DMD # 34918

USING OPEN SOURCE COMPUTATIONAL TOOLS FOR PREDICTING HUMAN METABOLIC STABILITY AND ADDITIONAL ADME/TOX PROPERTIES

Rishi R. Gupta, Eric M. Gifford, Ted Liston, Chris L.Waller, Moses Hohman, Barry A. Bunin and Sean Ekins

Pfizer Global Research and Development, Eastern Point Road, Groton, CT 06340, U.S.A. (RRG, EMG, TL, CW).

Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA. 94010, U.S.A. (BAB, MH, SE).

Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, U.S.A (SE).

Department of Pharmaceutical Sciences, University of Maryland, MD 21201, U.S.A (SE).

Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ)-Robert Wood Johnson Medical School, 675 Hoes lane, Piscataway, NJ 08854, U.S.A. (SE).

Running title page

a. Running Title: Open source tools for ADME/Tox

b. Corresponding Author:

Sean Ekins M.Sc., Ph.D, D.Sc.

Collaborative Drug Discovery,

601 Runnymede Ave, Jenkintown, PA 19046

Phone 215-687-1320

Email ekinssean@yahoo.com, sekins@collaborativedrug.com

c. Number pages

Text pages:	20
Tables:	6
Figures:	2
References:	63
Words in Abstract:	250
Words in Introduction:	703
Words in Discussion:	1810

d. Non standard abbreviations: ADME/Tox, absorption, distribution, metabolism, excretion and toxicity; CDK, chemistry development kit; HLM, human liver microsomes; MDCK, Madine Darby Canine Kidney; PCA, principal component analysis; P-gp, P-glycoprotein; PPV, positive predicted value; QSAR, quantitative structure activity relationship; RRCK, Russ Ralph Canine Kidney;

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Abstract

Ligand-based computational models could be more readily shared between researchers and organizations if they were generated with open source molecular descriptors (e.g. chemistry development kit, CDK) and modeling algorithms, as this would negate the requirement for proprietary commercial software. We initially evaluated open source descriptors and model building algorithms using a training set of approximately 50,000 molecules and a test set of approximately 25,000 molecules with human liver microsomal metabolic stability data. A C5.0 decision tree model demonstrated that CDK descriptors together with a set of SMARTS keys had good statistics (Kappa = 0.43, sensitivity = 0.57, specificity 0.91, positive predicted value (PPV) = 0.64) equivalent to models built with commercial MOE2D and the same set of SMARTS keys (Kappa = 0.43, sensitivity = 0.58, specificity 0.91, PPV = 0.63). Extending the dataset to ~193,000 molecules and generating a continuous model using Cubist with a combination of CDK and SMARTS keys or MOE2D and SMARTS keys confirmed this observation. When the continuous predictions and actual values were binned to get a categorical score we observed a similar Kappa statistic (0.42). The same combination of descriptor set and modeling method was applied to passive permeability and P-gp efflux data with similar model testing statistics. In summary, open source tools demonstrated comparable predictive results to commercial software with attendant cost savings. We discuss the advantages and disadvantages of open source descriptors and the opportunity for their use as a tool for organizations to share data precompetitively, avoiding repetition and assisting drug discovery.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Introduction

Problems associated with late stage failures of potent lead compounds in the pharmaceutical industry due to undesirable physicochemical properties has lead to a shift in the drug discovery protocols for well over the past decade. Pharmaceutical companies increasingly evaluate lead compounds for drug-like properties very early on in the discovery process using computational prediction methods that are based on statistical techniques utilizing experimental data from *in vitro* or physicochemical property assays (Ekins et al., 2000a). Well validated ligand-based *in silico* approaches are important and exist in the large pharmaceutical companies because these organizations have large diverse proprietary data sets, the financial resources for expensive commercial software and access to in-house computational, medicinal chemistry and high-throughput screening expertise. All these enablers are generally or in part lacking in academia, small biotechnology companies and non-profit neglected disease foundations.

The screening for absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) properties of molecules can be done using *in vitro* and *in vivo* methods, but they are not cost effective to perform for very large numbers of compounds. Instead, *in silico* techniques to predict these properties can be used and only those compounds that look likely to advance as lead molecules can be screened using *in vitro* and *in vivo* techniques. This can also lead to implementation of an Active Learning paradigm (Gupta and Gifford, 2009) where we can use a computational model to make decisions whether one wants to screen every compound or not (Figure 1). The primary limitation of such computational methods today is the absence of optimal training sets which adequately cover chemical space (as they use small literature datasets, low quality datasets or

combine disparate datasets). When using large proprietary datasets, the derived models are not publically available. The computational approaches are inherently only as good as the underlying data from which they are derived. If the models could be improved by leveraging more quality *in vitro* data and the methods could be widely used and understood by experimentalists as well as by computational scientists, it is evident that the results would be of enormous value to the entire drug discovery ecosystem, both industrial and academic.

Several ADME/Tox methods had been proposed at least a decade ago and the application or comparison of these programs has been extensively studied (Ekins et al., 2000b; Ekins et al., 2001a; Ekins, 2007; Villoutreix et al., 2007; Lagorce et al., 2008). The challenge is not only that sizeable drug discovery data (training sets) are lacking for model building, but there has been no mechanism to bring together isolated training sets, especially the very large proprietary data sets from different companies. If the sensitive intellectual property contained in the training sets could be obfuscated, pharmaceutical organizations would often want to share these models with collaborators and academics working on important neglected diseases, for example. There have been some efforts in understanding how chemical information can be shared without directly sharing structures, for example, fingerprints should be avoided and low levels of precision in numeric descriptors and feature count descriptors may be ‘fuzzy’ enough to protect the structure identity (Masek et al., 2008).

Software developed under the open source license provides important visibility into the implementation of descriptors and algorithms, so that computational chemists can verify the algorithm and suggest or actually contribute improvements (Guha et al., 2006).

