CHEMICAL LANGUAGE MODELS HAVE PROBLEMS WITH CHEMISTRY: A CASE STUDY ON MOLECULE CAPTIONING TASK

Veronika Ganeeva[1,2], **Kuzma Khrabrov**[3],
Artur Kadurin[3], **Andrey V. Savchenko**[1,4], **Elena Tutubalina**[2,3,4,5]
[1]Sber AI Lab, Moscow [2]HSE University, Moscow [3]AIRI, Moscow
[4]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow [5]Sber AI, Moscow

ABSTRACT

Drug discovery has been greatly enhanced through the recent fusion of molecular sciences and natural language processing, leading these research fields to significant advancements. Considering the crucial role of molecule representation in chemical understanding within these models, we introduce novel probing tests designed to evaluate chemical knowledge of molecular structure in state-of-the-art language models (LMs), specifically MolT5 and Chem+Text T5. These probing tests are conducted on a molecule captioning task to gather evidence and insights into the language models' comprehension of chemical information. By applying rules to transform molecular SMILES into equivalent variants, we have observed significant differences in the natural language descriptions generated by the LM for a given molecule depending on the exact transformation used.

1 INTRODUCTION

Several LMs, including BioBERT, BioGPT, and BioBART, have been developed for the biomedical and clinical domain (Lee et al., 2020; Phan et al., 2021). These models were traditionally pre-trained on textual data exclusively (Luo et al., 2022; Yuan et al., 2022; Savchenko et al., 2020). Widespread string-based molecular representations, e.g., SMILES (Weininger, 1988), allow LMs applications in the domain of medicinal chemistry (Schwaller et al., 2019; Irwin et al., 2022; Raffel et al., 2020b; Ross et al., 2022). Recently, novel cross-domain LMs were developed. In contrast with the aforementioned models, MolT5 (Edwards et al., 2022) and Text+Chem T5 (Christofidellis et al., 2023) are pre-trained on both chemical and textual data and tasks, e.g. the large C4 (Raffel et al., 2020a) corpus and 100 million SMILES strings from ZINC15 (Sterling & Irwin, 2015).

This paper contributes to a growing effort to understand better domain-specific capacities achieved by novel cross-domain LMs. Evaluation of LMs such as BioBERT is typically done with various downstream tasks (Miftahutdinov et al., 2021; Tutubalina et al., 2020; Sakhovskiy et al., 2021) or with probing tasks using knowledge graphs (Meng et al., 2022; Makarov et al., 2022), which reveal that biomedical LMs encounter challenges in capturing complex specialized domain terminology and lack awareness of synonyms (Sung et al., 2021). In this paper, we present novel probing tasks with chemical LMs. It is experimentally shown that the state-of-the-art models are vulnerable to even slight changes in molecule representations. The source code is publicly available at `https://github.com/ChemistryLLMs/SMILES-probing`.

2 MOLECULE CAPTIONING AND PROBING RULES

The molecule captioning task aims to generate a description for a given molecule. Edwards et al. (2022) proposed this task as a sequence-to-sequence translation task using the ChEBI-20 dataset. We propose several SMILES-based probing tests: 1. **canonicalization**: the transformation from a SMILES string to an RDKIT (Bento et al., 2020; Greg et al., 2022) canonical SMILES string. 2. **hydrogen**: The addition of explicit hydrogen atoms into SMILES string. 3. **kekulization**: the transformation from a SMILES string to a Kekulized SMILES string (i.e., the one where the aromatic

Table 1: Example of molecular SMILES from a test set and after our transformations.

AUGMENTATION	RESULT
original (w/o augmentations)	CC1=C2C=C(C=C(C2=CC=C1)C(=O)O)[O-]
canonicalization	Cc1cccc2c(C(=O)O)cc([O-])cc12
hydrogen	[CH3][c]1[cH][cH][cH][c]2[c]([C](=[O])[OH])[cH][c]([O-])[cH][c]12
kekulization	CC1=C2C=C([O-])C=C(C(=O)O)C2=CC=C1
cycles	CC1=C3C=C(C=C(C3=CC=C1)C(=O)O)[O-]

Table 2: Results on augmented sets. Non-canon. SMILES from the test set marked as "original".

Probing Test	MolT5-base		Chem+Text T5-base		MolT5-large		Chem+Text T5-augm	
	ROUGE-2	METEOR	ROUGE-2	METEOR	ROUGE-2	METEOR	ROUGE-2	METEOR
original	0.481	0.583	0.498	0.604	0.510	0.614	0.543	0.648
canonical	0.315	0.450	0.381	0.515	0.390	0.532	0.377	0.514
hydrogen	0.199	0.329	0.187	0.314	0.174	0.318	0.201	0.336
kekulization	0.333	0.475	0.413	0.574	0.405	0.546	0.410	0.546
cycles	0.417	0.540	0.483	0.600	0.566	0.603	0.4575	0.581

π-electrons are static between every second carbon). 4. **cycles**: valid replacement of cycle numerical identifiers with other random numbers. Their examples are provided in Table 1.

