Enhancing Cross Text-Molecule Learning by Self-Augmentation

Yinuo Jiang, Xiang Zhuang, Keyan Ding,
Qiang Zhang*, Huajun Chen*
College of Computer Science and Technology, Zhejiang University
ZJU-Hangzhou Global Scientific and Technological Innovation Center
{yinuoj,zhuangxiang,dingkeyan,qiang.zhang.cs,huajunsir}@zju.edu.cn

Abstract

The development of Large Language Models (LLMs) has greatly advanced the field of drug discovery, with the belief that natural language can enhance human control over molecule design. However, the scarcity of high-quality labeled data remains a challenge for cross text-molecule learning. Existing datasets are limited due to the difficulty of collecting precise molecule-description pairs. Although recent efforts have utilized pseudo data generated by LLMs for augmentation, the lack of specialized chemistry knowledge of LLMs and the absence of an effective high-quality data selector may introduce noise into the annotations, compromising the models' robustness. To address these challenges, this paper introduces a novel framework that interweaves model fine-tuning and data augmentation to overcome the scarcity of high-quality data. The proposed approach involves an iterative procedure where the model plays dual roles in annotating unlabeled data and sampling a subset of high-quality data until convergence is achieved, enhancing the model's understanding and adaptability. Additionally, a new dataset called SAPubChem-41 is presented, which comprises meticulously curated high-quality parallel molecule-description pairs designed specifically for fine-tuning purposes. This research provides an important contribution to the field by addressing the need for high-quality datasets and presenting an effective framework for cross text-molecule learning.

1 Introduction

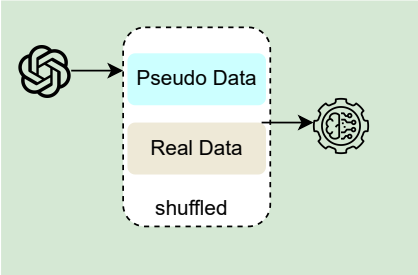
The emergence of Large Language Models (LLMs) has significantly propelled the development of drug discovery (Zhang et al., 2024). Recent progress in language models shed light on drug discovery, with the vision that humans can possess a higher-level control over molecule design facilitated by natural language. Edwards et al. (2022a) introduce two new tasks: molecule captioning (Mol2Cap) and text-based molecule generation (Cap2Mol), and present MolT5 (Edwards et al., 2022b) based on the T5 (Raffel et al., 2020) architecture to translate between molecule and text. Subsequently, diverse endeavors have been undertaken to tackle these challenges. Text+Chem T5 (Christofidellis et al., 2023) and BioT5 (Pei et al., 2024) are also T5-like models, which incorporate multi-task and multi-domain pretraining process and enable the bidirectional generation between different modalities in a single model. Some other methods employ the Generative Pretrained Transformer (GPT) architecture. MolXPT (Liu et al., 2023) proposes to wrap molecules in the sentences to make the pertaining corpus. More recently, MolCA(Liu et al., 2024) considers the 2D graph information for molecules and designs a projector to connect molecular content to language models. These studies mark the research efforts in the realm of cross text-molecule learning.

Despite rapid development in cross text-molecule modalities, there is still demand for high-quality labeled data in this field. Such data of superior quality is indispensable to capacitate models for downstream tasks. However, the available datasets relevant to the tasks are limited, a consequence of the arduous nature of collecting precise and enlightening parallel molecule-description pairs. This scarcity of annotated data hinders the model's ability to learn between the modalities of the molecule and the text. As shown in Figure 1, recent efforts by Chen et al. (2023) address the scarcity challenge by utilizing pseudo data generated by LLMs as augmentation. The problem is that the absence of specialized knowledge in chemistry may introduce noise into the data annotated by general Large Language Models, thereby compromising the robustness of models fine-tuned on such datasets. In spite of these commendable advance-

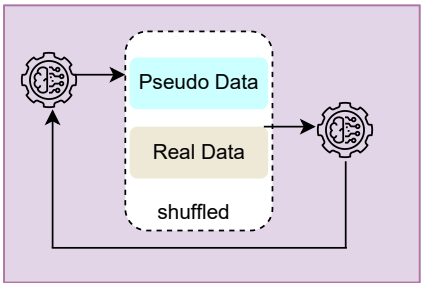
*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 9551–9565

ments, the accessibility of high-quality datasets for fine-tuning remains notably deficient.



(a) Conventional Data Augmentation Method



(b) Our Self-Augmentation Method

Figure 1: A comparison between a conventional data augmentation method referenced in Chen et al. (2023) and our iterative self-augmentation techniques is drawn herein. The original technique leverages Label Language Models (LLMs) to annotate unlabeled data; contrastingly, our proposed self-augmentation strategies use the model itself for annotating unlabeled data in tandem with an iterative training approach.

To address these challenges, this paper introduces a novel framework that interweaves model fine-tuning and data augmentation (See Fig. 1, aiming to rectify the high-quality data scarcity observed in the realm of cross text-molecule learning. Our proposed approach integrates the model into the process by assigning it dual roles: annotating unlabelled data and sampling a subset of high-quality data. This iterative procedure continues until convergence is achieved, thereby refining the model's understanding and adaptability. Additionally, we present a new, larger dataset SAPubChem-41. This meticulously curated dataset comprises a wealth of high-quality parallel molecule-description pairs meticulously designed for the specific purpose of fine-tuning. The model optimized using SAPubChem-41 has been evaluated across a total of 13 molecule-text benchmarks, exhibiting superior performance in comparison to the model optimized using the original dataset, ChEBI-20. Our key contributions can be summarized as:

- We present an innovative framework designed to augment cross-text-molecule learning through a self-augmentation strategy. This system leverages the model itself to annotate unlabelled data, choosing only samples that meet a high-quality standard.

- We introduce SAPubChem-41, a novel dataset consisting of both real data and high-quality pseudo-labeled data, augmented by the model itself. This makes a notable contribution to the available datasets in the field of cross-molecule-text modality.

- We show that within our iterative setting, self-augmented data significantly improves the performance of the model in downstream tasks with each epoch. Furthermore, these results tend to converge as the number of epochs increase. This process thereby successfully validates the efficacy of self-augmented data in enhancing model performance.

