MOLTEXTQA: A CURATED QUESTION-ANSWERING DATASET AND BENCHMARK FOR MOLECULAR STRUCTURE-TEXT RELATIONSHIP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in AI have significantly enhanced molecular representation learning, which is crucial for predicting molecule properties and designing new molecules. Despite these advances, effectively utilizing the vast amount of molecular data available in textual form from databases and scholarly articles remains a challenge. Recently, a large body of research has focused on utilizing Large Language Models (LLMs) and multi-modal architectures to interpret textual information and link it with molecular structures. Nevertheless, existing datasets often lack specificity in evaluation, as well as direct comparisons and comprehensive benchmarking across different models and model classes. In this work, we construct a dataset specifically designed for evaluating models on structure-directed questions and textual description-based molecule retrieval, featuring about 500,000 question-answer pairs related to approximately 240,000 molecules from PubChem. Its structure enhances evaluation specificity and precision through the use of multiple-choice answers. Moreover, we benchmark various architectural classes fine-tuned using this dataset, including multi-modal architectures, and large language models, uncovering several insights. Our experiments indicate that the BioT5 and MoleculeSTM models are the top performers in Molecule QA and Molecule Retrieval tasks respectively, achieving about 70% accuracy. We have made both the dataset and the fine-tuned models publicly available.

1 INTRODUCTION

Drug discovery is an extensive and laborious process, costing billions of dollars due to extensive property validations, lab tests, animal studies, and human trials (Dickson and Gagnon, 2009). In the field, much of the valuable information about molecules is documented in text form, accessible through an extensive array of public databases. This textual information includes a wide range of data, from drug details (Wishart et al., 2018) and toxicity reports (Fonger et al., 2014) to the methods of molecule extraction, as well as their physical characteristics (Kim et al., 2019), information available in patents, chemical reactions (Lowe, 2017), and applications in producing various goods like materials, fertilizers, perfumes, and insecticides (Dionisio et al., 2018).

While deep learning methods have significantly impacted drug discovery by predicting specific molecular properties such as solubility or energy, these approaches face challenges when deciphering the information encoded in textual form. For example, identifying the potential side effects of a drug might depend heavily on analyzing narrative case studies, and patient reports documented in FDA reports (U.S. FDA, 2024), clinical-trial documents(CTGov, 2024), etc which are rich in textual data but not easily quantifiable through traditional regression or classification methods. This necessitates the need for methods capable of interpreting molecular structures and their relationship with textual information and performing inference based on text. By bridging this gap, we can significantly enhance our ability to understand molecules and unlock a broader spectrum of textual knowledge that remains largely untapped. This improved understanding, in turn, enables literature-based discovery of novel materials and drugs with potentially significant benefits for science and medicine.

Very recently, there has been a surge in the development of models to decipher the complex relationships between molecular structures and textual descriptions (Liu et al. (2023a;b); Li et al. (2024);

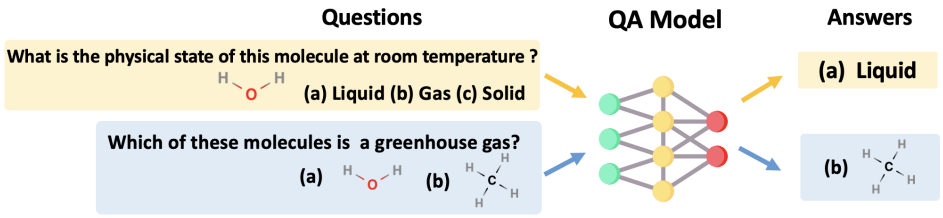

Figure 1: An example of a Question Answering task: In the first question, the objective is to infer certain information from a molecular structure. In the second question, the objective is to retrieve a molecule with properties that satisfy the prompt.

Edwards et al. (2022; 2021); Zeng et al. (2022)). Numerous methods have developed multi-modal frameworks, incorporating adaptations of models like CLIP Radford et al. (2021) and BLIP Li et al. (2023), which are specifically designed to learn correlations between visual content and text in the realm of molecular science Liu et al. (2023a;b). Additionally, the emergence of Scientific Large Language Models, such as Galactica Taylor et al. (2022), trained on vast troves of scientific data, represents a significant effort forward in harnessing computational power for molecular understanding and discovery.

Despite significant progress in model development, several challenges persist in the evaluation of these models. Existing datasets, such as those in Degtyarenko et al. (2007); Su et al. (2022); Liu et al. (2023a); Fang et al. (2024), often rely on free-form text generation or molecule/text retrieval tasks, which hinder the assessment of a model's ability to infer specific molecular properties. These datasets typically use generic prompts like "*Describe the molecule*," which fail to extract precise information. A more effective approach would involve using targeted questions, such as "*What is the physical state of the molecule at room temperature?*". Additionally, widely used evaluation metrics, such as the BLEU score in molecule captioning tasks, are inadequate for this context. Since answers to vague prompts like "*Describe the molecule*" can vary significantly, ranging from physical properties to industrial applications, BLEU's reliance on semantic similarity makes it unreliable. Further details on these methodological shortcomings are provided in section 2.2.

In this work, we have developed a comprehensive dataset consisting of over 500,000 question-and-answer (QA) pairs and small molecules represented by SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) sequences. These QAs are crafted from a rich base of textual data sourced from PubChem (Kim et al., 2019), encompassing a wide array of information such as chemical structures, physical properties, applications, and uses of molecules in drugs and biological pathways, as well as manufacturing details. We believe this dataset will significantly enhance the ability to infer information from molecular structures aid in the design of new molecules and improve the capability to evaluate these processes more effectively (See Figure 1). Our main contributions are:

1. We created a comprehensive dataset with about 500,000 QA pairs for 240,000 distinct molecule SMILES sequences across various categories, including multiple-choice answers to improve evaluation precision on specific areas including chemical information, biological information, physical properties, and more.
2. To ensure the reliability of our dataset, we implemented a comprehensive validation process that includes human annotation of a small subset to evaluate data accuracy.
3. We conduct an extensive evaluation of this dataset using state-of-the-art (SoTA) molecule-text multimodal models and recent advancements in large language models. Our analysis provides valuable insights into the advantages and limitations of current models, highlighting their performance in understanding and interpreting molecular data.

To the best of our knowledge, this work represents the first large-scale dataset and benchmarking effort dedicated to diverse question-answering methodologies for small molecules.

2 Related Work

This section overviews the key related works in molecule-text learning with a single model and separate encoders, language models and various benchmarking datasets. We first describe various models in 2.1, followed by datasets in 2.2.

2.1 Models Architectures in Molecule-Text learning

Molecule-Text learning with a Single Model: Initial models in molecule-text modeling combined molecules and texts using a unified encoder-only framework. Key works include KV-PLM (Zeng et al., 2022) and GPT-MolBERTa (Balaji et al., 2023), built on the foundations of SciBERT (Beltagy et al., 2019) and RoBERTa (Liu et al., 2019), respectively. KV-PLM fine-tunes SciBERT with SMILES sequences from 15,000 PubChem (Kim et al., 2019) descriptions, employing masked language modeling for representation learning. This model acts as a dual encoder for text and molecule representations. GPT-MolBERTa, using RoBERTa as its base, integrates descriptions from ChatGPT (OpenAI, 2022), which may introduce non-factual information, affecting reliability.

