MoleculeGPT: Instruction Following Large Language Models for Molecular Property Prediction

Weitong Zhang*
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
weightzero@cs.ucla.edu

Xiaoyun Wang
NVIDIA
Santa Clara, CA 95050
xiaoyunw@nvidia.com

Weili Nie
NVIDIA
Santa Clara, CA 95050
wnie@nvidia.com

Joe Eaton
NVIDIA
Santa Clara, CA 95050
featon@nvidia.com

Brad Rees
NVIDIA
Santa Clara, CA 95050
brees@nvidia.com

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

Harnessing textual information offers significant advantages in the drug design process, providing invaluable insights into complex molecular structures and facilitating molecule design based on textual instructions. With recent advancements in the utilization of Large Language Models (LLMs) for multi-modal data applications, we aim to leverage the capabilities of LLM for molecule property prediction tasks. We introduce MoleculeGPT, which is designed to provide answers to queries concerning molecular properties on the basis of molecular structure inputs. To train the MoleculeGPT, we have curated a new dataset from the raw molecule description in PubChem for instruction-following tasks. We evaluate the performance of MoleculeGPT on multiple-choice questions and several downstream tasks on molecule property prediction for drug design. Experimental results show that MoleculeGPT can generate responses that closely resemble human-level performance and demonstrate exceptional capabilities across diverse downstream tasks.

1 Introduction

Recent advances in Artificial Intelligence (AI) have significantly pushed the frontier of molecule design and drug discovery. Machine learning models, particularly deep neural networks, have found wide-ranging applications in various aspects of this field, including molecule property prediction (Thölke and Fabritiis, 2022; Schütt et al., 2017; Gasteiger et al., 2020), molecule generation (Karras et al., 2022), drug screening (Kumar and Zhang, 2018; Altalib and Salim, 2022), protein docking (Ketata et al., 2023) and structure-based drug design (Guan et al., 2023).

Most existing models primarily focus on the quantitative properties of the molecule. They utilize quantitative atom features, such as atom mass, formal charge to predict the properties of molecule, such as heat capacity, polarizability, or binding affinity. These closed-set predictions, can hardly adapt to new molecular properties when facing with new drug design task. However, considering the extensive history of development in chemistry and biology, harnessing textual information, such as

*Work done during the internship at NVIDIA

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

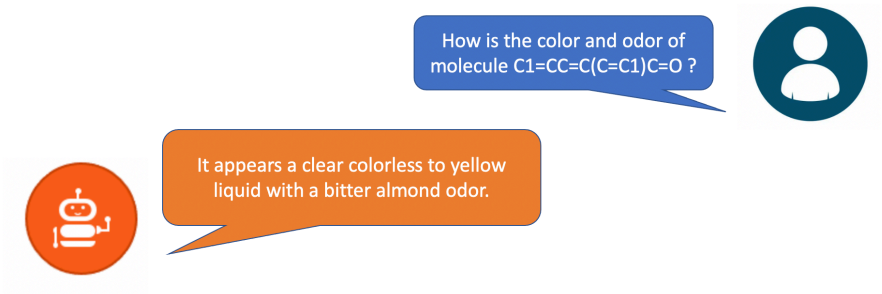

Figure 1: A language model generating instruction-following response given the molecule structure presented by `C1=CC=C(C=C1)C=O`.

annotations extracted from previous research articles, can greatly benefit the drug design process by providing valuable insights for creating novel molecules from the collective knowledge documented in prior literature. Harnessing textual information can also unlock the model's capabilities of open-set predictions that easily generalize to new tasks. To utilize this textual information, recent works mainly focus on enforcing alignment between the molecule structure and textual annotations. For example, both KV-PLM (Zeng et al., 2022) and MoleculeSTM (Liu et al., 2022) use contrastive learning to estimate the similarity score between the SMILES string and its paired textual annotations. However, for general molecule property prediction tasks, it is sometimes hard to generate a bunch of different annotations for these methods to compare the similarity. A more natural approach of leveraging the textual information is to **generate** corresponding responses given different molecule structures and instructions. An illustrative example is shown in Figure 1.

Recent advances in Large Language Models (LLMs), such as ChatGPT, GPT-4 (OpenAI, 2023), LLama (Touvron et al., 2023) and Vicuna (Chiang et al., 2023) make it possible to deliver the aforementioned output. In particular, MiniGPT-4 Zhu et al. (2023) and InstructBilp Dai et al. (2023) leverage multi-modal data and provide vivid image captions or responses to specific questions related to the image contents. Recent work (Liu et al., 2023b) leverages the capacity of LLMs to deliver molecule property analysis and text-guided molecule generation for specific drug design tasks, such as controlling toxicity. However, it would be more desirable to develop a more **general** instruction-following multi-modal language model, which can be readily fine-tuned for various downstream tasks.

