Practical model selection for virtual chemical screening

Shengchao Liu[1,2], Moayad Alnammi[1,2], Spencer S. Ericksen[3,4], Andrew Voter[5], James Keck[5], F. Michael Hoffmann[4,6], Scott A. Wildman[4], Anthony Gitter[1,2,3,7]

[1]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI; [2]Morgridge Institute for Research, Madison, WI; [3]Center for Predictive Computational Phenotyping, University of Wisconsin-Madison, Madison, WI; [4]Small Molecule Screening Facility, University of Wisconsin Carbone Cancer Center, 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

Virtual (computational) high-throughput chemical screening provides a strategy for prioritizing compounds for experimental screens. The optimal virtual screening algorithm depends on the dataset and evaluation strategy. We consider a wide range of ligand-based machine learning and docking-based approaches for virtual screening on two protein-protein interactions, SSB-PriA and RMI-FANCM, and present a strategy for choosing which algorithm is best for prospective compound prioritization. Our workflow identifies a random forest as the best algorithm for our targets over more sophisticated neural network-based models. The top 250 predictions from our random forest model recover 41 of the 84 active compounds from a library of 25,279 molecules assayed on SSB-PriA. We show that virtual screening methods that perform well in public datasets and synthetic benchmarks, like multi-task neural networks, do not always translate to wet lab prospective screening performance. In addition, we are exploring new machine learning ensembling strategies and chemical representations based on these results. Finally, we are experimentally testing whether the predictive performance generalizes when prioritizing millions of chemicals.

Acknowledgements:
The authors acknowledge GPU hardware from NVIDIA, support from the Center for High Throughput Computing, and funding from the Center for Predictive Computational Phenotyping NIH U54 AI117924, the University of Wisconsin Carbone Cancer Center Support Grant NIH P30 CA014520, the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education, and the Morgridge Institute for Research.