

MolPIF: a parameter interpolation flow model for molecule generation

Yaowei Jin^{1,†}, Junjie Wang^{1,2,†}, Yufan Tang³, Wenkai Xiang^{1, }, Duanhua Cao⁴, Dan Teng^{4, }, Zhehuan Fan⁴, Jiacheng Xiong⁴, Xia Sheng⁴, Chuanlong Zeng⁴, Duo An¹, Mingyue Zheng^{4, }, Shuangjia Zheng^{1,5}, Qian Shi^{1,*, }

¹Lingang Laboratory, Shanghai 200031, China

²School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

³College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China

⁴Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

⁵Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai 200240, China

*Corresponding author. Lingang Laboratory, 2380 Hechuan Road, Shanghai 200031, China. E-mail: shiqian@lglab.ac.cn

[†]These authors contributed equally to this work.

Associate Editor: Daisuke Kihara

Abstract

Motivation: Structure-based drug design (SBDD) has advanced with deep generative models, but bridging the gap between continuous atomic coordinates and discrete atom types remains a challenge. Current approaches, such as diffusion and flow matching models, often fail to unify these heterogeneous modalities, relying on separate strategies or ill-fitting Euclidean metrics for discrete variables. This lack of a consistent framework limits generative models' ability to capture the geometric and chemical structure of protein–ligand complexes.

Results: We present MolPIF, a parameter interpolation flow mechanism designed to unify the generation of continuous and discrete molecular variables. Unlike traditional flow models that operate in sample space, MolPIF interpolates between distributions in the parameter space, theoretically recovering Wasserstein-2 optimal transport for continuous coordinates and establishing Fisher–Rao geodesics for discrete atom types. We further incorporate a geometry-enhanced learning strategy to improve the capture of atomic contexts. Extensive evaluations on the CrossDocked2020 dataset demonstrate that MolPIF outperforms baselines in binding affinity, chemical validity, geometric fidelity, and chemical space coverage. Additionally, MolPIF exhibits versatility in lead optimization and offers flexible prior distribution selection (such as Laplace), establishing a robust paradigm for SBDD.

Availability and implementation: Source code is freely available at <https://github.com/BLEACH366/MolPIF>.

1 Introduction

Computer-aided drug design (CADD) is pivotal to drug discovery, spanning from target validation to preclinical evaluation (Ragoza *et al.* 2022). While structure-based drug design (SBDD) has demonstrated remarkable effectiveness in identifying lead compounds (Batool *et al.* 2019), traditional CADD still struggles with imbalanced datasets and the exhaustive exploration of vast chemical and conformational spaces. To address this, AI-driven molecular generation—inspired by successes in AIGC (Guo *et al.* 2025, Liu *et al.* 2025, Tian *et al.* 2026)—has emerged (Zang and Wang 2020, Jiang *et al.* 2021). These frameworks extract deep chemical insights from crystallographic data to navigate uncharted chemical territories, significantly accelerating the discovery of novel molecular structures.

Three-dimensional (3D) generative models have significantly advanced SBDD by incorporating protein pocket constraints and enabling end-to-end automation (Xie *et al.* 2022). For SBDD tasks in 3D space, the mainstream approaches primarily leverage autoregressive or diffusion-based generative frameworks (Gu *et al.* 2024, Li *et al.* 2025). However, these frameworks face distinct limitations: autoregressive models often suffer from mode collapse on unordered molecular data (Han *et al.* 2025), while diffusion and flow matching (FM) models struggle with the multimodal nature of atomic features (Song and Ermon 2019, Ho *et al.* 2020, Lipman *et al.* 2023). Specifically, molecular structures comprise heterogeneous variables—discrete (atom types), integer (formal charges), and continuous (spatial coordinates)—that require fundamentally different modeling methods

Received: 10 February 2026. Revised: 14 April 2026. Accepted: 14 May 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

(Song *et al.* 2023). The prevailing strategy involves extending diffusion and FM models to discrete domains, using advanced modeling techniques in order to achieve ideal performance (Cremer *et al.* 2025). Since the introduction of Bayesian Flow Networks (BFNs) (Graves *et al.* 2023) attempt to bridge this gap via parameter-space updates, they still model modalities separately. Consequently, the absence of a theoretically unified generative framework results in the persistent adverse effects of heterogeneous modalities on the quality of generated molecules.

To effectively unify discrete and continuous modeling, we introduce Parameter Interpolation Flow (PIF). By treating data as a superposition of Diracs, PIF interpolates distributional parameters between the data and a flexible prior. The model is optimized via Kullback-Leibler (KL) divergence to predict time-dependent parameters, enabling iterative refinement from the prior during inference. By shifting trajectories from sample to parameter space, PIF is highly versatile for both continuous and discrete data, allowing flexible prior selection and adaptation to various tasks without the need for complex closed-form derivations. PIF overcomes classical FM limitations: recovers 2-Wasserstein(W_2)-optimal transport for Gaussian priors and extends this optimality to other distributions. For discrete data with an exponential family prior, PIF leverages Fisher–Rao geometry, ensuring probability paths evolve along exponential geodesics, accurately capturing the data manifold’s structure.

Molecular Parameter Interpolation Flow (MolPIF) extends the PIF framework to molecular generation by learning the parameter spaces associated with atomic coordinates (modeled as Gaussian distributions) and atomic types (modeled as Dirichlet distributions). During training, we incorporate a geometry-enhanced learning strategy, inspired by prior work (Cui *et al.* 2024, Peng *et al.* 2026), to provide atomic-level contextual information of ligands to the model. This approach involves randomly masking subsets of atoms during training, enabling the model to dynamically optimize arbitrary atomic arrangements within a given molecular structure. As a result, MolPIF achieves superior performance in overall quality of the generated molecules.

Empirical evaluations conducted on the CrossDocked2020 dataset (Francoeur *et al.* 2020) demonstrate that MolPIF has comprehensive advantages across five key dimensions: (i) Advanced *de novo* generation capability, producing molecules with enhanced binding properties and chemical validity; (ii) Accurate geometric reproduction of molecular structural distributions, including rings, bond lengths, and bond angles; (iii) Comprehensive chemical space modeling, with substantial coverage of 2D structural features and accurate 3D conformational distribution, while extending shape diversity beyond the reference distribution; (iv) Flexible adaptation to different prior distributions, where we systematically compared and analyzed the advantages and limitations of using Gaussian versus Laplace distributions as priors for modeling atomic coordinates; (v) Effective lead optimization, demonstrating robust performance in enhancing drug candidate properties.

