Binding Mode Matters: Hotspot-Aware Drug Discovery via Explorative Preferences

Dingshuo Chen[1,2,4] **Hao Yang**[3] **Kuangqi Zhou**[4] **Zhixun Li**[5]
Qiang Liu[1,2] **Xiangxiang Zeng**[6] **Shu Wu**[1,2*] **Yaosen Min**[4*]
[1] Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University
[4] Zhongguancun Academy & Zhongguancun Institute of Artificial Intelligence
[5] The Chinese University of Hong Kong
[6]College of Computer Science and Electronic Engineering, Hunan University

Abstract

Discovering hit molecules requires not just high binding affinity, but also the identification of diverse binding modes, which are critical for experimental assays. Existing generative approaches have predominantly relied on optimizing a scalar docking score, obscuring the distinct contributions of key binding determinants. To this end, we introduce a paradigm shift by formulating target-based drug design as a multi-objective exploration task, where each objective explicitly corresponds to enhancing interactions with a specific hotspot. Here, we introduce `BindMol`, a novel generative framework driven by a customized multi-objective reinforcement learning algorithm. By incorporating explorative preferences during training, our approach efficiently uncovers molecules with diverse and desirable binding profiles. Empirical results demonstrate that `BindMol` facilitates the discovery of high-affinity compounds characterized by both structural novelty and diverse binding modes. Validated across target-based drug discovery and multi-property optimization tasks, our approach provides a versatile paradigm for goal-oriented drug discovery.

1 Introduction

Target-based drug design aims to identify ligands that not only bind to a target protein with high affinity but also engage specific residues to trigger a desired pharmacological response [11]. The recent surge in deep generative models has accelerated this process, offering a powerful toolkit that spans *de novo discovery* and *goal-oriented generation* [8]. Reinforcement learning, in particular, has emerged as a dominant paradigm for conditional generation [41, 35] and optimization [62, 59]. It generates molecular structures by maximizing a reward reflecting drug-likeness and bioactivity [41].

Despite these advances, a misalignment persists between the complexity of biochemical interactions and the simplified objectives used in RL training. Most existing approaches reduce the ligand-protein interaction to a single scalar metric, such as the docking score. While computationally convenient, such scalar measures often overlook the complexity of interactions: ***compounds targeting the same protein can adopt different binding modes, leading to distinct functional and selectivity profiles***. For instance,

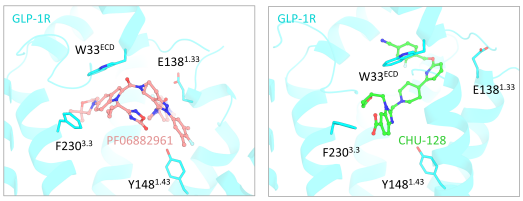


Figure 1: Illustration of two agonists of GLP-1R, highlighting their distinct binding modes within the same protein pocket.

*Corresponding author: Yaosen Min (`yaosenmin@zgci.ac.cn`), Shu Wu (`shu.wu@nlpr.ia.ac.cn`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

two well-known GLP-1 agonists exhibit similar affinities but engage distinct sub-regions, resulting in either broad-spectrum or biased signaling [60], as illustrated in Figure 1.

To address this limitation, we reformulate target-based drug discovery as a multi-objective optimization problem, where each objective corresponds to a distinct *hotspot* within the binding pocket. This conceptual shift naturally encourages the exploration of diverse binding modes by explicitly exposing trade-offs between interactions with different pocket regions. It also provides a more robust alternative to scalar objectives, which are known to be prone to reward hacking and suboptimal local solutions [13]. Specifically, we propose a novel reinforcement learning framework for drug discovery, termed `BindMol`. Its realization relies on three parts that bridge chemical validity, biological specificity, and algorithmic efficiency.

We explore the discrete chemical space using a fragment-based action space [59], where molecules are incrementally assembled in a *Lego-like* fashion. A compact yet informative fragment vocabulary is obtained via an energy-based model trained with noise contrastive estimation [21]. We theoretically ground fragment selection as a stability analysis problem on a thermodynamic potential surface, where high-scoring fragments act as stabilizing anchors for the ligand-protein complex. With a robust action space established, we next focus on the reward signal. Instead of flattening complex biological interactions into a docking score, we introduce a hotspot-aware vectorized reward by clustering the binding pocket into distinct *hotspots* based on reference structures, with each hotspot defining a separate objective. This design compels the agent to navigate trade-offs across sub-pockets, naturally leading to the discovery of diverse binding modes.

However, maximizing a vectorized reward is far more challenging than a scalar one [37, 1]. To enable precise control over multi-objective trade-offs, we establish *Envelope SAC*, a principled actor-critic framework that unifies the optimization of preference-aware critics and policies. This formulation significantly advances the theoretical scope of the envelope operator [58], moving beyond its value-based origins to the continuous domain. By constructing a utility envelope that aligns vectorized Bellman updates, our method efficiently approximates the convex hull of the Pareto frontier, ensuring robust performance across diverse preferences.

We empirically validate `BindMol` on binding affinity optimization tasks across five protein targets and the extensive GuacaMol benchmark. Our results demonstrate that `BindMol` not only generates valid, high-affinity compounds but also successfully discovers diverse binding modes that cover a broader range of the bioactive chemical space compared to state-of-the-art baselines. The main contributions of this work are three-fold:

❶ We propose the first molecular generation framework that treats binding-mode diversity as a *first-class* modeling objective, explicitly accounting for biologically grounded structural variations. This moves target-based design one step closer to treating molecular generation as an exploration of protein–ligand interaction patterns.

❷ We propose `BindMol`, a principled multi-objective reinforcement learning framework that integrates energy-based fragment scoring with hotspot-aware vectorized rewards, and develops Envelope SAC, an actor–critic algorithm capable of efficiently approximating the Pareto frontier under preference-conditioned optimization.

❸ We demonstrate the superiority of `BindMol` across diverse benchmarks. It outperforms existing methods in generating high-affinity ligands with novel binding modes across five protein targets and achieves state-of-the-art results on the comprehensive GuacaMol multi-objective benchmark.

2 Related Work

Fragment-based molecular generation Fragment-based molecular generation has emerged as an effective paradigm for drug discovery. Methods such as MARS [55], FREED [59], and MoLeR [38] assemble molecules from fragments to efficiently explore chemical space. Related approaches exploit informative substructures, including principal subgraph mining [30] and molecular rationales [27], to guide multi-objective generation. Incorporating protein structural context further improves generation quality, for example through transferable subpocket prototypes [61] or pocket-aware fragment assembly via diffusion [17]. Beyond generation, HN-GFN [63] enables sample-efficient multi-objective optimization, and recent antibiotic discovery results [31] demonstrate the practical impact of fragment-based strategies.

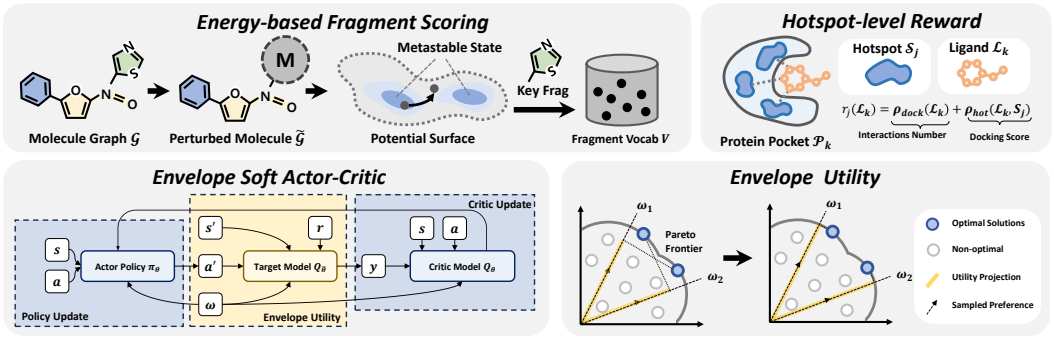


Figure 2: Overview of the `BindMol` consisting of three components: energy-based fragment scoring, Envelope SAC, and hotspot-level reward design. **(Top left)** Energy-based fragment scoring selects stabilizing fragments by an energy-based model trained via noise contrastive estimation. **(Bottom left)** Overview of the Envelope SAC workflow (double Q_θ models omitted for clarity). **(Top right)** Hotspot-aware rewards are computed from the sum of interaction counts and docking scores, followed by normalization. **(Bottom right)** Illustration of envelope utility: unlike scalarized RL without maximization over preference, our method identifies the optimal utility for each preference.

RL-driven molecular generation Deep reinforcement learning has been extensively applied to molecular design using both sequence- and graph-based formulations. Sequence-based policy optimization methods, including REINVENT [41] and its variants with curriculum learning [20], adapt SMILES generators toward target properties. Graph-based approaches [62] instead operate directly on molecular structures through chemistry-aware edit actions, often without pretraining. Fragment-based RL methods further address exploration efficiency and chemical validity by assembling molecules from learned vocabularies [59, 35]. More recently, structure-aware techniques such as AliDiff [18] explicitly align generative models with protein-binding preferences. Together, these works demonstrate the flexibility of RL for molecular design, while highlighting persistent challenges related to reward design and exploration.

Multi-objective reinforcement learning Multi-objective reinforcement learning (MORL) spans both tabular and deep settings, with single-policy and multi-policy formulations [10]. Tabular methods either optimize scalarized objectives [49] or maintain non-dominated value estimates, as in Pareto Q-learning [48]. In the deep setting, single-policy approaches such as EUPG [45] optimize user-defined utilities, while multi-policy methods dominate, including weight-conditioned models [58, 37], evolutionary approaches [57], return-conditioned models [44], and GPI frameworks [1]. These methods differ mainly in how preferences are incorporated and how the Pareto front is approximated.

3 Preliminaries

Markov decision process. We formalize a multi-objective reinforcement learning problem as a Multi-Objective Markov Decision Process (MOMDP), defined by

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \Omega \rangle$$

with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability distribution $\mathcal{P}(s'|s,a)$, vector reward function $\mathbf{r}$, and the space of preference Ω. If $\boldsymbol{\omega} \in \Omega$ is fixed to a single value, this MOMDP collapses into a standard MDP. On the other hand, if we consider all possible returns from an MOMDP, we have a Pareto frontier $\mathcal{F}^* := \{\hat{\boldsymbol{r}} \mid \nexists \hat{\boldsymbol{r}}' \geq \hat{\boldsymbol{r}}\}$, where the return $\hat{\boldsymbol{r}} := \sum_t \gamma^t \boldsymbol{r}(s_t, a_t)$.

Fragment-based molecular generation Following Yang et al. [59], each transition is driven by a triplet action $\mathbf{a}_t = (u_t, f_t, v_t)$, which attaches a chosen fragment f_t from the vocabulary to the current graph $\mathcal{G}_t$ at site u_t, using the fragment's site v_t. The generation policy $\pi_\theta(\mathbf{a}_t|s_t)$ is factorized into three hierarchical sub-policies:

$$\pi_\theta(\mathbf{a}_t \mid \mathcal{G}_t) = \mathbb{P}(u_t \mid \mathcal{G}_t) \cdot \mathbb{P}(f_t \mid u_t, \mathcal{G}_t) \cdot \mathbb{P}(v_t \mid f_t, u_t, \mathcal{G}_t). \tag{1}$$

Detailed descriptions of the molecular generator are provided in Appendix D.

4 Method

The three components of `BindMol` address complementary stages of one design problem. Energy-based fragment scoring (Section 4.1) constructs a compact, target-adapted molecular search space; hotspot-aware vector rewards (Section 4.2) formulate preference-controllable exploration of binding modes; and Envelope SAC (Section 4.3) optimizes the resulting conflicting objectives in the combinatorial generation process. The overall framework is illustrated in Figure 2.

4.1 Energy-Based Fragment Scoring

We formulate the identification of goal-aware fragments as a stability analysis problem within a learned thermodynamic potential. We hypothesize that high-activity molecules represent metastable states in a joint structure-activity landscape [7]. Consequently, critical substructures act as stabilizing anchors; their perturbation should yield a significant energy fluctuation, pushing the system towards an unstable manifold.

Let the joint state space be $\mathcal{Z} = \mathcal{G} \times \mathcal{Y}$, where $\mathcal{G} = (\mathcal{V}_g, \mathcal{E}_g)$ denotes the molecular graph with node feature matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}_g| \times d}$, and $y \in \mathcal{Y} \subset \mathbb{R}$ is the target property. We model the compatibility probability density as a Boltzmann distribution $p_\theta(\mathbf{z}) \propto e^{-E_\theta(\mathbf{z})}$, governed by a parameterized energy function $E_\theta : \mathcal{Z} \to \mathbb{R}$ [33]. The energy architecture employs a message-passing encoder $\Phi_G : \mathcal{G} \to \mathbb{R}^h$ and a scalar projection $\Phi_y : \mathcal{Y} \to \mathbb{R}^h$. Their shared output width is a compact implementation choice, not a mathematical requirement: Φ_G summarizes the graph, whereas Φ_y lifts the scalar property into a conditioning representation, and different widths could also be concatenated. The scalar potential is derived via a fused non-linear map Ψ:

$$E_\theta(\mathcal{G}, y) \coloneqq \Psi\left(\Phi_G(\mathbf{H}, \mathcal{E}_g) \,\|\, \Phi_y(y)\right), \tag{2}$$

where $\|$ denotes concatenation. To ensure the physical validity of the landscape, we train E_θ using Noise Contrastive Estimation (NCE) [21]. We define the training objective $\mathcal{J}(\theta)$ to maximize the discriminative margin between the empirical distribution $\mathbb{P}_{\text{data}}$ and a generated noise distribution $\mathbb{P}_{\text{noise}}$:

$$\mathcal{J}(\theta) = \mathbb{E}_{(\mathcal{G},y)\sim\mathbb{P}_{\text{data}},(\tilde{\mathcal{G}},\tilde{y})\sim\mathbb{P}_{\text{noise}}}\left[\log\left(1 + e^{E_\theta(\mathcal{G},y) - E_\theta(\tilde{\mathcal{G}},\tilde{y})}\right)\right]. \tag{3}$$

Here, $\mathbb{P}_{\text{noise}}$ is constructed by stochastically corrupting either the topological structure $\mathcal{E}_g$ or the property label y, forcing the model to assign global minima only to chemically valid and property-matched pairs. The empirical distribution may be formed from experimentally annotated target ligands or, for an underexplored target, from a public library scored by virtual screening. Thus, more target-relevant data improves the learned landscape, but the intended prerequisite is a reliable target structure and pocket definition rather than a large experimental activity set; a dataset-size analysis is provided in Section G.3.

