Transfer Learning on Small Biochemical Datasets: Attempts to Design an Intelligent Agent for Analyzing Human Voltage Gated Sodium Ion Channel Inhibitors

Lemuel Cardenas-Arriaga[1], Joshua Spayd[1], Jay Minsu Liu[2]

[1]Department of Computer Science, [2]Department of Chemistry, Stanford University

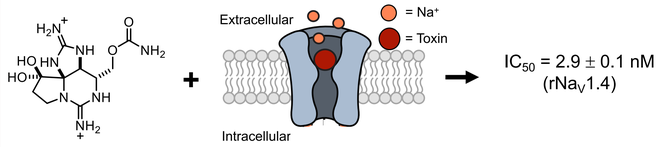
Motivation

- Voltage-gated sodium ion channels (Na_Vs) transmit nerve impulses.
- Na_V dysfunction is associated with several diseases, such as cardiac arrythmia, epilepsy, and chronic pain.
- Researchers are searching for Na_V - selective inhibitors.

Problem Definition

Task:

Predict binding affinities of modified bis-guanidinium neurotoxins **(the ligand)** to different isoforms of voltage-gated sodium ion channels **(the protein)**.



Dataset:

300 binding affinities of ligands to proteins collected from experimental electrophysiology recordings done by the Du Bois group.

- **40 unique ligands**
  - Saxitoxin, gonyautoxin, and tetrodotoxin scaffolds
- **70 unique proteins**
  - Human isoforms; rat isoforms; and single, double, and triply mutated versions

Challenges

- Small dataset = **overfitting**.
- How to featurize small molecules and proteins?
  - Challenges for geometric featurizer:
    - No three-dimensional structural files for most proteins in our dataset.
    - No three-dimensional docking poses for most ligands in our dataset.
  - Challenges for sequence-based featurizer:
    - Unintuitive modeling scheme.
    - Lose distance and interaction-based relationships.

Approaches

Baseline:

N-gram Naïve Bayes classifier:

[NH2+]=C1N[C@@H](CO C(NC)=O)[C @H](NC2=[N H2+])[C@]3(N2)N1CCC3

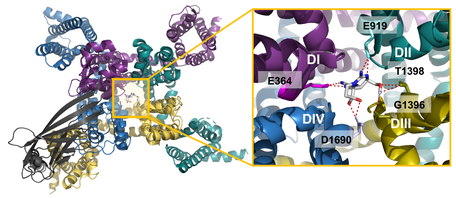
+

TQDYWEN CGEWIET TFKGWMD TSAGWDG

$\rightarrow \mathbb{1}\{IC_{50} \leq 500nM\}$

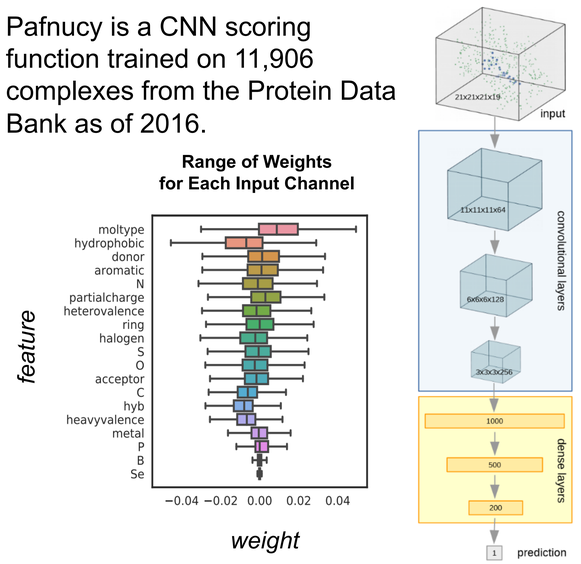
Oracle:

AutoDock Vina - industry-standard scoring and docking function:



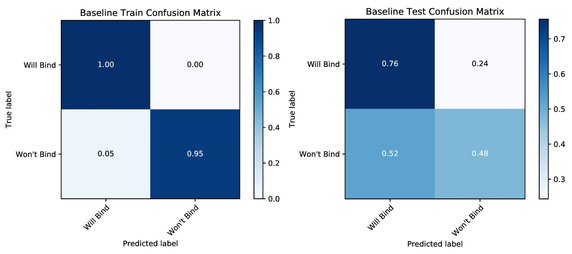
Transfer Learning:

Pafnucy is a CNN scoring function trained on 11,906 complexes from the Protein Data Bank as of 2016.

