Automatic Annotation Augmentation Boosts Translation between Molecules and Natural Language

Anonymous ACL submission

Abstract

Recent advancements in AI for biological research focus on integrating molecular data with natural language to accelerate drug discovery. However, the scarcity of high-quality annotations limits progress in this area. This paper introduces LA^3, a Language-based Automatic Annotation Augmentation framework that leverages large language models to augment existing datasets, thereby improving AI training. We demonstrate the effectiveness of LA^3 by creating an enhanced dataset, LACHEBI-20, where we systematically rewrite the annotations of molecules from an established dataset. These rewritten annotations preserve essential molecular information while providing more varied sentence structures and vocabulary. Using LACHEBI-20, we train LAMOLT5 based on a benchmark architecture to learn the mapping between molecular representations and augmented annotations.

Experimental results on text-based *de novo* molecule generation and molecule captioning demonstrate that LAMOLT5 outperforms state-of-the-art models. Notably, incorporating LA^3 leads to improvements of up to 301% over the benchmark architecture. Furthermore, we validate the effectiveness of LA^3 notable applications in *image*, *text* and *graph* tasks, affirming its versatility and utility.

1 Introduction

Artificial Intelligence (AI) has garnered increasing attention due to its transformative potential in broad real-world applications, including biology (Schwaller et al., 2019; Pei et al., 2023; Zhong and Mottin, 2023). Take a recent new trend as an example: researchers intend to jointly model SMILES (Weininger, 1988) strings and scientific text to obtain shared representations across the two modalities. For instance, Edwards et al. (2022) innovatively propose MOLT5, a model based on the T5 architecture (Raffel et al., 2020), pre-trained on ZINC (Sterling and Irwin, 2015) by predicting the masked text parts. Consequently, they fine-tune the model on CHEBI-20 (Edwards et al., 2021) to learn how to map between SMILES representations of molecules and their corresponding annotations (captions) to support molecular tasks.

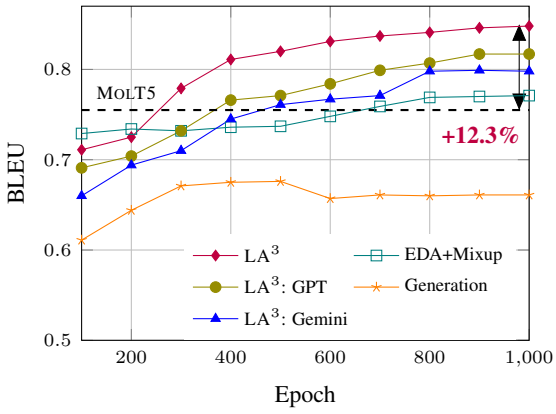


Figure 1: Molecule generation performance of LAMOLT5-Small with different LA^3 augmentations. Conventional augmentation (EDA (Wei and Zou, 2019), Mixup (Zhang et al., 2018)) and straightforward LLMs for data generation (Zhong et al., 2024a) fall behind.

However, the development of AI faces a fundamental setback: *the scarcity of high-quality annotated data*. For instance, molecular data annotation is often a costly and time-consuming process (DiMasi et al., 2016). This limitation restraints the development of AI approaches, as models grow in size and expressiveness, they require larger and more diverse annotated datasets to achieve high performance and generalisability (Devlin et al., 2019; He et al., 2023). Therefore, one viable alternative for AI in practice is to resort to *effective data augmentation strategies.*

Back to the example CHEBI-20 dataset, a set of studies attempted various solutions to address the scarcity limitation, following the MOLT5. Christofidellis et al. (2023), Liu et al. (2023b), Pei et al. (2023) and Pei et al. (2024) introduce extra

chemical databases and auxiliary tasks to train advanced models. However, the heavy dependencies on supplemental datasets limit their practical versatility, and the auxiliary task definition requires domain expertise. On the other hand, Li et al. (2024) use the general human knowledge embedded in Large Language Models (LLMs) to perform the molecule-caption translation tasks. Nevertheless, despite the widespread use of LLMs in literature review, table interpretation, *etc*, their application in biology is not straightforward (Laurent et al., 2024). The effectiveness of (Li et al., 2024) depends heavily on the specific retrieval strategy used, and its performance is surpassed by existing smaller models. Additionally, Zhong et al. (2024a) and our empirical studies in Figures 1,3 demonstrate the limitations of using LLMs for data augmentation.

This paper proposes an effective automatic pipeline (see Figure 2), LA^3, to effectively augment annotations of datasets with no human supervision. Once the augmented datasets are generated, existing methods can be conveniently re-trained for significant performance boosting.

We showcase the effectiveness of LA^3 by creating an enhanced dataset, LACHEBI-20, where we leverage the in-context learning (Liu et al., 2023a) capability of LLMs to rewrite the annotation of each molecule in CHEBI-20. These rewritten annotations preserve essential molecular information while providing more varied sentence structures and vocabulary (see our analysis in Section 4.4). After the annotation augmentation process, each molecule in LACHEBI-20 is accompanied by diverse annotations. Using these annotations, we proceed to train LAMOLT5, using the benchmark MOLT5 architecture, to support molecular tasks. During training, LAMOLT5 aims to learn a mapping function between the space of molecules and the augmented annotations, thereby enhancing the overall performance of the models.

We systematically evaluate the effectiveness of LAMOLT5 on challenging text-based *de novo* molecule generation and molecule captioning tasks. Through extensive experiments on the benchmark evaluation pipeline, we demonstrate that LAMOLT5 significantly elevates the performance of MOLT5, which was trained using the same architecture on the original CHEBI-20 dataset. Notably, LAMOLT5 achieves improvements of up to 301% on the molecule generation task and 9.51% on the molecule captioning task. Additionally, the small-size variant of LAMOLT5 ($77M$ parameters) outperforms the large-size variant of MOLT5 ($800M$ parameters) for the molecule generation task. Compared with other leading methods reported on the leaderboard, LAMOLT5 achieves new state-of-the-art performance with 99% fewer parameters. More importantly, LA^3 effectively boosts the performance of other applications, including *image captioning*, *text understanding* and *graph property prediction*, affirming its versatility.

Our contributions are as follows: **(1)** A fully automated pipeline for domain-specific applications where limited data availability restricts the effectiveness of existing technologies. **(2)** A set of lightweight open-source models tailored to address challenging molecular tasks. **(3)** Empirical studies demonstrating the necessity and effectiveness of LA^3 across multiple applications.

2 Related Work

Molecule Language Models. MLMs have recently seen significant advancements, leveraging NLP techniques to understand and generate molecules. Early works such as ChemBERTa (Chithrananda et al., 2020) and Text2Mol (Edwards et al., 2021) adapt transformer-based architectures for molecular representation learning. MolGPT (Bagal et al., 2022) and MOLT5 (Edwards et al., 2022) demonstrate the ability to predict molecular properties and generate novel compounds, highlighting the potential of language models in biological research. However, the effectiveness of MLMs is often constrained by the limited availability of annotated molecular data. Meanwhile, manual molecular data annotation is often a costly and time-consuming process, necessitating specialised equipment and extensive human labour (DiMasi et al., 2016).

Data Augmentation in MLMs. Data augmentation has emerged as a critical strategy to address the scarcity of high-quality datasets. Take the following studies on CHEBI-20, Christofidellis et al. (2023), Liu et al. (2023b), Pei et al. (2023) and Pei et al. (2024) introduce additional chemical databases (PubChem (Kim et al., 2023), Drugbank (Wishart et al., 2018), UniProt (Consortium, 2023), PubMed (White, 2020), *etc*) as to enrich model with extra knowledge, and design auxiliary tasks to train advanced models. However, these methods depend on supplemental datasets and the domain expertise required to shape the tasks. On the other hand, LLMs have experienced exponential growth in both size and capability in recent

years (Brown et al., 2020). A wide range of NLP applications have been reshaped by LLMs (Hendy et al., 2023; Zhong et al., 2024b). Notably, MolReGPT (Li et al., 2024) leverage the built-in general human knowledge of LLMs to perform the molecule-caption translation tasks. Despite the widespread use of LLMs for tasks like literature review and table interpretation (Achiam et al., 2023; Touvron et al., 2023), their application in biology remains challenging (Laurent et al., 2024). The effectiveness of MolReGPT is highly dependent on the chosen retrieval strategy, and it is often outperformed by smaller, existing models.

