A PREDICTIVE LIGAND-BASED BAYESIAN MODEL FOR HUMAN DRUG INDUCED LIVER INJURY

Sean Ekins, Antony J. Williams and Jinghai J. Xu

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. (SE)

Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587. (AJW)

Merck & Co., Inc., 126 E. Lincoln Ave, Rahway, NJ 07065. (JJX)

Non standard abbreviations:

ADME/Tox, absorption, distribution, metabolism, excretion and toxicity; DDI, drug-drug interactions; DILI, Drug Induced Liver Injury; ECFC_6, Extended connectivity functional class fingerprint of maximum diameter 6; HIAT, human hepatocyte imaging assay technology; PCA, principal component analysis; QSAR, quantitative structure activity relationship; ROC, Receiver Operator Curve XV, cross validated.

Abstract

Drug-induced liver injury (DILI) is one of the most important reasons for drug development failure at both pre-approval and post-approval stages. There has been increased interest in developing predictive *in vivo, in vitro* and *in silico* models to identify compounds that cause idiosyncratic hepatotoxicity. In the current study we applied machine learning, Bayesian modeling method with extended connectivity fingerprints and other interpretable descriptors. The model that was developed and internally validated (using a training set of 295 compounds) was then applied to a large test set relative to the training set (237 compounds) for external validation. The resulting concordance of 60%, sensitivity of 56%, and specificity of 67% were comparable to internal validation. The Bayesian model with ECFC_6 fingerprint and interpretable descriptors suggested several substructures that are chemically reactive and may also be important for DILI-causing compounds, e.g. ketones, diols and α-methyl styrene type structures. Using SMARTS filters published by several pharmaceutical companies we evaluated whether such reactive substructures could be readily detected by any of the published filters. It was apparent that the most stringent filters used in this study, like the Abbott alerts which captures thiol traps and other compounds, may be of utility in identifying DILI-causing compounds (sensitivity 67%). A significant outcome of the present study is that we provide predictions for many compounds that cause DILI by using the knowledge we have available from previous studies for computational approaches. These computational models may represent a cost effective selection criteria prior to costly *in vitro* or *in vivo* experimental studies.

Introduction

Pharmaceutical research must develop predictive approaches to decrease the late stage attrition of compounds in clinical trials. One approach to this is to optimize absorption, distribution, metabolism, distribution and toxicity (ADME/Tox) properties earlier which is now frequently facilitated by a panel of *in vitro* assays. The liver is highly perfused and the "first-pass" organ for any orally-administered xenobiotic, while it also represents a frequent site of toxicity of pharmaceuticals in humans (Lee, 2003; Kaplowitz, 2005). The physiological location and drug-clearance function of the liver dictate that for an orally-administered drug, the drug exposure or drug load that the liver experiences is higher than that being measured systemically in peripheral blood (Ito et al., 2002). Drug-metabolism in the liver can convert some drugs into highly reactive intermediates and which in turn can adversely affect the structure and functions of the liver (Kassahun et al., 2001; Park et al., 2005; Walgren et al., 2005; Boelsterli et al., 2006). Therefore, it is not surprising that drug-induced liver injury, DILI, is the number one reason why drugs are not approved and why some of them were withdrawn from the market after approval (Schuster et al., 2005).

We have previously assembled a list of approximately 300 drugs and chemicals with a classification scheme based on clinical data for hepatotoxicity, for the purpose of evaluating an *in vitro* testing methodology based on cellular imaging of human hepatocyte cultures (Xu et al., 2008). Since every drug can exhibit some toxicity at high enough exposure (i.e., the notion of "dose makes a poison" by Paracelsus), we previously tested a panel of orally administered drugs at multiples of the therapeutic C_{max} (maximum therapeutic concentration), taking into account the first-pass effect of the liver and other idiosyncratic toxicokinetic/toxicodynamic factors. It was found that the 100-fold C_{max} scaling factor represented a reasonable threshold to differentiate safe versus toxic drugs, for an orally dosed drug and with regard to hepatotoxicity (Xu et al., 2008). The overall concordance of the *in vitro* human hepatocyte imaging assay technology (HIAT), when applied to about 300 drugs and chemicals, is about 75% with regard to clinical hepatotoxicity, with very few false-positives (Xu et al., 2008). The reasonably high specificity and reasonable sensitivity of such an *in vitro* test system has made it especially attractive as part of a pre-clinical testing paradigm to select drug candidates with improved therapeutic index for clinical hepatotoxicity.

Obviously, using *in vitro* approaches still comes at a cost. Firstly the compound has to physically have been made and be available for testing, secondly the screening system is still relatively low throughput compared to any primary screens and as a result whole compound or vendor libraries cannot be cost effectively screened for prioritization. Thirdly, the screening system should be representative of the human organ including drug metabolism capability. Yet a fourth consideration is that the prediction of human therapeutic C_{max} is often imprecise prior to clinical testing in actual patients. A potential alternative may be to use the historic DILI data to create a computational model and then test it with an equally large set of compounds to ensure that there is enough confidence such that its predictions can be used as a prescreen prior to actual *in vitro* testing.

