Is Nomenclature Beneficial to Language Models for Chemistry?

Anonymous ACL submission

Abstract

Most existing research in language model applications for chemistry employs the Simplified Molecular Input Line Entry System (SMILES) nomenclature, designed to encode molecular structure in a string format as both input and output. In contrast, machine learning approaches using human-readable IUPAC (International Union of Pure and Applied Chemistry) nomenclature remain underexplored. IUPAC names are widely used in the chemical literature, providing opportunities to train large language models on a vast corpus with contextual expertise. We are motivated to compare these two nomenclatures across various language-molecule scenarios. We found that simply switching to IUPAC names in challenging downstream tasks such as molecular generation, captioning, and editing results in a performance improvement of up to 4 times. Additionally, catastrophic forgetting during fine-tuning is reduced by half when using IUPAC names compared to SMILES.

1 Introduction

Research for capturing domain knowledge achievable from the natural language of the vast scientific literature in chemistry is rapidly increasing. Integrating semantic supervision of natural language (Edwards et al., 2021; Liu et al., 2023a; Su et al., 2022; Luo et al., 2023) has been shown to unlock a variety of new capabilities, such as text-based molecule generation and molecule captioning. However, these approaches don't have the conversational capabilities of ChatGPT. Recently, the use of large language models (LLMs) that show impressive reasoning performance in chemistry is a promising research direction(Guo et al., 2024; Bran et al., 2023; Jablonka et al., 2024). Most existing language model applications for chemistry, including molecule generation models(Bagal et al., 2021; Liu et al., 2023c; Dobberstein et al., 2023), used nomenclatures such as SMILES (Weininger, 1988; Weininger et al., 1989), InChI (Heller et al., 2015), and SELFIES (Krenn et al., 2020), which are proposed for computer processing of molecular structures in text. However, these text representations generally do not benefit from large-scale pre-training since they are not widely used in the scientific literature written in natural language.

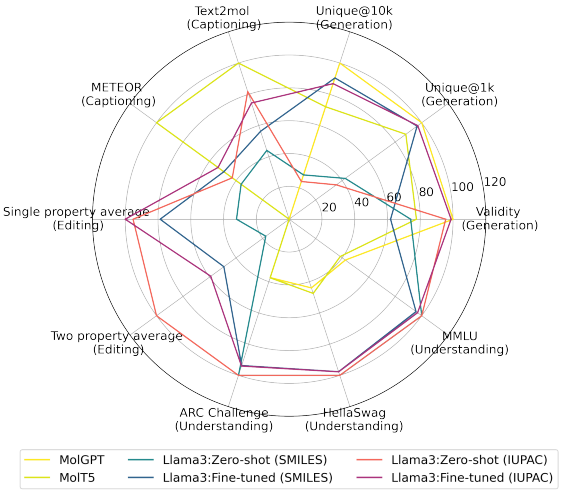


Figure 1: Overall performance of language models on chemical tasks.

The IUPAC nomenclature is a systematic method proposed in 1919 to standardize the naming of compounds and has been used in a variety of literature for a long time. However, IUPAC names for complex compounds are difficult to write and interpret accurately. Therefore, many chemical databases and software rely on SMILES and SELFIES, which are easily processed by computers, as the basic format. Conversion between SMILES and IUPAC was once a challenging problem, but recent advancements in language models have made it easier (Krasnov et al., 2021; Rajan et al., 2021). IUPAC names are widely used in the scientific literature, offering rich learning opportunities from rich corpora with domain knowledge compared to other

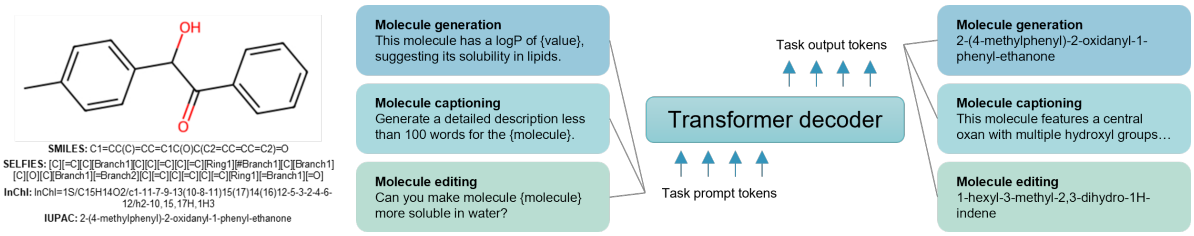


Figure 2: Example of chemical nomenclatures and overview of the experiment process.

nomenclatures. However, the benefits of integrating IUPAC into LLM haven't been fully explored. For example, Guo et al.(Guo et al., 2024) benchmarked several LLMs for various chemistry tasks only for SMILES and SELFIES.

In this study, We report the benefits of using IUPAC nomenclature. When using IUPAC names, LLM performances improved by up to 4 times on challenging tasks such as molecule generation, molecule captioning, and molecule editing (Fig.1). In particular, it learns efficiently from fine-tuning and forgets less during the training process.

2 Experiments

2.1 Tasks

We selected 3 challenging tasks in chemistry to evaluate the benefits of using IUPAC nomenclature in language models: Text-based molecule generation, molecule captioning, and molecule editing. For each task, we compare the zero-shot and fine-tuning performance of LLMs for SMILES and IUPAC. Table 1 contains descriptions, datasets, and metrics for each task. We also evaluate the catastrophic forgetting during training by evaluating benchmarks of LLMs. Detailed evaluation metrics for each task are in Appendix C.

Text-based molecule generation Text-based molecule generation tasks aim to generate candidate molecules with target properties. This task can be divided into value-specific generation and property-specific generation. For example, property-specific generation involves creating molecules that are non-flammable, have a specific color, or have specific functional groups. On the other hand, value-specific generation aims to find molecules that satisfy certain value of properties such as bandgap, logP, and TPSA.

Molecule captioning Molecule captioning aims to write text describing the structure and properties of a given molecule. It requires extracting patterns from given molecular representations and logically linking them by combining pre-trained chemical knowledge from text.

Molecule editing Molecular editing is a recently proposed chemical task. Generating similar molecules with modified properties, rather than creating them from scratch, is cost-effective for the chemical industry. This challenging task requires the ability to estimate a given molecule's properties and make predictions while preserving substructures.

2.2 Model

We performed our experiments by fine-tuning the Llama-3-8B-Instruct (Llama3), the latest variant of Llama family (Touvron et al., 2023a,b), and 3.8B Phi-3-mini-4k-instruct (Phi3) model (Abdin et al., 2024) on NVIDIA RTX 6000 Ada using 4-bit quantization and 8-bit optimizers with the low-rank adaptation (LoRA) technique(Hu et al., 2021). We consider the standard supervised fine-tuning (Dai and Le, 2015; Devlin et al., 2018) paradigm in full parameter space of LLMs. 512 and 0.0001 were used as LoRA rank and learning rate, respectively.