In another related work, Wang et al. (2024) utilise LLMs and human annotators to augment text for each image to improve contrastive learning of CLIP (Radford et al., 2021), yet these pipelines still require human supervision. Furthermore, Zhong et al. (2024a) reveal the limited capability of LLMs in understanding domain-specific data, *e.g.*, biology and physics, making LLM-based synthetic data generation challenging in many applications (see our analysis in Section 4.4). In contrast, our novel LA^3 pipeline is fully automated and tailored for domain-specific applications where data scarcity limits the effectiveness of current technologies.

3 Methodology

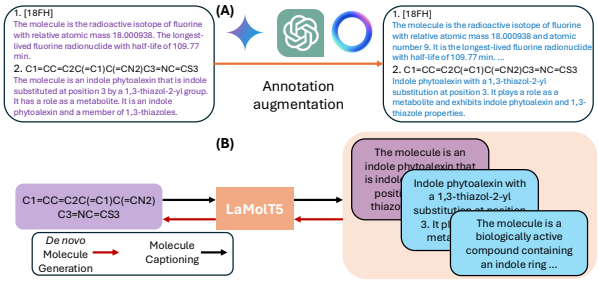


Figure 2: An example implementation of LA^3 for annotation augmentation (A) and training (B). Given molecules and their original annotations, we prompt LLMs to generate augmented annotations (LACHEBI-20) by rewriting the original annotations. Next, we train LAMOLT5 on LACHEBI-20 to learn a mapping function between the molecule's SMILES string and corresponding annotations.

To facilitate the practical re-implementation, we showcase details of LA^3, using CHEBI-20 (Edwards et al., 2021), a widely adopted dataset for molecular generation research (Section 3.1-3.3). In addition, Section 3.4 describes extensive implementations of LA^3 across broad applications in *image*, *text* and *graph* tasks.

3.1 Showcasing CHEBI-20

CHEBI-20. CHEBI-20 contains 33 010 molecular entities centred on chemical compounds. Each molecule is represented using a SMILES string (Weininger, 1988) and associated with a high-quality, manually annotated caption supporting various computational and experimental studies. Given a molecule, we formally represent it as $\mathcal{M} = (S, C)$, where S and C denote its SMILES string and associated caption. Examples are illustrated in Figure 2-(A). Consequently, CHEBI-20 can be formally represented as $\mathcal{D} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n\}$. $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ represents the SMILES string set and $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ the caption (annotation) set. The dataset is publicly available with a fixed split: $\mathcal{D}_{\text{TRAIN}}$ (80%), $\mathcal{D}_{\text{VALID}}$ (10%) and $\mathcal{D}_{\text{TEST}}$ (10%), allowing researchers to consistently train and evaluate their models.

Tasks. CHEBI-20 supports two molecular tasks: *(1)* text-based *de novo* molecule generation (GEN) and *(2)* molecule captioning (CAP). The goal of text-based *de novo* molecule generation is to train a model which can generate a variety of possible new molecules with desired properties as described in the text. Specifically, for CHEBI-20, we aim to learn a model $f_{\text{GEN}} : \mathcal{C} \rightarrow \hat{\mathcal{S}}$ by minimising the loss function value $\min_{\Psi} \mathcal{L}(\hat{\mathcal{S}}_{\text{TRAIN}}, \mathcal{S}_{\text{TRAIN}})$, where Ψ represents the set of trainable parameters of f_{GEN}. The target of molecule captioning is to generate descriptions of the components and chemical functionality of a molecule. Similarly, we aim to learn a model $f_{\text{CAP}} : \mathcal{S} \rightarrow \hat{\mathcal{C}}$, by minimising loss function value $\min_{\Phi} \mathcal{L}(\hat{\mathcal{C}}_{\text{TRAIN}}, \mathcal{C}_{\text{TRAIN}})$, where Φ represents the set of trainable parameters of f_{CAP}.

MOLT5. The molecule generation tasks can be considered translation tasks, and sequence-to-sequence models serve as solid solutions. One fundamental method in this category is MOLT5 (Edwards et al., 2022), an improved version of T5 (Raffel et al., 2020). MOLT5 initialise an encoder-decoder transformer model using public checkpoints of T5. The model is then pre-trained on a combined dataset of C4 (Raffel et al., 2020) and ZINC (Sterling and Irwin, 2015) for 1 million steps. Finally, it undergoes 50,000 steps of fine-tuning on CHEBI-20 for two molecular tasks. Since these tasks are formulated as sequence-to-sequence tasks, the model is trained using standard maximum likelihood, such as cross-entropy loss ($\mathcal{L}_{\text{CE}}$). Take the fine-tuning phase as an example. The parameters are optimised to match the generated text with

$\mathcal{D}_{\text{TRAIN}}$'s text:

$$\mathcal{L}_{\text{CE}}^{\text{GEN}}(\mathcal{D}_{\text{TRAIN}}) = -\frac{\sum_{i=1}^{n} \log p(S_{\text{TRAIN},i} \mid C_{\text{TRAIN},i})}{n},$$
$$\mathcal{L}_{\text{CE}}^{\text{CAP}}(\mathcal{D}_{\text{TRAIN}}) = -\frac{\sum_{i=1}^{n} \log p(C_{\text{TRAIN},i} \mid S_{\text{TRAIN},i})}{n}, \quad (1)$$

where n is the number of molecules in $\mathcal{D}_{\text{TRAIN}}$ and $p(S_{\text{TRAIN},i} \mid C_{\text{TRAIN},i})$ is the probability assigned by f_{GEN} to the i-th true SMILES string $S_{\text{TRAIN},i}$ given the i-th caption $C_{\text{TRAIN},i}$. This optimisation increases the probability of generating the correct outputs given the corresponding inputs. Following this way, MOLT5 provides three trained variants of varying sizes: MOLT5-Small ($77M$ parameters), MOLT5-Base ($250M$ parameters) and MOLT5-Large ($800M$ parameters).

3.2 Automatic Annotation Augmentation

As shown in Equation 1, the number of training instances directly affects the amount of information injected into the model. Edwards et al. (2022) discuss the potential limitations caused by the limited data in CHEBI-20. A recent breakthrough known as In-Context Learning (ICL) has enhanced the adaptability of LLMs by enabling them to acquire contextual knowledge during inference, eliminating the need for extensive fine-tuning (Clark et al., 2020). To harness ICL for CHEBI-20 augmentation, we first formulate a prompt for querying LLMs. The goal in prompt engineering is to find the correct way to formulate a question Q in such a way that an LLM (f_{LLM}) will return the corresponding answer A essentially represented as $A = f_{\text{LLM}}(Q)$. In this work, we design the prompt as shown in Appendix A. Precisely, the prompt consists of two components: *Instruction*: Provides general guidance to the LLM, clarifying its role in the conversation; *Message*: Tasks the LLM to rewrite the molecule caption, considering the chemical expertise and given information.

Given an instance $\mathcal{M}_i = (S_i, C_i)$ from $\mathcal{D}_{\text{TRAIN}}$, we can generate k rewritten captions $\{C_{i,1}, C_{i,2}, \ldots, C_{i,k}\}$ with multiple rounds of queries. This results in an augmented instance, $\mathcal{M}_i^+ = (S_i, C_{i,0}, C_{i,1}, C_{i,2} \ldots, C_{i,k})$, and an augmented dataset, LACHEBI-20 ($\mathcal{D}^+$). Specifically, $\mathcal{D}^+ = (\mathcal{D}_{\text{TRAIN}}^+, \mathcal{D}_{\text{VALID}}, \mathcal{D}_{\text{TEST}})$, where $\mathcal{D}_{\text{TRAIN}}^+ = \{\mathcal{M}_1^+, \mathcal{M}_2^+, \ldots\}$. Each SMILES string of $\mathcal{D}_{\text{TRAIN}}^+$ is associated with $k+1$ captions. This language augmentation process introduces diversity in sentence structure and vocabulary while preserving the essential knowledge about the molecules. In our experiments, we adopt two open-source LLMs (Llama 2-70b (Touvron et al., 2023) and Llama 3-70b (Touvron et al., 2023)) and two closed-source LLMs (GPT 3.5-turbo (Achiam et al., 2023) and Gemini Pro (Google, 2024)) to generate four rewritten captions. We demonstrate some generated example captions in Table 6 of Appendix B.

