Lang2Mol-Diff: A Diffusion-Based Generative Model for Language-to-Molecule Translation Leveraging SELFIES Molecular String Representation

Nguyen Doan Hieu Nguyen[†], Nhat Truong Pham[†], Duong Thanh Tran, Balachandran Manavalan[*]

Department of Integrative Biotechnology, Sungkyunkwan University, Republic of Korea
{ndhieunguyen, truongpham96, duongtt, bala2022}@skku.edu
[†]Equal contribution
[*]**Correspondence:** bala2022@skku.edu

Abstract

Generating *de novo* molecules from textual descriptions is challenging due to potential issues with molecule validity in SMILES representation and limitations of autoregressive models. This work introduces Lang2Mol-Diff, a diffusion-based language-to-molecule generative model using the SELFIES representation. Specifically, Lang2Mol-Diff leverages the strengths of two state-of-the-art molecular generative models: BioT5 and TGM-DLM. By employing BioT5 to tokenize the SELFIES representation, Lang2Mol-Diff addresses the validity issues associated with SMILES strings. Additionally, it incorporates a text diffusion mechanism from TGM-DLM to overcome the limitations of autoregressive models in this domain. To the best of our knowledge, this is the first study to leverage the diffusion mechanism for text-based *de novo* molecule generation using the SELFIES molecular string representation. Performance evaluation on the L+M-24 benchmark dataset shows that Lang2Mol-Diff outperforms all existing methods for molecule generation in terms of validity. Our code and pre-processed data are available at https://github.com/nhattruongpham/mol-lang-bridge/tree/lang2mol/.

1 Introduction

Molecules, the elementary constituents of all matter, play a pivotal role in dictating the properties and functionalities that govern our world. The immense scale of chemical space, estimated to encompass around 10^{33} molecules (Polishchuk et al., 2013), presents a significant challenge for traditional methods in finding new medicine, materials, and chemical processes. This has driven the exploration of artificial intelligence models for efficient molecule finding. A key advancement lies in the confluence of natural language and molecular representations such as SMILES (simplified molecular-input line-entry system) (Weininger, 1988) and SELFIES (SELF-referencIng Embedded Strings) (Krenn et al., 2020). These representations enable the seamless integration of natural language descriptions with corresponding molecular structures. By leveraging pre-trained language models and fine-tuning them on different benchmark datasets combining natural language and molecular string representations, researchers have successfully developed numerous downstream models capable of generating novel molecule structures based on textual descriptions outlining desired properties. Besides, the success of diffusion models in image generation has spurred their application to text generation, and more recently, to the domain of molecular representation.

In this research, we use the diffusion mechanism to address the limitations of autoregressive models, where errors from earlier predictions can propagate and magnify throughout the sequence and lead to inaccuracies, especially in long sequences. We also want to deal with validity issues in generating new molecules. The proposed method is a novel architecture that incorporates advancements in both the backbone model and the molecular representation. In essence, our key contributions are as follows:

- We employed SELFIES as the molecule presentation instead of SMILES for better validity in generating new molecules.
- This is the first study to leverage diffusion mechanism for text-based molecule generation using SELFIES molecular strings.

2 Related Work

2.1 Language Model-Based Approaches

The availability of molecular string representations like SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020) has transformed *de novo* molecule generation into a text-to-text task. Early approaches leveraged recurrent neural network

(RNN) architectures, such as those described in (Segler et al., 2018; Grisoni et al., 2020), achieving some success. However, the recent emergence of the text-to-text transfer transformer (T5) model (Raffel et al., 2020) as a powerful text-to-text model compared to RNN has led to the development of several successful models for this task, including MolT5 (Edwards et al., 2022a), Text+Chem T5 (Christofidellis et al., 2023), BioT5 (Pei et al., 2023), and BioT5+ (Pei et al., 2024). Additionally, transformer-based models like generative pre-trained transformer (GPT) (Brown et al., 2020) have been fine-tuned for this purpose, with MolXPT (Liu et al., 2023) serving as an example. Despite their advancements, autoregressive models exhibited limitations when dealing with long-term dependencies within the data. These models processed information one element at a time, leading to an inherent accumulation of errors. Additionally, autoregressive models were restricted by a fixed-size context window, limiting their ability to capture crucial relationships between elements that may reside far apart in the sequence. Consequently, these limitations could hinder the effectiveness of autoregressive models in tasks that necessitate understanding long-range dependencies within the data.

