UniMoT: Unified Molecule-Text Language Model with Discrete Token Representation

Shuhan Guo[1], Yatao Bian[2], Ruibing Wang[3], Nan Yin[4,†] , Quanming Yao[1,5]
[1]Department of Electronic Engineering, Tsinghua University
[2]Tencent AI Lab
[3]School of Mechanical Engineering, Northwestern Polytechnical University
[4]Hong Kong University of Science and Technology
[5]State Key Laboratory of Space Network and Communications, Tsinghua University
{guoshuhan, qyaoaa}@tsinghua.edu.cn
{yatao.bian, yinnan8911}@gmail.com
wrb5261@mail.nwpu.edu.cn

Abstract

The remarkable success of Large Language Models (LLMs) across diverse tasks has driven the research community to extend their capabilities to molecular applications. However, most molecular LLMs employ adapter-based architectures that fail to equally integrate molecule and text modalities and lack explicit supervision signals for the molecular modality. To address these issues, we introduce **UniMoT**, a **Uni**fied **Mo**lecule-**T**ext LLM adopting a tokenizer-based architecture that expands the vocabulary of LLMs with molecule tokens. Specifically, we introduce a Vector Quantization-driven tokenizer that incorporates a Q-Former to bridge the modality gap between molecule and text. This tokenizer transforms molecular structures into sequences of tokens exhibiting causal dependency, thereby encapsulating both high-level molecular features and textual information. Equipped with this tokenizer, UniMoT unifies molecule and text modalities under a shared token representation and an autoregressive training paradigm. This enables the model to process molecular structures as a distinct linguistic system and generate them in textual form. Through a four-stage training scheme, UniMoT functions as a multi-modal generalist capable of performing both molecule-to-text and text-to-molecule tasks. Extensive experiments demonstrate that UniMoT achieves state-of-the-art performance across a wide range of molecule comprehension and generation tasks.

1 Introduction

The incredible capabilities of Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023) have led to their widespread use as versatile tools for completing diverse real-world tasks. This success has sparked interest in Multi-modal LLMs (Zhan et al., 2024; Wu et al., 2023), which aim to enhance LLMs by enabling them to process multi-modal inputs and outputs. Prior research efforts (Liang et al., 2023; Tang et al., 2023; Fang et al., 2023; Cao et al., 2023; Liu et al., 2023b; Luo et al., 2023; Li et al., 2024) have focused on adapting LLMs to molecular tasks, resulting in the development of molecular LLMs. These molecular LLMs can analyze molecule structures (Luo et al., 2023; Liu et al., 2023b; Cao et al., 2023), address drug-related inquiries (Liang et al., 2023; Tang et al., 2023), assist in synthesis and retrosynthesis planning (Fang et al., 2023), support drug design (Fang et al., 2023), and more.

Prevalent molecular LLMs often use adapter-based architectures, such as linear projection Liang et al. (2023); Cao et al. (2023) or Q-Former Liu et al. (2023b); Li et al. (2024), to map molecule features into the LLM's semantic space (Figure 1a, Figure 1b). While effective in molecular comprehension and molecule-to-text generation, these models struggle with text-to-molecule generation. This is due to the reliance on adapters that require LLMs to directly generate SMILES strings Weininger (1988), a text-based representation of molecular structures. These architectures depend on strong alignment

† Corresponding author.

between SMILES and text, but as shown in Figure 1a and Figure 1b, molecule and text modalities are not treated equally, with insufficient supervision for the molecular side, making alignment difficult.

Discretizing continuous molecule features into discrete molecule tokens offers a promising solution for conducting both molecule-to-text and text-to-molecule generation tasks. By treating tokens from different modalities equally, we can predict the next molecule or text token in an autoregressive manner. However, directly discretizing molecule features poses several challenges: (i) This approach results in long sequences, with lengths equivalent to the number of atoms in a batch; (ii) Molecule tokens derived from molecule features lack left-to-right causal dependency, which conflicts with the unidirectional attention mechanism in LLMs; (iii) Molecule features lack textual information, hindering effective molecule-text interactions and alignment.

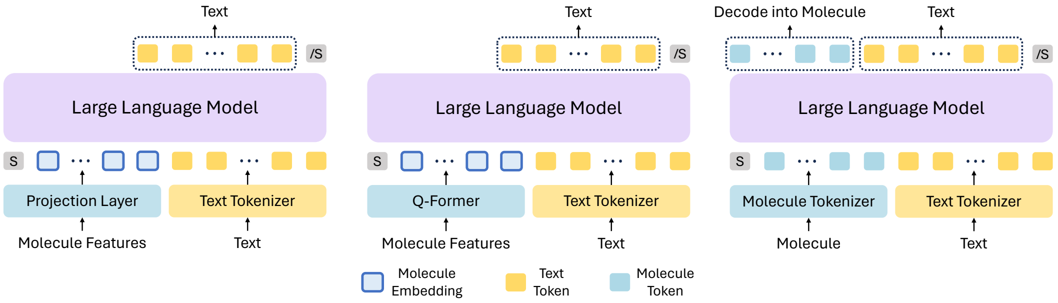
To this end, we present **UniMoT**, a **Uni**fied **Mo**lecule-**T**ext LLM that adopts a tokenizer-based architecture, integrating molecule comprehension and generation, as depicted in Figure 1c. A pivotal aspect of UniMoT's architecture is the molecule tokenizer for transforming molecules into molecule tokens. We introduce a Vector Quantization-driven (Van Den Oord et al., 2017) tokenizer, which incorporates a Q-Former (Li et al., 2023) to bridge the modality gap between molecules and text. Specifically, we incorporate causal masks for the queries, enabling the Q-Former to generate a causal sequence of queries compatible with the unidirectional attention in LLMs. The sequence of queries is subsequently quantized into a sequence of molecule tokens using a learnable codebook. The molecule tokens encapsulate high-level molecular and textual information, which are then aligned with the latent space of a pretrained generative model via an MLP adapter. Pretrained LLMs integrate the molecule tokenizer by treating molecule tokens as new words, constructing a molecule vocabulary via the learned codebook. We unify molecule and text representations with a shared next-token-prediction training paradigm, enabling effective molecule-text interactions. For molecule generation, UniMoT autoregressively produces molecule tokens, which are then decoded into molecules by the generative model.

Our contributions can be summarized as follows:

- We introduce a molecule tokenizer specifically designed for LLMs, enabling the tokenization of molecules into short sequences of molecule tokens with causal dependency. These tokens encapsulate high-level molecular and textual information and can be decoded into desired molecules during inference.
- We present UniMoT, a unified molecule-text LLM that adopts a tokenizer-based architecture instead of traditional adapter-based architectures. UniMoT unifies the modalities of molecule and text under a shared token representation and an autoregressive training paradigm. Following a four-stage training scheme, UniMoT effectively achieves molecule-text alignment.
- UniMoT exhibits remarkable capabilities in multi-modal comprehension and generation. Extensive experiments show that UniMoT achieves state-of-the-art performance across a wide range of molecule comprehension and generation tasks, while also offering a new perspective on molecule generation.

2 Related Works

Molecular Large Language Models. The recent emergence of Vision Large Language Models (VLLMs) Li et al. (2022; 2023); Liu et al. (2024a) has catalyzed advancements in molecular LLMs, which encompass both single modality and multi-modality approaches. In the single modality domain, researchers are exploring diverse molecule representations, such as 1D sequences like SMILES strings Irwin et al. (2022), 2D molecule graphs You et al. (2020), 3D geometric conformations You et al. (2020), and textual information from the literature Taylor et al. (2022); Beltagy et al. (2019). In the multiple modalities domain, various innovative approaches are being employed. MolT5 Edwards et al. (2022), a T5-based Raffel et al. (2020) model, is designed for SMILES-to-text and text-to-SMILES translations. Other works, such as MoMu Su et al. (2022), MoleculeSTM Liu et al. (2023a), and GIT-Mol Liu et al. (2024c), leverage cross-modal contrastive learning to align the representation spaces of molecules and text. Additionally, some studies Cao et al. (2023); Liang et al. (2023); Liu et al. (2023b); Li et al. (2024) use multi-modal learning architectures to develop molecular LLMs, which often adopt adapter-based architectures. However, these methods do not treat molecule and text



(a) Projection-Based Architecture. (b) Q-Former-Based Architecture. (c) Tokenizer-Based Architecture.

Figure 1: Comparisons among different molecular LLMs. 1a and 1b are adapter-based architectures that do not treat molecule and text modalities equally and lack a supervision signal for the molecule modality. 1c is our proposed tokenizer-based architecture, where molecules are presented in the same discrete token representation as that of text.

modalities equally and lack a supervision signal for the molecule modality, limiting model capacity and effectiveness.

Vector Quantization. Vector Quantization (VQ) (Gray, 1984) is a widely used technique in generative models. VQ-VAE (Van Den Oord et al., 2017) converts an image into a set of discrete codes within a learnable discrete latent space by learning to reconstruct the original image. VQ-GAN (Yu et al., 2021) enhances the generation quality by leveraging adversarial and perceptual objectives. In the context of molecules, VQ has been effectively applied to quantize molecule features. For example, DGAE (Boget et al., 2023) introduces a VQ model specifically for molecules, where molecules are encoded into discrete latent codes. Mole-BERT (Xia et al., 2022) uses VQ to rethink the pre-training of GNNs for molecular tasks. IMoLD (Zhuang et al., 2024) proposes using VQ to enhance invariant molecule representations, and VQSynergy (Wu et al., 2024) demonstrates the use of VQ for drug discovery.

3 Method

Our goal is to leverage LLMs' reasoning and generation abilities to enhance molecule and text comprehension. We achieve this by unifying their token representation and adopting the next-token-prediction training paradigm. As shown in Figure 2, we introduce a molecule tokenizer (Section 3.1) that transforms molecules into tokens by reconstructing input molecules. The resulting molecule sequence is concatenated with text to form a multi-modal sequence, enabling molecule-to-text and text-to-molecule autoregressive pretraining (Section 3.2), as illustrated in Figure 3. The LLM vocabulary is expanded with molecule tokens mapped from the learned codebook. We propose a four-stage training scheme (Section 3.3): Causal Q-Former pretraining, molecule tokenizer pretraining, unified molecule-text pretraining, and task-specific instruction tuning. Following this scheme, UniMoT effectively handles molecule comprehension and generation tasks.

3.1 Molecule Tokenizer for LLMs

Molecule Encoder. We represent the structural information of a molecule as a graph, denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of atoms and $|\mathcal{V}| = N$ is the number of atoms. The task of the molecule encoder is to extract molecule features that are context-aware and encompass diverse local neighborhood structural information. By employing a molecule encoder, we obtain molecule features $\mathbf{X} \in \mathbb{R}^{N \times F}$, where F denotes the dimensionality of the feature vector for each atom.

