Abstract

Virtual Screening is essential in the drug discovery process, as it reduces all of chemical space (~10^60) down to a reasonable number of testable compounds (~10^3). Our previous work, gnina, utilized convolutional neural networks to score protein-ligand binding poses in order to determine if a ligand would bind the protein. As protein-ligand binding affinity is dependent on its pose, we reason that there could be benefit to joint training on scoring the protein-ligand pose and predicting the binding affinity of that pose. We present here an extension to gnina, which simultaneously predicts a score for the protein-ligand complex and the affinity of said complex. Additionally we show the importance of training on docked poses, and testing on clustered cross-validated splits of the training data in order to obtain a model whose predictions are pose sensitive and generalizable to unseen data, and showing the importance of proper training data.

Background

Protein-ligand scoring provides a metric of binding strength between small molecules and target proteins; a critical subroutine of structure-based drug design. An ideal scoring function would correctly predict the binding affinity and correctly identify an accurate ligand pose for the protein. Convolutional neural networks are state-of-the-art in image recognition. Convolutional layers apply a small non-linear kernel function iteratively across the input to produce a feature map. More convolutions are then applied to these feature maps to recognize higher order features in the input.

Models

http://github.com/gnina/

Datasets

Smina docked and minimized poses are used for training.

Cross-Docked

Structures from Pocketome
2933 distinct pockets
22,767,352 non-redundant ligand poses
Affinity data for ~40% of ligands

Redocked

Subset of Cross-Docked
2933 distinct pockets
790,954 ligand poses
Affinity data for ~40%

Pose Sensitivity

To extend our previous models, we now perform joint training on the pose of a complex with a logistic loss (classification) AND a mean squared error L2 loss for affinity prediction (regression). Notably, we only penalize poor poses for over predicting the affinity of the complex.

Importance of Good Training Data

We observe that in general there is a left shift (IE more negative correlations) when joint training with the Pose and Affinity as expected.

Acknowledgements

This research was supported by R01GM108340 from the National Institute of General Medical Sciences and contributions from aigrant.org, Google Cloud, NVIDIA Corporation, the University of Pittsburgh Center for Simulation and Modeling, and the University of Pittsburgh Center for Research Computing.