Multi-modal Knowledge-enhanced Foundation Model for Generation, Retrieval, and Reasoning of Molecules and Text

Delvin Ce Zhang [1] **Menglin Yang** [1] **Hady W. Lauw** [2] **Hua Xu** [1] **Rex Ying** [1]

Abstract

Molecules are usually paired with text descriptions, revealing their chemical functions, properties, and relationships to others. Due to the large amount of molecules, it is important to automatically provide their text descriptions, such as molecule captioning. Given a molecule as query, we study the problem of generating and retrieving its text description in a cross-modal manner. Furthermore, molecules contain rich relations to others in the form of knowledge graphs, which capture auxiliary chemical and biomedical knowledge. However, existing works do not consider knowledge graphs when learning molecules. Knowledge graph contains complex relations among molecules, making the learning more challenging. To solve this problem, we propose a multi-modal foundation model for knowledge-enhanced molecule-to-text generation and retrieval. Specifically, we integrate molecular structure and knowledge graph in a nested layer-wise approach. With the proposed framework, molecule encoding and knowledge aggregation are fused into an iterative workflow. To unify molecular structures and texts in a fine-grained manner, we further propose cross-modal attention to integrate texts into different substructures of molecules. Experiments on generation and retrieval tasks verify the effectiveness of the model.

1. Introduction

Understanding molecules' chemical properties, functions, and their interactions with other molecules are crucial for chemical research. These chemical functions are usually represented in the form of natural language (Edwards et al., 2021). Fig. 1(a) illustrates a molecule with its corresponding text description. In addition to molecular structures, scientific literature contains important biochemical knowledge about molecules, and integrating such knowledge with molecule representation is crucial in understanding the molecules in various biochemical aspects. Thus, it is important to integrate auxiliary knowledge and automatically annotate them with natural language. For a previously unseen molecule, we aim to generate textual captions, while for an existing one, we retrieve its description. In this paper, we study the problem of generating and retrieving text given molecules as queries. Existing works formulate this problem as molecule captioning (Edwards et al., 2022), molecule-based text retrieval (Liu et al., 2023b), and molecule retrieval (Edwards et al., 2021), all of which are beneficial for downstream applications such as drug repurposing (Pushpakom et al., 2019).

Furthermore, knowledge of molecules has also been integrated in the form of knowledge graphs. Fig. 1(b) presents an example chemical egonetwork centered at the molecule acrylamide, extracted from ChEBI[1], a freely available online library containing rich ontologies among diverse molecules. Some chemical keywords that appeared in acrylamide's text description (colored in red) can be found by traversing the knowledge graph, thereby providing additional reliable knowledge from scientific research. Therefore, incorporating knowledge graph as auxiliary information could further improve the quality of molecule-to-text generation and retrieval. However, existing works mainly focus on molecular structure and text (Liu et al., 2023b; Edwards et al., 2022), and ignore the abundant abundant knowledge and connections in knowledge graph.

Challenges. To overcome the limitations of existing molecule-based generation and retrieval works, we must address two challenges as follows.

First, both molecule-based text generation and retrieval models (Edwards et al., 2021; 2022) only input molecular structures as queries and ignore the complementary abundant knowledge in chemical knowledge graph. However, it is an open question how to integrate multi-hop relationships in knowledge graph with molecule, since knowledge graph often contains redundant information and it is challenging to

[1]Yale University, New Haven, US [2]Singapore Management University, Singapore. Correspondence to: Rex Ying <rex.ying@yale.edu>.

Preliminary work. Do not distribute.

[1]https://www.ebi.ac.uk/chebi/

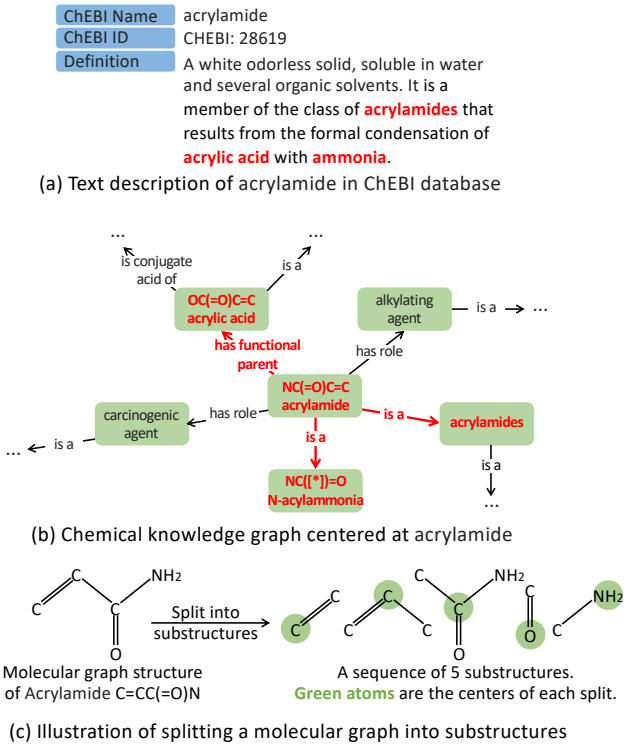


Figure 1. (a) Text description of molecule acrylamide. (b) Chemical knowledge graph centered at acrylamide. Some keywords in text (colored as red) appear in knowledge graph connections, but are difficult to infer only from acrylamide's molecular structure. (c) Illustration of splitting a molecular graph to substructures.

extract useful information and find corresponding evidence in the molecule.

Second, existing works on molecule-based text retrieval, such as MoleculeSTM (Liu et al., 2023b), fuse molecules and texts in a coarse-grained manner, i.e., encoding the entire molecular structure and text separately, then fusing them only at the output layer. Consequently, they cannot model the importance of each chemical keyword to different molecular substructures in a fine-grained manner. Though there exit some works (Zhang et al., 2022; Yasunaga et al., 2022) fusing knowledge graph and text in a layer-wise approach, they are not applicable for molecules as inputs, since molecules are graph structures while their input is text.

Approach. Motivated by above challenges, we design a foundation model MKMT, for **M**ulti-modal **K**nowledge enhanced **M**olecule-to-**T**ext generation and retrieval, which can be applied to different tasks with slight adjustments.

First, to capture high-order relations in knowledge graph, we reason over linked entities and inject the learned entity embeddings into each molecule encoding layer. We propose two aggregators with cross-attention to allow high-order knowledge to propagate to molecules and extract useful information from knowledge graph.

Second, for better integration of molecular structure and texts, we design cross-modal attention to fuse them at the word and molecular substructure level. Thus, different substructures could focus on their own chemical keywords and phrases in a fine-grained manner.