2 Background

2.1 Cross Text-Molecule Learning

Edwards et al. (2021) introduce a new task Text2Mol, which uses descriptions as search queries to retrieve the target molecules. Edwards et al. (2022a) first addresses the problem of cross-domain generation by linking natural language and chemistry, tackling tasks such as text-conditional de novo molecule generation and molecule captioning. Li et al. (2023a) combines retrieval-based prompt paradigm with LLMs like ChatGPT to achieve translation between molecule language and natural language. Christofidellis et al. (2023) propose the first multi-domain, multi-task language model that can solve a wide range of tasks in both the chemical and natural language domains. Liu et al. (2023) propose MolXPT, a GPT-based model pre-trained on molecule SMILES, biomedical text, and wrapped text. BioT5 (Pei et al., 2024) further exploits SELFIES for 100% robust molecular representations and discriminate structured knowledge from unstructured knowledge. To capitalize on these insights, we adopt an iterative framework to interweave model fitting and data augmentation, to mitigate data scarcity and further enhance the alignment between molecule and text representations.

2.2 Data Augmentation for Cross Text-Molecule Learning

Data augmentation is a regularization strategy utilized to enhance model performance by diversifying the available data through various techniques (Hernández-García and König, 2018). In the field of natural language processing (NLP), data augmentation has been widely employed to address data scarcity issues, leading to the proposal of various augmentation methods (Feng et al., 2021; Li et al., 2022). Textual data augmentation can be accomplished through simple rules, such as synonym replacement and word order modification (Zhang et al., 2015). Furthermore, the development of deep learning models has facilitated the generation of new text to augment the existing data (Szegedy et al., 2016; Wu et al., 2019; Anaby-Tavor et al., 2020; Yoo et al., 2021; Zhou et al., 2021; Li et al., 2023b; Dai et al., 2023). Acquiring high-quality annotated molecule-text datasets is prohibitively expensive, for many descriptions of corresponding molecules are scattered and inaccurate, leading to limited fine-tuning datasets for cross text-molecule learning.More data augmentation techniques are desired to further enhance cross text-molecule learning.Chen et al. (2023) pioneers to exploit LLMs' annotated pseudo label for domain adaption and data augmentation.However, while pseudo label as domain adaptation shows impressive performance compared to existing methods, pseudo label as data augmentation still struggles against label noise interfering model's performance. To address the problem and further improve the quality of augmented dataset, our work introduces a pioneering iterative framework that interweaves model fine-tuning and data augmentation.

3 Method

In this section, we describe a computational framework for jointly optimizing model performance and dataset quality.

3.1 Overall Framework

The literature demonstrates a substantial body of research focused on enhancing both model performance and dataset quality concurrently, as depicted in Figure 2. If we assume the existence of a subjective testing environment capable of gathering reliable molecule-caption pairs, then the core challenge becomes how to effectively sample a subset A from a large-scale pseudo-labeled dataset D that exhibits high confidence. In line with Li et al. (2023b), our approach adopts an iterative strategy, combining model fine-tuning and data augmentation operations defined below.

1. Model fine-tuning: Given a data distribution, find an approximate risk minimizer.
2. Data augmentation: Given a model, augment data and sample a new data distribution.
3. Data combination: Combine a set of distributions into a single distribution.

At the beginning of each iteration, a new model is finetuned on the latest dataset. This newly trained model is subsequently employed to annotate a set of unlabeled data, thereby generating a pseudo-labeled dataset. The sampler is then entrusted with the responsibility of discerning samples distinguished by their high quality from this pseudo-labeled dataset. The discerned insights are then harnessed to augment and refine the current dataset, thereby facilitating a continuous and iterative improvement process.

3.2 Self-Annotating

We gathered 200,000 unannotated SMILES of molecules from PubChem (Kim et al., 2023) and conducted deduplication to mitigate the risk of data leakage in the validation dataset. For each unlabeled SMILES, we executed inference on the model itself to produce a candidate caption, thereby obtaining the pseudo-labeled dataset.

ChatGPT Annotating vs Specialized Model Annotating Despite the remarkable performance of Large Language Models (LLMs) such as ChatGPT in various cross-modal tasks, they have not succeeded in surpassing specialized models in the molecule captioning task (Li et al., 2023a; Christofidellis et al., 2023). This suggests that specialized models serve as better annotators for unlabeled SMILES strings. Furthermore, as depicted in Figure 3(a), by leveraging the Text2Mol score (Edwards et al., 2022a) as a fundamental metric of data quality, which measures the similarity between SMILES strings and their captions, we evaluated the distribution of the real training set, MolReGPT annotated data, and self-annotated data. It is evident that the self-annotated data yields higher-quality data, closely resembling real data more than the MolReGPT annotated data.

Iterative Annotating We employ an iterative training approach to enhance the proficiency of annotators and produce captions of superior quality.

