

---

# Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective

---

Jiatong Li<sup>1\*</sup> Yunqing Liu<sup>1\*</sup> Wenqi Fan<sup>1</sup> Xiao-Yong Wei<sup>1,2</sup> Hui Liu<sup>3</sup> Jiliang Tang<sup>3</sup> Qing Li<sup>1</sup>

## Abstract

Molecule discovery plays a crucial role in various scientific fields, advancing the design of tailored materials and drugs. Traditional methods for molecule discovery follow a trial-and-error process, which are both time-consuming and costly, while computational approaches such as artificial intelligence (AI) have emerged as revolutionary tools to expedite various tasks, like molecule-caption translation. Despite the importance of molecule-caption translation for molecule discovery, most of the existing methods heavily rely on domain experts, require excessive computational cost, and suffer from poor performance. On the other hand, Large Language Models (LLMs), like ChatGPT, have shown remarkable performance in various cross-modal tasks due to their great powerful capabilities in natural language understanding, generalization, and reasoning, which provides unprecedented opportunities to advance molecule discovery. To address the above limitations, in this work, we propose a novel LLMs-based framework (**MolReGPT**) for molecule-caption translation, where a retrieval-based prompt paradigm is introduced to empower molecule discovery with LLMs like ChatGPT without fine-tuning. More specifically, MolReGPT leverages the principle of molecular similarity to retrieve similar molecules and their text descriptions from a local database to ground the generation of LLMs through in-context few-shot molecule learning. We evaluate the effectiveness of MolReGPT via molecule-caption translation, which includes molecule understanding and text-based molecule generation. Experimental results show that MolReGPT outperforms fine-tuned models like MolT5-base without any additional training. To the best of our

knowledge, MolReGPT is the first work to leverage LLMs in molecule-caption translation for advancing molecule discovery. Our implementation is available at: <https://github.com/phenixace/MolReGPT>

## 1. Introduction

Molecules are the fundamental building blocks of matter, comprising the intricate fabric of the world around us. As the foundation of all chemical compounds, molecules are composed of two or more atoms that are chemically bonded together, and they retain the unique chemical properties dictated by their specific structures (Xu et al., 2023). With a comprehensive understanding of molecules, scientists can effectively design materials, drugs, and products with tailored characteristics and functionalities, impacting a variety of crucial fields such as chemistry (Wang et al., 2023; Cuzucoli Crucitti et al., 2023; Weng et al., 2021), pharmacology (Patani & LaVoie, 1996; Anderson, 2003; Ding et al., 2019), material science (Curtarolo et al., 2013; Higuchi et al., 2023), and environmental science (Ali et al., 2023; Lv et al., 2023). One notable example is in the pharmaceutical industry during the COVID-19 pandemic, where the discovery of new molecules has the potential to revolutionize not only the development of groundbreaking treatments, therapies (Gupta et al., 2023), and vaccines against viruses but also a wide range of other diseases in the coming decade (Osamor et al., 2023).

However, traditional molecule discovery lies in the long, expensive, and failure-prone process that requires navigating a complex landscape of molecule structures and biological interactions, with limitations in scalability, precision, and data management (Hajduk & Greer, 2007). To overcome these challenges, computational technologies such as artificial intelligence (AI) have emerged as powerful tools to expedite the discovery of new molecules (Urbina & Ekins, 2022). Specifically, molecules can be represented as simplified molecular-input line-entry system (SMILES) strings (Weininger, 1988; Cao et al., 2022). As shown in Figure 1 (a), the structure of Phenol can be represented as a SMILES sequence, which is made of a Benzene ring and

---

\*Equal contribution <sup>1</sup>Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR China <sup>2</sup>Dept. of computer science, Sichuan University, China <sup>3</sup>Michigan State University, Michigan, USA. Correspondence to: Wenqi Fan <wenqifan03@gmail.com>, Qing Li <qing-prof.li@polyu.edu.hk>.

a Hydroxy. Such SMILES representations can be effectively processed by deep sequence models like Recurrent Neural Networks (Arús-Pous et al., 2019; Grisoni et al., 2020) and Transformers (Honda et al., 2019; Yoshikai et al., 2023). These AI-powered models enable researchers to understand molecular properties and functionalities and create promising compounds in a more efficient and cost-effective manner. For example, in order to generate new molecules and better understand them, a novel task that translates between molecules and natural language has been proposed by using language models like Text2Mol (Edwards et al., 2021) and MolT5 (Edwards et al., 2022). It consists of two sub-tasks: molecule captioning (Mol2Cap) and text-based molecule generation (Cap2Mol). More specifically, as shown in Figure 1 (b-c), the goal of *molecule captioning* is to generate a text caption to describe a SMILES string of the molecule for providing humans with a better understanding of molecule, while *text-based molecule generation* aims to generate the corresponding molecule (i.e., SMILES string) based on a given natural language description (e.g., properties and functional groups). Despite the impressive progress that has been made in the molecule-caption translation task, the majority of existing advanced approaches suffer from several limitations (Edwards et al., 2021; 2022; Su et al., 2022). First, the design of such model architecture in molecule-caption translation heavily relies on domain experts, which can significantly limit the development and deployment of AI-powered molecule discovery. Second, most existing methods follow the “pre-train&fine-tuning” paradigm for molecule-caption translation, which requires excessive computational costs. Third, existing approaches such as Text2Mol and MolT5 fall short in their inability to reason on complex tasks and generalize to unseen examples. Therefore, it is desired to design a novel paradigm for molecule-caption translation in molecule discovery.

Recently, Large Language Models (LLMs), scaling up their weights to the billion level, have achieved tremendous success not only in the field of Natural Language Processing (NLP) but also in some cross-modal areas like computer vision (Zhu et al., 2023), recommender systems (Bao et al., 2023), and molecule discovery (Edwards et al., 2022). For example, ChemGPT (Frey et al., 2022), a variant of GPT model with more than one billion parameters, is introduced to understand and generate small molecules in chemistry. Meanwhile, in addition to the impressive capabilities in natural language understanding and generation, LLMs also demonstrate their powerful generalization and reasoning capabilities (Rubin et al., 2022; Min et al., 2022), which can generalize to other unseen tasks by specific task context (In-Context Learning, ICL) without being fine-tuned and largely reduce computational cost. Therefore, LLMs provide unprecedented potential to advance molecule discovery, specifically the task of molecule-caption translation.

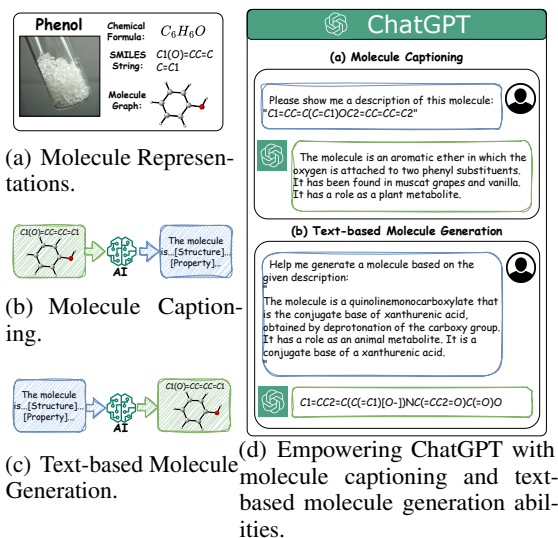


Figure 1. An illustration on translation between molecule and natural language in molecule discovery. (a) A molecule can be denoted as a chemical formula, SMILES string, and 2D molecule graph. (b) Molecule captioning aims to generate a text caption to describe a molecule for humans’ better understanding. (c) Given a text description, text-based molecule generation is used to generate a corresponding molecule. (d) Large language models (e.g., ChatGPT) can perform molecule captioning and text-based molecule generation with corresponding well-designed prompts.

Although building specific LLMs in molecule discovery has immense potential for advancing scientific research, we also face significant challenges. First, due to privacy and security concerns, many advanced large language models (e.g., ChatGPT and GPT4.0) are not publicly available, where LLMs’ architectures and parameters are not released publicly for fine-tuning in downstream tasks. Second, owing to their complex architectures and the extensive data required, training advanced LLMs requires significant computing resources, leading to high costs and substantial energy consumption. For instance, it has been reported that the cost of *one single training session* for GPT-3 exceeds 1 million. As a result, it is very challenging for us to re-design our own LLMs with pre-training and fine-tuning in specific downstream tasks. At last, it is crucial to design proper guidelines/prompts with high-quality few-shot examples to improve LLMs’ generalization and reasoning capabilities for molecule discovery.

To address such challenges, as the early exploration attempt to take advantage of the powerful capabilities of LLMs in the molecule discovery field, in this work, we propose a novel solution to teach LLMs with prompts for translating between molecules and natural language, as illustrated in Figure 1 (d).

More specifically, inspired by the latest ChatGPT, a retrieval-based prompt paradigm through in-context learning (ICL) is developed to conduct two sub-tasks (i.e., molecule captioning and text-based molecule generation) without fine-tuning the LLMs, where n-shot examples are retrieved to augment the prompt instances via BM25-based caption ranking and Morgan Fingerprints-based molecule ranking. Experiments show that MolReGPT can achieve Text2Mol scores of 0.560 in Mol2Cap generation and 0.571 in Cap2Mol generation, which surpasses the *fine-tuned* MolT5-base in both sub-tasks of molecule-caption translation. Notably, MolReGPT even outperforms MolT5 largely in text-based molecule generation, increasing the Text2Mol metric by 3%. Note that all of these improvements from our proposed method are achieved without any fine-tuning steps.