3.3 Training on Augmented Dataset

After generating k new captions for each molecule of the training dataset $\mathcal{D}_{\text{TRAIN}}$. We proceed to train a model based on LACHEBI-20 to perform the molecular tasks, *i.e.*, text-based *de novo* molecule generation (GEN) and molecule captioning (GEN). In this work, we initialise encoder-decoder transformer models using the available MOLT5 variants, as introduced in Section 3.1. We then train a novel model, LAMOLT5, using a cross-entropy loss:

$$\mathcal{L}_{\text{CE}}^{\text{GEN}}(\mathcal{D}_{\text{TRAIN}}^+) = -\frac{\sum_{i=1}^{n} \sum_{j=1}^{k+1} \log p(S_{\text{TRAIN},i} | C_{\text{TRAIN},i,j}^+)}{(k+1)n},$$
$$\mathcal{L}_{\text{CE}}^{\text{CAP}}(\mathcal{D}_{\text{TRAIN}}^+) = -\frac{\sum_{i=1}^{n} \sum_{j=1}^{k+1} \log p(C_{\text{TRAIN},i,j}^+ | S_{\text{TRAIN},i})}{(k+1)n}, \quad (2)$$

where $p(S_{\text{TRAIN},i} \mid C_{\text{TRAIN},i,j}^+)$ is the probability assigned by f_{GEN} to the i-th true SMILES string $S_{\text{TRAIN},i}$ given the i-th molecule's j-th caption $C_{\text{TRAIN},i,j}$. The critical addition to the MOLT5 is the augmented caption set $C_{\text{TRAIN},i,\{0,1,\ldots,k\}}^+$ for each SMILES string $S_{\text{TRAIN},i}$. For the molecule generation task, LAMOLT5 is trained to generate the correct SMILES string $S_{\text{TRAIN},i}$ by giving different caption inputs $C_{\text{TRAIN},i,\{0,1,\ldots,k\}}^+$. By introducing diversity into the training data, we enhance the effectiveness and robustness of the model in generating SMILES strings (validated in Section 4.2). Meanwhile, for the molecule captioning task, LAMOLT5 is trained to generate various captions for each SMILES string. Despite these captions having different sentence structures and vocabularies, they preserve the essential knowledge about the molecules. This training enhances the model's ability to generate more semantically precise and meaningful captions (validated in Section 4.3). Overall, this novel and effective methodology augments the biomedical dataset without re-

quiring human efforts, significantly contributing to the performance of LAMOLT5.

3.4 Extension to Broad Applications

To further demonstrate the versatility of LA^3, we extend its application to several additional datasets, including `ogbg-molbace`, `ogbg-molhiv`, `ogbg-molesol` (Hu et al., 2020), and `CC3M` (Sharma et al., 2018). These datasets support a variety of crucial tasks, *e.g.*, *image captioning*, *text understanding*, and *graph property prediction*. Due to the paper length constraints, the detailed implementations of LAMOLT5 on these additional datasets are presented in Appendix F. Experimental results derived from these implementations will be discussed in Section 4.5.

4 Experiments

4.1 Experimental Settings

Dataset. We use our generated LACHEBI-20 dataset for training LAMOLT5. One annotation is the original present in CHEBI-20, and the other two are LLM-generated. We adopt two conventional LLMs, GPT 3.5-turbo and Gemini Pro, to generate annotations in the main dataset from most experiments. We additionally experiment in Section 4.4 with annotations generated by two open-source LLMs, *e.g.*, Llama 2-70b and Llama 3-70b.

Baselines. We mainly consider three families of methods: **(1)** Methods included in the benchmark paper (Edwards et al., 2022), including RNN (Cho et al., 2014), Transformer (Vaswani et al., 2017), T5 (Raffel et al., 2020) and MOLT5 (Edwards et al., 2022). **(2)** The state-of-the-art methods (reported on the leaderboards (den, 2024; cap, 2024)) without additional datasets, including Text+Chem T5 (Christofidellis et al., 2023), TGM-DLM (Gong et al., 2024) and MolReGPT (Li et al., 2024). They rely on the same datasets as LAMOLT5, we report their results in Table 1, 3 **(3)** The state-of-the-art methods incorporating extra datasets. For the text-based *de novo* molecule generation task, we include BioT5 (Pei et al., 2023) and BioT5+ (Pei et al., 2024); for the molecule captioning task, we include BioT5, MolXPT (Liu et al., 2023b). For a fair comparison, we report their results in Figure 3. A detailed description of each baseline method can be found in Appendix D.

Training Setup. We train LAMOLT5-Small and -Base for as little as 1500 epochs and LAMOLT5-Large for 200 epochs. This project used $\sim$ 101500 H800 GPU hours. Detailed hyper-parameter settings and checkpoints are available online[1].

Evaluation Setup. We adopt the benchmark evaluation setup following (Edwards et al., 2022). For the text-based *de novo* molecule generation task, we employ: SMILES *BLEU* score (Papineni et al., 2002), *Levenshtein* distance (Miller et al., 2009), Fréchet ChemNet Distance (*FCD*) (Preuer et al., 2018), *MACCS FTS* (Durant et al., 2002), *RDK FTS* (Schneider et al., 2015) *Morgan FTS* (Rogers and Hahn, 2010), *Exact* score (Edwards et al., 2022), *Validity* (Edwards et al., 2022), and *Text2Mol* (Edwards et al., 2021). For the molecule captioning task, we measure the *BLEU* score, *ROUGE* (Lin, 2004), *METEOR* (Banerjee and Lavie, 2005), and *Text2Mol* of the generated annotation compared to the ground-truth. Detailed descriptions can be found in Appendix C.

4.2 Molecule Generation

Observation 1: LAMOLT5 significantly elevates the performance of MOLT5. The *de novo* molecule generation results across nine evaluation metrics in Table 1 reveal that LAMOLT5 achieves the highest performance on all measures. In addition, LAMOLT5 consistently delivers substantial enhancements over MOLT5, with improvements up to 301% in terms of Exact score, which measures the number of times the output corresponds to the ground truth. This consistent and notable performance underscores the effectiveness of LACHEBI-20 and LAMOLT5. Moreover, it illustrates the efficacy of our automatic annotation-augmentation pipeline, LA^3, on biomedical datasets.

Observation 2: The small $77M$ parameters LAMOLT5 outperforms the $800M$ MOLT5. The results of Table 1 indicate that the small-size variant of LAMOLT5 ($77M$) outperforms the large-size variant of MOLT5 ($800M$) in seven different evaluation metrics (Levenshtein, MACCS FTS, RDK FTS, Morgan FTS, FCD, Text2Mo, Validity) on the molecule generation task. On the other two evaluation metrics (BLEU and Exact), LAMOLT5-Small archives competitive performance compared to MOLT5-Large. LAMOLT5 achieves impressive results by leveraging the annotation-augmented dataset, LACHEBI-20, which introduces diversity in sentence structure and vocabulary while maintaining the core molecular knowledge.

[1]The augmented dataset and trained models are available at `https://anonymous.4open.science/r/LaMolT5-D3C3`.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.34	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	16.03	0.277	0.906
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	1.77	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	1.35	0.482	0.721
LaMolT5-Small	**0.852**	**0.287**	**16.009**	**0.891**	**0.805**	**0.741**	**0.384**	**0.594**	**0.950**
(Improvement)	+12.85%	+263.29%	+38.42%	+26.72%	+41.55%	+43.33%	+71.56%	+23.24%	+31.74%
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	1.43	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	1.16	0.496	0.772
LaMolT5-Base	**0.861**	**0.325**	**14.685**	**0.899**	**0.819**	**0.760**	**0.352**	**0.596**	**0.961**
(Improvement)	+11.96%	+301.23%	+39.95%	+24.69%	+39.29%	+43.67%	+69.66%	+20.16%	+24.48%
T5-Large	0.854	0.279	16.721	0.823	0.731	0.670	0.401	0.552	0.902
MolT5-Large	0.854	0.311	16.071	0.834	0.746	0.684	0.385	0.554	0.905
LaMolT5-Large	**0.856**	**0.328**	**15.666**	**0.892**	**0.816**	**0.754**	**0.371**	**0.593**	**0.962**
(Improvement)	+0.23%	+5.46%	+2.52%	+6.9%	+9.38%	+10.23%	+3.64%	+7.04%	+6.30%
Text+Chem T5	0.750	0.212	27.39	0.874	0.767	0.697	0.499	0.574	0.792
TGM-DLM	0.826	0.242	17.00	0.854	0.739	0.688	0.770	0.581	0.871
MolReGPT	0.857	0.280	17.14	0.903	0.805	0.739	0.410	0.593	0.899

Table 1: Text-based *de novo* molecule generation results for models without additional datasets. Models incorporating extra datasets are presented in Figure 3. Best performances are highlighted with an underline.

