CHEMSETS: How Capable Are Chemistry LLMs?

Christoph Bartmann[1]* **Mykyta Ielanskyi**[1]* **Johannes Schimunek**[1]
Philipp Seidl[1] **Günter Klambauer**[1,2] **Sohvi Luukkonen**[1]

[1]ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University, Linz, Austria
[2]NXAI GmbH, Linz, Austria
*Equal contribution.

Abstract

Large Language Models (LLMs) have demonstrated immense versatility and have been successfully adapted to tackle numerous problems in scientific domains. In chemistry, specialized LLMs have been recently developed for molecule structure tasks such as molecule name conversion, captioning, text-guided generation, and property or reaction prediction. However, evaluations of chemistry-focused LLMs remain inconsistent and often lack rigor: new models are typically assessed only on tasks they were explicitly trained for, while compared models have been trained on different sets of tasks. In addition, several proposed benchmarks introduce idiosyncratic features, e.g., task-specific input or output tags, and, thus, the LLMs' performance is highly sensitive to prompting strategies, answer formatting, and generation parameters, further complicating reproducible evaluation. To address these shortcomings, we perform a standardized and reproducible method comparison of chemical reasoning models on CHEMSETS, a flexible benchmark suite integrated into `lm-evaluation-harness`. CHEMSETS unifies existing benchmarks with newly designed symbolically verifiable tasks, thereby expanding both task diversity and difficulty. Through this evaluation, we establish a fair leaderboard and provide new insights into the limitations of recently proposed chemistry-aware LLMs. We show that current chemistry LLMs exhibit limited generalization beyond the specific tasks they were trained on. Remarkably, across chemical tasks, recent open-weight non-specialist reasoning models outperform specialist models.

1 Introduction

Large language models (LLMs) have emerged as versatile multi-task systems, capable of addressing a wide range of problems with a single model (Brown et al., 2020; Chowdhery et al., 2023; Hoffmann et al., 2022; Touvron et al., 2023). Advances in inference scaling, particularly chain- of-thought generation, have further improved performance on symbolic tasks such as mathematics and logic puzzles (DeepSeek-AI et al., 2025; Yang et al., 2025). In chemistry, the rise of LLMs has led to the development of several instruction-tuned models designed as easy-to-use tools for diverse chemical structure tasks (Zhang et al., 2024; Fang et al., 2024; Zhao et al., 2025c; Yu et al., 2024; Xia et al., 2025). Building on this momentum, the first chemistry-specialized reasoning models have appeared (Narayanan et al., 2025; Zhao et al., 2025b,a), with each new chemistry LLM (cLLM) claiming state-of-the-art performance across a range of chemical structure tasks.

Unfortunately, chemistry LLMs to date have been trained and *evaluated* on disparate sets of tasks, making direct model comparisons difficult and often unfair. New models are typically assessed on tasks included in or very similar to their own training data, while baseline models may not have seen those tasks at all during their training. In such situations, benchmarks are effectively

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

testing the "interpolation" capabilities of the new model against the "extrapolation" capabilities of the baselines. Moreover, evaluation pipelines often include idiosyncratic conventions (e.g., input/output tags), and performance is highly sensitive to the system prompt, decoding settings, and answer extraction – issues widely recognized in broader LLM evaluation (Liang et al., 2022; Cobbe et al., 2021; Hendrycks et al., 2021). Without a standardized protocol, formatting and extraction choices can artificially inflate performance for some models while unfairly penalizing others.

In this work, we address these challenges through the following three contributions:

- We introduce CHEMSETS, a standardized evaluation pipeline with robust model-specific answer extraction integrated into `lm-evaluation-harness` (Gao et al., 2024), to evaluate LLMs on chemical structure tasks. It integrates and standardizes two existing chemistry benchmarks, and
- introduces SymMolic, a benchmark set which solely focuses on symbolically solvable molecule structure tasks. This evaluation set was crafted to cover both a wide range of molecule complexities and property values for each task.
- We evaluate multiple domain-specific and general language LLMs in order to attain systematic insights into limitations and failure modes of current LLMs applied to chemistry.

2 CHEMSETS Benchmarks

CHEMSETS centers on *molecular structure reasoning* – tasks whose answers follow directly from graph-derived properties, where the molecular structure is typically represented in SMILES format (Weininger, 1988). By focusing on **symbolically verifiable tasks**, where correctness can be rigorously determined via deterministic algorithms, we enable reproducible model comparisons and ensure that performance reflects reasoning capability rather than artifacts.

Several existing benchmarks, while valuable, fall outside this scope of molecular structure reasoning with symbolically verifiable tasks. ChemBench (Mirza et al., 2024) and ChemEval (Huang et al., 2024) cover broad ranges of chemistry problems beyond reasoning on the chemical structure, including conceptual reasoning and literature comprehension. Similarly, ChemLLMBench (Guo et al., 2023), **Mol-Instructions** (Fang et al., 2024), and SMolInstruct (Yu et al., 2024) include tasks such as molecule captioning, property prediction, and retrosynthesis, which inherently depend on reference data or trained models for evaluation. These benchmarks are therefore not directly suited to measuring symbolic reasoning accuracy.

Within these constraints, CHEMSETS integrates three core datasets for which the majority of tasks are symbolically solvable: **ChemIQ**, **ether0**, and **SymMolic**. Table A1 outlines their task coverage, and Figure A2 shows the distribution of molecular complexity across datasets. We group the tasks under five categories: translation, constrained generation, feature counting, molecule comparisons, and reaction predictions.

ChemIQ was introduced by Runcie et al. (2025) and was originally used to evaluate the chemical reasoning capabilities of OpenAI's o3-mini series. It consists of eight tasks (five of which can be verified symbolically) spanning a wide range of task categories, totaling 796 open-ended questions with diverse expected output types. Many of the tasks do not require chemistry understanding but serve as useful sanity checks, testing whether a model can correctly parse and interpret SMILES notation - a crucial prerequisite for more advanced molecular structure reasoning tasks.

ether0 was introduced by Narayanan et al. (2025) as the evaluation set for their ether0 model and includes a subset of tasks the model was trained on. It comprises a total of 325 questions — spanning open-ended tasks and multiple-choice questions (MCQs) — all of which expect a SMILES string as the answer. The open-ended portion covers eight tasks, five of which are symbolically evaluable, with five falling under the category of constrained generation. The MCQs cover six property categories, none of which are symbolically evaluable.

SymMolic v0 is a dataset focused exclusively on symbolically verifiable tasks over molecular structure. SymMolic v0 consists of 1900 questions across 19 tasks from the translation and feature counting categories. One key design choice made for this benchmark is that, for each task, we intentionally cover a broad range of molecular complexities. This enables evaluation of a model's structural reasoning ability along two axes: the variety and difficulty of tasks, and the complexity of

Table 1: Results on CHEMSETS: For each model, average task accuracy per benchmark is reported. Error bars indicate standard errors across tasks. R denotes reasoning models. The highest value per column is marked in bold. Green color indicates the highest, and yellow within the standard deviation.