Causal Q-Former. We employ a Q-Former model introduced by BLIP-2 (Li et al., 2023) to generate queries $\mathbf{Z} = \{\boldsymbol{z}_i\}_{i=1}^{M} \in \mathbb{R}^{M \times d}$ containing high-level molecular and textual information, where M represents the number of queries and d denotes the dimension of queries. The Q-Former operates as a query-based transformer that utilizes learnable queries $\{\boldsymbol{z}_i\}_{i=1}^{M}$ to interact with molecule

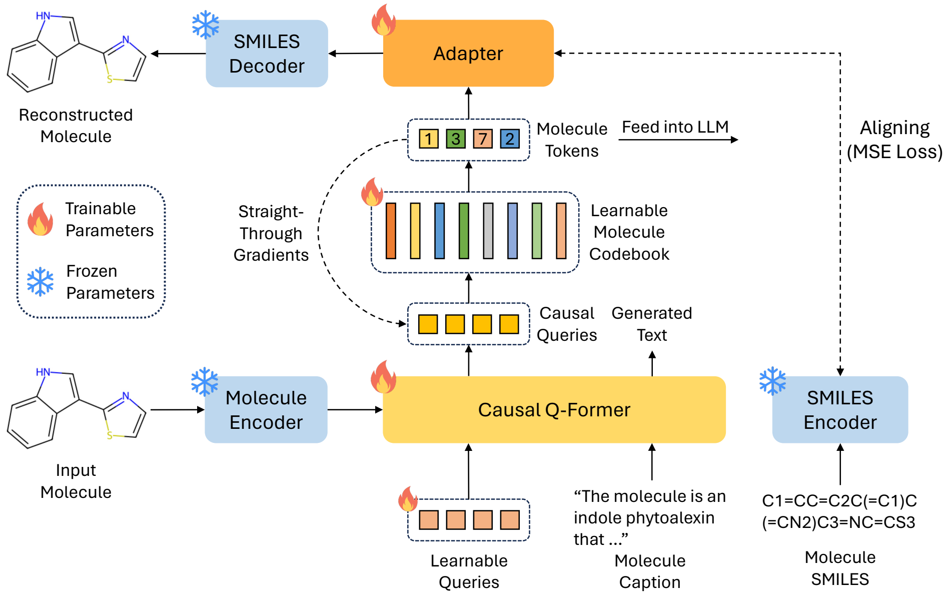


Figure 2: Illustration of our proposed molecule tokenizer. The tokenizer generates discrete molecule tokens, which can be fed into LLMs for downstream tasks. The generated molecule tokens can be decoded into molecules using the adapter and the SMILES decoder during inference.

features $\mathbf{X}$ extracted by the molecule encoder. Specifically, we incorporate causal masks into the queries, ensuring that they only interact with preceding queries. This ensures the sequence of queries maintains a causal dependency, aligning with the unidirectional requirements of LLMs operating on text sequence. Details regarding the Causal Q-Former can be found in Appendix A.

Vector Quantization. The Causal Q-Former converts molecules and text into a causal sequence of queries. Subsequently, the causal sequence of queries $\{\boldsymbol{z}_i\}_{i=1}^M$ is quantized into a causal sequence of molecule tokens $\{s_i\}_{i=1}^M$ by identifying the closest neighbor in a learnable codebook $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^K$, where K represents the size of the codebook. The codebook is randomly initialized and optimized during pretraining. Specifically, token s_i is determined as follows:

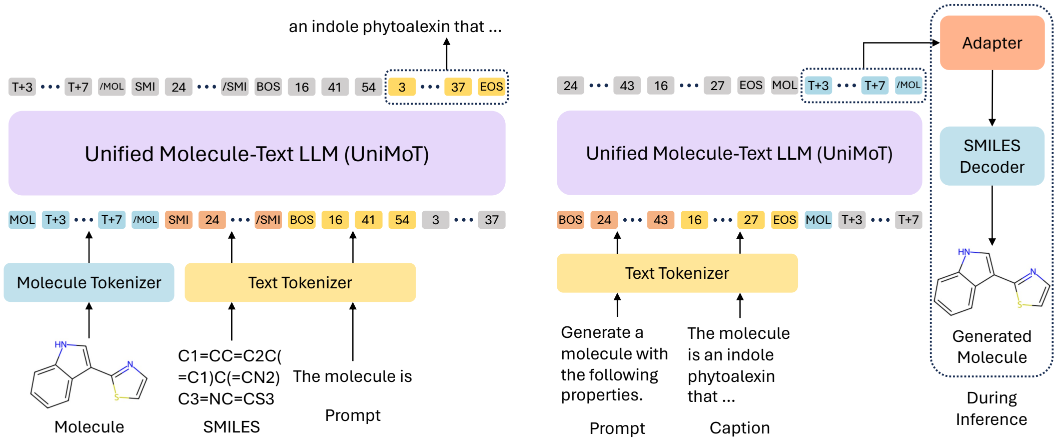
$$s_i = \text{argmin}_{j \in \{1, \cdots, K\}} \left\| \boldsymbol{z}_i - \boldsymbol{c}_j \right\|_2 , \quad \text{for} \quad i = 1, 2, \cdots, M. \tag{1}$$

Intuitively, the query $\boldsymbol{z}_i$ is quantized to the closest neighbor $\boldsymbol{c}_{s_i}$ in the codebook. As the vector quantization process is non-differentiable, we adopt the straight-through estimator (Bengio et al., 2013) to train the Causal Q-Former by copying the gradient from the molecule tokens to the queries, as shown in Figure 2. The resulting embeddings of molecule tokens $\{s_i\}_{i=1}^M$, denoted as $\mathbf{C} = \{\boldsymbol{c}_{s_i}\}_{i=1}^M$, are subsequently utilized for reconstructing molecules.

Molecule Reconstruction. An MLP adapter ψ needs to be trained to align the discrete latent space of molecule tokens with the continuous latent space of a molecular generative model for molecule reconstruction. This can be represented as $\mathbf{X}_R = \psi(\mathbf{C})$, where $\mathbf{X}_R$ denotes the embeddings for reconstruction. To achieve alignment, we minimize the Mean Squared Error (MSE) loss between $\mathbf{X}_R$ and the SMILES (Weininger, 1988) embeddings $\mathbf{X}_S$ produced by the pretrained SMILES encoder. Subsequently, we can reconstruct the molecule from $\mathbf{X}_R$ using the pretrained SMILES decoder. The training loss of the tokenizer is expressed as follows:

$$\mathcal{L}_{\text{Tokenizer}} = \|\mathbf{X}_R - \mathbf{X}_S\|_2^2 + \frac{1}{M}\sum_{i=1}^{M} \|\text{sg}\left[\boldsymbol{z}_i\right] - \boldsymbol{c}_{s_i}\|_2^2 + \frac{\beta}{M}\sum_{i=1}^{M} \|\text{sg}\left[\boldsymbol{c}_{s_i}\right] - \boldsymbol{z}_i\|_2^2 . \tag{2}$$

Here, the first term represents the alignment loss, the second term is a codebook loss aimed at updating the codebook embeddings, and the third term is a commitment loss that encourages the query to stay close to the chosen codebook embedding. $\text{sg}[\cdot]$ denotes the stop-gradient operator, and the hyperparameter β is set to 0.25.


an indole phytoalexin that ...
T+3 ··· T+7 /MOL SMI 24 ··· /SMI BOS 16 41 54 3 ··· 37 EOS
Unified Molecule-Text LLM (UniMoT)
MOL T+3 ··· T+7 /MOL SMI 24 ··· /SMI BOS 16 41 54 3 ··· 37
Molecule Tokenizer
Text Tokenizer
C1=CC=C2C(
=C1)C(=CN2)
C3=NC=CS3
The molecule is
Molecule
SMILES
Prompt
24 ··· 43 16 ··· 27 EOS MOL T+3 ··· T+7 /MOL
Unified Molecule-Text LLM (UniMoT)
BOS 24 ··· 43 16 ··· 27 EOS MOL T+3 ··· T+7
Text Tokenizer
Generate a molecule with the following properties.
The molecule is an indole phytoalexin that ...
Prompt
Caption
Adapter
SMILES Decoder
Generated Molecule
During Inference

(a) Molecule-to-Text Autoregression.

(b) Text-to-Molecule Autoregression.

Figure 3: Illustration of the multi-modal autoregressive pretraining on molecule-text datasets. UniMoT excels in multi-modal comprehension and generation tasks, enabled by the unified LM objective. T represents the size of the text vocabulary.

3.2 Unified Molecule-Text Language Model

Expanding Vocabulary. Employing the molecule tokenizer, a molecule can be tokenized into a molecule sequence $\{s_i\}_{i=1}^{M}$ with causal dependency. The molecule sequence can be concatenated with the text sequence to form a multi-modal sequence $\{u_i\}_{i=1}^{L}$, where L is the length of the multi-modal sequence. To facilitate the representation of the multi-modal sequence, we construct the molecule vocabulary $\mathcal{V}^m = \{\boldsymbol{v}_i^m\}_{i=1}^{K}$, which maintains the order of the molecule codebook $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^{K}$. Additionally, $\mathcal{V}^m$ includes several special tokens such as boundary indicators, e.g., [MOL] and [/MOL], to mark the beginning and end of the molecule sequence. Next, we merge the original text vocabulary $\mathcal{V}^t = \{\boldsymbol{v}_i^t\}_{i=1}^{T}$ with the molecule vocabulary $\mathcal{V}^m$. The unified molecule-text vocabulary $\mathcal{V} = \{\mathcal{V}^m, \mathcal{V}^t\}$ facilitates joint learning from molecules and text under a unified next-token-prediction objective. As the vocabulary is expanded, the corresponding embeddings and prediction layers also need to be extended, with the newly introduced parameters initialized randomly.

Unified Molecule-text Modeling. The multi-modal sequence $\{u_i\}_{i=1}^{L}$ is fed into the pretrained LLM for performing multi-modal autoregression. UniMoT adopts the general Language Modeling (LM) objective to directly maximize the log-likelihood of the data distribution:

$$\mathcal{L}_{\text{LM}} = -\sum_{u \in \mathcal{D}} \sum_{i \in \mathcal{I}} \log p\left(u_i \mid u_1, \cdots, u_{i-1}; \theta\right), \tag{3}$$

where $\mathcal{D}$ represents the dataset, $\mathcal{I}$ represents the set of indices of the generation target, and θ denotes the parameters of the LLM. The unification of representation and training paradigm for molecules and text enhances the abilities of LLMs to understand molecule-text interactions and alignment. UniMoT can interpret molecules similar to understanding a foreign language, and generate them as if they were text. We conduct autoregressive pretraining on molecule-to-text and text-to-molecule tasks to enhance the molecule comprehension and generation capabilities.