Our Approach In this paper, we propose MoleculeGPT, which provides an instruction-following response given the molecule structures. Specially, MoleculeGPT contains two branches, the **2D Graph Branch** and **1D SMILES Branch**, to digest both 2D molecular structure and 1D SMILES representation. The multi-modal representations are then processed by two Q-Formers respectively to fuse together as a soft prompt for LLM. To train the MoleculeGPT, we curate a new dataset that uses LLMs (vicuna-13b in our experiment) to process the raw annotations from PubChem (Kim et al., 2021) and obtain an effective instruction-following corpus. The generation quality can be improved using these generated corpora instead of using the original raw data annotations. To evaluate the generation quality, we propose multiple-choice question tasks similar to the CHEMIChoice (Zeng et al., 2022) for zero-shot comparison. Furthermore, we also showcase MoleculeGPT's good performance in downstream tasks related to molecule property prediction.

Concurrent to our work, ChatDrug Liang et al. (2023) also leverages the LLMs for understanding the molecule properties. They only use 2D graph representation and do not use the 1D SMILES string for the molecule representation. In contrast, we leverage the molecule structures from both 1D SMILES string and 2D graph. Using SMILES information, as discussed in (Zeng et al., 2022; Liu et al., 2022) can usually provide better performance than only using the graph structure. Additionally, we present a variety of numerical evaluations alongside the generative examples, offering a more comprehensive assessment of the model's performance.

2 Related Work

2.1 Multi-Modal Large Language Models

In this section, we review multi-modal LLM. Generally speaking, the multi-modal representation in LLMs can exist in both the input and the output of the model. In the first category, the multi-modal objects (e.g., images and texts) are fed into the LLMs, and the LLMs are trained to give responses given the multi-modal information. There are a series of works along this line. To mention a few, LLaVA (Liu et al., 2023a) uses formatted language to express the image information, MiniGPT4 (Zhu et al., 2023) and Instruct-BLIP (Dai et al., 2023) translate the image representation into the language domain using a Q-Former (Li et al., 2023) and then use the output as soft prompts for LLMs. In the second category, the multi-modal objects (e.g., images and texts) are generated by language models, usually by diffusion models. In particular, GILL (Koh et al., 2023) append the Stable Diffusion model to the end of LLMs to generate the desired images.

2.2 Using Text Information for Drug Discovery

The traditional method of text mining for drug discovery tasks is based on knowledge graph construction and retrieval (Roberts PM, 2008; Krallinger M, 2005). With the advancement of NLP, BERT(Devlin et al., 2019) and its variations are also used in drug discovery tasks. SMILES-BERT (Wang et al., 2019) uses SMILE strings as the input of BERT. Beltagy et al. (2019) is one of the most frequently used pre-trained language models in the biomedical domain, which is only trained on natural language data. KV-PLM, as introduced by Zeng et al. (2022), employs contrastive learning approach to estimate the similarity score between the SMILES representation of a molecular structure and the corresponding textual annotations describing that structure. MoleculeGPT (Liu et al., 2022) employs contrastive learning approach between the text annotations and the structure represented by both SMILES string and graph structure.

Following the rapid growth of language language models, several very recent works use LLMs for drug discovery. ChatDrug (Liu et al., 2023b) treats molecules' textual strings using their SMILES representation and enables researchers to modify the molecules or proteins to meet specific target properties. DrugChat (Liang et al., 2023) fuses the graph information into LLMs as soft prompt.

3 Methodology

In this section, we introduce MoleculeGPT as well as the pre-training dataset to train our MoleculeGPT.

3.1 MoleculeGPT Architecture

The architectural overview of MoleculeGPT is presented in Figure 2. MoleculeGPT consists of three components: **graph branch**, **SMILES branch** and **language model**. In a high-level view, the first two branches, which we will introduce later, will extract the structural information of the molecule and feed into the language model as soft prompts. We use `vicuna-7B-v1.5` as the language model which is fine-tuned based on Llama2 (Touvron et al., 2023). As in previous work like MiniGPT4 or InstructBlip, we freeze the LLM during the training, keeping it to generate a normal response if no structural information is given.