	Size	R	ChemIQ	ether0	SymMolic
Chemistry LLMs					
Llama-molinst (Fang et al., 2024)	8B		3.2 ± 0.6	0.6 ± 0.2	8.9 ± 0.5
ChemDFM-8B (Zhao et al., 2025c)	8B		1.1 ± 0.4	1.9 ± 0.4	3.3 ± 0.3
ChemDFM-13B (Zhao et al., 2025c)	13B		1.4 ± 0.3	0.9 ± 0.4	2.3 ± 0.2
ChemLLM-7B (Zhang et al., 2024)	7B		0.7 ± 0.3	0.4 ± 0.2	2.4 ± 0.2
LlaSMol-Mistral (Yu et al., 2024)	7B		1.6 ± 0.3	0.4 ± 0.2	3.6 ± 0.3
Txgemma-9b (Wang et al., 2025)	9B		2.6 ± 0.8	3.9 ± 0.8	0.7 ± 0.1
Txgemma-27b (Wang et al., 2025)	27B		4.0 ± 0.6	3.0 ± 0.6	3.0 ± 0.3
Ether0 (Narayanan et al., 2025)	24B	✓	13.1 ± 1.1	$\mathbf{45.9} \pm 2.2$	2.4 ± 0.3
Generalist LLMs					
Qwen3-8b (Yang et al., 2025)	8B	✓	12.0 ± 1.1	4.1 ± 0.7	19.3 ± 0.7
Qwen3-14b (Yang et al., 2025)	14B	✓	12.2 ± 1.2	3.7 ± 0.7	24.9 ± 0.8
Qwen3-32b (Yang et al., 2025)	32B	✓	22.6 ± 1.2	2.8 ± 0.6	28.2 ± 0.8
Qwen3-Think-30B* (Yang et al., 2025)	30B (A3B)	✓	31.7 ± 1.5	4.1 ± 0.9	34.8 ± 0.9
Qwen3-Think-235B* (Yang et al., 2025)	235B (A22B)	✓	65.5 ± 1.2	9.2 ± 1.3	50.1 ± 0.9
GPT-oss-20b-medium (OpenAI, 2025)	20B (A4B)	✓	20.8 ± 1.0	10.0 ± 1.1	33.8 ± 0.8
GPT-oss-20b-high (OpenAI, 2025)	20B (A4B)	✓	47.4 ± 1.0	13.5 ± 1.3	51.1 ± 0.8
GPT-oss-20b-high (OpenAI, 2025)	20B (A4B)	✓	47.4 ± 1.0	13.5 ± 1.3	51.1 ± 0.8
GPT-oss-120b-medium (OpenAI, 2025)	120B (A5B)	✓	36.9 ± 1.2	15.9 ± 1.5	43.1 ± 0.9
GPT-oss-120b-high (OpenAI, 2025)	120B (A5B)	✓	$\mathbf{65.6} \pm 1.1$	18.9 ± 1.5	$\mathbf{57.2} \pm 0.9$

* version 2507

the molecules involved. Furthermore, in the case of feature-counting tasks, we also ensure a diverse distribution of feature values. Additional details on the construction of SymMolic and the symbolic The verifiers used for each task are provided in Appendix A.1.

3 Evaluation & Leaderboard

Usability Features. We built our benchmark on top of lm-evaluation-harness (Gao et al., 2024). This enables a modular approach to tasks and model configurations. We provide default configurations with sampling parameters, preprocessor, and extractors specifically tailored for each model.

Symbolic Extraction. Fair comparison hinges on precise answer extraction. Recent work (Chandak et al., 2025; Shao et al., 2025) questions reported RLVR gains, arguing that models may learn format-following rather than new skills. In practice, adopting stronger extractors alone can inflate scores, while baselines without robust extraction are artificially deflated. To mitigate this, we use – for each model family – a system prompt and a matching extractor.

Task Verifiers. Each task is evaluated using a task-specific success metric and verifier. Success metric values range from 0 to 1, with metrics having higher values indicating better performance. For ChemIQ and ether0, we adopt the original metrics and verifiers as defined in their respective papers (Narayanan et al., 2025; Runcie et al., 2025). For the SymMolic tasks, the evaluation procedures are detailed in Section A.2.

Accuracy Metrics. We report the mean of 3 rollout attempts per question (Wang et al., 2022). Following standardized reporting practices advocated by HELM (Liang et al., 2022), we macro-average accuracies: task-level accuracy is averaged over questions; category- and dataset-level scores are the unweighted mean of task-level accuracies. For tasks present in multiple datasets (e.g., SMILES→IUPAC in ChemIQ and SymMolic), we recompute the task accuracy jointly over the union of questions.

Living Benchmark. The full leaderboard and per-task results is hosted online at CHEMSETS Leaderboard.

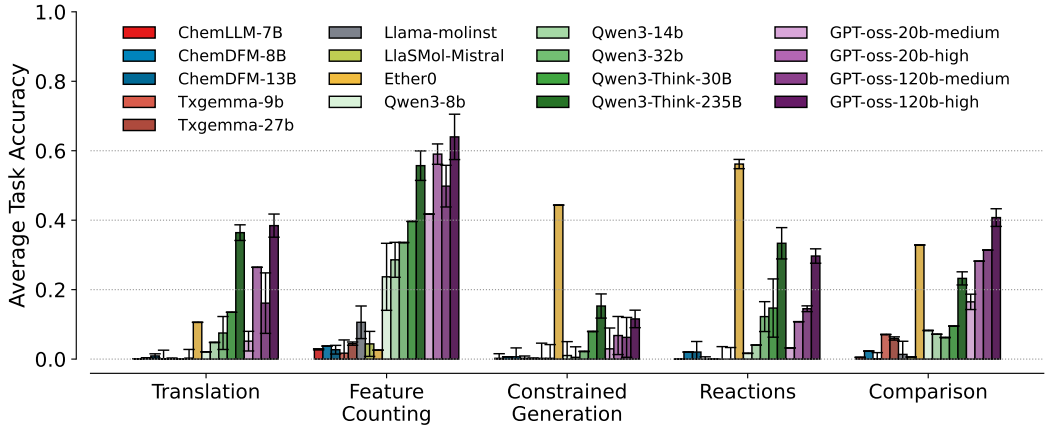

Figure 1: Performance comparison on CHEMSETS benchmark suite. Average task accuracy for five distinct categories of chemical reasoning tasks. Large generalist models, particularly from the Qwen3 and GPT-oss series, consistently outperform smaller, specialized chemical language models across a majority of categories. Error bars represent standard errors across tasks.

Following the model of the Open LLM Leaderboard (Fourrier et al., 2024), We will accept submissions of new chemistry LLMs (open weights or free API access). Authors provide (i) a config file to run their model in `lm-evaluation-harness`, and (ii) a short description of training tasks.