2.2 Diffusion Model-Based Approaches

The recent breakthroughs in image generation using diffusion models have paved the way for their exploration of text generation tasks. Diffusion-LM (Li et al., 2022) exemplified this exciting trend, demonstrating the potential of diffusion models for achieving controllable text generation. To address the limitations of autoregressive models, TGM-DLM (Gong et al., 2024) pioneered the application of Diffusion-LM in SMILES-based molecule generation. This work introduced the first diffusion language model for SMILES-guided molecule generation. However, due to its reliance on SMILES strings, TGM-DLM required a two-phase approach: an initial molecule generation phase followed by a correction phase. The necessity of the latter phase was questionable, as experimental results suggested that the correction phase did not lead to significant improvements in molecule validity.

3 Methodology

3.1 Overview of Lang2Mol-Diff

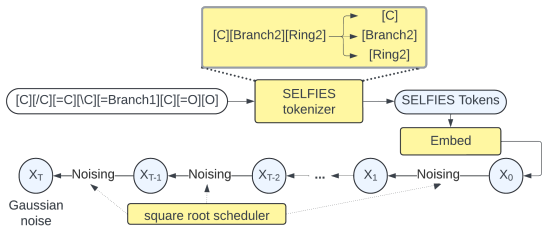
As discussed in Section 2, most existing language model-based methods suffered from limitations imposed by autoregressive nature. Therefore, we adopt a diffusion-based approach to address this challenge, enabling iterative and holistic content generation. To eliminate the need for a correction phase, a shortcoming identified in TGM-DLM (Gong et al., 2024) when using SMILES strings (Weininger, 1988), we leverage SELFIES strings (Krenn et al., 2020) for molecule representation, ensuring the inherent validity of generated molecules due to their superior ability to capture molecular structure. To achieve this, we exploit a pre-trained BioT5 (Pei et al., 2023) base model, which was fine-tuned for text-to-molecule tasks. This pre-trained model serves as the encoder for both SELFIES molecular strings and natural language text. We further incorporate embedding layers to construct a model that predicts molecule embeddings corresponding to Gaussian noise, drawing inspiration from the core concept of Diffusion-LM (Li et al., 2022). The overall architecture of our proposed approach is illustrated in Figure 1, which includes three main steps in the diffusion process: forward (Figure 1a), reverse (Figure 1b), and sampling (Figure 1c).

3.2 SELFIES Tokenizer

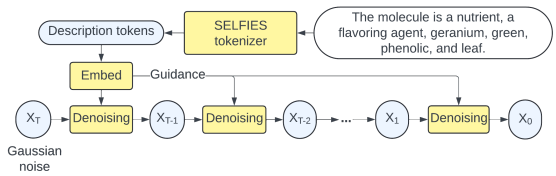
This work addresses the limitations of SMILES strings (Weininger, 1988) in terms of syntactic and semantic robustness, which can hinder the validity of molecules generated by deep learning models. For this reason, we opt for SELFIES representations (Krenn et al., 2020) due to their superior ability to capture molecular structure accurately. We leverage the tokenizer employed in BioT5 (Pei et al., 2023) to tokenize the text before passing it into the model. SELFIES string representation leverages brackets to encapsulate chemically meaningful atom groups, which are then individually tokenized as distinct SELFIES tokens. For instance, the SELFIES string *[C][Branch2][Ring2]* would be tokenized into *[C]*, *[Branch2]*, and *[Ring2]*.

3.3 Language Model-Based Encoder

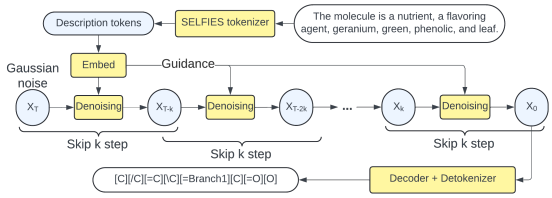
In contrast to TGM-DLM (Gong et al., 2024), which employed separate encoders for natural language texts and SMILES strings, namely SciBERT (Beltagy et al., 2019) for the first and uncased-



(a) Forward step in the diffusion process, noise is gradually added to the embed representation of molecular string (SELFIES) using a square root noise scheduler.



(b) Reverse step in the diffusion process, the model predicts the less noisy vector x_{t-1} from the noisier vector x_t.



(c) Sampling step in the diffusion process, we skip some steps to reduce the Markov chain so that the inferencing will be faster.

Figure 1: Illustration of Lang2Mol-Diff's diffusion process. X_0 is the embeddings of molecules tokenized by BioT5's tokenizer in SELFIES format. T is the number of diffusion steps. (a) Forward step, (b) Reverse step, (c) Sampling step.

