DOMAIN-AGNOSTIC MOLECULAR GENERATION WITH CHEMICAL FEEDBACK

Yin Fang[♣♠], **Ningyu Zhang**[♣♠*], **Zhuo Chen**[♣♠], **Lingbing Guo**[♣♠], **Xiaohui Fan**[♣], **Huajun Chen**[♣♠♡*]
[♣] College of Computer Science and Technology, Zhejiang University
[♠] ZJU-Ant Group Joint Research Center for Knowledge Graphs, Zhejiang University
[♡] Hangzhou Innovation Center, Zhejiang University
{fangyin, zhangningyu, zhuo.chen, lbguo, fanxh, huajunsir}@zju.edu.cn

ABSTRACT

The generation of molecules with desired properties has become increasingly popular, revolutionizing the way scientists design molecular structures and providing valuable support for chemical and drug design. However, despite the potential of language models in molecule generation, they face challenges such as generating syntactically or chemically flawed molecules, having narrow domain focus, and struggling to create diverse and feasible molecules due to limited annotated data or external molecular databases. To tackle these challenges, we introduce MOLGEN, a pre-trained molecular language model tailored specifically for molecule generation. Through the reconstruction of over 100 million molecular SELFIES, MOLGEN internalizes structural and grammatical insights. This is further enhanced by domain-agnostic molecular prefix tuning, fostering robust knowledge transfer across diverse domains. Importantly, our chemical feedback paradigm steers the model away from "*molecular hallucinations*", ensuring alignment between the model's estimated probabilities and real-world chemical preferences. Extensive experiments on well-known benchmarks underscore MOLGEN's optimization capabilities in properties such as penalized logP, QED, and molecular docking. Additional analyses confirm its proficiency in accurately capturing molecule distributions, discerning intricate structural patterns, and efficiently exploring the chemical space[1].

1 INTRODUCTION

Molecule generation – synthesizing and designing novel molecules with desirable properties – holds an important place in chemical science, with numerous applications in drug discovery (Wang et al., 2022). Generating molecules is challenging due to the immense and discrete nature of the molecular space, which, with an estimated size of 10^{33}, makes exhaustive searches impractical (Polishchuk et al., 2013). Early, deep generative models (Jin et al., 2020; Zang & Wang, 2020; Luo et al., 2021; Shi et al., 2020b) have emerged as one of the most promising tools for exploring the broader synthetically accessible chemical space. These models' ability to automatically generate chemically valid and structurally similar molecules has proven to be invaluable for tasks such as the inverse design of functional compounds (Flam-Shepherd et al., 2022).

Current deep generative models typically involve initial training of an unconditional generative model through a large set of existing molecules, and then use additional reward functions (Cao & Kipf, 2018; Popova et al., 2018; You et al., 2018; Popova et al., 2019; Shi et al., 2020b; Zang & Wang, 2020) or property predictors (Liu et al., 2018; Jin et al., 2019; Gómez-Bombarelli et al., 2018) to guide the synthesis of new molecules with desired properties. However, these approaches are limited by challenges in training due to the high variance of Reinforcement Learning (RL) (Xie et al., 2021), fixed-dimensional latent generation space (Wang et al., 2023), and expert-provided generation rules (Sun et al., 2022), which impede efficient exploration of the broader chemical space.

[*]Corresponding author.
[1]Code is available at https://github.com/zjunlp/MolGen.

Recent advancements in language models have demonstrated great potential for understanding complex molecular distributions (Flam-Shepherd et al., 2022). To gain a more profound comprehension of the underlying molecular structures and their representations, researchers have begun integrating SMILES (Weininger, 1988), a linear string notation for describing molecular structures, with pre-trained language models (PLMs) (Irwin et al., 2022). Despite their widespread use, several issues remain inadequately considered. **Firstly**, the brittleness of SMILES may lead to a high proportion of generated chemically invalid strings, either due to syntactic errors (e.g., not corresponding to molecular graphs) or fundamental chemical principle violations (e.g., exceeding the maximum number of inter-atomic valence bonds) (Krenn et al., 2020). **Secondly**, almost all previous studies have focused primarily on synthetic molecules, neglecting natural products (Du et al., 2022a). Notably, natural products, characterized by enormous scaffold diversity and structural complexity, exhibit a distinct distribution compared to synthetic molecules and confer additional challenges for numerous molecule generation applications such as drug discovery (Atanasov et al., 2021). **Thirdly**, pre-trained molecular language models often succumb to "*molecular hallucinations*". This refers to instances where the generated molecules structurally adhere to chemical rules, yet fail to demonstrate the anticipated chemical activity in practical applications. This occurs because, although the models assimilate a vast array of molecular structural representations during pre-training, yet they might not fully capture the complex relationships with real-world chemistry and biological properties. Some methods attempt to mitigate this issue by using supervised fine-tuning or external databases (Irwin et al., 2022; Wang et al., 2023), but they may constrain the direction of molecular optimization.

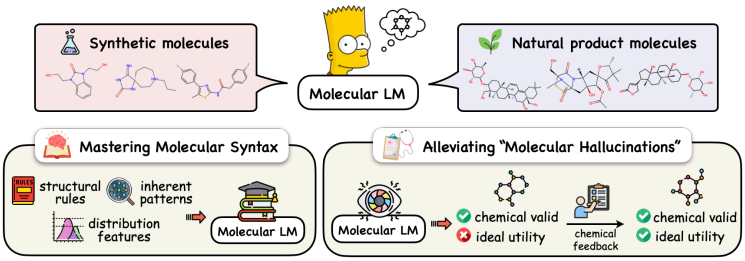


Figure 1: MOLGEN excels at generating chemically valid molecules with expected efficacy in both synthetic and natural product domains.

To tackle these challenges, we present MOLGEN, a novel pre-trained molecular language model designed for efficient molecule generation. As illustrated in Figure 1, our approach comprises: ***(i)* A two-stage domain-agnostic molecular pre-training.** First, we train bidirectional and auto-regressive Transformers (Vaswani et al., 2017) to reconstruct over 100 million corrupted molecular SELFIES (Krenn et al., 2020). This endows the model with a profound understanding of the structure, grammar, and intrinsic semantic information of SELFIES, an entirely robust molecular language, free from the predicaments of syntactic and semantic inconsistency often associated with conventional SMILES notation. Next, we leverage domain-agnostic molecular prefix tuning, enabling MOLGEN to harness knowledge transferable across diverse domains (i.e., synthetic and natural products), facilitating task adaptation. ***(ii)* A chemical feedback paradigm to alleviate "*molecular hallucinations*"**. By aligning the model's generative probabilities with real-world chemical preferences, MOLGEN learns to evaluate and rectify its molecular outputs, ensuring the generation of chemically valid molecules with genuine utility and anticipated properties.

Through extensive testing on both synthetic and natural product molecular datasets, we establish MOLGEN's capability in producing chemically valid molecules, navigating chemical spaces efficiently, and achieving notable optimization in properties like penalized logp, QED, and molecular docking. Our further analysis underscores MOLGEN's adeptness at understanding complex molecular distributions, recognizing meaningful substructures, and the efficacy of the chemical feedback mechanism, offering novel perspectives and tools to the molecular generation community.

2 METHODOLOGY

Figure 2 illustrates the general framework of MOLGEN. The pre-training process (§2.1) comprises two stages: molecular language syntax learning and domain-agnostic molecular prefix tuning. Then, a chemical feedback paradigm (§2.2) is introduced to align the PLM with the anticipated chemical preferences in the downstream phase.

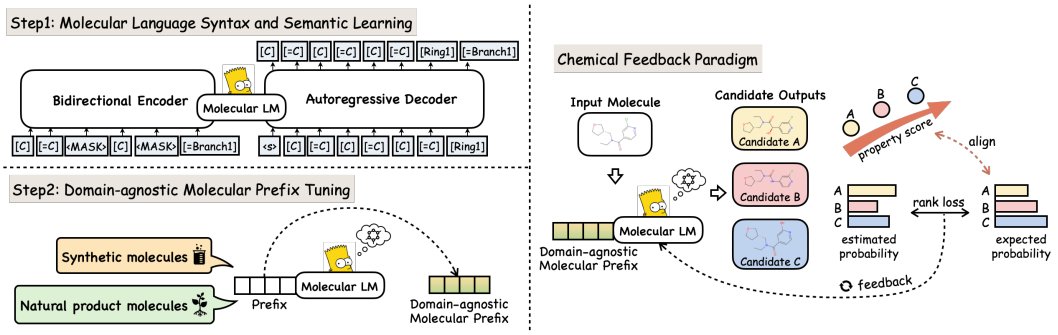


Figure 2: Overview of MOLGEN: pre-training (left) and downstream (right) stages.

2.1 DOMAIN-AGNOSTIC MOLECULAR PRE-TRAINING

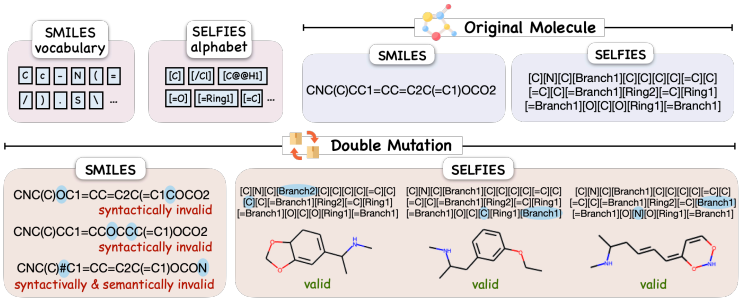


Figure 3: Random double mutations of SMILES and SELFIES derived from the same molecule, with blue markings indicating mutation locations. The likelihood of retaining a valid SMILES after a single mutation is 9.9%. For SELFIES, it's a consistent 100% (Krenn et al., 2020).

SMILES and SELFIES are two molecular languages that associate a token sequence with a molecular structure. SMILES denotes molecules as chains of atoms, encapsulating branches within parentheses and signifying ring closures with corresponding number pairs. Despite its longstanding prominence in cheminformatics, SMILES is fundamentally flawed in that it lacks a mechanism to ensure the validity of molecular strings in terms of syntax and physical principles (Krenn et al., 2020). Hence, we employ SELFIES (Krenn et al., 2022), a fully robust molecular language that guarantees every possible combination of symbols in the alphabet corresponds to a chemically sound graph structure. In contrast to SMILES, SELFIES overcomes syntactic invalidity by mapping each token to a specific structure or reference, effectively resolving issues such as unbalanced parentheses or ring identifiers, as depicted in Figure 3. MOLGEN boasts a compact and specialized vocabulary size of 185. While modest in size, this vocabulary is already sufficient to ensure that the language model learns meaningful representations (Rives et al., 2021).

