MoleculeQA: A Dataset to Evaluate Factual Accuracy in Molecular Comprehension

Anonymous ACL submission

Abstract

Large language models are playing an increasingly significant role in molecular research, yet existing models often generate erroneous information. Traditional evaluations fail to assess a model's factual correctness. To rectify this absence, we present MoleculeQA[1], a novel question answering (QA) dataset which possesses 62K QA pairs over 23K molecules. Each QA pair, composed of a manual question, a positive option and three negative options, has consistent semantics with a molecular description from authoritative corpus. MoleculeQA is not only the first benchmark to evaluate molecular factual correctness but also the largest molecular QA dataset. A comprehensive evaluation on MoleculeQA for existing molecular LLMs exposes their deficiencies in specific aspects and pinpoints crucial factors for molecular modeling. Furthermore, we employ MoleculeQA in reinforcement learning to mitigate model hallucinations, thereby enhancing the factual correctness of generated information.

1 Introduction

Large Language Models (LLMs) have practical applications in molecule research (Fabian et al., 2020; Bagal et al., 2021), specifically in the use of cross-modal models (Liu et al., 2023b; Cao et al., 2023). These models bridge the gap between molecular structures and natural language (Bran and Schwaller, 2023), helping experts understand the properties, potential applications, and acquisition methods of specific compounds, improving understanding of the complex molecular architecture and reducing experimental failure rates.

Although molecular LLMs are capable of generating task-relevant content, they often suffer from generating illusory statements. In the widely studied molecule captioning (mol2text) task, users expect the model to generate a comprehensive and detailed description of a given compound. In this task, existing benchmarks (Edwards et al., 2022; Liu et al., 2023b) typically employ metrics such as BLEU (Papineni et al., 2001) and ROUGE (Lin, 2004) to evaluate the performance of molecular LLMs. However, without examining the factual accuracy of these models, it is vague to justify how reliable they are. In Table 1, we provide several examples from the CheEBI-20 (Edwards et al., 2021) test dataset to illustrate this issue. Despite the plausible and fluent appearance of the generated content, there are numerous unnoticed inaccurate statements, which remain difficult to detect under the current lexical-based benchmarking approach.

CID	Ground-Truth	Generated
9810996	The molecule is a dipeptide composed of N-(3,3-dimethylbutyl)-L-aspartic acid and methyl L-phenylalanate units joined by a peptide linkage.	The molecule is a dipeptide obtained by formal condensation of the alpha-carboxy group of N-(3,3-dimethylbutyl)-L-phenylalanine with ethanol.
10129879	The molecule is the stable isotope of potassium with relative atomic mass 38.963707.	The molecule is the stable isotope of tellurium with relative atomic mass 124.904425.
5281034	A synthetic androgen, it was mainly used for the treatment of anaemias until being replaced by treatments with fewer side effects.	It has a role as a contraceptive drug, a progestin and a synthetic oral contraceptive.
15011611	It is found in Tripterygium wilfordii and Tripterygium hypoglaucum.	The molecule is an abietane diterpenoid isolated from the stem bark of Fraxinus sieboldiana.

Table 1: The accuracy of the state-of-the-art mol2text model in describing the given molecule is assessed using BioT5-base (Pei et al., 2023) inference results. The generated content is plausible and fluent, but comparison with the ground truth reveals several factual errors highlighted in red.

Counterfactual molecular generation content can lead to the following adverse consequences: 1) Misuse of deployed models can deceive and mislead ordinary users, reducing productivity. 2) Professionals may lower their expectations of deployed models when they recognize significant factual errorness, thus hindering positive applications. To avoid these repercussions, quantifying the level of comprehension that models have of molecule knowledge is valuable. However, expertise and professional knowledge are required for human to

[1] https://anonymous.4open.science/r/MoleculeQA

detect hallucinations in generated molecular text, which is extremely difficult with high cost.

To alleviate the absence of fine-grained factual correctness evaluation for molecular LLMs, we propose **MoleculeQA**, a comprehensive benchmark based on question-answer pairs covering various aspects including molecular property, source, structure, and application. MoleculeQA endeavors to provide reliable assessments of knowledge comprehension for molecular LLMs, and to offer potential solutions for mitigating model hallucinations.

Construction of MoleculeQA involves two main stages. 1) **Molecular Taxonomy Construction.** We utilize authoritative molecule description corpus as the source. Using a hybrid approach of rule-based and automated methods, we extract topics based on properties, sources, and other relevant aspects. After clustering and manual normalization, we gather the topics to build a hierarchical domain taxonomy that has broad coverage and strong expertise. 2) **Taxonomy-guided QA construction.** By converting each molecular description into several pairs of QA that align with the topics at different levels of taxonomy, we can create a QA benchmark that guarantees both granularity, breadth, and quality. MoleculeQA is not only the first factual evaluation benchmark in the molecular domain, but also the largest molecular QA dataset.

Based on MoleculeQA, we perform accuracy tests on various molecular LLMs. Our experimental results indicate that existing methods remain at a discernible remove from achieving a precise comprehension of molecules, and undercover several vital factors for molecule modeling. Furthermore, we utilize MoleculeQA to provide feedback for molecular LLMs' reinforcement learning, aiming to enhance the factual correctness of the models. Our contributions are summarized as follows:

- We reveal the factual inaccuracies in the content generated by existing LLMs in the molecule or chemistry domain, which have not been adequately detected by existing benchmarks.
- For comprehensive factual accuracy evaluation, we develop a domain taxonomy for molecule corpus and use it to create a high-quality question answering benchmark called MoleculeQA.
- Using MoleculeQA, we test a series of models. Based on our experimental outcomes, we identify specific deficiencies in molecular LLMs and summarize several critical factors for molecular understanding. We also attempt to use MoleculeQA as feedback for reinforcement learning to reduce model hallucinations.

2 Related Work

2.1 Molecule Understanding LLMs

Advancements in language models pre-trained with scientific corpora (Lee et al., 2019; Luo et al., 2022; Beltagy et al., 2019) have shown considerable success in molecular research. Recently, cross-modal models have emerged (Edwards et al., 2021; Luo et al., 2023a; Liu et al., 2023a), aiming to bridge the gap between molecular language (bio-sequence or structure) and natural language. Evaluation tasks for these models include seq2seq generation-based tasks (e.g., molecule captioning and text-based de novo molecule generation) and contrastive-based tasks (e.g., cross-modal retrieval). The corresponding models can be classified as generative models (e.g., MolT5 (Edwards et al., 2022), BioT5 (Pei et al., 2023)) and contrastive models (e.g., MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2022)).

Seq2seq tasks assess the model's translation ability between modalities. For text-to-molecule generation, metrics include molecule fingerprint similarity (e.g. Morgan-FTS (Schneider et al., 2015)), sequence-based metrics like BLEU (Papineni et al., 2001) and validity. Molecule captioning tasks rely on n-gram precision (BLEU), recall (ROUGE (Lin, 2004)), or both (METEOR (Banerjee and Lavie, 2005)) to measure lexical similarity but lack chemical knowledge comparison and factual correctness detection. Retrieval-type tasks align molecules with descriptions, but overlook fine-grained alignment between text snippets and substructures.

2.2 Domain-Specific QA

The Question Answering (QA) task serves as a quantitative measure for evaluating the reasoning and inference capabilities of intelligent systems. In the general domain, a large number of QA datasets have been constructed (Rajpurkar et al., 2016; Lai et al., 2017; Yang et al., 2018). In addition, specific domains such as medical (Jin et al., 2019, 2020; Pal et al., 2022), news (Nallapati et al., 2016; Trischler et al., 2016), and legal (Zheng et al., 2021; Zhong et al., 2019) have also developed standard QA datasets that are widely used by the community. QA datasets in specific domains can be classified into extraction-based (Pappas et al., 2018), generation-based (Savery et al., 2020), multi-choice (Pal et al., 2022) and Yes / No formats (Jin et al., 2019). QA pairs are constructed from various sources, including scientific articles (Jin et al., 2019), examination problems (Pal et al., 2022; Zaki et al., 2023), professional databases (Liang et al., 2023), and crowd-sourcing data (Wei et al., 2020; Hendrycks et al., 2020).

However, in the molecular domain, there is a scarcity of comprehensive, diverse, and high-quality QA datasets. Existing datasets like DrugChat (Liang et al., 2023) have limitations in terms of molecule features and simplistic answers. BioMedGPT (Luo et al., 2023b) transforms molecule caption task datasets into QA format, inheriting current evaluation issues like domain knowledge deficiency and excessive reliance on lexical similarity. Conversely, MoleculeQA constructs a domain taxonomy and derives QA pairs from descriptive texts, ensuring comprehensive, diverse, high-quality, and credible coverage.

3 MoleculeQA Dataset

3.1 Exposure of Factual Correctness Issue

In this subsection, we analyze the extent of factual correctness in the generated content of the molecule captioning (mol2text) models.

Setup. To evaluate the reliability of compound descriptions generated by these models, we categorize them into four different aspects: *Structure*, *Property*, *Application*, and *Source*. The aspects are derived from descriptions in PubChem (Kim et al., 2022), the largest molecule caption dataset currently available. PubChem includes specific sources for each molecule's description, such as Lotus (Mun et al., 2016) for source information, DrugBank (Wishart et al., 2017) for application details, CAMEO Chemicals (cam) for property descriptions, and multiple data repositories for structure information. The definitions of these main aspects are summarized in the Table 2 below.