3 Results and discussion

3.1 Molecule generation

To evaluate the LLM's ability to generate molecules according to nomenclatures, we used a pre- and post-fine-tuned model to generate molecules with constraints. The temperature was set to 0.8 to balance the basic probability distribution without being too strict.

We observe that the Llama3 model already demonstrates zero-shot performance in value-specific molecule generation (Table 2). The produced molecules are mostly valid, but the uniqueness of molecules is limited. Novelty cannot be calculated since the exact molecules included in the training corpus of the Llama3 model are unknown. For SMILES, it shows a lower MAD than IUPAC in the zero-shot setting, meaning it

better satisfies the given conditions. We assume that SMILES frequently appear with computable properties like logP in RDKit(RDKit, 2024), allowing related knowledge to be learned during pretraining. After fine-tuning, IUPAC also reaches a similar performance level. Through fine-tuning, we can create more diverse and valid molecules with performance comparable to task-specific language models. Recall that Llamol (Dobberstein et al., 2023) and MolGPT (Bagal et al., 2021) were trained on larger datasets, 13.1M and 1.9M, respectively, whereas in our study they were trained on 60k molecules. When IUPAC is used, the generated molecules are more valid compared to those generated using SMILES. For analysis of property-specific generation, please refer to Appendix E.

3.2 Molecule captioning

Table 4 shows the overall result of molecule captioning. ChEBI-20 (Edwards et al., 2021) test data The text2mol score of the original caption is 0.609. Most previous studies combined pre-trained MolT5 with a pre-trained multimodal encoder and then fine-tuned it on ChEBI-20. These models have high text2mol scores above 0.5. LLMs generate captions using IUPAC or SMILES molecular representations as input, respectively, with and without additional fine-tuning. All models exhibit better metrics when utilizing IUPAC in a zero-shot setting. After fine-tuning, other metrics increase for Llama3, but the Text2mol score slightly decreases, while the Phi3 model approaches the highest metric. Phi3 is trained on synthetic, "textbook-like" data, and llama3 is trained on publicly available documents. We found that using the IUPAC naming system consistently increased captioning performance in all zero-shot settings. For complete metrics and some case studies on the molecule captioning task, readers are referred to Appendix F and Figure 7. Creating expert-level evaluation metrics for chemistry is a challenging and open task. We further discuss the reliability of metrics through a case study of molecule captions generated by Phi3. In zero-shot settings, Phi3 cannot extract meaningful explanations from SMILES patterns. In contrast, when IUPAC is given as input, Phi3 successfully captures the structural information of the first molecule, a trisaccharide structure. Due to LLM's well-known hallucinations, it refers to non-existent esters. After fine-tuning, most of the original knowledge is lost and Phi3 focuses on adhering to the ground truth of the dataset. As a result,

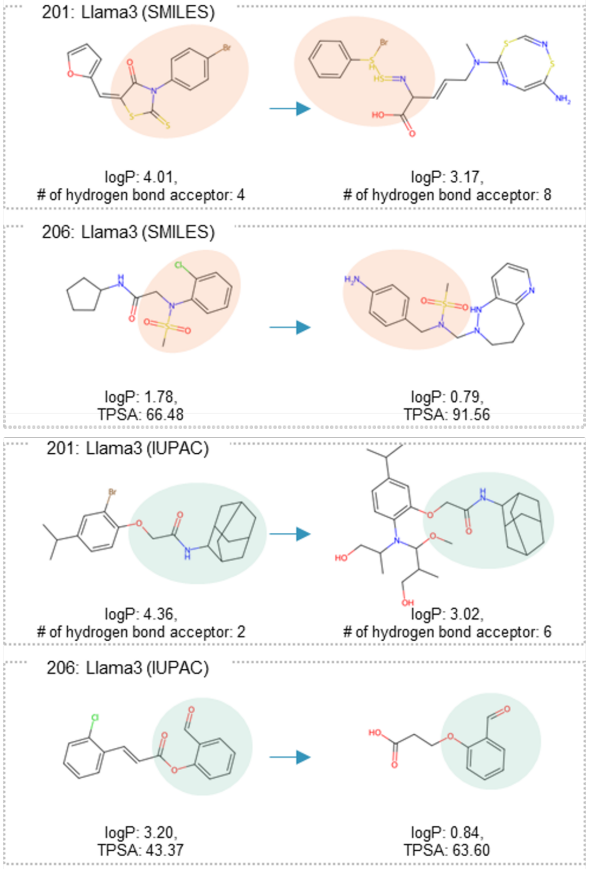


Figure 3: Visualization of molecule editing tasks. Llama3 using IUPAC preserves original substructures better than SMILES.

Phi3 achieves higher metrics due to higher token overlap, but it does not always mean better quality of generated captions.

3.3 Molecule editing

Even for molecular editing tasks, IUPAC consistently achieves higher metrics than SMILES. We illustrate the single- and multi-objective molecule editing results in Tables 5 and 6. If the generated molecule was not valid, it was considered a failure. The editing task itself was not fine-tuned, and instead, a model trained on the value-specific molecule generation task was used. While Llama3 is aware to distinguish hydrogen bond acceptors and donors from IUPAC names in a zero-shot setting, it appears to be unaware of other information such as logP and QED. As a result, it shows excellent performance in tasks 107 and 108 but shows similar performance to the random baseline in other tasks. In the case of the fine-tuned model, which are not trained on the hydrogen bond acceptors and doners, catastrophic forgetting occurs so that performance for them decreases while performance

increases for characteristics such as logP. Even in tasks that modifying two characteristics simultaneously, the zero-shot setting of llama3 using IUPAC shows a high success rate.

In the case study of molecular editing, we can see the advantages of IUPAC that are not apparent through metrics (Fig. 3). When using SMILES, even if editing is successful according to the desired conditions, we observed that the substructure of the original molecule is not preserved and the positions of the elements are mixed up. In contrast, when using IUPAC, the substructure of the molecule is defined in token units, so the detailed structure before and after editing is robustly preserved.

3.4 Catastrophic forgetting

Compared to developing small, specialized models for each task, the advantage of using a general-purpose language model is the flexibility to extend the task to a conversational agent. For example, Liu et al. (Liu et al., 2023b) combined conversational LLM, retrieval DB, and domain feedback to achieve high molecular editing performance by exchanging conversations repeatedly. To maintain these strengths, general-purpose language models must not lose their original knowledge even after they are fine-tuned for specialized tasks. Therefore, we evaluated several challenging NLP benchmarks after performing the molecule generation task in PubChem (Table 7). As a result, we found that using IUPAC achieved higher metrics when trained on the same number of data, thereby damaging the original knowledge less. The more data you train on, the wider the gap becomes. Therefore, when IUPAC is used as an input format, it is possible to maintain the flexibility of a general-purpose language model while achieving higher overall performance in specialized tasks, as shown in Fig. 4.

