

# De Novo Molecular Design Enabled by Direct Preference Optimization and Curriculum Learning

Junyu Hou✉

School of Computer Science, Nanjing University  
221220151@smail.nju.edu.cn

**Abstract.** De novo molecular design has extensive applications in drug discovery and materials science. The vast chemical space renders direct molecular searches computationally prohibitive, while traditional experimental screening is both time- and labor-intensive. Efficient molecular generation and screening methods are therefore essential for accelerating drug discovery and reducing costs. Although reinforcement learning (RL) has been applied to optimize molecular properties via reward mechanisms, its practical utility is limited by issues in training efficiency, convergence, and stability. To address these challenges, we adopt Direct Preference Optimization (DPO) from NLP, which uses molecular score-based sample pairs to maximize the likelihood difference between high- and low-quality molecules, effectively guiding the model toward better compounds. Moreover, integrating curriculum learning further boosts training efficiency and accelerates convergence. A systematic evaluation of the proposed method on the GuacaMol Benchmark yielded excellent scores. For instance, the method achieved a score of 0.883 on the Perindopril MPO task, representing a 6% improvement over competing models. And subsequent target protein binding experiments confirmed its practical efficacy. These results demonstrate the strong potential of DPO for molecular design tasks and highlight its effectiveness as a robust and efficient solution for data-driven drug discovery. The code is available at <https://github.com/SiliconMangrove/MolDPO>.

**Keywords:** De Novo Molecular Design, DPO, Curriculum Learning.

## 1 Introduction

De novo molecular design is one of the core tasks in fields such as catalyst design, energy materials design, and pharmaceutical research, aiming to generate novel molecules from scratch that satisfy specified physicochemical properties and biological activity requirements [21]. This process plays a pivotal role in new drug discovery, materials science, and synthetic chemistry. Traditional methods for candidate molecule screening and optimization typically rely on extensive experimental synthesis and biological assays, which are both time- and labor-intensive and require substantial financial investment [7]. Moreover, the chemical space is

astronomically vast—estimated to contain more than  $10^{60}$  potential molecules [4]—making exhaustive exploration by manual means virtually impossible. Consequently, computer-aided drug design (CADD) has emerged as a prominent research focus, leveraging mathematical models, statistical techniques, and advanced computational technologies to efficiently search and optimize within this enormous chemical space [28, 8, 31, 33].

In recent years, molecular conditional generation has played a critical role in drug development and materials design [22], and reinforcement learning (RL) has gradually been introduced into the field of molecular design [24, 25, 16]. By employing molecular scoring functions as reward signals, RL enables generative models to continuously adjust and improve their generation strategies toward predefined objectives—such as enhancing biological activity, optimizing physicochemical properties, and improving synthetic accessibility—thereby demonstrating enormous potential [30, 34].

However, RL-based methods still face several challenges: **(1) Convergence Challenges and Training Instability:** The high-dimensional and non-convex nature of molecular generation makes RL models prone to slow convergence and local optima. For instance, REINVENT [3] exhibits volatile policy updates and noisy rewards, requiring extensive training before reliably generating molecules that satisfy multiple objectives.

**(2) Exploration Inefficiency and Limited Coverage:** The vast chemical space limits the effectiveness of traditional RL approaches, which often get trapped in narrow structural regions. DrugEx [18], for example, generates molecules meeting specific activity criteria but lacks sufficient scaffold diversity.

**(3) Multi-Objective Optimization and Reward Design Challenges:** Designing effective reward functions for molecular optimization is complex and often requires empirical tuning. For example, to simultaneously maximize logP, TPSA, and structural similarity, some studies have employed multi-layered, non-linear composite reward models, increasing both implementation complexity and limiting generalizability across different tasks [15, 29, 27].

To address these challenges, we draw inspiration from two established methodologies in machine learning. Direct Preference Optimization (DPO), originally developed in NLP, has shown strong optimization capabilities in reinforcement learning tasks by leveraging paired samples to optimize likelihood differences, eliminating the need for explicit reward modeling [26, 6]. Meanwhile, Curriculum Learning, which gradually increases task complexity, has been adopted to enhance molecular generation [2, 11]. By starting with simpler tasks and progressively optimizing bioactivity, physicochemical properties, and synthetic feasibility, this approach improves model learning [23, 10]. We contend that integrating DPO with curriculum learning can both accelerate model convergence and substantially improve the overall performance of the generated molecules.