BERT (Devlin et al., 2018) for the latter, this work adopts a more efficient approach. As discussed in Subsection 3.1, we leverage a pre-trained BioT5 model's encoder (Pei et al., 2023) for encoding both tokenized SELFIES (Krenn et al., 2020) strings and natural language text. This unified encoder architecture offers several advantages. First, it allows us to finetune the pre-trained parameters of the BioT5 model, focusing training efforts on the latter layers specific to our task. This not only reduces computational cost but also potentially mitigates overfitting. Additionally, a single encoder streamlines the model architecture, enhancing overall efficiency.

3.4 Diffusion Process

3.4.1 Forward Step

This represents the initial stage of the diffusion process, which is shown in Figure 1a. Given a molecular string, denoted as M, the SELFIES tokenizer is utilized to perform tokenization, resulting in a list of tokens represented as $\{m_0, m_1, m_2, ..., m_{n-1}\}$ where n is the number of tokens. Subsequently, the BioT5 encoder is applied to convert these tokens into a vector representation, denoted as $Emb(M) \in R^{d_m \times n}$. Here, d_m signifies the embedding dimension, while n represents the length of the sequence. The initial matrix for the forward process, denoted as x_0, is generated by sampling from a Gaussian distribution with a mean centered at $Emb(M)$: $x_0 \sim \mathcal{N}(Emb(M), \sigma_0 I)$.

With the initial embedding of the molecular string x_0, the forward step in the diffusion process is initiated. This step involves the gradual introduction of noise to the embedding through the utilization of a noise scheduler, which uses the $squareroot$ function in our approach. The process continues until the embedding transforms entirely into pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$, where T represents the number of diffusion steps. The diffusion step from x_{t-1} to x_t is defined:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I) \quad (1)$$

where $\beta_t \in [0, 1]$ controls the amount of noise added to x_t at time step t.

3.4.2 Reverse Step

The objective of this step is to reverse the forward process, specifically by predicting the original vector x_0 from the Gaussian noise x_T. This involves continuously predicting the less noisy vector x_{t-1} from the comparatively noisier vector x_t. The proposed model is trained to perform this denoising step by calculating the loss between the original molecule embedding x_0 and the vector $\hat{x}_0$ predicted from x_t. Moreover, the denoising process refines the embedding vector under the guidance of the embedded description $Emb(D)$ extracted using pre-trained BioT5 (Pei et al., 2023) to create a relationship between the description and the generated molecular string at the last step in the reverse process. The loss function used in the training phase of the model is defined as:

$$\mathcal{L}(M, D) = \mathop{\mathbb{E}}_{q(x_{0:T}|M)} \left[\sum_{t=1}^{T} ||f_\theta(x_t, t, D) - x_0||^2 - logp_\theta(M|x_0) \right] \quad (2)$$

where f_θ is the proposed model with parameters θ; x_t, t and D are the molecule embedding vector

at time t, the time embedding and the description embedding, respectively. $p_\theta(M|x_0)$ represents the rounding process, where the embedding matrix is reverted to the original molecular string.

3.4.3 Sampling Step

The aforementioned training methodology enables the construction of a model with the ability to generate a molecular string given a textual description. This is accomplished through an iterative denoising process involving T steps, wherein a complete Gaussian noise vector undergoes denoising to obtain an embedding representative of the molecular string. The denoising process is guided by the accompanying text description. Subsequently, the generated embedding is decoded by removing padding and start/end tokens, then rounding and transforming it into tokens, resulting in the final molecular string representation. This approach is also known as the Denoising Diffusion Probabilistic Model (DDPM) technique (Ho et al., 2020). However, it is important to note that this process involves a Markov chain, resulting in a significant computational time requirement to obtain the final result. To deal with this problem, instead of iterating through all steps in the diffusion process, we skip k steps in the sampling step. This means we predict the less noisy vector x_{t-k} based on the noisier vector x_t instead of x_{t-k+1}.

