Published: 27 October 2025

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

RESEARCH-ARTICLE

CROP: Integrating Topological and Spatial Structures via Cross-View Prefixes for Molecular LLMs

JIANTING TANG, University of Science and Technology of China, Hefei, Anhui, China

YUBO WANG, University of Science and Technology of China, Hefei, Anhui, China

HAOYU CAO, University of Science and Technology of China, Hefei, Anhui, China

LINLI XU, University of Science and Technology of China, Hefei, Anhui, China

MM '25: Proceedings of the 33rd ACM International Conference on Multimedia (October 2025)
https://doi.org/10.1145/3746027.3754578
ISBN: 9798400720352

CROP: Integrating Topological and Spatial Structures via Cross-View Prefixes for Molecular LLMs

Jianting Tang
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
jiantingtang@mail.ustc.edu.cn

Yubo Wang
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
wyb123@mail.ustc.edu.cn

Haoyu Cao
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
caohaoyu@mail.ustc.edu.cn

Linli Xu*
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
Hefei, Anhui, China
linlixu@ustc.edu.cn

Abstract

Recent advances in molecular science have been propelled significantly by large language models (LLMs). However, their effectiveness is limited when relying solely on molecular sequences, which fail to capture the complex structures of molecules. Beyond sequence representation, molecules exhibit two complementary structural views: the first focuses on the *topological* relationships between atoms, as exemplified by the graph view; and the second emphasizes the *spatial* configuration of molecules, as represented by the image view. The two types of views provide unique insights into molecular structures. To leverage these views collaboratively, we propose the **CRO**ss-view **P**refixes (CROP) to enhance LLMs' molecular understanding through efficient multi-view integration. CROP possesses two advantages: (*i*) efficiency: by jointly resampling multiple structural views into fixed-length prefixes, it avoids excessive consumption of the LLM's limited context length and allows easy expansion to more views; (*ii*) effectiveness: by utilizing the LLM's self-encoded molecular sequences to guide the resampling process, it boosts the quality of the generated prefixes. Specifically, our framework features a carefully designed SMILES Guided Resampler for view resampling, and a Structural Embedding Gate for converting the resulting embeddings into LLM's prefixes. Extensive experiments demonstrate the superiority of CROP in tasks including molecule captioning, IUPAC name prediction and molecule property prediction.

CCS Concepts

• **Computing methodologies → Artificial intelligence**.

Keywords

Multimodal Large Language Models, Multimodal Fusion, Molecular Image, Molecular Graph

ACM Reference Format:
Jianting Tang, Yubo Wang, Haoyu Cao, and Linli Xu. 2025. CROP: Integrating Topological and Spatial Structures via Cross-View Prefixes for Molecular LLMs. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754578

1 Introduction

LLMs have exhibited remarkable proficiency across diverse domains [43]. In the chemical field, particularly in tasks such as molecule captioning and property prediction [34], LLMs have emerged as promising tools for streamlining research efforts. As a molecule's properties are fundamentally determined by its complex structure [10, 41], providing LLMs with accurate structural representations is essential for enhancing their molecular understanding. However, current LLMs primarily rely on sequence representations like SMILES [33] and SELFIES [16] for molecular tasks, which are inadequate for capturing complex molecular structures.

To address that, recent works [2, 23, 24, 30] have preliminarily explored integrating graph-based representations into LLMs, where molecules are modeled as graphs, with atoms as nodes and chemical bonds as edges. While these graph representations effectively capture *topological* relationships between atoms [19], they still exhibit limitations. Specifically, graph representations only encode topological relationships, allowing a single graph to correspond to infinite variety of node arrangements. This ambiguity makes it challenging to represent critical characteristics such as molecular spatial configuration and overall shape, as illustrated in Figure 1. Furthermore, when processing complex molecular graphs, graph-based approaches frequently encounter issues such as over-smoothing and over-squashing [3, 15], impeding the effective utilization of graph representations by LLMs.

As a typical representation of the molecular *spatial* view, molecular images provide complementary information about the spatial configuration and overall shape of molecules. This visual representation naturally encodes important structural features that are difficult to derive from graphs, such as symmetry planes, functional

[*] Corresponding author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754578

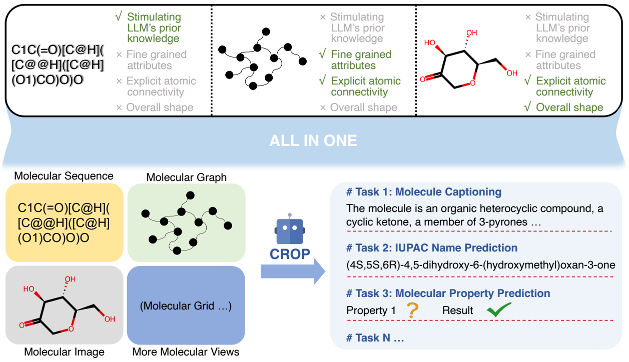

C1C(=O)[C@H](
[C@@H]([C@H]
(O1)CO)O)O
√ Stimulating LLM's prior knowledge
× Fine grained attributes
× Explicit atomic connectivity
× Overall shape
× Stimulating LLM's prior knowledge
√ Fine grained attributes
√ Explicit atomic connectivity
× Overall shape
HO
OH
O
O
OH
× Stimulating LLM's prior knowledge
× Fine grained attributes
√ Explicit atomic connectivity
√ Overall shape
ALL IN ONE
Molecular Sequence
Molecular Graph
C1C(=O)[C@H](
[C@@H]([C@H]
(O1)CO)O)O
Molecular Image
(Molecular Grid ...)
More Molecular Views
CROP
# Task 1: Molecule Captioning
The molecule is an organic heterocyclic compound, a cyclic ketone, a member of 3-pyrones ...
# Task 2: IUPAC Name Prediction
(4S,5S,6R)-4,5-dihydroxy-6-(hydroxymethyl)oxan-3-one
# Task 3: Molecular Property Prediction
Property 1 ?
Result ✓
# Task N ...

Figure 1: An overview of the strengths and weaknesses of each molecular view. Taking that into account, CROP comprehensively utilizes diverse views to enhance molecular understanding capabilities.

group positions and rigidity of molecules [42]. Benefiting from sustained advances in computer vision, some studies [11, 36, 48] that utilize images for molecular modeling have achieved impressive results in discriminative tasks, including molecular property prediction and drug target identification. Specifically, ImageMol [45] conducted 5 carefully crafted pretraining tasks on 10 million molecular images, outperforming existing sequence-based and graph-based methods on 10 regression tasks of GPCRs (G protein-coupled receptors) and 10 classification tasks of kinases in compound-protein binding prediction. While molecular images have demonstrated success in discriminative tasks, their application to generative tasks with LLMs remains largely unexplored. This work pioneers the integration of molecular visual representations to enhance LLM performance in generative tasks, such as molecule captioning and IUPAC name prediction.