4 Conclusion

We study the effect of using IUPAC nomenclature for language models on various challenging chemistry tasks. We find that an LLM using IUPAC nomenclature has the following unique advantages for chemistry.

Performance : Although the final performance may converge at the end if sufficient training resources are given, the training cost of LLM is an important aspect of LLM education. IUPAC performs better than SMILES in most tasks when investing the same training resources.

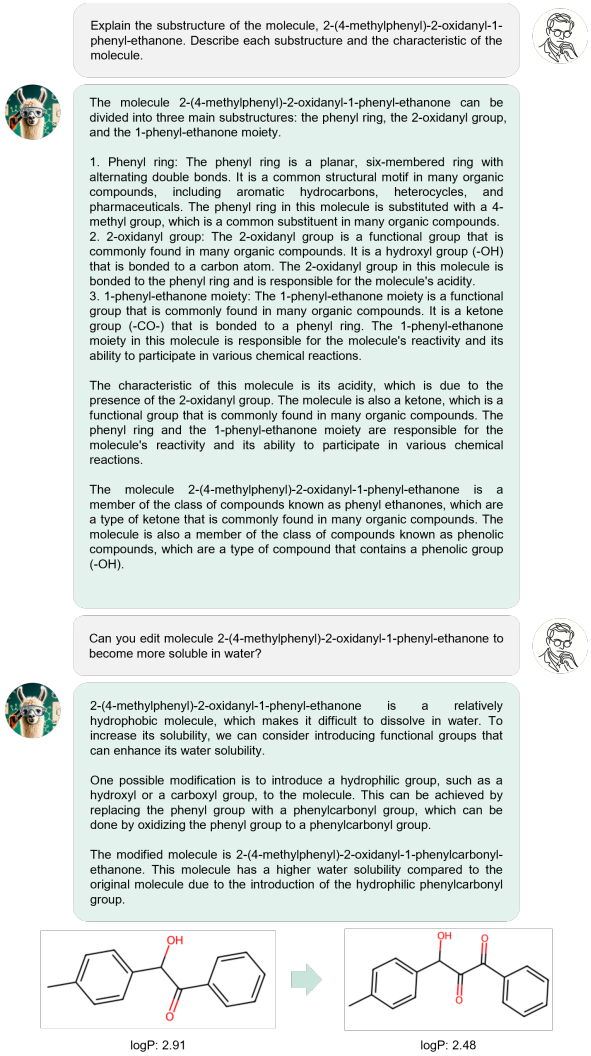


Figure 4: An example of open conversation with Llama3 performing molecule captioning and editing simultaneously.

Data efficiency : Acquiring high-quality labeled molecular data is challenging. According to LLM's scaling raw, general-purpose language models using IUPAC allows for a high level of generalization even with less data.

Accessibility : By using the IUPAC nomenclature, which is closer to natural language, practitioners unfamiliar with computational chemistry can access a vast knowledge base directly and interactively without using domain-expert conversion tools for other molecule representations such as SMILES.

Scalability : Using IUPAC minimizes forgetting, allowing the flexibility of the general-purpose language model to be leveraged for building a variety of specialized task pipelines in chemistry.

5 Limitations

We expect that a large language model trained using IUPAC names will be able to simultaneously perform the task of predicting molecular properties, but unfortunately, most of MoleculeNet's smiles could not be converted to valid IUPAC names, so we did not experiment as a fair comparison was not possible. At the current state of the art, one of the limitations is that the conversion between IUPAC SMILES relies entirely on neural network-based models. In addition, we did not compare the performance under equivalent conditions in which SMILES representation learns information from the surrounding context by controlling the placement of SMILES instead of IUPAC in the pre-training stage. If these transformations are performed properly, it is possible to achieve equal or better performance than IUPAC by learning expert knowledge from the grammar of SMILES. However, considering the cost and complexity of making these changes on several terakens of data, using IUPAC still has its advantages. Another limitation is that the model may be used to discover potentially dangerous molecules instead of beneficial molecules. In particular, molecular editing technologies and captioning capabilities can significantly lower the effort and cost barriers to synthesizing harmful molecules. Despite the above risks, we believe that the benefits to the chemical research community outweigh the disadvantages.

6 Acknowledgments

Y.J.P. was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government, Ministry of Science and ICT (MSIT) (No. 2021R1A6A3A01086766). The 05-Neuron supercomputer was provided by the Korea Institute of Science and Technology Information (KISTI) National Supercomputing Center for Y.J.P.

References

2024. Knowledgator/chemical-converters. Original-date: 2024-02-12T11:31:54Z.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219 [cs].

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. 2021. MolGPT: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076. Publisher: ACS Publications.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint*. ArXiv:2304.05376 [physics, stat].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint*. ArXiv:1803.05457 [cs].

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. *arXiv preprint*. ArXiv:1511.01432 [cs].

Nicola De Cao and Thomas Kipf. 2018. MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Niklas Dobberstein, Astrid Maass, and Jan Hamaekers. 2023. LLamol: A Dynamic Multi-Conditional Generative Transformer for De Novo Molecular Design. *arXiv preprint arXiv:2311.14407*.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280. Publisher: American Chemical Society.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and others. 2024. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36.

Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. 2015. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. 2022. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8867–8887. PMLR. ISSN: 2640-3498.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. 2023. MDM: Molecular Diffusion Model for 3D Molecule Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5105–5112. Number: 4.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pages 1–9. Publisher: Nature Publishing Group UK London.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR.

Lev Krasnov, Ivan Khokhlov, Maxim V. Fedorov, and Sergey Sosnin. 2021. Transformer-based artificial neural networks for the conversion between chemical notations. *Scientific Reports*, 11(1):14798. Publisher: Nature Publishing Group.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024. Publisher: IOP Publishing.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training Molecular Graph Representation with 3D Geometry. *arXiv preprint*. ArXiv:2110.07728 [cs, eess, q-bio].

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2023b. ChatGPT-powered Conversational Drug Editing Using Retrieval and Domain Feedback. *arXiv preprint*. ArXiv:2305.18090 [cs, q-bio].

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023c. MolXPT: Wrapping Molecules with Text for Generative Pre-training. *arXiv preprint*. ArXiv:2305.10688 [cs].

Daniel M. Lowe, Peter T. Corbett, Peter Murray-Rust, and Robert C. Glen. 2011. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling*, 51(3):739–753. Publisher: American Chemical Society.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. MolFM: A Multimodal Molecular Foundation Model. *arXiv preprint*. ArXiv:2307.09484 [physics, q-bio].

