DMD Fast Forward. Published on September 15, 2010 as doi:10.1124/dmd.110.035113

DMD # 35113

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)

Running title page

a. Running title: Bayesian DILI model

b. Corresponding Author:

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

Collaborations in Chemistry,

601 Runnymede Ave, Jenkintown, PA 19046. USA.

Phone 215-687-1320

Email ekinssean@yahoo.com

c. Number pages

Text pages:	16
Tables:	3
Figures:	2
References:	48
Words in Abstract:	249
Words in Introduction:	932
Words in Discussion:	1388

d. 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. These computational models may represent a cost effective selection criteria prior to *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).

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). 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, Supplemental sd files) 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 (due to the highest ROC) obtained 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 sd 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 *t*-test 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 77 molecules (mean molecular weight 427.05 ± 280. 31, range 94.11-1994.09) 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 increased to 80% and and sensitivity decreased to 50% (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). 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. 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. The previous 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).

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). When we used the Bayesian model with a test set 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, undesirable molecules, false positives and frequent hitters from their HTS screening libraries or to filter vendor compounds (Williams et al., 2009). 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 HIAT has 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.

Figure 1 A. ECFC_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.** ECFC_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).

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

Molecular descriptors generated in Discovery Studio 2.5.5 (Accelrys, San Diego, CA).

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

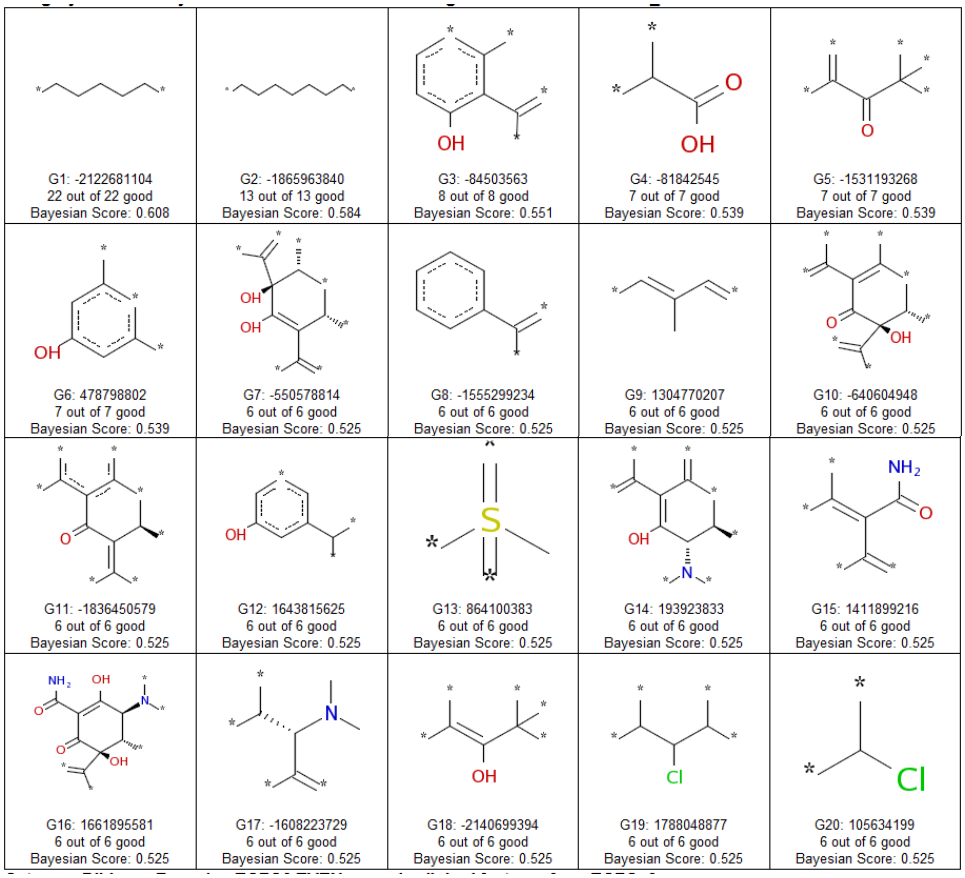
Table 4. Summary of SMARTS filtering for the combined DILI test and training set. The Abbott ALARM (Huth et al., 2005; Metz et al., 2007), Glaxo (Hann et al., 1999) and Blake SMARTS filter (Originally provided as a Sybyl script to Tripos by Dr. James Blake (Array Biopharma) while at Pfizer (Blake, 2005)) 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). True positive (TP), true negative (TN), false positive (FP) and false negative (FN)Concordance (prediction accuracy) = (TP+TN)/(TP+TN+FP+FN), Specificity = TN/(TN+FP), Sensitivity = TP/(TP+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			