Molecule-to-Text Autoregression. While structural information is embedded in molecule features and captured by the molecule tokens through the tokenizer, we also aim to incorporate sequential information of molecules for better comprehension. Therefore, we concatenate the molecule sequence $\{s_i\}_{i=1}^{M}$ with the SMILES (Weininger, 1988) sequence and a prompt to form the multi-modal input sequence $\{u_i\}_{i=1}^{L}$, as illustrated in Figure 3a. The corresponding molecule caption is used as the generation target.

Text-to-Molecule Autoregression. For molecule generation, a prompt and the molecule caption are concatenated, with a [MOL] token appended to signify the beginning of the molecule sequence, as illustrated in Figure 3b. The molecule sequence $\{s_i\}_{i=1}^{M}$ produced by the tokenizer is used as the generation target. During inference, given a prompt and the molecule caption, the output molecule sequence can be decoded into the desired molecule by the pretrained adapter and SMILES decoder.

3.3 Training Strategy

The training strategy for UniMoT is structured across four stages. Stage-1 focuses on Causal Q-Former pretraining with tailored objectives. In Stage-2, the molecule tokenizer is optimized using the frozen encoders and decoder. Stage-3 integrates the tokenizer with a language model for multi-modal comprehension and generation. Finally, Stage-4 fine-tunes UniMoT for specific tasks, aligning it with human instructions and optimizing performance for various molecular applications. More details regarding the training process can be found in Appendix C.

Stage-1: Causal Q-Former Pretraining. We connect the molecule encoder and Causal Q-Former, leveraging the pretrained MoleculeSTM molecule encoder (Liu et al., 2023a). The molecule encoder remains frozen while only the Causal Q-Former is updated. Both queries and text inputs are used, while only queries serve as input in subsequent stages. In our experiments, we utilize 16 queries. We employ three tailored objectives for the pretraining of the Causal Q-Former: Molecule-Text Contrastive Learning (MTC), Molecule-Text Matching (MTM), and Molecule-grounded Text Generation (MTG). The details of these objectives can be found in Appendix A.

Stage-2: Molecule Tokenizer Pretraining. We connect the Causal Q-Former with subsequent blocks and use the objective defined in Equation (2). We employ the pretrained ChemFormer (Irwin et al., 2022) as the generative model. Specifically, we leverage the SMILES encoder and the SMILES decoder provided by ChemFormer. The molecule codebook size is set to $K = 2048$. As shown in Figure 2, we keep the molecule encoder, the SMILES encoder, and the SMILES decoder frozen, while updating the Causal Q-Former, the learnable codebook, and the adapter.

Stage-3: Unified Molecule-Text Pretraining. We integrate the molecule tokenizer with the LLM using the unified vocabulary of molecule tokens and text tokens. We employ the LM objective defined in Equation (3) to pretrain the LLM. Pretraining involves molecule-to-text autoregression and text-to-molecule autoregression, aimed at enhancing UniMoT's multi-modal comprehension and generation capabilities. To enhance efficiency, we train the LLM using low-rank adaptation (LoRA) (Hu et al., 2021).

Stage-4: Task-Specific Instruction Tuning. UniMoT is fine-tuned on seven comprehension and generation tasks: molecular property prediction, molecule captioning, molecule-text retrieval, caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. We also utilize LoRA to improve efficiency. This stage ensures UniMoT can accurately interpret and respond to human instructions, making it versatile and effective for diverse molecular tasks.

4 Experiments

4.1 Molecule Comprehension Tasks

Molecular Property Prediction Task. Molecular property prediction aims to forecast a molecule's physical and chemical properties. For classification, we use eight binary classification datasets from MoleculeNet (Wu et al., 2018), where models predict "yes" or "no." We compare UniMoT with KV-PLM (Zeng et al., 2022), AttrMask (Hu et al., 2019), InfoGraph (Sun et al., 2019), MolCLR (Wang et al., 2021), GraphMVP (Liu et al., 2019), MoleculeSTM (Liu et al., 2023a), and InstructMol (Cao et al., 2023). The ROC-AUC (%) results on MoleculeNet are shown in Table 1, while regression task results are in Appendix D. Compared to graph learning methods and molecular LLMs like InstructMol (Cao et al., 2023), UniMoT consistently improves performance across all eight datasets, highlighting its robust molecular comprehension abilities.

Molecule Captioning Task. Molecule captioning involves generating detailed molecular descriptions. We compare UniMoT with MolT5 (Edwards et al., 2022), MoMu (Su et al., 2022), InstructMol (Cao et al., 2023), MolCA (Liu et al., 2023b), and 3D-MoLM (Li et al., 2024), using

Table 1: ROC-AUC (%) of molecular property prediction task (classification) on the MoleculeNet (Wu et al., 2018) datasets. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BBBP↑	Tox21↑	ToxCast↑	Sider↑	ClinTox↑	MUV↑	HIV↑	BACE↑
KV-PLM	70.50	72.12	55.03	59.83	89.17	54.63	65.40	78.50
AttrMask	67.79	75.00	63.57	58.05	75.44	73.76	75.44	80.28
InfoGraph	64.84	76.24	62.68	59.15	76.51	72.97	70.20	77.64
MolCLR	67.79	75.55	64.58	58.66	84.22	72.76	75.88	71.14
GraphMVP	68.11	**77.06**	65.11	60.64	84.46	74.38	77.74	80.48
MoleculeSTM	69.98	76.91	65.05	**60.96**	92.53	73.40	76.93	80.77
InstructMol (Vicuna-7B)	70.00	74.67	64.29	57.80	91.48	74.62	68.90	82.30
UniMoT (Llama-2-7B)	**71.37**	76.43	**65.78**	59.79	**92.89**	**75.97**	**78.49**	**83.69**

Table 2: Performance (%) of molecule captioning task on the PubChem (Kim et al., 2023) dataset. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-Small (T5-Small)	22.5	15.2	30.4	13.5	20.3	24.0
MolT5-Base (T5-Base)	24.5	16.6	32.2	14.0	21.4	26.1
MolT5-Large (T5-Large)	25.9	17.3	34.1	16.4	23.4	28.0
MoMu-Small (T5-Small)	22.9	16.0	31.0	13.7	20.8	24.4
MoMu-Base (T5-Base)	24.7	16.8	32.5	14.6	22.1	27.2
MoMu-Large (T5-Large)	26.3	18.0	34.8	16.9	24.8	28.7
InstructMol (Vicuna-7B)	18.9	11.7	27.3	11.8	17.8	21.3
MolCA (OPT-125M)	25.9	17.5	34.4	16.6	23.9	28.5
MolCA (OPT-1.3B)	28.6	21.3	36.2	21.4	29.7	32.6
3D-MoLM (Llama-2-7B)	30.3	22.5	36.8	22.3	31.2	33.1
UniMoT (Llama-2-7B)	**31.3**	**23.8**	**37.5**	**23.7**	**33.6**	**34.8**

BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) as evaluation metrics. UniMoT is assessed on PubChem (Kim et al., 2023) and ChEBI-20 (Edwards et al., 2022). Results for PubChem appear in Table 2, while ChEBI-20 results and examples are in Appendix D. ChEBI-20 replaces molecular names with “the molecule” to emphasize properties, but predicting names reflects structural understanding, so main experiments use PubChem.

As shown in Table 2, UniMoT significantly outperforms baselines on PubChem (Kim et al., 2023), making molecule captioning a strong test of comprehension. Our tokenizer-based architecture surpasses projection-based models (e.g., InstructMol (Cao et al., 2023)), Q-Former-based models (e.g., MolCA (Liu et al., 2023b), 3D-MoLM (Li et al., 2024)), and contrastive learning models (e.g., MoMu (Su et al., 2022)). This demonstrates that UniMoT achieves superior molecule-text alignment through autoregressive molecule-to-text and text-to-molecule pretraining.

4.2 Molecule Generation Tasks

We employ molecule generation tasks Fang et al. (2023), including caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. Caption-guided generation creates molecules from text, reagent prediction identifies suitable reagents, forward reaction prediction forecasts products, and retrosynthesis deconstructs target molecules. We compare UniMoT with Llama Touvron et al. (2023), Vicuna Chiang et al. (2023), Mol-Instructions Fang et al. (2023), and InstructMol Cao et al. (2023), using metrics such as Exact Match, BLEU Papineni et al. (2002), Levenshtein Distance Levenshtein et al. (1966), RDKit Landrum et al. (2006), MACCS Durant et al. (2002), and Morgan Fingerprint Similarities Morgan (1965) to evaluate structural similarity, along with Validity Kusner et al. (2017) for chemical correctness.

The Mol-Instructions Fang et al. (2023) benchmark is used to assess UniMoT’s performance, with results for caption-guided molecule generation in Table 3, and other tasks in Appendix D. Caption-guided generation is based on the PubChem Kim et al. (2023) dataset, while other tasks use USPTO Fang et al. (2023). Unlike baselines that generate SMILES strings before conversion, UniMoT directly produces molecule tokens, retrieves embeddings from the learned codebook, and decodes them into molecules. As shown in Table 3, UniMoT generates valid molecules with higher similarity to targets, as it treats molecule generation as a text-based process, unlike adapter-based architectures. This demonstrates strong generation capabilities and introduces a novel perspective on molecule generation tasks.

Table 3: Performance of molecule generation tasks on the Mol-Instructions Fang et al. (2023) benchmark, including caption-guided molecule generation and reagent prediction. **Bold** indicates the best performance, and underline indicates the second best performance.

Model	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Caption-guided Molecule Generation							
Llama	0.000	0.003	59.864	0.005	0.000	0.000	0.003
Vicuna	0.000	0.006	60.356	0.006	0.001	0.000	0.001
Mol-Instructions	0.002	0.345	41.367	0.231	0.412	0.147	1.000
MolT5	0.112	0.546	38.276	0.400	0.538	0.295	0.773
UniMoT	**0.237**	**0.698**	**27.782**	**0.543**	**0.651**	**0.411**	1.000

4.3 Ablation Studies

Cross-Modal Projector. We conducted an ablation study on the cross-modal projector, with the results on the molecule captioning task shown in Table 4. The linear projection demonstrated the worst performance, indicating that the molecule features lack textual information, thus hindering effective molecule-text interactions and alignment. Additionally, we compared the performance of a Q-Former with bidirectional self-attention to a Causal Q-Former with causal self-attention in the second and third rows. The results show that queries with causal dependency outperform those with bidirectional dependency. This demonstrates that input with left-to-right causal dependency aligns with the unidirectional attention mechanism in LLMs, leading to improved performance.