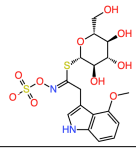
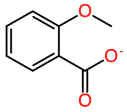
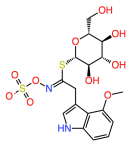
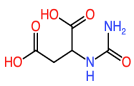
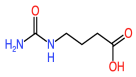
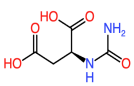
Caption	Ground Truth	MolT5-Small	LaMolT5-Small
(1): The molecule is an indolylmethylglucosinolate that is the conjugate base of 4-methoxyglucobrassicin, obtained by deprotonation of the sulfo group. It is a conjugate base of a 4-methoxyglucobrassicin.			
(2): The molecule is an N-carbamoylamino acid that is aspartic acid with one of its amino hydrogens replaced by a carbamoyl group. It has a role as a Saccharomyces cerevisiae metabolite, an Escherichia coli metabolite and a human metabolite. It is a N-carbamoyl-amino acid, an aspartic acid derivative and a C4-dicarboxylic acid. It is a conjugate acid of a N-carbamoylaspartate(2-).			

Table 2: Example of molecules generated with the same input descriptions. Generated SMILES strings are converted to molecule graphs for better visualisation.

4.3 Molecule Captioning

Observation 3: LaMolT5 generates coherent descriptions. The results in Table 3 highlight the superior performance of the LaMolT5 in the molecule captioning task. LaMolT5 excels in the Text2Mol metric, which provides a comprehensive assessment of the *semantic alignment* —throuch cosine similarity– between generated descriptions and their corresponding molecules. LaMolT5 variants achieve improvements up to 23% in LaMolT5-Small over the corresponding MolT5 variants. These results underscore LaMolT5 's enhanced ability to capture the intricate semantics of molecule descriptions, making it a highly effective model for this task.

Observation 4: LaMolT5 exhibits lower ROUGE score. LaMolT5 exhibits lower ROUGE scores than MolT5, as ROUGE emphasises exact n-gram overlaps, which do not fully capture the semantic accuracy of the generated text. The LaChEBI-20, by introducing variability in sentence structure and vocabulary, contributes to this discrepancy. During training, LaMolT5 prioritises achieving higher semantic coherence, potentially at the expense of exact word or phrase matches. As we argue above, although LaMolT5 excels in capturing the overall semantics of molecule descriptions, this focus leads to reduced performance on ROUGE scores.

4.4 Analysis

Comparison with state-of-the-art methods. Figure 3 demonstrates the performance of top-3 leaderboard SOTA models. LaMolT5 variants take the first two overall ranks. Notably, LaMolT5 establishes new SOTA results on molecule generation task. Although BioT5 outperforms LaMolT5 on the molecule captioning task, *BioT5 leverages external data*, offering an additional advantage. When comparing models that do not incorporate external knowledge, LaMolT5 achieves the best performance, solidifying its position as the top model

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑	Text2Mol↑
Ground Truth	1.000	1.000	1.000	1.000	1.000	1.000	0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.501	0.415	0.602	0.446	0.545	0.532	0.526
MolT5-Small	0.519	0.436	**0.620**	**0.469**	**0.563**	0.551	0.540
LAMOLT5-Small	**0.539**	**0.446**	0.610	0.446	0.538	**0.566**	**0.588**
(Improvement)	+3.85%	+2.29%	-1.61%	-4.90%	-4.44%	+2.72%	+8.89%
T5-Base	0.511	0.423	0.607	0.451	0.550	0.539	0.523
MolT5-Base	0.540	0.457	**0.634**	**0.485**	**0.578**	0.569	0.547
LAMOLT5-Base	**0.574**	**0.485**	**0.634**	0.478	0.564	**0.596**	**0.599**
(Improvement)	+6.30%	+6.13%	0.00%	-1.44%	-2.42%	+4.75%	+9.51%
T5-Large	0.558	0.467	0.630	0.478	0.569	0.586	0.563
MolT5-Large	0.594	0.508	0.654	0.510	0.594	0.614	0.582
LAMOLT5-Large	**0.602**	**0.521**	**0.655**	**0.512**	**0.598**	**0.634**	**0.597**
(Improvement)	+1.35%	+2.56%	+0.15%	+0.39%	+0.67%	+3.26%	+2.58%
Text+Chem T5	0.580	0.490	0.647	0.498	0.586	0.604	0.567
MolReGPT	0.607	0.525	0.634	0.476	0.562	0.610	0.585

Table 3: Molecule captioning results for models without additional datasets. Models incorporating extra datasets are presented in Figure 3. Best performances are highlighted with an underline.

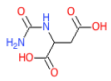
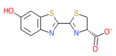
ID	Molecule	Ground Truth	MOLT5-Small	LAMOLT5-Small
1		The molecule is an N-carbamoylamino acid that is aspartic acid with one of its amino hydrogens replaced by a carbamoyl group. It has a role as a Saccharomyces cerevisiae metabolite, an Escherichia coli metabolite and a human metabolite. It is a N-carbamoyl-amino acid, an aspartic acid derivative and a C4-dicarboxylic acid. It is a conjugate acid of a N-carbamoylaspartate(2-).	The molecule is a member of the class of ureas that is urea in which one of the amino hydrogens is replaced by a carbamoyl group. It has a role as a metabolite. It is a N-acyl-amino acid and a member of ureas.	The molecule is a member of the class of ureas that is malonic acid in which one of the hydrogens attached to the nitrogen is substituted by a carbamoyl group. It has a role as an Escherichia coli metabolite and a mouse metabolite. It is a member of ureas and a member of ureas. It derives from a malonic acid. It is a conjugate acid of a N-carbamoylglycinate.
2		The molecule is the conjugate base of Photinus luciferin; major species at pH 7.3. It is a conjugate base of a Photinus luciferin. It is an enantiomer of an ent-Photinus luciferin(1-).	The molecule is a monocarboxylic acid anion that is the conjugate base of ent-Photinus luciferin, obtained by deprotonation of the carboxy group; major species at pH 7.3. It is a conjugate base of an ent-Photinus luciferin. It is an enantiomer of a Photinus luciferin(1-).	The molecule is a monocarboxylic acid anion that is the conjugate base of Photinus luciferin, obtained by deprotonation of the carboxy group; major species at pH 7.3. It is a conjugate base of a Photinus luciferin. It is an enantiomer of a Photinus luciferin(1-).

Table 4: Example of captions generated with the same input SMILES strings. Input SMILES strings are converted to molecule graphs for better visualisation.

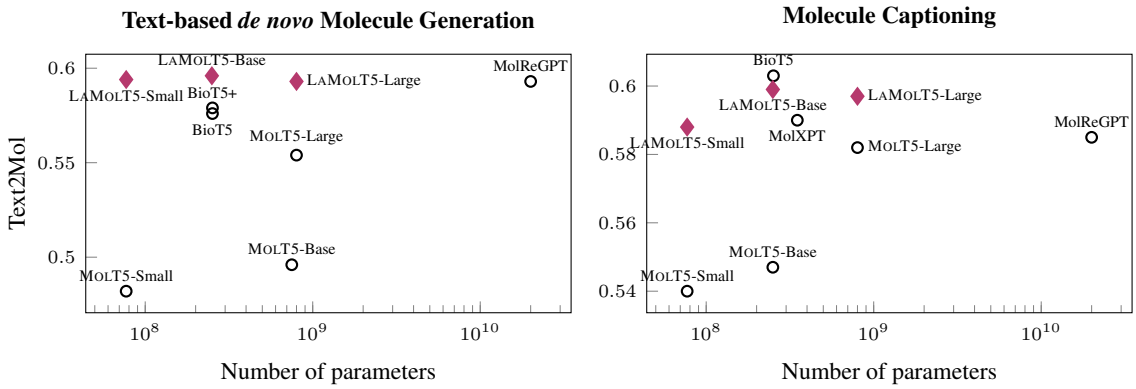


Figure 3: Performance vs. Number of parameters of LAMOLT5 and top-3 leaderboard state-of-the-art methods. Overall rank: LAMOLT5-Base (#1), LAMOLT5-Large (#2) and BioT5 (#3).

in this domain. Moreover, the small-size variant delivers highly competitive results with significantly fewer parameters than other leading models: LAMOLT5-Small has 99% fewer parameters than MolReGPT but delivers superior performance. This efficiency makes LAMOLT5-Small an attractive option for applications requiring high performance with reduced computational resources.

Performance with different augmentations. Figure 1 shows the performance of LAMOLT5 with different augmentation strategies during training. *(i)* The combination of two augmented annotations generated by LA^3 demonstrates a consistent improvement to using only one augmentation in performance, leading to more robust learning. *(ii)* Conventional data augmentation strategies, *e.g.*, EDA (Wei and Zou, 2019), Mixup (Zhang et al., 2018) are not feasible solutions in the context of biological data. We argue that the pre-training stage involved general knowledge about molecules, and easy text wrapping does not provide enough diversity to enhance the information. *(iii)* Relying on captions generated directly by LLMs not only fails to improve performance but actually degrades it. Specifically, including these directly generated annotations results in lower BLEU scores than the original MOLT5, underscoring that such annotations might introduce noise or lack the molecular knowledge needed for effective training. Such results echo the empirical studies of (Zhong et al., 2024a) and affirm the necessity of LA^3 design.

