mathbf{z},0} + t\mathbf{X}_{\mathbf{z},1} + \sigma\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathscr{N}(\mathbf{0}, \mathbf{I}), \tag{14}$$

with $\sigma = 0$ unless otherwise stated. Rather than directly regressing the velocity, the flow network predicts the clean target latent:

$$\widehat{\mathbf{X}}_{\mathbf{z},1} = f_{\phi}(t, \mathbf{X}_{\mathbf{z},t}), \qquad v_{\phi}(t, \mathbf{X}_{\mathbf{z},t}) = \frac{\widehat{\mathbf{X}}_{\mathbf{z},1} - \mathbf{X}_{\mathbf{z},t}}{1-t}. \tag{15}$$

The training objective is

$$\mathscr{L}_{\text{FM}} = \mathbb{E}_{(\mathbf{X}_{\mathbf{z},0}, \mathbf{X}_{\mathbf{z},1}) \sim \Pi^{\star}, t} \left[w(t) \left\| f_{\phi}(t, \mathbf{X}_{\mathbf{z},t}) - \mathbf{X}_{\mathbf{z},1} \right\|_{F,\Omega}^2 \right], \; w(t) = \min\{(1-t)^{-2}, w_{\max}\}. \tag{16}$$

We parameterize f_{ϕ} with a Diffusion Transformer using timestep conditioning through adaptive layer normalization with zero initialization. During sampling, we draw $\mathbf{X}_{\mathbf{z},0} \sim q_0$, integrate $d\mathbf{X}_{\mathbf{z},t}/dt = v_{\phi}(t, \mathbf{X}_{\mathbf{z},t})$, and decode the terminal latent with the frozen decoder. The decoder predicts canonical coordinates, atom types and bond types; the molecular graph is assembled from the predicted distributions and sanitized before evaluation. Generated coordinates are in the learned canonical frame. When an arbitrary global orientation is required, we apply uniform lifting by sampling $R \sim \mu$ and returning $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^{\star} R^{\top}$. Additional details on OT-CFM training, DiT architecture, self-conditioning and numerical integration are provided in Supplementary Note K.

3.7 Structure-based optimization by fixed-size latent search

For structure-based design, we use the pretrained flow generator as a frozen molecular prior and introduce target specificity only through black-box search over its initial latent distribution. The flow model, decoder and encoder are not retrained. For each target pocket $\mathscr{P}$, we fix the atom count N. The search space is therefore the initial latent tensor $\mathbf{X}_{\mathbf{z},0} \in \mathbb{R}^{N\times 64}$, or its padded representation in $\mathbb{R}^{N_{\max}\times 64}$. Thus, optimization changes where the generator starts in latent space, but not the molecular size, flow parameters or decoder parameters.

Let Φ_ϕ be the frozen flow map and ζ_{Dec} be the frozen molecular decoder. We parameterize the initial-state distribution as $p_\psi = \mathscr{N}(\mathbf{m}, \sigma^2\mathbf{C})$ over the valid latent entries and maximize

$$\max_{\psi} J(\psi) = \mathbb{E}_{\mathbf{X}_{\mathbf{z},0}\sim p_\psi}\left[\mathscr{F}\left(\zeta_{\mathrm{Dec}}\left(\Phi_\phi(\mathbf{X}_{\mathbf{z},0})\right);\mathscr{P}\right)\right]. \tag{17}$$

The fitness function $\mathscr{F}$ combines docking affinity, synthetic accessibility and invalid-molecule penalties:

$$\mathscr{F}(\mathscr{M};\mathscr{P}) = -E_{\mathrm{dock}}(\mathscr{M};\mathscr{P}) + \lambda_{\mathrm{SA}} S_{\mathrm{SA}}(\mathscr{M}) - \lambda_{\mathrm{inv}}\mathbf{1}\{\mathscr{M}\ \text{invalid}\}. \tag{18}$$

Here E_{dock} is the docking energy, S_{SA} is the normalized synthetic-accessibility score, and invalid decoded molecules receive a fixed penalty. Diversity is not included in the optimization objective; it is evaluated separately as an outcome of searching within a pretrained generative prior.