2 Materials and methods

2.1 Definitions and notations

Molecule generation based on receptor structure can be formulated as a conditional generation task. The input is a protein

binding site $\mathcal{P} = \{(\mathbf{x}_p^{(i)}, \mathbf{v}_p^{(i)})\}_{i=1}^{N_p}$, which contains N_p atoms with each $\mathbf{x}_p^{(i)} \in \mathbb{R}^3$ and $\mathbf{v}_p^{(i)} \in \mathbb{R}^{D_p}$ correspond to atom coordinates and atom features such as element types and amino acid types, respectively. The target output is a binding molecule $\mathcal{M} = \{(\mathbf{x}_m^{(i)}, \mathbf{v}_m^{(i)})\}_{i=1}^{N_m}$, where N_m is the number of atoms in molecule, $\mathbf{x}_m^{(i)} \in \mathbb{R}^3$ and $\mathbf{v}_m^{(i)} \in \mathbb{R}^{D_m}$. For brevity, we denote $\mathbf{p} = [\mathbf{x}_p, \mathbf{v}_p]$ ($\mathbf{x}_p \in \mathbb{R}^{N_p \times 3}$, $\mathbf{v}_p \in \mathbb{R}^{N_p \times D_p}$) and $\mathbf{m} = [\mathbf{x}_m, \mathbf{v}_m]$ ($\mathbf{x}_m \in \mathbb{R}^{N_m \times 3}$, $\mathbf{v}_m \in \mathbb{R}^{N_m \times D_m}$) as the concatenation of protein binding site and ligand atoms.

2.2 Parameter interpolation flow

Parameter Interpolation Flow (PIF) is a flow-based generative model that operates in the parameter space of probability distributions. Unlike conventional flow models that transform samples directly, PIF constructs an interpolation path between the parameters of a prior distribution θ_{prior} and a target data-driven distribution $\theta(\mathbf{x}_{\text{data}})$. Specifically, it learns a time-dependent parameter trajectory $\theta_t = f(t)\theta(\mathbf{x}_{\text{data}}) + (1-f(t))\theta_{\text{prior}}$, where $f(t)$ is a monotonic function ensuring $\theta_0 = \theta_{\text{prior}}$ and $\theta_1 = \theta(\mathbf{x}_{\text{data}})$.

During training, PIF optimizes the model Φ to predict $\theta(\mathbf{x}_{\text{data}})$ from samples drawn at intermediate θ_t . The prediction accuracy is evaluated via the KL divergence between the predicted distribution $p(\mathbf{x} | \theta_{t+\Delta t})$ and the true interpolated distribution $p(\mathbf{x} | \theta_{t+\Delta t})$. Specifically, $\theta_{t+\Delta t}$ is the true next-step parameter, calculated as a time-dependent interpolation of the prior and target distributions (Appendices A and B, available as [supplementary data](#) at *Bioinformatics* online). At inference, PIF generates samples through an iterative refinement process: starting from θ_{prior} , the model progressively updates parameters until converging to $\theta(\mathbf{x}_{\text{data}})$ (Fig. 1).

In classical continuous FM framework, the probability path is typically prescribed via linear interpolation in the sample space. While the resulting path corresponds to the optimal transport in the W_2 under such linear interpolation when source and target distributions are Gaussian, it fails to maintain such optimality for general distributions. To overcome this, PIF shifts the interpolation mechanism from the sample space to the distribution parameter space, thereby aligning the induced probability path with the intrinsic geometric structure of the distribution space. For continuous variables, PIF characterizes the W_2 -optimal transport structure between any two 1D distributions of the same type, offering a more flexible and theoretically grounded framework.

The advantage of PIF is particularly pronounced in discrete variable modeling. Existing discrete FM approaches (Austin *et al.* 2021, Gat *et al.* 2024, Zhao *et al.* 2025) often rely on W_2 distance, which assume an Euclidean structure that categorical variables do not possess. Applying these metrics to the probability simplex is equivalent to imposing a mean-squared-error-type loss, which often fails to capture the true curvature of the manifold—especially near the boundaries—and may lead to distorted interpolation paths and implausible intermediate distributions. In contrast, PIF lifts the problem to a statistical manifold. Under an exponential family prior, interpolating distribution parameters, PIF defines probability paths along exponential geodesics under the Fisher–Rao metric.

This information-geometric approach offers several advantages: offers a more direct and robust statistical interpretation that better reflects relative changes in probability mass (Cheng *et al.*

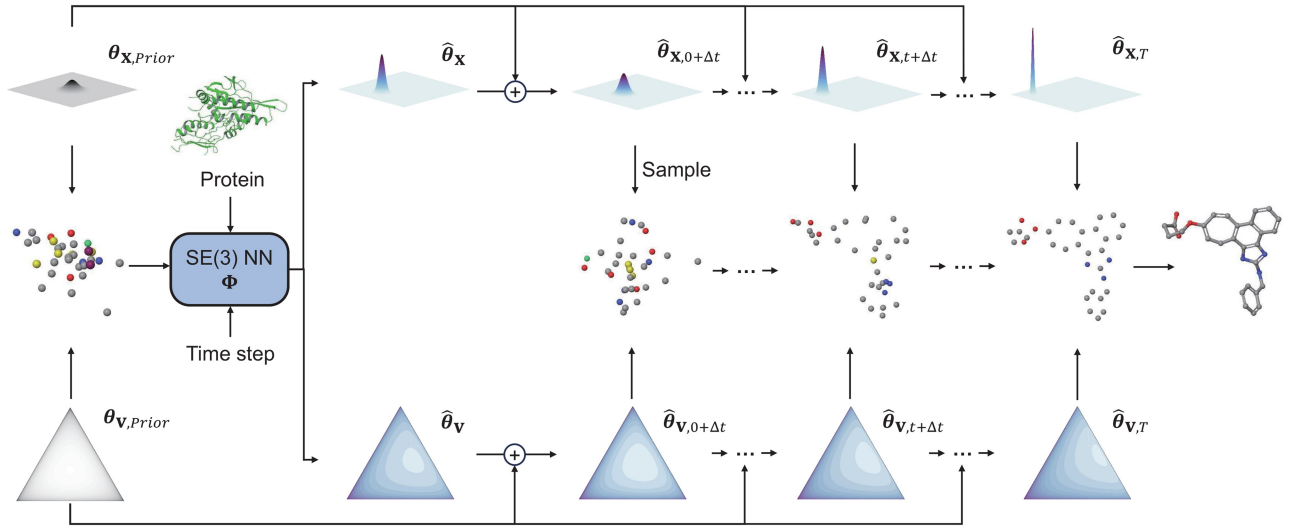


Figure 1 The sampling process of the MolPIF framework. Atomic coordinates $\mathbf{x}$ and types $\mathbf{v}$ are iteratively sampled and refined. An SE(3)-equivariant network updates distribution parameters by interpolating predictions with priors. At $t = 1$, final molecular structures are sampled from the predicted Dirac distributions and assembled via OpenBabel.

2024), ensures insensitivity to the specific choice of parameter representation, and is intrinsically linked to natural gradient dynamics on the statistical manifold, which contributes to improved training stability and efficiency (Boll et al. 2025, Cheng et al. 2024). Consequently, PIF provides a unified framework that recovers W_2 optimality in the continuous Gaussian case while establishing a metrically consistent and statistically natural foundation for discrete generative modeling.

