Unified Molecule-Text Language Model with Discrete Token Representation

Shuhan Guo[1] , **Yatao Bian**[2] , **Ruibing Wang**[3] , **Nan Yin**[4] , **Zhen Wang**[3] , **Quanming Yao**[1,*]
[1]Tsinghua University
[2]Tencent AI Lab
[3]Northwestern Polytechnical University
[4]Hong Kong University of Science and Technology
{guoshuhan, qyaoaa}@tsinghua.edu.cn, {yatao.bian, yinnan8911}@gmail.com,
wrb5261@mail.nwpu.edu.cn, w-zhen@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

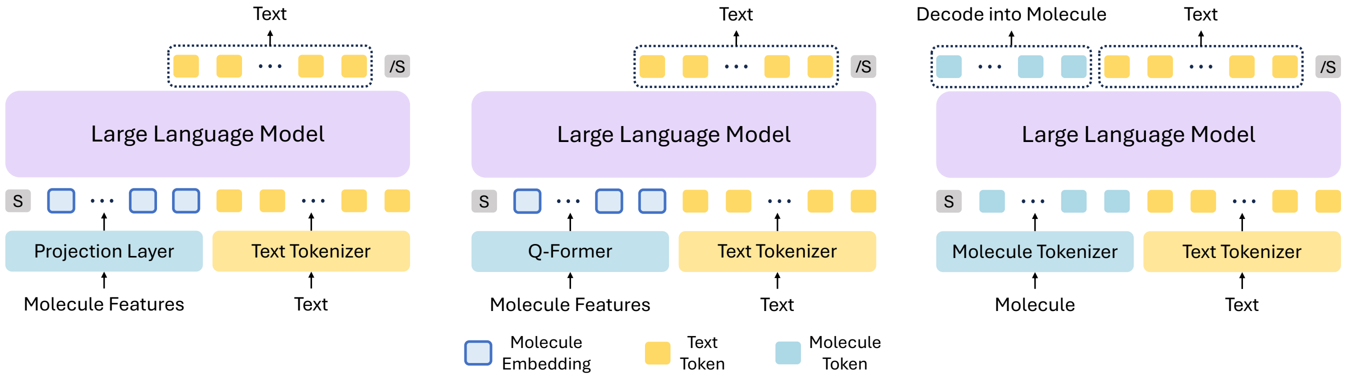
The incredible capabilities of Large Language Models (LLMs) [Brown *et al.*, 2020; Touvron *et al.*, 2023] have led to their widespread use as versatile tools for completing diverse real-world tasks. This success has sparked interest in Multi-modal LLMs [Zhan *et al.*, 2024], which aim to enhance LLMs by enabling them to process multi-modal inputs and outputs. In fields like molecular science, Multi-modal LLMs present new opportunities by seamlessly integrating molecular data with textual information, opening up fresh possibilities for more efficient and accurate research and development.Prior research efforts [Liang *et al.*, 2023; Fang *et al.*, 2023; Cao *et al.*, 2023; Liu *et al.*, 2023b; 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 [Liu *et al.*, 2023b; Cao *et al.*, 2023], address drug-related inquiries [Liang *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 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

[*] Corresponding author



(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.

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 can integrate the molecule tokenizer by treating molecule tokens as new words and constructing a molecule vocabulary through mapping the learned codebook. We adopt the unified discrete token representation for molecules and text, coupled with the unified next-token-prediction training paradigm of LLM. This unification of representation and training paradigm enables effective molecule-text interactions and alignment through molecule-to-text and text-to-molecule autoregressive pretraining. For molecule generation tasks, UniMoT generates molecule tokens in an autoregressive manner rather than producing SMILES strings, and these molecule tokens can then be decoded into molecules using 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 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 comprehension and generation tasks, while also offering a new perspective on molecule generation.

2 Related Works

Multi-modal Large Language Models. Multi-modal Large Language Models (LLMs): Current multi-modal LLMs are typically built on a pre-trained LLM backbone and can understand multiple modalities. LLaVA [Liu *et al.*, 2024a] connects the image encoder to the LLM using a simple linear projection, while BLIP-2 [Li *et al.*, 2023] extracts high-level features from images with CLIP [Radford *et al.*, 2021] and uses Q-Former to reduce image token counts. While these models excel at multi-modal comprehension, they often lack focus on multi-modal generation. To address this, recent work unifies multi-modal comprehension and generation, such as SEED-LLaMA [Ge *et al.*, 2023] and AnyGPT [Zhan *et al.*, 2024], which unify processing across different modalities. Inspired by these advances, we introduce a tokenizer-based architecture in the molecule-text domain, converting molecular features into tokens compatible with LLMs.

