CONTACT NAME
Edward Lowe: edward.w.lowe@vanderbilt.edu

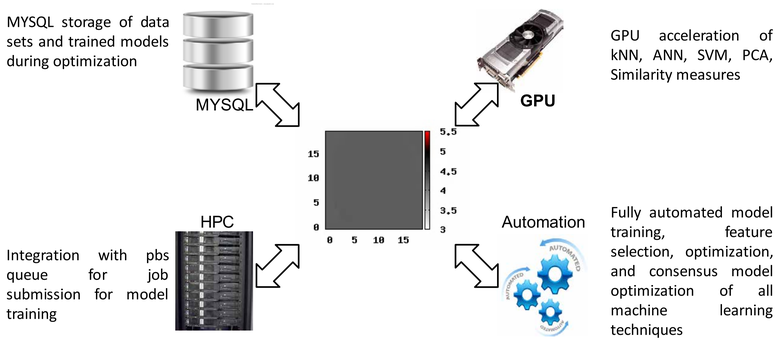
BCL::CHEMINFO – GPU-Accelerated Cheminformatics Suite for Probe Development and Drug Discovery

Edward W. Lowe, Jr., Mariusz Butkiewicz, and Jens Meiler (www.meilerlab.org)

Departments of Chemistry, Pharmacology, and Biomedical Informatics; Center for Structural Biology and Institute of Chemical Biology; Nashville, TN 37232, USA

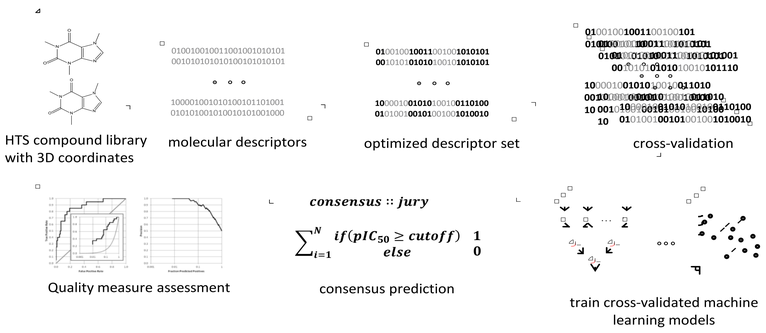
***Abstract* — With the rapidly increasing availability of High-Throughput Screening (HTS) data in the public domain, such as the PubChem database, methods for Ligand-Based Computer-Aided Drug Discovery (LB-CADD) or 'cheminformatics' have the potential to accelerate, reduce cost, and increases quality of probe development and drug discovery efforts. Prioritizing compounds for experimental screening from the 10^7 known and available drug-like compounds and for synthesis from the estimated space of 10^{30}-10^{60} small molecules is particularly important in the resource-limited environment of academia where often rare or neglected diseases are targeted. From a biomedical computational science and technology perspective, in a push-pull relation, increased public availability of large HTS data sets enables not only thorough benchmarking of existing LB-CADD methods but stimulates the development of innovative new LB-CADD tools that should than be applied in academic research. Here, we present such a tool. BCL::CHEMINFO is a cheminformatics framework featuring GPU acceleration, MYSQL integration, and automation of model optimization. This pipeline allows for the rapid construction of highly predictive quantitative structure activity relationship (QSAR) models for drug design. Here we present several current studies leveraging BCL::CHEMINFO against targets indicated in cancer (pancreatic), malaria, and neuroscience (schizophrenia, fragile X syndrome, Parkinson's).**

BCL::CHEMINFO Framework:



MATERIALS AND METHODS:

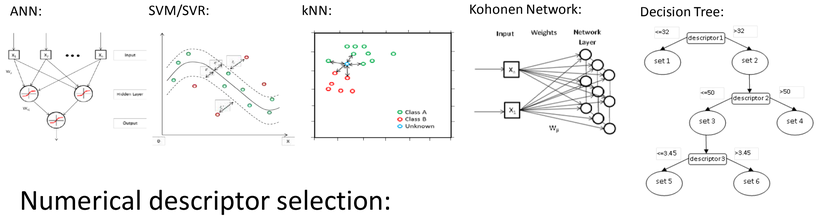
QSAR / QSPR model development:



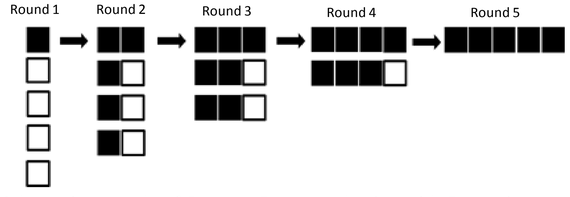
$$consensus :: jury$$

$$\sum_{i=1}^{N} \begin{matrix} if(pIC_{50} \geq cutoff) & 1 \\ else & 0 \end{matrix}$$

Applied Machine learning Algorithms:

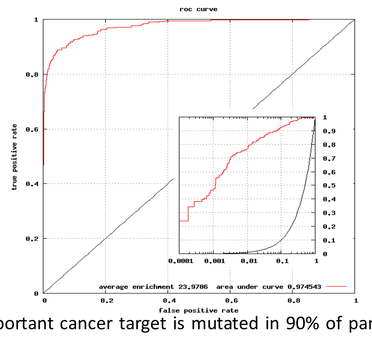


Numerical descriptor selection:

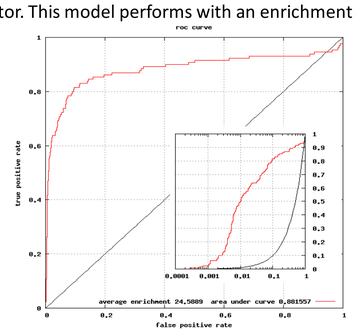


Schematic example of Forward Feature selection with 5 descriptor groups. models are trained for each machine learning technique during this process as $cv * \frac{n(n+1)}{2}$ where *cv* is the cross validation number and *n* is the number of feature categories (60).