Being the first of its kind to train language models utilizing SELFIES, our work necessitates a solid foundation for comprehending both the syntax and semantics of this language. To achieve a high-quality initialization for MOLGEN, we employ BART model (Lewis et al., 2020) during **the first stage of pre-training**, as shown in Figure 2. Firstly, we convert 100 million unlabeled molecules into SELFIES strings. The standardized representation of SELFIES facilitates the direct construction of an alphabet from the dataset, eliminating the need for a separate tokenizer to discern frequent substrings, thereby preventing the generation of nonsensical tokens. Secondly, we randomly select tokens from the original SELFIES string $S = \{s_1, \cdots, s_j, \cdots, s_l\}$ and replace them with a special token `[MASK]`. Finally, we encode the corrupted SELFIES using a bidirectional model and calculate the likelihood of S with a left-to-right autoregressive decoder. Formally, the cross-entropy between the decoder's output and the original input constitutes the reconstruction loss:

$$\mathcal{L}_{\text{ce}}(S) = -\sum_{j=1}^{l} \sum_{s} p_{\text{true}}\left(s|S, S_{<j}\right) \log p_\theta\left(s|S, S_{<j}; \theta\right), \tag{1}$$

where $S_{<j}$ denotes the partitial original sequence $\{s_0, \cdots, s_{j-1}\}$, s_0 is a pre-defined start token <s>. p_{true} refers to the one-hot distribution obtained under the standard maximum likelihood estimation:

$$p_{\text{true}}\left(s|S, S_{<j}\right) = \begin{cases} 1, & s = s_j \\ 0, & s \neq s_j \end{cases}. \tag{2}$$

Upon mastering the fundamental grammatical knowledge of SELFIES, we proceed to the second stage of pre-training, wherein we introduce the domain-agnostic molecular prefix as a domain instructor to

facilitate the transfer of knowledge across diverse domains. Unlike the conventional prefix-tuning approach, which exclusively updates the prefix matrices without altering the pre-trained model parameters (Mao et al., 2022; Li & Liang, 2021; He et al., 2022), we capitalize on its influence over the entire model's parameters to effectively bolster its ability to comprehend various domains.

We commence by prepending two sets of m tunable prefix vectors $\boldsymbol{P}_k, \boldsymbol{P}_v \in \mathbb{R}^{m \times d}$, shared among domains, to the keys and values of the multi-head attention at each layer. The output attention score for each head can be formulated as:

$$\text{head} = \text{Attn}\left(\boldsymbol{x}\boldsymbol{W}_q, [\boldsymbol{P}_k,\ \boldsymbol{X}\boldsymbol{W}_k], [\boldsymbol{P}_v, \boldsymbol{X}\boldsymbol{W}_v]\right), \tag{3}$$

where $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ denotes the input to a Transformer layer with length m, $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d_h}$ are project matrices that map inputs to queries, keys, and values, and $\boldsymbol{x} \in \mathbf{R}^d$ is a query vector.

Alternatively, the attention between $\boldsymbol{x}$ and $\boldsymbol{X}$ on head can be expressed as:

$$\begin{aligned}
\text{head} &= \text{softmax}\left(\boldsymbol{x}\boldsymbol{W}_q[\boldsymbol{P}_k,\ \boldsymbol{X}\boldsymbol{W}_k]^\top\right)\begin{bmatrix}\boldsymbol{P}_v \\ \boldsymbol{X}\boldsymbol{W}_v\end{bmatrix} = \text{softmax}\left(\boldsymbol{x}\boldsymbol{W}_q\begin{bmatrix}\boldsymbol{P}_k^\top \\ (\boldsymbol{W}_k)^\top(\boldsymbol{X})^\top\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{P}_v \\ \boldsymbol{X}\boldsymbol{W}_v\end{bmatrix} \\
&= \lambda(\boldsymbol{x})\ \text{softmax}\left(\boldsymbol{x}\boldsymbol{W}_q\boldsymbol{P}_k^\top\right)\boldsymbol{P}_v + (1-\lambda(\boldsymbol{x}))\ \text{softmax}\left(\boldsymbol{x}\boldsymbol{W}_q(\boldsymbol{W}_k)^\top(\boldsymbol{X})^\top\right)\boldsymbol{X}\boldsymbol{W}_v \\
&= \lambda(\boldsymbol{x})\underbrace{\text{Attn}\left(\boldsymbol{x}\boldsymbol{W}_q, \boldsymbol{P}_k, \boldsymbol{P}_v\right)}_{\text{attention of domain-agnostic molecular prefix}} + (1-\lambda(\boldsymbol{x}))\underbrace{\text{Attn}\left(\boldsymbol{x}\boldsymbol{W}_q, \boldsymbol{X}\boldsymbol{W}_k, \boldsymbol{X}\boldsymbol{W}_v\right)}_{\text{standard attention}},
\end{aligned} \tag{4}$$

where $\lambda(\boldsymbol{x})$ is a scalar representing the sum of normalized attention weights on the prefixes.

In this way, domain-agnostic molecular prefixes integrate domain knowledge into the original head attention through linear interpolation. These prefixes are trained simultaneously on different molecular domains, acting as a domain instructor that influences the parameters of the entire model, thereby enhancing the model's mastery of different molecular structural complexities and scaffold diversities.

2.2 Chemical Feedback Paradigm: Align PLM with Chemical Preference

After the pre-training stage, the model gains the capability to generate syntactically correct molecules. However, it may still suffer from "*molecular hallucination*". Consider a scenario where the model is employed to design novel drug molecules. It suggests a molecule with a unique cyclic structure, known to effectively bind with certain drug targets. In an attempt to boost structural robustness, the model introduces an additional side chain. However, this addition, despite seemingly increasing stability, actually interferes with the molecule's intended target interaction, leading to its ineffectiveness. This situation exemplifies "*molecular hallucination*", where the structural enhancements made by the model do not translate into functional success.

Definition 1. ***Molecular hallucinations*** *refer to molecules generated by language models that comply with chemical structural rules, yet fail to exhibit practical utility or the anticipated properties.*

Such hallucinations can hinder drug discovery efficiency, escalate costs, and compromise the real-world applicability of the model. Moreover, an abundance of hallucinated molecules may overshadow truly promising molecular structures. To alleviate "*molecular hallucinations*", we propose a strategy that can effectively gauge and rectify the quality of generated molecular structures. This chemical feedback paradigm ensures that produced molecules are not only syntactically correct but also of high practical utility. Specifically, as illustrated in Figure 2, we align the model's probabilistic rankings of diverse molecular responses with preference rankings observed in actual chemical contexts.

The measure of anticipated chemical preference, denoted as $\text{Ps}(\cdot)$, can be characterized in various ways; in this study, we define it based on the property score. Given a molecule $S = \{s_1, \cdots, s_l\}$, we can generate a set of candidate SELFIES $\mathcal{S}^*$ with distinct property scores using our pre-trained molecular language model. For each (S_i, S_j) pair in $\mathcal{S}^*$ that satisfies $\text{Ps}(S_i) > \text{Ps}(S_j)$, we expect:

$$p_{\text{true}}(S_i|S) > p_{\text{true}}(S_j|S), \quad \forall S_i, S_j \in \mathcal{S}^*,\ \text{Ps}(S_i) > \text{Ps}(S_j). \tag{5}$$

To incentivize the model to assign higher probabilities to candidates with desired properties, we utilize a rank loss (Liu et al., 2022). The rank loss arises when candidates with suboptimal properties obtain higher estimated probabilities compared to those with commendable properties:

$$\mathcal{L}_{\text{rank}}(S) = \sum_i \sum_{j>i} \max\left(0, f\left(S_j\right) - f\left(S_i\right) + \gamma_{ij}\right), \quad \forall i < j,\ \text{Ps}(S_i) > \text{Ps}(S_j), \tag{6}$$

where $\gamma_{ij} = (j - i) * \gamma$ represents the margin multiplied by the difference in rank between the candidates, and $f(S) = \sum_{t=1}^{l} \log p_\theta (s_t \mid S, S_{<t}; \theta)$ denotes the estimated log-probability provided by our pre-trained model with parameters θ. Consequently, we furnish chemical feedback to align the pre-trained model with the chemical preference, without necessitating any supplementary reference data. Unlike supervised fine-tuning, which may still be susceptible to hallucinations due to its reliance on ideal samples, chemical feedback equips the model with a broader perspective. It educates the model on both the commendable and the suboptimal, leading to more informed generation.

Nonetheless, fine-tuning the model solely with sequence-level coordination may diminish its generative capability. To ensure the model retains its generative prowess while optimizing for desired properties, we strike a balance by merging the sequence-level rank loss with token-level cross-entropy loss. The overall loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \alpha \mathcal{L}_{\text{rank}}, \tag{7}$$

where α is the weight of the rank loss. In practice, we leverage label smoothing (Szegedy et al., 2016) to transform the target distribution p_{true} (Eq. 2) in $\mathcal{L}_{\text{ce}}$ (Eq. 1) to a "soft" label, allocating probability mass β to other tokens in the alphabet of length N:

$$p_{\text{true}} (s|S, S_{<j}) = \begin{cases} 1 - \beta, & s = s_j \\ \frac{\beta}{N-1}, & s \neq s_j \end{cases}. \tag{8}$$

Overall, the cross-entropy loss serves as a normalization, complementing the rank loss by ensuring that the model allocates a balanced probability mass throughout the sequence. MolGen autonomously steer its learning and optimization paths based on the evaluations of molecules it generates. This cycle of generation and adjustment within the model epitomizes a self-reflective system, even as it incorporates an external scoring function to refine and validate its assessments.

3 Experiments

3.1 Experimental Setup

In the first stage of pre-training, we randomly select over 100 million unlabelled molecules from the publicly available ZINC-15 dataset (Sterling & Irwin, 2015), which is the same corpus used in Irwin et al. (2022). The chosen molecules meet specific criteria: they're reactive, available for purchase, have a molecular weight of $\leq$ 500 Daltons, and a LogP (octanol-water partition coefficient) of $\leq$ 5. The second stage includes 2.22 million molecules spanning both synthetic (Irwin et al., 2012; Polykovskiy et al., 2018) and natural product domains (Zhao et al., 2023). In the downstream tasks, as detailed in the following section, we thoroughly investigate the model's capabilities from two perspectives. More information on dataset and experimental procedures are in Appendices C and G.

3.2 Main Results

3.2.1 MolGen Captures Real-world Molecular Distributions

An essential capability for any molecular generation model is to capture the molecular distribution and generate diverse and realistic molecules. Such capabilities are paramount when constructing virtual libraries to advance computer-aided drug discovery endeavors (van Hilten et al., 2019). By leveraging a set of compounds, either manually or automatically selected, these models are designed to expand datasets significantly, all the while retaining the implicit structural and chemical patterns inherent to the reference set. In this section, we use seven well-established metrics, detailed in Appendix G, to evaluate the proficiency of models in generating molecules that conform to the distribution of real-world molecules. We generate 10,000 synthetic molecules following the setting in Polykovskiy et al. (2018), and 80,000 natural product molecules based on the pre-trained MolGen.

