Bridging the Gap Between Molecule and Textual Descriptions via Substructure-aware Alignment

Hyuntae Park[1*] **Yeachan Kim**[2*] **SangKeun Lee**[1,3]
[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]Division of Language & AI, Hankuk University of Foreign Studies, Seoul, Republic of Korea
[3]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
pht0639@korea.ac.kr, yeachan@hufs.ac.kr, yalphy@korea.ac.kr

Abstract

Molecule and text representation learning has gained increasing interest due to its potential for enhancing the understanding of chemical information. However, existing models often struggle to capture subtle differences between molecules and their descriptions, as they lack the ability to learn fine-grained alignments between molecular substructures and chemical phrases. To address this limitation, we introduce MolBridge, a novel molecule–text learning framework based on substructure-aware alignments. Specifically, we augment the original molecule–description pairs with additional alignment signals derived from molecular substructures and chemical phrases. To effectively learn from these enriched alignments, MolBridge employs substructure-aware contrastive learning, coupled with a self-refinement mechanism that filters out noisy alignment signals. Experimental results show that MolBridge effectively captures fine-grained correspondences and outperforms state-of-the-art baselines on a wide range of molecular benchmarks, underscoring the importance of substructure-aware alignment in molecule-text learning.[1]

1 Introduction

Recent advances in natural language processing (NLP) have transformed various scientific fields, with chemistry emerging as a prominent domain. Transformer-based models have demonstrated remarkable success in molecular tasks, such as drug discovery (Drews, 2000) and molecular property prediction (Wu et al., 2018), offering scalable alternatives to traditional wet-lab experiments (Beltagy et al., 2019; Wang et al., 2019). Among these advancements, Molecule-Text Models (MTMs) have been developed to bridge the gap between molecular structures and natural language, providing a symbolic interface for understanding complex chemical information (Edwards et al., 2022; Liu et al., 2023a,c).

Despite their potential, MTMs face a fundamental challenge: the severe **sparsity** of molecule-text alignment. Given the vast diversity of chemical structures, annotated datasets that explicitly pair molecules with their corresponding textual descriptions are extremely limited. This scarcity restricts the model's ability to generalize across chemical space, leading to biased representations that perform poorly on unseen compounds (Haghighatlari et al., 2020). More critically, it prevents MTMs from learning fine-grained correspondences between specific fragments (i.e., molecular substructures and corresponding chemical phrases), making them struggle to capture subtle differences between similar compounds, such as D-glutamate and L-glutamate (Zhang et al., 2025). These subtle differences reflect substructural variations, which are important because they often determine the core functionalities and chemical properties of the entire molecule (Wu et al., 2023).

Although some studies (Min et al., 2024; Zhang et al., 2025) have attempted to address this issue by introducing local alignments on the given pairs, these methods still suffer from several limitations. First, they rely heavily on **indirect alignment**, where local relations are inferred through feature similarity due to the lack of explicit fragment-level annotations. This absence of direct supervision can lead to incorrect or incomplete mappings, making it difficult for models to learn accurate local relationships between fragments. Second, they often suffer from **over-fragmented alignment**, where models attempt to align molecular tokens (e.g., SMILES characters like '=', '[]', '()') indiscriminately. Such token-level alignment introduces noise, causing the model to learn semantically meaningless fragments rather than chemically meaningful substructures. These limitations comprehensively hinder

[*]These authors contributed equally to this work.
[1]Our code and data are available at https://github.com/Park-ing-lot/MolBridge

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23459–23479
November 4-9, 2025

the ability of existing MTMs to achieve accurate fine-grained alignment and robust generalization.

In response, we propose MolBridge, a novel multimodal framework designed to learn fine-grained alignment between molecules and text through substructure-aware alignments. MolBridge begins by explicitly extracting substructures from molecules and chemical phrases from their corresponding descriptions. These fragments are then cross-linked to their semantically or chemically relevant counterparts: chemical phrases are associated with entire molecules, while substructures are connected to descriptions. To effectively learn from these enriched alignments, we introduce substructure-aware contrastive learning, which jointly considers both fragment-level and holistic molecule-text relations. This strategy encourages the model to capture meaningful substructural semantics while preserving consistency between molecules and their descriptions.

Building on the substructure-aware representations learned by MolBridge, we also introduce MolBridge-Gen, a generative variant of the framework that explicitly leverages local alignment signals derived from substructure–chemical phrase pairs identified by MolBridge. This extension enables the framework to generalize beyond discriminative tasks and effectively support generative scenarios, such as molecule captioning and generation, where fine-grained semantic understanding is essential (Xia et al., 2023).

We conduct comprehensive evaluations across core molecular tasks, including molecular property prediction (Wu et al., 2018), molecule–text retrieval (Zeng et al., 2022; Liu et al., 2023c), and generation tasks (Edwards et al., 2022), to thoroughly assess the effectiveness of MolBridge. Experimental results show that MolBridge consistently outperforms existing MTMs, achieving superior fine-grained alignment accuracy, higher retrieval performance, and enhanced generation quality. The contributions of this paper include the followings:

- We propose MolBridge, a novel framework for fine-grained molecule-text alignment, directly addressing the sparsity of alignment datasets through substructure-aware alignments.
- We introduce a substructure-aware contrastive learning, allowing the model to effectively capture fine-grained relations between molecules and text descriptions.
- We demonstrate that MolBridge consistently outperforms existing methods on diverse tasks, highlighting the significance of substructure-aware augmentations.

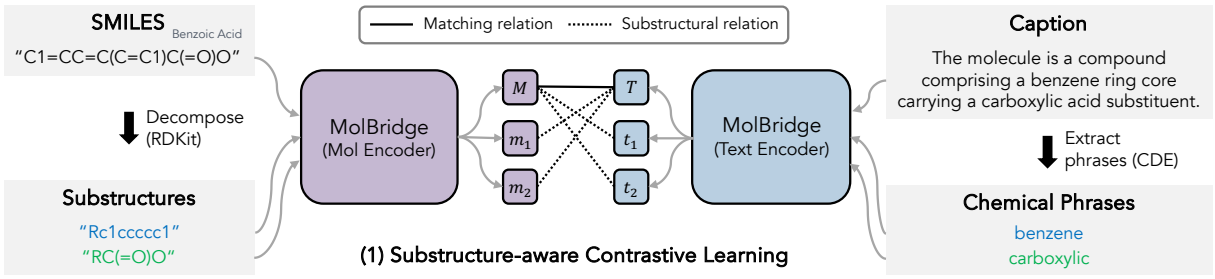
2 Related Works

2.1 Molecule-Text Multimodal Modeling

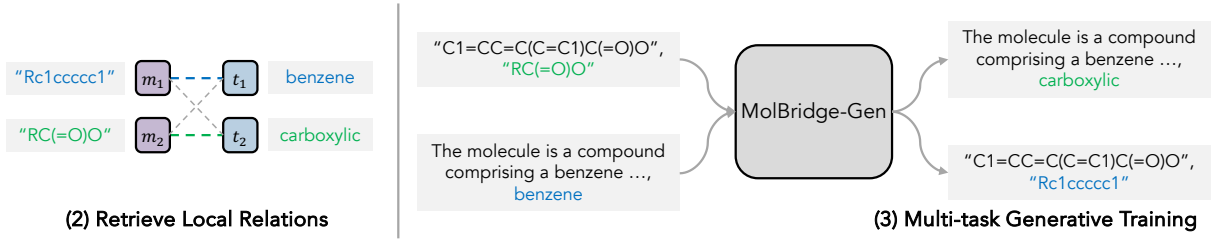
Various models have emerged for learning molecule-text representations. Early on, Zeng et al. (2022) attempted to leverage english textual knowledge to enhance molecule representation learning. Liu et al. (2023a) proposed MoleculeSTM, applying larger datasets and contrastive learning to improve alignment between molecules and text. To facilitate the translation between molecules and natural language, Edwards et al. (2022) proposed molecule captioning tasks along with a powerful baseline model, MolT5. Inspired by Raffel et al. (2020), they utilized corrupted spans replacement objective for pre-training to improve molecular understanding. Liu et al. (2023b) employed generative pre-training to reflect the importance of descriptions, while Liu et al. (2023c) introduced MolCA, a molecular text model that leverages a 2D graph for enhanced molecular understanding. Recently, large language model-based instruction-tuned generative models for molecule understanding Fang et al. (2024); Pei et al. (2024); Cao et al. (2025) have been proposed to enhance generalizability through multitask learning. However, these works only consider the global representations of molecules and text, overlooking finer-grained modal interactions. Unlike these approaches, we encourage the model to effectively learn compositional structures in both molecules and language.

2.2 Fine-grained Molecule-Text Representation Learning

Several studies have attempted to perform fine-grained molecule-text alignment. Yu et al. (2024); Feng et al. (2023) have focused on fine-grained alignment of molecular modalities and tailoring for prediction tasks. Unlike these paradigms, Atomas (Zhang et al., 2025) tackles the problem of cross-modal learning between molecular structures and textual descriptions. A key challenge is the scarcity of fine-grained expert annotations and the ambiguity in defining positive and negative pairs, since one textual property may relate to multiple substructures. To overcome this, the authors use clustering techniques to learn multi-scale representations and encourage consistency across different



(a) Training procedure of MolBridge. We first extract substructures and chemical phrases from SMILES and caption, respectively. We then augment our dataset by constructing additional positive pairs based on their substructural relationships and perform substructure-aware contrastive learning.



(b) Training procedure of MolBridge-Gen. After identifying local relations using MolBridge, we train the generative molecule-text models with them.