There have been many examples where computational quantitative structure activity relationship (QSAR) or machine learning methods have been used for predicting hepatotoxicity (Cheng and Dixon, 2003; Clark et al., 2004) or drug-drug interactions

(Ekins et al., 2000; Marechal et al., 2006; Ung et al., 2007; Zientek et al., 2010). One recent study used a small set of 74 compounds (33 of which were known to be associated with idiosyncratic hepatotoxicity and the rest were not) to create classification models based on linear discriminant analysis (LDA), artificial neural networks (ANN), and machine learning algorithms (OneR) (Cruz-Monteagudo et al., 2007). These modeling techniques were found to produce models with satisfactory internal cross-validation statistics (accuracy/sensitivity/specificity over 84%/78%/90%, respectively). These models were then tested on very small sets of compounds (6 and 13 compounds, respectively) with over 80% accuracy. A second study compiled a data set of compounds reported to produce a wide range of effects in the liver in different species then used binary QSAR models (248 active, 283 inactive) to predict whether a compound would be expected to produce liver effects in humans. The resultant support vector machine (SVM) models had good predictive power assessed by external 5-fold cross-validation procedures and 78% accuracy for a set of 18 compounds (Fourches et al., 2010). A third study created a knowledge-base with structural alerts from 1266 chemicals. Although not strictly a machine learning method the alerts created were used to predict 626 Pfizer compounds (ensitivity 46%, specificity 73% and concordance 56% for the latest version) (Greene et al., 2010).

A major limitation of these previous global models for DILI (and for many computational toxicology models) is their use of very small test sets in all cases. These studies also have not examined how well they could predict many sets of closely related compounds in which some show DILI and others do not, which is most likely the scenario facing us in the real world of pharmaceutical research. Another issue is the quality of the compound datasets used for model building and testing (Williams et al., 2009).

In the current study we have used a training set of 295 compounds and a test set of 237 molecules. In contrast to earlier studies we have used a Bayesian classification approach (Xia et al., 2004; Bender, 2005) with simple, interpretable molecular descriptors as well as extended connectivity functional class fingerprints of maximum diameter 6 (ECFC_6) (Jones et al., 2007) to classify compounds as DILI or non-DILI. We also use these descriptors to highlight chemical substructures that are important for DILI. In addition, we have applied chemical filters to all the 532 molecules in the test and training set as many pharmaceutical companies use SMARTS [SMiles ARbitrary Target Specification] queries which specify substructures of interest (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html). These approaches have enabled the removal of undesirable molecules, false positives and frequent hitters from their HTS screening libraries or to filter vendor compounds (Williams et al., 2009). Computational models or filters for DILI could be a valuable filter for selecting compounds for further synthesis and testing *in vitro* or *in vivo.*

Methods

Source of DILI data. We have greatly expanded our original DILI drug list of about 300 drugs and chemicals with the same classification scheme based on clinical data for hepatotoxicity (Xu et al., 2008). Our DILI positive drugs include those: 1) withdrawn from the market mainly due to hepatotoxicity (e.g., troglitazone (Parker, 2002)), 2) not marketed in the United States due to hepatotoxicity (e.g., nimesulide (Macia et al., 2002)), 3) receiving black box warnings from the FDA due to hepatotoxicity (e.g., dantrolene (Durham et al., 1984)), 4) marketed with hepatotoxicity warnings in their labels (e.g., zileuton (Watkins et al., 2007)), 5) others (mostly old drugs) that have well-known associations with liver injury and have a significant number (>10) of independent clinical reports of hepatotoxicity (e.g., diclofenac (Boelsterli, 2003)). Drugs that do not meet any of the above positive criteria are classified as DILI negatives. The expanded drug list and its DILI classifications were researched and collated at the same time as the original 300 drug list for *in vitro* testing. The expanded drug list includes 237 compounds which were previously not available for *in vitro* testing. However, since computational modeling does not require the physical availability of compounds, we have decided to use them as our relatively large test set for *in silico* modeling.

Training and test set curation. Assembling high quality data sets for the purpose of computational analysis can be very challenging. Commonly public data sources are used as trusted resources of information and without further validation and, as has been demonstrated or suggested in a number of previous studies, this is not appropriate ((Fourches et al., ; Williams et al., 2009) and references therein). The set of validated chemical structures utilized as the training and test data were assembled from the ChemSpider database (www.chemspider.com). The set of chemical names associated with the DILI set were searched against the ChemSpider database and the chemical compounds associated with manually curated chemical records were downloaded. This amounted to over 90% of the list of chemical names. For the remaining chemical names the associated structures in ChemSpider were then manually validated by checking various resources to assert the correct chemical structures. These included validation across multiple online resources (e.g., Dailymed, ChemIDPLus and Wikipedia) as well as the Merck Index to ensure consistency between the various resources. The test and training set (Supplemental Table 1) were also compared by Tanimoto similarity (Willett, 2003) with MDL keys to remove any compounds with a value of 1, indicative of them being identical but possessing different synonyms in each dataset.