Contributions. Our contributions are as follows. *First*, we propose a multi-modal foundation model MKMT, which automatically generates and retrieves text descriptions given molecules as queries. To capture high-order knowledge, we inject entity embeddings into each molecule encoding layer with two different aggregators. *Second*, to unify molecular structures and texts in a fine-grained manner, we propose cross-modal attention to integrate texts into different substructures of molecules. *Third*, we curated two multi-modal datasets. Each molecule is coupled with both its text description and up to four-hop knowledge graph connections. Code has been made available[2]. On both datasets, our model outperforms state-of-the-art baseline by around 10% ROUGE on molecule captioning.

2. Related Work

Molecule representation learning infers low-dimensional molecule embeddings. Some works input SMILES strings[3] to pre-trained language models to obtain molecule embeddings, e.g., MolBERT (Fabian et al., 2020), SMILES-BERT (Wang et al., 2019), ChemBERTa (Chithrananda et al., 2020), SMILES-Transformer (Honda et al., 2019), Molecule-Transformer (Shin et al., 2019). However, these models are not designed with textual descriptions or knowledge graphs. There are models designed with Graph Neural Networks to capture molecular graph structure, such as WLN (Jin et al., 2017), MPNN (Gilmer et al., 2017), MolGNN (Ishida et al., 2021), N-Gram Graph (Liu et al., 2019), etc. These models incorporate molecular graph structure and capture connectivity across atoms, but still ignore text descriptions or knowledge graph across molecules. DMP (Zhu et al., 2021), MM-Deacon (Guo et al., 2022), and EAGCN (Shang et al., 2021) introduce multiple views for molecule pre-training, but do not incorporate texts or chemical knowledge graph.

There are models that incorporate multiple types of data for multi-modal molecule representation learning. MolR (Wang et al.) and FusionRetro (Liu et al., 2023c) use chemical reaction to enhance molecule representations. GraphMVP (Liu et al.) and MoleculeSDE (Liu et al., 2023a) consider both 2D and 3D molecular graph structure. MoleculeSTM (Liu et al., 2023b) models both molecular graph and text

[2] https://www.dropbox.com/scl/fo/czi9d8wquhi4yhto6nkzu/h?rlkey=f6onrkctlyp0we5u8rdhxkvv2&dl=0

[3] SMILES denotes simplified molecular-input line-entry system, e.g., the SMILES string of Ethanol is CH3CH2OH.

descriptions. However, all these models do not incorporate knowledge graph structure.

Molecule-based text generation and retrieval aim to model both molecular structure and texts for generation and retrieval. Text2Mol (Edwards et al., 2021) and MoleculeSTM (Liu et al., 2023b) focus on text retrieval given molecules as queries, while MolT5 (Edwards et al., 2022) is designed for molecule captioning, i.e., text generation. However, all these models do not incorporate chemical knowledge graph to enhance text generation and retrieval. There are models for de novo molecule generation, including MolGPT (Bagal et al., 2021) and T5Chem (Lu & Zhang, 2022). We differentiate our work from them in two aspects. First, they are designed for molecule generation, while we emphasize text generation where molecules are inputs, not output generations. Second, they are not proposed with multi-modal data, while we focus on using knowledge graph to enhance text generation and retrieval. More broadly, information retrieval in chemistry has been researched (Krallinger et al., 2017b). Some works focus on knowledge-enhanced entity, relation, and event extraction from biomedical texts (Lai et al., 2021; Zhang et al., 2021; Li & Ji, 2019; Li et al., 2019). Though they work in biomedical domain, they deal with domain-specific texts but do not model molecular structure, and their knowledge graph is for text data, not for molecules. Some other works focus on molecule retrieval based on graph similarity (Qu et al., 2019) and substructure indexing (Kratochvíl et al., 2018). Chemical text mining (Krallinger et al., 2017a) is also a related area.

3. Problem Formulation and Preliminaries

In this section, we formulate the studied problem and introduce background knowledge.

Notations. We are given a molecule dataset with text descriptions and knowledge graph $\mathcal{S} = \{\mathcal{M}, \mathcal{D}, \mathcal{G}\}$. $\mathcal{M} = \{m_i\}_{i=1}^{M}$ is a set of M molecules. $\mathcal{D} = \{d_i\}_{i=1}^{M}$ is a set of text descriptions where each text d_i corresponds to molecule m_i. The original complete chemical knowledge graph is usually large with millions of nodes, in this paper we consider a subgraph of the original complete graph, i.e., $\mathcal{G} = \{G_i\}_{i=1}^{M}$ contains M subgraphs, or local knowledge graph, where each subgraph $G_i = \{(h, r, t)_j\}_{j=1}^{|G_i|}$ centers at molecule m_i and contains up to its four-hop neighbors. We use triple (h, r, t) to represent a connection where there is an edge with relation $r \in \mathcal{R}$ from head node $h \in \mathcal{V}$ to tail node $t \in \mathcal{V}$. $\mathcal{R}$ is a set of relations, and $\mathcal{V}$ is a set of entity nodes. The neighbors of a head node h are those tail nodes directly connected from h, i.e., $\mathcal{N}(h) = \{t|(h, r, t) \in G_i\}$. A molecule may have different relations to other entities. For example, *(Acrylamide,HasFunctionalParent,Acrylic Acid)* states the fact that molecule *Acrylamide* contains functional group *Acrylic Acid*.

High-order Connectivity. Exploiting high-order connectivity is important for high-quality molecule-based text generation and retrieval. We define P-order connectivity as a multi-relational path $e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} ... \xrightarrow{r_P} e_P$ where $e_{p-1} \xrightarrow{r_p} e_p$ denotes the p-th triple (e_{p-1}, r_p, e_p) on the path. We have $e_0 = m_i$, i.e., the starting entity on the knowledge graph is the current input query molecule m_i.

Molecule Captioning. Given both molecules $\mathcal{M}$ and auxiliary knowledge graph $\mathcal{G}$ as input, we aim to design a knowledge-enhanced language model f_{MolCap} with high-order connectivity, which generates text description $\hat{d}_i$ for each input query molecule m_i.

Molecule-based Text Retrieval. Given molecules $\mathcal{M}$ as well as their corresponding knowledge graph $\mathcal{G}$ and text description $\mathcal{T}$ as input, we similarly propose knowledge-enhanced model f_{TextRetr}, which produces a probability score $\hat{y}_{ij}$ that text d_j could be the description of molecule m_i.