Molecule-Text with Separate Encoders: Recent and more effective models used multi-modal strategies by integrating separate encoders for molecules and text. For instance, Text2Mol (Edwards et al., 2021) employs SciBERT, akin to KV-PLM, for text processing, while molecule representation benefits from Mol2Vec (Jaeger et al., 2018) tokens as initial inputs, utilizing contrastive training to synchronize molecule and text embeddings, mirroring CLIP's (Radford et al., 2021) approach. Conversely, MoMu (Su et al., 2022) introduces molecules as graphs, leveraging a Graph Isomorphism Network (GIN) (Xu et al., 2018) for molecule encoding and uses SciBERT for text processing. MoleculeSTM (Liu et al., 2023a) enhances its methodology by training with both SMILES strings and molecular graphs, initializing SMILES encoder weights from MegaMolBART (NVIDIA, 2021) and adopting a GIN model for graph representation, with text decoding also relying on SciBERT. All these methodologies use of contrastive learning to align text and molecule representations, gaining advantages from expansive datasets and diverse molecular encoding techniques. Alternatively, MolT5 (Edwards et al., 2022) has fine-tuned a T5 model(Raffel et al., 2020) to train separate SMILES and text encoders and decoders. Recently, MolCA (Liu et al., 2023b) utilize the BLIP model (Li et al., 2023), employing a GINE model (Brossard et al., 2020) for graph encoding and SciBERT for text, and text generation fine-tuned by LoRA optimization of the Galactica model (Taylor et al., 2022). 3DMolLM(Li et al., 2024) extended this framework to include question prompts for more directed QA.

Decoder-based Scientific Large Language Models: The Galactica model (Taylor et al., 2022), trained explicitly on scientific data, supports a variety of scientific tasks such as generating molecular and protein sequences, question answering, code generation, and mathematical problem-solving. We also evaluate popular large language models (LLMs) like GPT-3.5 (OpenAI, 2022) and LLaMA (Touvron et al., 2023) in our work, noting their strong performance across many tasks, including those in the scientific domain.

2.2 Corresponding Benchmark Datasets for Molecule-Text learning

All the methods discussed so far draw upon datasets sampled from various public sources. For example, KV-PLM introduced PCdes, utilizing PubChem to compile a dataset of 15,000 samples. Additionally, KV-PLM incorporated a set of multiple-choice questions (MCQs) akin to those in this work, albeit with a smaller scope of approximately 1,500 entries, which, in context, was considered a significantly large dataset. Text2Mol developed the CheBI-20 dataset by sampling from the Chemical Entities of Biological Interest (ChEBI) (Degtyarenko et al., 2007) database, resulting in a collection of approximately 20,000 molecular descriptions. MoMu extracted 50,000 captions from PubChem, while MoleculeSTM significantly expanded this approach by sampling over 300,000 captions from PubChem to create the PubChemSTM dataset. 3DMolLM (Li et al., 2024) has curated a set of Question-answering pairs from these captions using ChatGPT.

Limitations in existing datasets: We identify several limitations in previous dataset curation methodologies that justify the development of our current dataset and benchmarking efforts. Detailed discussions of these issues, with specific examples, are presented in Appendix.

1. **Lack of Specificity in Prompts and Question Diversity**: Existing datasets, such as PubChem-STM(Liu et al., 2023a), MoMu(Su et al., 2022), CheBI-20(Edwards et al., 2021), and PCDes(Zeng et al., 2022), typically consist of captions extracted directly from PubChem(Kim et al., 2019).The captions are free-form, containing diverse information about various tasks without categorization. Other datasets like InstructMol(Cao et al., 2023) and MolInstructions(Fang et al., 2024), which also derive captions from these sources, feature queries such as *"Describe the molecule."* Such

Dataset	Diverse Questions?	Multiple choice?	Model Benchmark ?	Factual validity ?	Anonymized names ?
PCDes (Zeng et al., 2022)	✗	✗	✗	✓	✗
CheBI-20 (Edwards et al., 2021)	✗	✗	✗	✓	✓
MoMu (Su et al., 2022)	✗	✗	✗	✓	✗
PubChemSTM (Liu et al., 2023a)	✗	✗	✗	✓	✓
MolInstructions (Fang et al., 2024)	✗	✗	✓	✓	✓
InstructMol (Cao et al., 2023)	✗	✗	✓	✓	✗
3DMolLM (Li et al., 2024)	✓	✗	✗	✗	✗
MolTextQA (ours)	✓	✓	✓	✓	✓

Table 1: **Different molecule captioning or instruction datasets.**

prompts are insufficient for extracting specific information about molecules. A more targeted approach should involve using directed, specific questions.

2. **Evaluation of Text Generation:** Models such as MolT5(Edwards et al., 2022), MoMu(Su et al., 2022), MolCA(Liu et al., 2023b), 3DMolLM(Li et al., 2024) are trained to generate text for an input smile. When questions are not sufficiently directed, metrics like BLEU or ROUGE scores fail to effectively evaluate the responses if the content of generated answers varies significantly, such as between physical properties and industrial uses. Furthermore, answers like "*The state of the molecule is water*" versus "*The stage of the molecule is gas*" may score similarly, despite their vast differences. Employing QA multiple-choice questions can help mitigate these issues by restricting the output space.
3. **Factual Correctness, Information Leakage, Dataset Scope:** Datasets like 3DMolLM(Li et al., 2024) include directed questions, but the reliability of their answers is questionable since they augment the original data with LLM-generated information. Additionally, 3DMolLM does not anonymize molecule names in questions, which undermines the assessment of a model's ability to learn from molecular structure alone, as queries such as "*What are the physical properties of Aspirin?*" provide extra hints to the model. Datasets like MoMu(Su et al., 2022) and 3DMolLM(Li et al., 2024), also limit their test set to entries with large descriptions, thus neglecting lesser studied molecules.
4. **Benchmarking Across Several Model Classes:** Different architectures such as MolT5, MoleculeSTM, 3DMolLM, and language models like Galactica and Llama, which vary in architecture and hence input formats, make head-to-head comparisons across these models challenging.

We summarize the differences in these datasets in Table 1. In this work, we introduce MolTextQA, a comprehensive dataset designed to benchmark molecule-text relationship learning. MolTextQA features directed question-answering on small molecules and incorporates multiple-choice questions to enhance accuracy. Additionally, we benchmark molecule-text models across architectural classes to facilitate direct comparisons. It is important to note that this dataset is not intended to replace existing resources and can be used with other datasets to supplement training. The dataset also seeks to enhance model evaluation and enable more direct comparisons between different methodologies.

3 Building the MolTextQA Dataset

3.1 Data Source

Our work primarily utilizes the PubChem library (Kim et al., 2019), a resource overseen by the National Center for Biotechnology Information (NCBI) for cheminformatics and drug discovery research. Compiling information from over 750 sources, this dataset supports diverse drug discovery tasks like property prediction and repurposing. Adopting Liu et al.'s approach, we crawl PubChem to extract descriptions and SMILES strings for small molecules, covering aspects from chemical properties to biological effects. This enabled an in-depth analysis of molecule characteristics, enriching our dataset with diverse questions and answers. PubChem is freely available and licensed for non-commercial purposes. Licensing terms are elaborated in the Appendix.

Splits	Molecules	Total QAs	Physical Properties	Chemical Information	Biological Information	Source	Application
Pretrain	213336	421227	39512	183936	38757	145590	13415
Train	20000	54842	3569	28115	10448	11862	848
Valid	2500	5754	331	2809	908	1620	86
Test	5000	11922	691	5941	1990	3140	160
Total	240836	493,742	45091	224468	52914	165856	14670

Table 2: **Dataset Statistics:** Distribution of questions across different splits and categories

Attribute	Data
SMILES sequence	CC(=O)C
Question	What is the physical state of the molecule at room temperature?
Options	(a) Liquid (b) Solid (c) Gas
Correct Option	(a) Liquid
Sentence	The physical state of the molecule is liquid.
SMILES options	(a) CH4 (b) CC(=O)C (c) C(=O)([O-])[O-].[Ca+2]
Correct SMILE	(b) CC(=O)C
PubChem ID	180
Category	Physical Properties

Table 3: **A sample datapoint**

3.2 Dataset overview

The dataset comprises about 500,000 questions related to more than 240,000 molecules, categorized into five distinct areas:

1. **Chemical Information** covers the chemical structure, functional groups, and chemical properties.
2. **Physical Properties** addresses the properties such as solubility, physical state, and odor.
3. **Biological Information** contains the molecules' role in biological pathways, drug applications, and drug toxicity.
4. **Source** details the molecules' origin and manufacturing processes.
5. **Application** describes application areas such as perfumes, fertilizers, and insecticides.

