

GRID-BASED EVOLUTIONARY ALGORITHM FOR MULTI-OBJECTIVE MOLECULE GENERATION ENHANCED BY REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Fragment-based drug discovery (FBDD) is limited by the need to construct and maintain static fragment libraries. To overcome these challenges, we propose a novel evolutionary framework. Our method starts with sample molecules that are fragmented and fed into a policy-decoupled architecture. This architecture utilizes reinforcement-learning-guided crossover and mutation operators to recombine and modify fragments. This approach expands the latent fragment space without relying on predefined libraries. By employing a grid-based fragment-masked crossover, our method enables combinatorial explorations and extends beyond conventional fragmentation patterns. In comparative experiments, our method outperforms recent state-of-the-art methods on most PMO benchmarks and target-protein docking tasks. Additionally, it achieves a low average synthetic accessibility (SA) score and maintains a structural novelty rate above 90%.

1 INTRODUCTION

Computer-aided molecular generation has emerged as a crucial pillar of contemporary drug discovery. It capitalizes on advancements in deep learning and evolutionary computation to navigate extensive chemical spaces, offering greater efficiency than traditional high-throughput screening and medicinal chemistry methods (Xue et al., 2025). Among these strategies, fragment-based drug design (FBDD) stands out by integrating low-molecular-weight fragments into lead compounds that exhibit high affinity, potency, and favorable drug-like properties (Maziarz et al., 2021). This methodology leverages the synthetic tractability and diverse binding modes of small fragments. It applies structure-guided optimization to achieve high precision. The balance between hit rate, chemical diversity, and synthetic feasibility makes FBDD a widely adopted approach (Geng et al., 2023).

Over the last decade, numerous implementations of FBDD have been proposed (Xie et al., 2021; Jensen, 2019; Tripp & Hernández-Lobato, 2023). The processes of these methods include: (1) constructing comprehensive fragment vocabularies via extraction techniques (Yang et al., 2021; Xie et al., 2021; Maziarz et al., 2021), (2) maintaining and expanding these vocabularies through fragment modification (Jensen, 2019; Tripp & Hernández-Lobato, 2023; Lee et al., 2024b;a), and (3) employing deep neural architectures to explore combinatorial assembly (Jin et al., 2020b; Maziarz et al., 2021; Kong et al., 2022; Geng et al., 2023; Yang et al., 2021; Lee et al., 2024b;a).

Nevertheless, they face several critical limitations: (i) reliance on predefined fragment libraries inherently constrains chemical exploration; (ii) construction and maintenance of these libraries incur significant inefficiencies (Wang et al., 2022); and (iii) integration of fragments into deep generative models often yields a black-box generation process, reduces the interpretability and complicates multi-property balancing (Angelo et al., 2023).

To address these limitations, we propose a Reinforcement Learning-Driven Grid-based Fragment-Masked Multi-objective Evolutionary Algorithm (RL-GFM). RL-GFM obviates the need to pre-define and maintain the fragment libraries by dynamically recombining and modifying molecules through variously designed crossover and mutation operators.

The main contributions of this paper can be summarized as follows:

- **RL-based crossover and mutation:** We employ a policy-decoupled architecture, wherein two actor-critic reinforcement learning (RL) agents specialize in operator coordination. By integrating multi-objective dominance information into policy backpropagation, these agents learn to select mutation rules for the mutation operator and pair elite parents for the crossover operator. This design avoids reliance on random-walk-like exploration, thereby mitigating the risk of unstable performance.
- **Grid-based fragment-masked crossover:** To transcend conventional fragmentation patterns, we partition molecules into spatial grid regions and enforce masked bond-cleavage rules. Reference points guide parent pairing to maximize inter-grid diversity and intra-grid convergence. Non-scaffold single bonds are randomly cleaved and recombined according to validated dissociation/recombination protocols. Constraining randomness to pharmacophore-compatible regions enables the algorithm to escape local optima and generate synthetically tractable, structurally novel candidates (achieving $\geq 90\%$ novelty).
- **Interpretability and Multi-objective optimization capabilities:** The core generation process of our method is grounded in a multi-objective evolutionary algorithm (MOEA), which inherently incorporates multi-objective optimization capabilities. To avoid sacrificing the primary objective, we employ distinct functional strategies tailored to different tasks. Moreover, by enabling explicit operations on SMILES strings (e.g., bond cleavage and recombination rules, Krenn et al. (2022)), the algorithm renders the generation process interpretable.

The remainder of this article is organized as follows. Section 2 reviews related work, Section 3 details the design and implementation of our RL-GFM framework, Section 4 evaluates our method on PMO and target-protein docking tasks, and Section 5 concludes with future directions.

2 RELATED WORK

FBDD has emerged as a prominent strategy in molecular generation. Early approaches primarily relied on constructing fragment libraries through extraction methods (Yang et al., 2021; Xie et al., 2021; Maziarz et al., 2021). However, the use of static libraries inherently restricts the chemical space that can be explored. Even recent FBDD-related methods Tan et al. (2022); Khemchandani et al. (2020) that integrate reinforcement learning or graph models still fail to address these three core limitations: they either rely on predefined fragment/graph templates (limitation (i)), lack interpretable assembly rules (limitation (iii)), or neglect efficient multi-property balancing (Mercado et al., 2021). This underscores the need for a dynamic, interpretable framework like our RL-GFM.