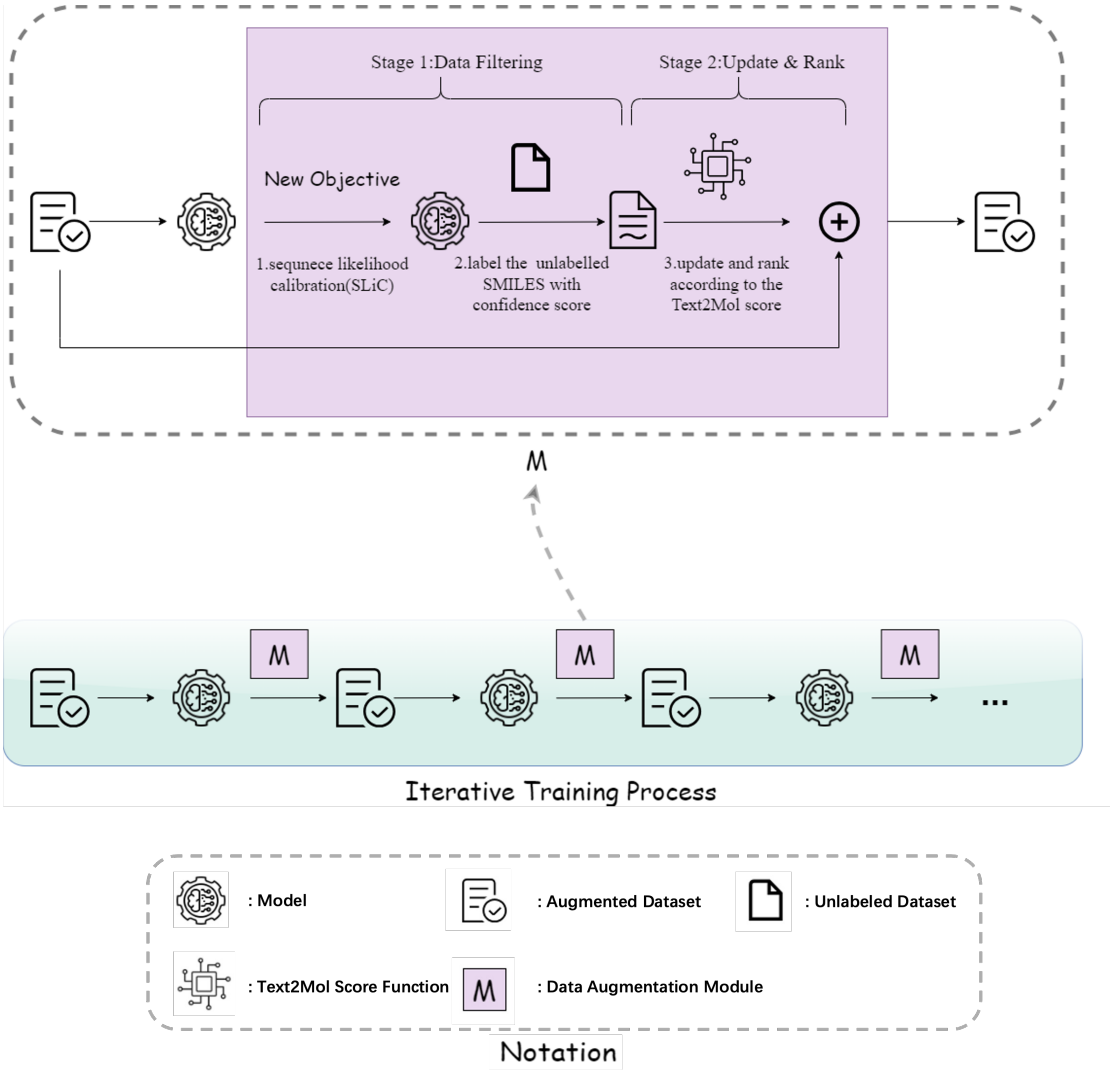
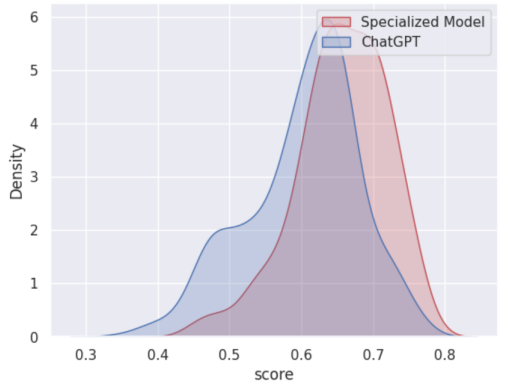


Figure 2: Overall framework. We interweave model fine-tuning and data augmentation iteratively. A core component of our framework, the Data Augmentation Module, is compartmentalised into two key elements: Self-Annotating and Self-Sampling. These elements underline the crux of our methodology, wherein the model autonomously performs the augmentation, subsequently enhancing its own performance.

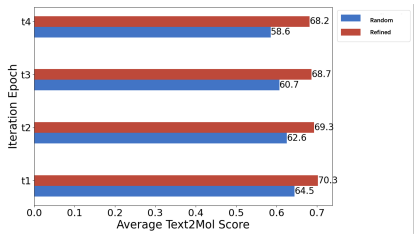
As depicted in Figure 3, during each iteration, the quality of the molecule-caption pairs is enhanced through the implementation of the iterative strategy.

3.3 Self-Sampling

A discerning subset sampling strategy is anticipated to effectively filter pseudo labels, thereby ensuring the attainment of high-quality labels for training the model. This refined approach aims to enhance the overall reliability of the model by mitigating the impact of unreliable or inaccurate labels in the training process. In contrast to prior work (Li et al., 2023b), wherein fine-tuning an external score model was employed to select high-quality pairs, we have found this approach to be both time-consuming and unreliable when applied to the domain of cross text-molecule learning. Notably, the susceptibility of the model to overfitting incorrect pseudo-labels is a concern, as detailed in section 4.4. To tackle this dilemma, we introduce a dual selection mechanism that not only proves to be more computationally efficient but also enhances reliability in ensuring the high quality of our augmented dataset. This mechanism involves two key steps. Firstly, leveraging the generation likelihood as a confidence score to filter out unreliable molecule-caption pairs. Secondly, updating and ranking the remaining molecule-caption pairs based on the Text2Mol score (Edwards et al., 2022a) This refined approach streamlines the sampling process, offering increased efficiency and



(a) Comparison of Data Quality Between LLMs Annotated Label and Self-Annotated Label



(b) Comparison of Data Quality in Relation to the Adoption or Non-Adoption of Iterative Strategy

Figure 3: In a manner akin to Chen et al. (2023) in their study, we employ the methodology proposed by Edwards et al. (2022) to evaluate the correlation between molecule-description pairs, utilizing it as a measure of data quality. The distributions are illustrated via Kernel Distribution Estimation. A noteworthy aspect to consider here is that a higher Text2Mol score typically implies a closer resemblance between the molecule and its description. In addition, the term "Density" in (a) pertains to the concentration of data within a specific region.

heightened confidence in the identification of high-quality pairs for dataset augmentation.

