Two Steps from Hell: Compositionality on Chemical LMs

Veronika Ganeeva[1*], **Kuzma Khrabrov**[1*], **Artur Kadurin**[1,3] and **Elena Tutubalina**[1,2,3]

[1]AIRI [2]Sber AI [3]ISP RAS Research Center for Trusted Artificial Intelligence

Correspondence: {khrabrov, tutubalina}@airi.net

Abstract

This paper investigates compositionality in chemical language models (ChemLLMs). We introduce STEP, a benchmark with compositional questions that reflect intricate chemical structures and reactions, to evaluate models' understanding of chemical language. Our approach focuses on identifying and analyzing compositional patterns within chemical data, allowing us to evaluate how well existing LLMs can handle complex queries. Experiments with state-of-the-art ChemLLMs show significant performance drops in compositional tasks, highlighting the need for models that move beyond pattern recognition. By creating and sharing this benchmark, we aim to enhance the development of more capable chemical LLMs and provide a resource for future research on compositionality in chemical understanding.

1 Introduction

Recent advances in large language models (LLMs) have significantly accelerated progress in computational chemistry, with applications ranging from molecular property prediction to reaction design and drug discovery (Schwaller et al., 2018, 2021; Livne et al., 2024; Kuznetsov et al., 2025). Domain-specific adaptations such as ChemLLMs built on architectures like T5 (Raffel et al., 2020) and LLaMA (Touvron et al., 2023) have enabled models to interface with molecular representations (e.g., SMILES (Weininger, 1988)) and operate effectively on datasets such as ZINC-15 (Sterling and Irwin, 2015b), PubChem (Kim et al., 2016), and USPTO-50KK (Lowe, 2012).

Despite these successes, a fundamental question remains underexplored: *Can chemical language models work with compositionally?* That is, can they combine known chemical concepts to solve novel, multi-step problems? This capability is crucial for tasks that require generalization beyond memorized patterns such as predicting the activity of the product of a previously unseen reaction or estimating the activity of a compound generated via hypothetical synthesis. Current benchmarks such as USPTO-50K (Lowe, 2012), and CHEBI-20 (Edwards et al., 2021) in chemical NLP largely focus on single-step tasks, e.g., predicting products from reactions, generating molecular descriptions, or estimating individual properties. While these tasks provide valuable evaluation signals, they do not capture the multi-faceted, compositional nature of real-world chemical reasoning. Consequently, it is unclear whether ChemLLMs truly understand chemical concepts or simply exploit surface-level correlations (Ganeeva et al., 2024a,b).

To address this gap, we introduce STEP (Structured Tasks for Evaluating and Promoting compositionality), a benchmark and framework designed to systematically evaluate compositional reasoning in chemical LLMs. STEP transforms standard datasets into two-step tasks that require chaining atomic reasoning steps, for example, predicting a reaction product and then describing its activity. By evaluating state-of-the-art ChemLLMs across these tasks, we identify substantial performance drops in compositional settings, revealing critical limitations in their generalization abilities.

Our contributions are as follows. We propose STEP, a benchmark that evaluates compositionality in ChemLLMs via structured tasks. We curate a dataset spanning several chemical subdomains, including synthesis, property prediction, and molecular description, and transform them into compositional queries. We showed that most models performed well on isolated tasks but struggled with compositional generalization, especially on out-of-distribution inputs.

By addressing the critical need for rigorous evaluation, our work advances the understanding com-

[*]These authors contributed equally to this work. The order of author names was randomly determined.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1042–1049
November 4-9, 2025

positionality by ChemLLMs, with implications for drug discovery, materials science, and beyond.

2 Methodology

Existing chemical language models (ChemLMs) are primarily evaluated on narrow, single-domain tasks (e.g., reaction prediction or property estimation). However, real-world chemical reasoning often requires integrating knowledge across domains—for example, predicting the environmental impact of a reaction product demands understanding both reaction mechanisms and physicochemical properties. To bridge this gap, we propose **STEPS (Structured Tasks for Evaluating and Promoting Compositionality)**, a framework that trains and evaluates ChemLMs on *compositional tasks* where complex questions are solved by combining simpler subtasks. Below, we formalize STEPS and its design principles.