We provide detailed dataset statistics across these QA categories in Table 2. Further, in Table 3, we depict a sample data point indicating different attributes. Each data point includes a question, options, a SMILES sequence, a set of candidate answers, a question category, the correct choice, and options in sentence form. Each datapoint also includes a set of candidate SMILES randomly sampled from the data and useful for the Molecule Retrieval task. A Tanimoto similarity threshold of 0.2 was applied during sampling to ensure that the selected molecules are sufficiently distinct from one another. The dataset is divided into test, valid, train and pretrain sets for different stages of model development and evaluation.

3.3 Data construction

In this work, we employed different LLMs, specifically Llama3-70B, Llama3-8B, and ChatGPT3.5, for constructing a question-and-answer dataset. The methodology involves passing molecular descriptions to the LLMs and prompting them to generate a set of questions with multiple-choice options that are semantically related to ensure challenging and informative questions. To prevent the possibility of inferring information solely from the molecule's common name, we specifically prompt the LLM to anonymize the common name of the molecule in the resulting QAs. This ensures that downstream models must rely on molecule structure alone for inference. Due to the expensive nature of the Llama3-70B API, the pretrain split was generated with Llama3-7B, while the other splits are generated using the Llama3-70B.

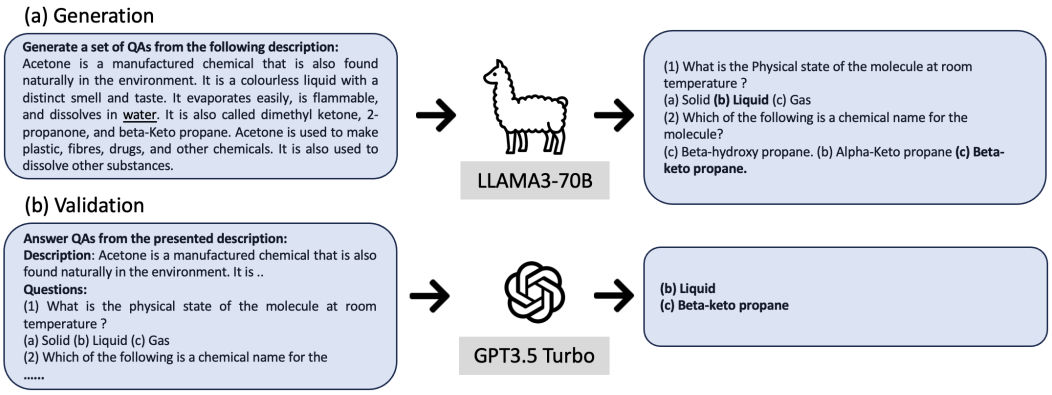

Figure 2: **Procedure for data generation:** (a) molecule captions are used to generate a set of QAs using Llama3-70B. (b) The generated QAs are validated for correctness using GPT 3.5.

Additionally, we instructed the LLMs to produce a short-sentence answer for each question, to facilitate text-generation tasks. This process is depicted in Figure 2(a). Each generated question was then categorized into one of five distinct categories to streamline the evaluation process. The specifics of the prompts utilized in this process and LLM versions are documented in the Appendix. It is important to note that LLMs are not utilized as sources of factual information; rather, only to transform data into a different format.

3.4 Dataset validation

To enhance the reliability of the generated question-and-answer content and mitigate the risk of fabricated information ("hallucinations") or uninformative, nonsensical questions, we implemented a two-pass validation process. First, following a methodology similar to that outlined by Es et al., we employed an alternative language model (GPT-3.5) to verify the accuracy of the QA pairs. In this step, the caption, question, and multiple-choice options were provided to the model, which was tasked with selecting the correct answer based on the given context. QA pairs where the model either failed to choose the correct answer or could not deduce it from the provided information were excluded from our dataset, ensuring that the content is grounded in the factual information from the captions. In the second stage, the remaining questions were passed to the LLM for an additional verification step. This phase focused on filtering out questions that were unanswerable, uninformative, or not useful for evaluating molecular characteristics and applications. Detailed descriptions of the prompts used in both stages can be found in the Appendix. Figure 2 depicts the data generation process.

3.5 Benchmarking Dataset efficacy

To assess the accuracy of language models in generating question-and-answer pairs, we manually reviewed a random sample of 400 QA pairs from the test set. Our evaluation focused on three key criteria: whether the question could be logically derived from the provided caption, the unambiguous correctness of the answer, and the relevance of the question—specifically, ensuring it avoids uninformative queries and contributes to meaningful chemical or biological insights based on the structure. We have found that 391 of the 400 samples satisfy this criteria. This corresponds to a greater than 96.13 percent accuracy on the entire dataset with a p-value of <0.05 under a hypergeometric test, indicating statistically significant performance. We elaborate the specifics of the hypothesis test and human verification in Appendix. We also include these samples with the supplementary material.

4 Benchmarking on MolTextQA

This section evaluates the MolTextQA dataset with various models, as detailed in subsection 4.1, which includes both molecule-text multi-modal and scientific language models. For model specifics and training details, see Appendix. The tasks and objectives for the benchmark are outlined in section

4.2. We then assess model performance in a zero-shot setting (4.3) and 4.4 covers model fine-tuning and their performance evaluation.

4.1 Benchmarked Models in Experiments

1. **Single Encoder Architectures:**

   SciBERT (Beltagy et al., 2019) leverages a BERT base, fine-tuned on scientific papers, especially from biomedical fields. It has superior performance on several scientific tasks like entity recognition and text classification.

   KV-PLM (Zeng et al., 2022) builds on SciBERT, trained further with molecule-text pairs from PubChem. It incorporates pre-training with SMILES and fine-tuning for text retrieval, employing a max hinge loss for improved prediction and retrieval.

   MolT5 (Edwards et al., 2022) is a T5 model fine-tuned for molecule captioning. It features an encoders and decoder architecture for both SMILES and captions.

   BioT5 Pei et al. (2023) is a T5 model similar to MolT5. However, the training dataset includes a larger training set and modalities including small molecules and proteins.

   BioT5 Plus Pei et al. (2024) is an updated BioT5 model including additional data sources and multi-task instruction tuning.

2. **Separate Encoder Architectures**

   MoleculeSTM (Liu et al., 2023a) employs a dual-encoder combining SciBERT (Beltagy et al., 2019) and a Graph Isomorphism Network (GIN) (Xu et al., 2018) for encoding text and chemical structures, respectively. It trains on PubChem data, using InfoNCE loss to refine molecule-text alignment.

   MoMu (Su et al., 2022) is similar to MoleculeSTM, yet introduces an extra contrastive loss between molecules in addition to the molecule-text contrastive loss.

3. **Large Language Models (Decoder-Only)**

   Llama (Touvron et al., 2023) includes autoregressive transformer models, fine-tuned for instruction adherence showing versatility in several general tasks from QA to code generation. We experiment with 4 model from the Llama2 and Llama3 series - Llama2-7B, Llama2-70B, Llama3-8B, Llama3-70B.

   GPT-3.5 Turbo (OpenAI, 2022), part of the GPT series, is optimized for human alignment, with broad application across numerous tasks.

   Galactica (Taylor et al., 2022) features autoregressive, decoder-only models trained on scientific content, effective in specialized biomedical datasets. We experimented with multiple sizes of the model - 125M, 1.3B, and 6.7B.

4.2 Tasks and Objectives

1. **Molecule QA task**: The first task, "Molecule QA", entails choosing the right answer from multiple options, where models are given a SMILES string or molecular graph of a molecule and a related question. The goal is to select the correct option, testing the model's capability to recognize and deduce molecular properties or characteristics from the molecule's structure.

   While large decoder-only models can be directly prompted to obtain answers to such queries, this approach is not feasible for the encoder-based models (e.g. KV-PLM, MoleculeSTM, MoMU) since they cannot generate text. For these models, this is framed as a sentence retrieval task from SMILES input, where each sentence is a "question + answer choice."