Given the complementary nature of topological and spatial structures of molecules, works [2, 24] that solely introduce graph views still fundamentally limit LLMs' molecular understanding capabilities. Our key insight is that integrating topological relationships from graphs and spatial configurations from images enables a comprehensive understanding of molecular structures.

Directly concatenating embeddings from graph and image views as input would result in excessive consumption of the LLM's limited context length, and this issue exacerbates as more views are introduced, as shown in Figure 2 (*arch1*). In fact, there is substantial information irrelevance and overlap among these embeddings. For instance, a significant portion of the image embeddings correspond to the blank areas in the molecular images, and all of these molecular views capture the overlapping information of atomic and bond types. Therefore, efficiently resampling key structural features from these molecular views becomes crucial. While independent resampling the graph and image embeddings can reduce input length (Figure 2, *arch2*), it still faces increased consumption of context length when accommodating additional views. Besides, the lack of guidance from chemical domain knowledge limits the effectiveness of resampling. Therefore, as shown in Figure 2 (*arch3*), we propose to jointly resample multiple structural views into fixed-length cross-view prefixes for LLMs. To boost the quality of the generated prefixes, we utilize the LLM's self-encoded SMILES as the resampling guidance, which are enriched with the LLM's prior chemical knowledge [22].

To this end, we propose CROP, an MLLM that demonstrates outstanding molecular understanding capabilities, augmented with multiple structural views, as illustrated in Figure 3. CROP partitions the LLM backbone into lower and upper segments. The whole forward propagation process is conducted as follows: (*1*) the LLM's lower segment processes SMILES strings to generate chemical knowledge-aware guidance, referred to as SMILES guidance; (*2*) the SMILES Guided Resampler adopts the SMILES guidance to resample molecular graphs and images jointly; (*3*) the Structural Embedding Gate converts the derived structural embeddings into fixed-length cross-view prefixes; (*4*) the LLM's upper segment processes both SMILES and prefixes to obtain a comprehensive understanding of molecules. This architecture substantially enhances the effectiveness of the resampling process. Meanwhile, injecting prefixes into multiple layers of the LLM allows for deep interaction with molecular structural information, facilitating a more accurate understanding of molecular structures.

In summary, our contributions are as follows:

- We identify the fundamental limitations of current molecular MLLMs that rely solely on the graph view, which captures only topological relationships. In this work, we propose leveraging the complementary topological and spatial information conveyed by molecular graph and image views to jointly advance the molecular understanding capabilities of LLMs.
- We propose CROP, an innovative and scalable MLLM architecture that can accommodate multiple structural views to jointly enhance molecular understanding while maintaining computational efficiency.
- Through extensive evaluation, we demonstrate CROP achieves significant performance gains across a wide range of tasks, including molecule captioning, IUPAC name prediction, and molecular property prediction, highlighting the superiority of our multi-view integration approach.

2 Related Work

2.1 Molecule Modeling

SMILES and SELFIES strings can be modeled by language models in a manner similar to text sequences. Models like KV-PLM [46], MolT5 [5] and Galactica [31] excel in molecule-related tasks by bridging SMILES and biomedical text. In molecular graph modeling, both Graph Neural Networks (GNNs) [29] and Graph Transformers [44] are widely employed. Pretraining is conducted at both the graph and node levels to capture the global and local information [28, 32, 49]. Molecular images offer distinct advantages in depicting the spatial configuration and overall shape of molecules. These images can be rendered by RDKit [1] or captured using physical microscopy [7]. Chemception [12], 2DConvNet [11] and DenseNet121 [48] are pioneering works utilizing molecular images for predicting chemical properties, compound toxicity, and contaminant reactivity respectively, demonstrating the potential of molecular images for promoting downstream tasks. Recently, ImageMol [45] conducts massive self-supervised pretraining on 10 million unlabeled molecules, outperforming sequence-based and

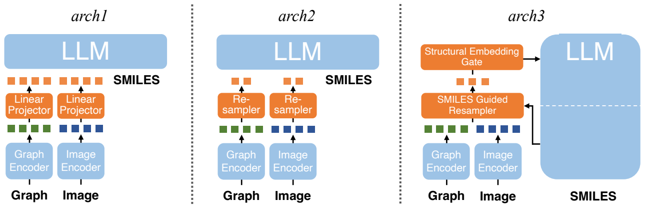


Figure 2: Comparison of three MLLM architectures with different ways to process embeddings from multiple molecular views. CROP (*arch3*) jointly resample graph and image views into fixed-length cross-view prefixes for LLMs, with LLM's prior chemical knowledge as resampling guidance, possessing both efficiency and effectiveness.

graph-based models across various benchmarks. Despite the success of molecular images in discriminative tasks, their potential to enhance the performance of LLMs in generative tasks remains largely unexplored.

2.2 Multimodal Large Language Models

MLLMs are capable of processing various modalities beyond text, most of which are tailored for natural modalities such as image [47] and audio [14, 39]. However, MLLMs designed for specialized modalities such as molecular graphs, images, and grids [37] have not been sufficiently explored, which inspires us to propose CROP, a specialized MLLM focusing on molecular modalities in the chemical field. BLIP-2 [17] and LLaVA [21] are two representative MLLM architectures, utilizing compressed and uncompressed multimodal embeddings, respectively. Considering the limited context length of LLMs and substantial information overlap among molecular views, we propose to derive efficient cross-view prefixes for LLMs, with LLM's prior chemical knowledge as resampling guidance.

2.3 MLLMs for Molecular Science

In the field of chemistry, beyond the commonly-used molecular sequence view, DrugChat [20], InstructMol [2], MolTC [8], GIT-Mol [22] and MolCA [24] introduce the graph view additionally to advance molecular understanding of LLMs. However, molecules inherently exhibit various structural views, and each of them exhibits distinct strengths and weaknesses. The model's performance remains limited when relying solely on a single graph view. In this research, we propose CROP to integrate the strengths of both graph and image views to collaboratively advance the molecular understanding capabilities. Moreover, our architecture enables seamless incorporation of additional structural views while maintaining computational efficiency.