2.1 Framework Design

STEPS operationalizes the principle of compositionality—the idea that complex expressions derive meaning from their constituent parts and the rules combining them. The framework decomposes chemical problems into two foundational task types (Figure 1): **Atomic tasks**, which are single-domain problems with deterministic answers (e.g., predicting a reaction product) and **Composite tasks**, which are multi-domain problems requiring sequential or parallel reasoning over atomic tasks (e.g., predicting a reaction product *and* its activity).

By systematically combining atomic tasks into novel composites, STEPS evaluates whether models can generalize to unseen combinations of domains, a key marker of compositional reasoning.

2.2 Task Taxonomy

Atomic Tasks ($\mathcal{T}_A$) Atomic tasks have unique, verifiable answers. Let $\mathcal{T}_A = \{T_1, T_2, \ldots, T_n\}$ denote tasks such as: **Reaction prediction**: Given reactants and conditions, output the product's SMILES notation (a string-based molecular representation); **Molecular captioning**: Generate a textual description of a molecule's properties; **InChI naming**: Produce the IUPAC-compliant InChI identifier for a molecule; **Heavy atom counting**: Predict the number of non-hydrogen atoms in a compound; **Activity prediction**: Determine biological activity metrics (e.g., IC50 values).

Composite Tasks ($\mathcal{T}_C$) We construct *compositional tasks* by chaining atomic tasks with explicit dependencies between steps. **One-step tasks**: Single atomic tasks (e.g., "Describe molecule M_1"). **Two-step tasks**: Chain of Reaction prediction task and another atomic task (e.g., "Predict the activity of reaction R product").

2.3 Evaluation Metrics

For each atomic task, we employ task-specific metrics to ensure rigorous assessment: **Reaction prediction**: Exact match of predicted SMILES strings (1 if correct, 0 otherwise). This ensures precise evaluation of molecular structure generation. **Molecular captioning**: ROUGE-L scores (F1) to measure lexical overlap between generated descriptions and ground truth. **InChI naming**: Exact match accuracy for InChI identifiers, as these follow strict IUPAC formatting rules. **Heavy atom counting**: Accuracy calculated as the percentage of exact matches for the number of non-hydrogen atoms. **Activity prediction**: Accuracy for the activity classification task (e.g., active/inactive). For compositional task we provide the metrics for the last step to compare.

3 Models and Datasets

Dataset Compilation To create a benchmark for compositional reasoning, we compile and modify existing chemical datasets: CHEBI-20, MoleculeACE, USPTO-50k, and PubChem (see Tab. 3). We randomly select 10,000 reactions, 3,000 description-reaction pairs and 10,000 reaction-properties pairs sharing the same molecules. By combining these datasets, we generate composite examples where the input consists of compositional questions (see Fig. 1).

Evaluated models We evaluated SoTA generative LLMs that were finetuned for chemical tasks and compile reaction prediction, molecule captioning or scietific QA tasks. This includes Text2Chem-standard and its augmented variant Text2Chem-augm, both based on a T5 architecture and introduced in (Christofidellis et al., 2023). Several models in the LLaMA family were also evaluated, including Chemical-LLaMA, LLaMA-3.2-1B-IT-Chemistry-Assistant, LLaMA-Finetuned-Chemistry, LLaMA-3.1-8B-Instruct-SFT-Chem, LLaMA-7B-Instruct-Base-Chem, and LLaMA-3.2-3B-IT-Chemistry, all based on the foundational architecture introduced in (Touvron et al., 2023).

