Generative Flows on Synthetic Pathway for Drug Design

Seonghwan Seo[1*], **Minsu Kim**[1], **Tony Shen**[2],
Martin Ester[2], **Jinkyoo Park**[1,3], **Sungsoo Ahn**[4], **Woo Youn Kim**[1,5*]
[1]KAIST [2]Simon Fraser University [3]OMELET [4]POSTECH [5]HITS

Abstract

Generative models in drug discovery have recently gained attention as efficient alternatives to brute-force virtual screening. However, most existing models do not account for synthesizability, limiting their practical use in real-world scenarios. In this paper, we propose RxnFlow, which sequentially assembles molecules using predefined molecular building blocks and chemical reaction templates to constrain the synthetic chemical pathway. We then train on this sequential generating process with the objective of generative flow networks (GFlowNets) to generate both highly rewarded and diverse molecules. To mitigate the large action space of synthetic pathways in GFlowNets, we implement a novel action space subsampling method. This enables RxnFlow to learn generative flows over extensive action spaces comprising combinations of 1.2 million building blocks and 71 reaction templates without significant computational overhead. Additionally, RxnFlow can employ modified or expanded action spaces for generation without retraining, allowing for the introduction of additional objectives or the incorporation of newly discovered building blocks. We experimentally demonstrate that RxnFlow outperforms existing reaction-based and fragment-based models in pocket-specific optimization across various target pockets. Furthermore, RxnFlow achieves state-of-the-art performance on CrossDocked2020 for pocket-conditional generation, with an average Vina score of –8.85 kcal/mol and 34.8% synthesizability. Code is available at https://github.com/SeonghwanSeo/RxnFlow.

1 Introduction

Structure-based drug discovery (SBDD) has emerged as a pivotal paradigm for early drug discovery, facilitated by the increasing accessibility of protein structure prediction tools [31] and high-resolution crystallography [37]. However, traditional brute-force virtual screening is computationally expensive [20], prompting the development of deep generative models that can bypass this inefficiency. In this context, various approaches such as deep reinforcement learning [61], variational autoencoders [63] generative adversarial network [48], and diffusion models [22, 23] have been proposed to directly sample candidate molecules against a given protein structure.

While generative models have shown success in molecular discovery with desirable biological properties, most overlook synthesizability which is a crucial factor for wet-lab validation [15]. One line to improve synthesizability is multi-objective optimization using cheap functions to estimate the synthesizability [12], but this is too simplified to reflect complex synthetic principles [9]. Other efforts aim to project molecules from generative models into a synthesizable space [17, 40], but chemical modifications in this process can degrade the optimized properties.

*Correspondence to: Seonghwan Seo, Woo Youn Kim <{shwan0106,wooyoun}@kaist.ac.kr>

AI for New Drug Modalities at NeurIPS 2024.

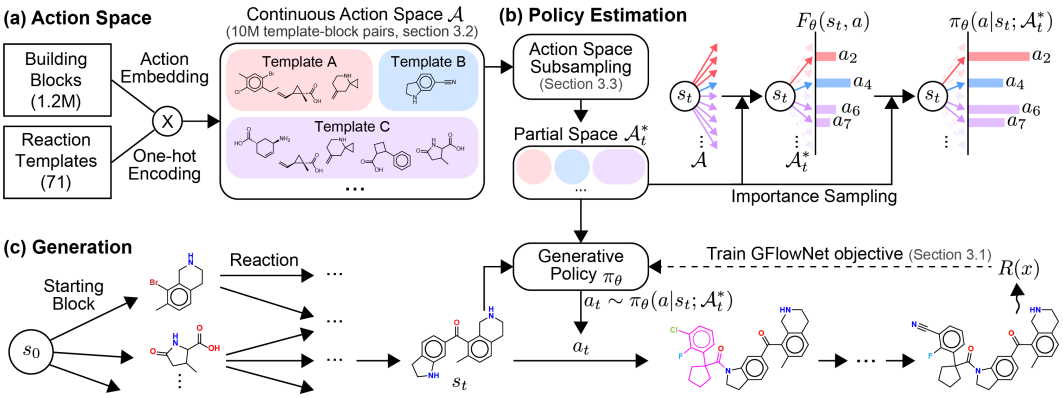

Figure 1: **Overview of RxnFlow. (a)** Synthetic action space which is represented in a continuous action space. Each colored box corresponds to a reaction template and the molecules in the box are reactant blocks. **(b)** Policy estimation using the action space subsampling in a manner of importance sampling. **(c)** Molecular generation process and model training.

To address this issue, recent works formulate the generation of synthetic pathways as a Markov decision process (MDP) for molecular design [19]. These approaches return a synthesizable molecule by assembling purchasable building blocks and reaction templates according to the generated synthetic pathway. Notably, the emergence of virtual libraries—created by combinatorially enumerating building blocks and reaction templates, such as Enamine REAL [21]—allows the generated molecules to be readily synthesizable on demand with their synthetic pathways. Recent studies [9, 33] advanced this approach by training the decision-making policy using the objective of generative flow networks [GFlowNets; 3]. This objective encourages the policy to sample in proportion to the reward function, enabling the retrieval of samples from a diverse range of modes.

Unlike atom-based or fragment-based models, the synthetic action spaces are massive and composed of millions of building blocks and tens of reaction templates. While the large action spaces offer opportunities to discover novel hit candidates by expanding an explorable chemical space [50], it incurs significant computational overhead. Thus, prior works have restricted the action spaces to trade off the size of the search space for efficiency. However, reducing the search space leads to a decrease in diversity and synthetic complexity. While one could compensate for the reduced number of building block candidates by adding more reaction steps, this leads to the increase in synthetic complexity, negatively influencing synthesizability, yield, and cost [8, 32].

In response to this challenge, we propose RxnFlow, a synthesis-oriented generative framework that allows training generative flows over a large action space to generate synthetic pathways for drug design. The distinctive features of this method are as follows. First, we introduce an *action space subsampling* (Figure 1) to handle massive action spaces without significant memory overhead, enabling us to explore a broader chemical space with fewer reaction steps than existing models. Then, we train the generative policy with a GFlowNet objective to sample both diverse and potent molecules from the expanded search space. We demonstrate that RxnFlow effectively generates drug candidates, outperforming existing reaction-based, atom-based, and fragment-based baselines across various SBDD tasks, while ensuring the synthesizability of generated drug candidates. We also achieve a new state-of-the-art Vina score, drug-likeness, and synthesizability on the CrossDocked2020 pocket-conditional generation benchmark [39].

Furthermore, we formulate an adaptable MDP (Figure 2) for consistent flow estimation on modified building block libraries, which can be highly practical in real-world applications. By combining the proposed MDP with action embedding [10], which represents actions in a continuous space instead of a discrete space, RxnFlow can achieve further objectives or incorporate newly discovered building blocks without retraining. We experimentally show that RxnFlow can achieve an additional solubility objective and behave appropriately for unseen building blocks. This capability makes RxnFlow highly adaptable to real-world drug discovery pipeline, where new objectives frequently arise [13] and building block libraries are continuously expanding [21].

2 Related Works

Structure-based drug discovery. The first type of SBDD involves **pocket-specific optimization** methods to enhance docking scores against a single pocket, including evolutionary algorithms [49], reinforcement learning (RL) [61], and GFlowNets [3]. However, these require individual optimizations for each pocket, limiting scalability. The second type is based on **pocket-conditional generation**, which generates molecules against arbitrary given pockets without additional training. This can be achieved by distribution-based generative models [48, 47, 23, 51] trained on protein-ligand complex datasets to model ligand distributions for given pockets. On the other hand, Shen et al. [54] formulated a pocket-conditioned policy for GFlowNets to generate samples from reward-biased distributions in a *zero-shot* manner.

Syntheis-oriented generative models. To ensure the synthesizability of generated molecules, synthesis-oriented *de novo* design approaches incorporate combinatorial chemistry principles into generative models. Bradshaw et al. [6] represented the synthetic pathways as directed acyclic graphs (DAG) for generative modeling. Horwood and Noutahi [25] formulated the synthetic pathway generation as an MDP and optimize the molecules with RL. Similarly, Gao et al. [17] employed a genetic algorithm to optimize synthesis trees to generate molecules with the desired properties. Seo et al. [53] proposed a conditional generative model to directly sample molecules with desired properties without optimization. Recently, Cretu et al. [9], Koziarski et al. [33] trained the models with the GFlowNets objective to generate diverse and potent molecules.

Action embedding for large action spaces. To handle large action spaces in the synthesis-oriented generation, Gottipati et al. [19] employed action embedding [10] that represents building blocks in a continuous action space with their chemical information. Later, Seo et al. [53], Koziarski et al. [33] experimentally demonstrated that it can enhance the model training and generative performance. The continuous action space provides the benefit of reducing the computational complexity for sampling from large space of actions and the memory requirement for parameterizing the categorical distribution over the large action space.

Generative Flow Networks. GFlowNets are a learning framework for a stochastic generative policy that constructs an object through a series of decisions, where the probability of generating each object is proportional to a given reward associated with that object [3]. Unlike other optimization methods that maximize rewards and often converge to a single solution, GFlowNets aim to sample a diverse set of high-rewarded modes, which is vital for novel drug design [54, 27]. To this end, the generative policy is trained using objectives such as flow matching [3], detailed balance [4], and trajectory balance [41]. Extending the GFlowNets to various applications is an active area of research, e.g., GFlowNets have been applied to designing crystal structures [44], phylogenetic inference [62], finetuning diffusion models [57], and causal inference [43].

3 Method

3.1 GFlowNet Preliminaries

GFlowNets [3] are the class of generative models that learn to sample objects $x \in \mathcal{X}$ proportional to a given reward function, i.e., $p(x) \propto R(x)$. This is achieved by sequentially constructing a compositional object x through a series of state transitions $s \to s'$, forming a trajectory $\tau = (s_0 \to \ldots \to s_n = x) \in \mathcal{T}$. The set of all complete trajectories from the initial state s_0 can be represented as a directed acyclic graph $\mathcal{G} = (\mathcal{S}, \mathcal{A})$ with a reachable state space $\mathcal{S}$ and an action space $\mathcal{A}$. Each action a induces a transition from the state s to the state s', expressed as $s' = T(s, a)$ and represented as $s \to s'$. Then, we define the *trajectory flow* $F(\tau)$, which flows along the trajectory $\tau = (s_0 \to \ldots \to s_n = x)$, as the reward of the terminal state, $R(x)$. The *edge flow* $F(s \to s')$, or equivalently $F(s, a)$, is defined as the total flow along the edge $a : s \to s'$:

$$F(s \to s') = F(s, a) = \sum_{\tau \in \mathcal{T} \text{ s.t. } (s \to s') \in \tau} F(\tau). \tag{1}$$

The *state flow* $F(s)$ for the intermediate state is defined as the total flow through the state s:

$$F(s) = \sum_{\tau \in \mathcal{T} \text{ s.t. } s \in \tau} F(\tau) = \underbrace{\sum_{(s'' \to s) \in \mathcal{A}} F(s'' \to s) = \sum_{(s \to s') \in \mathcal{A}} F(s \to s').}_{\text{Intermediate flow matching condition}} \tag{2}$$

In addition to the flow matching condition for intermediate states, there are two boundary conditions for the states. First, the flow of a terminal state x must equal the reward of the objective: $F(x) = R(x)$. Second, the *partition function*, Z, is equivalent to the sum of all trajectory flows and the sum of all rewards: $Z = F(s_0) = \sum_{\tau \in \mathcal{T}} F(\tau) = \sum_{x \in \mathcal{X}} R(x)$. These three conditions—one for intermediate states and two for boundary states—are known as the *flow matching* conditions and ensure that GFlowNets generate objectives proportional to their rewards.

To convert the flow network into a usable policy, we define the *forward policy* as the forward transition probability $P_F(s'|s)$ and the *backward policy* as the backward transition probability $P_B(s|s')$:

$$P_F(s'|s) := P(s \to s'|s) = \frac{F(s \to s')}{F(s)}, \quad P_B(s|s') := P(s \to s'|s') = \frac{F(s \to s')}{F(s')}. \tag{3}$$

3.2 Action Space for Synthetic Pathway Generation

Following Cretu et al. [9], we treated a chemical reaction as a forward transition and a synthetic pathway as a trajectory for molecular generation. For the initial state s_0, the model always chooses `AddFirstReactant` to sample a building block b from the entire building block set $\mathcal{B}$ as a starting molecule. For the later states s, the model samples actions among `ReactUni`, `ReactBi`, or `Stop`. When the action type is `ReactUni`, the model performs *in silico* uni-molecular reactions with an assigned reaction template $r \in \mathcal{R}_1$. When the action type is `ReactBi`, the model performs bi-molecular reactions with a reaction template $r \in \mathcal{R}_2$ and a reactant block b in the possible reactant set for the reaction template r: $\mathcal{B}_r \subseteq \mathcal{B}$. If `Stop` is sampled, the trajectory is terminated. To sum up, the allowable action space $\mathcal{A}(s)$ for the state s is:

$$\mathcal{A}(s) = \begin{cases} \mathcal{B} & \text{if } s = s_0 \\ \{\texttt{Stop}\} \cup \mathcal{R}_1 \cup \{(r, b) | r \in \mathcal{R}_2, b \in \mathcal{B}_r\} & \text{otherwise} \end{cases} \tag{4}$$

where unavailable reaction templates to the molecule of the state s are masked.