DMD Fast Forward. Published on September 15, 2010 as DOI: 10.1124/dmd.110.035113
This article has not been copyedited and formatted. The final version may differ from this version.

A



B

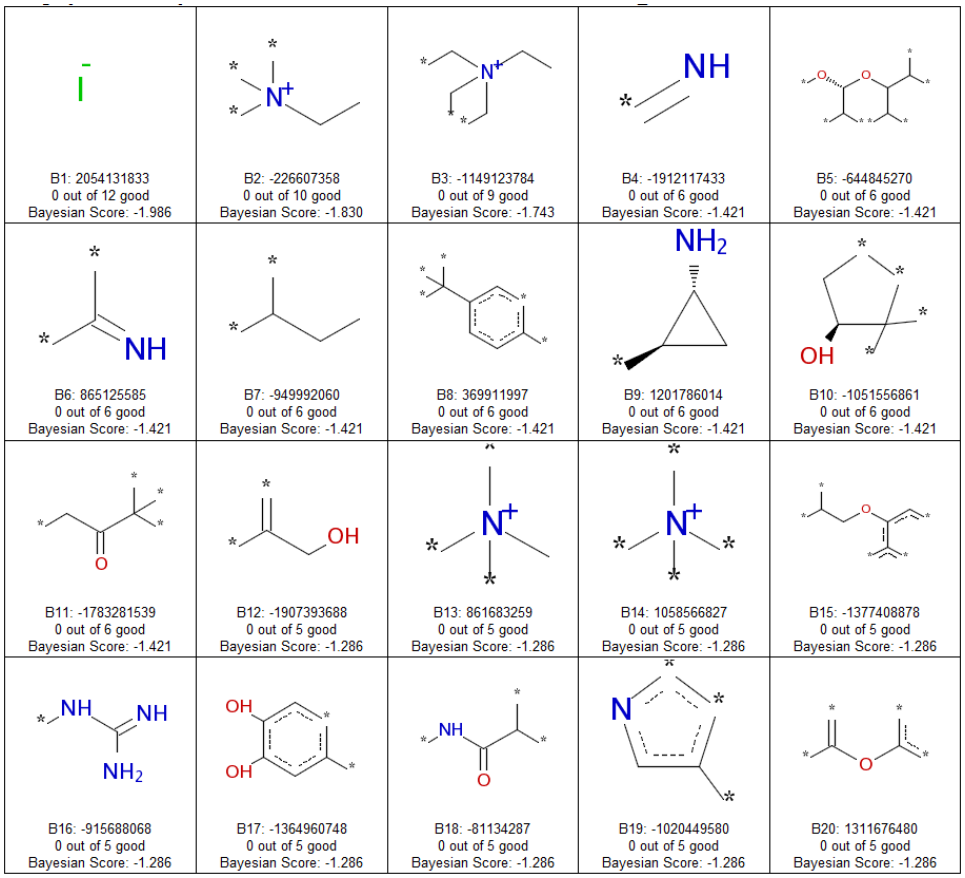
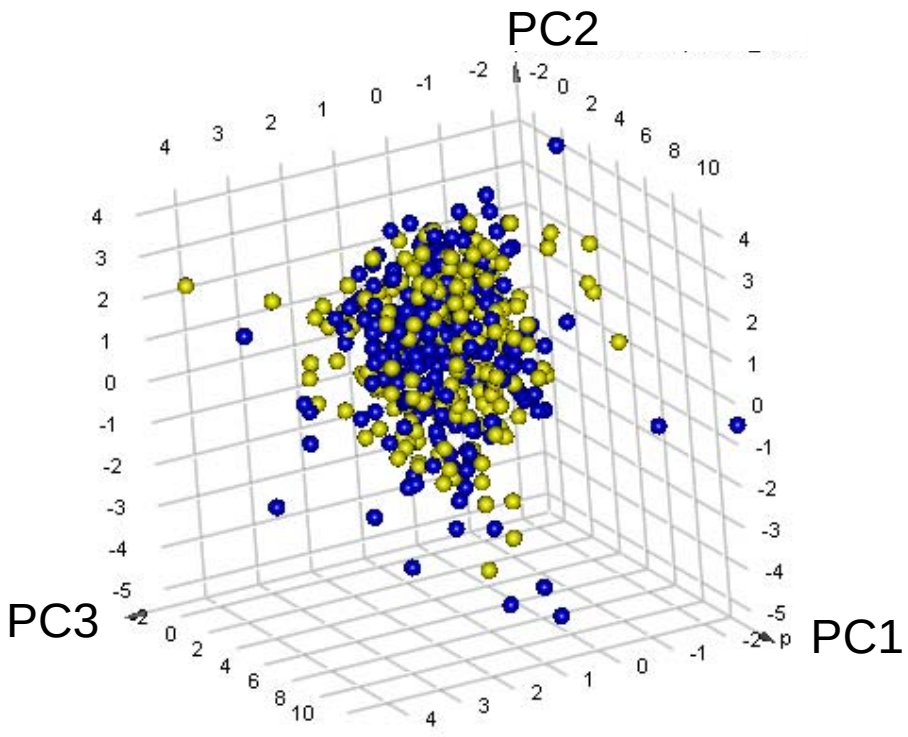