Upon convergence, we utilize E_θ to quantify the contribution of substructures. For a molecule decomposed into a fragment set $\{\mathcal{F}_k\}_{k=1}^K$, let $\mathcal{P}_{\mathcal{F}_k^\perp} : \mathbb{R}^{|\mathcal{V}_g| \times d} \to \mathbb{R}^{|\mathcal{V}_g| \times d}$ be an orthogonal projection operator that effectively filters out the signal from fragments $\mathcal{F}_k$ (i.e., masking). The structural importance $\mathcal{I}(\mathcal{F}_k)$ is defined as the variation in potential energy under this projection:

$$\mathcal{I}(\mathcal{F}_k) \coloneqq \delta E_\theta(\mathcal{G}, y; \mathcal{F}_k) = E_\theta(\mathcal{P}_{\mathcal{F}_k^\perp}(\mathbf{H}), \mathcal{E}_g, y) - E_\theta(\mathbf{H}, \mathcal{E}_g, y). \tag{4}$$

A high value of $\mathcal{I}(\mathcal{F}_k)$ implies that the fragment $\mathcal{F}_k$ is essential for maintaining the low-energy state associated with high bioactivity.

Theoretical Understanding. The following result connects the energy change induced by fragment masking to the path-integrated gradient of the learned energy function, providing a principled fragment-importance signal rather than a purely heuristic score.

Theorem 1 (Manifold Gradient Equivalence)**.** *Let $\gamma : [0, 1] \to \mathbb{R}^{|\mathcal{V}_g| \times d}$ be a linear path parameterizing the feature ablation from $\mathbf{H}$ to $\mathcal{P}_{\mathcal{F}_k^\perp}(\mathbf{H})$. Assuming E_θ is continuously differentiable, the discrete energy variation $\mathcal{I}(\mathcal{F}_k)$ is exactly the path integral of the directional derivatives along γ:*

$$\mathcal{I}(\mathcal{F}_k) = -\int_0^1 \langle \nabla_{\mathbf{H}} E_\theta(\gamma(t), y),\, \mathbf{1}_{\mathcal{F}_k} \odot \mathbf{H} \rangle_F \, dt, \tag{5}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, $\mathbf{1}_{\mathcal{F}_k}$ is the indicator mask for fragment $\mathcal{F}_k$, and $\odot$ is the Hadamard product.

In essence, Theorem 1 characterizes the learned energy-gap score as an integrated sensitivity measure along the masking path. Detailed proofs, error analysis, and conditions for ranking consistency are provided in Appendix E; pseudocode is given in Algorithm 1.

4.2 Hotspot-Based Multi-Objective Reward Design

To guide the agent in exploring diverse binding modes rather than converging to a single dominant interaction pattern, we formulate the molecular generation task as a multi-objective optimization problem. We decompose the binding pocket into distinct spatial subpockets based on structural information derived from reference ligand-protein complexes from Protein Data Bank (PDB) [2].

Let $\mathcal{D} = \{(\mathcal{P}_k, \mathcal{L}_k)\}_{k=1}^{K}$ denote a dataset of K reference complexes for the target protein, where each entry consists of a protein structure $\mathcal{P}_k$ and its corresponding ligand $\mathcal{L}_k$. We represent the geometry of each residue v by its side-chain centroid $\mathbf{c}_v$, and the ligand $\mathcal{L}_k$ by its set of heavy atom coordinates $\mathcal{A}(\mathcal{L}_k)$. For each complex, the binding site residue set $\mathcal{R}_k$ is identified via a centroid-to-atom distance threshold δ (e.g., 4Å):

$$\mathcal{R}_k = \left\{ v \in \mathcal{P}_k \mid \min_{\mathbf{a} \in \mathcal{A}(\mathcal{L}_k)} \|\mathbf{c}_v - \mathbf{a}\|_2 < \delta \right\}. \tag{6}$$

To maximize the distinctiveness of the objectives and reduce the bias toward common binding residues, we construct a *non-overlapping residue set* $\mathcal{V}$ by excluding the intersection of all binding sites from their union:

$$\mathcal{V} = \left(\bigcup_{k=1}^{K} \mathcal{R}_k \right) \setminus \left(\bigcap_{k=1}^{K} \mathcal{R}_k \right). \tag{7}$$

Excluding the intersection does not discard conserved interactions from optimization: the global docking term below is shared by every objective and maintains pressure toward overall affinity. Instead, the hotspot-specific terms use differentiating residues; otherwise, conserved contacts can correlate all reward dimensions and let the policy improve every objective through the same anchor. This set $\mathcal{V}$ serves as the basis for subpocket identification. We partition it into m spatial subpockets $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ using spectral clustering [50] on the Gaussian kernel of pairwise centroid distances. A full-union ablation is reported in Table 9, and clustering details are provided in Section C.

The reward function evaluates the generated molecule $\mathcal{M}$ based on both global binding affinity and local subpocket interactions. We define the reward vector $\boldsymbol{r} \in \mathbb{R}^m$, where the j-th component r_j corresponds to the j-th subpocket $\mathcal{S}_j$ and consists of two terms: $r_j(\mathcal{M}) = \rho_{\text{dock}}(\mathcal{M}) + \rho_{\text{hot}}(\mathcal{M}, \mathcal{S}_j)$. Here, ρ_{dock} represents the docking score normalized to a fixed range (e.g., $[0, 1]$). Docking is performed once for each completed molecule, rather than after every fragment action; PLIP [46] reuses that pose to compute ρ_{hot} without another docking call.

Let $\eta(v, \mathcal{M})$ be the count of non-covalent interactions between molecule $\mathcal{M}$ and residue v. We first compute the average interaction density within the subpocket, denoted as $\bar{\eta}_j = \frac{1}{|\mathcal{S}_j|} \sum_{v \in \mathcal{S}_j} \eta(v, \mathcal{M})$. This value is then clipped and normalized to encourage valid contacts without over-saturation: $\rho_{\text{hot}}(\mathcal{M}, \mathcal{S}_j) = \frac{1}{10} \min(\bar{\eta}_j, 10)$. This formulation ensures that the agent optimizes for molecules that are not only energetically favorable globally but also form dense, specific interaction patterns with the target hotspots.

4.3 Envelope Soft Actor-Critic

The reward formulation in Section 4.2 yields a vector $\boldsymbol{r} \in \mathbb{R}^m$, representing interactions across spatially disjoint subpockets. Reducing this vector to a static scalar via fixed weights prevents the agent from dynamically exploring trade-offs between different hotspots. To address this, we propose *Envelope SAC*, which learns a universal policy capable of approximating the entire Pareto frontier by conditioning on a preference vector $\boldsymbol{\omega} \in \Omega$. The training pipeline of Envelope SAC is described in Algorithm 2 in Section A.

Envelope target. We parameterize the critic as a vectorized function $\boldsymbol{Q}_\theta : \mathcal{S} \times \mathcal{A} \times \Omega \to \mathbb{R}^m$, which estimates the expected discounted return for each objective dimension independently. A key challenge in MORL is that the optimal policy for a specific preference query $\boldsymbol{\omega}$ is initially unknown. Standard off-policy updates, which use $a' \sim \pi_\phi(\cdot|s', \boldsymbol{\omega})$ to update $\boldsymbol{Q}_\theta(\cdot, \cdot, \boldsymbol{\omega})$, suffer from inefficiency as they ignore potentially superior actions discovered by the policy when conditioned on other preferences.

Given a batch of sampled preferences $\{\boldsymbol{\omega}_k\}_{k=1}^K$, we generate a candidate action set $\mathcal{A}_{\text{cand}} = \{a'_k \sim \pi_\phi(\cdot|s', \boldsymbol{\omega}_k)\}_{k=1}^K$ for the next state s'. Instead of restricting the update to the action generated by $\boldsymbol{\omega}_i$, we allow the critic to *"borrow"* the best action from the entire candidate pool. The envelope target $\boldsymbol{y}_i$ for preference $\boldsymbol{\omega}_i$ is computed as:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{r} + \gamma(1-d)\boldsymbol{Q}_{\bar{\theta}}(s', a'_{k^*}, \boldsymbol{\omega}_i), \\ \text{where } k^* &= \underset{k\in\{1,\dots,K\}}{\arg\max} \ \langle \boldsymbol{\omega}_i, \boldsymbol{Q}_{\bar{\theta}}(s', a'_k, \boldsymbol{\omega}_i)\rangle \,. \end{aligned} \tag{8}$$

By selecting the index k^* maximizing scalar utility for $\boldsymbol{\omega}_i$, this operator effectively approximates the convex hull of the Pareto front, accelerating convergence by leveraging preference diversity.

Homotopy-based critic update. The optimization of the critic presents a two-fold challenge: accurate estimation of the multi-dimensional return vector and maximization of the projected scalar utility. Training solely on scalarized utility leads to information loss regarding individual objectives, while training solely on vector outputs may fail to prioritize the specific preference query $\boldsymbol{\omega}$.

To resolve this, we employ a homotopy-based optimization $\mathcal{L}_Q(\theta)$ [54] that linearly interpolates between vector-wise regression and scalar-wise alignment using a dynamic coefficient β:

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{\mathcal{B}}\Big[\beta \cdot \underbrace{\big(\langle \boldsymbol{\omega}, \boldsymbol{Q}_\theta(s,a,\boldsymbol{\omega})\rangle - \langle \boldsymbol{\omega}, \boldsymbol{y}\rangle\big)^2}_{\mathcal{L}_{\text{util}}} + (1-\beta)\cdot \underbrace{\|\boldsymbol{Q}_\theta(s,a,\boldsymbol{\omega}) - \boldsymbol{y}\|_2^2}_{\mathcal{L}_{\text{vec}}}\Big]. \tag{9}$$

The coefficient $\beta \in [0,1]$ serves as a curriculum schedule. In the early training stage ($\beta \to 0$), the optimization prioritizes $\mathcal{L}_{\text{vec}}$ to ensure the critic accurately learns the structure of the objective space. As training progresses ($\beta \to 1$), the focus shifts to $\mathcal{L}_{\text{util}}$ to refine the policy alignment with the preferences, ensuring optimality on the Pareto frontier.

Preference-aware policy update. The actor $\pi_\phi(a|s,\boldsymbol{\omega})$ is trained to maximize the expected envelope utility subject to entropy regularization. By conditioning on $\boldsymbol{\omega}$, the policy learns a continuous mapping from preference weights to the optimal action manifold. The policy loss is defined as:

$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D},\boldsymbol{\omega}\sim\Omega}\Big[\alpha \log \pi_\phi(\tilde{a}|s,\boldsymbol{\omega}) - \min_{j=1,2} \big\langle \boldsymbol{\omega}, \boldsymbol{Q}_{\theta_j}(s,\tilde{a},\boldsymbol{\omega})\big\rangle\Big], \tag{10}$$

where $\tilde{a}$ is sampled via the reparameterization trick, and the minimum over two critics ($j=1,2$) is used to mitigate value overestimation [22]. This formulation ensures the agent not only adheres to the preference query but also maintains sufficient exploration to discover diverse molecular structures.

5 Experiments

To comprehensively validate the effectiveness of our proposed `BindMol`, we conduct comprehensive experiments on two important scenarios in drug discovery: (1) *binding affinity optimization* (Section 5.1), and (2) *similarity-based rediscovery* (Section 5.2), both of which can be formulated as multi-objective optimization tasks. We further demonstrate the effectiveness of our model design through ablation experiments and case studies in Section 5.3.

Experimental Settings We adhere to the task settings proposed by [35] for fair comparison. The policy model performs random sampling for the first 4,000 steps to collect experience, and parameter updates after 3,000 steps to ensure a sufficient replay buffer. Every method uses the same evaluation budget of 3,000 completed molecules and docking calls. We report *Novel hit ratio* and *Novel top 5% Docking Score* for hit discovery [59], area under the curve (AUC) of top-K value for rediscovery [16], and chemical-space coverage (#Circles) [56]. Please refer to Section B for experimental settings, evaluation protocols, baselines, and compute profiling.

5.1 Optimization of Binding Affinity

For binding affinity optimization, we aim to design novel, drug-like, and synthesizable binders to the specific target. Therefore, we validate `BindMol` on five protein targets, **parp1**, **fa7**, **5ht1b**, **braf**, and **jak2**, under the quantitative estimate of drug-likeness (QED) [3], synthetic accessibility (SA) [9], and novelty constraints following [34].

