Progress In Computational Toxicology

Sean Ekins[1,2]

[1]Collaborations in Chemistry , 5616 Hilltop Needmore Road, Fuquay Varina, NC27526, USA, NC. [2]Collaborative Drug Discovery, Burlingame, CA.

Abstract

Computational methods have been widely applied to toxicology across pharmaceutical, consumer product and environmental fields over the past decade. Progress in computational toxicology is now reviewed. A literature review was performed on computational models for hepatotoxicity (e.g. for drug-induced liver injury (DILI)), cardiotoxicity, renal toxicity and genotoxicity. In addition various publications have been highlighted that use machine learning methods. Several computational toxicology model datasets from past publications were used to compare Bayesian and Support Vector Machine (SVM) learning methods. The increasing amounts of data for defined toxicology endpoints have enabled machine learning models that have been increasingly used for predictions. It is shown that across many different models Bayesian and SVM perform similarly based on cross validation data. Considerable progress has been made in computational toxicology in a decade in both model development and availability of larger scale or 'big data' models. The future efforts in toxicology data generation will likely provide us with hundreds of thousands of compounds that are readily accessible for machine learning models. These models will cover relevant chemistry space for pharmaceutical, consumer product and environmental applications.

Methods

Datasets: Molecule datasets were extracted from various publications (Table 1).

Machine learning models: Laplacian-corrected Bayesian classifier models were developed using Discovery Studio 3.5 (Biovia, San Diego, CA). The models were generated using the following molecular descriptors: molecular function class fingerprints of maximum diameter 6 (FCFP_6), AlogP, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular fractional polar surface area which were all calculated from input sdf files. The resulting datasets were validated using leave-one-out cross-validation, 5 fold validation to generate the receiver operator curve area under the curve (ROC AUC). SVM, RP Forest and single tree models were built with the same molecular descriptors in Discovery Studio. For SVM models we calculated interpretable descriptors in Discovery Studio then used Pipeline Pilot to generate the FCFP_6 descriptors followed by integration with R. RP Forest and RP Single Tree models used the standard protocol in Discovery Studio. In the case of RP Forest models 10 trees were created with bagging. RP Single Trees had a minimum of 10 samples per node and a maximum tree depth of 20. In all cases, 5-fold cross validation (leave out 20% of the database) was used to calculate the ROC for the models generated.

The recently released CDD Models (https://www.collaborativedrug.com/) functionality in CDD Vault was used to build a Bayesian model (with FCFP_6 fingerprints only) for several datasets and used 3 fold cross validation AUC ROC.

Table 1. Comparison of SVM and Bayesian ADME/Tox classification models generated with the same molecular descriptors. NT not tested. Note DILI model used ECFC_6 fingerprint descriptors instead of FCFP_6. *For CDD Models only FCFP_6 descriptors were used; [1]SVM (regression) - 5 fold cross validation q^2 = 0.17; [2]SVM (regression) - 5 fold cross validation q^2 = 0.47.

Model	Reference	DS + R SVM 5 fold cross validation ROC	DS Bayesian 5 fold cross validation ROC	DS Bayesian Leave one out cross validation ROC	DS RP single tree	DS RP forest	CDD Bayesian 3 fold cross validation
DILI (N = 532)	(S. Ekins et al., 2010)	0.88	0.63	0.74	0.56	0.58	0.72*
PXR (N = 312)	(S. Kortagereet al., 2009)	0.81	0.78	0.84	0.69	0.72	-
$5HT_{2B}$ (N = 238)	(Chekmarev, et al., 2008)	0.83	0.82	0.87	0.70	0.70	-
$5HT_{2B}$ (N = 754)	(Hajjo et al., 2010)	0.86	0.91	0.91	0.79	0.82	0.91
hERG (N = 134)	(Chekmarev, et al., 2008)	0.82	0.71	0.74	0.68	0.73	-
hERG (N = 806)	(Wang, et al., 2012)	0.88	0.84	0.87	0.82	0.86	0.84
hERG (N = 305,616)	(Du, et al., 2011)	0.83	0.85	0.86	0.80	0.78	-
BBB (N = 1968)	(Martins, et al., 2012)	0.90	0.91	0.92	0.87	0.91	0.93
AMES (N = 6512)	(Hansen, et al., 2009)	0.86	0.84	0.84	0.81	0.84	0.83
Nephrotoxicity (N =104)	(Lin & Will, 2012)	0.53	0.64	0.65	0.52	0.56	-
Clearance iv (N = 512) [1]	(Gombar & Hall, 2013)	0.73	0.74	0.76	0.71	0.74	-
VDss iv (N = 556) [2]	(Gombar & Hall, 2013)	0.84	0.81	0.87	0.80	0.82	-

Figure 1. Ames Bayesian model built with 6512 molecules (Hansen et al., 2009). A. Features important for Ames actives. B. Features important for Ames inactives.