Bayesian machine learning model development. Laplacian-corrected Bayesian classifier models were generated using Discovery Studio. (Version 2.5.5., Accelrys, San Diego, CA) This approach employs a machine learning method with 2D descriptors (as described previously for other applications (Rogers et al., 2005; Hassan et al., 2006; Klon et al., 2006; Bender et al., 2007; Prathipati et al., 2008)) to distinguish between compounds that are DILI positive and those that are DILI negative. Preliminary work evaluated separately different functional class fingerprints (FCFP) (of size 0-20) descriptors alongside interpretable descriptors. FCFP_6 had approximately the highest receiver operator curve (ROC) for the leave-one-out for the DILI data. We then evaluated separately other fingerprint descriptors (e.g. elemental type fingerprints, ECFP; AlogP code path length fingerprint, LPFP), separately (ECFC_6, ECFP_6, EPFC_6, EPFP_6, FCFC_6, FPFC_6, FPFP_6 LCFC_6 and LPFC_6) ((Bender, 2005) descriptor naming conventions can be found within the help pages of Discovery Studio 2.5.5) . Several had ROC values > 0.8 while ECFC_6 is the focus of this study with the following interpretable descriptors:: ALogP, ECFC_6, Apol, logD, molecular weight, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rings, number of rotatable bonds, molecular polar surface area, molecular surface area. Wiener and Zagreb indices were calculated from an input sdf file using the "calculate molecular properties" protocol.

The "create Bayesian model" protocol was used for model generation. The theory behind this method has been described in more detail elsewhere (Zientek et al., 2010). A custom protocol for validation was also used in which 10%, 30% or 50% of the training set compounds were left out 100 times. The mean (±SD) of the calculated values were reported.

Comparison of training and test sets.

The interpretable descriptors described above were used to compare compounds of each class in the training and test sets using statistical comparisons performed with JMP (SAS Institute Cary, NC).

Principal Component Analysis (PCA) available in Discovery Studio version 2.5.5 was used to compare the molecular descriptor space for the test and training sets (using the descriptors of ALogP, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of rings, number of aromatic rings, and molecular fractional polar surface area). In each case, the respective test set and the training set compounds were combined and used to generate the PCA analysis.

For a comparison with recently launched drugs we extracted small-molecule drugs from 2006-2010 from the Prous Integrity database and went through a curation process similar to that described above. A number of these drugs were not “small molecules” appropriate for examination and modeling in this study and were immediately rejected. Structure validation resulted in a set of 75 molecules that were used for PCA and physicochemical property analysis.

SMARTS Filters. We used the 107 SMARTS filters in Discovery Studio 2.5.5 (Supplemental Text). The Abbott ALARM (Huth et al., 2005), Glaxo (Hann et al., 1999) and Pfizer LINT (also known as Blake filter (Blake, 2005)) SMARTS filter calculations were performed through the Smartsfilter web application kindly provided by Dr. Jeremy Yang (Division of Biocomputing, Dept. of Biochem and Molecular Biology, University of New Mexico, Albuquerque, NM, (http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter). This software identifies the number of compounds that pass or fail any of the filters implemented. Each filter was evaluated individually with the combined set of training and test compounds (N = 532).

Results

Bayesian Models. We initially evaluated the Bayesian model with multiple cross validation approaches then we evaluated the models with multiple external test sets which are more representative of chemical space coverage beyond the training set. The cross validated receiver operator curve area under the curve (XV ROC AUC) for the model with 295 molecules built with simple molecular descriptors alone was 0.86 and the best split was 0.17 with the ECFC_6 descriptors and interpretable descriptors (Supplemental data). By using the ECFC_6 descriptors, we can also identify those substructure descriptors that contribute to the DILI (Figure 1A) and those that are not present in compounds causing DILI (Figure 1B). The Bayesian model generated was also evaluated by leaving out either 10%, 30% or 50% of the data and rebuilding the model 100 times in order to generate the XV ROC AUC. In each case the leave out 10%, 30% or 50% testing AUC value was comparable to the leave-one-out approach and these values were very favorable indicating good model robustness (Table 1). The mean concordance > 57%, specificity > 61% and sensitivity > 52% did not seem to differ depending on the amount of data left out.

