Two Steps from Hell: Compositionality on Chemical LMs

Abstract

This paper investigates compositionality in chemical language models (ChemLLMs). We introduce STEP, a benchmark with compositional questions that reflect intricate chemical structures and reactions, to evaluate models' understanding of chemical language. Our approach focuses on identifying and analyzing compositional patterns within chemical data, allowing us to evaluate how well existing LLMs can handle complex queries. Experiments with state-of-the-art ChemLLMs show significant performance drops in compositional tasks, highlighting the need for models that move beyond pattern recognition. By creating and sharing this benchmark, we aim to enhance the development of more capable chemical LLMs and provide a resource for future research on compositionality in chemical understanding. This paper is accepted to EMNLP 2025 Findings.

1 Introduction

Recent advances in large language models (LLMs) have significantly accelerated progress in computational chemistry, with applications ranging from molecular property prediction to reaction design and drug discovery (Schwaller et al., 2018, 2021). Domain-specific adaptations such as ChemLLMs built on architectures like T5 (Raffel et al., 2020) and LLaMA (AI, 2023) have enabled models to interface with molecular representations (e.g., SMILES (Weininger, 1988)) and operate effectively on datasets such as ZINC-15 (Sterling and Irwin, 2015b), PubChem (Kim et al., 2016), and USPTO-50KK (Lowe, 2012).

Despite these successes, a fundamental question remains underexplored: *Can chemical language models work with compositionally?* That is, can they combine known chemical concepts to solve novel, multi-step problems? This capability is crucial for tasks that require generalization beyond memorized patterns such as predicting the solubility of the product of a previously unseen reaction or estimating the activity of a compound generated via hypothetical synthesis. Current benchmarks such as USPTO-50K (Lowe, 2012), and CHEBI-20 (Edwards et al., 2021) in chemical NLP largely focus on single-step tasks, e.g., predicting products from reactions, generating molecular descriptions, or estimating individual properties. While these tasks provide valuable evaluation signals, they do not capture the multi-faceted, compositional nature of real-world chemical reasoning. Consequently, it is unclear whether ChemLLMs truly understand chemical concepts or simply exploit surface-level correlations.

To address this gap, we introduce STEP (Structured Tasks for Evaluating and Promoting compositionality), a benchmark and framework designed to systematically evaluate compositional reasoning in chemical LLMs. STEP transforms standard datasets into two-step tasks that require chaining atomic reasoning steps, for example, predicting a reaction product and then describing its water solubility. By evaluating state-of-the-art ChemLLMs across these tasks, we identify substantial performance drops in compositional settings, revealing critical limitations in their generalization abilities.

Our contributions are as follows. We propose STEP, a benchmark that evaluates compositionality in ChemLLMs via structured tasks. We curate a dataset spanning several chemical subdomains, including synthesis, property prediction, and molecular description, and transform them into compositional queries. We showed that most models performed well on isolated tasks but struggled with compositional generalization, especially on out-of-distribution inputs.

By addressing the critical need for rigorous evaluation, our work advances the understanding compositionality by ChemLLMs, with implications for drug discovery, materials science, and beyond.