Discrete vs. Continuous Representation. We compared the performance of continuous causal embeddings and discrete tokens, quantized from causal embeddings, as inputs to LLMs in the third and fourth rows of Table 4. Continuous embeddings demonstrate better performance than discrete tokens in understanding molecules. This result is reasonable since the quantization process causes information loss in discrete tokens. However, we still use discrete token representation to facilitate the autoregressive training paradigm of LLMs, which supports the unification of comprehension and generation tasks. To achieve this unification, we unavoidably sacrifice some performance in comprehension tasks.

Additional ablation studies are presented in Appendix E.

Table 4: Ablation study on the projector and representation form.

Projector	Input to LLM	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Projection Layer	Molecule Emb.	19.3	12.1	27.9	12.3	18.1	21.5
Q-Former	Query Emb.	28.6	21.3	36.2	21.4	29.7	32.6
Causal Q-Former	Causal Emb.	32.8	25.2	39.2	24.8	35.3	36.5
Causal Q-Former	Causal Tokens	31.3	23.8	37.5	23.7	33.6	34.8

5 Conclusion

This work introduces UniMoT, a framework that unifies the modalities of molecules and text. By adopting a tokenizer-based architecture, UniMoT addresses previous limitations where the molecule and text modalities are not treated equally. The molecule tokenizer converts molecules into sequences of discrete tokens, embedding high-level molecular and textual information. The LLM vocabulary is expanded with molecule tokens mapped from a learned codebook. Through a four-stage training scheme, UniMoT has become a versatile multi-modal LLM, capable of handling both molecule-to-text and text-to-molecule tasks. Extensive empirical evaluations show that UniMoT achieves state-of-the-art performance across diverse molecule comprehension and generation tasks.

REFERENCES

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Yoann Boget, Magda Gregorova, and Alexandros Kalousis. Vector-quantized graph auto-encoder. *arXiv preprint arXiv:2306.07735*, 2023.

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

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*, pp. 6140–6157. PMLR, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Greg Landrum et al. Rdkit: Open-source cheminformatics, 2006.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*, 2024.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. Drugchat: towards enabling chatgpt-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907*, 2023.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Jinzhe Liu, Xiangsheng Huang, Zhuo Chen, and Yin Fang. Drak: Unlocking molecular insights with domain-specific retrieval-augmented knowledge in llms. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 255–267. Springer, 2024b.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171: 108073, 2024c.

Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

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

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*, 2023b.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

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

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

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

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models (2023). *arXiv preprint arXiv:2302.13971*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Y Wang, J Wang, Z Cao, and AB Farimani. Molclr: Molecular contrastive learning of representations via graph neural networks. arxiv 2021. *arXiv preprint arXiv:2102.10056*, 2021.

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, et al. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*, 2024.

Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126*, 2022.

David Weininger. 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, 1988.

Jiawei Wu, Mingyuan Yan, and Dianbo Liu. Vqsynery: Robust drug synergy prediction with vector quantization mechanism. *arXiv preprint arXiv:2403.03089*, 2024.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

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

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2022.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Xiang Zhuang, Qiang Zhang, Keyan Ding, Yatao Bian, Xiao Wang, Jingsong Lv, Hongyang Chen, and Huajun Chen. Learning invariant molecular representation in latent discrete space. *Advances in Neural Information Processing Systems*, 36, 2024.

A Details of Causal Q-Former

The Q-Former operates as a query-based transformer that utilizes learnable query vectors to interact with molecule features extracted by a frozen encoder. These queries are essential for extracting relevant information from the molecule features. The Q-Former comprises both a molecule transformer and a text transformer, sharing self-attention layers. The molecule transformer incorporates cross-attention layers between self-attention and feed-forward layers, while the text transformer architecture is based on BERT (Devlin et al., 2018). Q-Former employs a cross-attention mechanism where the query vectors selectively attend to different aspects of the molecule features, allowing the model to capture critical details necessary for understanding and generating textual descriptions of molecular properties.

Specifically, we incorporate causal masks into the queries, ensuring that they only interact with preceding queries. This ensures the sequence of queries maintains a causal dependency, aligning with the requirements of LLMs operating on text sequence. The Causal Q-Former is illustrated in Figure 4. We employ the Causal Q-Former to generate causal queries $\mathbf{Z} = \{\boldsymbol{z}_i\}_{i=1}^{M} \in \mathbb{R}^{M\times d}$ containing high-level molecular and textual information, where M represents the number of queries and d denotes the dimension of queries. Next, we introduce three tailored objectives MTC, MTM, and MTG for the pretraining of the Causal Q-Former.

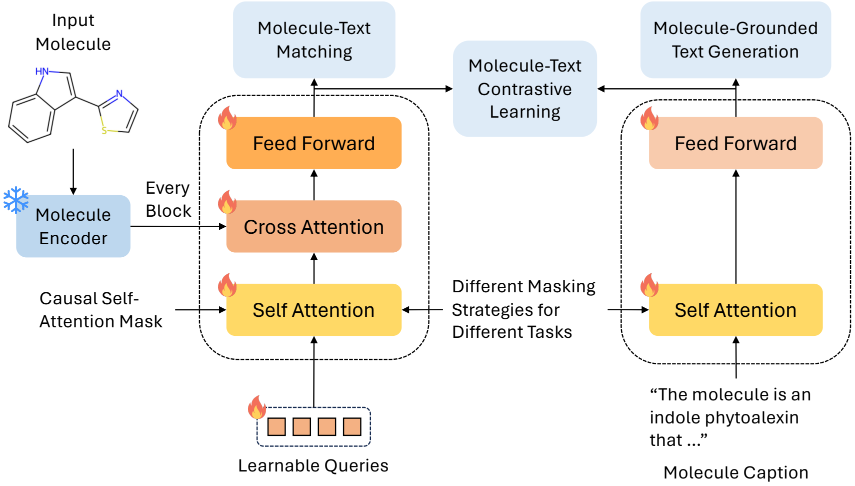

Input Molecule
HN
N
S
Molecule Encoder
Every Block
Causal Self-Attention Mask
Molecule-Text Matching
Feed Forward
Cross Attention
Self Attention
Learnable Queries
Molecule-Text Contrastive Learning
Different Masking Strategies for Different Tasks
Molecule-Grounded Text Generation
Feed Forward
Self Attention
"The molecule is an indole phytoalexin that ..."
Molecule Caption

Figure 4: Illustration of our proposed Causal Q-Former. The Causal Q-Former provides causal queries for subsequent blocks.

Molecule-Text Contrastive Learning (MTC) aims to align molecule and text features by maximizing their mutual information. This is achieved by maximizing the molecule-text similarity of positive pairs against that of negative pairs. We utilize the last query $\boldsymbol{z}_M$ of the query sequence $\{\boldsymbol{z}_i\}_{i=1}^{M}$ as the query representation, since the output query sequence is causal and the last query contains global information from the queries. For text representation, we use the output embedding of the [CLS] token, denoted as $\boldsymbol{y}$. The contrastive learning loss is expressed as follows:

$$\mathcal{L}_{\text{MTC}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp((\boldsymbol{z}_M^i)^T\boldsymbol{y}^i/\tau)}{\sum_{j=1}^{B}\exp((\boldsymbol{z}_M^i)^T\boldsymbol{y}^j/\tau)} - \frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp((\boldsymbol{y}^i)^T\boldsymbol{z}_M^i/\tau)}{\sum_{j=1}^{B}\exp((\boldsymbol{y}^i)^T\boldsymbol{z}_M^j/\tau)}, \qquad (4)$$

where B denotes the batch size, and τ represents the temperature parameter. Here, $\boldsymbol{z}_M^i$ and $\boldsymbol{y}^i$ refer to the i-th query representation and text representation in a batch, respectively.

Molecule-Text Matching (MTM) focuses on learning fine-grained alignment between molecule and text features. As queries $\{\boldsymbol{z}_i\}_{i=1}^{M}$ capture both molecular and textual information through cross-attention and self-attention layers respectively, we utilize the last query $\boldsymbol{z}_M$ as input to a binary classifier. This classifier predicts whether a given molecule-text pair is matched or unmatched. The corresponding loss function is formulated as follows:

$$\mathcal{L}_{\text{MTM}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\phi(\boldsymbol{z}_M \mid \mathbf{X}^i, \boldsymbol{t}^i))}{\sum_{j=1}^{B} \exp(\phi(\boldsymbol{z}_M \mid \mathbf{X}^i, \boldsymbol{t}^j)) + \sum_{j=1}^{B} \exp(\phi(\boldsymbol{z}_M \mid \mathbf{X}^j, \boldsymbol{t}^i))}, \tag{5}$$

where ϕ represents a binary classifier, and $\mathbf{X}^i$ and $\boldsymbol{t}^i$ denote the i-th input molecule features and input text in a batch, respectively.

Molecule-grounded Text Generation (MTG) focuses on generating textual descriptions given a molecule input. In this task, causal masks for queries are not applied since only textual output is required. However, causal masks are applied for text, allowing each text token to attend to its preceding text tokens and all queries, but not subsequent tokens. The Language Modeling (LM) loss function is applied to model the generation of text $\boldsymbol{t}^i$ conditioned on the molecule input $\mathbf{X}^i$, formulated as:

$$\mathcal{L}_{\text{MTG}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{L} \log p\left(t_j^i \mid t_1^i, \cdots, t_{j-1}^i, \mathbf{X}^i\right), \tag{6}$$

where t_j^i represents the j-th token in the text sequence $\boldsymbol{t}^i$. Here, $\mathbf{X}^i$ and $\boldsymbol{t}^i$ denote the i-th input molecule features and generated text in a batch, respectively.

The total loss for training the Causal Q-Former encompasses the three aforementioned objectives:

$$\mathcal{L}_{\text{Q-Former}} = \mathcal{L}_{\text{MTC}} + \mathcal{L}_{\text{MTM}} + \mathcal{L}_{\text{MTG}}. \tag{7}$$

B Details of Datasets

This section provides detailed information about the datasets used in evaluating the performance of UniMoT across various tasks. The datasets are utilized for molecular property prediction, molecule captioning, molecule-text retrieval, caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis task. Each dataset serves a unique purpose in assessing different capabilities of the model. We provide a comprehensive overview of datasets, including their types, associated tasks, descriptions, URLs for access, and licensing information.

