Enhancing Ligand Validity and Affinity in Structure-Based Drug Design with Multi-Reward Optimization

Munsun Jo, Seungbeom Lee, Dongwoo Kim

GSAI, POSTECH
{chom5621, slee2020, dongwookim}@postech.ac.kr

Abstract

Deep learning-based Structure-based drug design (SBDD) is a crucial approach in pharmaceutical research, aiming to generate ligand molecules with high binding affinity and desirable properties for protein targets. While recent generative models have demonstrated competitive performance in optimizing binding affinity and drug-likeness, they often neglect the validity of the generated ligands. As a result, many models produce invalid ligands, such as those with protein-ligand clashes or unfavorable conformation energy, limiting their practical application. To address this, we propose a multi-reward Direct Preference Optimization (DPO) method to fine-tune models by jointly optimizing binding affinity and validity. Experimental results demonstrate that our method generates more realistic ligands than baseline models and achieves higher binding affinity than the pretrained model. This advancement highlights the potential of multi-reward optimization in enhancing the applicability of generative models for pharmaceutical discovery.

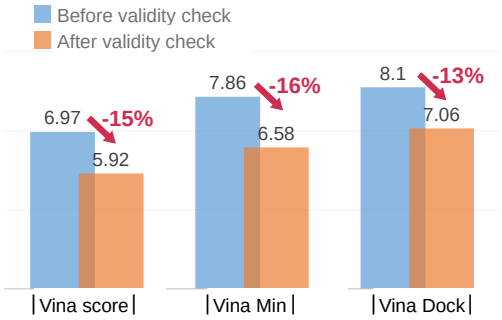

Figure 1: Absolute values of Vina score, Vian Min, and Vina Dock binding scores of the generated molecules with AliDiff (Gu et al. 2024) before and after the validity check. All three binding scores are significantly decreased by 15%, 16%, and 13%, with the valid molecules. We use absolute value for display purposes. The higher, the better.

Introduction

Designing ligand molecules with high binding affinity and favorable properties for protein structures, a process called structure-based drug design (SBDD), is a crucial aspect of pharmaceutical research. In recent years, deep learning approaches have framed SBDD as a conditional generative task, aiming to generate molecules tailored to specific proteins. These methods have demonstrated competitive performance, particularly in achieving strong protein-ligand binding affinities (Huang et al. 2024; Qu et al. 2024; Guan et al. 2023).

Conditional molecule generative models commonly focus on minimizing the atom-wise coordinate discrepancies between the reference and generated molecules. Their performance is usually evaluated based on the binding affinity and drug-likeness of the generated molecules, with their validity verified through simple valence rules. However, as shown in recent work (Buttenschoen, Morris, and Deane 2024), the concept of validity needs to be extended to the interaction between generated molecules and proteins. For example, although the generated molecule is valid alone, when the binding position is considered, there can be a crash between a molecule and protein, leading to an invalid binding position. In their work, the binding performances of existing models are significantly decreased when the affinity is measured on the molecules with valid positions. In our preliminary study, we find that one of the most advanced SBDD methods, named AliDiff (Gu et al. 2024), still suffers from the same issue. In Figure 1, we show the three binding scores of the AliDiff before and after the validity check. The binding scores decrease as we only consider the valid molecules.

Training a conditional generative model that satisfies high binding affinities with valid geometric positions is a challenging problem. At a molecular level, one may define a manifold of molecules to constrain the structure of molecules (Jing et al. 2022). When the protein and molecule interaction is considered, defining a proper manifold of the interaction seems impossible.

Inspired by the recent advances in reinforcement learning with human feedback, we tackle the problem by fine-tuning the pretrained generative models with feedback obtained from external software that can check the validity of the interaction. We propose a multi-reward direct preference optimization (DPO) that fine-tunes a model with multiple rewards to incorporate the feedback on validity while keeping the high binding affinity scores.

Experimental results show that our method generates more realistic ligands than baseline generative models. Furthermore, the fine-tuned model with our method can generate ligands with higher binding affinity than the pretrained model.