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

A number of open source software packages exist that calculate molecular descriptors (Melville and Hirst, 2007; Sykora and Leahy, 2008) or implement modeling algorithms (e.g. R). Some groups have also used open descriptors and open modeling algorithms to build quantitative structure activity relationship (QSAR) models (Guangli and Yiyu, 2006; Melville and Hirst, 2007; Guha, 2008) for mutagenicity, cytotoxicity, Caco-2 data as well as some drug targets. The datasets used have been relatively small to date (low thousands of molecules). While there are some open toolkits for cheminformatics and bioinformatics (Guha et al., 2006; Steinbeck et al., 2006; Spjuth et al., 2007; Spjuth et al., 2009) as well as proposed web services (Dong et al., 2007), no integrated open toolkit exists at the time of writing. In the current study we evaluate how some open descriptors and algorithms perform versus commercial software for generating ADME/Tox models with very large datasets produced at Pfizer.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Materials and Methods

Datasets. All compounds tested were synthesized at Pfizer as part of drug discovery projects. The dataset size generally exceeded sixty thousand compounds for each assay. The datasets were binned as per the guidance provided by experts in the Pharmacokinetics, Dynamics and Metabolism (PDM) business unit. In the work presented here we have primarily evaluated multiple datasets which all used Pfizer in-house compound screening. The datasets described in this work are: Human Liver Microsomal Stability (HLM), Passive Permeability (RRCK) and P-glycoprotein (P-gp) efflux activity (MDR). We also briefly present one case using literature solubility data.

The human liver microsomal stability dataset has over two hundred thousand compounds and covers a diverse range of chemistry as well as diversity coverage of therapeutic areas in which these compounds have been developed. This assay allows the measurement of the apparent intrinsic clearance (Cl_{int}) of a compound in human liver. The dynamic range of clearance is low (Cl_{int} < 13µl/min/mg), moderate (13 < Cl_{int} < 50µl/min/mg) and high (Cl_{int} > 50µl/min/mg) (Table 1). A 3 bin classification model as well as a continuous model on the full dataset was built. The distribution of the data in each class in this and the other datasets is shown in Table 2.

The permeability dataset which has over seventy thousand compounds is a cell based assay, which represents cellular passive apparent permeability (P_{app} as the end-point). The P_{app} values are rates (expressed as 10^{-6} cm/sec), the higher the value, the faster the compound crosses the cell monolayer. The dynamic range for P_{app} is described

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

as follows: P_{app}<2.5cm/sec (low), 2.5cm/sec < P_{app}< 10cm/sec (moderate), P_{app}>10cm/sec (high) (Table 1). For passive permeability a 3 bin classification model was built.

The P-gp efflux activity dataset has over sixty thousand diverse compounds. This is also a cell based assay, which is used to assess P-glycoprotein (P-gp) efflux activity, measured by cell line transfected with the human MDR-1 gene. The cell line is used in a bidirectional evaluation of permeability (apical to basolateral (A to B) and basolateral to apical (B to A)) generating a final (B to A)/(A to B) ratio which can be used to determine if there is asymmetry in the flux due to transporter activity. A compound is considered to be effluxed if its (B to A)/(A to B) ratio is 2.5 or greater in any of the individual cell lines (Table 1). For this particular endpoint a 2 bin classification model was derived. Thus these three rich datasets provide us with a very good starting point for our analysis where we can test the metric for the model prediction.

As a test for our strategy to demonstrate the applicability and equality of models from open source descriptors and modeling methods we also applied the same set of descriptors as applied to the above mentioned datasets and modeling methods to a public domain data set, namely Huuskonen's aqueous solubility dataset (Huuskonen, 2000). This publication describes the application of neural networks and linear regression with topological indices and E-state descriptors on a set of ~1300 diverse organic compounds.

Descriptors. We have used different descriptors such as the Pfizer modified Molecular Operating Environment 2D set (MOE2D, 2008) and CDK (http://cdk.sourceforge.net/) (Steinbeck et al., 2006) fingerprints etc. For each of the datasets the same descriptors were calculated i.e. for RRCK, HLM etc. we always calculated Pfizer modified MOE2D

descriptors (463), CDK descriptors (195) and SMARTS Keys (355). There was no descriptor pruning or feature selection performed because the *in silico* QSAR models discussed here are routinely updated with new screening data and recalculating descriptors for ~100,000 compounds each time would be too computationally expensive. To avoid issues where we need to add new descriptors to capture new chemotypes, we routinely calculate and test new descriptor sets and add them to the list as necessary.

355 SMARTS keys (a set of SMARTS strings used as count based descriptors put together very carefully by scientists at Pfizer, Inc.) cover a wide range of SMARTS-encoded substructural fragment/feature descriptors. Multiple internal studies at Pfizer have shown that addition of these type of descriptors to the physicochemical based descriptors provide better models than those built solely on physicochemical property based descriptors. We are also constantly evaluating new substructural fragments that could be added to the existing set of SMARTS keys.

Model building. We have made this study as exhaustive and comprehensive as possible by comparing results for a variety of modeling methods such as Random Forest (Liaw and Wiener, 2002), SVM (Chang and Lin, 2001) and Recursive Partitioning (RP) Forest (Scitegic Pipeline Pilot ver. 7.5.2, Accelrys, San Diego, CA) etc.

The key component of this work was to first build either a robust continuous or categorical model which would be able to deal with the diverse datasets. For continuous models the end point data was scaled by taking a base 10 logarithm (Log_{10}) of the data. Using the full data and descriptor matrix, a regression model was built using the Rulequest Cubist ((Rulequest, (Quinlan, 1991)) modeling algorithm or a categorical

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

model was built using the Rulequest C5.0 (Quinlan, 1993) modeling method or other modeling methods such as SVM, Random Forest or RP Forest.

The Rulequest Cubist algorithm can be defined as a piecewise linear modeling method with boosted trees (with an exception that the rules can overlap). It can also construct multiple models (committees) and can combine "rule-based" models with "instance-based" (nearest neighbor) models (http://rulequest.com/cubist-unix.html). The committee models are made up of several rule-based models. Each member of the committee predicts the target value for a case and the members' predictions are averaged to give a final prediction. In the present work we have used 5 instances i.e. the algorithm would look for five closest neighbors to our test compound in the dataset, 20 committees and the default setting of the rules (1000). An important fact to mention here would be that Rulequest Cubist makes the nearest neighbor search based on the Manhattan Distance (also known as city block distance) in the descriptor space. This combination of parameters was found to be the most optimal combination for superior prediction and computational efficiency by multiple in-house studies (results not shown).

For each of the categorical models discussed in this study the training and test sets were put together (Table 2) by using a maximum dissimilarity algorithm which allowed representative subsets of the larger datasets. It is always beneficial to test new techniques with a variety of datasets to make sure that the metric is not assay or end-point dependent where it may be successful in some and fail in others.

Model testing and evaluation. As a standard for determining the quality of the classification models built on datasets described above, the Kappa statistic (Kappa >0.4)

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

was used as a measure of the "predictability" of the model. The Kappa statistic (Carletta, 1996; Cohen, 2003) can be defined as an index which compares the agreement against that which might be expected by chance (Equation 1). Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement).

$$\kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$$ Equation 1.

where, Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance), then $\kappa \leq 0$.

