New Developments in ML-Based Drug Discovery

Wern Juin Gabriel Ong
gong@bowdoin.edu

Program in Molecular Therapeutics - Fox Chase Cancer Center
Philadelphia, PA 19111

Bowdoin College
Brunswick, ME 04011

January 13, 2021

Abstract

This presentation summarizes several notable papers presented at the Conference for Neural Information Processing 2020 relating to machine learning-based drug discovery and computational chemistry.

1 Introduction

This presentation goes over three papers showcased at the Conference for Neural Information Processing Systems 2020 at the Machine Learning for Molecules workshop. This is one of the largest and most significant machine learning conferences showcasing the latest work in the field. I will present three papers from the conference:

1. Jing, *et. al.* 'Learning from Protein Structure with Geometric Vector Perceptrons'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.

2. Nayak, *et. al.* 'Transformer based Molecule Encoding for Property Prediction'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.

3. Wieroch and Kirchmair. 'Deep Neural Network Approach to Predict Properties of Drugs and Drug-Like Molecules'. *ML For Molecules, 2020 Conference on Neural Information Processing Systems*. December 2020.

2 DNNs to Predict Properties of Drug-Like Molecules

The authors here attempt to design a neural network that will better predict molecular properties such as lipophilicity and solubility. Their model outperforms other property prediction models on MoleculeNet, a data set curated by the Pande Group at Stanford specifically for chemical property prediction. The authors tested on all their physical chemistry benchmarks (ESOL, Lipophilicity, and FreeSolv) and select biophysical and physiological benchmarks (BACE, BBBP, and ClinTox).

The authors use two feature extraction blocks that are then concatenated and fed into a dense neural network for regression and classification. The first block takes in a graph representation. This graph representation encodes features within the nodes and edges with the assumption that every atom interacts with other atoms, including itself. The first block uses graph attention layers to capture spatial information about the molecule. The second block takes in a vector representation of features extracted with RDKit and fed into fully connected layers. The outputs of the two blocks are concatenated and passed into an interpretation model with dropout to derive the output.

They split data 80-10-10 training to validation to test set and train 10 models, using the ensemble as the final model for testing. Regression tasks are assessed by RMSE while classification tasks are assessed by AUC-ROC.

The following figure presents the model's performance on the regression tasks. ESOL refers to the water solubility of the compound, FreeSolv refers to the free energy of hydration of the compound, and lipohilicity is the experimental results in a water/octanol solvent. In all tasks, the model gives a lower RMSE when compared to other models in the literature.

The following figure presents the model's performance on the classification tasks. BACE refers to the binding/non-binding identity against human β-secretase, BBBP refers to binary labels of blood-brain barrier permeability, and ToxCast refers to the toxicology of the molecule. Similar to the regression tasks, the model outperforms others in the literature.

3 Transformer-Based Molecular Encoding

As with the previous paper, this model attempts to demonstrate that a transformer-based molecule encoder will provide a richer featurization and make more accurate property predictions. Their approach leverages self-attention layers and transformer neural networks to extract features more efficiently on small data sets. Self attention layers were developed from natural language processing models and allow them to understand links between words. This architecture is used to demonstrate similarities between chemical features. Their approach overcomes the requirement for large data sets that arise in many of the latest models using message passing neural networks. Message passing neural networks are best understood as convolutions neural networks on graph representations. This complex task often requires large quantities of data not available in drug discovery data sets.

MoleculeNet is used as the test data set and the data is split 80-10-10, a final ensemble of 3 models drawn from random parameter seeds are used for final testing. As we can see the model achieves the lowest RMSE over other models reported in the literature. The authors also further test the model by feeding the model both labeled and unlabeled data. In these tests, only the given number of datapoints are labeled while other data points are fed into the model unlabeled. This tests the model's ability to perform in low data situations. A test set of 200 compounds is used in these experiments. Both panels demonstrate that the model outperforms GraphConv and MPNN in low data settings.

4 Geometric Vector Perceptrons for Protein Structure

This paper develops geometric vector perceptrons to better understand protein structure for protein design tasks. This allows the model to better understand both geometric features (arrangements of amino acids in space) and relational structure that arises from residue-residue interactions.

Previous models have typically used either graph neural networks or convolutional neural networks, each with their own shortcomings. Graph neural networks excel at relational tasks, falling short on caputuring geometric features. On the other hand, convolutional neural networks are able to capture large-scale geometric and structural features but are insufficient to fully understand some of these smaller scale interactions. Geometric vector perceptrons improve over the shortcomings of these previously mentioned models. Geometric vector perceptrons take in two objects as an input, first a vector of scalar features (those invariant under geometric operations such as mass) and a 3D grid with channels of the protein. These inputs are independently passed through a linear transformations that reduce the dimension of the inputs. This allows geometric vector perceptrons to remain invariant under 3D rotation and reflection.

The authors first test their model on a synthetic task for proof of concept. In these tasks three points are defined on a ball of radius 10 and the model is asked to estimate i) the distance between the centroid of the straight line triangle to the centroid of the sphere in the off-center task, ii) the perimeter task where the model is asked to define the perimeter of the triangle over the surface of the sphere, and iii) the combined task where the model predicts the difference between off-center and perimeter objectives. From the results, we can see that the CNN outperforms the GNN on the geometric task and the GNN outperforms the CNN on the relational task. Both the CNN and GNN perform poorly on the combined task. The geometric vector perceptron, however, outperforms both these models in all tasks.

The authors then test the model on a protein design task on the CATH data set. The authors' model designs protein sequences that are msot similar to the native sequence - as measured by native sequence recovery - and perplexity on held out protein sequences - which measures how well the model narrows down the choices for subsequent amino acids.