3.1.1 2D Graph Branch

The graph branch is designed to utilize the 2D representations of the molecule. In particular, the molecule graph is represented by a graph $\{\mathbf{v}_i\}_N, \{\mathbf{e}_{ij}\}_{N\times N}$ where N is the number of atoms in the molecule, $\mathbf{x}_i$ is the embedding of the atom type, and $\mathbf{e}_{ij}$ is the embedding of the bond type. We use GraphMVP (Liu et al., 2021) to aggregate the graph features to $\{\mathbf{y}_i\}_N$. Since GraphMVP is pre-trained by aligning the 2D molecule structures with the 3D molecule conformers, we explicitly utilize the 2D information as well as the 3D conformer information.

Given the output of the GraphMVP, we hypothesize that the structural information is fully embedded in the output features $\{\mathbf{y}_i\}_N$ thus we directly flatten it to a sequence $\mathbf{Y} \in \mathbb{R}^{N\times d}$ and ignore the graph structure. Then the sequence $\mathbf{Y}$ is fed into a Q-Former (Li et al., 2023) for translating to language

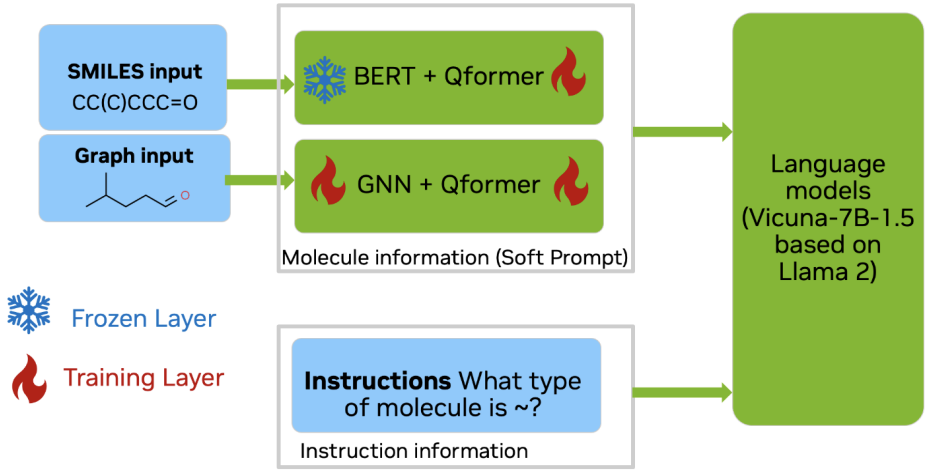

Figure 2: Overview of the network architecture, the green block indicates the network components, the blue block indicates the input or intermediate tensors.

modality. Unlike MiniGPT4, we need to train the Q-Former from scratch since the input is from the domain of molecule representations instead of vision representations.

3.1.2 1D SMILES Branch

The SMILES branch is designed to utilize the 1D SMILES string of the molecule. To do so, we adapt the ChemBerta-2 (Ahmad et al., 2022) pre-trained on multi-task regression using 77M compounds from PubChem. We also take the sequential output from ChemBerta and then fed it to Q-Former. The procedure is similar to the aforementioned graph branch. However, since the domain of SMILES representation and the graph representation are different, we use a different Q-Former instead of the one used after GraphMVP.

3.1.3 Merging into LLMs

The output sequences from two Q-Formers in graph branch and SMILES branch are then concatenated together serving as the soft prompt of the large language models. Compared with Liang et al. (2023), MoleculeGPT utilizes SMILES representation of the molecule, which usually provides better information. In addition, MoleculeGPT treats the graph output as a sequence and then processes with Q-Former, instead of taking the graph output as a single vector and using a linear layer to get the soft prompt. Therefore, it can provide a more comprehensive representation of the graph structure.