For continuous models R^2 (the square of the sample correlation coefficient between the outcome and the values being used for prediction) and Root Mean Squared Error (RMSE) were evaluated on test datasets as a quality measurement. R^2 provides information on the goodness of the fit i.e. how well the regression line approximates the real data points (Equation 2). The value is between 0 and 1 where 0 being no correlation and 1 being a perfect correlation or in other words $R^2=1$ indicates that the regression line perfectly fits the data.

$$R^2 = \left(1 - \frac{SS_{err}}{SS_{tot}}\right)$$ Equation 2

where,

SS_{err} is the residual sum of squares and
SS_{tot} is the total sum of squares

Root mean square error (RMSE) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the object being modeled or estimated. RMSE is a good measure of precision.

Chemical space analysis. A visualization of the chemical space covered by a test and training set can be created using principal component analysis (PCA) (Zientek et al., 2010). Such a visual chemical space map was generated for the HLM dataset by converting the CDK and SMARTS descriptors (total ~579 descriptors) into principal components (PC). The PCA component in Pipeline Pilot (Accelrys, San Diego, CA) was used for this calculation. First we calculated the PC's for the training set i.e. a matrix of ~193,000 compounds and 579 descriptors and then we calculated the PC's for the test set (2300 compounds and 579 descriptors). These PCs for the training and test set were used to produce a scatterplot with Spotfire (TIBCO, Somerville, MA).

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Results

Model testing and evaluation: human liver microsomal stability. Initially using approximately 50,000 molecules with human microsomal metabolic stability data (Table 2 and 3) with the C5.0 decision tree building software we have clearly demonstrated that a combination of open CDK and SMARTS keys are equivalent to models built with a combination of MOE2D and SMARTS keys based on the statistics presented. Recent data extended this model with over 193,000 molecules and testing on approximately 2,300 molecules (Table 4). We were able to build a regression model and a classification model. To make sure that each model is predictive we had split the data into a training set (80% of the total data) and a test set (20% of the total data) using venetian blind splitting method (Davis et al., 2006) which allows retention of the identical data distribution for the test set and the training set. As an example, to perform a 80-20 split, every 5^{th} compound in an activity sorted list can be added to the test set and the remaining data becomes the training set. Also, since HLM is a high throughput assay, we get about ~1500-2000 compounds screened every 2 weeks and hence the newly screened compound list was used as a blind test set. As shown in Table 4, the R^2 on the 20% test set is ~0.7 which is a relatively high correlation coefficient for a dataset of this magnitude. Also, RMSE was 0.291 which is good as the model didn't find too many errors in prediction that could be penalized. When we observe the performance of this model on the blind dataset (which usually have new chemotypes not represented in the training set or the 20% set-aside test set at all), we see a Pearson's correlation of 0.53 which is also a reasonable value for such a dataset as it probably lies partially outside of the applicability domain of the current training set used to train the model. When comparing these results

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

between the 2 different sets of descriptors we can clearly observe that there is no difference in the quality of the 2 models. The advantage of using CDK descriptors with SMARTS keys is that we have reduced our total descriptors from 818 (MOE2D and SMARTS) to 550, which allows us to reduce the dimensionality and hence significantly improve the speed of the calculations (for the predictions this scales linearly with number of descriptors) (Table 4). Another test was done to check the performance of the regression model against the categorical (or binned) model that was previously built on the same dataset. We binned the actual and predicted values and then compared the statistics for models built using the 2 sets of descriptors and as shown, the results were very similar, where we achieved a kappa statistic of 0.4 or higher (Table 4). This suggested comparable data could be generated between all open descriptors and commercial descriptors with HLM.

Model testing and evaluation: RRCK passive permeability. For RRCK passive permeability (Table 5) we built a categorical model which allowed a compound to be predicted as high risk, moderate risk or low risk based on the criteria provided to us by PDM domain experts. For in house applications we only use a continuous model but for this study we built only categorical models to compare the model performance for different combinations of modeling methods and descriptors (Table 5). The results are very promising for the C5.0 modeling method with MOE2D and SMARTS keys as descriptors. This was our baseline as a very similar continuous version of this model is what has been implemented in-house for the research community within Pfizer. The idea at this point was to provide an *in silico* model which is either equivalent or better than the

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

baseline model. Going through various combinations of modeling methods and descriptors we were able to identify the C5.0 modeling method when combined with CDK and SMARTS keys as descriptors performed approximately the same as our baseline model. Other model and descriptor combinations either did not complete due to memory intensiveness of the calculations or they were not comparable to the baseline.

Model testing and evaluation: P-gp efflux data. Another test case was chosen using P-gp efflux data where the data was segregated into 2 bins i.e. low risk and high risk as recommended by colleagues in the PDM group. As shown in the Table 6, a similar exercise was performed as described above for RRCK passive permeability where a baseline model and descriptor combination was chosen and then we tested a variety of modeling methods and descriptor combinations to find out a better or equivalent combination. Once again the results were encouraging as we observed that predictions for the C5.0 models built with either the MOE2D and SMARTS Keys or CDK and SMARTS Keys were about the same as the baseline model (Table 6).

There may be concerns that the Kappa values are not exactly similar between all the datasets but based on the dataset size and variation in the type of descriptors, the differences between Kappa for the baseline model and a comparable model is similar. Moreover with a reduced number of descriptors we always have the advantage of less computationally intensive calculations.

Model testing and evaluation: aqueous solubility dataset. Further proof of this modeling approach has used smaller datasets from the public domain for benchmarking

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

such as the Huuskonen aqueous solubility dataset for regression modeling (Huuskonen, 2000) (using over 1000 molecules for training and over 200 for testing) which suggested comparable data to that published $R^2 = 0.92$ (data not shown). This illustrates that we can use open descriptors and model building algorithms to build models with equivalent predictions to those with commercial descriptors for both large and small classification and continuous datasets.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Discussion

Computational QSAR models are primarily based on proprietary software, training data, and descriptors and are stored in proprietary file formats. These models are locked into a particular set of prerequisites and intellectual property restrictions that cannot be replicated anywhere else. The organizations best able to leverage QSAR modeling today are large pharmaceutical organizations, which have the resources to generate their own extremely large (100's of thousands of compounds with) high-quality, diverse training sets and are able to standardize on expensive proprietary modeling software across their organizations while then deploying their models on the intranet. One area of focus for at least a decade has been ADME/Tox modeling and high-throughput screening which has now resulted in the very large numbers of compounds and data available for modeling using machine learning methods, like those described here.