4 Results & Discussions

We evaluated 16 models across CHEMSETS under unified extraction protocols (Table 1, Figures 1, A3, A4, A5, and A6), and observed consistent differences between general-purpose LLMs and chemistry-specialized models, as well as clear effects of scale, architecture, and reasoning budget.

General-purpose LLMs, despite lacking chemistry-specific training, often outperform specialized chemistry models. On ChemIQ, GPT-oss-120b-high (65.6 ± 1.1) and Qwen3-235B-Think (65.5 ± 1.2) surpass all chemistry-focused models, which remain in the single-digit or low-teen range. This advantage holds even on SMILES-based tasks, where chemistry LLMs should excel, with general models maintaining 10 to 50 times higher accuracy.

Performance within model families is driven by scale and reasoning budget. Accuracy increases consistently with parameter count, MoE architectures outperform dense counterparts, and test-time reasoning nearly doubles accuracy in some cases. Together, these factors yield the strongest results, with Qwen3-235B-Think and GPT-oss-120b-high reaching comparable peak performance.

Among chemistry-specialized models, ether0 performs best (45.9 ± 2.2 on its benchmark), reflecting its training on SMILES outputs. However, it fails to generalize beyond this narrow setting, dropping to near-zero accuracy on tasks requiring natural language or non-SMILES outputs.

Task difficulty also varies: general LLMs achieve near-ceiling performance on simple counting tasks but remain below $40\,\%$ on translation or functional group identification. Chemistry-specialized models perform even worse on these tasks. This suggests that while some structured reasoning tasks are largely solved, more complex forms of chemical understanding remain open.

Conclusion. Evaluating 16 models, we find that general-purpose LLMs outperform chemistry-specialized models on the majority of the considered tasks, while scaling, mixture-of-experts, and reasoning-augmented variants yield the strongest results.

Our results highlight a central open question: as generalist models continue to advance, can domain-specific LLMs keep pace, or will their utility remain confined to narrow, in-domain tasks? By establishing a fair and reproducible evaluation suite, we hope CHEMSETS will help clarify this trajectory.

Limitations and Outlooks Our evaluation is restricted to open-weight models; closed-source systems (Runcie, 2025; Anthropic, 2024; DeepMind, 2025) remain untested. Although CHEMSETS introduces

new symbolically verifiable tasks, it also integrates existing benchmarks, so data contamination from pretraining corpora cannot be ruled out. The limited transparency about training sets raises the risk of molecule leakage; Reaction prediction tasks based on USPTO dataset (Lowe, 2012) are especially prone to such leakage, making generalization harder to assess. Finally, like any benchmark, CHEMSETS is static: tasks may saturate as models improve. To address this, we envision iterative releases with increasingly challenging symbolically verifiable tasks.

5 Acknowledgments

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects FWF AIRI FG 9-N (10.55776/FG9), AI4GreenHeatingGrids (FFG- 899943), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01), FWF Bilateral Artificial Intelligence (10.55776/COE12). We thank NXAI GmbH, Audi AG, Silicon Austria Labs (SAL), Merck Healthcare KGaA, GLS (Univ. Waterloo), TÜV Holding GmbH, Software Competence Center Hagenberg GmbH, dSPACE GmbH, TRUMPF SE + Co. KG.

References

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic model card, June 2024. URL https://www.anthropic.com/news/claude-3-family. Cited on page 4.

S. H. Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, June 1981. doi: 10.1021/ja00402a071. Cited on pages 9 and 12.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. Cited on page 1.

N. Chandak, S. Goel, and A. Prabhu. Incorrect Baseline Evaluations Call into Question Recent LLM-RL Claims. Notion Blog, 2025. URL https://safe-lip-9a8.notion.site/Incorrect-Baseline-Evaluations-Call-into-Question-Recent-LLM-RL-Claims-2012f1fbf0ee8094ab8ded1953c15a37?pvs=4. Cited on page 3.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. doi: 10.1145/3716846. Cited on page 1.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems, 2021. Cited on page 2.

G. DeepMind. Gemini 2.5 Pro. Google model page, 2025. URL https://deepmind.google/models/gemini/pro/. Cited on page 4.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. Cited on page 1.

C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho, and H. Ji. Translation between molecules and natural language, 2022. Cited on page 14.

Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models, March 2024. Cited on pages 1, 2, and 3.

C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. Open LLM Leaderboard v2, 2024. URL https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. Cited on page 4.

L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, et al. The Language Model Evaluation Harness, July 2024. URL https://zenodo.org/records/12608602. Cited on pages 2 and 3.

T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023. Cited on page 2.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset, 2021. Cited on page 2.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models, 2022. Cited on page 1.

Y. Huang, R. Zhang, X. He, X. Zhi, H. Wang, X. Li, F. Xu, D. Liu, H. Liang, Y. Li, et al. ChemEval: A Comprehensive Multi-Level Chemical Evaluation for Large Language Models, September 2024. Cited on page 2.

R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022. doi: 10.1088/2632-2153/ac3ffb. Cited on page 14.

S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, et al. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956. Cited on page 9.

G. Landrum and R. contributors. RDKit: Open-Source Cheminformatics Software, 2006. URL https://www.rdkit.org. Cited on page 9.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models, 2022. Cited on pages 2 and 3.

D. M. Lowe. *Extraction of Chemical Structures and Reactions from the Literature*. PhD thesis, University of Cambridge, 2012. Cited on page 5.

D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling*, 51(3):739–753, March 2011. doi: 10.1021/ci100384d. Cited on page 10.

A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, et al. Are large language models superhuman chemists?, 2024. Cited on page 2.

S. M. Narayanan, J. D. Braza, R.-R. Griffiths, A. Bou, G. Wellawatte, M. C. Ramos, L. Mitchener, S. G. Rodriques, and A. D. White. Training a scientific reasoning model for chemistry, 2025. Cited on pages 1, 2, and 3.

OpenAI. GPT-OSS: Open-Weight Model Release (GPT-OSS-20B, GPT-OSS-120B). Model documentation, August 2025. URL https://openai.com/index/introducing-gpt-oss/. Cited on page 3.

N. Runcie. GPT-5 achieves state-of-the-art chemical intelligence, 2025. URL https://www.blopig.com/blog/2025/08/gpt-5-achieves-state-of-the-art-chemical-intelligence/. Cited on page 4.

N. T. Runcie, C. M. Deane, and F. Imrie. Assessing the Chemical Intelligence of Large Language Models, May 2025. Cited on pages 2 and 3.

J. Schimunek, S. Luukkonen, and G. Klambauer. MHNfs: Prompting in-context bioactivity predictions for low-data drug discovery. *Journal of Chemical Information and Modeling*, 65(9): 4243–4250, 2025. doi: 10.1021/acs.jcim.4c02373. Cited on page 9.

R. Shao, S. S. Li, R. Xin, S. Geng, Y. Wang, S. Oh, et al. Spurious rewards: Rethinking training signals in RLVR, 2025. Cited on page 3.