Table 1 reveals the following observations: *(i)* MolGen demonstrates a remarkable ability to produce valid molecules without the need for additional valency checks, as required by JT-VAE (Jin et al., 2018). Since LIMO (Eckmann et al., 2022) also employs SELFIES, the generated molecules maintain 100% validity. However, the inherent complexity of natural product scaffolds presents a significant challenge for most models, resulting in a failure to produce valid molecules. The better performance of Chemformer (Irwin et al., 2022) can be attributed to its proficiency in learning SMILES grammar

Table 1: Molecular distribution learning performance on two molecule domains. The cells in highlight denote the best results garnered by MOLGEN and the peak performance achieved by the baselines.

MODEL	SYNTHETIC MOLECULES							NATURAL PRODUCT MOLECULES						
	Validity↑	Frag↑	Scaf↑	SNN↑	IntDiv↑	FCD↓	Novelty↑	Validity↑	Frag↑	Scaf ↑	SNN↑	IntDiv↑	FCD↓	Novelty↑
AAE	.9368	.9910	.9022	.6081	.8557	.5555	.7931	.0082	.9687	.2638	.3680	.8704	4.109	.9943
LATENTGAN	.8966	.9986	.8867	.5132	.8565	.2968	.9498	.9225	.2771	.0884	.5321	.6009	45.53	.9949
CHARRNN	.9748	.9998	.9242	.6015	.8562	.0732	.8419	.7351	.8816	.5212	.4179	.8756	2.212	.9792
VAE	.9767	.9994	.9386	.6257	.8558	.0990	.6949	.2627	.8840	.4563	.3950	.8719	4.318	.9912
JT-VAE	**1.000**	.9965	.8964	.5477	.8551	.3954	.9143	**1.000**	.8798	.5012	.3748	.8743	12.03	.9957
LIMO	**1.000**	.9562	.1073	.6125	.8544	.1532	.8956	**1.000**	.7242	.0005	.3416	.7726	31.84	.9962
CHEMFORMER	.9843	.9889	.9248	.5622	.8553	.0061	.9581	.9825	.9826	.4126	.5875	.8650	.8346	.9947
MOLGEN	**1.000**	**.9999**	**.9999**	**.9996**	**.8567**	**.0015**	**1.000**	**1.000**	**.9994**	**.8404**	**.8148**	**.8878**	**.6519**	**.9987**

during large-scale pre-training, highlighting the importance of pre-training. *(ii)* For the synthetic datasets, most models generate molecules with comparable fragments (Frag) and scaffolds (Scaf) to those of the reference molecules. MOLGEN excels at capturing substructure distributions in natural products, outperforming other models. *(iii)* MOLGEN exhibits the highest SNN and lowest FCD scores, indicating its excellent ability to master the dataset statistics in terms of both biological properties and topological structures. Moreover, its strong performance in IntDiv and Novelty metrics suggests that MOLGEN is well-suited for discovering new chemical structures and exploring unknown chemical space without overfitting. A visual comparison of the training set and generated molecules is presented in Appendix H.1.

3.2.2 MOLGEN MITIGATES MOLECULAR HALLUCINATIONS

Addressing the issue of "*molecular hallucinations*" has been a long-standing challenge in the realm of computer-aided molecular design. In this section, we delve into the prowess of MOLGEN in tackling this challenge and primarily focus on two types of experiments: **targeted molecule discovery** and **constrained molecular optimization**. Unlike the molecular distribution learning task, where we only rely on the pre-trained model, here we incorporate the chemical feedback paradigm to align the model with genuine chemical preferences. Specifically, we adopt the penalized logP (p-logP) (Jin et al., 2018), QED (Bickerton et al., 2012) and binding affinity to two protein targets as our optimization criteria, as detailed in Appendix G.

Table 2: Comparison of **QED and penalized logP maximization** methods on synthetic molecules. ♣ indicates output length limit (maximum molecule length of ZINC250K), while ♡ means no limit. The first row summarizes the top 3 property scores from the ZINC250K dataset.

	MODEL	PENALIZED LOGP			QED		
		1st	2nd	3rd	1st	2nd	3rd
	ZINC250K	4.52	4.30	4.23	0.948	0.948	0.948
♣	GCPN	7.98	7.85	7.80	**0.948**	0.947	0.946
♣	MOLDQN	11.80	11.80	11.80	**0.948**	0.943	0.943
♣	LIMO	10.50	9.69	9.60	0.947	0.946	0.945
♣	OURS	**30.51**	**28.98**	**28.95**	**0.948**	**0.948**	**0.948**
♡	JT-VAE	5.30	4.93	4.49	0.925	0.911	0.910
♡	GRAPHAF	12.23	11.29	11.05	**0.948**	**0.948**	0.947
♡	GRAPHDF	13.70	13.18	13.17	**0.948**	**0.948**	**0.948**
♡	MARS	44.99	44.32	43.81	**0.948**	**0.948**	**0.948**
	MOLGEN	**80.30**	**74.70**	**69.85**	**0.948**	**0.948**	**0.948**

Table 3: The top 3 **highest binding affinities** (i.e., lowest dissociation constants, K_D, as estimated with AutoDockGPU (Santos-Martins et al., 2021)) from a total of 10k generated molecules for each method.

MODEL	ESR1			ACAA1		
	1st	2nd	3rd	1st	2nd	3rd
GCPN	6.4	6.6	8.5	75	83	84
MOLDQN	373	588	1062	240	337	608
GRAPHDF	25	47	51	370	520	590
MARS	17	64	69	163	203	236
LIMO	0.72	0.89	1.4	37	37	41
MOLGEN	**0.13**	**0.35**	**0.47**	**3.36**	**3.98**	**8.50**

Targeted molecule discovery focuses on generating novel molecules with superior chemical properties. To evaluate model effectiveness, we first present the top-3 property scores of molecules generated on the synthetic dataset in Table 2, following conventions from prior studies (Shi et al., 2020b; Eckmann et al., 2022). It's essential to note that the p-logP score tends to increase linearly with molecule length (Xie et al., 2021; Eckmann et al., 2022). To ensure a fair comparison, we categorize the baselines into two groups. MOLGEN, due to its ability to handle variable-length output, is evaluated under both configurations.

In Table 2, MOLGEN outperforms all baselines in p-logP score and achieves comparable results for QED, indicating the effectiveness of the chemical feedback paradigm in promoting desired molecule probabilities. Further evidence of MOLGEN's capabilities can be found in the results for natural products in Appendix H.2. Given that a mere 0.701% of molecules in our reference set achieve a QED score above 0.9 (with a peak score of 0.9439, as detailed in Appendix C), MOLGEN's achievement of a 0.9478 score highlights its potential in drug discovery. Moreover, the model's ability to produce molecules with a p-logP score of 54.33, substantially exceeding the reference set's high of 17.69.

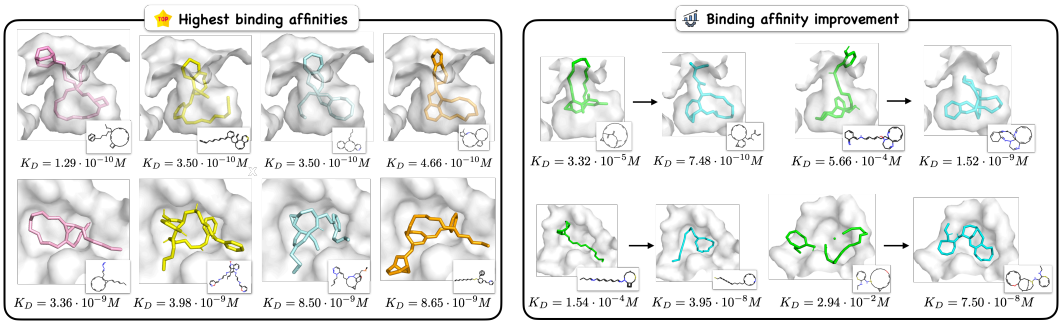


Figure 4: Optimizing ligand binding affinity using MolGen. (a) 3D visualization of ligands with the **highest binding affinities** docked against ESR1 (top row) and ACAA1 (bottom row). The protein pocket is displayed semi-opaquely, and the 2D molecular structure of the ligand is shown in the bottom right corner. (b) Examples of **binding affinity improvement** for protein targets ESR1 (top row) and ACAA1 (bottom row).

Moving beyond basic properties, we tackle a more realistic challenge: generating molecules with high binding affinity towards target proteins. Binding affinity quantifies the potency of interactions between a molecule and its intended protein target. Our investigations primarily target the binding sites of two human proteins: the estrogen receptor (PDB ESR1, UniProt P03372) and the peroxisomal acetyl-CoA acyl transferase 1 (PDB ACAA1, UniProt P09110). A detailed exploration of these proteins is available in Appendix G. As shown in Table 3, MolGen surpasses prior methods in enhancing binding affinities. Figure 4 (a) illustrates exemplary optimal ligands. To delve deeper into MolGen's optimization capability, we undertook an optimization for the 1,000 molecules with the lowest affinities for each protein receptor. Figure 4 (b) offers a comparative visualization of affinity advancements pre- and post-optimization, achieving overall relative improvements of 96.7% for ESR1 and 70.4% for ACAA1. These results illuminate MolGen's versatility in both targeted optimization of simpler properties and the more complex domain of molecular docking.

Table 4: Mean (and standard deviation) penalized logP improvement of generated molecules compared to inputs with different similarity constraints.

Model	Improvement	
	$\delta = 0.6$	$\delta = 0.4$
JT-VAE	0.28 (0.79)	1.03 (1.39)
GCPN	0.79 (0.63)	2.49 (1.30)
MolDQN	1.86 (1.21)	3.37 (1.62)
VSeq2Seq	2.33 (1.17)	3.37 (1.75)
VJTNN	2.33 (1.24)	3.55 (1.67)
GA	3.44 (1.09)	5.93 (1.41)
GraphAF	4.98 (6.49)	8.21 (6.51)
GraphDF	4.51 (5.80)	9.19 (6.43)
LIMO	1.80 (2.00)	3.60 (2.30)
Chemformer	2.48 (0.89)	3.56 (1.32)
RetMol	3.78 (3.29)	11.55 (11.27)
RT	2.21 (1.30)	3.16 (1.50)
MolGen	**12.08** (0.82)	**12.35** (1.21)

Constrained molecular optimization aims to modify a given molecule to improve desired properties while satisfying a similarity constraint (denoted as δ). Following previous studies (Jin et al., 2018; Shi et al., 2020b; Luo et al., 2021; Eckmann et al., 2022), we optimize 800 molecules from the ZINC250K dataset that exhibit the lowest p-logP scores. To assess the similarity between the optimized and original molecules, we utilize the Tanimoto similarity with Morgan fingerprints (Rogers & Hahn, 2010).

In Table 4, MolGen yields superior results under both similarity constraints, illustrating its prowess in scouring the proximate chemical space for molecules with higher property scores. MolGen's performance, surpassing models that employ additional reward functions, property predictors, and retrieval databases, confirms that equipping the model with the ability to discern chemical preference is instrumental in alleviating "*molecular hallucinations*".