3 Experimental Results

We evaluate two architectures: MolT5 (Edwards et al., 2022), and Chem+Text T5 (Christofidellis et al., 2023). We use both base (250M parameters) and large (780M) versions. The model is further fine-tuned on the ChEBI-20 dataset (Edwards et al., 2021), which consists of 33,010 pairs of molecule description split into 80% / 10% / 10% train / val / test sets. The URL links to pretrained models, data and our source code are available in Appendix A.

For evaluation, we use the following metrics: ROUGE-1, ROUGE-2, ROUGE-L (Chin-Yew, 2004), and METEOR (Banerjee & Lavie, 2005). Table 2 contains only the ROUGE-2 and METEOR metrics, as they change together, so some can be omitted from the brief discussion for the sake of simplicity (all metrics in Appendix B). We expected all augmentations to cause the captioning quality to decline. Our experimental results support this claim, as metrics on augmented datasets are significantly (Savchenko & Savchenko, 2019) lower than on original data. Although it showed slightly better results on average, MolT5-large fails in the **hydrogen** task. The Chem+Text T5-augm was trained on the dataset augmented with additional reactions. In contrast to MolT5-large, its performance drops on all the tasks except for **hydrogen**. The toughest task for all tested models is the **hydrogen** one: while a simple addition of explicit "H"s has no effect on the underlying molecular chemistry, it changes the SMILES representation drastically. In contrast, cycle renumbering does not affect the SMILES as much, and the metrics degradation is not so high. Two other augmentations (**canonicalization** and **kekulization**) may change the SMILES significantly but more often affect only a small part of it. While none of these augmentations transform the underlying molecular structures, they change the symbols with which these structures are represented and, broadly speaking, *known* by the model, which affects its performance. We present a qualitative analysis of predictions in Appendix C. In addition, we trained models on the augmented CHEBI-20 train set: each molecule has three augmentations (canonicalization, kekulization and explicit hydrogen). Results are slightly better: we have test models on the original test set. For example, the METEOR metric of Molt5-base is increased from 0.583 to 0.596.

4 Conclusion

In this paper, we introduced novel probing tasks with chemistry LMs. Our experiments demonstrated that the state-of-the-art models are vulnerable to changes in molecule representations, as was tested by several augmentations. All changes in symbolic representation have proven to cause a decline in performance, but the extent of this decline seems to be, most of the time, dictated by language processing rather than the underlying understanding of chemistry. This new information will allow the scientific community to better understand the domain-specific capabilities achieved by novel cross-domain LMs, such as chemical LMs, while keeping in mind their inner logic and the resulting weak spots that can hinder their usage.

ACKNOWLEDGMENTS

The work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of dated November 2, 2021 No. 70-2021-00142.

URM STATEMENT

The first author meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12:1–16, 2020.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*, 2023.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, 2021. URL https://aclanthology.org/2021.emnlp-main.47/.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Landrum Greg et al. RDKit: open-source cheminformatics, 2022. URL https://www.rdkit.org/.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

Ilya Makarov, Andrey Savchenko, Arseny Korovko, Leonid Sherstyuk, Nikita Severin, Dmitrii Kiselev, Aleksandr Mikheev, and Dmitrii Babaev. Temporal network embedding framework with causal anonymous walks representations. *PeerJ Computer Science*, 8:e858, 2022.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4798–4810, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.329. URL https://aclanthology.org/2022.acl-long.329.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, 37(21): 3856–3864, 2021.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020b. URL http://jmlr.org/papers/v21/20-074.html.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal bert-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pp. 39–43, 2021.

Andrey Savchenko, Anton Alekseev, Sejeong Kwon, Elena Tutubalina, Evgeny Myasnikov, and Sergey Nikolenko. Ad lingua: Text classification improves symbolism prediction in image advertisements. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 1886–1892, 2020.

Vladimir V Savchenko and Andrey V Savchenko. Criterion of significance level for selection of order of spectral estimation of entropy maximum. *Radioelectronics and Communications Systems*, 62(5):223–231, 2019.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. PMID: 31572784.

Teague Sterling and John J. Irwin. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559. URL https://doi.org/10.1021/acs.jcim.5b00559. PMID: 26479676.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4723–4734, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.388. URL https://aclanthology.org/2021.emnlp-main.388.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models. In *Proceedings of the 28th International conference on computational linguistics*, pp. 6710–6716, 2020.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988. URL https://api.semanticscholar.org/CorpusID:5445756.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 97–109, 2022.