R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science, 2022. Cited on page 14.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models, 2023. Cited on page 1.

E. Wang, S. Schmidgall, P. F. Jaeger, F. Zhang, R. Pilgrim, Y. Matias, J. Barral, D. Fleet, and S. Azizi. TxGemma: Efficient and Agentic LLMs for Therapeutics, 2025. Cited on page 3.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2022. Cited on page 3.

D. Weininger. 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, 1988. doi: 10.1021/ci00057a005. Cited on page 2.

G. P. Wellawatte, A. Seshadri, and A. D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science*, 13(13):3697–3705, 2022. Cited on page 9.

C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, et al. LiveBench: A Challenging, Contamination-Limited LLM Benchmark, April 2025. Cited on page 14.

Y. Xia, P. Jin, S. Xie, L. He, C. Cao, R. Luo, et al. NatureLM: Deciphering the Language of Nature for Scientific Discovery, February 2025. Cited on page 1.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, et al. Qwen3 technical report, 2025. Cited on pages 1 and 3.

B. Yu, F. N. Baker, Z. Chen, X. Ning, and H. Sun. LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset, August 2024. Cited on pages 1, 2, and 3.

Z. Zeng, B. Yin, S. Wang, J. Liu, C. Yang, H. Yao, et al. ChatMol: Interactive molecular discovery with natural language. *Bioinformatics*, 40(9):btae534, 2024. Cited on page 14.

D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, and Y. Li. ChemLLM: A Chemical Large Language Model, April 2024. Cited on pages 1, 3, and 14.

G. Zhao, S. Li, Z. Lu, Z. Cheng, H. Lin, L. Wu, H. Xia, H. Cai, W. Guo, H. Wang, et al. MolReasoner: Toward Effective and Interpretable Reasoning for Molecular LLMs, August 2025a. Cited on page 1.

Z. Zhao, B. Chen, Z. Wan, L. Chen, X. Lin, S. Yu, S. Zhang, D. Ma, Z. Zhu, D. Zhang, et al. ChemDFM-R: An Chemical Reasoner LLM Enhanced with Atomized Chemical Knowledge, July 2025b. Cited on page 1.

Z. Zhao, D. Ma, L. Chen, L. Sun, Z. Li, Y. Xia, et al. Developing ChemDFM as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4), 2025c. doi: 10.1016/j.xcrp.2025.102523. Cited on pages 1 and 3.

Contents

A SymMolic v0 Details **9**
A.1 Dataset Creation . . . 9
A.2 Verifiers . . . 10

B CHEMSETS Datasets Details **12**

C Models **14**
C.1 Scientific LLMs . . . 14
C.2 Generalist LLMs . . . 14
C.3 Model Nomenclature . . . 14

D Extended Results **15**
D.1 Effect of molecule complexity . . . 15
D.2 Results per Tasks . . . 16

A SymMolic v0 Details

An overview of the SymMolic task is provided in Table A1.

A.1 Dataset Creation

As all tasks in SymMolic are symbolically evaluable, data may come from any source because ground-truth answers can be derived from the molecular graph (except for IUPAC tasks, which require a provided ground truth). We selected PubChem (Kim et al., 2023) as it is a comprehensive and publicly accessible chemical database that covers a broad spectrum of molecules and associated bioactivity information (Schimunek et al., 2025). Its diversity and scale — currently over 119 million compounds and 295 million bioactivity records (Kim et al., 2023) — make it well suited for evaluating chemistry LLMs across heterogeneous chemical spaces. We first draw a random subset from PubChem to form a testing pool (we do not use the full database, as we are developing a comprehensive dataset suite with planned train/test/validation splits and reserve the remainder for future dataset releases). For this pool, we extract the SMILES and IUPAC name of each molecule and compute its Bertz complexity (Bertz, 1981) using RDKit (Landrum and contributors, 2006).

We argue that the difficulty of a chemistry reasoning task depends on both the question type and the complexity of the molecule of interest. Accordingly, we bin molecules into five complexity ranges: [0-100, 100-300, 300-600, 600-1000, 1000+], and sample molecules for each task from these bins. For each task, we sample 100 questions (see sampling strategy per task category below). To introduce variability in phrasing, each task has 15 question templates – a combination of manually authored templates and LLM-generated reformulations – which are sampled uniformly at random.

Translation. For the four translation tasks, we sample 20 molecules at random from each complexity bin once. The same set of molecules is used across all translation tasks, which allows for direct comparison of task difficulty. The SMILES and IUPAC names are already provided in the dataset, while for molecular formula tasks, the ground truth is computed using RDKit.

Feature counting. With the exception of `functional_group` – for which we adopt the exmol definition (Wellawatte et al., 2022) to enumerate present functional groups – we compute ground-truth feature counts for every molecule in the evaluation pools using RDKit-based symbolic solvers. Most feature-count distributions are highly skewed toward one or a few values. To mitigate this imbalance, rather than uniform sampling within each complexity bin, we use double inverse-frequency sampling: each candidate is weighted by the inverse frequency of its feature value and by the inverse frequency of its complexity bin, and sampled proportionally to the product (Algorithm 1). To avoid pathological outliers, for each task we discard molecules whose feature value occurs fewer than 1,000 times in the evaluation pool. The resulting feature-value histograms for SymMolic are shown in Figure A1

Algorithm 1 Feature Frequency-Complexity Bin Weighted Sampling

Require: Molecule dataset $\mathcal{D}$ with:
  f: discrete feature values
  c: complexity bin assignments
Require: Sample size $N = 100$, frequency threshold $\tau = 1000$
Compute value frequencies:
  $n_f \leftarrow$ count of each f in $\mathcal{D}$
  $n_c \leftarrow$ count of each c in $\mathcal{D}$
for each molecule $i \in \mathcal{D}$ **do**
  $W_{i,f} \leftarrow 1/n_{f_i}$ if $n_{f_i} > \tau$ else 0
  $W_{i,c} \leftarrow 1/n_{c_i}$
  $W_i \leftarrow W_{i,f} \times W_{i,c}$
end for
Normalize: $W_i \leftarrow W_i / \sum_j W_j$
$\mathcal{S} \leftarrow$ Select N molecules from $\mathcal{D}$ with probabilities W_i
return $\mathcal{S}$

