SCIENCE CHINA
Information Sciences

December 2024, Vol. 67, Iss. 12, 220109:1–220109:2
https://doi.org/10.1007/s11432-024-4243-0

• LETTER •
Special Topic: Large Multimodal Models

ChemDFM-X: towards large multimodal model for chemistry

Zihan ZHAO[1†], Bo CHEN[2†], Jingpiao LI[1,2], Lu CHEN[1,2*], Liyang WEN[2], Pengyu WANG[1,2], Zichen ZHU[1], Danyang ZHANG[1], Yansi LI[1], Zhongyang DAI[2], Xin CHEN[2*] & Kai YU[1,2*]

[1]*X-LANCE Lab, Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai* 200240*, China;*
[2]*Suzhou Laboratory, Suzhou* 215123*, China*

Received 16 May 2024/Revised 17 September 2024/Accepted 23 November 2024/Published online 10 December 2024

Chemistry, as a naturally multimodal discipline, plays a crucial role in various vital fields such as pharmaceutical research and material manufacturing. Therefore, research on artificial intelligence (AI) for chemistry has garnered increasing attention. Despite the rapid development, most of the chemical AI models today mainly focus on single tasks with unimodal input [1]. However, chemical data cover a wide range of modalities spanning from text description and molecular structure to image and spectrum, and chemical tasks take various forms ranging from property prediction to retrosynthesis. Although these unimodal specialist models can achieve state-of-the-art performance in the targeted tasks, they inherently cannot handle tasks even slightly different from those which they are trained on or their respective tasks when there is a slight alteration in the input modality. Therefore, the practical utility and assistance of these models in research and manufacturing are limited.

Nowadays, large language models (LLMs) and large multimodal models (LMMs) [2–4] have achieved impressive performance in a number of challenging fields, such as natural image inference [2], document analysis [3], and medical image reasoning [4]. Therefore, LMMs show great potential for building a cross-modal chemical general intelligence (CGI) system that can handle multiple tasks including inputs of multiple modalities simultaneously. However, most of the previous LMMs only focus on one non-text modality. Given the diversity of chemical modalities and the frequent co-occurrence of different modalities in practice, a single LMM that can handle multiple non-text modalities is needed for chemical LMMs to truly meet the requirements of chemists.

In this work, we detail our progress toward such a cross-modal chemical LLM and propose ChemDFM-X, a cross-modal dialogue foundation model for chemistry that can comprehend and interpret data of various chemical modalities and fulfill many downstream tasks with the same set of model weights. As shown in Figure 1, ChemDFM-X takes advantage of the pre-trained parameters of the ChemDFM [5] model and is continuously trained on various multi-modality data. Specifically, besides text and SMILES[1)] modalities which are already learned by ChemDFM, we choose five typical modalities that are both representative and meaningful in the field of chemistry.

One of the key challenges to achieving this goal is the absence of sufficient modality-aligned data. Large-scale experimental data are difficult to acquire, especially for the characterization modalities such as tandem mass spectra (MS2) and infrared spectra (IR), owing to the excessive expenditure of both experiments and quantum chemical calculations. We propose to leverage simplified approximate calculations and model predictions to obtain sub-optimal yet proximal results. In this way, we finally generate a multimodal instruction-tuning dataset containing 7.6M cross-modality data from 1.3M seed SMILES.

Benefiting from the instruction-tuning dataset, ChemDFM-X possesses the capabilities to comprehend and infer over various modalities including molecular graphs, conformations, images, and spectra. To the best of our knowledge, ChemDFM-X is the first demonstration of a cross-modality CGI system that can interpret chemical data of multiple modalities with the same sets of parameters while handling a wide variety of tasks. To demonstrate the prowess of ChemDFM-X, we conduct extensive experiments regarding the newly added modalities. The results demonstrate the strong capacity of ChemDFM-X for comprehending multi-modality input and exploiting inter-modality knowledge. Compared to LLMs and LMMs which only enable one or none of the chemical modalities, ChemDFM-X manages to handle most common modalities well and agilely leverages the knowledge of chemical materials and reactions learned from all the modalities to solve various practical chemical tasks with superior performances.

Modality selection. In the field of chemistry, there are primarily two kinds of modalities: structural modalities and characterization modalities.

Structural modalities mainly directly represent the connections and/or spatial arrangement of molecules and are

* Corresponding author (email: chenlusz@sjtu.edu.cn, mail.xinchen@gmail.com, kai.yu@sjtu.edu.cn)
† Zhao Z H and Chen B have the same contribution to this work.
1) Short for simplified molecular-input line-entry system, a linear representation of chemical molecule structures.

