Training Text-to-Molecule Models with Context-Aware Tokenization

Seojin Kim[1]*, **Hyeontae Song**[2]*†, **Jaehyun Nam**[3], **Jinwoo Shin**[3]
[1]Seoul National University, [2]Moloco Inc.,
[3]Korea Advanced Institute of Science and Technology (KAIST)
osikjs@snu.ac.kr, hyeontae3109@gmail.com,
{jaehyun.nam, jinwoos}@kaist.ac.kr

Abstract

Recently, text-to-molecule models have shown great potential across various chemical applications, e.g., drug-discovery. These models adapt language models to molecular data by representing molecules as sequences of atoms. However, they rely on atom-level tokenizations, which primarily focus on modeling local connectivity, thereby limiting the ability of models to capture the global structural context within molecules. To tackle this issue, we propose a novel text-to-molecule model, coined ***C**ontext-**A**ware **M**olecular **T5*** (**CAMT5**). Inspired by the significance of the substructure-level contexts in understanding molecule structures, e.g., ring systems, we introduce substructure-level tokenization for text-to-molecule models. Building on our tokenization scheme, we develop an importance-based training strategy that prioritizes key substructures, enabling CAMT5 to better capture the molecular semantics. Extensive experiments verify the superiority of CAMT5 in various text-to-molecule generation tasks. Intriguingly, we find that CAMT5 outperforms the state-of-the-art methods using only 2% of training tokens. In addition, we propose a simple yet effective ensemble strategy that aggregates the outputs of text-to-molecule models to further boost the generation performance. Code is available at https://github.com/Songhyeontae/CAMT5.git.

1 Introduction

Discovering molecules that match desired language descriptions has been a long-standing goal in chemistry since it is an essential ingredient for practical deployments like drug-discovery and material design (Su et al., 2022; Gong et al., 2024; Li et al., 2024). However, achieving such text-to-molecule generation poses a challenge because of the different structural modalities of language and molecules. To address this challenge, researchers have explored the fine-tuning of pre-trained language models with additional molecular data (Christofidellis et al., 2023; Chen et al., 2024), which is inspired by the recent success of language models in leveraging various domain knowledge, including chemical concepts (Taylor et al., 2022; Yu et al., 2024). Specifically, they treat each molecule as a sequence of tokens using string-based molecular representations such as SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020). Intriguingly, they show that these molecule-aware language models, i.e., text-to-molecule models, can be obtained by learning the text-conditional molecule distribution based on treating atoms as tokens of language models (Edwards et al., 2022; Pei et al., 2023).

However, it is yet underexplored *which* tokenization strategy for molecules is more effective for text-to-molecule models. Current state-of-the-art approaches (Edwards et al., 2022; Pei et al., 2023) adopt atom-level tokenization, where each atom is represented as a single token within the model's token space (Christofidellis et al., 2023; Liu et al., 2023; Pei et al., 2023). Even though they show remarkable performance as pioneering efforts, such atom-level tokenizations limit the models' ability to capture crucial global contextual patterns within molecules, focusing only on local connectivities (Xia et al., 2022; Liu et al., 2024; Luong and Singh, 2024). This leads to the question of *how to tokenize molecules in a context-preserving manner to train text-to-molecule models more effectively.*

Contribution. In this paper, we introduce a novel text-to-molecule model, coined ***C**ontext-**A**ware **M**olecular **T5*** (**CAMT5**), by proposing a context-enriched motif-level token space. Specifically, we draw inspiration from the following chemical prior—the structural context of molecules is more effectively captured through their substructure-level, i.e., motif-level, characteristics rather than the atom-level attributes (Jin et al., 2018, 2020;

* These authors contributed equally.
† Work done at KAIST.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22442–22460
November 4-9, 2025

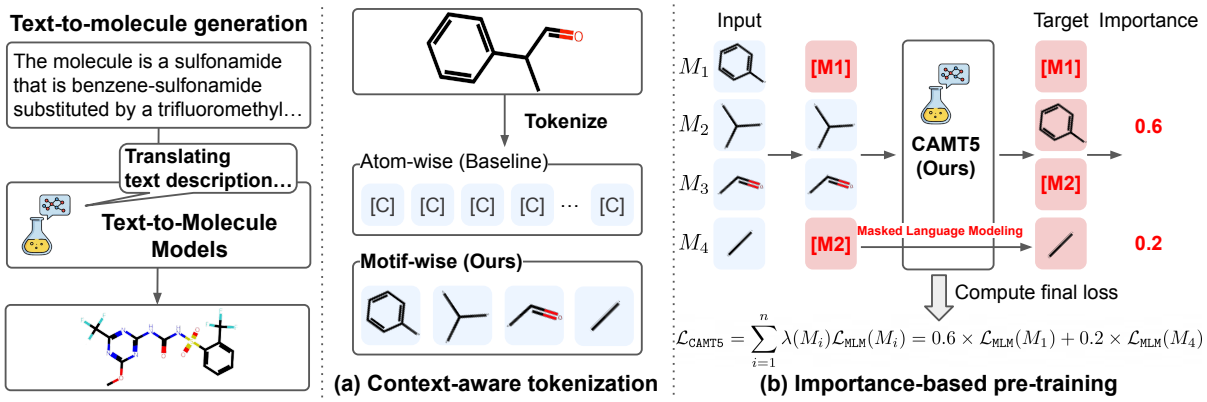


Figure 1: An overview of our proposed method. (a) Context-aware moecule tokenization: we train CAMT5 with a motif-level token sapce. (b) Importance-based pre-training: we priortize key motifs during pre-training.

Zhang et al., 2021; Kim et al., 2023). Consequently, we hypothesize that text-to-molecule models can be further improved by emphasizing the information on key motifs during the training phase. To this end, we propose a new motif-level tokens for text-to-molecule models and develop a novel training strategy that effectively leverages the relative importance of the individual motif-level tokens (see Figure 1 for description).

In particular, we carefully design the motif-level tokenization for CAMT5 to additionally alleviate two drawbacks in the tokenization strategies of previous text-to-molecule models. First, CAMT5 always generates a *valid* molecule, while MolT5 (Edwards et al., 2022) often generates an *invalid* token sequences that do not correspond to any molecule. Second, each of our motif-level tokens has a *unique* interpretation, while some of the tokens in BioT5 (Pei et al., 2023) have *multiple* interpretations, e.g., both an atom and the number of atoms in a ring preceding the token (Krenn et al., 2020), introducing semantic-level ambiguities for the models.

Consequently, to leverage our motif-level tokens effectively, we propose an importance-based training approach that prioritizes key motifs. Specifically, each token is assigned to an importance value derived from its constituent atoms, and the training loss is adjusted by weighting it according to this pre-defined importance. This loss design is made possible by our carefully designed motif-level tokens, each representing a unified chemical context (Kim et al., 2023; Luong and Singh, 2024), unlike atom-wise tokenizations (Weininger, 1988; Krenn et al., 2020) in previous text-to-molecule models.

We verify our method's effectiveness on popular benchmarks, e.g., ChEBI-20 (Edwards et al., 2021). On ChEBI-20, the state-of-the-art results are achieved using only 2% of the training tokens required by the previous best-performing baseline, BioT5 (Pei et al., 2023).[1] Specifically, CAMT5 improves the ratio of molecules that exactly match the description (Exact; higher is better) by 0.413 → 0.430, and those similar to the description (RDK; higher is better) by 0.801 → 0.840. We also show that a simple ensemble strategy utilizing CAMT5 further improves the overall performance, e.g., Exact by 0.430 → 0.472. Finally, we verify CAMT5's effectiveness in molecule modification.

2 Related works

Text-to-molecule models. Inspired by the recent advancements in language models (Raffel et al., 2020; Taylor et al., 2022; Achiam et al., 2023), significant efforts have been made to adapt these models for molecular applications, leading to the development of molecule-aware language models, i.e., text-to-molecule models (Edwards et al., 2021, 2022; Christofidellis et al., 2023; Liu et al., 2023; Pei et al., 2023; Chen et al., 2024). These approaches typically involve fine-tuning pre-trained language models, such as T5 (Raffel et al., 2020), using molecular data by representing molecules as sequences of atom-level tokens. For instance, existing models (Edwards et al., 2022; Christofidellis et al., 2023; Pei et al., 2023) leverage molecular representations like SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2020), which encode molecules as atom-level token sequences for text-to-molecule frameworks. As a result, they primarily focus on capturing local atom-wise connectivity, while overlooking the crucial global structural con-

[1]The performance of BioT5 (Pei et al., 2023) benefits from additional non-public high-quality molecular pre-training data, which is not available for us.