To address this, more recent methods (Tripp & Hernández-Lobato, 2023; Lee et al., 2024b;a) employ genetic algorithms with crossover and mutation operators to dynamically evolve fragment combinations. These approaches typically optimize either a single molecular property or combine multiple objectives into a single scalar function, often neglecting the inherent trade-offs among objectives (Xie et al., 2021). In contrast, Verhellen et al. (Verhellen, 2022) employed the Non-dominated Sorting Genetic Algorithm II (NSGA-II) algorithm to perform multi-objective molecular optimization. This approach decomposes parent molecules (SMILES strings) into fragments, generates offspring molecules through recombining and modifying these fragments. The generated molecules are selected by non-dominated sorting and crowding distance computations. Unlike methods that depend on predefined fragment vocabularies, this strategy leverages evolutionary principles to navigate the chemical space. Despite its strengths in multi-objective optimization and interpretability, this method remains vulnerable to issues such as stochasticity and premature convergence to local optima, which can hinder its ability to transcend existing fragment and assembly models.

To overcome these limitations, we propose the RL-GFM framework, which incorporates RL agents to guide search, a grid-based fragment-masked crossover mechanism to transcend existing fragments, and a core process governed by a MOEA for balancing multiple objectives. By integrating these components, RL-GFM delivers more consistent and robust performance in complex molecular discovery tasks.

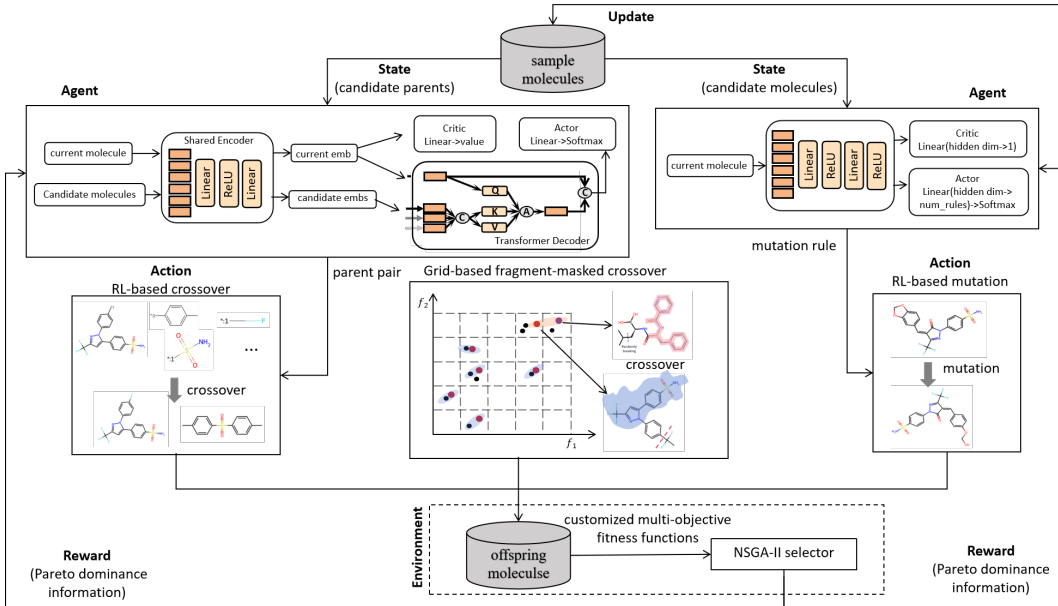


Figure 1: Workflow of the RL-GFM Algorithm. The population is initialized from ZINC250K; molecules are generated via three kinds of operators: grid-based fragment-masked crossover, RL-based crossover, and RL-based mutation. Finally, all molecules are filtered down to N molecules for the next generation by the NSGA-II selector paired with customized multi-objective functions..

3 METHODS

The proposed RL-GFM framework integrates adaptive policy control with fragment-guided structural exploration. It comprises three core components: (1) a grid-based fragment-masked crossover operator, which facilitates diversity-driven recombination; (2) RL-driven crossover and mutation operators, designed to enable targeted exploration; and (3) a Pareto-optimal solution selector paired with customized multi-objective functions to identify solutions with balanced objectives. The framework’s overall workflow is depicted in Figure 1.

3.1 GRID-BASED FRAGMENT-MASKED CROSSOVER

Grid-based strategies are commonly employed as a solution selection mechanism in multi-objective optimization, where partitioning the objective space into multiple grid cells serves to enhance solution diversity (Yang et al., 2013; Xu et al., 2023). Within RL-GFM, we adapt this grid-based strategy for parent pairing by defining two types of specialized reference points—grid ideal points and global ideal points—that guide the parent pairing process. Leveraging these paired parent molecules, we then implement a fragment-masked crossover mechanism, which further enhances chemical structural diversity and strengthens the algorithm’s capacity to explore the chemical space.

The grid-based parent pairing method organizes candidate solutions in the objective space to balance diversity and convergence. First, for each objective dimension f_k ($k = 1, 2, \dots, M$), the maximum observed value $f_k^{\max}$ is determined and the interval $[0, f_k^{\max}]$ is divided into five equal sub-intervals. Their Cartesian product then defines 5^M hypercubes (grid cells). Each solution x is assigned to a cell via $\mathcal{G}(x) = \bigotimes_{k=1}^M \left\lfloor \frac{f_k(x)}{f_k^{\max}/5} \right\rfloor$.