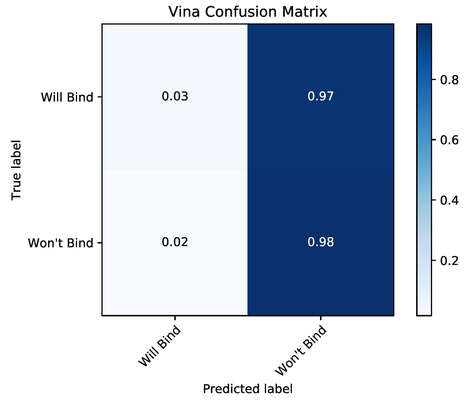


Results

Baseline:

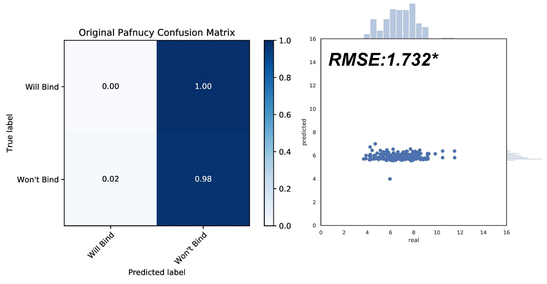


Oracle:

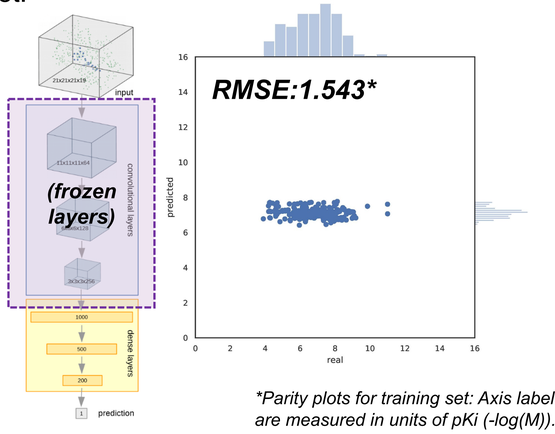


Transfer Learning:

Feature extraction: Used Pafnucy's featurizer and pretrained network to score our dataset.



Fine tuning: Froze convolution layers and retrained network starting from original weights using an 80:10:10 train/test/validation split of our dataset.



**Parity plots for training set: Axis labels are measured in units of pKi (-log(M)).*

Analysis

- Potential dataset translation issues:
  - Making point mutations to build every sodium channel from human 1.7 assumes that the channel pore retains its overall shape. In reality, mutations may alter the protein folding process and alter the architecture of the DEKA loop.
- Error metric analysis:
  - Baseline classifier clearly overfits training set.
  - Oracle achieves mediocre accuracy by predicting "not potent" for every datapoint.
  - RMSE becomes slightly better after fine-tuning. RMSE on Pafnucy's pretraining test set was 1.42.
- Future directions:
  - Analyze composition of original dataset in detail to achieve a more even train/test/validation split
  - K-fold cross validation
  - Selectively freeze fully-connected layers

Acknowledgements

Thank you to the CS221 teaching team and our mentor, Reid Pyrzant.

Scripts for running Pafnucy:

- https://gitlab.com/cheminfIBB/pafnucy

Project repository:

- https://github.com/jspayd/smiles-convert

References

[1] J. R. Walker, P. A. Novick, W. H. Parsons, M. McGregor, J. Zablocki, V. S. Pande, and J. Du Bois, Proceedings of the National Academy of Sciences 109, 18102 (2012).
[2] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, IFAC-PapersOnLine 48, 469 (2015).
[3] R. Wang, X. Fang, Y. Lu, and S. Wang, Journal of medicinal chemistry 47, 2977 (2004).
[4] Subramanian, G., Ramsundar, B., Pande, V., and Denny, R. A. (2016) Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. Journal of Chemical Information and Modeling 56, 1936–1949.
[5] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, Deep Learning for the Life Sciences (O'Reilly Media, 2019) https://www.amazon.com/ Deep-Learning-Life-Sciences-Microscopy/ dp/1492039837.
[6] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, Journal of medicinal chemistry 48, 4111 (2005).
[7] J. A. Lundbæk, P. Birn, A. J. Hansen, R. Søgaard, C. Nielsen, J. Girshman, M. J. Bruno, S. E. Tape, J. Egebjerg, D. V. Greathouse, et al., The Journal of general physiology 123, 599 (2004).
[8] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, stat 1050, 19 (2017).
[9] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, IEEE transactions on medical imaging 35, 1285 (2016).