We present the details of the Molecular Property Prediction Datasets below:

- **BBBP** (Wu et al., 2018): The Blood-Brain Barrier Penetration dataset predicts the ability of molecules to penetrate the blood-brain barrier.
- **Tox21** (Wu et al., 2018): This dataset is part of the Toxicology in the 21st Century initiative, used for toxicity prediction.
- **ToxCast** (Wu et al., 2018): Another toxicity prediction dataset with a broader range of biological assays.
- **Sider** (Wu et al., 2018): Side Effect Resource database, used for predicting drug side effects.
- **ClinTox** (Wu et al., 2018): Clinical Toxicity dataset for predicting clinical trial toxicity outcomes.
- **MUV** (Wu et al., 2018): Maximum Unbiased Validation dataset for virtual screening.
- **HIV** (Wu et al., 2018): Human Immunodeficiency Virus dataset for predicting anti-HIV activities.
- **BACE** (Wu et al., 2018): Beta-Secretase 1 dataset for predicting inhibitors of the BACE-1 enzyme, relevant for Alzheimer's research.
- **QM9** (Fang et al., 2023): The quantum mechanics properties dataset, where the objective is to predict key quantum mechanics properties of a given molecule, such as HUMO, LUMO, and the HUMO-LUMO gap.

We present the details of the Molecule Captioning Datasets below:

- **PubChem** (Kim et al., 2023): A large dataset of chemical molecules used for generating textual descriptions of molecular structures.

Table 5: Summary of datasets, their types, tasks, descriptions, URLs, and licenses used for evaluating UniMoT.

Dataset	Type	Tasks	Description	URL	License
BBBP	Classification	Molecular Property Prediction	Predicts blood-brain barrier penetration ability.	BBBP URL	CC-BY 4.0
Tox21	Classification	Molecular Property Prediction	Toxicity prediction using the Tox21 initiative data.	Tox21 URL	Public Domain
ToxCast	Classification	Molecular Property Prediction	Broad toxicity prediction with various biological assays.	ToxCast URL	Public Domain
Sider	Classification	Molecular Property Prediction	Predicts drug side effects.	Sider URL	CC-BY 4.0
ClinTox	Classification	Molecular Property Prediction	Clinical trial toxicity prediction.	ClinTox URL	Public Domain
MUV	Classification	Molecular Property Prediction	Virtual screening for unbiased validation.	MUV URL	CC-BY 4.0
HIV	Classification	Molecular Property Prediction	Predicts anti-HIV activity of molecules.	HIV URL	Public Domain
BACE	Classification	Molecular Property Prediction	Predicts inhibitors of the BACE-1 enzyme.	BACE URL	Public Domain
QM9	Regression	Molecular Property Prediction	Predicts various molecular properties such as atomization energy, dipole moment, etc.	QM9 URL	CC-BY 4.0
PubChem	Captioning, Retrieval, Generation	Molecule Captioning, Molecule-Text Retrieval, Caption-guided Molecule Generation	Generates descriptions, retrieves text / molecules based on input molecules / text, and guides molecule generation from captions.	PubChem URL	Public Domain
ChEBI-20	Captioning	Molecule Captioning	Generates detailed descriptions of molecular structures.	ChEBI-20 URL	CC-BY 4.0
PCdes	Retrieval	Molecule-Text Retrieval	Used for evaluating accuracy in molecule-text retrieval tasks.	PCdes URL	CC-BY 4.0
MoMu	Retrieval	Molecule-Text Retrieval	Dataset for molecule-text interaction and retrieval evaluation.	MoMu URL	CC-BY 4.0
USPTO	Generation	Reagent Prediction, Forward Reaction Prediction, Retrosynthesis	Provides data for predicting reagents, forward reaction outcomes, and retrosynthetic pathways.	USPTO URL	CC-BY 4.0

- **ChEBI-20** (Edwards et al., 2022): A subset of the Chemical Entities of Biological Interest database, provides structured and detailed descriptions of molecules.

We present the details of the Molecule-Text Retrieval Datasets below:

- **PubChem** (Kim et al., 2023): Used for both molecule-to-text (M2T) and text-to-molecule (T2M) retrieval tasks.
- **PCdes** (Zeng et al., 2022): Another dataset for evaluating M2T and T2M retrieval accuracy.
- **MoMu** (Su et al., 2022): Dataset specifically designed for molecule-text interactions and retrieval tasks.

We present the details of the Molecule Generation Datasets below:

- **Mol-Instructions** (Fang et al., 2023): This benchmark includes tasks such as caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. It is used to evaluate the model's ability to generate molecular structures based on textual descriptions and other related tasks.
- **PubChem** (Kim et al., 2023): Used for caption-guided molecule generation, generating molecular structures based on textual descriptions.
- **USPTO** (Fang et al., 2023): Used for reagent prediction, forward reaction prediction, and retrosynthesis, providing data for predicting reagents, reaction outcomes, and retrosynthetic pathways.

We summarize the datasets used for evaluating UniMoT in Table 5. It encompasses various types of datasets, including those for classification, regression, captioning, retrieval, and generation tasks. Each dataset is described in terms of its type, tasks it supports, a brief description of its content, its URL for access, and the license under which it is distributed. The licenses vary, with some datasets being in the public domain and others under CC-BY 4.0 license.

C Details of Training

Stage-1: Causal Q-Former Pretraining. During Stage-1, we only connect the molecule encoder and the Causal Q-Former, leaving out other blocks. We leverage the pretrained molecule encoder from MoleculeSTM (Liu et al., 2023a), which has undergone extensive contrastive learning with molecule-text pairs. We utilize the PubChem (Kim et al., 2023) dataset for pretraining, keeping the molecule encoder frozen while updating only the Causal Q-Former. Both queries and text serve as input to the Causal Q-Former, while only queries serve as input in subsequent stages. Inspired by BLIP-2 (Li et al., 2023), we employ three tailored objectives – Molecule-Text Contrastive Learning (MTC), Molecule-Text Matching (MTM), and Molecule-grounded Text Generation (MTG) – for the pretraining of the Causal Q-Former, as detailed in Appendix A.

The dimension of molecule features is set to 300. We use 16 queries, each with a dimension of 768. The size of $\mathbf{Z}$ (16×768) is much smaller than the size of molecule features $\mathbf{X}$ (e.g., 150×300). The Q-former is pretrained for 50 epochs. We adopt the AdamW optimizer with a weight decay of 0.05, and a cosine decay learning rate scheduler, with a minimal learning rate of 1e-5. The batch size is set to 64. The computational overhead for this pretraining is 20 GPU hours on 4 NVIDIA A100 GPUs.

Stage-2: Molecule Tokenizer Pretraining. We connect the Causal Q-Former with the subsequent blocks and train the molecule tokenizer using the objective defined in Equation (2). Following the approach of RetMol (Wang et al., 2022), we utilize SMILES strings (Weininger, 1988) to represent molecules, and employ the pretrained ChemFormer (Irwin et al., 2022) as the generative model. Specifically, we leverage the SMILES encoder and SMILES decoder components provided by ChemFormer. We utilize PubChem (Kim et al., 2023) and ChEBI-20 (Edwards et al., 2022) datasets, keeping the molecule encoder, SMILES encoder, and SMILES decoder frozen, while updating the Causal Q-Former, codebook, and adapter. Once optimized, the molecule tokenizer remains unchanged throughout the subsequent stages.

The molecule codebook size is set to $K = 2048$, and the dimension of codebook embedding is 768. The tokenizer is pretrained for 50 epochs. We adopt the AdamW optimizer with a weight decay of

0.05, and a cosine decay learning rate scheduler, with a minimal learning rate of 1e-5. The batch size is set to 64. The computational overhead for this pretraining is 40 GPU hours on 4 NVIDIA A100 GPUs.

Stage-3: Unified Molecule-Text Pretraining. We connect the molecule tokenizer with the LLM and employ the LM objective defined in Equation (3) to pretrain the LLM. We utilize Llama (Touvron et al., 2023) as the default LLM. To construct the unified molecule-text vocabulary, we merge 2048 molecule codes with the original text vocabulary. Pretraining the LLM involves molecule-to-text autoregression and text-to-molecule autoregression, aimed at enhancing UniMoT's multi-modal comprehension and generation capabilities. We utilize datasets PubChem (Kim et al., 2023) and ChEBI-20 (Edwards et al., 2022) for this purpose. To enhance efficiency, we train the LLM using LoRA (Hu et al., 2021).

The multi-modal LLM is pretrained for 10 epochs. We adopt the AdamW optimizer with a weight decay of 0.05, and a cosine decay learning rate scheduler, with a minimal learning rate of 1e-5. The batch size is set to 32. The computational overhead for this pretraining is 50 GPU hours on 4 NVIDIA A100 GPUs. To reduce CUDA memory usage, we integrate LoRA with the parameters set to $r = 8$, $\alpha = 32$, and dropout = 0.1. This integration is applied to the `k_proj`, `v_proj`, `q_proj`, and `o_proj` modules.

Stage-4: Task-Specific Instruction Tuning. We perform instruction tuning to align UniMoT with human instructions through supervised fine-tuning on seven tasks: molecular property prediction, molecule captioning, molecule-text retrieval, caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. For the molecular property prediction task, we utilize the quantum mechanics properties dataset (Fang et al., 2023) for regression prediction and the MoleculeNet (Wu et al., 2018) datasets for property classification. For the molecule captioning and molecule-text retrieval tasks, we employ datasets PubChem (Kim et al., 2023), PCdes (Zeng et al., 2022), and MoMu (Su et al., 2022). For the molecule generation tasks, we utilize the Mol-Instructions (Fang et al., 2023) benchmark to conduct instruction tuning. We fine-tune UniMoT for 10 epochs on each task using the same optimizer, learning rate scheduler, and LoRA configurations as in Stage-3 pretraining. Instruction samples for comprehension and generation tasks are shown in Table 6.

We have summarized the detailed training hyperparameters of UniMoT in Table 7.

D Details and More Results of Experiments

Molecular Property Prediction Task. Property prediction aims to anticipate a molecule's intrinsic physical and chemical properties based on its structural or sequential characteristics. In the regression task, we conduct experiments on the quantum mechanics properties dataset QM9 (Fang et al., 2023), where the objective is to predict key quantum mechanics properties of a given molecule, such as HUMO, LUMO, and the HUMO-LUMO gap. We compare UniMoT against several baselines, including Alpaca (Taori et al., 2023), Baize (Xu et al., 2023), Llama-2-7B (Touvron et al., 2023), Vicuna-13B (Chiang et al., 2023), Mol-Instructions (Fang et al., 2023), and InstructMol (Cao et al., 2023). Mean Absolute Error (MAE) serves as our evaluation metric. The performance of the regression task on the QM9 dataset is presented in Table 8. Compared to previous single-modal instruction-tuned LLMs and molecular LLMs, UniMoT exhibits further improvement on the regression task, showcasing its fundamental comprehension abilities in molecular contexts.