2.3 Molecule generation based on PIF

For molecular generation via PIF, we specify the distributions for atomic coordinates $\mathbf{x}$ and types $\mathbf{v}$ as Gaussian and Dirichlet (Stark et al. 2024), respectively:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \varepsilon^2 \mathbf{I}) \quad (1)$$

$$p(\mathbf{v}) = \text{Dir}(\mathbf{v}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K v_i^{\alpha_i - 1} \quad (2)$$

where $\boldsymbol{\mu} \in \mathbb{R}^3$, ε is a scalar, and $\boldsymbol{\alpha}$ is a K -dimensional vector, where K corresponds to the number of atom types, $B(\boldsymbol{\alpha})$ is the multivariate beta function. Accordingly, the parameters of the two distributions are denoted as $\boldsymbol{\theta}_\mathbf{x} = (\boldsymbol{\mu}, \varepsilon)$ and $\boldsymbol{\theta}_\mathbf{v} = \boldsymbol{\alpha}$, respectively.

To obtain the parameters corresponding to molecular data, we represent them in the form of Dirac distributions associated with the aforementioned two distribution types. In practice, we extend the definition of Gaussian distributions by setting the standard deviation to zero in the distribution parameters of continuous-variable Dirac distributions:

$$p(\mathbf{x}|\mathbf{x}_{\text{data}}) = \lim_{\varepsilon \rightarrow 0^+} \mathcal{N}(\mathbf{x}; \mathbf{x}_{\text{data}}, \varepsilon^2 \mathbf{I}) \quad \boldsymbol{\theta}_{\mathbf{x}, \text{data}} = (\mathbf{x}_{\text{data}}, 0) \quad (3)$$

$$p(\mathbf{v}|\mathbf{v}_{\text{data}}) = \text{Dir}(\mathbf{v}; \mathbf{v}_{\text{data}}) \quad \boldsymbol{\theta}_{\mathbf{v}, \text{data}} = \text{Onehot}(\mathbf{v}_{\text{data}}) \quad (4)$$

The formulation of the interpolation process is given as follows:

$$\boldsymbol{\theta}_{\mathbf{x}, t} = f(t) \boldsymbol{\theta}_{\mathbf{x}, \text{data}} + (1 - f(t)) \boldsymbol{\theta}_{\mathbf{x}, \text{prior}} \quad (5)$$

$$\boldsymbol{\theta}_{\mathbf{v}, t} = f(t) \boldsymbol{\theta}_{\mathbf{v}, \text{data}} + (1 - f(t)) \boldsymbol{\theta}_{\mathbf{v}, \text{prior}} \quad (6)$$

$$\boldsymbol{\theta}_{\mathbf{x}, \text{prior}} = (\mathbf{0}, \varepsilon_0) \quad (7)$$

$$\boldsymbol{\theta}_{\mathbf{v}, \text{prior}} = (1/K, 1/K, \dots, 1/K) \quad (8)$$

$$f(t) = 1 - \gamma^t \quad (9)$$

In the above equation, both ε_0 and γ are hyperparameters. The formulation of $f(t)$ encourages the model to focus more on learning fine-grained structures in the molecular data, thereby improving the quality of generation.

The loss function is formulated as follows (Soch et al. 2024):

$$L_{t-\Delta t, \mathbf{x}} = \frac{(1 - \gamma^t)^2}{2\gamma^t \varepsilon_0^2} \mathbb{E}_{p_{\text{data}}} [\|\hat{\boldsymbol{\theta}}_\mathbf{x}(1) - \boldsymbol{\theta}_\mathbf{x}^{(1)}\|^2] \quad (10)$$

$$L_{t-\Delta t, \mathbf{v}} = \mathbb{E}_{p_{\text{data}}} \left[\ln \frac{\Gamma(\boldsymbol{\theta}_{\mathbf{v}, t})}{\Gamma(\hat{\boldsymbol{\theta}}_{\mathbf{v}, t})} + (\hat{\boldsymbol{\theta}}_{\mathbf{v}, t} - \boldsymbol{\theta}_{\mathbf{v}, t})^T (\psi(\hat{\boldsymbol{\theta}}_{\mathbf{v}, t}) - \psi(1)) \right] \quad (11)$$

Here, $\Gamma(\mathbf{x})$ is the multivariate gamma function, $\psi(\mathbf{x})$ is the multivariate digamma function, both $\lambda_\mathbf{x}$ and $\lambda_\mathbf{v}$ are hyperparameters to adjust the weight of loss.

By conditioning on fixed substructure coordinates and atom types, the model generates the remaining molecular structure:

$$\boldsymbol{\theta}_{\mathbf{x}, t, \text{cond}} = \text{Concat}(\boldsymbol{\theta}_{\mathbf{x}, t}, (\mathbf{x}_{\text{cond}}, 0)) \quad (12)$$

$$\boldsymbol{\theta}_{\mathbf{v}, t, \text{cond}} = \text{Concat}(\boldsymbol{\theta}_{\mathbf{v}, t}, \mathbf{v}_{\text{cond}}) \quad (13)$$

MolPIF using Laplace prior is provided in Appendix A, available as [supplementary data](#) at *Bioinformatics* online.

2.4 Geometry-enhanced learning strategy

To improve the capture of atomic-level contextual information, we use a geometry-enhanced learning strategy based on a dynamic masking mechanism. During training, ligand atoms are stochastically masked (with activation probability P_m and

per-atom probability P_{am}) and the model is tasked with reconstructing their coordinates and types. This strategy forces the network to act as a structural imputer, learning to respect stringent local geometric constraints such as bond lengths and ring planarity.

2.5 Implementation details

Atom Featurization: Protein atoms were one-hot encoded by element (H, C, N, O, S, Se) and residue identity, plus a backbone indicator and region tag (arm/scaffold). Ligand atoms were encoded by element (C, N, O, F, P, S, Cl) and aromaticity.

Model & Training: MolPIF's architecture uses the UniTransformer (Qu *et al.* 2024) for equivariance and spatial encoding (see Appendix A, available as [supplementary data](#) at *Bioinformatics* online for the network structure). Protein–ligand graphs use K-Nearest-Neighbor-based edges. We set $\varepsilon_0 = 1$, $\gamma = 0.009$, $\beta_0 = 1$, the probability of enabling the masking mechanism $P_m = 0.3$, the probability of each atom being masked $P_{am} = 0.3$, and used 100 sample steps. Training took 24 h on a single NVIDIA 4090 GPU. For lead optimization, priors were initialized from unfixed region coordinates and atom types.

3 Results

3.1 Experimental setup

We used the CrossDocked2020 dataset (Francoeur *et al.* 2020) for training and evaluation. Following the preprocessing protocol AR (Luo *et al.* 2021), we retained only protein–ligand poses with an RMSD < 1 Å relative to experimental structures. To ensure generalizability, protein sequences were clustered at 30% identity using MMseqs2 (Steinegger and Söding 2017), splitting the data into 99 900 training pairs and 100 test proteins from unseen clusters. For each test protein, 100 molecules were sampled for comprehensive assessment.