Our major contributions are summarized as follows:

- We introduce a principle strategy based on LLMs to perform translation between molecules and natural language for molecule discovery. To the best of our knowledge, we are the first to investigate molecule-caption translation by employing LLMs.
- We develop a novel framework (MolReGPT) to empower LLMs like ChatGPT to perform molecule captioning and text-based molecule generation without being fine-tuned, where a retrieval-based prompt paradigm through in-context learning is developed to explicitly guide the generation process.
- Comprehensive experiments on a real-world dataset demonstrate the effectiveness of the proposed method on molecule captioning and text-based molecule generation tasks, surpassing even fine-tuned models such as T5-base and MolT5-base.

## 2. Related Work

In this section, we briefly review related work about molecule-caption translation tasks in molecule discovery as well as the advanced LLMs techniques.

### 2.1. Molecule Discovery

Molecule discovery plays a pivotal role across numerous scientific fields, driving advancements in the development of drug discovery and material design (Du et al., 2022). In recent decades, AI-powered approaches have emerged as mainstream techniques to revolutionize the process of molecule discovery (Hu et al., 2023; Fan et al., 2023). For instance, SMILES-based Variational Autoencoders (VAEs) methods such as ChemVAE (Gómez-Bombarelli et al., 2018), SD-VAE (Dai et al., 2018), and GrammarVAE (Kusner et al., 2017), employ a VAE-based model that encodes and decodes SMILES strings, generating new

molecules (strings) by decoding from a Gaussian prior. In terms of molecular string representation, existing studies have explored advanced deep representation methods from other fields, including Convolutional Neural Network (CNN) (Peng & Zhao, 2019; Le et al., 2019), Recurrent Neural Network (RNN) (Grisoni et al., 2020; Amabilino et al., 2020), and Transformer (Bagal et al., 2021; Wang et al., 2021).

More recently, as a new task in molecule discovery, Text2Mol (Edwards et al., 2021) is introduced to retrieve molecules using natural language descriptions as search queries, in which a paired dataset of molecules and their corresponding text descriptions are constructed, enabling the learning of a shared semantic embedding space for retrieval. KV-PLM (Zeng et al., 2022) develops a knowledgeable machine reading system pre-trained on a domain corpus, in which SMILES strings are inserted and link molecule structures with biomedical text. What’s more, a self-supervised learning framework MolT5 (Edwards et al., 2022) is proposed to pre-train on a substantial volume of unlabeled language text and SMILES strings, enhancing the molecule-caption translation task, such as molecule captioning and text-based molecule generation. MoMu (Su et al., 2022) bridges molecular graphs and natural language by pre-training molecular graphs and their semantically related text data through comparative learning.

### 2.2. Large Language Models

Large Language models (LLMs) have been a trending topic in recent years, with numerous studies exploring their capabilities and potential applications. One of the most well-known LLMs is the GPT family (Radford et al., 2018; 2019; Brown et al., 2020; Ouyang et al., 2022), which has played a pivotal role in advancing the field of generative language models. As a representative of the GPT family, ChatGPT is specifically fine-tuned for conversational purposes, which can generate impressively human-like responses (Leiter et al., 2023). In addition, other LLMs, such as LaMDA (Thoppilan et al., 2022), PaLM (Chowdhery et al., 2022), and Vicuna (Chiang et al., 2023), also show a decent performance.

The power of LLMs is far beyond language generation but also lies in their ability to learn from context, namely the ability of in-context learning. Several works have explored the utilization of in-context learning from various tasks, such as KATE (Liu et al., 2021) and AutoCoT (Zhang et al., 2022). These works show that through in-context learning, LLMs can adapt to new tasks based on the context provided in the input, eliminating the need for explicit fine-tuning on specific tasks.

In addition to NLP tasks, LLMs have also shown remarkable potential in various molecule discovery tasks, such as

molecule understanding (Bran et al., 2023; White, 2023). For instance, ChemBERTa (Chithrananda et al., 2020) leverages pre-training on an extensive corpus of chemical texts, enabling it to comprehend the structure and properties of chemical compounds. Another notable example is MoleculeSTM (Liu et al., 2022), which employs in-context learning in conjunction with LLMs. This approach facilitates a deeper understanding of the relationships between chemical structures and their corresponding textual descriptions. Furthermore, ChemGPT (Frey et al., 2022) represents a variant of the GPT model specifically trained on chemical data. Through the application of in-context learning, ChemGPT is capable of generating novel chemical structures and accurately predicting their properties. MolT5 (Edwards et al., 2022) shows that LLMs can perform the cross-modal transition task between molecule and text (i.e. molecule captioning task and text-based molecule generation task), which is one of the most closely related attempts to ours. Note that MolT5 needs to pre-train and fine-tune LLMs for translating between molecules and natural language, leading to huge computational costs. In this paper, we propose a novel framework to empower LLMs like ChatGPT to perform molecule captioning and text-based molecule generation without being fine-tuned.

### 3. MolReGPT

In this section, we aim to introduce our proposed method (MolReGPT) as a novel solution to empower molecule discovery for molecule-caption translation with LLMs like ChatGPT. We will first introduce an overview of the proposed framework, and then detail each model component.

#### 3.1. An Overview

Due to the huge computation costs, training or fine-tuning LLMs on the domain-specific corpus from the molecule discovery field is often infeasible in practice. To address such limitations, we investigate leveraging the great capabilities of LLMs without changing the LLMs, where we propose a novel framework (MolReGPT) to empower ChatGPT with the ability of molecule-caption translation for molecule discovery. More specifically, in order to improve the quality of guidelines/prompts, a retrieval-based prompt paradigm under in-context learning is introduced to teach ChatGPT to conduct two molecule-related tasks: *molecule captioning* (Mol2Cap) and *text-based molecule generation* (Cap2Mol). The framework of MolReGPT is shown in Figure 2, consisting of four main stages: Molecule-Caption Retrieval, Prompt Management, In-Context Few-Shot Molecule Learning, and Generation Calibration, following the workflow of pre-processing, querying, and post-processing. The first stage, Molecule-Caption Retrieval, is used to retrieve the  $n$  most similar examples (i.e., *few-shot examples*) from human-

annotated datasets (i.e., molecule-caption pairs database) for augmenting the prompt instances. The second stage is Prompt Management, which is executed to construct the system prompt as evidence for successive in-context learning. After that, both the system and the user input prompt are sent to query LLMs such as ChatGPT to perform In-Context Few-Shot Molecule Learning without fine-tuning LLMs for the molecule-caption translation task in the molecule discovery field. Valid responses are expected in the pre-defined JSON format, while there may be instances where the language models generate unexpected outputs. The last stage is Generation Calibration, which is deployed to calibrate the original responses for the validity of the outputs.

#### 3.2. Molecule-Caption Retrieval

In order to teach LLMs to handle the molecule-caption translation task (i.e., Mol2Cap and Cap2Mol) without fine-tuning LLMs, we propose to perform in-context learning with few-shot examples to prompt LLMs. In general,  $n$  examples are randomly selected from human-annotated datasets (i.e., molecule-caption pair database), providing a general task instruction to LLMs. However, such a naive solution often provides insufficient knowledge regarding the associations between natural language and molecules. To mitigate this issue, we propose incorporating retrieval methods into the selection of examples to complement the lack of domain-specific knowledge of LLMs in molecule discovery, specifically through the stage of Molecule-Caption Retrieval. These retrieval strategies are motivated by the similar property principle, in which molecules similar in structures tend to exhibit similar characteristics (Wang et al., 2016). Thus, similar captions containing the descriptions of molecule structures and properties are used to describe similar molecules. Therefore, by retrieving the most similar molecules or captions, we can utilize the corresponding molecule-caption pairs as examples to prompt LLMs.

However, the SMILES representation of molecules as a sequence structure can not reveal the actual 2-D graph topology of molecules. Hence, domain-specific methods are required for better molecular similarity calculation during the retrieval stage. Specifically, given a SMILES string representation for *molecule captioning* task, we introduce to use of Morgan Fingerprints (i.e., molecular structures representations) (Butina, 1999), to calculate molecular similarity using Dice similarity for molecule retrieval. In *text-based molecule generation task* for caption retrieval, BM25, which is widely used in information retrieval (Robertson et al., 2009), is proposed to compute similarity scores between captions of molecules, which mainly contain functional groups and properties of molecules. In both scenarios, top- $n$  molecule-caption pair examples are retrieved to serve as examples in the system prompt. Next, we will detail Morgan Fingerprints-based molecule retrieval and BM25-based