Molecule Captioning Task. The molecule captioning task involves generating a comprehensive description of a molecule. For this task, we compare UniMoT with several baselines: MolT5 (Edwards et al., 2022), MoMu (Su et al., 2022), InstructMol (Cao et al., 2023), MolCA (Liu et al., 2023b), and 3D-MoLM (Li et al., 2024). We adopt BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) as the evaluation metrics. The performance of UniMoT in the molecule captioning task on the ChEBI-20 (Edwards et al., 2022) dataset is presented in Table 9. Some concrete examples of molecule captioning task are presented in Table 10. From the results, it is evident that UniMoT consistently outperforms the baselines by a significant margin. These results underscore the effectiveness of the molecule tokenizer in providing molecule tokens with high-level molecular and textual information, thus enhancing molecule comprehension.

Table 6: Instruction samples for comprehension and generation tasks: molecular property prediction, molecule captioning, molecule-text retrieval, caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis.

Task	**Instruction**
Molecular Property Prediction (Regression)	Instruction: *Could you give me the LUMO energy value of this molecule?* (Optional: The SMILES sequence is: `SMILES`) Output: *0.0576.*
Molecular Property Prediction (Classification)	Instruction: *Evaluate whether the given molecule is able to enter the blood-brain barrier.* (Optional: The SMILES sequence is: `SMILES`) Output: *Yes.*
Molecule Captioning	Instruction: *Could you give me a brief overview of this molecule?* (Optional: The SMILES sequence is: `SMILES`) Output: *The molecule is an indole phytoalexin that ...*
Molecule-Text Retrieval	Instruction: *Retrieve relevant text for the given molecule.* (Optional: The SMILES sequence is: `SMILES`) Output: *The molecule is associated with ...*
Caption-Guided Molecule Generation	Instruction: *Create a molecule with the structure as described: The molecule is a primary arylamine that ...* Output: `SMILES` *of the molecule.*
Reagent Prediction	Instruction: *Please provide possible reagents based on the following chemical reaction.* `<REACTANT A> <REACTANT B> ...` » `<PRODUCTs>` Output: `SMILES` *of the reagents.*
Forward Reaction Prediction	Instruction: *With the provided reactants and reagents, propose potential products:* `<REACTANT A> <REACTANT B> ...  <REAGENT A> <REAGENT B> ...` Output: `SMILES` *of the products.*
Retrosynthesis	Instruction: *Please suggest potential reactants and reagents used in the synthesis of the products:* `<PRODUCTs>` Output: `SMILES` *of the reactants and reagents.*

Table 7: The detailed training hyperparameters of UniMoT.

Configuration	Q-Former Pretraining	Tokenizer Pretraining	LLM Pretraining
Molecule Encoder	MoleculeSTM	MoleculeSTM	MoleculeSTM
SMILES Encoder	-	ChemFormer	ChemFormer
SMILES Decoder	-	ChemFormer	ChemFormer
LLM Base	-	-	Llama-2-7B
Epoch	50	50	10
Optimizer	AdamW	AdamW	AdamW
Codebook Size	2048	2048	2048
Number of Queries	16	16	16
Query Emb. Dim.	768	768	768
Molecule Emb. Dim.	300	300	300
Batch Size	64	64	32
Minimal LR	1e-5	1e-5	1e-5
LR Scheduler	Cosine	Cosine	Cosine
Warm-up Steps	1000	1000	1000
Weight Decay	0.05	0.05	0.05
LoRA Config	-	-	$r = 8$, $\alpha = 32$, dropout $= 0.1$
Precision	bfloat16	bfloat16	bfloat16
GPU Usage	4 NVIDIA A100s	4 NVIDIA A100s	4 NVIDIA A100s
Training Time	20 GPU hours	40 GPU hours	50 GPU hours

Table 8: Mean Absolute Error (MAE) of molecular property prediction task (regression) on the QM9 (Fang et al., 2023) dataset. **Bold** indicates the best performance and underline indicates the second best performance. $\Delta\epsilon$ is the HOMO-LUMO energy gap.

Model	HOMO↓	LUMO↓	$\Delta\epsilon$ ↓	AVG↓
Alpaca (Llama-7B)	-	-	-	322.109
Baize (Llama-7B)	-	-	-	261.343
Llama-2-7B	0.7367	0.8641	0.5152	0.7510
Vicuna-13B	0.7135	3.6807	1.5407	1.9783
Mol-Instructions (Llama-7B)	0.0210	0.0210	0.0203	0.0210
InstructMol (Vicuna-7B)	0.0048	0.0050	0.0061	0.0050
UniMoT (Llama-2-7B)	**0.0042**	**0.0047**	**0.0055**	**0.0049**

Table 9: Performance (%) of molecule captioning task on the ChEBI-20 (Edwards et al., 2022) dataset. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
T5-Small	50.1	41.5	60.2	44.6	54.5	53.2
T5-Base	51.1	42.3	60.7	45.1	55.0	53.9
T5-Large	55.8	46.7	63.0	47.8	56.9	58.6
MolT5-Small (T5-Small)	51.9	43.6	62.0	46.9	56.3	55.1
MolT5-Base (T5-Base)	54.0	45.7	63.4	48.5	57.8	56.9
MolT5-Large (T5-Large)	59.4	50.8	65.4	51.0	59.4	61.4
MoMu-Small (T5-Small)	53.2	44.5	-	-	56.4	55.7
MoMu-Base (T5-Base)	54.9	46.2	-	-	57.5	57.6
MoMu-Large (T5-Large)	59.9	51.5	-	-	59.3	59.7
InstructMol (Vicuna-7B)	47.5	37.1	56.6	39.4	50.2	50.9
MolCA (OPT-125M)	61.6	52.9	67.4	53.3	61.5	63.9
MolCA (OPT-1.3B)	63.9	55.5	69.7	55.8	63.6	66.9
UniMoT (Llama-2-7B)	**66.4**	**58.3**	**72.2**	**58.4**	**66.4**	**70.3**

Molecule-Text Retrieval Task. The molecule-text retrieval task involves using a molecule to retrieve text (M2T) and using text to retrieve a molecule (T2M). We compare UniMoT with several baselines: Sci-BERT Beltagy et al. (2019), KV-PLM Zeng et al. (2022), MoMu Su et al. (2022), MoleculeSTM Liu et al. (2023a), MolCA Liu et al. (2023b), and 3D-MoLM Li et al. (2024). We report the performance of retrieval using a batch of 64 random samples and the entire test set, evaluated with the metrics of Accuracy and Recall@20. We use the checkpoint from Stage-1 of pretraining. UniMoT is evaluated on the datasets of PubChem Kim et al. (2023), PCdes Zeng et al. (2022), and MoMu Su et al. (2022). Performance on the PubChem dataset is shown in Table 11, while performance on the PCdes Zeng et al. (2022) and MoMu Su et al. (2022) datasets is shown in Table 12. UniMoT demonstrates superior performance over the baselines on molecule-text retrieval, particularly in molecule-to-text retrieval. This demonstrates that UniMoT has learned fine-grained alignment between molecules and text, and it can understand molecule-text interactions through the introduction of the Causal Q-Former.

Molecule Generation Tasks. Molecule generation tasks include caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. Caption-guided molecule generation involves creating molecular structures from textual descriptions, leveraging NLP and cheminformatics to interpret and translate descriptions into chemical structures. Reagent prediction focuses on identifying suitable reagents for given reactants and desired products, optimizing synthetic routes. Forward reaction prediction forecasts probable products from specific reactants and reagents, using knowledge of chemical reactivity. Retrosynthesis deconstructs target molecules into simpler starting materials. The results of forward reaction prediction and retrosynthesis deconstructs are presented in Table 13

In molecule generation tasks, evaluating the quality of generated molecules involves several metrics that measure different aspects of similarity and validity. Exact Match checks if the generated molecule is identical to the target molecule, offering a stringent criterion for precise replication but potentially overlooking chemically similar variants. The BLEU score Papineni et al. (2002), adapted from

Table 10: Examples of molecule captioning task on the ChEBI-20 dataset. We highlight in blue the text that accurately describes the molecule structures in the generated caption, ensuring alignment with the ground truth.

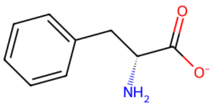
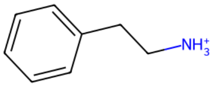
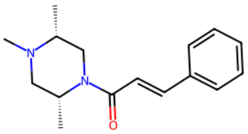
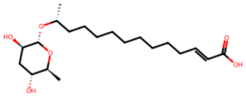
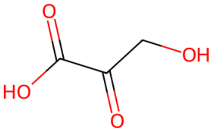
Molecule	Generated Molecule Caption	Ground Truth
	The molecule is an optically active form of phenylalaninate having D-configuration. It is a conjugate base of a D-phenylalanine. It is an enantiomer of a L-phenylalaninate.	The molecule is the D-enantiomer of phenylalaninate. It is a conjugate base of a D-phenylalanine. It is an enantiomer of a L-phenylalaninate.
	The molecule is an ammonium ion that is the conjugate acid of 2-phenylpropylamine arising from protonation of the primary amino function; major species at pH 7.3. It has a role as a human metabolite, an Escherichia coli metabolite and a mouse metabolite. It is a conjugate acid of a 2-phenylpropylamine.	The molecule is the cation obtained by protonation of the amino group of 2-phenylethylamine. It has a role as a human metabolite and an Escherichia coli metabolite. It is a conjugate acid of a 2-phenylethylamine.
	The molecule is an enamide obtained by the carboxy group of trans-cinnamic acid with the secondary amino group of (2S,5R)-1,2,5-trimethylpiperazine. It has a role as an Aspergillus metabolite. It is an alkaloid, a N-acylpiperazine, an enamide and a tertiary carboxamide. It derives from a trans-cinnamic acid.	The molecule is an enamide obtained by formal condensation of the carboxy group of trans-cinnamic acid with the secondary amino group of (2R,5R)-1,2,5-trimethylpiperazine. It has a role as an Aspergillus metabolite. It is a N-acylpiperazine, a N-alkylpiperazine, an alkaloid, an enamide and a tertiary carboxamide. It derives from a trans-cinnamic acid.
	The molecule is an (omega-1)-hydroxy fatty acid ascaroside obtained by formal condensation of the alcoholic hydroxy group of (10R)-10-hydroxylauric acid with ascarylopyranose (the alpha anomer). It is a metabolite of the nematode Caenorhabditis elegans. It has a role as a Caenorhabditis elegans metabolite. It is a monocarboxylic acid and an (omega-1)-hydroxy fatty acid ascaroside. It derives from an (11R)-11-hydroxylauric acid. It is a conjugate acid of an ascr18(1-).	The molecule is an (omega-1)-hydroxy fatty acid ascaroside obtained by formal condensation of the alcoholic hydroxy group of (10R)-10-hydroxyundecanoic acid with ascarylopyranose (the alpha anomer). It is a metabolite of the nematode Caenorhabditis elegans. It is a monocarboxylic acid and an (omega-1)-hydroxy fatty acid ascaroside. It derives from a (10R)-10-hydroxyundecanoic acid. It is a conjugate acid of an ascrblue18(1-).
	The molecule is a 2-oxo monocarboxylic acid that is pyruvic acid in which one of the methyl hydrogens is substituted by a 4-vinylcyclohex-2-en-1-yl group. It has a role as a plant metabolite. It derives from a pyruvic acid. It is a conjugate acid of a 4-[(1E)-4-vinylcyclohex-2-en-1-yl]pyruvate.	The molecule is a 2-oxo monocarboxylic acid that is pyruvic acid in which one of the methyl hydrogens has been replaced by a methylenecyclopropyl group. It has a role as a rat metabolite and a xenobiotic metabolite. It is a 2-oxo monocarboxylic acid, a member of cyclopropanes and an olefinic compound. It derives from a pyruvic acid.