We evaluate five baselines: AR (Luo *et al.* 2021) generates molecules atom-by-atom via a Markov Chain Monte Carlo (Geyer 1992) method on density grids; Pocket2Mol (Peng *et al.* 2022) uses an auto-regressive scheme for 3D positions and bonds; TargetDiff (Guan *et al.* 2023a) uses continuous and discrete diffusion for simultaneous generation; DecompDiff (Guan *et al.* 2023b) incorporates scaffold priors and bond diffusion with validity guidance; and MolCRAFT (Qu *et al.* 2024) utilizes the BFNs to bridge the continuous-discrete gap.

The performance is assessed via three categories: (i) Binding Affinity: Calculated via AutoDock Vina (Trott and Olson 2010) using Vina Score (kcal/mol, direct affinity), Vina Min (kcal/mol, post-relaxation affinity), and Vina Dock (kcal/mol, post-redocking affinity). (ii) Chemical Properties: Evaluated by QED (Bickerton *et al.* 2012) (drug-likeness), SA (Ertl and Schuffenhauer 2009) (synthetic ease), LogP [lipophilicity, target: −0.4 to 5.6 (Ghose *et al.* 1999)], Lipinski score (Lipinski *et al.* 1997) (Ro5 compliance), and Diversity (Div, average pairwise Tanimoto distance). (iii) Conformation Stability: Measured by Strain Energy (SE, kcal/mol) (Gu *et al.* 2021), Clash Ratio (CR) via PoseCheck (Harris *et al.* 2023), the Jensen-Shannon divergence

(JSD) of bond lengths (JS_{BL}), and angles (JS_{BA}) to quantify structural deviation from references.

3.2 Model evaluation

3.2.1 Evaluation of common properties for generated molecules

Table 1 summarizes the performance of MolPIF compared to baseline models across 10 000 generated molecules. MolPIF consistently demonstrated competitive binding affinity, drug-likeness, and conformational stability. Regarding binding affinity, under size-constrained conditions, MolPIF achieved the lowest mean values in Vina Score (−6.64), Vina Min (−7.41), and Vina Dock (−8.09), outperforming the best AR and TargetDiff baselines by 15.48% and 21.39%, respectively. A high Vina Score/Dock ratio (0.82) indicates strong structural consistency between initial generated conformations and docked poses. In terms of chemical properties, MolPIF exhibited superior drug-likeness, yielding the highest QED (0.59) and competitive SA (0.72) and Lipinski compliance. Furthermore, MolPIF showed excellent conformational stability, ranking first in the 25th and 50th percentiles of SE. Low JS values further confirm its accuracy in capturing local structural distributions. Case study of *de novo* task is given in Appendix C, available as [supplementary data](#) at *Bioinformatics* online.

3.2.2 Analysis of MolPIF on local geometries

We evaluated the local geometry of generated molecules by analyzing ring size frequencies (3–8 membered) and the Jensen-Shannon divergence (JSD) of bond lengths, angles, and torsions (see Appendix G, available as [supplementary data](#) at *Bioinformatics* online for JSD calculation). As shown in Table D1, available as [supplementary data](#) at *Bioinformatics* online, MolPIF effectively avoids unstable small rings (zero 3-membered and only 0.44% 4-membered rings), while predominantly producing 6-membered rings (76.88%)—closely matching the reference distribution. In contrast, AR models overproduce small rings, while diffusion-based models struggle to capture the prevalence of 6-membered rings. JSD analysis further confirms that parametric models, particularly MolPIF and MolCRAFT, achieve the lowest divergence from reference distributions, demonstrating superior capability in capturing fine-grained geometric features. Detailed JSD metrics are discussed in Section 3.2.4.

3.2.3 Analysis of MolPIF on chemical space distribution

To evaluate MolPIF's performance macroscopically, we analyzed its chemical space distribution using 2D (ECFP, RDKit) and 3D [USRCAT (Schreyer and Blundell 2012)] descriptors (Fig. 2a–c). While MolPIF-generated molecules comprehensively cover the 2D substructure space of the test set, they exhibit even more precise density alignment in 3D conformational space. This accurate reproduction of high-density regions indicates that MolPIF effectively captures authentic spatial distributions.

Beyond fingerprints, we characterized 3D geometry using Principal Moments of Inertia [PMI (Sauer and Schwarz 2003)] and Plane of Best Fit [PBF (Firth *et al.* 2012)]. The Normalized Principal Moment of Inertia (NPM) ternary plot (Fig. 2d and e) shows that MolPIF not only matches the test set's concentration

Table 1 Comparison of MolPIF and baseline models on the CrossDock test set across 10 000 generated molecules for *de novo* design.^a

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		Strain Energy (↓)			QED	SA	LogP	Lipinski	Div	JS (↓)		CR	Connected	
	Avg.	Med.	Avg.	Med.	Avg.	Med.	25%	50%	75%						BL	BA		Avg. (↓)	Avg. (↑)
Reference	-6.36	-6.46	-6.71	-6.49	-7.45	-7.26	34	107	196	0.48	0.73	0.89	4.27				0.17		
AR	-5.75	-5.64	-6.18	-5.88	-6.75	-6.62	259	595	2286	0.51	0.64	0.39	4.75	0.7	0.45	0.54	0.22		0.94
Pocket2Mol	-5.14	-4.70	-6.42	-5.82	-7.15	-6.79	102	189	374	<u>0.57</u>	0.76	1.51	4.88	0.74	0.37	0.43	0.56		0.96
TargetDiff	-5.47	-6.30	-6.64	-6.83	-7.80	-7.91	369	1243	13871	0.48	0.58	1.36	4.51	0.72	<u>0.26</u>	0.48	0.53		0.90
DecompDiff	-5.19	-5.27	-6.03	-6.00	-7.03	-7.16	115	421	1424	0.51	0.66	1.15	4.49	0.73	<u>0.26</u>	0.44	0.51		0.83
MolCRAFT	-6.55	-6.95	-7.21	-7.14	-7.67	-7.82	83	195	510	0.50	0.67	1.16	4.46	<u>0.73</u>	0.23	0.37	<u>0.26</u>		<u>0.97</u>
MolPIF	-6.64	-7.02	-7.41	-7.28	-8.09	-8.13	65	150	<u>375</u>	0.59	<u>0.72</u>	3.26	4.63	0.72	0.23	0.40	0.29		0.98

^a (↑)/(↓) indicates larger/smaller is better. Top-2 results are highlighted with bold and underlined.