Method	Token	Validity	Non-degeneracy
MolT5	Atom	✗	✓
BioT5	Atom	✓	✗
CAMT5 (Ours)	Motif	✓	✓

Table 1: Comparison of our molecular tokens with previous text-to-molecule models. Token denotes the information encoded in a single token. We mark Validity if a sequence of tokens always represents a valid molecule, and we mark Non-degeneracy if a single token corresponds to a unique molecular interpretation.

text of molecules (Zhang et al., 2021; Xia et al., 2022; Kim et al., 2023; Luong and Singh, 2024).

In addition, text-to-molecule models based on atom-level tokenizations come with additional drawbacks. First, SMILES-based models often generate *invalid* token sequences that violate the grammar (Weininger, 1988), which do not correspond to valid molecules. Second, SELFIES-based models introduce semantic-level ambiguities, i.e., *degeneracy*, in token interpretations (Krenn et al., 2020), leading to sub-optimal performance in modeling the token distribution. For example, the '[O]' token can be interpreted completely differently: an oxygen atom or an indicator of a ring system comprising six atoms preceding this token. To address the aforementioned limitations, we carefully design our context-enriched motif-level tokens, ensuring *validity* of the generated token sequences and *non-degeneracy* in token interpretations (see Table 1).

Context-aware molecule learning. Recent studies in the molecular domain have explored the concept of *context-aware* learning for molecules. For example, Zhang et al. (2021); Kim et al. (2023); Luong and Singh (2024) propose self-supervised learning frameworks that leverage motif-level context to derive chemically meaningful molecular embeddings. A notable approach in this line of work is context-aware molecule generation (Jin et al., 2018, 2020; Kong et al., 2022; Geng et al., 2023), which focuses on learning the distribution of motifs instead of individual atoms with specialized architectures. Intriguingly, they show superior performance in molecule generation by incorporating contextual patterns of motifs within molecules. In particular, t-SMILES (Wu et al., 2024) introduces a linearized representation of motifs using full binary tree structures. Therefore, they require additional grammar tokens to describe the full binary tree structures. In contrast, our motif tokens do not require any grammar tokens, enabling CAMT5 to concentrate solely on learning the relationships between motifs without being constrained by grammar representations (see Appendix D.5 for experimental comparison).

3 Method

In Section 3.1, we explain an overview of our problem of interest. In Section 3.2, we provide the description of our context-aware text-to-molecule model, CAMT5. In Section 3.3, we describe our confidence-based ensemble strategy.

3.1 Problem description

We define our problem of *text-to-molecule generation* as follows. Our goal is to train a text-to-molecule model f_θ such that $f_\theta(\mathbf{x}) = \mathbf{m}$, where $\mathbf{x}$ is a text description of the desired molecule and $\mathbf{m}$ is the corresponding target molecule (see Table 11 for an example). Recent works (Edwards et al., 2022; Pei et al., 2023) have shown that such f_θ can be obtained by fine-tuning a pre-trained language model with description-molecule pairs $\{\mathbf{x}_k, \mathbf{m}_k\}_{k=1}^N$, using the following objective:

$$\mathcal{L}(\theta; \mathbf{x}_k, \mathbf{m}_k) := \mathcal{L}_{\texttt{CE}}\Big(f_\theta(\mathbf{x}_k), \mathbf{m}_k\Big), \qquad (1)$$

where $\mathcal{L}_{\texttt{CE}}$ denotes cross-entropy loss, and $\mathbf{x}_k$ and $\mathbf{m}_k$ denote the k-th text description and the token sequences of the target molecule in the text-to-molecule model's token space, respectively.

Here, the choice of tokenization strategy for $\mathbf{m}_k$ plays a critical role in training an effective f_θ (Pei et al., 2023), as it directly influences how the sequence of tokens captures and represents the structural context of the original molecule. However, previous text-to-molecule models often overlook such importance, relying only on the local connectivity of atoms based on the atom-level tokenization methods, e.g., SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020). In contrast, our contribution resolves the drawbacks of previous tokenization strategies by incorporating substructure-level contextual patterns into the token space of text-to-molecule models. This allows us to represent a molecule in a context-aware manner.

3.2 CAMT5: Context-Aware Molecular T5

Context-aware molecule tokenization. We propose to construct the molecule token space of CAMT5 to effectively capture and reflect the structural context of molecules. To achieve this, we consider chemically meaningful fragments, i.e., motifs, as individual tokens in CAMT5. This approach differs from previous text-to-molecule models that

only rely on atom-level tokens (Edwards et al., 2022; Pei et al., 2023). Specifically, we consider the following set of atoms, i.e., a motif, as a single token: (1) atoms forming a ring structure and (2) atoms connected by a non-single bond (see Figure 1 for an example). Such groups of atoms are rigidly bound to each other and represent an important structural context, such as resonance (Anslyn and Dougherty, 2006). An atom not associated with (1) and (2) is considered as a single token.

We then propose to represent a molecule as a sequence of motif-level tokens, based on the order of the tree-search algorithm on a tree of motifs. Consider a molecule graph $G = (V, E)$ with the set of atoms V and edges E. We construct a tree of motifs $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{M_i\}_{i=1}^{n}$ is the set of n motifs with $M_i = (V_i, E_i)$, and $\mathcal{E}$ is the set of bonds between motifs. Here, $\mathcal{T}(G)$ effectively preserves all the information of the original molecule graph G, i.e., $V = \cup V_i$ and $E = \cup_i E_i \cup \mathcal{E}$, with context-enriched nodes by replacing atom-level nodes V with motif-level nodes $\mathcal{V}$, satisfying $|\mathcal{V}| \leq |V|$. Consequently, we obtain the sequence of motif tokens by enumerating $\mathcal{V}$ based on the order of the depth-first-search (DFS) algorithm, i.e., $\mathbf{m}_{\texttt{CAMT5}} = [M_1, ..., M_n]$. We then train our text-to-molecule model f_θ with $\{\mathbf{x}_k, \mathbf{m}_{\texttt{CAMT5},k}\}_{k=1}^{N}$ using the training objective in Eq. (1). Note that our method ensures the (1) *validity* of the generated token sequences since we do not introduce tokens that should appear as a pair, c.f., the branch tokens '(' and ')' in SMILES (Weininger, 1988). Also, our tokens are (2) *non-degenerate* by construction; a single token represents only a single motif, c.f., '[O]' as an oxygen atom or an indicator of a ring system comprising six atoms preceding this token in SELFIES (Krenn et al., 2020). We provide further details of our token space in Appendix A.

Our context-enriched tokenization plays a crucial role in discriminating the atoms within different structural contexts. For example, the aromatic carbons in a phenyl group (i.e., [C][=C][C][=C][C][=C][Ring1][=Branch1] in BioT5; Pei et al., 2023) and the aliphatic carbons (i.e., [C][C][C][C][C][C] in BioT5) differ significantly in chemical context, due to resonance and ring structure. However, previous text-to-molecule models do not distinguish the difference between the carbon atoms of each motif, regarding both carbons as the same [C] token. CAMT5 resolves this by assigning different tokens for the entire phenyl groups and the carbons in aliphatic carbons.

Importance-based pre-training. Previous state-of-the-art text-to-molecule models were pre-trained on vast amounts of tokens from unlabeled molecules (Liu et al., 2023; Chen et al., 2024). Notably, MolT5 (Edwards et al., 2021) and BioT5 (Pei et al., 2023) demonstrated the effectiveness of the masked language modeling pre-training objective (Raffel et al., 2020) in enriching the understanding of the molecular domain with unlabeled molecules.

In this paper, we advance the masked language modeling (Raffel et al., 2020) for our motif-level token space, focusing on key motifs during pre-training to better capture molecular structural context. To achieve this, we define an importance value $\lambda(M_i)$ for each $M_i \in \mathcal{V}$, reflecting the relative significance of motifs in a given molecule. Based on these pre-defined importance values, we train CAMT5 with the weighted training loss:

$$\mathcal{L}_{\texttt{CAMT5}} = \sum_{i=1}^{n} \lambda(M_i) \mathcal{L}_{\texttt{MLM}}(M_i), \tag{2}$$

where $\mathcal{L}_{\texttt{MLM}}$ denotes the masked language modeling loss. Here, we find that a simple choice of $\lambda(M_i)$, i.e., the number of atoms in M_i, efficiently and effectively improves the generation performance (see Appendix B for details on the definition of λ).