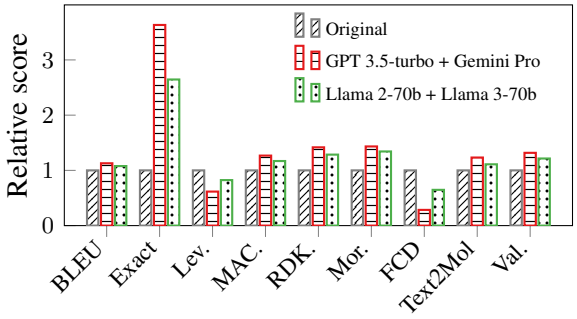


Figure 4: Molecule generation performance of MOLT5-Small and LAMOLT5-Small with captions generated by open-sourced and closed-sources LLMs.

Performance of LA^3 using different LLMs. Figure 4 show the molecule generation performance based on annotations augmented by different LLMs. LAMOLT5 trained on LACHEBI-20 annotations generated by both open- and closed-source LLMs consistently outperforms MOLT5. This highlights the versatility of our proposed annotation augmentation pipeline, LA^3, in practical applications. Expectedly, LAMOLT5 trained on closed-source LLMs (GPT 3.5-turbo + Gemini Pro) outperforms the one trained on open-source LLMs (Llama 2-70b + Llama 3-70b). As LLMs continue to improve in performance and in-context learning capabilities, LAMOLT5 can benefit directly from these advancements.

4.5 Broad Applications

Data Eval. Metric	Task	Model	Performance Orig.	 LA^3
ogbg-molbace ROC-AUC ↑	Class.	LM GNN	0.6163 0.7147	**0.6589** **0.7760**
ogbg-molhiv ROC-AUC ↑	Class.	LM GNN	0.5037 0.7376	**0.5562** **0.7641**
ogbg-molesol RMSE ↓	Reg.	LM GNN	2.2549 1.2561	**2.1811** **0.9301**
CC3M Accuracy ↑	Class.	CNN	15.8	**17.7**

Table 5: Results on *image*, *text* and *graph* tasks. LM: DeBERTa (He et al., 2023); GNN: GCN (Kipf and Welling, 2017); CNN: ViT-B/16 (Dosovitskiy et al., 2021).

Results of *image*, *text* and *graph* tasks. To further demonstrate the versatility of LA^3, we perform extended experiments on several additional datasets, including ogbg-molbace, ogbg-molhiv, ogbg-molesol, and CC3M, which support a variety of crucial tasks, *e.g.*, *image captioning*, *text understanding*, and *graph property prediction*. Due to the page limit, we describe the detailed implementations in Appendix F. Results in Table 5 show that LA^3 significantly enhances performance across these diverse applications. This improvement highlights LA^3 's potential to be a valuable tool in a wide range of AI tasks, offering substantial gains in accuracy and efficiency.

5 Concluding Discussion

This work proposes an automatic annotation augmentation pipeline, LA^3, designed to enhance annotated datasets and thereby boost the performance of AI approaches with minimal cost. We generate LACHEBI-20, an enriched biomedical dataset featuring diverse sentences and vocabulary while preserving essential molecular knowledge. This increased diversity is crucial for training LAMOLT5 models, leading to remarkable improvements in challenging molecular tasks. A set of ablation studies investigate the impact of LA^3 design and affirm its effectiveness. Furthermore, we expand the application of LA^3 to a wide range of datasets across different domains, including *image captioning*, *text understanding*, and *graph property prediction*. The observed significant improvements vindicate the remarkable versatility and utility of LA^3.

6 Limitations and Ethic Statement

Limitations. The language augmentation process relies on external LLMs, which introduce uncertainties because their robustness in other applications cannot be guaranteed. Additionally, results shown in Figure 4 demonstrate the impact of caption quality on molecular tasks, suggesting that developing techniques for filtering captions could be a valuable direction for future work. Moreover, while LLMs continue to improve in performance and ICL capabilities, LAMOLT5 can benefit from these advancements. However, the domain-specific knowledge embedded in LLMs remains relatively limited. Thus, exploring practical solutions to incorporate more comprehensive domain knowledge into LLMs for language augmentation is a promising future direction for enhancing LA^3.

Ethic Statement. Throughout our work, we did not utilise any private or sensitive information. The involved datasets are open-source, and outputs are available online to the community. However, it's essential to note that if any private information were to be inadvertently exposed to an LLM during internal pre-training and fine-tuning stages, LA^3 does not offer any privacy filtration mechanism. Therefore, there exists the potential for privacy concerns associated with the underlying model to manifest through the output provided by LA^3.

References

2024. Molecule captioning on chebi-20. Accessed: 2024-10-14.

2024. Text based de novo molecule generation. Accessed: 2024-10-14.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *CoRR*, abs/2307.09288.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 2020 Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1877–1901.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property predictio. *CoRR*, abs/2010.09885.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. ACL.

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 *Proceedings of the 2023 International Conference on Machine Learning (ICML)*, volume 202, pages 6140–6157. PMLR.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the 2020 International Joint Conferences on Artifical Intelligence (IJCAI)*, pages 3882–3890.

The UniProt Consortium. 2023. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. ACL.

Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. 2016. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of Health Economics*, 47:20–33.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias

Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 2021 International Conference on Learning Representations (ICLR)*.

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 Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 375–413. ACL.

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 (EMNLP)*, pages 595–607. ACL.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Training language models to follow instructions with human feedback. In *Proceedings of the 2023 Annual Conference on Neural Information Processing Systems (NeurIPS)*.

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

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 2023 International Conference on Learning Representations (ICLR)*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *CoRR*, abs/2302.09210.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *Proceedings of the 2020 Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 22118–22133.

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.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR)*.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2020. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Machine Learning: Science and Technology*, 1(4):45024.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *CoRR*, abs/2407.10362.

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 (TKDE)*.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*, 55(9):1–35.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. Molxpt: Wrapping molecules with text for generative pre-training. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1606–1616. ACL.

Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 2002 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. ACL.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. Biot5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1216–1240. ACL.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1102–1123. ACL.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. 2018. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 2021 International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763. PMLR.

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*, 21:140:1–140:67.

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.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 2018 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2556–2565. ACL.

Teague Sterling and John J. Irwin. 2015. ZINC 15 - ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 2017 Annual Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *CoRR*, abs/2401.00368.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6381–6387. ACL.

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.

Jacob White. 2020. Pubmed 2.0. *Medical Reference Services Quarterly*, 39(4):382–387.

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.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *Proceedings of the 2018 International Conference on Learning Representations (ICLR)*.

Zhiqiang Zhong and Davide Mottin. 2023. Knowledge-augmented graph machine learning for drug discovery: From precision to interpretability. In *Proceedings of the 2023 ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5841–5842. ACM.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024a. Benchmarking large language models for molecule prediction tasks. *CoRR*, abs/2403.05075.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024b. Harnessing large language models as post-hoc correctors. In *Findings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14559–14574. ACL.

A Proposed Prompt Details

In this work, we design the prompt as follows:

> **Instruction:** You are now a chemical specialist in rewriting captions for a molecule in SMILES format. Make sure those captions describe the given molecule correctly and precisely based on your two inputs (SMILES and Caption of it). Also, make sure your rewriting captions do not include the input SMILES. Write the response without using linebreaks, newlines, or special characters such as "\t" or "\n".
> **Message:** *SMILES string of target molecule:* [SMILES].
> *Description of the molecule:* [caption].
> *Task:* Rewrite the following molecule with its SMILES and caption. The newly rewritten caption should be elaborate, descriptive, and concise, highlighting the key structural features and biological activities of the molecule. Only output rewritten caption without any header and linebreak.
> *Answer:*

B Caption Augmentation Details

This section presents some examples of augmented captions. Particularly, we provide the SMILES string, the original caption of ChEBI-20 and four different captions rewritten by four LLMs, *i.e.*, two open-source LLMs (Llama 2-70b (Touvron et al., 2023) and Llama 3-70b (Touvron et al., 2023)) and two closed-source LLMs (GPT 3.5-turbo (Achiam et al., 2023) and Gemini Pro (Google, 2024)).