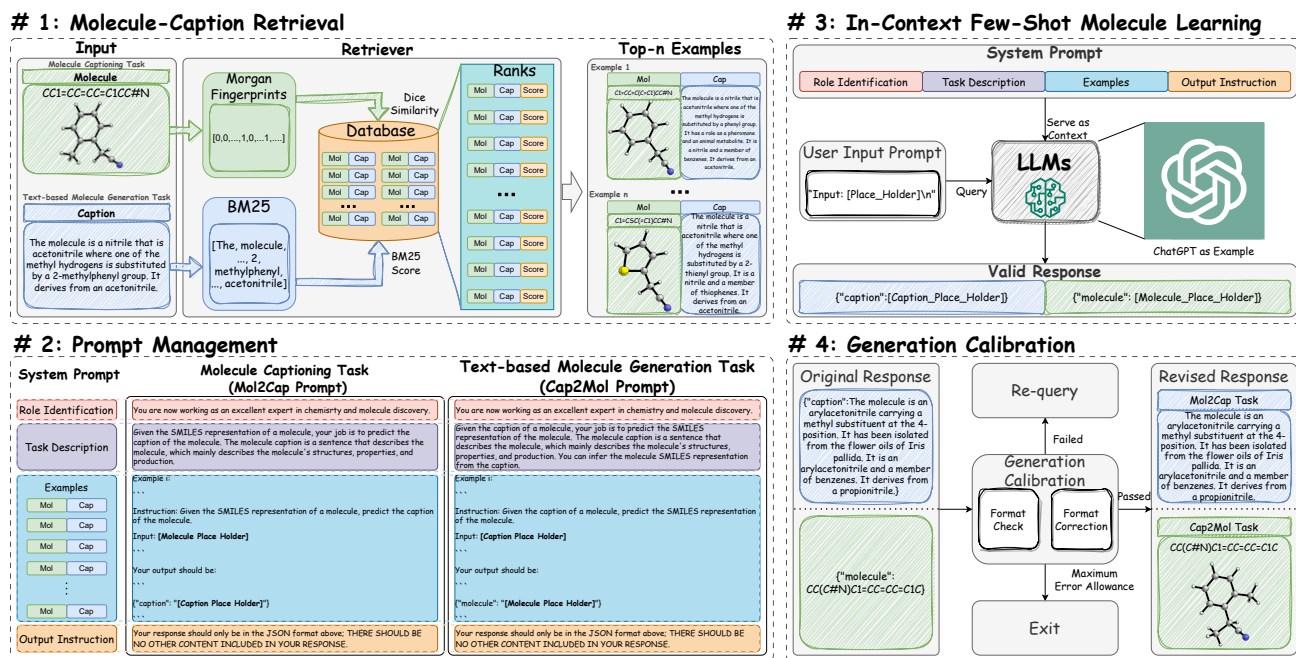


Figure 2. This diagram shows the workflow of MolReGPT. MolReGPT consists of four main stages. In stage 1, Molecule-Caption Retrieval is employed to find  $n$  best-matched examples from the local database. Then in stage 2, Prompt Management helps construct the system prompt with the retrieved molecule-caption pairs. Following this, LLMs perform In-Context Few-Shot Molecule Learning based on the provided system prompt and user input prompt. Finally, Generation Calibration is conducted to ensure desired output.

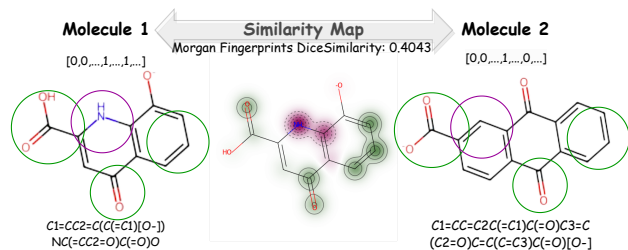


Figure 3. Illustrations of Morgan Fingerprints and Dice Similarity. The two molecules will first be transformed into the Morgan Fingerprints. Then, Dice similarity will be calculated. The green colour corresponds to sub-structures that contribute positively to the similarity score between the molecules, while the purple colour represents sub-structures that contribute negatively or have differences between the molecules.

caption retrieval.

### 3.2.1. MORGAN FINGERPRINTS-BASED MOLECULE RETRIEVAL

Molecular fingerprints are numerical representations of the chemical structures of molecules, which can be used for various computational objectives (Butina, 1999), such as similarity searching, property prediction, virtual screening,

and cluster analysis. One of the most representative molecular fingerprints is the Morgan Fingerprints (Morgan FTS), which is also known as circular fingerprints or extended-connectivity fingerprints (ECFP).

The key idea behind Morgan FTS is to capture the presence or absence of specific sub-structures or chemical fragments in a molecule. Morgan FTS follows a variant of the Morgan algorithm (Butina, 1999), which encodes the structural information of a molecule by representing its connectivity patterns in a circular manner. Morgan FTS is then generated by iteratively expanding a set of atoms from a central atom in the molecule, capturing the neighbouring atoms and their bond types at each expansion step. The process continues until a pre-defined radius is reached. The resulting fingerprint is a binary bit vector, where each bit represents the presence or absence of a particular substructure.

What's more, Morgan FTS has several advantages over other types of fingerprints, including their ability to handle molecules of varying sizes, resistance to small structural changes, and effectiveness in capturing structural similarities between molecules. To extract the Morgan Fingerprints, the SMILES representations of the molecules are converted into rdkit objects using the rdkit library<sup>1</sup>. Subsequently,

<sup>1</sup><https://www.rdkit.org/>

we apply Dice similarity (Dice, 1945), also known as the Dice coefficient, to measure the similarity between the input molecule and the molecules in the local database. Mathematically, it can be expressed as:

$$Dice(A, B) = \frac{2 * |A \cap B|}{|A| + |B|}, \quad (1)$$

where  $A$  and  $B$  are the Morgan Fingerprints of the two molecules.  $|A|$  and  $|B|$  represent the cardinality (i.e., number of sub-structures) of  $A$  and  $B$ .  $|A \cap B|$  denotes the number of sub-structures that are common to both  $A$  and  $B$ . Dice similarity ranges from 0 to 1, where the value of 0 indicates no overlap or similarity between the molecules, and the value of 1 represents complete overlap. As shown in Figure 3, Dice similarity can be calculated by comparing the Morgan fingerprints of the molecules. The similarity map shows the similarities and differences between the two molecules. The Dice similarity is particularly useful when dealing with imbalanced datasets or focusing on the agreement between positive instances (i.e., sub-structures present in both sets) rather than the overall agreement.

Compared to existing molecule embedding methods (Couptry & Pogány, 2022), Morgan FTS together with Dice similarity provides a distinctive advantage by explicitly indicating the similarities in detailed molecular structures, as these structures are usually directly stated in the molecule captions.

### 3.2.2. BM25-BASED CAPTION RETRIEVAL

BM25 is one of the most representative ranking approaches in information retrieval for calculating the relevance of the documents to the given query (Robertson et al., 2009). The idea is based on the term frequency-inverse document frequency (TF-IDF), which measures how often a term appears in a document (i.e., caption) and how rare it is in the corpus of documents (i.e., the local database) (Aizawa, 2003). In addition, BM25 considers the caption’s length and the position of the query terms in the caption.

In the Cap2Mol task, we use the input caption as the query sentence, while the captions in the local database (i.e., the training set), are served as the corpus of documents, where each caption represents a document. Mathematically, the formula of BM25 can be defined as follow:

$$score(Q, D) = \sum_{i=1}^N IDF(q_i) * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})}, \quad (2)$$

where  $D$  is the caption corpus and  $Q$  is the query caption.  $N$  is the number of query terms in the query caption,  $q_i$

is the  $i$ -th query term,  $IDF(q_i)$  is the inverse document frequency of  $q_i$ ,  $f(q_i, D)$  is the term frequency of  $q_i$  in  $D$ ,  $k_1$  and  $b$  are tuning parameters,  $|D|$  is the length of  $D$ , and  $avgdl$  is the average caption length in the corpus.

In caption retrieval, BM25 is applied to calculate the similarity scores between captions so that the relevant molecule structures described by captions can be learnt through retrieved molecule-caption pairs.

### 3.3. Prompt Management

System prompts and user input prompts are two important parts to form the task context. User prompts are usually more complex and contain essential instructions for task solving and format formalization, where user prompts are defined to formalize the user inputs. To help LLMs understand the task and generate desired outputs, Prompt Management is proposed to design the system prompt templates, which are further completed with the retrieved examples. As shown in stage 2 of Figure 2, the system prompts consist of four parts: Role Identification, Task Description, Retrieved Examples, and Output Instruction.

**Role Identification** aims to help LLMs identify the role of experts in the chemistry and molecule discovery domain. By establishing this role, the LLMs are encouraged to generate responses that align with the expertise expected in the specific domain.

**Task Description** provides a comprehensive explanation of the task’s content, ensuring that LLMs have a clear understanding of the specific task they need to address. It also includes critical definitions to clarify terms or concepts that are specialized in the molecule-caption translation task.

The next component of the system prompt is designed to define the user input prompt and incorporate the **Retrieved Examples**, which serve as the evidence for the molecule-caption translation task, allowing LLMs to leverage the information contained within few-shot examples to generate better responses.

Finally, **Output Instruction** specifies the desired format for the response. Here, we restrict the output to a JSON format. The choice of JSON format enables a quick and efficient validation of the LLMs’ response, ensuring that it adheres to the expected structure and facilitates further processing or analysis.

### 3.4. In-Context Few-Shot Molecule Learning

Since ChatGPT is treated as a *black-box* system, it is impossible for us to fine-tune the model’s parameters on task-specific datasets for translation between molecules and natural language captions. Besides, as the weights of LLMs continue to scale, it is infeasible to train and fine-tune these

foundation models with huge computational resources. To address the above limitations, recently, as an alternative to fine-tuning, in-context learning techniques provide great opportunities to teach ChatGPT to make predictions based on a few examples. In this work, we introduce in-context few-shot molecule learning to perform the Mol2Cap task and Cap2Mol task without fine-tuning ChatGPT. This stage is to utilize both the system prompt and user input prompt to query the LLMs after Molecule-Caption Retrieval and Prompt Management. In particular, the combination of the system prompt and user input prompt provides ChatGPT with a clear guideline (i.e., Mol2Cap and Cap2Mol prompts with a few examples) via in-context learning. The system prompt establishes the task framework of molecule-caption translation and molecule domain expertise, while the user prompt narrows the focus and directs the model’s attention to the specific user input. As a result, ChatGPT can learn how to perform the molecule-caption translation from the given task context, without the necessity to modify its parameters.