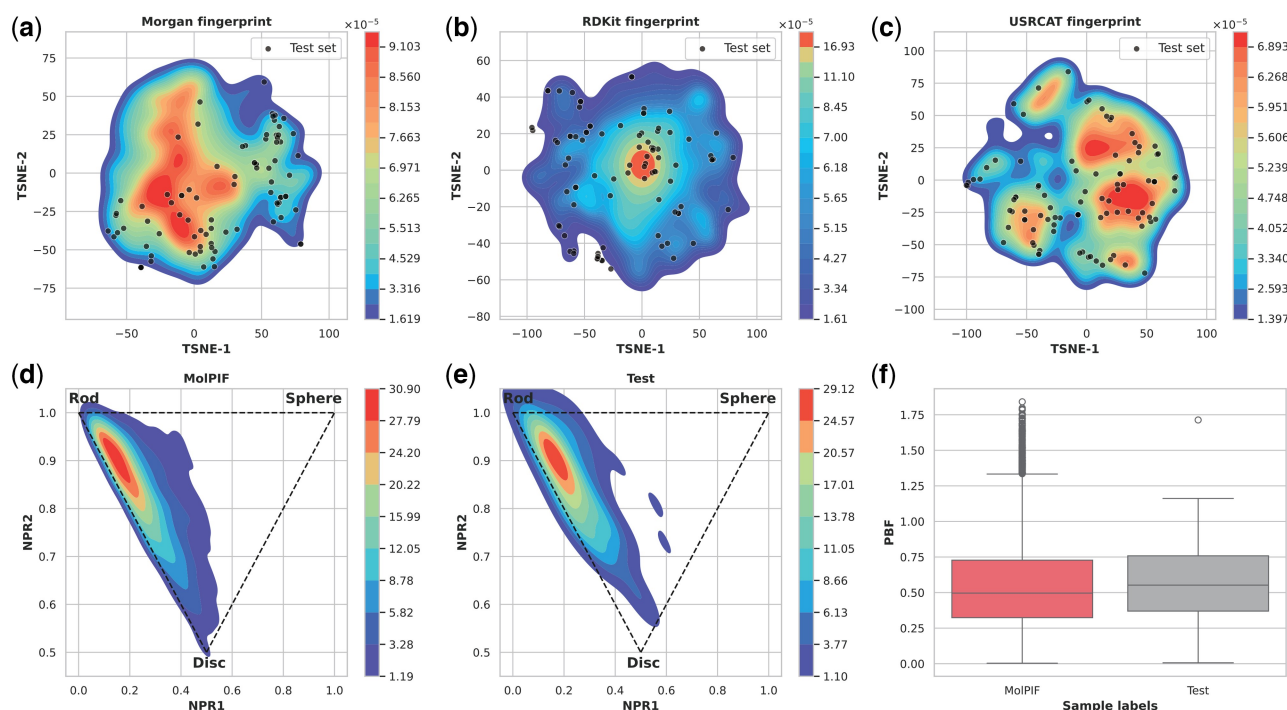


Figure 2 Chemical space and molecular shape distribution. (a–c) t-SNE visualization based on Morgan (a), RDKit (b), and USRCAT (c) fingerprints. (d, e) NPR shape distributions of generated (d) and reference (e) molecules. (f) PBF descriptor comparison between generated ($n = 10\,000$) and reference ($n = 100$) sets. Box plots indicate median, IQR, $1.5 \times \text{IQR}$ whiskers, and outliers.

near rod-shaped morphologies but also explores underrepresented disc- and sphere-shaped regions. Furthermore, the PBF distribution (Fig. 2f) shows strong agreement between MolPIF-generated molecules and the test set, confirming the model's ability to reproduce spatial planarity. Overall, MolPIF balances fidelity to structural trends with the exploration of novel molecular shapes.

3.2.4 Analysis of the prior distribution selection

We explored the flexibility of prior distributions by replacing the Gaussian distribution in MolPIF with a Laplace distribution [MolPIF(La)]. As shown in Table 2, while the Gaussian prior generally yields superior binding affinities (e.g. higher Vina scores), the Laplace prior excels in chemical properties (QED, LogP), molecular diversity (0.75), and substructure accuracy. Notably, MolPIF(La) achieved the lowest JSD in bond length distributions and captured distinct modes that other models missed (Fig. D1, available as [supplementary data](#) at *Bioinformatics* online, Tables D2–D4, available as [supplementary data](#) at *Bioinformatics* online). The mask module exhibited divergent effects: in the Gaussian model, it prioritized conventional properties over substructure precision; conversely, in the Laplace model, it significantly enhanced both binding metrics and substructure generation. Despite these trade-offs, MolPIF(La) still outperformed most baselines in Table 1.

These phenomena stem from the interaction between the prior's inductive bias and the mask module. While the Gaussian prior favors global smoothness, it conflicts with the local discontinuities introduced by masking, leading to over-smoothed

reconstructions. In contrast, the Laplace prior, with its heavy tails, aligns with the local sparsity of molecular substructures. This synergy enhances the model's ability to reconstruct masked regions with high fidelity. Consequently, MolPIF with a Laplace prior achieves superior structural realism compared to the Gaussian baseline. Furthermore, MolPIF maintains strong performance with non-conventional priors without the complex derivations required by diffusion models, offering significant flexibility for SBDD tasks.

3.2.5 Generalization ability of MolPIF

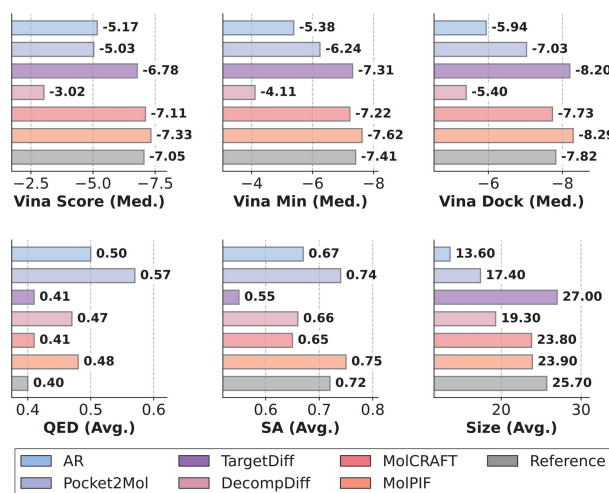
To rigorously evaluate model generalization, we adopt PoseBusters (Buttenschoen *et al.* 2024) as a challenging out-of-distribution (OOD) test set. Unlike the widely used CrossDock dataset, which contains noisy software-generated poses, PoseBusters consists exclusively of high-quality, wet-lab crystal structures released since 2021. To prevent information leakage, we follow (Qiu *et al.* 2025) and use MMseqs2 to discard any test proteins sharing $>30\%$ sequence identity with the training set, resulting in a rigorously curated subset of 180 test proteins. For each test protein, we generate 100 molecules using MolPIF and baseline models.