Related Work

Structure Based Drug Design Structure-based drug design (SBDD) aims to design ligand molecules with high binding affinity to a given protein structure and favorable molecular properties. Autoregressive-based models generate molecules by sequentially placing individual atoms or chemical groups within protein pockets (Luo et al. 2022; Peng et al. 2022). Based on the success in the vision domain, diffusion models have been introduced to generate ligand molecules by denoising the joint distribution of continuous atom positions and discrete atom types (Guan et al. 2023; Huang et al. 2024; Guan et al. 2024; Zhou et al. 2024). Recently, based on a new class of generative model called Bayesian flow networks (BFNs) (Graves et al. 2023), MolCRAFT shows improved binding affinity with more stable 3D structure (Qu et al. 2024).

Direct Preference Optimization Aligning pre-trained models with human preference shows remarkable efficacy in large language models (LLMs) (Rafailov et al. 2024; Ouyang et al. 2022) and text-to-image generative models (Wallace et al. 2023). The preference alignment has been extended to a scenario where one can find multiple rewards from different perspectives (Kim et al. 2024). In SBDD, ALiDiff (Gu et al. 2024) firstly applies the direct preference optimization in the SBDD task to improve binding scores. However, as shown in Figure 1, the validity of generated molecules with a single reward optimization is questionable. In this work, we tackle this problem through a multi-reward optimization. Our work is based on a recently proposed molecule generative model, MolCRAFT (Qu et al. 2024). As a preliminary, we introduce MolCRAFT with its backbone network, Bayesian flow networks (BFNs). We then propose a direct preference optimization (DPO) on BFN and multi-reward optimization with DPO.

Preliminaries

Problem formulation In SBDD, the generative model inputs target protein as a binding site defined as $\mathcal{P} = \{(\mathbf{x}_{\mathcal{P}}^{(i)}, \mathbf{v}_{\mathcal{P}}^{(i)})\}_{i=1}^{N_{\mathcal{P}}}$, where $\mathbf{x}_{\mathcal{P}}^{(i)} \in \mathbb{R}^3$ and $\mathbf{v}_{\mathcal{P}}^{(i)} \in \mathbb{R}^{D_{\mathcal{P}}}$ denote the i-th atom coordinates and type of $N_{\mathcal{P}}$ protein atoms, respectively. The molecules are defined as $\mathcal{M} = \{(\mathbf{x}_{\mathcal{M}}^{(i)}, \mathbf{v}_{\mathcal{M}}^{(i)})\}_{i=1}^{N_{\mathcal{M}}}$, where $\mathbf{x}_{\mathcal{M}}^{(i)} \in \mathbb{R}^3$ and $\mathbf{v}_{\mathcal{M}}^{(i)} \in \mathbb{R}^{D_{\mathcal{M}}}$ are coordinate and type vectors, respectively. In short, we denote each molecule as $\mathbf{m} = [\mathbf{x}, \mathbf{v}]$ where $[\cdot, \cdot]$ is the concatenation of $\mathbf{x} \in \mathbb{R}^{N_{\mathcal{M}} \times 3}$ and $\mathbf{v} \in \mathbb{R}^{N_{\mathcal{M}} \times K}$, given target protein as $\mathbf{p}$.

SBDD aims to discover a molecule $\mathbf{m}$ given protein and its binding site $\mathbf{p}$. A set of proteins and reference molecules is given as a training set. As a main performance metric, external software such as AutoDock Vina (Eberhardt et al. 2021) is used to measure the binding affinity between the protein and molecules. Along with the affinity score, various properties of the generated molecules can be used to measure their validity.

MolCRAFT and Bayesian Flow Networks MolCRAFT (Qu et al. 2024) uses Bayesian flow network (Graves et al. 2023) as a backbone for conditional molecule generative model given proteins. BFNs can be illustrated as an exchange of messages between a sender and a receiver distribution. Let $\mathbf{y} = [\mathbf{y}^{\mathbf{x}}, \mathbf{y}^{\mathbf{v}}]$ be a noise-injected version of original molecule $\mathbf{m}$. The sender p_S sends the noisy sample $\mathbf{y}$ to the receiver, similar to the forward diffusion process in a diffusion model, and the receiver p_R guesses the sender from known parameters, similar to the reverse denoising process.