3 Methodology

3.1 Problem Definition

In this work, we adopt SMILES as the sequence view to leverage the prior knowledge of LLMs, molecular graphs as the topological structure view to capture atomic connections, and molecular images as the spatial structure view to encode molecular configurations. Let $\boldsymbol{S} = (s_1, s_2, \ldots, s_l)$ be a molecule SMILES string tokenized based on characters, where l is the number of characters. Let $\boldsymbol{G} = (V, E)$ be a molecule graph, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of n graph nodes and E is the set of graph edges. Let $\boldsymbol{I}$ be a molecule image. Given a molecule's $\boldsymbol{S}$, $\boldsymbol{G}$ and $\boldsymbol{I}$, generative tasks such as molecule captioning and IUPAC name prediction require generating a corresponding text y. Classification tasks such as molecule property prediction require generating a probability distribution p, with cross-entropy loss used for optimization.

3.2 Model Architecture

Overview. As illustrated in Figure 3, CROP comprises four primary components: the LLM, molecule encoders, the SMILES Guided Resampler and the Structural Embedding Gate. The LLM is partitioned into lower and upper segments. The lower segment processes the SMILES to derive the SMILES guidance Z_S. Then, the SMILES Guided Resampler utilizes the SMILES guidance Z_S to jointly resample molecular graph embeddings Z_G and image embeddings Z_I derived from respective encoders, and output the structural embeddings Z. After that, the Structural Embedding Gate converts the structural embeddings Z into fixed-length cross-view prefixes $\hat{Z}$. Finally, the LLM's upper segment processes both the SMILES hidden states from the lower segment and prefixes $\hat{Z}$ to achieve a comprehensive understanding of molecules. Benefiting from the SMILES guidance Z_S, which is enriched with the LLM's prior chemical knowledge, the resampling possesses high effectiveness. Moreover, injecting cross-view prefixes $\hat{Z}$ into multiple layers of the LLM enables deep interaction with molecular structural information.

Molecule Encoders. The molecular graph encoder consists of five Graph Isomorphism Network (GIN) layers [38] initialized from moleculeSTM [23]. Molecular graphs are encoded into representations $Z_G \in \mathbb{R}^{n \times d}$, where each node representation contains local structural information of the neighboring subgraph. The molecular image encoder adopts the pretrained ResNet18 from ImageMol [45]. The original molecular image is encoded to a feature map with dimension $H \times W \times d$. This feature map is then flattened to $Z_I \in \mathbb{R}^{p \times d}$, where $p = HW$. These process can be formalized as:

$$\begin{aligned} Z_G &= GraphEncoder(\boldsymbol{G}), \\ Z_I &= ImageEncoder(\boldsymbol{I}). \end{aligned} \tag{1}$$

SMILES Guidance (SG). We adopt Galactica [31] as the LLM backbone of CROP, which is pretrained on an extensive chemical corpus and known for its strong proficiency in chemistry. Benefiting from this, we adopt the SMILES representations within Galactica as guidance, which are enriched with Galactica's prior chemical knowledge, to facilitate resampling key structural features from graphs and images. To obtain the SMILES guidance Z_S from the LLM and return cross-view prefixes $\hat{Z}$ to the LLM during a single forward propagation, we partition the LLM into lower and upper segments, consisting of b and u layers respectively.

We prepend w learnable vectors to all layers in the lower segment, enabling extensive interaction with SMILES hidden states via the LLM's attention layers. These fixed-length vectors then serve as the SMILES guidance $Z_S \in \mathbb{R}^{b \times w \times d}$. It is worth noting that the prepended vectors cannot directly perceive the SMILES tokens behind, due to the causal attention mechanism within the LLM. As illustrated in Figure 3 (a), with prepended vectors, the standard

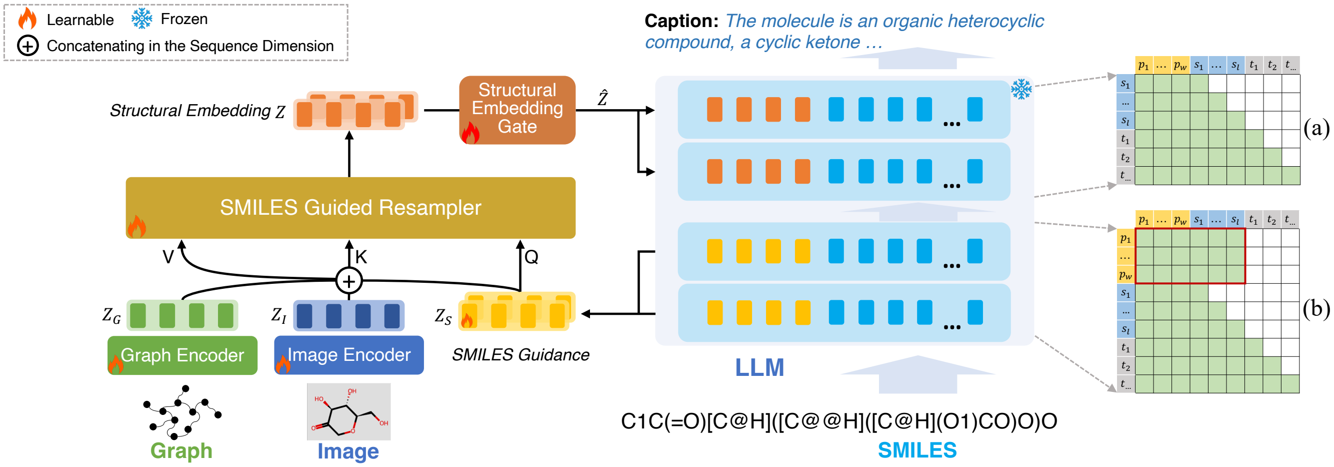

Learnable
Frozen
Concatenating in the Sequence Dimension
Structural Embedding Z
Structural Embedding Gate
$\hat{Z}$
Caption: The molecule is an organic heterocyclic compound, a cyclic ketone …
SMILES Guided Resampler
V
K
Q
Z_G
Z_I
Z_S
Graph Encoder
Image Encoder
SMILES Guidance
Graph
Image
LLM
C1C(=O)[C@H]([C@@H]([C@H](O1)CO)O)O
SMILES
p_1 … p_w s_1 … s_l t_1 t_2 $t_{…}$
(a)
(b)

Figure 3: The architecture of CROP. The SMILES Guided Resampler utilizes SMILES guidance Z_S derived from the LLM's lower segment to guide the resampling of graph and image views. The Structural Embedding Gate converts the derived structural embeddings Z into prefixes $\hat{Z}$ for the LLM's upper segment. (a) The standard attention mask of the LLM when handling prefixes, where p, s and t denote prefix, SMILES and plain text respectively. This is used in the LLM's upper segment. (b) The modified attention mask enabling prefixes to perceive SMILES tokens behind. This is employed in the LLM's lower segment.

attention calculation process can be formalized as:

$$\begin{aligned}&\text{AttentionLayer}(H, H_p, H_p)\\ =&\text{softmax}\left(\frac{(HW_Q)(H_pW_K)^T}{\sqrt{d_k}} + M\right)(H_pW_V),\end{aligned} \tag{2}$$

where $H \in \mathbb{R}^{l\times d}$ denotes SMILES hidden states, and $H_p \in \mathbb{R}^{(w+l)\times d}$ represents the concatenation of the prepended vectors and SMILES hidden states. W_Q, W_K and $W_V \in \mathbb{R}^{d\times d}$ are the query, key and value transform matrices respectively. M is the triangular causal attention mask, with $M_{i,j\leq i+w} = 0$ and $M_{i,j>i+w} = -\infty$.