Table 1: **Novel hit ratio (%) results.** The results are the means and the standard deviations of 3 runs.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
REINVENT [41]	0.480 (± 0.344)	0.213 (± 0.081)	2.453 (± 0.561)	0.127 (± 0.088)	0.613 (± 0.167)
Graph GA [24]	4.811 (± 1.661)	0.422 (± 0.193)	7.011 (± 2.732)	3.767 (± 1.498)	5.311 (± 1.667)
MORLD [25]	0.047 (± 0.050)	0.007 (± 0.013)	0.880 (± 0.735)	0.047 (± 0.040)	0.227 (± 0.118)
HierVAE [26]	0.553 (± 0.214)	0.007 (± 0.013)	0.507 (± 0.278)	0.207 (± 0.220)	0.227 (± 0.127)
RationaleRL [27]	4.267 (± 0.450)	0.900 (± 0.098)	2.967 (± 0.307)	0.000 (± 0.000)	2.967 (± 0.196)
FREED [59]	4.627 (± 0.727)	1.332 (± 0.113)	16.767 (± 0.897)	2.940 (± 0.359)	5.800 (± 0.295)
PS-VAE [30]	1.644 (± 0.389)	0.478 (± 0.140)	12.622 (± 1.437)	0.367 (± 0.047)	4.178 (± 0.933)
MOOD [34]	7.017 (± 0.428)	0.733 (± 0.141)	18.673 (± 0.423)	5.240 (± 0.285)	9.200 (± 0.524)
GDSS [28]	1.933 (± 0.208)	0.368 (± 0.103)	4.667 (± 0.306)	0.167 (± 0.134)	1.167 (± 0.281)
RetMol [53]	0.011 (± 0.016)	0.000 (± 0.000)	0.033 (± 0.027)	0.000 (± 0.000)	0.011 (± 0.016)
GEAM [35]	40.635 (± 0.763)	17.903 (± 1.220)	36.667 (± 1.929)	27.887 (± 1.584)	42.025 (± 1.187)
MOLLEO [51]	35.725 (± 0.825)	20.651 (± 1.573)	32.489 (± 0.533)	25.112 (± 1.837)	39.120 (± 1.025)
`BindMol` (ours)	**42.617** (± 0.733)	**24.580** (± 0.652)	**42.149** (± 0.125)	**30.767** (± 0.974)	**44.864** (± 0.982)

Table 2: **Novel top 5% docking score (kcal/mol) results.** The results are the means and the standard deviations of 3 runs. The best results are highlighted in bold.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
REINVENT [41]	-8.702 (± 0.523)	-7.205 (± 0.264)	-8.770 (± 0.316)	-8.392 (± 0.400)	-8.165 (± 0.277)
Graph GA [24]	-10.949 (± 0.532)	-7.365 (± 0.326)	-10.422 (± 0.670)	-10.789 (± 0.341)	-10.167 (± 0.576)
MORLD [25]	-7.532 (± 0.260)	-6.263 (± 0.165)	-7.869 (± 0.650)	-8.040 (± 0.337)	-7.816 (± 0.133)
HierVAE [26]	-9.487 (± 0.278)	-6.812 (± 0.274)	-8.081 (± 0.252)	-8.978 (± 0.525)	-8.285 (± 0.370)
RationaleRL [27]	-10.663 (± 0.086)	-8.129 (± 0.048)	-9.005 (± 0.155)	*No hit found*	-9.398 (± 0.076)
FREED [59]	-10.579 (± 0.104)	-8.378 (± 0.044)	-10.714 (± 0.183)	-10.561 (± 0.080)	-9.735 (± 0.022)
PS-VAE [30]	-9.978 (± 0.091)	-8.028 (± 0.050)	-9.887 (± 0.115)	-9.637 (± 0.049)	-9.464 (± 0.129)
MOOD [34]	-10.865 (± 0.113)	-8.160 (± 0.071)	-11.145 (± 0.042)	-11.063 (± 0.034)	-10.147 (± 0.060)
GDSS [28]	-9.967 (± 0.028)	-7.775 (± 0.039)	-9.459 (± 0.101)	-9.224 (± 0.068)	-8.926 (± 0.089)
RetMol [53]	-8.590 (± 0.475)	-5.448 (± 0.688)	-6.980 (± 0.740)	-8.811 (± 0.574)	-7.133 (± 0.242)
GEAM [35]	-12.536 (± 0.115)	-9.611 (± 0.019)	-12.385 (± 0.059)	-12.165 (± 0.092)	-11.822 (± 0.035)
MOLLEO [51]	-11.981 (± 0.125)	-8.921 (± 0.008)	-12.588 (± 0.039)	-11.365 (± 0.117)	-10.988 (± 0.019)
`BindMol` (ours)	**-12.930** (± 0.055)	**-10.189** (± 0.023)	**-13.028** (± 0.097)	**-12.497** (± 0.087)	**-12.011** (± 0.015)

Table 3: **#Circles of generated hit molecules.** The #Circles threshold is set to 0.75. The results are the means and the standard deviations of 3 runs. The best results are highlighted in bold.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
REINVENT [41]	44.2 (± 15.5)	23.2 (± 6.6)	138.8 (± 19.4)	18.0 (± 2.1)	59.6 (± 8.1)
MORLD [25]	1.4 (± 1.5)	0.2 (± 0.4)	22.2 (± 16.1)	1.4 (± 1.2)	6.6 (± 3.7)
HierVAE [26]	4.8 (± 1.6)	0.8 (± 0.7)	5.8 (± 1.0)	3.6 (± 1.4)	4.8 (± 0.7)
RationaleRL [27]	61.3 (± 1.2)	2.0 (± 0.0)	**312.7** (± 6.3)	1.0 (± 0.0)	**199.3** (± 7.1)
FREED [59]	34.8 (± 4.9)	21.2 (± 4.0)	88.2 (± 13.4)	34.4 (± 8.2)	59.6 (± 8.2)
PS-VAE [30]	38.0 (± 6.4)	18.0 (± 5.9)	180.7 (± 11.6)	16.0 (± 0.8)	83.7 (± 11.9)
MOOD [34]	86.4 (± 11.2)	19.2 (± 4.0)	144.4 (± 15.1)	50.8 (± 3.8)	81.8 (± 5.7)
GEAM [35]	123.0 (± 7.8)	79.0 (± 9.2)	144.3 (± 8.6)	84.7 (± 8.6)	118.3 (± 0.9)
MOLLEO [51]	74.2 (± 2.5)	56.1 (± 3.2)	125.3 (± 7.9)	68.1 (± 5.4)	95.2 (± 3.3)
`BindMol` (ours)	**136.7** (± 4.8)	**97.5** (± 3.3)	173.2 (± 7.1)	**103.2** (± 5.1)	132.3 (± 0.9)

Performance comparison. Empirical results are summarized in Tables 1 to 3. Our systematic study suggests the following trends: *(i) `BindMol` achieves state-of-the-art performance across all targets*, highlighting its ability to generate novel, drug-like, synthesizable compounds with strong binding affinity. In particular, for challenging targets such as *fa7*, our method attains a novel hit ratio above 24%, surpassing GEAM (17%) and far outperforming most competing methods, many of which approach zero. *(ii) Fragment-based reinforcement learning methods are especially effective at exploring high-reward regions.* FREED, MOOD, GEAM, and our `BindMol` all adopt fragment-based RL strategies and consistently produce lower docking scores. This advantage arises from a well-defined action space: fragment-level operations are limited yet high-quality, enabling more efficient exploration than token-based SMILES generation. *(iii) `BindMol` achieves higher #Circle*

Table 4: **GuacaMol MPO AUC Top-100 results.** The results are the means of 3 runs. We highlight the best- and the second-performing results in **boldface** and underlined, respectively.

Method	Benchmark							Average
	Amlodipine	Fexofenadine	Osimertinib	Perindopril	Ranolazine	Sitagliptin	Zaleplon	
REINVENT [41]	0.608	0.752	0.806	0.511	0.719	0.006	0.325	0.532
MolDQN [62]	0.230	0.431	0.636	0.125	0.018	0.000	0.002	0.206
Graph GA [24]	0.622	0.731	0.799	0.503	0.670	0.330	0.305	0.566
Graph MCTS [24]	0.385	0.522	0.655	0.219	0.694	0.117	0.165	0.394
DoG-Gen [4]	0.489	0.640	0.706	0.422	0.601	0.015	0.073	0.421
GPBO [47]	0.538	0.685	0.750	0.460	0.694	0.117	0.165	0.487
STONED [39]	0.593	0.777	0.799	0.472	0.738	0.351	0.307	0.577
SynNet [15]	0.533	0.720	0.759	0.512	0.690	0.007	0.223	0.492
DST [14]	0.469	0.690	0.742	0.425	0.579	0.017	0.089	0.430
GEAM [35]	0.627	0.786	0.817	0.514	0.714	0.408	0.386	0.607
Augmented Memory [19]	0.607	0.657	0.818	0.518	0.725	0.398	0.355	0.582
MOLLEO (MoleculeSTM) [51]	0.617	0.765	0.799	0.516	0.675	0.415	0.392	0.597
`BindMol` (Ours)	**0.671**	**0.842**	**0.845**	**0.533**	**0.748**	**0.489**	**0.418**	**0.649**

while maintaining a strong hit ratio, indicating that the generated molecules are not only effective but also diverse. This reflects the benefit of explorative preferences, which encourage RL to explore a wider range of action sequences and thereby cover chemical space more comprehensively. By contrast, RationaleRL achieves relatively high diversity on *5ht1b* and *jak2*, but its overly aggressive exploration leads to very low hit discovery rates.

Multi-residue exploration. We evaluate the ability of `BindMol` to explore diverse binding modes using the **Hypervolume** metric, which quantifies the dominated volume of the Pareto front with respect to a preference point. As multi-objective hotspot optimization typically involves three to five objectives, many existing methods become computationally prohibitive. We therefore compare against GEAM [35], a strong baseline for binding tasks, and GPI [1], a multi-objective reinforcement learning method scalable to 3–5 objectives. As shown in Figure 3, multi-objective RL methods achieve broader Pareto-front coverage than GEAM, reflecting improved balance across competing interaction hotspots. Our method consistently attains the most extensive coverage across all five objectives, spanning multiple trade-off directions. This uniform Pareto exploration indicates that `BindMol` avoids collapsing to a single optimization mode and instead generates molecules with more diverse binding modes.

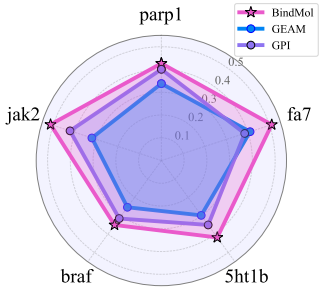


Figure 3: The Hypervolume comparison of `BindMol`.

Direct binding-mode diversity. Unlike #Circles and Hypervolume, which measure ligand-space and objective-space coverage, respectively, the PLIP-IFP metrics in Table 5 directly compare docking-predicted interaction patterns. `BindMol` yields more clusters and more balanced mode populations than GEAM on every target. We additionally select BRAF and JAK2 because they provide well-established, within-target pairs of type-I/DFG-in and type-II/DFG-out crystal complexes; the other targets lack equally clear paired mode labels. Reference-pair IFP/residue Jaccard similarities are 0.375/0.462 for BRAF and 0.188/0.250 for JAK2, and `BindMol` recovers all four crystallographic modes more frequently than the baselines.

Matched MORL comparison. We adapt GPI, C-MORL, and PreCo to the same molecular action space, hotspot reward vector, training protocol, docking budget, and three seeds. Table 5(c) summarizes the gains, with full per-target results in Table 8. Our claim is specific to preference-conditioned molecular exploration with 3–5 hotspot objectives, rather than general-purpose MORL superiority.

5.2 Optimization of Multi-property Objectives in GuacaMol

For similarity-based rediscovery, our goal is to generate molecules similar to the target compound and fulfill multiple required property profiles. Following [16], we select the multi-property optimization tasks from the GuacaMol benchmarks [6] for capability evaluation. Notably, unlike prior studies that scalarize multiple objectives into a single target, we treat them as a genuine multi-objective problem to better match our algorithm. To ensure fairness, we use each individual objective score prior to scalarization as the single entry of multi-objective reward.

Table 5: **Direct binding-mode and matched MORL evaluation.** IFP statistics are computed from full-pocket PLIP residue–interaction-type fingerprints; GEAM is the strongest baseline on all three statistics. Recovery denotes the percentage of generated hits with IFP Jaccard similarity ≥ 0.4 to the native crystal ligand. MORL results average the five targets; complete per-target results and protocols are in Section G.1.

(a) Interaction-pattern diversity

Metric	Method	PARP1	FA7	5HT1B	BRAF	JAK2
IFP clusters	GEAM	138.67	61.67	8.33	20.33	47.33
	`BindMol`	**192.00**	**94.33**	**11.67**	**35.00**	**77.33**
Entropy H	GEAM	2.93	1.85	0.24	0.68	1.52
	`BindMol`	**3.45**	**2.37**	**0.32**	**1.11**	**2.31**
Effective modes	GEAM	18.73	6.36	1.27	1.97	4.57
	`BindMol`	**31.40**	**10.71**	**1.37**	**3.03**	**10.10**

(b) Crystallographic-mode recovery (%)

Mode	`BindMol`	GEAM	FREED
BRAF DFG-in / I	**48.9**	20.3	11.7
BRAF DFG-out / II	**12.1**	6.4	0.0
JAK2 type-I	**1.3**	0.0	0.0
JAK2 type-II	**11.5**	5.7	2.8

(c) Matched MORL averages

Method	Hit (%)	DS
GPI	32.15	−10.60
C-MORL	32.35	−10.63
PreCo	32.87	−10.73
`BindMol`	**37.01**	**−12.13**

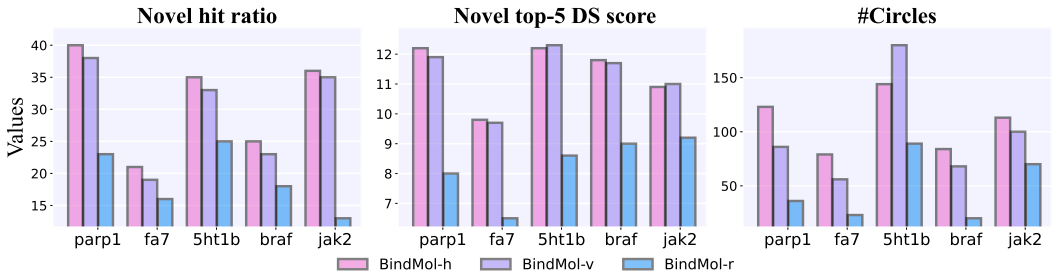


Figure 4: Performance of different ablation variants of `BindMol`. DS denotes Docking Score.