Using different noise factors $\alpha = [\alpha^{\mathbf{x}}, \alpha^{\mathbf{v}}]$ for the coordinate and type, the sender injects continuous noise for K atom types by

$$p_S(\mathbf{y}|\mathbf{m}, \mathbf{p}) = \mathcal{N}\big(\mathbf{y}^{\mathbf{x}}|\mathbf{x}, (\alpha^{\mathbf{x}})^{-1}\mathbf{I}\big) \times \prod_{i=1}^{N_{\mathcal{M}}} \mathcal{N}\big(\mathbf{y}^{\mathbf{v}(i)}|\alpha^{\mathbf{v}}(K\mathbf{e}_{\mathbf{v}}^{(i)} - 1), \alpha^{\mathbf{v}}K\mathbf{I}\big),$$

where $\mathbf{e}_{\mathbf{v}}^{(i)} \in \mathbb{R}^K$ is a length-K one-hot vector for i-th atom.

On the other hand, the receiver estimates the sender using the known parameters $\theta = [\theta^{\mathbf{x}}, \theta^{\mathbf{v}}]$ of Bayesian prior

$$p_R(\mathbf{y}|\theta, \mathbf{p}) = \mathcal{N}\big(\mathbf{y}^{\mathbf{x}}|\Phi^{\mathbf{x}}(\theta^{\mathbf{x}}, \mathbf{p}), (\alpha^{\mathbf{x}})^{-1}\mathbf{I}\big) \times \prod_{i=1}^{N_{\mathcal{M}}} \left(\sum_{k=1}^{K} \Phi^{\mathbf{v}(i)}(k|\theta^{\mathbf{v}}, \mathbf{p}) \cdot \mathcal{N}\big(\mathbf{y}^{\mathbf{v}(i)}|\alpha^{\mathbf{v}}(K\mathbf{e}_k^{(i)} - 1), \alpha^{\mathbf{v}}K\mathbf{I}\big) \right),$$

where Φ is a neural network estimating original molecule by $[\Phi^{\mathbf{x}}(\theta^{\mathbf{x}}, \mathbf{p}), \Phi^{\mathbf{v}}(\theta^{\mathbf{v}}, \mathbf{p})]$ and $\mathbf{e}_k^{(i)} \in \mathbb{R}^K$ is a length-K one-hot vector where the k-th dimension is 1.

Sampling in the parameter space is effectively done by Bayesian flow distribution p_F:

$$p_F(\theta_t|\mathbf{m}, \mathbf{p}; t) = \mathcal{N}\big(\mu|\gamma_t \cdot \Phi^{\mathbf{x}}(\theta^{\mathbf{x}}, \mathbf{p}), \gamma_t(1 - \gamma_t)\mathbf{I}\big) \cdot \mathbb{E}_{p_S^{\mathbf{v}}(\mathbf{y}_{t-1}^{\mathbf{v}}|\Phi^{\mathbf{v}}(\theta^{\mathbf{v}}, \mathbf{p}), \mathbf{p})} \left[\delta\left(\theta^{\mathbf{v}} - \text{softmax}(\mathbf{y}_{t-1}^{\mathbf{v}})\right)\right].$$

For the noise factor α at time t and a hyperparameter of input variance σ_1, $\gamma_t = 1 - \frac{\alpha_t}{2ln\sigma_1}$ is added to the predicted coordinates $\Phi^{\mathbf{x}}(\theta^{\mathbf{x}}, \mathbf{p})$ and for the atom type we sample the most probable one by using Dirac delta δ. The interested reader is referred to Qu et al. (2024) for a detailed description.

Direct Preference Optimization on BFNs

To improve the binding performance of the molecules while satisfying the validity of the molecules, we fine-tune the model with feedback obtained from validity-checking software. Since the feedback is non-differentiable, we consider the external validity score as a reward model and fine-tune the model with a direct preference optimization method.

Based on the pre-computed rewards, we construct a preference molecular dataset as $\mathcal{D} = \{(\mathbf{p}, \mathbf{m}^w, \mathbf{m}^l)\}$ where $\mathbf{p}$ is the protein and $\mathbf{m}^w, \mathbf{m}^l$ are winning-losing pair of molecules based on reward scores. Similar to Diffusion-DPO (Wallace et al. 2023), we can formulate the BFN-DPO

objective in SBDD with the preference dataset given a reference model p_{ref} and a target model p_ϕ as