Molecular Large Language Models. The recent emergence of Vision Large Language Models (VLLMs) [Li *et al.*, 2023] 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]. 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.*, 2024b], 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 modalities equally and lack a supervision signal for the molecule modality, limiting model capacity

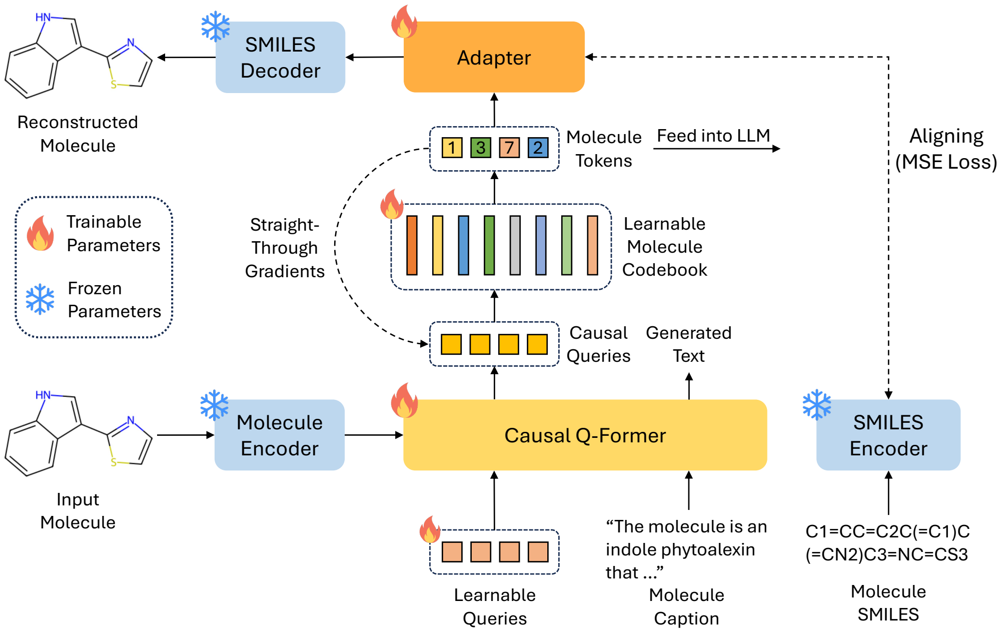


Figure 2: Illustration of our proposed molecule tokenizer. The tokenizer generates discrete molecule tokens, which can be fed into LLMs for downstream tasks. The generated molecule tokens can be decoded into molecules using the adapter and the SMILES decoder during inference.

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

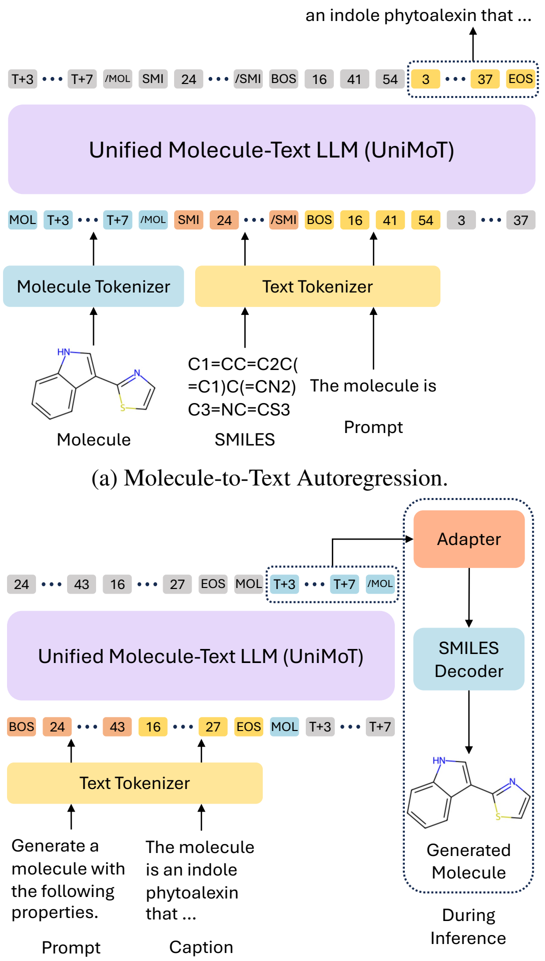
Our objective is to leverage the reasoning and generation capabilities of LLMs to enhance the comprehension and generation of molecule and text data. To achieve this, we focus on representing these modalities uniformly within the token representation, utilizing the next-token-prediction training paradigm of LLMs. As illustrated in Figure 2, we introduce a molecule tokenizer (Section 3.1) designed to transform molecules into molecule tokens by learning to reconstruct the input molecule. The molecule sequence can then be concatenated with the text sequence to form a multi-modal sequence, which is fed into an LLM for 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 introduce a four-stage training scheme for UniMoT (Section 3.3) comprising Causal Q-Former pretraining, molecule tokenizer pretraining, unified molecule-text pretraining, and task-specific instruction tuning. UniMoT is capable of performing molecule comprehension and generation tasks following the training scheme.