3.3 Flow Network on Action Space Subsampling

We propose a novel memory-efficient technique called the *action space subsampling*, that estimates the state flow $F_\theta(s)$ from a subset of the outgoing edge flows $F_\theta(s \to s')$ for forward policy estimation. First, we implement an auxiliary policy, termed *subsampling policy* $\mathcal{P}(\mathcal{A})$, which samples a subset of the action space $\mathcal{A}^* \subseteq \mathcal{A}$. This reduces both the memory footprint and the computational complexity from $\mathcal{O}(|\mathcal{B}||\mathcal{R}_2|)$ to $\mathcal{O}(|\mathcal{B}^*||\mathcal{R}_2|)$ with the controllable size $|\mathcal{B}^*|$. We then estimate the forward policy by importance sampling. In contrast to the parameterized forward policy, we formulate a fixed backward policy since it is hard to force invariance to molecule isomorphism [41]. Theoretical backgrounds are provided in Sec. A.

Subsampling policy. Subsampling policy $\mathcal{P}(\mathcal{A})$ performs uniform sampling for the initial state and importance sampling for the later states. For the initial state, the allowable action space $\mathcal{A}(s_0) = \mathcal{B}$ is homogeneous since all of them are `AddFirstReactant` actions. For the later state, the action space is comprised of one `Stop`, tens of `ReactUni` actions, and millions of `ReactBi` actions. To capture rare-type actions in the inhomogeneous space, we use all `Stop` and `ReactUni` actions. The partial action space $\mathcal{A}^*(s) \sim \mathcal{P}(\mathcal{A}(s))$ comprises the uniform subset $\mathcal{B}^* \subseteq \mathcal{B}$ or $\mathcal{B}_r^* \subseteq \mathcal{B}_r$:

$$\mathcal{A}^*(s) = \begin{cases} \mathcal{B}^* & \text{if } s = s_0 \\ \{\texttt{Stop}\} \cup \mathcal{R}_1 \cup \{(r, b) | r \in \mathcal{R}_2, b \in \mathcal{B}_r^*\} & \text{otherwise} \end{cases} \tag{5}$$

Forward policy. To estimate the forward policy $P_F(s'|s) = F(s \to s')/F(s)$ from the partial action space $\mathcal{A}^*$, we estimate the state flow $F(s)$ with a subset of outgoing edge flows $F(s \to s')$. Since

Algorithm 1 Training GFlowNets with action space subsampling

1: **input** Entire action space $\mathcal{A}$, Maximum trajectory length N
2: **repeat**
3: Sample partial action spaces $\mathcal{A}_0^*$, $\mathcal{A}_1^*$, ..., $\mathcal{A}_{N-1}^*$ from subsampling policy $\mathcal{P}(\mathcal{A})$
4: Sample trajectory τ from sampling policy $\pi_\theta(-|-;\mathcal{A}_{(-)}^*)$
5: Update model $\theta \leftarrow \theta - \eta\nabla\hat{\mathcal{L}}_{\text{TB}}(\tau)$
6: **until** model converges

we introduce importance sampling for action types, we weight the edge flow $F(s \to s')$ of edge $a : s \to s'$ according to the subsampling ratio:

$$w_a = w_{(s\to s')} = \begin{cases} |\mathcal{B}|/|\mathcal{B}^*| & \text{if } a \in \mathcal{B} \\ |\mathcal{B}_r|/|\mathcal{B}_r^*| & \text{if } a \text{ is } (r,b) \text{ where } r \in \mathcal{R}_2 \\ 1 & \text{if } a \text{ is } \texttt{Stop} \text{ or } a \in \mathcal{R}_1 \end{cases} \tag{6}$$

By weighting edge flows, we can estimate the state flow $\hat{F}_\theta(s;\mathcal{A}^*)$ as:

$$\hat{F}_\theta(s;\mathcal{A}^*) = \sum_{(s\to s')\in\mathcal{A}^*(s)} w_{(s\to s')} F_\theta(s \to s'), \tag{7}$$

which the estimated forward policy is $\hat{P}_F(s'|s;\mathcal{A}^*;\theta) = F_\theta(s \to s')/\hat{F}_\theta(s;\mathcal{A}^*)$.

Action embedding. In the standard implementation [3] of the flow function F_θ and its neural network ϕ_θ, the edge flow of the edge $a : s \to s'$ is computed with the corresponding action-specific parameter θ_a: $F_\theta(s \to s') = F_\theta(s,a) = \phi_{\theta_a}^{\text{flow}}(\phi_\theta^{\text{state}}(s))$.

However, large action spaces require numerous parameters which increase model complexity. To address this, we use an additional network $\phi_\theta^{\text{block}}$ for `AddFirstReactant` and `ReactBi`, which embeds the building block b into a continuous action space with its structural information:

$$F_\theta(s_0,b) = \phi_\theta^{\text{flow}}(\phi_\theta^{\text{state}}(s), \phi_\theta^{\text{block}}(b)), \quad F_\theta(s,(r,b)) = \phi_\theta^{\text{flow}}(\phi_\theta^{\text{state}}(s), \delta(r), \phi_\theta^{\text{block}}(b)) \tag{8}$$

where $\delta(r)$ is the one-hot encoding for a bi-molecular reaction template r.

GFlowNet training. In this work, we use the trajectory balance [TB; 41] as the training objective of GFlowNets from Eq. (9) and train models following Algorithm 1. The action space subsampling is performed for each transition $s_t \to s_{t+1}$: $\mathcal{A}_t^* \sim \mathcal{P}(\mathcal{A})$.

$$\hat{\mathcal{L}}_{\text{TB}}(\tau) = \left(\log \frac{Z_\theta \prod_{t=1}^n \hat{P}_F(s_t|s_{t-1};\mathcal{A}_{t-1}^*;\theta)}{R(x)\prod_{t=1}^n P_B(s_{t-1}|s_t)}\right)^2 \tag{9}$$

For online training, we use the sampling policy π_θ proportional to $\hat{P}_F(-|-;\mathcal{A}^*;\theta)$, given by:

$$\pi_\theta(s'|s;\mathcal{A}^*) = \frac{w_{(s\to s')}F_\theta(s \to s')}{\sum_{(s\to s'')\in\mathcal{A}^*(s)} w_{(s\to s'')}F_\theta(s \to s'')} \tag{10}$$

3.4 Joint Selection of Templates and Blocks

For bi-molecular reactions, existing synthesis-oriented methods [17, 9, 33] formulated a hierarchical MDP which selects a reaction template r first and then the corresponding reactant block b sequentially from Eq. (11). However, as shown in Figure 2, the probability of selecting each reaction template r is fixed after training in a hierarchical MDP, and this rigidity can lead to incorrect policy estimates in modified block libraries. Therefore, we formulate a non-hierarchical MDP that jointly selects reaction templates and reactant blocks (r,b) at once, as given by Eq. (12), resulting in more consistent estimates of forward policy P_F.

$$P_F(T(s,(r,b))|s;\theta) = \frac{F_\theta(s,r)}{\sum_{r'\in\mathcal{R}_1\cup\mathcal{R}_2\cup\{\texttt{Stop}\}} F_\theta(s,r')} \times \frac{F_\theta(s,(r,b))}{\sum_{b'\in\mathcal{B}_r} F_\theta(s,(r,b'))} \tag{11}$$

$$P_F(T(s,(r,b))|s;\theta) = \frac{F_\theta(s,(r,b))}{\sum_{r'\in\mathcal{R}_1\cup\{\texttt{Stop}\}} F_\theta(s,r') + \sum_{r'\in\mathcal{R}_2}\sum_{b'\in\mathcal{B}_{r'}} F_\theta(s,(r',b'))} \tag{12}$$

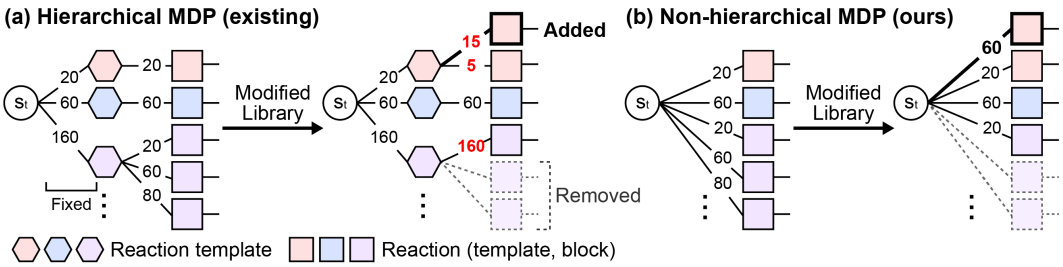

Figure 2: **Comparison of using modified building block library for the generation**: **(a)** a hierarchical MDP, and **(b)** a non-hierarchical MDP. More details are in Figure 7.

4 Experiments

Overview. We validate the effectiveness of RXNFLOW in two common SBDD tasks: pocket-specific optimization (Sec. 4.1) and pocket-conditional generation (Sec. 4.2). To the best of our knowledge, this is the first synthesis-oriented approach for pocket-conditional generation. We also investigate the applicability of RXNFLOW in real-world drug discovery pipelines where new further objectives may be introduced (Sec. 4.3) and the building block libraries are constantly expanded (Sec. 4.4). Lastly, we conduct an ablation study in Sec. 4.5 and a theoretical analysis in Sec. D.3.

Setup. We use the reaction template set constructed by Cretu et al. [9] including 13 uni- and 58 bi-molecular reaction templates. For the building blocks, we use 1.2M blocks from the Enamine comprehensive catalog. We use up to 3 reaction steps for generation following Enamine REAL Space [21], while SynFlowNet and RGFN allow 4 steps. For the subsampling policy, we set a sampling ratio of 1%. The experimental details are provided in Sec. C.

Synthesizability estimation. To assess the synthesizability of the generated compounds, we used the computationally intensive retrosynthetic analysis tool AiZynthFinder [18] with the Enamine building block library. We note that the molecule is identified as synthesizable only if it can be synthesized using the USPTO reactions [38] and given building blocks.

4.1 Pocket-Specific Optimization with GPU-accelerated Docking

Setup. Since GFlowNets sample a large number of molecules for online training, we employed a GPU-accelerated UniDock [59] with Vina scoring [56]. It is well known that docking can be hacked by increasing molecule size [46], so the appropriate constraints are required. We select QED [5] as a comprehensive molecular property constraint, QED>0.5, and set the reward function as $R(x) = w_1\text{QED}(x) + w_2\widehat{\text{Vina}}(x)$ where w_1, w_2 are used as the input of multi-objective GFlowNets [28] and are set to 0.5 for non-GFlowNet baselines. $\widehat{\text{Vina}}$ is a normalized docking score (Eq. (30)).

Each method generates up to 64,000 molecules for each of the 15 proteins in the LIT-PCBA dataset [55]. We then filter the molecules with the property constraint and select the top 100 diverse candidates based on the docking score, using a Tanimoto distance threshold of 0.5 to ensure structural diversity. The selected molecules are evaluated with the following metrics: **Hit ratio (%)** measures the fraction of *hits*, defined as the molecules that are identified as synthesizable by AiZynthFinder and having better docking scores than known active ligands [35]. **Vina (kcal/mol)** measures the average docking score. **Synthesizability (%)** is the fraction of synthesizable molecules. **Synthetic complexity**, which is highly correlated to yield and cost, is evaluated as the average number of synthesis steps [8].

Baselines. We perform comparisons to various synthetic-oriented approaches: genetic algorithm (**SynNet**) [17], conditional generative model (**BBAR**[2]) [53], and GFlowNets (**SynFlowNet**, **RGFN**) [9, 33]. For SynFlowNet and RGFN, we used 6,000 and 350 blocks, respectively, and set the maximum reaction step to 4 following the original papers. Moreover, we consider fragment-based GFlowNets (**FragGFN**) to analyze the effects of synthetic constraints on the performance. For FragGFN, we also consider additional synthesizability objectives with commonly-used synthetic accessibility score [SA; 12] (**FragGFN+SA**).