Molecular features important for DILI. Analysis of simple interpretable molecular properties between the compounds in the training set indicated that the mean ALogP was the only one statistically different between those that cause DILI and those that do not (Table 2). For the slightly smaller test set Apol, the number of rotatable bonds, the number of hydrogen bond acceptors, the number of hydrogen bond donors, molecular surface area, molecular polar surface area, and the Zagreb index were all significantly different between compounds that cause DILI and those that do not. Further molecular insights into the general properties of DILI forming compounds were obtained by using the ECFC_6 descriptor results from Discovery Studio to select molecules with a common substructure and analyze those that cause DILI from those that do not. As demonstrated in Figure 1A features such as long aliphatic chains (G1 and G2), phenols (G3), ketones (G5), diols (G7), α-methyl styrene (G8) (represents a polymer monomer), conjugated structures (G9), cyclohexenones (G10) and amides (G15) predominate.

Bayesian model validation. The Bayesian model was tested with 237 new compounds not present in the previous 295 training set (Supplemental Table 1). The concordance ~60%, specificity 67% and sensitivity 56% were comparable (Table 3) with internal validation (Table 1). A subset of 37 compounds (Supplemental Table 2) of most interest clinically (including similar compounds which were either DILI causing or not) showed similar testing values with a concordance greater than 63% (Table 2). Compounds of most interest can be defined as well-known hepatotoxic drugs (e.g., those hepatotoxic drugs cited elsewhere (FDA, 2009)), plus their less hepatotoxic comparators, if clinically available. These less hepatotoxic comparators are approved drugs that typically share a portion of the chemical core structure as the hepatotoxic ones (e.g., zolpidem versus alpidem, ibuprofen versus benoxaprofen, etc.). The purpose of this test set is to explore whether our *in silico* method can differentiate differences in DILI potential between or among closely related compounds, a scenario that is likely to be of most interest in real-world drug discovery and development efforts.

A PCA analysis using simple molecular descriptors showed that the training and test set covered overlapping or similar chemical space (Figure 2A). However, there were some distinct compounds like retinyl palmitate that were outside the training set (Figure 2B). Therefore, focusing in on compounds with a Tanimoto similarity greater than 0.7 left 28 compounds (Supplemental Table 3) whose Matthews correlation coefficient and concordance was similar to the complete test set. The specificity and sensitivity increased to 80% and 50%, respectively (Table 3) in this case.

SMARTS filtering We have also evaluated the training and test set compounds further by using various SMARTS filters which are used as alerts to remove undesirable compounds before *in vitro* screening (Williams et al., 2009). The hypothesis tested was whether the filters would predominantly remove compounds that caused DILI. Out of the four sets of independent filters tested the Abbott alerts had the highest concordance and sensitivity while the Glaxo filters had the highest specificity but lowest sensitivity and concordance (Table 4). It would appear that the Abbott Alerts retrieve two thirds of all the compounds causing DILI as they fail these alerts. The best statistics with filtering are lower than observed in Table 3 for the test sets with the Bayesian model.

Discussion

Pharmaceutical companies are keen to prevent late stage attrition due to adverse drug reactions or drug-drug interactions, and the earlier they are aware of a potentially problematic lead series, the sooner they can modify it and address the issue. In many ways this has been expedited and assisted by the increasing throughput of *in vitro* assays which are also used for the development of computational models (with particular focus on the liver due to its importance in first pass metabolism) (Ekins et al., 2003; O'Brien and de Groot, 2005). Idiosyncratic liver injury or drug induced liver injury are much harder to predict from the *in vitro* situation so we generally become aware of such problems once a drug reaches large populations in the clinic, which is too late. There have been efforts recently to use computational models to predict DILI or idiosyncratic hepatotoxicity. We are aware of at least three studies that tackled predicting DILI using either LDA, ANN, OneR (Cruz-Monteagudo et al., 2007), SVM (Fourches et al., 2010) or structural alerts (Greene et al., 2010). In the first two studies the models were tested with very small sets of compounds (<20) covering limited chemical space, while the third study used a large set of 626 proprietary compounds as the test set (Greene et al., 2010). In the current study we have carefully collated a training set of 295 compounds (of which 158 cause DILI) and a very large test set (relative to the training set) of 237 compounds (114 of these cause DILI) and used them to create and validate a Bayesian model.