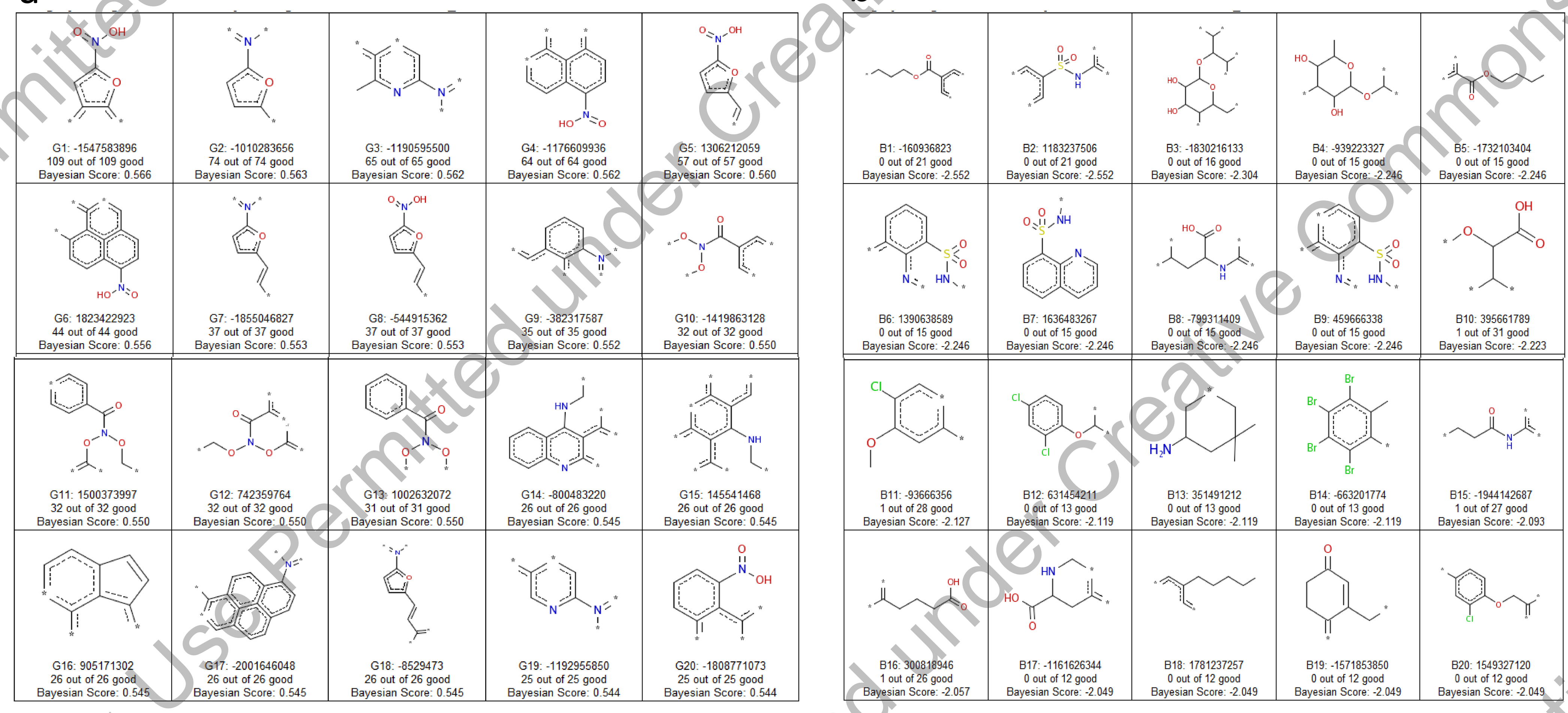
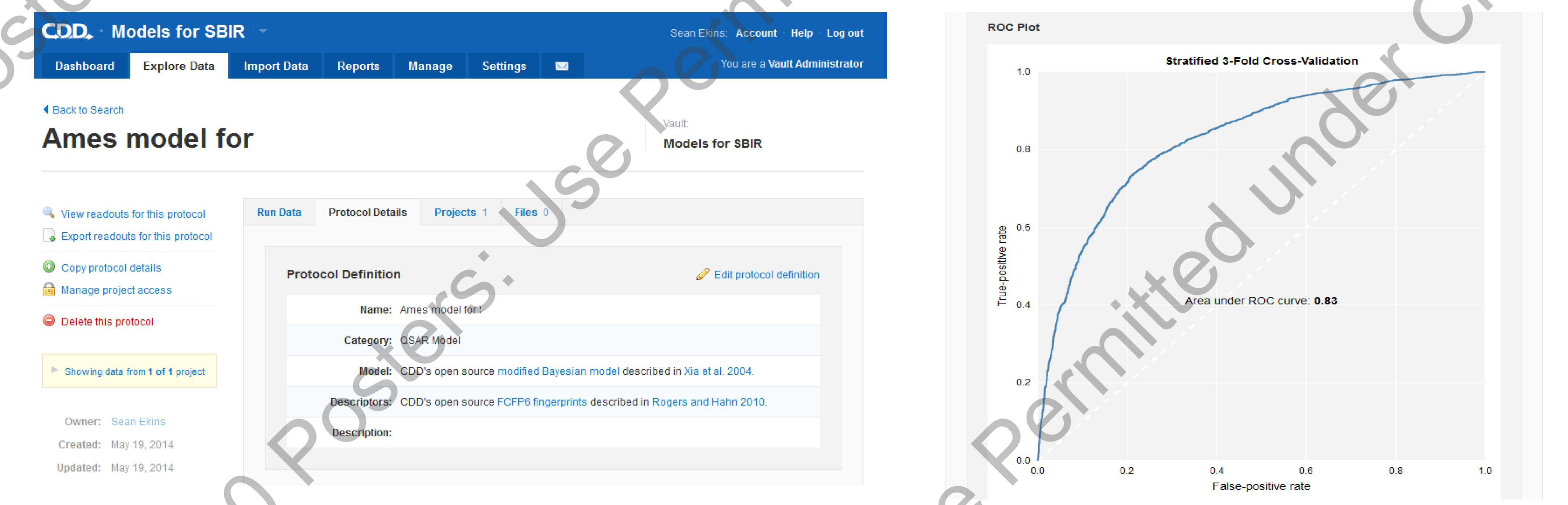