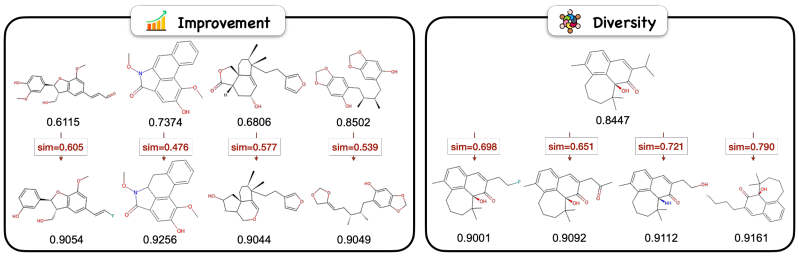


Figure 5: Illustrations of constrained optimization based on QED score within the natural products.

To further probe MolGen's capabilities, we expand our constrained optimization experiments to include QED scores for synthetic molecules and both properties for natural products. Figure 5 showcases examples of QED score optimization on natural products. These instances reveal that despite the complex molecular structure and elongated length of natural products, MolGen can elevate the property score whilst sustaining a degree of similarity between the input and the modified molecule. Moreover, MolGen preserves the diversity

of the generated molecules as it explores the nearby chemical space. Additional visual validations are provided in Appendix H.3.

3.3 A CLOSER LOOK AT MOLGEN

To dissect the potential of MOLGEN, we devise experiments from different perspectives.

3.3.1 PRE-TRAINING STAGE CAPTURES COMPLEX MOLECULAR CHARACTERISTICS

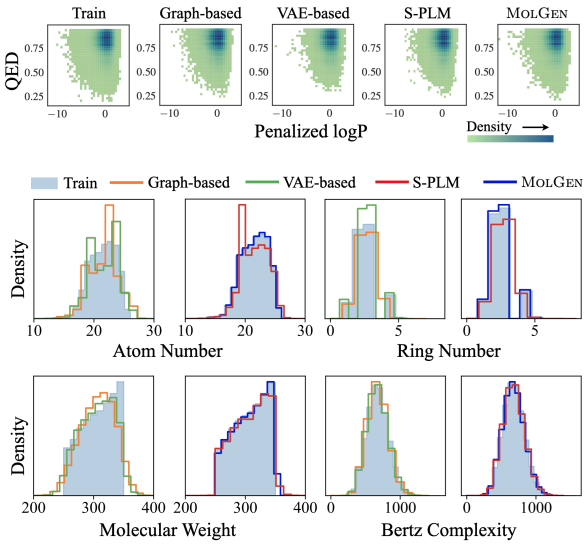


Figure 6: Comparative analysis of molecules generated by different models with respect to properties, atom counts, ring counts, molecular weights, and Bertz complexity (S-PLM stands for SMILES-based PLM).

To understand the differences in property distributions and molecular structures learned during the pre-training phase, we compare the pre-trained MOLGEN with the most popular deep generative GRAPH-based (Jin et al., 2018), VAE-based (Blaschke et al., 2018), and SMILES-based language models (Irwin et al., 2022). For this assessment, the training and generation configurations of all models align with the molecular distribution learning task on the synthetic MOSES dataset.

As shown in the 2D histograms of p-logP and QED scores in Figure 6, both VAE-based and SMILES-based PLMs tend to produce molecules with larger p-logP and QED scores than the training data. In comparison, the GRAPH-based model learns the main mode of p-logP in the training data, while MOLGEN exhibits a slightly superior performance - analogous outcomes are observed for QED. Furthermore, in terms of molecular topology, PLMs outperform others in perceiving atom numbers, ring numbers, and molecular weights, with MOLGEN producing a slightly closer match to the training distribution. All the models are proficient at picking up on molecular Bertz complexity. PLMs, particularly MOLGEN, demonstrate the capacity to capture the properties and structural attributes of the training molecules while maintaining generational diversity.

3.3.2 CHEMICAL FEEDBACK PARADIGM FACILITATES PROPERTY OPTIMIZATION

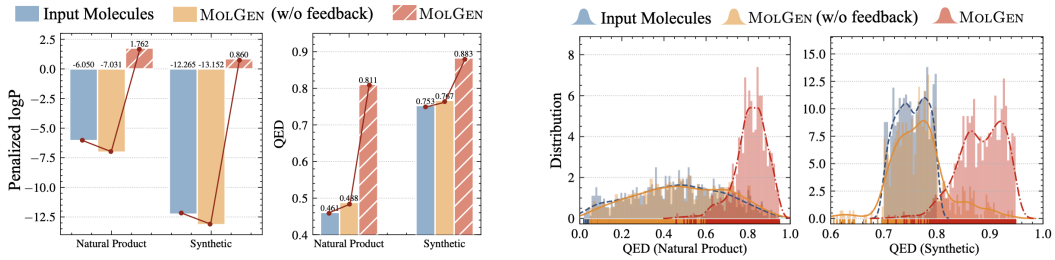


Figure 7: Property variations across different MOLGEN configurations.

As part of our investigation, we conduct an ablation study to examine the role of the chemical feedback paradigm in mitigating "*molecular hallucinations*". Starting from a batch of molecules from the domains of natural products and synthetic compounds, Figure 7 portrays the variations in property scores of molecules generated by different model configurations. A more comprehensive view of these variations is provided in Appendix H.2.

Without the chemical feedback, the PLM tends to generate molecules with property scores closely resembling those of the initial molecules. This can be attributed to the absence of a guiding signal, leaving the model to rely heavily on its learned patterns from the training data. However, once the

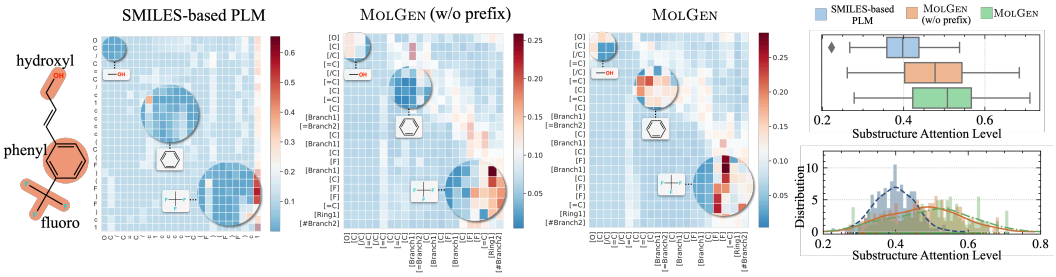


Figure 8: Meaningful substructure attention exploration. Visualization of attention maps learned by various PLM models for the same molecule (left), and substructure attention level of the three models (right). All models being compared are of a similar parameter scale for consistency.

chemical feedback mechanism is integrated, we witness an increase in property scores from the initial to the concluding groups. This underscores the pivotal role of chemical feedback: it furnishes the model with immediate feedback on its performance in generating molecules with the chemical preference, thus steering its outputs towards the desired objectives and alleviating the hallucinations.

3.3.3 MolGen Implicitly Comprehends Molecular Substructures

In this section, we investigate PLMs' ability to implicitly discern essential substructures when leveraging different molecular languages (SMILES and SELFIES). For a more intuitive comprehension, we visualize the attention weights of each token within an identical molecule. Specifically, we extract and normalize the attention weights from the final self-attention layer, as depicted in Figure 8.

The attention map generated by MolGen shows that the *fluoro* group garners the highest attention weights, followed by the *phenyl* and *hydroxyl* groups. This stems from the *fluoro* group's exceptional electron-capturing capabilities, significantly influencing the molecule's polarity. Meanwhile, the *phenyl* group constitutes a common organic functional group, and the *hydroxyl* group substantially impacts the intermolecular force between the molecule and water. Leveraging domain-agnostic molecular prefixes, MolGen directs its attention more efficiently towards these pivotal substructures. These prefixes, acting as domain instructors, enhance the model's adaptability across diverse molecular domains, steering attention away from less essential substructures. Conversely, SMILES-based PLM might divert attention to symbols or numbers devoid of intrinsic chemical significance. Evidently, by employing a precise vocabulary free from such distractions, MolGen maintains a clear and implicit understanding of molecular substructures. Further visualizations and analyses supporting this observation are available in Appendix F and H.4.

To objectively measure the model's focus on vital substructures, we propose a metric termed "Substructure Attention Level (SAL)". This metric is determined by the percentage of attention scores allocated to meaningful substructure tokens within a molecule. Higher SAL scores indicate a stronger focus on meaningful substructures. For effective evaluation, we intentionally select 200 molecules from PubChem, characterized by their simpler structures containing only 1-2 functional groups. This selection criterion ensures that the model's attention isn't diluted across excessively intricate structures, allowing for a clearer reflection of its focus on specific functional groups. The box and distribution plots in Figure 8 vividly depict the SAL of the three PLMs. In line with visualization results, both versions of MolGen surpass the SMILES-based PLM, underscoring MolGen's superior concentration on meaningful substructures. The prefix-enhanced MolGen exhibits a slight edge, highlighting the prefix's role in enhancing attentiveness.

4 Conclusion and Future Work

In this work, we propose MolGen, a pre-trained molecular language model specifically tailored for molecule generation. Our in-depth study on MolGen confirms its proficiency in generating molecules with chemical preferences while avoiding "*molecular hallucinations*". Furthermore, our model shows potential in identifying essential molecular substructures. Interesting future directions include: *i)* applying MolGen to other tasks such as retrosynthesis and reaction prediction (Shi et al., 2020a), *ii)* exploring multimodal pre-training like Edwards et al. (2022); Su et al. (2022); Fang et al. (2024), *iii)* incorporating additional sources of knowledge. We make our pre-trained model, code, and data publicly available, in the hope that our work will foster future research in the field.

ACKNOWLEDGMENTS

We would like to express gratitude to the anonymous reviewers for kind comments. This work was supported by the National Natural Science Foundation of China (No. 62206246), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Ningbo Natural Science Foundation (2021J190), CAAI-Huawei MindSpore Open Fund, Yongjiang Talent Introduction Programme (2021A-156-G), CCF-Baidu Open Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

REPRODUCIBILITY STATEMENT

All data, code, and model weights can be found in the Supplementary Materials. For a detailed description of the dataset, please refer to Appendix C. For specific experimental settings, please see Appendix G.

ETHICS STATEMENT

This study was carried out in strict accordance with ethical guidelines and best practices in research. The data utilized were sourced from publicly available datasets, and no proprietary or confidential data were used. This study does not involve any ethical issues.

REFERENCES

Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding deep molecular optimization with genetic exploration. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html.

Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. Molgpt: Molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.*, 62(9):2064–2076, 2022. doi: 10.1021/ACS.JCIM.1C00600. URL https://doi.org/10.1021/acs.jcim.1c00600.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 37(1-2): 1700123, 2018.

Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mac. Intell.*, 5(4):432–444, 2023. doi: 10.1038/S42256-023-00639-Z. URL https://doi.org/10.1038/s42256-023-00639-z.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *CoRR*, abs/1805.11973, 2018. URL http://arxiv.org/abs/1805.11973.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Lusine Khondkaryan, Karen Hambardzumyan, Zaven Navoyan, Hrant Khachatrian, and Armen Aghajanyan. Bartsmiles: Generative masked language models for molecular representations. *CoRR*, abs/2211.16349, 2022. doi: 10.48550/arXiv.2211.16349. URL https://doi.org/10.48550/arXiv.2211.16349.

Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *CoRR*, abs/2203.14500, 2022a. doi: 10.48550/ARXIV.2203.14500. URL https://doi.org/10.48550/arXiv.2203.14500.

Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *CoRR*, abs/2203.14500, 2022b. doi: 10.48550/ARXIV.2203.14500. URL https://doi.org/10.48550/arXiv.2203.14500.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 2197–2214. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.168. URL https://doi.org/10.18653/v1/2021.emnlp-main.168.

Peter Eckmann, Kunyang Sun, Bo Zhao, Mudong Feng, Michael K. Gilson, and Rose Yu. LIMO: latent inceptionism for targeted molecule generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5777–5792. PMLR, 2022. URL https://proceedings.mlr.press/v162/eckmann22a.html.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *EMNLP*, pp. 375–413. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.emnlp-main.26.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024. URL https://openreview.net/pdf?id=Tlsdsb6l9n.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):1–10, 2022.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=0RDcd5Axok.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768, 2012. doi: 10.1021/CI3001277. URL https://doi.org/10.1021/ci3001277.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.*, 3(1):15022, 2022. doi: 10.1088/2632-2153/AC3FFB. URL https://doi.org/10.1088/2632-2153/ac3ffb.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023a. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023b. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2328–2337. PMLR, 2018. URL http://proceedings.mlr.press/v80/jin18a.html.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi S. Jaakkola. Learning multimodal graph-to-graph translation for molecule optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xJAsA5F7.

Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4839–4848. PMLR, 2020. URL http://proceedings.mlr.press/v119/jin20a.html.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1(4):45024, 2020. doi: 10.1088/2632-2153/ABA947. URL https://doi.org/10.1088/2632-2153/aba947.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588, 2022. doi: 10.1016/J.PATTER.2022.100588. URL https://doi.org/10.1016/j.patter.2022.100588.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1945–1954. PMLR, 2017. URL http://proceedings.mlr.press/v70/kusner17a.html.

Youngchun Kwon, Seokho Kang, Youn-Suk Choi, and Inkoo Kim. Evolutionary design of molecules based on deep learning and a genetic algorithm. *Scientific reports*, 11(1):1–11, 2021.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.703. URL https://doi.org/10.18653/v1/2020.acl-main.703.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://doi.org/10.18653/v1/2021.acl-long.353.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. Constrained graph variational autoencoders for molecule design. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7806–7815, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/b8a03c5c15fcfa8dae0b03351eb1742f-Abstract.html.

Yixin Liu, Pengfei Liu, Dragomir R. Radev, and Graham Neubig. BRIO: bringing order to abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2890–2903. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.207. URL https://doi.org/10.18653/v1/2022.acl-long.207.

Youzhi Luo, Keqiang Yan, and Shuiwang Ji. Graphdf: A discrete flow model for molecular graph generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7192–7203. PMLR, 2021. URL http://proceedings.mlr.press/v139/luo21a.html.

Tengfei Ma, Jie Chen, and Cao Xiao. Constrained generation of semantically valid graphs via regularizing variational autoencoders. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7113–7124, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1458e7509aa5f47ecfb92536e7dd1dc7-Abstract.html.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 6253–6264. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.433. URL https://doi.org/10.18653/v1/2022.acl-long.433.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 1906–1919. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.173. URL https://doi.org/10.18653/v1/2020.acl-main.173.

AkshatKumar Nigam, Pascal Friederich, Mario Krenn, and Alán Aspuru-Guzik. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1lmyRNFvr.

Jie Pan. Large language model for molecular chemistry. *Nature Computational Science*, pp. 1–1, 2023.

Pavel G. Polishchuk, Timur I. Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.*, 27(8):675–679, 2013. doi: 10.1007/S10822-013-9672-4. URL https://doi.org/10.1007/s10822-013-9672-4.

Daniil Polykovskiy, Alexander Zhebrak, Benjamín Sánchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Sergey I. Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *CoRR*, abs/1811.12823, 2018. URL http://arxiv.org/abs/1811.12823.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. Molecularrnn: Generating realistic molecular graphs with optimized properties. *CoRR*, abs/1905.13372, 2019. URL http://arxiv.org/abs/1905.13372.

Vipula Rawte, Amit P. Sheth, and Amitava Das. A survey of hallucination in large foundation models. *CoRR*, abs/2309.05922, 2023. doi: 10.48550/ARXIV.2309.05922. URL https://doi.org/10.48550/arXiv.2309.05922.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, 118(15):e2016239118, 2021. doi: 10.1073/PNAS.2016239118. URL https://doi.org/10.1073/pnas.2016239118.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010. doi: 10.1021/CI100050T. URL https://doi.org/10.1021/ci100050t.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mac. Intell.*, 4(12):1256–1264, 2022. doi: 10.1038/S42256-022-00580-7. URL https://doi.org/10.1038/s42256-022-00580-7.

Diogo Santos-Martins, Leonardo Solis-Vasquez, Andreas F Tillack, Michel F Sanner, Andreas Koch, and Stefano Forli. Accelerating autodock4 with gpus and gradient-based local search. *Journal of chemical theory and computation*, 17(2):1060–1073, 2021.

Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1): 120–131, 2018.

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8818–8827. PMLR, 2020a. URL http://proceedings.mlr.press/v119/shi20d.html.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=S1esMkHYPr.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 3784–3803. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-EMNLP.320. URL https://doi.org/10.18653/v1/2021.findings-emnlp.320.

Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In Vera Kurková, Yannis Manolopoulos, Barbara Hammer, Lazaros S. Iliadis, and Ilias Maglogiannis (eds.), *Artificial Neural Networks and Machine Learning - ICANN 2018 - 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I*, volume 11139 of *Lecture Notes in Computer Science*, pp. 412–422. Springer, 2018. doi: 10.1007/978-3-030-01418-6_41. URL https://doi.org/10.1007/978-3-030-01418-6_41.

Teague Sterling and John J. Irwin. ZINC 15 - ligand discovery for everyone. *J. Chem. Inf. Model.*, 55(11):2324–2337, 2015. doi: 10.1021/ACS.JCIM.5B00559. URL https://doi.org/10.1021/acs.jcim.5b00559.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *CoRR*, abs/2209.05481, 2022. doi: 10.48550/ARXIV.2209.05481. URL https://doi.org/10.48550/arXiv.2209.05481.

Mengying Sun, Jing Xing, Han Meng, Huijun Wang, Bin Chen, and Jiayu Zhou. Molsearch: Search-based multi-objective molecular generation and property optimization. In Aidong Zhang and Huzefa Rangwala (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 4724–4732. ACM, 2022. doi: 10.1145/3534678.3542676. URL https://doi.org/10.1145/3534678.3542676.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Austin Tripp and José Miguel Hernández-Lobato. Genetic algorithms are strong baselines for molecule generation. *CoRR*, abs/2310.09267, 2023. doi: 10.48550/ARXIV.2310.09267. URL https://doi.org/10.48550/arXiv.2310.09267.

Niek van Hilten, Florent Chevillard, and Peter Kolb. Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.*, 59(2):644–651, 2019. doi: 10.1021/ACS.JCIM.8B00737. URL https://doi.org/10.1021/acs.jcim.8b00737.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Mingyang Wang, Zhe Wang, Huiyong Sun, Jike Wang, Chao Shen, Gaoqi Weng, Xin Chai, Honglin Li, Dongsheng Cao, and Tingjun Hou. Deep learning approaches for de novo drug design: An overview. *Current Opinion in Structural Biology*, 72:135–144, 2022.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard G. Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=vDFA1tpuLvk.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988. doi: 10.1021/CI00057A005. URL https://doi.org/10.1021/ci00057a005.

Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science*, 10(34):8016–8024, 2019.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. MARS: markov molecular sampling for multi-objective drug discovery. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=kHSu4ebxFXY.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *CoRR*, abs/2309.06794, 2023. doi: 10.48550/ARXIV.2309.06794. URL https://doi.org/10.48550/arXiv.2309.06794.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay S. Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp.

6412–6422, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d60678e8f2ba9c540798ebbde31177e8-Abstract.html.

Chengxi Zang and Fei Wang. Moflow: An invertible flow model for generating molecular graphs. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 617–626. ACM, 2020. doi: 10.1145/3394486.3403104. URL https://doi.org/10.1145/3394486.3403104.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023. doi: 10.48550/ARXIV.2309.01219. URL https://doi.org/10.48550/arXiv.2309.01219.

Hui Zhao, Yuan Yang, Shuaiqi Wang, Xue Yang, Kaicheng Zhou, Caili Xu, Xuyao Zhang, Jiajun Fan, Dongyue Hou, Xingxiu Li, Hanbo Lin, Ying Tan, Shanshan Wang, Xinyi Chu, Dongzhi Zhuoma, Fengying Zhang, Dianwen Ju, Xian Zeng, and Yu Zong Chen. NPASS database update 2023: quantitative natural product activity and species source database for biomedical research. *Nucleic Acids Res.*, 51(D1):621–628, 2023. doi: 10.1093/NAR/GKAC1069. URL https://doi.org/10.1093/nar/gkac1069.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

A AVAILABILITY OF MOLGEN

We have made MOLGEN accessible via **Hugging Face** in support of the broader scientific community[2,3,4]. It is noteworthy that MOLGEN is versatile enough to be applied to tasks beyond the three primary ones discussed in this paper, such as reaction prediction and retrosynthetic analysis. However, due to computational resource constraints, our experimentation is confined to the generation tasks within this study.

It's important to note that our generation task is different from 3D molecular generation. In 3D molecular generation, methods usually consider spatial conformations, bond angles, bond lengths, and other three-dimensional structural aspects of molecules. These approaches often use molecular force fields and molecular docking techniques to optimize the three-dimensional structures of generated molecules. In contrast, 2D molecular generation aims to create two-dimensional flat structures that capture the chemical composition, bond connectivity, and molecular topology of molecules. This approach places a stronger emphasis on molecular topology and chemical information, providing a representation of the molecule's overall structural arrangement and connectivity.

Our focus on 2D molecular generation is driven by several reasons. **Firstly**, 2D molecular representations capture essential chemical information and structural features, making them highly interpretable and suitable for various downstream applications such as virtual screening and drug design. **Secondly**, 2D molecular generation offers computational efficiency and scalability, enabling us to explore a larger chemical space and generate a higher number of diverse molecules within a reasonable time frame. **Lastly**, while 3D molecular generation is valuable for studying molecular interactions and binding modes, it often requires complex optimization techniques and is computationally more demanding. By concentrating on 2D molecular generation, we can achieve a balance between generating chemically relevant molecules and efficiently exploring chemical space for various property optimizations. We leave the incorporation of 3D conformation information into molecular design for our future work.

B LIMITATIONS AND POTENTIAL ISSUES

While our model, MOLGEN, achieves significant advancements in molecule generation, it is important to acknowledge some of its limitations, which open avenues for future research.

Computational Efficiency: The process of training and fine-tuning MOLGEN, especially with large datasets, can be computationally intensive, which may limit its usage in scenarios with limited computational resources.