Figure 1: Illustration of our framework to learn fine-grained alignments.

levels of granularity. Similarly, Min et al. (2024) utilized optimal transport to conduct fragments at the atom, motif, and global levels within the embedding space. Unlike these methods for implicit local alignment, Li et al. (2024c) proposed an explicit local alignment approach; however, it heavily relies on costly large language models. In this work, we address the challenge of explicit fine-grained alignment by learning fragment-level representations enriched with substructural cues, such as identifying whether a chemical phrase is part of a molecule or a substructure is mentioned in the description, and discovering fine-grained correspondences between these fragments.

3 Methodology

In this section, we introduce MolBridge, a novel framework designed to learn fine-grained alignments between molecules and its descriptions. MolBridge begins with a sparse molecule–text alignment dataset, from which it extracts chemical substructures and phrases to form explicit alignment pairs (§3.1). These augmented pairs then guide structure-aware contrastive learning, enhanced by a self-refinement procedure (§3.2). Finally, we extend the learned representations to various molecular tasks by either directly adopting MolBridge or combining it with generative models (§3.3). The overall procedure is illustrated in Figure 1.

3.1 Substructure Alignment Augmentation

Given the vast diversity of chemical space, existing alignment datasets remain severely limited, constraining a model's ability to generalize across novel molecules (Haghighatlari et al., 2020). To overcome this challenge, we exploit the inherent substructural relations between molecules and their textual descriptions. Let $S = (M, T)$ denote a molecule–text pair, where M is represented by its SMILES (Weininger, 1988) string and T is the corresponding caption, we aim to enrich the following alignments:

Substructures (m) to Description (T) The chemical properties of molecules are largely determined by their constituent substructures (e.g., functional groups, ring systems), indicating the potential alignment between the substructure and original molecule's description. To capture these relations, we decompose each original molecule into a set of substructures using established fragmentation methods, including BRICS (Degen et al., 2008), RECAP (Lewell et al., 1998) fragmentation methods[2]. We then link each substructure m to the original description, yielding a set of substructure–text alignments $(m, T) \in S_m$. These enriched pairs are subsequently integrated into the original alignment set S.

[2]We compare various decomposition strategies in Table 7 of the Appendix.

Chemical Phrases (t) to Molecules (M) Additionally, we align chemical phrases extracted from each description with the corresponding molecule. This step is motivated by the observation that molecular captions often contain non-informative tokens—such as articles, pronouns, and filler words—that are unrelated to chemical properties and may introduce noise into the alignment process (Radford et al., 2021; Messina et al., 2021). To this end, we extract chemical phrases t from the original caption T using two approaches: (i) ChemDataExtractor (Mavracic et al., 2021), a chemistry-specific phrase extractor, and (ii) a large language model–based method[3]. As with substructures, each phrase is then linked back to the original molecule, yielding molecule–phrase alignments $(M, t) \in S_t$. Notably, these chemical phrases can also provide contextual cues for understanding molecular substructures, as many phrases directly reference functional groups (e.g., hydroxyl aromatic), thereby enhancing substructure-aware alignment. These enriched pairs are finally added to the alignment set S.

3.2 Substructure-aware Contrastive Learning with Self-Refinement

Based on the augmented datasets, we train MTMs to learn fine-grained alignments. Considering that the augmented alignments are structured in one-to-many relations, we introduce a substructure-aware contrastive learning that explicitly aligns (i) Molecules with both their original descriptions and extracted chemical phrases (ii) Descriptions with both their original molecules and associated chemical substructures. However, the augmented alignment set potentially involves the incorrect alignment. To address this, the substructure-aware training is built on the self-refinement process.

Substructure-aware Contrastive Learning Given a mini-batch of molecule-side inputs x_m^i and text-side inputs x_t^j, we encode them with modality-specific encoders $f_m(\cdot)$ and $f_t(\cdot)$. We define the pairwise similarity:

$$\sigma_{i,j} = \exp\left(\frac{1}{\tau} \cdot \frac{f_m(x_m^i) \cdot f_t(x_t^j)}{\|f_m(x_m^i)\| \|f_t(x_t^j)\|}\right), \quad (1)$$

where τ is a learnable temperature parameter, and we use the hidden state of the first token to compute similarity.

For each anchor i in standard contrastive learning, let $\mathcal{P}(i)$ be its set of positive matches (i.e., aligned) drawn from our augmented alignments, and let $\mathcal{U}(i)$ be the set that includes contrastive examples in batches, excluding substructure-phrase pairs to avoid false negatives from semantically related examples. We then optimize the in-batch contrastive loss:

$$\mathcal{L}_{\text{mol2txt}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}(i)|}\log\frac{\sum_{j\in\mathcal{P}(i)}\sigma_{i,j}}{\sum_{k\in\mathcal{U}(i)}\sigma_{i,k}},$$

$$\mathcal{L}_{\text{txt2mol}} = -\frac{1}{N}\sum_{j=1}^{N}\frac{1}{|\mathcal{P}(j)|}\log\frac{\sum_{i\in\mathcal{P}(j)}\sigma_{i,j}}{\sum_{k\in\mathcal{U}(j)}\sigma_{k,j}}, \quad (2)$$

where $\mathcal{L}_{\text{mol2txt}}$ denotes the contrastive loss when molecules serve as anchors and its associated text units (descriptions or phrases) as positives, while $\mathcal{L}_{\text{txt2mol}}$ denotes the converse loss, with text units as anchors and molecular structures as positives. The total loss for substructure-aware contrastive learning is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{txt2mol}} + \mathcal{L}_{\text{mol2txt}} \quad (3)$$

Self-Refinement Although we consider potential relationships between molecules and their descriptions, some incorrect associations may still arise. To detect and remove these low-quality pairs, we embed our contrastive training within an iterative self-refinement loop.

To obtain signals for erroneous relations, we introduce a relation classification loss to the total loss (Eq. (3)):

$$\mathcal{L}_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{3} y_{i,c}\log p_{i,c}(f_m(x_m^i)\oplus f_t(x_t^i)), \quad (4)$$

where $y_{i,c}$ is the ground-truth indicator for class $c \in \{S, S_m, S_t\}$, and $p_{i,c}$ is the model's predicted probability that pair (x_m^i, x_t^i) belongs to class c. Inspired by the observation that *models learn clean samples before noisy ones* (Arazo et al., 2019), we discard any pairs that are misclassified in all of a set of predefined epochs. This filtering ensures that subsequent training focuses on higher-quality alignment signals.

3.3 MolBridge with Generative Models

For molecular generative tasks (e.g., molecule captioning, molecule generation), we need to train

[3]We compare these two methods in Section 5 of the Appendix and, for our main experiments, adopt ChemDataExtractor due to its superior cost-effectiveness

Methods	# Params	Text to Molecule				Molecule to Text			
		R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
1D SMILES + 2D Graph									
MoMu (Su et al., 2022)	111M	4.90	14.48	20.69	10.33	5.08	12.82	18.93	9.89
MolFM (Luo et al., 2023)	138M	16.14	30.67	39.54	23.63	13.90	28.69	36.21	21.42
MolFM (fine-tuned) (Luo et al., 2023)	138M	29.39	50.26	58.49	39.34	29.76	50.53	58.63	39.56
MolCA (Liu et al., 2023c)	111M	35.09	62.14	69.77	47.33	37.95	66.81	74.48	50.80
1D SMILES									
MoleculeSTM (Liu et al., 2023a)	120M	35.80	-	-	-	39.50	-	-	-
Atomas-base (Zhang et al., 2025)	271M	39.08	59.72	66.56	47.33	37.88	59.22	65.56	47.81
Atomas-large (Zhang et al., 2025)	825M	49.08	68.32	73.16	57.79	46.22	66.02	72.32	55.52
MolBridge w/o augmentation	155M	23.89	48.91	57.53	35.30	27.30	51.86	60.34	38.47
MolBridge	155M	**50.45**	**70.83**	**76.11**	**59.63**	**52.76**	**73.54**	**78.55**	**62.25**

Table 1: Zero-shot molecule-text retrieval performance on PCDes test set (scaffold split). w/o augmentation refers to the model trained without substructure alignment augmentation. Baseline results are from Zhang et al. (2025).

Methods	T2M		M2T	
	R@1	R@20	R@1	R@20
1D SMILES + 2D Graph				
MoMu-S (Su et al., 2022)	40.8	86.1	40.9	86.2
MoMu-K (Su et al., 2022)	41.6	87.8	41.8	87.5
MoleculeSTM (Liu et al., 2023a)	44.3	90.3	45.8	88.4
MolCA (Liu et al., 2023c)	66.0	93.5	66.6	94.6
1D SMILES				
SciBERT (Beltagy et al., 2019)	37.5	85.2	39.7	85.8
KV-PLM (Zeng et al., 2022)	37.7	85.5	38.8	86.0
MolBridge	**70.9**	**95.6**	**75.0**	**97.4**

Table 2: Zero-shot molecule-text retrieval performance on Pubchem324k test set. Baseline results are from Liu et al. (2023c).

generative models with the translation objective. While the previous augmented datasets provide valuable insights into the fine-grained alignment between molecules and descriptions, they are not directly applicable to this translation task, which demands one-to-one mappings at the same semantic level (i.e., molecule-to-description (M, T) or substructure-to-phrases (m, t)).

Substructure–phrase Relations Accurately identifying these one-to-one relations is challenging due to the absence of explicit supervision linking substructures and phrases. To address this, we leverage the pre-trained MolBridge, which has been trained on the previously augmented datasets. This model inherently captures fine-grained associations between substructures and phrases through the substructure-aware alignment.