3.3 Confidence-based ensemble

We propose a simple yet effective confidence-based ensemble method to further improve the generation quality of our CAMT5. Specifically, we leverage the outputs of other text-to-molecule models, which often use *different* token space, e.g., SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020). Here, we note that recent ensemble strategies (Jiang et al., 2024; Sukhbaatar et al., 2024) only work on the models with the same token space, and thus are not applicable to text-to-molecule models with different tokenizations.

To tackle this issue, we define the *confidence* $C(\mathbf{m}_i; f_i, \mathbf{x})$ as the average log-likelihood of the generated tokens, and treat it as a proxy for the quality measure of the generated molecules, i.e.,

$$\begin{aligned} C(\mathbf{m}_i; f_i, \mathbf{x}) &= \frac{\sum_{j=1}^{K_i} \texttt{log} P_{f_i}([T_j] | \mathbf{x}, [T_1 ..., T_{j-1}])}{K_i} \\ &= -\mathcal{L}_{\texttt{CE}}(f_i(\mathbf{x}), \mathbf{m}_i), \end{aligned}$$

where f_i is the i-th text-to-molecule model and $\mathbf{m}_i = [T_1, ..., T_{K_i}]$ be the generated K_i tokens

Model	#Params.	Representation	Train Tokens	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Valid. ↑
RNN	56M	SMILES	-	0.005	0.591	0.400	0.362	0.542
Transformer	76M	SMILES	-	0.000	0.480	0.320	0.217	0.906
$\text{T5}_{\texttt{small}}$	77M	SMILES	-	0.064	0.704	0.578	0.525	0.608
$\text{T5}_{\texttt{base}}$	248M	SMILES	-	0.069	0.731	0.605	0.545	0.660
$\text{T5}_{\texttt{large}}$	783M	SMILES	-	0.279	0.823	0.731	0.670	0.902
$\text{MolT5}_{\texttt{small}}$	77M	SMILES	66B	0.079	0.703	0.568	0.517	0.721
$\text{MolT5}_{\texttt{base}}$	248M	SMILES	66B	0.081	0.721	0.588	0.529	0.772
$\text{MolT5}_{\texttt{large}}$	783M	SMILES	66B	0.311	0.834	0.746	0.684	0.905
GPT-3.5-turbo	>175B	SMILES	-	0.019	0.705	0.462	0.367	0.802
MolReGPT	>175B	SMILES	-	0.139	0.847	0.708	0.624	0.887
MolXPT	350M	SMILES	1.8B	0.215	0.859	0.757	0.667	0.983
$\text{BioT5}^{*}_{\texttt{base}}$	252M	SELFIES	69B	0.413	**0.886**	0.801	0.734	**1.000**
$\text{MolT5}^{\dagger}_{\texttt{base}}$	248M	SMILES	1.6B	0.326	0.847	0.797	0.720	0.950
$\text{BioT5}^{\dagger}_{\texttt{base}}$	252M	SELFIES	1.6B	0.344	0.842	0.773	0.664	**1.000**
$\text{CAMT5}_{\texttt{small}}$ (Ours)	103M	**Motif (Ours)**	1.6B	0.391	0.874	0.827	0.727	**1.000**
$\text{CAMT5}_{\texttt{base}}$ (Ours)	286M	**Motif (Ours)**	1.6B	0.422	0.882	0.834	0.742	**1.000**
$\text{CAMT5}_{\texttt{large}}$ (Ours)	836M	**Motif (Ours)**	1.6B	**0.430**	0.885	**0.840**	**0.749**	**1.000**

Table 2: Quantitative results of the text-to-molecule generation task in the CheBI-20 (Edwards et al., 2021) benchmark. `small`, `base` and `large` denote that the model is derived from the T5-small, T5-base and T5-large (Raffel et al., 2020), respectively. #Params denotes the number of parameters in each text-to-molecule model. Train Tokens refers to the number of molecule-related pre-training tokens. $*$ denotes that the model is pre-trained with an additional non-public high-quality molecular dataset, which is not available for us. † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours. We highlight the best score in bold.

Model	#Params.	Representation	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Valid. ↑
$\text{MolT5}^{\dagger}_{\texttt{base}}$	248M	SMILES	0.151	0.578	0.523	0.417	0.793
$\text{BioT5}^{\dagger}_{\texttt{base}}$	252M	SELFIES	0.132	0.695	0.624	0.458	**1.000**
$\text{CAMT5}_{\texttt{base}}$ (Ours)	286M	**Motif (Ours)**	**0.196**	**0.738**	**0.679**	**0.528**	**1.000**

Table 3: Quantitative results of the text-to-molecule generation task in PCDes (Zeng et al., 2022). † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours. We bold the best score.

from f_i to the given description $\mathbf{x}$. Then, we define the confidence-based ensemble $f_{\texttt{ens}}$ with $f_1, ..., f_n$ as follows:

$$f_{\texttt{ens}}(\mathbf{x}) = \mathbf{m}_k, \text{ where } k = \texttt{argmax}_i \, C(\mathbf{m}_i; f_i, \mathbf{x}). \tag{3}$$

We note that this ensemble strategy is particularly useful in practical scenarios. Previously, people simply chose the best-performing model among the existing text-to-molecule models, ignoring other on-average under-performing models. However, when the selected model is not *confident* in a certain text description, other models may provide more confident alternatives. In this case, our confidence-based ensemble strategy can be applied to further improve the performance of the best-performing model, i.e., CAMT5, with the help of other existing models, i.e., MolT5 and BioT5.

4 Experiments

We verify the effectiveness of CAMT5 through extensive experiments. In Section 4.1, we explain our experimental setups. Section 4.2 presents the text-to-molecule generation results on the ChEBI-20 and PCDes benchmarks. In Section 3.3, we present the results of our confidence-based ensemble strategy. In Section 4.4, we show the text-conditional molecule modification task results. In Section 4.5, we provide ablation studies on components of CAMT5. We provide additional experimental results and analyses in Appendix D.

4.1 Experimental setup

Baselines. We consider the recently proposed state-of-the-art text-to-molecule models: MolT5 (Edwards et al., 2022), MolReGPT (Li et al., 2024), MolXPT (Liu et al., 2023), and BioT5 (Pei et al., 2023). These models are based on atom-wise tokenization, i.e., SMILES and SELFIES.

Datasets. We evaluate the text-to-molecule generation performance of text-to-molecule models on two popular benchmarks, ChEBI-20 (Edwards et al., 2021) and PCDes (Zeng et al., 2022). In addition, we construct a new dataset of 34k description-

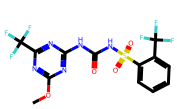
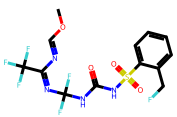
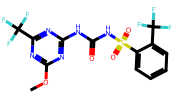
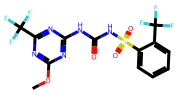
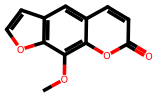
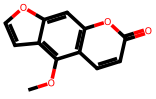
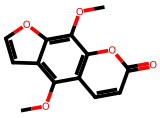
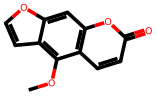
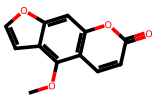
Description	MolT5	BioT5	**CAMT5 (Ours)**	**Ensemble (Ours)**	Target
The molecule is a sulfonamide that is benzene-sulfonamide substituted by a trifluoromethyl...	 RDK: **1.00** Confidence: **-5E-4**	 RDK: 0.50 Confidence: -2E-3	 RDK: 0.92 Confidence: -8E-4	 RDK: **1.00** Confidence: **-5E-4**	
It is a 5-metho-xyfurocoumarin that is psoralen substituted by a methoxy group at position 5. It has...	 RDK: 0.73 Confidence: -5E-4	 RDK: **1.00** Confidence: **-6E-5**	 RDK: 0.82 Confidence: -1E-3	 RDK: **1.00** Confidence: **-6E-5**	

Table 4: Visualizations of the confidence-based ensemble on CheBI-20 (Edwards et al. 2021; the first row) and PCDes (Zeng et al. 2022; the second row). We visualize the cases that other models, i.e., MolT5 and BioT5, help our CAMT5 through ensemble when the confidence (maximally 0.00) of our generated molecule is relatively low. We report the confidences and the RDK scores below each visualization. Our ensemble strategy selects the molecule with the highest confidence as the output of ensemble (see Eq. (3)). We bold the highest score.