The formula below describes the differences between fine-tuning and in-context learning. Let  $L$  be the model of ChatGPT,  $m$  be the molecule,  $c$  be the molecule caption, and  $\theta$  be the parameters of  $L$ . The fine-tuning process can be formulated as:

$$c = L(m; \theta_m^*), \quad (3)$$

$$m = L(c; \theta_c^*), \quad (4)$$

where  $\theta_m^*$  and  $\theta_c^*$  are the updated parameters after being fine-tuned on the entire training set ( $\theta_m^*$  for Mol2Cap and  $\theta_c^*$  for Cap2Mol).

In contrast, the In-Context Few-Shot Molecule Learning process can be defined as:

$$c = L(p_m(m); \theta), \quad (5)$$

$$m = L(p_c(c); \theta), \quad (6)$$

where  $p_m(\cdot)$  and  $p_c(\cdot)$  are the Prompt Management templates that transform the original user input (molecules  $p_m$  or captions  $p_c$ ) into system prompts with the user input prompts for querying ChatGPT, and  $\theta$  is the original parameters without being fine-tuned.

It is apparent that the fine-tuning methods require additional model training for the sub-tasks of molecule-caption translation. In contrast, in-context few-shot molecule learning only needs to switch the prompt templates, which is much more efficient for deployment. Through the way of in-context few-shot molecule learning, valid and meaningful responses are expected, which contain the generated captions or molecules in our pre-defined JSON formats.

### 3.5. Generation Calibration

Despite specifying the desired output format, LLMs (e.g., ChatGPT) can occasionally produce unexpected responses, including incorrect output formats and denial of answering. To address these issues, a generation calibration mechanism is introduced to validate the response from ChatGPT.

In Generation Calibration, we first check the format of original responses by parsing them into JSON objects. If the parsing process fails, indicating a deviation from the expected format, several pre-defined format correction strategies, such as Regular Matching (Thompson, 1968), are introduced to correct the format and extract the desired output from the response. If the original response successfully passes the format check or can be calibrated using the format correction strategies, it is considered valid and accepted as a final response. However, if the original response fails the format check and cannot be corrected within the predefined strategies, we initiate re-queries. Notably, there is a special case for re-queries. When the original response reports the "Exceed Maximum Input Length Limitation" error, we will remove the longest example in the re-query phase until the query length meets the length limitation. The re-query process involves making additional queries to the LLMs until a valid response is obtained or until the maximum error allowance is reached. This maximum error allowance is set to ensure that the system does not get stuck in an endless loop and instead provides a suitable response to the user within acceptable bounds.

By employing the generation calibration stage, we can mitigate unexpected deviations from the desired output format and ensure that the final responses align with the expected format and requirements.

## 4. Experiment

In this section, we aim to evaluate the feasibility and effectiveness of the proposed method MolReGPT by conducting comprehensive experiments on the molecule-caption translation task. Additionally, ablation studies are conducted to investigate the impact of different retrieval methods and the number of selected examples. These investigations aim to provide deeper insights into the performance and capabilities of MolReGPT.

### 4.1. Experimental Settings

We first introduce the basic experimental settings. In this work, we use ChatGPT through the OpenAI API<sup>2</sup> with backend model **GPT-3.5-turbo**, which can not be fine-tuned in our tasks. Besides, we will provide an overview of the data and metrics employed in this section.

<sup>2</sup><https://openai.com/blog/openai-api>

#### 4.1.1. DATASET

The research on molecule-caption translation is still in the early stage, and there is only one public dataset ChEBI-20 (Edwards et al., 2021), which contains 33,010 molecule-caption pairs. To ensure consistency, we adhere to the data split process as used in MolT5 (Edwards et al., 2022), dividing the dataset into 80/10/10% train/validation/test splits. For our method evaluation, we focus on the test split while utilizing the training set as the local database to retrieve n-shot examples through in-context learning.

#### 4.1.2. EVALUATION METRICS

In terms of evaluation metrics, we align with the metrics adopted in MolT5 for comparison (Edwards et al., 2022). By adopting these metrics, we ensure consistency and enable a meaningful and fair assessment of the performance of our method.

- **Mol2Cap Metrics.** In the Mol2Cap task, natural language generation metrics like BLEU, ROUGE, and METEOR scores are applied to assess the proximity of the generated output to the ground truth. Here, BLEU and ROUGE scores evaluate the n-gram precision, measuring the alignment between the generated structures and the reference structures, while METEOR is a recall-oriented metric that accounts for both exact matches and paraphrases between the generated and reference structures. Additionally, we incorporate *Text2Mol*, a task-specific metric that employs pre-trained models to quantify the structural similarity between the generated and reference molecules using their SMILES representations (Edwards et al., 2021). This metric provides further insights into the quality and relevance of the generated output in terms of the underlying molecular structures.
- **Cap2Mol Metrics.** Since the SMILES representation of molecules exhibits a sequence structure, natural language metrics can be directly applied for evaluation. Thus, BLEU and the Exact Match scores are calculated as initial assessments. Furthermore, molecule-specified metrics are also reported, including Levenshtein distance, validity, and three molecule fingerprints scores - MACCS FTS, RDK FTS, and Morgan FTS. These metrics provide valuable insights into the quality, validity, and structural characteristics of the generated molecules. Finally, the *Text2Mol* metric is also discussed here to highlight the relevance between the generated molecules and the input molecule captions.

Note that smaller values of Levenshtein score and FCD indicate better generation performance in the molecule genera-

tion task, while other evaluation metrics positively correlate to the performance.

#### 4.1.3. BASELINES

It is worth mentioning that there are limited baselines for translating between molecule captioning and text-based molecule generation. Specifically, the following baselines are selected for performance evaluation:

- **Transformer** (Vaswani et al., 2017). This method is the most representative language architecture to process natural language. A vanilla Transformer model with six encoder and decoder layers, directly trained on ChEBI-20. Note that this model is not pre-trained, making it simple and easy to implement.
- **T5-base** (Raffel et al., 2020). T5 is pre-trained on the Colossal Clean Crawled Corpus (C4), but no domain knowledge is specifically fed for pre-training. In this work, the base version of T5 is directly fine-tuned on ChEBI-20 for molecule discovery.
- **MolT5-base** (Edwards et al., 2022). This model is pre-trained on a large corpus with both language texts and SMILES strings so that it can have a prior understanding of the two domains. More specifically, the base version of MolT5 was pre-trained on the Colossal Clean Crawled Corpus (C4) and ZINC-15 datasets and further fine-tuned on task-specific dataset ChEBI-20.

Note that these baselines are required to fine-tune the model on the public dataset ChEBI-20, specifically tailored to the molecule-caption translation task.

## 4.2. Performance Comparison of Molecule-Caption Translation

We present the results of each sub-task within the molecule-caption translation task, incorporating both quantitative analysis and detailed examples for comparison. In addition, Figures 7, 5 and 4 illustrate specific examples that demonstrate the differences among various models, providing a visual understanding of their performance.

#### 4.2.1. MOLECULE CAPTIONING (MOL2CAP)

Given a molecule’s SMILES representation, Mol2Cap aims to generate natural language for describing the molecule to enable humans to understand molecular structure, properties, and functionalities in a more efficient and cost-effective manner. Table 1 illustrates the performance comparison of 10-shot MolReGPT (GPT-3.5-turbo) with other advanced methods for the Mol2Cap task, offering an overview of the results. Notably, our method can achieve comparable



ROUGE scores to MolT5-base and T5-base while surpassing all selected baselines in the remaining metrics without being fine-tuned on ChEBI-20 dataset. Furthermore, we obtain the following observations.

First, GPT-3.5-turbo is not explicitly trained or fine-tuned for molecule-caption translation tasks so it has poor zero-shot performance. However, with the instruction of 10-shot MolReGPT, GPT-3.5-turbo achieves significantly improved results that gain an improvement of 60% to the zero-shot case and 2.4% to MolT5-base under the Text2Mol metric, indicating that our proposed method can teach ChatGPT to effectively learn the Mol2Cap task from the system prompt.

Second, limited by the number of examples, MolReGPT only gains limited insights from the distribution of molecule captions. The model’s predictions for captions heavily rely on its internal factual knowledge and the contextual information provided by the system prompt, which means that common patterns may not be as apparent and can not be captured from the selected  $n$  examples. As a result, although our 10-shot MolReGPT achieves a 0.560 Text2Mol score, which is higher than MolT5’s 0.547, MolReGPT in turn gets lower ROUGE scores compared to MolT5. However, it is crucial to note that the captions generated by 10-shot MolReGPT with lower ROUGE scores are not entirely incorrect. In fact, the highest Text2Mol score serves as a reliable indicator of the generation quality and highlights the better relevance between the generated molecules and the molecule captions.

Figure 4 lists examples of molecule captioning results to compare the performance among different models. From the given examples, we note that our MolReGPT can generate captions that contain key information about the input molecule. And more importantly, the generated captions are better in grammar and easy for humans to understand.