To obtain these relations, we begin by extracting substructures and phrases from the original training dataset. Each substructure is then paired with candidate phrases, and the relevance of each pair is evaluated using the MolBridge score—defined as the cosine similarity between substructure and phrase embeddings. Only substructure–phrase pairs with scores exceeding a predefined threshold τ are retained, ensuring high-quality alignment. If no valid pair is found for a given molecule, the original pair is excluded from training, as it lacks sufficient alignment signals.

Training MolBridge-Gen The resulting substructure–phrase pairs are used to train MolBridge-Gen, a generative model optimized using a conditional generation loss in a multi-task setting, following Christofidellis et al. (2023). For example, in the molecule captioning task, MolBridge-Gen is trained to simultaneously generate full captions from the complete molecular representation and chemical phrases from the substructures. This dual-generation strategy ensures that the model learns both the original context of the molecule and the fine-grained details of its substructures. Detailed prompt templates used for pre-training are provided in Table 21 in the Appendix.

4 Experiments

In this section, we verify the efficacy of MolBridge through extensive experiments and analyses aimed at answering the following questions:

- Can MolBridge capture fine-grained alignments more effectively than existing MTMs? (§4.2)
- Can MolBridge transfer its learned representations to uncover diverse structure–property relationships in downstream tasks? (§4.3)
- Can the relations identified by MolBridge yield interpretable alignments that support effective translation between molecules and text? (§4.4)

Method	BBBP	Tox21	ToxCast	ClinTox	MUV	HIV	BACE	SIDER	Avg.
MoleculeSTM (Liu et al., 2023a)	70.6	75.7	65.2	86.6	65.7	77.0	82.0	63.7	73.3
MolFM (Luo et al., 2023)	72.9	77.2	64.4	79.7	76.0	78.8	83.9	64.2	74.6
MoMu (Su et al., 2022)	70.5	75.6	63.4	79.9	70.6	75.9	76.7	60.5	71.6
MolCA-SMILES (Liu et al., 2023c)	70.8	76.0	56.2	89.0	-	-	79.3	61.1	-
Atomas (Zhang et al., 2025)	73.7	77.9	66.9	93.2	76.3	**80.6**	83.1	64.4	77.0
MolBridge	**77.6**	**84.7**	**70.3**	**94.8**	**76.8**	77.8	**84.5**	**66.9**	**79.2**

Table 3: Results for molecular property prediction tasks (ROC-AUC) on MoleculeNet benchmark. **Bold** indicates the best results.

4.1 Experimental Settings

Dataset. For training MolBridge, we collect the descriptions for 431,877 molecules following previous works (Liu et al., 2023a,c), removing any data overlapping with downstream task datasets to prevent data leakage. The augmented dataset contains approximately 2M pairs. We train MolBridge-Gen with 32,455 pairs of data that are estimated to contain local relations as described in Section 3.3. When decomposing molecules into substructures, we set a maximum number of atoms to 100 due to its high computational complexity. We extract chemical phrases from all captions in the dataset. For evaluation, we perform zero-shot molecule-text retrieval tasks on the PubChem324k (Liu et al., 2023c) and PCdes (Zeng et al., 2022) datasets, molecule captioning tasks on the ChEBI-20 (Edwards et al., 2021) dataset, and molecule property prediction tasks using the MoleculeNet benchmark (Wu et al., 2018). Details of the implementation are provided in Appendix A.

4.2 Zero-shot Molecule-Text Retrieval

Settings. We report zero-shot retrieval performance using Recall at 1/5/10/20, which measures the proportion of relevant results found within the top 1, 5, 10, or 20 positions, a performance metric for information retrieval systems (Manning et al., 2008). We also report the Mean Reversed Rank (MRR) (Voorhees, 1999), which measures how effectively a retrieval model ranks relevant items by averaging the inverse rank positions of the first correct result across multiple queries.

Results. We evaluated retrieval performance on three datasets, as summarized in Tables 1 and 2. On the PCDes scaffold test set, MolBridge (155M) achieves substantial performance gains compared to both Atomas-base (271M) and Atomas-large (825M), despite having significantly fewer parameters. Specifically, MolBridge shows average improvements of 11.1%p and 14.2%p over Atomas-base, and 2.2%p and 6.8%p over Atomas-large in text-to-molecule and molecule-to-text retrieval, respectively[4]. This result demonstrates (i) the efficiency and effectiveness of MolBridge in capturing fine-grained molecule–text alignments with a more compact architecture (ii) prior methods relying solely on implicit alignment signals may learn incorrect fragment correspondences.

Further, removing our substructural alignment augmentation leads to a noticeable drop in performance, validating its critical role in guiding fragment-level representation learning. Table 2 reports the performance on the PubChem324k test set, where MolBridge again outperforms all baselines, including those utilizing 2D molecular graphs, demonstrating its effectiveness in accurately linking molecular structures with natural language descriptions.

4.3 Molecular Property Prediction

Settings. Following Zhang et al. (2025), we evaluate MolBridge on eight classification datasets from MoleculeNet. We use the scaffold split provided by DeepChem (Ramsundar et al., 2019), and we report the ROC-AUC scores. We jointly train the MolBridge molecule encoder and text encoder to see whether our proposed framework empowers the fine-grained understanding of molecules. We compare our model with multimodal methods.

Results. Table 3 shows the results of eight property prediction tasks. Although MolBridge is much smaller than the previous fine-grained alignment method, it achieves a 2.2%p improvement in performance. This demonstrates that the substructural relation–based alignment approach enhances MolBridge's ability to capture fine-grained structural information. Because property prediction often requires distinguishing subtle differences between similar molecules (Park et al., 2024), the perfor-

[4]Results on the original PCDes split are shown in Table 14 in the Appendix.

Method	# Params	BLEU-2↑	BLUE-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-large (Edwards et al., 2022)	783M	0.594	0.508	0.654	0.510	0.594	0.614
Text+Chem T5 (Christofidellis et al., 2023)	220M	0.625	0.542	0.682	0.543	0.622	0.648
MolReGPT (GPT-4) (Li et al., 2024b)	-	0.607	0.525	0.634	0.476	0.562	0.610
Atomas-base (Zhang et al., 2025)	271M	0.632	0.545	0.685	0.545	0.626	-
MolReFlect (Li et al., 2024c)	7B	0.617	0.539	0.657	0.510	0.593	0.623
MolBridge-Gen-small	82M	0.625	0.542	0.686	0.549	0.629	0.649
MolBridge-Gen-base	248M	**0.674**	**0.605**	**0.724**	**0.609**	**0.676**	**0.693**

Table 4: Results of molecule captioning task on CheBI-20 test set. **Bold** and underlined indicate the best and second-best results, respectively. Full comparison is in Table 22 in the Appendix.

Method	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
MolT5-large (Edwards et al., 2022)	0.854	0.318	16.32	0.889	0.813	0.750	0.958
Text+Chem T5 (Christofidellis et al., 2023)	0.853	0.322	16.87	0.901	0.816	0.757	0.943
MolReGPT (GPT-4) (Li et al., 2024b)	0.857	0.280	17.14	0.903	0.805	0.739	0.899
Atomas-large (Zhang et al., 2025)	**0.874**	**0.387**	**12.70**	0.914	0.841	0.788	**0.980**
MolReFlect (Li et al., 2024c)	0.886	0.430	13.99	0.916	0.828	0.775	0.981
MolBridge-Gen-small	0.827	0.266	16.88	0.898	0.820	0.751	0.947
MolBridge-Gen-base	0.842	0.358	15.66	**0.918**	**0.854**	**0.798**	0.956

Table 5: Results of molecule generation task on CheBI-20 test set. **Bold** and underlined indicate the best and second-best results, respectively. Full comparison is in Table 23 in the Appendix.

mance gain suggests that our method enables the model to learn precise substructural representations, thereby capturing nuanced relationships and effectively transferring molecular knowledge to a wide range of prediction tasks.

4.4 Molecule Captioning & Generation

Settings. We evaluate MolBridge-Gen on molecule captioning and generation tasks using the CheBI-20 dataset. To assess the quality of generated captions, we use evaluation metrics that include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2014) scores. For the de novo molecule generation task, we employ BLEU to measure the percentage of predictions that exactly match the true labels (Exact Match; EM), Levenshtein distance (Miller et al., 2009) for string similarity, Validity for grammatical correctness of the generated molecules, and molecular fingerprint-based similarity measures such as MACCS FTS (Durant et al., 2002), RDK FTS (Schneider et al., 2015), and Morgan FTS (Rogers and Hahn, 2010) to compare similarity with the original molecules.[5]

Results. Tables 4 and 5 present the results for molecule captioning and generation. Despite being based on MolT5-small/base, both MolBridge-Gen-small/base show improvements over MolT5-large and other baselines. In the captioning task, MolBridge-Gen-base achieves the highest scores in ROUGE scores and METEOR, outperforming all baseline models, including the much larger 7B MolReFlect. This suggests that learning substructure–phrase relationships enables more fine-grained understanding of molecular content, even with smaller models. Moreover, these improvements demonstrate that the substructure–phrase pairs identified by MolBridge for training MolBridge-Gen are indeed effective in learning fine-grained alignments between molecules and textual descriptions, thereby enhancing the model's generative capabilities.[6]

In the molecule generation task, MolBridge-Gen-base achieves the best fingerprint-based similarity scores across MACCS, RDK , and Morgan metrics, and outperforms Atomas-base, which conducts an implicit fine-grained alignment, while reaching comparable performance to Atomas-large. This indicates that explicitly modeling local structure–text relationships enables the model to generate molecules that are semantically aligned with input descriptions. These suggest that our explicit fine-grained alignment strategy enhances the model's capacity to encode and decode chemically meaningful information, leading to improved performance in both captioning and generation tasks, even with smaller sizes.[7]

[5]Additional experimental results on PubChem324k are in Appendix G.