[2]Since BBAR requires labeled training data with QED and docking score, we perform docking with random 62,720 ZINC molecules for training and evaluate 1,280 samples according to the reported splitting ratio.

Table 1: **Hit ratio.** Mean and standard deviation over 4 runs. The best results are in bold.

		Hit Ratio (%, ↑)				
Category	Method	ADRB2	ALDH1	ESR_ago	ESR_antago	FEN1
Fragment	FragGFN	4.00 (± 3.54)	3.75 (± 1.92)	0.25 (± 0.43)	0.25 (± 0.43)	0.25 (± 0.43)
	FragGFN+SA	5.75 (± 1.48)	4.00 (± 1.58)	0.25 (± 0.43)	0.00 (± 0.00)	0.00 (± 0.00)
Reaction	SynNet	45.83 (± 7.22)	25.00 (± 25.00)	0.00 (± 0.00)	0.00 (± 0.00)	50.00 (± 0.00)
	BBAR	21.25 (± 5.36)	18.25 (± 1.92)	3.50 (± 1.12)	2.25 (± 1.09)	11.75 (± 2.59)
	SynFlowNet	52.75 (± 1.09)	57.00 (± 6.04)	30.75 (± 10.03)	11.25 (± 1.48)	53.00 (± 8.92)
	RGFN	46.75 (± 6.87)	39.75 (± 8.17)	4.50 (± 1.66)	1.25 (± 0.43)	19.75 (± 4.32)
	RxnFlow	**60.25** (± 3.77)	**63.25** (± 3.11)	**71.25** (± 4.15)	**46.00** (± 7.00)	**65.50** (± 4.09)

Table 2: **Vina.** Mean and standard deviation over 4 runs. The best results are in bold.

		Average Vina Docking Score (kcal/mol, ↓)				
Category	Method	ADRB2	ALDH1	ESR_ago	ESR_antago	FEN1
Fragment	FragGFN	-10.19 (± 0.33)	-10.43 (± 0.29)	-9.81 (± 0.09)	-9.85 (± 0.13)	-7.67 (± 0.71)
	FragGFN+SA	-9.70 (± 0.61)	-9.83 (± 0.65)	-9.27 (± 0.95)	-10.06 (± 0.30)	-7.26 (± 0.10)
Reaction	SynNet	-8.03 (± 0.26)	-8.81 (± 0.21)	-8.88 (± 0.13)	-8.52 (± 0.16)	-6.36 (± 0.09)
	BBAR	-9.95 (± 0.04)	-10.06 (± 0.14)	-9.97 (± 0.03)	-9.92 (± 0.05)	-6.84 (± 0.07)
	SynFlowNet	-10.85 (± 0.10)	-10.69 (± 0.09)	-10.44 (± 0.05)	-10.27 (± 0.04)	-7.47 (± 0.02)
	RGFN	-9.84 (± 0.21)	-9.93 (± 0.11)	-9.99 (± 0.11)	-9.72 (± 0.14)	-6.92 (± 0.06)
	RxnFlow	**-11.45** (± 0.05)	**-11.26** (± 0.07)	**-11.15** (± 0.02)	**-10.77** (± 0.04)	**-7.66** (± 0.02)

Table 3: **Synthesizability.** Mean and standard deviation over 4 runs. The best results are in bold.

		Percentage of Synthesizable Molecules (%, ↑)				
Category	Method	ADRB2	ALDH1	ESR_ago	ESR_antago	FEN1
Fragment	FragGFN	4.00 (± 3.54)	3.75 (± 1.92)	1.00 (± 1.00)	3.75 (± 1.92)	0.25 (± 0.43)
	FragGFN+SA	5.75 (± 1.48)	6.00 (± 2.55)	4.00 (± 2.24)	1.00 (± 0.00)	0.00 (± 0.00)
Reaction	SynNet	54.17 (± 7.22)	50.00 (± 0.00)	50.00 (± 0.00)	25.00 (± 25.00)	50.00 (± 0.00)
	BBAR	21.25 (± 5.36)	19.50 (± 3.20)	17.50 (± 1.50)	19.50 (± 3.64)	20.00 (± 2.12)
	SynFlowNet	52.75 (± 1.09)	57.00 (± 6.04)	53.75 (± 9.52)	56.50 (± 2.29)	53.00 (± 8.92)
	RGFN	46.75 (± 6.86)	47.50 (± 4.06)	50.25 (± 2.17)	49.25 (± 4.38)	48.50 (± 6.58)
	RxnFlow	**60.25** (± 3.77)	**63.25** (± 3.11)	**71.25** (± 4.15)	**66.50** (± 4.03)	**65.50** (± 4.09)

Table 4: **Synthetic complexity**. Mean and standard deviation over 4 runs. The best results are in bold.

		Average Number of Synthesis Steps (↓)				
Category	Method	ADRB2	ALDH1	ESR_ago	ESR_antago	FEN1
Fragment	FragGFN	3.60 (± 0.10)	3.83 (± 0.08)	3.76 (± 0.20)	3.76 (± 0.16)	3.34 (± 0.18)
	FragGFN+SA	3.73 (± 0.09)	3.66 (± 0.04)	3.66 (± 0.07)	3.67 (± 0.21)	3.79 (± 0.19)
Reaction	SynNet	3.29 (± 0.36)	3.50 (± 0.00)	3.00 (± 0.00)	4.13 (± 0.89)	3.50 (± 0.00)
	BBAR	3.60 (± 0.17)	3.62 (± 0.19)	3.76 (± 0.04)	3.72 (± 0.11)	3.59 (± 0.14)
	SynFlowNet	2.64 (± 0.07)	2.48 (± 0.07)	2.60 (± 0.25)	2.45 (± 0.09)	2.56 (± 0.29)
	RGFN	2.88 (± 0.21)	2.65 (± 0.09)	2.78 (± 0.19)	2.91 (± 0.23)	2.76 (± 0.17)
	RxnFlow	**2.42** (± 0.23)	**2.19** (± 0.12)	**1.95** (± 0.20)	**2.15** (± 0.18)	**2.23** (± 0.18)

Results. The results for the first five targets are shown in Tables 1 and 2, and additional results for the 10 remaining targets are reported in Sec. D.1. RxnFlow outperforms the baselines across all test proteins, demonstrating that the expanded sample space with the large action space enabled the model to generate more potent and diverse molecules. Additionally, as shown in Tables 3 and 4, RxnFlow ensures the synthesizability of the generated molecules more effectively than the other synthesis-oriented methods and GFlowNets which employ the same reactions as ours. These results support our primary assertion that existing synthesis-oriented approaches using more synthetic steps with smaller building block libraries can increase overall synthesis complexity and reduce synthesizability. Furthermore, FragGFN+SA does not show meaningful improvement in synthesizability, implying that optimization of a cheap synthesizability estimation function is suboptimal.

Table 5: **CrossDocked2020 benchmark**. We report the average and median values over the average properties for each test pocket. The best results are in bold and the second ones are in underlined. We denote the Validity as Valid, Synthesizability as Synthesiz, and Diversity as Div. Reference means the known binding active molecules in the test set.

		Valid (↑)	Vina (↓)		QED (↑)		Synthesiz (↑)		Div (↑)	Time (↓)
Category	Model	Avg.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Avg.
Reference	-	-	-7.71	-7.80	0.48	0.47	36.1%	-	-	-
Atom	Pocket2Mol	98.3%	-7.60	-7.16	0.57	0.58	29.1%	22.0%	0.83	2504[a]
	TargetDiff	91.5%	-7.37	-7.56	0.49	0.49	9.9%	3.2%	**0.87**	3428[a]
	DecompDiff	66.0%	-8.35	-8.25	0.37	0.35	0.9%	0.0%	0.84	6189[a]
Fragment	TacoGFN	**100.0%**	-8.24	-8.44	**0.67**	**0.67**	1.3%	1.0%	0.67	**4**[b]
Reaction	**RxnFlow**	**100.0%**	**-8.85**	**-9.03**	**0.67**	**0.67**	**34.8%**	**34.5%**	0.81	**4**

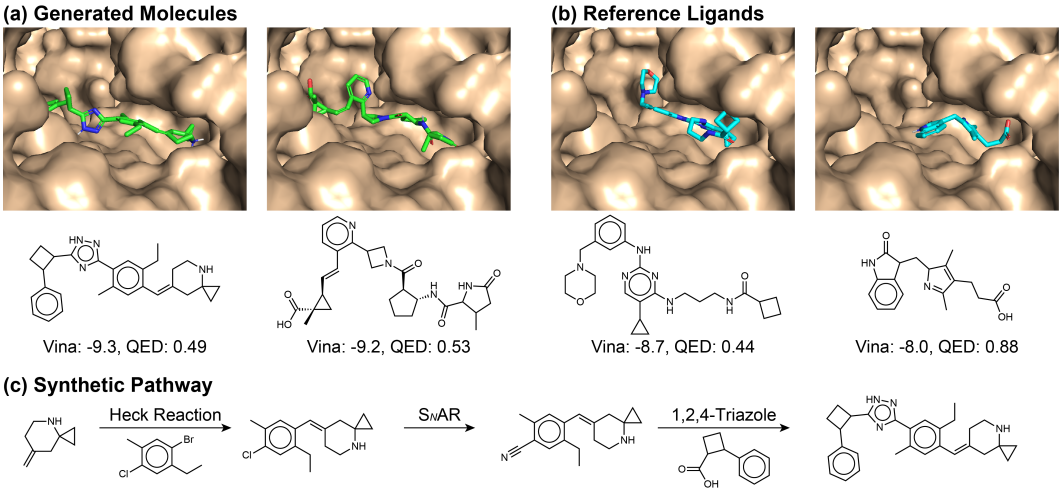

Figure 3: **Visualization of generated molecules** in a zero-shot manner. **(a-b)** Docking results of generated molecules and known reference ligands of TBK1 (PDB Id: 1FV, SU6). **(c)** Generative trajectory, which is the generated synthetic pathway of the left molecule in (a).

Furthermore, it is noteworthy that RxnFlow outperformed FragGFN which does not consider synthesizability. This improvement can be attributed to two key factors. First, the Enamine building block library is specifically curated for drug discovery, limiting the search space $\mathcal{X}$ to a drug-like chemical space and simplifying the optimization complexity. Second, RxnFlow needs shorter trajectories compared to FragGFNs since it assembles molecules with large building blocks instead of atoms or small fragments. This is beneficial for trajectory balance objectives where stochastic gradient variance tends to increase over longer trajectories [41].

4.2 Zero-Shot Sampling via Pocket Conditioning

Setup. We extend our works to a pocket-conditional generation problem to design binders for arbitrary pockets without additional training oracles [48, 36, 47, 22, 23, 51]. To address this challenge, we follow TacoGFN [54], which is a fragment-based GFlowNet for pocket-conditional generation. Since it requires more training oracles to learn pocket-conditional policies than target-specific generation, TacoGFN used pre-trained proxies that predict docking scores against arbitrary pockets using pharmacophore representation [52]. Since RxnFlow explicitly considers synthesizability, we exclude the SA score from the TacoGFN's reward function as described in Sec. C.2.

We generate 100 molecules for each of the 100 test pockets in the CrossDocked2020 benchmark [14] and evaluate them with the following metrics: **Vina (kcal/mol)** measures the average docking score from QuickVina [1]. **QED** measures the average drug-likeness of molecules. **Synthesizability (%)** is the fraction of synthesizable molecules. **Diversity** measures an average pairwise Tanimoto distance of ECFP4 fingerprints [42]. Moreover, we report the **Validity (%)** which is the percentage

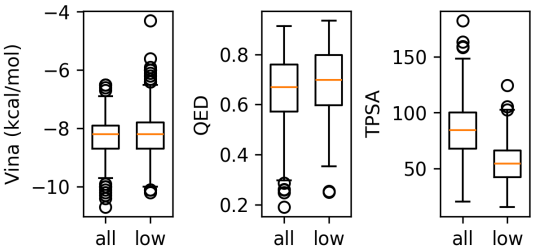

Figure 4: **Property distribution** of sampled molecules with "all" building blocks and "low"-TPSA building blocks. Vina score was calculated against the KRAS-G12C target.

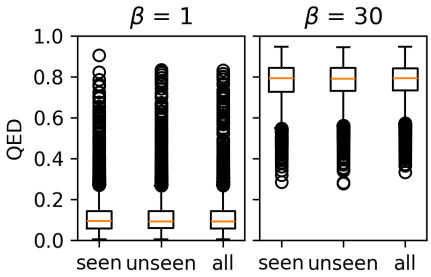

Figure 5: **QED reward distribution** of generated molecules for each of the "seen", "unseen", and "all" blocks. Additional results are in Figure 8.

of unique RDKit-readable molecules, and **Time (sec.)** which is the average sampling time to generate 100 molecules.