In this study, we first employ traditional autoregressive training to train a prior model, which is then assigned to four agent models. These agent models are responsible for sampling and constructing paired samples, which are subsequently trained using DPO and curriculum learning to optimize the molecular

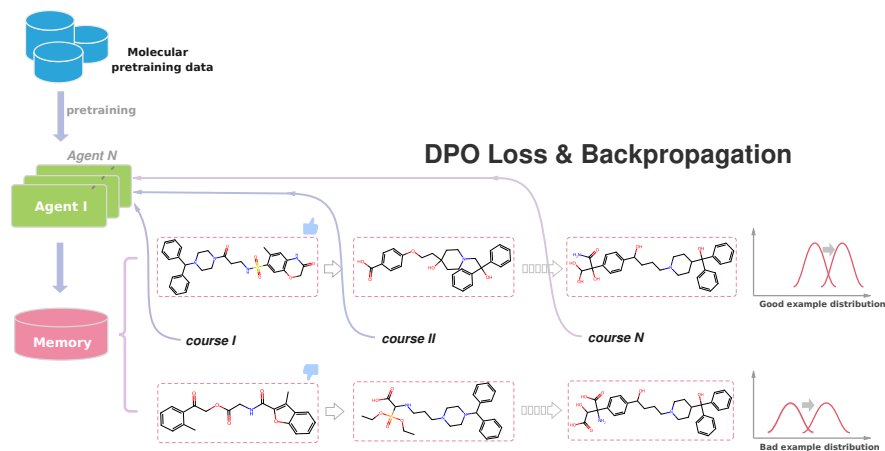


Fig. 1: **Structure of the DPO+Curriculum Learning model.** The model is initially pre-trained, followed by optimization using Direct Preference Optimization. As curriculum learning progresses, the molecular scores of the collected compounds steadily increase while the distinction between superior and inferior molecules gradually narrows. Ultimately, the process yields molecules that meet the predefined quality criteria.

generation process, ultimately yielding molecules that satisfy the desired properties. The overall architecture of our proposed method is illustrated in Figure 1. We evaluate our method on the Guacamol benchmark [5] and target protein binding experiments, where experimental results demonstrate that our approach achieves superior performance across multiple evaluation metrics, validating its effectiveness in molecular conditional generation tasks.

The primary objective of this study is to investigate the application of DPO combined with curriculum learning in molecular conditional generation, aiming to improve the molecular generation process through a more efficient and stable approach. By integrating these two methods, we aim to improve molecular discovery and optimization efficiency while providing strong technological support for drug development. Our main contributions are threefold:

- We propose a novel de novo molecular design framework that combines Direct Preference Optimization (DPO) with curriculum learning.
- Our method has achieved high scores on the GuacaMol benchmark and demonstrated outstanding performance in target protein docking experiments.
- The proposed framework exhibits strong potential for scalability in terms of multi-objective optimization, training stability, and computational efficiency.

## 2 Related Work

In recent years, computationally driven de novo molecular design has witnessed rapid advancements, primarily evolving along three key directions: **(1) continuous optimization of reinforcement learning frameworks**, **(2) cross-domain adaptation of large language models (LLMs)**, and **(3) efficiency improvements in preference learning paradigms**. These innovations have enhanced chemical space exploration and multi-objective optimization, laying a stronger foundation for drug discovery.

### 2.1 Reinforcement Learning-Based Molecular Generation

Reinforcement learning (RL)-based generative models have established a systematic paradigm for molecular design[19, 34]. Among them, REINVENT[20] integrates recurrent neural networks (RNNs) with policy gradient algorithms to enable targeted chemical space exploration. Another notable approach[24] leverages prior knowledge to constrain the reward function, optimizing molecular properties while maintaining synthetic feasibility. However, traditional RL approaches face challenges when handling high-dimensional chemical spaces, including inefficient policy updates and susceptibility to local optima.

### 2.2 Curriculum Learning Strategies

To enhance training efficiency in complex tasks, researchers have introduced curriculum learning frameworks into de novo molecular design. Guo et al. [10] proposed a strategy that gradually increases task difficulty: the model initially focuses on generating simpler chemical structures, thereby establishing a solid foundation, and then progressively tackles more challenging optimization tasks. This staged approach not only accelerates convergence but also improves the diversity and quality of the generated molecules, demonstrating the effectiveness of curriculum learning in refining generative models for molecular design.

### 2.3 Large Language Models for Molecular Generation

Large Language Models (LLMs) have recently been applied to molecular design, offering novel strategies for molecule generation. Liu et al. [17] explored the adaptation of ChatGPT for molecular tasks, demonstrating its ability to capture chemical patterns and generate valid molecular representations through language modeling. Similarly, Hu et al. [12] introduced MolRL-MGPT, which integrates a GPT-based generative strategy with reinforcement learning to enhance molecular diversity and optimize target-directed properties. These studies highlight the promising potential of LLMs to provide scalable and effective approaches for molecular design.

## 2.4 Preference Optimization in Molecular Design

Preference learning techniques provide an efficient pathway for strategy optimization in molecular generation. Rafailov et al. proposed direct preference optimization (DPO)[26], which employs implicit reward modeling to bypass the complexity of explicit reward function design in traditional RL. This paradigm has recently been successfully adapted to molecular design[9, 6]: Widatalla et al. utilized experimental data to construct preference pairs, enabling DPO to directly optimize protein stability[32]. Experimental results show that ProteinDPO performs exceptionally well in protein stability prediction and demonstrates strong generalization capabilities for large proteins and multi-chain complexes. This suggests that it has effectively learned transferable insights from its biophysical alignment data.