In cases where a grid cell contains only a single solution, that solution is reassigned to its nearest non-empty neighboring grid. For each non-empty grid G_i , we compute the Euclidean distance $d(x) = \sqrt{\sum_{k=1}^M f_k(x)^2}$ for all $x \in G_i$, and designate the solution with the maximum $d(x)$ as the grid’s ideal point $\mathcal{I}_i$. The pairing mechanism is implemented hierarchically: locally, each solution within G_i is paired with its grid’s ideal point to preserve diversity; globally, all grid ideal points

are paired with a global ideal point $\mathcal{I}_g$ (selected from the grid with the highest index) to promote convergence. This dual-level strategy ensures both intra-grid diversity and inter-grid convergence, thereby allocating resources in proportion to grid density (see Figure 2a).

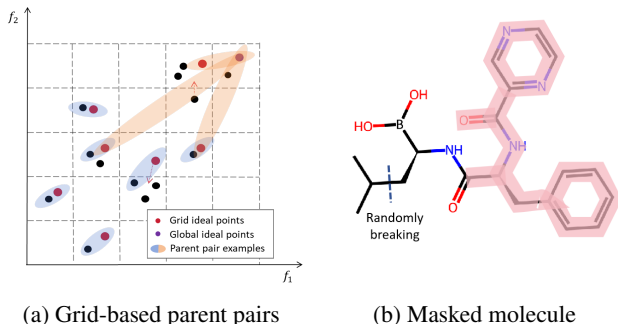


Figure 2: Grid-based fragment-masked crossover. (a) Each solution within a grid is paired with its respective grid’s ideal point; all grid ideal points are then paired with the global ideal point. (b) Bonds internal to the scaffold or those connecting the scaffold and side chains are assigned a mask value of 0 (highlighted in pink), while bonds within the side chains are assigned a mask value of 1 (not highlighted), where a random cut can be made.

The fragment-masked crossover mechanism is introduced to balance structural preservation and chemical diversity, building on the grid-based parent pair approach. Inspired by Li et al. (2024)(Li et al., 2024), this process begins by dividing each molecule into two functional components, as illustrated in Figure 2b. The protected scaffold (highlighted in pink) comprises the core structural motifs essential for molecular stability and bioactivity, while the side chains (not highlighted) represent modifiable peripheral substituents. Bonds are categorized according to their roles: those internal to the scaffold or connecting the scaffold and side chains are assigned a mask of 0, whereas bonds within the side chains receive a mask of 1. A randomized cleavage is then applied to the regions with mask 1, splitting each molecule into fragments. Fragments from grid-based paired parent molecules are recombined, enabling exploration of under-explored regions of chemical space. By protecting the core scaffold and exposing peripheral substituents for randomized bond cleavage, this mechanism maintains synthetic accessibility and bioactivity while enabling controlled chemical diversity.

3.2 RL-BASED CROSSOVER & MUTATION

In our framework, two specialized reinforcement learning (RL) policies are introduced to guide molecular evolution. The goal is to train an actor (the policy, denoted as π) to select optimal actions (crossover or mutation), and a critic (the value function, denoted as V) to evaluate the quality of the current molecular state. Both policies are optimized using an actor-critic method, where their parameters, θ_a for the actor and θ_c for the critic, are updated based on a reward signal r_i derived from the quality of the generated offspring.

RL-based Crossover Policy: The crossover actor $\pi^\times(a_{i,j} \mid s_i; \theta_a)$ is a transformer-decoder with cross-attention designed to select an appropriate parent molecule $a_{i,j}$ to cross with the current molecule s_i . First, a shared two-layer MLP encodes the Morgan fingerprint (FP($\cdot$)) of the current molecule (s_i) and each candidate parent ($a_{i,j}$) into latent vectors. The decoder then computes pairwise compatibility scores via cross-attention, and a softmax function yields the selection probabilities. The crossover critic is a single linear layer applied to the shared encoder output. It is formulated as $V^\times(s_i; \theta_c) = w_c^\top \text{Enc}(s_i)$, where $\text{Enc}(s_i)$ is the latent vector for state s_i produced by the encoder.

RL-based Mutation Policy: The mutation actor $\pi^\mu(a_{i,j} \mid s_i; \theta_a)$ is a two-layer ReLU-MLP that outputs a probability distribution over a set of predefined mutation operations $a_{i,j} \in 1^P$. Its hidden state h_i is computed as $h_i = \text{ReLU}(W_2, \text{ReLU}(W_1, \text{FP}(s_i)))$, and the policy is $\pi^\mu(a_{i,j} \mid s_i) = \text{softmax}(W_3, h_i)$. Similarly, its critic $V^\mu(s_i; \theta_c) = w_c^{\mu\top} h_i$ is a linear layer operating on h_i .

Loss Formulation: Both policies are optimized via a unified actor-critic framework. For each policy type $* \in \{\times, \mu\}$, let θ_a denote the policy-specific actor parameters and θ_c represent shared

critic parameters. At each training iteration, the actor samples an action $a_{i,j}$ from the candidate set $\{a_{i,j}\}_{j=1}^Q$ following the policy $\pi^*(a_{i,j} | s_i; \theta_a)$, where s_i is the molecular state and $a_{i,j}$ corresponds to either a crossover partner or mutation operation. The loss function for policy $*$ is:

$$\mathcal{L}^*(\theta_a, \theta_c) = \frac{1}{N} \sum_{i=1}^N \left[-\log \pi^*(a_{i,j} | s_i; \theta_a) \cdot A_i^* + (V^*(s_i; \theta_c) - r_i)^2 \right] \quad (1)$$

where $A_i^* = r_i - V^*(s_i; \theta_c)$ is the advantage function. The reward r_i is defined as:

$$r_i = \begin{cases} \text{GMean}(\mathbf{f}_{\text{off},i}), & \text{if } i\text{-th offspring is Pareto-optimal,} \\ -\epsilon, & \text{if } i\text{-th offspring is non-Pareto-optimal,} \\ -\text{GMean}(\mathbf{f}_{\text{off},i}), & \text{if same parent pair is used.} \end{cases} \quad (2)$$

with $\text{GMean}(\mathbf{f}) \triangleq (\prod_{k=1}^n f_k)^{1/n}$ and $\epsilon = 10^{-6}$. GMean is chosen to avoid single-objective dominance. This design synergizes with Pareto optimality by quantifying the "dominance strength" of Pareto solutions, guiding the RL policy to prioritize multi-objective balanced parent pairs. The small value of ϵ avoids overwhelming the policy update with excessive penalty signals for non-Pareto solutions, while still serving as a weak constraint to guide exploration away from low-quality molecules.