3.2 Pretraining Dataset

We collect the annotations and descriptions from PubChem (Kim et al., 2021) database. PubChem database contains about 112M molecules. Following Liu et al. (2022), we collect the annotations for more than 300k molecules with annotations. A sample of the collected annotations is

> *Benzaldehyde is an arenecarbaldehyde that consists of benzene bearing a single formyl substituent; the simplest aromatic aldehyde and parent of the class of benzaldehydes. It has a role as a flavouring agent, a fragrance, an odorant receptor agonist, a plant metabolite, an EC 3.5.5.1 (nitrilase) inhibitor and an EC 3.1.1.3 (triacylglycerol lipase) inhibitor. Benzaldehyde is a natural product found in Nepeta nepetella, Xylopia aromatica, and other organisms with data available.*

Several issues exist from the raw annotation collected from PubChem database that prevent it from being efficiently used to train MoleculeGPT. Therefore, we apply multiple methods to clean and refine the data. First, as suggested by Liu et al. (2022), we replace all the molecule name with

token '$\sim$' to prevent the network learning the complex molecule names. Second, it is hard for the language model to predict the origin, source, and literature related to the molecule given the molecule structure (e.g., *Benzaldehyde is a natural product found in Nepeta nepetella, Xylopia aromatica, and other organisms with data available.*). Thus, we split the annotations into sentences and remove the sentences containing geographic information, citations and other hard-to-predict part.

In addition to the aforementioned issue also mentioned in Liu et al. (2022), when using the above corpus to train the language model, a unified instruction '*Give me a description of the molecule*' would be very rough. To generate diverse responses given different instructions about the structure, usage, and toxicity of the molecule, we adapt another large language model to propose questions given the annotations. In particular, we use the `vicuna-13B-v1.3` to generate the questions, where the input prompt is

> *Propose a question regarding the molecule '$\sim$' whose answer is: $\sim$ is an alpha-CH2-containing aldehyde:*

More than 130k pairs of instruction-following on 80k molecules are generated via this procedure. Some samples of the generated instruction-following pairs are presented below; we defer more examples to Appendix A.

> *Instruction: What type of molecule is $\sim$?*
> *Response: $\sim$ is an alpha-CH2-containing aldehyde.*

The molecule structure is then preprocessed to get SMILES string and 2D graph networks for feeding into the neural networks.

3.3 Training Objective of MoleculeGPT

We train MoleculeGPT using the cross-entropy loss with the desired response, following the standard practice for fine-tuning language models. To construct the input context, we concatenated the embeddings of the input instructions and the molecule, denoted by $\mathcal{H}$. Given the output response $x_{i\,L}$ as the training data, we trained MoleculeGPT by minimizing the objective function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{L} \log \Pr_{\boldsymbol{\theta}}(x_i | \mathcal{H}, x_{1:i}).$$

Here, $\Pr_{\boldsymbol{\theta}}$ is parameterized by the autoregressive model in `vicuna-7b-v1.5`. As suggested in Liu et al. (2022), the molecule dataset (130k in our setting) is usually much smaller than the vision data set (129M in Li et al. (2023)). Therefore, we leverage the pretrained network. In particular, we freeze ChemBerta-2 for SMILES feature extraction but finetune the GraphMVP, which is common in practice (Liu et al., 2022, 2021; Wang et al., 2022). Two Q-Formers on the graph branch and SMILES branch are trained from scratch since they are serving different domains.

4 Experiments

We present the generation results and several numerical experiments in this section. We use the dataset described in Section 3.2 to pretrain the model. In particular, the training dataset is constructed by 80% molecules in the original dataset and the testing dataset contains the rest 20% molecules. We train MoleculeGPT on 8 V100 GPUs for 100 epochs with batch size 64 on the training dataset.

Samples of instruction-following responses in the test dataset are presented below. It's obvious that MoleculeGPT can well capture the structural information and get accurate answers on questions related to molecule structure, even if MoleculeGPT has never seen the structure before. Regarding the queries about the physical property, such as the color and odor, the response is less accurate. This aligns with the intuition that physical properties are usually harder to predict given the molecule structure. We defer more examples to Appendix B

> *Instruction: What is the chemical structure of the molecule represented by $\sim$?*
> *Ground Truth: $\sim$ is an amino alcohol and a secondary alcohol*
> *Response: $\sim$ is a secondary amino compound and a primary amino compound*

Instruction: What is the physical appearance and odor of the molecule ?
Ground Truth: appears as a colorless liquid with a slight ammonia-like odor
Response: appears as a clear colorless to yellow liquid with a fishy odor

4.1 Multiple-Choice Questions

To provide a numerical evaluation for the generation performance, we conduct the multiple-choice questions. Specifically, for each instance in the test dataset, which consists of a description paired with its corresponding molecular structure, we create an M-choice question. In this setup, MoleculeGPT is tasked with selecting the most precise description from the available options. To construct these options, we randomly sample $M-1$ descriptions from the test dataset. An example of the prompt and the output from MoleculeGPT is:

Human: In the following four choices, which one better describes the molecule $\sim$?
A. $\sim$ is an aldimine and a triol.
B. $\sim$ is a selective and almost irreversible inhibitor of thrombin, both free and clot-bound, by blocking its active site.
C. $\sim$ is an acyl monophosphate and a 2,3-bisphosphoglyceric acid.
D. $\sim$ is a gonadotropin releasing hormone agonist that is used to treat central precocious puberty in children and endometriosis in women
AI: $\sim$ is an aldimine and a triol

We evaluate the accuracy through comparing the normalized log-likelihood of all M choices: suppose the prompt containing the options and molecule structure is $\mathcal{H}_{\text{MCQ}}$ and the m-th choice is tokenized into sequence $\{x_{m,i}\}_{L_m}$, where L_m demotes the length of the tokens. Then the (normalized) loss for m-th choice is defined by

$$\mathcal{L}_m = \sum_{i=1}^{L_m} \log \mathbb{P}(x_i|\mathcal{H}_{\text{MCQ}}, x_{0:i})/L_m. \tag{4.1}$$

This normalization is commonly used in evaluating LLMs in general questions (Robinson and Wingate, 2023) to encourage choosing the longer options. MoleculeGPT is considered making a correct choice when the loss on the correct option is lowest among M choices.

The accuracy is presented in Table 1, where we also include the CHEMIChoice task result from Zeng et al. (2022). Our evaluation task presents greater complexity compared to CHEMIChoice, and this heightened difficulty can be attributed to two distinct factors. Firstly, due to the limitation of context `vicuna-7b-v1.5` can handle, we divide lengthy descriptions into multiple shorter segments, resulting in options that are inherently less informative compared to those in the CHEMIChoice task. Secondly, whereas CHEMIChoice allows for the selection of incorrect answers from molecules that differ significantly from the correct choice, we do not incorporate this aspect into our evaluation, thereby increasing the challenge of our task.

Method	Tasks	M	Accuracy ($\%, \uparrow$)
Random Guess	-	4	25%
KV-PLM (Zeng et al., 2022)	CHEMIChoice	4	**83.1**
Sci-BERT (Beltagy et al., 2019)	CHEMIChoice	4	81.6
BERT (Devlin et al., 2019)	CHEMIChoice	4	32.3
Human Response (Zeng et al., 2022)	CHEMIChoice	4	76.5
MoleculeGPT	Ours	2	78.3
MoleculeGPT	Ours	3	70.8
MoleculeGPT	Ours	4	61.4

Table 1: The accuracy on different choices of M and tasks, the numbers for CHEMIChoice dataset is copied from Zeng et al. (2022)

.

Several observations can be made from the results. First, MoleculeGPT significantly does better than random baseline. This suggests that MoleculeGPT can indeed leverage the molecule structure to infer the best description of the molecule. Second, there is a gap between the MoleculeGPT and Human Response and KV-PLM. We hypothesize this gap is because a more challenging task due

to the limitation of the context length limitation. We expect the performance would be significantly better if we employ larger models, like `vicuna-30b-v1.5`.

4.2 Downstream Tasks on MoleculeNet Benchmarks

We present the results of downstream tasks on MoleculeNet (Ramsundar et al., 2019) to showcase MoleculeGPT's capabilities in assisting with specific tasks related to drug design. MoleculeNet contains several benchmarks designed for testing machine learning methods of molecular properties. We test the classification tasks including Tox21, BBBP, HIV and SIDER. To mitigate LLMs to answer the (binary) classification tasks, the input prompt is designed as

BBBP dataset: *Can molecule $\sim$ bypass the Blood-Brain Barrier?*

Following the method in Section 4.1, we compare the loss for the positive response and negative response. We present the prompts for the BBBP task as follows and defer the prompts for the rest of the data set to Appendix C.

Positive: *The molecule $\sim$ can bypass the Blood-Brain Barrier*
Negative: *The molecule $\sim$ can not bypass the Blood-Brain Barrier*

We finetune MoleculeGPT for 10 epochs on 80% of the dataset. As Liu et al. (2022); Zeng et al. (2022), we report the AuROC for these tasks in Table 2. It is evident that MoleculeGPT's performance consistently outperforms other baseline methods. There are two reasons leading to this results. First, MoleculeGPT uses both SMILES string and Graph representation of the molecule, providing a multi-modal molecule features. Second, MoleculeGPT utilizes ChemBerta-2 and GraphMVP as pretrained components, which have already been trained on these tasks. The structure and pretraining of MoleculeGPT can fully **leverage** this information thus providing a better result.