## 3 Methodology

Our molecular generation framework is built upon three core technical components integrated through a structured training pipeline: (1) Pretraining establishes chemical validity by learning SMILES syntax from large-scale datasets; (2) Direct Preference Optimization (DPO) replaces reward modeling with contrastive learning to align generation with target objectives; (3) Curriculum Learning introduces progressive difficulty levels for gradual chemical space exploration. To synergistically combine these components, we design a two-stage training procedure: pretraining initializes molecular priors, followed by DPO fine-tuning guided by curriculum-constructed preference pairs. The following subsections detail each component.

### 3.1 Pretrain on large molecular dataset

In this study, we adopt the same model architecture as MolRL-MGPT by building a multi-agent GPT model with 8 layers and 8 attention heads [12]. Two distinct prior models were pre-trained on different datasets: one on the GuacaMol dataset (a subset of ChEMBL) for benchmark evaluation and another on the ZINC dataset (containing approximately 100 M molecules) for general-purpose molecular generation tasks [13]. The primary objective during pretraining is to enable the model to deeply learn the syntax rules of SMILES representations, thereby allowing it to generate valid SMILES structures one character at a time and effectively capture the distribution of the chemical space. This pretraining strategy not only demonstrates the model’s capability in efficiently generating chemically valid molecules but also lays a solid foundation for subsequent task-specific optimization and performance enhancement.

To rigorously train our prior models on SMILES representations, we employ an autoregressive framework that decomposes each sequence into a series of incremental prediction tasks. Let  $S = (c_1, c_2, \dots, c_L)$  denote a SMILES sequence of length  $L$ , where each  $c_i$  represents a character from the SMILES vocabulary

$\mathcal{V}$ . In our autoregressive training approach, we generate training pairs  $(x_i, y_i)$  for  $i = 1, 2, \dots, L - 1$ , where

$$x_i = (c_1, c_2, \dots, c_i) \quad \text{and} \quad y_i = (c_1, c_2, \dots, c_i, c_{i+1}). \quad (1)$$

The model parameterized by  $\theta$  learns a conditional probability distribution  $P_\theta(c_{i+1} \mid c_1, \dots, c_i)$  such that the joint probability of the sequence can be expressed as:

$$P_\theta(S) = \prod_{i=1}^L P_\theta(c_i \mid c_1, \dots, c_{i-1}), \quad (2)$$

with the convention that  $P_\theta(c_1 \mid \cdot) = P_\theta(c_1)$ .

The training objective is to minimize the cross-entropy loss over the entire training set, which for a single sequence is given by:

$$\mathcal{L}(\theta) = - \sum_{i=1}^{L-1} \log P_\theta(c_{i+1} \mid c_1, \dots, c_i). \quad (3)$$

This objective encourages the model to assign high probabilities to the correct next character at each step. Parameter updates are performed using gradient descent:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta), \quad (4)$$

where  $\eta$  is the learning rate. This autoregressive framework enables the model to learn the syntax rules of SMILES representations, thereby generating valid molecular structures character by character.

The model trained on the Guacamol dataset for approximately 3 hours, achieving an 97% validity rate for the generated molecular structures, demonstrating that it had effectively learned the SMILES generation rules. The model trained on the ZINC dataset for approximately 70 hours, achieving a 99.6% validity rate for the generated molecules, further validating its generalization capability. All pretraining experiments were conducted on a single A100 GPU.

### 3.2 DPO for Molecular Optimization

Building upon the pretrained molecular generation capability, we introduce Direct Preference Optimization (DPO) to align molecular generation with chemical preferences. DPO is a contrastive learning approach that optimizes the generation policy without explicitly modeling a reward function. Instead of using reinforcement learning with human feedback (RLHF) methods that first train a reward model and then optimize the policy using algorithms like PPO, DPO directly optimizes the policy by enforcing preference constraints.

The training data consists of triplets  $(x, y_w, y_l)$ , where:  $x$  represents the input.  $y_w$  is the preferred (or "winning") response.  $y_l$  is the less preferred (or "losing") response.

DPO is built upon the idea that an optimal policy  $\pi^*$  should satisfy the following preference ratio constraint:

$$\frac{\pi^*(y_w|x)}{\pi^*(y_l|x)} = \exp(r(y_w|x) - r(y_l|x))$$

where  $r(y|x)$  is an implicit reward function that ranks different responses. Instead of explicitly learning this reward function, DPO directly optimizes the policy ratio by defining the following log preference probability:

$$\log \sigma(\beta \cdot (\log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x)))$$

where  $\sigma(z) = \frac{1}{1+e^{-z}}$  is the sigmoid function, and  $\beta$  is a temperature hyper-parameter controlling sensitivity to preference differences.