Model	**Exact** ↑	**MACCS** ↑	**RDK** ↑	**Morgan** ↑	**Valid.** ↑
Results on the CheBI-20 benchmark.					
MolT5	0.326	0.847	0.797	0.720	0.950
BioT5	0.413	0.886	0.801	0.734	**1.000**
CAMT5	0.430	0.885	0.840	0.749	**1.000**
Ensemble	**0.472**	**0.902**	**0.860**	**0.781**	**1.000**
Results on the PCDes benchmark.					
MolT5	0.151	0.578	0.523	0.417	0.793
BioT5	0.132	0.695	0.624	0.458	**1.000**
CAMT5	0.196	0.738	0.679	0.528	**1.000**
Ensemble	**0.213**	**0.755**	**0.695**	**0.554**	**1.000**

Table 5: Quantitative results of our confidence-based ensemble on the CheBI-20 (Edwards et al., 2021) and PCDes (Zeng et al., 2022) benchmarks. We report the ensemble results based on the best-performing models of MolT5, BioT5, CAMT5 in Table 2 and 3, respectively. We highlight the best score in bold.

molecule pairs from the PubChem database, ensuring no overlap with the molecules in ChEBI-20 or PCDes. This dataset is used to train our CAMT5, as well as †-marked MolT5 and BioT5 models (see Table 2). Further details are provided in Appendix C.

Training setup. Following the previous practices (Edwards et al., 2021; Pei et al., 2023), we pre-train text-to-molecule models with publically available uni-modal datasets, i.e., C4 (Raffel et al., 2020) for the text corpus and ZINC-15 (Sterling and Irwin, 2015) for the molecule corpus. We note that the previous models, e.g., MolT5 and BioT5, are trained with different datasets, which limits a genuine comparison of proposed methods. For example, the official BioT5 model benefits from an additional non-public pre-training dataset. To alleviate this issue, we have aligned the pre-training and fine-tuning configurations of each model and marked †, e.g., as shown in Table 2. We provide further details of experimental setups in Appendix B.

Metrics. For an extensive evaluation of text-to-molecule generation, we utilize various metrics that reflect the quality of the generated molecules. The detailed description of metrics are as follows:

- **Exact**: The ratio of the generated molecules that exactly match with the target molecule.
- **MACCS/RDK/Morgan Fingerprint Tanimoto Similarity (MACCS/RDK/Morgan)**: Metrics that measure the fingerprint-level similarity between the generated molecule and the target molecule. MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010) fingerprints are used. If the generated token sequence is not a valid molecule, we set this score as 0. For each dataset, we report the average scores of the generated molecules in each dataset.
- **Validity (Valid.)**: The ratio of the generated token sequences that are valid molecules.[2]

4.2 Main experiments

Table 2 and 3 summarize the quantitative results of the text-to-molecule generation tasks in ChEBI-20 (Edwards et al., 2021) and PCDes (Zeng et al., 2022), respectively. In both benchmarks, our CAMT5 consistently outperforms the baseline text-to-molecule models by generating desirable molecules corresponding to the text description.

[2]For BioT5 and CAMT5, Validity is guaranteed to be 1.0 due to the characteristics of used token representations.

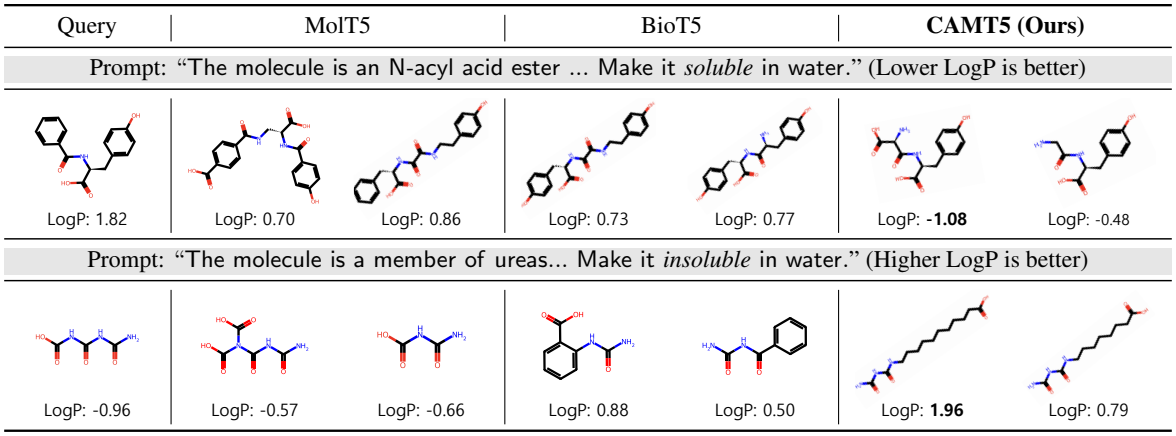


Table 6: Qualitative results of molecule modification on ChEBI-20 (Edwards et al., 2021). We visualize the generated molecules with respect to the prompt with an additional chemical condition, i.e., solubility in water. We report the LogP score below each visualization. Molecules with lower LogP values are more soluble in water. For the best-performing models of MolT5, BioT5, and CAMT5, we report the top-2 molecules that match the property description among 30 generated molecules based on temperature sampling with τ=0.5. We set the best score in bold.

	"Soluble in water"		"Insoluble in water"	
Model	**MACCS ↑**	**ΔLogP ↑**	**MACCS ↑**	**ΔLogP ↓**
MolT5	0.351	1.96	0.268	-1.49
BioT5	0.357	2.01	0.286	-1.71
CAMT5	**0.441**	**2.26**	**0.378**	**-1.98**

Table 7: Quantitative results of molecule modification on ChEBI-20 (Edwards et al., 2021). We average the scores of top-2 molecules from the test set descriptions.

Results on ChEBI-20. In the ChEBI-20 benchmark (Edwards et al., 2021), CAMT5 highly outperforms the state-of-the-art text-to-molecule model, BioT5 (Pei et al., 2023), which leverages an additional non-public high-quality pre-training dataset. For example, CAMT5 shows superior performance in generating molecules that exactly match the given text descriptions, improving the Exact score by 0.413 → 0.430. Furthermore, CAMT5 generates molecules more similar to the given description, achieving higher fingerprint similarity-based scores, e.g., 0.801 → 0.840 and 0.734 → 0.749 in the RDK and Morgan similarity scores, respectively. Notably, CAMT5 achieves these improvements with only 2% of molecule-related pre-training tokens compared to BioT5, underscoring the superiority of our molecule tokenization and importance-based pre-training strategy.

For an extensive comparison with baselines in a fair setup, we also provide the results under the same training datasets and configurations as our CAMT5 (denoted by $\text{MolT5}^{\dagger}_{\texttt{base}}$ and $\text{BioT5}^{\dagger}_{\texttt{base}}$). Within this setup, our model of a similar size, i.e., $\text{CAMT5}_{\texttt{base}}$, demonstrates a significant performance improvement, achieving the Exact score by 0.344 → 0.422. Moreover, it is noteworthy that even with the smaller variant, i.e., $\text{CAMT5}_{\texttt{small}}$, our method consistently outperforms both $\text{MolT5}^{\dagger}_{\texttt{base}}$ and $\text{BioT5}^{\dagger}_{\texttt{base}}$ across all evaluated metrics. These results underscore CAMT5's strong efficacy in generating desired molecules and establish it as a promising approach for text-to-molecule generation tasks.

Results on PCDes. Table 3 shows that CAMT5 is also effective in the more challenging PCDes (Zeng et al., 2022) benchmark, with improvements such as 0.151 → 0.196 in the Exact score and 0.624 → 0.679 in the RDK score. This highlights the robustness and applicability of our CAMT5 across various text-to-molecule generation tasks.

