Scrutinizing Deep Learning: A Virtual Screening Case Study

Shengchao Liu[1], Moayad Alnammi[1], Scott Wildman[2], Spencer Ericksen[2,3], Haozhen Wu[2,4], Andrew Voter[5], James Keck[5], Michael Hoffmann[2,6], Anthony Gitter[1,3,7,8]

[1]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI; [2]Small Molecule Screening Facility, University of Wisconsin Carbone Cancer Center, Madison, WI; [3]Center for Predictive Computational Phenotyping, University of Wisconsin-Madison, Madison, WI; [4]Department of Statistics, University of Wisconsin-Madison, Madison, WI; [5]Department of Biomolecular Chemistry, University of Wisconsin-Madison, Madison, WI; [6]McArdle Laboratory for Cancer Research, University of Wisconsin-Madison, Madison, WI; [7]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI; [8]Morgridge Institute for Research, Madison, WI

In a drug discovery pipeline, once a disease-relevant protein target has been identified, researchers face the daunting task of identifying chemical compounds that effectively modulate that target. Experimental phenotypic screening of thousands or millions of small molecules is time-consuming and expensive, whereas virtual (computational) screening can provide a small set of promising molecules that are more likely to be active towards the target protein. It acts as a pre-processing step for filtering the extremely large number of candidate chemicals. Among virtual screening methods, deep learning has become popular recently. It can benefit from fully exploring complex, non-linear relationships among chemicals' features. Our goal is to critically evaluate deep learning versus established virtual screening methods to see if the hype translates to real-world utility in this domain. We focus on the SSB-PriA target, a protein-protein interaction, and analyze four classes of virtual screening methods: influence relevance voter, structure-based docking, single-task learning, and multi-task learning. We compare these methods in a real-world setting by assessing their ability to prioritize active compounds in an untested set. We also argue that the most popular evaluation metric in this domain, area under the ROC curve, can be misleading and compare it with other evaluation metrics, showing which provide real-world value. Moreover, we present a user-friendly framework for virtual screening tasks based on Keras, a neural network library built on top of Theano and Tensorflow.

Acknowledgements:
The authors acknowledge GPU hardware from NVIDIA and support from the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education, the Morgridge Institute for Research, and the NIH BD2K grant U54 AI117924.