The final DPO objective function is:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(\beta \cdot (\log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x)))]$$

In our molecular generation task, there is no explicit input  $x$ . Furthermore, whereas other DPO tasks often require human annotation of preference samples, which is costly, our task does not rely on human annotations to determine response quality. Instead, we leverage a chemical computation library to evaluate the quality of generated molecules, thereby streamlining the preference learning process.

### 3.3 Curriculum Learning for Structured Molecular Optimization

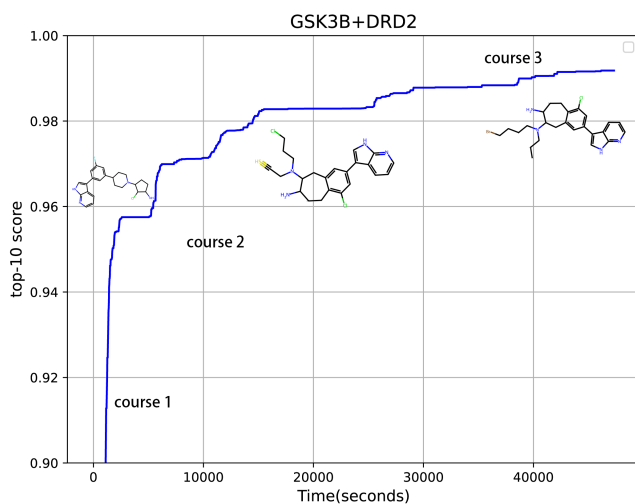
To address the challenge of learning complex chemical spaces, we integrate DPO with curriculum learning. Curriculum learning is a machine learning strategy inspired by the human learning process, where the model begins with simple tasks and gradually progresses to more complex ones. By organizing training samples in order of increasing difficulty, the model builds a solid foundation with easier examples, leading to more efficient learning, better generalization, and ultimately enhanced performance on challenging tasks.

Aligned with this progressive approach, our pair construction process incrementally increases the difficulty of the learning task. Initially, the score gap between the superior and inferior samples is large, making it straightforward for the model to distinguish high-quality molecules. As training advances, this gap is gradually reduced, requiring the model to discern more subtle differences. This strategy reinforces the curriculum learning paradigm and refines the model’s fine-grained discrimination of molecular quality, ultimately enhancing the validity of the generated compounds.

Furthermore, we adopt a multi-stage learning mechanism to enhance model performance. Specifically, during training, high-quality molecules collected by the model are stored in memory. Subsequently, all agents are reinitialized to the pre-trained model and continue training by constructing new sample pairs from the high-scoring molecules in memory. The primary objective of this strategy is to mitigate potential biases introduced during early exploration, preventing the model from converging to suboptimal solutions. By leveraging previously identified high-quality molecules, the model can effectively restart its learning process

in a more optimized direction, ultimately improving the quality of generated molecules.

As illustrated in the figure 2, our training process is divided into three stages. In the first stage, the model learns the fundamental requirements of the task and rapidly identifies high-scoring molecules from the vast chemical space. In the second stage, the model fine-tunes molecular scaffolds to further refine its understanding. Finally, in the third stage, the model modifies functional groups based on the optimal molecular scaffolds stored in memory, further optimizing molecular structures. This multi-stage approach significantly enhances molecular design efficiency, leading to a remarkable score of 0.993 on the GSK3B+DRD2 task.



**Fig. 2: In the GSK3B+DRD2 docking experiment, the model achieved good performance through curriculum learning.** In Course 1, the model learns the fundamental requirements of the task. In Course 2, it fine-tunes the molecular scaffold. In Course 3, it adjusts functional groups to optimize molecular structures.



### 3.4 Training Procedure

---

**Algorithm 1** Direct Preference Optimization (DPO) Training
 

---

```

1: procedure DPO_TRAIN( $F_{\text{score}}, k$ )
  Initialize:
2:   Load pretrained prior  $p_{\text{ref}} \leftarrow \text{GPT}(\theta_{\text{prior}})$ 
3:   Initialize agents  $\{\pi_i\}_{i=1}^N$  with  $\theta_i \sim \mathcal{N}(0, 0.02)$ 
4:   for  $t \leftarrow 1$  to  $T$  do
5:     AgentLoop
6:       for each agent  $\pi_i$  do
7:         Sampling
8:          $\mathcal{D}_i \leftarrow \text{SampleSMILES}(\pi_i, m_{\text{batch}})$ 
9:          $\mathbf{s} \leftarrow F_{\text{score}}(\mathcal{D}_i)$ 
10:         $\mathcal{M} \leftarrow \text{UpdateMemory}(\mathcal{D}_i, \mathbf{s})$ 
11:       EndSampling
12:       POSITIVE SELECTION:
           $\mathbf{x}^w \sim p_{\mathcal{M}}(x) \propto \exp(s/\tau) // \text{top-weighted historical samples}$ 
13:       NEGATIVE SELECTION:
           $\mathbf{x}^l \sim \text{Uniform}(\mathcal{D}_t) // \text{current batch negatives}$ 
14:       Compute log-ratios for each sample:
          
$$\log r_{\theta}(x) = \log \pi_{\theta}(x) - \log \pi_{\text{ref}}(x)$$

15:       Optimize loss:
          
$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[ \log \sigma \left( \beta(r_{\theta}(x^w) - r_{\theta}(x^l)) \right) \right]$$

16:       Gradient step:  $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\text{DPO}}$ 
17:     end for
18:   EndAgentLoop
19:   Log max  $\mathcal{M}.\mathbf{s}$ , top $_k$ -mean( $\mathcal{M}.\mathbf{s}$ )
20: end for
21: return Top $_k(\mathcal{M}.\mathbf{x}, \mathcal{M}.\mathbf{s})$ 
22: end procedure

