Brendan: A Deep Convolutional Network for Representing Latent Features of Protein-Ligand Binding Poses

Thomas Lau, Ron Dror

Department of Computer Science and Statistics

Abstract

Molecular "fingerprints" (feature vectors) are often used in computational drug discovery to predict Protein-Ligand binding affinity. However, these fingerprints are based on chemical descriptors that are hand-tailored to match quantum mechanical data, making the development and choice of individual fingerprint features extremely difficult and arbitrary. In this paper, we introduce a deep convolutional network, Brendan, that allows us to learn the latent features of Protein-Ligand binding poses by training on 13,000 raw crystallographic poses from PDBBind.

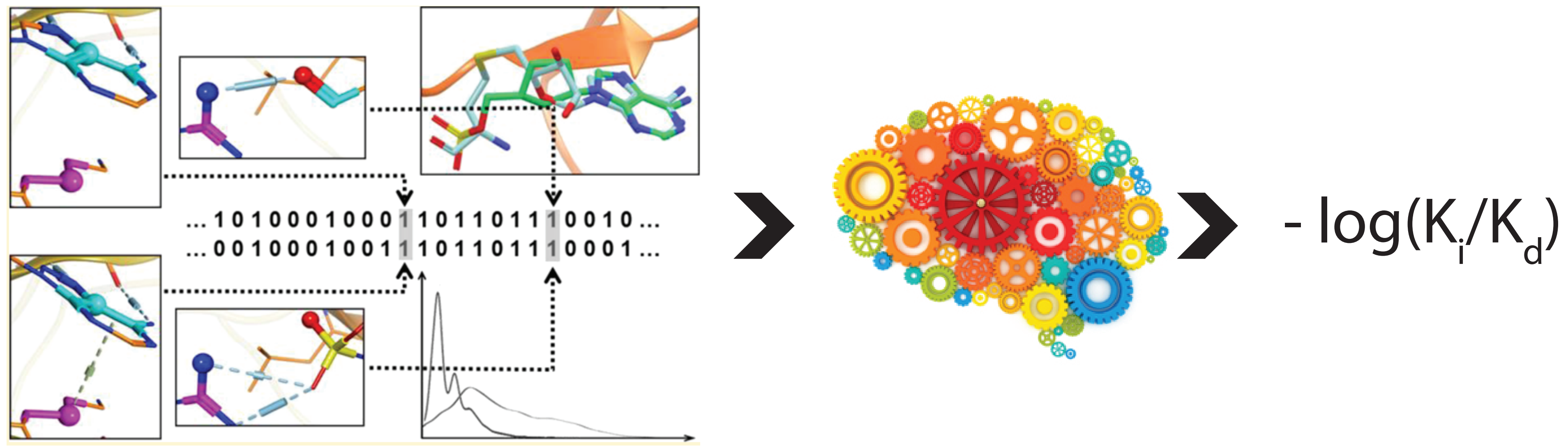
...1010001000110110111100l0...
...0010001001110110111l0001...
- log(K_i/K_d)

Figure 1. Classical QSAR Approaches use hand-crafted fingerprint vectors and standard machine learning methods (random forests) to predict chemical features

Although other approaches have attempted to use deep learning to predict the chemical properties of molecules, we are the first to (1) accurately predict the properties of protein-ligand complexes using a (2) flexible spatial representation of the complex of interest. We show that these methods can be used for downstream applications.