4.3 Confidence-based ensemble

In Table 5, we report the quantitative results of the selected molecules from our confidence-based ensemble strategy (see Eq.(3)). In this experiment, we construct an ensemble model based on the state-of-the-art text-to-molecule models, e.g., MolT5 (Edwards et al., 2022), BioT5 (Pei et al., 2023), and our CAMT5. During ensemble, we make sure that the generated molecules are all valid, by ignoring the output from MolT5 when it does not correspond to a valid molecule. We note that this does not incur an additional computational overhead, since verifying the validity of the generated output does not require computational cost. Overall, our ensemble strategy significantly improves the performance of existing text-to-molecule models, e.g., 0.430 → 0.472 and 0.196 → 0.213 in the Ex-

Model	Importance	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Valid. ↑
$\text{MolT5}^{\dagger}_{\text{base}}$	-	0.326	0.847	0.797	0.720	0.950
$\text{BioT5}^{\dagger}_{\text{base}}$	-	0.344	0.842	0.773	0.664	**1.000**
$\textbf{CAMT5}_{\text{base}}$	✗	0.397	0.868	0.819	0.725	**1.000**
$\textbf{CAMT5}_{\text{base}}$	✓	**0.422**	**0.882**	**0.834**	**0.742**	**1.000**

Table 8: Quantitative results on the CheBI-20 (Edwards et al., 2021) benchmark. † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours. We mark Importance if the importance-based pre-training strategy (see Eq. (2)) is applied. We bold the highest score.

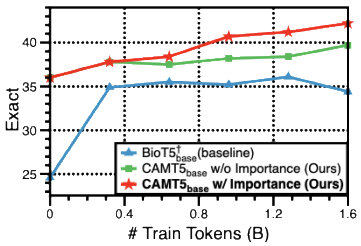


Figure 2: Performance varying the number of pre-training tokens.

act score on the CheBI-20 and PCDes benchmarks, respectively. In Table 4, we provide some examples where our CAMT5 is not quite confident in its output, and other models, i.e., MolT5 and BioT5, generate more confident molecules. In these cases, the ensemble strategy selects the molecules generated by MolT5 or BioT5, which are indeed more similar to the target molecules. In summary, our ensemble strategy effectively leverages on-average underperforming models, i.e., MolT5 and BioT5, to further improve the output of the best-performing model, i.e., CAMT5, through our carefully designed confidence-based ensemble strategy.

4.4 Text-conditional molecule modification

In this section, we verify the potential of our CAMT5 in the context of *modifying* molecules based on additional text prompt conditions. To achieve molecule modification, we consider a text-to-molecule model f, where $f(\mathbf{x}) = \mathbf{m}$ maps a molecule description $\mathbf{x}$ to its corresponding molecule $\mathbf{m}$. Then, we slightly alter the description $\mathbf{x}$ by appending an additional condition prompt, such as $\mathbf{x}' = \mathbf{x} +$ "Make it *insoluble* in water". The resulting modified molecule $\mathbf{m}' = f(\mathbf{x}')$ is expected to (1) maintain structural similarity to the original molecule $\mathbf{m}$ and (2) faithfully capture the additional conditional text in $\mathbf{x}'$. Although previous studies have considered the modification of molecules based on numerical value conditions (Chen et al., 2021; Zhu et al., 2024b), the exploration of molecule modification conditioned on textual descriptions remains relatively under-explored (Zhu et al., 2024a), despite its practical potential.

In Table 6, we consider the descriptions in the ChEBI-20 test set where MolT5, BioT5, and CAMT5 each generate the same molecule represented in the Query column. We then generate molecules with additional condition prompt in addition to the original description. The results show that our CAMT5 demonstrates meaningful modification capabilities by excelling in two key aspects: (1) preserving the critical substructures of the original molecule in the Query column, such as the N-acyl group, and (2) effectively incorporating additional prompts, as evidenced by the resulting LogP values. We hypothesize that this improvement stems from our unique motif-level tokenization strategy, which is advantageous for incorporating motifs closely related to molecular properties.

In Table 7, we compare the performance of each model based on (i) the MACCS similarity to the original target molecule, and (ii) ΔLogP, defined as the difference in LogP values between the generated molecule and the target molecule. The results show that CAMT5 generates molecules that are structurally closer to the target molecule, while also reflecting the intended property condition.

4.5 Ablation studies

In this section, we verify the effectiveness of the core components of CAMT5, context-aware tokenization and importance-based pre-training strategy. As demonstrated in Table 8 and Figure 2, our CAMT5 without importance-based training (i.e., the third row of the table) already improves the previous best-performing models, e.g., MolT5 (Edwards et al., 2022) and BioT5 (Pei et al., 2023). In other words, this result shows the superiority of our tokenization compared to the previous atom-wise tokenizations, such as SMILES (Weininger, 1988) in MolT5 and SELFIES (Krenn et al., 2020) in BioT5. Furthermore, CAMT5 with the importance-based pre-training strategy (i.e., the last row of the table) significantly outperforms the model pre-trained with the conventional masked language modeling (Raffel et al., 2020) objective (i.e., the third row of the table). This result underscores that guiding the text-to-molecule model to focus more on key substructures is largely advantageous in learning the text-conditional molecule distribution. Overall, these results demonstrate that our care-

fully designed context-aware tokenization and the importance-based pre-training strategy play crucial roles in understanding molecules, and thus improving text-to-molecule generation performance.

5 Conclusion

We propose CAMT5, a new text-to-molecule model with chemistry-specialized tokenization. Specifically, we adapt pre-trained language models by utilizing motif-level tokenization and importance-based training strategy to better understand the chemical structural context of molecules. In addition, we propose a confidence-based ensemble technique to further improve the quality of the generated molecules from CAMT5, using other text-to-molecules. We hope that our work further accelerates future research on domain-specific adaptations of pre-trained language models.

Limitations

In this work, we mainly focus on improving the token space of text-to-molecule models, which is a crucial yet under-explored problem in text-to-molecule models. An interesting future direction would be applying our tokenization to train advanced text-to-molecule models, e.g., leveraging pseudo-data (Chen et al., 2024), diffusion-based generation (Chang and Ye, 2024), and multi-task language modeling (Christofidellis et al., 2023), which are originally based on the previous atom-wise tokenization schemes, e.g., SMILES (Weininger, 1988). We believe that those works will further benefit from our carefully designed context-aware tokenization.

Acknowledgments

This work was supported by the National Supercomputing Center with supercomputing resources including technical support KSC-2024-CRE-0362, and partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST); RS-2022-II220959, Few-shot Learning of Causal Inference in Vision and Language for Decision Making).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Eric V Anslyn and Dennis A Dougherty. 2006. *Modern physical organic chemistry*. University science books.

Jinho Chang and Jong Chul Ye. 2024. Ldmol: A text-to-molecule diffusion model with structurally informative latent space surpasses ar models. *arXiv preprint arXiv:2405.17829*.

Yuhan Chen, Nuwa Xi, Yanrui Du, Haochun Wang, Jianyu Chen, Sendong Zhao, and Bing Qin. 2024. From artificially real to real: Leveraging pseudo data from large language models for low-resource molecule discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21958–21966.

Ziqi Chen, Martin Renqiang Min, Srinivasan Parthasarathy, and Xia Ning. 2021. A deep generative model for molecule optimization via one fragment modification. *Nature machine intelligence*, 3(12):1040–1049.

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 *International Conference on Machine Learning*, pages 6140–6157. PMLR.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. 2008. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503.

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. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

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.

Zijie Geng, Shufang Xie, Yingce Xia, Lijun Wu, Tao Qin, Jie Wang, Yongdong Zhang, Feng Wu, and Tie-Yan Liu. 2023. De novo molecular generation via connection-aware motif mining. In *International Conference on Learning Representations*.

Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024. Text-guided molecule generation with diffusion language model. *arXiv preprint arXiv:2402.13040*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pages 4839–4848. PMLR.

Seojin Kim, Jaehyun Nam, Junsu Kim, Hankook Lee, Sungsoo Ahn, and Jinwoo Shin. 2023. Fragment-based multi-view molecular contrastive learning. In *Workshop on"Machine Learning for Materials"ICLR 2023*.

Seojin Kim, Jaehyun Nam, Sihyun Yu, Younghoon Shin, and Jinwoo Shin. 2024. Data-efficient molecular generation with hierarchical textual inversion. *arXiv preprint arXiv:2405.02845*.

Xiangzhe Kong, Wenbing Huang, Zhixing Tan, and Yang Liu. 2022. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35:2550–2563.