There is also considerable discussion about how to evaluate computational models in other areas (Group, 2004; Dearden et al., 2009) yet there are no clear standard methods for evaluating model robustness for ADME/Tox properties. Widely accepted approaches for model validation and testing such large leave out groups at random many times (also known as X-Fold Cross Validation, where X is the times you want to leave a subset out to test the model (e.g. leave out 20% or leave out 50% 100 times)) or external test sets (Tetko et al., 2008; Zheng et al., 2009) are commonly used. We could define acceptable models (depending on the endpoint) that could predict correct classes >70% or correlations for an external test set that were statistically significant using a Pearson or

Spearman Coefficient. We have described several metrics of model quality that could also be used including the Kappa value.

In this study we have focused on models for some key ADME/Tox endpoints (namely, metabolic stability in HLM, passive permeability and P-gp efflux), for which Pfizer has very large proprietary datasets. There have been several computational models of metabolic stability in the literature. For example, a recursive partitioning model containing 875 molecules with HLM metabolic stability was used to predict and rank the clearance of 41 drugs (Ekins, 2003). A k-nearest neighbour model of metabolic stability data using human S9 homogenate for 631 diverse molecules was able to adequately classify metabolism of a further set of over 100 molecules (Shen et al., 2003). Partial Least Squares Regression (PLS) QSAR models developed with molecular structure descriptors from QikProp (Jorgensen, 2004) and DiverseSolutions software were used with a set of 130 calcitriol analogs (Jensen et al., 2003). The latter model was used to select 20 molecules were selected for *in vitro* testing with 85 % success rate (Jensen et al., 2003). To our knowledge the current study with over ~193,000 compounds is likely the largest validated metabolic stability model to date. While there have been a vast number of models generated for ADME/Tox and physicochemical properties such as solubility (Cheng and Merz, 2003; Lind and Maltseva, 2003; Yamashita et al., 2006) even these models have not used such large numbers of compounds. Passive Permeability has also been the focus for extensive modeling for many years initially based on Caco-2 or MDCK cell data (Segarra et al., 1999; Ekins et al., 2000b; Ren and Lien, 2000; Stenberg et al., 2000; Ekins et al., 2001b). These models generally do not take account of the role of efflux transporters so there has been a parallel effort to build various computational

models for the major transporters in the intestine such as P-gp (Ekins et al., 2002; Xue et al., 2004; Pleban et al., 2005; Chang et al., 2006; Ekins et al., 2007). This study also likely uses the largest available training sets available in the industry to our knowledge for these two endpoints. Our model predictions could be combined to provide a more complete picture of absorption. Ultimately other efflux and uptake transporters (Ekins et al., 2007) should also be modeled to account for outliers.

The results of this study may provide a starting point for a validated universal framework for enabling the sharing of ADME/Tox models and facilitating their use for making predictions by third parties, without the requirement of sharing sensitive molecule structure data. It remains to be seen how well the descriptors used mask the structure identity and further studies will require assessment of this like that performed with other descriptors (Masek et al., 2008). While we have not described the actual sharing of models or the format for doing this, it would require testing to ensure compatibility and reproducibility between laboratories. These open models could be integrated into other software that would allow their selective sharing with selected users. One could readily integrate open molecular descriptors from the CDK, an LGPL Java cheminformatics library that is used in a wide variety of academic and commercial tools (Steinbeck et al., 2006) with modeling algorithms from the GPLed statistical software package, R (http://www.r-project.org). The CDK supports integration with R (Guha et al., 2006; Guha, 2008) so these two tools provide a promising starting point. An alternative open algorithm source is Weka (Frank et al., 2004) which is also widely used.

To be of further value such models will require measures of prediction confidence and applicability domain which will assist the user in interpreting model predictions. A

combination of Tanimoto similarity, PCA, clustering, Mahalanobis distance have been used to determine prediction confidence (Sheridan et al., 2004; Dimitrov et al., 2005; Ekins et al., 2006a; Tetko et al., 2006; Chekmarev et al., 2008; Kortagere et al., 2008; Tetko et al., 2008; Chekmarev et al., 2009; Kortagere et al., 2009). A prediction should not be provided if the test molecule is too far away from a training set member as defined by the user based on a combination of distance as well as similarity metric of choice. A prototype measure of prediction confidence is already in place for continuous models developed in this study, (Gupta and Gifford, 2009). This confidence metric has a sound statistical foundation where the metric captures both error in prediction as well as distance (similarity) to the neighbors in the chemical space as defined by the descriptor used in the model. We had previously used this confidence metric to establish an *in silico* screening strategy which also leads to active learning implementation (data not shown). The idea behind this work is to use *in silico* models to screen compounds that must get *in vitro* data and prevent compounds that do not need to be screened if an *in silico* model can already predict the value with very high confidence (Figure 1). This would allow significant cost savings by not screening each compound. A recent study by us suggested an approximately 30% saving in *in vitro* testing by implementing computational models (Zientek et al., 2010). Clearly we should also be cognizant of the chemical space coverage of the model. For example we have visualized the large dataset of > 193,000 HLM data using a PCA analysis (Figure 2), the majority of the > 2000 test compounds overlap the training set but there are compounds that could be considered outside of the training set and their removal may improve predictions.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Part of the problem with some QSAR approaches is that a model output is often not inherently interpretable by other researchers. When the models are black boxes, the outputs will not be widely embraced. There have also been efforts made at ADME data visualization (Stoner et al., 2004a; Stoner et al., 2004b; Ekins et al., 2006b; Maniyar et al., 2006; Yamashita et al., 2006; Yamashita et al., 2008). Expanding these approaches to show outputs from multiple computational models in a color coded or symbolic manner will require significant innovation to balance information complexity with intuitive graphical representations. Development of truly novel, simple and interpretable visualization methods is not trivial, yet long overdue and this could be based around open source ADME/Tox models developed in a similar manner to those described in the current study. Some thought should also be given as to how the ADME/Tox models are used e.g. early stages of compound library development to minimize the number of compounds synthesized and improve their ADME/Tox properties, versus later in discovery to try to solve problems with these same properties.

This work could be greatly expanded in future to build a community of models with data from a consortium of leading industry and academic partners by validating their quality and predictability. This would represent precompetitive data and we would need to ensure that molecular structures could not be distinguished or reverse engineered from the training sets and descriptors upon which the models were built (Masek et al., 2008). ADME/Tox programs have been traditionally limited in using the same very small datasets from the literature or combining datasets from different groups. These datasets also only cover a small region of chemical space focused on drug-like molecules which tend to be compliant with the rule of 5 (Lipinski et al., 1997). This may not be ideal as it

could be too restrictive, considering there are many example of FDA approved drugs that fail these rules and others. Thus, there is a need for building models using data from various pharmaceutical and biotechnology companies, and then securely sharing the models with collaborators or groups designated by the user. The advantage of using such data from pharmaceutical and biotech companies is that they have generally screened orders of magnitude more data (e.g. tens to hundreds of thousands of compounds under standardized conditions) than is in the public domain and thus have far better coverage of chemistry space. There is of course a trade off here between small local models useful for lead optimization (and limited chemical coverage), and large global models that may be useful for library screening and filtering (greater chemical coverage but less likelihood of distinguishing differences between similar structures).