4. Model Architecture

In this section, we discuss the technical details of our proposed model. See Figs. 2–3 for an overview of model architecture. Based on the two tasks formulated in Section 3, we first present how we aggregate high-order connectivity in knowledge graph into molecular structure encoding for molecule captioning task. We then describe how to design a cross-modal attention to unify texts and molecular structures in a fine-grained manner for molecule-based text retrieval.

4.1. Molecule Captioning

4.1.1. Pre-training

Given an input query molecule m_i, we follow previous works (Edwards et al., 2021) and use an existing tool RDKit (Landrum et al., 2016) to split its molecular graph structure into a sequence of S *substructures*, i.e., $m_i = \{m_{i,s}\}_{s=1}^{S}$. Fig. 1(c) illustrates how to split a complete molecule into a sequence of five substructures, which is analogous to splitting a textual sentence into a sequence of words. $m_i = \{m_{i,s}\}_{s=1}^{S}$ is the input to the language model encoder. We use $\mathbf{H}_i^{(l)} = [\mathbf{h}_{i,\text{CLS}}^{(l)}, \mathbf{h}_{i,1}^{(l)}, ..., \mathbf{h}_{i,S}^{(l)}] \in \mathbb{R}^{(S+1)\times K}$ to represent output embeddings of all substructure tokens in m_i after the l-th encoding layer. Here $\mathbf{h}_{i,\text{CLS}}^{(l)}$ is the [CLS] token, and K is the dimension of embeddings.

To pre-train the molecule language encoder with substructures as input, we use ZINC (Sterling & Irwin, 2015), a large set of molecules. Specifically, we randomly select 3 million molecules from ZINC, and use RDKit (Landrum et al., 2016) to split each into a sequence of substructures. We adopt masked language modeling as pre-training objective.

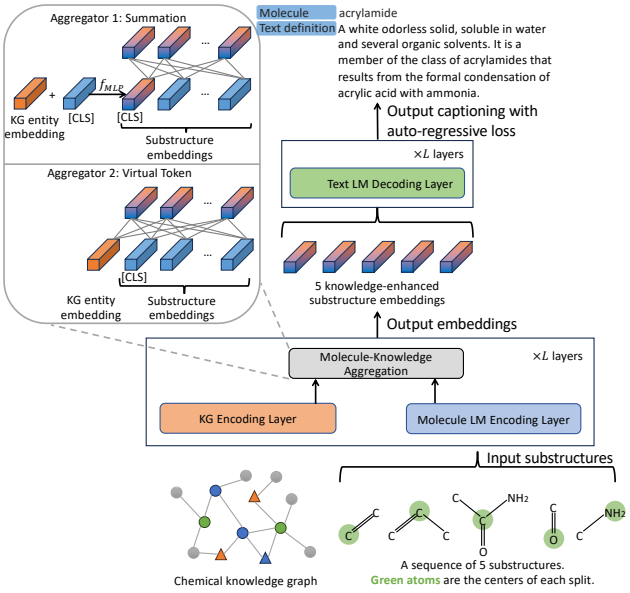


Figure 2. Illustration of molecule captioning. Each molecule is split into a sequence of substructures, which are fed into molecule encoder. Each molecule encoding layer is augmented by knowledge graph with two different aggregators. The augmented substructure embeddings are then input to the decoding layer for molecule captioning.

4.1.2. Fine-tuning

High-order Knowledge Propagation. To capture high-order connectivity on knowledge graph, we first use knowledge graph neural network (KGNN) to infer entity embedding of the input query molecule by

$$\mathbf{e}_i = \text{KGNN}(m_i, G_i). \tag{1}$$

Specifically, to make presentation clearer, we briefly describe the architecture of KGNN. Given an entity h, we capture its first-order connectivity by

$$\mathbf{e}^{(z)}_{\mathcal{N}(h)} = \sum_{(h,r,t)\in\mathcal{N}(h)} \alpha(h,r,t)\mathbf{e}^{(z)}_t. \tag{2}$$

Here superscript $(^{(z)})$ denotes the z-th graph convolutional layer. $\alpha(h,r,t)$ is attention score, representing the importance of neighbor t conditioned on relation r. We compute $\alpha(h,r,t)$ by a knowledge-aware attention below.

$$\begin{aligned}\tilde{\alpha}(h,r,t) &= (\mathbf{W}^{(z)}_r \mathbf{e}^{(z)}_t)^\top \tanh(\mathbf{W}^{(z)}_r \mathbf{e}^{(z)}_h + \mathbf{e}_r),\\ \alpha(h,r,t) &= \frac{\exp(\tilde{\alpha}(h,r,t))}{\sum_{(h,r',t')\in\mathcal{N}(h)}\exp(\tilde{\alpha}(h,r',t'))}\end{aligned} \tag{3}$$

Finally, we aggregate neighbor embedding $\mathbf{e}^{(z)}_{\mathcal{N}(h)}$ and current entity embedding $\mathbf{e}^{(z)}_h$ by

$$\mathbf{e}^{(z+1)}_h = \text{LeakyReLU}\big(\mathbf{W}^{(z)}(\mathbf{e}^{(z)}_h + \mathbf{e}^{(z)}_{\mathcal{N}(h)})\big). \tag{4}$$

We repeat Eqs. 2–4 for a maximum of Z layers and obtain $\mathbf{e}_i = \mathbf{e}^{(Z)}_i$, i.e., the entity embedding of the input query molecule m_i.

Molecule-Knowledge Aggregation. Since our goal is to use chemical knowledge graph to enhance molecule captioning, we aim to aggregate entity embedding $\mathbf{e}_i$ and molecular substructure embedding $\mathbf{H}^{(l)}_i$ at each encoding layer. Here we propose two different aggregators.

1. Summation Aggregator. We add entity embedding learned from knowledge graph to the [CLS] token embedding and apply a non-linear transformation as follows.

$$\tilde{\mathbf{h}}^{(l)}_{i,\text{CLS}} = \mathbf{W}_2\Big(\text{LeakyReLU}\big(\mathbf{W}_1(\mathbf{h}^{(l)}_{i,\text{CLS}} + \mathbf{e}_i)\big)\Big). \tag{5}$$

Since language model encoder learns contextualized token embeddings, the entity embedding added to [CLS] token will be propagated to other tokens to fully integrate high-order knowledge into molecular structure. The aggregated [CLS] embedding $\tilde{\mathbf{h}}^{(l)}_{i,\text{CLS}}$, together with other tokens $\tilde{\mathbf{H}}^{(l)}_i = [\tilde{\mathbf{h}}^{(l)}_{i,\text{CLS}}, \mathbf{h}^{(l)}_{i,1}, ..., \mathbf{h}^{(l)}_{i,S}]$, is passed to the next language model encoding layer. We execute this summation aggregator at every encoding layer to fully integrate high-order knowledge into molecular structures.