Kaushalya Madhawa, Katushiko Ishiguro, Kosuke Nakago, and Motoki Abe. 2019. GraphNVP: An Invertible Flow Model for Generating Molecular Graphs. *arXiv preprint*. ArXiv:1905.11600 [cs, stat].

Jiashun Mao, Jianmin Wang, Kwang-Hwi Cho, and Kyoung Tai No. 2023. iupacGPT: IUPAC-based large-scale molecular pre-trained model for property prediction and molecule generation.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. 2019. MolecularRNN: Generating realistic molecular graphs with optimized properties. *arXiv preprint*. ArXiv:1905.13372 [cs, q-bio, stat].

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741. Publisher: American Chemical Society.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. STOUT: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics*, 13(1):34.

Organization RDKit. 2024. RDKit.

David Rogers and Mathew Hahn. 2010. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754. Publisher: American Chemical Society.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get Your Atoms in Order—An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *Journal of Chemical Information and Modeling*, 55(10):2111–2120. Publisher: American Chemical Society.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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. Publisher: ACS Publications.

David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences*, 29(2):97–101. Publisher: ACS Publications.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation. *arXiv preprint*. ArXiv:2203.02923 [cs, q-bio].

Chengxi Zang and Fei Wang. 2020. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 617–626, New York, NY, USA. Association for Computing Machinery.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862. Publisher: Nature Publishing Group UK London.

A Related works

A.1 Text-based molecular representation learning

Research into controlling molecules using natural language has recently been attracting attention. Molecules can be broadly divided into two modalities(Zeng et al., 2022): internal information represented by SMILES(Weininger, 1988; Weininger et al., 1989), a 2D or 3D graph representation that describes the structure of the molecule, and text-based explanation that describes external information such as the functional characteristics of the molecule.

KV-PLM(Zeng et al., 2022) applies BERT-based masked token prediction training to heterogeneous

data consisting of SMILES strings and biomedical text descriptions. Text2Mol(Edwards et al., 2021) performs cross-modality search by performing contrastive learning between molecular graphs and text data. MolT5(Edwards et al., 2022) was trained to perform translation between SMILES and text annotations of molecules. MoMu(Su et al., 2022) showed that the contrast learning model between modalities could be extended to molecular caption writing and molecule generation tasks by introducing an additional projection layer and connecting it with pre-trained models such as MolT5 and MoFlow(Zang and Wang, 2020) . MoleculeSTM(Liu et al., 2023a) has also been extended to zero-shot text-based molecular editing tasks based on a pre-trained contrastive learning model. This work demonstrates the potential of LLMs for more realistic drug discovery tasks.

A.2 Text-based molecular generation

Various approaches have been attempted to create a molecule generation model depending on the modality of the molecule. In graph-based models, conditional generation to design molecules with desired properties is challenging. JT-VAE(Jin et al., 2018) based on molecular graph generates molecules in two iterative steps, utilizing Bayesian optimization for conditional generation. MolGAN(De Cao and Kipf, 2018) is an implicit, likelihood-free generative model for small molecular graphs that uses GANs on graph-structured data. This uses reinforcement learning to find molecule with desired properties. Flow-based models such as GraphNVP(Madhawa et al., 2019) and MoFlow(Zang and Wang, 2020) learn the molecule generation process through mapping to an invertible latent space. Optimizations along the latent space can be used to generate molecular graphs with specific desired properties without any expert/domain knowledge. Diffusion-based generation models that have been actively studied recently mainly focus on 3D molecule generation(Xu et al., 2022; Hoogeboom et al., 2022; Huang et al., 2023).

SMILES-based autoregressive molecule generation models have also been actively studied. MolecularRNN(Popova et al., 2019) sequentially generates each character of SMILES. MolGPT(Bagal et al., 2021) performs on par with other previously proposed modern machine learning frameworks for molecular generation in terms of generating valid, unique, and novel molecules. MolXPT(Liu et al., 2023c) detect the molecule names in each sequence and replace them with the corresponding SMILES. Llamol(Dobberstein et al., 2023) trains a 15 million parameter model that is modified from the Llama-2(Touvron et al., 2023b) architecture to generate a SMILES representation that satisfies given characteristics. iupacGPT(Mao et al., 2023), learned from 97M molecules, showed an equivalent level of molecule generation ability to SMILES using IUPAC names instead of SMILES.

The most closely related work to ours is (Hernandez et al., 2021) which explored the scaling for knowledge transfer by comparing finetuning with training from scratch. Our study is orthogonal to theirs with significant differences as our key focus is understanding the scaling of different factors for LLM finetuning, rather than the transfer.

B Training data

PubChem We downloaded 1 million molecules from PubChem. We cleaned the data according to several conditions.

- 1. Structures that RDKit could not parse were removed.
- 2. Limited to molecules with a total charge of 0.
- 3. The number of heavy atoms is limited to 30 or less (This represents approximately 75% of the total).

After this process, approximately 0.6 million, or 591,575 molecules remained. Afterward, logP, SA score, QED, TPSA, and molecular weight were calculated from the SMILES representations using RDKit.

ChEBI-20 We use ChEBI-20 (Edwards et al., 2021) as a training dataset for text-based property-specific molecule generation and molecule captioning. This dataset consists of 33,010 molecules with SMILES, IUPAC, and their description. We separate it into 80/10/10 train/validation/test splits, respectively.

C Evaluation

C.1 Molecule generation

We measured the following metrics to evaluate the performance of the molecule generation task. All metrics of value-specific molecule generation were calculated statistically after generating 10k molecules.

Task	Task type	Fine-tuning	Dataset	Metrics
Text-based molecule design (Value-specific)	Generation	Fine-tuning	PubChem	validity, uniqueness, etc.
Text-based molecule design (property-specific)	Generation	Fine-tuning	ChEBI-20	validity, uniqueness, etc.
Molecule editing	Generation	Zero-shot	PubChem	Success rate
Molecule captioning	Generation	Fine-tuning/zero-shot	ChEBI-20	BLEU, Text2mol score, etc.

Table 1: The statistics of tasks, datasets, the number of samples, and evaluation metrics

- Validity: the fraction of generated molecules that are valid. We use RDkit for the validity check of molecules. Validity measures how well the model has learned the SMILES grammar and the valency of atoms.

- Uniqueness: the fraction of validly generated molecules that are unique. Low uniqueness highlights repetitive molecule generation and a low level of distribution learning by the model.

- Novelty: the fraction of valid unique generated molecules that are not in the training set. Low novelty is a sign of overfitting. We do not want the model to memorize the training data.

- Mean absolute deviation (MAD): the deviation between property values of generated molecules and the given target property value. The lower MAD indicates a better understanding of the connection between the properties and the molecule.

In general, novelty is recommended to be measured as it is an indicator of overfitting that determines whether the model remembers the data. However, checking for duplicates across hundreds of thousands of training data pools would be an overwhelming effort. Additionally, in the zero-shot setting, it is not possible to determine what data the model was exposed to during pre-training, so it was not measured in this study.