[6]More analysis of molecule captioning is in Appendix E.

[7]Generated examples are shown in Figure 7 and 8.

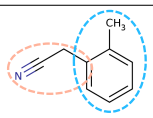
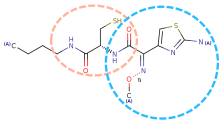
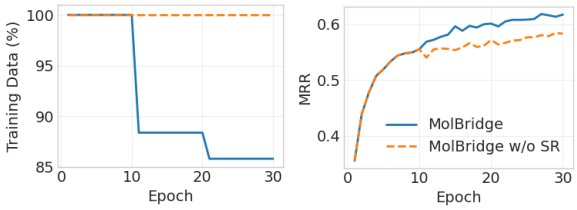
Molecule	Phrases found from PubChem	Phrases found by MolBridge	Phrases found by MolBridge w/o aug.
 CC1=CC=CC=C1CC#N	The molecule is a nitrile that is acetonitrile where one of the methyl hydrogens is substituted by a 2-methylphenyl group. It derives from an acetonitrile.	The molecule is a nitrile that is acetonitrile where one of the methyl hydrogens is substituted by a 2-methylphenyl group. It derives from an acetonitrile.	The molecule is a nitrile that is acetonitrile where one of the methyl hydrogens is substituted by a 2-methylphenyl group. It derives from an acetonitrile.
 CCCCNC(=O)[C@H](CS)NC(=O)/C (=N\\OC)/C1=CSC(=N1)N	The molecule is an amino acid amide that is a carboxamide obtained by formal condensation between N-butyl-L-cysteinamide and (2Z)-2-(2-amino-1,3-thiazol-4-yl)-2-(methoxyimino)acetic acid. It is a member of 1,3-thiazoles, an oxime O-ether, an amino acid amide and a L-cysteine derivative.	The molecule is an amino acid amide that is a carboxamide obtained by formal condensation between N-butyl-L-cysteinamide and (2Z)-2-(2-amino-1,3-thiazol-4-yl)-2-(methoxyimino)acetic acid. It is a member of 1,3-thiazoles, an oxime O-ether, an amino acid amide and a L-cysteine derivative.	None

Figure 2: Examples of chemical phrase retrieval. Identified substructures (dashed circles) and retrieved phrases from PubChem, MolBridge, and MolBridge w/o augmentation.



(a) Percentage of train data. (b) Performance (MRR; M2T).

Figure 3: Analysis of the impact of Self-Refinement (SR). We evaluate both MolBridge and MolBridge trained without SR on PCDes scaffold test set.

Phrases Extractor	T2M		M2T	
	R@1	MRR	R@1	MRR
ChemDataExtractor	**34.38**	**45.73**	**36.45**	**47.82**
GPT-4 + MolT5	23.89	34.99	27.33	38.88

Table 6: Evaluation results on PCDes scaffold test set with different phrase extractors for MolBridge.

Decompose Method	T2M		M2T	
	R@1	MRR	R@1	MRR
BRICS	**34.38**	**45.73**	**36.45**	**47.82**
RECAP	19.35	30.27	21.05	32.53

Table 7: Evaluation results on PCDes scaffold test set with different substructure extractors for MolBridge.

5 Analysis

In this section, we analyze the fine-grained alignment capability of MolBridge. Further analyses on model choice, ablation study, and error analysis on the molecular scale and complexity are provided in Appendix B, C, and F.

Case study. To investigate whether the local relations discovered by MolBridge accurately capture meaningful substructure–phrase correspondences, we conduct a qualitative evaluation. Figure 2 illustrates representative examples comparing retrieved chemical phrases from MolBridge, a baseline model trained without augmentation, and ground-truth phrases curated from the PubChem database. Each molecule is annotated with dashed circles indicating identified substructures, and the corresponding phrases are highlighted with matching colors. In both cases, MolBridge retrieves phrases that closely align with the ground truth, demonstrating its ability to identify valid local relationships. In contrast, the MolBridge trained without augmentation retrieves either irrelevant or no phrases. These results suggest that substructure-aware alignments enables the model to learn more precise and semantically meaningful mappings between molecular and textual fragments.

Impact of self-refinement. To assess the effectiveness of the self-refinement process in MolBridge, we compare retrieval performance between models trained with and without refinement, as shown in Figure 3. We observe that the refined model consistently outperforms the baseline over training epochs, despite filtering out approximately 15% of the training data in two stages. The majority of the removed pairs are substructure–caption relations, particularly cases where the entire molecule is incorrectly treated as a substructure during decomposition. As illustrated in Figure 4 in the Appendix, such cases include molecules misidentified as their own substructures, leading to invalid substructural relations, as well as overly generic captions that fail to capture substructure-level semantics. These results indicate that removing noisy supervision signals during training helps construct a cleaner and more informative dataset, ultimately improving model robustness and alignment quality.

Analysis on Fragment Extractor In our search for optimal tools to extract substructures and phrases, we employ various techniques and analyze

Method	Ranking↓
MolT5-large (Edwards et al., 2022)	2.4 (0/3/2)
Atomas-base (Zhang et al., 2025)	2.2 (1/2/2)
MolBridge-Gen-base	**1.4** (4/0/1)

Table 8: Human evaluation results. Ranking refers to the average ranking of human evaluation. The numbers in brackets indicate the counts of ranks 1, 2, and 3.

their differences by training MolBridge for three epochs. The retrieval performances are reported in Table 6 and 7.

For molecular decomposition, we use RECAP (Degen et al., 2008) and BRICS (Lewell et al., 1998). The key difference between them lies in the number of bond types considered, with BRICS enabling MolBridge to explore a broader range of substructures and yield more positive results. For phrase extraction, ChemDataExtractor (CDE) often produces incomplete outputs, so we design an LLM-based extractor (Li et al., 2024b). Using GPT-4 (OpenAI, 2023) on 10K sampled captions and fine-tuning MolT5-large, we expand phrase coverage. However, this reduces diversity, making CDE comparatively more effective. These findings underscore the importance of collecting diverse substructures and phrases for fine-grained alignment.

Human evaluation Following the human evaluation settings of Zhang et al. (2025), we conduct experiments in which annotators are asked to rank five randomly selected captions for each method according to their closeness to the ground truth. Specifically, we compare MolBridge-Gen-base against MolT5-large and Atomas-base with three human annotators. Table 8 reports the average rankings. MolBridge-Gen-base achieves the highest average ranking, placing first in 4 out of 5 generated captions in our human evaluation. More details, including evaluation instructions and generated examples, are provided in Appendix H.

6 Conclusion

We have presented MolBridge, a substructure-aware framework for fine-grained molecule–text alignment. By explicitly aligning molecular substructures with corresponding chemical phrases, MolBridge enables fragment-level representation learning that better captures subtle differences between molecules and their descriptions. Our experiments show that MolBridge consistently outperforms existing models across retrieval, property prediction, and generation tasks, demonstrating the effectiveness of leveraging localized alignment signals in multimodal molecular learning.

Limitations

While MolBridge demonstrates strong performance across multiple molecule–text tasks, several limitations remain.

Reliance on fragment extractors. MolBridge relies on external extractors to identify molecular substructures and chemical phrases for alignment. To address potential noise and inaccuracies introduced during this process, we incorporate a self-refinement mechanism that filters unreliable alignment signals and contributes meaningfully to the robustness of MolBridge. Even with this mechanism, the alignment quality is still influenced by the choice of extractor, as can be seen in Appendix 5. Developing more customized or domain-specific extractors could further improve the precision of fragment-level alignment in future work.

Limited structural exploration. MolBridge operates solely on 1D SMILES representations yet consistently outperforms models that directly utilize 2D molecular graphs. This result highlights the effectiveness of our fragment-level alignment approach. Nevertheless, incorporating additional structural information such as 2D topology or 3D conformations could provide complementary benefits (Liu et al., 2023c; Xiao et al., 2024). Extending MolBridge in this direction may further enhance its ability to model complex spatial relationships in molecular data.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3613–3618.

Steven H Bertz. 1981. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2025. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 354–379.

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

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157.

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.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

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,*, pages 4171–4186.

Jurgen Drews. 2000. Drug discovery: a historical perspective. *science*, 287(5460):1960–1964.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(5):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*, pages 375–413.

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.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*.

Shikun Feng, Lixin Yang, Wei-Ying Ma, and Yanyan Lan. 2023. Unimap: Universal smiles-graph representation learning. *CoRR*, abs/2310.14216.

Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. 2020. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. *Chem*, 6(7):1527–1542.

Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. 1998. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences*, 38(3):511–522.

Jiatong Li, Wei Liu, Zhihao Ding, Wenqi Fan, Yuqiang Li, and Qing Li. 2024a. Large language models are in-context molecule learners. *CoRR*, abs/2403.04197.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024b. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Trans. Knowl. Data Eng.*, 36(11):6071–6083.

Jiatong Li, Yunqing Liu, Wei Liu, Jingdi Lei, Di Zhang, Wenqi Fan, Dongzhan Zhou, Yuqiang Li, and Qing Li. 2024c. Molreflect: Towards in-context fine-grained alignments between molecules and texts. *CoRR*, abs/2411.14721.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Comput. Biol. Medicine*, 171:108073.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023a. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nat. Mac. Intell.*, 5(12):1447–1457.

