Convolutional Neural Networks for Protein-Ligand Scoring

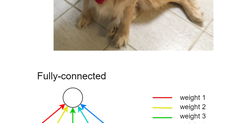
Matt Ragoza[1,2], Jasmine Collins[1,2], Noah Bastola[2,3], Dr. David Koes[4]

[1]Department of Neuroscience, [2]Department of Computer Science, [3]Department of Biological Sciences, [4]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15213

Introduction

Computational approaches are invaluable to the drug development pipeline for their low cost and high throughput compared to experimental assays. Though methods exist for scoring protein-ligand interactions for tasks like virtual screening and lead optimization, they face a number of issues. One of the major challenges is deciding what features are most important to use for scoring. Many current approaches require features to be chosen beforehand. Consequently, these approaches do not fully utilize the vast amount of structural and cheminformatic data that is publicly available. A scoring function that learned from data the types of chemical interactions that best predict binding would eliminate the need for arbitrary feature selection. Deep learning is a successful strategy for empirically learning the most influential in image recognition tasks, and the similarly spatial nature of chemical structures suggests that this approach can be adapted for protein-ligand scoring. Drawing from the demonstrated ability of hierarchical models to extract abstract features from images for classification, we introduce a novel approach for learning protein-ligand scoring functions using convolutional neural networks.

Background



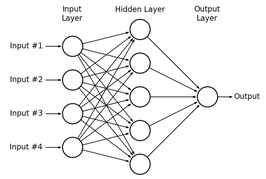
One possible pose of a small molecule binding to a protein target.

Protein-ligand scoring involves developing a metric of binding strength between small molecules and target proteins. This has a wide array of uses such as **virtual screening**, which filters large databases of candidate molecules for potential hits. Once hits have been found, scoring can also be used for **lead optimization** which aids in increasing affinity and selectivity of leads. Some protein-ligand scoring models can even generate completely new compounds that would maximize their output score, called **de novo design**.

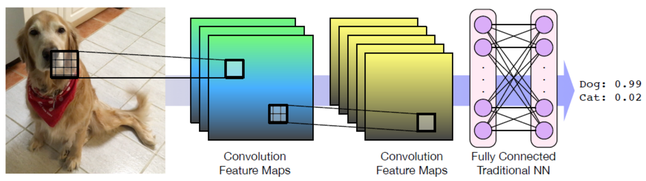
Machine learning strategies have been used to treat the scoring function as a classification problem between "good" and "bad" binding states, though they often require manually selecting properties that the model uses for discrimination, for example pairwise interactions and global counts of typical chemical interactions. However, other machine learning models can learn the most important features directly from data.

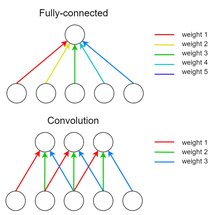
Neural networks are a supervised machine learning algorithm inspired by the nervous system. A basic network consists of an input layer, one or more hidden layers, and an output layer of interconnected nodes. Each hidden node computes a **feature** that is a function of the weighted input it receives from the nodes of the previous layer.

Input data are fed forward through the network, and a prediction is output by the last layer. A neural network is trained by iteratively updating its **weights** by minimization of an **objective function**, for example, the mean squared deviation between predictions and their ground truth labels.



A simple 3-layer neural network.



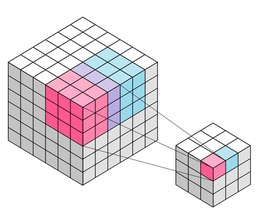


Within the last decade **convolutional neural networks** have become the state-of-the-art in image classification. Convolutional layers only have connection weights to small spatial subsets of the previous layer, and apply these **weight kernels** across the entire input to produce **feature maps**.

The fact that convolutional layers learn local features and apply them across the entire input space leads to faster training and improved accuracy on data with spatial structure. Though originally intended for classifying images, convolution works just the same in three or more dimensions.

The rise of GPU computing in combination with other advances has made training networks with many more layers feasible, leading to the surge of research in **deep learning**. Each successive layer in a deep neural network learns features at a higher level of abstraction.

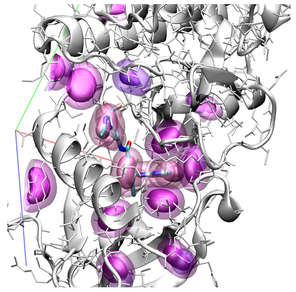
Taking advantage of the success of deep convolutional networks in a domain that bears many similarities to the task of protein-ligand scoring could lead to more accurate models that eliminate the need to select features manually.

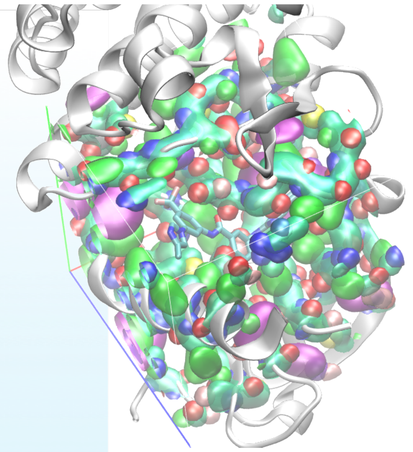


Methods

Dataset

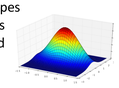
- Community Structure-Activity Resource (CSAR)
  - CSAR version 2010 with 2011 update
  - 337 targets, eliminate weak binders
  - Crystal structures → Known correct poses
- Docking and labelling with AutoDock Vina
  - Poses < 2Å RMSD → 327 actives
  - Poses > 4Å RMSD → 300 decoys
  - Eliminate ambiguous poses
  - 627 total docked poses
- Data augmentation by rotation
  - 1 random non-axial rotation
  - 24 unique axial rotations
  - x48 total increase in training examples
  - 30,096 total augmented poses