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.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*.

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, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2024. Rethinking tokenizer and decoder in masked graph modeling for molecules. *Advances in Neural Information Processing Systems*, 36.

Kha-Dinh Luong and Ambuj K Singh. 2024. Fragment-based pretraining and finetuning on molecular graphs. *Advances in Neural Information Processing Systems*, 36.

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*.

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.

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.

Teague Sterling and John J Irwin. 2015. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, and 1 others. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. 2009. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl_2):W623–W633.

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.

Juan-Ni Wu, Tong Wang, Yue Chen, Li-Juan Tang, Hai-Long Wu, and Ru-Qin Yu. 2024. t-smiles: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1):4993.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. 2022.

Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

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.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. 2021. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. 2024a. 3m-diffusion: Latent multi-modal diffusion for language-guided molecular structure generation. In *First Conference on Language Modeling*.

Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, Tingjun Hou, Jian Wu, and 1 others. 2024b. Sample-efficient multi-objective molecular optimization with gflownets. *Advances in Neural Information Processing Systems*, 36.

Appendix: Training Text-to-Molecule Models with Context-Aware Tokenization

A Context-aware tokenization details

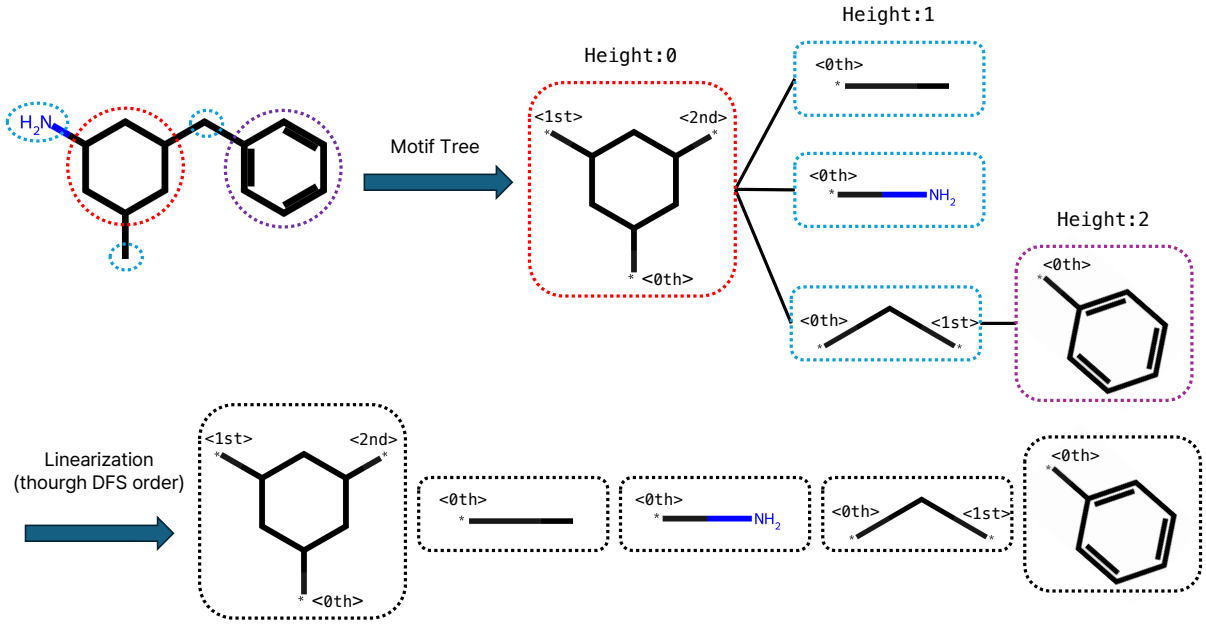


Figure 3: Details of our proposed tokenization scheme in CAMT5: (1) atoms forming a ring structure, (2) atoms connected by a non-single bond, and (3) an atom not associated with (1) and (2) is considered as a single token.

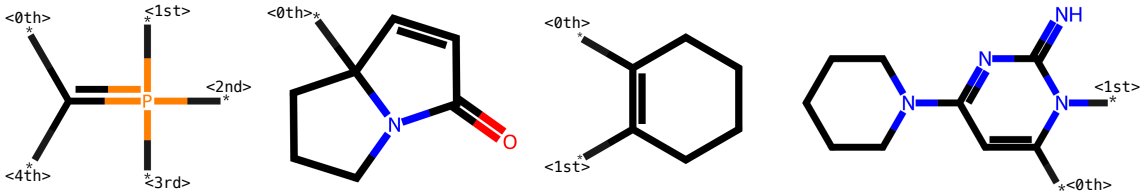


Figure 4: Visualizations of some context-aware motif tokens in CAMT5.

In Figure 3, we visualize the details of our context-aware tokenization scheme. For each motif-level token M_i, there may exist several $v \in V_i$ where $(u, v) \in \mathcal{E}$ for some $u \in V$, i.e., a single motif which is connected to several motifs in $\mathcal{T}$ (see the second token of CAMT5 in Figure 1 for an example). Therefore, we additionally store the order of fragmented bonds when traversing the motif tree. In the fragmented bonds in height 0, the marked number denotes the order of bonds that are connected to children in the linearized token sequence. In the fragmented bonds in height more than 0, the bond marked with zero is connected to its parent. Other bonds are connected to its children by their marked order, starting with 1. We also store the stereo information, e.g., tetrahedral or E-Z, in each token. We utilize this order when converting the sequence of tokens to a molecule. For a given sequence of tokens, we convert the sequence to a molecule by the exactly inverse consequences of the construction of the token sequences. If there exist unvisited fragmented bonds, we simply ignore them, i.e., we consider them to be connected to a hydrogen atom, not to other motif tokens. The number of motif tokens introduced in our CAMT5 is 24,735 in the ChEBI-20 and PCDes benchmarks.

We visualize some of our motif tokens in Figure 4. We note that our choice of tokenization strategy is largely different from previous molecule representations. For example, t-SMILES (Wu et al., 2024) constructs a full binary tree to construct a tree of motifs. However, such construction requires additional grammar representations, e.g., dummy nodes, to ensure the full binary trees. In contrast, our representation does not impose such restrictions, e.g., our motif token can have several children and thus does not require any grammar tokens.

B Experimental details

Details on token importance. We simply use the number of atoms in each token as the importance value $\lambda(M_i)$:

$$\lambda(M_i) = \texttt{Softmax}(\log(A_i + 1)),$$

where A_i denotes the number of atoms in each motif token. For special tokens, e.g., mask tokens, the atom count is set to 0.

Details on pre-training. For each text-to-molecule model, i.e., MolT5†, BioT5†, and CAMT5, we use a general text corpus (Colossal Clean Crawled Corpus (Raffel et al., 2020)) and a molecule corpus (ZINC-15 (Sterling and Irwin, 2015)). Specifically, each model is pre-trained on 1.6B of molecule-related tokens. Training is conducted with a batch size of 16 per GPU across 4 GPUs, with each batch containing an equal mix of text and molecule data. The training runs for 100k steps. We use AdamW with RMS scaling as the optimizer, and apply cosine annealing for the learning rate schedule. Gradients are clipped at 30.0. The base learning rate is 2e-3, and the number of warm-up steps is 1000. The maximum input length for pre-training is 512. Except for our CAMT5, the pre-training loss is the conventional masked language modeling loss from Raffel et al. (2020).[3]

Details on fine-tuning. We fine-tune each model with description-molecule data pairs in the ChEBI-20 (Edwards et al., 2021) and the PCDes (Zeng et al., 2022) benchmarks. Additionally, we utilize 34k text-molecule pairs extracted from PubChem (Wang et al., 2009), ensuring that no molecules overlap with those in the benchmarks. Each model is trained based on the objective in Eq. (1) with the corresponding molecule token representations. We fine-tune the models in 50k steps with a batch size of 48, applying cosine annealing and gradient clipping at 30.0. We select the best model by varying the learning rate within [1e-3, 2e-3].

Computing resources. In our experiments, we use Intel(R) Xeon(R) Gold 6326Y CPU @ 2.90GHz. We use A6000 48GB GPUs for pre-training and a single NVIDIA GeForce RTX 3090 GPU for fine-tuning.

[3] Our importance-based pre-training strategy is not applicable for models with atom-level tokenization, since their tokens represent a single atom.