Data Filtering First, we employ the model itself to filter out a set of low-quality pairs. Specifically, we utilize the generation likelihood as a measure of the quality of the generated sequence. However, previous studies (Liu and Liu, 2021; Liu et al., 2022) have underscored that the correlation between sequence probability and its quality for Maximum Likelihood Estimation (MLE) trained models can be low, due to the presence of deterministic (one-point) target distribution issues. Furthermore, sequence likelihood estimation becomes noisier when the decoded sequences of models deviate from the exposed training data distribution, thereby exacerbating the problem of exposure bias (Ranzato et al., 2016). To address these intricate challenges, we introduce the Sequence Likelihood Calibration (SLiC) stage. This stage, depicted in Figure 1, serves to mitigate the aforementioned issues and diminish the gap between sequence likelihood and its associated quality. Following the methodology of Zhao et al. (2022), we conduct further fine-tuning of the model, introducing a new objective that encompasses the following loss functions:

$$\mathrm{L}_{\mathrm{rank}}^{\mathrm{cal}} = \max\Big(0, \beta - \log P_\theta(\hat{\mathbf{y}}_+ \mid \mathbf{x}) + \log P_\theta(\hat{\mathbf{y}}_- \mid \mathbf{x})\Big),$$

$$\mathbf{L}_{\mathrm{ce}}^{\mathrm{reg}} = \sum_t -\log P_\theta\left(\bar{y}_t \mid \overline{\mathbf{y}}_{t-1}, \mathbf{x}\right),$$

$$\mathrm{L}_{\mathrm{kl}}^{\mathrm{reg}} = \textstyle\sum_t P_\theta\left(\bar{y}_t \mid \overline{\mathbf{y}}_{t-1}, \mathbf{x}\right) \log \frac{P_\theta\left(\bar{y}_t|\overline{\mathbf{y}}_{t-1},\mathbf{x}\right)}{P_{\theta_{ft}}\left(\bar{y}_t|\overline{\mathbf{y}}_{t-1},\mathbf{x}\right)}.$$

Given the context $\mathbf{x}$, target $\overline{\mathbf{y}}$,and positive and negative candidates pairs $\hat{\mathbf{y}}_+, \hat{\mathbf{y}}_-$, $P_\theta(\mathbf{y} \mid \mathbf{x})$ denotes the generation sequence likelihood. **Rank loss** L_{rank}^{cal} optimizes the ranking order of positive and negative candidate pairs. **Cross entropy loss** L_{ce}^{reg} is the standard fine-tuning MLE objective . **KL divergence loss** L_{kl}^{reg} directly minimizes the probability distribution distance between the calibrated model and the fine-tuned model at each token on the observed target sequence.

Following the SLiC stage, we regard the generation likelihood as one of our metrics for assessing the quality of pseudo labels. In our experiments, we discard pairs with log-likelihood equal to negative infinity and retain the remaining pairs as high-quality pairs.

Update and Rank Secondly, for those high-quality pairs, we further refine the data based on the similarity score$\mathcal{F}_{\text{text2mol}}$ proposed by Edwards et al. (2022a). At iteration t, we update the caption if $\mathcal{F}_{\text{text2mol}}(m, c_t) > \mathcal{F}_{\text{text2mol}}(m, c_{t-1})$,then we rank the updated dataset according to $\mathcal{F}_{\text{text2mol}}$,selecting the top $t \times 2k$ pairs to form the new Augmented Dataset A_t.

4 Experiments

In order to validate the efficacy of utilizing our framework, we have undertaken a series of comprehensive experiments.

4.1 Experimental Setup

Dataset We employ the ChEBI-20 (Edwards et al., 2021) to serve as our original dataset. Additionally, we gather 200,000 unannotated molecules

Info	SAPubChem-41	ChEBI-20
Train	34,407	26,407
Validation	3,301	3,301
Test	3,300	3,300
$\mathcal{L}_{\text{SMILES}}$	69.96	81.56
$\mathcal{L}_{\text{Description}}$	41.59	52.88

Table 1: Comparison of our SAPubChem-41 dataset with the original dataset ChEBI-20. L_{SMILES} denotes the average length of SMILES while $L_{\text{Description}}$ denotes the average word count per description.

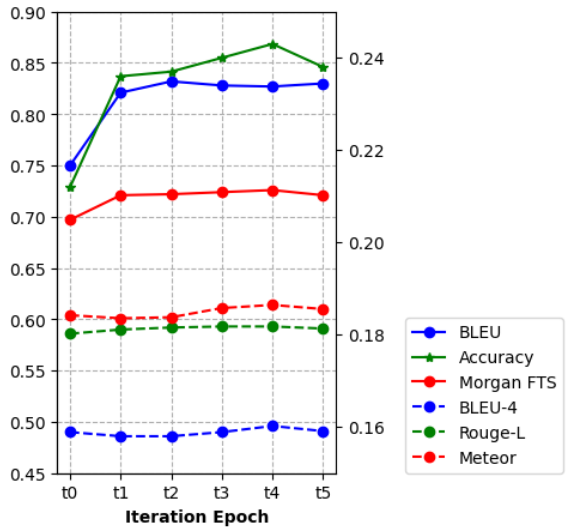


Figure 4: The results of the iteration procedure. The dashed lines denote the metrics associated with the molecule captioning task, while the solid lines signify those related to the molecule generation task.

from the PubChem database (Kim et al., 2023). Subsequently, we conduct a rigorous filtering procedure to exclude any molecules present in downstream datasets, ensuring the elimination of potential overlaps between the newly acquired molecules and those already present in the ChEBI-20 dataset. For our experimental dataset augmentation, we set the iteration step to 2000. At each iteration stage, the dataset size expands by 2000 from the preceding iteration. A detailed comparison between our final augmented dataset, SAPubChem-41, and the original dataset, ChEBI-20, is presented in Table 1.

Base Model To prove the effectiveness of our framework, we use the Text+Chem T5 (Christofidellis et al., 2023) as the base model for simulating the iterative process.