$$\mathcal{L}_{\text{BFN-DPO}} = - \mathbb{E}_{\substack{(\mathbf{p},\mathbf{m}_0^w,\mathbf{m}_0^l)\sim\mathcal{D},(\mathbf{m}_{1:T}^w,\mathbf{m}_{1:T}^l)\sim p_\phi,\theta_{0:T}^w\sim p_F(\cdot|\mathbf{m}_{0:T}^w),\\ \theta_{0:T}^l\sim p_F(\cdot|\mathbf{m}_{0:T}^l),\mathbf{y}_{0:T-1}^w\sim p_R(\cdot|\theta_{1:T}^w),\mathbf{y}_{0:T-1}^l\sim p_R(\cdot|\theta_{1:T}^l)}} \left[\log\sigma\left(\beta\log\frac{p_\phi(\mathbf{y}_{0:T-1}^w|\theta_{1:T}^w)}{p_{\text{ref}}(\mathbf{y}_{0:T-1}^w|\theta_{1:T}^w)} - \beta\log\frac{p_\phi(\mathbf{y}_{0:T-1}^l|\theta_{1:T}^l)}{p_{\text{ref}}(\mathbf{y}_{0:T-1}^l|\theta_{1:T}^l)}\right)\right],$$

where β is a hyperparameter balancing between p_{ref} and p_ϕ. The model predicts the winning sample $\mathbf{y}_{t-1}^w$ over the losing sample $\mathbf{y}_{t-1}^l$, with the parameters θ_t^w and θ_t^l each sampled from Bayesian flow distribution p_F at time t. We omit the conditioning on protein $\mathbf{p}$ for brevity.

Expanding the equation further, we can show that the objective can be represented as a summation of the atom coordinates and atom type preference losses as follows:

$$\mathcal{L}_{\text{BFN-DPO}} = -\mathbb{E}_{(\mathbf{p},\mathbf{m}_0^w,\mathbf{m}_0^l)\sim\mathcal{D},t\sim[0,T],\theta_t^w\sim p_F(\cdot|\mathbf{m}_t^w),\theta_t^l\sim p_F(\cdot|\mathbf{m}_t^l)} \left[\mathcal{L}_{t-1}^{\mathbf{x}} + \mathcal{L}_{t-1}^{\mathbf{v}}\right].$$

The coordinate loss is designed to directly train the model output $\Phi^{\mathbf{x}}(\theta_t^{\mathbf{x}})$ to approximate the clean coordinate values $\mathbf{x}_0$. The noise factor α at time t is multiplied during training.

$$\mathcal{L}_{t-1}^{\mathbf{x}} = -\mathbb{E}_{(\mathbf{p},\mathbf{x}_0^w,\mathbf{x}_0^l)\sim\mathcal{D},t\sim[0,T],\theta_t^{\mathbf{x},w}\sim p_F(\cdot|\mathbf{x}_t^w),\theta_t^{\mathbf{x},l}\sim p_F(\cdot|\mathbf{x}_t^l)} \left[\log\sigma\left(-\frac{\alpha_t^{\mathbf{x}}\beta T}{2}\left(\|\mathbf{x}_0^w - \Phi_\phi^{\mathbf{x}}(\theta_t^{\mathbf{x},w})\|^2 - \|\mathbf{x}_0^w - \Phi_{\text{ref}}^{\mathbf{x}}(\theta_t^{\mathbf{x},w})\|^2 - \|\mathbf{x}_0^l - \Phi_\phi^{\mathbf{x}}(\theta_t^{\mathbf{x},l})\|^2 + \|\mathbf{x}_0^l - \Phi_{\text{ref}}^{\mathbf{x}}(\theta_t^{\mathbf{x},l})\|^2\right)\right)\right]$$

Similarly, the type loss is designed to directly train the model output $\Phi^{\mathbf{v}}(\theta_t^{\mathbf{v}})$ to approximate the clean atom types $e_{\mathbf{v}_0}$, which is the $N_{\mathcal{M}} \times K$ matrix with each row as K-dimensional one-hot vector.

