Mathematical AI for molecular data analysis

Kelin XIA（Nanyang Technological University）

Abstract

Artificial intelligence (AI) based molecular data analysis has begun to gain momentum due to the great advancement in experimental data, computational power and learning models. However, a major issue that remains for all AI-based learning models is the efficient molecular representations and featurization. Here we propose advanced mathematics-based molecular representations and featurization (or feature engineering).

Molecular structures and their interactions are represented as various simplicial complexes (VR complex, Neighborhood complex, Dowker complex, and Hom-complex), hypergraphs, and Tor-algebra-based models. Molecular descriptors are systematically generated from various persistent invariants, including persistent homology, persistent Ricci curvature, persistent spectral, and persistent Tor-algebra. These features are combined with machine learning and deep learning models, including random forest, CNN, RNN, Transformer, BERT, and others. They have demonstrated great advantage over traditional models in drug design and material informatics.