2. Virtual Token Aggregator. We consider entity embedding $\mathbf{e}_i$ as one more virtual token and concatenate it with existing $(S+1)$ tokens as follows.

$$\tilde{\mathbf{H}}^{(l)}_i = (\mathbf{e}_i || \mathbf{H}^{(l)}_i) = [\mathbf{e}_i, \mathbf{h}^{(l)}_{i,\text{CLS}}, \mathbf{h}^{(l)}_{i,1}, ..., \mathbf{h}^{(l)}_{i,S}]. \tag{6}$$

Here $(\cdot||\cdot)$ is concatenation operation. After concatenation, $\tilde{\mathbf{H}}^{(l)}_i$ contains information from both molecular substructures and their associated high-order knowledge. $\tilde{\mathbf{H}}^{(l)}_i$ is passed to the next language encoding layer, where we propose *asymmetric* Multi-Head Attention (MHA^{asym}). $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are computed as follows.

$$\mathbf{Q}^{(l)} = \mathbf{H}^{(l)}_i \mathbf{W}^{(l)}_Q, \quad \mathbf{K}^{(l)} = \tilde{\mathbf{H}}^{(l)}_i \mathbf{W}^{(l)}_K, \quad \mathbf{V}^{(l)} = \tilde{\mathbf{H}}^{(l)}_i \mathbf{W}^{(l)}_V. \tag{7}$$

This multi-head attention is asymmetric, where $\mathbf{K}$ and $\mathbf{V}$ are augmented by knowledge graph embedding, but $\mathbf{Q}$ is not. This design has been used by previous works (Yang et al., 2021; Jin et al., 2022) to avoid knowledge graph information being overwritten by molecular structure, and has been shown to be superior to original symmetric multi-head attention (Vaswani et al., 2017). Integrating asymmetric multi-head attention into language encoding, we have

$$\begin{aligned}\mathbf{H}^{(l)'}_i &= \text{LN}(\mathbf{H}^{(l)}_i + \text{MHA}^{asym}(\mathbf{H}^{(l)}_i, \tilde{\mathbf{H}}^{(l)}_i)),\\ \mathbf{H}^{(l+1)}_i &= \text{LN}(\mathbf{H}^{(l)'}_i + \text{MLP}(\mathbf{H}^{(l)'}_i)).\end{aligned} \tag{8}$$

Here $\text{LN}(\cdot)$ is layer normalization, and $\text{MLP}(\cdot)$ is multi-layer perceptron (Bishop & Nasrabadi, 2006). The output

token embeddings $\mathbf{H}_i^{(l+1)} \in \mathbb{R}^{(S+1)\times K}$ have the same dimension as the input sequence $\mathbf{H}_i^{(l)}$. This one-layer encoding propagates knowledge entities to every molecular substructure token. Similarly, we implement this virtual token aggregator with asymmetric MHA to unify high-order knowledge and molecular structures.

Text Generation. Having obtained the fused molecule representations $\mathbf{H}_i^{(L)}$ augmented by auxiliary knowledge graph, we feed them into decoder to generate textual descriptions $\hat{d}_i$ for the input query molecule m_i. Since our decoder is a standard architecture, we do not discuss its specific details here, but represent it as a function $\hat{d}_i = f_{\text{dec}}(\mathbf{H}_i^{(L)})$. We have auto-regressive loss $\mathcal{L}_{\text{AutoReg}}(\hat{d}_i, d_i)$ (Radford et al., 2019; Achiam et al., 2023) between the prediction and the ground-truth text d_i.

Knowledge Graph Embedding. We also aim to preserve knowledge graph structure, so that entity and relation embeddings are effective representations. Here we employ TransR (Lin et al., 2015), a widely used knowledge graph embedding method, though other methods are also applicable. For a given triple (h, r, t), its energy score is $\phi_r(\mathbf{e}_h, \mathbf{e}_t) = ||\mathbf{W}_r\mathbf{e}_h + \mathbf{e}_r - \mathbf{W}_r\mathbf{e}_t||_2^2$. Here $\mathbf{W}_r \in \mathbb{R}^{K\times K}$ is relation-specific transformation matrix. A lower energy score indicates that the triple is more likely to be true, and vice versa. For each input query molecule m_i, the loss function of knowledge graph embedding component is

$$\mathcal{L}_{\text{KGE}}(m_i) = -\log\Big(\sigma\big(-\phi_r(\mathbf{h}_{i,\text{CLS}}^{(L)}, \mathbf{e}_t) + \frac{1}{N}\sum_{(i,r,t')}\phi_r(\mathbf{h}_{i,\text{CLS}}^{(L)}, \mathbf{e}_{t'})\big)\Big) \tag{9}$$

Here we consider the embedding of [CLS] token as the final representation of the input molecule. N is the number of negative samples, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is sigmoid function.

Combining text generation and knowledge graph embedding, we have the final loss where λ_{KGE} is a hyperparameter.

$$\mathcal{L}_{\text{MolCap}} = \sum_{m_i\in\mathcal{M}} \mathcal{L}_{\text{AutoReg}}(\hat{d}_i, d_i) + \lambda_{\text{KGE}}\mathcal{L}_{\text{KGE}}(m_i). \tag{10}$$

4.2. Molecule-based Text Retrieval

In this subsection, we discuss molecule-based text retrieval, which uses above knowledge-enhanced molecule encoder as building block. With slight adjustments, we here predict a probability score for the input molecule-text pair.