2. **Molecule Retrieval task:** The second task, known as the "Molecule Retrieval Task," reverses the domains. The model is given a sentence description of a molecule and must identify the correct SMILES string from among a list of candidates. This task evaluates the model's ability to accurately retrieve structures, useful for generating new molecules with target properties. For decoder-only LLMs, the task involves presenting the model with a sentence and a set of SMILES options, from which the model is prompted to retrieve the correct SMILES string. For encoder models, the task remains similar: retrieving the appropriate SMILES strings based on a sentence.

In both settings, the performance is measured by Accuracy, which is defined as the proportion of times the correct option is picked as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

4.3 RESULTS ON ZERO-SHOT INFERENCE

Model	Entire dataset	Physical Properties	Chemical Info	Biological Info	Sources	Uses
Molecule QA						
SciBERT	21.26	20.16	21.52	19.02	22.36	22.46
KV-PLM	29.86	27.90	32.03	26.61	28.51	26.74
MoleculeSTM	44.68	30.17	48.02	30.90	51.28	31.02
MoMu	44.93	28.84	49.60	30.05	50.31	27.81
gpt3.5	40.30	44.73	38.54	43.29	40.24	47.59
llama3-8b	16.16	20.03	16.08	17.27	14.07	27.27
llama3-70b	58.24	63.28	57.07	56.91	59.58	66.31
llama2-7b	24.57	22.70	25.31	24.86	23.46	24.60
llama2-70b	28.79	30.57	27.11	36.15	26.39	37.43
Random	20.69	20.55	22.78	21.11	20.18	19.49
Molecule Retrieval						
SciBERT	21.22	21.36	21.93	21.12	19.84	22.99
KV-PLM	48.53	47.80	60.01	43.09	30.47	54.01
MoleculeSTM	67.38	49.13	76.98	53.02	63.57	54.55
MoMu	66.02	45.79	76.51	51.12	61.51	50.27
gpt3.5	38.31	39.92	47.02	31.10	26.36	36.90
llama3-8b	20.96	22.16	21.93	20.12	19.68	16.04
llama3-70b	52.72	41.26	70.32	38.14	32.63	39.57
llama2-7b	18.58	19.36	19.24	17.57	17.97	16.04
llama2-70b	20.35	16.56	21.93	18.22	19.93	15.51
Random	20.28	20.43	21.12	19.84	20.58	19.17

Table 4: **Zero-Shot Setting Accuracy**:The Molecule QA task requires selecting the correct option from a set, given a SMILES string and a related question. The Molecule Retrieval task involves choosing the correct SMILES from candidates, based on a molecular property description.

The results for the Molecule QA and Molecule Retrieval tasks under zero-shot conditions are outlined in Table 4. These evaluations involved models that were not trained on the question-answering dataset introduced in this study. However, models such as MoleculeSTM, MoMu, and KV-PLM, which were trained using a dataset similar to the one presented here, which may partly explain their effectiveness. To avoid data leakage, MoleculeSTM and MoMu were retrained, explicitly excluding test samples from their training sets. For the LLMs, the public checkpoints were prompted with QA tasks; details of these prompts are provided in the Appendix. Results for the Galactica, MolT5, and BioT5 models are omitted in zero-shot scenarios as they are not aligned for question-answering tasks.

In the Molecule QA task, which requires answering questions based on SMILES inputs, both decoder-only LLMs and multimodal architectures exhibited similar levels of performance. Here, Llama3-70B emerged as the top performer. MoleculeSTM, MoMu, and GPT3.5 showed comparable results. In contrast, smaller models such as Llama-7b and SciBERT approximated random guessing performance, highlighting their limited applicability. Notably, LLMs performed better in predicting physical properties and uses (e.g., appearance, odor), whereas multimodal architectures excelled in processing chemical information. This suggests that LLMs are adept at handling diverse data types, while multimodal systems effectively leverage structural data inherent in molecular representations useful for inferring chemical information.

For the Molecule Retrieval task, multimodal architectures, particularly MoleculeSTM, significantly outperformed LLMs, accurately retrieving SMILES strings in over 66% of instances. The marginal superiority of MoleculeSTM over MoMu suggests that 3D pretraining initialization may enhance

its retrieval capabilities. Multimodal architectures consistently demonstrated higher accuracy in identifying chemical properties compared to other types similar to the MoleculeQA task. Among the LLMs, only Llama3-70B achieved notable performance, with an accuracy exceeding 50%.

4.4 Results on Finetuned models

Finally, we discuss the outcomes of finetuning our models in Table 5 using the training subset from the proposed dataset. We finetuned the Llama3-8B, Llama2-7B, Galactica and MolT5 models, alongside the top multimodal architectures, MoMu and MoleculeSTM, by fine-tuning them with this subset. For MolT5, we have fine-tuned both the pretrained checkpoint and the checkpoints specifically for SMILES generation or caption generation. We did not fine-tune larger LLMs due to the high costs associated with training. We faced challenges in fine-tuning the architecture proposed in 3DMolLM (Li et al., 2024), which are discussed in the Appendix.

Model	Entire dataset	Physical Properties	Chemical Info	Biological Info	Sources	Uses
Molecule QA						
MoleculeSTM	65.14	68.62	61.86	65.35	69.93	71.12
MoMu	65.08	70.76	60.69	66.65	70.56	71.66
Llama3-8b	60.41	64.35	58.67	64.10	60.73	55.08
Llama2-7b	41.84	42.06	43.97	41.64	38.21	37.43
Galactica-125m	43.97	43.39	43.13	43.58	46.29	37.43
Galactica-1.3b	60.98	62.62	58.60	62.41	64.85	48.66
Galactica-6.7b	69.01	70.36	65.73	72.99	72.52	65.24
Molt5-large	34.15	30.57	34.27	38.34	32.56	26.74
Molt5-large-s2c	34.69	47.26	31.16	37.89	36.49	31.55
BioT5	75.10	81.04	73.70	73.79	77.51	68.45
BioT5-plus	71.16	69.35	68.08	72.59	78.42	72.73
Random	20.69	20.55	22.78	21.11	20.18	19.49
Molecule Retrieval						
MoleculeSTM	65.27	59.95	72.39	54.57	60.17	62.03
MoMu	63.6	56.34	70.52	53.27	59.23	57.75
Llama3-8b	20.6	19.76	20.67	20.07	20.9	22.46
Llama2-7b	20.58	20.43	20.03	20.77	21.49	21.39
Galactica-125m	21.62	28.04	21.57	20.92	20.09	31.02
Galactica-1.3b	22.17	29.64	22.45	21.67	19.71	31.02
Galactica-6.7b	22.30	30.44	22.6	22.22	19.28	33.16
Molt5-large	23.54	39.79	23.89	23.36	18.18	41.18
Molt5-large-c2s	23.00	32.31	23.86	21.32	19.87	29.95
BioT5	23.34	37.63	21.29	21.14	17.26	33.16
BioT5-plus	22.30	31.24	21.286	19.46	17.75	27.72
Random	20.28	20.43	21.12	19.84	20.58	19.17

Table 5: **Finetuning performance:** Accuracy of different models in the finetuning setting, in both Molecule QA and Molecule Retrieval tasks.

Upon fine-tuning, all LLMs have exhibited reasonable performance in the molecule QA task. Most notably, the Llama3-8B model showed a 45% improvement in accuracy, surpassing the best performing zero-shot Llama3-70B model from the previous section. The BioT5 model emerged as the best performer overall with a 75% accuracy, suggesting that they can be effectively fine-tuned for question answering tasks. However, this also highlights room for improvement, which future works should focus on. Additionally, the size of the LLMs appears to be advantageous, indicating that scaling could further enhance performance. The results overall indicate that LLMs can be effective in infering molecular properties. On the other hand, multimodal architectures also demonstrated a considereable improvement across categories, increasing performance by 20%. However, the MolT5 models, while outperforming random benchmarks, lagged behind other models. The architecture

of MolT5 is similar to BioT5; however, the notable performance gap between them highlights the advantages of BioT5's multi-modal training approach.