Aspect	**Definition**
Structure	Details about architecture, composition, and interaction of atoms within a molecule.
Property	Physical, biological or chemical property in various environments or reactions.
Application	The utilization of a molecular compound in various applications and scenarios.
Source	The natural or synthetic origin, as well as the production context related to a molecule.

Table 2: **Evaluation Aspects** of description about molecules.

We randomly sample 100 molecule&caption samples from the ChEBI-20 test set and take MolT5, MoMu, and BioT5 models to generate descriptions for each molecule. Both ground truth and generated content are manually classified based on four aspects. We evaluate the models' descriptions in each aspect against the ground truth, with two trained domain experts judging them as **correct** (if the generated content matches the ground truth), **miss** (if the ground truth has a corresponding aspect description but it was completely missing in the generated content), or **error** (if there is a clear factual inconsistency with the ground truth).

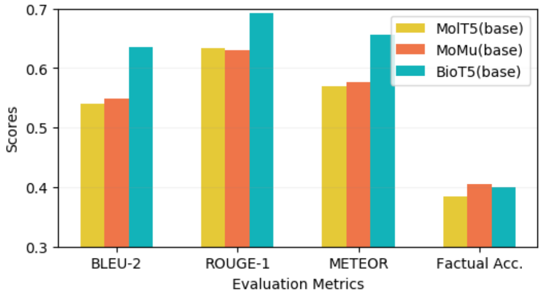


Figure 1: The performance of three representative models on the traditional metrics for the molecule caption task (e.g. BLEU etc.) and the factual accuracy metric we defined.

Model	Structure	Property	Application	Source
MolT5-base	63/0/34	1/4/3	7/15/8	20/10/30
MoMu-base	63/0/34	1/4/3	5/16/9	19/ 8/33
BioT5-base	62/0/35	2/3/3	9/12/9	16/13/31

Table 3: **Human Assessment of Model Generated Molecular Descriptions** based on 4 aspects, with the counts presented according to error / miss / correct.

Results. In Figure 1, we assess the content generated by the model using traditional lexical-based metrics (BLEU, ROUGE, METEOR), as well as their factual accuracy on the selected subset. We define factual accuracy as the ratio of correct predictions to the total number of slots, serving as an average metric to evaluate the reliability of the generated content. Despite the progress in training methodologies, models have exhibited incremental improvements in lexical similarity metrics (such as a 17.6% increase in BLEU-2). Nevertheless, there is no discernible improvement in the dependability of the generated content, with factual accuracy persisting at 0.4. In our detailed factual performance analysis (Table 3), we observe that models often omit application-related details and relevant properties. The generated descriptions about *Structure* show a significant discrepancy rate of more than 63% compared to ground truth. This challenges the credibility of expert model-generated content, which warrants further scrutiny.

3.2 Domain Taxonomy Construction

Taxonomy frameworks organize concepts or entities within a domain hierarchically, aiding in

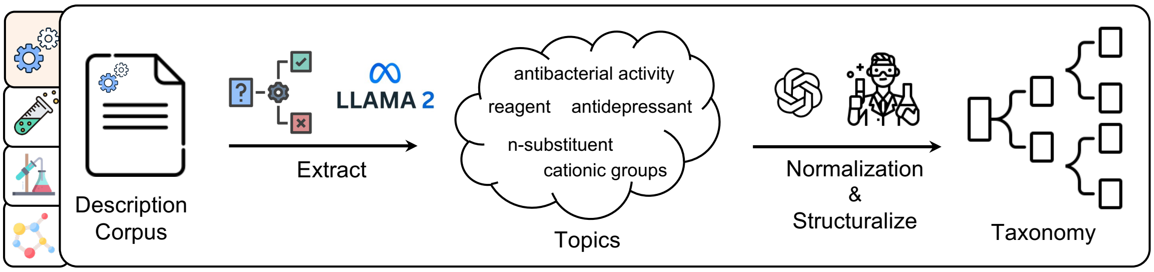
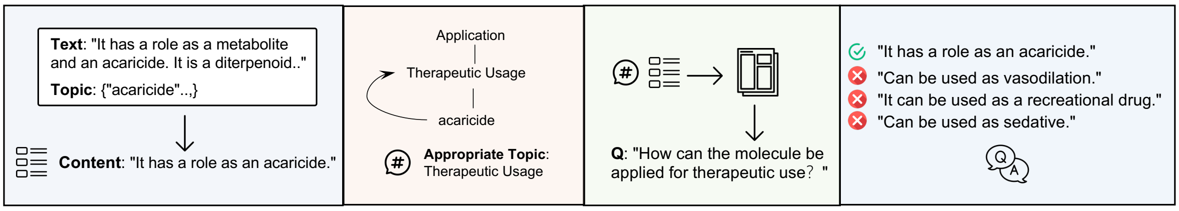


Figure 2: The process of constructing a molecular domain taxonomy. The procedures involve the selection of the information source, extraction of topics, normalization and structuralization of topics, and hierarchical clustering by domain experts.


Text: "It has a role as a metabolite and an acaricide. It is a diterpenoid.."
Topic: {"acaricide"..}
Content: "It has a role as an acaricide."
Application
Therapeutic Usage
acaricide
Appropriate Topic: Therapeutic Usage
Q: "How can the molecule be applied for therapeutic use? "
"It has a role as an acaricide."
"Can be used as vasodilation."
"It can be used as a recreational drug."
"Can be used as sedative."
Q
A

(a) Extract Content (b) Reassign topic (c) Design question (d) Collect Pos. & Neg. answers

Figure 3: The process of constructing a molecular domain taxonomy. The procedures involve the selection of the information source, extraction of topics, normalization and structuralization of topics, and hierarchical clustering by domain experts.

the organization of domain-specific queries (Liu et al., 2012) and ensuring the quality of comprehension domain knowledge and constructing question-answering pairs. We adhere to established procedures for the construction of domain taxonomies, as illustrated in Figure 2.

Information Source. Considering the data quality, we choose the most widely used ChEBI-20 dataset as our molecular description corpus. To mitigate the class imbalance issue in ChEBI-20, primarily dominated by structural information, we include additional sources like T3DB (Wishart et al., 2014), FDA Pharm Classes, and DrugBank. We employ a pre-trained text classifier to perform an initial coarse-grained division of the corpus based on the four aspects we defined above, which serve as the first-level nodes in our taxonomy.

Topics Extraction. We further employ a hybrid approach combining rule-based and few-shot prompting methods to extract topics and their corresponding original text from the corpus, formatting the $(topic, text)$ pairs. Subsequently, to mitigate lexical noise and uncontrolled granularity within the 1K topics collected, we utilize GPT-4 (OpenAI, 2023b) with a few-shot prompt-based approach to accomplish an initial semantic aggregation.

Topics Normalization & Structuralization. Next, domain experts intervene to perform rule-based topic merging and concept splitting manually. Finally, the remaining 587 topics are hierarchically clustered by human experts, resulting in a three-level molecular domain taxonomy. An overview of this taxonomy can be found in the Appendix. The leaf nodes represent specific molecule characteristics and are the narrowest topics/concepts, while non-leaf nodes represent broader concepts.

3.3 MoleculeQA Construction

Based on the taxonomy in 3.2, we develop a 4-step procedure to extract questions and answers from molecular descriptions to construct MoleculeQA. The whole workflow is displayed in Fig 3.

Content Extraction & Reassign Topic. With $(topic, text)$ pairs annotated in 3.2, a reasonable notion is to query molecules by topic, but content related to a specific topic can be over-brief to be queried. For example, for the molecule `CID:5479113`, the content of topic `acaricide` is `It has a role as an acaricide.` Without enough information, it is difficult to justify which species of mites this molecule is effective. However, it can be queried from a coarser granularity like `Therapeutic Usage`, the parent topic of `acaricide`.

To select a suitable topic for querying, we first use an agent to extract content related to the topic from text. A rule-based program is employed to verify the content, and, in cases where specific details about a given topic are unavailable, we replace the topic with its parent topic until the level of granularity is appropriate for querying purposes.

Question Design. We invite two annotators to design questions for topics based on the extracted contents. For example, contents for topic `inhibitor` include `It is a protein synthesis inhibitor` and `It is a mitotic`

Taxonomy	Reference Description	Extracted Question	Positive Answer	Negative Answer
Property→ Antiviral activity	It has been shown to *exhibit inhibitory effects on the viral neuraminidases from two influenza viral strains, H1N1 and H9N2*.	Which kind of **antiviral activity** does this molecule have/exhibit?	It exhibits inhibitory effects on the viral neuraminidases from two influenza viral strains, H1N1 and H9N2.	It is used for the treatment of cytomegalovirus (CMV) retinitis in AIDS patients.
Structure→ Backbone	The molecule is a heparan sulfate composed of a backbone of *repeating beta-D-glucuronosyl-(1->4)- N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages*.	Which kind of **backbone** does this molecule have?	It has a backbone of repeating beta-D-glucuronosyl-(1->4)-N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages	It has a backbone of repeating alpha-L-iduronosyl-(1->4)-N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages.

Table 4: Examples of automatically generated QA instances. *blue* stands for reference locations, red for factual errors.

`inhibitor`, annotators may design `Which kind of inhibitor is this molecule?`. For each topic, annotators discuss choosing the better design as its final question and make sure each question can be answered using the molecular descriptions.

Pos. Options Collection. For the positive options, since formal extracted contents may be rigid and can't be directly used as answers, we leverage the in-context learning capability of ChatGPT (OpenAI, 2023a) to generate appropriate positive options via few-shot prompting.

Neg. Options Collection. For the same question, we take positive options from other molecules as negative candidates for each molecule. To eliminate illegal negatives, we merge synonymous options and remove overlapping options. Then we adopt BioT5 (Pei et al., 2023) to encode all candidates and choose candidates with similar semantics to the positive option as negatives. Several generated QA instances are shown in Table 4.

Data Split. We split molecules in MoleculeQA into train/dev/test sets by scaffolds to divide molecules with similar structures into the same sets as suggested in (Hu et al., 2019), making the QA task more challenging yet realistic.

Quality Control. To provide reliable factual evaluation, LLM and human efforts are combined to ensure MoleculeQA's quality. We convert each QA instance into natural language using templates and assess its logical and semantic consistency with the original description using ChatGPT. This process is repeated 3 times to minimize variations. With taxonomy guidance, the number of disqualified samples is minimal and can be manually resolved.

Human Evaluation. We assign one annotator [2] to evaluate the reliability of the test split and receive error rate lower than 1%. Finally, we randomly sample 100 cases and assign two annotators to evaluate the quality of QA samples. The annotators assess the **Consistency** between the question and the correct option with the reference caption text, as well as **Discrimination** between the positive and negative options. Human evaluation results can be found in Table 5. The high consistency and discrimination metrics, along with a satisfactory level of agreement (Cohen kappa) among annotators, validate the quality and reliability of our MoleculeQA.

Metric	Annotator 1	Annotator 2	Agreement (κ)
Consistency	99.0	99.0	1.0
Discrimination	97.0	96.0	0.85

Table 5: Evaluation for the generated QAs quality.

3.4 Data Analysis

Data Statistics. In Table 6, we present the number of QA samples and the coverage of topics in MoleculeQA in comparison to several popular biomolecular and chemistry-related benchmarks (Wei et al., 2020; Yue et al., 2023; Hendrycks et al., 2020; Lu et al., 2022). We observe that MoleculeQA is both the first benchmark focused on evaluating molecular factual knowledge and the largest scale QA dataset in the molecular field.

Benchmarks	# QA	Sophistication
MMLU(Chem)	534	College, High school, Medicine
MMMU(Chem)	638	Inorganic, Organic, Physical
ScienceQA	867	Solution, Reaction, Molecule
ChemistryQA	4,500	Reaction, Molecule, Physics
MoleculeQA	61,574	Structure, Source, Property, Application

Table 6: Number of samples and topics coverage compared to popular related benchmarks.

The train, development, and test split consists of 49,993, 5,795 & 5,786 QA samples. The general statistics of the dataset are summarized in Table 7.

Aspects	Structure	Property	Application	Source	Total
# Train	32,176	4,838	1,917	11,062	49,993
# Dev	3,314	698	558	1,225	5,795
# Test	3,113	731	599	1,343	5,786
Avg. Q Tokens	7.96	9.02	7.90	7.00	7.74
Avg. A Tokens	9.50	10.98	11.93	7.96	9.42

Table 7: MoleculeQA dataset statistics, where Q and A represent the Question and Answer respectively.

Data Distribution. Fig 4 provides the visualized distribution of MoleculeQA. All topics in our taxonomy are queried in MoleculeQA for a comprehensive, fine-grained factual evaluation. Inherited from ChEBI-20, QA pairs in the *Structure* aspect

[2] All annotators are doctoral students engaged in molecule research, with at least six months of professional experience.

account for approximately two-thirds of the whole MoleculeQA. While topics within each aspect have relatively balanced sample numbers.

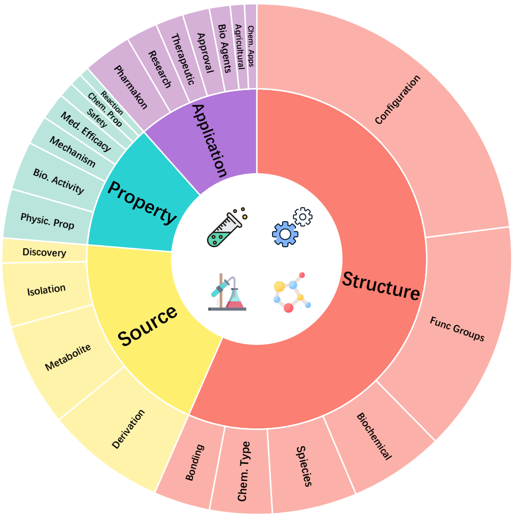


Figure 4: An overview of MoleculeQA topics distribution. Four coarse-grained aspects occupy the inner circle, and in the outer circle we list finer-grained non-leaf topics.

4 Experiment

4.1 Baseline Models

The main purpose of baseline experiments is to investigate current models' performance in answering multiple-choice questions related to molecular knowledge. We categorize models based on whether their base LLMs are adequately trained on a large-scale biomolecular corpus as follows:

Molecular LLM, represented by MolT5 (Edwards et al., 2022), MoMu (Su et al., 2022), BioT5 (Pei et al., 2023), MolCA (Liu et al., 2023b) and BioMedGPT-LM-7B (Luo et al., 2023b). These models undergo incremental training stages with extensive molecular modality data (e.g. SMILES or SELFIES strings), biomedical academic papers, and molecule-description pairs.

General LLM, represented by T5 (Raffel et al., 2019), OPT (Zhang et al., 2022), GALACTICA (Taylor et al., 2022), BLOOM (Scao, 2022), Pythia (Biderman et al., 2023), LLama-2 (Touvron et al., 2023b), along with its instruction fine-tuned derivatives, such as Vicuna (Chiang et al., 2023) and Mol-Instruction-7B (Fang et al., 2023).

Large-scale Universal Models. We evaluate the large-scale, state-of-the-art LLMs in few-shot settings, including open-access models such as Mixtral 8×7B (Jiang et al., 2024), and OpenAI's GPT family, specifically GPT-3.5 (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b) accessed via API [3].

4.2 Evaluation Setups

We follow training approaches and hyper-parameters in the **original** papers for respective methods. Details about training configuration and few-shot examples are provided in Appendix A.3. Training approaches in our evaluation include:

Full Fine-tuning: All model parameters are updated, including the base LLMs, structure encoders, and projectors for molecule-language alignment.

LoRA-based Fine-tuning: The base LLMs are tuned by low-rank adaptation (Hu et al., 2021), and structure encoders are also trainable.

Few-shot Setting: We sample 10 QA examples from four aspects respectively to prompt LLMs with task definition and contextual information.

The main metric of MoleculeQA is the **accuracy**, which is defined as the ratio of correctly answered samples among all test samples. We present the accuracy in four aspects as well as the total accuracy.

4.3 Main Results

We summarize the benchmarking results in Table 8:

- **Comparison over four aspects.** Achieving the highest accuracy on *Source* is generally more feasible for each model, whereas addressing *Property* and *Application* presents notable difficulties, with no method surpassing a 50% accuracy rate. This phenomenon may be ascribed to the comparatively smaller data scale within these domains.
- **Molecular LLMs v.s. General LLMs.** Molecular LLMs demonstrate better performance, with a minimum total accuracy over 51%. By contrast, other than T5s, decoder-only General LLMs fail to achieve a total accuracy exceeding 50%, whether fully fine-tuned or tuned with LoRA.
- **T5 series comparison.** Among T5-based methods, T5 demonstrates superior performance compared to MolT5 (e.g., T5-base surpasses MolT5-base in total accuracy by 5.1%) contradicting their performance on molecule caption tasks. BioT5 combines bio-molecular texts and databases for molecular pretraining, achieving higher total accuracy than T5 (+ 6.5%).
- **Decoder-only LLMs comparison.** Among Llama-based models, BioMedGPT-7B achieves the best performance with incremental pre-training, while Mol-Instruction fine-tuned by instructions has slight improvement than Llama and Vicuna. With the similar size of the base model (7B) and LoRA parameters, the performance ranking among different models is as

[3] https://api.openai.com/v1/chat/completions

Model	# Trainable Params	Implementation	Structure	Source	Property	Application	Total
Random	–	–	24.41	22.30	23.04	24.57	24.03
Molecular LLM							
MolT5-small	80M	full ft	49.59	64.18	46.51	40.90	51.69
MolT5-base	250M	full ft	58.01	65.85	45.14	42.24	55.39
MoMu-small	82M	full ft	52.71	63.44	44.87	40.57	52.96
MoMu-base	252M	full ft	61.58	65.30	43.78	43.07	57.43
BioT5-base	252M	full ft	65.98	69.24	**49.11**	40.73	62.03
MolCA-125M	100M	LoRA ft	65.54	67.34	45.77	40.33	60.30
MolCA-1.3B	110M	LoRA ft	**71.12**	**70.98**	47.81	**43.17**	**64.79**
BioMedGPT-LM-7B	40M	LoRA ft	54.19	60.01	38.85	40.90	52.23
General LLM							
T5-small	60M	full ft	55.51	64.41	45.42	38.56	54.55
T5-base	220M	full ft	**60.42**	**66.42**	45.83	**43.74**	**58.24**
OPT-125M	125M	full ft	38.58	55.92	41.04	28.73	42.93
OPT-350M	331M	full ft	44.39	60.83	**46.24**	40.57	48.05
GALACTICA-6.7B	12.5M	LoRA ft	32.35	41.92	31.05	28.21	33.96
BLOOM-7.1B	27.5M	LoRA ft	35.01	47.51	31.46	33.56	37.31
Pythia-6.9B	29.4M	LoRA ft	42.79	58.90	38.58	39.07	45.61
Mol-Instruction-7B	40M	LoRA ft	37.46	47.36	32.69	29.88	38.37
Llama-2-7B-chat	40M	LoRA ft	28.75	39.84	31.33	27.71	31.54
Llama-2-13B-chat	63M	LoRA ft	34.37	43.86	31.05	29.72	35.67
Vicuna-v1.5-7B	40M	LoRA ft	34.89	44.15	34.20	31.55	36.61
Vicuna-v1.5-13B	63M	LoRA ft	37.01	43.19	30.64	31.55	37.07
Large-scale Universal Models							
Mixtral-8×7B-Instruct-v0.1	–	10-shot	23.32	31.87	32.89	29.96	27.79
GPT-3.5-1106-turbo	–	10-shot	25.60	37.60	28.04	32.22	29.29
GPT-4-1106-preview	–	10-shot	**60.94**	**50.19**	**35.57**	**43.91**	**53.47**

Table 8: We report the accuracy (%) results on MoleculeQA test set under different aspects (**Best** for model-wise).

follows: Pythia > BLOOM > GALACTICA > Llama, which may provide a reference for molecular base model selection. Increasing model size (e.g. 7B→13B) also receives mild accuracy gain.

- **Single v.s. Multiple modalities.** Both MoMu and MolCA are models that jointly incorporate molecular 2D graph modality and textual information. They demonstrate improvements over their base models (MolT5 and GALACTICA respectively) that solely rely on 1D-text modality.
- **Large-scale Universal Models.** The utilization of highly advanced models, such as GPT-4, has potential in the field of molecular research. In a 10-shot scenario, GPT-4 demonstrates accuracy comparable to certain specialized models. However, the performance of smaller models declined sharply, which may be attributed to the lack of their emergent abilities((Wei et al., 2022)).

5 Analysis

We propose the following research questions (RQs) for the molecular domain to guide our analysis:

- **RQ1**: Are existing LLMs powerful enough for application in practical molecular scenarios?
- **RQ2**: What factors are crucial for enhancing LLMs' ability for molecule comprehension?
- **RQ3**: Can MoleculeQA be adopted to alleviate the hallucinations in molecular LLMs?

5.1 In-depth Performance Analysis (RQ1)

We draw a preliminary conclusion from Table 8 that existing LLMs' comprehension of molecules is far from satisfactory: When confronted with aspects of Property and Application, pivotal for real-world applications, evaluated models consistently fail to achieve commendable accuracy. To more thoroughly assess the methods' level of comprehension across various molecular aspects, we plot T5-base and BioT5's accuracy over each subcategory in our taxonomy in Fig. 5. We find that in aspects of *Source* and *Structure*, two models exhibit consistent performance, with accuracy exceeding 40% across all categories. But on sub-topics like *Agricultural Chemical* and *Approval status*, two models perform notably sub-optimal. Various accuracy on different topics can serve as a confidence coefficient for related model applications.

5.2 Crucial Factor Attribution (RQ2)

We summarize the following crucial factors for improving molecular comprehension ability:

Molecular Corpora. The two T5 variants, MolT5 and BioT5, displays divergent outcomes. MolT5 performs worse compared to T5, while BioT5 demonstrates improved performance. This divergence can be attributed to the differences in their training corpora, specifically in terms of scale and diversity. Similarly, decoder-only mod-

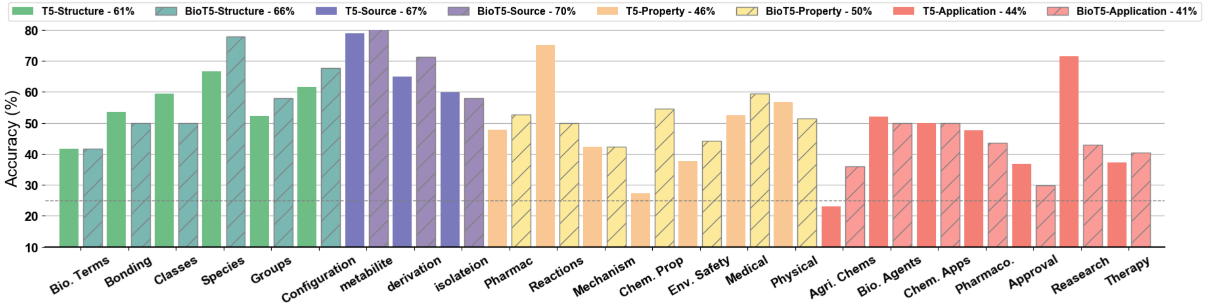


Figure 5: Accuracy of different finer topics under 4 coarse-grained aspects on the MoleculeQA test set. We select BioT5- and T5-base as representatives of Molecular LLM and General LLM, respectively, represented by solid and dashed bars.

els also exhibit this phenomenon: BioMedGPT (4.2M bio-molecular papers) > Mol-Instruction (1M molecular-oriented instruction samples) > Vicuna (70K general instruction samples) > Llama (General corpus). The above findings emphasize the importance of large, diverse, and high-quality molecular corpus for improving performance.

Modality Modeling Strategy. We investigate which modality modeling strategies can more effectively facilitate molecular modeling. (1) **Modality learning:** There is a significant performance gap between LoRA-based methods and methods employing multi-modal fusion or full fine-tuning, which underscores that an adequate scale of trainable parameters is necessary to master the molecule modalities. (2) **Multi-modal fusion:** MolCA and MoMu demonstrate that fusing molecular graphs into the semantic space of LLMs is viable. However, although they both deploy GIN as graph encoder, in comparison to MoMu's linear adaptation, MolCA's Q-Former (Li et al., 2022) graph adapter achieves a much more significant improvement.

5.3 Hallucination Alleviation (RQ3)

Reinforcement Learning (RL) from feedback has widespread applications for mitigating hallucinations (Yu et al., 2024; Gunjal et al., 2024). However, this method is rarely applied to molecule caption (Gkoumas and Liakata, 2024). To verify the feasibility of this approach, we adopt MoleculeQA to provide feedback to optimize the fine-tuned molecule caption models: Given a QA pair of molecule x, we designate the positive option as the preferred output y_w and one negative option as the dis-preferred output y_l, and employ Direct Preference Optimization (Rafailov et al., 2023) (DPO) as the RL strategy to optimize model π:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log\sigma\left(\beta\log\left(\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}\right) - \beta\log\left(\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right)\right],$$

where π_θ is the policy model parameterised by θ, π_{ref} is the fine-tuned model as the reference, β is a hyper-parameter and σ is the Sigmoid function.

We convert QA instances into mol2text format and remove molecules in ChEBI-20's test set, after optimizing MolT5 and BioT5 on this corpus with DPO, we evaluate their factual correctness like Section 3.1 and report the result in Table 9.

Model	Structure	Property	Application	Source
MolT5-base	63/0/34	1/4/3	7/15/8	20/10/30
MolT5-base-DPO	59/0/38	0/2/6	10/13/7	17/ 8/35
BioT5-base	62/0/35	2/3/3	9/12/9	16/13/31
BioT5-base-DPO	57/1/39	1/2/5	11/10/9	14/14/33

Table 9: **Comparison about Factual Correctness.** We manually evaluate two optimized models on the same 100 cases. An intuitive comparison is provided in Table 12.

The result indicates that, two models are guided by counterfactual negative options to discern correct/incorrect fine-grained molecular facts, and to generate descriptions that align better with ground truth across most aspects, except for *Application*. We attribute this to the small scale of *Application*.

6 Conclusion and Future Work

In conclusion, this paper addresses the absence of evaluation for factual correctness in Large Language Models (LLMs) within the molecular domain. By organizing molecular descriptions into a taxonomy and constructing QA pairs through human and LLM efforts, we introduce MoleculeQA, a novel dataset for molecular factual question answering. Our evaluation reveals shortcomings of existing models, emphasizing critical factors for molecular comprehension and providing guidance for molecular LLMs' development. We also make preliminary attempt to alleviate the hallucinations in molecular LLMs based on MoleculeQA. Looking forward, we propose three future directions: (1) Design a powerful molecular model based on our analysis. (2) Investigate more and better methods to apply MoleculeQA for the optimization of molecular LLMs. (3) Incorporate additional data sources to enrich MoleculeQA's comprehensiveness.

Limitations

We conclude our limitations into the following aspects: (1) Imbalanced data distribution across different aspects, notably with *Structure* and *Source* data dominating the majority. This skew results from the overall prevalence of structural and source-related information in the data sources. To address this, future efforts will focus on introducing more data related to properties and applications while expanding topic coverage and diversity, all while safeguarding against data leakage. (2) Absence of full fine-tuning for large models: Under the constraint of computational resources, we fail to fully fine-tune LLMs with 7B parameters and above, leading us to opt for adaptation-based fine-tuning methods. (3) We only conduct a preliminary attempt to alleviate the issue of model hallucinations, the potential of MoleculeQA is left for further exploration.

Potential Risks

Although MoleculeQA offers a viable approach for factual assessment in the molecular domain with reliable data quality, there remains a risk of misuse. Evaluations on this dataset may not accurately represent a model's comprehension over all molecules. MoleculeQA could potentially be leveraged to furnish a veneer of reliability for models with underlying risks.

References

CAMEO Chemicals. https://cameochemicals.noaa.gov/.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2021. MolGPT: Molecular generation using a transformer-decoder model. *Journal of chemical information and modeling*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Stella Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *ArXiv*, abs/2304.01373.

Andrés M Bran and Philippe Schwaller. 2023. Transformers and large language models for chemistry and drug discovery. *ArXiv*, abs/2310.06083.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Carl N. Edwards, T. Lai, Kevin Ros, Garrett Honke, and Heng Ji. 2022. Translation between molecules and natural language. *ArXiv*, abs/2204.11817.