Performance comparison. Empirical results for MPO tasks are presented in Table 4 and Table 12. It can be observed that `BindMol` significantly outperforms all baseline models across all MPO tasks, which further demonstrates the effectiveness of the proposed Envelope SAC algorithm. On the two rediscovery tasks of *Fexofenadine* and *Sitagliptin*, it achieves relative improvements of 7.1% and 17.8% over the second-best models, respectively. Moreover, on the most challenging *Zaleplon* rediscovery task, `BindMol` attains a record-breaking score of 0.418. It is worth noting that MOLLEO (MoleculeSTM), a framework that applies genetic algorithms over large-scale molecular corpora with language models, is still outperformed by our method on rediscovery tasks. We exclude other variants of MOLLEO (e.g., those using GPT-4 as the base LLM) from comparison due to their unfair advantage in this setting.

5.3 Ablation Studies

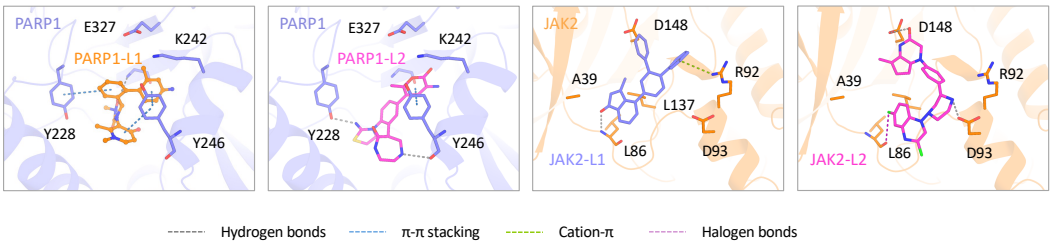


Figure 5: Visualization of representative binding modes on two protein complexes: *parp1* and *jak2*.

We perform ablation studies to isolate the contributions of the key components in our framework. Specifically, we remove energy-based fragment scoring and construct the action space from a default fragment library (*`BindMol-r`*); ablate the hotspot-aware reward and use only target-wise docking

scores for optimization (*BindMol-h*); and remove the envelope target, reducing Envelope SAC to standard SAC with vector-valued rewards (*BindMol-v*). As shown in Figure 4, removing any component leads to consistent performance degradation in both binding affinity and chemical-space coverage. The energy-based vocabulary most directly improves affinity-oriented search, explaining the larger decline of *BindMol-r* on docking metrics. The hotspot reward serves a different role: removing it reduces BRAF IFP clusters from 35.00 ± 1.41 to 25.67 ± 1.53 and effective modes from 3.03 ± 0.08 to 2.20 ± 0.08 (Table 10), directly supporting its contribution to interaction-pattern diversity. The envelope update further improves preference-conditioned exploration, as reflected by *BindMol-v*.

5.4 Case Studies

To provide an intuitive understanding of the binding modes exploration of BindMol, we cluster the generated molecules according to the number of protein–ligand interactions across different targets using k-means ($k = 2$). For each cluster, we select and visualize representative samples to examine their interaction patterns. Specifically, we perform clustering analysis on the *parp1* and *jak2* targets, resulting in four protein–ligand complexes for visualization, as shown in Figure 5.

In the *parp1* complexes, the two ligands exhibit distinct binding modes: PARP1-L1, with a polyaromatic scaffold, is dominated by aromatic stacking with Y228 and Y246, whereas PARP1-L2, featuring a heteroaromatic scaffold, relies more heavily on hydrogen bonding with diminished stacking. In the *jak2* complexes, the differences are also evident. JAK2-L1 adopts a stacking- and cation–π-driven mode, while JAK2-L2 introduces heteroatoms and halogen substitution that favor additional hydrogen and halogen bonds, reducing dependence on R92.

6 Conclusion

In this work, we revisit target-based drug discovery from the perspective of binding-mode diversity and argue that optimizing a single scalar objective is insufficient to capture the complexity of protein–ligand interactions. To this end, we propose BindMol, a multi-objective reinforcement learning framework that explicitly models hotspot-level interactions and their inherent trade-offs. By combining energy-based fragment scoring, biologically grounded vectorized rewards, and the *Envelope SAC* algorithm for preference-aware optimization, BindMol enables efficient exploration of diverse, high-affinity binding modes within the same target. Experiments across multiple targets and benchmarks demonstrate consistent improvements over existing approaches.

Broader impacts. Given that our application tasks fall within the molecular domain, improper use of methods for tasks such as molecular property prediction may result in significant deviations. This could impact subsequent applications of the molecules in drug development or materials design, especially in predicting properties like toxicity and stability. We recommend further experimental validation of key molecules after using the model to ensure the reliability of the results.

Acknowledgements. This work is jointly supported by National Science and Technology Major Project (2023ZD0120901) and Zhongguancun Academy (XTS0047, C20250208).

References

[1] Alegre, L. N., Bazzan, A. L., Roijers, D. M., Nowé, A., and da Silva, B. C. Sample-efficient multi-objective learning via generalized policy improvement prioritization. *arXiv preprint arXiv:2301.07784*, 2023.

[2] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[3] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

[4] Bradshaw, J., Paige, B., Kusner, M. J., Segler, M., and Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis dags. *Advances in neural information processing systems*, 33:6852–6866, 2020.

[5] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[6] Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[7] Bussi, G. and Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020.

[8] Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., and Blundell, T. L. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6(6):589–604, 2024.

[9] Ertl, P. and Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1: 1–11, 2009.

[10] Felten, F., Alegre, L. N., Nowe, A., Bazzan, A., Talbi, E. G., Danoy, G., and C da Silva, B. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23671–23700, 2023.

[11] Ferreira, L. G., Dos Santos, R. N., Oliva, G., and Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

[12] Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[13] Fromer, J. C. and Coley, C. W. Computer-aided multi-objective optimization in small molecule discovery. *Patterns*, 4(2), 2023.

[14] Fu, T., Gao, W., Xiao, C., Yasonik, J., Coley, C. W., and Sun, J. Differentiable scaffolding tree for molecular optimization. *arXiv preprint arXiv:2109.10469*, 2021.

[15] Gao, W., Mercado, R., and Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. *arXiv preprint arXiv:2110.06389*, 2021.

[16] Gao, W., Fu, T., Sun, J., and Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems*, 35:21342–21357, 2022.

[17] Ghorbani, M., Gendelev, L., Beroza, P., and Keiser, M. J. Autoregressive fragment-based diffusion for pocket-aware ligand design. *arXiv preprint arXiv:2401.05370*, 2023.

[18] Gu, S., Xu, M., Powers, A., Nie, W., Geffner, T., Kreis, K., Leskovec, J., Vahdat, A., and Ermon, S. Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in Neural Information Processing Systems*, 37:44040–44063, 2024.

[19] Guo, J. and Schwaller, P. Augmented memory: sample-efficient generative molecular design with reinforcement learning. *Jacs Au*, 4(6):2160–2172, 2024.

[20] Guo, J., Fialková, V., Arango, J. D., Margreitter, C., Janet, J. P., Papadopoulos, K., Engkvist, O., and Patronov, A. Improving de novo molecular design with curriculum learning. *Nature Machine Intelligence*, 4(6):555–563, 2022.

[21] Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

[22] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

[23] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7): 1757–1768, 2012.

[24] Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

[25] Jeon, W. and Kim, D. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific Reports*, 10, 12 2020.

[26] Jin, W., Barzilay, R., and Jaakkola, T. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, pp. 4839–4848. PMLR, 2020.

[27] Jin, W., Barzilay, R., and Jaakkola, T. Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning*, pp. 4849–4859. PMLR, 2020.

[28] Jo, J., Lee, S., and Hwang, S. J. Score-based generative modeling of graphs via the system of stochastic differential equations. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[29] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Kong, X., Huang, W., Tan, Z., and Liu, Y. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35:2550–2563, 2022.

[31] Krishnan, A., Valeri, J. A., Jin, W., Donghia, N. M., Sieben, L., Luttens, A., Zhang, Y., Modaresi, S. M., Hennes, A., Fromer, J., et al. A generative deep learning approach to de novo antibiotic design. *Cell*, 2025.

[32] Landrum, G. et al. Rdkit: Open-source cheminformatics software. 2016. *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit*, 149(150):650, 2016.

[33] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[34] Lee, S., Jo, J., and Hwang, S. J. Exploring chemical space with score-based out-of-distribution generation. *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[35] Lee, S., Kreis, K., Veccham, S. P., Liu, M., Reidenbach, D., Paliwal, S., Vahdat, A., and Nie, W. Molecule generation with fragment retrieval augmentation. *arXiv preprint arXiv:2411.12078*, 2024.

[36] Lee, S., Kreis, K., Veccham, S. P., Liu, M., Reidenbach, D., Peng, Y., Paliwal, S., Nie, W., and Vahdat, A. Genmol: A drug discovery generalist with discrete diffusion. *arXiv preprint arXiv:2501.06158*, 2025.

[37] Lu, H., Herman, D., and Yu, Y. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Maziarz, K., Jackson-Flux, H., Cameron, P., Sirockin, F., Schneider, N., Stiefl, N., Segler, M., and Brockschmidt, M. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.

[39] Nigam, A., Pollice, R., Krenn, M., dos Passos Gomes, G., and Aspuru-Guzik, A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical science*, 12(20):7079–7090, 2021.

[40] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

[41] Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[42] Pandey, M., Subbaraj, G., Cherkasov, A., Ester, M., and Bengio, E. Pretraining generative flow networks with inexpensive rewards for molecular graph generation. *arXiv preprint arXiv:2503.06337*, 2025.

[43] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[44] Reymond, M., Bargiacchi, E., and Nowé, A. Pareto conditioned networks. *arXiv preprint arXiv:2204.05036*, 2022.

[45] Roijers, D., Steckelmacher, D., and Nowé, A. Multi-objective reinforcement learning for the expected utility of the return. In *Adaptive Learning Agents Workshop 2018*, 2018.

[46] Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

[47] Tripp, A., Simm, G. N., and Hernández-Lobato, J. M. A fresh look at de novo molecular design benchmarks. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[48] Van Moffaert, K. and Nowé, A. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

[49] Van Moffaert, K., Drugan, M. M., and Nowé, A. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pp. 191–199. IEEE, 2013.

[50] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[51] Wang, H., Skreta, M., Ser, C.-T., Gao, W., Kong, L., Strieth-Kalthoff, F., Duan, C., Zhuang, Y., Yu, Y., Zhu, Y., et al. Efficient evolutionary search over chemical space with large language models. *arXiv preprint arXiv:2406.16976*, 2025.

[52] Wang, M. Y. Deep graph library: Towards efficient and scalable deep learning on graphs. In *ICLR workshop on representation learning on graphs and manifolds*, 2019.

[53] Wang, Z., Nie, W., Qiao, Z., Xiao, C., Baraniuk, R., and Anandkumar, A. Retrieval-based controllable molecule generation. In *International Conference on Learning Representations*, 2023.

[54] Watson, L. T. and Haftka, R. T. Modern homotopy methods in optimization. *Computer Methods in Applied Mechanics and Engineering*, 74(3):289–305, 1989.

[55] Xie, Y., Shi, C., Zhou, H., Yang, Y., Zhang, W., Yu, Y., and Li, L. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

[56] Xie, Y., Xu, Z., Ma, J., and Mei, Q. How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules. *International Conference on Learning Representations*, 2023.

[57] Xu, J., Tian, Y., Ma, P., Rus, D., Sueda, S., and Matusik, W. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pp. 10607–10616. PMLR, 2020.

[58] Yang, R., Sun, X., and Narasimhan, K. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

[59] Yang, S., Hwang, D., Lee, S., Ryu, S., and Hwang, S. J. Hit and lead discovery with explorative rl and fragment-based molecule generation. *Advances in Neural Information Processing Systems*, 34:7924–7936, 2021.

[60] Zhang, X., Belousoff, M. J., Zhao, P., Kooistra, A. J., Truong, T. T., Ang, S. Y., Underwood, C. R., Egebjerg, T., Šenel, P., Stewart, G. D., et al. Differential glp-1r binding and activation by peptide and non-peptide agonists. *Molecular Cell*, 80(3):485–500, 2020.

[61] Zhang, Z. and Liu, Q. Learning subpocket prototypes for generalizable structure-based drug design. In *International Conference on Machine Learning*, pp. 41382–41398. PMLR, 2023.

[62] Zhou, Z., Kearnes, S., Li, L., Zare, R. N., and Riley, P. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.