Baselines. We compare RxnFlow with state-of-the-art distribution learning-based models trained on a synthesizable drug set: an autoregressive model (**Pocket2Mol** [47]) and diffusion models (**TargetDiff** [22], **DecompDiff** [23]). We also perform a comparison with an optimization-based **TacoGFN** [54]. For a fair comparison with the distribution learning-based approaches, we used the docking proxy trained on CrossDocked2020.

Result. As shown in Table 5, RxnFlow achieves significant improvements in drug-related properties. In particular, RxnFlow attains the best average docking score of -8.85 kcal/mol, significantly outperforming TacoGFN at -8.24 kcal/mol and the state-of-the-art diffusion model (DecompDiff) at -8.35 kcal/mol while achieving high drug-like properties. Moreover, RxnFlow ensures the synthesizability comparable to known active ligands, outperforming the fragment-based TacoGFN trained on the SA score objective and the distributional learning-based models trained on synthesizable drug molecules. Figure 3 illustrates generated molecules against TANK-binding kinase 1 (TBK1) which is not included in the training set.

An important finding is that RxnFlow maintains high structural diversity (0.81) despite the typical trade-off between optimization power and diversity [16]. This is a significant improvement over the fragment-based TacoGFN, which scored 0.67 and is comparable to the distributional learning-based models that range from 0.83 to 0.87. We attribute this enhancement to our action space, which contains chemically diverse building blocks, in contrast to the small and limited fragment sets used in fragment-based GFlowNets. This suggests that our model can effectively balance the potency and diversity of generated molecules.

4.3 Introducing Additional Objective without Retraining

In drug discovery, new objectives often arise during the research process, such as enhancing solubility, reducing toxicity, or improving selectivity [13, 30]. These additional objectives typically not only require retraining models but also increase the optimization complexity. In this context, RxnFlow can achieve some additional objectives by simply introducing constraints to MDP without retraining thanks to the non-hierarchical action space (Sec. 3.4).

As shown in Figure 4, we explore the scenario of adding a solubility objective to a pre-trained GFlowNet in Sec. 4.2. Specifically, we target the generation of hydrophobic molecules by restricting the building blocks with topological polar surface area (TPSA) in the bottom 15% ("low") and sampled 500 molecules for the KRAS-G12C mutant (PDB Id: 6oim). While there are slight differences in the QED distributions due to the correlation between TPSA and QED, the generated molecules are more hydrophobic and retain similar overall reward distributions.

4.4 Scaling Action Space without Retraining

The building block libraries for drug discovery continue to grow, from 60,000 in 2020 to over 1.2 million blocks today, to enhance chemical diversity and novelty [21]. However, existing generative

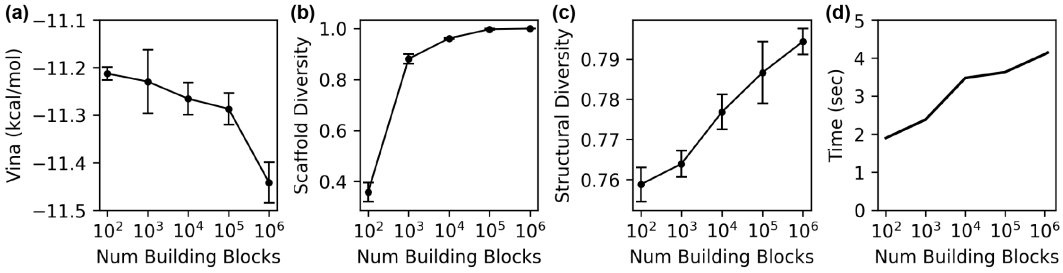

Figure 6: **Optimization power, diversity, and generation time** according to building block library size. **(a-c)** Average and standard deviation of properties of the top-1000 high-affinity molecules over 4 runs on pocket-specific generation. **(a)** Average docking score. **(b)** The uniqueness of Bemis-Murcko scaffolds. **(c)** Average Tanimoto distance. **(d)** Average runtime to generate 100 molecules over the CrossDocked2020 test pockets in a zero-shot manner.

models require retraining to accommodate newly discovered building blocks, limiting their scalability and adaptability. On the other hand, RXNFLOW can integrate new building blocks without retraining by understanding the chemical context of actions through action embedding. We first divide the 1M-sized block library ("all") into two sets: 500,000 blocks for training ("seen") and the remaining blocks ("unseen"). After training with the QED objective on various reward exponent settings (R^β), we generate 5,000 molecules from each set ("seen", "unseen", and "all"). Figure 5 shows that the reward distributions of samples are nearly identical, demonstrating that RXNFLOW performs robustly with unseen building blocks. This result highlights the generalization ability and scalability of RXNFLOW, a significant advantage for real-world applications.

4.5 Ablation Study: the effect of building block library size

Expanding the size of building block libraries provides an opportunity to discover more diverse and potent drug candidates [21]. In Sec. 4.1, RXNFLOW outperforms SynFlowNet and RGFN which use smaller block libraries, but differences in model architectures may have contributed to these results. To isolate the effect of the building block library size, we conduct an ablation study using partial libraries with a pocket-specific optimization task against the kappa-opioid receptor (PDB Id: 6b73), as illustrated in Figure 6(a-c). The results indicate that increasing the library size enhances both optimization power, in terms of docking scores, and diversity, in terms of a higher number of unique Bemis-Murcko scaffolds [2] and an increased Tanimoto diversity of the generated molecules. Additionally, as shown in Figure 6(d), the generation time only doubles on the 10,000-fold larger action space, highlighting the efficiency of RXNFLOW. These results demonstrate the forte of RXNFLOW in navigating a broader chemical space to discover novel drug candidates by overcoming the computational limitations for expanding the action space.

5 Conclusion

In this work, we introduce RXNFLOW, a synthesis-oriented generative framework designed to explore broader chemical spaces, thereby enhancing both diversity and potency for drug discovery. Our framework efficiently handles massive action spaces to expand the search space without significant computational or memory overhead by employing a novel action space subsampling technique. RXNFLOW can effectively identify diverse drug candidates with desired properties and synthetic feasibility by learning the objective of generative flow networks on synthetic pathways. Additionally, by formulating a non-hierarchical MDP, RXNFLOW can model generative flows on modified action spaces, allowing it to achieve additional objectives and incorporate newly discovered building blocks without retraining. These results highlight the potential of RXNFLOW as a practical and versatile solution for real-world drug discovery.

Acknowledgments and Disclosure of Funding

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF), grant-funded by the Ministry of Science and ICT (NRF-2023R1A2C2004376, RS-2023-00257479).

References

[1] A. Alhossary, S. D. Handoko, Y. Mu, and C.-K. Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

[2] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

[3] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

[4] Y. Bengio, S. Lahlou, T. Deleu, E. J. Hu, M. Tiwari, and E. Bengio. Gflownet foundations. *The Journal of Machine Learning Research*, 24(1):10006–10060, 2023.

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

[6] J. Bradshaw, B. Paige, M. J. Kusner, M. Segler, and J. M. Hernández-Lobato. A model to search for synthesizable molecules. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] A. Button, D. Merk, J. A. Hiss, and G. Schneider. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature machine intelligence*, 1 (7):307–315, 2019.

[8] C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen. Scscore: synthetic complexity learned from a reaction corpus. *Journal of chemical information and modeling*, 58(2):252–261, 2018.

[9] M. Cretu, C. Harris, J. Roy, E. Bengio, and P. Liò. Synflownet: Towards molecule design with guaranteed synthesis pathways. *arXiv preprint arXiv:2405.01155*, 2024.

[10] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

[11] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

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

[13] E. A. Fink, J. Xu, H. Hübner, J. M. Braz, P. Seemann, C. Avet, V. Craik, D. Weikert, M. F. Schmidt, C. M. Webb, et al. Structure-based discovery of nonopioid analgesics acting through the α2a-adrenergic receptor. *Science*, 377(6614):eabn7065, 2022.

[14] P. G. Francoeur, T. Masuda, J. Sunseri, A. Jia, R. B. Iovanisci, I. Snyder, and D. R. Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

[15] W. Gao and C. W. Coley. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 60(12):5714–5723, 2020.

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

[17] W. Gao, R. Mercado, and C. W. Coley. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FRxhHdnxt1.

[18] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, and E. Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

[19] S. K. Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, S. Blackburn, K. Thomas, C. Coley, J. Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International conference on machine learning*, pages 3668–3679. PMLR, 2020.

[20] D. E. Graff, E. I. Shakhnovich, and C. W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical science*, 12(22):7866–7881, 2021.

[21] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, and Y. S. Moroz. Generating multibillion chemical space of readily accessible screening compounds. *Iscience*, 23(11), 2020.

[22] J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng, and J. Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

[23] J. Guan, X. Zhou, Y. Yang, Y. Bao, J. Peng, J. Ma, Q. Liu, L. Wang, and Q. Gu. Decompdiff: Diffusion models with decomposed priors for structure-based drug design. *ICML*, 2023.

[24] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, and G. Schneider. Dogs: reaction-driven de novo design of bioactive compounds. *PLoS computational biology*, 8(2):e1002380, 2012.

[25] J. Horwood and E. Noutahi. Molecular design in synthetically accessible chemical space via deep reinforcement learning. *ACS omega*, 5(51):32984–32994, 2020.

[26] R. Huey, G. M. Morris, S. Forli, et al. Using autodock 4 and autodock vina with autodocktools: a tutorial. *The Scripps Research Institute Molecular Graphics Laboratory*, 10550(92037):1000, 2012.

[27] M. Jain, E. Bengio, A. Hernandez-Garcia, J. Rector-Brooks, B. F. Dossou, C. A. Ekbote, J. Fu, T. Zhang, M. Kilgour, D. Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pages 9786–9801. PMLR, 2022.

[28] M. Jain, S. C. Raparthy, A. Hernández-Garcıa, J. Rector-Brooks, Y. Bengio, S. Miret, and E. Bengio. Multi-objective gflownets. In *International conference on machine learning*, pages 14631–14653. PMLR, 2023.

[29] B. Jing, S. Eismann, P. Suriana, R. J. L. Townshend, and R. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

[30] T. Joshi, P. Sharma, T. Joshi, H. Pundir, S. Mathpal, and S. Chandra. Structure-based screening of novel lichen compounds against sars coronavirus main protease (mpro) as potentials inhibitors of covid-19. *Molecular Diversity*, 25:1665–1677, 2021.

[31] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[32] H. Kim, K. Lee, C. Kim, J. Lim, and W. Y. Kim. Dfrscore: deep learning-based scoring of synthetic complexity with drug-focused retrosynthetic analysis for high-throughput virtual screening. *Journal of Chemical Information and Modeling*, 64(7):2432–2444, 2023.

[33] M. Koziarski, A. Rekesh, D. Shevchuk, A. van der Sloot, P. Gaiński, Y. Bengio, C.-H. Liu, M. Tyers, and R. A. Batey. Rgfn: Synthesizable molecular generation using gflownets. *arXiv preprint arXiv:2406.08506*, 2024.

[34] G. Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

[35] S. Lee, J. Jo, and S. J. Hwang. Exploring chemical space with score-based out-of-distribution generation. In *International Conference on Machine Learning*, pages 18872–18892. PMLR, 2023.

[36] M. Liu, Y. Luo, K. Uchino, K. Maruhashi, and S. Ji. Generating 3d molecules for target protein binding. *arXiv preprint arXiv:2204.09410*, 2022.

[37] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

[38] D. Lowe. Chemical reactions from US patents (1976-Sep2016), 6 2017. URL https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

[39] S. Luo, J. Guan, J. Ma, and J. Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

[40] S. Luo, W. Gao, Z. Wu, J. Peng, C. W. Coley, and J. Ma. Projecting molecules into synthesizable chemical spaces. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=scFlbJQdm1.

[41] N. Malkin, M. Jain, E. Bengio, C. Sun, and Y. Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

[42] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2): 107–113, 1965.

[43] T. Nguyen, A. Tong, K. Madan, Y. Bengio, and D. Liu. Causal inference in gene regulatory networks with gflownet: Towards scalability in large systems. *arXiv preprint arXiv:2310.03579*, 2023.

[44] T. M. Nguyen, S. A. Tawfik, T. Tran, S. Gupta, S. Rana, and S. Venkatesh. Hierarchical gflownet for crystal structure generation. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*, 2023.

[45] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3:1–14, 2011.

[46] Y. Pan, N. Huang, S. Cho, and A. D. Mackerell. Consideration of molecular weight during compound selection in virtual target-based database screening. *Journal of chemical information and computer sciences*, 43(1):267–272, 2003.

[47] X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng, and J. Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pages 17644–17655. PMLR, 2022.

[48] M. Ragoza, T. Masuda, and D. R. Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022.