To investigate the second hypothesis, which posits that pretraining MoleculeGPT can enhance the utilization of multi-modal information from its pretrained components, we conducted an ablation study. Specifically, we trained the Q-Formers from scratch for 10 epochs using the same training dataset while keeping ChemBerta-2 and GraphMVP frozen. As depicted in Figure 2, the F1 scores for all 12 tasks in the Tox21 dataset exhibited a significant decrease (on average, from more than 80% to less than 20%. This result suggests that training MoleculeGPT solely on these specific tasks, without prior pretraining, leads to severe overfitting instead of utilizing the molecule structure for effective prediction.

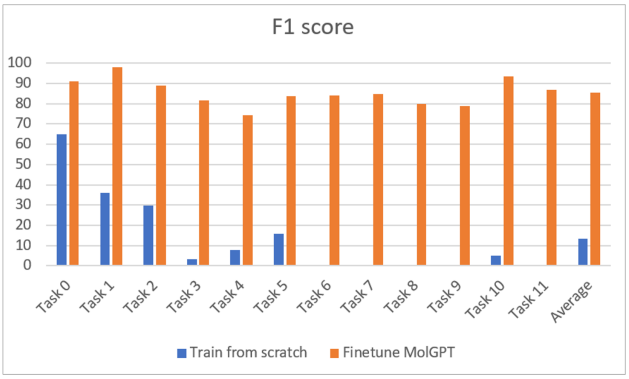

Figure 3: Comparison of training MoleculeGPT from scratch (blue, by keeping the ChemBerta-2 and GraphMVP frozen) and from pretraining methods (orange)

5 Discussion and Limitations

In this paper, we propose MoleculeGPT to generate instruction-following texts about the property of the molecule, given the structure of the molecule. We curate a new data set that takes advantage of the power of LLMs to improve data quality. Several experiments show that the trained model achieves near-human performance on multiple-choice questions and has superior performance on downstream tasks. Future work includes understanding why graph representations can potentially help improve the response quality and fine-tuning/testing our models on more challenging tasks, including the downstream tasks.

Model \ AuROC %, ↑	Tox21	BBBP	HIV	SIDER
KV-PLM (Zeng et al., 2022)	72.12	70.50	65.40	59.83
GraphMVP (Liu et al., 2021)	77.06	68.11	77.74	60.64
MoleculeSTM + SMLIES (Liu et al., 2022)	75.71	70.75	77.02	63.70
MoleculeSTM + Graph (Liu et al., 2022)	76.91	69.98	73.40	60.96
MoleculeGPT	**83.05**	**86.51**	**81.07**	**75.57**

Table 2: AuROC for MoleculeNet benchmark tasks. The numbers of benchmark algorithms are from Liu et al. (2022).

There are indeed some limitations in our model. First, the current model only **utilizes** the molecule structure, however, as discussed in Section 2.1, it would be more beneficial and challenging to **generate** the desired molecule structure given the text input, or even **editing** the molecule structure to incorporate the desired properties. Second, the current response generated by the model is straightforward, as the example shows. It would be more interesting and helpful for drug-design researchers if the model could output the **reasoning** for the output. We believe that this could be achieved using the **chain-of-thought** (Wei et al., 2023) idea and our work can lead to many follow-up works reaching this goal.

References

AHMAD, W., SIMON, E., CHITHRANANDA, S., GRAND, G. and RAMSUNDAR, B. (2022). Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* .

ALTALIB, M. K. and SALIM, N. (2022). Similarity-based virtual screen using enhanced siamese deep learning methods. *ACS Omega* **7** 4769–4786.

BELTAGY, I., LO, K. and COHAN, A. (2019). Scibert: Pretrained language model for scientific text. In *EMNLP*.

CHIANG, W.-L., LI, Z., LIN, Z., SHENG, Y., WU, Z., ZHANG, H., ZHENG, L., ZHUANG, S., ZHUANG, Y., GONZALEZ, J. E., STOICA, I. and XING, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

DAI, W., LI, J., LI, D., TIONG, A. M. H., ZHAO, J., WANG, W., LI, B., FUNG, P. and HOI, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning.

DEVLIN, J., CHANG, M.-W., LEE, K. and TOUTANOVA, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

GASTEIGER, J., GROSS, J. and GÜNNEMANN, S. (2020). Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*.