As shown in Fig. 3, MolPIF maintains robust performance within this OOD setting. Regarding binding affinity, drug-likeness, and synthetic accessibility, the molecules generated by MolPIF even surpass the reference molecules. This level of performance was not achieved by any of the baselines in the OOD scenario, making MolPIF the sole model to do so in this experiment. Furthermore, the molecular size data confirms that the

Table 2 Comparison of MolPIF variants on the CrossDock test set across 10 000 generated molecules for *de novo* design.^a

Methods	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		Strain Energy (↓)			QED	SA	LogP		Lipinski Div		JS (↓)		CR
	Avg.	Med.	Avg.	Med.	Avg.	Med.	25%	50%	75%			Avg. (↑)	Avg. (↑)	Avg. (↑)	Avg. (↑)	BL	BA	
MolPIF	-6.64	-7.02	-7.41	-7.28	-8.09	-8.13	65	150	375	0.59	0.72	3.26	<u>4.63</u>	0.72	0.23	<u>0.40</u>	<u>0.29</u>	
MolPIF (w/o mask)	-6.78	-6.99	-7.28	-7.18	-7.90	-7.95	<u>73</u>	<u>173</u>	467	0.55	0.70	2.28	4.49	0.72	<u>0.20</u>	0.39	0.24	
MolPIF(La)	-5.17	-6.02	-6.43	-6.70	-7.64	-7.76	77	203	691	0.55	0.70	2.45	4.51	<u>0.73</u>	0.16	0.40	0.39	
MolPIF (La w/o mask)	-4.75	-6.10	-6.28	-6.68	-7.84	-7.90	81	183	<u>417</u>	<u>0.56</u>	0.72	3.42	4.65	0.75	0.26	0.47	0.36	

^a (↑)/(↓) indicates larger/smaller is better. Top-2 results are highlighted with bold and underlined.


Figure 3 Comparison of MolPIF and baseline models on the PoseBusters OOD test set across 18 000 generated molecules for *de novo* design.

superior binding affinity is driven by more rational structural design rather than a mere increase in atom count. These results demonstrate the strong generalization ability of MolPIF, which effectively captures underlying structure-activity relationships and generates high-quality molecules even in challenging OOD scenarios.

3.2.6 Ablation study

Table 3 illustrates the effectiveness of each component within the MolPIF framework. We evaluate the choice of the time-dependent function $f(t)$ within our formulation $\theta_t = f(t)\theta_{data} + (1 - f(t))\theta_{prior}$. Comparing linear ($f(t) = t$), quadratic ($f(t) = t^2$), and exponential ($f(t) = 1 - \gamma^t$) schedules, empirical results show that $f(t) = 1 - \gamma^t$ yields the best performance. This suggests that a convex $f(t)$ is advantageous, as it reduces the proportion of the high-noise stage and helps the model capture more accurate molecular structures. Furthermore, a comparison of different decay rates ($\gamma \in \{0.005, 0.009, 0.02\}$) confirms that $\gamma = 0.009$ is the optimal setting for our model.

To isolate the improvements introduced by PIF, we compare it against standard flow matching, which typically uses a

Gaussian distribution with linear interpolation for atomic coordinates and a categorical distribution with linear interpolation for atom types. Using the identical UniTransformer architecture and training data, PIF clearly outperforms the standard flow matching baseline. Additionally, comparative experiments on toy datasets (see Appendix B, available as supplementary data at Bioinformatics online) consistently demonstrate PIF's superiority over common generative frameworks under strictly controlled settings, further validating our architectural design.

3.2.7 The performance of MolPIF in lead optimization

MolPIF enables atom-level lead optimization by allowing users to fix specific atoms while generating complementary substituent regions. We validated this capability using targets 1umd, 3ZCW (Alexandar *et al.* 2022), and 6KZZ (Ushiyama *et al.* 2020). For 1umd (Fig. 4a), following the CBGBench (Lin *et al.* 2025) protocol (see Appendix F, available as supplementary data at Bioinformatics online), MolPIF demonstrated versatility in both extending scaffold groups (fragment/side-chain modification) and integrating discrete fragments (linker/scaffold operations). This atom-level control facilitates the rapid modification of physicochemical properties. For 3ZCW, we adopted the same side-chain fixation strategy as Delete (Chen *et al.* 2025). Alternatively, for 6KZZ we relaxed the atomic constraints compared to DeepFrag (Green and Durrant 2021) to better demonstrate lead optimization potential. 100 optimized molecules have been generated for each target. In scaffold hopping tasks for 3ZCW (Fig. 4b), 25.58% of MolPIF-optimized ligands surpassed the reference Vina score (-9.93). In fragment growth tasks for 6KZZ (Fig. 4c), 63.33% of generated molecules outperformed the reference score (-8.22). Across all subtasks, MolPIF consistently produced a higher proportion of molecules with superior binding affinity and 3D spatial compatibility compared to Delete and DeepFrag (Tables D5 and D6, available as supplementary data at Bioinformatics online). These results underscore MolPIF's broad applicability and potential to enhance lead optimization.

4 Conclusion

We introduce Parameter Interpolation Flow (PIF), a novel generative framework that operates in the parameter space of

Table 3 Ablation studies of MolPIF on the CrossDock test set across 10 000 generated molecules for *de novo* design.^a

Methods		$f(t)=t$	$f(t)=t^2$	$\gamma=0.02$	$\gamma=0.005$	F.M.	Best
V.S.	Avg. (↓)	-6.72	-6.54	-6.34	-6.1	-5.66	-6.64
	Med. (↓)	<u>-6.84</u>	-6.51	-6.72	-6.53	-6.3	-7.02
V.M.	Avg. (↓)	-7.18	-6.93	-6.98	-6.89	-6.67	-7.41
	Med. (↓)	<u>-7.06</u>	-6.7	-6.96	-6.92	-6.73	-7.28
V.D.	Avg. (↓)	-7.89	-7.70	-7.78	-7.68	-7.67	-8.09
	Med. (↓)	<u>-7.97</u>	-7.70	-7.77	-7.87	-7.81	-8.13
Q.	Avg. (↑)	0.55	0.48	<u>0.57</u>	<u>0.57</u>	0.54	0.59
S.	Avg. (↑)	0.69	0.64	<u>0.74</u>	0.78	0.71	0.72
S.E.	25% (↓)	99.62	229.31	81	64	91	65
	50% (↓)	369.94	758.55	232	148	216	<u>150</u>
	75% (↓)	1247.60	2469.64	718	363	514	<u>375</u>

^a (↑)/(↓) indicates larger/smaller is better. Top-2 results are highlighted with **bold** and underlined. Note: V.S. = Vina Score, V.M. = Vina Min, V.D. = Vina Dock, Q. = QED, S. = SA, S.E. = Strain Energy, F.M. = Flow Matching.