Methodology

A variety of methods were implemented for Brendan and compared to state of the art performance (both hand-crafted and learned):

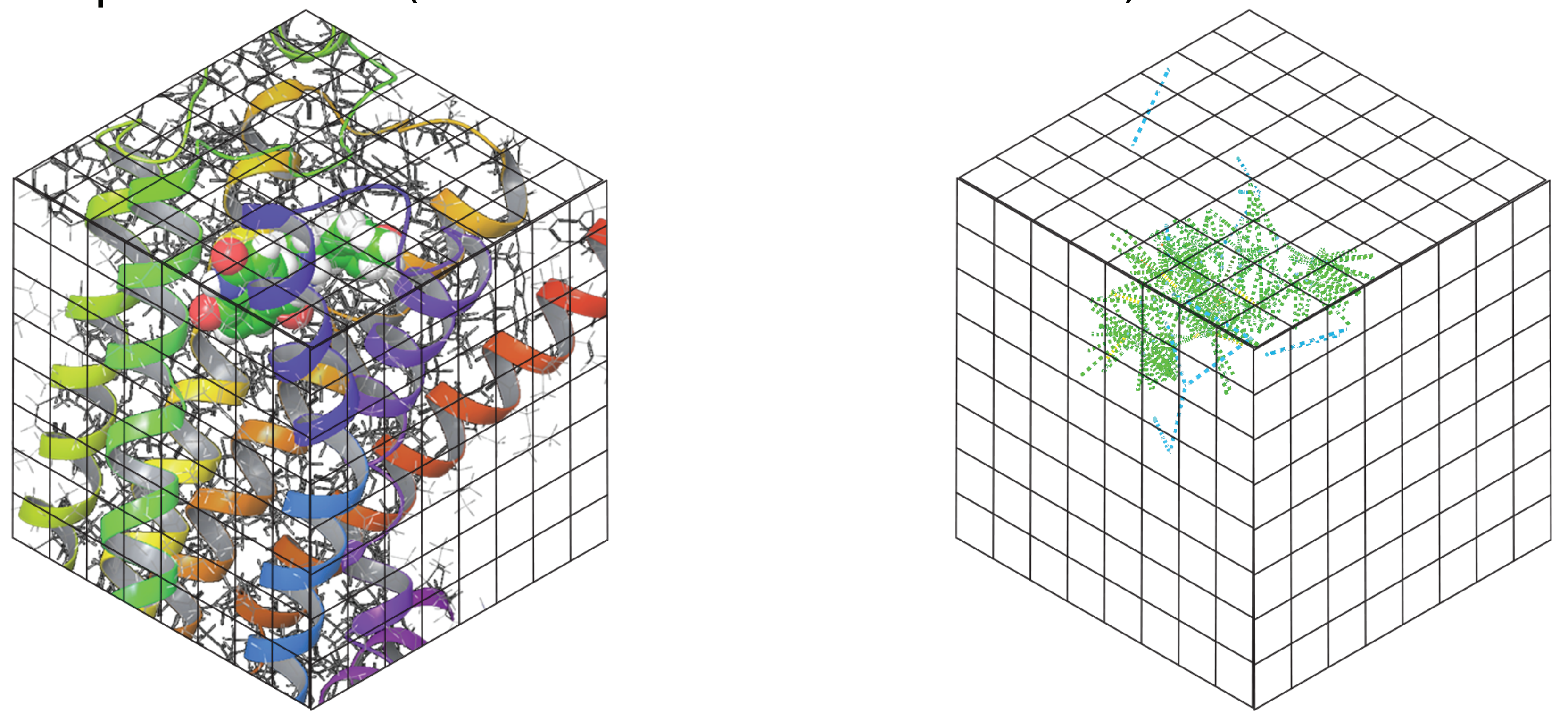

Figure 2. Novel 3D Convolutional architecture consisting of (left) SPLIF hashed input and (right) ligand-protein interaction energy both into 1x1x1 Angstrom voxels, creating in total, a 20x20x20 Angstrom input cube.