1. **SMILES string:** [18FH]
   Original caption: The molecule is the radioactive isotope of fluorine with relative atomic mass 18.000938. The longest-lived fluorine radionuclide with half-life of 109.77 min.
   GPT 3.5-turbo: The molecule is the radioactive isotope of fluorine with relative atomic mass 18.000938 and atomic number 9. It is the longest-lived fluorine radionuclide with half-life of 109.77 min. It is used in positron emission tomography (PET) for imaging the brain, heart, and other organs.
   Gemini Pro: This molecule is a radioactive isotope of fluorine with a relative atomic mass of 18.000938. It is the longest-lived fluorine radionuclide with a half-life of 109.77 minutes.
   Llama 2-70b: This molecule is a radioactive isotope of fluorine, with a relative atomic mass of 18.000938. It is the longest-lived fluorine radionuclide, with a half-life of 109.77 minutes. It has a number of applications in medical imaging and research, including positron emission tomography (PET) scans and fluorodeoxyglucose (FDG) tracers.
   Llama 3-70b: This molecule is fluorine-18 a radioactive isotope of fluorine with a relative atomic mass of 18.000938 and a half-life of 109.77 minutes making it the longest-lived fluorine radionuclide.

2. **SMILES string:** CCN(CC)CCN1C(=O)CN=C(C2=C1C=CC(=C2)Cl)C3=CC=CC=C3F
   Original caption: The molecule is a 1,4-benzodiazepinone that is 1,3-dihydro-2H-1,4-benzodiazepin-2-one substituted by a 2-(diethylamino)ethyl group, 2-fluorophenyl group and chloro group at positions 1, 5 and 7, respectively. It is a partial agonist of GABAA receptors and used for the treatment of insomnia. It has a role as a sedati gve, an anticonvulsant, a GABAA receptor agonist and an anxiolytic drug. It is a 1,4-benzodiazepinone, an organochlorine compound, a member of monofluorobenzenes and a tertiary amino compound.
   GPT 3.5-turbo: A 1,4-benzodiazepinone with a 1,3-dihydro-2H-1,4-benzodiazepin-2-one core, this molecule features a 2-(diethylamino)ethyl group, 2-fluorophenyl group, and chloro group at positions 1, 5, and 7, respectively. As a partial agonist of GABAA receptors, it exhibits sedative, anticonvulsant, and anxiolytic effects. Its therapeutic applications include the treatment of insomnia, making it a valuable pharmaceutical agent in managing sleep disorders.
   Gemini Pro: The molecule is a 1,4-benzodiazepinone derivative with a 2-(diethylamino)ethyl side chain and substitutions of a 2-fluorophenyl group at position 5 and a chloro group at position 7. Known for its partial agonist activity on GABAA receptors, it is commonly used as a sedative, anticonvulsant, and anxiolytic agent for the treatment of insomnia. Its chemical structure includes a diazepine ring fused to a benzene ring with specific substituents, making it a potent thera-

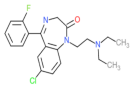
Molecule	Caption - Original	Caption - GPT 3.5-turbo	Caption - Gemini Pro
	The molecule is the radioactive isotope of fluorine with relative atomic mass 18.000938. The longest-lived fluorine radionuclide with half-life of 109.77 min.	The molecule is the radioactive isotope of fluorine with relative atomic mass 18.000938 and atomic number 9. It is the longest-lived fluorine radionuclide with half-life of 109.77 min. It is used in positron emission tomography (PET) for imaging the brain, heart, and other organs.	This molecule is a radioactive isotope of fluorine with a relative atomic mass of 18.000938. It is the longest-lived fluorine radionuclide with a half-life of 109.77 minutes.
	The molecule is a 1,4-benzodiazepinone that is 1,3-dihydro-2H-1,4-benzodiazepin-2-one substituted by a 2-(diethylamino)ethyl group, 2-fluorophenyl group and chloro group at positions 1, 5 and 7, respectively. It is a partial agonist of GABAA receptors and used for the treatment of insomnia. It has a role as a sedati gve, an anticonvulsant, a GABAA receptor agonist and an anxiolytic drug. It is a 1,4-benzodiazepinone, an organochlorine compound, a member of monofluorobenzenes and a tertiary amino compound.	A 1,4-benzodiazepinone with a 1,3-dihydro-2H-1,4-benzodiazepin-2-one core, this molecule features a 2-(diethylamino)ethyl group, 2-fluorophenyl group, and chloro group at positions 1, 5, and 7, respectively. As a partial agonist of GABAA receptors, it exhibits sedative, anticonvulsant, and anxiolytic effects. Its therapeutic applications include the treatment of insomnia, making it a valuable pharmaceutical agent in managing sleep disorders.	The molecule is a 1,4-benzodiazepinone derivative with a 2-(diethylamino)ethyl side chain and substitutions of a 2-fluorophenyl group at position 5 and a chloro group at position 7. Known for its partial agonist activity on GABAA receptors, it is commonly used as a sedative, anticonvulsant, and anxiolytic agent for the treatment of insomnia. Its chemical structure includes a diazepine ring fused to a benzene ring with specific substituents, making it a potent therapeutic compound with diverse pharmacological properties.
	The molecule is an indole phytoalexin that is indole substituted at position 3 by a 1,3-thiazol-2-yl group. It has a role as a metabolite. It is an indole phytoalexin and a member of 1,3-thiazoles.	Indole phytoalexin with a 1,3-thiazol-2-yl substitution at position 3. It plays a role as a metabolite and exhibits indole phytoalexin and 1,3-thiazole properties.	The molecule is a biologically active compound containing an indole ring substituted at position 3 with a 1,3-thiazol-2-yl group. It acts as an indole phytoalexin and plays a role as a metabolite. Its structure consists of fused aromatic rings and a sulfur-containing heterocycle, making it a unique and potentially potent agent in biological systems.

Table 6: Example of molecules and their captions generated by GPT 3.5-turbo and Gemini Pro in LACHEBI-20.

peutic compound with diverse pharmacological properties.

Llama 2-70b: This molecule is a 1,4-benzodiazepinone, a type of sedative drug that selectively activates GABAA receptors. It has a 2-(diethylamino)ethyl group, 2-fluorophenyl group, and chloro group attached at positions 1, 5, and 7, respectively. Its role as a partial agonist of GABAA receptors makes it useful for treating insomnia, and it also has anticonvulsant, anxiolytic, and sedative properties. This compound belongs to the class of organochlorine compounds and is a tertiary amino compound.

Llama 3-70b: This molecule is a 1,4-benzodiazepinone featuring a 2-(diethylamino)ethyl substituent at position 1, a 2-fluorophenyl group at position 5, and a chloro group at position 7, functioning as a partial GABAA receptor agonist used to treat insomnia, and possessing sedative, anticonvulsant, and anxiolytic properties.

3. **SMILES string:** C1=CC=C2C(=C1)C(=CN2)C3=NC=CS3

Original caption: The molecule is an indole phytoalexin that is indole substituted at position 3 by a 1,3-thiazol-2-yl group. It has a role as a metabolite. It is an indole phytoalexin and a member of 1,3-thiazoles.

GPT 3.5-turbo: Indole phytoalexin with a 1,3-thiazol-2-yl substitution at position 3. It plays a role as a metabolite and exhibits indole phytoalexin and 1,3-thiazole properties.

Gemini Pro: The molecule is a biologically active compound containing an indole ring substituted at position 3 with a 1,3-thiazol-2-yl group. It acts as an indole phytoalexin and plays a role as a metabolite. Its structure consists of fused aromatic rings and a sulfur-containing heterocycle, making it a unique and potentially potent agent in biological systems.

Llama 2-70b: This indole phytoalexin features a unique molecular structure where an indole ring is substituted at position 3 by a 1,3-thiazol-2-yl group, resulting in a distinct chemical compound with significant biological activity. As a metabolite, it plays a crucial role in various physiological processes and is also a member of the 1,3-thiazole class of compounds, which are characterized by their sulfur-containing heterocyclic ring.

Llama 3-70b: This molecule is a substituted indole featuring a 1,3-thiazol-2-yl group at

position 3, classified as an indole phytoalexin and a metabolite, also belonging to the class of 1,3-thiazoles.