4 Experiments

4.1 Dataset

Our study employs the "split_train" split of the L+M-24 extra dataset (Edwards et al., 2022b, 2024) for training and the "split_valid" of the L+M-24 dataset (Edwards et al., 2022b, 2024) for evaluation. Each dataset comprises paired SMILES strings (Weininger, 1988) representing molecules and their corresponding descriptive captions. The training dataset was augmented from the original L+M-24 dataset by creating 4 additional captions for each existing molecule based on the randomly chosen available templates. Therefore, there are a lot of duplicated samples within the training split of the dataset. We first remove all of them to improve data efficiency. As a pre-processing step, we converted the SMILES strings to SELFIES representations (Krenn et al., 2020) using Python's *selfies* package [1]. It is important to note that a small portion of the molecules could not be successfully converted to a SELFIES format. These inconvertible molecules were excluded due to their negligible impact on the overall dataset size. On the other hand, the evaluation dataset is kept as original. A summary of the final training dataset is provided in Table 1. Some molecules that cannot be converted from SMILES to SELFIES are displayed in Table 2.

Split	Original	After removing duplicated samples	After converting to SELFIES
train	634,320	533,953	533,949
valid	33,696	33,696	33,696

Table 1: Summary of train and validation splits of L+M-24 dataset. From the original dataset, we filtered out duplicated samples, then we converted the SMILES strings to SELFIES strings to get the final dataset.

4.2 Implementation Details

We choose the maximum length of the tokens for the tokenizer to be 256. Consistent with the pre-trained BioT5 model (Pei et al., 2023), our approach leverages its established vocabulary. As detailed in the BioT5 paper, this vocabulary is segmented into distinct domains: molecules, proteins, and text. We specifically utilize BioT5's molecular vocabulary, encompassing 35,073 tokens, to ensure compatibility with the SELFIES string representation of molecules. This selection facilitates the efficient processing of molecule-related information within our model. The final Lang2Mol-Diff model architecture possesses approximately 218 million parameters. We opted for a diffusion schedule with T of 2,000 steps and a total training regime of 400,000 steps. The AdamW optimizer (Loshchilov and Hutter, 2017) was utilized with a learning rate of 0.5×10^{-5}. The training process ran for approximately 60 hours on a single NVIDIA GeForce RTX 4090 GPU with a batch size of 16. In the sampling step for inference, we set the step skipped k to be 10.

Because the L+M-24 dataset (Edwards et al., 2024) utilizes SMILES strings (Weininger, 1988) and established evaluation metrics are calculated based on this format, it is more precise for evaluation in this format of molecular string presentation. Moreover, to facilitate a fair comparison with existing research, we opt for SMILES over SELFIES representations (Krenn et al., 2020) for the evaluation phase. However, this decision is premised on the assumption that the SMILES molecules

[1] https://github.com/aspuru-guzik-group/selfies

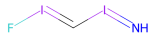
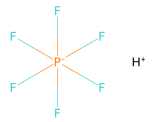
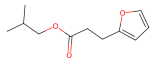
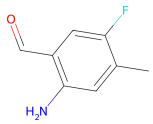
SMILES	SELFIES	Caption	Image
N=IC=IF	Error	The molecule is a nutrient.	
F[P-](F)(F)(F)(F)F.[H+]	Error	When heated to decomp, it emits highly toxic fumes of hydrogen fluoride and phosphoxides.	
CC(C)COC(=O)CCc1ccco1	[C][C][Branch1][C][C][C][O][C][=Branch1][C][=O][C][C][C][=C][C][=C][O][Ring1][Branch1]	The molecule is a nutrient.	
Cc1cc(N)c(C=O)cc1F	[C][C][=C][C][Branch1][C][N][=C][Branch1][Ring1][C][=O][C][=C][Ring1][=Branch2][F]	The molecule is a prmt5 inhibitor.	

Table 2: Some samples in the L+M-24 dataset after being converted from SMILES strings to SELFIES strings, 2 of them cannot be converted and are removed from training and evaluation splits of the final dataset.

within the dataset are all valid and unique representations, which means a molecule corresponds to a unique SMILES molecular string. To verify this assumption, we leverage the RDKit cheminformatics toolkit [2] to confirm that all SMILES strings are canonicalized.

4.3 Results and Discussion

The evaluation results of our model compared to other approaches are displayed in Table 3. Notably, our proposed method achieves BLEU, Levenshtein, MACCS FTS, RDK FTS, Morgan FTS, FCD, and Validity scores of 54.28, 55.87, 60.64, 33.21, 32.78, 38.09, and 100.00, respectively. More specifically, our proposed method outperforms all state-of-the-art methods regarding validity. Compared to Meditron-7B, our proposed method improves Morgan FTS by 15.96%. Regarding the Levenshtein metric, our proposed method is better than MolT5-Small by 0.47%. The result statistics of the proposed model cannot yet outperform the existing methods. It might be because it was not pre-trained on a large enough dataset as other models before being fine-tuned on the L+M-24 dataset. Besides, the model was trained with the default model configuration used for TGM-DLM (Gong et al., 2024), which may be incompatible and not fully optimized.