Molecule-Text Aggregation. As in molecule captioning, we input a molecule with its substructures $m_i = \{m_{i,s}\}_{s=1}^S$ to the encoder, and apply either summation or virtual token aggregator to incorporate high-order knowledge. In addition, specifically for molecule-based text retrieval task, we further input m_i's text description $d_i = \{d_{i,w}\}_{w=1}^W$ with a sequence of W words to a separate language model encoder. We use $\mathbf{Z}_i^{(l)} = [\mathbf{z}_{i,\text{CLS}}^{(l)}, \mathbf{z}_{i,1}^{(l)}, ..., \mathbf{z}_{i,W}^{(l)}] \in \mathbb{R}^{(W+1)\times K}$ to represent output embeddings of all word tokens after the l-th encoding layer. To unify molecular structures and texts in a fine-grained manner, we propose cross-modal attention. For each substructure embedding $\mathbf{h}_{i,s}^{(l)}$ where $s = \text{CLS}, 1, ..., S$,

$$\begin{aligned}\tilde{\beta}(s,w) &= \text{LeakyReLU}(\mathbf{b}^\top(\mathbf{h}_{i,s}^{(l)}||\mathbf{z}_{i,w}^{(l)})),\\ \beta(s,w) &= \frac{\exp(\tilde{\beta}(s,w))}{\sum_{w'=1}^W \exp(\tilde{\beta}(s,w'))}.\end{aligned} \tag{11}$$

$\mathbf{b} \in \mathbb{R}^{2K}$ is learnable parameter. We have attention score between a substructure and each word, so that we can model the importance of different chemical keywords to a certain substructure in a fine-grained manner. Finally, we aggregate word tokens and combine them with molecular substructures.

$$\tilde{\mathbf{h}}_{i,s}^{(l)} = \mathbf{W}_2'\Big(\text{LeakyReLU}\big(\mathbf{W}_1'(\mathbf{h}_{i,s}^{(l)} + \sum_w \beta(s,w)\mathbf{z}_{i,w}^{(l)})\big)\Big). \tag{12}$$

We repeat this process for every molecular substructure, and obtain the aggregated embedding $\tilde{\mathbf{H}}_i^{(l)}$. Since we still use knowledge graph to enhance text retrieval task, now $\tilde{\mathbf{H}}_i^{(l)}$ contains molecular structure, high-order knowledge, and text description. We input $\tilde{\mathbf{H}}_i^{(l)}$ to the next molecule encoding layer. Thus, we inject both knowledge graph and text description into molecule encoder in a layer-wise manner.

After a maximum of L encoding layers, we take [CLS] token embeddings $\mathbf{h}_i = \mathbf{h}_{i,\text{CLS}}^{(L)}$ and $\mathbf{z}_i = \mathbf{z}_{i,\text{CLS}}^{(L)}$ as the final molecule and text representations, respectively. Their matching score is $p_{ii} = \mathbf{h}_i^\top\mathbf{z}_i$. Finally, we have

$$\mathcal{L}_{\text{TextRetr}} = \sum_{m_i\in\mathcal{M}} \lambda_{\text{KGE}}\mathcal{L}_{\text{KGE}}(m_i) - \log\frac{\exp(p_{ii})}{\exp(p_{ii}) + \sum_j \exp(p_{ij})} \tag{13}$$

Here j denotes negative samples. In our implementation, we use in-batch negative sampling (Liu et al.) to save computational cost. We also add knowledge graph embedding loss (Eq. 9) to preserve knowledge graph structure. After training, for molecule-text pair (m_i, d_j), we are able to infer their matching score $p(d_j|m_i) \propto \exp(\mathbf{h}_i^\top\mathbf{z}_j)$. For a molecule m_i, we predict its most plausible text description by $\arg\max_{d_j\in\mathcal{D}} p(d_j|m_i)$.

5. Experiments

The goal of experiments is to evaluate the performance of molecule captioning and molecule-based text retrieval.

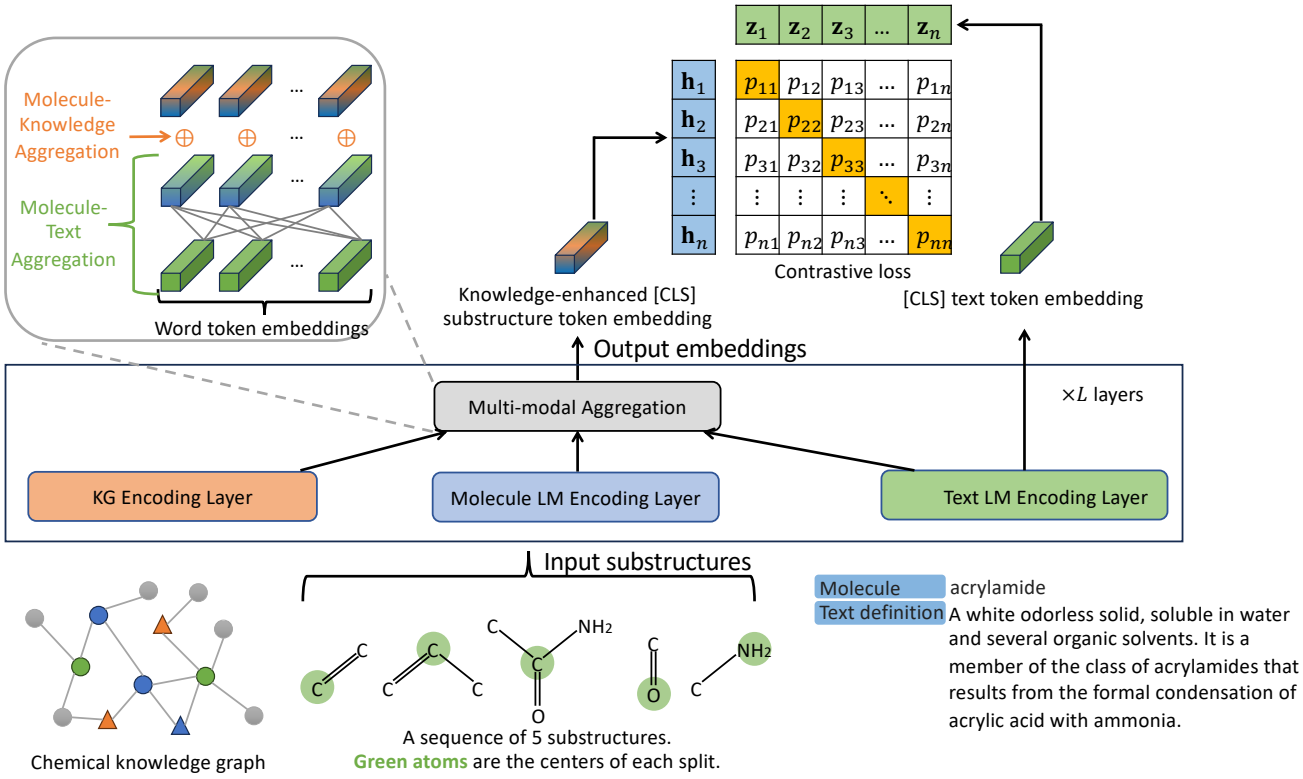


Figure 3. Illustration of molecule-based text retrieval. We input molecules, chemical knowledge graph, and text descriptions into the model. After molecule-text and molecule-knowledge aggregations, we obtain final molecule representations, which are compared with text representations, with a contrastive loss. On the contrary, molecule captioning uses an auto-regressive loss.