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

[50] A. A. Sadybekov, A. V. Sadybekov, Y. Liu, C. Iliopoulos-Tsoutsouvas, X.-P. Huang, J. Pickett, B. Houser, N. Patel, N. K. Tran, F. Tong, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature*, 601(7893):452–459, 2022.

[51] A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes, M. Welling, M. Bronstein, and B. Correia. Structure-based drug design with equivariant diffusion models, 2023.

[52] S. Seo and W. Y. Kim. Pharmaconet: Accelerating structure-based virtual screening by pharmacophore modeling. *arXiv preprint arXiv:2310.00681*, 2023.

[53] S. Seo, J. Lim, and W. Y. Kim. Molecular generative model via retrosynthetically prepared chemical building block assembly. *Advanced Science*, 10(8):2206674, 2023.

[54] T. Shen, M. Pandey, and M. Ester. Tacogfn: Target conditioned gflownet for structure-based drug design. *arXiv preprint arXiv:2310.03223*, 2023.

[55] V.-K. Tran-Nguyen, C. Jacquemard, and D. Rognan. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60(9):4263–4273, 2020.

[56] O. Trott and A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[57] S. Venkatraman, M. Jain, L. Scimeca, M. Kim, M. Sendera, M. Hasan, L. Rowe, S. Mittal, P. Lemos, E. Bengio, et al. Amortizing intractable inference in diffusion models for vision, language, and control. *arXiv preprint arXiv:2405.20971*, 2024.

[58] S. Wang, J. Witek, G. A. Landrum, and S. Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of chemical information and modeling*, 60(4):2044–2058, 2020.

[59] Y. Yu, C. Cai, J. Wang, Z. Bo, Z. Zhu, and H. Zheng. Uni-dock: Gpu-accelerated docking enables ultralarge virtual screening. *Journal of chemical theory and computation*, 19(11): 3336–3345, 2023.

[60] S. Yun, M. Jeong, S. Yoo, S. Lee, S. Y. Sean, R. Kim, J. Kang, and H. J. Kim. Graph transformer networks: Learning meta-path graphs to improve gnns. *Neural Networks*, 153:104–119, 2022.

[61] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[62] M. Zhou, Z. Yan, E. Layne, N. Malkin, D. Zhang, M. Jain, M. Blanchette, and Y. Bengio. Phylogfn: Phylogenetic inference with generative flow networks. *arXiv preprint arXiv:2310.08774*, 2023.

[63] W. Zhung, H. Kim, and W. Y. Kim. 3d molecular generative framework for interaction-guided drug design. *Nature Communications*, 15(1):2688, 2024.

A Theoretical Analysis

In this section, we provide the theoretical background of action subsampling. We define $\mathcal{U}$ as the uniform subsampling policy, i.e., $\mathcal{B}^* \sim \text{Uniform}(\{\mathcal{B}^* \subseteq \mathcal{B} \mid |\mathcal{B}^*| = k\})$.

Bias of log forward policy. In this section, we prove that the log forward policy estimation is unbiased (Eq. (7)) with the following weights:

$$w_a = \begin{cases} |\mathcal{B}|/|\mathcal{B}^*| & \text{if } a \in \mathcal{B} \\ |\mathcal{B}_r|/|\mathcal{B}_r^*| & \text{if } a \text{ is } (r,b) \text{ and } r \in \mathcal{R}_2 \\ 1 & \text{if } a \text{ is } \texttt{Stop} \text{ or } a \in \mathcal{R}_1 \end{cases}$$

For readability, we express the edge flow F and forward policy P_F where the action a is $s \to s'$ as follows:

$$F_\theta(s \to s'; \theta) = F_\theta(s, a; \theta), \quad P_F(s'|s;\theta) = P_F(a|s;\theta)$$

Then, the forward policy P_F and the estimated policy $\hat{P}_F$ (Eq. (7)) can be rewritten as follows:

$$P_F(s'|s;\theta) = P_F(a|s;\theta) = \frac{F_\theta(s,a)}{F_\theta(s)} = \frac{F_\theta(s,a)}{\sum_{a' \in \mathcal{A}(s)} F_\theta(s,a')}$$
$$\hat{P}_F(s'|s;\mathcal{A}^*;\theta) = \hat{P}_F(a|s;\mathcal{A}^*;\theta) = \frac{F_\theta(s,a)}{\hat{F}_\theta(s;\mathcal{A}^*)} = \frac{F_\theta(s,a)}{\sum_{a' \in \mathcal{A}^*(s)} w_{a'} F_\theta(s,a')}$$

The expectation of the estimated initial state flow $\hat{F}_\theta(s_0;\mathcal{A}^*) = \sum_{a \in \mathcal{A}^*} w_a F_\theta(s_0, a)$ is given by

$$\mathbb{E}_{\mathcal{A}^* \sim \mathcal{P}(\mathcal{A})}[\hat{F}_\theta(s_0;\mathcal{A}^*)] = \mathbb{E}_{\mathcal{B}^* \sim \mathcal{U}(\mathcal{B})}\left[\frac{|\mathcal{B}|}{|\mathcal{B}^*|} \sum_{b \in \mathcal{B}^*} F_\theta(s_0, b)\right] = \sum_{b \in \mathcal{B}} F_\theta(s_0, b) = F_\theta(s_0), \tag{13}$$

For a later state $s \neq s_0$, the expectation of the state flow $\hat{F}_\theta(s;\mathcal{A}^*)$ is given by

$$\begin{aligned} \mathbb{E}_{\mathcal{A}^* \sim \mathcal{P}(\mathcal{A})}[\hat{F}_\theta(s;\mathcal{A}^*)] &= \sum_{a \in \mathcal{R}_1 \cup \{\texttt{Stop}\}} F_\theta(s,a) + \sum_{r \in \mathcal{R}_2} \mathbb{E}_{\mathcal{B}_r^* \sim \mathcal{U}(\mathcal{B}_r)}\left[\frac{|\mathcal{B}_r|}{|\mathcal{B}_r^*|} \sum_{b \in \mathcal{B}_r^*} F_\theta(s,(r,b))\right] \\ &= \sum_{a \in \mathcal{R}_1 \cup \{\texttt{Stop}\}} F_\theta(s,a) + \sum_{r \in \mathcal{R}_2} \sum_{b \in \mathcal{B}_r} F_\theta(s,(r,b)) \\ &= P_F(s'|s;\theta) \end{aligned} \tag{14}$$

Variance of log forward policy. We define the standard deviation of the probability $P_F(-|s;\theta)$ as $\sigma_\theta(-|s)$. For the initial state s_0, the variance of $\log \hat{P}_F(s'|s_0;\mathcal{A}^*;\theta)$ is given by

$$\begin{aligned} &\text{Var}_{\mathcal{A}^* \sim \mathcal{P}(\mathcal{A})}\left[\log \hat{P}_F(s'|s_0;\mathcal{A}^*;\theta)\right] \\ &= \text{Var}_{\mathcal{B}^* \sim \mathcal{U}(\mathcal{B})}\left[\cancel{\log F_\theta(s_0 \to s')} - \log\left(\frac{\cancel{|\mathcal{B}|}}{\cancel{|\mathcal{B}^*|}} \sum_{b \in \mathcal{B}^*} F_\theta(s_0, b)\right)\right] \\ &= \text{Var}_{\mathcal{B}^* \sim \mathcal{U}(\mathcal{B})}\left[\log \sum_{b \in \mathcal{B}^*} P_F(b|s_0;\theta)\right] \qquad \text{(by normalizing with } F_\theta(s_0)) \\ &\approx \frac{|\mathcal{B}|^2(|\mathcal{B}| - |\mathcal{B}^*|)}{|\mathcal{B}^*|(|\mathcal{B}|-1)} \sigma_\theta(-|s_0)^2 \quad \text{(by Eq. (24))} \end{aligned} \tag{15}$$

For the later state $s \neq s_0$, the variance of $\log \hat{P}_F(s'|s;\mathcal{A}^*;\theta)$ is:

$$
\begin{aligned}
&\mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log \hat{P}_F(s'|s;\mathcal{A}^*;\theta)\right] \\
&= \mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log\left(\sum_{a\in\mathcal{R}_1\cup\{\texttt{Stop}\}} P_F(a|s;\theta) + \sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|}{|\mathcal{B}_r^*|}\sum_{b\in\mathcal{B}_r^*} P_F((r,b)|s;\theta)\right)\right] \\
&\approx \frac{\mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\cancel{\sum_{a\in\mathcal{R}_1\cup\{\texttt{Stop}\}} P_F(a|s;\theta)} + \sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|}{|\mathcal{B}_r^*|}\sum_{b\in\mathcal{B}_r^*} P_F((r,b)|s;\theta)\right]}{\mathbb{E}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\sum_{a\in\mathcal{R}_1\cup\{\texttt{Stop}\}} P_F(a|s;\theta) + \sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|}{|\mathcal{B}_r^*|}\sum_{b\in\mathcal{B}_r^*} P_F((r,b)|s;\theta)\right]^2} \\
&= \frac{1}{\cancel{\left(\sum_{a\in\mathcal{A}(s)} P_F(a|s;\theta)\right)^2}}\mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A}(s))}\left[\sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|}{|\mathcal{B}_r^*|}\sum_{b\in\mathcal{B}_r^*} P_F((r,b)|s;\theta)\right] \\
&= \sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|^2(|\mathcal{B}_r|-|\mathcal{B}_r^*|)}{|\mathcal{B}_r^*|(|\mathcal{B}_r|-1)}\sigma_\theta((r,-)|s)^2 \quad \text{(by Eq. (22))}
\end{aligned} \tag{16}
$$

Finally, we get the variance of $\log \hat{P}_F(\tau;\theta)$ where $\tau = (s_0 \to ... \to s_n = x)$:

$$
\begin{aligned}
&\mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log \hat{P}_F(\tau;\mathcal{A}^*;\theta)\right] \\
&\qquad\approx \frac{|\mathcal{B}|^2(|\mathcal{B}|-|\mathcal{B}^*|)}{k|\mathcal{B}^*|(|\mathcal{B}|-1)}\sigma_\theta(-|s_0)^2 + \sum_{t=1}^{n-1}\sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|^2(|\mathcal{B}_r|-|\mathcal{B}_r^*|)}{|\mathcal{B}_r^*|(|\mathcal{B}_r|-1)}\sigma_\theta((r,-)|s_t)^2
\end{aligned} \tag{17}
$$

Expectation of trajectory balance loss. Due to the variance of forward policy, there is a bias of the trajectory balance loss equal to the variance of $\log \hat{P}_F(\tau)$:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}[\hat{\mathcal{L}}_{\text{TB}}(\tau)] \\
&= \mathbb{E}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log\frac{Z_\theta \hat{P}_F(\tau;\theta)}{R(x)P_B(\tau|x)}\right]^2 + \mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log\frac{Z_\theta \hat{P}_F(\tau;\theta)}{R(x)P_B(\tau|x)}\right] \\
&= \mathcal{L}_{\text{TB}}(\tau) + \mathrm{Var}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\log \hat{P}_F(\tau;\theta)\right] \qquad (18) \\
&\approx \mathcal{L}_{\text{TB}}(\tau) + \frac{|\mathcal{B}|^2(|\mathcal{B}|-|\mathcal{B}^*|)}{|\mathcal{B}^*|(|\mathcal{B}|-1)}\sigma_\theta(-|s_0)^2 + \sum_{t=1}^{n-1}\sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|^2(|\mathcal{B}_r|-|\mathcal{B}_r^*|)}{|\mathcal{B}_r^*|(|\mathcal{B}_r|-1)}\sigma_\theta((r,-)|s_t)^2 \qquad (19)
\end{aligned}
$$

Therefore, the gradient of trajectory balance loss is:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}^*\sim\mathcal{P}(\mathcal{A})}\left[\nabla_\theta \hat{\mathcal{L}}_{\text{TB}}(\tau)\right] \\
&\approx \nabla_\theta \mathcal{L}_{\text{TB}}(\tau) + \frac{|\mathcal{B}|^2(|\mathcal{B}|-|\mathcal{B}^*|)}{|\mathcal{B}^*|(|\mathcal{B}|-1)}\nabla_\theta\sigma_\theta(-|s_0)^2 + \sum_{t=1}^{n-1}\sum_{r\in\mathcal{R}_2}\frac{|\mathcal{B}_r|^2(|\mathcal{B}_r|-|\mathcal{B}_r^*|)}{|\mathcal{B}_r^*|(|\mathcal{B}_r|-1)}\nabla_\theta\sigma_\theta((r,-)|s_t)^2
\end{aligned} \tag{20}
$$

Given that $\sigma_\theta(\cdot)$, which is the standard deviation of probability to select actions, is inversely proportional to $|\mathcal{B}|$ or $|\mathcal{B}_r|$, the bias is highly dependent on the subsampling size (e.g. $|\mathcal{B}^*|$) rather than the subsampling ratio (e.g. $|\mathcal{B}^*|/|\mathcal{B}|$). Moreover, we can decrease the bias by MC sampling for state flow estimation, and the bias is reversely proportional to the number of samples k. However, in the same computational cost (proportional to $\mathcal{O}(k \times \sum_r |\mathcal{B}_r^*|)$), using the larger partial action spaces without MC sampling is more precise than using smaller partial action spaces with multiple MC samples. In Sec. D.3, we expermentally show that the bias is relatively small to trajectory balance loss with the toy experiment.