Moreover, we also compare the molecules generated by our proposed method with MolT5-Based and MolT5-Large models. Some generated molecules of MolT5 (Edwards et al., 2022a), which has been fine-tuned on L+M-24 dataset (Edwards et al., 2022b, 2024), and Lang2Mol-Diff are shown in Table 4. The empirical findings demonstrate that our proposed method exhibits a higher level of novelty compared to MolT5 in the generation of molecules. Although the input description is different, they share some important keywords, which leads to the identical generation of molecules using MolT5. On the other hand, Lang2Mol-Diff generates differently for two distinct input descriptions.

5 Conclusion

This work presents Lang2Mol-Diff, a novel diffusion-based language-to-molecule generative model that addresses the challenges of *de novo* molecule generation from textual descriptions. By leveraging the strengths of BioT5 for accurate tokenization of the SELFIES representation and incorporating a text diffusion mechanism inspired by

[2] https://github.com/rdkit/rdkit

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Validity↑
Ground Truth	100.00	100.00	0.00	100.00	100.00	100.00	0.0	100.00
MolT5-Small	56.56	0.00	56.34	64.22	58.10	37.44	NaN	80.52
MolT5-Base	68.38	0.00	44.79	76.03	65.23	47.46	NaN	100.00
MolT5-Large	56.42	0.00	55.40	75.70	65.01	39.51	17.52	99.44
Meditron-7B	69.40	0.00	46.49	77.16	69.34	16.82	2.46	99.63
Ours	54.28	0.00	55.87	60.64	33.21	32.78	38.09	100.00

Table 3: Text-guided molecule generation results on L+M-24 validation split. Data is taken from the report on the L+M-24 dataset (Edwards et al., 2024).

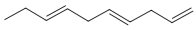
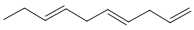
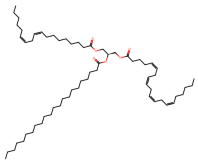
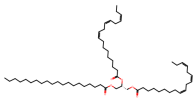
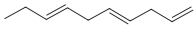
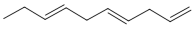
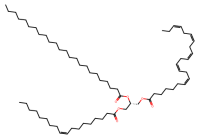
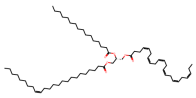
Input	MolT5-Base	MolT5-Large	Ours	Ground truth
The molecule is a nutrient and fat storage, and it impacts pancreatitis. The molecule is a thyroxine treatment that impacts cardiovascular disease, metabolic syndrome, and atherosclerosis.				
It impacts pancreatitis, cardiovascular disease, and metabolic syndrome. The molecule is a nutrient and a fat storage, it impacts atherosclerosis, and is thyroxine treatment.				

Table 4: Comparative visualization of *de novo* generated molecules across models. We use the Hugging Face's Inference API to collect the outputs of MolT5-Base and MolT5-Large.

TGM-DLM, Lang2Mol-Diff overcomes the limitations of SMILES-based approaches and autoregressive models. Extensive evaluation on the benchmark dataset confirms Lang2Mol-Diff's superior performance in generating valid molecules compared to the current state-of-the-art methods. This achievement paves the way for more reliable and robust methods for *de novo* molecule generation based on textual descriptions.

Our proposed model presents opportunities for future research and improvement. One promising direction for enhancement involves pre-training the model on a larger dataset, which would enable it to learn more meaningful representations and enhance its generalization capabilities. Furthermore, exploring alternative configurations such as adjusting the model's architecture and fine-tuning hyperparameters holds potential for optimizing performance and overcoming existing limitations. Pursuing these avenues is expected to refine the model and further optimize its ability to generate new molecules with improved outcomes.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (RS-2024-00344752). This research was supported by the Department of Integrative Biotechnology, Sungkyunkwan University (SKKU) and the BK21 FOUR Project. This work was supported by the Korea Bio Data Station (K-BDS) with computing resources including technical support.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *Inter-*

national Conference on Machine Learning, pages 6140–6157. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022a. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022b. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+M-24: Building a dataset for Language + Molecules @ ACL 2024. *arXiv preprint arXiv:2403.00791*.

Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024. Text-guided molecule generation with diffusion language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):109–117.

Francesca Grisoni, Michael Moret, Robin Lingwood, and Gisbert Schneider. 2020. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3):1175–1183.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*.

Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. 2013. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27:675–679.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.