The advantage term A_i^* prioritizes actions yielding higher returns than the critic’s baseline estimate, while the mean squared error (MSE) term $(V^*(s_i; \theta_c) - r_i)^2$ trains the critic to improve value predictions.

Crossover and mutation: The crossover policy determines parent pair molecules, which are then split into fragments and randomly recombined to form new molecules. For mutation, specific mutation rules are selected by the policy; these rules modify the molecules to facilitate exploration of chemical novelty. Their synergistic operation establishes an evolutionary loop consistent with the process described in (Verhellen, 2022).

3.3 MULTI-OBJECTIVE OPTIMIZATION

NSGA-II selector: After generating molecules via crossover and mutation, these resulting molecules are filtered using the NSGA-II selector. This selector leverages fast non-dominated sorting and crowding-distance computation to enable efficient traversal of the multi-objective search space; its pseudocode is provided in Appendix 1.4.

Customized objective functions: We employ two strategies: *complete attribute separation* and *adaptive weighting scheme*. The former guarantees that molecules exhibiting improved performance in each attribute, albeit with the risk of sacrificing primary objective. By contrast, the adaptive weighting scheme treats the primary objective as an independent target while defining composite targets for secondary objectives, wherein adaptive parameters dynamically regulate their trade-offs.

On the PMO benchmark tasks, the adaptive weighting scheme is formulated as follows:

$$[Obj_1, Obj_2] = [\text{score}, \lambda \times \text{score} + \widehat{SA}], \lambda = 1 - \frac{1}{1 + \exp(-k(x - \text{max_calls}/2))} \quad (3)$$

where score denotes the PMO task score; λ is computed via a sigmoid decay function; and $k = 0.001$ is an attenuation coefficient. The synthetic accessibility $\widehat{SA}$ score is normalized as $\widehat{SA} = \frac{10 - SA}{9}$, as defined in (Lee et al., 2024a).

For docking tasks, we adopt the complete attribute separation approach, defining a tri-objective function as:

$$[Obj_1, Obj_2, Obj_3] = [\widehat{DS}, \text{QED}, \widehat{SA}] \quad (4)$$

with $\widehat{DS} = -\frac{\text{clip}(DS, -20, 0)}{20}$ in accordance with (Lee et al., 2024a). This explicit separation prevents any single factor from dominating the selection process.

Table 1: Baseline methods employed for experimental comparison, classified by representation modalities (columns) and optimization algorithms (rows).

	SMILES	SELFIES	Graph (fragment)
EA	Mol-GA/NSGA-II	STONED/GA+D	Graph-GA
BO			GP-BO
SBM	Genetic-GFN/GEGL		MARS
HC	LSTM-HC		
RL	REINVENT	SELFIES-REINVENT	MORLD/GEAM/RationaleRL/FREED
VAE			JT-VAE/HierVAE/PS-VAE
DM			MOOD
Transformer	RetMol		f-RAG

In the PMO benchmark, we prioritize task-specific optimization above all else, with SA serving as a secondary consideration; accordingly, we employ an adaptive weighting scheme for the evaluation function. By contrast, in the molecular docking task, we regard docking score, SA, and Quantitative Estimate of Drug-likeness (QED) as equally critical, and thus adopt a fully decoupled attribute-separation approach.

4 EXPERIMENTS

In this section, we extensively evaluate RL-GFM across multiple molecular optimization tasks designed to reflect diverse challenges in drug discovery.

Table 1 provides a comprehensive overview of the algorithms used in the experimental evaluation. The methodologies include three molecular representation modalities: SMILES strings, SELFIES strings (Krenn et al., 2020), and graph (fragment)-based representations. Additionally, eight distinct optimization strategies are employed: Evolutionary Algorithms (EA), Bayesian Optimization (BO), Score-Based Modeling (SBM), Hill Climbing (HC), Reinforcement Learning (RL), Variational Autoencoders (VAE), Diffusion Models (DM), and Transformer.

For all tasks, molecules are drawn from a standard repository (ZINC250k (Irwin et al., 2012)). All molecules are filtered according to the rules in (Verhellen, 2022).

4.1 EXPERIMENTS ON PMO BENCHMARK

Setup. We evaluate RL-GFM’s performance across all 23 PMO benchmark tasks. Following the benchmark’s standard protocol, we cap the number of oracle queries at 3,000 (10,000 for baselines according to Lee et al. (2024a)) and assess optimization efficacy using the area under the curve (AUC) of the average property scores for the top-10 molecules as a function of the number of oracle calls. Beyond optimization, we assess three critical drug-discovery criteria—diversity, novelty, and synthesizability (SA)—of the generated compounds. Diversity is quantified using the Therapeutics Data Commons (TDC) library (Huang et al., 2021). Novelty is defined as the proportion of generated molecules whose maximum Tanimoto similarity to any molecule in ZINC250k is below 0.4. SA is measured by the standard Synthetic Accessibility Score, where lower scores indicate easier synthesis. All the experiment configuration followed (Lee et al., 2024a) unless otherwise noted.