Fig 1

DMD Fast Forward. Published on September 15, 2010 as DOI: 10.1124/dmd.110.035113
This article has not been copyedited and formatted. The final version may differ from this version.

A.



B.

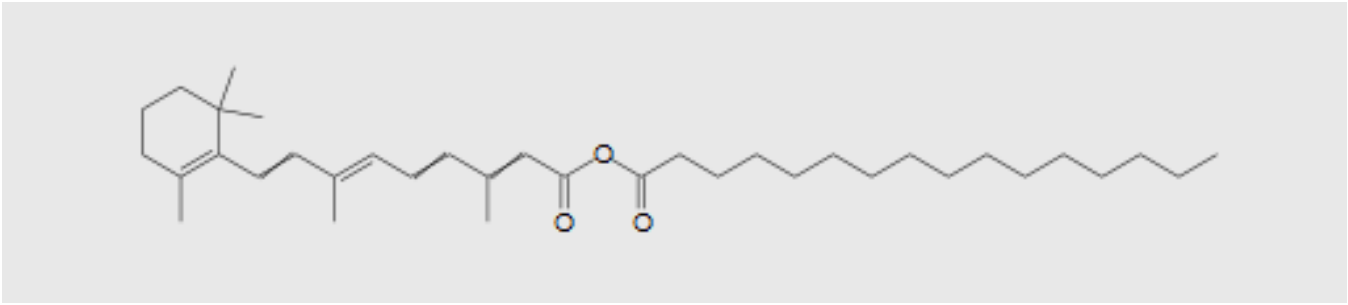


Fig 2

Drug Metabolism and Disposition

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)

Supplemental data for Bayesian model

Output from discovery studio

Leave-one-out Cross-Validation Results

This model was built using 295 samples, and validated using a leave-one-out cross-validation. Each sample was left out one at a time, and a model built using the results of the samples, and that model used to predict the left-out sample. Once all the samples had predictions, a ROC plot was generated, and the area under the curve (**XV ROC AUC**) calculated.

Best Split was calculated by picking the split that minimized the sum of the percent misclassified for category members and for category nonmembers, using the cross-validated score for each sample. Using that split, a contingency table is constructed, containing the number of true positives (**TP**), false negatives (**FN**), false positives (**FP**), and true negatives (**TN**).

Output	XV ROC AUC	Best Split	TP/FN FP/TN	# in Category
DILI new Bayesian ECFC6 EVEN more des II	0.860	0.167	115/43 17/120	158

Enrichment Results

Back to Top

This model was built using 295 samples, and validated using a leave-one-out cross-validation. Each sample was left out one at a time, and a model built using the results of the samples, and that model used to predict the left-out sample. Once all the samples had predictions, an enrichment plot was generated, and the percentage of true category members captured at a particular percentage cutoff. (For example, in a column labeled "1%" would be the percentage of true category members (e.g., actives) that were found in the top 1% of the list, when sorted by the model score.)