In order to enable the prepended vectors to perceive the SMILES tokens behind, as illustrated in Figure 3 (b), we modify the standard attention calculation process in the lower segment, which can be formalized as:

$$\begin{aligned}&\text{AttentionLayer}(H_p, H_p, H_p)\\ =&\text{softmax}\left(\frac{(H_pW_Q)(H_pW_K)^T}{\sqrt{d_k}} + M'\right)(H_pW_V).\end{aligned} \tag{3}$$

where M' denotes the modified attention mask, with $M'_{i\leq w,j\leq w+l} = 0$, $M'_{i\leq w,j>w+l} = -\infty$, $M'_{i>w,j\leq i} = 0$ and $M'_{i>w,j>i} = -\infty$.

SMILES Guided Resampler (SGR). By leveraging the SMILES guidance, SGR proceeds to jointly resample lengthy graph and image embeddings into fixed-length structural embeddings, which integrate the strengths of both molecular views.

SGR consists of multiple transformer layers. Graph embeddings Z_G, image embeddings Z_I and SMILES guidance Z_S are concatenated along the sequence dimension to serve as keys and values. The SMILES guidance Z_S, enriched with LLM's prior chemical knowledge, serve as queries:

$$Keys, Values = [Z_G, Z_I, Z_S] \in \mathbb{R}^{b\times(n+p+w)\times d}, \tag{4}$$

$$Queries = Z_S \in \mathbb{R}^{b\times w\times d}, \tag{5}$$

$$Z = SGR(Queries, Keys, Values) \in \mathbb{R}^{b\times w\times d}. \tag{6}$$

where the resampling process is conducted through the cross attention mechanism in SGR's transformer layers, producing compact structural embeddings Z.

Beyond molecular graph and image views, SGR can further accommodate more structural views by simply concatenating their embeddings with $[Z_G, Z_I, Z_S]$, and serve as keys and values. This design ensures the extensibility and flexibility of SGR.

Structural Embedding Gate (SEG). Considering the lower and upper segment of the LLM may contain different number of layers, namely $b \neq u$, we propose the SEG. SEG can flexibly convert structural embeddings $Z \in R^{b\times w\times d}$ into fixed-length cross-view prefixes $\hat{Z} \in R^{u\times w\times d}$, which are then prepended to the LLM's upper segment. These prefixes empower the LLM to understand molecular structures accurately and comprehensively.

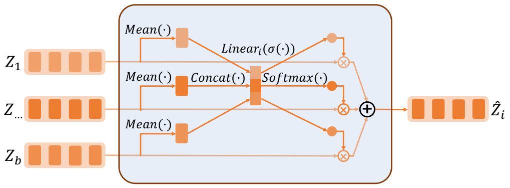


Figure 4: The architecture of the Structural Embedding Gate for the LLM's Layer$_i$, which sums b groups of $Z_j \in \mathbb{R}^{w\times d}$ with the weighted vector $P_i \in \mathbb{R}^b$ to obtain the prefixes $\hat{Z}_i \in \mathbb{R}^{w\times d}$.

Table 1: Molecule captioning results on PubChem324k and CheBI-20 datasets. Bold denotes the best performance.

Dataset	Model	Modalities	TrainableParams	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
PubChem324k	GPT-4o	S	- , 3-shot	16.5	7.1	30.8	11.3	23.1	22.5
	Llama3$_{Instruct8B}$	S	- , 3-shot	14.4	6.4	30.3	12.8	24.7	20.3
	BioT5	S	252M, full ft	42.9	34.3	53.1	39.8	47.5	48.7
	MolT5-Large	S	780M, full ft	30.2	22.2	41.5	25.9	34.8	36.6
	MoMu-Large	$S+G$	782M, full ft	31.1	22.8	41.8	25.7	36.7	36.2
	MolCA$_{Galac1.3B}$	$S+G$	100M, LoRA ft	38.7	30.3	50.2	35.9	44.5	45.6
	CROP$_{Galac1.3B}$	S	71M, LoRA ft	36.6	29.1	48.7	34.8	43.1	43.5
	CROP$_{Galac1.3B}$	$S+G$	71M, LoRA ft	43.4	35.4	54.0	40.1	48.7	49.5
	CROP$_{Galac1.3B}$	$S+I$	71M, LoRA ft	43.1	34.9	53.5	39.7	48.3	49.1
	CROP$_{Galac1.3B}$	$S+G+I$	71M, LoRA ft	**44.9**	**36.7**	**54.8**	**41.1**	**49.5**	**50.8**
CheBI-20	GPT-4o	S	- , 3-shot	21.9	9.8	35.9	13.6	26.7	27.6
	Llama3$_{Instruct8B}$	S	- , 3-shot	20.9	9.4	35.8	15.5	28.9	25.4
	MolReGPT$_{GPT4}$	S	RAG, 10-shot	60.7	52.5	63.4	47.6	56.2	61.0
	MolXPT	S	350M, full ft	59.4	50.5	66.0	51.1	59.7	62.6
	BioT5	S	252M, full ft	63.5	55.6	69.2	55.9	63.3	65.6
	MoMu-Large	$S+G$	782M, full ft	59.9	51.5	-	-	59.3	59.7
	GIT-Mol	$S+G$	210M, LoRA ft	35.2	26.3	57.5	48.5	56.0	53.3
	InstructMol	$S+G$	- , LoRA ft	47.5	37.1	56.6	39.4	50.2	50.9
	MolCA$_{Galac1.3B}$	$S+G$	110M, LoRA ft	62.0	53.1	68.1	53.7	61.8	65.1
	CROP$_{Galac1.3B}$	S	71M, LoRA ft	58.4	49.6	67.0	51.3	60.3	62.7
	CROP$_{Galac1.3B}$	$S+G$	71M, LoRA ft	63.8	55.8	69.0	55.0	63.5	66.1
	CROP$_{Galac1.3B}$	$S+I$	71M, LoRA ft	62.8	54.2	68.9	54.7	62.8	65.8
	CROP$_{Galac1.3B}$	$S+G+I$	71M, LoRA ft	**64.6**	**56.2**	**69.8**	**55.9**	**63.9**	**66.7**