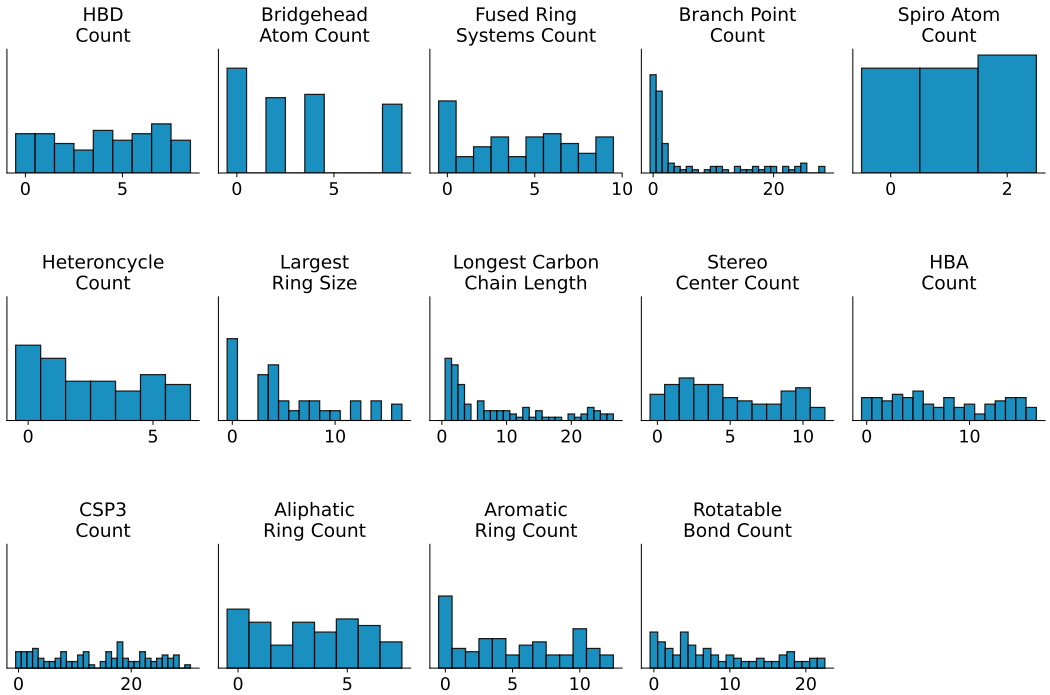

Figure A1: Feature value distributions for each feature counting task in SymMolic, except Function Groups Identification, as it's not a distribution of integers.

A.2 Verifiers

Translation. For `x2formula` tasks, an answer is considered correct if the molecular composition matches that of the reference answer, meaning the order of atom types in the formula is not important. For `smiles2iupac`, we adopt the same verifier as in **ChemIQ**: the generated IUPAC name is evaluated using the Open Parser for Systematic IUPAC Nomenclature (OPSIN) API (Lowe et al., 2011). An IUPAC name is accepted as correct if it can be parsed into the intended structure. For `iupac2smiles`, correctness is determined by verifying that the generated SMILES corresponds to the same molecular structure as the reference SMILES associated with the input IUPAC name.

Feature Counting. For all feature counting tasks, except functional groups identification, the output is an integer for which the ground truth can be deterministically obtained from the molecular structure using RDKit. The generated answer must exactly match the reference value. For functional group

identification the output is a set of strings, extracted from a string in the format `"primary carbon, carboxylic acid, carboxylic acid derivative"`, must match the reference set obtained with adaptation of the exmol functional group definitions.

B CHEMSETS Datasets Details

Table A1 lists all tasks included in the CHEMSETS datasets, with information on their symbolic evaluability, task category, number of questions, and expected output type.

Table A1: Overview of tasks in each benchmark dataset. SE = Symbolically evaluable.

	Task	SE	Category[a]	N	Output Type
ChemIQ	SMILES to IUPAC	✓	T	200	String (IUPAC)
	Shortest Path	✓	FC	108	Integer
	Carbon Counting	✓	FC	50	Integer
	Ring Counting	✓	FC	48	Integer
	NMR Elucidation	✗	CG	76	String (SMILES)
	Reaction prediction	✗	R	90	String (SMILES)
	Atom Mapping	✓	C	184	List of integer tuples
	Free-Wilson Analysis	✓	C	40	Float
ether0	IUPAC to SMILES	✓	T	25	String (SMILES)
	Solubility edit	✓	CG	25	String (SMILES)
	SMILES completion	✓	CG	25	String (SMILES)
	Formula to SMILES	✓	CG	15	String (SMILES)
	Functional groups to SMILES	✓	CG	10	String (SMILES)
	Organism Elucidation	✗	CG	25	String (SMILES)
	Reaction prediction	✗	R	25	String (SMILES)
	Retrosynthesis prediction	✗	R	25	String (SMILES)
	Property Selection[b]	✗	C	150	String (SMILES)
SymMolic	SMILES to IUPAC	✓	T	100	String (IUPAC)
	IUPAC to SMILES	✓	T	100	String (SMILES)
	SMILES to Formula	✓	T	100	String (Formula)
	IUPAC to Formula	✓	T	100	String (Formula)
	Alipatic Ring Counting	✓	FC	100	Integer
	Aromatic Ring Counting	✓	FC	100	Integer
	Branch Point Counting	✓	FC	100	Integer
	Bridgehead Counting	✓	FC	100	Integer
	sp3 Carbon Counting	✓	FC	100	Integer
	Fusen Ring Counting	✓	FC	100	Integer
	HBA Counting	✓	FC	100	Integer
	HBD Counting	✓	FC	100	Integer
	Heterocycle Counting	✓	FC	100	Integer
	Largest Ring Size	✓	FC	100	Integer
	Longest Carbon Chain Length	✓	FC	100	Integer
	Rotable Bond Counting	✓	FC	100	Integer
	Spiro Atom Counting	✓	FC	100	Integer
	Stereo Center Counting	✓	FC	100	Integer
	Function Groups Identification	✓	FC	100	String[c]

[a] Categories: **T**: Translation, **FC**: Feature Counting, **CG**: Constrained Generation, **R**: Reactions, **C**: Comparisons.
[b] MCQ across 6 different property categories.
[c] List of functional groups (e.g., `"primary carbon, alcohol"`)

For each question in the CHEMSETS datasets, we compute the Bertz complexity (Bertz, 1981) for the reference molecule. For translation and feature counting tasks, as well as retrosynthesis prediction, the reference molecule is the molecule given in the question. For forward reaction prediction and NMR elucidation tasks, the reference molecule is the correct answer. The constrained generation tasks in **ether0** do not have a single correct answer molecule, but for each question, the authors provide a reference molecule that fulfills all the constraints. The only task excluded is the SMILES completion task from **ether0**, as no valid reference molecule was provided. Figure A2 shows the molecule complexity distribution for each task in CHEMSETS. SymMolic contains, on average, more complex molecules than **ether0** and **ChemIQ**. Each task in SymMolic spans a similar range of molecular complexities. However, the median complexity varies substantially between tasks, ensuring a good representation of different feature counts in each task.