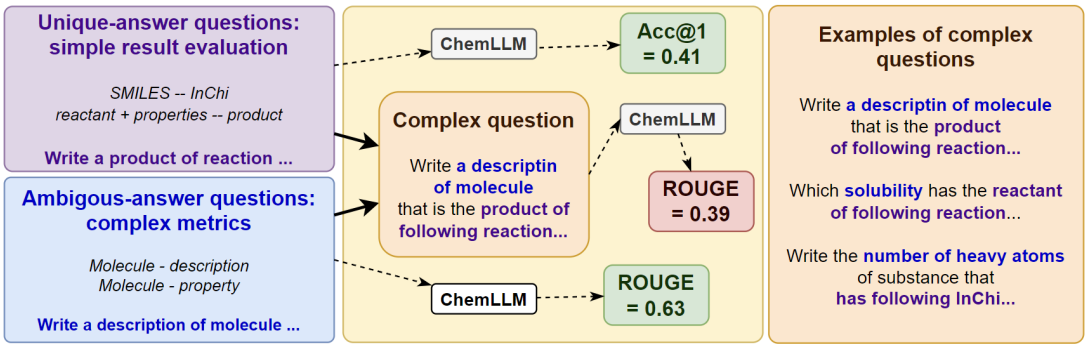

Figure 1: Our benchmark provides compositional questions for one-to-one, one-to-many and many-to-many tasks. By combining one-step tasks we create 2-steps compositional questions which enrich the data for model evaluation and training.

Among these, we further fine-tuned LLaMA-3.2-3B-IT-Chemistry on our STEP benchmark to obtain LLaMA-3.2-3B-IT-Chemistry-2Step, a compositional reasoning–oriented variant introduced in this work. In addition to LLaMA variants, we included ChemLLM-7B-Chat (Zhang et al., 2024), a dialogue-oriented chemistry model; ChemQwen2-vL (Bai et al., 2023), a large-scale model from the Qwen series; and ChemDFM-v1.5-8B (Zhao et al., 2024), a domain-specific foundation model for drug discovery and molecular property prediction. Full list of models is provided in Tab. 2. We evaluate these models to understand their compositional reasoning and identify areas for improvement in the development of ChemLLMs.

4 Evaluation results

This section presents the results of our evaluation of various ChemLLMs across a range of compositional reasoning tasks. The findings are summarized in Table 1, which reports performance metrics on one-step and compositional tasks commbine of reaction prediction, molecular description, property estimation, and multi-step generation.

Two-Step Generation and Compositionality Tasks requiring two-step generation, such as predicting the product of a reaction and then generating its description (‘react+desc‘), generally resulted in lower performance compared to single-step tasks. For example, while Text+Chem T5-standard achieved an accuracy of 45.03% on reaction prediction (‘react‘), its performance dropped significantly to 22.07% on the two-step task (‘react+desc‘). This trend highlights the challenge of maintaining accuracy across multiple compositional steps, where errors accumulate due to the sequential nature of the task. The drop in performance underscores the importance of compositional reasoning in handling multi-step tasks, that indicating a reliance on surface-level pattern recognition rather than compositional understanding.

Molecular Captions and Linguistic Alignment Molecular captions generated by the models were evaluated using ROUGE metrics. As expected, T5-based models performed best in this category, achieving ROUGE scores of 63.03% and 60.46%, respectively. These models were specifically trained for descriptive tasks, aligning well with the expected format and linguistic style. In contrast, LLaMA-based models produced semantically appropriate descriptions but deviated from the expected format, resulting in lower scores. For instance, LLaMA-7b-instruct-base-chem achieved a ROUGE score of 53.67%, which is competitive but still lower than T5-based models. This discrepancy can be attributed to the free-form nature of LLaMA’s outputs, which may lack the structured format preferred by evaluation metrics. Despite this, the semantic correctness of LLaMA’s descriptions suggests that it treats "Give me a description of this molecule" as a form of scientific question answering (QA), demonstrating versatility even without task-specific training.