In property-specific molecule generation, we measure the similarity between the generated molecule and the original molecule and the similarity between the description text and the generated molecule using the following metrics.

- String similarity: BLEU (Papineni et al., 2002), Exact, and Levenshtein distance(Miller et al., 2009) are used to measure whether accuracy by comparing the strings of generated molecules.

- Molecular feature similarity: MACCS FTS(Durant et al., 2002), RDK FTS(Schneider et al., 2015), Morgan FTS(Rogers and Hahn, 2010), and FCD(Preuer et al., 2018) measure similarity by comparing the features of the generated molecule and the original molecule.

- String-Molecule similarity: Text2mol score is designed to measure the similarity between the text description and the molecule by comparing the latent representation of each branch of the pre-trained multimodal model.

C.2 Molecule captioning

Molecule captioning. We utilize the ChEBI-20 (Edwards et al., 2021) dataset with 33,010 molecule-description pairs. We follow the original 8:1:1 train/validation/test split. Evaluation metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for string-similarity and Text2Mol score (Edwards et al., 2021) for text-molecule similarity

C.3 Molecule editing

Measure the success rate of introducing new molecules that satisfy predefined properties from given molecules. Generating an invalid molecule is considered a failure. We wanted to use MoleculeSTM as a baseline, which proposed this task, but since their dataset does not support IUPAC, so we experimented with 200 molecules randomly selected from PubChem.

C.4 Catastrophic forgetting

We measured ARC challenge(Clark et al., 2018), HellaSwag(Zellers et al., 2019), and MMLU(Hendrycks et al., 2021), which are benchmarks for measuring the comprehensive performance of LLM across extensive tasks before and after training, to quantify forgotten knowledge during fine-tuning.

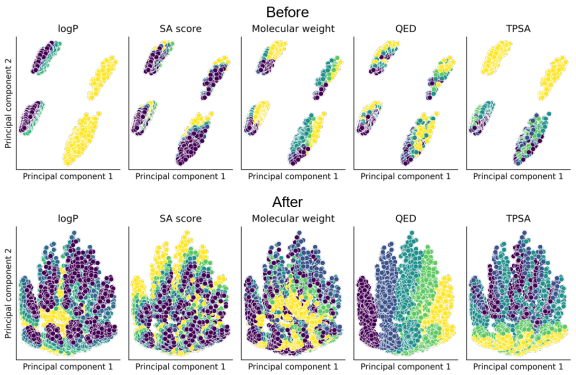


Figure 5: Visualization of hidden representations of prompts containing target properties. Five values were selected for each properties, and different colors were assigned according to the values.

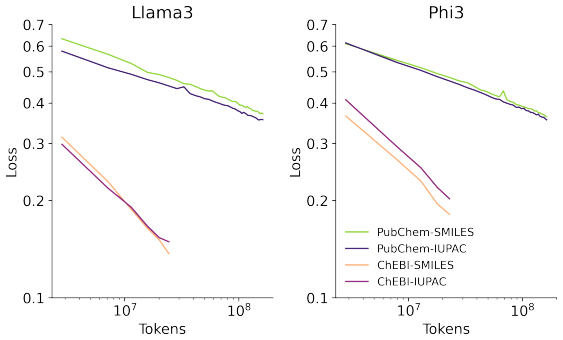


Figure 6: Learning dynamics of large language models with different nomenclature.

D Visualization of prompt representation

After training, the hidden representation of the molecule generation condition prompt is visualized in Fig. 5 using principal component analysis (PCA). This contains probability information for generating the next molecular token based on the given condition and thus represents the interface between the text representation of the condition and the molecular representation. Prompts for visualization were selected and assembled from a pool of five values for each property condition, therefore a total of $5^5 = 3125$.

We assume that the model before training tends to generate specific tokens regardless of the generation conditions, which leads to low uniqueness since the hidden representation of the model before training is strongly clustered. On the other hand, after training, the model sorts the molecular feature conditions according to their value terms and tries to generate more diverse tokens.

E Learning dynamics of LLMs

While conducting fine-tuning to evaluate property-specific molecule generation performance, we found that the learning dynamics were significantly different between training datasets, ChEBI-20 and PubChem. The starting loss on ChEBI-20 data is lower than 0.6 and decreases much faster than value-specific molecule generation. Molecules in ChEBI-20 data contain a wider variety of special symbols such as '/' and '@' than PubChem's typical SMILES to describe their three-dimensional stereochemical structures. This may cause the overall difficulty of the dataset to increase with the addition of more complex symbols or cause the difficulty to decrease with the inclusion of conformational information. However, a starting loss of 0.6 or less has approximately a 55% chance of correctly predicting the next token, according to negative log-likelihood calculations. Therefore, we believe that we cannot rule out the possibility that the pre-training data of LLM contains ChEBI-20 data.

F Examples of generated captions

We provide examples of molecule captions generated by Phi3 model.

G Performance metrics

This section contains Tables of metrics for all tasks studied in this paper. In the case of molecule generation work using IUPAC, due to limitations in tools that can directly parse IUPAC names, the metric was evaluated after conversion to SMILES. Although there are relatively accurate conversion tools such as OPSIN(Lowe et al., 2011), the IUPAC to SMILES conversion took too much time, so we used a neural network-based translation tool(noa, 2024) capable of GPU parallel calculation. This conversion tool was evaluated on 50k molecules extracted from PubChem and achieved an accuracy of 99.7%, justifying its usage. What is impressive is that failure in the conversion process also shows high validity even though it was considered a failure in validation.