#### 4.2.2. TEXT-BASED MOLECULE GENERATION (CAP2MOL)

Given a natural language description (e.g., properties and functional groups), the goal of Cap2Mol is to generate the corresponding molecule (i.e., SMILES string) for molecule discovery. Results of the text-based molecule generation task are presented in Table 2. Comparing all these baselines, 10-shot MolReGPT significantly enhances the capabilities of GPT-3.5-turbo, leading to the best overall performance. In molecular evaluation metrics like MACCS FTS, RDK FTS, and Morgan FTS, MolReGPT helps GPT-3.5-turbo gain a significant 15% increase in Text2Mol score compared to MolT5-base. Considering the molecule fingerprints scores, our 10-shot MolReGPT also gets an average of 18% improvement compared to MolT5-base. Besides, MolReGPT also achieves the highest exact match score, generating 13.9% molecules that are completely correct to

the ground truth. Remarkably, these impressive results are achieved without additional training or fine-tuning steps.

Furthermore, it is worth noting that the original weights of T5 are primarily for natural language generation, which means it has to be fine-tuned separately to fit the two sub-tasks in this study. Unfortunately, MolT5 does not tackle this issue, as it continues to treat the two sub-tasks of the molecule-caption translation task as separate tasks. Switching between the two sub-tasks in MolT5 requires using a different model class and reloading the weights, which makes it technically inefficient. Besides, treating these sub-tasks as independent overlooks the potential knowledge transfer between them. In contrast, MolReGPT enables a single foundation LLMs to solve both the two sub-tasks simultaneously, providing a comprehensive solution for LLMs to address molecule-related tasks.

Figure 5 lists examples of text-based molecule generation results to compare the performance among different models. From the given examples, it is clear that our MolReGPT can generate structures more similar to the ground truth.

### 4.3. Ablation Study

In addition to the experiments above, we also perform ablation studies to analyze the critical factors that influence the performance of MolReGPT. We first examine the impact of different retrieval strategies employed for retrieving  $n$ -shot examples. Subsequently, we investigate the influence of the number of selected examples, denoted as  $n$ , including zero-shot results to ensure that GPT-3.5-turbo is not already trained to handle the molecule-caption translation task. These ablation studies shed light on the key aspects contributing to the performance of MolReGPT.

#### 4.3.1. IMPACT OF RETRIEVAL STRATEGIES

Retrieval strategies play a key role in guiding LLMs to perform molecule-caption translation tasks for MolReGPT. More similar examples are retrieved, and more valuable information could be contained for in-context few-shot molecule learning. For each sub-task in the molecule-caption translation task, we choose three retrieval strategies for comparison. The detailed results are shown in Table 3 and Table 4.

**Molecule Captioning (Mol2Cap).** In the Mol2Cap task, we compare the performance of three retrieval strategies: Random, BM25, and Morgan FTS (adopted in MolReGPT). The Random strategy involves retrieving  $n$  random examples, while BM25 applies a character-level BM25 algorithm to the molecule SMILES representations.

As shown in Table 3, among the three retrieval strategies, Morgan FTS shows the best performance with the same value of  $n$ , outperforming BM25 by 37% in the Text2Mol

Table 1. The performance of molecule captioning on ChEBI-20 dataset. Experimental results for Transformer, T5-base, and MolT5-base are retrieved from (Edwards et al., 2022). The **best** scores are in bold, and the second-best scores are underlined.

| Methods                            | BLEU-2 $\uparrow$ | BLEU-4 $\uparrow$ | ROUGE-1 $\uparrow$ | ROUGE-2 $\uparrow$ | ROUGE-L $\uparrow$ | METETOR $\uparrow$ | Text2Mol $\uparrow$ |
|------------------------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|---------------------|
| Transformer (Edwards et al., 2022) | 0.061             | 0.027             | 0.204              | 0.087              | 0.186              | 0.114              | 0.057               |
| T5-base (Edwards et al., 2022)     | 0.511             | 0.423             | 0.607              | <u>0.451</u>       | <u>0.550</u>       | 0.539              | 0.523               |
| MolT5-base (Edwards et al., 2022)  | <u>0.540</u>      | <u>0.457</u>      | <b>0.634</b>       | <b>0.485</b>       | <b>0.578</b>       | <u>0.569</u>       | <u>0.547</u>        |
| GPT-3.5-turbo (zero-shot)          | 0.103             | 0.050             | 0.261              | 0.088              | 0.204              | 0.161              | 0.352               |
| GPT-3.5-turbo (10-shot MolReGPT)   | <b>0.565</b>      | <b>0.482</b>      | <u>0.623</u>       | 0.450              | 0.543              | <b>0.585</b>       | <b>0.560</b>        |

Table 2. Text-based molecule generation results on CheBI-20. Experimental results for Transformer, T5-base, and MolT5-base are retrieved from (Edwards et al., 2022). The **best** scores are in bold, and the second-best scores are underlined.

| Method                             | BLEU $\uparrow$ | EM $\uparrow$ | Levenshtein $\downarrow$ | MACCS FTS $\uparrow$ | RDk FTS $\uparrow$ | Morgan FTS $\uparrow$ | FCD $\downarrow$ | Text2Mol $\uparrow$ | Validity $\uparrow$ |
|------------------------------------|-----------------|---------------|--------------------------|----------------------|--------------------|-----------------------|------------------|---------------------|---------------------|
| Transformer (Edwards et al., 2022) | 0.499           | 0.000         | 57.66                    | 0.480                | 0.320              | 0.217                 | 11.32            | 0.277               | <b>0.906</b>        |
| T5-base (Edwards et al., 2022)     | 0.762           | 0.069         | 24.950                   | <u>0.731</u>         | <u>0.605</u>       | <u>0.545</u>          | 2.48             | <u>0.499</u>        | 0.660               |
| MolT5-base (Edwards et al., 2022)  | <u>0.769</u>    | <u>0.081</u>  | <b>24.458</b>            | 0.721                | 0.588              | 0.529                 | <u>2.18</u>      | 0.496               | 0.772               |
| GPT-3.5-turbo (zero-shot)          | 0.489           | 0.019         | 52.13                    | 0.705                | 0.462              | 0.367                 | 2.05             | 0.479               | 0.802               |
| GPT-3.5-turbo (10-shot MolReGPT)   | <b>0.790</b>    | <b>0.139</b>  | <u>24.91</u>             | <b>0.847</b>         | <b>0.708</b>       | <b>0.624</b>          | <b>0.57</b>      | <b>0.571</b>        | <u>0.887</u>        |

metric. Besides, the ROUGE-L score achieved by Morgan FTS is almost doubled compared to the Random or BM25 retrieval strategies. The use of Morgan FTS with Dice similarity shows a better estimation of the structural similarity between molecules by comparing unique structural features like functional groups. These features are usually revealed in molecule captions with detailed descriptions. In this case, retrieving similar molecules by Morgan FTS could effectively guide the LLM to learn the associations between molecule structures and caption descriptions, resulting in more accurate and desired outputs.

**Text-based Molecule Generation (Cap2Mol).** In the Cap2Mol task, we also employ three retrieval strategies: Random, SentenceBert, and BM25 (adopted in MolReGPT). The Random strategy still retrieves  $n$  random examples, while SentenceBert encodes captions as numerical vectors to compute their semantic similarity.

As shown in Table 4, we find that BM25 is better in the Cap2Mol task, despite the fact that SentenceBert has outperformed BM25 in many classical NLP text retrieval datasets. When  $n$  changes from 1 to 10,  $n$ -shot BM25 always achieves better BLEU, Exact Match, Levenshtein, and fingerprints scores than  $n$ -shot SentenceBert. As shown in the input caption (stage#1) of Figure 2, the input molecule captions tend to use phrases with dashes (-) like "2-methylphenyl" to connect the structure details of the molecule. Understanding such phrases plays a crucial role in generating correct molecule structures. In this case, retrieving similar texts while precisely matching these details significantly contributes to performance improvement. In contrast, SentenceBert, as a neural method, encodes an entire caption into a 1-D embedding vector, focusing more on semantic similarity rather than specific details. Consequently, BM25 is chosen as the retrieval strategy of MolReGPT in text-based

molecule generation.

All in all, in both sub-tasks, compared to random selection, the other retrieval strategies used in this paper can help improve  $n$ -shot generation results. These strategies enhance the overall performance metrics, underscoring the importance of thoughtful retrieval strategy design for achieving performance improvement in MolReGPT.

#### 4.3.2. IMPACT OF THE NUMBER OF EXAMPLES FOR IN-CONTEXT LEARNING

In this subsection, we study how the number of examples contained in the system prompt through in-context learning affects the performance.

**Zero-shot Performance.** In the zero-shot scenario, where no extra examples are included in the prompt for guiding LLMs for learning molecule-caption translation tasks, we utilize two special spans, '[MOLECULE\_MASK]' and '[CAPTION\_MASK]', to inform the LLMs of the desired output format, as shown in Figure 6 (in Appendix). In this case, the output of the LLMs can be conveniently filtered and further processed to satisfy the desired output specifications for molecule discovery.