Finding a quality dataset is still an issue since most of the experimental studies on large datasets to derive ADME/Tox properties are still performed by pharmaceutical companies and the data is inaccessible (Ekins and Williams, 2010). Some forums such as www.cheminformatics.org, QSAR world (http://www.qsarworld.com/), www.opentox.org, www.openqsar.com. are making an effort to collect these datasets as an open repository for chemoinformatics data as well as toolkits for models and descriptors e.g. CDK (Steinbeck et al., 2006) and Mold2 (Hong et al., 2008).

The beneficiaries of open ADME/Tox models would be those in academia, foundations (e.g. working on neglected diseases like tuberculosis and malaria) and or pharmaceutical companies which could avoid duplicative testing and cover more chemical space. This could certainly result in improved predictions and greater applicability of such models for use by groups with compounds of interest, but with no

idea of their ADME properties and ultimately predict likely issues before they become major hurdles to a project. This study suggests a new approach to sharing ADME/Tox models built using widely available open descriptors and algorithms.

References

Carletta J (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* **22:**249-254.

Chang C, Bahadduri PM, Polli JE, Swaan PW and Ekins S (2006) Rapid Identification of P-glycoprotein Substrates and Inhibitors. *Drug Metab Dispos* **34:**1976-1984.

Chang CC and Lin CJ (2001) LIBSVM: A library for support vector machines.

Chekmarev D, Kholodovych V, Kortagere S, Welsh WJ and Ekins S (2009) Predicting Inhibitors of Acetylcholinesterase by Regression and ClassificationMachine Learning Approaches with Combinations of Molecular Descriptors. *Pharm Res* **26:**2216-2224.

Chekmarev DS, Kholodovych V, Balakin KV, Ivanenkov Y, Ekins S and Welsh WJ (2008) Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem Res Toxicol* **21:**1304-1314.

Cheng A and Merz KM, Jr. (2003) Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. *J Med Chem* **46:**3572-3580.

Cohen CM (2003) A path to improved pharmaceutical productivity. *Nat Rev Drug Discov* **2:**751-753.

Davis RA, Charlton AJ, Oehlschlager S and Wilson JC (2006) Novel feature selection method for genetic programming using metabolomic 1H NMR data. *Chemo Intell Lab Sys* **81:**50-59.

Dearden JC, Cronin MT and Kaiser KL (2009) How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ Res* **20:**241-266.

Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J and Mekenyan O (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. *J Chem Inf Model* **45:**839-849.

Dong X, Gilbert KE, Guha R, Heiland R, Kim J, Pierce ME, Fox GC and Wild DJ (2007) Web service infrastructure for chemoinformatics. *J Chem Inf Model* **47:**1303-1307.

Ekins S (2003) In silico approaches to predicting metabolism, toxicology and beyond. *Biochem Soc Trans* **31:**611-614.

Ekins S (2007) *Computational Toxicology: risk assessment for pharmaceutical and environmental chemicals*. John Wiley and Sons, Hoboken, NJ.

Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Sorokina S, Bugrim A and Nikolskaya T (2006a) A Combined Approach to Drug Metabolism and Toxicity Assessment. *Drug Metab Dispos* **34:**495-503.

Ekins S, de Groot M and Jones JP (2001a) Pharmacophore and three dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab Dispos* **29:**936-944.

Ekins S, Durst GL, Stratford RE, Thorner DA, Lewis R, Loncharich RJ and Wikel JH (2001b) Three dimensional quantitative structure permeability relationship analysis for a series of inhibitors of rhinovirus replication. *J Chem Inf Comput Sci* **41:**1578-1586.

Ekins S, Ecker GF, Chiba P and Swaan PW (2007) Future directions for drug transporter modelling. *Xenobiotica* **37:**1152-1170.

Ekins S, Kim RB, Leake BF, Dantzig AH, Schuetz E, Lan LB, Yasuda K, Shepard RL, Winter MA, Schuetz JD, Wikel JH and Wrighton SA (2002) Application of three dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates. *Mol Pharmacol* **61:**974-981.

Ekins S, Ring BJ, Grace J, McRobie-Belle DJ and Wrighton SA (2000a) Present and future in vitro approaches for drug metabolism. *J Pharmacol Toxicol Methods* **44:**313-324.

Ekins S, Shimada J and Chang C (2006b) Application of data mining approaches to drug delivery. *Adv Drug Deliv Rev* **58:**1409-1430.

Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA and Wikel JH (2000b) Progress in predicting human ADME parameters in silico. *J Pharmacol Toxicol Methods* **44:**251-272.

Ekins S and Williams AJ (2010) Precompetitive Preclinical ADME/Tox Data: Set It Free on The Web to Facilitate Computational Model Building to Assist Drug Development. *Lab on a Chip* **10:**13-22.

Frank E, Hall M, Trigg L, Holmes G and Witten IH (2004) Data mining in bioinformatics using Weka. *Bioinformatics* **20:**2479-2481.

Group QE (2004) The report from the expert group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the principles for the validation of (Q)SARs. . **49:**206.

Guangli M and Yiyu C (2006) Predicting Caco-2 permeability using support vector machine and chemistry development kit. *J Pharm Pharm Sci* **9:**210-221.

Guha R (2008) Flexible Web service infrastructure for the development and deployment of predictive models. *J Chem Inf Model* **48:**456-464.

Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J and Willighagen EL (2006) The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model* **46:**991-998.

Gupta RR and Gifford EM (2009) Automated Compound Submission and Active Learning Using HT-ADME in silico Models, in: *Bio-IT World Conference and Exposition*, Boston.

Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R and Tong W (2008) Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* **48:**1337-1344.

Huuskonen J (2000) Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci* **40:**773-777.

Jensen BF, Sorensen MD, Kissmeyer AM, Bjorkling F, Sonne K, Engelsen SB and Norgaard L (2003) Prediction of in vitro metabolic stability of calcitriol analogs by QSAR. *J Comput Aided Mol Des* **17:**849-859.

Jorgensen WL (2004) The many roles of computation in drug discovery. *Science* **303:**1813-1818.