Figure 2. Ames Bayesian model built using CDD Models showing ROC for 3 fold cross validation. Note only FCFP_6 descriptors were used



Results and Discussion

- Several literature datasets were selected to build computational models for hepatotoxicity, cardiotoxicity, renal toxicity, genotoxicity and other ADME properties of interest using multiple machine learning methods.
- There is good agreement in the ROC values across different methods using cross validation (SVM perform slightly better Bayesian in 7 of 12 models) (Table 1).
- Can the models be combined or a consensus used for predictions?
- The Bayesian methods are interpretable by defining toxicophores in the fingerprint features (Fig 1).
- REACH and other new legislation are likely to require even more computational testing of compounds for which "there is insufficient information on the hazards that they pose to human health and the environment".
- There needs to be a first tier prioritization to filter molecules of interest to pharmaceutical, consumer products and environmental researchers.
- Efforts to make these computational toxicology models accessible may aid in their utilization.
- Within a decade we have progressed from models like those for hERG with a handful of molecules to over 300,000 molecules.
- The challenge ensure such large scale models can be used to maximal effect and influence (Ekins 2014).
- A major hurdle is awareness of what models exist, where they reside and how to utilize them for decision making. Therefore inclusion in database such as the CDD Vault may be important (Fig 2).

In conclusion, 'big data' models will promote computational toxicology and further focus *in vitro* and *in vivo* testing on verification of selected predictions.

Future efforts to externally test the models in Table 1 is critical and I welcome interest in assessing them.

References

- Chekmarev, D. S., Kholodovych, V., Balakin, K. V., Ivanenkov, Y., Ekins, S., & Welsh, W. J. (2008). *Chem Res Toxicol, 21*, 1304-1314.
- Du, F., Yu, H., Zou, B., Babcock, J., Long, S., & Li, M. (2011). *Assay Drug Dev Technol, 9*, 580-588
- Ekins, S. (2014). J Pharmacol Toxicol Meth, 69, 115-140.
- Ekins, S., Williams, A. J., & Xu, J. J. (2010). *Drug Metab Dispos, 38*, 2302-2308.
- Gombar, V. K., & Hall, S. D. (2013). *J Chem Inf Model, 53*, 948-957.
- Hajjo, R., et al., J Med Chem, 2010. **53**(21): p. 7573-86.
- Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., & Muller, K. R. (2009). *J Chem Inf Model, 49*, 2077-2081.
- Kortagere, S., Chekmarev, D., Welsh, W. J., & Ekins, S. (2009). *Pharm Res, 26*, 1001-1011.
- Lin, Z., & Will, Y. (2012). *Toxicol Sci, 126*, 114-127.
- Martins, I. F., Teixeira, A. L., Pinheiro, L., & Falcao, A. O. (2012). *J Chem Inf Model, 52*, 1686-1697.
- Wang, S., Li, Y., Wang, J., Chen, L., Zhang, L., Yu, H., & Hou, T. (2012). *Mol Pharm, 9*, 996-1010.

Acknowledgments

S.E. gratefully acknowledges Biovia (formerly Accelrys) for providing Discovery Studio and Dr. Katalin Nadassy for her assistance in using PipelinePilot. S.E. also acknowledges S.C. Johnson & Son, Inc. for discussions and support. NCAT, NIH funded the CDD Models development (9R44TR000942-02, Ekins PI) and CDD colleagues are acknowledged for assistance developing this software.

Collaborations In Chemistry

A division of Collaborations In Communications, Inc.