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 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 **??**.

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


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

randomly initialized and optimized during pretraining. Specifically, token s_i is determined as follows:

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

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:

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

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.

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:

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

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

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 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.

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 **??**.

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 **??**.

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. The goal of molecular property prediction is to forecast a molecule's intrinsic physical and chemical properties. For the classification task, we incorporate eight binary classification datasets from MoleculeNet [Wu *et al.*, 2018]. Models are tasked with generating a single prediction ("yes" or "no"). We compare UniMoT with the following baselines: 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 the MoleculeNet datasets are shown in Table 1. The performance of the regression task of molecular property prediction is provided in Appendix **??**. Compared to traditional graph learning methods and molecular LLMs like InstructMol [Cao *et al.*, 2023], UniMoT demonstrates consistent improvements across the eight datasets, indicating its robust molecule comprehension abilities.

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 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.

Molecule Captioning Task. The molecule captioning task involves generating a comprehensive description of a molecule. 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]. BLEU [Papineni *et al.*, 2002], ROUGE [Lin, 2004], and METEOR [Banerjee and Lavie, 2005] are adopted as evaluation metrics. UniMoT is evaluated for molecule captioning on the PubChem [Kim *et al.*, 2023] and ChEBI-20 [Edwards *et al.*, 2022] datasets. Performance on the PubChem dataset is shown in Table 2, while the performance on the ChEBI-20 dataset and some concrete examples are presented in Appendix **??**. The ChEBI-20 dataset replaces molecular names with "the molecule" to focus on properties. However, predicting molecular names reflects the model's structural understanding, so we conducted the main experiments on PubChem.

From Table 2, we observe that UniMoT consistently outperforms the baselines by a significant margin on the PubChem [Kim *et al.*, 2023] dataset. This task is more complex than classification or regression, providing a robust measure of the model's molecule comprehension abilities. Notably, our proposed tokenizer-based architecture surpasses the projection-based architecture (such as InstructMol [Cao *et al.*, 2023]), Q-Former-based architecture (such as MolCA [Liu *et al.*, 2023b] and 3D-MoLM [Li *et al.*, 2024]), and models trained with contrastive learning strategies (such as MoMu [Su *et al.*, 2022]). This demonstrates that the tokenizer-based architecture achieves better molecule-text alignment through autoregressive molecule-to-text and text-to-molecule pretraining compared to other architectures. Details and More Results of Experiments can be found in Appendix **??**.

4.2 Molecule Generation Tasks

We employ molecule generation tasks, which encompass caption-guided molecule generation [Fang *et al.*, 2023], reagent prediction [Fang *et al.*, 2023], forward reaction prediction [Fang *et al.*, 2023], and retrosynthesis [Fang *et al.*, 2023]. Caption-guided molecule generation involves generating molecular structures based on textual descriptions. Reagent prediction entails determining suitable reagents given reactants and products. Forward reaction prediction involves predicting probable products given specific reactants and reagents. Retrosynthesis involves deconstructing a target molecule into simpler starting materials. We compare UniMoT with the following baselines: Llama [Touvron *et al.*, 2023], Vicuna [Chiang *et al.*, 2023], Mol-Instructions [Fang *et al.*, 2023], and InstructMol [Cao *et al.*, 2023]. The metrics used to evaluate molecule generation tasks include Exact Match, BLEU [Papineni *et al.*, 2002], Levenshtein Distance [Levenshtein and others, 1966], RDKit Fingerprint Similarity [Landrum and others, 2006], MACCS Fingerprint Similarity [Durant *et al.*, 2002], and Morgan Fingerprint Similarity [Morgan, 1965]. These metrics evaluate structural similarity between generated and target molecules, along with Validity [Kusner *et al.*, 2017], which assesses the proportion of chemically valid molecules generated.