Table 11: Performance (%) of molecule-text retrieval task on the PubChem (Kim et al., 2023) dataset. **Bold** indicates the best performance and underline indicates the second best performance.

Model	Retrieval in batch				Retrieval in test set			
	M2T (%)		T2M (%)		M2T (%)		T2M (%)	
	Acc↑	R@20↑	Acc↑	R@20↑	Acc↑	R@20↑	Acc↑	R@20↑
Sci-BERT	85.3	98.7	84.2	98.4	41.7	87.3	40.2	86.8
KV-PLM	86.1	98.6	85.2	98.5	42.8	88.5	41.7	87.8
MoMu (Sci-BERT)	87.6	99.2	86.4	99.4	47.3	90.8	48.1	89.9
MoMu (KV-PLM)	88.2	99.4	87.3	99.4	48.5	91.6	49.5	90.7
MoleculeSTM	90.5	99.6	88.6	99.5	52.7	92.9	53.2	92.5
MolCA (OPT-1.3B)	92.6	99.8	91.3	99.5	67.9	94.4	68.6	93.3
3D-MoLM (Llama-2-7B)	93.5	**100.0**	**92.9**	**99.6**	69.1	95.9	**70.1**	**94.9**
UniMoT (Llama-2-7B)	**93.6**	**100.0**	92.7	99.4	**69.5**	**96.3**	69.8	94.4

Table 12: Accuracy (%) of molecule-text retrieval task on the PCdes (Zeng et al., 2022) and MoMu (Su et al., 2022) datasets. **Bold** indicates the best performance and underline indicates the second best performance. We report the performance of retrieval using a batch of 64 random samples and the entire test set.

(a) Accuracy (%) of molecule-text retrieval task on the PCdes (Zeng et al., 2022) dataset.

Model	Retrieval in batch		Retrieval in test set	
	M2T (%)	T2M (%)	M2T (%)	T2M (%)
Sci-BERT	62.6	61.8	60.7	60.8
KV-PLM	77.9	65.0	75.9	64.3
MoMu (Sci-BERT)	80.6	77.0	79.1	75.5
MoMu (KV-PLM)	81.1	80.2	80.2	79.0
MoleculeSTM	86.2	83.9	84.6	85.1
MolCA (OPT-1.3B)	91.4	88.4	90.5	87.6
3D-MoLM (Llama-2-7B)	92.3	**89.6**	91.2	**88.5**
UniMoT (Llama-2-7B)	**92.6**	89.4	**91.6**	88.3

(b) Accuracy (%) of molecule-text retrieval task on the MoMu (Su et al., 2022) dataset.

Model	Retrieval in batch		Retrieval in test set	
	M2T (%)	T2M (%)	M2T (%)	T2M (%)
Sci-BERT	1.4	1.6	0.3	0.3
KV-PLM	1.5	1.3	0.5	0.3
MoMu (Sci-BERT)	45.7	40.0	43.3	43.4
MoMu (KV-PLM)	46.2	38.5	43.7	43.5
MoleculeSTM	81.8	81.9	75.8	74.5
MolCA (OPT-1.3B)	83.7	84.3	88.6	87.3
3D-MoLM (Llama-2-7B)	84.9	85.4	89.9	88.7
UniMoT (Llama-2-7B)	**85.4**	**85.6**	**90.3**	**89.0**

Table 13: Performance of molecule generation tasks on the Mol-Instructions Fang et al. (2023) benchmark, including forward reaction prediction and retrosynthesis. **Bold** indicates the best performance, and underline indicates the second best performance.

Model	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Forward Reaction Prediction							
Llama	0.000	0.020	42.002	0.001	0.002	0.001	0.039
Vicuna	0.000	0.057	41.690	0.007	0.016	0.006	0.059
Mol-Instructions	0.045	0.654	27.262	0.313	0.509	0.262	1.000
InstructMol	0.536	0.967	10.851	0.776	0.878	0.741	1.000
UniMoT	**0.611**	**0.980**	**8.297**	**0.836**	**0.911**	**0.807**	1.000
Retrosynthesis							
Llama	0.000	0.036	46.844	0.018	0.029	0.017	0.010
Vicuna	0.000	0.057	46.877	0.025	0.030	0.021	0.017
Mol-Instructions	0.009	0.705	31.227	0.283	0.487	0.230	1.000
InstructMol	0.407	0.941	13.967	0.753	0.852	0.714	1.000
UniMoT	**0.478**	**0.974**	**11.634**	**0.810**	**0.909**	**0.771**	1.000
Reagent Prediction							
Llama	0.000	0.003	28.040	0.037	0.001	0.001	0.001
Vicuna	0.000	0.010	27.948	0.038	0.002	0.001	0.007
Mol-Instructions	0.044	0.224	23.167	0.237	0.364	0.213	1.000
InstructMol	0.129	0.610	19.664	0.444	0.539	0.400	1.000
UniMoT	**0.167**	**0.728**	**14.588**	**0.549**	**0.621**	**0.507**	1.000

Table 14: Ablation study on the LLM architecture and adaptation strategy.

Architecture	Adaptation	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Galactica-125M	LoRA	28.7	21.5	34.2	21.1	30.3	31.0
Galactica-1.3B	LoRA	30.2	22.8	36.0	22.4	32.2	33.2
Mistral-7B	LoRA	32.0	24.2	38.0	24.0	34.1	35.2
Llama-2-7B	LoRA	31.3	23.8	37.5	23.7	33.6	34.8
Llama-2-7B	FFT	32.0	24.6	38.3	24.3	34.7	35.6
Llama-2-13B	LoRA	31.8	24.3	38.0	24.1	34.4	35.3

machine translation, measures the overlap of n-grams (short sequences of atoms or bonds) between generated and target molecules, thus assessing partial similarities. Levenshtein Distance Levenshtein et al. (1966) evaluates the minimum number of edits needed to transform the generated molecule into the target, providing insight into structural changes required. RDKit Landrum et al. (2006), MACCS Durant et al. (2002), and Morgan Morgan (1965) Fingerprint Similarities compare the generated and target molecules based on various molecular fingerprinting methods, which capture different aspects of molecular structure and properties. The Validity Kusner et al. (2017) metric assesses the proportion of chemically valid molecules generated, ensuring that the output consists of plausible chemical structures. Together, these metrics offer a comprehensive evaluation framework, balancing exact matches with structural and chemical validity.

E Additional Ablation Studies

LLM Architecture and Adaptation. We conducted a comparison of molecule captioning performance across various LLM architectures and adaptation strategies, as illustrated in Table 14. Our experiments show that UniMoT performs well across multiple LLM architectures, including Galactica Taylor et al. (2022) and Mistral Jiang et al. (2023) series, demonstrating its robustness and generalizability. The experiments also indicate that scaling up the LLM to 13B or adopting a full fine-tuning (FFT) strategy yields only marginal improvements in performance compared to using Llama-2-7B with LoRA. While larger models and FFT strategy might offer slight gains in performance, they come at a significant cost in terms of efficiency.

Table 15: Ablation study on the codebook size.

Architecture	Codebook Size	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Llama-2-7B	512	28.7	20.5	33.2	20.7	29.6	30.2
Llama-2-7B	1024	29.5	21.3	34.5	21.8	30.9	31.5
Llama-2-7B	2048	**31.3**	**23.8**	**37.5**	**23.7**	**33.6**	**34.8**
Llama-2-7B	4096	31.1	23.6	37.1	23.5	33.2	34.3

Table 16: Performance of UniMoT and MolCA using comparable model sizes on the molecule captioning task using the PubChem dataset.

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolCA (Galactica-125M)	25.9	17.5	34.4	16.6	23.9	28.5
MolCA (Galactica-1.3B)	28.6	21.3	36.2	21.4	29.7	32.6
MolCA (Llama-2-7B)	28.2	21.0	33.5	20.9	30.0	30.8
UniMoT (Galactica-125M)	28.7	21.5	34.2	21.1	30.3	31.0
UniMoT (Galactica-1.3B)	30.2	22.8	36.0	22.4	32.2	33.2
UniMoT (Llama-2-7B)	31.3	23.8	37.5	23.7	33.6	34.8

Codebook Size. We conducted experiments with different molecule codebook sizes and reported the performance on the molecule captioning task. The performance is shown in Table 15. The results demonstrate that the codebook size of 2048 consistently provides the best performance. This choice balances model complexity and performance. A larger codebook could capture more subtle interactions between molecules and text. However, there may be some codes that are not often used. A smaller codebook may result in nearby embeddings being assigned the same code, which reduces the granularity of the representation.

Models with Comparable Sizes. We conducted a comprehensive performance comparison between UniMoT and MolCA (Liu et al., 2023b) using models of comparable sizes, as detailed in Table 16. The results show that UniMoT consistently outperforms MolCA across various LLM architectures, including Galactica-125M, Galactica-1.3B, and LLaMA-2-7B. This consistent performance highlights the effectiveness of UniMoT in handling molecule-to-text tasks, further validating the superiority of tokenizer-based architecture over adapter-based architecture. The tokenizer-based architecture can achieve better molecule-text alignment through autoregressive molecule-to-text and text-to-molecule pretraining compared to other architectures.

Query Size. We also conducted an ablation study to evaluate the performance of UniMoT with different query sizes, as presented in Table 17. The results indicate that increasing the query size leads to improved performance, with the best performance achieves at a query size of 32. However, this larger query size also demands significantly more training time and memory. Therefore, for a more efficient balance between performance and resource consumption, we opt to use a query size of 16, which still offers strong performance while being more computationally feasible.