Kortagere S, Chekmarev D, Welsh WJ and Ekins S (2009) Hybrid scoring and classification approaches to predict human pregane X receptor activiators. *Pharm Res* **26:**1001-1011.

Kortagere S, Chekmarev DS, Welsh WJ and Ekins S (2008) New predictive models for blood brain barrier permeability of drug-like molecules. *Pharm Res* **25:**1836-1845.

Lagorce D, Sperandio O, Galons H, Miteva MA and Villoutreix BO (2008) FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* **9:**396.

Liaw A and Wiener M (2002) Classification and regression by random forest. *R News* **2/3:**18-22.

Lind P and Maltseva T (2003) Support vector machines for the estimation of aqueous solubility. *J Chem Inf Comput Sci* **43:**1855-1859.

Lipinski CA, Lombardo F, Dominy BW and Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* **23:**3-25.

Maniyar DM, Nabney IT, Williams BS and Sewing A (2006) Data Visualization during the Early Stages of Drug Discovery. *J Chem Inf Model* **46:**1806-1818.

Masek BB, Shen L, Smith KM and Pearlman RS (2008) Sharing chemical information without sharing chemical structure. *J Chem Inf Model* **48:**256-261.

Melville JL and Hirst JD (2007) TMACC: interpretable correlation descriptors for quantitative structure-activity relationships. *J Chem Inf Model* **47:**626-634.

Pleban K, Kaiser D, Kopp S, Peer M, Chiba P and Ecker GF (2005) Targeting drug-efflux pumps -- a pharmacoinformatic approach. *Acta Biochim Pol* **52:**737-740.

Quinlan JR (1991) Improved estimated for the acccuracy of small disjuncts. *Machine Learn* **6:**93-98.

Quinlan JR (1993) C4.5: Program for Machine Learning, Morgan Kaufmann, Los Altos.

Ren S and Lien EJ (2000) Caco-2 cell permeability vs human gastrointestinal absorption: QSPR analysis. *Prog Drug Res* **54:**1-23.

Segarra V, Lopez M, Ryder H and Palacios JM (1999) Prediction of drug permeability based on Grid calculations. *QSAR* **18:**474-481.

Shen M, Xiao Y, Golbraikh A, Gombar VK and Tropsha A (2003) Development and validation of k-nearest neighbour QSPR models of metabolic stability of drug candidates. *J Med Chem* **46:**3013-3020.

Sheridan RP, Feuston BP, Maiorov VN and Kearsley SK (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J Chem Inf Comput Sci* **44:**1912-1928.

Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C, Torrance G, Wagener J, Willighagen EL, Steinbeck C and Wikberg JE (2009) Bioclipse 2: a scriptable integration platform for the life sciences. *BMC Bioinformatics* **10:**397.

Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C and Wikberg JE (2007) Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* **8:**59.

Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R and Willighagen EL (2006) Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* **12:**2111-2120.

Stenberg P, Luthman K and Artursson P (2000) Virtual screening of intestinal permeability. *J Controlled Release* **65:**231-243.

Stoner CL, Cleton A, Johnson K, Oh DM, Hallak H, Brodfuehrer J, Surendran N and Han HK (2004a) Integrated oral bioavailability projection using in vitro screening data as a selection tool in drug discovery. *Int J Pharm* **269:**241-249.

Stoner CL, Gifford E, Stankovic C, Lepsy CS, Brodfuehrer J, Prasad JV and Surendran N (2004b) Implementation of an ADME enabling selection and visualization tool for drug discovery. *J Pharm Sci* **93:**1131-1141.

Sykora VJ and Leahy DE (2008) Chemical Descriptors Library (CDL): a generic, open source software library for chemical informatics. *J Chem Inf Model* **48:**1931-1942.

Tetko IV, Bruneau P, Mewes HW, Rohrer DC and Poda GI (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today* **11:**700-707.

Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D and Varnek A (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* **48:**1733-1746.

Villoutreix BO, Renault N, Lagorce D, Sperandio O, Montes M and Miteva MA (2007) Free resources to assist structure-based virtual ligand screening experiments. *Curr Protein Pept Sci* **8:**381-411.

Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF and Chen YZ (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci* **44:**1497-1505.

Yamashita F, Hara H, Ito T and Hashida M (2008) Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. *J Chem Inf Model* **48:**364-369.

Yamashita F, Itoh T, Hara H and Hashida M (2006) Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J Chem Inf Model* **46:**1054-1059.

Zheng X, Ekins S, Rauffman J-P and Polli JE (2009) Computational models for drug inhibition of the Human Apical Sodium-dependent Bile Acid Transporter. *Mol Pharm* **6:**1591-1603.

Zientek M, Stoner C, Ayscue R, Klug-McLeod J, Jiang Y, West M, Collins C and Ekins S (2010) Integrated in silico-in vitro strategy for addressing cytochrome P450 3A4 time-dependent inhibition. *Chem Res Toxicol* **23:**664-676.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Footnotes Page

a). CDD funding from the Bill and Melinda Gates Foundation [Grant#49852 "Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing"].

b). Send reprint requests to: Sean Ekins, Collaborative Drug Discovery, 601 Runnymede Avenue, Jenkintown, PA 19046. Email ekinssean@yahoo.com, sekins@collaborativedrug.com

c). Competing Financial Interest: SE, MH, BAB are employees or consultants of Collaborative Drug Discovery, Inc.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Figure legend

Figure 1. This schematic describes the "*in silico* screening" methodology where one can use a predictive *in silico* model to pre-screen the compounds before they can go through the actual *in vitro* screening. This also allows implementation of the "Active Learning" paradigm.

Figure 2. The chemical space of the HLM training dataset (red circles) and the test set of 2300 compounds (blue circles) as described by a PCA plot using the descriptors described in the Materials and Methods. We have constrained the 579 descriptors (or 579 dimensional space) into 3 PC's and therefore the total variance explained is low. The total variance explained by the 3 PCs for the training set was 0.251. The total variance explained by the 3 PCs for the test set was 0.309.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Tables

Table 1. A summary of the classification bins for each endpoint/assay discussed in this work.

BIN **Assay**	**Low Risk** **(low)**	**Moderate Risk** **(moderate)**	**High Risk (high)**
Human Liver Microsomal Stability	$CL_{int} < 9.2$	$9.2 < CL_{int} < 48$	$CL_{int} > 48$
RRCK Permeability	RRCK < 2.5	2.5 < RRCK < 10	RRCK > 10
MDR P-gp Efflux*	MDR < 2.5		MDR > 2.5

* For MDR only 2 bins are provided. Bins were formed on the basis of guidance provided by domain experts in the Pharmacokinetics Dynamics and Metabolism (PDM) research group.