[63] Zhu, Y., Wu, J., Hu, C., Yan, J., Hou, T., Wu, J., et al. Sample-efficient multi-objective molecular optimization with gflownets. *Advances in Neural Information Processing Systems*, 36:79667–79684, 2023.

A Algorithms

Algorithm 1: Energy-Based Fragment Scoring

Input: Molecular dataset $\mathcal{D}$, property labels $\mathcal{Y}$, batch size B, learning rate η
Initialize energy network E_θ parametrized by GNN, and empty vocabulary $\mathcal{V} \leftarrow \emptyset$;
// Phase 1: Energy Landscape Learning
repeat
  Sample batch of positive pairs $B_{\text{real}} = \{(G_i, y_i)\}_{i=1}^{B}$ from $\mathcal{D}$;
  for *each sample* (G, y) *in* B_{real} **do**
    Sample noisy property $\tilde{y} \sim p_{\text{uniform}}(\mathcal{Y})$;
    Generate corrupted graph $\tilde{G}$ by masking random node features: $\mathbf{H}_{\tilde{G}} \leftarrow \mathbf{H}_G \odot \mathbf{M}_{\text{rand}}$;
    Compute energy for real and noisy pairs:;
    $E_{\text{real}} \leftarrow E_\theta(G, y)$; $E_{\text{noise}} \leftarrow E_\theta(\tilde{G}, \tilde{y})$;
    Compute InfoNCE loss to shape the landscape:;
    $\mathcal{L}_{\text{NCE}} \leftarrow \log\left(1 + \exp(E_{\text{real}} - E_{\text{noise}})\right)$;
  Update parameters θ by minimizing the batch loss: $\theta \leftarrow \theta - \eta \nabla_\theta \sum \mathcal{L}_{\text{NCE}}$;
until *Convergence*;
// Phase 2: Goal-Aware Fragment Extraction
Filter high-activity molecules $\mathcal{D}_{\text{high}} \subset \mathcal{D}$;
for *each molecule* $G \in \mathcal{D}_{high}$ **do**
  Decompose G into candidate fragments $\{F_1, \ldots, F_k\}$ (e.g., via BRICS);
  for *each fragment* F_j **do**
    Calculate Importance Score (Energy Gap):;
    $\mathcal{I}(F_j) \leftarrow E_\theta(G_{\setminus F_j}, y_{\text{high}}) - E_\theta(G, y_{\text{high}})$;
    if $\mathcal{I}(F_j) > \tau$ **then**
      Add fragment to vocabulary: $\mathcal{V} \leftarrow \mathcal{V} \cup \{F_j\}$;
Output: Goal-Aware Fragment Vocabulary $\mathcal{V}$

Algorithm 2: Envelope Soft Actor-Critic

Input: Preference sampling distribution Ω, batch size B, number of sampled preferences K
Initialize policy π_ϕ, critics $\boldsymbol{Q}_{\theta_1,\theta_2}$, target critics $\boldsymbol{Q}_{\bar{\theta}_1,\bar{\theta}_2}$, and replay buffer $\mathcal{D}$;
repeat
  Interact with the environment using $\pi_\phi(\cdot|s, \boldsymbol{\omega}_{\text{env}})$, and store transitions $(s, a, \boldsymbol{r}, s', d)$ in $\mathcal{D}$;
  for *each gradient step* **do**
    Sample minibatch $\mathcal{B} = \{(s, a, \boldsymbol{r}, s', d)\}$ from $\mathcal{D}$ and preferences $\{\boldsymbol{\omega}_i\}_{i=1}^{K} \sim \Omega$;
    // Critic Update
    Generate candidate actions for next state: $\mathcal{A}_{\text{cand}} = \{a'_k \sim \pi_\phi(\cdot|s', \boldsymbol{\omega}_k)\}_{k=1}^{K}$;
    for *each preference* $\boldsymbol{\omega}_i$ **do**
      Select envelope action index:
      $k^* \leftarrow \arg\max_{k \in \{1,\ldots,K\}} \left\langle \boldsymbol{\omega}_i,\ \min_{j=1,2} \boldsymbol{Q}_{\bar{\theta}_j}(s', a'_k, \boldsymbol{\omega}_i) \right\rangle$;
      Compute envelope target vector:
      $\boldsymbol{y}_i \leftarrow \boldsymbol{r} + \gamma(1-d)\ \min_{j=1,2} \boldsymbol{Q}_{\bar{\theta}_j}(s', a'_{k^*}, \boldsymbol{\omega}_i)$;
    Update critics $\theta_{1,2}$ by minimizing the homotopy loss:

$$\mathcal{L}_Q = \mathbb{E}_{\mathcal{B},\boldsymbol{\omega}}\Big[\beta\big(\langle \boldsymbol{\omega}, \boldsymbol{Q}_\theta(s,a,\boldsymbol{\omega})\rangle - \langle \boldsymbol{\omega}, \boldsymbol{y}\rangle\big)^2 + (1-\beta)\|\boldsymbol{Q}_\theta(s,a,\boldsymbol{\omega}) - \boldsymbol{y}\|_2^2\Big]$$

    // Policy Update
    Sample actions $\tilde{a} \sim \pi_\phi(\cdot|s, \boldsymbol{\omega})$ using reparameterization;
    Update policy ϕ by minimizing:
    $\mathcal{L}_\pi = \mathbb{E}_{\mathcal{B},\boldsymbol{\omega}}\Big[\alpha \log \pi_\phi(\tilde{a}|s,\boldsymbol{\omega}) - \min_{j=1,2} \langle \boldsymbol{\omega}, \boldsymbol{Q}_{\theta_j}(s,\tilde{a},\boldsymbol{\omega})\rangle\Big]$
    Update target critics: $\bar{\theta}_j \leftarrow \rho\bar{\theta}_j + (1-\rho)\theta_j$ for $j = 1, 2$;
until *Convergence*;

B Experimental Settings

B.1 Computing Infrastructures

Software infrastructures. All of the experiments are implemented in Python 3.10, with the following supporting libraries: PyTorch 2.1.0 [43], PyG 2.0.4 [12], DGL 2.4.0 [52], Gym 0.21.0 [5], Open Babel 3.1.1 [40], PLIP 2.3.1 [46], and RDKit 2022.03.1 [32].

Hardware infrastructures. We conduct all experiments on a computer server with 8 NVIDIA A800 GPUs and 2 Intel Xeon Gold 6348 CPUs (with 56 cores and 112 threads).

B.2 Implementation Details

Following [59, 16, 35], we base our experimental setup on the ZINC250k dataset [23], using the standard split for training and testing. For our generative model, the training process begins with an initial fragment vocabulary of 300 fragments, which is dynamically expanded to a maximum size of 1000 by adding up to 50 new fragments per cycle. Any fragments that cause RDKit sanitization errors are filtered out.

The fragment-based generator employs Envelope SAC for training, where episode starts from a benzene ring scaffold with three attachment points (*ortho-*, *meta-*, *para-*). An episode terminates if the heavy atom count of molecule exceeds 40. To ensure sufficient exploration, the policy model samples molecules randomly for the first 4,000 steps to populate the replay buffer, while parameter updates commence after 3,000 steps. For fair evaluation, molecules generated during this initial exploration phase are included in the final results. The fragment modification module utilizes a genetic algorithm with a population size of 100 and a mutation rate of 0.1.

All neural networks use a hidden dimension of 64 and are trained with a batch size of 256. Both the value and policy networks are optimized using the Adam optimizer [29] with a learning rate of 1×10^{-4}. For exploration in our multi-objective setting, we sample 5 preference vectors per batch. Moreover, we observe that as the number of objectives increases, the training benefits from a higher ratio of off-policy updates (i.e., utilizing more samples from the replay buffer relative to on-policy samples). We hypothesize that in high-dimensional objective spaces, the model requires more frequent re-visitation of past action episodes to effectively learn the complex trade-offs and stabilize convergence.

Oracle budget and runtime. Each target-based method is evaluated with the same budget of 3,000 completed molecules and docking calls. Docking is invoked only after a molecule is completed; the resulting pose is reused by PLIP, so hotspot profiling requires no additional docking call. On the same server, representative training times are 2.3, 9.7, 11.6, 12.4, and 13.2 hours for REINVENT, FREED, MOOD, GEAM, and `BindMol`, respectively. REINVENT benefits from its lightweight RNN, while `BindMol` remains comparable to fragment-based baselines. Preference vectors are replicated across the batch for parallel actor/critic evaluation rather than processed in a Python loop. With fixed poses, multi-threaded PLIP processes 10^4 molecules within minutes, so its overhead is small relative to docking.

B.3 Evaluation Protocols

We adopt several well-defined metrics to evaluate our framework:

- For binding affinity optimization, we evaluate the generated molecules following [59]. A molecule is considered a *hit* if it achieves a docking score lower than the median of experimentally known actives, while also satisfying QED > 0.5 and SA < 5. To assess affinity optimization, we report the ***Novel top 5% Docking Score*** (kcal/mol), which is the average docking score of the best-performing 5% unique and novel hits. We also measure the ***Novel hit ratio***, defined as the fraction of hits that are novel.
- For rediscovery tasks, we report the area under the curve (AUC) of top-K average property value versus the number of oracle calls (***AUC top-K***) following [16] for fair comparison.
- We further report ***#Circles*** [56] for both of the tasks, which quantifies the coverage of chemical space by the generated hits.

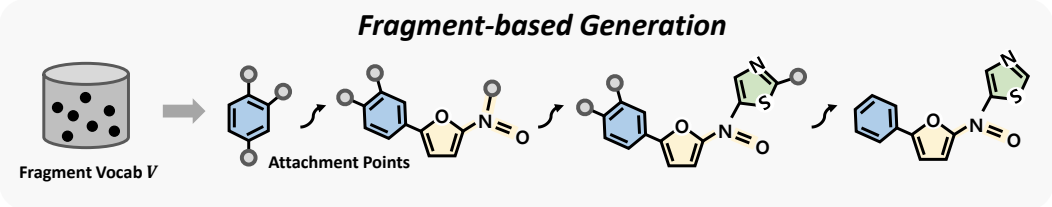


Figure 6: Fragment-based generation sequentially attaches fragments from a fragment vocabulary to a partial graph until termination.

B.4 Baselines

Binding affinity tasks. To comprehensively evaluate the effectiveness of our approach, we select 11 competitive molecular generation models as baselines: REINVENT [41], Graph GA [24], MORLD [25], HierVAE [26], RationaleRL [27], FREED [59], PS-VAE [30], MOOD [34], GDSS [28], RetMol [53], GEAM [35] and MOLLEO [51]. These baselines span both SMILES- and graph-based modalities, and encompass diverse training paradigms including distribution learning, genetic algorithms, and reinforcement learning. Note that our model is trained from scratch, rather than fine-tuned from a large pretrained molecular generative model. To ensure a fair and controlled comparison under this setting, we do not include approaches that rely heavily on large-scale pretraining [42, 36].

Multi-property tasks in GuacaMol. We select 12 molecular generation models with strong performance on the GuacaMol benchmark as baselines: REINVENT [41], MolDQN [62], Graph GA [24], Graph MCTS [24], DoG-Gen [4], GPBO [47], STONED [39], SynNet [15], DST [14], GEAM [35], Augmented Memory [19], and MOLLEO (MoleculeSTM) [51]. These methods also span various molecular modalities and training strategies.

C Spatial Clustering of Hotspot Residues

Given the non-overlapping residue set $\mathcal{V}$ defined in Section 4.2, our goal is to further partition these residues into a set of spatially coherent subpockets that capture distinct local interaction patterns within the binding site. The clustering is performed purely based on the three-dimensional geometry of residues, without relying on ligand-specific distance thresholds. Each residue $v \in \mathcal{V}$ is represented by a three-dimensional centroid $\mathbf{c}_v$, computed from its side-chain atoms to emphasize interaction-relevant geometry. Backbone atoms (N, C_α, C, O) are excluded whenever side-chain atoms are available; for residues lacking side chains, the centroid is computed over all atoms of the residue. This representation yields a set of residue-level points embedded in $\mathbb{R}^3$.

We construct a pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, where each entry

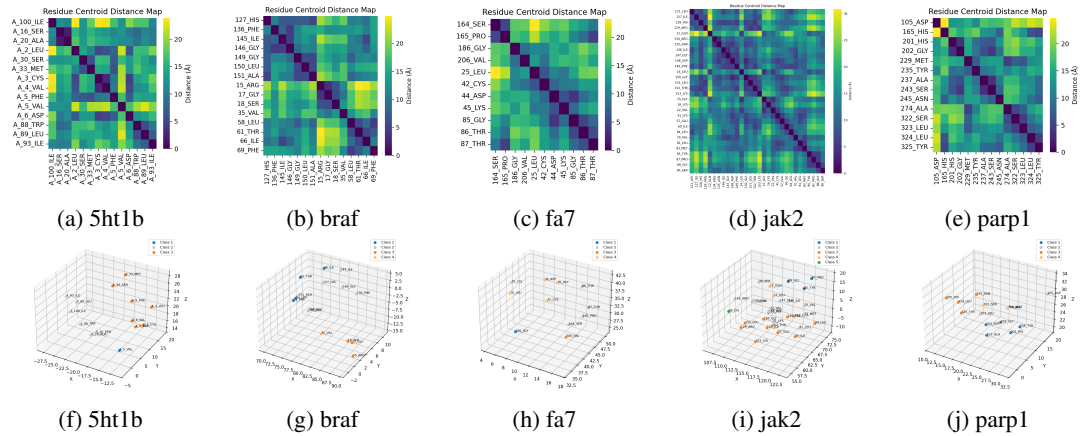
$$D_{uv} = \|\mathbf{c}_u - \mathbf{c}_v\|_2, \qquad u, v \in \mathcal{V},$$

quantifies the spatial proximity between residues. Based on this distance matrix, we define a Gaussian affinity graph with weights

$$W_{uv} = \exp\left(-\frac{D_{uv}^2}{\sigma^2}\right),$$

where σ controls the spatial scale at which residue interactions are considered local. The diagonal entries of $\mathbf{W}$ are set to zero. To identify spatially separated but internally compact residue groups, we apply spectral clustering to the affinity graph. Specifically, we construct the unnormalized graph Laplacian $\mathbf{L} = \mathbf{D}_W - \mathbf{W}$, where $\mathbf{D}_W$ is the diagonal degree matrix of $\mathbf{W}$. The eigenvectors corresponding to the m smallest eigenvalues of $\mathbf{L}$ are extracted and row-normalized to form a low-dimensional spectral embedding of residues in $\mathcal{V}$.