GUAN, J., ZHOU, X., YANG, Y., BAO, Y., PENG, J., MA, J., LIU, Q., WANG, L. and GU, Q. (2023). DecompDiff: Diffusion models with decomposed priors for structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*. PMLR.

KARRAS, T., AITTALA, M., AILA, T. and LAINE, S. (2022). Elucidating the design space of diffusion-based generative models.

KETATA, M. A., LAUE, C., MAMMADOV, R., STÄRK, H., WU, M., CORSO, G., MARQUET, C., BARZILAY, R. and JAAKKOLA, T. S. (2023). Diffdock-pp: Rigid protein-protein docking with diffusion models.

KIM, S., CHEN, J., CHENG, T., GINDULYTE, A., HE, J., HE, S., LI, Q., SHOEMAKER, B. A., THIESSEN, P. A., YU, B. ET AL. (2021). Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research* **49** D1388–D1395.

KOH, J. Y., FRIED, D. and SALAKHUTDINOV, R. (2023). Generating images with multimodal language models. *NeurIPS* .

KRALLINGER M, V. A., ERHARDT RA (2005). Text-mining approaches in molecular biology and biomedicine.

KUMAR, A. and ZHANG, K. Y. J. (2018). Advances in the development of shape similarity methods and their application in drug discovery. *Frontiers in Chemistry* **6**.

LI, J., LI, D., SAVARESE, S. and HOI, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* .

LIANG, Y., ZHANG, R., ZHANG, L. and XIE, P. (2023). Drugchat: Towards enabling chatgpt-like capabilities on drug molecule graphs. *TechRxiv* .

LIU, H., LI, C., WU, Q. and LEE, Y. J. (2023a). Visual instruction tuning.

LIU, S., NIE, W., WANG, C., LU, J., QIAO, Z., LIU, L., TANG, J., XIAO, C. and ANANDKUMAR, A. (2022). Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789* .

LIU, S., WANG, H., LIU, W., LASENBY, J., GUO, H. and TANG, J. (2021). Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*.

LIU, S., WANG, J., YANG, Y., WANG, C., LIU, L. and HONGYU GUO, C. X. (2023b). Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090* .

OPENAI (2023). Gpt-4 technical report.

RAMSUNDAR, B., EASTMAN, P., WALTERS, P., PANDE, V., LESWING, K. and WU, Z. (2019). *Deep Learning for the Life Sciences*. O'Reilly Media.

ROBERTS PM, H. W. (2008). Information needs and the role of text mining in drug development. Pac Symp Biocomput.

ROBINSON, J. and WINGATE, D. (2023). Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

SCHÜTT, K. T., KINDERMANS, P.-J., SAUCEDA, H. E., CHMIELA, S., TKATCHENKO, A. and MÜLLER, K.-R. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. NIPS'17, Curran Associates Inc., Red Hook, NY, USA.

THÖLKE, P. and FABRITIIS, G. D. (2022). Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*.

TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., RODRIGUEZ, A., JOULIN, A., GRAVE, E. and LAMPLE, G. (2023). Llama: Open and efficient foundation language models.

WANG, S., GUO, Y., WANG, Y., SUN, H. and HUANG, J. (2019). Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. BCB '19, Association for Computing Machinery, New York, NY, USA.

WANG, Y., WANG, J., CAO, Z. and BARATI FARIMANI, A. (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4** 279–287.

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E., LE, Q. and ZHOU, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

ZENG, Z., YAO, Y., LIU, Z. and SUN, M. (2022). A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications* **13**.

ZHU, D., CHEN, J., SHEN, X., LI, X. and ELHOSEINY, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* .

A Example of Generated Instruction-Following Data Pairs

Below are randomly picked generated instruction-following data pairs using `vicuna-13b-v1.3`.

Instruction: What is the chemical classification of the molecule represented by ∼?
Response: ∼ is an organosilicon compound.

Instruction: What are the properties of the molecule ∼?
Response: ∼ appears as a colorless volatile liquid with an agreeable odor

Instruction: What is the role of the molecule ∼ in various biological processes?
Response: ∼ has a role as a metabolite, an antineoplastic agent, an antitubercular agent and an antimalarial

Instruction: What are the properties exhibited by the molecule ∼?
Response: ∼ exhibits cytotoxic, antimalarial, and antitubercular activities.