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Table 2. Dataset size and data distribution for the HLM, RRCK and MDR assays that were used to build the categorical models. The datasets are divided into training and test datasets that were used to build the model and test the predictivity of the *in silico* model respectively.

Assay	**Training set** **Number in high, moderate and low categories**	**Test set** **Number in high, moderate and low categories**
Human Liver Microsomal Stability	20,445, 18,545, 10,978,	10,236, 9466, 5282
RRCK Permeability	13,887, 6656, 4446	13,491, 6924, 4580
MDR P-gp Efflux	10,820, 14,175	7972, 10,441

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Table 3. Details on the human liver microsomal data modeling grid showing performance of various modeling methods versus descriptors. Some combinations were not evaluated as it was apparent from the combination of CDK descriptors and the modeling methods that the results are not going to be equivalent or better than the baseline model. CDK = chemistry development kit, PPV = positive predicted value, RP = recursive partitioning, SVM = Support vector machine.

	SVM	RP Forest Uni Class	RP Forest	C5.0
CDK	Kappa = 0.14 Sensitivity = 0.11 Specificity = 0.96 PPV = 0.43	Kappa = 0.16 Sensitivity = 0.54 Specificity = 0.70 PPV = 0.33	Kappa = 0.11 Sensitivity = 0.85 Specificity = 0.33 PPV = 0.25	Kappa = 0.39 Sensitivity = 0.54 Specificity = 0.91 PPV = 0.61
MOE2D and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.43 Sensitivity = 0.58 Specificity = 0.91 PPV = 0.63 (Baseline)
CDK and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.43 Sensitivity = 0.58 Specificity = 0.91 PPV = 0.63

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Table 4. Summary of results for human liver microsomal (HLM) models generated with very large training sets and different molecular descriptors. PPV = positive predicted value.

HLM Model with CDK and SMARTS Keys:	**HLM Model with MOE2D and SMARTS Keys**
• Number of Descriptors: 578 Descriptors • Number of Training Set compounds: 193,650 • Cross Validation Results: 38,730 compounds • Training R^2: 0.79 • 20% Test Set R^2: 0.69 **Blind Data Set (2310 compounds):** • R^2 = 0.53 • RMSE = 0.367 **Continuous → Categorical:** • κ = 0.40 • Sensitivity = 0.16 • Specificity = 0.99 • PPV = 0.80 **Time (sec/compound): 0.252**	• Number of Descriptors: 818 Descriptors • Number of Training Set compounds: 193,930 • Cross Validation Results: 38,786 compounds Training R^2: 0.77 • 20% Test Set R^2: 0.69 **Blind Data Set (2310 compounds):** • R^2 = 0.53 • RMSE = 0.367 **Continuous → Categorical:** • κ = 0.42 • Sensitivity = 0.24 • Specificity = 0.987 • PPV = 0.823 **Time (sec/compound): 0.303**

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Table 5. Details on the RRCK data modeling grid i.e. performance of various modeling methods versus descriptors. Some combinations were not evaluated as it was apparent from the combination of CDK descriptors and the modeling methods that the results are not going to be equivalent or better than the baseline model. CDK = chemistry development kit, PPV = positive predicted value, RP = recursive partitioning, SVM = Support vector machine.

	SVM	RP Forest Uni Class	RP Forest	C5.0
CDK	Kappa = 0.12 Sensitivity = 0.11 Specificity = 0.98 PPV = 0.51	Kappa = 0.18 Sensitivity = 0.44 Specificity = 0.77 PPV = 0.30	Kappa = 0.15 Sensitivity = 0.54 Specificity = 0.65 PPV = 0.26	Kappa = 0.47 Sensitivity = 0.59 Specificity = 0.93 PPV = 0.67
MOE2D and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.53 Sensitivity = 0.64 Specificity = 0.94 PPV = 0.72 (Baseline)
CDK and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.50 Sensitivity = 0.62 Specificity = 0.94 PPV = 0.68

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

Table 6. Details on the MDR data modeling grid i.e. performance of various modeling methods versus descriptors. Some combinations were not evaluated as it was apparent from the combination of CDK descriptors and the modeling methods that the results are not going to be equivalent or better than the baseline model. One thing to note here is that unlike HLM and RRCK, the MDR dataset was divided into 2 bins only. CDK = chemistry development kit, PPV = positive predicted value, RP = recursive partitioning, SVM = Support vector machine.

	SVM	RP Forest Uni Class	RP Forest	C5.0
CDK	Kappa = 0.31 Sensitivity = 0.79 Specificity = 0.52 PPV = 0.68	Kappa = 0.36 Sensitivity = 0.68 Specificity = 0.67 PPV = 0.73	Kappa = 0.33 Sensitivity = 0.64 Specificity = 0.70 PPV = 0.73	Kappa = 0.62 Sensitivity = 0.85 Specificity = 0.77 PPV = 0.83
MOE2D and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.67 Sensitivity = 0.86 Specificity = 0.80 PPV = 0.85 (Baseline)
CDK and SMARTS Keys	Not evaluated	Not evaluated	Not evaluated	Kappa = 0.65 Sensitivity = 0.86