After analyzing the zero-shot results of GPT-3.5-turbo in Tables 3 and 4, we can observe that OpenAI did include SMILES strings in their training corpus because it can generate basically valid SMILES representations of molecules based on zero-shot prompts, achieving a 0.802 validity score and a 0.479 Text2Mol score in molecule generation. However, it is important to note that the zero-shot results exhibit a performance level similar to a vanilla Transformer model. This observation provides evidence that GPT-3.5-turbo is not specifically trained on the molecule-caption translation task, thereby alleviating concerns regarding potential infor-

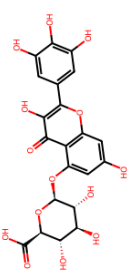
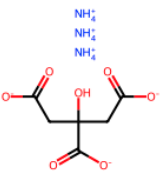
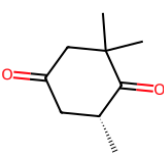
| Input   | Transformer  | MolT5  | Ours   | Ground Truth  |
|---|--|--|--|---|
| <p>1</p>   | <p>the molecule is the stable isotope of molybdenum with relative atomic mass 94. 905842, 15. 9 atom percent natural abundance and nuclear spin 5 / 2.</p> | <p>The molecule is a tetrasaccharide derivative consisting of a beta-D-galactopyranosyl residue attached to the mannose via an alpha-(2-&gt;3)-linkage. It is a member of pyranoses, a tetrasaccharide derivative and an amino cyclitol. It derives from a beta-D-Gal-(1-&gt;3)-beta-D-Glcp-(1-&gt;3)-beta-D-Galp.</p> | <p>The molecule is a myricetin O-glucuronide that is myricetin with a beta-D-glucosiduronic acid residue attached at the 3-position. It has a role as a metabolite. It is a myricetin O-glucuronide, a pentahydroxyflavone and a monosaccharide derivative.</p>  | <p>The molecule is a myricetin O-glucuronide that is myricetin with a beta-D-glucosiduronic acid residue attached at the 5-position. It has a role as a metabolite. It is a myricetin O-glucuronide, a pentahydroxyflavone, a member of flavonols and a monosaccharide derivative.</p>            |
| <p>2</p>   | <p>the molecule is the stable isotope of hydrogen with relative atomic mass 1. 007825, 3. 4 atom percent natural abundance and nuclear spin 1 / 2.</p>     | <p>The molecule is an alkanesulfonic acid comprising equimolar amounts of (-)-c1ccsc1=S. It has a role as a food emulsifier and a mammalian metabolite. It is an alkanesulfonic acid and a member of octa-2,4-dien-1-ols. It derives from a hydride of an equimolar amount of a (-)-c1ccn.</p>                         | <p>The molecule is a citrate salt in which two of the three carboxy groups are deprotonated and associated with ammonium ions as counter-cations. It has a role as a buffer. It is an ammonium salt and a citrate salt.</p>  | <p>The molecule is a citrate salt in which all three carboxy groups are deprotonated and associated with ammonium ions as counter-cations. It has a role as a food emulsifier and a buffer. It is an ammonium salt and a citrate salt.</p>  |
| <p>3</p>  | <p>the molecule is the stable isotope of molybdenum with relative atomic mass 94. 905842, 15. 9 atom percent natural abundance and nuclear spin 5 / 2.</p> | <p>The molecule is a member of the class of chromanes that is chromane which is substituted by a hydroxy group at position 3 and a methyl group at position 2. It is a member of chromanes and a hexacyclic triterpenoid. It derives from a chromane. It is a conjugate acid of a dehydrocytidine(1-).</p>             | <p>The molecule is a member of the class of hydroxycyclohexanones that is bornane-2,6-dione bearing an additional methyl substituent at position 3 and a gem-dimethyl group at position 2 (the 4R,6R-dia stereomer). It has a role as a bacterial metabolite. It is a cyclic 1,3-diketone, a hydroxycyclohexanone and a bornane monoterpene. It derives from a hydride of a bornane.</p> | <p>The molecule is a member of the class of cyclohexanones that is cyclohexane-1,4-dione carrying a gem-dimethyl group at position 2 and an additional methyl substituent at position 6 (the R-enantiomer). It has a role as a bacterial metabolite. It derives from a cyclohexane-1,4-dione.</p> |

Figure 4. Examples of molecule captions generated by different models, where SMILES strings are converted to molecule graphs for better visualization. Based on the same input molecule graph, our MolReGPT can generate accurate and natural captions to describe the structure, properties, and even the functions of the molecule. In contrast, Transformer generates meaningless captions that are far from the ground truth. Captions generated by MolT5-base seem better but still have many typo errors.

mation leakage.

**Few-shot Performance.** Table 3 and Table 4 list the comprehensive details of the experimental results, while Figure 8 (a) and Figure 8 (b) illustrate how the Text2Mol score changes when the number of examples increases.

Normally, the performance should improve as the number of examples, denoted as  $n$ , increases, as more examples provide additional knowledge for the task at hand. However, due to the input length limitation of LLMs, it is impossible to contain a large number of examples in the system prompt. Therefore, for few-shot scenarios, we choose four different values 1, 2, 5, and 10.

Tables 3 and 4 illustrate that performance generally improves as  $n$  increases in the system prompt through in-context learning. Significant performance enhancements are observed as  $n$  changes from 0 to 10. Taking Morgan FTS and BM25 as examples, in caption generation, we see

remarkable increases from 0.050 to 0.482, 0.204 to 0.543, and 0.352 to 0.560 in BLEU-4, ROUGE-L, and Text2Mol scores, respectively. Besides, BM25 improves molecule generation from 0.489 to 0.790 in the BLEU score and 0.479 to 0.571 in the Text2Mol score.

Besides, it is also interesting to notice that when  $n$  increases from 5 to 10, the Text2Mol metrics almost keep the same. This could be the problem of the maximum input length limitation of LLMs. To fit the input length limitation, we would remove the longest examples to degrade the  $n$ -shot generation to  $(n-1)$ -shot generation. As  $n$  increases, there is a higher possibility of exceeding the input length limitation. In this case, unless the maximum input length of the LLM is expanded, the performance will finally converge when  $n$  continues to grow.

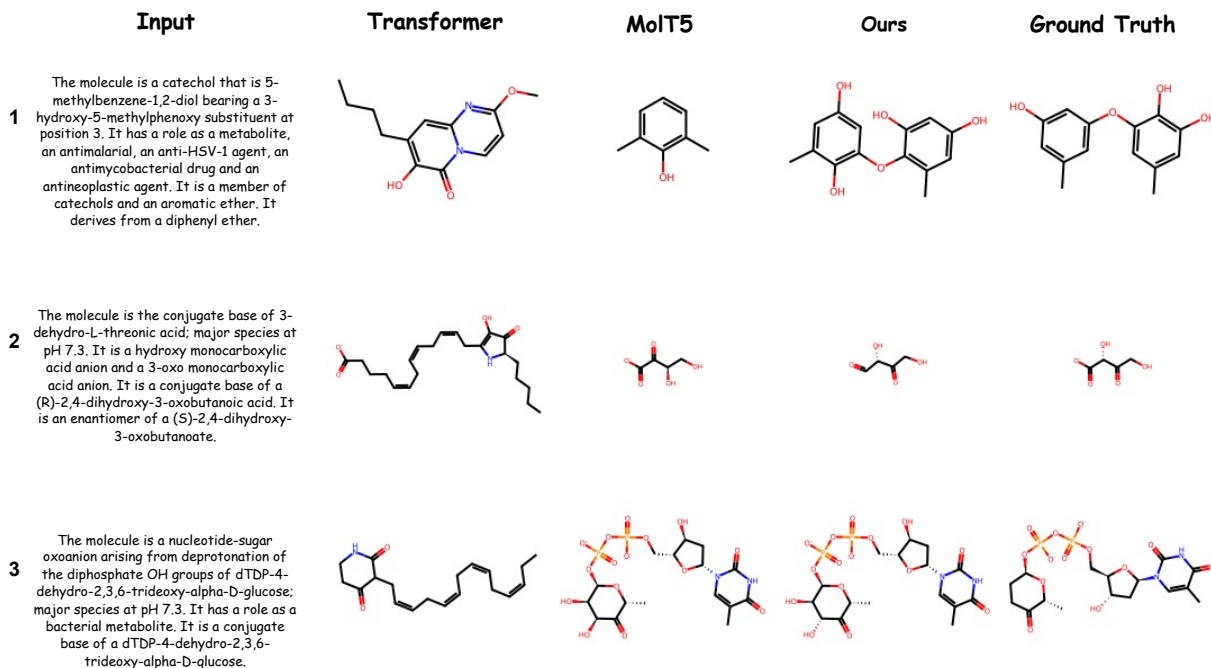


Figure 5. Examples of molecules generated by different models, where SMILES strings are converted to molecule graphs for better visualization. Based on the same input caption, our MolReGPT can generate accurate molecule graphs described by the caption. In contrast, Transformer generates quite different molecules compared to the ground truth. Compared to Transformer, molecules generated by MolT5-base are closer to the ground truth but still miss so many details.

## 5. Conclusion

In this work, we propose MolReGPT, a general retrieval-based prompt paradigm that empowers molecule discovery with LLMs like ChatGPT under In-Context Few-Shot Molecule Learning. MolReGPT leverages the molecular similarity principle to retrieve examples from a local database, guiding LLMs in generating n-shot outputs without fine-tuning. Our method is focused and evaluated on the task of molecule-caption translation, including molecule captioning (Mol2Cap) and text-based molecule generation (Cap2Mol). Specifically, BM25 is applied to retrieve similar molecule captions, while Morgan Fingerprints and Dice similarity are adopted to retrieve similar molecules. Experimental results show that our proposed MolReGPT can empower ChatGPT to achieve 0.560 and 0.571 Text2Mol scores in molecule captioning and molecule generation, respectively. The performance surpasses fine-tuned models like MolT5-base in both molecule understanding and text-based molecule generation aspects and is even comparable to the fine-tuned MolT5-large. To conclude, MolReGPT provides a novel and versatile paradigm to deploy LLMs in molecule discovery through in-context learning, which greatly reduces the cost of domain transfer and explores the potential of LLMs in molecule discovery.