This table shows the output name, the percentage of samples that are in that particular category, the number of category members, and the percentage of true members found. Percentages that are less than 100% are in **bold**.

Output	Category %	1%	5%	10%	25%	50%	75%	90%	95%	99%
DILI new Bayesian ECFC6 EVEN more des II	53.559%	**1.9%**	**8.9%**	**17.7%**	**43%**	**75.9%**	**94.9%**	**99.4%**	**100%**	**100%**

Percentile Results

Back to Top

This table shows, for each model, the cutoff needed to capture a particular percentage of the good samples. For each cutoff, it shows below the estimated percentages of false positives and true negatives for the non-good samples. This table is designed to help you pick the cutoff value that best balances your desire to capture as many good samples as possible, while keeping the number of false positives at a minimum.

The rates shown in this table are estimates derived from the cross-validated data; the actual numbers you would find on your own data may vary.

Cutoff which lead to 10% or greater false positives are displayed in **bold** for ease of identification.

Model Name	99%	95%	90%	70%	50%	30%	10%	5%	1%
DILI new Bayesian ECFC6 EVEN more des II	-11.190 **62%**/38%	-7.008 **48%**/52%	-4.739 **40%**/60%	-2.230 **32%**/68%	-2.230 **18%**/82%	7.924 9%/91%	10.433 6%/94%	12.703 4%/96%	16.884 2%/98%

Category Statistics Results

Back to Top

This table shows, for each category, statistics derived from the cross-validated predictions of the model built for that category as applied to members of that category and non-members of that category. For each group, the number of members/nonmembers (N) is given; the mean prediction for each subset (Mean); and the estimate standard deviation of the predictions for each subset (StdDev).

(Categories with one or no members do not have a mean and standard deviation, as there are too few predictions upon which to base them during cross-validation. Also, occasionally categories may contain many duplicate or highly-similar compounds which predict close or identical values, causing them to have unusually low standard deviation values. These low values may be adjusted at time of use of these standard deviations for predicting, for example, percentile results.)

Output	Category N	Category Mean (±StdDev)	Noncategory N	Noncategory Mean (±StdDev)
DILI new Bayesian ECFC6 EVEN more des II	158	2.85 (±5.97)	137	-7.81 (±11.33)

Non-validated Models Results

Back to Top

Training Data Information

Back to Top

The properties used to provide the variables were: **ALogP**; **ECFC_6**; **Apol**; **logD**; **Molecular_Weight**; **Num_AromaticRings**; **Num_H_Acceptors**; **Num_H_Donors**; **Num_Rings**; **Num_RotatableBonds**; **Molecular_PolarSurfaceArea**; **Molecular_SurfaceArea**; **Wiener**; **Zagreb**

The test to identify "good" samples is:

```
property("DILI_Bins_Binary *") is defined AND property("DILI_Bins_Binary *") = 1;
```

You can extend this model by adding your own training data to it to create a new model, but because the original training data is no longer available, you will not be able to re-validate the new model. This extending is done using the *New Model from Old* component. The new training samples must already have the appropriate properties as specified above (though properties that can be calculated-on-demand will be). The "good" samples must be marked so that they will be correctly identified by the aforementioned test.

Model Construction Information

Back to Top

Model construction information:

Post-processing was performed to remove low-information bins. Low-information bins are those who have: normalized estimates in the range [-0.05, 0.05].

For each property, the following table gives the original number of bins (*Original*), the number removed due to too few samples (*TooFew*), the number removed due to a poor normalized estimate (*Noninformative*), and the final number of bins saved in the model (*Final*).