```

---

Integrating DPO with curriculum learning, we design a two-stage training protocol, as illustrated in Algorithm 1. The process begins with pre-training to obtain the prior model, followed by reinforcement fine-tuning using DPO and curriculum learning.

In the pre-training phase, the model learns to generate valid SMILES strings and capture the chemical space distribution, forming a prior model.

During reinforcement fine-tuning, the agents—initialized from the prior generates molecules, which are evaluated by a task-specific scoring function. Prefer-

ence pairs are then constructed by selecting the top-k highest-scoring molecules (k varies across agents) as “preferred samples” and randomly sampling lower-quality ones as “dispreferred samples”. These pairs are used to optimize the policy via DPO, progressively aligning the model’s generation strategy with the target objectives.

As training progresses, the scores of the generated molecules steadily improve while the gap between preferred and dispreferred samples narrows, indicating an increase in training difficulty. Initially, the model makes broad scaffold-level adjustments to identify promising frameworks; as scaffolds stabilize, it shifts to fine-tuning functional groups to further optimize molecular properties.

## 4 Experiments

To validate the effectiveness of our model, we designed and conducted a series of experiments, including the Guacamol benchmark evaluation, target protein binding experiments, and impact analysis. The experimental results demonstrate that our model is not only capable of handling classical molecular design tasks but also performs exceptionally well in tasks that are more closely aligned with real-world drug discovery. Furthermore, the impact analysis, which examines model performance under different parameter settings, helping us identify the optimal parameter settings.

### 4.1 GuacaMol benchmark

**Guacamol Introduction** Guacamol Benchmark, proposed by BenevolentAI in 2019 [5], is a standardized framework for evaluating molecular generation models in terms of diversity, synthetic feasibility, and goal-directed optimization. It comprises 20 tasks covering key challenges in molecular design.

These tasks can be broadly categorized into rediscovery and similarity-based optimization, isomer generation, and molecular property balancing. Additionally, multi-parameter optimization (MPO) tasks focus on improving physicochemical properties of known drugs, while SMARTS-constrained tasks enforce structural constraints. Lastly, scaffold hopping and decorator hopping tasks assess the model’s ability to modify core structures and substituents.

**Baselines** To comprehensively evaluate our approach, we compare it against several representative baselines: **SMILES LSTM** [29]: An LSTM-based model trained via maximum likelihood estimation to generate SMILES strings. **Graph GA** [14]: A graph-based genetic algorithm that optimizes molecular structures through crossover and mutation. **Reinvent** [3]: A model combining recurrent neural networks with reinforcement learning, using reward functions to enhance both bioactivity and physicochemical properties. **GEGL** [1]: An approach that integrates graph neural networks with reinforcement learning to directly optimize molecular graphs. **MolRL-MGPT** [12]: A hybrid model that fuses GPT-based generative strategies with reinforcement learning to boost molecular diversity and target-specific performance.

**Experimental Details** First of all, we pre-trained the model using the training set provided by Guacamol. The pre-training was conducted for 15 epochs on a single A100 GPU over a duration of 3 hours. After pre-training, the model achieved a molecular validity of **97%**, demonstrating its high accuracy in molecular structure generation.

Subsequently, we further trained the model on 20 tasks from the Guacamol benchmark to evaluate its performance across different objectives. The **hyperparameter** settings used in this stage were as follows: **Batch size** = 50, **n\_steps** = 1000, **num\_agents** = 4, **Learning rate** = 1e-4, **Memory size** = 1000.

The complete training of the 20 benchmark tasks required 60 hours, whereas MolRL-MGPT took 400 hours under the same conditions (both on a single A100 GPU). Our model demonstrated a training speed nearly 6 times faster, highlighting the advantages of DPO’s stable training and faster convergence, which significantly reduces training costs.

Table 1: Scores of DPO&CL and baselines on the GuacaMol benchmark. (All task scores are rounded to three decimal places.)