Recently computational Bayesian models were developed for time-dependent inhibition of CYP3A4 using over 2000 molecules for filtering of compounds that must be screened *in vitro* due to this activity (Zientek et al., 2010). The Bayesian approach has also been used for modeling the apical sodium dependent bile acid transporter to identify inhibitors (Zheng et al., 2009) and for modeling inhibitory activity of a large set of compounds (>200,000) against Mycobacterium Tuberculosis in whole cells (Ekins et al., 2010). In our experience the Bayesian method can generate classifiers with good enrichments and classification accuracy for an external test set. In this study internal testing of the Bayesian model resulted in internal ROC scores (> 0.85) and specificity (> 61%), concordance (> 57%) and sensitivity (> 52%) (Table 1). Using the ECFC_6 descriptors we found that numerous of the fingerprints with high Bayesian scores and present in many DILI compounds, appeared to be reactive in nature which could cause time dependent inhibition of CYPs for example (Zientek et al., 2010) or be precursors for metabolites (Kassahun et al., 2001) that are reactive and may covalently bind to proteins. However, it is puzzling why long aliphatic chains may be important for DILI (Figure 1A) other than being generally hydrophobic and perhaps enabling increased accumulation. It is possible they may be hydroxylated, then form other metabolites that are in turn reactive. Further analysis of simple molecular descriptors calculated for the test and training sets showed only differences in ALogP for the training set while many descriptors were significantly different in the test set (e.g. DILI causing compounds have less molecular branching as measured by the Zagreb index and lower sum of atomic polarizabilities (Apol)) but not ALogP (Table 2). It was not until we used the Bayesian model with a test set that we could appreciate its potential utility. In this case for the whole dataset we saw concordance (~60%) and specificity (~67%) and sensitivity (~56%), comparable to internal testing (Table 3). When we focused on a very small subset of compounds of clinical interest the concordance increased. When we narrowed down the dataset to only those molecules with > 70% similar to the training set (N = 28) based on the Tanimoto similarity (with MDL Keys descriptors) the specificity increased above 80% and concordance increased slightly to ~64%. Such an increase in concordance statistics is analogous to that observed with other computational chemistry predictions, as it simply and effectively narrows the applicability domain to molecules that would be expected to be better predicted (Ekins et al., 2006). We have also evaluated the overlap of the training and test set chemical space using PCA (Figure 2A), an approach we have used previously (Zientek et al., 2010) that shows that many of the molecules in the test set cover similar chemical space to the training set, while there are some compounds that may be outliers like retinyl palmitate (Figure 2B), in this case it was correctly predicted as causing DILI. We have compared how these 532 compounds relate to a set of 77 recently launched small-molecule drugs from the period 2006-2010 extracted from the Prous Integrity database (Supplemental Figure 1). Again we find these molecules are distributed throughout the combined training and test set, representative of overlap which is also suggested from the mean physicochemical property values (Supplemental Table 4 compared with Table 2). These combined analyses would suggest that the test and training set used for the DILI model is representative of current medicinal chemistry efforts.

A further approach we have taken based on the output of the Bayesian model fingerprint descriptors (which suggested many reactive substructures) was to use published SMARTS filters which many groups have routinely used to remove reactive compounds in vendor compound screening libraries. For example REOS from Vertex (Walters and Murcko, 2002), filters from GSK (Hann et al., 1999), BMS (Pearce et al., 2006),Abbott (Huth et al., 2005; Huth et al., 2007; Metz et al., 2007) and others (Blake, 2005) have all been described. These latter SMARTS filters in particular detect thiol traps and redox active compounds. More recently, an academic group has published an extensive series of over 400 substructural features for removal of Pan Assay INterference compoundS (PAINS) from screening libraries (Baell and Holloway, 2010). In only one case in our study with the filters from Abbott (Huth et al., 2005; Metz et al., 2007) did we see a concordance or sensitivity value that was similar to that observed previously with the Bayesian model. This would suggest that these SMARTS may be useful as a pre-screen to remove potential DILI causing compounds alongside the Bayesian models which perform better.

In summary, we present the first large scale testing of a machine learning model for DILI that uses a similarly sized training and test sets. Our model may have utility in identifying compounds with a potential to cause human DILI. The overall concordance of the model is lower (~60-64% depending on test set size) than that observed previously for the *in vitro* HIAT (75% (Xu et al., 2008)). Our test-set statistics are similar to those reported elsewhere using structural alerts (Greene et al., 2010). The compounds that are scored to be DILI positive by our model, if still of high therapeutic interest, could be further tested by combined *in vitro* and *in vivo* testing, as they have sufficient sensitivity and very high specificity (Xu et al., 2008). By providing all of our structural and DILI classification data, the research community should now have a foundation for testing and benchmarking future computational models as well as generating predictions for DILI with new compounds. In conclusion, a significant outcome of this study is that we can enhance the predictive accuracy of models to identify compounds that cause DILI by using the knowledge we have available currently from compounds already evaluated (in the literature) to build a computational model. Such models alongside alerts based on undesirable substructures ((Greene et al., 2010) or those in this study), could be used to either filter or flag early stage molecules for this potential liability and could be evaluated in future studies. It is also feasible that combinations of such computational approaches may also be of utility to identify DILI causing compounds.

Acknowledgments

S.E. gratefully acknowledges Dr. Maggie A.Z. Hupcey for chemistry discussions, Accelrys Inc. for providing Discovery Studio 2.1 and 2.5.5., and the reviewers for their excellent suggestions.

