MACHINE LEARNING HARNESSES MOLECULAR DYNAMICS TO DEVELOP THERAPEUTIC STRATEGIES FOR ALZHEIMER'S DISEASE AND CHRONIC PAIN

Evan N. Feinberg, Stanford University
2017–2018 Graduate Fellow

EXECUTIVE SUMMARY

The arc of drug discovery entails a multiparameter optimization problem spanning vast length scales. The key parameters range from solubility (angstroms) to protein–ligand binding (nanometers) to *in vivo* toxicity (meters). Through feature learning—instead of feature engineering—deep neural networks promise to outperform both traditional physics-based and knowledge-based machine learning models for predicting molecular properties pertinent to drug discovery. To this end, we developed the PotentialNet family of graph convolutions. These models are designed for and achieve state-of-the-art performance for protein–ligand binding affinity. We further validated these deep neural networks by setting new standards of performance in several ligand-based tasks. Finally, we introduced a cross-validation strategy based on structural homology clustering that can more accurately measure model generalizability, which crucially distinguishes the aims of machine learning for drug discovery from standard machine learning tasks.

RESEARCH CHALLENGE

Most FDA-approved drugs are small organic molecules that elicit a therapeutic response by binding to a target biological macromolecule. Once bound, small molecule ligands either inhibit the binding of other ligands or allosterically adjust the target's conformational ensemble. Binding is thus crucial to any behavior of a therapeutic ligand. To maximize a molecule's therapeutic effect, its affinity—or binding free energy—for the desired targets must be maximized while simultaneously minimizing its affinity for other macromolecules. Historically, scientists have used both cheminformatics- and structure-based approaches to model ligands and their targets, and most machine learning approaches use domain expertise-driven features.

More recently, deep neural networks (DNNs) have been translated to the molecular sciences. Training most conventional DNN architectures requires vast amounts of data. For example, ImageNet currently contains over 14 million labeled images. In contrast, the largest publicly available data sets for the properties of drug-like molecules include PDBBind 2017, with a little over 4,000 samples of protein–ligand co-crystal structures and associated binding affinity values; Tox21 with nearly 10,000 small molecules and associated toxicity endpoints; QM8 with around 22,000 small molecules and associated electronic properties; and ESOL with a little over 1,000 small molecules and associated solubility values. This scarcity of high-quality scientific data necessitates innovative neural architectures for molecular machine learning.

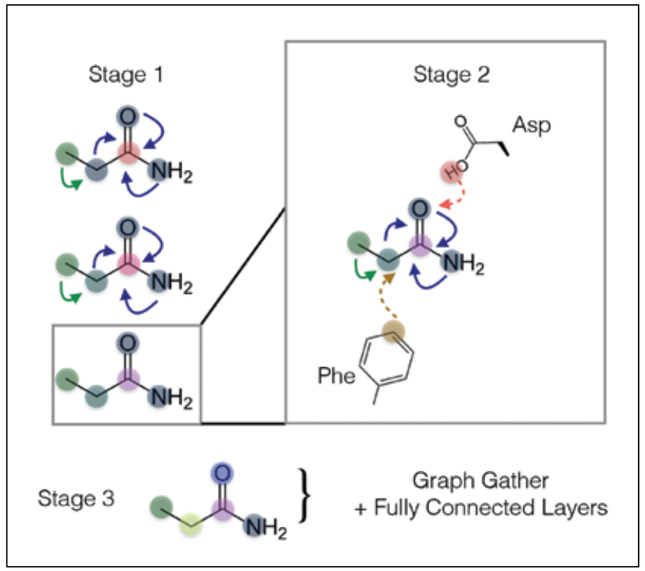

Figure 1: Visual depiction of multistaged spatial gated graph neural network. Stage 1 entails graph convolutions over only bonds, which derives new node (atom) feature maps. Stage 2 entails both bond-based and spatial distance-based propagation of information. In the final stage, a graph gather operation is conducted over the ligand atoms.

METHODS & CODES

In this project, we generalized a graph convolution to include both intramolecular interactions and noncovalent interactions between different molecules. In particular, we described a staged gated graph neural network, which distinguishes the derivation of differentiable bonded atom types from the propagation of information between different molecules. We implemented the models in PyTorch, a cutting-edge deep-learning framework. We trained and evaluated our models on publicly available data sets, including Tox21 for toxicity, ESOL for solubility, and PDBBind for protein–ligand affinity.

RESULTS & IMPACT

Spatial Graph Convolutions exhibit state-of-the-art performance in affinity prediction. Whether based on linear regression, random forests, or other classes of DNNs, RF-Score, X-Score, and TopologyNet are machine learning models that explicitly draw upon traditional physics-based features. Meanwhile, the Spatial Graph Convolutions presented here use a more principled deep-learning approach. Input features are only basic information about atoms, bonds, and distances. This framework does not use traditional hand-crafted features such as hydrophobic effects, π-stacking, or hydrogen bonding. Instead, higher-level interaction "features" are learned through intermediate graph convolutional neural network layers. In light of the continued importance and success of ligand-based methods in drug discovery, we benchmarked PotentialNet on several ligand-based tasks: electronic property (multitask), solubility (single task), and toxicity prediction (multitask). We observed statistically significant performance increases for all three prediction tasks. A potentially step change improvement was observed for the QM8 challenge, which also reinforced the value of the concept of stages that privilege bonded from nonbonded interaction.

WHY BLUE WATERS

The Blue Waters supercomputer, in particular the many GPU nodes, as well as the outstanding staff, were quite important in enabling us to run massively parallel hyperparameter searches to train the optimal deep-learning models for drug discovery tasks.

PUBLICATIONS & DATA SETS

Farimani, A.B., E.N. Feinberg, and V.S. Pande, Binding Pathway of Opiates to μ Opioid Receptors Revealed by Unsupervised Machine Learning. arXiv preprint *arXiv:1804.08206* (2018).

Feinberg, E.N., et al., Spatial Graph Convolutions for Drug Discovery. *arXiv preprint arXiv:1803.04465* (2018)

Feinberg, E.N., et al., Machine Learning Harnesses Molecular Dynamics to Discover New μ Opioid Chemotypes. *arXiv preprint arXiv:1803.04479* (2018).Feinberg, E.N., V.S. Pande, A.B. Farimani, and C.X. Hernandez, Kinetic Machine Learning Unravels Ligand-Directed Conformational Change of μ Opioid Receptor. *Biophysical Journal*, 114:3 (2018), p.56a.

As a fifth-year PhD candidate in biophysics at Stanford University, Evan N. Feinberg worked under the direction of Vijay S. Pande and Kerwyn C. Huang. He received his degree in September 2018.