Datasets. We create two multi-modal datasets ChEBIKG and PubChemKG with molecules, text descriptions, and knowledge graphs. ChEBI[4] is a publicly available online library containing molecules, texts, and ontologies. Specifically, we create a dataset containing 139,572 molecules with corresponding text descriptions. For each molecule, we further extract up to 4-hop ontologies as its knowledge graph structure. In total, we have 150,414 entities and 11 relations on knowledge graph. PubChem[5] is another public library. We similarly create another dataset with three modalities. In total, we have 7,814 molecules with texts, as well as 13,703 entities and 11 relations on knowledge graph. See supplementary material for data preprocessing details.

Implementation Details. We set the number of knowledge graph convolutional layers to 2. The number of negative samples N for knowledge graph embedding loss is 10. Knowledge graph regularizer λ_{KGE} is 1. The number of molecular encoding layers L is 12. We train the model for 50 epochs with learning rate of 10^{-5}, minibatch size of 16, and Adam optimizer (Kingma & Ba, 2014).

[4] https://www.ebi.ac.uk/chebi/
[5] https://pubchem.ncbi.nlm.nih.gov/

5.1. Molecule Captioning

In this subsection, we evaluate the performance of molecule captioning on evaluative metrics. We differ the discussion on molecule-based text retrieval to the next subsection.

Evaluation Protocol. For each dataset, we split 80% molecules with texts and knowledge graphs as training data (among which 10% are further reserved for validation). The remaining 20% are testing data. After optimization, we input testing molecules and their knowledge graphs to infer the generated text descriptions, and compare them against the corresponding ground-truth ones. We use BLUE-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and Meteor as evaluation metrics.

Baselines. We follow previous works (Edwards et al., 2022) and consider RNN (Cho et al., 2014), Transformer (Vaswani et al., 2017), and the recently proposed MolT5 (Edwards et al., 2022) as baselines. We also include GPT4 (Achiam et al., 2023). We further remove knowledge graph from our model, MKMT (w/o KG). For our model, we have MKMT-Sum (summation aggregator) and MKMT-VT (virtual token aggregator).

Table 1. Experimental results on molecule captioning.

Model	ChEBIKG						PubChemKG					
	BLUE-2	BLUE-4	ROUGE-1	ROUGE-2	ROUGE-L	Meteor	BLUE-2	BLUE-4	ROUGE-1	ROUGE-2	ROUGE-L	Meteor
RNN	0.3235	0.2339	0.4310	0.3271	0.4264	0.4572	0.2085	0.1464	0.4571	0.3783	0.4426	0.4267
Transformer	0.0828	0.0397	0.1706	0.1048	0.1625	0.2341	0.1339	0.1001	0.2237	0.1579	0.2138	0.2814
MolT5	0.3072	0.2192	0.3943	0.3191	0.4123	0.4612	0.1903	0.1252	0.4265	0.3586	0.4101	0.4810
GPT4	0.1220	0.1056	0.1906	0.1366	0.1656	0.2346	0.1378	0.1045	0.2655	0.1839	0.2103	0.2965
MKMT (w/o KG)	0.3122	0.2126	0.5029	0.4338	0.5020	0.5159	0.2402	0.1864	0.5208	0.4507	0.5191	0.5175
MKMT-Sum	0.2948	0.2013	0.5394	0.4507	0.5392	0.5182	0.2603	0.2000	0.5301	0.4566	0.5388	0.5261
MKMT-VT	**0.3269**	**0.2384**	**0.5538**	**0.4705**	**0.5538**	**0.5200**	**0.2851**	**0.2134**	**0.5469**	**0.4729**	**0.5447**	**0.5251**

Table 2. Experimental results on molecule-based text retrieval.

Model	ChEBIKG						PubChemKG					
	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10
Text2Mol	50.8654	0.0530	0.0121	0.0302	0.0502	0.1012	50.4440	0.0509	0.0102	0.0288	0.0467	0.1017
MoleculeSTM	4.7351	0.5470	0.4286	0.5324	0.6463	0.8104	4.5048	0.6120	0.4792	0.6814	0.7384	0.8620
KV-PLM	16.3520	0.3204	0.2095	0.3258	0.3974	0.5371	18.0805	0.3946	0.2469	0.3885	0.4913	0.5536
MKMT (w/o KG)	3.8111	0.6287	0.4861	0.7133	0.8155	0.9245	4.8401	0.5966	0.4575	0.6782	0.7697	0.8848
MKMT-Sum	3.3953	0.6566	0.5193	0.7412	0.8387	0.9388	**4.2956**	0.6417	0.5144	0.7172	**0.8113**	**0.9066**
MKMT-VT	**3.2317**	**0.6737**	**0.5387**	**0.7607**	**0.8544**	**0.9466**	4.3650	**0.6488**	**0.5247**	**0.7224**	0.8062	0.9022

Result. The result is shown at Tables 1. Among our models, the virtual token aggregator tends to outperform summation aggregator, potentially because virtual token aggregator considers entity embedding as one more token, and the multi-head attention helps differentiate the importance of entity embedding to different substructure tokens, while summation aggregator directly adds entity embedding to [CLS] token and does not have attention. MKMT (w/o KG) correspond to our model with no knowledge graph structure. Our models outperform it, which verifies that knowledge graph indeed brings useful information to enhance molecule captioning.

5.2. Molecule-based Text Retrieval

Evaluation Protocol. We split the datasets the same as molecule captioning. After optimization, we infer embeddings for testing molecules. For evaluation, for each molecule, in addition to its corresponding ground-truth text description, we randomly sample 99 negative texts. Thus in total we have 100 texts as candidates. We compare the similarity between each testing molecule embedding and 100 text embeddings, and produce a ranking for texts. We following existing works (Wang et al.) and use MR, MRR, Hit Rate (@ 1, 3, 5, 10) as evaluation metrics.

Baselines. We consider Text2Mol (Edwards et al., 2021), MoleculeSTM (Liu et al., 2023b), KV-PLM (Zeng et al., 2022) as baselines. In addition, we further remove knowledge graph from our model, MKMT (w/o KG).

Result. Tables 2 summarize the results on two datasets. MoleculeSTM produces satisfying results among baseline models, since its molecule encoder is based on pre-trained 3D molecular graph structure. After removing knowledge graph structure from our models, the performance drops, which again verifies the usefulness of high-order knowledge. Similarly, between our models, virtual token aggregator tends to outperform summation aggregator.