We utilize the Mol-Instructions [Fang *et al.*, 2023] benchmark to evaluate the generation capabilities of UniMoT. The results of caption-guided molecule generation and reagent prediction are presented in Table 3, and the results of other tasks are in Appendix **??**. The caption-guided molecule generation task, the reverse of molecule captioning, is conducted using the PubChem [Kim *et al.*, 2023] dataset, while the other tasks utilize the USPTO [Fang *et al.*, 2023] dataset. As the baselines generate SMILES strings and then convert them to molecules, UniMoT directly leverages the generated molecule tokens and obtains their embeddings from the learned codebook. These embeddings can be decoded to desired molecules through the pretrained adapter and SMILES decoder. As shown in Table 3, UniMoT generates valid molecules with a higher degree of similarity to the target molecules compared to the baselines. This is because UniMoT can generate molecules as if they were text, which is fundamentally different from adapter-based architectures. UniMoT demonstrates strong generation capabilities and offers a new perspective on these tasks.

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 4a. 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

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
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 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.

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

(a) Ablation study on the projector and representation form.

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

(b) Ablation study on the codebook size.

Table 4: Ablation studies on the molecule captioning task using the PubChem dataset.

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 4a. 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.

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 4b. 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. More ablation studies are presented in Appendix **??**.

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. Although UniMoT excels in molecule-to-text and text-to-molecule tasks, it has yet to be extensively tested on more complex tasks like molecule editing, which require precise structural modifications. Additionally, limited annotated data in the molecular domain restricts UniMoT's training, hindering its ability to fully learn and generalize molecular structures and properties. To improve its effectiveness, addressing data scarcity is crucial. Furthermore, expanding evaluations to include a broader range of real-world scenarios will offer a more comprehensive understanding of the model's robustness and generalizability.

Acknowledgments

Q. Yao is supported by National Natural Science Foundation of China (under Grant No.92270106), Beijing Natural Science Foundation (under Grant No.4242039), and CCF-Zhipu Large Model Innovation Fund (CCF-Zhipu202402)

References

[Banerjee and Lavie, 2005] 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*, pages 65–72, 2005.

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

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

[Brown *et al.*, 2020] 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.

[Cao *et al.*, 2023] 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.

[Chiang *et al.*, 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. *https://vicuna. lmsys. org*, 2(3):6, 2023.

[Durant *et al.*, 2002] 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.

[Edwards *et al.*, 2022] 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.

[Fang *et al.*, 2023] 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.

[Ge *et al.*, 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.

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

[Hu *et al.*, 2019] 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.

[Hu *et al.*, 2021] 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.

[Irwin *et al.*, 2022] 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.

[Kim *et al.*, 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.

[Kusner *et al.*, 2017] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*. PMLR, 2017.

[Landrum and others, 2006] Greg Landrum et al. Rdkit: Open-source cheminformatics, 2006.

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

[Li *et al.*, 2023] 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*. PMLR, 2023.

[Li *et al.*, 2024] 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.

[Liang *et al.*, 2023] 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.

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

[Liu *et al.*, 2019] 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.

[Liu *et al.*, 2023a] 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, 2023.

[Liu *et al.*, 2023b] 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*, 2023.

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

[Liu *et al.*, 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, 2024.

[Morgan, 1965] 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.

[Papineni *et al.*, 2002] 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*, pages 311–318, 2002.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Raffel *et al.*, 2020] 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.

[Su *et al.*, 2022] 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.

[Sun *et al.*, 2019] 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.

[Taylor *et al.*, 2022] 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.

[Touvron *et al.*, 2023] 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.

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

[Wang *et al.*, 2021] 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.

[Weininger, 1988] 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.

[Wu *et al.*, 2018] 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.

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

[Xia *et al.*, 2022] 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.

[You *et al.*, 2020] 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.

[Yu *et al.*, 2021] 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.

[Zeng *et al.*, 2022] 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.

[Zhan *et al.*, 2024] 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.

[Zhuang *et al.*, 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.