Metrics Following the previous studies (Edwards et al., 2022a; Li et al., 2023a; Christofidellis et al., 2023; Chen et al., 2023), we evaluate the results with following metrics:

- **Molecular Captioning**: BLEU-2 and BLEU-4(Papineni et al., 2002) are metrics used to evaluate the quality of machine-generated text by comparing it to a reference text, with a higher score on the BLEU metric indicating a higher level of similarity between the generated text and reference text. ROUGE-1, ROUGE-2, and ROUGE-L(Lin, 2004) is similar to BLEU, while computing the recall-overlap of unigrams, bigrams, and longest common subsequences between the generated and reference texts. METEOR (Banerjee and Lavie, 2005) is a metric that uses a combination of unigram precision, recall, and a synonym-matching component to evaluate the generated text against the reference text, which is designed to be more sensitive to fluency, meaning, and structure than BLEU.

- **Text-based Molecule Generation**: BLEU and the Exact Match scores are calculated as basic assessments. Molecule-specified metrics including Levenshtein distance(Levenshtein et al., 1966), validity(Edwards et al., 2022a), and three molecule fingerprints scores - MACCS FTS(Durant et al., 2002), RDK FTS(Schneider et al., 2015), and Morgan FTS(Rogers and Hahn, 2010) are calculated to provide valuable insights into the quality, validity, and structural characteristics of the generated molecules.

Implementation We conducted our experiments on Ubuntu 22.04 using RTX 4090(24GB) * 4 with CUDA 11.7.1. Our primary dependencies are Python 3.11.4, PyTorch 2.0.1, Transformers 4.31.0, and Numpy 1.24.3. We initialize the three pre-trained models using public checkpoints. For fine-tuning, we adopted the configuration in MolT5 (Edwards et al., 2022b), with a learning rate of 1e-3, 50,000 fine-tuning steps, weight decay of 0.1, batch size of 32, random seed of 42, and 1000 warm-up steps. For evaluation, we use a greedy search with a maximum generation length of 512 during generation. The evaluation metrics script is derived from MolT5 (Edwards et al., 2022b). For other hyper-parameters, we relied on the default settings of the T5ForConditionalGeneration class in Huggingface.

Dataset	Text+Chem T5			Ada-T5			MolT5		
	BL	RG	MET	BL	RG	MET	BL	RG	MET
ChEBI-20	0.490	0.498	0.604	0.316	0.369	0.464	0.457	**0.485**	0.569
SAPubChem-41	**0.496**	**0.504**	**0.614**	**0.350**	**0.410**	**0.493**	**0.464**	0.484	**0.572**

Table 2: Results of different models for molecular captioning on ChEBI-20, SAPubChem-41 datasets. The **best** scores are in bold. **BL**:BLEU-4, **RG**:ROUGE-2, **MET**:METEOR

Dataset	Text+Chem T5			Ada-T5			MolT5		
	Morgan	Acc	Val	Morgan	Acc	Val	Morgan	Acc	Val
ChEBI-20	0.697	0.212	0.792	0.672	0.182	**0.886**	0.529	0.081	0.772
SAPubChem-41	**0.726**	**0.243**	**0.937**	**0.699**	**0.213**	0.854	0.582	**0.093**	**0.788**

Table 3: Results of different models for molecular generation on ChEBI-20, SAPubChem-41 datasets. The **best** scores are in bold. **Morgan**:Morgan FTS, **Acc**:Accuracy, **Val**:Validity

4.2 Performance Comparison of Molecule-Caption Translation

Molecular Captioning Table 2 presents the results of the molecule captioning task, utilizing the original ChEBI-20 dataset and the augmented SAPubChem-41 dataset to optimize three distinct pre-trained models. Across all measurement metrics, it is observable that the performance of all models using our augmented SAPubChem-41 dataset is generally superior to that utilizing the original ChEBI-20 dataset.

Text-based Molecule Generation Table 3 presents the results of the text-based molecule generation task, utilizing the original ChEBI-20 dataset and the augmented SAPubChem-41 dataset to optimize three distinct pre-trained models.Notably, the performance of all models that utilize our augmented SAPubChem-41 dataset generally surpasses those that use the original ChEBI-20 dataset. The accuracy of the generated molecule, as well as its similarity to the ground truth molecule, have been significantly improved following the self-augmentation of the dataset. This evidences substantial enhancements in performance when integrated with our self-rewarding framework.

4.3 Iterative Procedure

We utilize our framework on the Text+Chem T5, taking the ChEBI-20 as the original dataset. This approach facilitates an augmented dataset while simultaneously enhancing the model's performance. It is critical to note that our intention in these experiments is not necessarily to achieve a state-of-the-art model. Rather, our prime objective is to examine the efficacy of our framework, which is designed to optimize both the model and the pseudo-labeled dataset concurrently. Figure 4 illustrates the iterative process in which dashed lines represent the molecule captioning task and solid lines signify the text-based molecule generation task. We report the complete results in Table 4 and 5. There are a few observations: First, it is noticeable that for both tasks the outcome of iteration 5 mostly falls short of that of iteration 4 in all metrics. This suggests that the training process reaches convergence on iteration 4 for both tasks. Consequently, we adopt the augmented dataset obtained at the fourth iteration as our conclusive dataset. Second, for the text-based molecule generation task, there is a consistent upward trend across all pertinent metrics. This suggests that an increased quantity of high-quality pseudo-labeled data contributes to the optimization of the model in the molecule generation process. This trend underscores the necessity of sufficient high-quality data in cross text-molecule learning. Lastly, when it comes to the molecule captioning task, the general trend in relevant metrics indicates a rise. Initially, there is a minor decrease in the BLEU and Meteor metrics, which is counterweighed by an enhancement in Rouge. As the volume of high-quality pseudo-labeled data increases, all metrics demonstrate improvement. This trend suggests that the quantity of high-quality pseudo-labeled data is also a significant factor.

4.4 Ablation Study

We perform further ablation studies to validate the effectiveness of the key component in our framework, i.e., the self-sampling strategy.