	SMILES- LSTM	GraphGA	Reinvent	GEGL	MolRL- MGPT	<b>DPO&amp;CL</b>
Celecoxib rediscovery	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Troglitazone rediscovery	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.552	<b>1.000</b>	<b>1.000</b>
Thiothixene rediscovery	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Aripiprazole similarity	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Albuterol similarity	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Mestranol similarity	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
C <sub>11</sub> H <sub>24</sub>	0.993	0.971	0.999	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
C <sub>9</sub> H <sub>10</sub> N <sub>2</sub> O <sub>2</sub> PF <sub>2</sub> Cl	0.879	0.982	0.877	<b>1.000</b>	0.939	<b>1.000</b>
Median molecules 1	0.438	0.406	0.434	<b>0.455</b>	0.449	<b>0.455</b>
Median molecules 2	0.422	0.432	0.395	<b>0.437</b>	0.422	0.422
Osimertinib MPO	0.907	0.953	0.889	<b>1.000</b>	0.977	0.990
Fexofenadine MPO	0.959	0.998	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Ranolazine MPO	0.855	0.920	0.895	0.933	0.939	<b>0.950</b>
Perindopril MPO	0.808	0.792	0.764	0.833	0.810	<b>0.883</b>
Amlodipine MPO	0.894	0.894	0.888	0.905	<b>0.906</b>	<b>0.906</b>
Sitagliptin MPO	0.545	<b>0.891</b>	0.539	0.749	0.823	0.838
Zaleplon MPO	0.669	0.754	0.590	0.763	0.790	<b>0.797</b>
Valsartan SMARTS	0.978	0.990	0.095	<b>1.000</b>	0.997	0.994
deco hop	0.996	<b>1.000</b>	0.994	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
scaffold hop	0.998	<b>1.000</b>	0.990	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>
Total	17.340	17.983	16.350	17.627	18.052	<b>18.235</b>

As shown in Table 1, our model achieved the best performance on multiple tasks, with its overall score surpassing all other baselines. Specifically, our

method achieved top performance on 16 out of 20 benchmark tasks, demonstrating strong competitiveness in molecular generation. Compared to GEGL, our model exhibited higher stability across diverse tasks, achieving consistently superior performance rather than excelling in only a subset of cases.

The GuacaMol benchmark [5] consists of 20 tasks designed to comprehensively evaluate molecular generative models across rediscovery, similarity, isomer generation, scaffold hopping, and multi-parameter optimization (MPO). The molecules used in these tasks span diverse chemical classes, including drug-like compounds such as Celecoxib and Aripiprazole, small aliphatic molecules like  $C_{11}H_{24}$ , and highly specific drug targets such as Osimertinib and Ranolazine. This diversity ensures a rigorous assessment of the model’s ability to generalize across different molecular properties and structural features. Most of the tasks require not only the reproduction of known actives but also the generation of novel molecules that satisfy specific structural constraints, pharmacophoric features, or physicochemical profiles.

In terms of dataset size, GuacaMol includes over 1.5 million molecules drawn from the ChEMBL database, filtered to ensure drug-likeness and chemical feasibility. The challenge lies in generating valid, novel, and property-compliant molecules from this vast and structurally heterogeneous chemical space.

Our model’s effectiveness is particularly evident in chemically challenging tasks. For instance, in the Perindopril MPO task, which requires optimizing multiple pharmacokinetic and structural constraints simultaneously, our method achieved a score of 0.883, significantly outperforming existing methods by a margin of 0.05. This indicates strong optimization capabilities in complex molecular design settings. Another compelling example is the Ranolazine MPO task, where MolRL-MGPT—previously the best-performing model—surpassed its closest competitor by only 0.006. However, our model further improved upon MolRL-MGPT by an additional 0.011, demonstrating its ability to break through performance plateaus and generate high-quality, novel candidates in narrow, high-density regions of the chemical space.

These results highlight that our model not only performs well on standard rediscovery and similarity tasks but also maintains robustness and adaptability when facing multi-objective constraints and structurally diverse target molecules. Moreover, these findings validate the effectiveness of the DPO method and curriculum learning in guiding the generative process toward chemically meaningful and practically valuable solutions, laying a strong foundation for future applications in real-world drug discovery.

## 4.2 Molecular Generation for High Binding Affinity to Target Proteins

In this experiment, we utilized a prior model pretrained on the ZINC dataset. The evaluation was conducted on six tasks: JNK3, GSK3B, DRD2, and their pairwise combinations (JNK3+GSK3B, JNK3+DRD2, GSK3B+DRD2).

The model performance was evaluated using the oracle function provided by TDC. For multi-objective optimization tasks, we used the arithmetic mean of

the individual target scores as the final score to assess the overall performance of the generated molecules across multiple targets.