A.1 Theoretical backgrounds

Variance of uniformly sampled subset For the set $\mathcal{X}$ with a size of n, we define its uniformly sampled subset with size m as $\mathcal{X}' \sim \mathcal{P}(\mathcal{X})$. For the function $F(\mathcal{X}) = \sum_{x \in \mathcal{X}} f(x)$, we define the unbiased estimation:

$$\hat{F}(\mathcal{X}) = \frac{n}{m} F(\mathcal{X}') = \frac{n}{m} \sum_{x \in \mathcal{X}'} f(x)$$

When the mean and standard deviation of $f(x)$ is μ_f and σ_f, variance of $\hat{F}(\mathcal{X})$ is:

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X}) \right] \\
&= \mathbb{E}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X})^2 \right] - \mathbb{E}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X}) \right]^2 \\
&= \frac{1}{{}^n C_m} \sum_{\mathcal{X}' \in \mathcal{P}(\mathcal{X})} \left(\frac{n}{m} \sum_{x \in \mathcal{X}'} f(x) \right)^2 - F(\mathcal{X})^2 \\
&= \frac{1}{{}^n C_m} \frac{n^2}{m^2} \left({}^{n-1} C_{m-1} \sum_{i=1}^{n} f(b_i)^2 + {}^{n-2} C_{m-2} \sum_{i=1}^{n} \sum_{j>i}^{n} 2 f(b_i) f(b_j) \right) - \left(\sum_{x \in \mathcal{X}} f(x) \right)^2 \\
&= \frac{n}{m} \sum_{i=1}^{n} f(x_i)^2 + \frac{2n(m-1)}{m(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} f(x_i) f(x_j) - \left(\sum_{i=1}^{n} f(x_i)^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} f(x_i) f(x_j) \right) \\
&= \frac{n-m}{m(n-1)} \left((n-1) \sum_{i=1}^{n} f(x_i)^2 - 2 \sum_{i=1}^{n} \sum_{j>i}^{n} f(x_i) f(x_j) \right) \\
&= \frac{n-m}{m(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} (f(x_i) - f(x_j))^2 \\
&= \frac{n^2(n-m)}{m(n-1)} \sigma_f^2
\end{aligned} \tag{21}$$

For MC sampling with k samples, the variance is:

$$\mathrm{Var}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X}) \right] = \frac{n^2(n-m)}{km(n-1)} \sigma_f^2 \tag{22}$$

The variance of $\log \hat{F}(\mathcal{X})$ is:

$$\mathrm{Var}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\log \hat{F}(\mathcal{X}) \right] \approx \frac{\mathrm{Var}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X}) \right]}{\mathbb{E}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\hat{F}(\mathcal{X}) \right]^2} = \frac{(n-m)}{km(n-1)} (\sigma_f / \mu_f)^2 \tag{23}$$

When the $F(\mathcal{X}) = 1$, the variance is:

$$\mathrm{Var}_{\mathcal{X}' \sim \mathcal{P}(\mathcal{X})} \left[\log \hat{F}(\mathcal{X}) \right] \approx \frac{m^2(n-m)}{km(n-1)} \sigma_f^2 \tag{24}$$

B RXNFLOW Architecture

B.1 Forward policy

For the model architecture, we used a graph transformer [60] as the backbone f_θ following Bengio et al. [3] and a multi-layer perceptron (MLP) for action embedding g_θ. The graph embedding dimension is d_1, and the building block embedding dimension is d_2. The molecular graph for a state is s, and the GFlowNet condition vector is c which includes a reward exponent and multi-objective optimization weights [28]. $a\|b$ means the concatenation of two feature vectors a and b.

Initial block selection. For the first action, the model always selects `AddFirstReactant` action which selects $b \in \mathcal{B}$ for the starting molecule with $\text{MLP}_{\texttt{AddFirstReactant}} : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$.

$$F_\theta(s_0, b, c) = \text{MLP}_{\texttt{AddFirstReactant}}(f_\theta(s_0, c)\|g_\theta(b)) \tag{25}$$

Reaction selection. For the later states $s \neq s_0$, the model calculates the logits for the `Stop` action, `ReactUni` actions $r_1 \in \mathcal{R}_1$, and `ReactBi` actions $(r_2, b) \in \mathcal{R}_2 \times \mathcal{B}$.

The logit for `Stop` is calculated by $\text{MLP}_{\texttt{Stop}} : \mathbb{R}^{d_1} \to \mathbb{R}$:

$$F_\theta(s, \texttt{Stop}, c) = \text{MLP}_{\texttt{Stop}}(f_\theta(s, c)). \tag{26}$$

The logit for a `ReactUni` action $r_1 \in \mathcal{R}_1$ is calculated by $\text{MLP}^{r_1}_{\texttt{ReactUni}} : \mathbb{R}^{d_1} \to \mathbb{R}$:

$$F_\theta(s, r_1, c) = \text{MLP}^{r_1}_{\texttt{ReactUni}}(f_\theta(s, c)). \tag{27}$$

Finally, the logit for a `ReactBi` action (r_2, b) is calculated by the one-hot embedding of reaction template $\delta(r_2) : \{0, 1\}^{|\mathcal{R}_2|}$ and $\text{MLP}_{\texttt{ReactBi}} : \mathbb{R}^{d_1+|\mathcal{R}_2|+d_2} \to \mathbb{R}$:

$$F_\theta(s, (r_2, b), c) = \text{MLP}_{\texttt{ReactBi}}(f_\theta(s, c)\|\delta(r_2)\|g_\theta(b)). \tag{28}$$

Pocket conditioning. For a pocket-conditional generation, the model uses a *K*-NN pocket residual graph $\mathcal{G}^P$ and encodes GVP-GNN [29] according to Shen et al. [54]. The pocket conditions are included in the GFlowNet condition vector c.

B.2 Backward policy.

Malkin et al. [41] introduced a uniform backward policy for small-scale drug discovery. However, in a directed acyclic graph (DAG) on synthetic pathways, where each state has numerous outgoing edges but few incoming ones, there is a significant imbalance in the number of trajectories along each incoming edge, depending on the distance from the initial state. In a uniform backward policy, where the flow of all incoming edges is equal, this imbalance diminishes the flow of trajectories that reach a state via shorter routes, i.e., those with fewer reaction steps. To facilitate shorter synthetic pathways as trajectories, we set the backward transition probability proportional to the expected number of trajectories along each incoming edge of the state, $a : s'' \to s$:

$$P_B(s''|s) = P((s'' \to s) \in \tau | s \in \tau) = \frac{\sum_{\tau \in \mathcal{T}: \tau = (s_0 \to \ldots \to s'' \to s)} |\mathcal{A}(s)|^{N-|\tau|}}{\sum_{\tau \in \mathcal{T}: \tau = (s_0 \to \ldots \to s)} |\mathcal{A}(s)|^{N-|\tau|}} \tag{29}$$

where N is the maximum length of trajectories.

B.3 Building block representation for action embedding.

To represent building blocks, we used both physicochemical properties, chemical structural properties, and topological properties. For physicochemical properties, we used 8 molecular features: molecular weight, the number of atoms, the number of H-bond acceptors/donors (HBA, HBD), the number of aromatic/non-aromatic rings, LogP, and TPSA. For chemical properties, we used the MACCS fingerprint [11], which represents the composition of the chemical functional groups in the molecule. For topological information, we used the Morgan ECFP4 fingerprint [42] with a dimension of 1024, which is widely used in fingerprint-based deep learning researches. All properties are calculated with RDKit [34].

B.4 Non-hierarchical Markov decision process

Table 6: Comparison with SynFlowNet and RGFN. RGFN investigates scaling up to 64,000 building blocks, but their experimental validation and proof-of-principle implementations use only 350.

Methods	Hierarchical	Action Embedding	`ReactUni`	Num Blocks	Max Reaction Steps
Enamine REAL	-	-	◯	>1M	3
SynFlowNet	◯	×	◯	6,000	4
RGFN	◯	◯	×	350∼64,000	4
RxnFlow	×	◯	◯	>1M	3

While RxnFlow shares similar rules with SynFlowNet [9] and RGFN [33], there are several major differences as described in Table 6. In particular, we formulate the non-hierarchical MDP instead of the hierarchical MDP for bi-molecular reactions.

To describe the difference, we supposed the situation that the toxicity is observed in molecules containing a functional group $-NR_3$, thereby excluding blocks with the function group from the building block library. In trained hierarchical MDP, the modification of the block library cannot change the probability of reaction templates, leading to the overestimation or underestimation of edge flows (the red color of Figure 7(a)). However, in the non-hierarchical MDP, the exclusion of some reactants does not affect actions that share the same reaction template (Figure 7(b)).

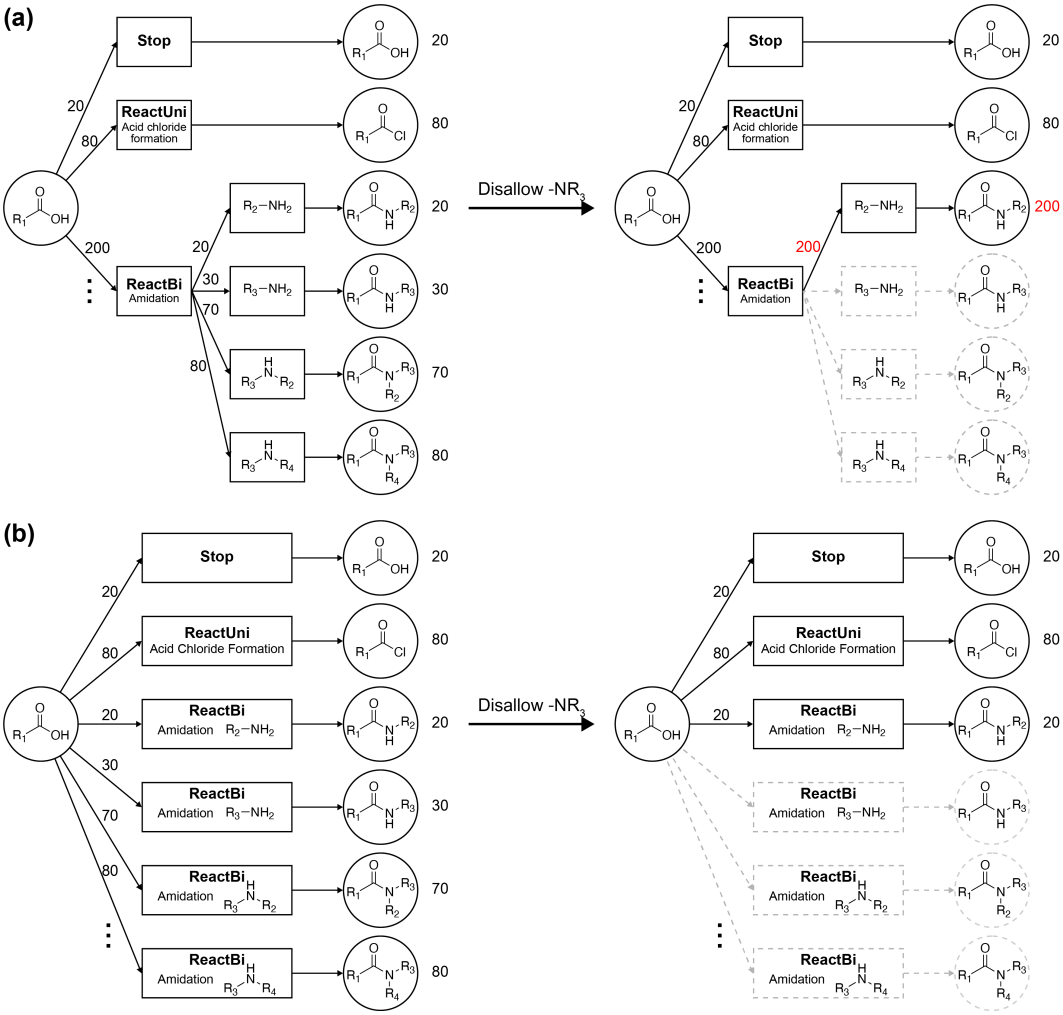

Figure 7: **Illustration of a situation** where an additional objective is introduced: excluding building blocks (reactants) containing the functional group $-NR_3$. The Gray dashed lines mean masked actions. **(a)** GFlowNet on hierarchical MDP. **(b)** GFlowNet on non-hierarchical MDP.