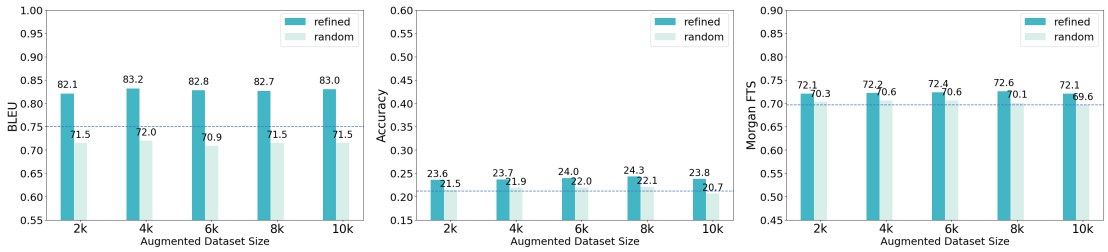
To demonstrate the significance of pseudo-

Iteration	BLEU↑	Accuracy↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
t0	0.750	0.212	27.39	0.874	0.767	0.697	0.792
t1	0.821	0.236	19.31	0.885	0.788	0.721	0.931
t2	**0.832**	0.237	18.69	0.885	0.788	0.722	**0.938**
t3	0.828	0.240	19.28	0.886	0.788	0.724	0.934
t4	0.827	**0.243**	**18.59**	**0.887**	**0.792**	**0.726**	0.937
t5	0.830	0.238	19.03	0.883	0.786	0.721	0.937

Table 4: The iterative results of molecule generation task. The best scores are in bold.

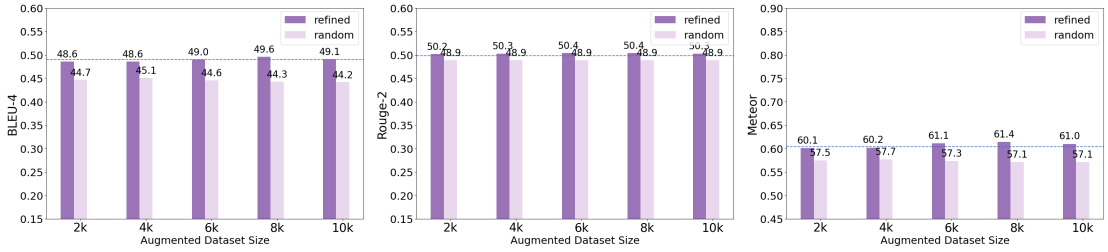
Iteration	BLEU-2↑	BLEU-4↑	Rouge-1↑	Rouge-2↑	Rouge-L↑	Meteor↑
t0	0.580	0.490	0.647	0.498	0.586	0.604
t1	0.574	0.486	0.651	0.502	0.590	0.601
t2	0.580	0.486	**0.653**	0.503	0.592	0.602
t3	0.583	0.490	**0.653**	**0.504**	**0.593**	0.611
t4	**0.584**	**0.496**	**0.653**	**0.504**	**0.593**	**0.614**
t5	0.583	0.491	0.652	0.503	0.591	0.610

Table 5: The iterative results of molecule captioning task. The best scores are in bold.



(a) BLEU (b) Accuracy (c) Morgan FTS

Figure 5: Results of molecule generation task using different sampling strategy. The blue line marks the result of model optimization on the original dataset.



(a) BLEU-4 (b) Rouge-2 (c) Meteor

Figure 6: Results of molecule captioning task using different sampling strategy. The blue line marks the result of model optimization on the original dataset.

labeled data selection, we conducted a comparative analysis of the performance of all intermediate models. We compared the use of randomly-selected data of the same size with our refined data. As depicted in Figure 5 and 6, our refined data consistently outperformed the randomly-selected data across all metrics at each iteration for both tasks. Notably, when optimizing the model on randomly-selected data, the results were generally inferior to the original model on most metrics. This indicates that the model tends to overfit to noisy pseudo-labels. These findings emphasize the critical role of the self-sampling strategy in our framework.

5 Conclusion

In this paper, we have explored and grappled with the challenge of a deficiency in the accessibility of high-quality datasets for model fine-tuning in the field of cross text-molecule learning, by developing a novel framework. This innovative framework efficiently combines model fine-tuning with data augmentation, integrating the model in such a way that it expands its roles to include annotating unlabeled data and sampling a subset of high-quality data. The paper further introduces a high-quality dataset, the SAPubChem-41, expressly designed for fine-tuning purposes. The experiments' results were impressive as the model optimized using SAPubChem-41 consistently outperformed the model optimized using the original dataset, ChEBI-20. Going forward, this paper's findings present crucial steps toward addressing high-quality data scarcity for fine-tuning and ensuring further advancement in the exciting field of cross text-molecule learning.

6 Limitations and Future Works

One limitation of our framework is that it only considers a 1D representation of molecules. The integration of additional representations, such as molecule graphs, is left as future work. Additionally, our data augmentation module currently only supports sample-level labels. It is expected that more quality measurements from different levels will be incorporated to provide a comprehensive perspective on quantifying the dataset quality.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62302433, U23A20496), Zhejiang Provincial "Jianbing" "Lingyan" Research and Development Program of China (2024C01135), Zhejiang Provincial Natural Science Foundation of China (LQ24F020007), New Generation AI Development Plan for 2030 of China (2023ZD0120802) and CCF-Tencent Rhino-Bird Fund (RAGR20230122).

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yuhan Chen, Nuwa Xi, Yanrui Du, Haochun Wang, Chen Jianyu, Sendong Zhao, and Bing Qin. 2023. From artificially real to real: Leveraging pseudo data from large language models for low-resource molecule discovery.

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

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, et al. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.

Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

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

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022b. Translation between molecules and natural language. In *EMNLP*, pages 375–413. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Alex Hernández-García and Peter König. 2018. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2023. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2023a. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

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*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2024. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2024. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Nadine Schneider, Roger A Sayle, and Gregory A Landrum. 2015. Get your atoms in order an open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of chemical information and modeling*, 55(10):2111–2120.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2022. Calibrating sequence likelihood improves conditional language generation.

Jing Zhou, Yanan Zheng, Jie Tang, Jian Li, and Zhilin Yang. 2021. Flipda: Effective and robust data augmentation for few-shot learning. *arXiv preprint arXiv:2108.06332*.

A Datasets Information

SAPubChem-41 The SAPubChem-41 datasets comprises 33k real data sourced from CHEBI-20 and 8k high-quality pseudo labelled data sourced from self-augmentation on Text+ChemT5. Representative examples of pseudo-labelled part are provided in Table 7. In addition, to further elucidate the advantages of iterative strategy on model performance, we have documented representative examples of updated captions in Table 6.