Table 2: The scores of the generated molecules on JNK3, GSK3 $\beta$ , DRD2, and pairwise combination tasks.

	top 1	top 10 mean	top 100 mean
<b>JNK3</b>	1.000	1.000	1.000
<b>Gsk3B</b>	1.000	1.000	1.000
<b>DRD2</b>	1.000	1.000	1.000
<b>JNK3+GSK3B</b>	0.944	0.943	0.938
<b>JNK3+DRD2</b>	0.925	0.925	0.920
<b>GSK3B+DRD2</b>	0.993	0.992	0.989

As shown in the table 2, our model successfully generated molecules with strong binding potential to JNK3, GSK3B, and DRD2, demonstrating its effectiveness in molecular generation tasks. These results indicate that our model not only performs well on the Guacamol benchmark but also excels in real-world drug discovery tasks, providing a solid foundation for future research in molecular design and generation.

### 4.3 Impact Analysis

This study suggests that multiple factors, including the learning rate, the number of agents, the sampling-to-training ratio, and the DPO parameter  $\beta$ , may influence model performance. Through preliminary analysis, we identified that the number of agents and the sampling-to-training ratio have a particularly significant impact. To validate this hypothesis, we conducted a systematic impact analysis experiment focusing on these two key parameters. The experimental results demonstrate that appropriately adjusting the number of agents and the sampling-to-training ratio can significantly enhance model performance, providing valuable theoretical insights and practical guidance for further model optimization.

**Agents Num** Experimental results in figure 3 indicate that the model achieves optimal performance when employing 2 to 4 agents. When the number of agents is too small, the model’s expressive capacity is limited, making it difficult to effectively learn complex patterns. Conversely, an excessive number of agents may introduce redundant information and increase optimization complexity, thereby lowering the model’s performance upper bound and slowing down convergence. Moreover, the number of agents also affects the score distribution of molecules stored in memory: fewer agents lead to a more concentrated distribution, whereas a larger number of agents result in a more dispersed distribution. Due to differences in the learning configurations of individual agents, their capabilities

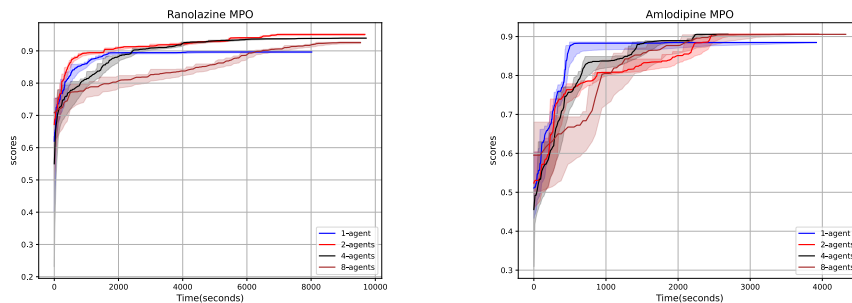


Fig. 3: Model performance on Ranolazine MPO and Amlodipine MPO tasks under different numbers of agents. (The curve represents the Top-10 score, while the shaded region indicates the score distribution of the top 100 molecules.)

diverge as training progresses, leading to an increasingly diverse molecular distribution. This broader distribution facilitates the optimization of DPO training and enhances the model’s generalization capability.

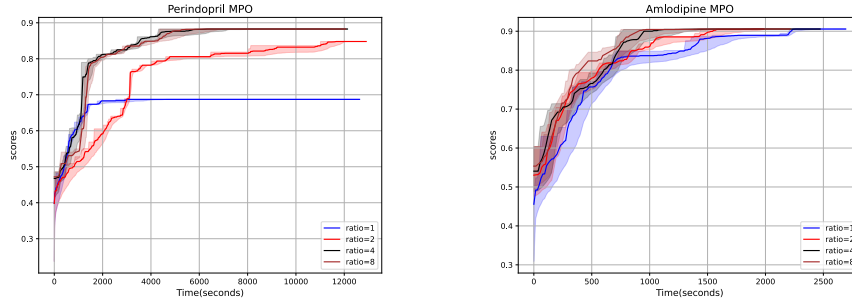


Fig. 4: Model performance on Perindopril MPO and Amlodipine MPO tasks under different Sampling-to-Training Ratios. (The curve represents the Top-10 score, while the shaded region indicates the score distribution of the top 100 molecules.)

**Sampling-to-Training Ratio** As shown in the figure 4, increasing the sampling-to-training ratio within a certain range can improve the model’s performance upper bound and accelerate convergence. If the ratio is too small, the weights assigned to a few high-quality molecules obtained by chance become excessively large, causing the model to shift towards them—even when they do not represent the optimal direction—ultimately leading to a suboptimal solution. Conversely,

if the ratio is too large, the model remains in the sampling phase for an extended period, collecting a large number of redundant molecules while lacking sufficient training. This results in inadequate gradient updates and significantly prolongs the model’s convergence time.