For the Molecule Retrieval tasks, all LLMs performed close to randomly, indicating that this architecture might not be well-suited for molecule generation tasks, though it can be useful for inferring properties. In contrast, multimodal architectures consistently outperformed LLMs. Notably, while these architectures showed superior performance in chemical properties in zero-shot settings, fine-tuning enabled them to excel across categories. Despite significant gains in tasks like Physical properties (by over 10%), this has also led to a slight decline in performance for Chemical properties and consequently, overall. This performance dip suggests that embeddings may not fully capture diverse types of information, leading to trade-offs. We also observe a similar trend between BioT5 and BioT5-plus, where BioT5-plus demonstrates improved performance in the "Sources" and "Uses" categories but shows a decline in accuracy for "Physical" and "Chemical Properties." This highlights the need for improved modeling strategies to achieve robust performance across all tasks.

It is important to note that these fine-tuning efforts utilized only a small portion of the full dataset, raising questions about potential outcomes if the entire dataset were employed. Given the success of LLMs in Molecule QA tasks and multimodal architectures in Molecule Retrieval tasks, we speculate that an architecture fine-tuned on QA tasks, which integrates elements from both types of architectures and employs strategies to capture diverse sorts of information, could excel in both scenarios. Exploring this possibility will be a focus of our future research.

5 Conclusion

Our work introduces MolTextQA, a novel dataset featuring over 500,000 QA pairs related to small molecule structures, covering a broad spectrum of molecular properties and applications. We have benchmarked MolTextQA against a diverse array of large language models and state-of-the-art multimodal architectures, analyzing their strengths and weaknesses. We also see potential in extending our approach to include other scientific modalities, like proteins, to widen the dataset's applicability. We aspire for MolTextQA to become a foundational resource for the development of more efficient molecule-text foundation models. We anticipate its application in drug discovery, materials science, and other fields.

We acknowledge a few limitations in the dataset creation process. First, the dataset exhibits minor inaccuracies, as reported in Section 3.5, which could potentially complicate model training. Future work will focus on implementing more rigorous validation strategies to enhance data precision. Additionally, the dataset includes a broad spectrum of questions, a small fraction of which may appear straightforward, such as identifying the polarity of a molecule—an aspect that only requires a basic knowledge of chemistry. While this diversity of questions benefits model training, it also suggests we could refine our selection process to ensure each question more effectively serves the dataset's purpose. Furthermore, the dataset covers data across a wide range of general categories. The exploration of data acquisition and fine-tuning of specialized datasets for niche applications is an area that requires further investigation.

6 Ethics Statement

All data, code, and models used in this research have been sourced from public domains that are freely distributable, ensuring our adherence to ethical standards of transparency and accessibility. We elaborate on the licensing of the dataset source (PubChem) in Appendix B. Given the long history of research within the biomedical and cheminformatics communities, our work aligns with established practices. We do not foresee any ethical concerns regarding our work.

7 Reproducibility

The dataset introduced in this work, along with the code used to generate the benchmark, is publicly available in the project repository (`https://anonymous.4open.science/r/MolTextQA-D688/`). Further details on the prompts utilized for dataset generation can be found in Appendix D, while Appendix F provides comprehensive information on the models employed in the benchmarking process

REFERENCES

scipy.stats.hypergeom — Hypergeometric Distribution. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.hypergeom.html.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, et al. Gpt-molberta: Gpt molecular features language model for molecular property prediction. *arXiv preprint arXiv:2310.03030*, 2023.

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

Rémy Brossard, Oriel Frigo, and David Dehaene. Graph convolutions that can finally model local structure. *arXiv preprint arXiv:2011.15069*, 2020.

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

CTGov. ClinicalTrials.gov, 2024. URL https://clinicaltrials.gov/.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1): D344–D350, 2007.

Michael Dickson and Jean Paul Gagnon. The cost of new drug discovery and development. *Discovery medicine*, 4(22):172–179, 2009.

Kathie L Dionisio, Katherine Phillips, Paul S Price, Christopher M Grulke, Antony Williams, Derya Biryol, Tao Hong, and Kristin K Isaacs. The chemical and products database, a resource for exposure-relevant data on chemicals in consumer products. *Scientific data*, 5(1):1–9, 2018.

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

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.26.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024. URL https://openreview.net/pdf?id=Tlsdsb6l9n.

George Charles Fonger, Pertti Hakkinen, Shannon Jordan, and Stephanie Publicker. The national library of medicine's (nlm) hazardous substances data bank (hsdb): background, recent enhancements and future plans. *Toxicology*, 325:209–216, 2014.

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

Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 3d-molm: Towards 3d molecule-text interpretation in language models. In *ICLR*, 2024. URL https://openreview.net/forum?id=xI4yNlkaqh.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

Daniel Lowe. Chemical reactions from us patents (1976-sep2016). *Figshare https://doi. org/10.6084/m9. figshare*, 5104873(1), 2017.

NVIDIA. Megamolbart. https://github.com/NVIDIA/MegaMolBART, 2021.

OpenAI. Chatgpt-3.5, 2022. URL https://openai.com/. [Software]. Available: https://openai.com/.

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

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*, 2024.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

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

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

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

U.S. FDA. FDALabel: Full-text search of drug product labeling, 2024. URL `https://www.fda.gov/science-research/bioinformatics-tools/fdalabel-full-text-search-drug-product-labeling`.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

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

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

A Dataset Description

The MolTextQA dataset will be made publicly available upon acceptance and will be hosted with a permanent availability using a DOI identifier. The dataset is organized in a CSV file format, which is a standard and widely used format in machine learning applications. The dataset is available at - `https://anonymous.4open.science/r/MolTextQA-D688/`

A.1 Dataset Structure

Each data point in the dataset contains the following fields:

- **CID**: The PubChem Identifier of the molecule.
- **QID**: The identifier of the question within a CID.
- **Category**: The category of the data point, following this convention:
  1. Physical properties
  2. Chemical information
  3. Biological uses
  4. Sources
  5. General applications
- **Sentence**: A sentence summary of the question and answer.
- **Question**: The actual question asked.
- **Options**: A set of options for the answer, of which one is correct.
- **Correct_option**: The index (1-based) of the correct option.
- **Retrieval_options**: A set of PubChem IDs used for molecule retrieval from the sentence task.
- **Retrieval_correct**: The correct option in the retrieval task.

For benchmarking and further details regarding the application of this dataset in machine learning tasks, please refer to the project repository. The model weights will be made available upon acceptance and all the results are reproducible. The predictions file and the codes used for generating the results in the benchmark are also available in the repository.

- Dataset and Benchmark Repository: `https://anonymous.4open.science/r/MolTextQA-D688/`

A.2 Intended Uses of the Dataset

The dataset is primarily intended to be used for molecule-text relationship learning. The task of molecule-text learning has been gaining increasing attention in recent research. However, the current datasets and developed models do not enable structured inference, and evaluation is not precise. The MolTextQA dataset addresses these challenges by offering a question-answering format with multiple-choice answers. Questions are based on a small molecule input, with answers provided in textual sentence or multiple-choice format. The dataset is intended for applications in fields such as drug discovery, retrosynthesis, and the discovery of materials like fertilizers, pesticides, and perfumes.

A.3 License

The MolTextQA dataset will be distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

A.4 Rights and Responsibilities

The authors bear all responsibility in case of violation of rights associated with the dataset.

B DATASET SOURCE

The primary source for building this dataset is PubChem (Kim et al., 2019). PubChem consolidates chemical data from multiple public sources. A comprehensive list of these sources is accessible at `https://pubchem.ncbi.nlm.nih.gov/sources/`. The licensing terms on PubChem are stated as: "Works produced by the U.S. government are not subject to copyright protection in the United States. Any such works found on National Library of Medicine (NLM) Web sites may be freely used or reproduced without permission in the U.S. Please acknowledge NLM as the source of the information by including the phrase "Courtesy of the National Library of Medicine" or "Source: National Library of Medicine." More details on the licensing terms can be found at `https://www.nlm.nih.gov/web_policies.html`.