Molecule Retrieval. For completeness, here we also conduct molecule retrieval given textual description as query. We compare to the same set of baseline models and use the same evaluation metrics. Results are shown in Table 3. Similarly, our models consistently outperform baseline models, due to the advantage of knowledge graph. Virtual token aggregator performs better than summation aggregator, since virtual token aggregator applies multi-head attention and differentiates the importance of different substructures.

5.3. Model Analysis

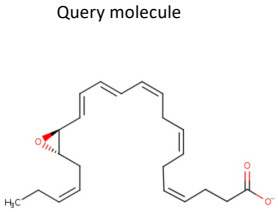
In this section, we perform case studies and ablation analyses to gain a better understanding of our model.

Case Study. To better understand what captions our model generates for input molecules, here we conduct case study in Fig. 4(a). Here, we show an illustrative query molecules. Compared to ground-truth text, MKMT-VT generally produces more accurate captions than MKMT-Sum. MolT5 does not produce as accurate captions as our models. For example, for the first molecule, both our models correctly describe it as an anion, while MolT5 does not have this keyword. In Fig. 4(b) we show case study on molecule-based text retrieval. Our model correctly retrieves the correct text at the first place, while MoleculeSTM puts it on the third place.

Different Number of Convolutional Layers. We analyze the performance of our model with different number of graph convolutional layers Z. The results are summarized in Fig. 5(a). When $Z = 1$, we cannot fully capture high-order

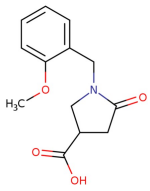
Table 3. Experimental results on molecule retrieval.

Model	ChEBIKG						PubChemKG					
	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10	MR	MRR	Hits @ 1	Hits @ 3	Hits @ 5	Hits @ 10
Text2Mol	50.8398	0.0529	0.0116	0.0302	0.0494	0.1011	50.4779	0.0501	0.0090	0.0250	0.0480	0.1030
MoleculeSTM	38.5910	0.0703	0.0194	0.0521	0.0836	0.1842	40.5038	0.0869	0.0295	0.0810	0.1194	0.2104
KV-PLM	46.7102	0.0693	0.0184	0.0392	0.0599	0.1400	46.0912	0.0692	0.0135	0.0498	0.0843	0.1623
MKMT (w/o KG)	35.8724	0.0894	0.0229	0.0628	0.1029	0.2049	39.0526	0.1085	0.0325	0.0971	0.1310	0.2306
MKMT-Sum	**31.3744**	0.0940	0.0248	0.0685	0.1129	0.2147	37.0940	0.1135	**0.0429**	0.1030	0.1472	0.2521
MKMT-VT	33.9591	**0.0942**	**0.0258**	**0.0754**	**0.1176**	**0.2199**	**32.6321**	**0.1189**	0.0384	**0.1075**	**0.1542**	**0.2655**



Query molecule	Ground-truth text	MKMT-Sum	MKMT-VT	MolT5
	This molecule is a docosanoid anion that is the conjugate base of (16S,17S)-epoxy-(4Z,7Z,10Z,12E,14E,19Z)-docosahexaenoic acid, obtained by deprotonation of the carboxy group; major species at pH 7.3.	This molecule is a polyunsaturated fatty acid anion that is the conjugate base of (5Z,8Z,11Z,14Z,16Z,19Z)-docosahexaenoic acid, obtained by deprotonation of the carboxy group; major species at pH 7.3.	This molecule is a docosanoid anion that is the conjugate base of (4Z,7Z,10Z,13Z,16Z,19Z)-docosapentaenoic acid, obtained by deprotonation of the carboxy group; major species at pH 7.3.	The molecule is a dihydroxydocosahexaenoate that is the conjugate base of (4Z,7Z,10Z,12E,14S,16Z,19Z,21R)-dihydroxydocosahexaenoic acid, obtained by deprotonation of the carboxy group; major species at pH 7.3.

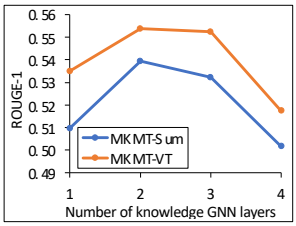
(a) Case study on molecule captioning where correct captionings are colored as green, and wrong captionings are colored as red.



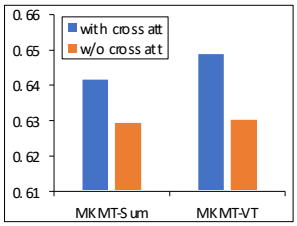
Query molecule	Retrieved text descriptions by KMKT-VT	Retrieved text descriptions by MoleculeSTM
	1. This molecule is an organosulfur compound (√). 2. this molecule is a substituted aniline and an aromatic ether (×). 3. this molecule is a carbonyl compound (×). 4. this molecule appears as white amorphous lumps or a crystalline mass with a faint odor of bitter almonds (×). 5. this molecule is an alpha - substituted cyanoacetate ester and an ethyl ester (×).	1. This molecule is an organic hydroxy compound (×). 2. this molecule is a primary amine (×). 3. This molecule is an organosulfur compound (√). 4. this molecule is a substituted aniline and an aromatic ether (×). 5. this molecule is a carbonyl compound (×).

(b) Case study on molecule-based text retrieval.

Figure 4. Case study on (a) molecule captioning and (b) molecule-based text retrieval.



(a) Number of knowledge GNN layers



(b) Effect of cross-modal attention

Figure 5. Ablation analysis of our model on PubChemKG dataset.

knowledge, leading to inferior results. When we increase Z to 2, we discover an improvement, since more neighbors on knowledge graph are incorporated to enhance the performance. However, an overly high value of Z hurts the result, since further neighbors with noise are modeled.

Effect of Cross-modal Attention. One key component in our model is the modeling of fine-grained integration between molecular structure and text by cross-modal attention. To investigate its effect on model performance, we remove it from the complete model and report the results at Fig. 5(b). We discover that after removing the cross-model attention, the performance drops, since we cannot model the importance between each substructure and chemical keyword, resulting in a coarse-grained integration.

6. Conclusion

We propose a multi-modal knowledge-enhanced model for molecule-based generation and retrieval. To incorporate knowledge graph, we propose two aggregators to integrate it with molecules in a layer-wise manner. Experiments verify the effectiveness of the proposed model.

7. Broader Impact

This paper presents work whose goal is to advance the field of molecule representation learning. A potential societal impact of our work is to help discover chemical properties and functions of newly discovered molecules. We assume the texts of molecules are correct. If some texts provide misleading information, they may influence the training.