## 5 Conclusion

This study proposes a molecule generation method based on DPO and curriculum learning, and achieves favorable experimental results on the Guacamol benchmark and several target tasks. The experiments demonstrate that the proposed method has significant advantages in tasks that generate molecules with specified properties.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Ahn, S., Kim, J., Lee, H., Shin, J.: Guiding deep molecular optimization with genetic exploration. *Advances in neural information processing systems* **33**, 12008–12021 (2020)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: *Proceedings of the 26th annual international conference on machine learning*. pp. 41–48 (2009)
3. Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K., Patronov, A.: Reinvent 2.0: an ai tool for de novo drug design. *Journal of chemical information and modeling* **60**(12), 5918–5922 (2020)
4. Bohacek, R.S., McMartin, C., Guida, W.C.: The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews* **16**(1), 3–50 (1996)
5. Brown, N., Fiscato, M., Segler, M.H., Vaucher, A.C.: Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **59**(3), 1096–1108 (2019)
6. Cheng, X., Zhou, X., Yang, Y., Bao, Y., Gu, Q.: Decomposed direct preference optimization for structure-based drug design. *arXiv preprint arXiv:2407.13981* (2024)
7. DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics* **47**, 20–33 (2016)
8. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4**(2), 268–276 (2018)
9. Gu, S., Xu, M., Powers, A., Nie, W., Geffner, T., Kreis, K., Leskovec, J., Vahdat, A., Ermon, S.: Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in Neural Information Processing Systems* **37**, 44040–44063 (2025)
10. Guo, J., Fialková, V., Arango, J.D., Margreitter, C., Janet, J.P., Papadopoulos, K., Engkvist, O., Patronov, A.: Improving de novo molecular design with curriculum learning. *Nature Machine Intelligence* **4**(6), 555–563 (2022)

11. Hacothen, G., Weinshall, D.: On the power of curriculum learning in training deep networks. In: International conference on machine learning. pp. 2535–2544. PMLR (2019)
12. Hu, X., Liu, G., Zhao, Y., Zhang, H.: De novo drug design using reinforcement learning with multiple gpt agents. *Advances in Neural Information Processing Systems* **36**, 7405–7418 (2023)
13. Irwin, J.J., Shoichet, B.K.: Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**(1), 177–182 (2005)
14. Jensen, J.H.: A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science* **10**(12), 3567–3572 (2019)
15. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. In: International conference on machine learning. pp. 2323–2332. PMLR (2018)
16. Jin, W., Barzilay, R., Jaakkola, T.: Multi-objective molecule generation using interpretable substructures. In: International conference on machine learning. pp. 4849–4859. PMLR (2020)
17. Liu, S., Wang, J., Yang, Y., Wang, C., Liu, L., Guo, H., Xiao, C.: Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090* (2023)
18. Liu, X., Ye, K., van Vlijmen, H.W., Emmerich, M.T., IJzerman, A.P., van Westen, G.J.: Drugex v2: de novo design of drug molecules by pareto-based multi-objective reinforcement learning in polypharmacology. *Journal of cheminformatics* **13**(1), 85 (2021)
19. Liu, X., Ye, K., van Vlijmen, H.W., IJzerman, A.P., van Westen, G.J.: Drugex v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. *Journal of Cheminformatics* **15**(1), 24 (2023)
20. Loeffler, H.H., He, J., Tibo, A., Janet, J.P., Voronov, A., Mervin, L.H., Engkvist, O.: Reinvent 4: Modern ai-driven generative molecule design. *Journal of Cheminformatics* **16**(1), 20 (2024)
21. Mandal, S., Mandal, S.K., et al.: Rational drug design. *European journal of pharmacology* **625**(1-3), 90–100 (2009)
22. Meyers, J., Fabian, B., Brown, N.: De novo molecular design and generative models. *Drug discovery today* **26**(11), 2707–2715 (2021)
23. Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E., Stone, P.: Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research* **21**(181), 1–50 (2020)
24. Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H.: Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **9**, 1–14 (2017)
25. Popova, M., Isayev, O., Tropsha, A.: Deep reinforcement learning for de novo drug design. *Science advances* **4**(7), eaap7885 (2018)
26. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36** (2024)
27. Sanchez-Lengeling, B., Aspuru-Guzik, A.: Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**(6400), 360–365 (2018)
28. Schneider, G., Fechner, U.: Computer-based de novo design of drug-like molecules. *Nature reviews Drug discovery* **4**(8), 649–663 (2005)



29. Segler, M.H., Kogej, T., Tyrchan, C., Waller, M.P.: Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* **4**(1), 120–131 (2018)
30. Simm, G., Pinsler, R., Hernández-Lobato, J.M.: Reinforcement learning for molecular design guided by quantum mechanics. In: *International Conference on Machine Learning*. pp. 8959–8969. PMLR (2020)
31. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al.: A deep learning approach to antibiotic discovery. *Cell* **180**(4), 688–702 (2020)
32. Wadatalla, T., Rafailov, R., Hie, B.: Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv* pp. 2024–05 (2024)
33. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al.: Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology* **37**(9), 1038–1040 (2019)
34. Zhou, Z., Kearnes, S., Li, L., Zare, R.N., Riley, P.: Optimization of molecules via deep reinforcement learning. *Scientific reports* **9**(1), 10752 (2019)