C Experimental details

C.1 Action space details

Reaction templates. In this work, we used the reaction template set constructed by Cretu et al. [9] from two public collections Hartenfeller et al. [24], Button et al. [7]. The entire reaction template set includes 71 reaction templates, 13 for uni-molecular reactions and 58 for bi-molecular reactions. In *in silico* reactions with bi-molecular reaction templates, the products depend on the order of the input reactants. To ensure the consistency of the action, we consider two templates for each bi-molecular template according to the order of reactants, i.e. $|\mathcal{R}_1| = 13, |\mathcal{R}_2| = 116$. We note that our template set does not contain templates that have the same first and second reactant patterns.

Building blocks. We used the Enamine comprehensive catalog with 1,309,385 building blocks released on 2024.06.10 [21]. We filtered out building blocks that are not RDKit-readable, have no possible reactions, or contain unallowable atoms[3], resulting in 1,193,871 remaining blocks.

C.2 GFlowNet Training details

To minimize optimization performance influencing factors, we mostly followed the standard GFlowNet's model architecture and hyperparameters[4] except for some parameters in Table 7. All experiments were performed on a single NVIDIA RTX A4000 GPU.

Table 7: **Default hyperparameters** used in RXNFLOW training.

Hyperparameters	Values
Minimum trajectory length	2 (minimum reaction steps: 1)
Maximum trajectory length	4 (maximum reaction steps: 3)
GFN temperature β	Uniform(0, 64)
Train random action probability	0.05 (5%)
Action space subsampling ratio	1%
Building block embedding size	64

For action space subsampling, we randomly subsample 1% actions for `AddFirstReactant` and each bi-molecular reaction template $r \in \mathcal{R}_2$. However, for bi-molecular reactions with small possible reactant block sets $\mathcal{B}_r \in \mathcal{B}$, the memory benefit from the action space subsampling is small while a variance penalty is large. Therefore, we set the minimum subsampling size to 100 for each bi-molecular reaction, and the action space subsampling is not performed when the number of actions is smaller than 100.

The number of actions for each action type is imbalanced, and the number of reactant blocks ($\mathcal{B}_r$) for each bi-molecular reaction template r is also imbalanced. This can make some rare action categories not being sampled during training. We empirically found that `ReactBi` action were only sampled during 20,000 iterations (1.28M samples) in a toy experiment that uses one bi-molecular reaction template and 10,000 building blocks in some random seeds. Therefore, we set the random action probability as the default of 5%, and the model uniformly samples each action category in the random action sampling. This prevents incorrect predictions by ensuring that the model experiences trajectories including rare actions. We note that this random selection is only performed during model training.

[3]The allowable atom types are B, C, N, O, F, P, S, Cl, Br, I.

[4]Default hyperparameters in https://github.com/recursionpharma/gflownet/blob/trunk/src/gflownet/tasks/seh_frag.py

C.3 Experimental details

Pocket-specific optimization. For pocket-specific optimization with GPU-accelerated UniDock, we normalize the Vina and SA scores for multi-objective optimization:

$$\widehat{\text{Vina}}(x) = -0.1 \max(\text{Vina}(x), 0), \quad \widehat{\text{SA}}(x) = (10 - \text{SA}(x))/9. \tag{30}$$

For FragGFN, we set the maximum trajectory length to 9, and for SynFlowNet and RGFN, we used the same hyperparameters as our framework except for the maximum trajectory length and building block library size. According to their default setting, we set a the maximum trajectory length to 5 rather than 4, and we randomly sampled 350 building blocks for RGFN and 6000 building blocks for SynFlowNet. Moreover, we do not perform action subsampling for SynFlowNet and RGFN.

Pocket-conditional generation in a zero-shot manner. We used the modified version of the TacoGFN's reward function and training set. Since we don't need to optimize SA [12], we excluded the SA term from the reward functions:

$$r_{\text{affinity}}(x) = \begin{cases} 0 & \text{if } 0 \leq \text{Proxy}(x) \\ -0.04 \times \text{Proxy}(x) & \text{if } -8 \leq \text{Proxy}(x) \leq 0 \\ -0.2 \times \text{Proxy}(x) - 1.28 & \text{if } -13 \leq \text{Proxy}(x) \leq -8 \\ 1.32 & \text{if Proxy}(x) \leq -13 \end{cases}$$

$$r_{\text{QED}}(x) = \begin{cases} \text{QED}(x)/0.7 & \text{if QED}(x) \leq 0.7 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{Reward}(x) = \frac{r_{\text{affinity}}(x) \times r_{\text{QED}}(x)}{\sqrt[3]{\text{HeavyAtomCounts}(x)}} \tag{31}$$

For hyperparameters, we set the pocket embedding dimension to 128 and the training GFN temperature to Uniform(0, 64) which are used in TacoGFN. We trained the model with 40,000 oracles whereas TacoGFN is trained for 50,000 oracles.

Introducing further objectives without retraining. For the restricted block library (TPSA<30), we set the action space subsampling ratio as 10% for both `AddFirstReactant` and `ReactBi`, and we set that as 1% for an entire library.

Scaling action space without retraining. For action space subsampling, we set the subsampling ratio as 2% for both `AddFirstReactant` and `ReactBi` for the "seen" and "unseen" libraries. For the "all" library ("seen" + "unseen"), we set the subsampling ratio as 1%.

Ablation study. We set different subsampling ratios according to the building block library size. For 100-sized, 1k-sized, and 10k-sized libraries, we do not perform the action space subsampling. We set an action space subsampling ratio of 10% for a 100k-sized one and 1% for a 1M-sized one.

C.4 Softwares

Molecular docking software. For a fair comparison with the baseline model, we used UniDock [59] for target-specific generation and QuickVina 2.1 [1] for SBDD. The initial ligand conformer is generated with srETKDG3 [58] in RDKit [34]. For QuickVina, we converted the molecule format to pdbqt with OpenBabel [45] and AutoDock Tools [26]. To set up an exhaustive search, we set the search mode to *balance* for UniDock and the exhaustiveness to 8 for QuickVina. We kept the seed fixed at 1 throughout the ETKDG and the whole docking process.

Docking proxy. We used the QuickVina proxy proposed by Shen et al. [54] which is implemented in PharmacoNet [52]. We used a proxy model trained on the CrossDocked2020 training set rather than the model trained on the ZINCDock15M training set.

Synthetic accessibility estimation. To evaluate the synthetic accessibility of molecules, we used the retrosynthesis planning tool AiZynthFinder [18]. AiZynthFinder uses MCTS to find synthesis paths and estimate the number of steps, search time, success rate, and synthetically accessible score as metrics to indicate synthesis complexity or synthesizability.

C.5 Baselines

SynNet. For SynNet [17], we perform multi-objective optimization with the following reward function:

$$R(x) = 0.5\text{QED}(x) + 0.5\widehat{\text{Vina}}(x). \tag{32}$$

According to the standard setting for optimization, we set the number of offspring to 512 and the number of oracles to 125. To use pre-trained models, we used SynNet's template set (91 templates) instead of a template set (71 templates) used in our work.

RGFN. Since the code of RGFN is not released, we reimplement the RGFN.

BBAR. Since BBAR [53] allows multi-conditional generation, we directly used QED and docking scores without any processing. We split 64,000 ZINC20 molecules according to the reported splitting of BBAR: 90% for the training set, 8% for the validation set, and 2% for the test set (in our case, the number of sampling molecules). We performed UniDock for training and validation set to prepare the label of the molecules. Since BBAR requires the desired property value, we used the average docking score of the top 100 diverse modes from our model.

Pocket2Mol, TargetDiff, DecompDiff, TacoGFN. We followed the reported generative setting to generate 100 molecules for each CrossDocked test pocket. We set the center of the pockets with the reference ligands in the CrossDocked2020 database. We reuse reported runtime in Shen et al. [54], which is measured on NVIDIA A100 for Pocket2Mol, TargetDiff, and DecompDiff, and NVIDIA RTX3090 for TacoGFN.

C.6 LIT-PCBA Pockets

Table 8 describes the protein information used in pocket-specific optimization with UniDock, which is performed on Sec. 4.1.

Table 8: **The basic target information** of the LIT-PCBA dataset and PDB entry used in this work.

Target	PDB Id	Target name
ADRB2	4ldo	Beta2 adrenoceptor
ALDH1	5l2m	Aldehyde dehydrogenase 1
ESR_ago	2p15	Estrogen receptor α with agonist
ESR_antago	2iok	Estrogen receptor α with antagonist
FEN1	5fv7	FLAP Endonuclease 1
GBA	2v3d	Acid Beta-Glucocerebrosidase
IDH1	4umx	Isocitrate dehydrogenase 1
KAT2A	5h86	Histone acetyltransferase KAT2A
MAPK1	4zzn	Mitogen-activated protein kinase 1
MTORC1	4dri	PPIase domain of FKBP51, Rapamycin
OPRK1	6b73	Kappa opioid receptor
PKM2	4jpg	Pyruvate kinase muscle isoform M1/M2
PPARG	5y2t	Peroxisome proliferator-activated receptor γ
TP53	3zme	Cellular tumor antigen p53
VDR	3a2i	Vitamin D receptor

D Additional results

D.1 Additional results for Pocket-specific generation task

We reported the additional results of Sec. 4.1 for the remaining 10 pockets on the LIT-PCBA benchmark.

Table 9: **Hit ratio** (%). Average and standard deviation for 4 runs. The best results are in bold.

		Hit ratio (%, ↑)				
Category	Method	GBA	IDH1	KAT2A	MAPK1	MTORC1
Fragment	FragGFN	5.00 (± 4.24)	4.50 (± 1.66)	1.25 (± 0.83)	0.75 (± 0.83)	0.00 (± 0.00)
	FragGFN+SA	3.00 (± 1.00)	4.50 (± 4.97)	1.50 (± 0.50)	2.00 (± 1.73)	0.00 (± 0.00)
Reaction	SynNet	50.00 (± 0.00)	50.00 (± 0.00)	29.17 (± 18.16)	37.50 (± 21.65)	0.00 (± 0.00)
	BBAR	17.75 (± 2.28)	19.50 (± 1.50)	18.75 (± 1.92)	16.25 (± 3.49)	0.00 (± 0.00)
	SynFlowNet	58.00 (± 4.64)	59.00 (± 4.06)	55.50 (± 10.23)	47.25 (± 6.61)	0.00 (± 0.00)
	RGFN	48.00 (± 1.22)	43.00 (± 2.74)	49.00 (± 1.22)	38.00 (± 4.12)	0.00 (± 0.00)
	RxnFlow	**66.00** (± 1.58)	**64.00** (± 5.05)	**66.50** (± 2.06)	**63.00** (± 4.64)	0.00 (± 0.00)
		OPRK1	PKM2	PPARG	TP53	VDR
Fragment	FragGFN	0.50 (± 0.50)	7.25 (± 1.92)	0.75 (± 0.43)	4.25 (± 1.64)	0.00 (± 0.00)
	FragGFN+SA	0.50 (± 0.87)	4.50 (± 1.50)	1.00 (± 0.71)	2.25 (± 1.92)	0.00 (± 0.00)
Reaction	SynNet	0.00 (± 0.00)	0.00 (± 0.00)	33.33 (± 20.41)	8.33 (± 14.43)	0.00 (± 0.00)
	BBAR	2.50 (± 1.12)	20.00 (± 0.71)	10.50 (± 2.69)	14.00 (± 3.94)	0.00 (± 0.00)
	SynFlowNet	23.50 (± 5.94)	50.75 (± 1.09)	53.50 (± 5.68)	55.50 (± 9.94)	0.00 (± 0.00)
	RGFN	2.50 (± 2.06)	34.75 (± 6.57)	29.00 (± 6.52)	37.00 (± 6.60)	0.00 (± 0.00)
	RxnFlow	**72.25** (± 2.05)	**62.00** (± 3.24)	**65.50** (± 4.03)	**67.50** (± 2.96)	**1.75** (± 0.83)

Table 10: **Vina**. Average and standard deviation for 4 runs. The best results are in bold.