For efficient data retrieval, the PubChem Power User Gateway offers abstracts of compound records in XML format. This facilitates the extraction and analysis of chemical information by enabling users to search for molecular descriptions and their unique PubChem Compound Identifier (CID). This CID is then used to fetch the Simplified Molecular-Input Line-Entry System (SMILES) representation for each compound listed in PubChem. For utilizing the PubChem Power User Gateway, visit `https://pubchem.ncbi.nlm.nih.gov/docs/power-user-gateway`. This approach for obtaining PubChem data is also followed by (Fang et al., 2024; Cao et al., 2023; Li et al., 2024), all publicly available.

C LIMITATIONS OF EXISTING DATASETS

In this section, we discuss some limitations of existing datasets, illustrating each point with concrete examples.

Lack of Specificity in Prompts and Question Diversity: Existing datasets such as PubChemSTM (Liu et al., 2023a), MoMu (Su et al., 2022), CheBI-20 (Edwards et al., 2021), and PCDes (Zeng et al., 2022) predominantly consist of captions derived from PubChem (Kim et al., 2019). These captions are free-form and include diverse information across various tasks without specific categorization. For instance, consider the following captions from PubChem:

- Caption A: *This molecule is a natural product found in Carica papaya.*
- Caption B: *It is an N-glycosyl compound, a ribose triphosphate, a pyrimidone, and an aminopyrimidine.*

Datasets such as InstructMol (Cao et al., 2023) and MolInstructions (Fang et al., 2024), which also source their captions from similar databases, pose queries like *'Describe the molecule.'* Given the broad range of information in the captions—from molecule manufacturing, to physical properties, to chemical structures, to drug toxicity, to industrial applications—the queries remain insufficiently structured for detailed inference. In contrast, our proposed dataset includes specific queries such as *Is this molecule denser than water?*, *Does this molecule contain a mannose ring?*, or *Is this molecule an antibiotic or an analgesic?*

This issue also affects retrieval models such as MoleculeSTM or MoMu, where the challenge is compounded by the necessity for a single molecule embedding to retrieve both Caption A and Caption B. This task is difficult as these captions semantically reside in distinct spaces.

Factual Correctness: Datasets such as 3DMolLM(Li et al., 2024) employ data augmentation from LLMs in their data generation procedure, raising concerns about the overall accuracy of the dataset. Moreover, these datasets generate five questions for each data point, irrespective of whether sufficient information exists. There is no validation process to determine the reliability of the data, leading to the generation of numerous unreliable questions. Consider the following example from PubChem: `https://pubchem.ncbi.nlm.nih.gov/compound/10008613`

The provided caption is:

> $(1S,2S,4R,5R,6R,9R,10S,11R,12R,16R,18S,21R)-2,9,10,11-tetrahydroxy-4,6,12,17,17-pentamethyl-18-[(2S,3R,4S,5R)-3,4,5-trihydroxyoxan-2-yl]oxyhexacyclo[11.9.0.0^{1},21.0^{4},12.0^{5},10.0^{1}6,21]docos-13-en-8-one$ is a natural product found in Actaea yunnanensis and Actaea cimicifuga with data available.

The generated questions in the 3DMolLM dataset are:

- *What is the SMILES code of the molecule?*, which is trivial as SMILES sequence is part of the input.
- *What is the chemical name of this molecule?*,
- *What are some of the functional groups present in this molecule?*,
- *What are the physical properties of this molecule?*, speculating about properties not detailed in the caption.
- *What is the potential biological significance of this molecule?*, hypothesizing about the biological activity without supporting data.

These questions illustrate the challenge of relying on LLM-generated content without appropriate validation, leading to questions that speculate beyond available data.

In contrast, our current work employs a two-stage procedure to validate and filter data generated with LLMs. We also manually evaluate the dataset and estimate its overall accuracy at less than 0.05.

Since not much information is available on most molecules, our dataset averages about two questions per molecule, enhancing factual reliability.

Information Leakage: The 3DMolLM dataset does not anonymize molecule names, resulting in many questions that include either the common or chemical name of the molecule, thereby providing unintended hints. For example, a question in the dataset - *"What is the main component of Lobaplatin that gives it its anticancer properties?"* offers additional clues that may influence the evaluation process, complicating assessing a model's ability to learn from the molecule sequence or structure alone.

Evaluation Limited to Samples with Available Data: Datasets like MoMu, 3DMolLM, and MolCA limit their test sets to molecules accompanied by captions of more than 20 words. This approach inherently biases the evaluation towards well-known and extensively studied molecules. Conversely, the dataset presented in this work, while including these well-documented data points, also augments the test set with molecule captions chosen randomly, not by length. This strategy helps to provide a more balanced and representative evaluation of the model's capabilities across a wider range of molecular data.

Benchmarking Across Several Model Classes: Existing datasets often exhibit limitations in their benchmarking scope. For instance, MolInstructions has been evaluated solely using LLMs, while 3DMolLM employs only Llama as an additional baseline. Other models like KV-PLM and MoMu are benchmarked against only a subset of available models. Similarly, MoleculeSTM and MolT5 lack direct comparisons in their evaluations. In contrast, our approach aims to extensively benchmark and compare models across different architectural classes. This broader evaluation is facilitated by the structure of our dataset, which includes questions and multiple-choice options, allowing for a more comprehensive assessment of model performance.

D PROMPTS FOR DATA GENERATION AND INFERENCE

In this section, we discuss the various prompts for large language models in this work

- In Figure 3, we depict the prompt used to generate QA data (section 3.3). The LLM is provided with the input of a description of a molecule and a set of QAs is generated.
- In Figures 4 and 5, we depict the prompt used for validating the generated data with an LLM(section 3.4). In the first stage, the LLM is provided with the generated question, answer, and molecule description, and tasked with inferring the correct option based on this input. In the next stage, the LLM is given the filtered questions and prompted to evaluate their relevance in relation to the molecular characteristics.
- In Figures 6 and 7, we discuss the prompts used for inference from LLMs(section 4.3), and also for fine-tuning LLMs(section 4.4).

System Text:
""" Your task is to generate a set of questions about a molecule given its description. Each question should contain 5 multiple choice options. The correct answer should be an integer between 1 and 5. Also categorize each question into 1. Physical Properties, 2. Chemical structure information/properties 3. Biological or therepeutic information 4. Origin/Molecule Synthesis, 5. Applications.
Follow these strict rules:
1.All questions should be factually grounded in the caption, using the same terminology, and do not include any information not present in the caption. The focus of the questions should be the molecule.
2.The options should be thematically similar but should be discriminative enough and EXACTLY one of the options is correct. Avoid options like “all” or “none”.
3.Anonymize the actual name of the molecule in the QAs, and refer to it as “ molecule”.
The output should only include json parsable text (with no additional text), in the following format:
[{“question 1”:question, “options”:[option1, option2,..], “correct”:1, “sentence”: The molecule is option1, “category”:2 }, {“question”: .., .. :.. },..]"""
User Text:
""" Acetone is a manufactured chemical that is also found naturally in the environment. It is a colourless liquid with a distinct smell and taste. It evaporates easily, is flammable, and dissolves in water. It is also called dimethyl ketone, 2-propanone, and beta-Keto propane. """

Figure 3: **Prompt used for QA generation:** The LLM is provided with the input of a description of a molecule and is prompted to generate a set of QAs is generated

System Text:
""" You are given a description about molecule and a list of questions following with multiple choice options. For each question, provide the index of a single correct option. If the answer cannot be inferred from the description or the correct option is not available, output 0 for that question. The output be a list of integers (between 1 to 5) and strictly do not include any other text before or after the answer.
Example Output: [answer_idx_for_q1, answer_idx_for_q2, ...].
"""
User Text:
""" Description: Acetone is a manufactured chemical that is also found naturally in the environment. It is a colourless liquid with a distinct smell and taste. It evaporates easily, is flammable, and dissolves in water. It is also called dimethyl ketone, 2-propanone, and beta-Keto propane.
Questions:
Question 1: What is an alternative names for the molecule?
Options: ["Isopropanol", "Methyl ethyl ketone", "Beta-Keto propane", "Toluene", "Acetic acid"]
Question 2: What is the evaporation characteristic of the molecule?
Options: ["Does not evaporate", "Evaporates under high temperature", "Evaporates easily", "Sublimates instead of evaporating", "Evaporates only in vacuum"] """

Figure 4: **Prompt used for validation:** The LLM is provided with a a generated Question and options, and the description of the question it was generated from. The LLM is then prompted to identify the correct option.