				Specificity = 0.78 PPV = 0.84

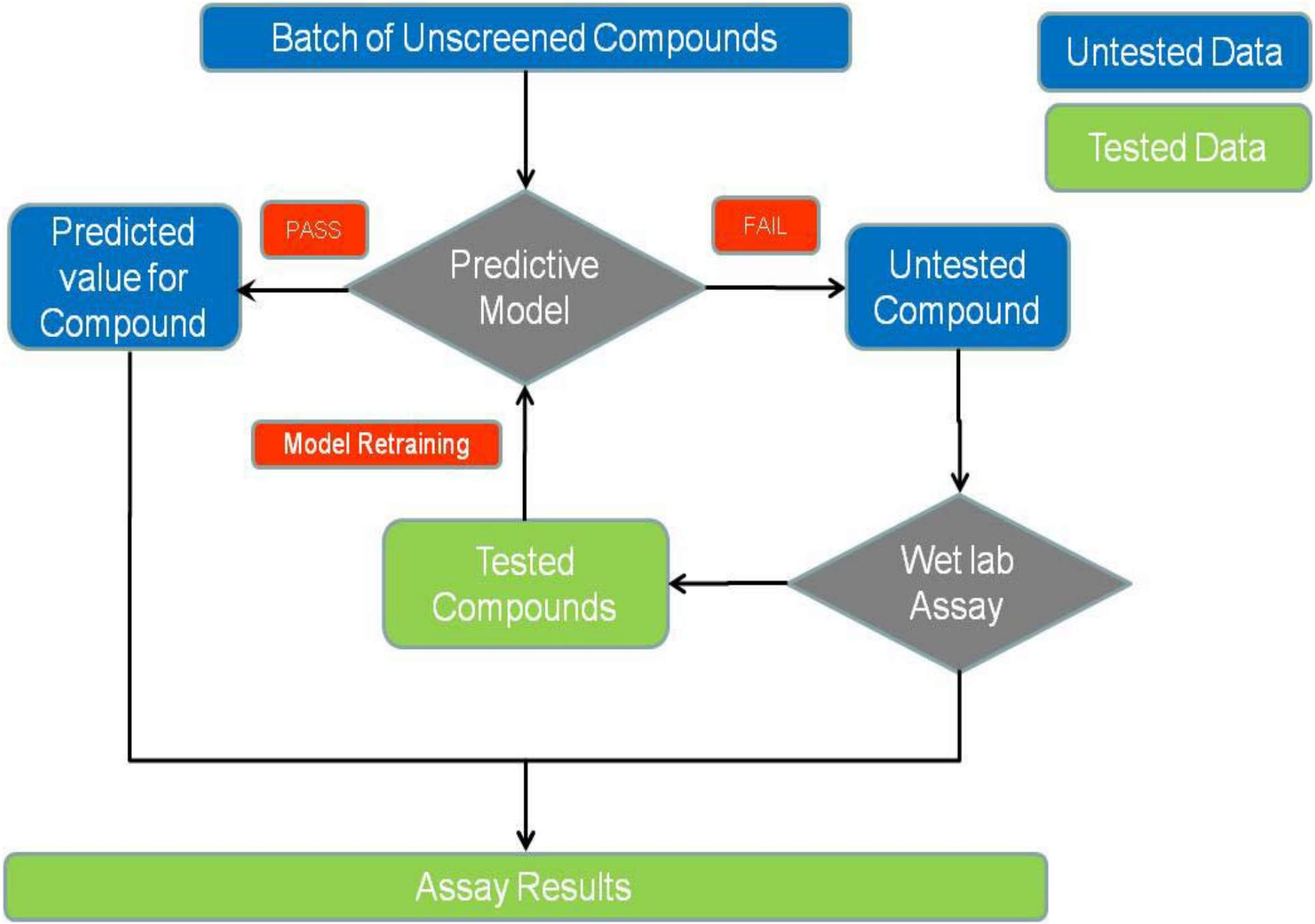

Fig 1

DMD Fast Forward. Published on August 6, 2010 as DOI: 10.1124/dmd.110.034918
This article has not been copyedited and formatted. The final version may differ from this version.

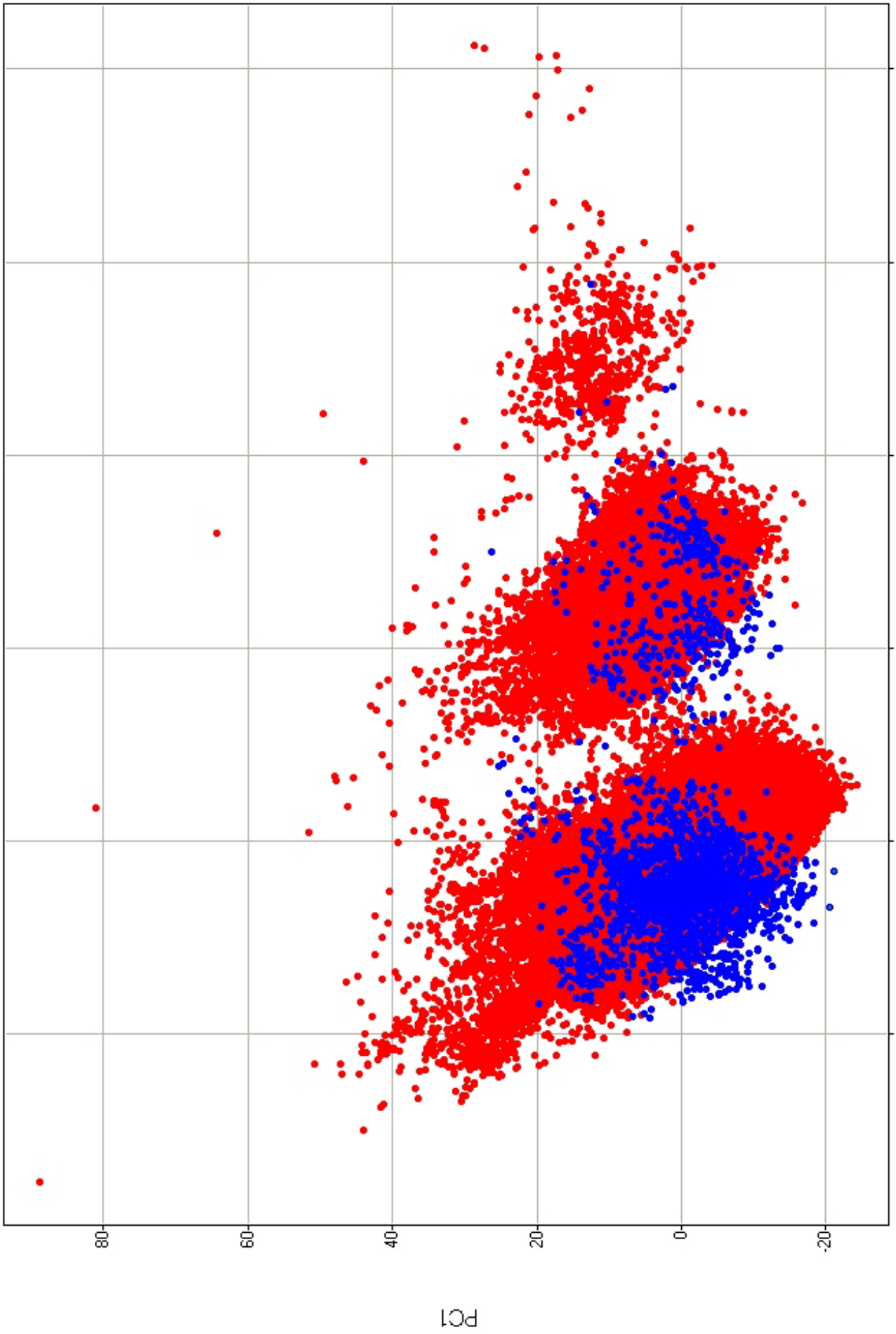

Fig 2