Baseline. We selected nine algorithms from Table 1 as our baseline set: the six top-performing methods reported by the PMO benchmark—REINVENT (Olivecrona et al., 2017), Graph-GA (Jensen, 2019), SELFIES-REINVENT, GP-BO (Tripp et al., 2021), STONED (Nigam et al., 2021), and LSTM-HC (Brown et al., 2019)—together with two recent state-of-the-art approaches, f-RAG (Lee et al., 2024a) and Genetic-GFN (Kim et al., 2024), and the highly effective GA-based Mol-GA (Tripp & Hernández-Lobato, 2023).

Results. The main experimental results are summarized in Figure 3. For visual clarity, the figure compares RL-GFM against the five top-performing methods from the original PMO benchmark. Our method demonstrates highly competitive performance across the 23 benchmark tasks. The complete statistical results for the comparison against all nine state-of-the-art baselines are detailed in the Appendix 1.5. Our proposed algorithm achieves superior performance in the majority of cases, attaining the highest cumulative score of 18.007 across all benchmarks, representing a

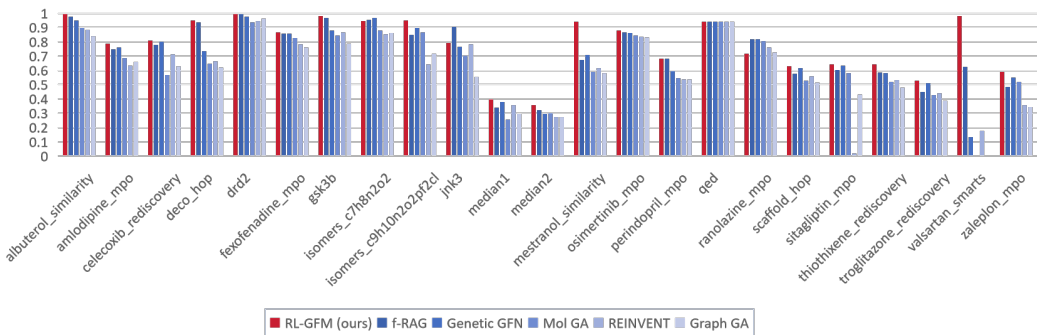


Figure 3: Statistical analysis of top-10 PMO AUC scores.

6.4% improvement over the second-best performing method (f-RAG) and 47.3% over the weakest baseline (LSTM-HC). Notably, RL-GFM exhibits particular dominance in similarity-based tasks such as albuterol_similarity (0.995 vs 0.977 for f-RAG) and mestranol_similarity (0.939 vs 0.708 for Genetic-GFN), suggesting enhanced capability in structural analog generation. The method also excels in complex multi-property optimization (MPO) challenges, including amlodipine_mpo, fexofenadine_mpo and zaleplon_mpo, where it outperforms all baselines by significant margins. A remarkable achievement is observed in the valsartan_smarts task, where RL-GFM attains near-perfect performance (0.979) compared to the zero performance of most algorithms, such as Mol-GA (0.000), Graph-GA (0.000), and GP-BO (0.000), demonstrating exceptional SMARTS pattern matching capabilities.

The comprehensive evaluation of molecular design efficiency across 23 PMO benchmarks reveals the trade-offs between chemical diversity, structural novelty, and SA. As shown in Table 2, RL-GFM achieves the best balance among these metrics, with the highest average diversity (0.573) and novelty (0.912), while maintaining a competitive SA score of 2.332. This tripartite optimization underscores RL-GFM’s ability to explore uncharted chemical space without sacrificing practical synthesizability, a core challenge in de novo drug design.

Table 2: Average top-100 diversity, novelty, and SA scores across 23 tasks (best in bold).

Metric	RL-GFM (ours)	Genetic GFN	Mol GA	REINVENT
Average diversity $\uparrow$	0.573	0.443	0.491	0.468
Average novelty $\uparrow$	0.912	0.724	0.845	0.540
Average SA score $\downarrow$	2.332	3.770	4.605	3.207

4.2 EXPERIMENTS ON DOCKING SCORE

Setup. We evaluated RL-GFM on a battery of multi-objective molecular design tasks that seek to maximize binding affinity against a given protein target while simultaneously preserving favorable QED and SA. Binding affinities were approximated by docking scores (in kcal/mol) computed with QuickVina 2 (Alhossary et al., 2015) against five clinically relevant targets—PARP1, FA7, 5HT1B, BRAF, and JAK2—where more negative scores denote stronger predicted interactions. To ensure practical relevance, we quantified QED and SA using the methods described in (Bickerton et al., 2012) and (Ertl & Schuffenhauer, 2009). Consistent with prior work (Lee et al., 2024a), we focused on the top 5% of novel candidates by docking score, with maximum Tanimoto similarity to any molecule in ZINC250k below 0.40, docking score below the median of known actives, QED exceeding 0.50, and SA under 5.0. In all cases, docking scores are reported as mean $\pm$ standard deviation over three independent runs, with smaller (more negative) values indicating tighter predicted binding. The Novel hit ratio (%) quantifies the proportion of unique and novel hits within the generated molecular corpus.