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 61st Annual Meeting of the Association for Computational Linguistics*, pages 1606–1616.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. *CoRR*, abs/2307.09484.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*.

Juraj Mavracic, Callum J. Court, Taketomo Isazawa, R. Stephen Elliott, and Jacqueline M. Cole. 2021. Chemdataextractor 2.0: Autopopulated ontologies for materials science. *J. Chem. Inf. Model.*, 61(9):4280–4289.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23.

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.

Zijun Min, Bingshuai Liu, Liang Zhang, Jia Song, Jinsong Su, Song He, and Xiaochen Bo. 2024. Exploring optimal transport-based multi-grained alignments for text-molecule retrieval. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 2317–2324.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Jun-Hyung Park, Hyuntae Park, Yeachan Kim, Woosang Lim, and SangKeun Lee. 2024. Moleco: Molecular contrastive learning with chemical language models for molecular property prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track*, pages 408–420.

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 Association for Computational Linguistics*, pages 1216–1240.

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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

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

Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. 2019. *Deep Learning for the Life Sciences*. O'Reilly Media. `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mac. Intell.*, 4(12):1256–1264.

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. *J. Chem. Inf. Model.*, 55(10):2111–2120.

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. *CoRR*, abs/2209.05481.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436.

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.

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.

Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and 1 others. 2023. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585.

Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z. Li. 2023. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*.

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G. Honavar. 2024. Geomclip: Contrastive geometry-text pre-training for molecules. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1250–1256.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. 2024. Multimodal molecular pretraining via modality blending. In *The Twelfth International Conference on Learning Representations*.

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.

Yikun Zhang, Geyan Ye, Chaohao Yuan, Bo Han, Long-Kai Huang, Jianhua Yao, Wei Liu, and Yu Rong. 2025. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *The Thirteenth International Conference on Learning Representations*.

Appendix

A Implementation Details

We initialize MolBridge with two specialized encoders: MoLFormer-XL (Ross et al., 2022)[8] for SMILES input and SciBERT (Beltagy et al., 2019)[9] for textual input. Noisy supervision signals are filtered out every 10 epochs. We explored learning rates from the set {1e-5, 3e-5, 1e-4, 5e-4} and batch sizes from {32, 64, 128, 256, 512}, and report results from the best-performing configurations. The sequence length for both SMILES and text inputs is fixed to 256 tokens. For MolBridge, we use a learning rate of 2e-4 and a batch size of 256 during pretraining. For MolBridge-Gen, which is built upon MolT5 (Edwards et al., 2022), we pre-train the model with a learning rate of 5e-4 and fine-tune it with 1e-4. A batch size of 128 is used for both pretraining and fine-tuning of MolBridge-Gen.

For molecular property prediction using MolBridge trained over 10 epochs, we conducted a full grid search over the learning rates and batch sizes mentioned above to select the best-performing configuration for each task. All models are trained for 50 epochs using canonicalized SMILES and the AdamW optimizer. We apply gradient accumulation to handle large batch sizes across four NVIDIA A5000 GPUs. To extract structure–phrase pairs for generative training, we empirically set the MolBridge score threshold τ to 0.3. For retrieval tasks, we adopt a zero-shot setting to ensure a fair comparison with previous works (Liu et al., 2023c; Zhang et al., 2025).

B Analysis on Model Choice

We aim to assess the impact of molecular and scientific literature understanding on the encoders used as the backbone of MolBridge. To this end, we initialize the model with ChemBERTa (Chithrananda et al., 2020), which has been reported to show lower molecular property prediction performance compared to MoLFormer-XL, and with BERT (Devlin et al., 2019), trained on general domain texts. We train the models for three epochs, as these results show a similar tendency to final results under the same training objectives. The retrieval results on the PCDes scaffold test set in Table 9 show decreases of 20%p and 12.8%p in MRR for text-to-molecule and molecule-to-text retrieval, respectively. This underscores the critical role of molecular understanding and scientific literature comprehension in molecule and text alignment.

SMILES Encoder	Text Encoder	T2M R@1	T2M MRR	M2T R@1	M2T MRR
MoLFormer-XL	SciBERT	**34.38**	**45.73**	**36.45**	**47.82**
ChemBERTa	BERT	15.57	25.73	23.42	34.98

Table 9: Evaluation results on PCDes scaffold test set with different model choices for MolBridge.

C Ablation study

To validate the effectiveness of each component in the MolBridge framework, we conduct an ablation study by training the same model for three epochs under different objective configurations and augmentation settings. The results, summarized in Table 10, show that each proposed component contributes meaningfully to overall performance.

First, we observe that removing our augmentation strategy causes a dramatic drop in retrieval performance, with an average decrease of 21.8%p in MRR compared to the full model. This highlights the importance of explicitly modeling substructural relationships for fine-grained alignment.

When we remove either the substructure-caption or the molecule-phrase pairs, the performance still improves relative to the MolBridge w/o augmentation. This indicates that even partial fragment-level supervision is beneficial. Among the two, the absence of molecule-phrase alignment leads to a larger drop, which suggests that identifying diverse and accurate phrases plays an important role in learning meaningful semantic correspondences.

We also find that removing multi-positive contrastive learning leads to a 4.28%p decrease in average MRR. This result supports the assumption that a single molecule or caption can correspond to multiple relevant fragments, and confirms that the proposed objective effectively captures such compositional relationships.

Lastly, we examine the effect of removing the type classification loss used during the self-refinement process. The results show a slight drop in molecule-to-text retrieval performance, although a marginal improvement is observed in the reverse direction. Despite this, the overall benefit of the self-refinement mechanism remains clear, as it enables the model to filter noisy alignment signals during training and contributes to the stability and

[8] ibm-research/MoLFormer-XL-both-10pct
[9] allenai/scibert_scivocab_uncased

Methods	Text to Molecule				Molecule to Text			
	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
MolBridge	**34.38**	58.84	67.86	45.73	**36.45**	**61.38**	**70.33**	**47.82**
MolBridge w/o type classification	34.31	**59.87**	**68.23**	**46.01**	35.12	60.57	69.23	46.93
MolBridge w/o multi-positive contrastive learning	29.27	54.73	63.85	41.03	32.14	58.30	67.56	43.96
MolBridge w/o substructure-caption pairs	23.92	49.42	60.11	35.83	26.23	51.19	61.68	38.02
MolBridge w/o molecule-phrase pairs	19.48	42.43	52.62	30.55	22.22	48.51	59.51	34.51
MolBridge w/o augmentation	13.97	34.65	45.24	23.96	14.73	37.82	49.48	25.96

Table 10: Ablation study of MolBridge on PCDes scaffold test set. Each model is trained over 3 epochs, as we observed that these results show a similar tendency to the final results.

Method	τ	# Pairs	BLEU-2↑	BLUE-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-base	-	-	0.549	0.457	0.635	0.481	0.576	0.580
Atomas-base	-	-	0.632	0.545	0.686	0.545	0.626	-
MolBridge-Gen-base	0.2	163k	0.629	0.547	0.687	0.551	0.631	0.651
MolBridge-Gen-base	0.3	32k	**0.674**	**0.605**	**0.724**	**0.609**	**0.676**	**0.693**
MolBridge-Gen-base	0.4	5k	0.543	0.447	0.619	0.463	0.560	0.567

Table 11: Molecule captioning performance of MolBridge-Gen with different cosine similarity thresholds on ChEBI-20.

Method	τ	# Pairs	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
MolT5-base	-	-	0.854	0.318	16.32	0.889	0.813	0.750	0.958
Atomas-base	-	-	**0.868**	0.343	**13.76**	0.908	0.827	0.773	0.971
MolBridge-Gen-base	0.2	163k	0.834	0.278	16.11	0.901	0.822	0.758	0.956
MolBridge-Gen-base	0.3	32k	0.842	**0.358**	15.66	**0.918**	**0.854**	**0.798**	0.956
MolBridge-Gen-base	0.4	5k	0.783	0.173	21.82	0.853	0.750	0.670	0.943

Table 12: Molecule generation performance of MolBridge-Gen with different cosine similarity thresholds on ChEBI-20.

robustness of representations, as discussed in Section 5.

D Analysis on Cosine Similarity Threshold

We evaluate the effect of different threshold values (τ) on the generative performance of our model. Table 11 and 12 show the results of pre-training and fine-tuning MolBridge-Gen on the ChEBI-20 dataset using three different thresholds. All models were trained for the same number of steps to ensure a fair comparison.

Our experiments indicate that a threshold of 0.3 yields the best overall performance. When the threshold is set to 0.2, MolBridge-Gen achieves results comparable to the Atomas, suggesting our approach remains robust. However, at a threshold of 0.4, performance drops to the level of, or slightly below, the backbone of our model (MolT5). We attribute this to the substantial reduction in the number of original pairs containing fragment matches, leading to overfitting due to limited training data.

E Analysis on Pretraining Strategy for Generation

We investigate the effect of our pretraining strategy on molecule captioning by comparing MolBridge-Gen with its baseline, as shown in Table 13. To this end, we fine-tune both the original MolT5-base and the MolT5-base pretrained on our curated dataset without augmentation.