Compositionality in Property Estimation Property estimation tasks, such as predicting the number of heavy atoms (‘atoms‘) or activity potential (‘activity‘), yielded consistently high performance across models. For example, LLaMA-finetuned-chemistry achieved an RMSE of 85.03% for atom prediction and 75.3% for activity estimation. These results indicate that short-answer tasks, which require concise numerical outputs, are less prone to errors and thus better suited for evaluating model performance. The success of these models on property estimation tasks reflects their ability to handle

Model	react	desc	react+desc	inchi	react+inchi	activity	react+activity	atoms	react+atoms
Text+Chem T5-standard	45.03	63.03	22.07	39.86	41.6	50.73	44.09	60.08	65.73
Text+Chem T5-augm	40.34	60.46	20.2	39.61	41.3	50.45	43.87	59.83	65.48
chemical-LLaMA	25.76	38.09	31.8	54.97	39.03	61.67	46.9	71.24	66.27
LLaMA-3.2-1b-it-chemistry-assistant	23.85	33.85	28.8	48.45	35.09	75.3	41.7	85.03	80.1
LLaMA-finetuned-chemistry	29.73	44.94	39.9	48.91	35.42	75.75	42.1	85.38	80.43
LLaMA-3.1-8B-Instruct-sft-chem	29.71	48.62	40.3	48.12	34.87	74.9	41.45	84.78	79.8
LLaMA-7b-instruct-base-chem	32.79	51.34	39.77	51.35	37.12	78.5	44.8	88.27	83.4
LLaMA-3.2-3b-it-chemistry	29.48	50.18	41.74	48.21	34.95	75.05	41.52	84.85	79.93
LLaMA-3.2-3b-it-chemistry-2step	30.41	53.67	41.79	48.78	35.31	75.60	41.95	85.27	80.31
ChemLLM-7B-Chat	38.55	61.86	49.91	46.78	36.89	73.2	43.1	87.65	81.3
ChemQwen2-vL	39.79	65.40	51.76	45.92	37.45	72.6	44.5	88.13	82.46
ChemDFM-v1.5-8B	37.5	50.89	47.88	53.05	32.87	80.4	38.9	82.78	76.29

Table 1: Model performance on STEP tasks: one-step vs. two-step evaluation. Cell shading intensity reflects performance, darker blue indicates better scores.

simple compositional structures. However, their performance on more complex tasks, such as multi-step generation, suggests that true compositional reasoning remains a challenge.

Impact of Training Data on Compositionality Models pre-trained on diverse scientific corpora, demonstrated stronger generalization capabilities compared to those trained on narrower datasets. For example, LLaMA-3.1-8B-Instruct-sft-chem achieved a balanced performance across tasks, scoring 48.62% on molecular descriptions (‘desc‘) and 40.3% on two-step generation (‘react+desc‘). This highlights the importance of data diversity in improving model robustness and compositional reasoning. In contrast, models trained on specialized datasets, such as Text+Chem T5-standard, excelled in specific tasks but struggled with broader applications. For instance, while Text+Chem T5-standard achieved the highest ROUGE score (63.03%) for molecular descriptions, its performance on two-step generation (‘react+desc‘) was subpar (22.07%). This trade-off between specialization and generalization underscores the need for training strategies that balance task-specific expertise with broad applicability.

Fine-tuning on compositional tasks The LLaMA-3.2-3b-it-chemistry model (3 billion parameters) was selected for training experiments due to its compact size and competitive baseline performance across tasks, we call the resulting model LLaMA-3.2-3b-it-chemistry-2step. The results are shown in Table 1. We see several improvements in 2-step tasks, but the difference is marginal. The latter suggest the need for better fine-tuning and reasoning techniques.

Challenges with Out-of-Distribution Inputs Across all models, performance dropped significantly when presented with out-of-distribution inputs. For example, ChemQwen2-vL achieved a high ROUGE score of 65.40% for molecular descriptions (‘deck‘) but struggled with multi-step tasks (‘react+desc‘: 51.76%). This indicates a reliance on pattern recognition rather than true compositional reasoning, as models fail to generalize beyond their training distribution.

Linguistic Explanation of Results The discrepancies in performance can often be attributed to linguistic factors. Models trained on specific formats (e.g., T5 for molecular descriptions) exhibit better alignment with evaluation metrics, whereas models which generate free-form text, may produce semantically correct but stylistically divergent outputs. For example, LLaMA-7b-instruct-base-chem achieved a high ROUGE score of 53.67% for molecular captions, but its deviation from the expected format decreased its score. This highlights the importance of aligning model outputs with evaluation criteria to ensure fair comparisons.