Baselines. To benchmark performance, we compared RL-GFM against sixteen contemporary generative methods spanning variational autoencoders (JT-VAE (Jin et al., 2018), HierVAE (Jin

Table 3: Novel top 5% docking score (kcal/mol) results. The results are the means and standard deviations of 3 independent runs. The results for compared methods are taken from Lee et al. (Lee et al. 2024a,b). Lower is better, and the best results are highlighted in bold.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
JT-VAE	-9.482 $\pm$ 0.132	-7.683 $\pm$ 0.048	-9.382 $\pm$ 0.332	-9.079 $\pm$ 0.069	-8.885 $\pm$ 0.026
REINVENT	-8.702 $\pm$ 0.523	-7.205 $\pm$ 0.264	-8.770 $\pm$ 0.316	-8.392 $\pm$ 0.400	-8.165 $\pm$ 0.277
Graph GA	-10.949 $\pm$ 0.532	-7.365 $\pm$ 0.326	-10.422 $\pm$ 0.670	-10.789 $\pm$ 0.341	-10.167 $\pm$ 0.576
MORLD	-7.532 $\pm$ 0.260	-6.263 $\pm$ 0.165	-7.869 $\pm$ 0.650	-8.040 $\pm$ 0.337	-7.816 $\pm$ 0.133
HierVAE	-9.487 $\pm$ 0.278	-6.812 $\pm$ 0.274	-8.081 $\pm$ 0.252	-8.978 $\pm$ 0.525	-8.285 $\pm$ 0.370
GA+D	-8.365 $\pm$ 0.201	-6.539 $\pm$ 0.297	-8.567 $\pm$ 0.177	-9.371 $\pm$ 0.728	-8.610 $\pm$ 0.104
MARS	-9.716 $\pm$ 0.082	-7.839 $\pm$ 0.018	-9.804 $\pm$ 0.073	-9.569 $\pm$ 0.078	-9.150 $\pm$ 0.114
GEGL	-9.329 $\pm$ 0.170	-7.470 $\pm$ 0.013	-9.086 $\pm$ 0.067	-9.073 $\pm$ 0.047	-8.601 $\pm$ 0.038
RationaleRL	-10.663 $\pm$ 0.086	-8.129 $\pm$ 0.048	-9.005 $\pm$ 0.155	<i>No hit found</i>	-9.398 $\pm$ 0.076
FREED	-10.579 $\pm$ 0.104	-8.378 $\pm$ 0.044	-10.714 $\pm$ 0.183	-10.561 $\pm$ 0.080	-9.735 $\pm$ 0.022
PS-VAE	-9.978 $\pm$ 0.091	-8.028 $\pm$ 0.050	-9.887 $\pm$ 0.115	-9.637 $\pm$ 0.049	-9.464 $\pm$ 0.129
MOOD	-10.865 $\pm$ 0.113	-8.160 $\pm$ 0.071	-11.145 $\pm$ 0.042	-11.063 $\pm$ 0.034	-10.147 $\pm$ 0.060
RetMol	-8.590 $\pm$ 0.475	-5.448 $\pm$ 0.688	-6.980 $\pm$ 0.740	-8.811 $\pm$ 0.574	-7.133 $\pm$ 0.242
Genetic GFN	-9.227 $\pm$ 0.644	-7.288 $\pm$ 0.433	-8.973 $\pm$ 0.804	-8.719 $\pm$ 0.190	-8.539 $\pm$ 0.592
GEAM	-12.891 $\pm$ 0.158	-9.890 $\pm$ 0.116	-12.374 $\pm$ 0.036	-12.342 $\pm$ 0.095	-11.816 $\pm$ 0.067
f-RAG	-12.945 $\pm$ 0.053	-9.899 $\pm$ 0.205	-12.670 $\pm$ 0.144	-12.390 $\pm$ 0.046	-11.842 $\pm$ 0.316
RL-GFM (ours)	-12.960 $\pm$ 0.01	-10.310$\pm$0.03	-13.343$\pm$0.2	-12.671$\pm$0.05	-12.090$\pm$0.07

Table 4: Novel hit ratio (%) results. The results are the means and the standard deviations of 3 runs. The results for the baselines are taken from Lee et al. (Lee et al. 2024b). The best results are highlighted in bold.

Method	Target protein				
	parp1	fa7	5ht1b	braf	jak2
REINVENT	0.480 $\pm$ 0.344	0.213 $\pm$ 0.081	2.453 $\pm$ 0.561	0.127 $\pm$ 0.088	0.613 $\pm$ 0.167
Graph GA	4.811 $\pm$ 1.661	0.422 $\pm$ 0.193	7.011 $\pm$ 2.732	3.767 $\pm$ 1.498	5.311 $\pm$ 1.667
MORLD	0.047 $\pm$ 0.050	0.007 $\pm$ 0.013	0.880 $\pm$ 0.735	0.047 $\pm$ 0.040	0.227 $\pm$ 0.118
HierVAE	0.553 $\pm$ 0.214	0.007 $\pm$ 0.013	0.507 $\pm$ 0.278	0.207 $\pm$ 0.220	0.227 $\pm$ 0.127
RationaleRL	4.267 $\pm$ 0.450	0.900 $\pm$ 0.098	2.967 $\pm$ 0.307	0.000 $\pm$ 0.000	2.967 $\pm$ 0.196
FREED	4.627 $\pm$ 0.727	1.332 $\pm$ 0.113	16.767 $\pm$ 0.897	2.940 $\pm$ 0.359	5.800 $\pm$ 0.295
PS-VAE	1.644 $\pm$ 0.389	0.478 $\pm$ 0.140	12.622 $\pm$ 1.437	0.367 $\pm$ 0.047	4.178 $\pm$ 0.933
MOOD	7.017 $\pm$ 0.428	0.733 $\pm$ 0.141	18.673 $\pm$ 0.423	5.240 $\pm$ 0.285	9.200 $\pm$ 0.524
GEAM	40.567 $\pm$ 0.825	20.711 $\pm$ 1.873	38.489 $\pm$ 0.350	27.900 $\pm$ 1.822	42.950 $\pm$ 1.117
RL-GFM (ours)	43.250 $\pm$ 1.041	35.570 $\pm$ 1.547	70.070 $\pm$ 5.051	42.667 $\pm$ 2.245	42.557 $\pm$ 3.482