A LINKS TO THE EVALUATED MODELS AND RESOURCES

- Publicly available source code of our experiments: will be added in the camera-ready version
- MolT5 is available at https://huggingface.co/laituan245/molt5-base-smiles2caption;
- Text+Chem T5 is available at https://huggingface.co/GT4SD/multitask-text-and-chemistry-t5-base-standard;
- The ChEBI-20 dataset used for experiments: https://github.com/blender-nlp/MolT5/tree/main/ChEBI-20_data;
- Evaluation framework luna-nlg is available at https://pypi.org/project/luna-nlg/;
- RDKit: https://www.rdkit.org

B FULL EVALUATION RESULTS

Table 3 contains the complete results of our experiments, namely, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR metrics for molecule captioning on our datasets.

Table 3: Evaluation results on augmented datasets. We refer to a non-canonicalized SMILES from the test set as "original".

Probing Test	**Chem+Text T5-base**				**MolT5-base**			
	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
original	0.647	0.498	0.586	0.604	0.633	0.481	0.574	0.583
canonical	0.487	0.381	0.487	0.515	0.493	0.315	0.435	0.450
hydrogen	0.306	0.187	0.306	0.314	0.372	0.199	0.324	0.329
kekulization	0.516	0.413	0.517	0.574	0.512	0.333	0.451	0.475
cycles	0.574	0.483	0.575	0.600	0.579	0.417	0.520	0.540
Probing Test	**Chem+Text T5-large**				**MolT5-large**			
	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
original	0.682	0.543	0.622	0.648	0.654	0.510	0.594	0.614
canonical	0.543	0.377	0.484	0.514	0.557	0.390	0.494	0.532
hydrogen	0.376	0.201	0.325	0.336	0.348	0.174	0.297	0.318
kekulization	0.577	0.410	0.515	0.546	0.571	0.405	0.507	0.546
cycles	0.614	0.458	0.552	0.581	0.627	0.477	0.566	0.603

C QUALITATIVE ANALYSIS

Several examples of the outputs of the models are shown in Table 4. For SMILES from Pubchem, the model managed to remember the description of the molecule exactly. In canonical form, the model did not understand that the ring structure is aromatic and incorrectly captioned it as a cyclohexane. Besides that, it incorrectly placed substitutions in positions 1 and 4. Hydrogen augmentation made the model think there is deuterium in the molecule, an isotope of hydrogen. Additionally, it describes the ring structure as a furan instead of benzene. After kekulization, the model managed to caption the basic chemistry of the molecule but didn't mention any phis/chem properties, such as boiling point or color presented in the original description. Cycle augmentation made the model unsure about chemistry: "It is a methylcatechol and a monomethoxybenzen." These two molecules are similar to some extent but still different. In this case, all the augmentation except kekulization led to a wrong chemical description.

Table 4: Examples of SMILES and descriptions generated by Chem+Text T5-base.

Probing Test	**Input**	**Output**
original	`CC1=CC=CC=C1OC`	The molecule is a monomethoxybenzene that is o-cresol in which phenolic hydroxy group has been converted to the corresponding methyl ether. A 'green' solvent (b.p. 171 °C) and food flavour ingredient, it is found in mastic oils, virgin olive oils and frankincense. It has a role as a polar aprotic solvent and a flavouring agent. It is a monomethoxybenzene, a volatile organic compound and a member of toluenes. It derives from an o-cresol.
canonical	`COc1ccccc1C`	The molecule is a cycloalkane that is cyclohexane substituted by a methyl group at position 1 and a methoxy group at position 4. It has a role as a metabolite. It is a cycloalkane and a volatile organic compound. It derives from a hydride of a cyclohexane.
hydrogen	`[CH3][O][c]1[cH][cH][cH][cH][c]1[CH3]`	The molecule is a deuterated compound that is is an isotopologue of cis-2,3,6-trimethyltetrahydrofuran. It is a deuterated compound, an isotopologue of cis-2,3,6-trimethyltetrahydrofuran.
kekulization	`COC1=CC=CC=C1C`	The molecule is a monomethoxybenzene that is methoxybenzene substituted by a methyl group at position 2. It has a role as a plant metabolite.
cycles	`CC5=CC=CC=C5OC`	The molecule is a methylcatechol that is toluene in which the hydrogen at position 2 has been replaced by a methoxy group. It is a methylcatechol and a monomethoxybenzene. It derives from a methyl toluene.
Pubchem	`CC1=CC=CC=C1OC`	2-methylanisole is a monomethoxybenzene that is o-cresol in which phenolic hydroxy group has been converted to the corresponding methyl ether. A 'green' solvent (b.p. 171 °C) and food flavour ingredient, it is found in mastic oils, virgin olive oils and frankincense. It has a role as a polar aprotic solvent and a flavouring agent. It is a monomethoxybenzene, a volatile organic compound and a member of toluenes. It is functionally related to an o-cresol.