System Text:
""" You are given a description about a molecule, and a set of question-answer pairs generated from it using an LLM. The objective of the questions is to test the understanding of the molecule and its properties (such as structure, manufacturing, chemical/physical properties, biological applications etc). Your task is to filter out any questions that are not relevant to the molecule, or that are not useful for testing the understanding of the molecule. Also provide a brief explanation for each question you filter out.
"""
User Text:
""" Description: Acetone is a manufactured chemical that is also found naturally in the environment. It is a colourless liquid with a distinct smell and taste. It evaporates easily, is flammable, and dissolves in water. It is also called dimethyl ketone, 2-propanone, and beta-Keto propane.
Questions:
Question 1: What is an alternative names for the molecule? Answer: “Beta-Keto propane”
"""

Figure 5: **Prompt used for validation step 2:** The LLM is then provided with the questions from the previous stage and tasked with evaluating their relevance in the context of deciphering molecular characteristics

Prompt for Inference (or) Finetuning:
""" You are given a SMILES string of a molecule, a question about the molecule and a set of candidate options. Output the index of the option that best answers the question. Do not include additional text in the output.
Question: What is an alternative names for the molecule?
SMILES string: CC(=O)C
Options: (1) "Isopropanol" (2) "Methyl ethyl ketone" (3) "Beta-Keto propane" (4) "Toluene" (5) "Acetic acid"
"""

Figure 6: **Prompt for Molecule QA inference:** The LLM is provided with an input SMILES string, and a set of options, and is prompted to identify the correct option.

Prompt for Inference (or) Finetuning:
""" You are given a sentence describing a molecule. Choose the SMILES string that best describes the sentence. Output the index of the best correct option only and nothing else. Do not include additional text in the output.
Sentence: The molecule is also called as Beta-Keto propane
Options: (1) C(=O)O (2) CCC(=O)C (3) CC(C)O (4) CC(=O)C (5) CC(=O)O """

Figure 7: **Prompt for Molecule Retrieval inference:** The LLM is provided with an input sentence about a SMILES string, and a set of options, and is prompted to identify the correct SMILES seqeuce.

E Dataset efficacy evaluation

To assess the overall reliability of our dataset, we conducted a benchmark by randomly sampling 400 data points from the test split. The data points were evaluated on three criteria: whether the question could be logically derived from the provided caption, the unambiguous correctness of the answer, and the relevance of the question—specifically, ensuring it avoids uninformative queries and contributes to meaningful chemical or biological insights based on the structure. Out of the sampled data, 391 points met these criteria. To understand the implications of this result for the entire dataset, we calculated the p-value using a hypergeometric distribution (sci). A hypergeometric test is used to measure the probability of obtaining a specific number of successes in a given number of draws from a finite population containing a certain amount of successes. With parameters k=400, n=391, N=11,922 (i.e the size of all test split), and K=0.961*11,922, we found a p-value of <0.05. This result suggests that the dataset is over 96.1 percent accurate, demonstrating a high level of reliability. Details of the random test samples used for this evaluation can be accessed through the project's repository.

To illustrate our evaluation process, we present representative examples of both accepted and rejected cases:

E.1 Accepted Examples

- **PubChem ID: 21580808**
  Question: What is the molecule resulting from?
  Options:

  1. Protonation of the oxygen of the primary amino group of sotalol
  2. Protonation of the nitrogen of the secondary amino group of sotalol
  3. Deprotonation of the nitrogen of the primary amino group of sotalol
  4. Protonation of the oxygen of the secondary hydroxyl group of sotalol
  5. Deprotonation of the oxygen of the primary hydroxyl group of sotalol

  Accepted because: The question addresses specific chemical modifications with clearly distinguishable options.

- **PubChem ID: 47528**
  Question: What is the mechanism of action of the molecule on vascular smooth muscles?
  Options:

  1. Membrane depolarization
  2. Membrane hyperpolarization
  3. Increased transmembrane sodium conductance
  4. Increased intracellular concentration of cyclic AMP
  5. Reduced transmembrane potassium conductance

  Accepted because: The question relates to structure-function relationships with distinct, non-overlapping answer choices.

- **PubChem ID: 1711945**
  Question: Where is the molecule naturally found?
  Options:

  1. Tilia platyphyllos
  2. Tilia tomentosa
  3. Sargassum natans
  4. Sargassum micracanthum
  5. Sargassum flavescens

  Accepted because: The question has a single, verifiable correct answer among distinct options.

E.2 Rejected Examples

- **PubChem ID: 54671008**
  Question: When did the molecule receive FDA approval?
  Options:
  1. 10 October 2006
  2. 12 October 2007
  3. 12 October 2008
  4. 10 October 2009
  5. 12 October 2010

  Rejection Rationale: Relies on temporal metadata information rather than molecular properties, which cannot be inferred from structure.
- **PubChem ID: 10129**
  Question: What type of odor does the molecule have?
  Options:
  1. Strong
  2. Mild
  3. Sweet
  4. Pungent
  5. Unpleasant

  Rejection Rationale: Options lack clear differentiation and are potentially overlapping.
- **PubChem ID: 101562486**
  Question: What is the general class of biomolecules to which the molecule belongs?
  Options:
  1. Carbohydrate
  2. Lipid
  3. Oligopeptide
  4. Nucleic acid
  5. Heterocycle

  Rejection Rationale: Multiple options could be technically correct.
- **PubChem ID: 53361968**
  Question: What characteristic may make the molecule a desirable therapy?
  Options:
  1. It is less expensive
  2. It is less likely to generate resistance
  3. It is only for treatment-naive patients
  4. It is only for PI-experienced patients
  5. It is only for HIV-2 infections

  Rejection Rationale: Addresses clinical outcomes not directly inferrable from structure.

F Baseline Models, Training, and Finetuning

In this section, we expand on the specific details of different models used to evaluate the MolQA dataset.

F.1 Model Details

F.1.1 Single Encoder Architectures:

Scibert: SciBERT (Beltagy et al., 2019) is an encoder model, that leverages a pre-trained BERT framework, subsequently fine-tuned on a substantial corpus of scientific papers, predominantly from the biomedical domain (constituting 85% of its training data). This specialization makes SciBERT a useful baseline for our study. It has a robust performance across various scientific tasks, including named entity recognition and text classification.

KV-PLM: KV-PLM(Zeng et al., 2022) is a single encoder model derived from SciBERT, further trained on molecule-text pairs sourced from PubChem. The training process begins with pre-training, during which SMILES sequences are appended to molecular descriptions to form the training data. The model employs a masking strategy where certain tokens representing both molecular structures and biomedical text are masked at random. The model's task is to predict these masked tokens based on the surrounding context. Following pre-training, KV-PLM undergoes fine-tuning for text retrieval tasks. In this phase, the model learns to accurately retrieve specific text descriptions based on SMILES sequence inputs, utilizing a max hinge loss function. This loss is given by:

$$\begin{aligned}\mathcal{L}_{\mathrm{MH}} = & \max_{\mathbf{d}'}\left[\alpha + \mathrm{s}\left(\mathbf{m}, \mathbf{d}'\right) - \mathrm{s}(\mathbf{m}, \mathbf{d})\right] \\ & + \max_{\mathbf{m}'}\left[\alpha + \mathrm{s}\left(\mathbf{m}', \mathbf{d}\right) - \mathrm{s}(\mathbf{m}, \mathbf{d})\right], \end{aligned} \tag{1}$$

here $\mathcal{L}_{\mathrm{MH}}$ represents the max hinge loss, $\mathbf{m}$ and $\mathbf{d}$ denotes the molecule (SMILES sequence) and its corresponding text description, respectively. The terms $\mathbf{d}'$ and $\mathbf{m}'$ refer to a negative text and molecule that do not match the original pairing and the function $\mathrm{s}(\mathbf{m}, \mathbf{d})$ calculates the similarity score between a molecule and a document.