Because docking, graph decoding and sanitization are non-differentiable, we optimize p_ψ using LRA-CMA-ES (49), following the derivative-free latent-search strategy used in evolutionary structure-based design (50). At each iteration, λ candidate initial states are sampled from p_ψ, passed through the frozen flow and decoder, docked against $\mathscr{P}$, scored by $\mathscr{F}$, and used to update $(\mathbf{m}, \sigma, \mathbf{C})$. After the final iteration, we sample molecules from the optimized distribution and return the top-scoring valid candidates. CMA-ES settings, docking settings and final sampling details are given in Supplementary Note L.

4 Data availability

All datasets used in this study are publicly available. The pretraining corpus of approximately 19 million molecules and 209 million conformations was obtained from the Uni-Mol release (25). Molecular property prediction was evaluated on MoleculeNet (20) (https://moleculenet.org) and on the ADMET benchmark of the Therapeutics Data Commons (44) (https://tdcommons.ai). Three-dimensional molecule generation was evaluated on GEOM-DRUG (45). Structure-based optimization used the CrossDocked2020 dataset, processed following the protocol of previous research (51). Representation-quality and latent-space diagnostics used the PDBbind v2020 database (52) (http://www.pdbbind.org.cn). The processed data splits and preprocessing scripts used to reproduce our results are available in the code repository listed under Code availability.

5 Code availability

The source code for THEMOL, including model implementation, training, generation and optimization pipelines, and evaluation scripts, is publicly available at https://github.com/themolsubmission/TheMol.

References

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

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

3. 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).

4. Schütt, K. *et al.* Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. neural information processing systems* **30** (2017).

5. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *International Conference on Learning Representations*.

6. Satorras, V. G., Hoogeboom, E. & Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, 9323–9332 (PMLR, 2021).

7. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059* (2023).

8. Vonessen, C., Harris, C., Cretu, M. & Liò, P. Tabasco: A fast, simplified model for molecular generation with improved physical quality. *arXiv preprint arXiv:2507.00899* (2025).

9. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).

10. Schoenberg, I. J. Remarks to maurice frechet's article "sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert". *Annals Math.* 724–732 (1935).

11. Adams, K., Pattanaik, L. & Coley, C. W. Chiro: Chirality-aware 3d molecule encoder. *arXiv preprint arXiv:2111.08638* (2021).

12. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478* (2021).

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

14. Joshi, C. K. *et al.* All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965* (2025).

15. Xu, Y. *et al.* Quotient-space diffusion models. *arXiv preprint arXiv:2604.21809* (2026).

16. Chakraborty, S. & Narayanan, H. Generating an equidistributed net on a sphere using random rotations. *Discret. & Comput. Geom.* **67**, 231–257 (2022).

17. Hewitt, J. & Liang, P. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, 2733–2743 (2019).

18. Papaspiliopoulos, O. High-dimensional probability: An introduction with applications in data science (2020).

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

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

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

22. 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).

23. 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).

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

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

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

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

28. 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).

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

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

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

32. 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).

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

34. 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).

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

36. 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).

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

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

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

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

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

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

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

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

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

46. 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).

47. 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).

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

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

50. Reidenbach, D. Evosbdd: Latent evolution for accurate and efficient structure-based drug design. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations* (2024).

51. 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).

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

6 Author contributions

S.C. conceived the study, developed the THEMOL framework, implemented the models and conducted the experiments. H.K., Y.K., S.L. and I.H. conducted theoretical analysis. S.P. supervised the project, acquired funding and provided critical revisions. S.C. wrote the manuscript with input from all authors. All authors reviewed and approved the final manuscript.

7 Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf