Chem2Text: A Multimodal AI for General-Purpose Molecule Question Answering

Anonymous ACL submission

Abstract

The structural attributes of chemical compounds are crucial for understanding their functions in drug discovery, but laboratory-based elucidation is resource-intensive. Current AI models often cannot handle open-ended queries necessary for hypothesis testing. We introduce Chem2Text, an interactive chatbot that interface NLP with multimodal knowledge to analyze chemical structures. Trained on 300,000 question-answer pairs from PubMed abstracts, Chem2Text generates hypotheses by providing relevant answers about compound properties. Benchmarking shows superior performance in general-purpose molecular question-answering compared to existing models, positioning Chem2Text as a versatile tool for pharmaceutical research.

1 Introduction

Predicting chemical compound properties is crucial in drug discovery, traditionally relying on resource-intensive lab analyses. Advances in computational methods, particularly NLP and multi-modal models, now enable efficient modeling of chemical structures like SMILES representation into dense, information-rich embeddings . NLP models excel in tasks like classification, regression, and generation (Jararweh et al., 2024; Ahmad et al., 2022; Edwards et al., 2022; Liu et al., 2023b), speeding up the identification of compound properties and their biological effects. However, existing AI models lack interactivity and the ability to handle complex, open-ended queries, limiting their use for real-time, iterative hypothesis testing by biologists and chemists.

To address this, we introduce *Chem2Text*, a bimodal AI that integrates SMILES-based molecular data with NLP. We also devised 300,000 complex PubMed-derived question-answer for compounds. Leveraging this data, Chem2Text builds on the LLaVA multimodal reasoning model to provide real-time, interactive analysis of chemical properties, ability to answer complex and open-ended queries, enabling users to generate actionable insights for laboratory research.

2 Related works

Instruction tuning has significantly enhanced the capabilities of large language models (LLMs), allowing them to excel in tasks like summarization, translation, question answering, and zero-shot learning across various domains. Recently, it has been extended to multimodal models, combining text with other data types like images and protein sequences for tasks such as captioning and molecular property prediction. In drug discovery, instruction-tuned models have shown promise in addressing challenges like molecular interactions and property prediction, with both task-specific and broader general-purpose models emerging in this space. Related works is described in details in Appendix **A**.

3 Methodology

3.1 Chem2Text-QA Dataset Collection

Available datasets on compound's question and answer such as Chebl-20 (Edwards et al., 2021) and others (Fang et al., 2024) are mainly focused on specific compound properties and molecule captioning. In order to create a 300K comprehensive compound's question and answer dataset (Chem2Text-QA) which covers wide range of questions about researches that have been carried out on compounds, we utilized information regarding compounds found in published articles in the PubMed Central (PMC) database (Consortium, 2015). The data collection procedures are discussed in Appendix B.

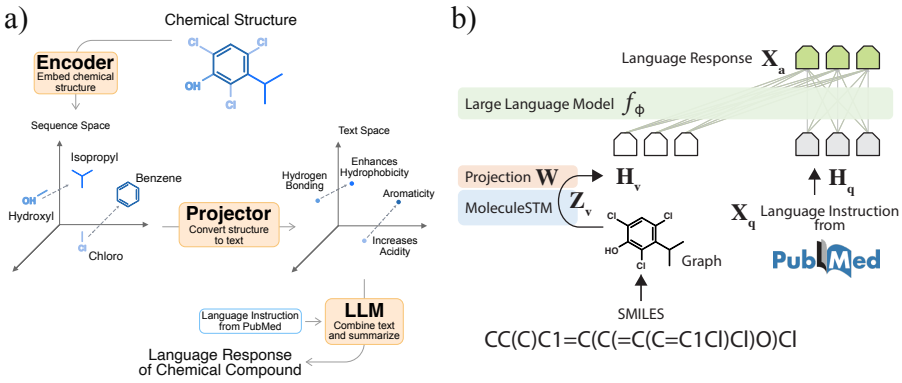

Figure 1: Overview of our proposed methodology. a) **Chem2Text** generates descriptive text from chemical structures (i.e. SMILES) by combining pre-trained SMILES encoder and language models. Compound tokens are first embedded and projected into the language token space, then combined with the text (i.e. the instruction) tokens, to predict compound properties and attributes. b) The LLaVA-based architecture adapted to the chemical structures and languages.

3.2 Data Processing

3.2.1 SMILES to Graph Construction

Chemical compounds in the dataset were initially represented as SMILES (Simplified Molecular Input Line Entry System) strings. To model the molecular structures, each SMILES string was converted into a graph using RDKit (Landrum et al., 2013), with:

- **Nodes:** Each atom in the molecule is treated as a node. Features associated with the atoms, such as atomic number and chirality, are extracted and encoded as node attributes.
- **Edges:** Bonds between atoms are represented as edges. Each bond is characterized by attributes such as bond type (e.g., single, double, triple) and bond direction, which are encoded as edge attributes.

3.3 Training

The training process for the model consists of two stages: (i) pretraining for alignment between molecular graph data and textual information, and (ii) fine-tuning to adapt the model for compound's question answering.

3.3.1 Stage 1: Alignment Pretraining

The molecular graphs are represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes atoms (nodes) and $\mathcal{E}$ represents bonds (edges). The graph is encoded into a molecular embedding $\mathbf{Z}_v$ using a MoleculeSTM (Liu et al., 2023b) graph encoder ϕ_g:

$$\mathbf{Z}_v = \phi_g(\mathcal{G})$$

where $\mathbf{Z}_v \in \mathbb{R}^{d \times T_1}$ represents the graph embedding of the molecular tokens where d is the dimension size and T_1 is the number of tokens.

Simultaneously, the instruction/question $\mathbf{X}_q$, given as natural language input, is tokenized and embedded using the LLaMA-3, ϕ_t:

$$\mathbf{H}_q = \phi_t(\mathbf{X}_q; \theta_\mathbf{t})$$

where $\mathbf{H}_q \in \mathbb{R}^{k \times T_2}$ represents the token embeddings of the instruction, with k being the embedding dimension and T_2 the number of tokens.

To align the two modalities, the molecular graph embedding $\mathbf{Z}_v$ is projected into the same space as the token embeddings via a learnable projection matrix W:

$$\mathbf{H}_v = W\mathbf{Z}_v$$

where $\mathbf{H}_v \in \mathbb{R}^{k \times T_1}$ is the projected graph embedding. This embedding is concatenated with the instruction token embeddings to form a joint embedding:

$$\mathbf{H} = [\mathbf{H}_v; \mathbf{H}_q]$$

where $\mathbf{H} \in \mathbb{R}^{k \times (T_1 + T_2)}$ is the combined input that is fed to the decoder, which is kept frozen at this stage.

3.3.2 Stage 2: Fine-Tuning Stage

In the second stage, we fine-tune the model to develop a general-purpose molecule question-answering (QA) model. The combined embedding matrix $\mathbf{H}$, consisting of the concatenated molecular embedding $\mathbf{H}_v$ and the instruction/question token embeddings $\mathbf{H}_q$, is passed through the LLaMA-3 decoder:

$$\mathbf{X}_a = f_\phi(\mathbf{H}; \theta)$$

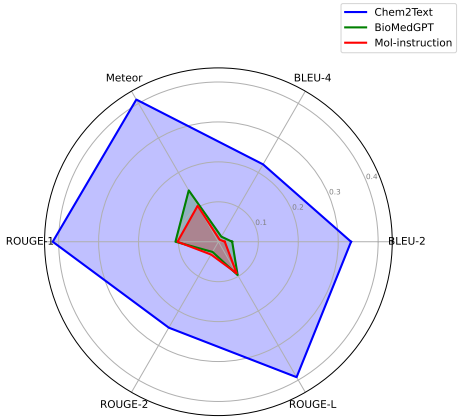

Figure 2: Evaluation of baselines in the question and answering task on our Chem2Text-QA dataset.

where $\mathbf{H}_v$ represents the generated response, and θ are the trainable parameters of the LLaMA-3 decoder. The combined input $\mathbf{H}$ allows the model to jointly process both the molecular and textual information for answering a broad set of molecular-related questions such as describing molecular properties, analyzing molecular interactions, or predicting molecular behaviors. The fine-tuning process minimizes a general-purpose loss function $\mathcal{L}_{\text{QA}}$, which is the negative log-likelihood of generating the correct token sequence for the response $\mathbf{X}_a$:

$$\mathcal{L}_{\text{QA}} = -\sum_{i=1}^{L} \log p(r_i|\mathbf{H}, r_{<i}; \theta)$$

where $p(r_i|\mathbf{H}, r_{<i}; \theta)$ is the probability of generating the next token r_i in the response sequence $\mathbf{X}_a$, given the combined input $\mathbf{H}$, previous tokens $r_{<i}$, and the model parameters θ. The length of the response is denoted by L.

Unlike the pretraining stage, in which the LLaMA-3 decoder was frozen, in this fine-tuning stage, the weights of both the projection layer and LLaMA-3 decoder θ were updated. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is employed during fine-tuning to update the weights efficiently.

4 Results

4.1 Benchmarking on ChEBI-20 test Dataset

We benchmarked **Chem2Text** against several baseline models, including those specifically designed for molecular captioning and instruction-tuned language models (Table 1), on the ChEBI-20 dataset—an established benchmark for molecule captioning tasks, where models generate descriptions for input compound structure. Chem2Text outperforms other models in the *instruction-tuning-based LLM* category, such as **GPT-3.5-turbo** and **BioMedGPT-10B**, across evaluation metrics (Table 1).

Molecule captioning models like **MoT5-base**, **MolFM**, and **MolXPT** are only optimized for one specific task. Therefore, our model (trained with only 5 epochs) does not yet match their performance in this task.

However, these molecule captioning models fail to generalize and lack the versatility to answer broader, open-ended queries, such as protein interactions or disease relevance (Table 2). In these areas, Chem2Text outperforms. Designed as a *general-purpose molecule question-answering model*, it handles a wider range of queries, going beyond simple molecule caption generation as illustrated below.

4.2 Benchmarking on Chem2Text-QA Dataset

Chem2Text-QA (section 3.1) was created to evaluate models' ability to handle general-purpose molecule question-answering tasks, where both the SMILES representation of a chemical compound and a corresponding instruction or question are provided as inputs. Chem2Text was benchmarked against baseline models **Mol-instructions** and **BioMedGPT** that are capable answering open-ended question in Chem2Text-QA dataset. Unfortunately, specialized models like **MolT5** and other molecule captioning models could not be evaluated, as they only accept SMILES inputs and are not equipped to process additional instructions. **InstructMol** was also excluded from evaluation because its its model weights and training data were unavailable at the time. **Chem2Text** demonstrated significantly better performance than the other tested models across various metrics, highlighting its ability to handle general-purpose molecule question-answering tasks (2). Table 2 demonstrate performance on randomly selected QA examples.

5 Conclusion

In this work, we presented **Chem2Text**, a general-purpose molecule question-answering model that integrates SMILES-based chemical representations with natural language processing (NLP). Chem2Text was designed to overcome the limitations of existing models that specialize in either

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-base (Edwards et al., 2022)	0.540	0.457	0.634	0.485	0.568	0.569
MolFM (MoT5-base) (Luo et al., 2023a)	0.585	0.498	0.653	0.508	0.594	0.607
MolXPT (Liu et al., 2023c)	0.594	0.505	0.660	0.511	0.597	0.626
GIT-Mol-graph (Liu et al., 2024)	0.290	0.210	0.540	0.445	0.512	0.491
GIT-Mol-SMILES (Liu et al., 2024)	0.264	0.176	0.477	0.374	0.451	0.430
GIT-Mol (graph+SMILES) (Liu et al., 2024)	0.352	0.263	0.575	0.485	0.560	0.430
Text+Chem T5-augm-base (Christofidellis et al., 2023)	0.625	0.542	0.682	0.543	0.622	0.648
instruction-tuning-based LLM						
GPT-3.5-turbo (zero-shot) (Brown et al., 2020)	0.103	0.050	0.261	0.088	0.204	0.161
BioMedGPT-10B (Luo et al., 2023b)	0.234	0.141	0.386	0.206	0.332	0.308
Mol-Instructions (Fang et al., 2024)	0.249	0.171	0.331	0.203	0.289	0.271
Chem2Text-M (Ours)	0.349	0.235	0.468	0.274	0.402	0.398
Chem2Text-F (Ours)	0.0	0.0	0.0	0.0	0.0	0.0

Table 1: Results of molecular description generation task on the test split of ChEBI-20.

Question	Compound	Ground Truth	Model Outputs
What type of effect does the compound have on blood vessels?	carboprost	It has a vasoconstriction effect.	**Mol-instruction:** C[=C][Branch2][Ring1][Branch2]... **BioMedGPT:** The molecule has a role as a platelet aggregation inhibitor. It is a prostaglandin E2, a monoepoxy fatty acid and a prostaglandin. It is a conjugate acid of a prostaglandinE2(1-). **Chem2text:** It causes vasoconstriction.
Does the enzyme that catalyzes the conversion of the compound use NADPH for the reaction?	Dihomomethionine	Yes, it is NADPH-dependent.	**Mol-instruction:** [I][C][=C][C][=C][Branch1][=Branch2][N][C][C][N][... **BioMedGPT:** 100% (molecule)major species at pH 7.3 is the enantiomer with R-configuration at the chiral centre. It is a tautomer of a (2R)-2-amino-3-hydroxy-3-methylbutanoate. It is a tautomer of a (2R)-2-amino-3-hydroxy-3-methylbutanoate. It is a tautomer of a (2R)-2-amino-3-hydroxy-3-methylbutanoate. It is a tautomer of a (2R)-2-amino-3-hydroxy-3-methylbutano. **Chem2text:** Yes, the enzyme uses NADPH as a cofactor.
What type of chemical is the compound?	L-threonine	The compound is an amino acid.	**Mol-instruction:** The compound is a member of the class of morpholines that is morpholine substituted by a carboxy group at position 3. It is functionally related to a morpholine. **BioMedGPT:** The molecule is a 3-hydroxy-3-methylglutaryl-CoA. It is a tautomer of a 3-hydroxy-3-methylglutarate. **Chem2text:** The compound is an amino acid.
What happens to autophagic functions when the compound is present?	palmitoleoyl-CoA(4-)	Autophagic functions are reduced.	**Mol-instruction:** Autophagic functions of [C@@H1][C@@H1][Branch1][Ring2][O]... **BioMedGPT:** The autophagic functions of the compound are not known. A vaccine for use in humans is not available. **Chem2text:** Autophagic functions are impaired.

Table 2: Comparison of outputs for various models on different molecular questions. When questions are asked, Mol-instructions and BioMedGPT responds with the compound summary majority of the time. Due to design limitations, MoT5-base, MolFM, and MolXPT output SMILES strings as answers to queries instead of descriptive text, which reflects their optimization for specific task rather than general-purpose question answering.

molecular description generation or specific downstream tasks, by offering a versatile system capable of answering open-ended questions about chemical compounds.

Our evaluation on two datasets, *ChEBI-20* and our newly constructed *Chem2Text-QA* dataset, demonstrates the utility of Chem2Text in generating accurate descriptions and answering complex queries related to chemical compounds. While specialized models like **MolT5**, **MolFM**, and **MolXPT** excelled at molecule captioning tasks, Chem2Text proved to be more versatile, outperforming other instruction-tuning-based LLM models like **BioMedGPT** and **Mol-Instructions** across various evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Ganesan, 2018), and METEOR (Banerjee and Lavie, 2005). This showcases its strength in both generating precise molecule descriptions and addressing broader scientific questions.

6 Limitations and Ethics

6.1 Data Use and Privacy

The Chem2Text-QA dataset used in this project was derived from publicly available articles in the PubMed Central (PMC) database. All data utilized in this research adheres to the usage guidelines set forth by the source databases, and no private or proprietary information was accessed or used.

6.2 Environmental Impact

Training large-scale AI models can be resource-intensive and have an environmental impact due to the energy required for computation. We have made efforts to minimize the environmental footprint of this research by using efficient training practices, including the use of pre-trained models and leveraging low-rank adaptation (LoRA) to reduce the need for extensive full-model fine-tuning. We encourage the wider research community to remain mindful of the environmental cost associated with large-scale model training.

6.3 Limitations

Chem2Text, while promising, has several limitations. One key issue is hallucination, especially when the model is asked questions unrelated to the compound or context. To address this, future research will focus on generating more diverse, balanced, and well-structured training data to minimize output bias, improve task understanding, and enhance overall performance. We are also exploring methods to constrain outputs, using filtering tools and probabilistic thresholds to limit responses to within relevant boundaries, reducing the risk of incorrect or speculative answers.

Another limitation is the model's inability to handle follow-up questions, which hinders its capacity to provide more specific answers. Integrating this capability is a priority for future iterations. Additionally, Chem2Text lacks interpretability—a critical drawback in biomedical applications. We are considering the use of the Chain of Thought technique to enable stepwise reasoning, which could provide more transparent insights into how the model arrives at its conclusions.

Chem2Text is also unable to generate chemical structures, meaning it cannot directly be used for compound design with specific properties. Enabling the model to output SMILES would significantly expand its utility. While our test set was carefully excluded from training, we recognize that, as with any model trained on large datasets, data leakage cannot be entirely ruled out (Kaufman et al., 2012). This is an inherent challenge in evaluating models trained on vast and diverse data.

Furthermore, Chem2Text inherits some limitations from its underlying pretrained language decoder and chemical encoders. Fine-tuning alone is unlikely to fully address these endemic issues, which may affect performance. Lastly, while Chem2Text is designed to assist researchers in understanding chemical compounds and generating hypotheses, it should not be seen as a definitive source of truth. Its outputs are constrained by the quality and breadth of its training data, and it may not always provide accurate or scientifically validated answers, particularly for novel or highly specific queries. Users must critically evaluate the model's outputs and cross-reference them with verified scientific literature. We present these limitations not as flaws, but as areas for future improvement, ensuring that Chem2Text can evolve into a more reliable and comprehensive tool for chemical research.

6.4 Future Directions and Ethical Considerations

As AI continues to play a larger role in scientific discovery, it is essential to consider the ethical implications of integrating such models into research workflows. We encourage future researchers to remain vigilant about potential biases in training data, ensure proper validation of model outputs, and use AI tools responsibly within their respective domains.

Overall, we believe that Chem2Text can positively contribute to scientific research, but it must be used as a complement to expert analysis and with an understanding of its limitations.

References

M.J. Buehler A. Ghafarollahi. 2024. Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, DOI: 10.1039/D4DD00013G.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv*, (arXiv:2209.01712). ArXiv:2209.01712 [cs, q-bio].

Jean-Baptiste Alayrac, Adria Recasens, Romana Schneider, and et al. 2022. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv: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. *Preprint*, arXiv:2311.16208.

Xi Chen, An Wang, Jiannan Tao, and et al. 2022. Pali: A jointly-scaled multilingual language-image model for vision-language tasks. *arXiv preprint arXiv:2209.06794*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. *Preprint*, arXiv:2301.12586.

Europe PMC Consortium. 2015. Europe pmc: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43(D1):D1042–D1048.

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

Nan Du, Barret Zoph, Noam Shazeer, Justin Gilmer, Yiming Wang, Marie Pellat, Chen Wang, Adam Roberts, Sharan Narang, Aakanksha Chowdhery, et al. 2021. Glam: Efficient scaling of language models with mixture-of-experts. *arXiv preprint arXiv:2112.06905*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *Preprint*, arXiv:2204.11817.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. 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.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *Preprint*, arXiv:2306.08018.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

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

Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A. Grotjahn, Elizabeth Villa, Le Song, and Pengtao Xie. 2024. Multi-modal large language model enables protein function prediction. *bioRxiv*.

Ala Jararweh, Oladimeji Macaulay, David Arredondo, Olufunmilola M Oyebamiji, Yue Hu, Luis Tafoya, Yanfu Zhang, Kushal Virupakshappa, and Avinash Sahu. 2024. LitGene: a transformer-based model that uses contrastive learning to integrate textual information into gene representations. *bioRxiv*.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Greg Landrum et al. 2013. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Haotian Liu, Chunyuan Lin, Fangyun Zeng, and et al. 2023a. Llava: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*.

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

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

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

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023a. Molfm: A multimodal molecular foundation model. *Preprint*, arXiv:2307.09484.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023b. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Adam Roberts, Noam Shazeer, Sharan Narang, et al. 2022. T5x: A modular and scalable framework for sequence-to-sequence models. https://github.com/google-research/t5x.

Victor Sanh, Albert Webson, Colin Raffel, Rowan Bashforth, Adam Roberts, Swaroop Mishra, Gokul Barik, Kabir Sardhana, Linting Xue, Noah Constas, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Rami Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Amarnag Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Theo Bos, Lloyd Baker, Yuanzhong Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2210.11416*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. 2023. Drugassist: A large language model for molecule optimization. *arXiv preprint arXiv:2401.10334*.

Rowan Zellers and et al. 2021. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Qiang Zhang, Jiazhang Lian, and Huajun Chen. 2022. Ontoprotein: Protein pretraining with gene ontology embedding. In *International Conference on Learning Representations*.

A Related Work

Instruction tuning has emerged as a powerful fine-tuning paradigm for enhancing the capabilities of LLMs and enabling them to perform a wide range of tasks. These models have excelled in a variety of tasks such as summarization (Raffel et al., 2020; Lewis et al., 2020; Zhang et al., 2019), translation (Raffel et al., 2020; Lewis et al., 2020; Wei et al., 2022), question answering (Khashabi et al., 2020; Raffel et al., 2020; Sanh et al., 2021), zero-shot capabilities (Brown et al., 2020; Ouyang et al., 2022; Wei et al., 2022; Sanh et al., 2021; Thoppilan et al., 2022; Du et al., 2021), and multilingual task generalization (Xue et al., 2021; Roberts et al., 2022). Recent efforts also succeeded in extending instruction tuning to include other modalities, where models are instructed to perform reasoning tasks across modalities. For instance, vision-language models (Liu et al., 2023a; Li et al., 2023; Chen et al., 2022; Alayrac et al., 2022; Zellers and et al., 2021) integrate text and vision modalities to perform tasks like captioning, retrieval, and visual question-answering.

Due to its versatility and impact, multimodal instruction tuning has also been applied to other scientific fields. For instance, several studies explored the potential integration of protein sequences and language modalities, enabling the prediction of complex protein properties (Zhang et al., 2022; Huo et al., 2024; A. Ghafarollahi, 2024; Lv et al., 2024). Molecule-language models have also gained more attention and have been applied to a variety of instruction datasets. These models aim to address significant challenges in different drug discovery tasks such as molecular interactions, property prediction, molecule description generation, etc. Several models were introduced that are either task-specific models trained for downstream tasks such as translation and captioning across the language and the compound structure modalities (Edwards et al., 2022; Liu et al., 2023c, 2024; Christofidellis et al., 2023), and instruction-tuned LLMs that perform wider ranges of tasks based on the instruction (Fang et al., 2024; Cao et al., 2023; Ye et al., 2023).

Hyperparameter	Pre-training	Fine-tuning
Number of Epochs	3	5
Per-device Batch Size	16	8
Total Optimization Steps	10,782	16,810
Learning Rate	2×10^{-3}	8×10^{-5}
Number of Trainable Parameters	1.2M	171M
Gradient Checkpointing	Enabled	
Max Sequence Length	2048 tokens	
Optimizer	AdamW	
Precision	bf16 (Mixed Precision)	
Learning Rate Scheduler	Cosine	
Gradient Accumulation Steps	1 step	
Warmup Ratio	0.03	

Table 3: Hyperparameters used for pretraining and fine-tuning stages in Chem2Text.

B Data Collection

B.0.1 Compound Selection

We began by compiling a list of chemical compounds that were of interest for the study. This list was sourced from ChebI Database (Degtyarenko et al., 2007), a total of 30000 compound names were identified and used for further processing.

To create our dataset, we queried the PMC database for articles explicitly mentioning target compounds in their abstracts, using the NCBI Entrez utilities to retrieve up to 10 abstracts per compound. The following process was adopted from LLaVA (Liu et al., 2023a). The abstracts were extracted, preprocessed to remove irrelevant sections, and standardized for analysis. We then utilized the Llama-3 language model to generate question-and-answer pairs based on the content of each abstract. The model was instructed to produce context-relevant, general questions about the compound's functions and attributes, ensuring that the questions were not compound-specific but focused on general characteristics. If the abstract lacked sufficient information for valid questions, no output was generated for that entry.

C Experiment Setup

We conducted the experiments for training and fine-tuning the Chem2Text model using two NVIDIA H100 GPUs. During pretraining, the model was trained for 3 epochs with a batch size of 16 per GPU, resulting in a total batch size of 32. Key hyperparameters included a learning rate of 2×10^{-3}, a warmup ratio of 0.03, and gradient accumulation steps set to 1. We employed **bf16 mixed precision** to optimize memory usage and accelerate the training process. Pretraining consisted of 10,782 total optimization steps, with model checkpoints saved every 8,000 steps.

Following pretraining, the model was fine-tuned for 10 epochs with a reduced batch size of 8 per GPU, resulting in a total batch size of 16. The learning rate was decreased to 8×10^{-5}, and the same warmup ratio of 0.03 was maintained. The total number of optimization steps during fine-tuning was 16,810, and checkpoints were saved at the end of each epoch. Table 3 enumerates the parameters used for both experiments.