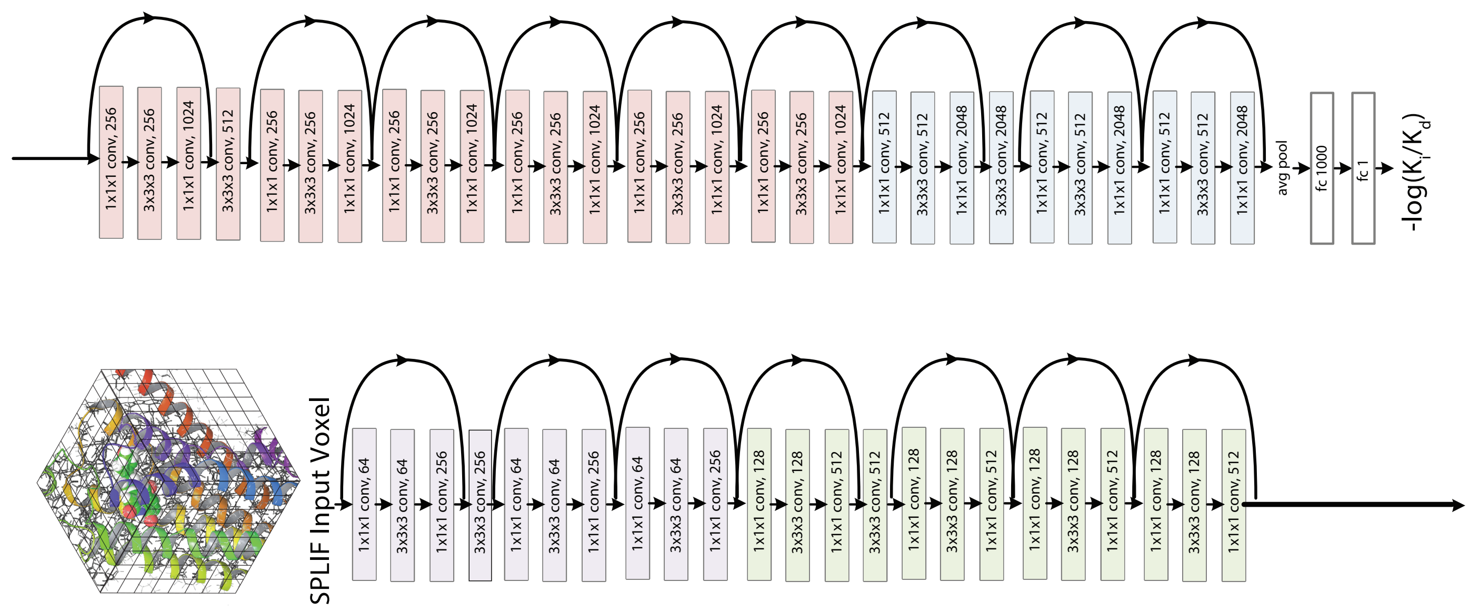

Figure 3. Brendan 3D Convolutional architecture usinga ResNet inspired structure. Crystallographic structures from PDBBind are regressed on binding affinity.

Novel graph convolutional methods were also implemented using a GS32-P4-GC64-P4-FC512-FC1 architecture as proposed in Defferrard et al. These methods were are compared to the Duvenaud et al. graph convolutional method implemented in DeepChem.