5 Conclusion

The results of our study reveal that current Chemical LLMs struggle with compositional reasoning, particularly in multi-step tasks and out-of-distribution (OOD) scenarios. While specialized models excel in specific tasks, general-purpose models demonstrate adaptability but falter when integrating information across multiple steps. Key findings include significant performance drops in multi-step tasks, highlighting a reliance on surface-level pattern recognition rather than true compositional understanding. These insights highlight the need for better architectures and training to im-

prove compositional reasoning, with a promising direction being the integration of external tools into LLM prompts for hybrid reasoning.

Limitations

First, we evaluated models that are publicly available on HuggingFace (HF) (details in Appendix A). Notably, there are other popular models such as Chemformer (https://github.com/MolecularAI/Chemformer), Molformer (https://github.com/IBM/molformer), and T5Chem (https://github.com/HelloJocelynLu/t5chem), which could not be integrated due to the lack of HF checkpoints. Second, the evaluated models predominantly focus on compositional reasoning within sequence-based representations of molecules, but it is crucial to explore other formats in the future, such as 3D structures, which also hold significant importance for chemical reasoning. Third, we emphasize that the evaluated models were developed for research purposes and may contain unintended biases; any molecules generated by these models should undergo thorough evaluation through standard clinical testing.

Ethics Statement

The models and datasets used in this work are publicly available for research purposes. The incorporation of AI into applied chemistry introduces a variety of risks and ethical dilemmas. First, the direct implementation of AI-generated predictions—particularly those involving potentially hazardous or dangerous compounds—without rigorous validation could result in human injuries, casualties, or damage to laboratory facilities. Second, the absence of proper oversight could lead to the misuse of chemical language models and AI in general, potentially facilitating the production of dangerous or illegal chemical compounds, with significant ethical and societal consequences. To address these concerns, it is essential to develop and implement robust ethical guidelines for the development and deployment of AI in chemistry. Additionally, fostering transparency in model design, training data, and evaluation methodologies can help mitigate potential risks and ensure responsible use of these technologies.

Acknowledgements

This work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

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.

Veronika Ganeeva, Kuzma Khrabrov, Artur Kadurin, Andrey Savchenko, and Elena Tutubalina. 2024a. Chemical language models have problems with chemistry: A case study on molecule captioning task. In *The Second Tiny Papers Track at ICLR 2024*.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. 2024b. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013, Miami, Florida, USA. Association for Computational Linguistics.

Anna Gaulton, Anne Hersey, Michael Nowotka, et al. 2017. The chembl database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954.

Anna Gaulton, Anne Hersey, Michal Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo-Meullenet, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María P. Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. 2016. The chembl database in 2017. *Nucleic Acids Research*, 45:D945 – D954.

Wengong Jin et al. 2020. Uspto-mit dataset for reaction prediction. *Journal of Chemical Information and Modeling*, 60(10):4649–4657.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Maksim Kuznetsov, Airat Valiev, Alex Aliper, Daniil Polykovskiy, Elena Tutubalina, Rim Shayakhmetov, and Zulfat Miftahutdinov. 2025. nach0-pc: Multitask language model with molecular point cloud encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24357–24365.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, et al. 2024. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389.

Daniel Mark Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge.

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.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. 2021. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Constantine Bekas, and Alpha Albert Lee. 2018. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5:1572 – 1583.

Teague Sterling and John J. Irwin. 2015a. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337. PMID: 26479676.

Teague Sterling and John J Irwin. 2015b. Zinc 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

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. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. 2022. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23):5938–5951.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2012. Pubtator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 40(W1):W518–W522.

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

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024. Chemdfm: Dialogue foundation model for chemistry. *Preprint*, arXiv:2401.14818.