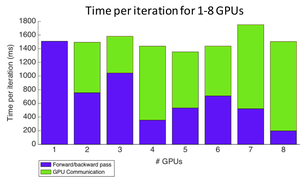




Input Format

- Voxel grid centered at active site
- 34 atom type channels
  - 16 receptor atom types
  - 18 ligand atom types
- 24Å^3 Gaussian atom grid
  - 0.5Å resolution
  - 49^3 points
- 34*49*49*49 = 4,000,066 values

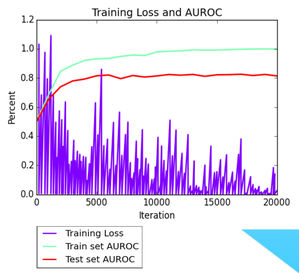




Our large models fail to benefit from multi-GPU training due to large communication overheads. Reducing these overheads is a current research goal.

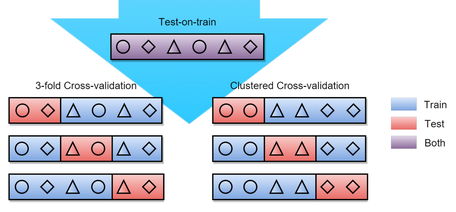
Training

- Caffe Deep Learning Framework
- Layer-wise model definition
  - N-dimensional input layer
  - Convolutional layers
  - Non-linear layers (rectified linear units)
  - Dropout regularization
  - Fully-connected layers
  - Softmax (convert to probabilities)
  - Multinomial logistic loss (2-class)
- Hyperparameter search
  - Learning rate
    - Initial value
    - Rate and shape of decrease
  - Weight decay
- Performance
  - Mini-batch parallelism (batch size=10)
  - Multi-GPU support (# GPUs=2)



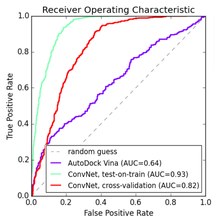
Model Evaluation

- Receiver operator characteristic (ROC)
  - False positive vs. true positive rate
  - Area under ROC curve (AUC)
- Test-on-train
  - Control for overfitting
- 3-fold cross-validation
  - Split targets into 3 balanced folds
  - Train 3 models, leaving one fold out
  - Combine performance on test sets
- Clustered cross-validation
  - Split target families into folds
    - E.g. Kinases

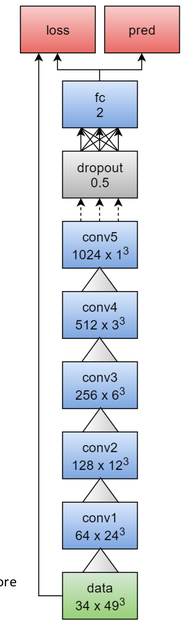


Results

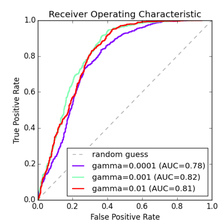
Model Performance



- Best model achieved peak AUC of 0.82 by clustered cross-validation
  - Compare to AUC of 0.64 using the Vina docking score
- Model architecture characterized by repeated down-sampling convolution layers
  - Reduce each spatial dimension
  - Double the number of feature maps
  - Repeat until dimensions are consumed
- Final 1024 convolutional features
  - Receptive fields cover entire input structure
  - Regularized with 50% dropout
- 2-class output probabilities
  - 1024 features fully-connected to 2 outputs
  - Softmax converts them to probabilities

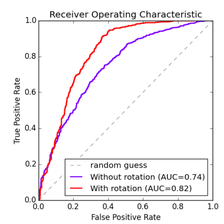


Hyperparameter Selection



- Hyperparameters affect training convergence
  - Ranges of values plateau to similar AUC
  - Parameters at either extreme inhibit learning
- Rate of learning rate decrease (gamma)
  - Fast decrease → slow convergence
  - Slow decrease → oscillations
- Weight magnitude penalization (weight decay)
  - 0.001 helped avoid overfitting
  - 0.005 prevented any learning

Effect of Data Augmentation



- No data augmentation
  - Peak cross-validation AUROC of 0.74
  - Near-perfect test-on-train accuracy; overfitting
- Data augmentation by axial and non-axial rotation
  - Peak cross-validation AUROC of 0.82
  - Test-on-train accuracy took longer to overfit
  - x48 more examples → x48 more iterations before an example is seen twice

Conclusion

We have demonstrated that convolutional neural networks can classify active protein-ligand interactions 28% more accurately compared to AutoDock Vina. Though searching for the optimal parameters can be challenging, the capability of a well-trained model is evident in its high performance on test data with little preprocessing. Though there is some benefit gained from meticulously searching for hyperparameters that maximize training convergence, the improvements witnessed from simple data augmentation outweighed the benefits from finely-tuned hyperparameters as long as the hyperparameters were in a sensible range. Other transformations such as translation and mirroring could be applied to further augment the amount of training data and potentially lead to even higher accuracy. In addition, the fairly small size of the original CSAR dataset implies that accuracy and generalizability have the potential to improve even more given a larger initial dataset. As the amount of publicly accessible structural data increases, the accuracy and generalizability of convolutional neural net models will only benefit. We will continue to test these models on larger datasets and determine what parameters and methods allow the best performance, with the long-term goal of incorporating them into tools that will increase the rate and quality of drug discovery.

Acknowledgements

This work is supported by R01GM108340 from the National Institute of General Medical Sciences with additional aid from the University of Pittsburgh Student Government Board. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.