## References

- Aizawa, A. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1): 45–65, 2003.
- Ali, O. A. A., Khan, M. U., Asghar, M. A., Mahmoud, S. F., El-Bahy, S. M., Baby, R., and Janjua, M. R. S. A. A new cyano (-cn) free molecular design perspective for constructing carbazole-thiophene based environmental friendly organic solar cells. *Physica B: Condensed Matter*, pp. 414630, 2023.
- Amabilino, S., Pogány, P., Pickett, S. D., and Green, D. V. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *Journal of Chemical Information and Modeling*, 60(12):5699–5713, 2020.
- Anderson, A. C. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.
- Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11(1):1–13, 2019.



Table 3. N-shot Molecule Captioning results on ChEBI-20 dataset. The **best** scores are in bold, and the second-best scores are underlined.

| Method               | BLEU-2 $\uparrow$ | BLEU-4 $\uparrow$ | ROUGE-1 $\uparrow$ | ROUGE-2 $\uparrow$ | ROUGE-L $\uparrow$ | METETOR $\uparrow$ | Text2Mol $\uparrow$ |
|----------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|---------------------|
| zero-shot            | 0.103             | 0.050             | 0.261              | 0.088              | 0.204              | 0.161              | 0.352               |
| 1-shot (random)      | 0.236             | 0.131             | 0.335              | 0.135              | 0.257              | 0.253              | 0.372               |
| 1-shot (BM25)        | 0.243             | 0.150             | 0.350              | 0.156              | 0.278              | 0.262              | 0.394               |
| 1-shot (Morgan FTS)  | 0.506             | 0.416             | 0.547              | 0.372              | 0.473              | 0.499              | 0.529               |
| 2-shot (random)      | 0.273             | 0.158             | 0.357              | 0.154              | 0.278              | 0.284              | 0.371               |
| 2-shot (BM25)        | 0.287             | 0.188             | 0.380              | 0.185              | 0.307              | 0.297              | 0.397               |
| 2-shot (Morgan FTS)  | 0.547             | 0.460             | 0.592              | 0.425              | 0.520              | 0.559              | 0.548               |
| 5-shot (random)      | 0.297             | 0.178             | 0.376              | 0.173              | 0.300              | 0.305              | 0.366               |
| 5-shot (BM25)        | 0.311             | 0.213             | 0.398              | 0.205              | 0.327              | 0.317              | 0.405               |
| 5-shot (Morgan FTS)  | <u>0.562</u>      | <u>0.478</u>      | <u>0.609</u>       | <u>0.446</u>       | <u>0.540</u>       | <u>0.583</u>       | <u>0.559(6)</u>     |
| 10-shot (random)     | 0.295             | 0.181             | 0.389              | 0.185              | 0.310              | 0.329              | 0.369               |
| 10-shot (BM25)       | 0.326             | 0.227             | 0.413              | 0.221              | 0.342              | 0.333              | 0.408               |
| 10-shot (Morgan FTS) | <b>0.565</b>      | <b>0.482</b>      | <b>0.623</b>       | <b>0.450</b>       | <b>0.543</b>       | <b>0.585</b>       | <b>0.559(8)</b>     |

Table 4. N-shot Molecule Generation results on ChEBI-20 dataset. The **best** scores are in bold, and the second-best scores are underlined.

| Method                 | BLEU $\uparrow$ | EM $\uparrow$ | Levenshtein $\downarrow$ | MACCS FTS $\uparrow$ | RDk FTS $\uparrow$ | Morgan FTS $\uparrow$ | FCD $\downarrow$ | Text2Mol $\uparrow$ | Validity $\uparrow$ |
|------------------------|-----------------|---------------|--------------------------|----------------------|--------------------|-----------------------|------------------|---------------------|---------------------|
| zero-shot              | 0.489           | 0.019         | 52.13                    | 0.705                | 0.462              | 0.367                 | 2.05             | 0.479               | 0.802               |
| 1-shot (random)        | 0.525           | 0.027         | 51.86                    | 0.716                | 0.475              | 0.373                 | 1.67             | 0.482               | 0.821               |
| 1-shot (SentenceBert)  | 0.687           | 0.066         | 35.89                    | 0.796                | 0.609              | 0.511                 | 0.85             | 0.541               | 0.839               |
| 1-shot (BM25)          | 0.706           | 0.074         | 33.38                    | 0.799                | 0.620              | 0.526                 | 0.84             | 0.540               | 0.842               |
| 2-shot (random)        | 0.529           | 0.026         | 49.87                    | 0.720                | 0.479              | 0.380                 | 1.71             | 0.483               | 0.824               |
| 2-shot (SentenceBert)  | 0.642           | 0.048         | 40.98                    | 0.770                | 0.560              | 0.463                 | 1.01             | 0.557               | 0.841               |
| 2-shot (BM25)          | 0.748           | 0.101         | 28.89                    | 0.827                | 0.668              | 0.578                 | 0.67             | 0.519               | 0.860               |
| 5-shot (random)        | 0.552           | 0.028         | 49.26                    | 0.720                | 0.476              | 0.382                 | 1.60             | 0.481               | 0.832               |
| 5-shot (SentenceBert)  | 0.758           | 0.095         | 28.34                    | 0.824                | 0.659              | 0.568                 | 0.71             | 0.558               | 0.871               |
| 5-shot (BM25)          | 0.771           | 0.121         | 26.78                    | 0.836                | 0.686              | 0.599                 | 0.60             | 0.564               | 0.882               |
| 10-shot (random)       | 0.564           | 0.029         | 49.11                    | 0.723                | 0.486              | 0.386                 | 1.46             | 0.484               | 0.846               |
| 10-shot (SentenceBert) | 0.767           | 0.098         | 27.46                    | 0.831                | 0.672              | 0.585                 | 0.63             | 0.562               | <b>0.890</b>        |
| 10-shot (BM25)         | <b>0.790</b>    | <b>0.139</b>  | <b>24.91</b>             | <b>0.847</b>         | <b>0.708</b>       | <b>0.624</b>          | <b>0.57</b>      | <b>0.571</b>        | <u>0.887</u>        |

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.

Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., and He, X. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. *arXiv preprint arXiv:2305.00447*, 2023.

Bran, A. M., Cox, S., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Butina, D. Unsupervised data base clustering based on daylight’s fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

Cao, Y., Yang, Z.-Q., Zhang, X.-L., Fan, W., Wang, Y., Shen, J., Wei, D.-Q., Li, Q., and Wei, X.-Y. Identifying the kind

behind smiles—anatomical therapeutic chemical classification using structure-only representations. *Briefings in Bioinformatics*, 23(5):bbac346, 2022.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Coupry, D. E. and Pogány, P. Application of deep metric learning to molecular graph similarity. *Journal of Cheminformatics*, 14(1):1–12, 2022.

Curtarolo, S., Hart, G. L., Nardelli, M. B., Mingo, N., Sanvito, S., and Levy, O. The high-throughput highway to computational materials design. *Nature materials*, 12(3): 191–201, 2013.

- Cuzzucoli Crucitti, V., Ilchev, A., Moore, J. C., Fowler, H. R., Dubern, J.-F., Sanni, O., Xue, X., Husband, B. K., Dundas, A. A., Smith, S., et al. Predictive molecular design and structure–property validation of novel terpene-based, sustainably sourced bacterial biofilm-resistant materials. *Biomacromolecules*, 2023.
- Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.
- Dice, L. R. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- Ding, B., Weng, Y., Liu, Y., Song, C., Yin, L., Yuan, J., Ren, Y., Lei, A., and Chiang, C.-W. Selective photoredox trifluoromethylation of tryptophan-containing peptides. *European Journal of Organic Chemistry*, 2019(46):7596–7605, 2019.
- Du, Y., Fu, T., Sun, J., and Liu, S. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.
- 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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.26>.
- Fan, W., Liu, C., Liu, Y., Li, J., Li, H., Liu, H., Tang, J., and Li, Q. Generative diffusion models on graphs: Methods and applications. *arXiv preprint arXiv:2302.02591*, 2023.
- Frey, N., Soklaski, R., Axelrod, S., Samsi, S., Gomez-Bombarelli, R., Coley, C., and Gadepally, V. Neural scaling of deep chemical models. *chemrxiv*, 2022.
- Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.
- Grisoni, F., Moret, M., Lingwood, R., and Schneider, G. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3):1175–1183, 2020.
- Gupta, Y., Savytskyi, O. V., Coban, M., Venugopal, A., Pleqi, V., Weber, C. A., Chitale, R., Durvasula, R., Hopkins, C., Kempaiah, P., et al. Protein structure-based in-silico approaches to drug discovery: Guide to covid-19 therapeutics. *Molecular Aspects of Medicine*, 91:101151, 2023.
- Hajduk, P. J. and Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery*, 6(3):211–219, 2007.
- Higuchi, A., Sung, T.-C., Wang, T., Ling, Q.-D., Kumar, S. S., Hsu, S.-T., and Umezawa, A. Material design for next-generation mRNA vaccines using lipid nanoparticles. *Polymer Reviews*, 63(2):394–436, 2023.
- 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.
- Hu, W., Liu, Y., Chen, X., Chai, W., Chen, H., Wang, H., and Wang, G. Deep learning methods for small molecule drug discovery: A survey. *IEEE Transactions on Artificial Intelligence*, 2023.
- Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.
- Le, N. Q. K., Yapp, E. K. Y., Ou, Y.-Y., and Yeh, H.-Y. imotor-cnn: Identifying molecular functions of cytoskeleton motor proteins using 2d convolutional neural network via chou’s 5-step rule. *Analytical biochemistry*, 575:17–26, 2019.
- Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., and Eger, S. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*, 2023.
- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.
- 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. *arXiv preprint arXiv:2212.10789*, 2022.
- Lv, Z., Li, W., Wei, J., Ho, F., Cao, J., and Chen, X. Autonomous chemistry enabling environment-adaptive electrochemical energy storage devices. *CCS Chemistry*, 5(1):11–29, 2023.
- Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, 2022.