References

Baell JB and Holloway GA (2010) New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem* **53:**2719-2740.

Bender A (2005) Studies on Molecular Similarity, University of Cambridge, Cambridge.

Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S and Jenkins JL (2007) Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2:**861-873.

Blake JF (2005) Identification and evaluation of molecular properties related to preclinical optimization and clinical fate. *Med Chem* **1:**649-655.

Boelsterli UA (2003) Diclofenac-induced liver injury: a paradigm of idiosyncratic drug toxicity. *Toxicol Appl Pharmacol* **192:**307-322.

Boelsterli UA, Ho HK, Zhou S and Leow KY (2006) Bioactivation and hepatotoxicity of nitroaromatic drugs. *Curr Drug Metab* **7:**715-727.

Cheng A and Dixon SL (2003) In silico models for the prediction of dose-dependent human hepatotoxicity. *J Comput Aided Mol Des* **17:**811-823.

Clark RD, Wolohan PR, Hodgkin EE, Kelly JH and Sussman NL (2004) Modelling in vitro hepatotoxicity using molecular interaction fields and SIMCA. *J Mol Graph Model* **22:**487-497.

Cruz-Monteagudo M, Cordeiro MN and Borges F (2007) Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. *J Comput Chem*.

Durham JA, Gandolfi AJ and Bentley JB (1984) Hepatotoxicological evaluation of dantrolene sodium. *Drug Chem Toxicol* **7:**23-40.

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

Ekins S, Berbaum J and Harrison RK (2003) Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab Dispos* **31:**1077-1080.

Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Hohman M and Bunin B (2010) A Collaborative Database And Computational Models For Tuberculosis Drug Discovery. *Mol BioSystems* **6:**840-851.

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

FDA U (2009) Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation.

Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ and Tropsha A (2010) Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chem Res Toxicol* **23:**171-183.

Fourches D, Muratov E and Tropsha A Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* **50:**1189-1204.

Greene N, Fisk L, Naven RT, Note RR, Patel ML and Pelletier DJ (2010) Developing Structure-Activity Relationships for the Prediction of Hepatotoxicity. *Chem Res Toxicol* **23:**1215-1222.

Hann M, Hudson B, Lewell X, Lifely R, Miller L and Ramsden N (1999) Strategic pooling of compounds for high-throughput screening. *J Chem Inf Comput Sci* **39:**897-902.

Hassan M, Brown RD, Varma-O'brien S and Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* **10:**283-299.

Huth JR, Mendoza R, Olejniczak ET, Johnson RW, Cothron DA, Liu Y, Lerner CG, Chen J and Hajduk PJ (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J Am Chem Soc* **127:**217-224.

Huth JR, Song D, Mendoza RR, Black-Schaefer CL, Mack JC, Dorwin SA, Ladror US, Severin JM, Walter KA, Bartley DM and Hajduk PJ (2007) Toxicological evaluation of thiol-reactive compounds identified using a la assay to detect reactive molecules by nuclear magnetic resonance. *Chem Res Toxicol* **20:**1752-1759.

Ito K, Chiba K, Horikawa M, Ishigami M, Mizuno N, Aoki J, Gotoh Y, Iwatsubo T, Kanamitsu S, Kato M, Kawahara I, Niinuma K, Nishino A, Sato N, Tsukamoto Y, Ueda K, Itoh T and Sugiyama Y (2002) Which concentration of the inhibitor should be used to predict in vivo drug interactions from in vitro data? *AAPS PharmSci* **4:**E25.

Jones DR, Ekins S, Li L and Hall SD (2007) Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). *Drug Metab Dispos* **35:**1466-1475.

Kaplowitz N (2005) Idiosyncratic drug hepatotoxicity. *Nat Rev Drug Discov* **4:**489-499.

Kassahun K, Pearson PG, Tang W, McIntosh I, Leung K, Elmore C, Dean D, Wang R, Doss G and Baillie TA (2001) Studies on the metabolism of troglitazone to reactive intermediates in vitro and in vivo. Evidence for novel biotransformation pathways involving quinone methide formation and thiazolidinedione ring scission. *Chem Res Toxicol* **14:**62-70.

Klon AE, Lowrie JF and Diller DJ (2006) Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model* **46:**1945-1956.

Lee WM (2003) Drug-induced hepatotoxicity. *N Engl J Med* **349:**474-485.

Macia MA, Carvajal A, del Pozo JG, Vera E and del Pino A (2002) Hepatotoxicity associated with nimesulide: data from the Spanish Pharmacovigilance System. *Clin Pharmacol Ther* **72:**596-597.

Marechal JD, Yu J, Brown S, Kapelioukh I, Rankin EM, Wolf CR, Roberts GC, Paine MJ and Sutcliffe MJ (2006) In silico and in vitro screening for inhibition of cytochrome P450 CYP3A4 by co-medications commonly used by patients with cancer. *Drug Metab Dispos* **34:**534-538.