The results in Table 13 indicate that initial pretraining enhances generation quality, yielding consistent improvements across all evaluation metrics. More importantly, MolBridge-Gen, which is trained on only 32k molecule–caption pairs augmented with local alignment signals discovered by MolBridge (Figure 5), surpasses all other settings by a substantial margin. In particular, it achieves gains of 9.4 to 11.0 points in BLEU-2, BLEU-4, and METEOR scores over the original MolT5 model, despite using far fewer training examples. These findings demonstrate that the local structure–language correspondences captured by MolBridge provide more informative supervision than large-scale pretraining without alignment, under-

Method	# Pairs	BLEU-2↑	BLUE-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-base	-	0.549	0.457	0.635	0.481	0.576	0.580
MolT5-base + inital training	432k	0.580	0.495	0.657	0.516	0.600	0.611
MolBridge-Gen-base	32k	**0.674**	**0.605**	**0.724**	**0.609**	**0.676**	**0.693**

Table 13: Analysis on pre-training approach of MolBridge-Gen on CheBI-20 test set. Initial training refers to the training of the model before fine-tuning with our curated dataset without augmentation described in Section 3.3.

Methods	T2M		M2T	
	R@1	R@20	R@1	R@20
1D SMILES + 2D Graph				
MoMu-S (Su et al., 2022)	-	75.5	-	79.1
MoMu-K (Su et al., 2022)	-	79.0	-	80.2
MoleculeSTM (Liu et al., 2023a)	35.8	77.0	39.5	80.4
MolCA (Liu et al., 2023c)	46.0	82.3	48.1	85.6
1D SMILES				
SciBERT (Beltagy et al., 2019)	-	60.8	-	60.7
KV-PLM (Zeng et al., 2022)	-	64.3	-	75.9
MolBridge	**54.2**	**86.7**	**57.0**	**88.6**

Table 14: Zero-shot molecule-text retrieval performance on PCDes test set. The results of baselines are borrowed from (Liu et al., 2023c).

scoring the advantage of explicitly modeling fine-grained relationships for molecule captioning.

F Error Analysis on Molecular Scale and Complexity

To better understand the behavior of MolBridge-Gen, we conduct an error analysis on the ChEBI-20 dataset by dividing the test set according to the median values of molecular scale (atom count) and molecular complexity (BertzCT (Bertz, 1981)). We evaluate our model on low/high scale and low/high complexity subsets for both molecule captioning and molecule generation tasks. Results are shown in Table 15, 16, 17, and 18. We find that MolBridge-Gen makes more errors on molecules with lower scale and lower complexity, while performance is higher for larger and more complex molecules. This suggests that complex and large molecules contain richer compositional relationships between substructures and language, which our model is designed to capture.

G Evaluation on PubChem324k

To further assess the generalizability of MolBridge-Gen beyond ChEBI-20, we evaluate the model on the PubChem324k dataset (Liu et al., 2023c) for both molecule captioning and molecule generation tasks. As shown in Tables 19 and 20, MolBridge-Gen achieves strong and consistent performance across different models and evaluation metrics.

H Details of Human Evaluation

We invited three NLP experts as annotators, specifically those with prior publications or project experience in related domains. The annotators participated on a voluntary basis, and no payment was provided. For the evaluation, we randomly sampled five molecules. For each molecule, annotators were given the molecule structure, its ground-truth caption, and three generated captions produced by MolT5-large, Atomas-base, and MolBridge-Gen-base. The generated captions were presented in randomized order, and annotators were instructed to "rank the three models (Model 1, Model 2, Model 3) according to their relevance to the ground-truth caption." The five evaluation examples used in this study are shown in Figure 6.

MolBridge-Gen-base	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
Low scale set	0.660	0.589	0.715	0.598	0.667	0.683
High scale set	**0.687**	**0.620**	**0.735**	**0.622**	**0.687**	**0.704**
Original set	0.674	0.605	0.724	0.609	0.676	0.693

Table 15: Evaluation results on the ChEBI-20 test set with respect to molecular scale (Molecule Captioning).

MolBridge-Gen-base	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
Low complexity set	0.667	0.595	0.717	0.599	0.670	0.686
High complexity set	**0.680**	**0.614**	**0.732**	**0.620**	**0.683**	**0.701**
Original set	0.674	0.605	0.724	0.609	0.676	0.693

Table 16: Evaluation results on the ChEBI-20 test set with respect to molecular complexity (Molecule Captioning).

MolBridge-Gen-base	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
Low scale set	**0.844**	**0.438**	**6.538**	0.900	0.822	0.777	**0.988**
High scale set	0.831	0.271	25.506	**0.938**	**0.892**	**0.821**	0.921
Original set	0.842	0.358	15.660	0.918	0.854	0.798	0.956

Table 17: Evaluation results on the ChEBI-20 test set with respect to molecular scale (Molecule Generation).

MolBridge-Gen-base	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
Low complexity set	**0.860**	**0.433**	**6.382**	0.909	0.837	0.795	**0.992**
High complexity set	0.827	0.282	24.948	**0.927**	**0.873**	**0.801**	0.919
Original set	0.842	0.358	15.660	0.918	0.854	0.798	0.956

Table 18: Evaluation results on the ChEBI-20 test set with respect to molecular complexity (Molecule Generation).

Method	#Params	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-base (Edwards et al., 2022)	248M	0.301	0.209	0.403	0.251	0.338	0.356
MolCA-Galactica-1.3B (Liu et al., 2023c)	1.3B	0.387	0.303	0.502	0.359	0.445	0.456
ICMA-Mistral-7B (Li et al., 2024a)	7B	0.416	0.345	0.505	0.367	0.453	0.464
MolBridge-Gen-base	248M	**0.479**	**0.421**	**0.596**	**0.486**	**0.554**	**0.549**

Table 19: Molecule captioning results on the PubChem324k dataset.

Method	#Params	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
ICMA-Mistral-7B (Li et al., 2024a)	7B	0.526	0.163	62.25	0.799	0.678	0.573	0.935
Atomas-large (Zhang et al., 2025)	825M	0.734	–	28.186	0.773	0.637	0.535	0.945
MolBridge-Gen-base	248M	**0.742**	**0.200**	**27.294**	**0.829**	**0.740**	**0.642**	0.934

Table 20: Molecule generation results on the PubChem324k dataset.

Task	Template
SMILES-to-caption	Provide a whole description of this molecule: <input>
Caption-to-SMILES	Provide a molecule based on this description: <input>
Substructure-to-phrase	Provide a keyword of this substructure: <input>
phrase-to-substructure	Provide a substructure based on this keyword: <input>

Table 21: Prompt templates that are used for our multi-task pre-training of MolBridge-Gen.

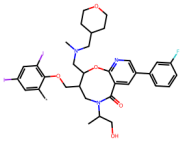
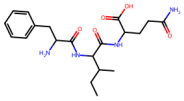
Filtered	Graph	SMILES & Description
Example 1		This molecule is a phenylpyridine. CC1CN(C(C)CO)C(=O)c2cc(-c3cccc(F)c3)cnc2OC1CN(C)CC1CCOCC1
Example 2		This molecule is a peptide. CCC(C)C(NC(=O)C(N)Cc1ccccc1)C(=O)NC(CCC(N)=O)C(=O)O

Figure 4: Examples of filtered augmented pairs.

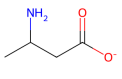
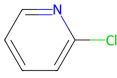
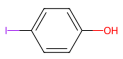
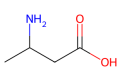
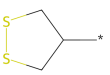
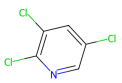
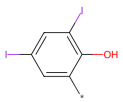
Phrase	3-Aminobutyrate	Dithiolanes	Chloropyridine	Iodophenol
Relevant Molecule	 CC(CC(=O)[O-])N (3-Aminobutyrate)	 C1CSSC1 (1,2-Dithiolane)	 C1=CC=NC(=C1)Cl (2-Chloropyridine)	 C1=CC(=CC=C1O)I (4-Iodophenol)
Retrieved Substructure	 C1=CC=C(C=C1)C(=O)O	 *C1CSSC1	 *c1nc(Cl)c(Cl)cc1Cl	 *c1cc(I)cc(I)c1O

Figure 5: Retrieved substructures using chemical phrase by MolBridge along with the phrase-relevant molecules found in PubChem.