A Models

Model	Size	Molecule Captioning	Reaction Prediction	Property Prediction	Scientific Q&A
Text2Chem-standard	223M	✓	✓	×	×
Text2Chem-augm	223M	✓	✓	×	×
chemical-llama	6.74B	×	✓	✓	✓
llama-3.2-1b-it-chem-assist	1.24B	×	×	×	✓
llama-finetuned-chemistry	8.03B	×	✓	✓	✓
Llama-3.1-8B-Instruct-sft-chem	8.03B	×	✓	✓	✓
llama-7b-instruct-base-chem	6.22B	×	×	×	✓
Llama-3.2-3b-it-chemistry	1.24B	×	×	×	✓
ChemLLM-7B-Chat	7.74B	×	×	×	✓
ChemQwen2-vL	2.21B	×	×	×	✓
ChemDFM-v1.5-8B	8.03B	✓	✓	✓	✓

Table 2: Summary of Chemical Language Models. The ✓ and × symbols indicate whether the model was trained on the corresponding task or not.

Dataset	Size	Task
CHEBI-20	33,000 Molecules	Molecule Captioning
USPTO-50k	55,000 Reactions	Reaction Prediction
PubChem	119M Compounds 328M Substances	Molecule properties
MoleculeACE	35,000 Molecules	Activity prediction

Table 3: Summary of used datasets. MoleculeACE was proposed in (Van Tilborg et al., 2022).

Text2Chem is a T5-architecture model designed for molecule captioning and reaction prediction. Built upon pre-trained language models,

Text2Chem is fine-tuned on annotated datasets derived from scientific articles and patents. By leveraging domain-specific annotations, Text2Chem demonstrates strong performance in molecule description generation, reactions, and properties within textual data. The model was trained on a large corpus of chemical patents, scientific papers, and datasets such as USPTO-MIT (Jin et al., 2020) and ChEMBL (Gaulton et al., 2017), making it well-suited for chemistry-specific applications.

Chemical-LLaMA is a variant of the LLaMA series fine-tuned for chemistry-specific tasks. By incorporating domain knowledge through pre-training on chemical corpora, Chemical-LLaMA demonstrates strong performance in tasks such as reaction prediction, molecule captioning, property prediction, and scientific question answering. The model is fine-tuned on a combination of general-purpose text data and chemistry-specific datasets, ensuring its versatility in handling both generic and domain-specific queries.

Llama-2-Finetuned-Chemistry is a fine-tuned version of Llama-2 for chemistry-related tasks. Building upon the robust capabilities of Llama-2, this model incorporates domain-specific knowledge through fine-tuning on chemistry-specific datasets. It is particularly effective in tasks such as reaction outcome prediction, molecular property estimation, and scientific question answering, making it a valuable tool for evaluating compositional reasoning in the chemistry domain.

Meta-Llama-3.1-8B-Instruct-sft-re-chemprot-1209 is an instruction-tuned version of the LLaMA 3.1 8B series, further refined for chemistry-specific tasks. This model incorporates supervised fine-tuning (SFT) and reinforcement learning techniques to improve performance. Trained on a combination of general-purpose text data and chemistry-specific datasets, it includes additional fine-tuning on benchmarks such as ChemProt (Wei et al., 2012). As a result, it excels in tasks requiring precise instructions, such as reaction prediction, property prediction, and scientific reasoning.

The **Text+Chem T5-Standard** model is a domain-specific adaptation of the T5 (Text-to-Text Transfer Transformer) architecture (Raffel et al., 2020), tailored for chemical applications. It leverages the standard T5 architecture with encoder-decoder capabilities, enabling it to handle tasks such as molecular property prediction, reaction generation, and chemical text summarization. This variant uses the base configuration of T5, featuring approximately **220 million parameters**. The model is pre-trained on a combination of general-purpose text corpora and chemistry-specific datasets, ensuring strong performance in both linguistic and chemical domains. [Link]

The **Text+Chem T5-Augm** model extends the capabilities of the standard T5 architecture by incorporating augmentations specifically designed for chemical data. These augmentations include specialized tokenizers for SMILES strings and additional training on chemical reaction datasets (Christofidellis et al., 2023). With around **3 billion parameters**, this augmented variant demonstrates superior performance in multi-step reasoning tasks compared to its standard counterpart. The model's enhanced architecture allows it to better capture complex relationships between functional groups and reaction pathways. [Link]