$$\mathcal{L}_{t-1}^{\mathbf{v}} = -\mathbb{E}_{(\mathbf{p},\mathbf{v}_0^w,\mathbf{v}_0^l)\sim\mathcal{D},t\sim[0,T],\theta_t^{\mathbf{v},w}\sim p_F(\cdot|\mathbf{v}_t^w),\theta_t^{\mathbf{v},l}\sim p_F(\cdot|\mathbf{v}_t^l)} \left[\log\sigma\left(-\beta T\left(\left(\ln p_S^{\mathbf{v}}(.|e_{\mathbf{v}_0^w}) - \ln p_R^{\mathbf{v}}(.|\Phi_\phi^{\mathbf{v}}(\theta_t^{\mathbf{v},w}))\right)\right.\right.\right.$$
$$\left.\left.\left. - \left(\ln p_S^{\mathbf{v}}(.|e_{\mathbf{v}_0^w}) - \ln p_R^{\mathbf{v}}(.|\Phi_{\text{ref}}^{\mathbf{v}}(\theta_t^{\mathbf{v},w}))\right) - \left(\ln p_S^{\mathbf{v}}(.|e_{\mathbf{v}_0^l}) - \ln p_R^{\mathbf{v}}(.|\Phi_\phi^{\mathbf{v}}(\theta_t^{\mathbf{v},l}))\right)\right.\right.\right.$$
$$\left.\left.\left. + \left(\ln p_S^{\mathbf{v}}(.|e_{\mathbf{v}_0^l}) - \ln p_R^{v}(.|\Phi_{\text{ref}}^{\mathbf{v}}(\theta_t^{\mathbf{v},l}))\right)\right)\right)\right]$$

Multi-Reward Direct Preference Optimization

Although one can directly fine-tune the model with a BFN-DPO for the validity reward, fine-tuning a model with the validity reward may have a negative influence on the performance of binding affinity. To mitigate the problem, we use a multi-reward DPO method to fine-tune the model with both validity and affinity rewards.

We use the multi-reward DPO approach proposed in Kim et al. (2024). We here include the details of the multi-reward DPO for completeness. Without loss of generality, assume that we have m different rewards for a given pair of a molecule and protein. To perform a multi-reward DPO, an average of the rewards is used as the final reward of a given molecule. However, since the scale of each reward is different, a softmax function can be applied to normalize the value of each reward. Specifically, let $(\mathbf{m})_{i=1}^B$ be a B-batch of generated molecules given protein and $r_i^{(j)}$ be j-th reward of the i-th molecule. The reward is normalized via the softmax function as

$$\hat{r}_i^{(j)} = \frac{\exp\left(r_i^{(j)}/\tau_j\right)}{\sum_{i=1}^B \exp\left(r_i^{(j)}/\tau_j\right)},$$

where τ_j is a temperature parameter.

To further penalize the cases where the rewards disagree significantly, Kim et al. (2024) proposes an uncertainty-regularized ensemble. With the uncertainty regularized ensemble, the final reward for molecule i becomes

$$\bar{r}_i = \mu_{\hat{r}_i} - \gamma\frac{1}{m}\sum_{j=1}^m(\hat{r}_i^{(j)} - \mu_{\hat{r}_i})^2,$$

where $\mu_{\hat{r}_i} = \frac{1}{m}\sum_{j=1}^m \hat{r}_i^{(j)}$ and γ is a hyperparameter controlling the penalty.

We finally use a variant of DPO, denoted as E^{2}PO (Gu et al. 2024), developed to prevent overfitting to the winning data samples. $\mathcal{L}_{\text{BFN-DPO}}$ alone has a risk of over-optimization when it keeps greedily optimizing to winning samples. Instead, the following objective leads to more stable learning by adding the second term to prevent overfitting.

$$\mathcal{L}_{\text{BFN-E}^2\text{PO}} = -\mathbb{E}[\sigma(\bar{r}_w - \bar{r}_l)\mathcal{L}_{\text{BFN-DPO}} + (1 - \sigma(\bar{r}_w - \bar{r}_l))(2 - \mathcal{L}_{\text{BFN-DPO}})],$$

where w and l index the winning and losing cases.

Experiment

Experimental setting

We compare our multi-reward DPO method to the pretrained model without DPO and the single-reward DPO that optimizes solely for binding affinity or validity. We use Vina Dock and strain energy as rewards for binding affinity and validity, respectively. We fine-tune MolCRAFT (Qu et al. 2024) trained on PDBbind dataset (Liu et al. 2017) to apply DPO. We choose MolCRAFT as a backbone due to its superior performance in generating valid molecules (cf., Table 2). For each protein, win-lose pairs are selected from molecules generated by the pretrained model, and DPO is applied for fine-tuning. To ensure only valid molecules are considered, valid molecules are filtered from 10,000 generated samples using the method proposed in PoseBuster (Buttenschoen, Morris, and Deane 2024), and we report performance on the *valid* molecules.

