From Exact Match Dogma to Chemical Plausibility: Rethinking Single-Step Retrosynthesis Benchmarks

Arkadii Lin[1], Maksim Vendin[1], Bogdan Zagribelnyy[1], Ivan Ilin[1], Maksim Kuznetsov[2], Nikita Bondarev[1], Vladimir Aladinskiy[1], Alex Aliper[1], Alex Zhavoronkov[1,2,3]

[1] *Insilico Medicine AI Limited, Level 6, Unit 08, Block A, IRENA HQ Building, Masdar City, Abu Dhabi, UAE*

[2] *Insilico Medicine Canada Inc., 3710-1250 Ren´ e-L´ evesque Blvd W, Montreal, Quebec, H3B 4W8, Canada*

[3] *Insilico Medicine Hong Kong Ltd., Unit 310, 3/F, Building 8W, Phase 2, Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong, HongKong SAR, China.*

Computer-Aided Synthesis Planning (CASP) is critical for assessing the synthetic feasibility of small molecules in drug discovery. While Large Language Models (LLMs) show significant promise in automating the Single-Step RetroSynthesis (SSRS) part in CASP programs [1], objectively evaluating SSRS solutions' real-world, out-of-domain (OOD) performance remains a recognized bottleneck [2]. Current evaluation protocols heavily rely on Top-K exact-match accuracy [3], an approach that implicitly assumes a single ground-truth disconnection. This methodology provides limited signals of chemical plausibility [4] and fails to account for alternative valid retrosynthetic routes, reaction-center validity, and functional-group compatibility.

To overcome the limitations of exact-match paradigms, we introduce the ChemCensor (CC) Score [5], a precedent-based deterministic quantitative metric designed to evaluate SSRS predictions through the extraction and intercompatibility analysis of reaction centers and functional groups. Utilizing this framework, we developed CREED (Comprehensive Reactant Exhaustive Enumeration Dataset), a dataset comprising approximately 6.4 million verified reaction records, with multiple per-target transformation alternatives, generated from around 3 thousand expert-coded reaction templates. Furthermore, we established URSA-expert-2026, an expert-annotated benchmark of 100 completely novel molecular structures designed to rigorously test SSRS capabilities.

Comprehensive benchmarking shows many LLMs excel on standard tests like USPTO-50K due to data leakage. Consequently, the significant drop in performance on the stricter URSA-expert-2026 benchmark reveals the true capabilities of LLMs in the SSRS task. To achieve robust OOD synthesis planning capabilities, we fine-tuned Qwen3-8B, a language model which had one of the lowest baseline CC Scores, on the CREED. The proposed C3LM (Chemistry Constraint–Consistent Language Model) model achieves state-of-the-art CC Scores across both benchmarks, demonstrating superior generalization over existing general-purpose foundation and chemical specialist models.

Bibliography:

[1] Xuan-Vu, N.; Armstrong, D.; Wehrbach, M.; Bran, A.M.; Jončev, Z.; Schwaller, P. Synthelite: Chemist-aligned and feasibility-aware synthesis planning with LLMs. arXiv (2025).
[2] Maziarz, K.; Tripp, A.; Liu, G.; Stanley, M.; Xie, S.; Gaiński, P.; Seidl, P.; Segler, M.H.S. Faraday Discuss. 256 (2025) 568–586.
[3] Segler, M.H.S.; Waller, M.P. Chem. – Eur. J. 23 (2017) 5966–5971.
[4] Morgunov, A.; Batista, V.S. Procrustean Bed for AI-Driven Retrosynthesis: A Unified Framework for Reproducible Evaluation. arXiv (2025).
[5] Zagribelnyy, B.; Ilin, I.; Kuznetsov, M.; Bondarev, N.; Schutski, R.; MacDougall, T.; Shayakhmetov, R.; Miftakhutdinov, Z.; Mizera, M.; Aladinskiy, V.; Aliper, A.; Zhavoronkov, A. When Single Answer Is Not Enough: Rethinking Single-Step Retrosynthesis Benchmarks for LLMs. arXiv (2026).