Contribution ID: **52** Type: **not specified**

Thinking like a CHEMIST: Combined Heterogeneous Embedding Model Integrating Structure and Tokens

Tuesday, 20 May 2025 15:27 (12 minutes)

Representing molecular structures effectively in chemistry remains a challenging task, with both string- and graph-based approaches commonly employed. Language models and graph-based models are extensively utilized within this domain, consistently achieving state-of-the-art results across an array of tasks. However, the prevailing practice of representing chemical compounds in the SMILES format – used by most data sets and many language models – presents notable limitations as a training data format. In this study, we present a novel approach that decomposes molecules into substructures and computes descriptor-based representations for these fragments, providing more detailed and chemically relevant input for model training. We train a language model on this substructure and descriptor data and propose a bimodal architecture that integrates this language model with graph-based models including RoBERTa, Graph Isomorphism Networks (GIN), Graph Convolutional Networks (GCN) and Graphormer. Our framework shows notable improvements over traditional methods in various tasks such as Quantitative Structure-Activity Relationship (QSAR) prediction.

Primary author: REKUT, Nikolai (MIPT)

Presenter: REKUT, Nikolai (MIPT)

Session Classification: 20-Машинное обучение и нейросети

Track Classification: Машинное обучение и нейросети