4. **SMILES string:** CN(CC(=O)O)N=O
   Original caption: The molecule is a nitrosamine that is sarcosine in which the hydrogen attached to the nitrogen has been replaced by a nitroso group. It has a role as a carcinogenic agent. It is a non-proteinogenic amino acid derivative and a nitrosamine. It derives from a sarcosine.
   GPT 3.5-turbo: The molecule is a nitrosamine that is sarcosine in which the hydrogen attached to the nitrogen has been replaced by a nitroso group. It has a role as a carcinogenic agent. It is a non-proteinogenic amino acid derivative and a nitrosamine. It derives from a sarcosine. This molecule is a nitrosamine containing a nitroso group attached to a sarcosine derivative. It is known for its carcinogenic properties and is classified as a non-proteinogenic amino acid derivative.
   Gemini Pro: Sarcosine N-nitrosamide, a nitrosamine with a nitroso group replacing the hydrogen atom on the nitrogen atom of the sarcosine molecule, exerts carcinogenic effects. This non-proteinogenic amino acid derivative exhibits a potential role as a carcinogen and is derived from sarcosine.
   Llama 2-70b: This nitrosamine is a modified form of sarcosine, where the hydrogen atom attached to the nitrogen has been replaced by a nitroso group, resulting in a distinct chemical structure. As a non-proteinogenic amino acid derivative, it exhibits unique properties and functions, and its presence has been linked to carcinogenic effects.
   Llama 3-70b: This molecule is a carcinogenic nitrosamine derived from sarcosine where a nitroso group replaces the hydrogen attached to the nitrogen atom making it a non-proteinogenic amino acid derivative.

5. **SMILES string:** CC(=O)N[C@@H]1[C@H]([C@@H]([C@H](O[C@H]1O)CO)O[C@@H]2[C@@H]([C@H]([C@@H]([C@H](O2)CO)O)O[C@@H]3[C@H]([C@H]([C@@H]([C@H](O3)CO)O)O)O)NC(=O)C)O
   Original caption: The molecule is an amino trisaccharide consisting of alpha-D-mannopyranose, 2-acetamido-2-deoxy-alpha-D-glucopyranose and 2-acetamido-2-deoxy-beta-D-glucopyranose residues joined in sequence by (1->3) and (1->4) glycosidic bonds. It is a glucosamine oligosaccharide, an amino trisaccharide and a member of acetamides. It derives from an alpha-D-Manp-(1->3)-alpha-D-GlcpNAc and an alpha-D-GlcpNAc-(1->4)-beta-D-GlcpNAc.
   GPT 3.5-turbo: The molecule is an amino trisaccharide consisting of alpha-D-mannopyranose, 2-acetamido-2-deoxy-alpha-D-glucopyranose and 2-acetamido-2-deoxy-beta-D-glucopyranose residues joined in sequence by (1->3) and (1->4) glycosidic bonds. It is a glucosamine oligosaccharide, an amino trisaccharide and a member of acetamides. It derives from an alpha-D-Manp-(1->3)-alpha-D-GlcpNAc and an alpha-D-GlcpNAc-(1->4)-beta-D-GlcpNAc. The molecule is a complex amino trisaccharide composed of alpha-D-mannopyranose, 2-acetamido-2-deoxy-alpha-D-glucopyranose, and 2-acetamido-2-deoxy-beta-D-glucopyranose residues linked together via (1->3) and (1->4) glycosidic bonds. This glucosamine oligosaccharide is a member of acetamides and plays a crucial role in various biological processes due to its unique sugar composition and glycosidic linkages.
   Gemini Pro: This molecule is an amino trisaccharide composed of an alpha-D-mannopyranose unit, a 2-acetamido-2-deoxy-alpha-D-glucopyranose unit, and a 2-acetamido-2-deoxy-beta-D-glucopyranose unit. These units are linked by (1->3) and (1->4) glycosidic bonds. This amino trisaccharide is a member of the acetamides family and is derived from an alpha-D-Manp-(1->3)-alpha-D-GlcpNAc and an alpha-D-GlcpNAc-(1->4)-beta-D-GlcpNAc.
   Llama 2-70b: This molecule is a glucosamine oligosaccharide composed of three residues: alpha-D-mannopyranose, 2-acetamido-2-deoxy-alpha-D-glucopyranose, and 2-acetamido-2-deoxy-beta-D-glucopyranose. These residues are joined in sequence by (1->3) and (1->4) glycosidic bonds. It is an amino trisaccharide, derived from an alpha-D-Manp-(1->3)-alpha-D-GlcpNAc and an alpha-D-GlcpNAc-(1->4)-beta-D-GlcpNAc.
   Llama 3-70b: This molecule is a trisaccharide composed of alpha-D-mannopyranose, 2-acetamido-2-deoxy-

alpha-D-glucopyranose, and 2-acetamido-2-deoxy-beta-D-glucopyranose residues linked by (1->3) and (1->4) glycosidic bonds, classified as a glucosamine oligosaccharide and an acetamide derivative.

C Evaluation Setup

Following the benchmark settings of CHEBI-20, we train LAMOLT5 on the training dataset of LACHEBI-20 and evaluate it on the test dataset. Since we are considering two molecular tasks: text-based *de novo* molecule generation and molecule captioning, we employ two evaluation metric sets.

To evaluate the molecule generation task, we employ eight metrics following previous work (Edwards et al., 2022): SMILES **BLEU** score (Papineni et al., 2002), **Levenshtein** distance (Miller et al., 2009), Fréchet ChemNet Distance (**FCD**) (Preuer et al., 2018), **MACCS FTS** (Durant et al., 2002), **RDK FTS** (Schneider et al., 2015) **Morgan FTS** (Rogers and Hahn, 2010), **Exact** score (Edwards et al., 2022), and **Validity** (Edwards et al., 2022). Notably, there are three fingerprint metrics: MACCS FTS, RDK FTS and Morgan FTS. FTS stands for fingerprint Tanimoto similarity. MACCS, RDK, and Morgan are each fingerprinting methods for molecules. The fingerprints of two molecules are compared using Tanimoto similarity, and the average similarity over the evaluation matrix is reported. Additionally, we report exact SMILES string matches *i.e.*, Levenshtein distance and SMILES BLEU score. Exact score and Validity are the percentage of generated molecules that exactly match the ground truth and the percentage of generated strings that are valid.

To evaluate the molecule captioning task, we employ three natural language generation metrics, *e.g.*, Caption **BLEU** score (Papineni et al., 2002), **ROUGE** (Lin, 2004), and **METEOR** (Banerjee and Lavie, 2005). BLEU measures the precision of n-grams between generated and reference texts, ROUGE evaluates recall and precision of overlapping units such as n-grams or word sequences, and METEOR combines precision, recall, and synonym matching for a more holistic evaluation of text generation quality.

Furthermore, the cross-modal evaluation metric **Text2Mol** (Edwards et al., 2021) aims to train a retrieval model to rank molecules given their text descriptions. Unlike traditional metrics that rely on words or n-grams, the ranking function of Text2Mol uses cosine similarity between the ground truth molecule/description and the generated description/molecule, respectively. It can offer a more integrated assessment to measure the comprehensive semantics of molecule/description. Therefore, we adopt this metric as an essential assessment to understand the effectiveness of different models.

D Baseline Models Description

This section presents brief descriptions of baseline models included in this work.

RNN (Cho et al., 2014). It introduces a novel approach for improving statistical machine translation through the use of Recurrent Neural Networks (RNNs). They propose an encoder-decoder architecture that learns continuous-space representations for phrases. The encoder processes an input phrase and compresses it into a fixed-dimensional vector, while the decoder uses this vector to generate the target phrase. This method allows for better handling of variable-length input sequences and capturing long-term dependencies in phrases, leading to significant improvements in translation quality compared to traditional models.

Transformer (Vaswani et al., 2017). It introduces the Transformer model, a novel neural network architecture designed for sequence transduction tasks, such as machine translation. The Transformer model relies entirely on attention mechanisms to capture dependencies between input and output without using recurrent or convolutional layers. This self-attention mechanism allows for greater parallelization and better handling of long-range dependencies compared to previous models. They demonstrate that the Transformer achieves state-of-the-art performance on translation tasks, significantly improving both training speed and translation quality.

T5 (Raffel et al., 2020). It presents the Text-to-Text Transfer Transformer (T5), a model designed to unify various NLP tasks by converting all tasks into a text-to-text format. They explore the capabilities of transfer learning within this framework, demonstrating that the same model, training objective, hyperparameters, and architecture can be applied across a wide range of NLP tasks. By pre-training on a massive and diverse dataset and fine-tuning specific tasks, T5 achieves state-of-the-art performance on numerous benchmarks. Additionally, they propose the CHEBI-20 dataset.

MOLT5 (Edwards et al., 2022). It explores the novel concept of bridging the gap between molecular representations and natural language descriptions. They propose a model, MOLT5, that translates molecular structures into textual descriptions and vice versa. This interdisciplinary approach leverages advances in natural language processing and cheminformatics, using techniques such as neural networks to encode and decode information between these two domains. MOLT5 is the fundamental model that motivates our work.