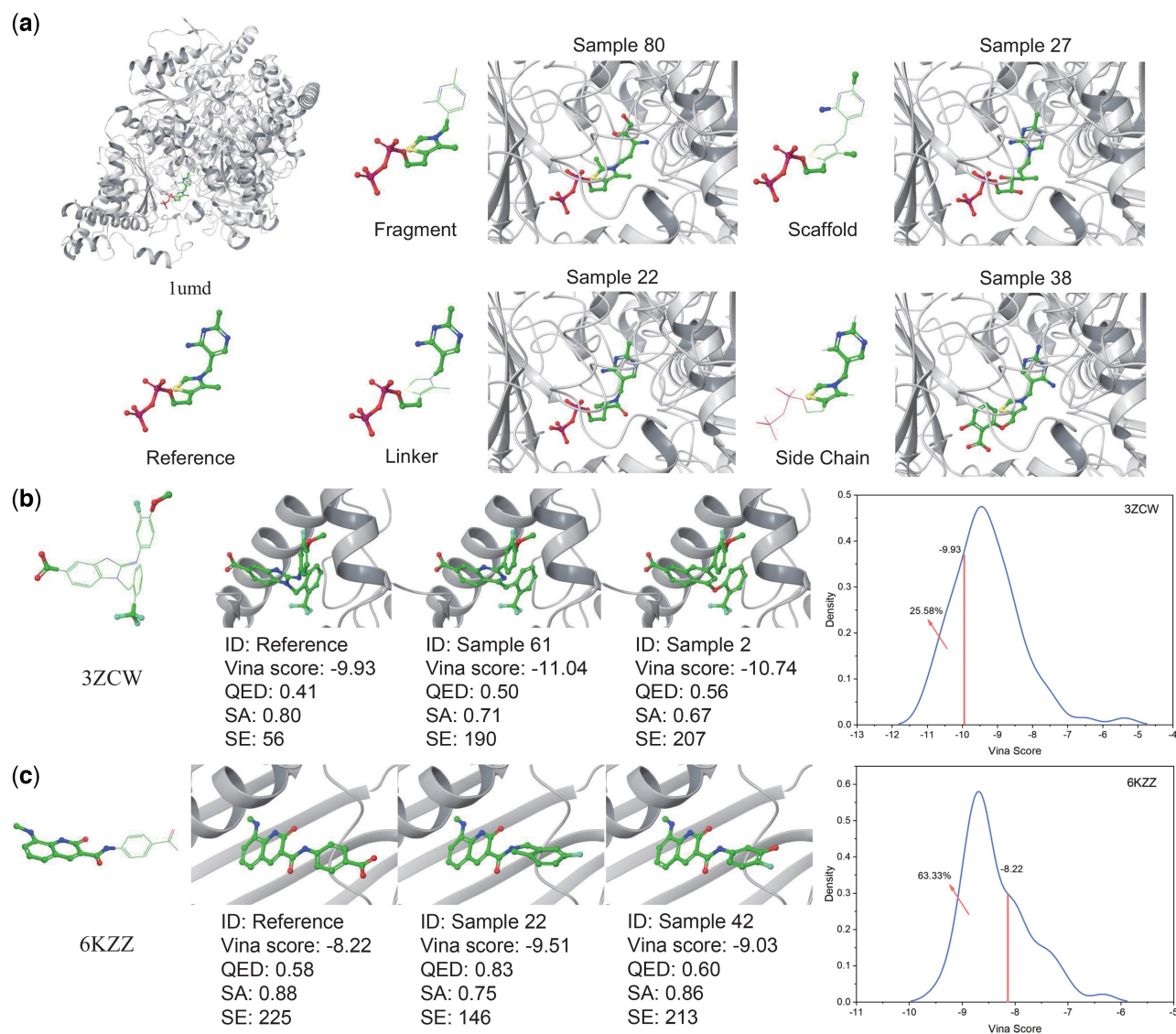


Figure 4 Case study of generated molecules in lead optimization scenarios. (a) 3D structures of selected MolPIF-generated molecules for target 1umld across four optimization scenarios. (b, c) 3D structures of MolPIF-generated molecules for targets 3ZCW (scaffold hopping) and 6KZZ (fragment growth) alongside reference molecules, accompanied by their respective Vina score distributions.

probability distributions, enabling smooth transformations between prior and target distributions for both continuous and discrete data. PIF overcomes the limitations of traditional flow models, diffusion models, and Bayesian flow networks by interpolating between distributions in the parameter space rather than sample space.

Theoretically, we demonstrate that PIF recovers W_2 -optimal transport for continuous variables and establishes Fisher–Rao geodesics for discrete data under exponential family priors, preserving the intrinsic manifold structure.

We apply PIF in MolPIF for SBDD. Our empirical evaluations on the CrossDocked2020 dataset confirm that MolPIF achieves superior performance in generating molecules with high chemical validity, accurate geometries, and strong binding affinities. Chemical space analysis confirms the model's ability to replicate reference distributions and expand structural diversity beyond known chemical territories. Our analysis of prior distributions (Gaussian versus Laplace) further illuminated the impact of prior selection on generation quality and structural refinement. MolPIF also shows promising results in lead optimization, generating novel candidates with enhanced docking scores and preserved critical binding characteristics.

Despite its efficacy, MolPIF has several limitations. First, it treats the protein binding pocket as a rigid structure, omitting the induced-fit dynamics crucial in physiological environments. Second, the framework currently lacks explicit attribute-guided mechanisms for fine-grained lead optimization. Lastly, although MolPIF is computationally efficient, further reducing the sampling steps through advanced solvers or distillation is necessary for its application in ultra-large-scale virtual screening.

In summary, PIF offers a new paradigm for generative modeling in drug discovery. By bridging the gap between heterogeneous modalities within a unified theoretical framework, MolPIF enables a more consistent and accurate exploration of the vast chemical space. Future work will explore extending PIF to more complex distribution priors and broader applications in molecular and materials science, as well as the attribute-guided mechanism under the MolPIF framework.

Author contributions

Yaowei Jin (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Junjie Wang (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Yufan Tang (Data curation [supporting], Validation [supporting], Visualization [supporting], Writing—review & editing [equal]), Wenkai Xiang (Conceptualization [supporting], Formal analysis [supporting], Validation [supporting], Writing—original draft [supporting]), Duanhua Cao (Supervision [supporting], Validation [supporting]), Dan Teng (Supervision [supporting], Validation [supporting]), Zhehuan Fan (Supervision [supporting], Validation [supporting]), Jiacheng Xiong (Supervision [supporting], Validation [supporting]), Xia Sheng (Supervision [supporting],

Validation [supporting]), Chuanlong Zeng (Supervision [supporting], Validation [supporting]), Duo An (Supervision [supporting], Validation [supporting]), Mingyue Zheng (Methodology [supporting], Supervision [supporting], Validation [supporting]), Shuangjia Zheng (Formal analysis [supporting], Supervision [supporting], Validation [supporting]), and Qian Shi (Conceptualization [supporting], Formal analysis [supporting], Funding acquisition [lead], Investigation [supporting], Methodology [supporting], Project administration [lead], Resources [lead], Supervision [lead], Validation [supporting], Writing—original draft [supporting], Writing—review & editing [supporting])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by Shanghai Rising-Star Program [23QD1400600] and National Key Research and Development Program of China [2022YFC3400504].

Data availability

Our model weights, configuration files, and generated molecules are publicly available at <https://drive.google.com/drive/folders/1VBGnHyThNHpdALgppOeKCKomwfl6oXde>. The code of MolPIF is freely available at <https://github.com/BLEACH366/MolPIF>, and is archived at Zenodo: <https://doi.org/10.5281/zenodo.16925025>.