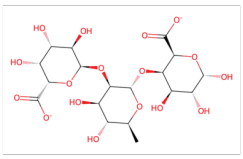
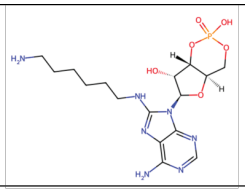
Molecule		
MolFM	The molecule is a polyanionic polymer obtained by deprotonation of the carboxy groups of [(2R,3S)-4-deoxy-alpha-L-rhamnosyl-(1->4)-alpha-D-galacturonosyl]-L-iduronic acid; major species at pH 7.3. It is a polyanionic polymer and a carbohydrate acid derivative anion.	The molecule is a 3',5'-cyclic purine nucleotide that is 3',5'-cyclic AMP in which the hydrogen at position 2 on the purine fragment is replaced by a 6-aminohexylamino group. It is a 3',5'-cyclic purine nucleotide, an adenyl ribonucleotide, a primary amino compound and an aryl amine. It derives from a 3',5'-cyclic AMP.
Ground Truth	The molecule is a polyanionic polymer obtained by deprotonation of the carboxy groups of alpha-D-galacturonosyl-[(1->2)-alpha-L-rhamnosyl-(1->4)-alpha-D-galacturonosyl]n; major species at pH 7.3. It is a carbohydrate acid anion and a polyanionic polymer.	The molecule is a 3',5'-cyclic purine nucleotide that is 3',5'-cyclic AMP in which the hydrogen at position 2 on the purine fragment is replaced by a 6-aminohexylamino group. It is a 3',5'-cyclic purine nucleotide, an adenyl ribonucleotide, a primary amino compound and a secondary amino compound. It derives from a 3',5'-cyclic AMP.
Phi3: Fine-tuned (IUPAC)	The molecule is a trisaccharide derivative that is the carbohydrate acid derivative anion formed by deprotonation of the carboxy groups of 3-deoxy-alpha-D-manno-oct-2-ulosonic acid; major species at pH 7.3. It is a conjugate base of a 3-deoxy-alpha-D-manno-oct-2-ulosonic acid. It is a trisaccharide derivative and a carbohydrate acid derivative anion. It is a conjugate base of a 3-deoxy-alpha-D-manno-oct-2-ulosonic acid.	The molecule is a 3',5'-cyclic purine nucleotide that is 3',5'-cyclic AMP in which the hydrogen at position 2 on the purine fragment is replaced by a 6-aminohexylamino group. It has a role as a protein kinase agonist. It is a 3',5'-cyclic purine nucleotide, a primary amino compound and a primary aminoalkylamino compound. It derives from a 3',5'-cyclic AMP.
Phi3: Zero-shot (SMILES)	This molecule is a pharmaceutical compound, specifically a derivative of a flavonoid, with a complex polyhydroxy structure. It features a central carbon ring, interconnected with additional rings and hydroxyl groups. The molecule possesses multiple hydroxyl groups, indicating potential antioxidant properties. It may serve as a precursor for various pharmaceutical agents due to its structural complexity and potential biological activity.	This molecule, a pharmaceutical derivative, features a complex, bicyclic structure with a quinazoline core. It possesses a tertiary amine, a phosphate ester, and a quaternary ammonium group. The compound exhibits potential as a diuretic and antihypertensive agent, derived from the quinazoline scaffold.
Phi3: Zero-shot (IUPAC)	This molecule is a complex carbohydrate featuring multiple hydroxyl groups and a carboxylate ester functional group. It possesses a symmetrical structure with stereochemistry at each chiral center, resulting in a specific three-dimensional arrangement. The molecule is a derivative of a trisaccharide, with a glucose unit linked to a galactose unit via a 1,2-beta-glycosidic bond. The presence of multiple hydroxyl groups and the carboxylate ester group suggest potential roles in energy storage and structural support in biological systems.	This molecule is a complex tetrahydrofuro[3,2-d][1,3,2]dioxaphosphinin with a fused furofuran ring. It features a purine base, an amino group, and a hydroxyl group. The stereochemistry indicates chiral centers at carbon 4, 6, and 7, with specific configurations. The molecule's structure suggests potential roles in biochemical processes, possibly as a phosphate-containing compound in nucleic acid metabolism. Derived forms may include salts or esters, depending on the functional groups' interactions with other molecules.

Figure 7: Examples of molecule captioning. We highlight segments included in the ground truth in red, information not included in the ground truth in blue, and information misunderstood in purple.

Condition Type	Interval	Model	Novelty [%] ↑	Uniqueness @ 1k [%] ↑	Uniqueness [%] ↑	Validity [%] ↑	MAD ↓
Unconditional	-	Llamol	97.58	100.0	100.0	99.49	
	-	MolGPT	79.7	100.0	100.0	99.4	
	-	Llama3:Zero-shot (SMILES)	-	25.30	14.95	87.90	
	-	Llama3:Zero-shot (IUPAC)	-	**42.70**	**26.47**	**97.29**	
	-	Llama3:Fine-tuned (SMILES)	**100.0**	96.70	**90.14**	60.79	
	-	Llama3:Fine-tuned (IUPAC)	99.95	**97.70**	87.23	**97.94**	
	-	Phi3:Zero-shot (SMILES)	-	**34.1**	**21.08**	64.56	
	-	Phi3:Zero-shot (IUPAC)	-	28.20	15.82	**71.88**	
	-	Phi3:Fine-tuned (SMILES)	**98.37**	31.3	**20.29**	63.67	
	-	Phi3:Fine-tuned (IUPAC)	96.59	**29.40**	15.69	**70.94**	
LogP	{2, 4, 6}	Llamol	97.45	100.0	99.82	99.61	0.194
	{2, 4, 6}	MolGPT	100.0	99.8	99.8	97.1	0.23
	{2, 4, 6}	Llama3:Zero-shot (SMILES)	-	52.00	37.26	81.74	**0.73**
	{2, 4, 6}	Llama3:Zero-shot (IUPAC)	-	**57.27**	**40.16**	**96.45**	2.85
	{2, 4, 6}	Llama3:Fine-tuned (SMILES)	**100.0**	**96.27**	**90.15**	65.84	**0.66**
	{2, 4, 6}	Llama3:Fine-tuned (IUPAC)	99.94	95.40	85.74	**97.2**7	0.86
	{2, 4, 6}	Phi3:Zero-shot (SMILES)	-	**54.77**	**37.68**	57.25	**0.96**
	{2, 4, 6}	Phi3:Zero-shot (IUPAC)	-	23.87	11.34	**97.30**	1.82
	{2, 4, 6}	Phi3:Fine-tuned (SMILES)	**99.95**	**53.10**	**38.12**	57.36	**0.94**
	{2, 4, 6}	Phi3:Fine-tuned (IUPAC)	97.63	23.53	11.52	**97.19**	1.85
SAScore	{2, 3, 4}	Llamol	97.41	100.0	99.94	99.70	0.099
	{2, 3, 4}	MolGPT	97.0	100.0	99.5	97.7	0.13
	{2, 3, 4}	Llama3:Zero-shot (SMILES)	-	**38.73**	**26.28**	**86.71**	**0.78**
	{2, 3, 4}	Llama3:Zero-shot (IUPAC)	-	20.93	16.01	65.95	1.66
	{2, 3, 4}	Llama3:Fine-tuned (SMILES)	**99.96**	**95.30**	**90.66**	70.43	0.59
	{2, 3, 4}	Llama3:Fine-tuned (IUPAC)	99.93	94.83	87.87	**96.68**	**0.54**
	{2, 3, 4}	Phi3:Zero-shot (SMILES)	-	**48.47**	**32.33**	66.98	0.86
	{2, 3, 4}	Phi3:Zero-shot (IUPAC)	-	11.17	5.21	**80.65**	**0.81**
	{2, 3, 4}	Phi3:Fine-tuned (SMILES)	**99.61**	**47.63**	**32.04**	67.57	0.86
	{2, 3, 4}	Phi3:Fine-tuned (IUPAC)	94.45	11.43	5.17	**80.39**	**0.78**

Table 2: Value-specific molecule generation performance metrics for various models. Comparing IUPAC and SMILES, the better one is written in bold.