Carl N. Edwards, Chengxiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Conference on Empirical Methods in Natural Language Processing*.

Benedek Fabian, Thomas Edlich, H'el'ena Gaspar, Marwin H. S. Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *ArXiv*, abs/2011.13230.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *ArXiv*, abs/2306.08018.

Dimitris Gkoumas and Maria Liakata. 2024. Feedback-aligned mixed llms for machine language-molecule translation. *arXiv preprint arXiv:2405.13984*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv: Learning*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *ArXiv*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Y. Zaslavsky, Jian Zhang, and Evan E. Bolton. 2022. Pubchem 2023 update. *Nucleic acids research*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. *Cornell University - arXiv,Cornell University - arXiv*.

Hugo Laurenccon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo Gonz'alez Ponferrada, Huu Nguyen, Jorg Frohberg, Mario vSavsko, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, So maieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, S. Longpre, Sebastian Nagel, Leon Weber, Manuel Sevilla Muñoz, Jian Zhu, Daniel Alexander van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa Etxabe, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Trung Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset. *ArXiv*, abs/2303.03915.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Peng Xie. 2023. DrugChat: Towards enabling chatgpt-like capabilities on drug molecule graphs. *ArXiv*, abs/2309.03907.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *ArXiv*, abs/2212.10789.

Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Knowledge Discovery and Data Mining*.

Zequn Liu, W. Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Yang Zhang, and Tie-Yan Liu. 2023a. MolXPT: Wrapping molecules with text for generative pre-training. *ArXiv*, abs/2305.10688.

Zhiyuan Liu, Sihang Li, Yancheng Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Conference on Empirical Methods in Natural Language Processing*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*.

Yi Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023a. MolFM: A multimodal molecular foundation model. *ArXiv*, abs/2307.09484.

Yi Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023b. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *ArXiv*, abs/2308.09442.

Terry Mun, Asger Bachmann, Vikas Gupta, Jens Stougaard, and Stig U Andersen. 2016. emphLotus base: An integrated information portal for the model legume emphLotus japonicus. *Sci Rep*, 6:39447.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Conference on Computational Natural Language Learning*.

OpenAI. 2023a. ChatGPT: A language model for conversational ai.

OpenAI. 2023b. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *International Conference on Language Resources and Evaluation*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *ArXiv*, abs/2310.07276.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. arxiv 2023. *arXiv preprint arXiv:2305.18290*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7.

Teven Le Scao. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get your atoms in order - an open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of chemical information and modeling*, 55 10:2111–20.

T. Sterling and John J. Irwin. 2015. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55:2324 – 2337.

Bing Su, Dazhao Du, Zhao-Qing Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Haoran Sun, Zhiwu Lu, and Ji rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *ArXiv*, abs/2209.05481.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A machine comprehension dataset. In *Rep4NLP@ACL*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zhuoyu Wei, Wei Ji, Xiubo Geng, Yining Chen, Baihua Chen, Tao Qin, and Daxin Jiang. 2020. Chemistryqa: A complex question answering dataset from chemistry.

David Scott Wishart, David Arndt, Allison Pon, Tanvir Sajed, Anchi Guo, Yannick Djoumbou, Craig Knox, Michael Wilson, Yongjie Liang, Jason R. Grant, Yifeng Liu, Seyed Ali Goldansaz, and Stephen M. Rappaport. 2014. T3db: the toxic exposome database. *Nucleic Acids Research*, 43:D928 – D934.

David Scott Wishart, Yannick Djoumbou Feunang, Anchi Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2017. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46:D1074 – D1082.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv*, abs/2311.16502.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. 2023. MaScQA: A question answering dataset for investigating materials science knowledge of large language models. *ArXiv*, abs/2308.09115.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*.

Haoxiang Zhong, Chaojun Xiao, Cunchao Tu, T. Zhang, Zhiyuan Liu, and Maosong Sun. 2019. JEC-QA: A legal-domain question answering dataset. *ArXiv*, abs/1911.12011.

A Appendix

A.1 Data Sources and License

As depicted in Table 11, we elaborate on the origins and legal permissions associated with each data component utilized in the development of the MoleculeQA. This encompasses both biomolecular data and textual descriptions. Thorough scrutiny was conducted on all data origins to confirm compatibility with our research objectives and subsequent utilization. Proper and accurate citation of these data sources is consistently maintained throughout the paper.

A.2 Details about Taxonomy

We present the overall hierarchical structure of the taxonomy upon which MoleculeQA is based in Figure 6. Additionally, Table 10 provides details regarding the subtopics and part of leaf topics encompassed within each of the four aspects: *Structure*, *Source*, *Property*, and *Application*.

A.3 Experimental Setup Details

A.3.1 Baselines

The following parts will individually introduce the models we evaluated in this study and the approaches used for implementation.

T5 (Raffel et al., 2019) is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks for which each task is converted into a text-to-text format. We directly fine-tuned it on MoleculeQA dataset from public checkpoints [4] with three different model sizes: small, base and large. It's important to note

[4] https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md#t511

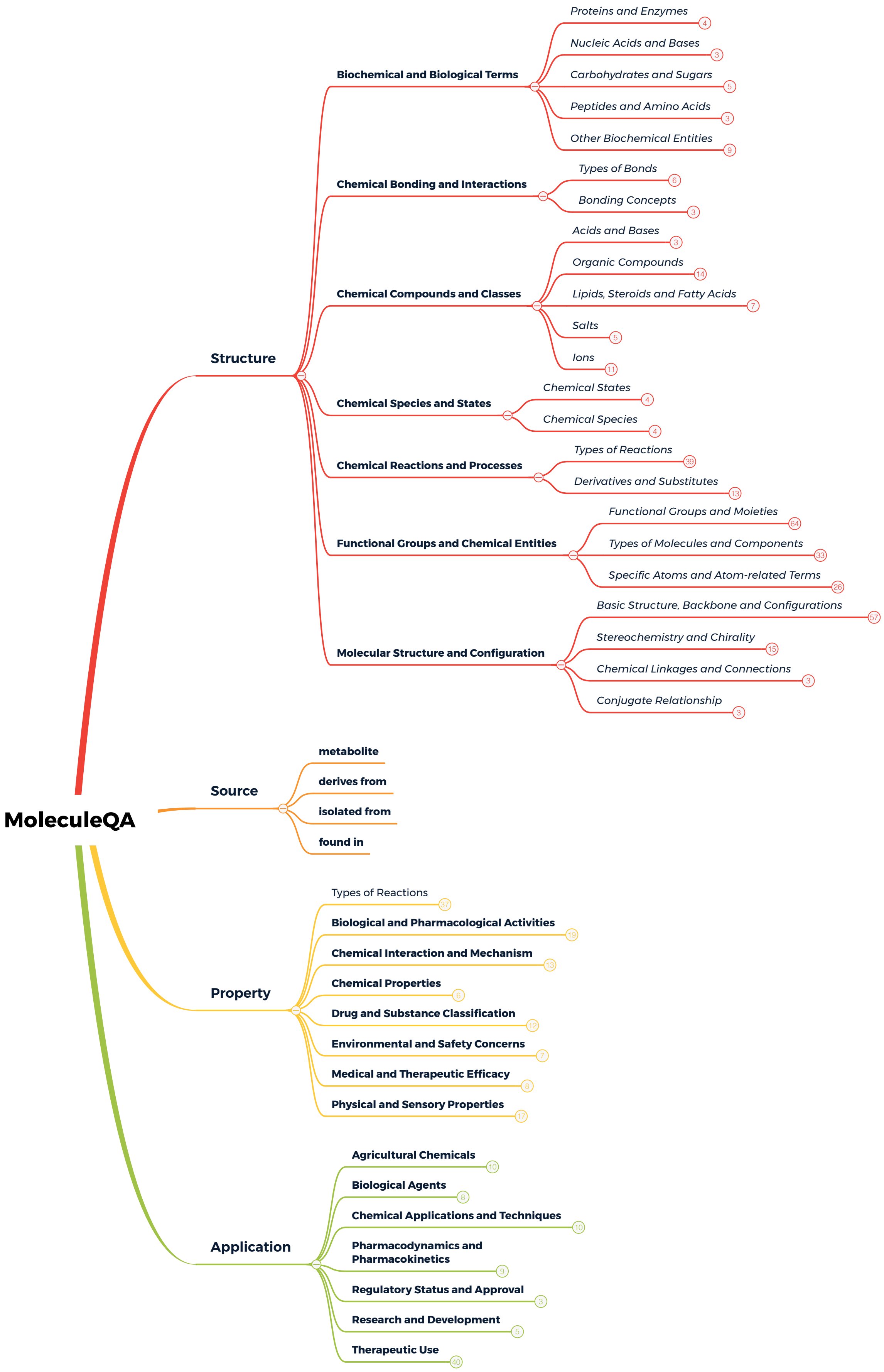

MoleculeQA
Structure
Biochemical and Biological Terms
Proteins and Enzymes
4
Nucleic Acids and Bases
3
Carbohydrates and Sugars
5
Peptides and Amino Acids
3
Other Biochemical Entities
9
Chemical Bonding and Interactions
Types of Bonds
6
Bonding Concepts
3
Chemical Compounds and Classes
Acids and Bases
3
Organic Compounds
14
Lipids, Steroids and Fatty Acids
7
Salts
5
Ions
11
Chemical Species and States
Chemical States
4
Chemical Species
4
Chemical Reactions and Processes
Types of Reactions
39
Derivatives and Substitutes
13
Functional Groups and Chemical Entities
Functional Groups and Moieties
64
Types of Molecules and Components
33
Specific Atoms and Atom-related Terms
26
Molecular Structure and Configuration
Basic Structure, Backbone and Configurations
57
Stereochemistry and Chirality
15
Chemical Linkages and Connections
3
Conjugate Relationship
3
Source
metabolite
derives from
isolated from
found in
Property
Types of Reactions
37
Biological and Pharmacological Activities
19
Chemical Interaction and Mechanism
13
Chemical Properties
6
Drug and Substance Classification
12
Environmental and Safety Concerns
7
Medical and Therapeutic Efficacy
8
Physical and Sensory Properties
17
Application
Agricultural Chemicals
10
Biological Agents
8
Chemical Applications and Techniques
10
Pharmacodynamics and Pharmacokinetics
9
Regulatory Status and Approval
3
Research and Development
5
Therapeutic Use
40

Figure 6: The overarching structure of the MoleculeQA taxonomy comprises multiple aspects and subtopics arranged hierarchically to categorize various facets of molecular factual knowledge. Due to space constraints, we did not elaborate on all leaf topics.

ASPECT	SUB TOPICS	LEAF TOPICS
Property	*Biological and Pharmacological Activities*	"antimicrobial activity", "anti-neoplastic activity", "antioxidant activity", "enzyme inhibition", "ion channel activity", "receptor activity"...
	Reaction Types	"acetylation", "condensation", "dehydrogenation", "epoxidation", "glycosylation", "hydroxylation", "oxidation", "phosphorylation", "reduction"...
	Chemical Interaction and Mechanism	"action", "affinity", "binding", "conversion", "decomposition", "duration", "formation", "mechanism", "reaction/binding", "receptor affinity", "selectivity"...
	Chemical Properties	"chemical nature", "sensitivity", "ph value", "stability", "valence", "reactivity"
	Environmental and Safety Concerns	"bio-accumulation", "xenobiotic", "cell permeability", "teratogenic agent", "environmental contaminant", "resistance", "safety concerns"
	Medical and Therapeutic Efficacy	"analgesic activity", "anti-inflammatory activity", "antimalarial activity", "anti-mycobacterial", "carcinogenicity", "medical effects", "potency"...
	Physical and Sensory Properties	"abundance", "atomic mass", "boiling point", "color", "half-life", "odor", "optical activity", "physical state", "solubility", "taste", "volatileness"...
Application	*Agricultural Chemicals*	"fungicide", "herbicide", "insecticide", "disease control", "herbicide safener", "synthetic auxin", "phytoestrogen"...
	Biological Agents	"antibiotic", "antifungal drug", "antibacterial drug", "antiprotozoal", "antiviral drug", "nematicide", "acaricide", "antiseptic"...
	Chemical Applications and Techniques	"reagent", "indicator", "detection", "derivatisation agent", "fluorescent dye", "production", "chromatographic reagent", "tracer", "solvent", "food additive"...
	Pharmacodynamics and Pharmacokinetics	"inhibitor", "antagonist", "prodrug", "modulator", "sympathomimetic agent", "allergen", "sodium channel blocker", "ligand", "agonist"...
	Regulatory Status and Approval	"approval", "withdrawn from market", "registered in"...
	Research and Development	"experimental", "biomarker", "clinical development", "testing"...
	Therapeutic Use	"anti-arrhythmia drug", "anti-allergic agent", "anti-asthmatic drug", "anticoronaviral agent", "anti-neoplastic agent", "anti-ulcer drug", "anti-HIV agent", "orphan drug", "recreational drug", "vasodilator"...
Source	*found in*	"found in"
	metabolite	"metabolite"
	derives from	"derives"
	isolated from	"isolated"
Structure	*Biochemical and Biological Terms*	"active metabolite", "alkaloid" "coenzyme a", "enzyme", "epitope", "fatty acyl coa", "glucoside", "hapten", "nucleobase", "oligosaccharide", "sphingoid base", "substrate"...
	Chemical Bonding and Interactions	"glycosidic bond", "disulfide bonds", "double bond", "exocyclic double bond", "peptide bond", "c=c double bond", "bond", "connection", "attachment"...
	Chemical Compounds and Classes	"acid", "alcohol", "amine", "cation", "dimer", "enamide", "hydrochloride", "ion", "lactam", "polyphenol", "salt", "phosphate", "sulfate", "oxoanion", "zwitterion"...
	Chemical Species and States	"anhydrous form", "heptahydrate form", "oxidation state", "hydrate", "major microspecies", "deoxygenated", "major species", "microspecies"...
	Functional Groups and Chemical Entities	"acyl group", "alcohol group", "alkyl group", "anilino group", "carbamoyl group", "chloro group", "epoxy group", "ester group", "fatty acyl group", "hydrazino group", "hydroperoxy group", "isopropyl substituent", "keto group", "methyl group", "oxo group", "pentyl group", "phosphate group", "primary hydroxy group", "s-acyl component", "s-methyl group", "sulfo group", "thiol group"...
	Molecular Structure and Configuration	"alpha-branch", "alpha-carbon", "backbone", "branch", "bridge", "core", "composition", "configuration", "linked group", "n-substituent", "oh groups", "omega-hydroxy", "position", "prenyl units", "terminal", "terminal group", "glycosyl fragment", "repeating unit", "sequence", "subcomponents", "side chain", "nucleus", "sugar fragment", "unit"...

Table 10: Taxonomy of *Property*, *Application*, *Structure* and *Source* aspects in MoleculeQA. **Leaf Topics** correspond to the most granular concepts, while **Sub Topics** aggregate leaf topics further. The table presents only a subset of leaf topics.

DATA SOURCES	LICENSE URL	LICENSE NOTE
PubChem	https://www.nlm.nih.gov/web_policies.html	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.
FDA Pharm Classes	https://www.fda.gov/about-fda/about-website/website-policies	Unless otherwise noted, the contents of the FDA website (www.fda.gov), both text and graphics, are not copyrighted. They are in the public domain and may be republished, reprinted and otherwise used freely by anyone without the need to obtain permission from FDA. Credit to the U.S. Food and Drug Administration as the source is appreciated but not required.
Drug Bank	https://creativecommons.org/licenses/by-nc/4.0/legalcode	Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to: reproduce and Share the Licensed Material, in whole or in part, for NonCommercial purposes only; and produce, reproduce, and Share Adapted Material for NonCommercial purposes only.
ChEBI	https://creativecommons.org/licenses/by/4.0/	You are free to: Share — copy and redistribute the material in any medium or format. Adapt — remix, transform, and build upon the material for any purpose, even commercially.
LOTUS	https://lotus.nprod.net/	LOTUS is one of the biggest and best-annotated resources for natural products occurrences available free of charge and without any restriction.
CAMEO Chemicals	https://cameochemicals.noaa.gov/help/reference/terms_and_conditions.htm?d_f=false	CAMEO Chemicals and all other CAMEO products are available at no charge to those organizations and individuals (recipients) responsible for the safe handling of chemicals.
Toxin-Toxin-Target Database (T3DB)	http://www.t3db.ca/	T3DB is offered to the public as a freely available resource. Use and re-distribution of the data, in whole or in part, for commercial purposes requires explicit permission of the authors and explicit acknowledgment of the source material (T3DB) and the original publication.

Table 11: Data resources and licenses utilized in data collection for MoleculeQA.

that the original T5 pre-training does not incorporate any specific knowledge related to the domain of molecules.

MolT5 (Edwards et al., 2022) undergoes joint training on molecule SMILES from the ZINC-15 dataset (Sterling and Irwin, 2015) and a general corpus from the C4 dataset (Raffel et al., 2019), enabling MolT5 to acquire prior knowledge in both of these domains. It contains three different sizes: small, base, and large. In the experiment, we utilized pre-trained model checkpoints of various sizes [5] released by the authors. Subsequently, we conducted full fine-tuning on the MoleculeQA train set, followed by evaluating on the test set.

MoMu (Su et al., 2022) is pre-trained using molecular 2D graphs and their semantically related textual data (crawled from published Scientific Citation Index papers) via contrastive learning. We adopted MoMu-K pre-trained checkpoints [6] where the text encoder is initialized with the weights of KV-PLM (Zeng et al., 2022). Following the original methodology, we injected encoded graph features into MolT5-base & large and conducted fine-tuning on MoleculeQA.

BioT5 (Pei et al., 2023) as a comprehensive pre-training framework, builds upon the methodology of MolT5 while enhancing cross-modal integration into biology through chemical knowledge and natural language associations. It leverages SELFIES for robust molecular representations and extracts knowledge from the surrounding context of bio-entities in unstructured biological literature. We utilized the official base version pre-trained checkpoint [7] and converted the MoleculeQA data into the corresponding format for fine-tuning.

MolCA (Liu et al., 2023b) facilitates a language model (LM), such as Galactica, in comprehending both text- and graph-based molecular contents through its cross-modal projector. This projector, implemented as a Q-Former, serves to bridge the representation space of a graph encoder with the text space of an LM. Additionally, MolCA employs a uni-modal adapter to enable efficient adaptation of the LM to downstream tasks. We conducted pre-training, including both stage 1 and stage 2, on the 125M and 1.3B versions, based on the official code

[5] https://huggingface.co/laituan245/molt5-small, https://huggingface.co/laituan245/molt5-base/, https://huggingface.co/laituan245/molt5-large/

[6] https://github.com/ddz16/MoMu?tab=readme-ov-file#pretrain

[7] https://huggingface.co/QizhiPei/biot5-base

and cleaned data [8]. Subsequently, we performed finetuning on MoleculeQA.