References

- Alexandar SP, Yennamalli RM, Ulaganathan V. Coarse grained modelling highlights the binding differences in the two different allosteric sites of the human kinesin eg5 and its implications in inhibitor design. *Comput Biol Chem* 2022;**99**:107708.
- Austin J, Johnson DD, Ho J *et al*. Structured denoising diffusion models in discrete state-spaces. *Adv Neural Inf Process Syst* 2021;**34**:17981–93.
- Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. *Int J Mol Sci* 2019;**20**:2783.
- Bickerton GR, Paolini GV, Besnard J *et al*. Quantifying the chemical beauty of drugs. *Nat Chem* 2012;**4**:90–8.
- Boll B, Gonzalez-Alvarado D, Petra S *et al*. Generative assignment flows for representing and learning joint distributions of discrete data. *J Math Imaging Vis* 2025;**67**:34.
- Buttenschoen M, Morris GM, Deane CM. Posebusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci* 2024;**15**:3130–9.
- Chen S, Zhang O, Jiang C *et al*. Deep lead optimization enveloped in protein pocket and its application in designing potent

- and selective ligands targeting LTK protein. *Nat Mach Intell* 2025;**7**:448–58.
- Cheng C, Li J, Peng J *et al.* Categorical flow matching on statistical manifolds. *Adv Neural Inf Process Syst* 2024;**37**:54787–819.
- Cremer J, Irwin R, Tibo A *et al.* FLOWR: flow matching for structure- and interaction-aware de novo ligand generation. In: *Proceedings of the ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2025.
- Cui T, Tang C, Su M *et al.* Geometry-enhanced pretraining on interatomic potentials. *Nat Mach Intell* 2024;**6**:428–36.
- Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* 2009;**1**:8.
- Firth NC, Brown N, Blagg J. Plane of best fit: a novel method to characterize the three-dimensionality of molecules. *J Chem Inf Model* 2012;**52**:2516–25.
- Francoeur PG, Masuda T, Sunseri J *et al.* Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J Chem Inf Model* 2020;**60**:4200–15.
- Gat I, Remez T, Shaul N *et al.* Discrete flow matching. *Adv Neural Inf Process Syst* 2024;**37**:133345–85.
- Geyer CJ. Practical Markov chain Monte Carlo. *Statist Sci* 1992;**7**:473–83.
- Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *J Comb Chem* 1999;**1**:55–68.
- Graves A, Srivastava RK, Atkinson T *et al.* Bayesian flow networks. arXiv, arXiv:230807037, 2023, preprint: not peer reviewed.
- Green H, Durrant JD. Deepfrag: an open-source browser app for deep-learning lead optimization. *J Chem Inf Model* 2021;**61**:2523–9.
- Gu S, Smith MS, Yang Y *et al.* Ligand strain energy in large library docking. *J Chem Inf Model* 2021;**61**:4331–41.
- Gu S, Xu M, Powers A *et al.* Aligning target-aware molecule diffusion models with exact energy optimization. *Adv Neural Inf Process Syst* 2024;**37**:44040–63.
- Guan J, Qian WW, Peng X *et al.* 3D equivariant diffusion for target-aware molecule generation and affinity prediction. In: *Proceedings of the Eleventh International Conference on Learning Representations*, 2023a.
- Guan J, Zhou X, Yang Y *et al.* Decompdiff: diffusion models with decomposed priors for structure-based drug design. In: *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202, 2023b, 11827–46.
- Guo D, Yang D, Zhang H *et al.* Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 2025;**645**:633–8.
- Han J, Liu J, Jiang Y *et al.* Infinity ∞ : Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, 15733–44.
- Harris C, Didi K, Jamasb AR *et al.* Benchmarking generated poses: how rational is structure-based drug design with generative models? arXiv, arXiv:230807413, 2023, preprint: not peer reviewed.
- Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 2020;**33**:6840–51.
- Jiang D, Hsieh C-Y, Wu Z *et al.* A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *J Med Chem* 2021;**64**:18209–32.
- Li G, Jiang C, Gao Z *et al.* Molecule generation for target protein binding with hierarchical consistency diffusion model. *Chem Sci* 2025;**16**:18263–77.
- Lin H, Zhao G, Zhang O *et al.* Cbgbench: fill in the blank of protein-molecule complex binding graph. In: *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.
- Lipinski CA, Lombardo F, Dominy BW *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997;**23**:3–25.
- Lipman Y, Chen RT, Ben-Hamu H *et al.* Flow matching for generative modeling. In: *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.
- Liu X, Li Z, Gu Q. Cachequant: comprehensively accelerated diffusion models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, 23269–80.
- Luo S, Guan J, Ma J *et al.* A 3d generative model for structure-based drug design. *Adv Neural Inf Process Syst* 2021;**34**:6229–39.
- Peng X, Guo F, Guo R *et al.* Unified modeling of 3D molecular generation via atomic interactions with PocketXMol. *Cell* 2026;**189**:1904–22.e28.
- Peng X, Luo S, Guan J *et al.* Pocket2mol: efficient molecular sampling based on 3d protein pockets. In: *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162, 2022, 17644–55.
- Qiu K, Song Y, Fan Z *et al.* Piloting structure-based drug design via modality-specific optimal schedule. In: *Proceedings of the 42nd International Conference on Machine Learning*, Vol. 267, 2025, 50619–44.
- Qu Y, Qiu K, Song Y *et al.* Molcraft: structure-based drug design in continuous parameter space. In: *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235, 2024, 41749–68.
- Ragoza M, Masuda T, Koes DR. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chem Sci* 2022;**13**:2701–13.
- Sauer WH, Schwarz MK. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J Chem Inf Comput Sci* 2003;**43**:987–1003.
- Schreyer AM, Blundell T. Uscat: real-time ultrafast shape recognition with pharmacophoric constraints. *J Cheminform* 2012;**4**:27.
- Soch J, Saritaş K, Maja *et al.* StatProofBook/StatProofBook. github.io: The Book of Statistical Proofs (Version 2023). Geneva, Switzerland: Zenodo, 2024. <https://doi.org/10.5281/zenodo.4305949>
- Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. In: *Advances in Neural Information Processing Systems*, 2019, Vol. 32, 11895–907.
- Song Y, Gong J, Xu M *et al.* Equivariant flow matching with hybrid probability transport for 3D molecule generation. *Adv Neural Inf Process Syst* 2023;**36**:549–68.
- Stark H, Jing B, Wang C *et al.* Dirichlet flow matching with applications to DNA sequence design. In: *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235, 2024, 46495–513.

- Steinegger M, Söding J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8.
- Tian Z, Jin Y, Liu Z *et al*. AudioX: a unified framework for anything-to-audio generation. In: *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026.
- Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61.
- Ushiyama F, Amada H, Takeuchi T *et al*. Lead identification of 8-(methylamino)-2-oxo-1, 2-dihydroquinoline derivatives as dna gyrase inhibitors: hit-to-lead generation involving thermodynamic evaluation. *ACS Omega* 2020;**5**:10145–59.
- Xie W, Wang F, Li Y *et al*. Advances and challenges in de novo drug design using three-dimensional deep generative models. *J Chem Inf Model* 2022;**62**:2269–79.
- Zang C, Wang F. Moflow: an invertible flow model for generating molecular graphs. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, **2020**, 617–26.
- Zhao L, Ding X, Yu L *et al*. Unified discrete diffusion for categorical data. *J Mach Learn Res* 2025;**26**:1–49.