MolT5: MolT5 (Edwards et al., 2022) is an encoder-decoder model, built by fine-tuning a T5 (Raffel et al., 2020) model. The model is trained in two stages. First, the the model is trained with masked language modeling objective, to encode and decode SMILES string and molecule captions. Next, the model is fine-tuned to generate SMILES strings or captions from the captions or the SMILES strings input respectively. The dataset used for fine-tuning is CheBI-20 (Edwards et al., 2021)).

F.1.2 Multimodal Architectures:

MoleculeSTM: The paper (Liu et al., 2023a) introduces a framework that uses a dual-encoder to extract and align representations of text and molecules. The framework employs a Graph Isomorphism Network (GIN) (Xu et al., 2018) to encode chemical data, represented by f_c, and SciBert to encode textual data, denoted as f_t. The GIN model is initialized from GraphMVP (Liu et al., 2021), which does multi-view pretraining between the 2D topologies and 3D geometries from the GEOM dataset (Axelrod and Gomez-Bombarelli, 2022). The components are trained end-to-end on a dataset that contains molecules and their descriptions sourced from PubChem. The model's learning process is governed by the InfoNCE loss:

$$\begin{aligned}\mathscr{L}_{\text{InfoNCE}} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}_c, \boldsymbol{x}_t}\Bigg[\log \frac{\exp\left(E\left(\boldsymbol{x}_c, \boldsymbol{x}_t\right)\right)}{\exp\left(E\left(\boldsymbol{x}_c, \boldsymbol{x}_t\right)\right) + \sum_{\boldsymbol{x}_{t'}} \exp\left(E\left(\boldsymbol{x}_c, \boldsymbol{x}_{t'}\right)\right)} \\ + \log \frac{\exp\left(E\left(\boldsymbol{x}_c, \boldsymbol{x}_t\right)\right)}{\exp\left(E\left(\boldsymbol{x}_c, \boldsymbol{x}_t\right)\right) + \sum_{\boldsymbol{x}_{c'}} \exp\left(E\left(\boldsymbol{x}_{c'}, \boldsymbol{x}_t\right)\right)}\Bigg] \end{aligned} \tag{2}$$

Here $\boldsymbol{x}_c$ and $\boldsymbol{x}_t$ represent the input chemical structure and textual description, respectively. f_c, f_t represent the chemical and text representation model, and p_c, p_t represents chemical, text projection matrices. The function $E(\boldsymbol{x}_c, \boldsymbol{x}_t) = \langle p_c \circ f_c(\boldsymbol{x}_c), p_t \circ f_t(\boldsymbol{x}_t)\rangle$ calculates the similarity. The goal

is to distinguish between correctly matched chemical-text pairs $(\boldsymbol{x}_c, \boldsymbol{x} * t)$ and mismatched pairs $(\boldsymbol{x}c, \boldsymbol{x}t', \boldsymbol{x} * c', \boldsymbol{x}_t)$, enhancing the model's ability to map chemical structures to their descriptive texts accurately.

MoMu: MoMu (Su et al., 2022) is another dual-encoder model similar to MoleculeSTM, leveraging both SciBERT for textual data and a Graph Isomorphism Network (GIN) for chemical structures. Similar to MoleculeSTM, MoMu employs a GIN network but is initialized with random weights, which are then trained using a contrastive loss mechanism akin to that used in MoleculeSTM. This baseline comparison underscores the potential enhancements 3D pretraining brings to the model's ability to capture complex molecular structures.

F.1.3 Large Language Models

Llama: Llama (Touvron et al., 2023) is a series of decoder-only, autoregressive transformer models trained on a large general corpus. Llama demonstrates exceptional performance in various tasks, including common sense reasoning, closed-book QA, mathematical reasoning, and code generation. Llama has been fine-tuned on a select instructional dataset to follow human instructions effectively. Given its extensive application range, assessing Llama's chemical understanding capabilities is of interest. We have experimented with the smallest and the largest versions of the Llama-2 and Llama-3 series of models.

GPT 3.5 Turbo We further benchmark the GPT-3.5 Turbo (OpenAI, 2022) model, a member of the Generative Pre-trained Transformer series.

Galactica: We also benchmark Galactica (Taylor et al., 2022), a set of scientific language models that are autoregressive, decoder-only similar to the previous models. These models are trained to recognize and understand a wide range of scientific information, such as chemical structures represented by SMILES strings, sequences of amino acids, computer code, and mathematical equations. The dataset used contains a large collection of scientific documents and research papers. It is shown to be effective on specialized biomedical datasets like PubMedQA and MedMCQA. For this study, we use Galactica models of different sizes, with 125M, 1.3B, and 6.7B parameters.

F.2 Zero-shot Inference:

Large Language Models: For the Llama series of models, zero-shot inference was performed using the API service offered by Microsoft Azure AI services. For GPT-3.5, zero-shot inference is performed using the OpenAI platform. For Galactica, model weights are obtained from the Hugging Face library.

Multi-modal Architectures: For SciBERT and KV-PLM, the model weights are initialized from the MoleculeSTM official repository: `https://github.com/chao1224/MoleculeSTM`. We retrained MoMu and MoleculeSTM, due to potential data leakage issues by removing the samples in the pertaining set that overlap with the test set. 30 Epochs of training on the pre-training dataset were performed for MoleculeSTM and MoMu, which took about 40 hours on an NVIDIA RTA A6000 GPU. The code and model parameters were obtained from Molecule STM's official repository. THe learning rate used was 1e-5 and a batch size of 45. For MoMu, we used the augmentation probability of 0.2. A temperature of 0.1 is used for both the models.

F.3 Finetuning:

Galactica: 3 Epochs of training on the finetuning dataset were performed by the other models. LoRA (Hu et al., 2021) was used to finetune the query and value vectors of Galactica 125M, 1.3B, and 6.7B. Training and evaluation took around 2.5 hours, 5 hours, and 30 hours respectively on a NVIDIA A100 80GB PCIe. The learned rate used was 2e-5, a weight decay of 0.01, and a batch size of 8 using Huggingface's Trainer for causal language modeling (because the base OPT model is a decoder-only model)[1]. The LoRA parameters are $r = 16$, $\alpha = 32$ and a lora_dropout of 0.05.

[1] `https://huggingface.co/docs/transformers/en/tasks/language_modeling`

MoleculeSTM and MoMu: The finetuning was performed in the same setting as training, as described in Section F.2. Both the models are fine-tuned for 3 epochs, at a learning rate of 1e-15. The total finetuning time is about 1 hour.

Llama2-7B and Llama3-8B models: Both the models are trained for 3 epochs using LoRA on an NVIDIA A100 80GB PCIe. The total training time is approximately 30 hours for each model. The Lora parameters used are $r = 16$, $\alpha = 32$ and a lora_dropout of 0.1. The learning rate is 1e-5 and the weight decay used is 1e-4.

MolT5, BioT5 and BioT5 plus: The models are finetuned for 10 epochs on an NVIDIA A100 80GB PCIE. The approximate training time for the MolT5 model is 3.5 hours, while it is 6 hours for the BioT5 models. These models are trained with a learning rate of 2e-5, a weight decay of 0.01 and a batch size of 8 using the Huggingface's trainer.

F.4 A NOTE ON 3DMOLLM

Additionally, we attempted to fine-tune the BLIP-like model, as proposed in the 3DMolLM paper (Li et al., 2024), on the MolTextQA dataset. This model methodology involves projecting the 3D structure of a molecule into the space of LLM tokens and utilizing these tokens for text generation. Unfortunately, we encountered challenges during fine-tuning using the default configurations provided in the code repository. Specifically, we were unable to finetune the model to generate answers in the required format, which impeded our ability to perform any meaningful analysis.