Evaluation metrics

We evaluate the generated molecules by our method on binding affinity, conformation validity, and molecular properties. Following the setup used in Ragoza, Masuda, and Koes (2022), we use Vina Score, Vina Min, and Vina Dock measured with AutoDock Vina (Eberhardt et al. 2021) for binding affinity. The Vina Score measures the binding affinity of

Reward	Binding Affinity						Conformation Stability				Drug-like Properties	
	Vina Score (↓)		Vina Min (↓)		Vina Dock (↓)		SE (↓)			Clash (↓)	SA (↑)	QED (↑)
	Avg.	Med.	Avg.	Med.	Avg.	Med.	25%	50%	75%	Avg.	Avg.	Avg.
No	-6.23	-6.87	-7.02	-7.07	-7.60	-7.80	3.36	9.20	19.10	7.52	0.67	0.50
Vina Dock	-6.27	-6.74	-7.14	-7.10	-7.77	-7.86	2.80	8.11	17.95	8.67	0.65	0.47
Strain Energy	-	-	-	-	-	-	1.68	4.57	**9.37**	7.82	0.67	0.47
Vina Dock + Strain Energy	**-6.98**	**-7.23**	**-7.88**	**-7.80**	**-8.58**	**-8.68**	**1.24**	**4.09**	9.67	**7.46**	**0.72**	**0.57**

Table 1: Performance comparison across different reward methods. We mark the best, and the second-best performances in **bold** and underline, respectively. (↓)/(↑) indicate whether a smaller/ larger number is better, respectively. (-) refers to the NaN value of Binding Affinity, which AutoDock Vina fails to calculate due to the ligand and protein being too far apart.

Method	Auto-regressive		Diffusion					BFN	
	AR (Luo et al. 2022)	Pocket2Mol (Peng et al. 2022)	TargetDiff (Guan et al. 2023)	DecompDiff (Guan et al. 2024)	DecompOpt (Zhou et al. 2024)	IPDiff (Huang et al. 2024)	AliDiff (Gu et al. 2024)	MolCRAFT (Qu et al. 2024)	Ours
Validity	59.1%	72.4%	50.6%	71.9%	48.8%	25.0%	18.6%	74.9%	**82.5%**

Table 2: The proportion of valid molecules for eight deep learning based SBDD models. We categorize the models into three methods: Auto-regressive, diffusion, and BFN.

# rewards	Binding Affinity		Conformation Stability	Reward Average		
	Vina Min Med.(↓)	Vina Dock Med.(↓)	SE Med.(↓)	Winning-sample	Losing-samples	Difference
Dock + Energy rewards	-7.80	-8.68	4.09	0.39	0.05	0.33
All rewards	-7.25	-8.63	4.40	0.19	0.07	0.12

Table 3: Comparison between two rewards and seven rewards. A model fine-tuned with two rewards shows better binding affinity scores than the model with all seven rewards.

the generated molecule as is. Vina Min measures the affinity after optimizing the molecular structure without changing the docking position. Vina Dock measures the binding affinity after re-docking the molecule; hence, reconsider the molecule's position and orientation. Low scores indicate stronger protein-ligand binding affinity. We report average and median scores for all Vina metrics. We report strain energy (SE) measuring molecule conformation energy and protein-ligand clash measuring the number of potential overlaps between ligand atoms and protein (Harris et al. 2023). For Drug-like properties, we use synthetic accessibility (SA) (Ertl and Schuffenhauer 2009) and quantitative estimation of drug-likeness (QED) (Bickerton et al. 2012).

Results

Quality of ligands Table 1 shows our multi-reward method outperforms the original model. When compared with the fine-tuning model with a single reward, the multi-reward approach outperforms the single-reward models except for 75% of SE. The result presents that our multi-reward optimization can generate more realistic drug-like ligands with high binding affinity. Although the single reward model with Vina Dock can still improve the binding affinity, it fails to reduce the clashes between protein and ligand. On the other hand, the multi-reward approach can satisfy both objectives successfully.