Comparison with Existing Datasets In Table 8 we present a detailed comparison of our SAPubChem-41 datasets with existing limited datasets with parallel molecule-description pairs,including:

- **ChEBI-20** (Edwards et al., 2021): consists of 33k molecule-caption pairs with caption exploiting ChEBI(Degtyarenko et al., 2007)annotations and more than 20 words.
- **PCdes** (Zeng et al., 2022): consists of 15k substances in PubChem which have names, SMILES and corresponding paragraphs of property descriptions.
- **DrugBank-23**(Chen et al., 2023): consists of 23k compounds from DrugBank (Wishart et al., 2018) with corresponding description's length longer than 20 words.

B Pre-trained Model Information

Table 9 presents the information of three different pre-trained models utilized in our experiments.

C Experiments

C.1 Performance Comparison

Here we present the complete results of performance comparison for both molecular captioning task and molecule generation task in Table 10&Table 11.

C.2 Ablation Study

Here we present the complete results of both molecular captioning task and molecule generation task using different sampling strategy in Table 12 & Table 13.

C.3 Additional Experiments

To further substantiate the efficacy of our framework and the exceptional quality of the SAPubChem-41 dataset, we have conducted a series of experiments. These were performed with the goal of optimizing three distinct pre-trained models, utilizing both our augmented dataset, SAPubChem-41, and other prevalent public datasets which include ChEBI-20 and PCdes. The findings from these experiments are reported in Table 14 & Table 15. It can be noted that across both task categories, models that were optimized using our self-augmented dataset, SAPubChem-41, generally exhibited superior performance in comparison to those optimized with other existing datasets.

SMILES	[C@H]1(O[C@H]([C@@H](O)C)C(O)O)[C@@H](O)[C@H](O)[C@H](O)[C@H](C)O1
previous caption	The molecule is a glycosylglucose consisting of D-galactopyranose and D-glucopyranose residues joined in sequence by a (1->4) glycosidic bond. It derives from a D-galactopyranose and a D-glucopyranose.(0.6611)
updated caption	The molecule is a deoxygalactose that is D-galactopyranose in which the hydroxy group at position 3 has been replaced by a methyl group. It is a deoxygalactose and a methyl glycoside. It derives from a D-galactopyranose.(0.6817)
SMILES	CC([C@@H](C)O)(C)OC-
previous caption	The molecule is an ether in which the stereocentres at positions 2 and 3 both have S-configuration. It is an ether and a secondary alcohol.(0.6610)
updated caption	The molecule is an ether in which the stereocentres at positions 2 and 3 both have S-configuration. It is an ether and a secondary alcohol. It derives from a hydride of an oxepane.(0.6704)
SMILES	CI
previous caption	The molecule is an iodoalkane that is ethane in which one of the hydrogens is substituted by iodine. It has a role as a human metabolite. It is an iodoalkane and a member of iodos. It derives from a hydride of an ethane.(0.7479)
updated caption	The molecule is an iodoalkane that is ethane in which one of the hydrogens is substituted by iodine. It has a role as a metabolite. It is an iodoalkane and a volatile organic compound. It derives from a hydride of an ethane.(0.7671)
SMILES	N([N+])=C
previous caption	The molecule is an organic cation resulting from the protonation of the nitrogen of nitric acid. It is a conjugate acid of a nitric acid.(0.7479)
updated caption	The molecule is a hydracid and a one-carbon compound. It is a conjugate acid of a hydrazine. It derives from a hydride of a hydrazine.(0.7599)
SMILES	[C@H]1(C)[C@@](CO)(O)[C@@H]-(O[C@@H](C)O)[C@H](OC(C)C)O1
previous caption	The molecule is a spiro-epoxide resulting from the formal epoxidation of the hydroxy group at position 2 of D-fructofuranose. It is a spiro-epoxide and a glycoside. It derives from a D-fructofuranose.(0.6014)
updated caption	The molecule is a deoxygalactose that is alpha-D-galactopyranose in which the hydroxy group at position 2 has been replaced by a methyl group. It is a deoxygalactose and a tertiary alcohol.(0.6839)

Table 6: Five examples of the comparison of previous caption when adopting our iterative strategy.Corresponding Text2Mol Score of each molecule-caption pairs is present in the bracket.

SMILES	Caption
C@H](C[C@H](C(/C=C/C)=O)C)([C@@H]([C@H](/C=C/C(=O)O)C)O[C@H]1[C@H](O)[C@@H]([NH+])C[C@@H](C)O1)C	The molecule is an organic cation that is the conjugate acid of 1D-myo-inositol, obtained by protonation of the tertiary amino group; major species at pH 7.3. It is an ammonium ion derivative and an organic cation. It is a conjugate acid of a 1D-myo-inositol.
[O-][Mn]	The molecule is a monovalent inorganic anion obtained by deprotonation of manganese. It is a manganese oxoanion and a monovalent inorganic anion. It is a conjugate base of a manganese.
C[C@@H](C(OC)OC)C=C	The molecule is an ether in which the stereocentres at positions 2 and 3 both have S-configuration. It is an ether and an alicyclic compound.
O(C([C@H](C)N)=O)[C@H]([C@H](O)C)[C@H](C)O	The molecule is an amino cyclitol that is scyllo-inositol in which the hydroxy group at position 2 has been replaced by an amino group. It has a role as a bacterial metabolite. It is an amino cyclitol and a primary amino compound. It derives from a scyllo-inositol.
C(C)(C)(C)Br	The molecule is a bromoalkane that is ethane substituted by a bromo group at position 2. It has a role as a metabolite. It derives from a hydride of an ethane.
C[Se]C	The molecule is an organoselenium compound that is selenium substituted by a methyl group at position 2. It has a role as a metabolite. It derives from a selenium.

Table 7: Five examples of SAPubChem-41

Info	SAPubChem-41	ChEBI-20	PCdes	DrugBank-23
Train	34,407	26,407	10,500	17,109
Validation	3,301	3,301	1,500	3,667
Test	3,300	3,300	3,000	3,666
$\mathcal{L}_{\text{SMILES}}$	69.96	81.56	56.47	54.11
$\mathcal{L}_{\text{Description}}$	41.59	52.88	72.47	65.04
Data source	PubChem			DrugBank

Table 8: Details about the existing datasets and ours (SAPubChem-41). L_{SMILES} denotes the average length of SMILES while $L_{\text{Description}}$ denotes the average word count per description.