Specifically, we sum b groups of $Z_j \in \mathbb{R}^{w \times d}$ with the weighted vector $P_i \in \mathbb{R}^b$ to obtain the prefixes $\hat{Z}_i \in \mathbb{R}^{w \times d}$ for Layer$_i$ in the LLM's upper segment, as shown in Figure 4. The calculation is formalized as follows:

$$V = \mathrm{Concat}(\mathrm{Mean}(Z_1); \dots; \mathrm{Mean}(Z_b)), \tag{7}$$

$$P_i = \mathrm{Softmax}(\mathrm{Linear}_i(\sigma(V))), \tag{8}$$

$$\hat{Z}_i = P_i Z. \tag{9}$$

where $Z_{j=1,2,\dots,b} \in \mathbb{R}^{w \times d}$ indicates the jth group of embedding in Z. Mean$(\cdot)$ denotes mean pooling along the sequence dimension, yielding Mean$(Z_j) \in \mathbb{R}^d$. Concat$(\cdot)$ concatenates embeddings along the hidden dimension, producing $V \in \mathbb{R}^{bd}$. $\sigma(\cdot)$ is the activation function. Linear$_i(\cdot)$ has an input dimension of bd and an output dimension of b, thus $P_i \in \mathbb{R}^b$.

SEG can flexibly convert any b groups of structural embeddings into specified u groups of cross-view prefixes, thus the LLM can be partitioned arbitrarily, where the lower segment is unimodal and the upper segment is multimodal.

3.3 Training Strategies

As Galactica is primarily pretrained on SMILES, we adopt SMILES as the molecular sequence view in this research to fully stimulate Galactica's prior chemical knowledge. The training process consists of two stages. In the pretraining stage (Stage 1), CROP performs the molecule captioning task conditioned on molecular SMILES, graphs, and images. Molecular graphs and images could be obtained using the RDKit toolkit according to SMILES. Common data augmentation techniques for images are applied to enhance the LLM's ability to leverage the structural information in molecular images. The primary objective of this stage is to establish initial alignment between multiple views and the LLM. Therefore, we freeze the molecular graph encoder, image encoder and the LLM, and focus on training the bridging modules, including the SMILES guidance, SGR and SEG. In the fine-tuning stage (Stage 2), to pursue optimal performance on downstream tasks, in addition to the aforementioned bridging modules, we unfreeze the molecular graph encoder and image encoder and utilize LoRA [13] to fine-tune the LLM.

To conduct a fair comparison with baselines and to analyze the contributions of molecular graphs and images respectively, we develop four variants of CROP by selectively including different structural representations alongside SMILES. Specifically, by masking graph embeddings Z_G or image embeddings Z_I, we develop CROP$_{(S+G+I)}$, CROP$_{(S+G)}$ and CROP$_{(S+I)}$. When all structural views are excluded, CROP$_{(S)}$ reduces to the original Galactica operating only on SMILES input.

4 Experiments

4.1 Experimental Setup

Datasets. We pretrain CROP on PubChem324k's pretraining subset [24], which contains about 300k molecule-text pairs of relatively low-quality. For the molecule captioning task, we evaluate

Table 2: IUPAC name prediction results on PubChem324k (Baseline results are from [24]).

Model	Modalities	TrainableParams	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
GPT-4o [26]	S	- , 3-shot	42.6	26.1	40.5	13.4	32.6	41.1
Llama3$_{Instruct8B}$ [4]	S	- , 3-shot	31.9	16.0	28.9	5.0	22.3	26.8
BioT5 [27]	S	252M, full ft	79.4	72.6	75.4	55.7	69.5	75.8
GIT-Mol [22]	$S+G$	210M, LoRA ft	58.3	51.7	54.5	32.6	50.2	55.7
MolCA$_{Galac1.3B}$ [24]	$S+G$	100M, LoRA ft	75.0	66.6	69.6	48.2	63.4	72.1
CROP$_{Galac1.3B}$	S	71M, LoRA ft	74.5	65.9	68.7	47.3	62.5	71.4
CROP$_{Galac1.3B}$	$S+G$	71M, LoRA ft	80.8	73.2	77.5	57.5	72.0	78.2
CROP$_{Galac1.3B}$	$S+I$	71M, LoRA ft	80.6	72.7	77.1	56.9	71.4	77.9
CROP$_{Galac1.3B}$	$S+G+I$	71M, LoRA ft	**81.5**	**74.3**	**78.5**	**58.6**	**72.9**	**78.8**

Table 3: Molecule property prediction results on 6 datasets in MoleculeNet. The scaffold splits [40] are adopted. Baseline results are from their original papers. ROC-AUC scores are calculated across 5 random seeds.

Model	Modalities	Tox21 ↑	ToxCast ↑	Sider ↑	ClinTox ↑	BBBP ↑	Bace ↑	Mean
KV-PLM [46]	S	72.1±1.0	55.0±1.7	59.8±0.6	-	70.5±0.5	78.5±2.7	67.2
Mole-BERT [35]	G	76.8±0.5	64.3±0.2	62.8±1.1	78.9±3.0	71.9±1.6	80.8±1.4	72.6
MoMu [30]	G	75.6±0.3	63.4±0.5	60.5±0.9	79.9±4.1	70.5±2.0	76.7±2.1	71.1
GIT-Mol [22]	$S+G$	75.9±0.5	**66.8±0.5**	63.4±0.8	88.3±1.2	**73.9±0.6**	81.1±1.5	74.9
MolCA$_{Galac1.3B}$ [24]	$S+G$	77.2±0.5	64.5±0.8	63.0±1.7	89.5±0.7	70.0±0.5	79.8±0.5	74.0
CROP$_{Galac1.3B}$	S	72.6±0.6	58.3±0.7	63.3±1.5	90.8±1.8	71.0±1.5	81.0±1.2	72.8
CROP$_{Galac1.3B}$	$S+G$	76.2±0.4	60.9±0.6	65.1±0.7	93.7±0.9	71.2±1.0	83.5±0.5	75.1
CROP$_{Galac1.3B}$	$S+I$	75.4±0.6	60.7±0.4	66.2±1.2	92.0±1.4	72.2±0.8	82.9±0.8	74.9
CROP$_{Galac1.3B}$	$S+G+I$	**77.5±0.2**	61.4±0.4	**67.3±0.7**	**94.6±0.6**	72.6±0.6	**84.2±0.7**	**76.3**

CROP's performance on the standard PubChem324k and CheBI-20 [5] datasets. For the IUPAC name prediction task, we evaluate CROP's performance on the standard PubChem324k dataset. For the molecule property prediction task, we evaluate CROP's performance on various sub-datasets in MoleculeNet [34]: Tox21, ToxCast, Sider, ClinTox, BBBP, and Bace, with the scaffold splits [40] adopted.