C Dataset details

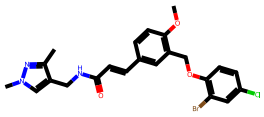
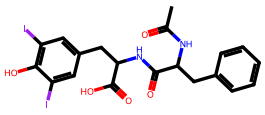
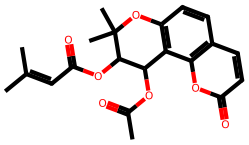
PubChem			
3-[3-[(2-bromo-4-chlorophenoxy)methyl]-4-methoxyphenyl]-N-[(1,3-dimethyl-4-pyrazolyl)methyl]-2-propenamide is a member of cinnamamides and a secondary carboxamide.		N-acetyl-L-phenylalanyl-L-diiodotyrosine is a dipeptide composed of N-acetyl-L-phenylalanine and 3,5-diiodo-L-tyrosine joined together by a peptide linkage. It is a synthetic substrate for pepsin. It is functionally related...	
Disuccinimidyl suberate is a N-hydroxysuccinimide ester resulting from the formal condensation of the two carboxy groups of suberic acid with the hydroxy group of 1-hydroxypyrrolidine-2,5-dione. It is a noncleavable and...		(10-Acetyloxy-8,8-dimethyl-2-oxo-9,10-dihydropyrano[2,3-f]chromen-9-yl) 3-methylbut-2-enoate is a member of coumarins.	

Table 9: Visualizations of our description-molecule pairs collected from Pubchem database (Wang et al., 2009).

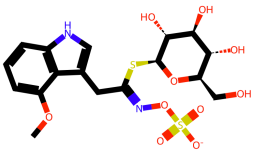
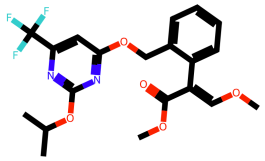
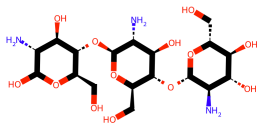
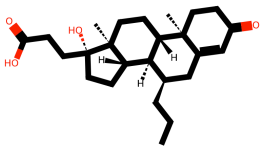
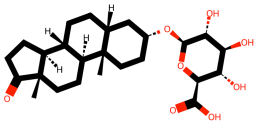
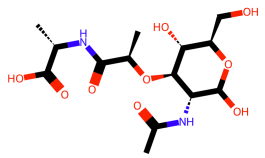
ChEBI-20		PCDes	
The molecule is an indolylmethylglucosinolate that is the conjugate base of 4-methoxyglucobrassicin, obtained by deprotonation of the sulfo group. It is a conjugate base of a 4-methoxyglucobrassicin.		It is a member of pyrimidines, an organofluorine acaricide, a methyl ester, an enoate ester and an enol ether. It has a role as a mitochondrial cytochrome-bc1 complex inhibitor.	
The molecule is an amino trisaccharide comprising of three 2-amino-2-deoxy-D-glucopyranose units joined by beta-(1->4) linkages. It has a role as a marine metabolite and a eukaryotic metabolite.		It is a spironolactone derivative and a potent aldosterone antagonist on mineralocorticoid biosynthesis with diuretic activity . As an aldosterone antagonist, it may inhibit sodium resorption in the collecting duct and may eventually lead to diuresis.	
The molecule is a steroid glucosiduronic acid. It has a role as a human metabolite and a mouse metabolite. It derives from a 3alpha-hydroxy-5beta-androstan-17-one.		It is an L-alanine derivative consisting of an N-acetyl-D-muramoyl group attached to L-alanine via an amide linkage. It is a glyco-amino acid and a L-alanine derivative. It is a conjugate acid of a N-acetyl-D-muramoyl-L-alaninate.	

Table 10: Visualizations of description-molecule pairs in the ChEBI-20 (Edwards et al., 2021) and PCDes (Zeng et al., 2022) datasets.

The ChEBI-20 dataset consists of 33,008 description-molecule pairs, split into 26,407/3,301/3,300 pairs for train/validation/test (Pei et al., 2023). The PCDes dataset contains more challenging 15,000 description-molecule pairs, split into 10,500/1,500/3,000 pairs for train/validation/test (Zeng et al., 2022). Both are derived from qualified description-molecule pairs in the open-sourced PubChem database (Wang et al., 2009), where each text description describes the corresponding molecule's structure and chemical properties. In Table 9, we provide some visualizations of our curated 34k text-molecule pairs, which are introduced in Section 4.1. In Table 10, we visualize some description-molecule pairs of our main benchmark dataset, i.e., the ChEBI-20 (Edwards et al., 2021) and PCDes (Zeng et al., 2022) benchmarks.

D Additional experiments and analyses

D.1 Qualitative results

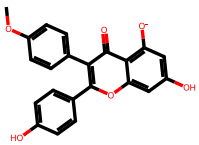
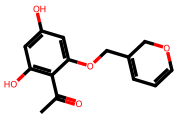
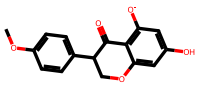
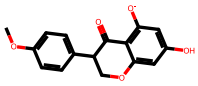
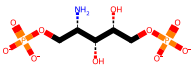
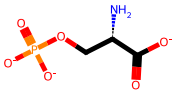
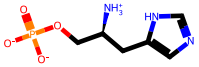
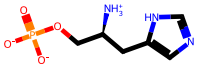
Description	MolT5	BioT5	**CAMT5 (Ours)**	Target
The molecule is conjugate base of 2,3-dihydrobiochanin A arising from selective deprotonation...	 RDK: 0.29	 RDK: 0.57	 RDK: **1.00**	
It is conjugate base of L-histidinol phosphate having an anionic phosphate and a catoinic amino group...	 RDK: 0.37	 RDK: 0.50	 RDK: **1.00**	

Table 11: Qualitative results of the text-to-molecule generation task in the CheBI-20 (Edwards et al., 2021) benchmark (the first row) and PCDes (Zeng et al., 2022) benchmark (the second row). For the best-performing models of MolT5, BioT5, and CAMT5, we visualize the generated molecules with respect to the given description. We report the RDK score (Schneider et al., 2015) between the generated and ground truth molecules below each visualization. We set the highest score in bold.

In Table 11, we provide some visualizations of the generated molecules from each text-to-molecule model. From these visualizations, we observe that our CAMT5 effectively generates molecules that contain crucial motifs of the target molecules, e.g., imidazole in the second row, which further demonstrates the importance of our motif-level tokenization scheme in CAMT5.

D.2 Results based on T5-large models

Model	**Exact** ↑	**MACCS** ↑	**RDK** ↑	**Morgan** ↑	**Validity** ↑
$\text{MolT5}^{\dagger}_{\text{large}}$	0.351	0.872	0.820	0.746	0.963
$\text{BioT5}^{\dagger}_{\text{large}}$	0.375	0.855	0.790	0.688	**1.000**
$\text{CAMT5}_{\text{large}}$ (Ours)	**0.430**	**0.885**	**0.840**	**0.749**	**1.000**

Table 12: Quantitative results on the ChEBI-20 (Edwards et al., 2021) benchmark. † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours. We highlight the best score in bold.

In Table 12, we report the results of the text-to-molecule models derived from the T5-large (Raffel et al., 2020) backbone model. Our $\text{CAMT5}_{\text{large}}$ outperforms the previous state-of-the-art text-to-molecule models, improving the Exact score by 0.375 → 0.430.

D.3 Performance varying the size of molecules

# Atoms	Model	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
$(0, 30]$	$\text{MolT5}^{\dagger}_{\text{base}}$	0.365	0.848	0.789	0.708	0.964
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.346	0.830	0.760	0.650	**1.000**
	CAMT5$_{\text{base}}$	**0.439**	**0.861**	**0.808**	**0.717**	**1.000**
$(30, 70]$	$\text{MolT5}^{\dagger}_{\text{base}}$	0.278	0.873	0.840	0.763	0.939
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.341	0.884	0.829	0.721	**1.000**
	CAMT5$_{\text{base}}$	**0.397**	**0.913**	**0.870**	**0.775**	**1.000**
$(70, \infty)$	$\text{MolT5}^{\dagger}_{\text{base}}$	0.194	0.826	0.811	0.749	0.854
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.291	0.924	0.887	0.761	**1.000**
	CAMT5$_{\text{base}}$	**0.369**	**0.951**	**0.931**	**0.843**	**1.000**

Table 13: Performance on ChEBI-20 (Edwards et al., 2021) grouped by number of atoms in molecules. † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours.