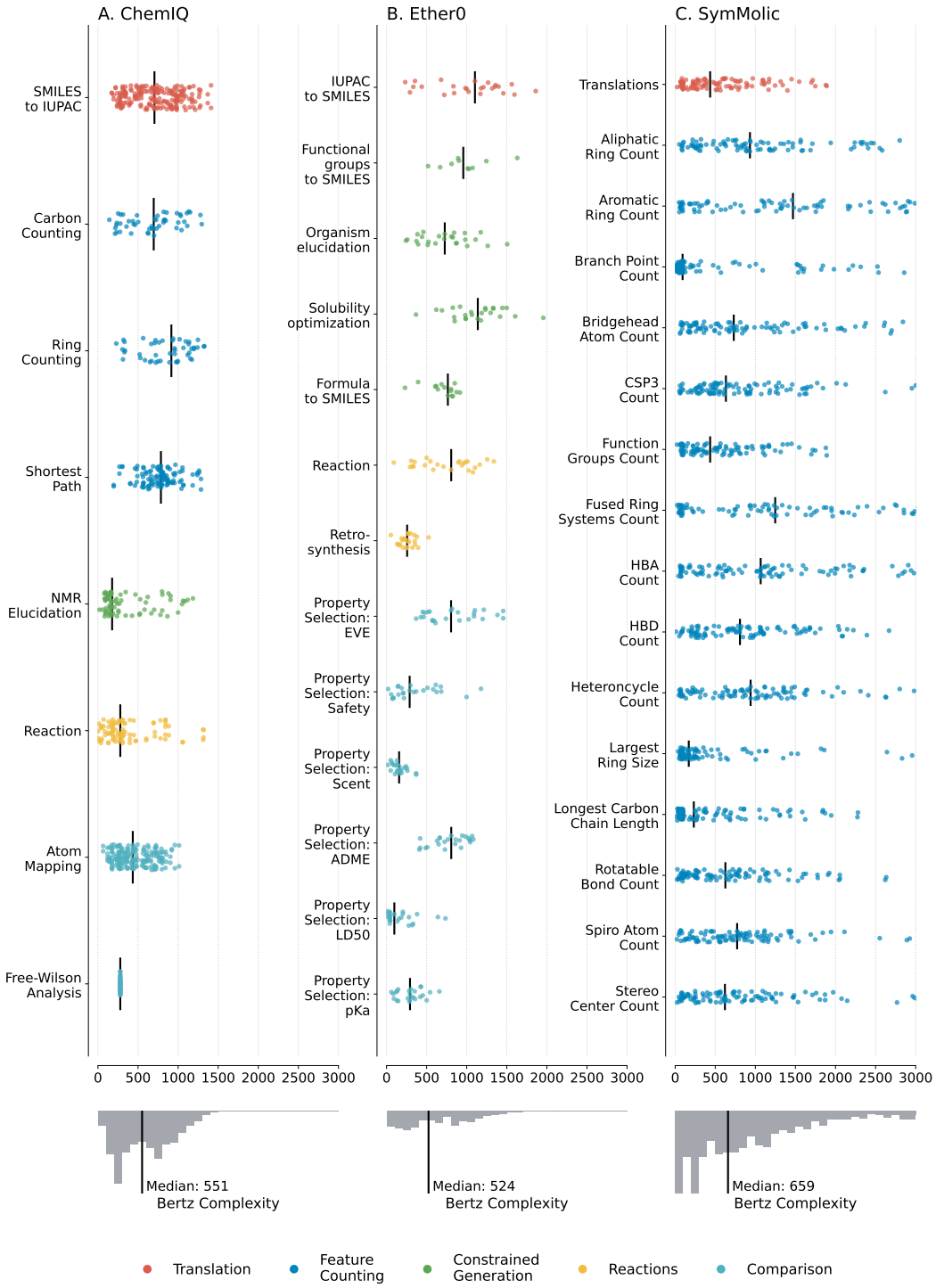

Figure A2: Bertz molecular complexity across the CHEMSETS benchmark sets. Black bars indicate median values per task. All translation tasks in SymMolic share the same set of molecules; their distribution is therefore shown only once under "Translation." Note that SymMolic contains 92 molecules with Bertz complexity above 3000, which are not displayed in the figure. At the bottom, the distributions are aggregated across all tasks within each dataset.

C Models

CHEMSETS is designed to benchmark modern chemistry reasoning models across diverse chemical–understanding tasks. The suite targets instruction-tuned, general-purpose LLMs that can follow free-form prompts and produce verifiable final answers, rather than narrow sequence-to-sequence translators.

C.1 Scientific LLMs

Many chemistry LMs have been trained primarily for molecule-text translation – captioning (`smiles2text`) and text-guided generation (`text2smiles`) – rather than open-ended reasoning (Edwards et al., 2022; Irwin et al., 2022; Zeng et al., 2024). Because these tasks are absent from CHEMSETS and differ substantially from our evaluation format, such models are not well-matched without additional instruction tuning or extraction layers. We nevertheless probed representative systems; for example, ChemLLM (Zhang et al., 2024), trained broadly but evaluated in a strict Q&A style, performed poorly on our open-response prompts and is among the worst performing model on all benchmarks in CHEMSETS (table 1). Consequently, our main comparisons center on instruction-tuned LLMs suitable for free-form reasoning , evaluated with the model-specific prompts, preprocessors and extractors. We also exclude research prototypes without released checkpoints or a stable inference API; for instance, Galactica (Taylor et al., 2022) is not instruction-tuned for our prompt+extractor protocol, precluding a reproducible comparison.

C.2 Generalist LLMs

We restrict our evaluation to **open-weight generalist models**. Model selection was guided by the LiveBench leaderboard as of 15/08/2025 (White et al., 2025). We include two leading model families: (i) **Qwen3**, covering both the largest reasoning-optimized variant (Qwen3-235B-A22B-Thinking-2507) as well as smaller scales (Qwen3-8B, Qwen3-32B, Qwen3-30B-A3B-Thinking-2507), and (ii) the **GPT-oss** series, spanning medium- and high-reasoning variants at different parameter scales (GPT-oss-20B, GPT-oss-120B). All models were executed on our local GPU cluster under unified inference settings.

Closed-source systems (e.g., GPT-5, Claude 3, Gemini 2.5 Pro) were not included, as our benchmark focuses on transparent and fully reproducible evaluation.

C.3 Model Nomenclature

For clarity and consistency throughout this work, we use abbreviated model names. Table A2 provides a comprehensive mapping between our abbreviations, the official model names from their respective publications, and their HuggingFace repository identifiers to ensure reproducibility.

Table A2: Model nomenclature mapping our abbreviations to official names and HuggingFace identifiers.