References

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J. R., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A. A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D. P., Mu, T., Murati, M., Murk, O., M'ely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Long, O., O'Keefe, C., Pachocki, J. W., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Pokorny, M., Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M. D., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B. D., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N. A., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Bagal, V., Aggarwal, R., Vinod, P., and Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *Machine Learning for Molecules Workshop at NeurIPS 2020*, 2020.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724. Association for Computational Linguistics, 2014.

Edwards, C., Zhai, C., and Ji, H. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, 2021.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022.

Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Guo, Z., Sharma, P., Martinez, A., Du, L., and Abraham, R. Multilingual molecular representation learning via contrastive pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3441–3453, 2022.

Honda, S., Shi, S., and Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

Ishida, S., Miyazaki, T., Sugaya, Y., and Omachi, S. Graph neural networks with multiple feature extraction paths for chemical property estimation. *Molecules*, 26(11):3125, 2021.

Jin, B., Zhang, Y., Meng, Y., and Han, J. Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks. In *The Eleventh International Conference on Learning Representations*, 2022.

Jin, W., Coley, C., Barzilay, R., and Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., and Valencia, A. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761, 2017a.

Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., and Valencia, A. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761, 2017b.

Kratochvíl, M., Vondrášek, J., and Galgonek, J. Sachem: a chemical cartridge for high-performance substructure search. *Journal of cheminformatics*, 10(1):1–11, 2018.

Lai, T., Ji, H., Zhai, C., and Tran, Q. H. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6248–6260, 2021.

Landrum, G. et al. Rdkit: Open-source cheminformatics software, 2016. *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit*, 149(150):650, 2016.

Li, D. and Ji, H. Syntax-aware multi-task graph convolutional networks for biomedical relation extraction. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pp. 28–33, 2019.

Li, D., Huang, L., Ji, H., and Han, J. Biomedical event extraction based on knowledge-driven tree-lstm. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1421–1430, 2019.

Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*.

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

Liu, S., Du, W., Ma, Z.-M., Guo, H., and Tang, J. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pp. 21497–21526. PMLR, 2023a.

Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, Dec 2023b. ISSN 2522-5839. doi: 10.1038/s42256-023-00759-6. URL https://doi.org/10.1038/s42256-023-00759-6.

Liu, S., Tu, Z., Xu, M., Zhang, Z., Lin, L., Ying, R., Tang, J., Zhao, P., and Wu, D. Fusionretro: molecule representation fusion via in-context learning for retrosynthetic planning. In *International Conference on Machine Learning*, pp. 22028–22041. PMLR, 2023c.

Lu, J. and Zhang, Y. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*, 62(6):1376–1387, 2022.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

Qu, J., Sun, P., Li, X., Wang, B., Lu, X., Tang, Z., and Zhang, C. A retrieval system of medicine molecules based on graph similarity. *IEEE MultiMedia*, 26(4):17–27, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Shang, C., Liu, Q., Tong, Q., Sun, J., Song, M., and Bi, J. Multi-view spectral graph convolution with consistent edge attention for molecular modeling. *Neurocomputing*, 445:12–25, 2021.

Shin, B., Park, S., Kang, K., and Ho, J. C. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, pp. 230–248. PMLR, 2019.

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Li, W., Jin, X., Cho, K., Ji, H., Han, J., and Burke, M. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smilesbert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pp. 429–436, 2019.

Yang, J., Liu, Z., Xiao, S., Li, C., Lian, D., Agrawal, S., Singh, A., Sun, G., and Xie, X. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810, 2021.

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P. S., and Leskovec, J. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323, 2022.

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

Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C., and Leskovec, J. Greaselm: Graph reasoning enhanced language models for question answering. In *International Conference on Representation Learning (ICLR)*, 2022.

Zhang, Z., Parulian, N. N., Ji, H., Elsayed, A. S., Myers, S., and Palmer, M. Fine-grained information extraction from biomedical literature based on knowledge-enriched abstract meaning representation. In *Proc. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. Dual-view molecule pre-training. *arXiv preprint arXiv:2106.10234*, 2021.

A. Dataset Preprocessing

Here we describe the details of dataset preprocessing.

ChEBI[6] is a publicly available online library containing molecules, text descriptions, and ontologies. Specifically, it provides molecule-text pair on its website. In total, it has 141 webpages with each page containing around 1,000 molecule-text pair. An example page (the first page) can be found here. In each webpage, there is a "URL" field, containing the URL of the current molecule in ChEBI library. We use this URL to obtain the SMILES string of the current molecule in ChEBI library. There is another "String" field in the webpage, containing the text description of the molecule. We process each webpage and obtain totally 139,572 molecule-text pairs. In order to obtain knowledge graph structure, for each molecule we search its profile using its URL in ChEBI library. For each molecule profile in ChEBI library, there is a "ChEBI Ontology" field, representing the connections between the current molecule and others. Specifically, we crawl the outgoing ontologies up to 4 hops for each molecule. Finally, we obtain knowledge graph structure with 150,414v entities and 11 relations.

PubChem[7] is another publicly available molecule library. Since existing work (Liu et al., 2023b) has already created a dataset with molecule-text pairs, we directly use their preprocessed dataset. However, their dataset does not have knowledge graph structure, and PubChem library does not have molecule ontologies as well. Thus, for each molecule we search its SMILES string in ChEBI library to obtain its profile. For those molecules with a valid ChEBI profile, we follow the same steps as above to crawl their knowledge graph structure. In total, we have 7,814 molecule-text pairs, 13,703 entities and 11 relations on knowledge graph.

Since the text descriptions of molecules start with the molecule name, for both datasets, we follow previous work (Edwards et al., 2022) and replace the molecule name with "This molecule".

B. GPT-4 Input Prompt

Here we present the input to GPT-4 for molecule captioning task. We aim to input both molecule SMILES strings and knowledge graph structure to GPT-4, so that it can capture multi-modal data for captioning. We have tried multiple different input prompts and discovered that the following prompt produces the best result. "A query molecule Acrylamide has SMILES string NC(=O)C=C. Acrylamide is a N-Acrylammonia and Acrylamides, has role Carcinogenic agent and Alkylating agent, has functional parent Acrylic acid. Return a paragraph to describe Acrylamide."

Here we use Acrylamide as an example to illustrate how to construct the prompt. For simplicity and clarity purpose, we incorporate 1-hop knowledge triples for illustration, and in our experiments we use 2-hop knowledge triples to construct the prompt.

[6] https://www.ebi.ac.uk/chebi/
[7] https://pubchem.ncbi.nlm.nih.gov/