et al., 2020a), PS-VAE (Kong et al., 2022)), reinforcement-learning and policy-gradient approaches (REINVENT (Olivecrona et al., 2017), RationaleRL (Jin et al., 2020b), GEGL (Ahn et al., 2020), FREED (Yang et al., 2021), MORLD (Jeon & Kim, 2020), MARS (Xie et al., 2021)), genetic-based optimizers (Graph GA (Jensen, 2019), GA + D (Nigam et al., 2020)), as well as more recent methods (RetMol (Wang et al., 2022), MOOD (Lee et al., 2023), Genetic GFN (Kim et al., 2024), GEAM (Lee et al., 2024b), f-RAG (Lee et al., 2024a)).

Results. From the results represented in Table 3, our RL-GFM method achieves the most favorable mean docking scores on all five targets, consistently surpassing the previous best-performing baselines (f-RAG and GEAM). Against PARP1, RL-GFM reaches -12.96 ± 0.01 kcal/mol, marginally improving on f-RAG’s -12.945 ± 0.053 kcal/mol; the improvement is more pronounced on FA7 (-10.31 ± 0.03 vs. -9.899 ± 0.205 kcal/mol), 5HT1B (-13.34 ± 0.20 vs. -12.670 ± 0.144 kcal/mol), BRAF (-12.67 ± 0.05 vs. -12.390 ± 0.046 kcal/mol), and JAK2 (-12.09 ± 0.07 vs. -11.842 ± 0.316 kcal/mol). These gains—ranging from 0.015 kcal/mol on PARP1 to 0.67 kcal/mol on 5HT1B—demonstrate that RL-GFM systematically identifies higher-affinity chemotypes than the compared baselines.

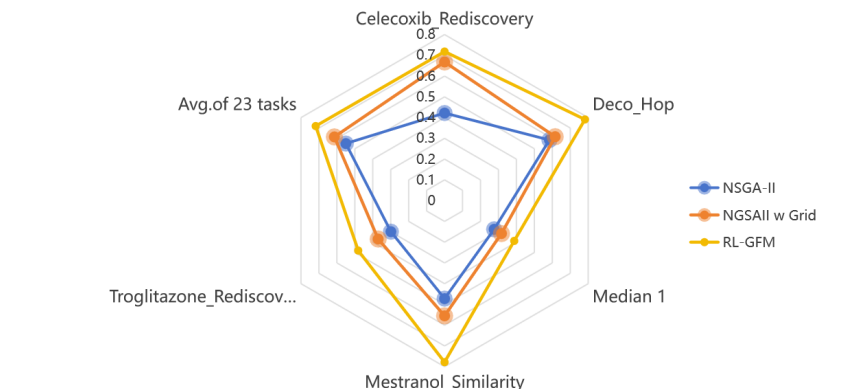


Figure 4: Ablation Study Results. Performance Comparison of NSGA-II Baseline, NSGA-II with Grid-Based Fragment-Masked Crossover (NSGA-II w Grid), and the Full RL-GFM Algorithm on several PMO tasks

Furthermore, we evaluated the model’s ability to discover novel chemical matter, measured by the novel hit ratio (%) presented in Table 4. RL-GFM demonstrates exceptional performance, achieving the highest hit ratio on four of the five targets. The improvements over the next-best baseline, GEAM, are particularly striking for 5HT1B, where RL-GFM achieves a 70.070 ± 5.05 ratio compared to GEAM’s 38.489 ± 0.350 , and for FA7 (35.570 ± 1.547 vs. 20.711 ± 1.873). For the JAK2 target, RL-GFM (42.557 ± 3.482) performs on par with the leading baseline (42.950 ± 1.117). This confirms that our method not only refines molecules towards high-scoring regions but is also highly effective at exploring the chemical space to identify new and diverse molecular structures.

In summary, RL-GFM not only matches or exceeds the state-of-the-art on small-molecule optimization benchmarks but also delivers the strongest predicted binding affinities in structure-based docking tasks, with excellent reproducibility across independent runs.

4.3 ABLATION STUDY

To validate the contributions of RL-GFM’s core components, we conducted ablation experiments within 2000 function calls across diverse PMO tasks with three experimental groups: NSGA-II (baseline) (Verhellen, 2022), NSGA-II w Grid (NSGA-II integrated with grid-based fragment-masked crossover), and the full RL-GFM.

Results in Figure 4 indicate that NSGA-II w Grid outperforms the NSGA-II baseline across all experimental tasks, including the average score of 23 PMO tasks, confirming that the grid-based crossover effectively enhances chemical space exploration efficiency. Further, the full RL-GFM achieves an additional performance improvement compared to NSGA-II w Grid. This demonstrates that RL-driven operators enable targeted exploration of high-potential molecular regions. Overall, the ablation study verifies the necessity of each component.

