ORIGINAL ARTICLE

Addressing Imbalanced Retrosynthesis Datasets with Metric Transformation and Subclass Mapping

Hung Nguyen | Milo Roucairol | Tristan Cazenave

LAMSADE, Université Paris Dauphine - PSL, Paris, France

Abstract

In this article, we investigate the impact of Proxy Anchor Loss, Transformer architecture and subclass mapping on predicting the retrosynthesis of Simplified Molecular Input Line Entry System (SMILES) chemical compounds.

We demonstrate that combining the Attention mechanism with Proxy Anchor Loss effectively enhances classification by capturing both local and global contexts and differentiating between classes, while subclass mapping further improves performance on highly imbalanced datasets.

Our approach, which requires no prior chemical knowledge, achieves promising results on the USPTO-FULL dataset, with accuracies of 53.4%, 83.8%, 90.6%, and 97.5% for top-1, top-5, top-10, and top-50 predictions, respectively.

We further validate the practical application of our approach by correctly predicting the retrosynthesis pathways for 63 out of 100 randomly selected compounds from the ChEMBL database and for 39 out of 60 compounds from a sample data set used by Bayer's chemists.