Input	COC1=C2C(=CC3=C1OCO3)OC=C(C2=O)C4=CC=CC=C4O
Target	The molecule is a hydroxyisoflavone that is isoflavone substituted by a hydroxy group at position 2', a methoxy group at position 5 and a methylenedioxy group across positions 6 and 7 respectively. It has a role as a plant metabolite. It is a hydroxyisoflavone and a methoxyisoflavone. It derives from an isoflavone.
Ours	The molecule is a member of the class of psoralens that is psoralen substituted by a methoxy group at position 7 and a hydroxy group at position 4'. It has been isolated from Pisonia aculeata. It has a role as a metabolite and a plant metabolite. It is a member of psoralens and an aromatic ether. It derives from a psoralen.
Atomas	The molecule is a member of the class of 7-hydroxyisoflavones that is isoflavone substituted by a methoxy group at position 7 and a hydroxy group at position 4'. It has been isolated from Pisonia aculeata. It has a role as a plant metabolite. It is a member of 7-hydroxyisoflavones and a member of 4'-hydroxyisoflavones. It derives from an isoflavone.
MolT5	The molecule is a member of the class of 7-hydroxyisoflavones that is 7-hydroxyisoflavone and in which the phenyl group at position 3 is replaced by a 1,3-benzodioxol-5-yl group. It has a role as an antiprotozoal drug and a plant metabolite. It is a member of benzodioxoles and a member of 7-hydroxyisoflavones. It is a conjugate acid of a pseudobaptigenin(1-).
Input	C[S+](CCC(=O)C(=O)[O-])C[C@@H]1[C@H]([C@H]([C@@H](O1)N2C=NC3=C(N=CN=C32)N)O)O
Target	The molecule is a sulfonium betaine that is the conjugate base of S-adenosyl-4-methylthio-2-oxobutanoic acid, arising from deprotonation of the carboxy group. It has a role as a Saccharomyces cerevisiae metabolite. It is a sulfonium betaine and a carboxylic acid anion. It derives from a 4-methylthio-2-oxobutanoate. It is a conjugate base of a S-adenosyl-4-methylthio-2-oxobutanoic acid.
Ours	The molecule is a sulfonium betaine that is the conjugate base of S-adenosyl-4-methylthio-2-oxobutanoic acid, arising from deprotonation of the carboxy group. It has a role as a Saccharomyces cerevisiae metabolite. It is a sulfonium betaine and a carboxylic acid anion. It derives from a 4-methylthio-2-oxobutanoate. It is a conjugate base of a S-adenosyl-4-methylthio-2-oxobutanoic acid.
Atomas	The molecule is a sulfonium betaine that is the conjugate base of S-methyl-3-oxopropanoic acid, obtained by deprotonation of the carboxy group; major species at pH 7.3. It is a conjugate base of a S-methyl-3-oxopropanoic acid.
MolT5	The molecule is a sulfonium betaine obtained by deprotonation of the carboxy group of S-adenosyl-L-methionine. Major microspecies at pH 7.3 It is a L-alpha-amino acid zwitterion and a sulfonium betaine. It is a conjugate base of a S-adenosyl-L-methionine.
Input	CC1=CC(=CC(=C1)O)C
Target	The molecule is a member of the class of phenols that phenol substituted by methyl groups at positions 3 and 5. It has a role as a xenobiotic metabolite. It derives from a hydride of a m-xylene.
Ours	The molecule is a member of the class of phenols that is p-xylene substituted by methyl groups at positions 3 and 5. It has a role as a volatile oil component and an animal metabolite. It derives from a p-xylene.
Atomas	The molecule is a member of the class of phenols that is phenol substituted by a methyl group at position 3 and a hydroxy group at position 4. It has a role as a bacterial xenobiotic metabolite. It is a member of phenols and a member of guaiacols. It derives from a guaiacol.
MolT5	The molecule is a 5-alkylresorcinol in which the alkyl group is specified as methyl. It has a role as an Aspergillus metabolite. It is a 5-alkylresorcinol and a dimethylresorcinol.
Input	C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O[C@H]([C@H](CO)O)[C@@H]([C@@H](CO)O)O)O)O)O)O
Target	The molecule is a glycosyl alditol consisting of D-glucitol in which the hydroxy group at position 3 has been converted into the corresponding beta-D-glucopyranosyl derivative.
Ours	The molecule is a glycosyl alditol consisting of beta-D-galactopyranose and D-galactitol residues joined in sequence by a (1->4) glycosidic bond. It derives from a beta-D-galactose and a galactitol.
Atomas	The molecule is a disaccharide that is D-glycero-alpha-D-manno-heptopyranose in which the hydroxy group at position 3 has been converted into the corresponding alpha-D-galactopyranoside. It is an alpha-D-galactoside and a glycosylgalactose. It derives from a D-glycero-alpha-D-manno-heptopyranose.
MolT5	The molecule is an alpha-D-glucoside consisting of D-glucitol having an alpha-D-glucosyl residue attached at the 4-position. Used as a sugar substitute. It has a role as a metabolite, a laxative and a sweetening agent. It derives from an alpha-D-glucose and a D-glucitol.
Input	CC1=CN(C(=O)NC1=O)[C@H]2C[C@@H]([C@H](O2)COP(=O)(O)OP(=O)(O)O[C@H]3[C@@H]([C@H]([C@@H](O3)[C@@H](C)O)O)O)O
Target	The molecule is a dTDP-sugar having beta-D-fucofuranose as the sugar component. It has a role as a metabolite. It is a conjugate acid of a dTDP-beta-D-fucofuranose(2-).
Ours	The molecule is a dTDP-sugar having beta-L-rhamnose as the sugar component. It has a role as a metabolite. It derives from a dTDP-L-rhamnose. It is a conjugate acid of a dTDP-beta-L-rhamnose(2-).
Atomas	The molecule is a dTDP-sugar having alpha-D-glucose as the sugar component. It has a role as a bacterial metabolite. It is a dTDP-sugar and a secondary alcohol. It derives from an alpha-D-glucose. It is a conjugate acid of a dTDP-alpha-D-glucose(2-).
MolT5	The molecule is a dTDP-sugar having alpha-D-glucopyranose as the sugar portion. It has a role as an Escherichia coli metabolite and a mouse metabolite. It is a conjugate acid of a dTDP-alpha-D-glucose(2-).

Figure 6: Examples used for human evaluation.

Input	CCCCCCCCCCCCCCCCCOC[C@H](COP(=O)([O-])OCC[N+](C)(C)C)OC(=O)CCC(=O)[O-]
Target	The molecule is an anionic phospholipid obtained by deprotonation of the free carboxy group of 1-hexadecyl-2-succinyl-sn-glycero-3-phosphocholine; major species at pH 7.3. It is an anionic phospholipid and a monocarboxylic acid anion. It is a conjugate base of a 1-hexadecyl-2-succinyl-sn-glycero-3-phosphocholine.
Ours	The molecule is an anionic phospholipid obtained by deprotonation of the free carboxy group of 1-hexadecyl-2-glutaryl-sn-glycero-3-phosphocholine; major species at pH 7.3. It is an anionic phospholipid and a monocarboxylic acid anion. It is a conjugate base of a 1-hexadecyl-2-glutaryl-sn-glycero-3-phosphocholine.
MolT5	The molecule is a 2-acyl-1-alkyl-sn-glycero-3-phosphocholine in which the alkyl and the acyl groups at positions 1 and 2 are specified as hexadecyl and succinyl respectively; major species at pH 7.3. It is a conjugate base of a 1-hexadecyl-2-succinyl-sn-glycero-3-phosphocholine.
Input	CC(=O)N[C@@H](CCCN=C(N)N)C(=O)[O-]
Target	The molecule is an N-acyl-L-alpha-amino acid anion arising from deprotonation of the carboxy group of N(alpha)-acetyl-L-arginine; major species at pH 7.3. It is a conjugate base of a N(alpha)-acetyl-L-arginine.
Ours	The molecule is an N-acyl-L-alpha-amino acid anion arising from deprotonation of the carboxy group of N(alpha)-acetyl-L-arginine; major species at pH 7.3. It is a conjugate base of a N(alpha)-acetyl-L-arginine.
MolT5	The molecule is a monocarboxylic acid anion that is the conjugate base of N-acetyl-L-arginine, obtained by deprotonation of the carboxy group; major species at pH 7.3. It is a conjugate base of a N-acetyl-L-arginine.
Input	CCCCCCCCCCOS(=O)(=O)[O-]
Target	The molecule is an organosulfate oxoanion that is the conjugate base of decyl hydrogen sulfate. Isolated from Daphnia pulex, it induces morphological changes of phytoplankton Scenedesmus gutwinskii. It has a role as a kairomone, a Daphnia pulex metabolite and a marine metabolite. It is a conjugate base of a decyl hydrogen sulfate.
Ours	The molecule is an organosulfate oxoanion that is the conjugate base of decyl hydrogen sulfate. Isolated from Daphnia pulex, it induces morphological changes of phytoplankton Scenedesmus gutwinskii. It has a role as a Daphnia pulex metabolite, a kairomone and a marine metabolite. It is a conjugate base of a decyl hydrogen sulfate.
MolT5	The molecule is an organosulfate oxoanion that is the conjugate base of octyl hydrogen sulfate. It has been isolated from Daphnia pulex and has been shown to cause morphological changes in the green alga Scenedesmus gutwinskii. It has a role as a kairomone and a Daphnia pulex metabolite. It is a conjugate base of an octyl hydrogen sulfate.
Input	CC(=O)N[C@@H](CC(=O)[O-])C(=O)N[C@@H](CCC(=O)[O-])C(=O)N[C@@H](CCC(=O)[O-])C(=O)[O-
Target	The molecule is a peptide anion obtained by deprotonation of the four carboxy groups of Ac-Asp-Glu-Glu; major species at pH 7.3. It is a conjugate base of an Ac-Asp-Glu-Glu.
Ours	The molecule is a peptide anion obtained by deprotonation of the four carboxy groups of Ac-Asp-Glu-Glu; major species at pH 7.3. It is a conjugate base of an Ac-Asp-Glu-Glu.
MolT5	The molecule is a peptide anion obtained by deprotonation of the carboxy groups of N-acetyl-L-gamma-glutamyl-L-glutamic acid; major species at pH 7.3. It is a conjugate base of a N-acetyl-L-gamma-glutamyl-L-glutamic
Input	CSCCCCCCCCC(C(=O)O)N(O)O
Target	The molecule is an N,N-dihydroxy-alpha-amino acid having a 9-thiadecyl substituent at the 2-position. It derives from a hexahomomethionine. It is a conjugate acid of a N,N-dihydroxyhexahomomethioninate
Ours	The molecule is an N,N-dihydroxy-alpha-amino acid having a 9-thiadecyl substituent at the 2-position. It derives from a hexahomomethionine. It is a conjugate acid of a N,N-dihydroxyhexahomomethioninate
MolT5	The molecule is an N,N-dihydroxy-alpha-amino acid having a 7-thiaoctyl substituent at the 2-position. It derives from a tetrahomomethionine. It is a conjugate acid of a N,N-dihydroxytetrahomomethioninate

Figure 7: Examples of generated captions from input SMILES.