Implementation Details. CROP is pretrained for 20 epochs and finetuned for 100 epochs on all downstream task datasets respectively. The best-performing model on the validation set is selected for testing. To save the context length of the LLM, we experimentally determine the prefix length w as 10. We analyze the performance of CROP under different partition settings and identify the optimal partition for each task. Specifically, we set $b = 12, u = 12$ in the molecule captioning and molecule property prediction tasks, and set $b = 6, u = 18$ in the IUPAC name prediction task.

4.2 Evaluation on Downstream Tasks

Molecule Captioning. This task requires generating a description about the molecule's properties, structures, biological activity, and other characteristics. In the setting with sequence-only input, leading LLMs such as MolT5 [6], MolXPT [25] and BioT5 [27] are included as baselines. Besides, we report the performance of GPT-4o and Llama3-Instruct in a few-shot setting to showcase the current progress of mainstream general-purpose LLMs in the field of chemistry. MolReGPT [18] is based on GPT4 with the retrieval-augmented generation (RAG) technique. In the setting with multi-view input, advanced MLLMs such as MoMu [30], GIT-Mol [22], InstructMol [2], and MolCA [24] are included as baselines. Among them, InstructMol employs the LLaVA architecture [21] and MolCA employs the BLIP-2 [17] architecture.

The results are reported at Table 1. Beyond the original SMILES view, when utilizing both graph and image views additionally, CROP$_{(S+G+I)}$ achieves the best performance. This highlights the importance of integrating multiple structural views to provide more accurate molecular information, which can mitigate the limitations of relying on a single structural view. When only adopt the graph view additionally, CROP$_{(S+G)}$ still outperforms other models by a large margin with fewer trainable parameters, highlighting CROP's architectural superiority. The superiority primarily stems from two aspects. First, CROP introduces SMILES guidance when jointly re-sampling multiple structural views, thereby effectively leveraging Galactica's prior chemical knowledge. Second, the generated cross-view prefixes are prepended to all layers in the upper segment of the LLM, enabling it to fully interact with the molecular structural information. In contrast, models such as InstructMol or MolCA provide structural information from the input embedding layer. Detailed analysis is provided in the Ablation Studies.

IUPAC Name Prediction. The IUPAC systematic names [9] are standardized molecular identifiers that take aspects, such as molecular functional groups and substituents, into account. This

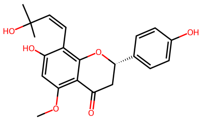
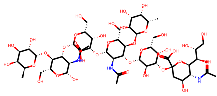
	Ground Truth	CROP($\boldsymbol{S}+\boldsymbol{G}$)	CROP($\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I}$)
	The molecule is a monomethoxyflavanone that is (2S)-flavanone substituted by a methoxy group at position 5, hydroxy groups at positions 7 and 4' and a 3-hydroxy-3-methylbut-1-en-1-yl group at position 8…	The molecule is a monomethoxyflavanone that is (2S)-flavanone substituted by a methoxy group at position 5, a hydroxy group at position 7, a hydroxy group at position 8 and a 3-hydroxy-3-methylbut-1-en-1-yl group at position 6…	The molecule is a monomethoxyflavanone that is (2S)-flavanone substituted by a methoxy group at position 5, hydroxy groups at positions 7 and 4' and a 3-hydroxy-3-methylbut-1-en-1-yl group at position 8…
	The molecule is a branched amino heptasaccharide consisting of a linear sequence of … beta-D-galactosyl and N-acetyl-beta-D-glucosamine residues linked respectively (2->3), (1->3), (1->3) and (1->4), to each N-acetyl-beta-D-glucosamine residue of which is also linked (1->4) an alpha-L-fucosyl residue…	The molecule is a branched amino heptasaccharide consisting of a linear sequence of … beta-D-galactosyl and N-acetyl-beta-D-glucosaminyl residues, linked (2->3), (1->4), (1->3) and (1->3), to each N-acetyl-beta-D-glucosaminyl residue of which is also linked an alpha-L-fucosyl residue…	The molecule is a branched amino heptasaccharide consisting of a linear sequence of … beta-D-galactosyl and N-acetyl-beta-D-glucosamine residues linked respectively (2->3), (1->3), (1->3) and (1->4), to each N-acetyl-beta-D-glucosamine residue of which is also linked an alpha-L-fucosyl residue…

Figure 5: The captions generated by CROP$_{(\boldsymbol{S}+\boldsymbol{G})}$ and CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ on example molecules. CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ provides more accurate descriptions of substituent positions, types, and the connectivity of branched structures in molecules.

task aims at predicting IUPAC name strings from other molecular representations, thus requiring accurately understanding molecular complex structures. CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ outperforms MolCA by 6.5 in BLEU-2 score, as shown in Table 2.

Molecule Property Prediction. This task involves judging a molecule's potential toxicity and other properties based on its structure. Cross-view prefixes and SMILES hidden states from the LLM's last layer are passed through a separate linear head for classification. As shown in Table 3, on six datasets in MoleculeNet, CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ outperforms GIT-Mol by 1.4 ROC-AUC scores on average.

4.3 Ablation Studies

Effectiveness of Integrating Topological and Spatial Structural Views. The structures of molecules can be categorized into two types: topological and spatial structures. However, previous MLLMs primarily focus on the graph view, which emphasizes topological relationships, while neglecting the image view, which excels at capturing molecular spatial configurations. In this section, we highlight the benefits of jointly considering the two types of structural information conveyed in molecular graphs and images respectively. As shown in Table 1, compared to CROP$_{(\boldsymbol{S})}$, CROP$_{(\boldsymbol{S}+\boldsymbol{G})}$ achieves a 6.8 BLEU-2 score improvement on PubChem324k and a 5.4 BLEU-2 score improvement on CheBI-20. Similarly, CROP$_{(\boldsymbol{S}+\boldsymbol{I})}$ improves the BLEU-2 score by 6.5 on PubChem324k and 4.4 on CheBI-20. This demonstrates that incorporating either molecular graphs or images independently can enhance the LLM's comprehension of molecular structures. Furthermore, when both graph and image views are introduced, CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ outperforms CROP$_{(\boldsymbol{S}+\boldsymbol{G})}$ and CROP$_{(\boldsymbol{S}+\boldsymbol{I})}$ by a large margin, demonstrating the limitations of relying solely on a single structural view. For a more detailed analysis, we compare the quality of captions generated by CROP$_{(\boldsymbol{S}+\boldsymbol{G})}$ and CROP$_{(\boldsymbol{S}+\boldsymbol{G}+\boldsymbol{I})}$ for molecules of different complexity, which is measured by the length of SMILES. As illustrated in Figure 6 (Left), molecular images enhance the performance of CROP particularly on more complex molecules.