Finally, k-means clustering is performed on this embedding space to partition $\mathcal{V}$ into m disjoint subsets $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$, each representing a spatial hotspot within the binding pocket. These subpockets serve as the structural basis for defining preference-conditioned reward components in Section 4.2. The visualization of distance matrices and clustered hotspots are provided in Figure 7.



(a) 5ht1b (b) braf (c) fa7 (d) jak2 (e) parp1

(f) 5ht1b (g) braf (h) fa7 (i) jak2 (j) parp1

Figure 7: Visualization results for five protein targets. The first row shows residue distance matrices, and the second row shows clustered residues (hotspots) for each target.

D Fragment-based Generator

Following [59], we frame the molecular generation as a sequential decision process. The vivid illustration of this fragment-based generator is provided in Figure 6.At time step t, the state is the partially built graph $\mathcal{G}_t$. Each transition is achieved by a triplet action

$$\mathbf{a}_t = (u_t, f_t, v_t),$$

where $u_t \in \mathcal{U}(\mathcal{G}_t)$ is an attachment site in the current graph, $f_t \in \mathcal{S}$ is a fragment drawn from the vocabulary $\mathcal{S}$, and $v_t \in \mathcal{V}(f_t)$ is an attachment site on the chosen fragment. Attaching f_t to $\mathcal{G}_t$ at (u_t, v_t) yields $\mathcal{G}_{t+1}$.

Formally, we define three action components governed by sub-policies:

$$\mathbb{P}_1(u_t \mid \mathcal{G}_t) = \Pi_1\big(\mathcal{H}(\mathcal{G}_t), \mathcal{H}_{\text{site}}(\mathcal{G}_t)\big), \tag{11}$$

$$\mathbb{P}_2(f_t \mid u_t, \mathcal{G}_t) = \Pi_2\big(\phi(u_t), \Phi(\mathcal{S})\big), \tag{12}$$

$$\mathbb{P}_3(v_t \mid f_t, u_t, \mathcal{G}_t) = \Pi_3\big(\psi(f_t), \psi_{\text{site}}(f_t)\big), \tag{13}$$

where $\mathcal{H}(\mathcal{G}_t) \in \mathbb{R}^d$ is a global embedding of the partial graph, $\mathcal{H}_{\text{site}}(\mathcal{G}_t) = \{h_u : u \in \mathcal{U}(\mathcal{G}_t)\}$ gives embeddings of candidate sites, and $\phi(u_t) := h_{u_t}$ denotes the embedding of the chosen site. The fragment vocabulary is represented by $\Phi(\mathcal{S}) = \{\phi_b : b \in \mathcal{S}\}$, a collection of precomputed fragment descriptors. Each fragment f is further encoded as a global embedding $\psi(f) \in \mathbb{R}^d$ and its attachment sites as $\psi_{\text{site}}(f) = \{g_v^{(f)} : v \in \mathcal{V}(f)\}$.

Here Π_1, Π_2, Π_3 are neural scoring functions that map their inputs to logits; applying a softmax yields categorical distributions $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3$ over the respective action spaces. The overall policy can be factorized as

$$\pi(\mathbf{a}_t \mid \mathcal{G}_t) = \mathbb{P}_1(u_t \mid \mathcal{G}_t)\, \mathbb{P}_2(f_t \mid u_t, \mathcal{G}_t)\, \mathbb{P}_3(v_t \mid f_t, u_t, \mathcal{G}_t). \tag{14}$$

Inspired by [35], we introduce new fragments by applying graph genetic operations (mutation and crossover) on generated molecules [24] and incorporating novel substructures back into $\mathcal{S}$. When the vocabulary grows too large, only the top-ranked fragments are retained.

E Theoretical Proofs and Analysis

This appendix provides a comprehensive theoretical treatment of the energy-based fragment scoring framework introduced in Section 4.1. Our goal is to move beyond heuristic feature importance by grounding our scoring mechanism in the rigorous principles of differential geometry and calculus on manifolds.

Specifically, we address a fundamental question: *How does the discrete removal of a chemical substructure relate to the continuous gradient information learned by the neural network?* To answer

this, we model the ablation process as a trajectory through the feature space and analyze the energy dynamics along this path. We rigorously establish that our discrete ablation score is mathematically equivalent to the path integral of gradients (Integrated Gradients), thereby satisfying the axiom of completeness in attribution theory. Furthermore, we provide error bounds to justify the stability of this metric against local perturbations.

E.1 Preliminaries: Geometry of the Feature Space

To formalize our analysis, we treat the molecular representation not merely as a matrix of numbers, but as a vector in a high-dimensional Hilbert space.

Let $\mathcal{H} = \mathbb{R}^{|\mathcal{V}| \times d}$ denote the Hilbert space of node features equipped with the Frobenius inner product. For any two feature matrices $\mathbf{A}, \mathbf{B} \in \mathcal{H}$, the inner product is defined as:

$$\langle \mathbf{A}, \mathbf{B} \rangle_F := \text{Tr}(\mathbf{A}^\top \mathbf{B}) = \sum_{i,j} A_{ij} B_{ij}. \tag{15}$$

The induced norm $\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle_F}$ provides a measure of the signal magnitude. The energy function $E_\theta(\mathbf{H}, y)$ maps a point in this feature space (conditioned on property y) to a scalar potential. The gradient $\nabla_{\mathbf{H}} E_\theta(\mathbf{H}, y)$ serves as a vector field over $\mathcal{H}$, indicating the direction of steepest energy ascent.

Definition 1 (Fragment Subspace Projection)**.** Given a fragment $\mathcal{F} \subseteq \mathcal{V}$, we define the binary indicator mask $\mathbf{1}_\mathcal{F} \in \{0, 1\}^{|\mathcal{V}| \times d}$ such that entries corresponding to nodes in $\mathcal{F}$ are 1, and 0 otherwise. Geometrically, the operation $\mathbf{1}_\mathcal{F} \odot \mathbf{H}$ represents the orthogonal projection of the molecule's features onto the subspace spanned by the fragment. Conversely, the ablation is a projection onto the orthogonal complement, realized by $(\mathbf{1} - \mathbf{1}_\mathcal{F}) \odot \mathbf{H}$.

E.2 Proof of Theorem 1: Manifold Gradient Equivalence

One might intuitively expect that the importance of a fragment is proportional to the gradients of its atoms. However, in deep non-linear networks, gradients can saturate or vanish. The "Energy Gap" method we propose (computing $E_{\text{original}} - E_{\text{ablated}}$) captures the global effect of the fragment. Theorem 1 bridges the gap between this global discrete difference and the local continuous gradients.

Theorem 1. *Let $\gamma : [0, 1] \to \mathbb{R}^{|\mathcal{V}| \times d}$ be a linear path parameterizing the feature ablation from $\mathbf{H}$ to $\mathcal{P}_{\mathcal{F}^\perp}(\mathbf{H})$. Assuming E_θ is continuously differentiable, the discrete energy variation $\mathcal{I}(\mathcal{F})$ is exactly the path integral of the directional derivatives along γ:*

$$\mathcal{I}(\mathcal{F}) = -\int_0^1 \langle \nabla_{\mathbf{H}} E_\theta(\gamma(t), y),\, \mathbf{1}_\mathcal{F} \odot \mathbf{H} \rangle_F \, dt. \tag{16}$$

E.2.1 Proof Derivation

The proof relies on constructing a virtual trajectory connecting the intact molecule to the ablated version and applying the Fundamental Theorem of Calculus along this path.

Lemma 1 (Path Dynamics)**.** *We define the ablation path $\gamma(t)$ connecting the original molecule state $\mathbf{H}$ (at $t = 0$) to the ablated state (at $t = 1$) as:*

$$\gamma(t) = \mathbf{H} - t(\mathbf{1}_\mathcal{F} \odot \mathbf{H}), \quad t \in [0, 1]. \tag{17}$$

The velocity vector (tangent) of this path is constant and points in the negative direction of the fragment features:

$$\frac{d}{dt}\gamma(t) = -(\mathbf{1}_\mathcal{F} \odot \mathbf{H}). \tag{18}$$

Proof of Lemma 1. By definition, the ablated state is $\mathbf{H} - \mathbf{1}_\mathcal{F} \odot \mathbf{H}$. The linear interpolation is:

$$\gamma(t) = (1 - t)\mathbf{H} + t(\mathbf{H} - \mathbf{1}_\mathcal{F} \odot \mathbf{H}) \tag{19}$$

$$= \mathbf{H} - t(\mathbf{1}_\mathcal{F} \odot \mathbf{H}). \tag{20}$$

Differentiating strictly with respect to the scalar parameter t yields the tangent vector $\gamma'(t) = -(\mathbf{1}_\mathcal{F} \odot \mathbf{H})$. This confirms the path is a straight line in $\mathcal{H}$. □

Proof of Theorem 1:

Let $g(t) = E_\theta(\gamma(t), y)$ represent the scalar energy evaluated along the ablation trajectory. Since E_θ is a composition of smooth functions (neural network layers), it is differentiable. We apply the Chain Rule to analyze the rate of energy change. The derivative of the composite function $g(t)$ is the inner product of the gradient of the outer function (the energy field) and the derivative of the inner function (the path velocity):

$$\frac{dg}{dt} = \left\langle \nabla_{\gamma(t)} E_\theta(\gamma(t), y),\ \frac{d\gamma}{dt} \right\rangle_F . \tag{21}$$

Substituting the result from Lemma 1, the rate of change becomes:

$$\frac{dg}{dt} = \langle \nabla_{\mathbf{H}} E_\theta(\gamma(t), y),\ -(\mathbf{1}_{\mathcal{F}} \odot \mathbf{H}) \rangle_F . \tag{22}$$

By the linearity of the inner product, we pull the negative sign out:

$$\frac{dg}{dt} = -\langle \nabla_{\mathbf{H}} E_\theta(\gamma(t), y),\ \mathbf{1}_{\mathcal{F}} \odot \mathbf{H} \rangle_F . \tag{23}$$

The total change in energy, which defines our importance score $\mathcal{I}(\mathcal{F})$, is obtained by accumulating these infinitesimal changes from the initial state to the final ablated state:

$$\mathcal{I}(\mathcal{F}) = E_\theta(\gamma(1), y) - E_\theta(\gamma(0), y) \tag{24}$$

$$= \int_0^1 \frac{dg}{dt}\, dt \tag{25}$$

$$= -\int_0^1 \langle \nabla_{\mathbf{H}} E_\theta(\gamma(t), y),\ \mathbf{1}_{\mathcal{F}} \odot \mathbf{H} \rangle_F \, dt. \tag{26}$$

Remark 1 (Theoretical Implication). The integral form derived above corresponds to *Integrated Gradients (IG)* applied to the fragment subspace. Theorem 1 proves that our "Energy Gap" score is not a heuristic; it is an exact implementation of IG. This ensures that our method satisfies the *Completeness Axiom*: the sum of attributions equals the total function output difference. Consequently, our method correctly accounts for non-linear saturation effects that would otherwise mislead a simple gradient-based analysis.

E.3 Error Analysis and Bounds

While Theorem 1 provides an exact equality, computing the integral requires evaluating the neural network multiple times along the path, which can be computationally expensive. In practice, one might approximate this using only the gradient at the initial point (the original molecule). Here, we analyze the theoretical error introduced by such a first-order approximation. This analysis tells us how "trustworthy" the local gradients are for estimating global fragment importance.