Chemical-LLaMA is a domain-adapted version of Meta's LLaMA series (Touvron et al., 2023), fine-tuned for chemical applications. Built upon the foundation of LLaMA's causal language modeling architecture, this model contains approximately **7 billion parameters** and is trained on a diverse set of chemical datasets, including ZINC (Sterling and Irwin, 2015a) and PubChem (Kim et al., 2016). Its architecture supports zero-shot and few-shot learning, making it highly versatile for tasks like molecular design and reaction prediction. [Link]

This model, **LLaMA-3.2-1B-IT-Chemistry-Assistant**, is a lightweight variant of the LLaMA series, specifically optimized for interactive chemistry tasks. With approximately **1 billion parameters**, it balances computational efficiency with performance, making it suitable for real-time applications such as chatbots or virtual assistants in chemistry education and research. The model is fine-tuned on Italian-annotated chemical datasets, enhancing its multilingual capabilities (Touvron et al., 2023). [Link]

LLaMA-Finetuned-Chemistry is a fine-tuned version of the original LLaMA model, adapted for chemical tasks through extensive finetuning on domain-specific datasets (Touvron et al., 2023). Featuring **13 billion parameters**, this model excels in tasks requiring deep understanding of chemical concepts, such as retrosynthesis and drug discovery. Its architecture retains the robustness of the LLaMA series while incorporating specialized knowledge from sources like ChemBL (Gaulton et al., 2016). [Link]

The **Meta-LLaMA-3.1-8B-Instruct-ChemProt** model is an instruction-tuned variant of LLaMA, specifically designed for chemical protein interaction tasks. With approximately **8 billion parameters**, it is trained on curated datasets related to cheminformatics and biochemistry, enabling it to predict interactions between small molecules and proteins with high accuracy (Touvron et al., 2023). This model is particularly useful for drug-target interaction studies. [Link]

LLaMA-7B-Instruct-Base-Chem is a 7-billion-parameter model derived from the LLaMA series, fine-tuned for general chemical tasks using instruction-based learning (Touvron et al., 2023). It combines the strengths of LLaMA's large-scale pretraining with domain-specific instructions, allowing it to perform well in tasks like molecular property prediction and reaction classification. The model's versatility makes it a popular choice for researchers working across various subfields of chemistry. [Link]

The **LLaMA-3.2-3B-IT-Chemistry** model is a compact, Italian-language variant of LLaMA, designed for multilingual chemical applications. With approximately **3 billion parameters**, it is optimized for tasks involving chemical nomenclature and descriptions in Italian (Touvron et al., 2023). This model bridges the gap between linguistic and chemical knowledge, making it valuable for educational and professional settings where multilingual support is required. [Link]

ChemLLM-7B-Chat is a conversational model built on a 7-billion-parameter architecture, specifically designed for interactive chemical discussions (Zhang et al., 2024). Trained on a mix of scientific literature and dialogue datasets, this model excels in generating human-like responses to chemical queries. Its focus on conversational AI makes it ideal for use cases such as virtual labs, tutoring systems, and collaborative research environments. [Link]

ChemQwen2-vL is part of the Qwen series developed by Alibaba Cloud, with a specialization in chemical applications (Bai et al., 2023). This large-scale model contains over **17 billion parameters** and is trained on extensive chemical datasets, including reaction databases and material science corpora. Its architecture supports advanced reasoning tasks, such as predicting reaction outcomes and designing novel compounds. The "vL" variant emphasizes visual and linguistic integration, enabling it to process multimodal inputs. [Link]

ChemDFM-v1.5-8B is a domain-specific model developed for drug discovery and formulation modeling, featuring approximately **8 billion parameters** (Zhao et al., 2024). Based on the DFM (Drug Formulation Modeling) framework, this model integrates deep learning techniques with chemical knowledge graphs to enhance predictive accuracy. Its architecture is optimized for handling complex molecular structures and predicting formulation stability, making it a powerful tool for pharmaceutical research. [Link]