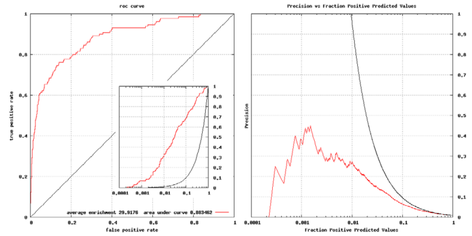
RESULTS:



K-Ras Inhibition: This important cancer target is mutated in 90% of pancreatic cancers, 50% of colon cancers, and 40% of breast cancers. It has long been targeted but has been classified as undruggable by most researchers in the field. BCL::CHEMINFO is being leveraged to construct models trained on NMR fragment-screening results. The models show high predictive power as is illustrated by the receiver operator characteristic curve (ROC) in which the magnitude of the slope of the initial curve is indicative of the predictive power of the model whereas a slope of 1 indicates a random predictor. This model performs with an enrichment of 24.

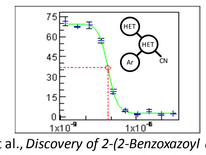


Replication Protein A: This cancer target affects all cancer cells by preventing their DNA repair machinery from working against chemotherapeutics. Inhibition would have a profound affect on the effectiveness of chemotherapeutics. In collaboration with Professor Chazin, renowned structural biologist, this project improves initial molecules identified by experimental nuclear magnetic resonance methods by optimizing the interactions required for binding to the target by utilizing. Enrichment achieves is 25.

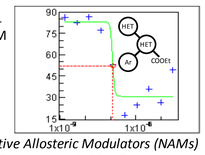


Malaria: This tropical parasitic disease causes high fevers, flu-like symptoms, and anemia. Annually, there are 250 million cases of fever symptoms and 1 million deaths, often in children. This parasite digests hemoglobin, found in the blood, and releases heme. The parasite crystallizes this heme to hemozoin to prevent heme toxicity which would kill the parasite. BCL::CHEMINFO is being leveraged to design inhibitors of this crystallization process which will ultimately kill the parasite indirectly. Both internal HTS data as well as publically available data are being utilized in this project (experimental data on over 1 million molecules). Enrichment achieved is currently 33.

VU0240790-4
EC_{50} = 75 nM



VU0360620-1
EC_{50} = 124 nM



Mueller, R., et al., *Discovery of 2-(2-Benzoxazoyl amino)-4-Aryl-5-Cyanopyrimidine as Negative Allosteric Modulators (NAMs) of Metabotropic Glutamate Receptor 5 (mGlu5): From an Artificial Neural Network Virtual Screen to an In Vivo Tool Compound.* ChemMedChem, 2012. **7**(3): p. 406-414.

Metabotropic Glutamate Receptors: This target belongs to a family of receptors known as G-Protein Coupled Receptors (GPCRs). There is little experimental structural data for these receptors. These particular GPCRs are indicated in many neurological disorders such as schizophrenia, Parkinson's, and Fragile X Syndrome. An experimental HTS was performed on 180k molecules at the Vanderbilt Center for HTS using mGluR subtype 4 and subtype 5 as targets. Negative and positive modulation of these receptors has differing effects, both of which can be desired depending upon the neurological condition. Initial computational work on this system using BCL::CHEMINFO has led to the development of molecules which are now used ***in vivo*** as probes. The high performance models generated enabled the elucidation of novel chemical structures completely different from anything known to elicit the desired modulation. Specifically, a virtual HTS was performed *in silico* and molecules suggested. This data set of biological results has recently been updated as the Conn lab has performed further HTS screens. The models are updated iteratively as new biological data becomes available to continuously improve performance.

GPU Speed-ups:

ML Method	12k Molecules	71k Molecules	210k Molecules
ANN	109	115	114
SVM	35	29	32
KNN	18	29	68

Similarity Measure	3500 Molecules	5500 Molecules	1000 Molecules
Tanimoto	265	267	170
Cosine	235	**283**	175
Dice	260	269	195
Euclidean	138	186	230
Manhattan	100	108	160

GPU-Acceleration: The GPU acceleration achieved in this work using the GTX 480 is enabling rapid production of highly predictive models which would otherwise be computationally prohibitive by traditional methods. The entire workflow has seen speed-ups of orders of magnitude allowing more thorough cross validation and feature selection methods. This directly translates, as is evidenced by our work on mGluR, the elucidation of novel chemical entities which can elicit the desired effects on biological targets of interest. This technology is changing the way drug discovery is performed.

ACKNOWLEDGEMENT:

This work is supported by 1R21MH082254 and 1R01MH090192 to Jens Meiler. Edward W. Lowe, Jr. acknowledges NSF support through the CI-TraCS Fellowship (OCI-1122919). The authors thank the Advanced Computing Center for Research & Education at Vanderbilt University of hardware support.