Metz JT, Huth JR and Hajduk PJ (2007) Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J Comput Aided Mol Des* **21:**139-144.

O'Brien SE and de Groot MJ (2005) Greater than the sum of its parts: combining models for useful ADMET prediction. *J Med Chem* **48:**1287-1291.

Park BK, Kitteringham NR, Maggs JL, Pirmohamed M and Williams DP (2005) The role of metabolic activation in drug-induced hepatotoxicity. *Annu Rev Pharmacol Toxicol* **45:**177-202.

Parker JC (2002) Troglitazone: the discovery and development of a novel therapy for the treatment of Type 2 diabetes mellitus. *Adv Drug Deliv Rev* **54:**1173-1197.

Pearce BC, Sofia MJ, Good AC, Drexler DM and Stock DA (2006) An empirical process for the design of high-throughput screening deck filters. *J Chem Inf Model* **46:**1060-1068.

Prathipati P, Ma NL and Keller TH (2008) Global Bayesian models for the prioritization of antitubercular agents. *J Chem Inf Model* **48:**2362-2370.

Rogers D, Brown RD and Hahn M (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* **10:**682-686.

Schuster D, Laggner C and Langer T (2005) Why drugs fail--a study on side effects in new chemical entities. *Curr Pharm Des* **11:**3545-3559.

Ung CY, Li H, Yap CW and Chen YZ (2007) In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol Pharmacol* **71:**158-168.

Walgren JL, Mitchell MD and Thompson DC (2005) Role of metabolism in drug-induced idiosyncratic hepatotoxicity. *Crit Rev Toxicol* **35:**325-361.

Walters WP and Murcko MA (2002) Prediction of 'drug-likeness'. *Adv Drug Del Rev* **54:**255-271.

Watkins PB, Dube LM, Walton-Bowen K, Cameron CM and Kasten LE (2007) Clinical pattern of zileuton-associated liver injury: results of a 12-month study in patients with chronic asthma. *Drug Saf* **30:**805-815.

Willett P (2003) Similarity-based approaches to virtual screening. *Biochem Soc Trans* **31:**603-606.

Williams AJ, Tkachenko V, Lipinski C, Tropsha A and Ekins S (2009) Free Online Resources Enabling Crowdsourced Drug Discovery. *Drug Discovery World* **Winter**.

Xia XY, Maliski EG, Gallant P and Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. *J Med Chem* **47:**4463-4470.

Xu JJ, Henstock PV, Dunn MC, Smith AR, Chabot JR and de Graaf D (2008) Cellular imaging predictions of clinical drug-induced liver injury. *Toxicol Sci* **105:**97-105.

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.

Footnotes Page

a). Send reprint requests to: Sean Ekins, Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046. Email ekinssean@yahoo.com

b). Competing Financial Interest: SE consults for various pharmaceutical and software companies including Merck although he did not receive any payment for this study. JJX is currently employed by Merck, previously employed by Pfizer, and has stock ownership in both companies as well as other biopharmaceutical companies.

c) The structures of all compounds in the test and training sets as well as the set of recently approved drugs are available in sdf format online and the Bayesian model protocols used in Discovery Studio are available from the authors upon request.

Table 1. Results of internal validation of Bayesian model for DILI

Cross validated results (Mean ± SD) for Bayesian model building (ROC = Receiver operator curve).

Concordance (prediction accuracy) = (TP+TN)/(TP+TN+FP+FN), Specificity = TN/(TN+FP), Sensitivity = TP/(TP+FN)

true positive (TP), true negative (TN), false positive (FP) and false negative (FN)

	External ROC Score	Internal ROC Score	Concordance (%)	Specificity (%)	Sensitivity (%)
leave out 10% x 100	0.62 ± 0.08	0.86 ± 0.01	58.48 ± 8.31	65.45 ± 15.22	52.83 ± 12.92
leave out 30% x 100	0.62 ± 0.05	0.86 ± 0.03	59.23 ± 4.35	65.15 ± 9.18	54.21 ±9.69
leave out 50% x 100	0.60 ± 0.04	0.85 ± 0.04	57.63 ± 3.87	61.81 ± 10.57	54.20 ± 9.83

Table 2. Mean physicochemical properties for the 295 DILI training set molecules and 237 test set molecules

