Enhancing Generalization Performance of Molecular Property Prediction via Graph Latent Diffusion Autoencoder

Daiki Koge
Niigata University

In machine learning-based drug discovery and materials design, construction of effective molecular representation is a fundamental problem. The effective molecular representation numerically from chemical structure can facilitate generalization of molecular property prediction. Given the vast space of potential organic compounds and limited availability of labeled data with specific molecular properties, enhancing the generalization performance of molecular property prediction is key for accelerating drug discovery and material design. In this work, we introduce the graph latent diffusion autoencoder (Graph LDA), a deep molecular generative model that combines a graph-transformer-based variational autoencoder and latent diffusion model, designed for constructing molecular latent representation. To assess the generalization performance of Graph LDA, we compared our model with several existing models using the widely applicable information criterion (WAIC) and the widely applicable Bayesian information criterion (WBIC). The results showed that Graph LDA achieved strong generalization performance across multiple property prediction tasks, including quantum chemical properties such as HOMO energy, physicochemical properties such as solubility, and biological activities. Moreover, Graph LDA consistently outperformed existing models in most of these tasks. Furthermore, we empirically demonstrated that the superior generalization performance of Graph LDA is attributable to the smoothness and multimodality of its learned molecular latent representation.