© Science China Press 2024

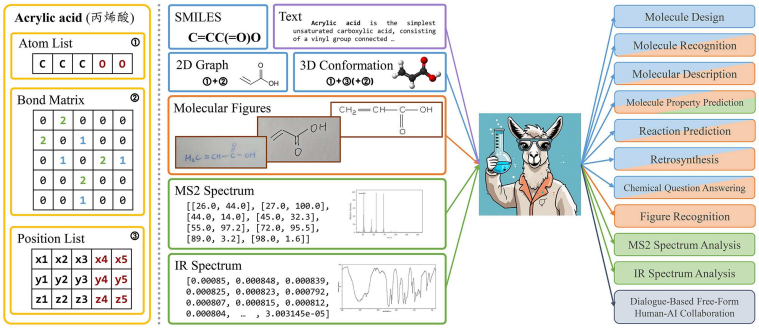

Figure 1 (Color online) Overview of ChemDFM-X. Different modalities involved in chemical tasks are distinguished by colors. Structural modalities are marked in blue, images orange, and spectra green. The text modality is marked in purple in the input and omitted in the output. The dialogue-based free-form human-AI collaboration may involve any feasible modalities and is marked in gray.

usually used for reaction inference or theoretical calculation. Among the structural modalities, two modalities are introduced to ChemDFM-X: two-dimensional molecular graphs and three-dimensional molecular conformations.

Characterization modalities mainly imply the partial properties and substructure information of molecules. As the characterization results of molecules, their data are usually sequences of data points with information implicitly hidden among them. One typical usage of characterization modal data is to identify unknown substances. Due to the high expense of chemical experiments, the amount of real experimental data of these modalities is very limited, which significantly hinders the development of AI models. In this work, we manage to construct a great amount of data of MS2 and IR, two of the most widely used characterization methods, through approximate calculation and model prediction. ChemDFM-X is then trained on these two characterization modalities.

We also introduce the image modality, including molecular images and reaction images, to ChemDFM-X, as images are the most convenient data used by human researchers. Please refer to Figure 1 for examples of these modalities.

Structure of ChemDFM-X. Generally speaking, our ChemDFM-X incorporates the typical "LLM decoder + modality encoder" framework widely used by current LMMs [2, 3]. Considering the significant differences in data formats among the different modalities, we incorporate separate modality encoders and corresponding projection modules for each modality. We use one of the advanced chemical LLMs, ChemDFM, as the "LLM decoder" to leverage its promising comprehension capabilities of natural and chemical languages. We freeze the parameters of ChemDFM and construct instruction-tuning data for each modality to train its modality encoder and corresponding projection module. For the modality where the experimental data are expensive to acquire, we take advantage of either approximate calculation or task-specific model prediction to reduce the expense and obtain sufficient data. In total, 7.6M instruction-tuning data is constructed. More details of the training of ChemDFM-X are introduced in Appendix A.

Through this "separate encoders + unified decoder" design, the separate encoders enable ChemDFM-X to obtain knowledge and information from different modalities, while the LLM decoder provides the capabilities to aggregate and analyze information from different modalities within a simple generative framework.

Experimental results. Introductions of evaluation tasks, detailed evaluation results, and more analyses are shown in Appendix B. Overall, ChemDFM-X outperforms chemical LLMs that can only accept SMILES and natural language as input, as well as LMMs for the corresponding modalities that are applicable. This indicates that ChemDFM-X possesses strong prowess in handling multimodal and inter-model contents to tackle chemical problems.

Conclusion. We proposed ChemDFM-X, a large multimodal model for chemistry. ChemDFM-X is a generalist model that has the ability to understand five of the most commonly used modalities in the field of chemistry, including structural modalities, image modalities, and characterization modalities. The evaluation results show that ChemDFM-X possesses the capabilities to comprehend chemical data in all five non-text modalities. This assists ChemDFM-X in outperforming other generalist models in a series of common chemical tasks and demonstrates the practical value of ChemDFM-X in chemistry research. It also has the potential for dealing with inputs of multiple modalities simultaneously, which is powerful in reaction-related tasks and will be further studied in our future work.

Acknowledgements This work was supported by National Science and Technology Major Project (Grant No. 2023ZD0120703), National Natural Science Foundation of China (Grant Nos. U23B2057, 62106142, 62120106006), and Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Xia J, Zhu Y, Du Y, et al. A systematic survey of chemical pre-trained models. In: Proceedings of International Joint Conference on Artificial Intelligence, 2023
2 Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2024. 26296–26306
3 Hu A, Shi Y, Xu H, et al. mPLUG-PaperOwl: scientific diagram analysis with the multimodal large language model. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 6929–6938
4 Yang L, Xu S, Sellergren A, et al. Advancing multimodal medical capabilities of Gemini. 2024. ArXiv:2405.03162
5 Zhao Z, Ma D, Chen L, et al. ChemDFM: dialogue foundation model for Chemistry. 2024. ArXiv:2401.14818