Assumption 1 (Lipschitz Smoothness)**.** We assume the energy landscape is smooth. Specifically, the gradient $\nabla_{\mathbf{H}} E_\theta$ is L-Lipschitz continuous over the domain $\mathcal{H}$. This means there exists a constant $L > 0$ representing the maximum curvature of the energy surface, such that for any $\mathbf{A}, \mathbf{B} \in \mathcal{H}$:

$$\|\nabla_{\mathbf{H}} E_\theta(\mathbf{A}, y) - \nabla_{\mathbf{H}} E_\theta(\mathbf{B}, y)\|_F \leq L \|\mathbf{A} - \mathbf{B}\|_F. \tag{27}$$

Theorem 2 (Lagrangian Error Bound)**.** *The deviation of the exact discrete importance score from its first-order Taylor approximation is quadratically bounded by the energy norm of the fragment features. Specifically:*

$$|\mathcal{I}(\mathcal{F}) - (-\langle \nabla_{\mathbf{H}} E_\theta(\mathbf{H}, y), \mathbf{1}_{\mathcal{F}} \odot \mathbf{H} \rangle_F)| \leq \frac{L}{2} \|\mathbf{1}_{\mathcal{F}} \odot \mathbf{H}\|_F^2. \tag{28}$$

Proof. Let $\Delta\mathbf{H} = -(\mathbf{1}_{\mathcal{F}} \odot \mathbf{H})$. By the fundamental theorem of calculus derived in the previous section:

$$\mathcal{I}(\mathcal{F}) = \int_0^1 \langle \nabla E_\theta(\mathbf{H} + t\Delta\mathbf{H}, y), \Delta\mathbf{H} \rangle_F \, dt. \tag{29}$$

We aim to compare this to the linear approximation $\langle \nabla E_\theta(\mathbf{H}, y), \Delta\mathbf{H} \rangle_F$. To do so, we add and subtract the gradient at the origin point inside the integral:

$$\mathcal{I}(\mathcal{F}) = \int_0^1 \langle \nabla E_\theta(\mathbf{H}, y), \Delta\mathbf{H} \rangle_F \, dt + \int_0^1 \langle \nabla E_\theta(\mathbf{H} + t\Delta\mathbf{H}, y) - \nabla E_\theta(\mathbf{H}, y), \Delta\mathbf{H} \rangle_F \, dt. \tag{30}$$

The first term is constant with respect to t and simplifies to the linear approximation:

$$\int_0^1 \langle \nabla E_\theta(\mathbf{H}, y), -(\mathbf{1}_\mathcal{F} \odot \mathbf{H}) \rangle_F \, dt = -\langle \nabla E_\theta(\mathbf{H}, y), \mathbf{1}_\mathcal{F} \odot \mathbf{H} \rangle_F. \tag{31}$$

Let $\mathcal{R}$ denote the second term (the remainder/error). We bound $|\mathcal{R}|$ using the Cauchy-Schwarz inequality and Assumption 1:

$$|\mathcal{R}| = \left| \int_0^1 \langle \nabla E_\theta(\mathbf{H} + t\Delta\mathbf{H}, y) - \nabla E_\theta(\mathbf{H}, y), \Delta\mathbf{H} \rangle_F \, dt \right| \tag{32}$$

$$\leq \int_0^1 \|\nabla E_\theta(\mathbf{H} + t\Delta\mathbf{H}, y) - \nabla E_\theta(\mathbf{H}, y)\|_F \, \|\Delta\mathbf{H}\|_F \, dt \tag{33}$$

$$\leq \int_0^1 L\|t\Delta\mathbf{H}\|_F \|\Delta\mathbf{H}\|_F \, dt \tag{34}$$

$$= L\|\Delta\mathbf{H}\|_F^2 \int_0^1 t \, dt. \tag{35}$$

Computing the standard scalar integral $\int_0^1 t \, dt = \frac{1}{2}$, and substituting $\Delta\mathbf{H}$, we obtain the final bound:

$$|\mathcal{R}| \leq \frac{L}{2} \|\mathbf{1}_\mathcal{F} \odot \mathbf{H}\|_F^2. \tag{36}$$

□

Interpretation: The error bound $\frac{L}{2}\|\mathbf{1}_\mathcal{F} \odot \mathbf{H}\|_F^2$ implies that the linear approximation is most accurate for (1) "flat" energy landscapes where L is small, or (2) fragments with small feature norms. For large, complex fragments, the non-linearity (captured by $\mathcal{R}$) becomes significant, further justifying our use of the exact Energy Gap method rather than simple gradient maps.

E.4 Consistency of Gradient-Based Ranking

Finally, we utilize the previous results to justify the consistency between the rigorous ablation score and simpler gradient-based metrics often used in saliency maps. In many practical scenarios, we are interested in the *ranking* of fragments rather than their exact energy contribution.

Corollary 1 (Ranking Alignment under Homogeneity)**.** *Suppose the fragment features satisfy a local homogeneity condition where the gradient is aligned with the feature vector. Then, ranking fragments by the magnitude of the energy gap $|\mathcal{I}(\mathcal{F})|$ is asymptotically equivalent to ranking by the summed gradient norms.*

Proof. From Theorem 2, if the remainder term is small relative to the linear term, we have:

$$|\mathcal{I}(\mathcal{F})| \approx |\langle \nabla_\mathbf{H} E_\theta(\mathbf{H}, y), \mathbf{1}_\mathcal{F} \odot \mathbf{H} \rangle_F| \, . \tag{37}$$

This term can be decomposed into individual node contributions:

$$\left| \sum_{i \in \mathcal{F}} \nabla_{\mathbf{x}_i} E_\theta(\mathbf{H}, y)^\top \mathbf{x}_i \right| . \tag{38}$$

Assuming alignment ($\nabla^\top \mathbf{x} \propto \|\nabla\|$), this is proportional to $\sum_{i \in \mathcal{F}} \|\nabla_{\mathbf{x}_i} E_\theta\|_2$. Thus, the computationally expensive ablation score $\mathcal{I}(\mathcal{F})$ and the efficient gradient norm score induce the same ordering of fragments. This result provides theoretical grounding for using gradient-based screening as a preliminary filter before precise energy-based scoring. □

F Limitations and future works

While BindMol demonstrates strong performance in exploring diverse binding modes, our current reward formulation relies on simplified interaction counts and static docking poses. First, treating all interaction types equally—without weighting them by energetic contribution—prioritizes interaction diversity over fine-grained affinity modeling. Assigning biochemically accurate weights is non-trivial and target-dependent, yet future iterations could leverage learnable or physics-based weighting schemes to better capture residue-specific importance. Second, like most structure-based generative models, our reliance on rapid static docking (e.g., QVina) may occasionally reward suboptimal or physically strained poses. The PLIP-IFP analyses therefore quantify *docking-predicted* interaction modes rather than experimentally confirmed binding poses, and IFP cluster counts may retain some sample-size dependence despite the complementary entropy statistics. Crystallographic-mode recovery is limited to BRAF and JAK2 because only these targets provide directly comparable paired mode labels. Future work should incorporate receptor ensembles, rigorous energy minimization, and experimental structural validation.

G Additional Empirical Results

G.1 Interaction-level diversity and crystallographic-mode recovery

For every qualified docked complex, we construct a binary full-pocket PLIP fingerprint whose features are residue–interaction-type pairs. Pairwise Jaccard distances are clustered with the Butina algorithm at distance cutoff 0.5. We report the number of recurring clusters; entropy $H = -\sum_c p_c \log p_c$, where p_c is the fraction of hits in cluster c; and the effective number of modes $\exp(H)$. Cluster count measures breadth, whereas entropy and effective modes also account for population balance. Because cluster count can depend on the number of qualified hits, we interpret all three statistics jointly.

Table 6: **Full-pocket interaction-fingerprint diversity.** Results are means and standard deviations over three runs. The best result for each target is highlighted in bold.

Target	Method	IFP clusters	IFP entropy H	Effective modes $\exp(H)$
PARP1	REINVENT	52.33 ± 4.04	1.62 ± 0.09	5.05 ± 0.45
	FREED	102.67 ± 6.11	2.47 ± 0.08	11.83 ± 0.95
	RationaleRL	96.33 ± 5.51	2.42 ± 0.10	11.25 ± 1.12
	GEAM	138.67 ± 7.02	2.93 ± 0.07	18.73 ± 1.26
	BindMol	$\mathbf{192.00 \pm 6.20}$	$\mathbf{3.45 \pm 0.07}$	$\mathbf{31.40 \pm 0.96}$
FA7	REINVENT	21.33 ± 3.21	0.77 ± 0.09	2.16 ± 0.20
	FREED	44.67 ± 4.51	1.43 ± 0.08	4.18 ± 0.34
	RationaleRL	41.33 ± 4.04	1.38 ± 0.10	3.97 ± 0.40
	GEAM	61.67 ± 5.51	1.85 ± 0.07	6.36 ± 0.45
	BindMol	$\mathbf{94.33 \pm 1.53}$	$\mathbf{2.37 \pm 0.06}$	$\mathbf{10.71 \pm 0.34}$
5HT1B	REINVENT	3.67 ± 0.58	0.07 ± 0.03	1.07 ± 0.03
	FREED	6.33 ± 0.58	0.17 ± 0.04	1.19 ± 0.05
	RationaleRL	6.00 ± 1.00	0.15 ± 0.05	1.16 ± 0.06
	GEAM	8.33 ± 0.58	0.24 ± 0.04	1.27 ± 0.05
	BindMol	$\mathbf{11.67 \pm 0.58}$	$\mathbf{0.32 \pm 0.02}$	$\mathbf{1.37 \pm 0.03}$
BRAF	REINVENT	7.33 ± 1.53	0.18 ± 0.06	1.20 ± 0.07
	FREED	13.67 ± 2.08	0.41 ± 0.08	1.51 ± 0.12
	RationaleRL	12.67 ± 1.53	0.38 ± 0.09	1.46 ± 0.13
	GEAM	20.33 ± 2.52	0.68 ± 0.07	1.97 ± 0.14
	BindMol	$\mathbf{35.00 \pm 1.41}$	$\mathbf{1.11 \pm 0.08}$	$\mathbf{3.03 \pm 0.08}$
JAK2	REINVENT	14.33 ± 2.08	0.39 ± 0.08	1.48 ± 0.12
	FREED	31.67 ± 3.51	0.97 ± 0.09	2.64 ± 0.24
	RationaleRL	28.67 ± 3.06	0.92 ± 0.10	2.51 ± 0.26
	GEAM	47.33 ± 4.04	1.52 ± 0.08	4.57 ± 0.37
	BindMol	$\mathbf{77.33 \pm 2.16}$	$\mathbf{2.31 \pm 0.05}$	$\mathbf{10.10 \pm 0.35}$

For crystallographic-mode recovery, we use BRAF and JAK2 because both provide well-established pairs of type-I/DFG-in and type-II/DFG-out structures within the same target. Specifically, the

references are BRAF 4MNF (DFG-in/type-I) and 5HI2 (DFG-out/type-II), and JAK2 6VGL (type-I) and 7TEU (type-II/DFG-out). The remaining targets do not provide equally clear and directly comparable paired labels. For each reference conformation, we re-dock generated hits into that receptor and compare their PLIP IFPs with the native ligand IFP; similarity ≥ 0.4 defines recovery.

Table 7: **Crystallographic-mode recovery and reference separation.** The left panel reports recovery rates; the right panel compares the two native references for each target using the same full-pocket representation.

Target	Crystal mode	PDB	BindMol	GEAM	FREED
BRAF	DFG-in/type-I	4MNF	**48.9%**	20.3%	11.7%
BRAF	DFG-out/type-II	5HI2	**12.1%**	6.4%	0.0%
JAK2	type-I	6VGL	**1.3%**	0.0%	0.0%
JAK2	type-II/DFG-out	7TEU	**11.5%**	5.7%	2.8%

	BRAF	JAK2
IFP Jaccard	0.375	0.188
Shared IFP features	$6/16$	$3/16$
Residue Jaccard	0.462	0.250
Shared residues	$6/13$	$3/12$

Both native-pair IFP similarities are below the recovery threshold, so recovery does not merely match a conserved interaction core. Absolute rates remain conservative because training uses a fixed receptor while evaluation uses independent crystal structures with induced-fit, side-chain, DFG-loop, and αC-helix differences. These are docking-predicted interaction modes, not experimentally confirmed poses.

G.2 Matched MORL comparisons and hotspot ablations

All MORL baselines use the same fragment action space, vector reward, training protocol, docking-oracle budget, and three random seeds as BindMol.

Table 8: **Matched MORL comparison.** Means and standard deviations over three runs.

Novel hit ratio (%)

Method	PARP1	FA7	5HT1B	BRAF	JAK2
GPI	37.38 ± 1.12	19.84 ± 0.91	37.67 ± 0.74	27.53 ± 1.08	38.31 ± 1.16
C-MORL	38.74 ± 0.95	20.92 ± 0.83	36.46 ± 0.62	27.38 ± 0.89	38.27 ± 1.04
PreCo	38.81 ± 0.88	21.67 ± 0.72	37.93 ± 0.51	26.84 ± 1.03	39.12 ± 0.92
BindMol	$\mathbf{42.62 \pm 0.73}$	$\mathbf{24.58 \pm 0.65}$	$\mathbf{42.15 \pm 0.13}$	$\mathbf{30.77 \pm 0.97}$	$\mathbf{44.86 \pm 0.98}$

Novel top-5% docking score

Method	PARP1	FA7	5HT1B	BRAF	JAK2
GPI	-10.98 ± 0.14	-9.01 ± 0.11	-11.02 ± 0.16	-10.31 ± 0.13	-10.67 ± 0.12
C-MORL	-11.85 ± 0.10	-8.92 ± 0.08	-11.24 ± 0.12	-10.63 ± 0.11	-10.51 ± 0.09
PreCo	-11.24 ± 0.08	-9.17 ± 0.05	-11.61 ± 0.11	-10.88 ± 0.10	-10.73 ± 0.06
BindMol	$\mathbf{-12.93 \pm 0.06}$	$\mathbf{-10.19 \pm 0.02}$	$\mathbf{-13.03 \pm 0.10}$	$\mathbf{-12.50 \pm 0.09}$	$\mathbf{-12.01 \pm 0.02}$

Table 9: **Effect of retaining conserved residues in every hotspot objective.** Novel hit ratio (%) over three runs.

Method	PARP1	FA7	5HT1B	BRAF	JAK2
Full union (including intersection)	35.23 ± 0.63	18.33 ± 0.89	35.21 ± 0.53	24.21 ± 0.96	38.67 ± 0.91
Non-overlapping residues	$\mathbf{42.62 \pm 0.73}$	$\mathbf{24.58 \pm 0.65}$	$\mathbf{42.15 \pm 0.13}$	$\mathbf{30.77 \pm 0.97}$	$\mathbf{44.86 \pm 0.98}$

The shared docking component protects affinity-driving core contacts in both variants. Removing the intersection only from the hotspot-specific term reduces correlation among objectives and encourages differentiating subpocket interactions, explaining the consistent advantage in Table 9.

