Toward More Generalizable Reaction Prediction: From USPTO Bias to SMARTS-Based Reasoning

Ozer, D. [*, 1, 2], Da Mota, B. [2], Cauchy, T. [3], Gutowski, N. [2], Lamprier, S. [2],
* lead presenter
[1] derin.ozer@univ-angers.fr
[2] *Univ Angers, LERIA, SFR MATHSTIC F-49000 Angers, France*
[3] *Univ Angers, CNRS, MOLTECH-ANJOU, SFR MATRIX F-49000 Angers, France*

Abstract: Recent advances in Natural Language Processing (NLP) have reshaped Computer-Aided Synthesis Planning (CASP) by framing chemical reaction prediction as a sequence-to-sequence problem over molecular string representations such as SMILES [1]. This paradigm has enabled the direct use of language models in chemistry, leading to remarkable benchmark performances [2] on datasets like USPTO [3], a large corpus of reactions extracted from patents. However, our analysis reveals that the USPTO dataset is both industrially biased and chemically incomplete, excluding many fundamental transformations essential for practical synthesis. By systematically evaluating language models on simple, pharmaceutically relevant reactions, we demonstrate where and how these benchmark-driven systems fail to generalize beyond the data they were trained on [4].

Rather than addressing this limitation through further dataset expansion, we propose a complementary perspective grounded in chemical knowledge [5]. The Broad Reaction Set (BRS) provides 20 generic reaction templates written in SMARTS, designed to capture broad reactivity classes absent from patent data. Building on this foundation, ProPreT5, a T5-based model adapted for chemistry, is the first language model capable of directly handling and applying SMARTS reaction templates. To further enhance generalization, a SMARTS-level augmentation strategy is introduced to inject structural diversity into reaction patterns.

Together, these contributions highlight how integrating symbolic chemical knowledge with language modeling offers a promising direction toward interpretable and generalizable synthesis prediction. This work emphasizes the need for chemically meaningful benchmarks and closer interdisciplinary dialogue between NLP and chemistry to ensure real-world applicability.

Bibliography:

[1] P. Schwaller *et al.*, “Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction,” *ACS Cent. Sci.*, vol. 5, no. 9, pp. 1572–1583, Sep. 2019, doi: 10.1021/acscentsci.9b00576.

[2] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, “Chemformer: a pre-trained transformer for computational chemistry,” *Mach. Learn. Sci. Technol.*, vol. 3, no. 1, p. 015022, Jan. 2022, doi: 10.1088/2632-2153/ac3ffb.

[3] D. M. Lowe, “Extraction of chemical structures and reactions from the literature,” Oct. 2012, doi: 10.17863/CAM.16293.

[4] D. Ozer, Nicolas Gutowski, Benoit Da Mota, Sylvain Lamprier, and Thomas Cauchy, “Rethinking NLP for Chemistry: A Critical Look at the USPTO Benchmark,” presented at the Empirical Methods in Natural Language Processing (EMNLP), Suzhou, Nov. 2025. [Online]. Available: https://ngutowski.fr/nlp4chem.pdf

[5] D. Ozer, S. Lamprier, T. Cauchy, N. Gutowski, and B. Da Mota, “A Transformer Model for Predicting Chemical Products from Generic SMARTS Templates with Data Augmentation,” presented at the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Athens, Nov. 2025. doi: 10.48550/arXiv.2503.05810.