Model	**BLEU ↑**	**Exact ↑**	**Levenshtein ↓**	**MACCS FTS ↑**	**RDK FTS ↑**	**Morgan FTS ↑**	**FCD ↓**	**Text2Mol ↑**	**Validity ↑**
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	**0.652**	**0.005**	**38.09**	**0.591**	**0.400**	**0.362**	**4.55**	**0.409**	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	**0.906**
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	2.89	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	2.49	0.482	0.721
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
T5-Large	**0.854**	0.279	16.721	0.823	0.731	0.670	1.22	0.552	0.902
MolT5-Large	**0.854**	**0.311**	**16.071**	**0.834**	**0.746**	**0.684**	**1.20**	**0.554**	**0.905**
Llama:Zero-shot (SMILES)	0.322	0.003	59.75	0.573	0.316	0.275	19.40	0.387	0.214
Llama:Zero-shot (IUPAC)	0.230	0.011	63.33	0.440	0.256	0.204	**19.00**	0.256	0.863
Llama3:Fine-tuned (SMILES)	**0.688**	**0.075**	**37.13**	**0.798**	**0.606**	**0.550**	20.00	**0.547**	0.652
Llama3:Fine-tuned (IUPAC)	0.362	0.055	47.37	0.698	0.520	0.430	19.44	0.462	**0.891**
Phi3:Zero-shot (SMILES)	0.256	0.001	65.69	0.439	0.206	0.159	**16.55**	0.268	0.243
Phi3:Zero-shot (IUPAC)	0.212	0.003	67.68	0.394	0.219	0.150	20.00	0.194	0.842
Phi3:Fine-tuned (SMILES)	**0.554**	0.017	**51.55**	**0.696**	**0.480**	**0.416**	19.81	**0.480**	0.510
Phi3:Fine-tuned (IUPAC)	0.314	**0.025**	55.58	0.572	0.370	0.280	19.86	0.355	**0.863**

Table 3: Property-specific molecule generation performance metrics for various models on different metrics.

Decoder	Encoder	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
MolT5-small	MolT5-small (Edwards et al., 2022)	0.519	0.436	0.620	0.469	0.563	0.551	0.540
	MoMu (Su et al., 2022)	0.532	0.445	0.621	0.469	0.564	0.557	0.543
	GraphMVP (Liu et al., 2022)	0.540	0.449	0.619	0.465	0.560	0.562	0.553
	MolFM (Luo et al., 2023)	0.542	0.452	0.623	0.469	0.562	0.564	0.557
MolT5-base	MolT5-base (Edwards et al., 2022)	0.540	0.457	0.634	0.485	0.578	0.569	0.547
	MoMu (Su et al., 2022)	0.549	0.462	0.630	0.479	0.575	0.576	0.558
	GraphMVP (Liu et al., 2022)	0.577	0.491	0.651	0.505	0.592	0.599	0.570
	MolFM (Luo et al., 2023)	0.585	0.498	0.653	0.508	0.594	0.607	0.576
Llama-3:Zero-shot (SMILES)	-	0.104	0.025	0.253	0.058	0.171	0.206	0.241
Llama-3:Zero-shot (IUPAC)	-	**0.140**	**0.049**	**0.276**	**0.069**	**0.171**	**0.244**	**0.447**
Llacha-3:Fine-tuned (IUPAC)	-	0.290	0.188	0.410	0.222	0.343	0.317	0.407
Phi-3:Zero-shot (SMILES)	-	0.080	0.013	0.214	0.039	0.147	0.172	0.225
Phi-3:Zero-shot (IUPAC)	-	**0.131**	**0.044**	**0.261**	**0.056**	**0.167**	**0.220**	**0.453**
Phi-3:Fine-tuned (IUPAC)	-	0.316	0.250	0.461	0.321	0.403	0.511	0.569
GPT-3.5-turbo (SMILES)	-	0.102	0.028	0.217	0.051	0.155	0.165	0.336
GPT-3.5-turbo (IUPAC)	-	**0.125**	**0.048**	**0.245**	**0.059**	**0.163**	**0.221**	**0.451**
GPT-4o (SMILES)	-	0.093	0.021	0.215	0.039	0.139	0.180	0.434
GPT-4o (IUPAC)	-	**0.133**	**0.052**	**0.257**	**0.056**	**0.161**	**0.239**	**0.488**

Table 4: Molecule captioning results on the test split of ChEBI-20.

Single Target Properties	Δ	**Random (MoleculeSTM)**	**MoleculeSTM (SMILES)**	**MoleculeSTM (Graph)**	**Random (Ours)**	**Llama3:Zero-shot (SMILES)**	**Llama3:Zero-shot (IUPAC)**	**Llama3:Fine-tuned (SMILES)**	**Llama3:Fine-tuned (IUPAC)**
101 more soluble in water	0	35.33 ± 1.31	61.87 ± 2.47	**67.86 ± 4.37**	49.67 ± 3.33	24.37 ± 7.30	50.15 ± 4.70	50.17 ± 3.86	**61.55 ± 3.21**
	0.5	11.04 ± 2.40	49.02 ± 1.84	**54.44 ± 3.99**	38.83 ± 2.02	18.79 ± 5.47	41.07 ± 5.03	42.26 ± 5.00	**48.77 ± 3.28**
102 less soluble in water	0	43.36 ± 3.06	52.71 ± 1.67	**64.79 ± 2.76**	**50.17 ± 2.75**	20.15 ± 2.49	46.67 ± 5.77	29.25 ± 4.63	34.81 ± 0.79
	0.5	19.75 ± 1.56	47.17 ± 1.37	**48.70 ± 2.04**	**36.83 ± 2.47**	12.90 ± 4.00	30.93 ± 7.98	21.73 ± 2.83	25.36 ± 2.23
103 more like a drug	0	38.06 ± 2.57	36.52 ± 2.46	**39.97 ± 4.32**	**49.83 ± 1.61**	10.24 ± 6.80	44.39 ± 18.53	25.13 ± 2.35	35.81 ± 2.42
	0.1	5.27 ± 0.24	8.11 ± 0.82	**14.06 ± 3.18**	**33.83 ± 1.26**	6.56 ± 7.21	29.61 ± 21.78	15.08 ± 2.85	24.09 ± 1.12
104 less like a drug	0	36.96 ± 2.25	58.59 ± 1.10	**77.62 ± 2.80**	50.50 ± 1.32	13.53 ± 8.98	42.09 ± 5.10	49.77 ± 4.55	**60.26 ± 1.01**
	0.1	6.16 ± 1.87	11.55 ± 0.90	**54.22 ± 3.01**	33.50 ± 3.50	10.52 ± 9.37	33.28 ± 7.24	38.36 ± 2.05	**45.59 ± 2.20**
105 higher permeability	0	25.23 ± 2.13	**61.87 ± 1.76**	59.84 ± 0.78	44.17 ± 0.76	7.45 ± 5.50	15.87 ± 16.72	43.50 ± 1.80	**51.09 ± 1.90**
	10	17.41 ± 1.43	47.45 ± 1.88	**50.42 ± 2.73**	32.17 ± 1.89	5.93 ± 7.10	9.21 ± 7.99	34.00 ± 2.29	**38.66 ± 3.90**
106 lower permeability	0	16.79 ± 2.54	31.76 ± 0.97	**40.35 ± 1.87**	**52.67 ± 4.51**	8.94 ± 4.06	37.78 ± 10.72	29.77 ± 1.61	49.80 ± 6.09
	10	11.02 ± 0.71	29.37 ± 0.96	**31.71 ± 1.47**	**41.83 ± 2.02**	5.24 ± 2.48	21.48 ± 11.18	21.41 ± 3.22	40.52 ± 5.10
107 more hydrogen bond acceptors	0	12.64 ± 1.64	34.52 ± 5.26	**37.35 ± 7.09**	45.17 ± 2.84	19.02 ± 2.67	**76.19 ± 21.82**	32.61 ± 4.18	40.35 ± 6.90
	1	6.09 ± 0.01	16.13 ± 1.62	**16.13 ± 7.63**	32.67 ± 1.89	12.42 ± 7.38	**35.71 ± 18.90**	22.41 ± 3.08	26.40 ± 5.29
108 more hydrogen bond donors	0	2.97 ± 0.61	3.00 ± 0.86	**7.69 ± 0.56**	35.00 ± 4.36	13.92 ± 4.66	**55.93 ± 16.04**	25.80 ± 1.28	31.41 ± 1.47
	1	0.00 ± 0.00	1.00 ± 0.86	**3.23 ± 5.27**	10.67 ± 2.31	10.55 ± 2.05	**24.37 ± 10.72**	10.29 ± 2.03	10.15 ± 5.31