- Osamor, V. C., Ikekanam, E., Bishung, J., Abiodun, T., and Ekpo, R. H. Covid-19 vaccines: Computational tools and development. *Informatics in Medicine Unlocked*, pp. 101164, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Patani, G. A. and LaVoie, E. J. Bioisosterism: a rational approach in drug design. *Chemical reviews*, 96(8):3147–3176, 1996.
- Peng, S.-P. and Zhao, Y. Convolutional neural networks for the design and analysis of non-fullerene acceptors. *Journal of Chemical Information and Modeling*, 59(12):4993–5001, 2019.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.
- 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.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- Robertson, S., Zaragoza, H., et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.
- Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.
- Thompson, K. Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422, 1968.
- Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- Urbina, F. and Ekins, S. The commoditization of ai for molecule design. *Artificial Intelligence in the Life Sciences*, 2:100031, 2022.
- 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, J., Hsieh, C.-Y., Wang, M., Wang, X., Wu, Z., Jiang, D., Liao, B., Zhang, X., Yang, B., He, Q., et al. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence*, 3(10):914–922, 2021.
- Wang, Z., Liang, L., Yin, Z., and Lin, J. Improving chemical similarity ensemble approach in target prediction. *Journal of cheminformatics*, 8:1–10, 2016.
- Wang, Z., Liu, T., Peng, H., and Fang, Y. Advances in molecular design and photophysical engineering of perylene bisimide-containing polyads and multichromophores for film-based fluorescent sensors. *The Journal of Physical Chemistry B*, 127(4):828–837, 2023.
- Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- Weng, Y., Ding, B., Liu, Y., Song, C., Chan, L.-Y., and Chiang, C.-W. Late-stage photoredox c–h amidation of n-unprotected indole derivatives: Access to n-(indol-2-yl) amides. *Organic Letters*, 23(7):2710–2714, 2021.
- White, A. D. The future of chemistry is language. *Nature Reviews Chemistry*, pp. 1–2, 2023.
- Xu, J., Li, Y., Yang, J., Zhou, S., and Situ, W. Plasma etching effect on the molecular structure of chitosan-based hydrogels and its biological properties. *International Journal of Biological Macromolecules*, pp. 123257, 2023.
- Yoshikai, Y., Mizuno, T., Nemoto, S., and Kusuhara, H. Difficulty in learning chirality for transformer fed with smiles. *arXiv preprint arXiv:2303.11593*, 2023.
- 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, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M.  
Minigt-4: Enhancing vision-language understanding  
with advanced large language models. *arXiv preprint  
arXiv:2304.10592*, 2023.



## A. Appendix

You are now working as an excellent expert in chemisrty and drug discovery.

Given the SMILES representation of a molecule, your job is to predict the caption of the molecule. The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production.

**Task Format:**  
 ...  
 Instruction: Given the SMILES representation of a molecule, predict the caption of the molecule.  
 Input: [MOLECULE\_MASK]  
 ...  
 Your output should be:  
 ...  
 {"caption": "[CAPTION\_MASK]}"  
 ...

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

You are now working as an excellent expert in chemisrty and drug discovery.

Given the caption of a molecule, your job is to predict the SMILES representation of the molecule. The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production. You can infer the molecule SMILES representation from the caption.

**Task Format:**  
 ...  
 Instruction: Given the caption of a molecule, predict the SMILES representation of the molecule.  
 Input: [CAPTION\_MASK]  
 ...  
 Your output should be:  
 ...  
 {"molecule": "[MOLECULE\_MASK]}"  
 ...

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Figure 6. System Prompt for zero-shot Molecule-Caption translation. The main structure of zero-shot prompts is almost the same as that of few-shot prompts. The main difference lies in that the **Example** part in few-shot prompts is changed to **Task Format** to pre-define the input and output format. To avoid information leaks, we use "[CAPTION\_MASK]" and "[MOLECULE\_MASK]" to denote the position of captions and molecules.

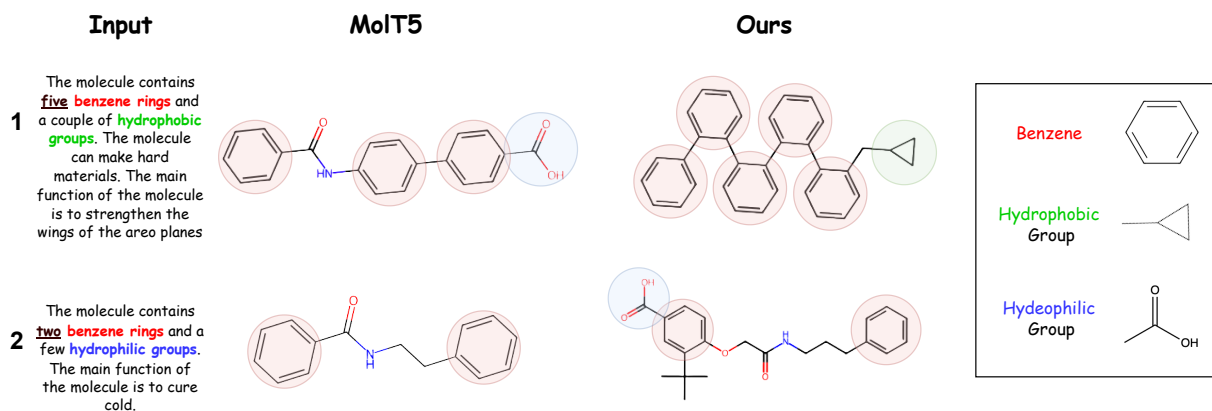
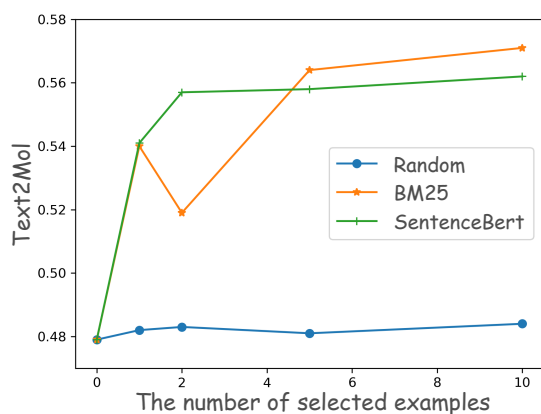
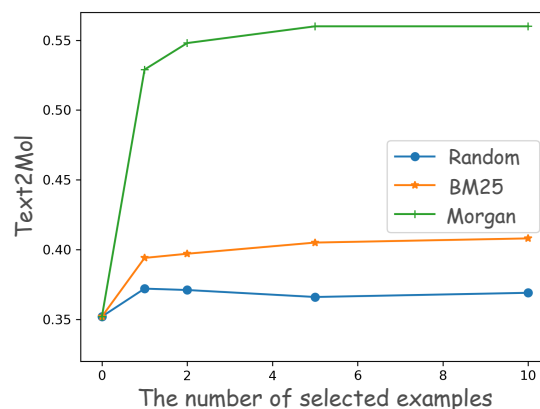


Figure 7. Illustrations of molecule graphs generated by MolT5 and our MolReGPT, given customized inputs. Notably, the key points in Example 1 highlight the **five benzene rings** and **hydrophobic groups** in the structure, which are correctly generated by our MolReGPT. In contrast, the results of MolT5 generate the incorrect number of **benzene rings** and contain a few **hydrophilic groups**. In example 2, both generations give the correct number of benzene rings, while MolReGPT generates more hydrophilic groups, which are closer to our input caption.



(a) Text2Mol metric comparison of caption retrieval strategies with respect to the change of the number of selected examples in the *Cap2Mol* task.



(b) Text2Mol metric comparison of molecule retrieval strategies with respect to the change of the number of selected examples in the *Mol2Cap* task.

Figure 8. The trend of the Text2Mol metric with respect to the number of examples (i.e.,  $n$ ). Basically, as  $n$  increases, the  $n$ -shot performance is also improved. However, when  $n$  increases from 1 to 2, we see a clear performance drop in molecule generation, which is possibly the reason that the noise brought by the added examples exceeds the information gain they could bring. Besides, in caption generation, we see a remarkable increase by comparing Morgan Fingerprints to other retrieval strategies, showing the superiority of Morgan Fingerprints-based molecule retrieval. It is also interesting to notice that when  $n$  increases from 5 to 10, the Text2Mol metrics almost keep the same. This is the problem of the maximum input length limitation of LLMs. To fit the input length limitation, we would remove the longest examples to degrade the  $n$ -shot generation to  $(n-1)$ -shot generation. As  $n$  increases, there is a higher possibility of exceeding the input length limitation. In this case, unless the maximum input length of the LLMs is expanded, the performance will finally converge when  $n$  continues to grow.