Model Interpretability: Though MOLGEN exhibits prowess in generating molecules with designated properties and discerning vital molecular substructures, the opacity of transformer-based models complicates the understanding of the explicit rationale behind its determinations.

Applicability Limitations: A salient limitation of MOLGEN is its exclusive support for single-target optimization. The chemical feedback paradigm, whilst proficient in managing single-target molecular properties, may grapple with multiple targets. Disparate rankings for multiple objectives could engender ambiguity in the model's optimization trajectory, potentially culminating in less-than-optimal solutions. Future endeavors could investigate methodologies to adapt the chemical feedback paradigm to accommodate and prioritize diverse objectives.

Generality Limitations: In a bid to assess the versatility of MOLGEN, we extended our investigations to reaction prediction. Our fine-tuned model, devoid of any reliance on reaction templates, registered a 71.4% accuracy in predicting products from a pool of 39,990 reaction samples. While this underscores the model's capability to predict reactions to a certain degree, it's noteworthy that MOLGEN is not inherently structured for this task, thereby potentially curtailing its performance. Consequently, future research could consider designing a model architecture or training paradigm that concurrently and systematically accommodates reaction prediction, molecule generation, and other tasks.

[2] https://huggingface.co/zjunlp/MolGen-large
[3] https://huggingface.co/zjunlp/MolGen-large-opt
[4] https://huggingface.co/spaces/zjunlp/MolGen

C DATA INFORMATION

This section provides further information regarding the dataset employed in our study. The division of the molecular dataset into "synthetic" and "natural product" domains is to effectively explore and understand molecules of varying complexities and origins. The "synthetic" domain encompasses artificially synthesized chemical molecules tailored for specific needs, e.g., drug development. On the other hand, the "natural product" domain covers molecules naturally occurring, which are pivotal in biological activities and often provide insights for drug development. Natural product molecules generally exhibit greater structural complexity and diversity, often resulting from the myriad of unique chemical structures produced through natural biological processes. This classification helps us better understand the unique challenges and features of each domain.

In our research, we follow the methodologies of prior works (Polykovskiy et al., 2018) for distribution learning, where the baselines focus on synthetic molecule generation. Building upon this foundation, we have extended our scope by including the generation of natural products as a new and more challenging task. This expansion not only enhances the complexity of the tasks we address but also broadens the applicability of our model to a wider range of molecular structures encountered in various scientific domains.

For the natural product dataset, we sourced 30,926 compounds from the Natural Product Activity & Species Source Database (NPASS) [5] (Zhao et al., 2023). Out of these, we arbitrarily chose 30,126 molecules for training and reserved 800 molecules for testing, utilizing the same sets for all ensuing molecule generation tasks.

The characteristics of our datasets are depicted in Appendix Table 1. It is apparent that the natural product dataset manifests a distinctive distribution in comparison to the synthetic dataset, characterized by a broader spectrum of p-logP scores and reduced QED scores. This underscores the augmented complexity intrinsic to the optimization of natural product properties.

Appendix Table 1: Data statistics.

DATASET	LENGTH			PENALIZED LOGP			QED		
	MIN	MAX	MEAN	MIN	MAX	MEAN	MIN	MAX	MEAN
MOSES	13	55	35	-10.241	3.329	-0.027	0.191	0.948	0.807
ZINC250K	8	72	37	-22.189	5.073	-0.622	0.117	0.948	0.732
NATURAL PRODUCT	2	436	55	-51.083	17.691	-2.186	0.005	0.944	0.438

D RELATED WORK

D.1 DEEP GENERATIVE MODELS

In the last decade, significant strides have been made in the field of deep molecule generation (Gómez-Bombarelli et al., 2018). An array of molecular graph-based generative methods have surfaced (Ma et al., 2018; Simonovsky & Komodakis, 2018; Jin et al., 2020; Zang & Wang, 2020; Luo et al., 2021), while another branch has treated this task as a sequence generation problem with a preference for SMILES Kusner et al. (2017); Gómez-Bombarelli et al. (2018); Segler et al. (2018); Kwon et al. (2021). Based on them, existing approaches can be broadly categorized into four venues. **Bayesian Optimization** (Gómez-Bombarelli et al., 2018; Jin et al., 2018; Winter et al., 2019) learns a continuous latent space of molecules and optimizes the target properties by navigating through this space, but it often demands a protracted evaluation time to optimize the objective function (Du et al., 2022b). **Reinforcement Learning** approaches utilize an agent to select actions (e.g., adding substructures) in an explicit chemical space to enhance desired properties (Cao & Kipf, 2018; Popova et al., 2018; You et al., 2018; Popova et al., 2019; Shi et al., 2020b; Zang & Wang, 2020). However, these methods can suffer from high variance (Xie et al., 2021). An alternative approach is to employ a **Variational Auto-Encoder** (Simonovsky & Komodakis, 2018; Jin et al., 2019; Gómez-Bombarelli et al., 2018; Liu et al., 2018), but its performance heavily relies on the quality of the fixed-dimensional latent space. **Genetic Algorithms** (Jensen, 2019; Ahn et al., 2020; Nigam et al., 2020; Tripp & Hernández-Lobato, 2023) leverage predefined mutation and crossover rules to generate molecules.

[5] https://bidd.group/NPASS/

Despite their flexibility, obtaining the necessary prior knowledge and rules can be a challenge, hindering the efficiency of search process.

D.2 Pre-trained Language Models

Just as the syntax of natural languages enforces a grammatical structure that facilitates the connection between words in specific ways, biological symbols also amalgamate in precise structural manners. PLMs have emerged as an intuitive solution for molecule generation, and several pioneers have already begun to harness SMILES-based language models, yielding promising performance (Bagal et al., 2022; Irwin et al., 2022; Flam-Shepherd et al., 2022; Ross et al., 2022; Chilingaryan et al., 2022; Pan, 2023). To date, the only publicly available PLM capable of tackling molecule generation tasks is Chemformer (Irwin et al., 2022), which follows BART (Lewis et al., 2020) to corrupt SMILES and optimize a reconstruction loss for pre-training. Expanding on the foundation laid by Chemformer, RetMol (Wang et al., 2023) incorporates external retrieval data to further improve the synthesis of molecules. Nonetheless, SMILES imposes and circumscribes grammatical rules, leading to a significant number of sequences within the appropriate character set not belonging to well-defined molecules. Additionally, the paucity of annotated or reference data may constrain the optimization direction of molecules. Diverging from those approaches, MolGen is pre-trained using SELFIES, which is immune to syntactic and semantic obstacles while permitting facile adaptation to different domains by sharing knowledge among model parameters via domain instruction. Moreover, it autonomously aligns with the objective of producing desirable molecules without the need for external annotated data.

D.3 Hallucination

In the field of Natural Language Processing (NLP), "hallucination" refers to generating text or responses that, while grammatically correct, fluent, and natural, deviate from the provided source inputs or lack factual accuracy (Maynez et al., 2020; Dziri et al., 2021; Shuster et al., 2021; Ji et al., 2023a; Rawte et al., 2023; Ye et al., 2023). Hallucinations are typically categorized into several types: Input-conflicting hallucinations (where the model's output deviates from the user's input), Context-conflicting hallucinations (where the output conflicts with information previously generated), and Fact-conflicting hallucinations (where the output contradicts established world knowledge) (Zhang et al., 2023). The causes of these hallucinations are varied, including biases in training data, the model's lack of access to real-time information, or the inherent limitations of the model in comprehending and generating contextually accurate responses. (Ji et al., 2023b; Zhang et al., 2023; Rawte et al., 2023).

The concept of "hallucination" is not restricted to the domain of NLP. Its adaptation in fields like molecular science, as seen in the term "*molecular hallucination*", reflects a similar disconnect between structural validity and functional accuracy. In this context, "*molecular hallucination*" refers to molecules generated by language models that are chemically valid but fail to exhibit desired properties or functionalities. In essence, these molecules, although structurally sound, do not meet the specific chemical criteria or functional expectations set for them, similar to how text generated by a language model might be grammatically correct but deviate from the intended message or content of the source input. This analogy aims to convey the concept of "unfulfilled potential" or "misleading outcomes" in molecular generation.

E Compared Baselines

In this section, we expound upon the baselines employed for comparison in our experiments. These baselines are reproduced using their open-source codes under identical experimental conditions. The baselines include:

- JT-VAE (Jin et al., 2018), a Variational Autoencoder (VAE)-based generative model that constructs a molecular graph by generating a scaffold junction tree and assembling its nodes.
- GCPN (You et al., 2018), a Reinforcement Learning (RL)-based method that crafts a molecule by optimizing a reward comprising adversarial loss and molecular property objectives.

- MOLGQN (Zhou et al., 2019), an RL-based approach that capitalizes on double Q-learning and chemical domain knowledge.
- MARS (Xie et al., 2021), a Markov Chain Monte Carlo sampling approach that employs an adaptive fragment-editing proposal distribution with Graph Neural Networks (GNN).
- GRAPHAF (Shi et al., 2020b), an autoregressive flow model that sequentially adds edges and nodes to generate molecular graphs.
- GRAPHDF (Luo et al., 2021), a normalizing flow model utilizing a discrete latent variable model and is fine-tuned with RL.
- LIMO (Eckmann et al., 2022), a VAE-based model leveraging a variational autoencoder-generated latent space.
- CHEMFORMER (Irwin et al., 2022), a pre-trained molecular language model operating on SMILES representations.
- RETMOL (Wang et al., 2023), a retrieval-based framework predicated on CHEMFORMER that incorporates a task-specific retrieval database to guide the generative model towards creating new molecules that fulfill the desired design criteria.
- RT (Born & Manica, 2023), a Transformer-based model pre-trained on SELFIES that generate molecules by inputting expected molecular property values along with a given molecular scaffold (with the generated molecules incorporating this scaffold), or to predict molecular property values based on an input molecule.

F COMPARISON WITH SMILES-BASED PLM

In this section, we delineate the disparities between two molecular language models, Chemformer (Irwin et al., 2022) and MOLGEN. For fairness, we select the large version of Chemformer for comparison in our paper, given its analogous size to MOLGEN. Both models leverage a pre-training dataset of 100 million molecules from the ZINC-15 dataset (Sterling & Irwin, 2015). MOLGEN boasts a more compact and specialized vocabulary size of 185, as opposed to Chemformer's expansive vocabulary of 523. This allows MOLGEN to more effectively encapsulate critical molecular substructure information.

Moreover, we present a more detailed discussion concerning SELFIES and SMILES.

