Enhancements in Modelling with Reaxys Reaction Data Tailored to AI and ML Applications

Timur Madzhidov

The use of AI and ML models for chemical reactions requires accurate data curation and preparation. Reaxys reaction data is considered the best source of reaction data and the standard for modeling. However, data curation involves a significant amount of manual work and subject matter expertise to fetch correct data, homogenize chemical object representation, reduce noise level, and delete duplicated and incorrect information. While some publications have addressed the reaction data curation pipeline, no comprehensive approach for Reaxys reaction curation has been proposed. Our study presents an optimized approach to Reaxys reaction data representation for better AI and ML modeling. We developed a dataset of 25M single step reactions ready for AI and ML applications by interpreting Reaxys fields, improving data representation, correcting structural representation of organometallics, and detecting and correcting concurrent reactions. We show how such data representation optimization influenced helped to improve the modelling outcomes. We discuss how it influenced the number and coverage of retrosynthesis template generation and the outcome of retrosynthesis predictions, the quality of modeling of reactions and their characteristics.

Our study highlights the importance of data curation and preparation in AI and ML modeling and aims to improve the efficiency of data processing, allowing data scientists and chemoinformaticians to focus on building and refining AI and ML models.