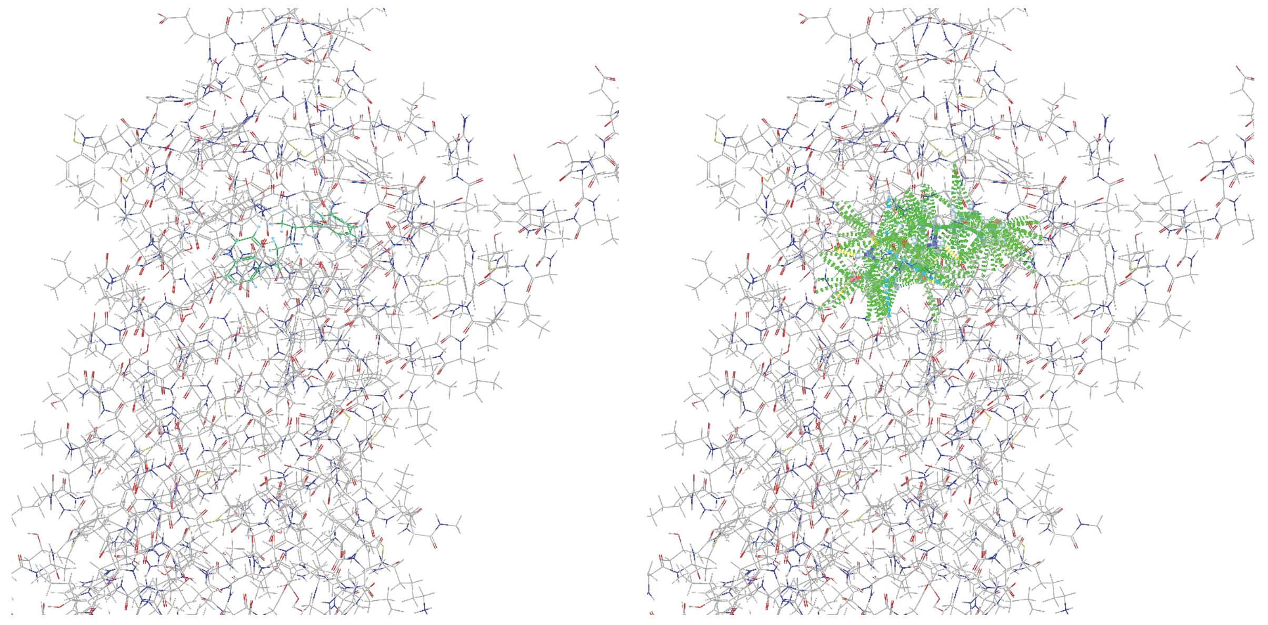

Figure 4. (Left) Latent features are learned from a ligand-protein graph, given precomputed atom features and a binary adjacency matrix. (Right) Latent features are learned from a ligand-protein graph with a real-valued adjacency matrix.

Graph convolutions in DeepChem (benchmark) are preformed using a naive graph convolution f($H^{(l)}$, A) = $\sigma(AH^{(l)}W^{(l)})$ where A is a binary adjacency matrix. Graph convolutions in our method use the Chebyshev polynomial approximation to the symmetric normalization f($H^{(l)}$, A) = $\sigma(D^{-1/2}AD^{-1/2}H^{(l)}W^{(l)})$ where A is the weighted (real-valued) adjacency matrix as in Defferard et al. Chebyshev polonomial approximations are shown to give a higher validation accuracy on graph datasets than non-parameterized graph convolutions. Adjacency weights are given by bonded and unbonded energies from the OPLS v3 force field from Schrodinger.

Results

PDBBind: Regression on -log(K_i/K_d); R^2 Performance

Methodology	Train	Valid	Test
ECFP	0.373	0.361	0.337
ECFP Grid	0.960	0.488	0.471
SPLIF Grid	0.971	0.501	0.497
Interaction Grid	0.915	0.402	0.348
Naive Graph Convolution	0.193	0.196	0.189
Atom Convolution	.962	?	0.562
Brendan Graph Convolution	0.916	0.567	0.503

Figure 5. Performance of our novel methods against state of the art benchmarks in DeepChem

Preliminary results of regression on the full set of PDBBind crystals shows a significant performance from previous graph convolutional based chemical learning methods. Although Brendan is slightly beat out by atomistic convolutions, the two methods are quite similar and with further refinement, Brendan should be able to at least match the performance of atomistic convolutions.

Conclusion

Brendan represents a new paradigm of a end-to-end differentiable deep learning model for chemical learning. Our preliminary results show that good similarity between ligand-protein interactions can significantly improve the ranking of docked poses. With accurate graph based convolutions, we can finally explore models such as triplet networks and GANs to optimize chemical data.