- *Inherent Robustness*: Although chemical tools like RDKit (Landrum, 2013) can externally validate SMILES strings, the representation itself doesn't inherently ensure grammatical or chemical correctness. In contrast, the construction principle of SELFIES ensures a surjective mapping to molecular graphs.
- *Generative Capabilities*: Flam-Shepherd et al. (2022) provides further evidence by comparing the generative capabilities of deep models using SMILES and SELFIES. SELFIES consistently outperforms SMILES in validity, uniqueness, and novelty across tasks. Notably, SELFIES excels with longer and more complex molecules, whereas using SMILES becomes challenging due to increased character requirements and the heightened risk of encountering errors.
- *Quantitative Experiments*: Our paper includes quantitative experiment outcomes. Table 1 and Figure 6 encompass comparative analyses of SMILES and SELFIES from distribution learning and molecule generation perspectives. Note that the MolGen version in this comparison does not use the chemical feedback mechanism.
- *About SMILES*: We respect and recognize SMILES's significant contributions as a molecular descriptor. Our inclination towards SELFIES is motivated by its inherent validity in molecular generation and its simpler vocabulary, ideal for molecular language pretraining.

G EXPERIMENT DETAILS AND METRICS

In this section, we elucidate the evaluation metrics, training procedures, and hyperparameters utilized for each task and dataset within our experiments. MOLGEN is implemented using Pytorch and

trained on 6 Nvidia V100 GPUs. The specific experimental settings and parameters are presented in Appendix Table 2.

Appendix Table 2: Hyper-parameter settings.

HYPER-PARAMETERS	VALUE
maximum sequence length	{55, 148, 436}
learning rate	{1e-5, 3e-5, 1e-4}
batch size	{8, 32, 64, 200, 256}
weight of rank loss α	{1,3,5}
prefix length	5

G.1 TWO-STAGE PRE-TRAINING

In the first stage of pre-training, we train a Seq2seq model to learn the structure, grammar, and intrinsic semantic information of SELFIES. To efficiently share parameters and knowledge, during the second stage of pre-training, we train the domain-agnostic molecular prefixes across two molecular domains. It is noteworthy that the pre-training objectives in both the first and second stages are aligned. Subsequently, we initialize the prefixes for each task with the pre-trained prefixes and optimize them for that particular task.

We utilize the LAMB optimizer, employing a linear warm-up of the learning rate for the first 180,000 gradient updates, succeeded by a linear decay for the remaining training steps. This process comprised 600 million steps with a batch size of 256 molecules per GPU.

G.2 MOLECULAR DISTRIBUTION LEARNING

We outline the metrics employed to evaluate the performance of the generative models in our experiments, encompassing:

- *Validity*, which gauges the proportion of generated molecules adhering to valence rules.
- *Fragment similarity (Frag)*, comparing the distribution of BRICS fragments in the generated and reference sets. For instance, the Frag metric will be high if the molecules in both sets share similar fragments. Conversely, if some fragments are over- or under-represented (or entirely absent) in the generated set, the metric will be low.
- *Scaffold similarity (Scaff)* comparing the frequencies of Bemis–Murcko scaffolds (comprising all molecule's linker fragments connecting rings and ring structures) in the generated and reference sets. Specifically, if the model seldom produces a specific chemotype from a reference set, the metric will be low.
- *Similarity to the nearest neighbor (SNN)*, which measures the average Tanimoto similarity between a molecule from the generated set and its nearest neighbor molecule in the reference dataset. If the generated molecules deviate significantly from the manifold of the reference set, then the similarity to the nearest neighbor will be low.
- *Internal diversity (IntDiv)*, assessing the chemical diversity of generated molecules by calculating the average Tanimoto coefficient within the generated set.
- *Fréchet ChemNet Distance (FCD)*, considering chemically and biologically pertinent information about molecules. It can discern if the generated molecules share similar biological and chemical properties with real molecules.
- *Novelty*, measuring the percentage of the generated molecules that are not present in the training set and assessing the ability to explore the unknown chemical space.

To obtain the results detailed in Table 1, MOLGEN is trained using the AdamW optimizer with a batch size of 200 for the MOSES dataset and 32 for the natural product dataset on 6 Nvidia V100 GPUs for 100 epochs. A linear warm-up of 20000 steps was also employed.

G.3 TARGETED MOLECULE DISCOVERY & CONSTRAINED MOLECULAR OPTIMIZATION

G.3.1 SIMPLE PROPERTIES

We utilize properties such as p-logP and QED as they are commonly used benchmarks in the field.

- *P-logP* refers to the logP score penalized by ring size and synthetic accessibility.
- *QED* estimates the drug-likeness of a molecule quantitatively.

For computation, p-logP and QED scores are calculated by empirical prediction models, and we employ the script based on the official implementation (Shi et al., 2020b) for comparability.

G.3.2 PROTEIN TARGETS

Binding affinity pertains to the strength of the interaction between a drug-like molecule and its target protein. We focus on optimizing binding affinity for two human proteins:

- *Human Estrogen Receptor (ESR1)*: A well-studied protein targeted by drugs for breast cancer treatment. This choice is driven by its clinical relevance and the availability of numerous known binding molecules, enabling effective comparison with generated compounds. MolGen utilizes solely the crystal structure of the protein (PDB 1ERR) for docking calculations and binding site information, without access to additional data on known binders.
- *Human Peroxisomal Acetyl-CoA Acyl Transferase 1 (ACAA1)*: Despite lacking known binders, this enzyme possesses a crystal structure (PDB 2IIK) featuring a potential drug-binding pocket. Identified via the Structural Genomics Consortium, this protein is recognized as a potentially disease-relevant target, possessing a documented crystal structure but devoid of known binding molecules.

The determination of binding affinity employs AutoDockGPU (Santos-Martins et al., 2021).

For both targeted molecule discovery and constrained molecular optimization tasks, we employ the chemical feedback paradigm to align the PLM with the optimization objectives. Initially, we use the pre-trained MOLGEN to generate 30 candidates for each data sample in synthetic compounds and 8 candidates for natural products. We then train the model on 6 Nvidia V100 GPUs for 100 epochs. The batch size is set to 6 for both the synthetic and natural product datasets. We utilize the AdamW optimizer, incorporating a linear warm-up of 20,000 steps.

H ADDITIONAL VISUALIZATION OF MOLECULES GENERATED BY MOLGEN

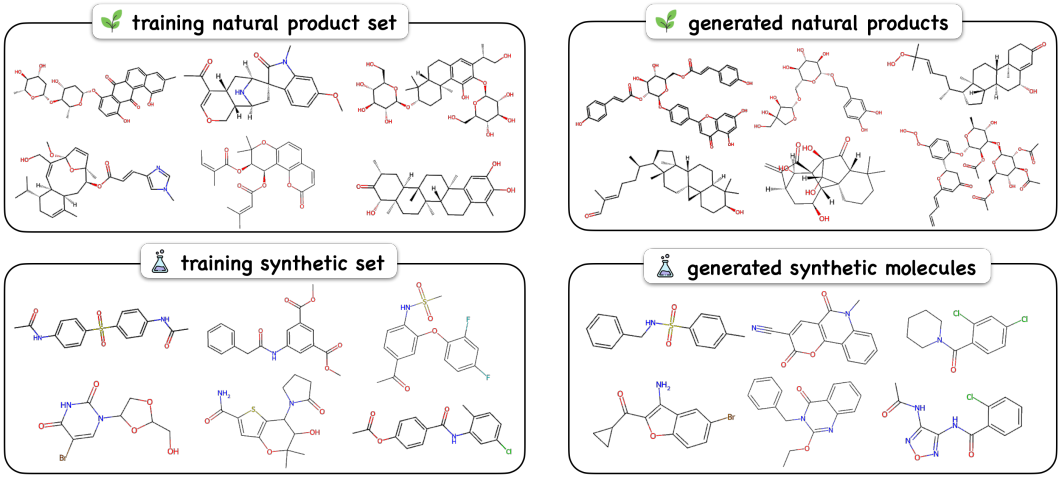
H.1 MOLECULAR DISTRIBUTION LEARNING

In this section, we furnish additional visual insights underscoring the prowess of MOLGEN in the realm of distribution learning. Appendix Figure 1 offers a comparative view of molecules from the training set and those generated by MOLGEN for both natural product and synthetic datasets.

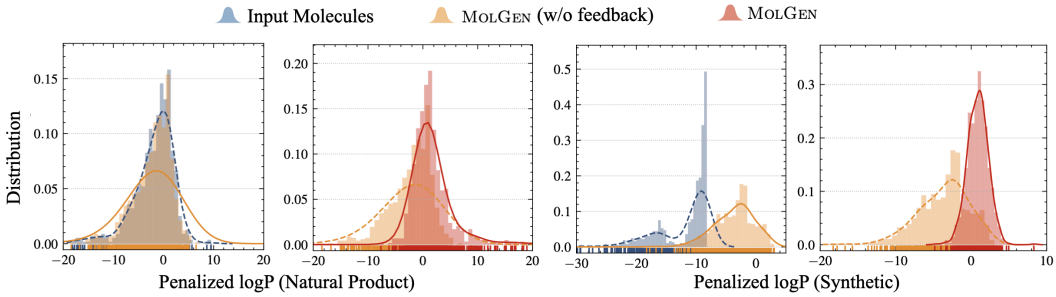
The illustrated molecules provide a visual representation of how effectively MOLGEN is able to capture and reproduce the structural characteristics of molecules from different domains. This is particularly noteworthy given the substantial structural variation between molecules in the natural product and synthetic datasets. The ability of MOLGEN to generate molecules that so closely resemble the training set highlights its capability to learn and reproduce the underlying distribution of molecular structures across diverse chemical domains.

H.2 TARGETED MOLECULE DISCOVERY

In this section, we present additional visualizations to further substantiate our claims made in the main text. To provide a more nuanced understanding of the changes in molecular properties depicted in Figure 7, we illustrate the distribution dynamics of the p-logP score in Appendix Figure 2. This reaffirms that our pre-trained MOLGEN model effectively learns the distribution of molecular



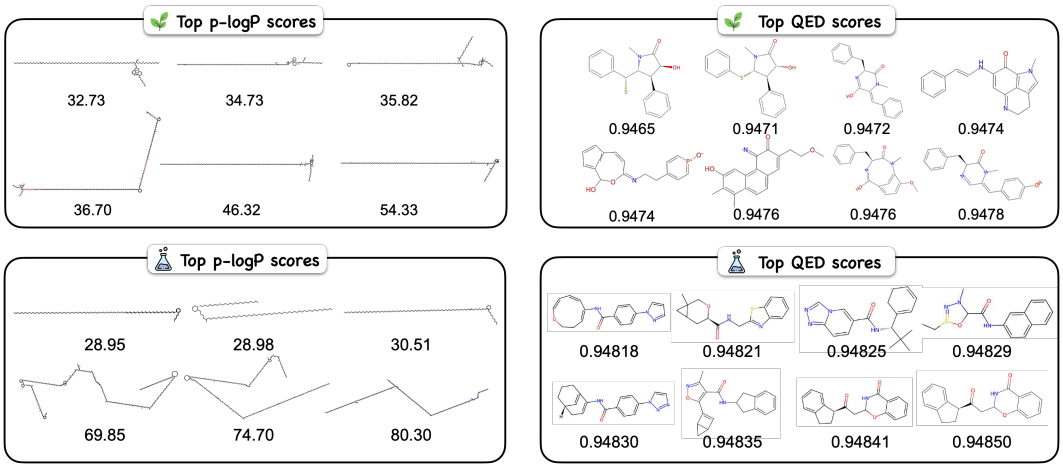
Appendix Figure 1: Comparison of visualizations of training and generated molecules



Appendix Figure 2: Property variations across different MolGen configurations.

properties, and that the chemical feedback paradigm enhances the property scores of the generated molecules, aligning them closer to the desired attributes.

