CGR-SMILES: A Compact and Universal Sequence Representation for Chemical Reaction Modeling

Abstract

The ability to predict chemical reactivity is central to advances in drug discovery and materials science. With the high cost and time demands of experimental studies, machine learning (ML) offers a powerful alternative, aiming to enable high-throughput and routine reaction screening. A key challenge in applying ML to chemical reactions is their representation in a machine-readable format.

Capturing them in a manner that is both compact and chemically meaningful remains non-trivial. Unlike for images or text, there is no single canonical molecular representation. Commonly used options include fingerprint-based descriptors, graph-based representations, and sequence-based formats, each introducing its own assumptions.

For graph-based approaches, the Condensed Graph of Reaction (CGR) is regarded as the state of the art [1]. The CGR superimposes reactants and products into a single graph, rather than treating them separately. It explicitly encodes the chemically relevant transformation while avoiding redundancy: most single-step reactions alter only a small region of the molecules, leaving the majority unchanged. The CGR collapses this invariant information into a compact form. In contrast, sequence-based models largely rely on reaction SMILES, which list reactants and products sequentially: `reactant(s)>>product(s)`. While simple and widely adopted, this representation encodes transformations only implicitly and duplicates unchanged molecular regions, thereby limiting efficiency and generalization during model training.

Inspired by the CGR, we introduce CGR-SMILES, a novel sequence-based representation of chemical reactions, and show that it offers comparable advantages in the sequence domain. We define the representation in accordance to the following key criteria. First, it is universally applicable: every single-step chemical reaction of the form `reactant(s)>>product(s)` can be expressed as a CGR-SMILES. Second, the mapping between a reaction SMILES and its corresponding CGR-SMILES is bijective, ensuring one-to-one correspondence. Third, the representation is compact, avoiding redundant duplication of molecular information while retaining full expressivity. And finally, to facilitate adoption, we provide an accessible software interface, enabling straightforward integration into computational workflows.

To contextualize our approach, we examine ReactSeq, a recent sequence-based representation that also encodes reactions in a local-change-aware manner [2]. However, its applicability is limited, as it is confined to synthesis reactions, containing only a single product molecule.

We are evaluating CGR-SMILES across several datasets and downstream tasks to asses its universal applicability. We hypothesize that by reducing redundancy and explicitly encoding transformation changes, learning efficiency and predictive performance can be improved. To this end, we are reproducing results for reaction property prediction, forward prediction, and retrosynthesis prediction by retraining state-of-the-art models on both reaction SMILES and CGR-SMILES. Preliminary experiments on activation barrier height prediction show that CGR-SMILES achieves performance on par with ReactSeq, with both surpassing reaction SMILES baselines.

We believe CGR-SMILES can serve as a foundation for more expressive and efficient reaction modeling, and we look forward to exploring its broader applications in chemical machine learning.

References

[1] E. Heid and W. H. Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *ChemRxiv*, 2021. Preprint, not peer-reviewed.

[2] Jiacheng Xiong, Wei Zhang, Yinquan Wang, Jiatao Huang, Yuqi Shi, Mingyan Xu, Manjia Li, Zunyun Fu, Xiangtai Kong, Yitian Wang, Zhaoping Xiong, and Mingyue Zheng. Bridging chemistry and artificial intelligence by a reaction description language. *Nature Machine Intelligence*, 7(5):782–793, 2025.