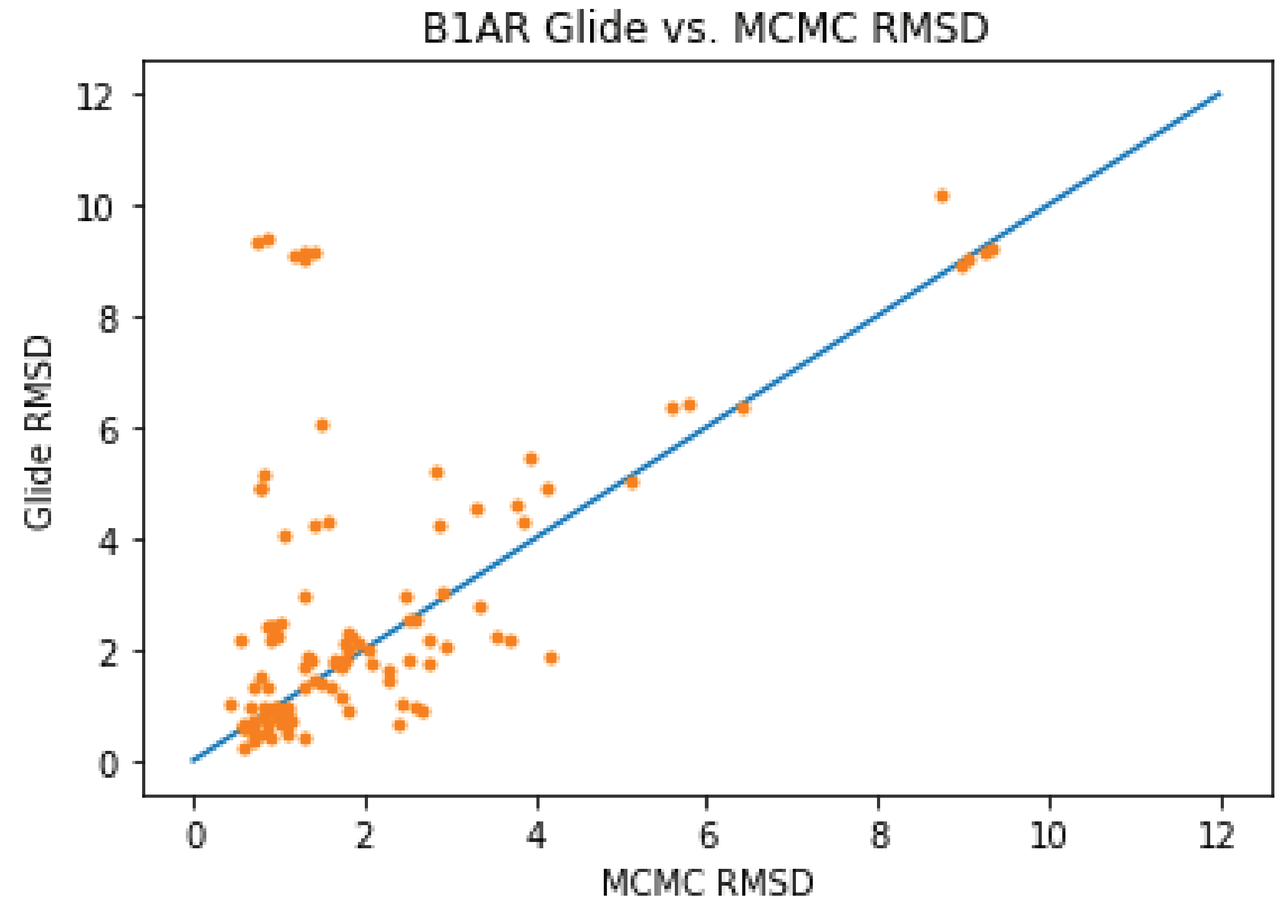

Figure 6. Our preliminary results show that fingerprint based similarity metrics can improve the performance of docking.