		Average Vina Docking Score (kcal/mol, ↓)				
Category	Method	GBA	IDH1	KAT2A	MAPK1	MTORC1
Fragment	FragGFN	-8.76 (± 0.46)	-9.91 (± 0.32)	-9.27 (± 0.20)	-8.93 (± 0.18)	-10.51 (± 0.31)
	FragGFN+SA	-8.92 (± 0.27)	-9.76 (± 0.64)	-9.14 (± 0.43)	-8.28 (± 0.40)	-10.14 (± 0.30)
Reaction	SynNet	-7.60 (± 0.09)	-8.74 (± 0.08)	-7.64 (± 0.38)	-7.33 (± 0.14)	-9.30 (± 0.45)
	BBAR	-8.70 (± 0.05)	-9.84 (± 0.09)	-8.54 (± 0.06)	-8.49 (± 0.07)	-10.07 (± 0.16)
	SynFlowNet	-9.27 (± 0.06)	-10.40 (± 0.08)	-9.41 (± 0.04)	-8.92 (± 0.05)	-10.84 (± 0.03)
	RGFN	-8.48 (± 0.06)	-9.49 (± 0.13)	-8.53 (± 0.11)	-8.22 (± 0.15)	-9.89 (± 0.06)
	RxnFlow	**-9.62** (± 0.04)	**-10.95** (± 0.05)	**-9.73** (± 0.03)	**-9.30** (± 0.01)	**-11.39** (± 0.09)
		OPRK1	PKM2	PPARG	TP53	VDR
Fragment	FragGFN	-10.28 (± 0.15)	-11.24 (± 0.27)	-9.54 (± 0.12)	-7.90 (± 0.02)	-10.96 (± 0.06)
	FragGFN+SA	-9.58 (± 0.44)	-10.83 (± 0.34)	-9.19 (± 0.29)	-7.61 (± 0.27)	-10.66 (± 0.61)
Reaction	SynNet	-8.70 (± 0.36)	-9.55 (± 0.14)	-7.47 (± 0.34)	-5.34 (± 0.23)	-10.98 (± 0.57)
	BBAR	-9.84 (± 0.10)	-11.39 (± 0.08)	-8.69 (± 0.10)	-7.05 (± 0.09)	-11.07 (± 0.04)
	SynFlowNet	-10.34 (± 0.07)	-11.98 (± 0.12)	-9.40 (± 0.05)	-7.90 (± 0.10)	-11.62 (± 0.13)
	RGFN	-9.61 (± 0.11)	-10.96 (± 0.18)	-8.53 (± 0.07)	-7.07 (± 0.06)	-10.86 (± 0.11)
	RxnFlow	**-10.84** (± 0.03)	**-12.53** (± 0.02)	**-9.73** (± 0.02)	**-8.09** (± 0.06)	**-12.30** (± 0.07)

Table 11: **Synthesizability**. Average and standard deviation for 4 runs. The best results are in bold.

Category	Method	Percentage of Synthesizable Molecules (%, ↑)				
		GBA	IDH1	KAT2A	MAPK1	MTORC1
Fragment	FragGFN	5.00 (± 4.24)	4.50 (± 1.66)	1.25 (± 0.83)	0.75 (± 0.83)	2.75 (± 1.30)
	FragGFN+SA	3.00 (± 1.00)	4.50 (± 4.97)	1.50 (± 0.50)	3.25 (± 1.48)	3.50 (± 2.50)
Reaction	SynNet	50.00 (± 0.00)	50.00 (± 0.00)	45.83 (± 27.32)	50.00 (± 0.00)	54.17 (± 7.22)
	BBAR	17.75 (± 2.28)	19.50 (± 1.50)	18.75 (± 1.92)	16.25 (± 3.49)	18.75 (± 3.90)
	SynFlowNet	58.00 (± 4.64)	59.00 (± 4.06)	55.50 (± 10.23)	47.25 (± 6.61)	57.00 (± 7.58)
	RGFN	48.00 (± 1.22)	43.00 (± 2.74)	49.00 (± 1.22)	42.00 (± 3.00)	44.50 (± 4.03)
	RXNFLOW	**66.00** (± 1.58)	**64.00** (± 5.05)	**66.50** (± 2.06)	**63.00** (± 4.64)	**70.50** (± 2.87)
		OPRK1	PKM2	PPARG	TP53	VDR
Fragment	FragGFN	2.50 (± 2.29)	8.75 (± 3.11)	0.75 (± 0.43)	4.25 (± 1.64)	3.50 (± 2.18)
	FragGFN+SA	3.25 (± 1.79)	9.75 (± 2.28)	1.25 (± 1.09)	2.25 (± 1.92)	3.75 (± 2.77)
Reaction	SynNet	54.17 (± 7.22)	50.00 (± 0.00)	54.17 (± 7.22)	29.17 (± 18.16)	45.83 (± 7.22)
	BBAR	13.75 (± 3.11)	20.00 (± 0.71)	15.50 (± 2.29)	18.50 (± 3.28)	12.25 (± 3.34)
	SynFlowNet	56.50 (± 7.63)	50.75 (± 1.09)	53.50 (± 5.68)	55.50 (± 9.94)	53.50 (± 1.80)
	RGFN	48.00 (± 2.55)	48.50 (± 3.20)	47.00 (± 5.83)	53.25 (± 3.63)	46.50 (± 2.69)
	RXNFLOW	**72.25** (± 2.05)	62.00 (± 3.24)	**65.50** (± 4.03)	**67.50** (± 2.96)	**66.75** (± 2.28)

Table 12: **Synthetic complexity**. Average and standard deviation for 4 runs. The best results are in bold.

Category	Method	Average Number of Synthesis Steps (↓)				
		GBA	IDH1	KAT2A	MAPK1	MTORC1
Fragment	FragGFN	3.94 (± 0.11)	3.74 (± 0.10)	3.78 (± 0.09)	3.72 (± 0.18)	3.84 (± 0.18)
	FragGFN+SA	3.94 (± 0.15)	3.84 (± 0.23)	3.66 (± 0.18)	3.69 (± 0.21)	3.94 (± 0.08)
Reaction	SynNet	3.38 (± 0.22)	3.38 (± 0.22)	3.46 (± 0.95)	3.50 (± 0.00)	3.29 (± 0.36)
	BBAR	3.71 (± 0.12)	3.68 (± 0.02)	3.63 (± 0.05)	3.73 (± 0.05)	3.77 (± 0.09)
	SynFlowNet	2.48 (± 0.18)	2.61 (± 0.13)	2.45 (± 0.37)	2.81 (± 0.24)	2.44 (± 0.27)
	RGFN	2.77 (± 0.20)	2.97 (± 0.15)	2.78 (± 0.10)	2.86 (± 0.19)	2.92 (± 0.06)
	RXNFLOW	**2.10** (± 0.08)	**2.16** (± 0.11)	**2.29** (± 0.05)	**2.29** (± 0.11)	**2.05** (± 0.09)
		OPRK1	PKM2	PPARG	TP53	VDR
Fragment	FragGFN	3.82 (± 0.13)	3.71 (± 0.12)	3.73 (± 0.24)	3.73 (± 0.23)	3.75 (± 0.06)
	FragGFN+SA	3.62 (± 0.12)	3.84 (± 0.21)	3.71 (± 0.04)	3.66 (± 0.05)	3.67 (± 0.25)
Reaction	SynNet	3.29 (± 0.36)	3.50 (± 0.00)	3.29 (± 0.36)	3.67 (± 0.91)	3.63 (± 0.22)
	BBAR	3.70 (± 0.17)	3.61 (± 0.05)	3.72 (± 0.13)	3.65 (± 0.05)	3.77 (± 0.16)
	SynFlowNet	2.49 (± 0.33)	2.62 (± 0.10)	2.56 (± 0.12)	2.51 (± 0.27)	2.55 (± 0.09)
	RGFN	2.81 (± 0.12)	2.82 (± 0.10)	2.82 (± 0.18)	2.64 (± 0.10)	2.84 (± 0.18)
	RXNFLOW	**2.00** (± 0.09)	**2.34** (± 0.19)	**2.21** (± 0.06)	**2.12** (± 0.12)	**2.12** (± 0.12)

D.2 Additional results for scaling action space without retraining

We reported the additional results for additional reward exponent settings (R^β).

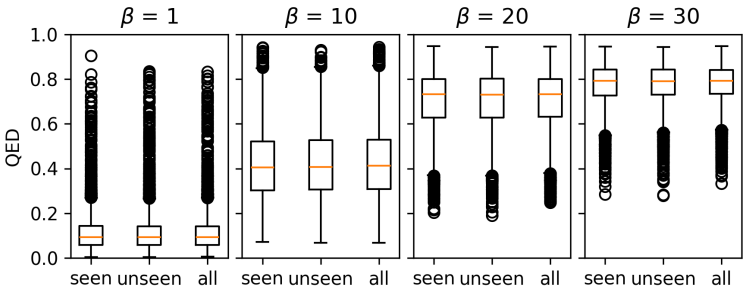

Figure 8: **QED reward distribution** of generated molecules.

D.3 Theoretical analysis

To assess the impact of action space subsampling in GFlowNet training, we conduct a toy experiment using a simplified setup with 10,000 blocks, one uni-molecular reaction template, one bi-molecular reaction template, and the QED objective. We used the Hell-Volhard-Zelinsky reaction as a uni-molecular reaction and the Amide reaction as a bi-molecular reaction, which are illustrated in Fig. 9. We used a minimum trajectory length of 1, max trajectory length of 2, constant GFN temperature of 1.0, and learning rate decay of 3,000 for P_F and $\log Z$. For the GFlowNet sampler, we used the same weights of the proxy model, i.e. EMA factor of 0. We performed optimization for 30,000 oracles with a batch size of 64.

As shown in Figures 10(a) and 10(b), we compare a baseline GFlowNet trained without subsampling ("base") to models using various subsampling ratios and Monte Carlo (MC) sampling. The differences in $\log Z_\theta$ are relatively small (<0.005) across all settings, and increasing MC samples for state flow estimation F_θ further reduced the bias. In Figures 10(c) and 10(d), we also evaluate the bias in the trajectory balance loss ($\hat{\mathcal{L}}_{\text{TB}}$) and its gradient norm ($\|\nabla_\theta \hat{\mathcal{L}}_{\text{TB}}\|$) during training, finding negligible differences compared to the true values. These results indicate that our importance sampling reweighting approach effectively mitigates bias from action space subsampling, enabling efficient and accurate policy estimation.

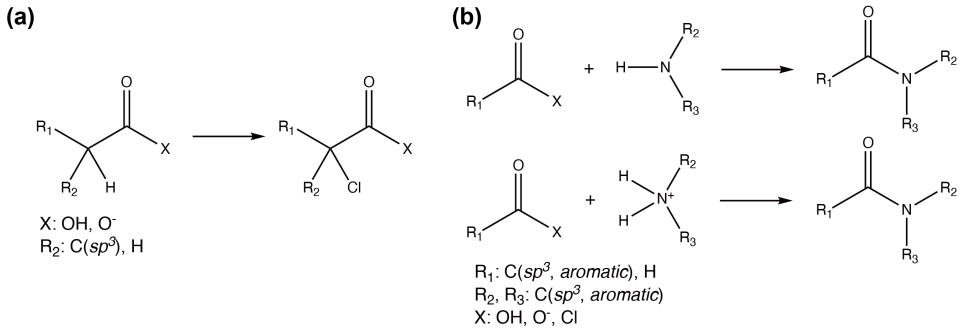

Figure 9: **Reaction templates employed in toy experiments**. **(a)** Hell-Volhard-Zelinsky reaction. **(b)** Amide reaction.

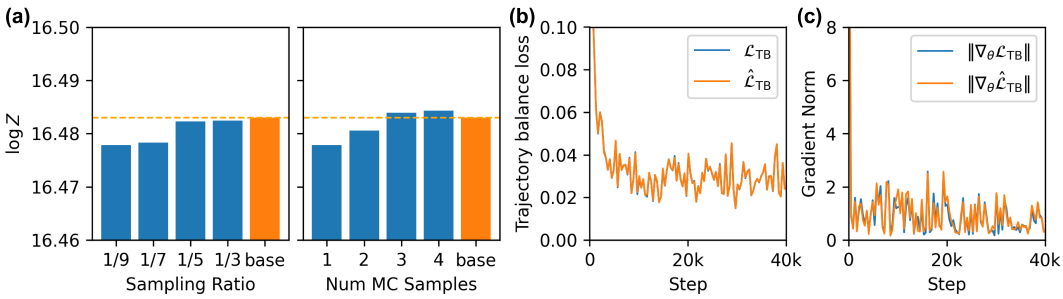

Figure 10: **Bias estimation. (a)** $\log Z_\theta$ according to the action space subsampling ratio (left) and the number of MC samples where the subsampling ratio is 1/9 (right). **(b)** The trajectory balance loss ($\mathcal{L}_{\text{TB}}, \hat{\mathcal{L}}_{\text{TB}}$) where the subsampling ratio is 1/9 under 4 MC samples. **(c)** The loss gradient norms ($\|\nabla_\theta \mathcal{L}_{\text{TB}}\|, \|\nabla_\theta \hat{\mathcal{L}}_{\text{TB}}\|$) where the subsampling ratio is 1/9 under 4 MC samples.