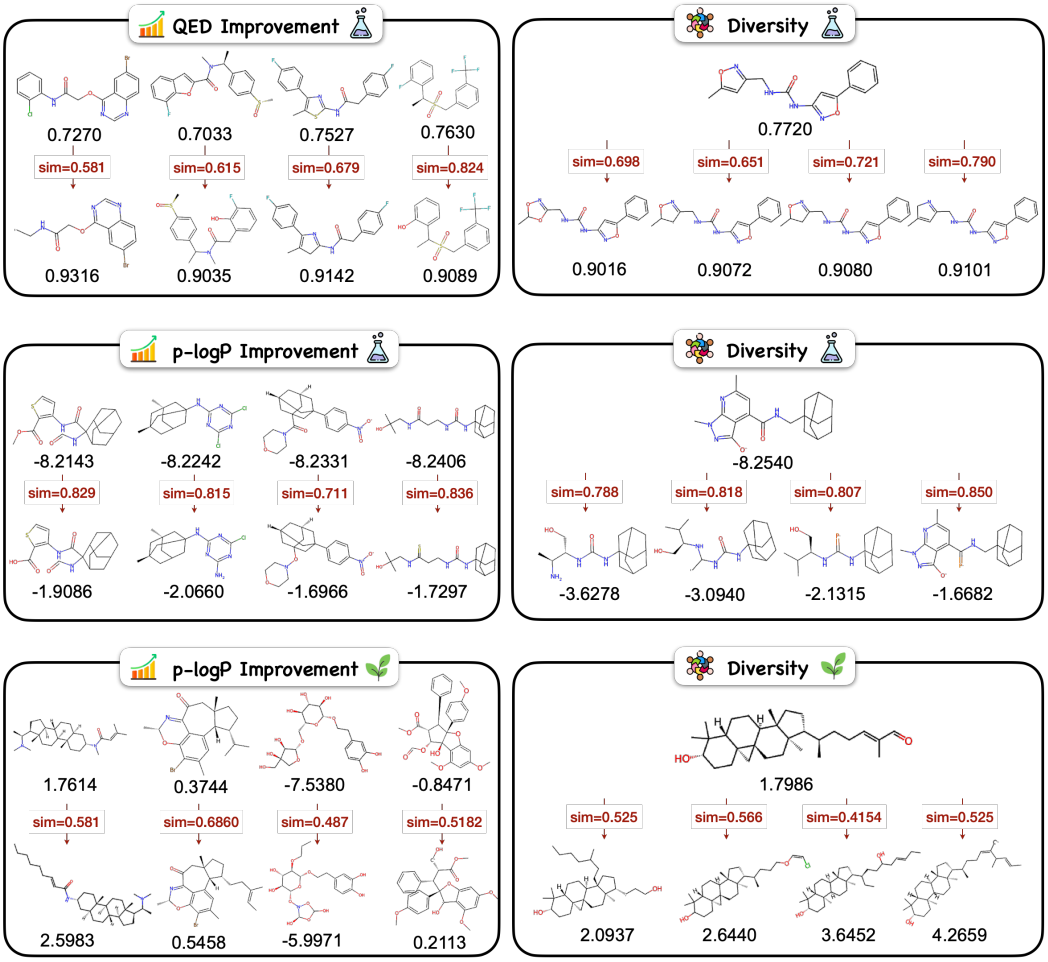
Moreover, we display the molecules with the highest scores from both data sources in Appendix Figure 3. From this, we can deduce that although ultra-long carbon chains are frequently observed in molecules with high p-logP scores, MolGen is capable of maintaining high scores while preserving structural diversity. Furthermore, MolGen is adept at generating molecules with high QED scores while retaining the structural features characteristic of various domains.



Appendix Figure 3: Samples of molecules with high penalized logP and QED scores generated by MolGen in natural products (left) and synthetic (right) domains.

H.3 CONSTRAINED MOLECULAR OPTIMIZATION

In Appendix Figure 4, we provide more illustrations of constrained optimization examples for both QED and p-logP scores. These examples further highlight the proficiency of MOLGEN in optimizing molecular properties while maintaining their fundamental structures. Moreover, MOLGEN demonstrates remarkable performance even in more challenging tasks of optimizing natural products, underlining its exceptional ability to navigate and explore a broader chemical space.

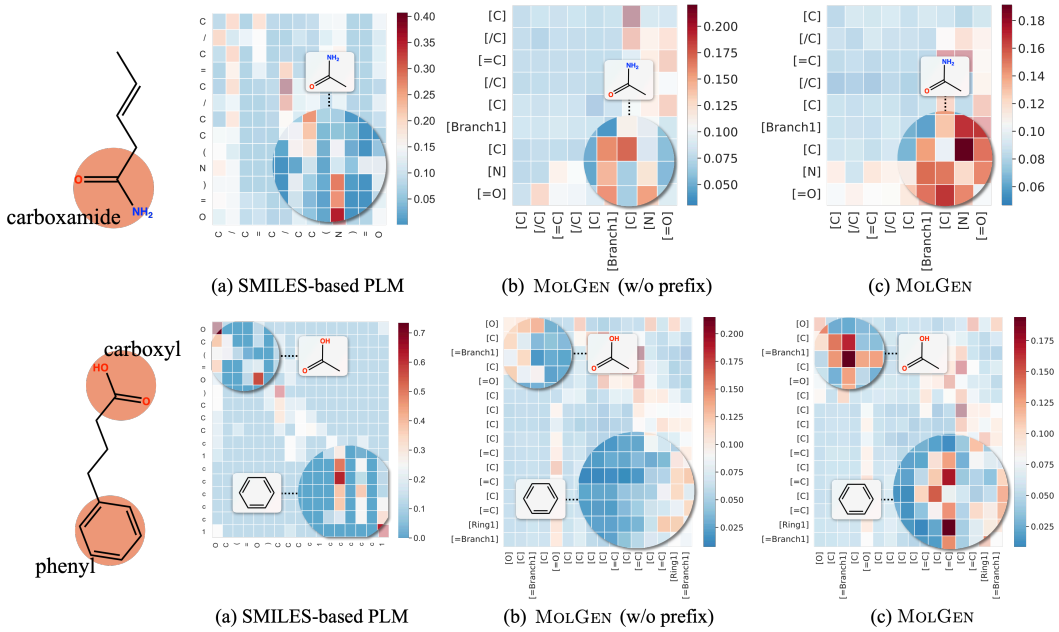


Appendix Figure 4: Further illustrations of constrained optimization using MOLGEN. Left: Examples of constrained molecular optimization for synthetic molecules (top two rows) and natural products (bottom row). Right: Optimized molecules derived from the same starting molecule, showcasing the diversity of outputs.

H.4 MORE MOLECULE SAMPLES OF ATTENTION VISUALIZATION

Lastly, we evaluate the representation capabilities of different PLMs by visualizing the attention weights of each token within an identical molecule, using the same setting as shown in Figure 8.

As depicted in Appendix Figure 5, MOLGEN allocates more attention to chemically significant functional groups like *carboxamide*, which demands high energy to break, and *carboxyl*, which exhibits strong polarity. In contrast, the attention mechanism in SMILES-based PLM tends to scatter across less relevant tokens, thereby diluting its focus. This demonstrates the advantage of the fine-grained vocabulary of MOLGEN, which can accurately identify and concentrate on pivotal structural components within molecules.



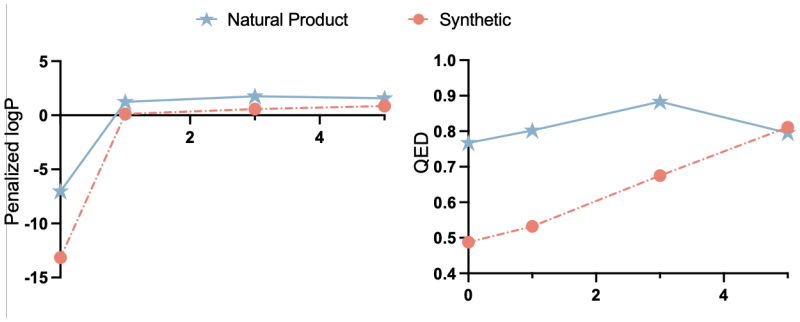
(a) SMILES-based PLM (b) MolGen (w/o prefix) (c) MolGen

(a) SMILES-based PLM (b) MolGen (w/o prefix) (c) MolGen

Appendix Figure 5: Visualization of the learned attention map.

Furthermore, when we compare MolGen with MolGen (w/o prefix), we observe that the former exhibits a more focused attention pattern. It is more adept at homing in on chemically meaningful substructures and avoids unnecessary dispersion of attention. This suggests that the incorporation of domain-agnostic molecular prefixes in MolGen effectively guides the model's attention towards regions of significance in the molecules, thus enhancing its ability to discern vital chemical structures.

H.5 More ablation studies



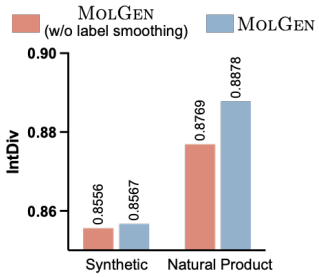
Appendix Figure 6: The impact of the hyperparameter α.

We then explore the impact of the hyperparameter α. As illustrated in Figure 6, an α value of 0 indicates no chemical feedback. When α is increased to 3, the model demonstrates a superior ability to optimize molecules compared to an α of 1. However, increasing α to 5 does not necessarily lead to better performance than at an α of 3. During actual operation, it is necessary to adjust parameters according to specific requirements. Based on experience, setting α to either 3 or 5 is recommended.

Additionally, we investigate the impact of label smoothing on the diversity of molecules generated by the model, employing IntDiv as the metric. IntDiv assesses the chemical diversity of generated molecules by calculating the average Tanimoto coefficient within the generated set.

As shown in Appendix Figure 7, the model with label smoothing does not overly rely on singular, frequently occurring patterns learned from the training data. Consequently, this enhances the diversity and creativity of the molecules generated.

To investigate the impact of prefix tuning on the model, we present the mean (and standard deviation) of penalized logP improvement for molecules generated compared to inputs under varying similarity



Appendix Figure 7: The impact of label smoothing.

Appendix Table 3: The impact of prefix tuning.

Model	Improvement	
	$\delta = 0.6$	$\delta = 0.4$
MolGen (w/o prefix)	11.63 (0.18)	10.23 (1.47)
MolGen	↑0.45 12.08 (0.82)	↑2.12 12.35 (1.21)

constraints, as detailed in Appendix Table 3. The incorporation of prefix tuning has resulted in enhanced molecule optimization performance. Taken together with Figure 8, the implementation of domain-agnostic prefix tuning not only enables the model to more effectively adapt to downstream tasks but also improves its interpretability.