Some examples are provided in Figure 5, where it can be observed that CROP, by incorporating images, is able to more accurately describe the substituent positions and the spatial interconnectivity of units in complex molecules. Therefore, in this work, we also reveal the effectiveness of molecular images in generative tasks.

Effectiveness of SGR. The effectiveness of SGR stems from its use of SMILES guidance to guide the resampling process, which is derived from the LLM and enriched with prior chemical knowledge. To validate the role of SMILES guidance in boosting the quality of the derived cross-view prefixes, we replace SGR with an alternative resampler, in which the SMILES guidance is substituted with learnable vectors which are randomly initialized. This variant is referred to as CROP$_{\text{w/o SGR}}$. Without the need to obtain SMILES guidance during the forward propagation, the LLM does not need to be partitioned into two segments, allowing us to prepend the cross-view prefixes to all layers of the LLM. We compare the performance of CROP and CROP$_{\text{w/o SGR}}$ when adopting both graph and image views. As shown in Table 4, CROP consistently outperforms CROP$_{\text{w/o SGR}}$ across diverse tasks. In addition, we also find that

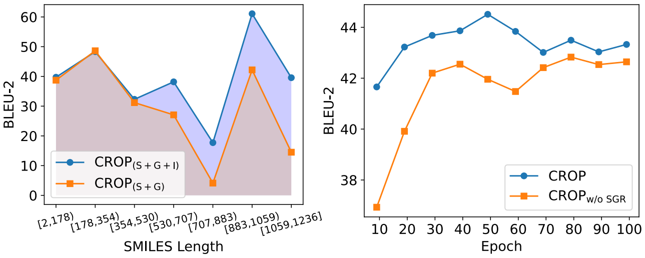


Figure 6: (Left) The distinct BLEU-2 scores in different SMILES length ranges on the PubChem324k molecule captioning dataset. (Right) The distinct metric curves of CROP and CROP$_{\text{w/o SGR}}$ on the PubChem324k molecule captioning validation set during finetuning.

Table 4: Ablation results on the molecule captioning, IUPAC name prediction and molecule property prediction tasks. All experiments are conducted with S, G, I as input simultaneously. The second lines denote CROP$_{\text{w/o SGR}}$ and the third lines denote CROP$_{\text{w/o SEG}}$.

CROP		Molecule Captioning					
SGR	SEG	BLEU$_2$	BLEU$_4$	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	METEOR
✓	✓	**44.9**	**36.7**	**54.8**	**41.1**	**49.5**	**50.8**
	✓	42.4	33.5	52.4	38.9	48.1	48.1
✓		44.2	36.2	54.0	40.7	48.4	50.2
CROP		IUPAC Name Prediction					
SGR	SEG	BLEU$_2$	BLEU$_4$	ROUGE$_1$	ROUGE$_2$	ROUGE$_L$	METEOR
✓	✓	**81.5**	**74.3**	**78.5**	**58.6**	**72.9**	**78.8**
	✓	78.5	71.8	76.4	56.2	71.2	77.3
✓		80.7	73.6	77.9	58.1	72.3	78.4
CROP		Molecule Property Prediction					
SGR	SEG	Tox21	ToxCast	Sider	ClinTox	BBBP	Bace
✓	✓	**77.5**	**61.4**	**67.3**	**94.6**	**72.6**	**84.2**
	✓	75.4	60.2	66.4	92.1	71.6	82.7
✓		76.2	60.8	66.9	94.0	72.1	83.4

CROP converges much faster particularly in the initial stage, and reaches the optimum state earlier than CROP$_{\text{w/o SGR}}$ during the fine-tuning process, as illustrated in Figure 6 (Right). These demonstrate the effectiveness of SGR when utilizing SMILES guidance.

Effectiveness of SEG. SEG enables converting any b groups of structural embeddings $Z \in \mathbb{R}^{b \times w \times d}$ into specified u groups of cross-view prefixes $\hat{Z} \in \mathbb{R}^{u \times w \times d}$, thus allowing the LLM to be partitioned arbitrarily. However, when $b = u$, an alternative is to directly prepend Z into the LLM's upper segment, namely discarding the SEG. In this condition, we compare the performance of CROP and CROP$_{\text{w/o SEG}}$, where Z passes through and bypasses SEG, respectively. As shown in Table 4, CROP outperforms CROP$_{\text{w/o SEG}}$ across diverse tasks. Due to the varying structural features conveyed by each group of Z, SEG combines these groups to generate cross-view prefixes $\hat{Z}$, enabling each prefix to encapsulate comprehensive structural information and thereby enhancing the LLM's ability to understand molecular structures more accurately.

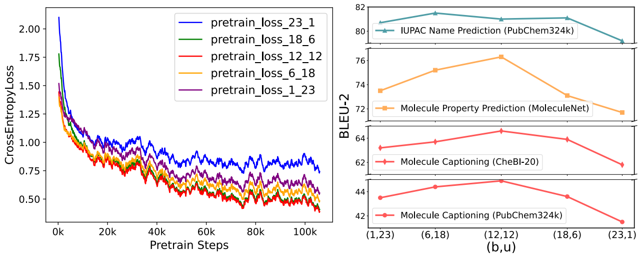


Figure 7: The experimental results of different partition settings for the LLM. (Left) The pre-training loss curves. (Right) The test results of four molecular tasks.

Table 5: The comparison in terms of computational cost across four models on the PubChem324k molecule captioning dataset. The metrics include the average length of cross-view prefixes, average FLOPs for these prefixes and training time on the PubChem324k molecule captioning dataset.