In Table 13, we analyze the generation performance on the ChEBI-20 benchmark grouped by the number of atoms in molecules. Our tokenization consistently outperforms other methods when working with both small and large molecules. This is likely due to the fact that our tokenization successfully incorporates both local and global molecular information.

D.4 Performance on atom-level descriptions

Description	Model	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
'chlor'	$\text{MolT5}^{\dagger}_{\text{base}}$	0.244	0.781	0.677	0.589	0.944
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.216	0.738	0.638	0.524	**1.000**
	CAMT5$_{\text{base}}$	**0.258**	**0.793**	**0.702**	**0.603**	**1.000**
'fluoro'	$\text{MolT5}^{\dagger}_{\text{base}}$	0.223	0.790	0.709	0.610	0.961
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.204	0.738	0.644	0.523	**1.000**
	CAMT5$_{\text{base}}$	**0.262**	**0.816**	**0.715**	**0.622**	**1.000**
'phospho'	$\text{MolT5}^{\dagger}_{\text{base}}$	0.371	0.911	0.872	0.824	0.976
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.510	0.910	0.854	0.799	**1.000**
	CAMT5$_{\text{base}}$	**0.614**	**0.952**	**0.916**	**0.865**	**1.000**
'sulf'	$\text{MolT5}^{\dagger}_{\text{base}}$	0.387	0.856	0.779	0.707	0.955
	$\text{BioT5}^{\dagger}_{\text{base}}$	0.300	0.831	0.744	0.632	**1.000**
	CAMT5$_{\text{base}}$	**0.424**	**0.882**	**0.825**	**0.740**	**1.000**

Table 14: Performance on ChEBI-20 (Edwards et al., 2021) containing specific atom-level descriptions. † denotes that the model is trained with the same training configuration, e.g., training dataset, as ours.

In Table 14, we evaluate the generation performance on ChEBI-20 that include atom-level descriptions. CAMT5 consistently outperforms MolT5 and BioT5 across various atom-level descriptions such as 'chlor', 'fluoro', 'phospho', and 'sulf', demonstrating its robustness in handling atom-specific information.

D.5 Comparison with alternative tokenization strategies

Tokenization	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
t-SMILES (Wu et al., 2024)	0.025	0.700	0.636	0.475	0.997
BRICS (Degen et al., 2008)	0.216	0.808	0.765	0.633	1.000
Ours	**0.391**	**0.874**	**0.827**	**0.727**	**1.000**

Table 15: Comparison of tokenization strategies on ChEBI-20 (Edwards et al., 2021) using the models derived from T5-small (Raffel et al., 2020).

In Table 15, we compare our tokenization strategy with previously proposed motif-aware tokenizations i.e, t-SMILES (Wu et al., 2024) and BRICS (Degen et al., 2008), following their official implementations. The result shows that our tokenization strategy achieves superior performance across all metrics.

D.6 Analysis on linearization algorithms in tokenization

Algorithm	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
Breadth-First Search (BFS)	0.368	0.858	0.808	0.707	1.000
Depth-First Search (DFS)	**0.391**	**0.874**	**0.827**	**0.727**	**1.000**

Table 16: Ablation of linearization algorithms in our tokenization strategy on ChEBI-20 (Edwards et al., 2021) using the models derived from T5-small (Raffel et al., 2020).

In Table 16, we compare the linearization algorithms used in our tokenization strategy (see Figure 4). We adopt depth-first search (DFS) as our traversal strategy, which is a common linearization algorithm in molecular serialization such as SMILES (Weininger, 1988). To verify whether this choice is indeed effective, we compare DFS with breadth-first search (BFS), another popular traversal method. The result shows that DFS consistently outperforms BFS across all evaluation metrics. We think that the nature of DFS traversal, which sequentially explores long, connected motif chains, facilitates the model's ability to capture the underlying structural patterns of molecules.

D.7 Ablation on token importance

Importance	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
Frequency of atoms	0.390	0.867	0.816	0.719	**1.000**
Frequency of motifs	0.281	0.811	0.758	0.638	**1.000**
Number of atoms	**0.391**	**0.874**	**0.827**	**0.727**	**1.000**

Table 17: Ablation on the training objective in importance-based training on ChEBI-20 (Edwards et al., 2021) using the models derived from T5-small (Raffel et al., 2020).

In Table 17, we compare the results varying the definition of imporatnce in Eq. (2), i.e., frequency of atoms and frequency of motifs. Specifically, frequency-based importance is defined as the inverse frequency of atoms or motifs in the training data prioritizing rare components. We find that our original choice of importance, i.e., the number of atoms, is the most effective among the candidates.

D.8 Ensemble performance analysis

MolT5	BioT5	CAMT5 (Ours)	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
✓	✓	✗	0.443	0.889	0.841	0.760	**1.000**
✓	✗	✓	0.455	0.891	0.849	0.768	**1.000**
✗	✓	✓	0.462	**0.902**	0.857	0.772	**1.000**
✓	✓	✓	**0.472**	**0.902**	**0.860**	**0.781**	**1.000**

Table 18: Performance of ensemble with different combinations of models on ChEBI-20 (Edwards et al., 2021). We report the results based on the best-performing model of MolT5, BioT5, CAMT5 in Table 2, respectively.

In Table 18, we compare the ensemble result based on different combinations of models. Intriguingly, the result shows that even the model with inferior performance, e.g., MolT5, is useful when it is ensembled with other models, i.e., the performance is improved by 0.462 → 0.472.

D.9 Data-efficient molecular generation

Importance	Active. ↑	FCD ↑	NSPDK ↑	Valid. ↑	Unique. ↑	Novelty ↑
MolT5 (Edwards et al., 2022)	11.2	18.7	0.020	70.4	87.2	**100**
BioT5 (Pei et al., 2023)	11.6	17.0	0.019	**100**	**96.8**	**100**
CAMT5 (Ours)	**12.0**	**16.3**	**0.018**	**100**	96.4	**100**

Table 19: Results on data-efficient generation (Kim et al., 2024) on the HIV dataset (Wu et al., 2018).

In Table 19, we compare the results on data-efficient molecular generation, which an important application of text-to-molecule models. CAMT5 outperforms other text-to-molecule models, verifying the potential of CAMT5 to other molecule-related applications.

D.10 Statistical analysis

Model	Exact ↑	MACCS ↑	RDK ↑	Morgan ↑	Validity ↑
MolT5 (Edwards et al., 2021)	0.278 ± 0.009	0.836 ± 0.006	0.779 ± 0.005	0.696 ± 0.005	0.946 ± 0.004
BioT5 (Pei et al., 2023)	0.324 ± 0.010	0.841 ± 0.002	0.771 ± 0.003	0.658 ± 0.004	**1.0** ± 0.0
CAMT5 (Ours)	**0.388** ± 0.003	**0.871** ± 0.003	**0.823** ± 0.004	**0.723** ± 0.003	**1.0** ± 0.0

Table 20: Comparison of the mean and standard deviation values based on the text-to-molecule models derived from T5-small (Raffel et al., 2020). The results are calculated over 3 runs with different seeds.

In Table 20, we report the mean and standard deviation values based on 3 independently trained text-to-molecule models. CAMT5 shows superior performance across the evaluation metrics, consistently achieving the higher average scores compared to the baselines.

D.11 Computational cost

Cost	Fine-tuning cost (hrs)	Memory (GB)	Inference cost (sec)
MolT5 (Edwards et al., 2021)	20	3.95	0.65
BioT5 (Pei et al., 2023)	19	3.95	0.61
CAMT5 (Ours)	**15**	**2.79**	**0.30**

Table 21: Comparison of computational costs among the models derived from T5-base (Raffel et al., 2020).

In Table 21, we provide the computational cost, including training costs, memory requirements, and inference costs. By treating multiple atoms as a single motif token, CAMT5 significantly reduces the token sequence length, leading to improved training and inference efficiency.

E Impact statement

This work will accelerate improvements in the field of text-to-molecule models, which will affect various applications such as drug discovery and material design. However, malicious usage of text-to-molecule models (including our models) may lead to a potential threat of generating harmful chemicals. We believe that safeguarding these models is an important future research direction, which is also widely studied in various domains (Achiam et al., 2023). We used AI assistants in coding and draft refinement, e.g., grammar check.