BioMedGPT-LM-7B (Luo et al., 2023b) It is a large generative language model based on Llama2 in the biomedical domain. It was fully fine-tuned from the Llama2-7B-Chat with millions of biomedical papers from the S2ORC corpus (Lo et al., 2020). We directly apply the LoRA finetuning method on the checkpoint [9] provided by the official source.

OPT (Zhang et al., 2022) is a series of open-sourced large causal language models which perform similar in performance to GPT-3 (Brown et al., 2020). For comparison with fully fine-tuned T5 series models, we opted to fully fine-tune OPT-125M, -350M, and -1.3B size models on MoleculeQA. In our implementation, we referred to the interfaces provided by Hugging Face [10].

GALACTICA (Taylor et al., 2022) is a large language model (LLM) for Science: trained on over 48 million papers, textbooks, reference material, compounds, proteins and other sources of scientific knowledge. We selected GALACTICA-125M, -1.3B, and -7.1B versions of the model [11] and conducted fine-tuning using LoRA on MoleculeQA.

Pythia (Biderman et al., 2023) is an open suite of large language models, all trained on public data in the same order. These models vary in size, ranging from 70M to 12B parameters. They were trained on the Pile dataset, which is constructed from 22 diverse high-quality subsets. We opted to conduct finetuning based on LoRA on the standard versions of Pythia-410M, -1B, -2.8B, -6.9B, and -12B sizes models [12].

BLOOM (Scao, 2022) is an autoregressive large language model, trained to continue text from a prompt on vast amounts of text data using industrial-scale computational resources. It was trained on the ROOTS (Laurenccon et al., 2023) corpus, a dataset comprising hundreds of sources in 46 natural and 13 programming languages (59 in total). For model scaling evaluation, we chose to conduct finetuning based on LoRA on the BLOOM-560M, -1.1B, -1.7B, -3B, and -7.1B sizes versions of the model [13]. Subsequently, we provided the results on the MoleculeQA test set.

LLaMA-2 (Touvron et al., 2023b) is a collection of large language models with parameters ranging from 7 billion to 70 billion. The model architecture remains largely unchanged from that of LLaMA-1 models (Touvron et al., 2023a), but 40% more data was used to train the foundational models. Specifically, Llama 2 includes pre-trained and fine-tuned models optimized for dialogue applications, termed Llama 2-Chat. We opted to utilize the LLaMA-2-Chat 7B and 13B models [14] and transformed MoleculeQA into instruction samples for LoRA fine-tuning.

Vicuna-v-1.5 (Chiang et al., 2023) is an open-source chatbot that has been trained by fine-tuning LLaMA on over 150K user-shared conversations collected from ShareGPT.com. Preliminary evaluation, conducted with GPT-4 as the judge, demonstrates that the Vicuna series achieves competitive performance when compared to OpenAI ChatGPT, while also outperforming other models such as LLaMA. We selected the v1.5 series models and conducted LoRA Finetuning on both the 7B and 13B versions [15].

Mol-Instructions-7B (Fang et al., 2023) is a low-rank adapter designed for LLaMA-2 base LLM, specifically trained on molecule-oriented instructions sourced from the Mol-Instructions dataset. We utilize the version tailored for LLaMA-2-Chat [16], merging the adapter back to the base LLM before proceeding with LoRA fine-tuning.

Mixtral-8×7B (Jiang et al., 2024) is a Sparse Mixture of Experts (SMoE) language model consisting of a decoder-only architecture. Its feedforward block selects from a set of 8 distinct groups of parameters. Notably, it is recognized as the most robust open-weight model currently available, licensed under Apache 2.0. We adopt a locally deployed approach for conducting few-shot prompting inference.

GPT-3.5-turbo and GPT-4. For closed-source models such as OpenAI GPT Family GPT-3.5-turbo (OpenAI, 2023a) and GPT-4 (OpenAI,

[8] https://github.com/acharkq/MolCA

[9] https://huggingface.co/PharMolix/BioMedGPT-LM-7B

[10] https://huggingface.co/docs/transformers/model_doc/opt

[11] https://huggingface.co/models?other=galactica

[12] https://huggingface.co/models?other=pythia

[13] https://huggingface.co/docs/transformers/model_doc/bloom

[14] https://huggingface.co/docs/transformers/model_doc/llama2

[15] https://huggingface.co/lmsys/vicuna-7b-v1.5, https://huggingface.co/lmsys/vicuna-13b-v1.5

[16] https://huggingface.co/zjunlp/llama2-molinst-molecule-7b

2023b), we employ batch inference via APIs for conducting few-shot prompt inference. This approach significantly enhances evaluation efficiency and reduces overhead.

A.3.2 Hyper-parameters

For MolT5, MoMu, T5, and BioT5, we employed the original codebases and hyper-parameters provided in the respective papers for full fine-tuning. Specifically, these models were trained on a single NVIDIA 48GB A6000 GPU. Except for BioT5, which had a learning rate set to 1e-3, the learning rates for all other models were set to 1e-4. All models underwent fine-tuning for 100 epochs on the training set, and the checkpoint with the best performance on the development set was selected for evaluation on the test set.

For MolCA, we utilized the author's recently updated dataset (excluding any data leakage concerns) and conducted pre-training stage 1 and stage 2 training on 2 NVIDIA 48GB A6000 GPUs. We maintained consistency with the training hyper-parameters provided in the original paper. Subsequently, we fine-tuned pre-trained checkpoints of different sizes on MoleculeQA, with a total batch size set to 16. The 125M model was trained on a single GPU card, while the 1.3B model was trained on two GPU cards. The fine-tuning total epochs were set to 100 for all versions.

For full fine-tuning of the OPT series, we conducted training on 4 A6000 GPUs for the 125M and 350M versions and 8 GPUs for the 1.3B version. The total batch size was set to 256, and the learning rates were set to 3e-4 and 2e-4 for the respective versions. All other hyper-parameters were kept consistent with those specified in the original paper. We performed full fine-tuning for 60 epochs, as we observed over-fitting phenomena when exceeding 50 epochs.

For the remaining experiments based on LoRA tuning, we employed the Alpaca-LoRA codebase for instruction fine-tuning. Except for the 13B size model trained on 8 A6000 GPUs, all other models were trained on 4 GPUs. The total batch size was set to 400, with gradient accumulation and learning rate adjusted according to the model size (typically set to 3e-4). We set the total training epochs to 20.

Regarding the LoRA configuration, we utilized the PEFT [17] library for implementation. We set LoRA's rank r as 16, α as 16, dropout rate as 0.05, and applied LoRA to all modules of `["q/k/v/o_proj", "gate_proj", "down/up_proj"]` (adjusting module names if necessary based on actual implementation). Equivalent trainable parameters are reported in Table 8.

We implemented DPO using the alignment-handbook [18] library and retained all the hyper-parameters.

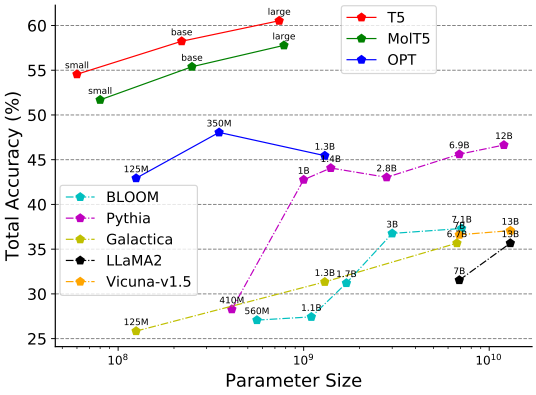


Figure 7: Model parameter size vs. total accuracy on MoleculeQA test set. Solid lines denote full fine-tune models, and dashed lines represent LoRA fine-tune.

A.4 Scaling Law for Molecular LLMs

In Fig. 7 and Fig. 8, we depict the variations in overall accuracy and aspect-specific accuracy of several models over increasing model scale.

For fully fine-tuned models, we conduct comparisons between T5-based models (T5 and MolT5) and decoder-only models (represented by OPT). To validate whether adaptively fine-tuning can adapt general LLMs to acquire molecular knowledge, we compared models such as BLOOM, Pythia, and the LLaMA2-series models using LoRA fine-tuning.

We observe a pronounced scaling effect across different training methods and model architectures, with the scale effect being more evident in the full fine-tuning approaches. This observation is consistent with previous analysis about the scale of parameters and indicates that scaling up model size is a promising way to enhance molecular modeling.