Model	BLEU-2	Avg. Prefixes (Count)	Avg. FLOPs (Billions)	Train. Time (Hours)
Galactica	36.6	-	-	5.37
CROP$_{arch1}$	42.8	288	353.52	7.14
CROP$_{arch2}$	41.5	20	24.20	5.86
CROP$_{arch3}$	**44.9**	**10**	**14.05**	**5.53**

Impact of Different LLM Partitions. Galactica consists of 24 layers in total. We compare five CROP variants with (b, u) set to $(1, 23)$, $(6, 18)$, $(12, 12)$, $(18, 6)$, and $(23, 1)$, respectively. As illustrated in Figure 7, CROP$_{b=12,u=12}$ performs best on the PubChem324k and CheBI-20 molecule captioning datasets. Additionally, CROP$_{b=6,u=18}$ performs best on the PubChem324k IUPAC name prediction dataset. Increasing b provides more groups of SMILES guidance for SGR, facilitating the resampling process. Conversely, increasing u allows more LLM layers to leverage cross-view prefixes, enhancing the utilization of structural information. There is a trade-off between b and u for different tasks, with generally better performance achieved when b and u are nearly balanced.

Efficiency Analysis. We compare the performance of CROP with three different architectures, as shown in Figure 2 (*arch1*, *arch2* and *arch3*), under the setting of inputting $\boldsymbol{S}$, $\boldsymbol{G}$ and $\boldsymbol{I}$ simultaneously. Galactica serves as the common baseline, processing only $\boldsymbol{S}$. The results are shown in Table 5. Compared to Galactica, three variants of CROP achieve significant improvements. Among them, CROP$_{arch3}$ achieves superior performance with shorter cross-view prefixes, fewer additional FLOPs, and smaller training time overhead. Specifically, compared to CROP$_{arch1}$, CROP$_{arch3}$ reduces the length of cross-view prefixes by 96.5%, the additional average number of FLOPs by 96.0% and the training time by 22.5%. This demonstrates the architectural advantages of CROP$_{arch3}$, including utilizing SMILES guidance during the resampling to boost the quality of derived cross-view prefixes, and prepending prefixes to multiple LLM layers to promote structural information utilization.

5 Conclusion

In this work, we identify the fundamental limitations of relying solely on the molecular graph view, and propose CROP, an innovative and scalable MLLM architecture that can integrate both topological and spatial structural views to jointly advance molecular understanding while maintaining computational efficiency. We primarily explore enhancing LLMs by integrating molecular graphs and images, which are representative topological and spatial views respectively, and highlight the impressive effectiveness of molecular images for enhancing the performance of LLMs in generative tasks. In future research, we will consider fine-tuning CROP on large-scale molecular instruction datasets, and integrating more representative molecular views into the CROP.

Acknowledgments

This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA0490000) and the National Natural Science Foundation of China (62276245).

References

[1] A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. 2020. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* 12 (2020), 1–16.

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

[3] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 3438–3445.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

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

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

[7] Ray F Egerton et al. 2005. *Physical principles of electron microscopy*. Vol. 56. Springer.

[8] Junfeng Fang, Shuai Zhang, Chang Wu, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, Xiang Wang, and Xiangnan He. 2024. Moltc: Towards molecular relational modeling in language models. *arXiv preprint arXiv:2402.03781* (2024).

[9] Henri A Favre and Warren H Powell. 2013. *Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013*. Royal Society of Chemistry.

[10] Nikita Fedik, Roman I. Zubatyuk, Maksim Kulichenko, Nicholas Lubbers, Justin S. Smith, Benjamin Tyler Nebgen, Richard A. Messerly, Ying Wai Li, Alexander I. Boldyrev, Kipton Barros, Olexandr Isayev, and Sergei Tretiak. 2022. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry* 6 (2022), 653 – 672. https://api.semanticscholar.org/CorpusID:251771384

[11] Michael Fernandez, Fuqiang Ban, Godwin Woo, Michael Hsing, Takeshi Yamazaki, Eric LeBlanc, Paul S Rennie, William J Welch, and Artem Cherkasov. 2018. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *Journal of chemical information and modeling* 58, 8 (2018), 1533–1543.

[12] Garrett B Goh, Charles Siegel, Abhinav Vishnu, Nathan O Hodas, and Nathan Baker. 2017. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689* (2017).

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

[14] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23802–23804.

[15] Nicolas Keriven. 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems* 35 (2022), 2268–2281.

[16] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* 1, 4 (2020), 045024.

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

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

[19] Zhen Li, Mingjian Jiang, Shuang Wang, and Shugang Zhang. 2022. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today* 27, 12 (2022), 103373.

[20] Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023. DrugChat: towards enabling ChatGPT-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907* (2023).

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023).

[22] Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine* 171 (2024), 108073.

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

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

[25] Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688* (2023).

[26] OpenAI. 2024. Hello GPT-4o.

[27] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276* (2023).

[28] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems* 33 (2020), 12559–12571.

[29] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.

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

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

[32] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. *ICLR (Poster)* 2, 3 (2019), 4.

[33] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28, 1 (1988), 31–36.

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

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

[36] Hongxin Xiang, Shuting Jin, Xiangrong Liu, Xiangxiang Zeng, and Li Zeng. 2023. Chemical structure-aware molecular image representation learning. *Briefings in Bioinformatics* 24, 6 (2023), bbad404.

[37] Liangxu Xie, Lei Xu, Shan Chang, Xiaojun Xu, and Li Meng. 2020. Multitask deep networks with grid featurization achieve improved scoring performance for protein–ligand binding. *Chemical biology & drug design* 96, 3 (2020), 973–983.

[38] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[39] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19323–19331.

[40] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* 59, 8 (2019), 3370–3388.

[41] Kevin Yang, Kyle Swanson, Wengong Jin, Connor W. Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian P. Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, T. Jaakkola, Klavs F. Jensen, and Regina Barzilay. 2019. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* 59 (2019), 3370 – 3388. https://api.semanticscholar.org/CorpusID:198986021

[42] Jiacai Yi, Chengkun Wu, Xiaochen Zhang, Xinyi Xiao, Yanlong Qiu, Wentao Zhao, Tingjun Hou, and Dongsheng Cao. 2022. MICER: a pre-trained encoder–decoder architecture for molecular image captioning. *Bioinformatics* 38, 19 (2022), 4562–4572.

[43] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).

[44] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems* 34 (2021), 28877–28888.

[45] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. 2022. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* 4, 11 (2022), 1004–1016.

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

[47] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2024. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101* (2024).

[48] Shifa Zhong, Jiajie Hu, Xiong Yu, and Huichun Zhang. 2021. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chemical Engineering Journal* 408 (2021), 127998.

[49] Dan-Hao Zhu, Xin-Yu Dai, and Jia-Jun Chen. 2021. Pre-train and learn: Preserving global information for graph neural networks. *Journal of Computer Science and Technology* 36 (2021), 1420–1430.