Abbreviation Used	Model Name	HuggingFace Identifier
Chemistry-Specialized Models		
Llama-molinst	Mol-Instructions	zjunlp/llama3-instruct-molinst-molecule-8b
ChemDFM-8B	ChemDFM	OpenDFM/ChemDFM-v1.0-8B
ChemDFM-13B	ChemDFM	OpenDFM/ChemDFM-v1.0-13B
ChemLLM-7B	ChemLLM	AI4Chem/ChemLLM-7B-Chat
LlaSMol-Mistral	LlaSMol	osunlp/LlaSMol-Mistral-7B
Txgemma-9b	TxGemma	google/txgemma-9b-chat
Txgemma-27b	TxGemma	google/txgemma-27b-chat
Ether0	Ether0	futurehouse/ether0
General-Purpose Models		
Qwen3-8b	Qwen3-8B	Qwen/Qwen3-8B
Qwen3-14b	Qwen3-14B	Qwen/Qwen3-14B
Qwen3-32b	Qwen3-32B	Qwen/Qwen3-32B
Qwen3-Think-30B	Qwen3-30B-A3B	Qwen/Qwen3-30B-A3B-Thinking-2507
Qwen3-Think-235B	Qwen3-235B-A22B	Qwen/Qwen3-235B-A22B-Thinking-2507
GPT-oss-20b-medium*	GPT-oss-20b	openai/gpt-oss-20b
GPT-oss-20b-high*	GPT-oss-20b	openai/gpt-oss-20b
GPT-oss-120b-medium*	GPT-oss-120b	openai/gpt-oss-120b
GPT-oss-120b-high*	GPT-oss-120b	openai/gpt-oss-120b

* Medium/high refer to reasoning effort settings, not separate models

D Extended Results

D.1 Effect of molecule complexity

Figure A3 shows the average accuracy of GPT-oss-120B-high for each task category as a function of the molecule complexity in the question. As we suspected when designing SymMolic, the difficulty of chemical reasoning tasks depends on both the question type and the complexity of the molecule of interest. For each task category, the accuracy decreases with the increasing complexity of the molecules. This demonstrates the risk of drawing conclusions about the difficulty of chemical structure reasoning without controlling for the complexity of the molecules in the questions. It also highlights the usefulness of SymMolic, as it was explicitly designed to cover a wide range of molecule complexities in a balanced manner.

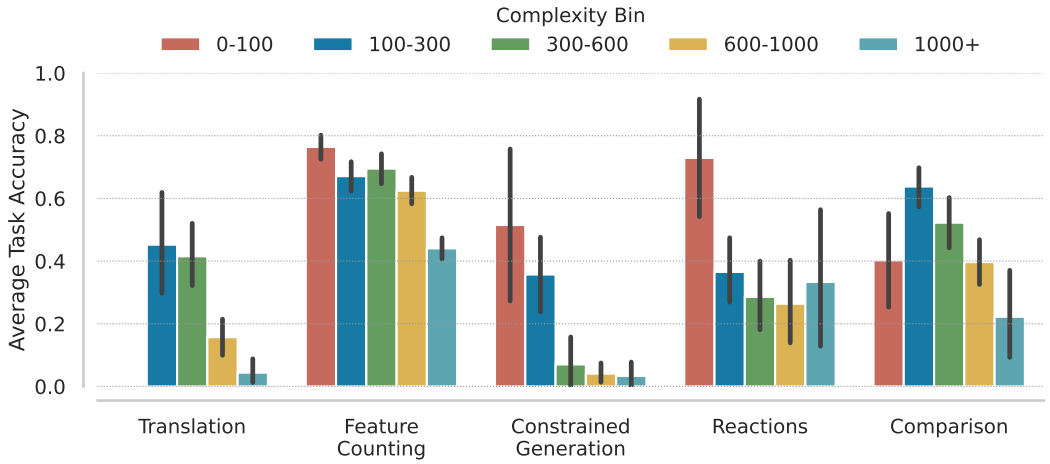

Figure A3: Average accuracy of GPT-oss-120B-high across the different task categories as a function of binned Bertz molecule complexity.

D.2 Results per Tasks

Figures A4, A5, and A6 show the per-task performance of all models, for the ChemIQ, ether0, and SymMolic evaluation sets, respectively.

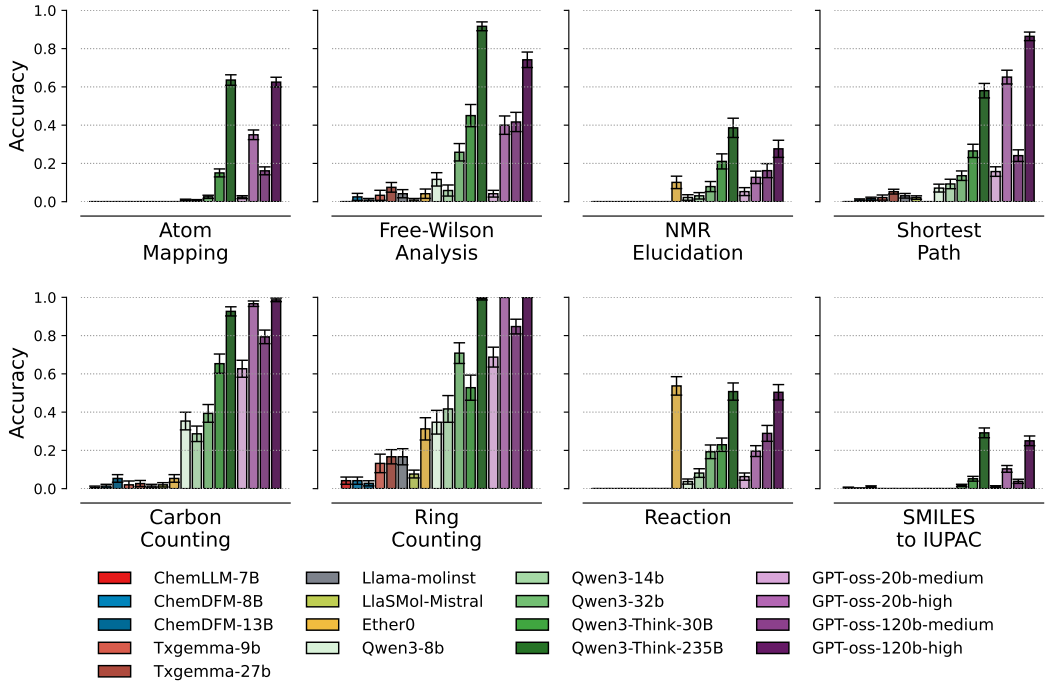

Figure A4: Average accuracy of all models for ChemIQ tasks.