Model	Suffix	Parameters (M)
Text+Chem T5	base	220
Ada-T5	-	220
MolT5	base	220

Table 9: Model Size

model	dataset	BLEU-2	BLEU-4	Rouge-1	Rouge-2	Rouge-L	Meteor
Text+Chem T5	ChEBI-20	0.580	0.490	0.647	0.498	0.586	0.604
	SAPubChem-41	0.584	0.496	0.653	0.504	0.593	0.614
Ada-T5	ChEBI-20	0.424	0.316	0.543	0.369	0.483	0.464
	SAPubChem-41	0.444	0.350	0.575	0.410	0.514	0.493
MolT5	ChEBI-20	0.540	0.457	0.634	0.485	0.578	0.569
	SAPubChem-41	0.551	0.464	0.638	0.484	0.585	0.572

Table 10: The complete results of different models for molecular captioning on ChEBI-20, SAPubChem-41 datasets.

size	dataset	BL	Acc	Lev	MACCS	RDK	Morgan	Val
Text+Chem T5	ChEBI-20	0.750	0.212	27.39	0.874	0.767	0.697	0.792
	SAPubChem-41	0.827	0.243	18.59	0.887	0.792	0.726	0.937
Ada-T5	ChEBI-20	0.699	0.182	27.48	0.869	0.753	0.672	0.886
	SAPubChem-41	0.714	0.213	26.52	0.879	0.772	0.699	0.854
MolT5	ChEBI-20	0.769	0.081	24.49	0.721	0.588	0.529	0.772
	SAPubChem-41	0.775	0.093	33.16	0.814	0.668	0.582	0.788

Table 11: The complete results of different models for molecular generation on ChEBI-20, SAPubChem-41 datasets.

size	dataset	BLEU-2	BLEU-4	Rouge-1	Rouge-2	Rouge-L	Meteor
2k	random	0.534	0.447	0.641	0.489	0.579	0.575
	refined	0.574	0.486	0.651	0.502	0.590	0.601
4k	random	0.537	0.451	0.641	0.489	0.579	0.577
	refined	0.580	0.486	0.653	0.503	0.592	0.602
6k	random	0.532	0.446	0.640	0.489	0.577	0.573
	refined	0.583	0.490	0.653	0.504	0.593	0.611
8k	random	0.527	0.443	0.640	0.489	0.579	0.571
	refined	0.584	0.496	0.653	0.504	0.593	0.614
10k	random	0.526	0.442	0.639	0.489	0.579	0.571
	refined	0.583	0.491	0.652	0.503	0.591	0.610

Table 12: Full results of molecule captioning task using different sampling strategy.

size	dataset	BLEU	Accuracy	Levenshtein	MACCS FTS	RDK FTS	Morgan FTS	Validity
2k	random	0.715	0.215	23.82	0.873	0.766	0.703	0.868
	refined	0.821	0.236	19.31	0.885	0.788	0.721	0.931
4k	random	0.720	0.219	23.79	0.876	0.768	0.706	0.862
	refined	0.832	0.237	18.69	0.885	0.788	0.722	0.938
6k	random	0.709	0.220	24.38	0.876	0.767	0.706	0.858
	refined	0.828	0.240	19.28	0.886	0.788	0.724	0.934
8k	random	0.715	0.221	24.27	0.875	0.763	0.701	0.855
	refined	0.827	0.243	18.59	0.887	0.792	0.726	0.937
10k	random	0.715	0.207	24.37	0.870	0.759	0.696	0.864
	refined	0.830	0.238	19.03	0.883	0.786	0.721	0.937

Table 13: Full results of molecule generation task using different sampling strategy.

model	dataset	BLEU-2	BLEU-4	Rouge-1	Rouge-2	Rouge-L	Meteor
Text+Chem T5	ChEBI-20	0.580	0.490	0.647	0.498	0.586	0.604
	PCdes	0.352	0.266	0.439	0.274	0.373	0.382
	SAPubChem-41	**0.584**	**0.496**	**0.653**	**0.504**	**0.593**	**0.614**
Ada-T5	ChEBI-20	0.424	0.316	0.543	0.369	0.483	0.464
	PCdes	0.289	0.188	0.428	0.243	0.367	0.324
	SAPubChem-41	**0.444**	**0.350**	**0.575**	**0.410**	**0.514**	**0.493**
MolT5	ChEBI-20	0.540	0.457	0.634	**0.485**	0.578	0.569
	PCdes	0.165	0.078	0.290	0.118	0.233	0.204
	SAPubChem-41	**0.551**	**0.464**	**0.638**	0.484	**0.585**	**0.572**

Table 14: The complete results of different models for molecular captioning on SAPubChem-41 and other existing datasets.The **best** scores are in bold.

size	dataset	BL	Acc	Lev	MACCS	RDK	Morgan	Val
Text+Chem T5	ChEBI-20	0.750	0.212	27.39	0.874	0.767	0.697	0.792
	PCdes	0.614	0.105	30.43	0.697	0.544	0.459	0.849
	SAPubChem-41	**0.827**	**0.243**	**18.59**	**0.887**	**0.792**	**0.726**	**0.937**
Ada-T5	ChEBI-20	0.699	0.182	27.48	0.869	0.753	0.672	0.886
	PCdes	0.579	0.089	38.99	0.778	0.600	0.492	**0.947**
	SAPubChem-41	**0.714**	**0.213**	**26.52**	**0.879**	**0.772**	**0.699**	0.854
MolT5	ChEBI-20	0.769	0.081	**24.49**	0.721	0.588	0.529	0.772
	PCdes	0.476	0.009	53.27	0.635	0.432	0.330	0.711
	SAPubChem-41	**0.775**	**0.093**	33.16	**0.814**	**0.668**	**0.582**	**0.788**

Table 15: The complete results of different models for molecular generation on SAPubChem-41 and other existing datasets.The **best** scores are in bold.