Validity of ligands Note that the result in Table 1 is measured over the valid molecules. We also report the proportion of valid molecules in Table 2. Our method generates significantly more valid ligands compared to other generative models. AliDiff, a fine-tuned IPDiff (Huang et al. 2024) with Vina Dock as a single reward, is known to achieve the best binding performance so far. However, the model fails to generate valid molecules in many cases.

Many-rewards optimization Our multi-reward approach allows the ensemble of more than two rewards, making it possible to fine-tune models with many evaluation metrics as rewards. To test the performance of the many-rewards model, we fine-tune the original model with all seven rewards reported in Table 1. We report the binding affinity of two and seven rewards models in Table 3. The result shows that a fine-tuned model with seven rewards performs worse than the model with two rewards. To investigate this, we analyze the reward values of winning and losing samples. Table 3 shows that the average difference between winning and losing cases with seven rewards is much closer than that of the two rewards model. We conjecture that the small difference between winning and losing cases with seven rewards introduces conflicts in the fine-tuning process, making it difficult for the model to optimize all properties.

Conclusion

In this work, we propose a multi-reward direct preference optimization method that fine-tunes generative models by optimizing both binding affinity and ligand validity simultaneously. Our experiments show that the multi-reward DPO outperforms baseline and single-reward models, achieving higher binding affinity and generating more valid molecules.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337955 and RS-2023-00217286).

References

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

Buttenschoen, M.; Morris, G. M.; and Deane, C. M. 2024. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9): 3130–3139.

Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; and Forli, S. 2021. AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8): 3891–3898.

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

Graves, A.; Srivastava, R. K.; Atkinson, T.; and Gomez, F. 2023. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*.

Gu, S.; Xu, M.; Powers, A.; Nie, W.; Geffner, T.; Kreis, K.; Leskovec, J.; Vahdat, A.; and Ermon, S. 2024. Aligning Target-Aware Molecule Diffusion Models with Exact Energy Optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Guan, J.; Qian, W. W.; Peng, X.; Su, Y.; Peng, J.; and Ma, J. 2023. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *International Conference on Learning Representations*.

Guan, J.; Zhou, X.; Yang, Y.; Bao, Y.; Peng, J.; Ma, J.; Liu, Q.; Wang, L.; and Gu, Q. 2024. DecompDiff: diffusion models with decomposed priors for structure-based drug design. *arXiv preprint arXiv:2403.07902*.

Harris, C.; Didi, K.; Jamasb, A. R.; Joshi, C. K.; Mathis, S. V.; Lio, P.; and Blundell, T. 2023. Benchmarking Generated Poses: How Rational is Structure-based Drug Design with Generative Models? *arXiv preprint arXiv:2308.07413*.

Huang, Z.; Yang, L.; Zhou, X.; Zhang, Z.; Zhang, W.; Zheng, X.; Chen, J.; Wang, Y.; Bin, C.; and Yang, W. 2024. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *The Twelfth International Conference on Learning Representations*.

Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; and Jaakkola, T. 2022. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35: 24240–24253.

Kim, K.; Jeong, J.; An, M.; Ghavamzadeh, M.; Dvijotham, K.; Shin, J.; and Lee, K. 2024. Confidence-aware Reward Optimization for Fine-tuning Text-to-Image Models. *arXiv preprint arXiv:2404.01863*.

Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; and Wang, R. 2017. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2): 302–309.

Luo, S.; Guan, J.; Ma, J.; and Peng, J. 2022. A 3D Generative Model for Structure-Based Drug Design. arXiv:2203.10446.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peng, X.; Luo, S.; Guan, J.; Xie, Q.; Peng, J.; and Ma, J. 2022. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 17644–17655. PMLR.

Qu, Y.; Qiu, K.; Song, Y.; Gong, J.; Han, J.; Zheng, M.; Zhou, H.; and Ma, W.-Y. 2024. MolCRAFT: Structure-Based Drug Design in Continuous Parameter Space. *arXiv preprint arXiv:2404.12141*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization. arXiv:2311.12908.

Zhou, X.; Cheng, X.; Yang, Y.; Bao, Y.; Wang, L.; and Gu, Q. 2024. DecompOpt: Controllable and Decomposed Diffusion Models for Structure-based Molecular Optimization. *arXiv preprint arXiv:2403.13829*.