Instruction: What type of molecule is ∼?
Response: ∼ is an aminoglycoside

Instruction: What is ∼ functional related to?
Response: ∼ is functionally related to a propionic acid and a loxoprofen

Instruction: What is the structure of the molecule represented by the symbol ∼?
Response: ∼ is a monocarboxylic acid that is propionic acid in which one of the hydrogens at position 2 is substituted by a 4-[(2-hydroxycyclopentyl)methyl]phenyl group (the 2S,1'R,2'S-stereoisomer)

B Example of Generated Response from MoleculeGPT

Below are randomly picked generated response given the instructions and the molecule structures

Molecule SMILES: `CCCCC(=O)N`
Instruction: What is ∼ functional related to?
Ground Truth: ∼ is functionally related to a valeric acid
Response: ∼ is functionally related to a valeric acid

Molecule SMILES: `CC(C)C(=O)C(=O)O`
Instruction: What is the molecular structure of the compound represented by the symbol ∼?
Ground Truth: ∼ is a 2-oxo monocarboxylic acid and a branched-chain keto acid
Response: ∼ is a 2-oxo monocarboxylic acid that is pentanoic acid [C5H9NO2-2] substituted with a methyl group at position 2

Molecule SMILES: `CC(C)C(=O)C(=O)O`
Instruction: What is the metabolic role of the molecule ∼ in living organisms?
Ground Truth: ∼ is a metabolite of isoleucine in man, animals and bacteria
Response: ∼ has a role as a metabolic intermediate in living organisms

Molecule SMILES: `CC(C)C(=O)C(=O)O`
Instruction: What is the primary role of the molecule ∼ in clinical medicine?
Ground Truth: ∼ is used as a clinical marker for maple syrup urine disease
Response: ∼ is a component of clinical medicine that is a fatty acid

Molecule SMILES: `CC(C)C(=O)C(=O)O`
Instruction: What is the chemical structure of the molecule represented by ∼?
Ground Truth: ∼ is the alpha-keto acid analogue of isoleucine
Response: ∼ is a deoxyribose that is valine 2 a 2-oxo compound

Molecule SMILES: `CC(C)C(=O)C(=O)O`

Instruction: What is the chemical formula and structure of the molecule represented by ∼?
Ground Truth: ∼ is a 2-oxo monocarboxylic acid that is valeric acid carrying oxo- and methyl substituents at C-2 and C-3, respectively
Response: ∼ is a 2-oxo monocarboxylic acid that is pentanoic acid [1177-12-2H] with a keto substituent at C-2

Molecule SMILES: `C1=CC(=C(C=C1Cl)O)Cl`
Instruction: What is the chemical structure of the molecule represented by ∼?
Ground Truth: ∼ is a dichlorophenol with the chloro substituents at positions 2 and 5
Response: ∼ is a dichlorophenol that is phenol in which the hydrogens at positions 3 and 4 are replaced by chlorines

Molecule SMILES: `C1=CC=C2C(=C1)C=CC(=O)N2`
Instruction: What is the role of the molecule ∼ in bacterial metabolism?
Ground Truth: ∼ has a role as a bacterial xenobiotic metabolite
Response: ∼ has a role as a bacterial xenobiotic metabolite

C Prompts used for MoleculeNet datasets

Below are the prompts to finetune Tox21, HIV and Sider dataset

Tox21 dataset. There are 12 separate tasks whose prompt are

1. *Does ∼ interact with the Androgen Receptor (NR-AR)?*
2. *Does ∼ interact with the Androgen Receptor Ligand Binding Domain (NR-AR-LBD)?*
3. *Does ∼ interact with the Aryl Hydrocabon Receptor (NR-AhR)?*
4. *Does ∼ interact with the Aromatase (NR-Aromatase)?*
5. *Does ∼ interact with the Estrogen Receptor (NR-ER)?*
6. *Does ∼ interact with the Estrogen Receptor Ligand Binding Domain (NR-ER-LBD)?*
7. *Does ∼ interact with the Peroxisome Proliferator-Activated Receptor Gamma (NR-PPAR-gamma)?*
8. *Does ∼ interact with the Antioxidant Response Element (SR-ARE)?*
9. *Does ∼ interact with the ATAD5 (SR-ATAD5)?*
10. *Does ∼ interact with the Heat Shock Response Element (SR-HSE)?*
11. *Does ∼ interact with the Matrix Metalloproteinase (SR-MMP)?*
12. *Does ∼ interact with the Tumor Protein p53 (SR-p53)?*

HIV dataset. The prompt for HIV dataset is

Is molecule ∼ active against HIV-1 protease?

Sider dataset. The universal prompt for all 27 tasks in Sider dataset is

Does ∼ cause any adverse effect?