04-6

Improving Molecular Property Prediction using Self-supervised Learning

Laurent Dillard[1]
laurent.dillard@elix-inc.com

[1] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Self-supervised learning, transfer learning, molecular property prediction, graph neural networks,

Fast computation of molecular properties holds great potential to boost the efficiency of drug discovery pipelines. In recent years, there has been a surge of interest in developing deep learning models for such applications. A popular choice of architecture to process molecular data are Graph Neural Networks (GNNs). However, as with most deep learning models, training GNNs faces the challenge of gathering large amounts of labeled data. Since labels mostly come from experimental results, the data collection process is both time-consuming and costly. On the other hand, it is easy to access large databases of molecules making it attractive for approaches that do not rely explicitly on labels. Self-supervised learning techniques leverage large amounts of unlabeled data to train models on pretext tasks for which labels can be generated from the raw data. These pretext tasks help the model learn to extract useful feature representations from the data and the trained model can then be fine tuned on downstream tasks. Recently, self-supervised learning techniques have been applied to a growing number of fields, including chemistry. In this work, we introduce a self-supervised framework for GNNs tailored specifically for molecular property prediction. Our framework uses three different pretext tasks, each focusing on a different scale of molecules (atoms, fragments and complete molecules). For the atom and molecule level tasks, a predefined list of fragments is used to encode atom contexts and molecule labels to train the model in a classification setting. For the fragment level task, molecules are decomposed into several fragments and the model is trained to recognize which fragments belong to the same molecule through a binary classification task. Using a subset of ZINC15 molecule database[1] as the pretraining dataset, we evaluate the efficiency of our framework on the MoleculeNet[2] benchmark datasets as well as ADME datasets collected from the literature[3-6]. Our results show that self-supervised learning can successfully improve performance compared to training from scratch, especially in low data regimes. The improvement varies depending on the dataset and model architecture reaching up to +2.6% in area under the curve (AUC) for classification tasks and up to +7% in coefficient of determination (R2) for regression tasks.

[1] Shoichet Brian K Irwin John J. Zinc–a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 2005.
[2] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. Moleculenet: A benchmark for molecular machine learning. CoRR,abs/ 1703.00564, 2017.
[3] Dominique Douguet. Data sets representative of the structures and experimental properties of fda-approved drugs.ACS Medicinal ChemistryLetters, 9(3):204–209, 2018.
[4] Shuangquan Wang, Huiyong Sun, Hui Liu, Dan Li, Youyong Li, and Tingjun Hou. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches.Molecular Pharmaceutics, 13(8):2855–2866, 2016. PMID:27379394.
[5] Mark Wenlock and Nicholas Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds.
[6] Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep learning for drug-induced liver injury. Journal of Chemical Information and Modeling, 55(10):2085–2093, 2015. PMID:26437739.