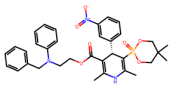
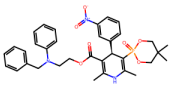
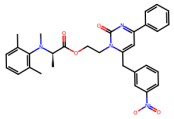
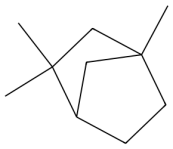
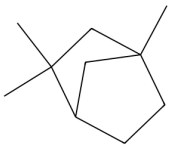
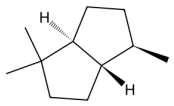
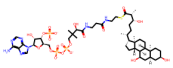
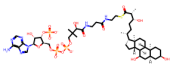
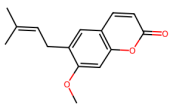
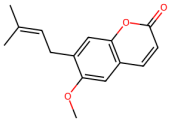
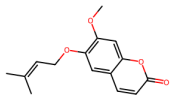
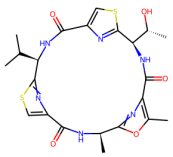
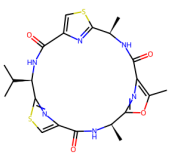
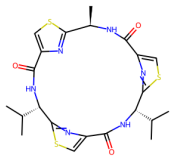
Input	The molecule is a 2-[benzyl(phenyl)amino]ethyl 5-(5,5-dimethyl-2-oxido-1,3,2-dioxaphosphinan-2-yl)-2,6-dimethyl-4-(3-nitrophenyl)-1,4-dihydropyridine-3-carboxylate that has (S)-configuration. It is a blocker of L- and T-type Ca(2+) channels. It has a role as a calcium channel blocker. It is an enantiomer of a (R)-efonidipine.					
Target & Predicted	Target		Ours (MACCS FTS=1.0)		MolT5 (MACCS FTS=0.26)	
Input	The molecule is a monoterpene that is bicyclo[2.2.1]heptane substituted by methy groups at positions 1, 3 and 3. It is a monoterpene, a terpenoid fundamental parent and a carbobicyclic compound.					
Target & Predicted	Target		Ours (MACCS FTS=1.0)		MolT5 (MACCS FTS=0.26)	
Input	The molecule is an acyl-CoA(4-) arising from deprotonation of phosphate and diphosphate OH groups of (24R,25R)-3alpha,7alpha,24-trihydroxy-5beta-cholestan-26-oyl-CoA; major species at pH 7.3. It is a conjugate base of a (24R,25R)-3alpha,7alpha,24-trihydroxy-5beta-cholestan-26-oyl-CoA.					
Target & Predicted	Target		Ours (MACCS FTS=1.0)		MolT5 (MACCS FTS=-1)	Invalid
Input	The molecule is a member of the class of coumarins in which the coumarin ring is substituted at positions 6 and 7 by a 3-methylbut-2-en-1-yl group and a methoxy group, respectively. A natural product found in Citropsis articulata. It has a role as a plant metabolite and an anticoagulant. It is a member of coumarins and an aromatic ether. It derives from a 7-demethylsuberosin.					
Target & Predicted	Target		Ours (MACCS FTS=0.92)		MolT5 (MACCS FTS=0.66)	
Input	The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.					
Target & Predicted	Target		Ours (MACCS FTS=0.86)		MolT5 (MACCS FTS=0.61)	

Figure 8: Visualized examples of generated SMILES from input caption.

Method	# Params	BLEU-2↑	BLUE-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
1D SELFIES + 2D Graph							
Mol-Instructions (Fang et al., 2024)	7B	0.249	0.171	0.311	0.203	0.239	0.271
InstructMol-GS (Cao et al., 2025)	7B	0.475	0.371	0.566	0.394	0.502	0.509
1D SELFIES + IUPAC name							
BioT5+ (Pei et al., 2024)	252M	0.666	0.591	0.710	0.584	0.650	0.681
1D SMILES + 2D Graph + 2D Image							
GIT-Mol (Liu et al., 2024)	700M	0.352	0.263	0.575	0.485	0.560	0.533
1D SMILES + 2D Graph + Knowledge Graph							
MolFM-small (Luo et al., 2023)	136M	0.542	0.452	0.623	0.469	0.562	0.564
MolFM-base (Luo et al., 2023)	296M	0.585	0.498	0.653	0.508	0.594	0.607
1D SMILES + 2D Graph							
MoMu-small (Su et al., 2022)	82M	0.532	0.445	-	-	0.564	0.557
MoMu-base (Su et al., 2022)	252M	0.549	0.462	-	-	0.575	0.576
MoMu-large (Su et al., 2022)	780M	0.599	0.462	-	-	0.593	0.597
MolCA (Galactica-125M) (Liu et al., 2023c)	125M	0.612	0.526	0.674	0.521	0.606	0.636
MolCA (Galactica-1.3B) (Liu et al., 2023c)	1.3B	0.620	0.531	0.681	0.537	0.618	0.651
ICMA (Galatica-125M) (Li et al., 2024a)	125M	0.636	0.565	0.674	0.536	0.615	0.648
ICMA (Mistral-7B) (Li et al., 2024a)	7B	0.651	0.581	0.686	0.550	0.625	0.661
1D SMILES + Context Examples							
MolReFlect (Li et al., 2024c)	7B	**0.676**	**0.608**	0.703	0.571	0.644	0.680
1D SMILES							
MolT5-small (Edwards et al., 2022)	77M	0.532	0.445	0.627	0.477	0.583	0.543
MolT5-base (Edwards et al., 2022)	248M	0.540	0.457	0.634	0.485	0.578	0.569
MolT5-large (Edwards et al., 2022)	783M	0.594	0.508	0.654	0.510	0.594	0.614
Text+Chem T5 (Christofidellis et al., 2023)	220M	0.625	0.542	0.682	0.543	0.622	0.648
MolXPT (Liu et al., 2023b)	350M	0.594	0.505	0.660	0.511	0.597	0.626
MolReGPT (GPT-3.5) (Li et al., 2024b)	-	0.565	0.482	0.623	0.450	0.543	0.585
MolReGPT (GPT-4) (Li et al., 2024b)	-	0.607	0.525	0.634	0.476	0.562	0.610
Atomas-base (Zhang et al., 2025)	271M	0.632	0.545	0.685	0.545	0.626	-
MolReFlect w/o Examples (Li et al., 2024c)	7B	0.617	0.539	0.657	0.510	0.593	0.623
MolBridge-Gen-small	82M	0.625	0.542	0.686	0.549	0.629	0.649
MolBridge-Gen-base	248M	0.674	0.605	**0.724**	**0.609**	**0.676**	**0.693**

Table 22: Results of molecule captioning task on CheBI-20 test set.

Method	BLEU↑	EM↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	Validity↑
1D SELFIES + IUPAC name							
BioT5+ (Pei et al., 2024)	0.872	**0.522**	12.77	0.907	0.835	0.779	**1.000**
1D SMILES + 2D Graph + 2D Image							
GIT-Mol (Liu et al., 2024)	0.756	0.051	26.32	0.738	0.582	0.519	0.928
1D SMILES + 2D Graph + Knowledge Graph							
MolFM-small (Luo et al., 2023)	0.803	0.169	20.86	0.834	0.721	0.662	0.859
MolFM-base (Luo et al., 2023)	0.822	0.210	19.45	0.854	0.758	0.758	0.892
1D SMILES + 2D Graph							
MoMu-small (Su et al., 2022)	0.800	0.150	21.45	0.818	0.709	0.651	0.858
MoMu-base (Su et al., 2022)	0.815	0.183	20.52	0.847	0.737	0.678	0.863
ICMA (Galatica-125M) (Li et al., 2024a)	0.836	-	21.48	0.893	0.809	0.743	0.825
ICMA (Mistral-7B) (Li et al., 2024a)	0.855	-	18.73	0.916	0.837	0.789	0.891
1D SMILES + Context Examples							
MolReFlect (Li et al., 2024c)	**0.903**	0.510	**11.84**	**0.929**	**0.860**	**0.813**	0.977
1D SMILES							
MolT5-small (Edwards et al., 2022)							
MolT5-base (Edwards et al., 2022)	0.779	0.082	25.19	0.788	0.662	0.602	0.787
MolT5-large (Edwards et al., 2022)	0.854	0.318	16.32	0.889	0.813	0.750	0.958
Text+Chem T5 (Christofidellis et al., 2023)	0.853	0.322	16.87	0.901	0.816	0.757	0.943
MolXPT (Liu et al., 2023b)	-	0.215	-	0.859	0.757	0.667	0.983
MolReGPT (GPT-3.5) (Li et al., 2024b)	0.790	0.139	24.91	0.847	0.708	0.624	0.887
MolReGPT (GPT-4) (Li et al., 2024b)	0.857	0.280	17.14	0.903	0.805	0.739	0.899
Atomas-base (Zhang et al., 2025)	0.868	0.343	13.76	0.908	0.827	0.773	0.971
Atomas-large (Zhang et al., 2025)	0.874	0.387	12.70	0.914	0.841	0.788	0.980
MolReFlect w/o Examples (Li et al., 2024c)	0.886	0.430	13.99	0.916	0.828	0.775	0.981
MolBridge-Gen-small	0.827	0.266	16.88	0.898	0.820	0.751	0.947
MolBridge-Gen-base	0.842	0.358	15.66	0.918	0.854	0.798	0.956

Table 23: Results of text-based de novo molecule generation on CheBI-20 test set.