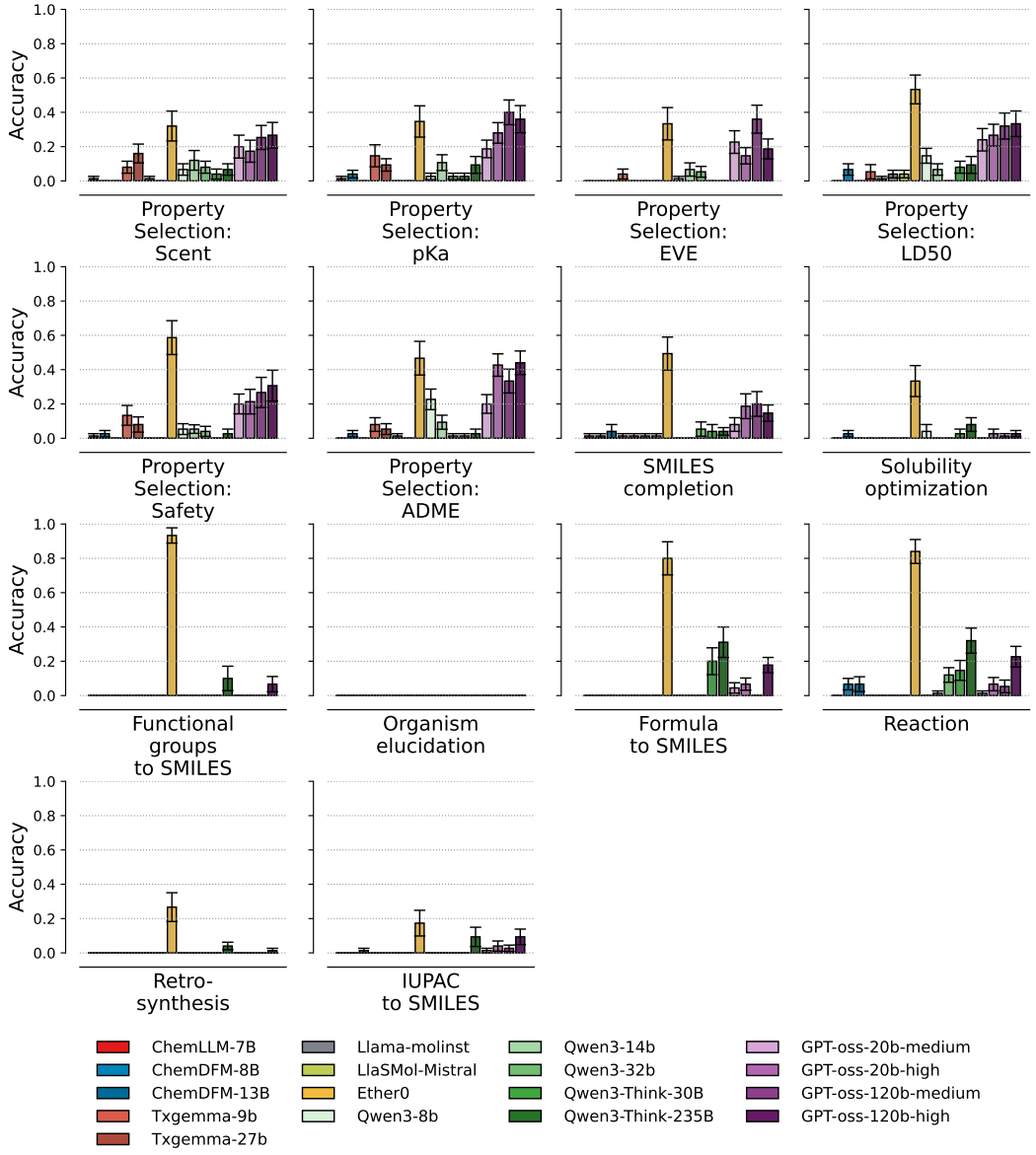

Figure A5: Average accuracy of all models for ether0 tasks.

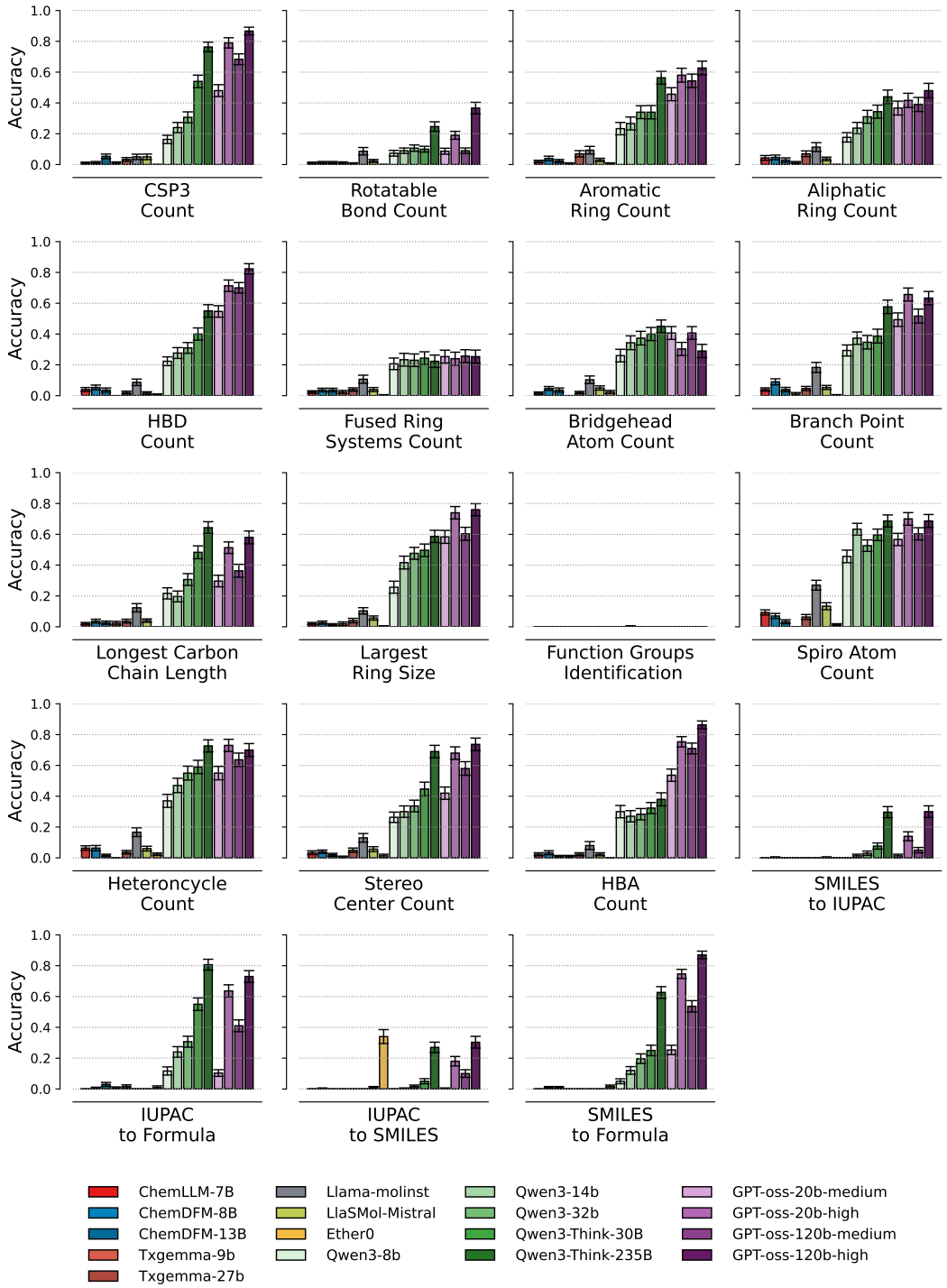

Figure A6: Average accuracy of all models for SymMolic tasks.