Table 17: Performance of UniMoT with different query sizes on the molecule captioning task using the PubChem dataset.

Architecture	Query Size	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Llama-2-7B	4	25.1	18.3	30.2	18.5	26.1	27.3
Llama-2-7B	8	29.5	21.3	34.5	21.8	30.9	31.5
Llama-2-7B	16	31.3	23.8	37.5	23.7	33.6	34.8
Llama-2-7B	32	32.2	24.9	38.2	24.4	34.7	35.7

Table 18: ROC-AUC (%) of molecular property prediction task (classification) on the MoleculeNet (Wu et al., 2018) datasets. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BBBP↑	Tox21↑	ToxCast↑	Sider↑	ClinTox↑	MUV↑	HIV↑	BACE↑
KV-PLM	70.50	72.12	55.03	59.83	89.17	54.63	65.40	78.50
EdgePred	67.30	76.00	64.10	60.40	64.10	74.10	76.30	79.90
AttrMask	67.79	75.00	63.57	58.05	75.44	73.76	75.44	80.28
InfoGraph	64.84	76.24	62.68	59.15	76.51	72.97	70.20	77.64
MolCLR	67.79	75.55	64.58	58.66	84.22	72.76	75.88	71.14
GraphMVP	68.11	77.06	65.11	60.64	84.46	74.38	77.74	80.48
GraphCL	69.70	73.90	62.40	60.50	76.00	69.80	**78.50**	75.40
Mole-BERT	71.90	76.80	64.30	62.80	78.90	**78.60**	78.20	80.80
MoMu-S	70.50	75.60	63.40	60.50	79.90	70.50	75.90	76.70
MoMu-K	70.10	75.60	63.00	60.40	77.40	71.10	76.20	77.10
MoleculeSTM	69.98	76.91	65.05	60.96	92.53	73.40	76.93	80.77
ChemBERTa	64.30	72.80	-	-	73.30	-	62.20	-
GIT-Mol	**73.90**	75.90	**66.80**	**63.40**	88.30	-	-	81.08
InstructMol (Vicuna-7B)	70.00	74.67	64.29	57.80	91.48	74.62	68.90	82.30
MolCA (OPT-1.3B)	70.00	**77.20**	64.50	63.00	89.50	-	-	79.80
UniMoT (Llama-2-7B)	71.37	76.43	65.78	59.79	**92.89**	75.97	78.49	**83.69**

F Additional Related Work

Multi-modal Large Language Models. With the rapid advancement of Large Language Models (LLMs), current multi-modal LLMs are typically built on a pre-trained LLM backbone and equipped with the ability to understand multiple modalities. LLaVA (Liu et al., 2024a) uses a simple linear projection to connect the image encoder with the LLM backbone. In contrast, BLIP-2 (Li et al., 2023) uses CLIP (Radford et al., 2021) to extract high-level features from images and employs a Q-Former to reduce the number of image tokens. These models demonstrate strong multi-modal comprehension abilities but often overlook the important aspect of multi-modal generation. Consequently, recent research has focused on unifying multi-modal comprehension and generation within a single model, enabling the generation of multi-modal tokens. Emu (Sun et al., 2023) and Emu2 (Sun et al., 2024) introduce a unified autoregressive objective: predicting the next multi-modal element by regressing visual embeddings or classifying text tokens. CM3Leon (Yu et al., 2023) and Chameleon (Team, 2024) train token-based autoregressive models on mixed image and text data. SEED-LLaMA (Ge et al., 2023) proposes a new image tokenizer aligned with the LLMs' embedding space. AnyGPT (Zhan et al., 2024) and MIO (Wang et al., 2024) construct any-to-any multi-modal language models that use discrete tokens for unified processing across various modalities. Inspired by these developments in multi-modal LLMs, we introduce a tokenizer-based architecture in the molecule-text domain. This architecture discretizes molecule features into tokens compatible with LLMs, enabling molecules to be processed alongside text tokens.

G Experimental Results with Additional Baselines

We aim to enhance the molecule comprehension and generation experiments in the main text by including additional baselines such as EdgePred (Hu et al., 2019), GraphCL (You et al., 2020), Mole-BERT (Xia et al., 2022), MoMu (Su et al., 2022), ChemBERTa (Chithrananda et al., 2020), GIT-Mol (Liu et al., 2024c), MolCA (Liu et al., 2023b), Text+Chem T5 (Christofidellis et al., 2023), and DRAk (Liu et al., 2024b). These baselines are presented in Tables 18, 19, 20, and 21.

H Broader Impacts

The development of UniMoT, a unified model for molecule and text modalities, has significant potential to positively impact various fields. UniMoT can streamline the drug discovery process by enabling efficient molecule generation and optimization based on textual descriptions. In material

Table 19: Performance (%) of molecule captioning task on the PubChem (Kim et al., 2023) dataset. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-Small (T5-Small)	22.5	15.2	30.4	13.5	20.3	24.0
MolT5-Base (T5-Base)	24.5	16.6	32.2	14.0	21.4	26.1
MolT5-Large (T5-Large)	25.9	17.3	34.1	16.4	23.4	28.0
MoMu-Small (T5-Small)	22.9	16.0	31.0	13.7	20.8	24.4
MoMu-Base (T5-Base)	24.7	16.8	32.5	14.6	22.1	27.2
MoMu-Large (T5-Large)	26.3	18.0	34.8	16.9	24.8	28.7
InstructMol (Vicuna-7B)	18.9	11.7	27.3	11.8	17.8	21.3
MolCA (OPT-125M)	25.9	17.5	34.4	16.6	23.9	28.5
MolCA (OPT-1.3B)	28.6	21.3	36.2	21.4	29.7	32.6
3D-MoLM (Llama-2-7B)	30.3	22.5	36.8	22.3	31.2	33.1
UniMoT (Llama-2-7B)	**31.3**	**23.8**	**37.5**	**23.7**	**33.6**	**34.8**

Table 20: Performance (%) of molecule captioning task on the ChEBI-20 (Edwards et al., 2022) dataset. **Bold** indicates the best performance and underline indicates the second best performance.

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
T5-Small	50.1	41.5	60.2	44.6	54.5	53.2
T5-Base	51.1	42.3	60.7	45.1	55.0	53.9
T5-Large	55.8	46.7	63.0	47.8	56.9	58.6
MolT5-Small (T5-Small)	51.9	43.6	62.0	46.9	56.3	55.1
MolT5-Base (T5-Base)	54.0	45.7	63.4	48.5	57.8	56.9
MolT5-Large (T5-Large)	59.4	50.8	65.4	51.0	59.4	61.4
MoMu-Small (T5-Small)	53.2	44.5	-	-	56.4	55.7
MoMu-Base (T5-Base)	54.9	46.2	-	-	57.5	57.6
MoMu-Large (T5-Large)	59.9	51.5	-	-	59.3	59.7
Text+Chem T5 (T5-Small)	56.0	47.0	63.8	48.8	58.0	58.8
Text+Chem T5 (T5-Base)	62.5	54.2	68.2	54.3	62.2	64.8
InstructMol (Vicuna-7B)	47.5	37.1	56.6	39.4	50.2	50.9
MolCA (OPT-125M)	61.6	52.9	67.4	53.3	61.5	63.9
MolCA (OPT-1.3B)	63.9	55.5	69.7	55.8	63.6	66.9
UniMoT (Llama-2-7B)	**66.4**	**58.3**	**72.2**	**58.4**	**66.4**	**70.3**

Table 21: Performance of molecule generation tasks on the Mol-Instructions (Fang et al., 2023) benchmark, including caption-guided molecule generation, reagent prediction, forward reaction prediction, and retrosynthesis. **Bold** indicates the best performance, and underline indicates the second best performance.

Model	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Caption-guided Molecule Generation							
Llama	0.000	0.003	59.864	0.005	0.000	0.000	0.003
Vicuna	0.000	0.006	60.356	0.006	0.001	0.000	0.001
Mol-Instructions	0.002	0.345	41.367	0.231	0.412	0.147	1.000
DRAk-K	0.104	0.515	32.641	0.455	0.600	0.326	1.000
MolT5	0.112	0.546	38.276	0.400	0.538	0.295	0.773
UniMoT	**0.237**	**0.698**	**27.782**	**0.543**	**0.651**	**0.411**	1.000
Reagent Prediction							
Llama	0.000	0.003	28.040	0.037	0.001	0.001	0.001
Vicuna	0.000	0.010	27.948	0.038	0.002	0.001	0.007
Mol-Instructions	0.044	0.224	23.167	0.237	0.364	0.213	1.000
DRAk-K	0.049	0.487	22.87	0.238	0.331	0.207	1.000
InstructMol	0.129	0.610	19.664	0.444	0.539	0.400	1.000
UniMoT	**0.167**	**0.728**	**14.588**	**0.549**	**0.621**	**0.507**	1.000
Forward Reaction Prediction							
Llama	0.000	0.020	42.002	0.001	0.002	0.001	0.039
Vicuna	0.000	0.057	41.690	0.007	0.016	0.006	0.059
Mol-Instructions	0.045	0.654	27.262	0.313	0.509	0.262	1.000
DRAk-K	0.254	0.778	18.649	0.602	0.741	0.546	1.000
InstructMol	0.536	0.967	10.851	0.776	0.878	0.741	1.000
UniMoT	**0.611**	**0.980**	**8.297**	**0.836**	**0.911**	**0.807**	1.000
Retrosynthesis							
Llama	0.000	0.036	46.844	0.018	0.029	0.017	0.010
Vicuna	0.000	0.057	46.877	0.025	0.030	0.021	0.017
Mol-Instructions	0.009	0.705	31.227	0.283	0.487	0.230	1.000
DRAk-K	0.319	0.793	20.779	0.625	0.758	0.565	1.000
InstructMol	0.407	0.941	13.967	0.753	0.852	0.714	1.000
UniMoT	**0.478**	**0.974**	**11.634**	**0.810**	**0.909**	**0.771**	1.000

science, it can aid in discovering new materials with desirable properties. Additionally, UniMoT can enhance research collaboration between chemists, biologists, and data scientists by integrating molecular and textual data, leading to comprehensive research insights and innovative solutions.

This paper does not pose any ethical concerns. The study does not involve human subjects and follows proper procedures for dataset releases. There are no potentially harmful insights, methodologies, or applications. Additionally, there are no conflicts of interest or sponsorship concerns. Discrimination, bias, and fairness issues are not applicable. Privacy and security matters have been appropriately addressed, legal compliance has been maintained, and research integrity has been upheld.