5 CONCLUSION

In this study, we introduce RL-GFM, a novel framework that integrates innovative crossover and mutation operators within an NSGA-II-based optimization architecture to guide molecular design—thereby mitigating key limitations of fragment-based drug discovery (FBDD), including reliance on predefined and curated libraries, inadequate interpretability, and suboptimal multi-objective balancing. Notably, the RL-driven coordination of crossover and mutation operators enhances adaptive decision-making in navigating trade-offs between objectives, while the grid-based fragment-masked crossover augments both the efficiency and breadth of chemical space exploration. Future work will focus on reconciling RL-based guidance with the inherent stochastic exploration properties of evolutionary algorithms to further refine the balance between targeted search precision and structural diversity in complex chemical spaces.

REFERENCES

- Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding deep molecular optimization with genetic exploration. *Advances in neural information processing systems*, 33:12008–12021, 2020.
- Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.
- Jaqueline S Angelo, Isabella A Guedes, Helio JC Barbosa, and Laurent E Dardenne. Multi-and many-objective optimization: present and future in de novo drug design. *Frontiers in Chemistry*, 11:1288626, 2023.
- G Richard Bickerton, Gaia V Paolini, J  r  my Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.
- Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.
- Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.
- Zijie Geng, Shufang Xie, Yingce Xia, Lijun Wu, Tao Qin, Jie Wang, Yongdong Zhang, Feng Wu, and Tie-Yan Liu. De novo molecular generation via connection-aware motif mining. *arXiv preprint arXiv:2302.01129*, 2023.
- Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.
- John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.
- Jan H Jensen. 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.
- Woosung Jeon and Dongsup Kim. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports*, 10(1):22104, 2020.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pp. 4839–4848. PMLR, 2020a.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pp. 4849–4859. PMLR, 2020b.
- Yash Khemchandani, Stephen O’Hagan, Soumitra Samanta, Neil Swainston, Timothy J Roberts, Danushka Bollegala, and Douglas B Kell. Deepgraphmolgen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *Journal of cheminformatics*, 12(1):53, 2020.
- Hyeonah Kim, Minsu Kim, Sanghyeok Choi, and Jinkyoo Park. Genetic-guided gflownets: Advancing in practical molecular optimization benchmark. *arXiv e-prints*, pp. arXiv–2402, 2024.
- Xiangzhe Kong, Wenbing Huang, Zhixing Tan, and Yang Liu. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35:2550–2563, 2022.

- Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.
- Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.
- Seul Lee, Jaehyeong Jo, and Sung Ju Hwang. Exploring chemical space with score-based out-of-distribution generation. In *International Conference on Machine Learning*, pp. 18872–18892. PMLR, 2023.
- Seul Lee, Karsten Kreis, Srimukh Veccham, Meng Liu, Danny Reidenbach, Saeed Paliwal, Arash Vahdat, and Weili Nie. Molecule generation with fragment retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:132463–132490, 2024a.
- Seul Lee, Seanie Lee, Kenji Kawaguchi, and Sung Ju Hwang. Drug discovery with dynamic goal-aware fragments. In *International Conference on Machine Learning*, pp. 26731–26751. PMLR, 2024b.
- Kun Li, Xiantao Cai, Jia Wu, Bo Du, and Wenbin Hu. Fragment-masked molecular optimization. *arXiv preprint arXiv:2408.09106*, 2024.
- Krzysztof Maziarsz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.
- Rocío Mercado, Esben J Bjerrum, and Ola Engkvist. Exploring graph traversal algorithms in graph-based molecular generation. *Journal of Chemical Information and Modeling*, 62(9):2093–2100, 2021.
- AkshatKumar Nigam, Pascal Friederich, Mario Krenn, and Alan Aspuru-Guzik. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. In *International Conference on Learning Representations*, 2020.
- AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical science*, 12(20):7079–7090, 2021.
- Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9:1–14, 2017.
- Youhai Tan, Lingxue Dai, Weifeng Huang, Yinfeng Guo, Shuangjia Zheng, Jinping Lei, Hongming Chen, and Yuedong Yang. Drlinker: deep reinforcement learning for optimization in fragment linking design. *Journal of Chemical Information and Modeling*, 62(23):5907–5917, 2022.
- Austin Tripp and José Miguel Hernández-Lobato. Genetic algorithms are strong baselines for molecule generation. *arXiv preprint arXiv:2310.09267*, 2023.
- Austin Tripp, Gregor NC Simm, and José Miguel Hernández-Lobato. A fresh look at de novo molecular design benchmarks. In *NeurIPS 2021 AI for Science Workshop*, 2021.
- Jonas Verhellen. Graph-based molecular pareto optimisation. *Chemical Science*, 13(25):7526–7535, 2022.
- Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126*, 2022.
- Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. Mars: Markov molecular sampling for multi-objective drug discovery. In *International Conference on Learning Representations*, 2021.

- Ying Xu, Huan Zhang, Lei Huang, Rong Qu, and Yusuke Nojima. A pareto front grid guided multi-objective evolutionary algorithm. *Applied Soft Computing*, 136:110095, 2023.
- Zhe Xue, Chenwei Sun, Wenhao Zheng, Jiancheng Lv, and Xianggen Liu. Targetsa: adaptive simulated annealing for target-specific drug design. *Bioinformatics*, 41(1):btac730, 2025.
- Shengxiang Yang, Miqing Li, Xiaohui Liu, and Jinhua Zheng. A grid-based evolutionary algorithm for many-objective optimization. *IEEE transactions on evolutionary computation*, 17(5):721–736, 2013.
- Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, and Sung Ju Hwang. Hit and lead discovery with explorative rl and fragment-based molecule generation. *Advances in Neural Information Processing Systems*, 34:7924–7936, 2021.