Text+Chem T5 (Christofidellis et al., 2023). Text+Chem T5 is a novel multi-task, multi-domain language model designed to bridge the gap between natural language and chemical language tasks. Built on the T5 architecture, it is specifically designed to handle tasks spanning both textual and chemical domains. This model can effectively translate between natural and chemical languages, enabling it to perform a variety of tasks such as chemical reaction prediction (forward and retrosynthesis), text-conditional de novo molecule generation, molecular captioning, and paragraph-to-action conversion for chemical procedures.

TGM-DLM (Gong et al., 2024). TGM-DLM employs a Transformer-based architecture with cross-attention to incorporate textual guidance. It is trained using two objectives: denoising embeddings with text guidance and recovering uncorrupted SMILES strings from deliberately corrupted ones. This training strategy enhances the model's ability to generate valid and relevant molecular structures. The model demonstrates superior performance compared to autoregressive models like MolT5-Base, achieving this without additional data resources or pre-training.

MolReGPT (Li et al., 2024). MolReGPT is a novel framework leveraging LLMs like GPT to advance molecule discovery through molecule-caption translation. Unlike traditional methods, which rely heavily on domain experts, computational resources, or domain-specific pre-training, MolReGPT uses ICL few-shot learning. This approach enables LLMs to perform molecule understanding and text-based molecule generation by retrieving and learning from similar molecules and their descriptions from a local database.

BioT5 (Pei et al., 2023). It is a pre-training framework designed to enhance drug discovery by integrating molecules, proteins, and natural language. This framework addresses limitations in current models, such as generating invalid molecular SMILES, underutilising contextual information, and treating structured and unstructured knowledge equally. BioT5 utilises SELFIES (Krenn et al., 2020) for robust molecular representations and extracts relevant knowledge from the context surrounding bio-entities in unstructured biological literature.

MolXPT (Liu et al., 2023b). It is a unified language model that integrates text and molecular representations for enhanced molecular modelling. MolXPT leverages the success of Generative Pre-trained Transformers (GPT) by pre-training on SMILES sequences wrapped in relevant textual context. This involves detecting molecule names in text, replacing them with corresponding SMILES, and thus allowing mutual information exchange between text and molecule representations.

BioT5+ (Pei et al., 2024). BioT5+ is designed to bridge the gap between molecular data and textual descriptions in biological research. Building upon the BioT5 framework, BioT5+ introduces several innovations, including the integration of IUPAC nomenclature for molecules, which enhances its ability to understand molecular structures in both scientific literature and formal representations like SMILES and SELFIES. By employing multi-task instruction tuning, BioT5+ can generalise across diverse biological tasks, such as classification, regression, and generation, making it versatile for applications ranging from molecule property prediction to drug discovery.

E Additional Experimental Results and Discussion

Anecdotal molecule generation examples. Table 2 shows some example molecules generated by MOLT5-Small and LAMOLT5-Small, and the ground-truth molecules from LACHEBI-20. From these examples, we can find that LAMOLT5-Small can generate accurate molecules similar to the ground truth, while MOLT5-Small is making mistakes. (1) is an interesting case since (*i*) its ground truth SMILES string has a long length, 88 characters for which LAMOLT5-Small is able to generate an exact match; (*ii*) indicates that LAMOLT5-Small can understand crystalline solids, like indolylmethylglucosinolate, in the annotation. In another interesting case, (2), LAMOLT5-Small not only understands chemical compounds but also can understand chemical treatments, *e.g.*, replace-

ment, mentioned in the annotation. These examples showcase the superiority of LAMOLT5 for the text-based *de novo* molecule generation task.

Anecdotal molecule captioning examples. Table 4 shows examples of molecule descriptions generated by MOLT5-Small and LAMOLT5-Small, alongside the ground-truth descriptions. LAMOLT5-Small can generate more accurate and detailed descriptions that align closely with the ground truth, whereas MOLT5-Small often misses important semantic details. In (1), MOLT5-Small omits critical details about the specific role and structure. In contrast, LAMOLT5-Small correctly identifies the molecule as derived from malonic acid and mentions its role as an Escherichia coli metabolite and mouse metabolite. These examples show the superiority of LAMOLT5-Small in generating detailed and accurate molecule descriptions, making it a more effective model for the molecule captioning task.

F Additional Experiments

Section 3.1 and Section 3.2 demonstrate a detailed example implementation of LA^3 on the CHEBI-20 for challenging molecular generation tasks. To further demonstrate the versatility of LA^3, we extend its application to several additional datasets, `ogbg-molbace` (Hu et al., 2020), `ogbg-molhiv` (Hu et al., 2020), `ogbg-molesol` (Hu et al., 2020), and `CC3M` (Sharma et al., 2018). These datasets support a variety of crucial tasks, such as *image captioning*, *text understanding*, and *graph property prediction*.

F.1 Dataset and Task

We consider four benchmark datasets, which contain *image*, *text* and *graph* data.

1. `ogbg-molbace`. The `ogbg-molbace` dataset provides quantitative (IC_{50}) and qualitative (binary label) binding results for a set of inhibitors of human b-secretase 1 (BACE-1). All data are experimental values reported in the scientific literature over the past decade, some with detailed crystal structures available. **Task:** `ogbg-molbace` merged a collection of 1,522 compounds with their 2D structures and binary labels, built as a classification task.
2. `ogbg-molhiv`. The HIV dataset was introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, which tested the ability to inhibit HIV replication for 41,127 compounds. Screening results were evaluated and placed into three categories: confirmed inactive (CI), confirmed active (CA) and confirmed moderately active (CM). We further combine the latter two labels, making it a classification task between inactive (CI) and active (CA and CM). **Task:** As we are more interested in discovering new categories of HIV inhibitors based on the available *text* and *graph* structure information.
3. `ogbg-molesol`. `ogbg-molesol` is a small dataset consisting of water solubility data for 1,128 compounds. **Task:** We intend to estimate solubility directly from chemical *graph* structures (as encoded in *text* SMILES strings).
4. `CC3M`. `CC3M` is a large-scale dataset comprising around 3.3 million image-caption pairs. It is designed for automatic image captioning tasks and represents a significant step forward in terms of the variety and volume of data compared to previous datasets like MS-COCO. **Task:** We follow the settings of (Fan et al., 2023) to train CLIP model (Radford et al., 2021) and test it on ImageNet (Deng et al., 2009).

F.2 Automatic Annotation Augmentation

Given `ogbg-molbace`, `ogbg-molhiv`, and `ogbg-molesol` datasets, we first generate descriptions following the instruction of (Zhong et al., 2024a). Consequently, we query LLMs to augment these descriptions as described in Section 3.2. The prompt is designed as follows:

> **Instruction:** You are now a chemical specialist in rewriting descriptions for a molecule in SMILES format. Make sure those descriptions describe the given molecule correctly and precisely based on your two inputs (SMILES and Description of it). Also, make sure your rewriting captions do not include the input SMILES.
> **Message:** *SMILES string of target molecule:* [SMILES].
> *Description of the molecule:* [description].
> *Task:* Rewrite the following molecule with its SMILES and description. The newly rewritten caption should be elaborate, descriptive, and concise, highlighting the key structural features and biological activities of the molecule. Only output rewritten caption without any header and linebreak.
> *Answer:*

CC3M has available annotations for each image. We leverage LLMs to augment their annotations using this prompt:

> **Instruction:** You are now a specialist in rewriting descriptions for an image. Make sure those descriptions describe the given image correctly and precisely.
> **Message:** *Description of the image:* [description].
> *Task:* Rewrite the following description. The newly rewritten caption should be elaborate, descriptive, and concise, highlighting the key knowledge of the molecule. Only output rewritten caption without any header and linebreak.
> *Answer:*

In this paper, we utilise two closed-source LLMs (GPT 3.5-turbo (Achiam et al., 2023) and Gemini Pro (Google, 2024)) to generate two rewritten annotations for the above-mentioned datasets.

F.3 Training on Augmented Dataset

After obtaining the augmented datasets (ogbg-molbace, ogbg-molhiv, and ogbg-molesol), we simply combine three annotations of each molecule as the input features and integrate them within the LM and GNN models. Other training implementations follow the instruction of (Zhong et al., 2024a). About the CC3M dataset, we follow the implementation of (Fan et al., 2023) to integrate the augmented annotations with the CLP model and evaluate them.

Results in Table 5 show that LA^3 significantly enhances performance across these diverse applications. This improvement highlights LA^3 's potential to be a valuable tool in a wide range of AI tasks, offering substantial gains in accuracy and efficiency.