Table 10: **Interaction-level contribution of the hotspot reward on BRAF.**

Method	IFP clusters	IFP entropy H	Effective modes
BindMol-h	25.67 ± 1.53	0.89 ± 0.03	2.20 ± 0.08
BindMol	$\mathbf{35.00 \pm 1.41}$	$\mathbf{1.11 \pm 0.08}$	$\mathbf{3.03 \pm 0.08}$

G.3 Sensitivity to energy-model dataset size

We train the energy model on stratified 10%, 25%, 50%, 75%, and 100% subsets of the empirical BRAF and JAK2 distributions, preserving the property-label distribution in each subset. We then evaluate the downstream fragment vocabulary under the same generation protocol.

Table 11: **Dataset-size sensitivity of energy-based fragment scoring.** Means and standard deviations over three runs.

Metric	Target	10%	25%	50%	75%	100%
Novel hit ratio (%)	BRAF	15.24 ± 0.76	20.83 ± 0.84	26.17 ± 0.91	29.24 ± 0.95	$\mathbf{30.77 \pm 0.97}$
	JAK2	30.76 ± 1.03	34.68 ± 1.01	40.02 ± 0.99	43.46 ± 0.98	$\mathbf{44.86 \pm 0.98}$
Novel top-5% docking score	BRAF	-10.35 ± 0.04	-11.14 ± 0.06	-11.89 ± 0.07	-12.28 ± 0.08	$\mathbf{-12.50 \pm 0.09}$
	JAK2	-10.08 ± 0.04	-10.76 ± 0.03	-11.43 ± 0.03	-11.81 ± 0.02	$\mathbf{-12.01 \pm 0.02}$

Performance improves and begins to saturate with more target-relevant data. This supports the intended interpretation that such data are beneficial rather than an all-or-nothing prerequisite: for underexplored targets, a virtually screened public library can supply the scored empirical distribution before directed generation.

Table 12: **GuacaMol MPO #Circles results.** The results are the means of 3 runs. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Benchmark						
	Amlodipine	Fexofenadine	Osimertinib	Perindopril	Ranolazine	Sitagliptin	Zaleplon
REINVENT [41]	303.7	343.3	<u>452.3</u>	318.3	253.3	**398.3**	275.3
Graph GA [24]	258.7	333.3	270.3	278.7	<u>364.7</u>	306.3	272.7
STONED [39]	303.7	330.3	301.3	301.0	316.7	326.3	280.3
GEAM-static [35]	412.0	397.7	315.3	318.0	256.7	233.0	267.0
GEAM [35]	<u>424.0</u>	<u>502.0</u>	435.0	**377.3**	295.3	257.0	<u>336.0</u>
`BindMol` (Ours)	**480.5**	**542.1**	**475.5**	<u>374.8</u>	**405.2**	<u>352.7</u>	**362.6**

G.4 Sensitivity of preference sampling size

As mentioned in Section 4.3, the maximization over $\boldsymbol{\omega}_i$ requires approximation. It is sampled by first drawing i.i.d. standard normal vectors from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, applying an element-wise absolute value to ensure non-negativity, and then performing ℓ_1 normalization. This ensures that the resulting vectors are uniformly distributed on the probability simplex. These weights are then replicated across the batch dimension to enable efficient preference-conditioned training. However, since replicating preference vectors for a batch of PyG graphs increases memory usage, very dense sampling would significantly increase the GPU footprint. In practice, we therefore sample $\boldsymbol{\omega}_i$ five times, which provides a good balance between computational efficiency and performance. To further demonstrate the effect of sampling size k, we have conducted a sensitivity analysis and the results are shown in Table 13.

Table 13: Sensitivity analysis of preference sampling size k.

Novel hit ratio (%)	parp1	fa7	5ht1b	braf	jak2
$k = 1$	35.025	18.277	36.272	19.333	39.047
$k = 3$	38.903	21.667	39.622	25.887	42.887
$k = 5$ (`BindMol`)	42.617	24.580	42.149	30.767	44.964
$k = 10$	43.622	24.891	42.277	31.047	45.219
$k = 20$	45.667	25.891	44.239	33.127	46.975

It can be observed that with samping size increase, the performance of novel hit ratio on five targets steadily increase. However, due to the memory burden, we set sampling size to $k = 5$ as the final configuration to achieve the trade-off between performance and efficiency.

G.5 Molecular weight distribution analysis

To investigate whether our per-residue interaction rewards unintentionally incentivize the generation of larger molecules, we conducted a comparative analysis of the molecular weight (MW) distributions

produced by BindMol and several representative baselines. For each method, we sample an equal number of generated molecules for the target jak2 and compute their MW values using RDkit. As an external reference, we also included the MW of experimentally validated ligand extracted from the PDB (PDB id: 7F7W). The results are shown in Figure 8.

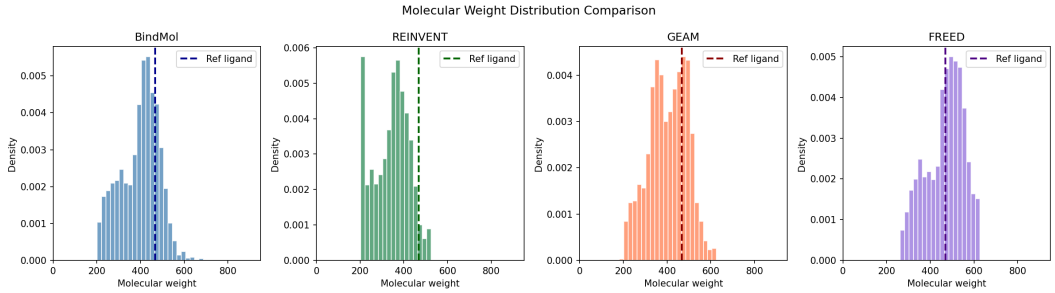


Figure 8: The comparison of molecular weight of generated molecules on jak2 target using different competitive generative approaches.

As shown in Figure 8, the MW distribution of BindMol closely aligns with those of the baseline methods and is broadly consistent with the distribution observed in PDB ligands. These results indicate that our per-hotspot interaction rewards do not systematically drive the model toward generating heavier molecules. The similarity between generated molecules and reference ligands further suggests that we operate within a chemically reasonable MW range that is typical for bioactive small molecules.

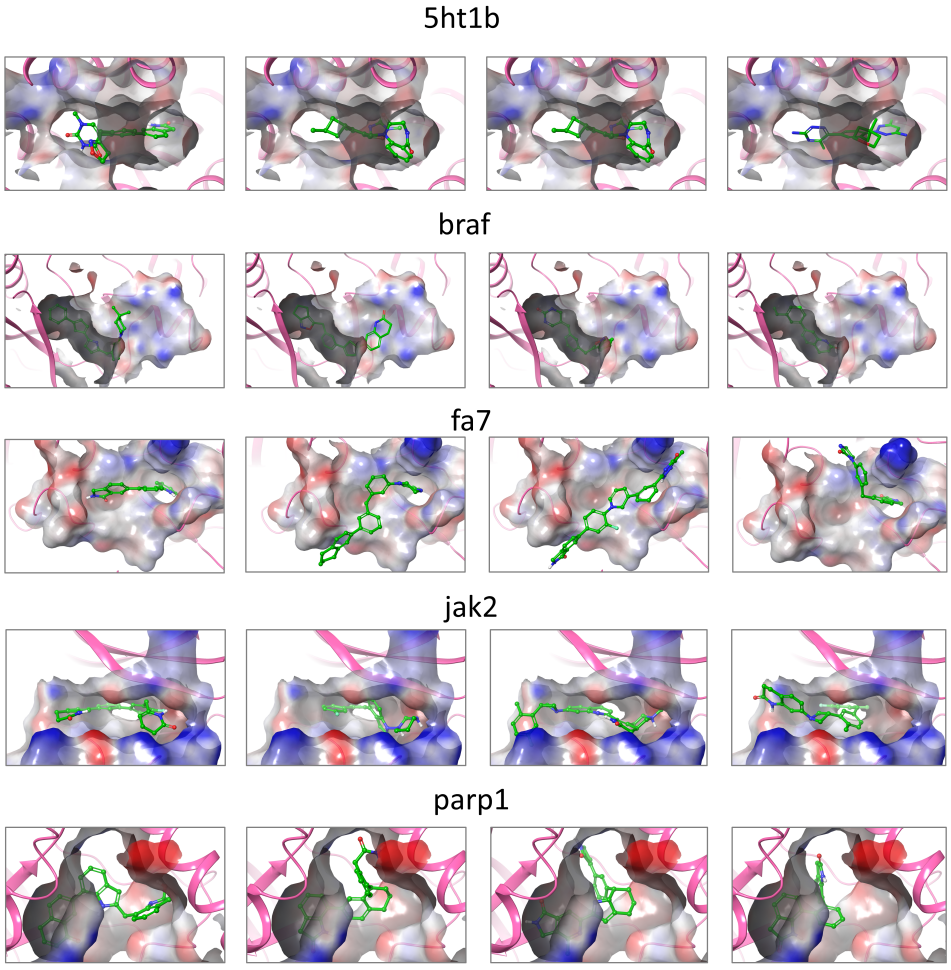


Figure 9: 3D visualization of generated binders of five targets

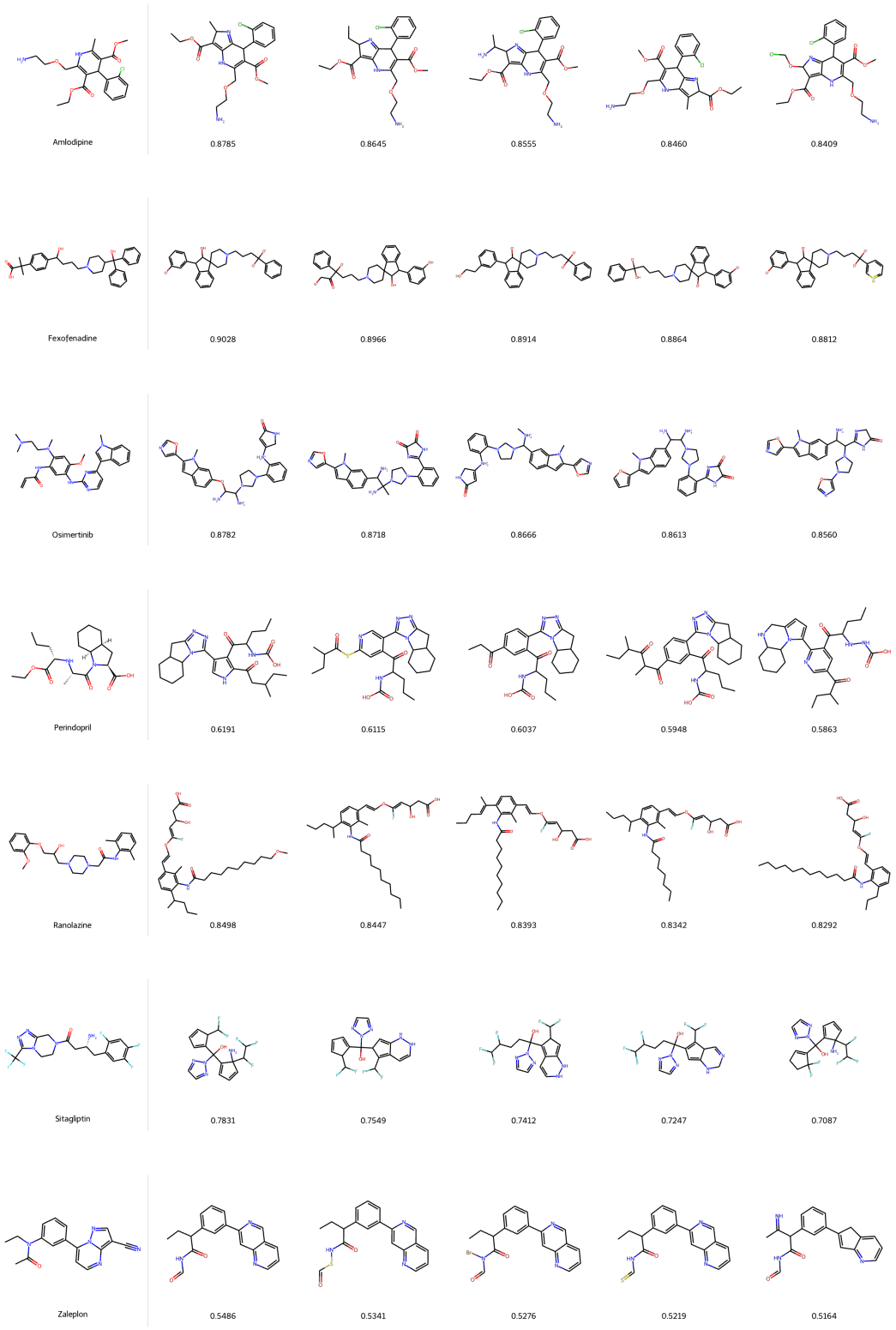


Figure 10: The rediscovery molecules of the GuacaMol MPO tasks and the examples of the generated top-5 molecules. The scores are provided at the bottom of each generated molecule.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: It accurately reflect our contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Section F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please refer to Section E.2 for complete proof of Theorem 1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the code (.zip file) in the supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code (.zip file) in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 5 and Section B for detailed experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the error bars in section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computational resources in Section B.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide the discussion of broader impacts in section F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly respect and credit the existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are only used for polishing the manuscript and grammar.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.