Table 5: Results on single-objective molecule editing are evaluated based on the hit ratio of the property change.

Two Target Properties	Δ	**Random (MoleculeSTM)**	**MoleculeSTM (SMILES)**	**MoleculeSTM (Graph)**	**Random (Ours)**	**Llama3:Zero-shot (SMILES)**	**Llama3:Zero-shot (IUPAC)**	**Llama3:Fine-tuned (SMILES)**	**Llama3:Fine-tuned (IUPAC)**
201 more soluble in water and more hydrogen bond acceptors	0 – 0	9.88 ± 1.03	**27.87 ± 3.86**	27.43 ± 3.41	31.50 ± 0.50	8.76 ± 4.83	**60.83 ± 1.44**	24.59 ± 5.07	33.14 ± 4.37
	0.5 – 1	0.23 ± 0.33	8.80 ± 0.04	**11.10 ± 1.80**	17.50 ± 3.61	6.23 ± 3.57	**45.83 ± 12.33**	15.56 ± 3.05	18.82 ± 4.52
202 less soluble in water and more hydrogen bond acceptors	0 – 0	2.99 ± 0.38	**8.55 ± 2.75**	8.21 ± 0.81	17.83 ± 3.01	5.18 ± 5.86	**23.28 ± 9.57**	8.19 ± 2.01	12.06 ± 2.29
	0.5 – 1	0.22 ± 0.31	2.93 ± 0.30	**3.10 ± 0.32**	**8.83 ± 1.15**	1.28 ± 2.22	7.41 ± 6.42	3.01 ± 2.64	4.48 ± 1.38
203 more soluble in water and more hydrogen bond donors	0 – 0	2.28 ± 1.15	33.51 ± 4.08	**49.23 ± 1.71**	19.33 ± 3.75	8.41 ± 4.60	**43.72 ± 7.30**	19.97 ± 1.82	24.16 ± 2.20
	0.5 – 1	0.00 ± 0.00	9.98 ± 1.03	**23.94 ± 1.09**	5.83 ± 0.29	4.89 ± 4.29	**21.47 ± 12.48**	7.55 ± 0.50	9.86 ± 3.41
204 less soluble in water and more hydrogen bond donors	0 – 0	0.69 ± 0.58	**17.03 ± 2.75**	14.42 ± 3.43	**13.17 ± 0.29**	4.66 ± 4.19	5.16 ± 4.51	5.37 ± 0.75	8.57 ± 0.91
	0.5 – 1	0.00 ± 0.00	2.59 ± 1.14	**3.84 ± 0.71**	**2.00 ± 1.00**	1.88 ± 1.63	0.00 ± 0.00	1.34 ± 0.76	1.66 ± 1.05
205 more soluble in water and higher permeability	0 – 0	5.06 ± 1.21	35.69 ± 3.19	**39.74 ± 2.26**	16.50 ± 0.87	3.90 ± 3.62	12.50 ± 21.65	**23.58 ± 3.26**	19.62 ± 4.00
	0.5 – 1	1.16 ± 0.68	19.15 ± 0.73	**22.66 ± 1.90**	7.50 ± 2.29	0.00 ± 0.00	0.00 ± 0.00	**15.05 ± 2.13**	9.06 ± 2.32
206 more soluble in water and lower permeability	0 – 0	12.17 ± 1.05	**44.35 ± 0.68**	30.87 ± 0.62	35.17 ± 5.20	9.68 ± 3.17	**65.48 ± 6.63**	28.11 ± 1.85	37.13 ± 1.8
	0.5 – 10	6.23 ± 2.31	**28.67 ± 2.22**	20.06 ± 1.26	23.50 ± 3.77	6.12 ± 1.91	**60.71 ± 5.19**	18.18 ± 3.47	26.15 ± 2.82

Table 6: Results on double-objective molecule editing are evaluated based on the hit ratio of the property change.

Model	**Fine-tuning dataset**	**Data type**	**ARC Challenge**	**HellaSwag**	**MMLU**
MolT5-base (Edwards et al., 2022)	-	-	0.1988	0.2744	0.2465
MolGPT (Bagal et al., 2021)	-	-	0.1980	0.2541	0.2704
Llama3	-	-	**0.5299**	**0.5776**	**0.6385**
Llama3	50k PubChem molecules	SMILES	0.4966	0.5640	0.6110
Llama3	50k PubChem molecules	IUPAC	**0.4983**	**0.5645**	**0.6185**
Llama3	500k PubChem molecules	SMILES	0.3259	0.4583	0.3448
Llama3	500k PubChem molecules	IUPAC	**0.3942**	**0.5050**	**0.4946**

Table 7: Various benchmark results with few-shot learning performance of different models