A.5 Examples of Hallucination Alleviation

Here, we present selected examples (Table 12) comparing the originally generated captions with those refined using the DPO method discussed in Section 5.3

[17] `https://github.com/huggingface/peft`

[18] `https://github.com/huggingface/alignment-handbook`

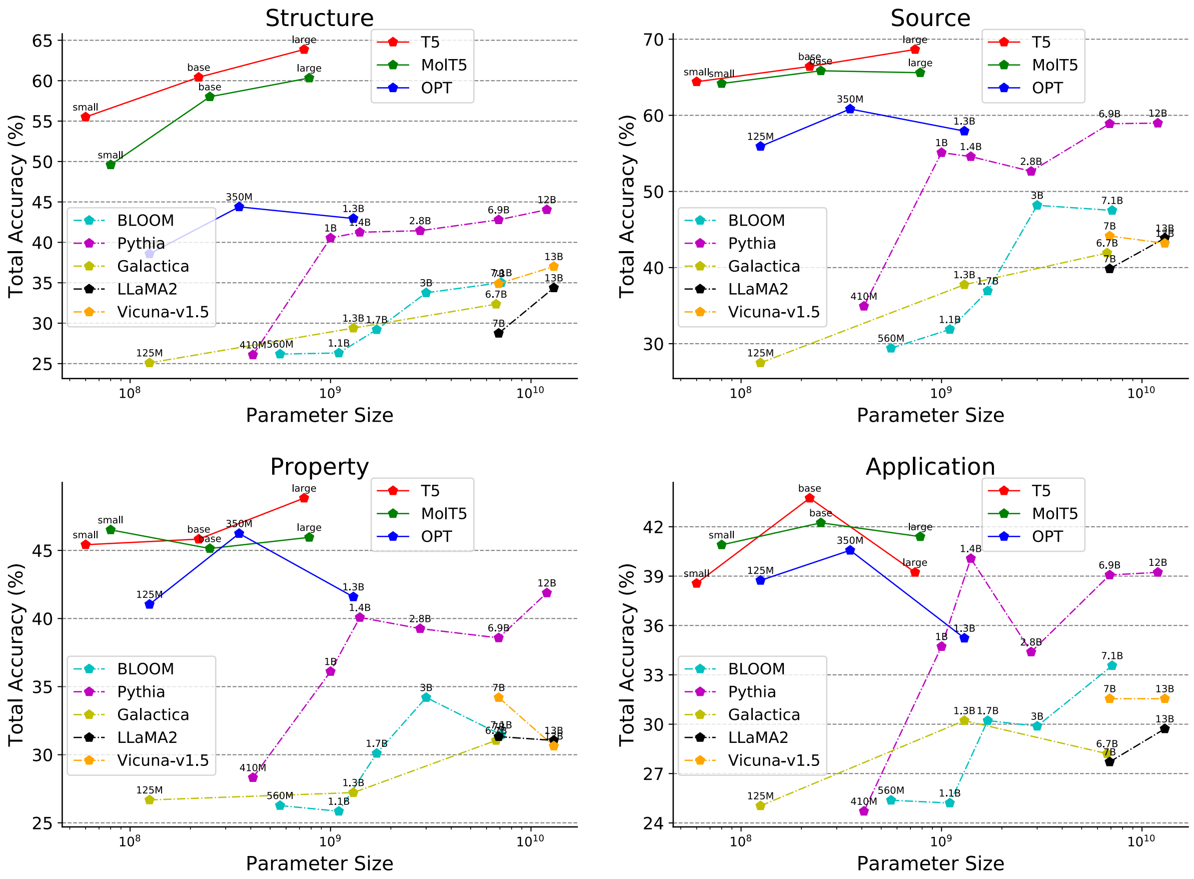


Figure 8: Model parameter size vs. Accuracy in four aspects.

A.6 Prompt of Different Tasks for LLM

In the construction process of MoleculeQA, we deploy LLMs to finish the following tasks: (1) Corpus classification; (2) Topic extraction; (3) Answer generation; (4) Semantic consistency validation. We report the definitions and task contexts, which are components of prompts for LLMs, of these tasks in Table 13.

A.7 Few-Shot Details and Prompt Exhibition

We introduce details about our few-shot setting: For each aspect, we select a representative and various samples from different topics as examples to construct an aspect-specific 10-shot prompt. We demonstrate selected samples and the format of prompt in *Source* aspect in Table 14.

CID	Ground-Truth	Original	DPO Refined
5280695	The molecule is a monomethoxyflavone that is the 8-O-methyl derivative of gossypetin. It is a pentahydroxyflavone, a member of flavonols and a monomethoxyflavone. It derives from a gossypetin. It is a conjugate acid of a 3’,4’,5,7-pentahydroxy-8-methoxyflavon-3-olate.	The molecule is a monomethoxyflavone that is quercetin methylated at position 7. It has a role as a metabolite, an antioxidant and an anti-inflammatory agent. It is a monomethoxyflavone, a tetrahydroxyflavone and a 7-hydroxyflavonol. It derives from a quercetin.	The molecule is a 7-hydroxyflavonol that is quercetin substituted by a methoxy group at position 8. It has a role as a plant metabolite. It is a tetrahydroxyflavone, a 7-hydroxyflavonol, a monomethoxyflavone and a member of flavonols. It derives from a quercetin.
91972219	The molecule is an amino disaccharide that consists of N-acetyl-D-glucosamine having an N-acetyl-6-O-sulfo-beta-D-galactosaminyl residue attached at O-4. It is an amino disaccharide, a glucosamine oligosaccharide and a galactosamine oligosaccharide.	The molecule is a linear amino disaccharide that consists of N-acetyl-beta-D-glucosamine and N-acetyl-D-galactosamine units connected in sequence by (1->4) links. It has a role as an epitope. It is an oligosaccharide sulfate, an amino disaccharide and a glucosamine oligosaccharide.	The molecule is an amino disaccharide consisting of N-acetylglucosamine having an O-sulfated N-acetylglucosaminyl residue attached at the 4-position. It has a role as an epitope. It is an amino disaccharide, a glucosamine oligosaccharide and a galactosamine oligosaccharide.
53239760	The molecule is a fifteen-membered glycopeptide comprising glycyl, glutaminyl, alanyl, glycyl, 3-(1,3-thiazol-4-yl)alanyl, (5R)-5-(beta-D-galactopyranosyloxy)lysyl, glycyl. alpha-glutamyl, glutaminyl, glycyl, prolyl, lysyl, glycyl, alpha-glutamyl and threonine residues coupled in sequence.	The molecule is a fifteen-membered glycopeptide comprising glycyl, glutaminyl, alanyl, glycyl, 3-(1,3)-alpha-D-glutamyl, (5R)-5-(beta-D-galactopyranosyloxy)lysyl, (5R)-5-(beta-D-glucopyranosyloxy)lysyl, glycyl. alpha-glutamyl, glutaminyl, glutaminyl, prolyl, lysyl, glycyl, alpha-glutamyl and threonyl residues coupled in sequence.	The molecule is a fifteen-membered glycopeptide comprising glycyl, glutaminyl, alanyl, glycyl, 3-(1,3-thiazol-4-yl)alanyl, (5R)-5-(beta-D-galactopyranosyloxy)lysyl, glycyl. alpha-glutamyl, glutaminyl, glycyl, prolyl, lysyl, glycyl, alpha-glutamyl and threonine residues coupled in sequence.

Table 12: Comparison between the original generated captions of BioT5-base and those generated by DPO-refined BioT5-base discussed in Section 5.3. Errors relative to the ground truth are highlighted in red. Notably, the model trained using data transformed by MoleculeQA and the DPO method exhibits fewer factual inaccuracies.

Task	Definition	Task Context
Corpus Classification	Classify molecular descriptions from the data source into one of four aspects.	You are a research assistant for molecular research. Please help me to classify some corpus. Four kinds of content are included in this corpus : The first is Source, which describes... The second is...
Topic Extraction	Extract attributes of molecules in specific aspect from original descriptions.	You are a chemical research assistant, you are familiar with description text of molecules, you need to help me extract molecules’ Source information, which describes...
Answer Generation	Generate answer for given question with original description	You are a chemistry research assistant, and I need you to complete the following task: You will be given a detailed description of a molecule and a question, please extract specific information from the given description to answer the question...
Semantic Consistency Validation	Check if generated answer has consistent semantic with original description.	You are a chemistry research assistant, and I need you to complete the following task: You will be given a description of a molecule and a sentence transcribed from it, please justify whether their semantics are consistent...

Table 13: Definition and context for each task. We prompt LLMs to finish these tasks for MoleculeQA construction.

```
messages = [{"role":"system", "content": f"""
You are a chemistry research assistant, and I'd like to test your professional ability on molecule understanding, please complete the following task:

You are provided with the SMILES representation of a molecule and asked a question about the molecule's source-related knowledge (Source means the natural or synthetic origin, as well as the production context related to a molecule), with four options given. Three of these options do not describe the given molecule, and you must select the correct option.

Here are several examples to show how to finish the Question Answering task:

###

Example 1:

Molecular SMILES: C1=CC(=CC=C1/C=CC(=O)O[C@@H]([C@H](C(=O)O)O)C(=O)O)O

Question: Which molecule does this molecule derive from?

Choices:

A: It derives from a meso-tartaric acid and a cis-4-coumaric acid.

B: It derives from a meso-tartaric acid and a cis-caffeic acid.

C: It derives from a cyanidin cation and a cis-4-coumaric acid.

D: It derives from a cis-vaccenic acid and an oleic acid.

Answer: A

###


###

Example 2:

Molecular SMILES: COC1=C(C=C(C=C1)C=O)OC

Question: Where this molecule can be found?

Choices:

A: It can be found in leaves and fruit of cowberry Vaccinium vitis-idaea, grape seeds and beer.

B: It can be found in peppermint, ginger, raspberry, and other fruits.

C: It can be found in edible vegetables, grains, and fruits.

D: It can be found in grape seeds, in Hibiscus cannabinus (kenaf) root and bark, in apple and in cacao.

Answer: B

###

...


Notice that here are some rules you need to follow:

1. Your answer for each question should be one of A/B/C/D, which corresponds to the four options.

2. For my convenience, please give me a list of ANSWERs for the given instances in the format 'Answer X: ...', without any other information.

"""}

{ "role":"user", "content": f"""

Please give me your choices for these instances in the above examples' styles. No other information is required.

Instance ID: <Instance ID>

Molecular SMILES: <Instance SMILES>

Question: <Instance Question>

Choices: <Instance Choices>

"""}

]
```

Table 14: An illustration depicting the process of constructing few-shot in-context-learning prompts for MoleculeQA test set inference with GPT-4-like large-scale universal models.