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

Keywords: Life Sciences, Physical Sciences, Transfer Learning, Neural Networks

Lemuel Antonio Cardenas-Arriaga (lemuel), Jay Liu (jliu99), Joshua Spayd (jspayd)

Voltage-gated sodium channels (Na_Vs) modulate membrane permeability to sodium ions and facilitate crucial intercellular communication. Their dysfunction is implicated in a variety of diseases, including epilepsy, cardiac arrhythmia, and chronic pain. Nine human isoforms of Na_Vs (Na_V1.1–1.9) have been distinguished, but as we currently lack the ability to image Na_V isoforms in live cells, many channelopathy studies have been tabled until a proper tracking agent can be developed. Luckily, some eukaryotic Na_Vs are susceptible to a class of highly potent mono- and bis guanidinium neurotoxins, including (+)-saxitoxin (STX), tetrodotoxin (TTX), batrachotoxin (BTX), saxitoxinethanoic acid (SEA), and zetekitoxin (ZTX). [1] This class of neurotoxins reversibly disables Na_V function with high potency (STX has an IC50 = 2.9 ± 0.1 nM against rat Na_V1.4 (rNa_V1.4)) and represents a scaffold upon which chemists have attempted to synthesize small molecule Na_V imaging agents, mutant-selective ligands, and isoform-specific blockers.

Unfortunately, not all Na_Vs are susceptible to toxin inhibition; for instance, Na_V 1.5, 1.8, and 1.9 are TTX-resistant. Attempts to intelligently design around these limitations have been frustrated by the lack of information pertaining to the binding pose of toxins in certain hNa_Vs (due to the difficulty of acquiring crystal structures of transmembrane receptors and the imprecision of homology modeling). Traditionally, binding models have been proposed from clues garnered through mutant cycle analysis of Na_V targets and modified toxin derivatives, but this is a labor-intensive endeavor in organic synthesis and site-directed mutagenesis. A parallel problem emerges when we consider advancements in organic synthesis that have accelerated the modification and study of previously inaccessible complex natural products like 11-SEA and ZTX, STX congenerics that hold significant promise as potential therapeutics. Understanding their binding pose within hNa_Vs seems to be a prerequisite for engaging in the design of Na_V isoform-specific inhibitors.

Justin Du Bois is a faculty member in Stanford's Department of Chemistry that has done significant work on this problem. His group works on synthesizing sets of toxin derivatives in order to assess their potency towards various Na_V isoforms and mutants. Our project proposes to train a supervised learning algorithm on a dataset from the Du Bois lab and arrive at a model that is able to predict the potency of candidate bis guanidinium toxin derivatives against Na_Vs 1.1-1.9.

Du Bois' dataset consists of about 300 datapoints. The inputs are the molecular structure of an artificial neurotoxin (encoded as a SMILES string) and the candidate three-dimensional coordinates of the protein to which the neurotoxin is experimentally bound (encoded in a mol2 file). The output is a binary indicator of potent (1) or not potent (0). Since this is a binary classification problem, we will use standard error reporting metrics like ROC-AUC plots and confusion matrices to evaluate the success of our algorithm.

As a baseline, we implemented a unigram naive bayes classifier using SMILES strings and protein amino acid sequences as features. Our classifier only managed to achieved an accuracy of 47% on the training set (very close to random guessing). These results indicate that unigram featurization of string input is insufficient to capture relationships in chemical space and is likely to underfit the data.

Having observed these results for the baseline, we attempted to design an oracle that could fit the training data very well using more intricate featurization. To this end, we featurized our dataset using tuples of (protein sequence, unigram SMILES string), (protein sequence, bigram SMILES string), and (protein sequence, 3-gram SMILES string). We wrote a basic linear classifier to fit the data by minimizing the hinge loss, and we trained it with stochastic gradient descent. We used an 80:20 train/test split. Our oracle achieved a surprisingly high accuracy: by iteration 49, it reported a training set error of 0.41% and a test set error of 24%. Because our dataset is well-balanced, the accuracy is not a result of guessing the same label for every datapoint. However, we did notice that while the error decreased over the course of training, it started very low from iteration 1. Moreover, the final error varies quite a bit with repeated runs. The first point could be explained by the fact that the algorithm is overfitting the dataset. The second point is indicative of the effects of working with a small dataset. Because we have less than 300 datapoints, the initial random train/test partition can have a significant effect on the final results of the training.

The performance gap between the baseline and oracle is encouraging. It indicates that this is a problem space in which we could potentially design a good predictive algorithm. However, it is also important to recognize the limitations of our initial tests. Our classifiers are trained on n-gram featurization of the strings, with no regard for the chemical identity of the compounds, much less the intermolecular interactions (hydrogen bonding, Van der Waals' forces, ionic bonding) that chemists usually rely on to make predictions. We may be able to achieve a higher accuracy within this dataset because all the ligands and protein targets we are currently considering are within the same family of compounds. However, this algorithm may struggle to generalize to new data or achieve relevance/repute within the chemical community. Thus, there is still significant room for improvement in the way of featurization.

These questions and potential problems are of significant interest to not only the Du Bois lab, but to the biochemical community in general, where the immensely promising predictive capacity of ML is tempered by the unique challenges that such a problem space present. Specifically, biochemical datasets are often small, disaggregated, and imbalanced, and methods of representation for complex natural products and their protein targets remain an open problem.

Proposed solutions to such problem fall into two classes. The first is to leverage existing optimization techniques to run machine learning algorithms on the small dataset. Some of the techniques include multiple runs for model development, surrogate data analysis for model validation, and k-fold cross validation. These methods were successfully leveraged by Shaikhina et al. to design a neural network and decision tree capable of predicting the likelihood of antibody-mediated kidney transplant rejection from only 35 bone specimens and 80 kidney transplants. [2]

The second option is to supplement a small dataset with a much larger dataset of protein-ligand interactions. The interactions that determine binding between proteins and their small molecules are common across most protein-ligand binding pairs. Thus, even though the larger dataset may not contain information specific to saxitoxin and sodium channels, machine learning models may still be able to learn enough about atomic interactions to predict accurate binding affinities of candidate STX derivatives to their protein targets. Such large target/ligand datasets exist and have been successfully curated by multiple sources. [3] This approach has been utilized successfully by Pande et al. to quantitatively predict the binding affinities of inhibitors to human -secretase 1 (BACE-1). [4, 5]

The existence of accurate, open-source models trained on large target/ligand datasets may preclude further explorations in regression/classification model development, particularly when considering the application of our project to specific settings such as the sodium channel-STX binding subspace. Thus, rather than designing a new classification algorithm, we propose to address our challenges in this project by applying transfer learning to one such three-dimensional convolutional neural network, Pafnucy, and fine-tuning it into a model capable of making accurate predictions on a smaller dataset without overfitting. Pafuncy is trained on 11,000 mol2 files of ligands bound to proteins, files that can easily be generated from our dataset with AutoDock Vina. [8] Transfer learning has been applied extensively by researchers working with ImageNet, an extremely large database of images, to classify images of interest. [9] In addition to classifying candidate ligands, the convolutional layers of this network might be able to extract key features and functional groups responsible for the enhanced potency of the mono and bis guanidinium toxin class and ultimately enable selection within large drug discovery datasets (like ChEMBL and NCI) for hits against the Na_V isoforms for which candidate binders have yet to be identified. The realization of these goals could critically streamline synthetic organic efforts by narrowing the set of derivatives of interest in ongoing toxin-channel binding studies and by proposing new leads in the development of isoform-specific channel blockers.

[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).