Descriptor	Training set DILI – (N = 137)	Training set DILI + (N = 158)	Test Set DILI – (N = 84)	Test set DILI + (N = 153)
ALogP	1.31 ± 3.24	1.89 ± 2.47 *	1.49 ± 3.07	2.09 ± 2.56
Apol	12644.0 ± 6478.29	12178.1 ± 6061.78	14401.3 ± 6419.16	12711.8 ± 7124.28 *
LogD	0.65 ± 3.43	1.23 ± 2.45	0.80 ± 3.07	1.46 ± 2.69
MW	355.67 ± 186.93	184.83 ± 184.83	398.56 ± 183.56	361.54 ± 201.89
Number of rotatable bonds	5.17 ± 4.35	4.47 ± 4.04	5.74 ± 3.17	4.81 ± 4.04 *
Number of rings	2.63 ± 1.51	2.51 ± 1.53	2.80 ± 1.75	2.45 ± 1.72
Number of aromatic rings	1.27 ± 1.04	1.36 ± 1.00	1.58 ± 1.14	1.39 ± 1.11
Number of H bond acceptors	5.20 ± 4.06	4.97 ± 3.61	6.49 ± 4.07	5.08 ± 3.81 **
Number of H bond donors	2.51 ± 2.82	2.09 ± 2.38	2.57 ± 2.52	1.88 ± 1.96 *
Molecular surface area	352.68 ± 180.92	332.88 ± 183.78	386.34 ± 177.07	342.62 ± 197.55 *

Molecular polar surface area	102.17 ± 92.83	96.48 ± 74.51	125.60 ± 78.23	97.80 ± 74.76 **
Wiener Index	2383.90 ± 6919.65	1919.01 ± 5230.99	2667.27 ± 3562.05	2280.12 ± 4890.95
Zagreb Index	122.38 ± 69.64	115.48 ± 64.32	136.52 ± 70.87	115.82 ± 76.90 *

* t-test $p < 0.05$

** t-test $p <0.01$

Table 3. Results of external validation of Bayesian model for DILI

The results were for the complete test set true positive (TP) =86, true negative (TN) =56, false positive (FP) = 28 and false negative (FN) = 67.

For the subset of most interest TP = 13, TN = 10, FP = 5 and FN = 8. For the compounds > 70 % similar to the training set TP = 9, TN = 8, FP = 2 and FN = 9.

Matthews correlation coefficient (TPxTN-FPxFN)/((TP+FN)(TP+FP)(TN+FP)(TN+FN))^0.5

Concordance (prediction accuracy) = (TP+TN)/(TP+TN+FP+FN), Specificity = TN/(TN+FP), Sensitivity = TP/(TP+FN)

Test Set (N)	Matthews correlation coefficient	Concordance (%)	Specificity (%)	Sensitivity (%)
Complete test set (N = 237)	0.22	59.91	66.67	56
Subset of most interest (N = 37)	0.28	63.88	66.67	61.9
Compounds > 70% similar to training set (N = 28)	0.29	60.71	80.00	50

Table 4. Summary of SMARTS filtering for the combined DILI test and training set. The Abbott ALARM (Huth et al., 2005; Metz et al., 2007), Glaxo (Hann et al., 1999) and Blake SMARTS filter (Originally provided as a Sybyl script to Tripos by Dr. James Blake (Array Biopharma) while at Pfizer (Blake, 2005)) calculation were performed through the Smartsfilter web application, (Dr. Jeremy Yang) Division of Biocomputing, Dept. of Biochem & Mol Biology, University of New Mexico, Albuquerque, NM, (http://pangolin.health.unm.edu/tomcat/biocomp/smartsfilter). Concordance (prediction accuracy) = (TP+TN)/(TP+TN+FP+FN), Specificity = TN/(TN+FP), Sensitivity = TP/(TP+FN)

true positive (TP), true negative (TN), false positive (FP) and false negative (FN)

Filters / DILI class	Molecules Passing filter	Molecules failing filter	Concordance	Specificity	Sensitivity
			(%)	(%)	(%)
Blake (Pfizer) total	283	249	50.7	54.7	47.9
DILI –ve	121	100			
DILI +ve	162	149			
Glaxo total	458	74	44.2	86.4	14.1
DILI –ve	191	30			

DILI +ve	267	44			
Abbott total	192	340	55.8	40.3	66.9
DILI –ve	89	132			
DILI +ve	103	208			
Accelrys total	276	256	47.9	49.8	46.6
DILI –ve	110	111			
DILI +ve	166	145			

Figure 1 A. ECFP_6 descriptors: features important for DILI. Each panel shows the naming convention for each fragment, the numbers of molecules it is present in that are active and the Bayesian score for the fragment..**1B.** ECFP_6 descriptors: features absent from DILI compounds. Each panel shows the naming convention for each fragment, the numbers of molecules it is present in that are active and the Bayesian score for the fragment.

Figure 2. Analysis of DILI training and test set by PCA. A. PCA plot. Yellow = test set, blue = training set. The following descriptors were used with Discovery Studio 2.5.5: ALogP, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of rings, number of aromatic rings, and molecular fractional polar surface area. 0.82 % of the variance was explained with the first three principal components. B. Retinyl palmitate (O15-hexadecanoylretinoic acid), the top left yellow compound in the PCA plot (A).