Property	Original	TooFew	Noninformative	Final
ALogP	11	0	1	10
ECFC_6	7094	0	437	6657
Apol	11	0	1	10
logD	11	0	2	9
Molecular_Weight	11	0	0	11
Num_AromaticRings	5	0	2	3
Num_H_Acceptors	8	0	3	5
Num_H_Donors	6	0	1	5

Num_Rings	5	0	3	2
Num_RotatableBonds	9	0	2	7
Molecular_PolarSurfaceArea	11	0	1	10
Molecular_SurfaceArea	11	0	2	9
Wiener	10	0	2	8
Zagreb	10	0	2	8

Drug Metabolism and Disposition

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)

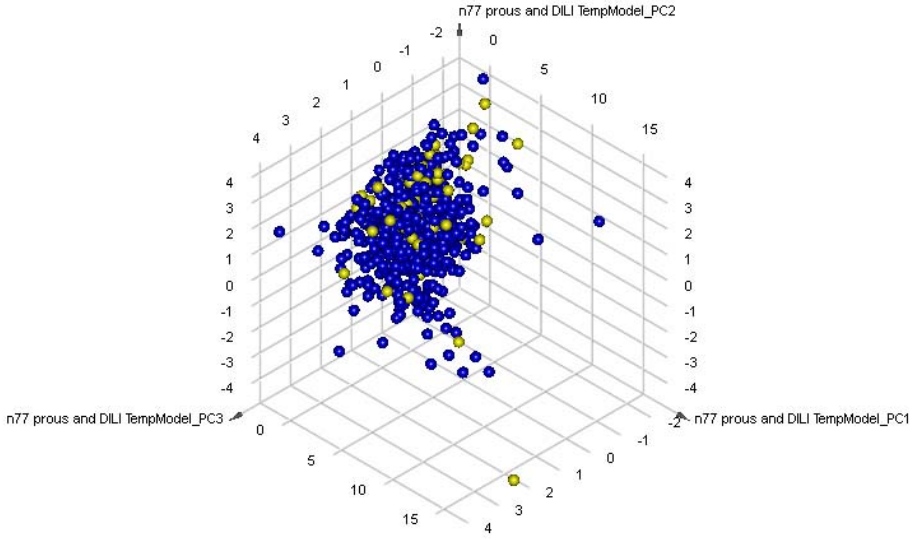
Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587. (AJW)

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

Supplemental Figure 1.

PCA analysis for DILI combined training and test set compared with a set of 77 recently approved drugs.

532 compounds in the DILI training and test set (N= 532, blue) were compared with 77 recently approved drugs obtained from the Prous database. 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. 0.83 % of the variance was explained with the first three principal components. One outlier approved drug molecule is Sugammadex (bottom, yellow).



Drug Metabolism and Disposition

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)

Supplemental Table 2. Subset of compounds of most interest. False =DILI negative (0), True = DILI Positive (1).

Compound Name	DILI Bayesian Prediction	Human DILI Bin (1=positive; 0=negative)	DILI Bayesian score	DILI Bayesian Closest Similarity	DILI Bayesian Closest Sample	True Positive	True Negative	False Positive	False Negative
Alpidem	TRUE	1	2.69	0.31	Sample 249	1			
Benoxaprofen	TRUE	1	4.55	0.57	Sample 126	1			
Bromfenac	TRUE	1	2.50	0.48	Sample 295	1			
Candesartan	FALSE	0	-6.75	0.37	Sample 270		1		
Ciglitizone	FALSE	1	-8.23	0.48	Sample 27				1
Dilevalol	TRUE	1	2.74	1.00	Sample 157				
Entacapone	TRUE	0	3.84	0.32	Sample 204			1	
Flunoxaprofen	TRUE	0	2.11	0.57	Sample 126			1	
Ibufenac	FALSE	1	-3.01	0.43	Sample 53				1

Ibuprofen	FALSE	0	-0.81	0.50	Sample 126		1			
Irbesartan	FALSE	0	-5.62	0.24	Sample 227		1			
Ketoprofen	TRUE	0	7.35	0.55	Sample 126			1		
Losartan	FALSE	1	-8.59	0.29	Sample 270					1
Lumiracoxib	TRUE	1	5.33	0.55	Sample 173	1				
Pemoline	TRUE	1	0.29	0.41	Sample 115	1				
Pirprofen	TRUE	1	3.65	0.54	Sample 126	1				
Tasosartan	FALSE	1	-2.87	0.24	Sample 197					1
Tolcapone	TRUE	1	7.71	0.35	Sample 162	1				
Troleandomycin	FALSE	1	-14.09	0.46	Sample 106					1
Ximelagatran	FALSE	1	-1.55	0.35	Sample 165					1
Zolpidem	TRUE	0	1.19	0.28	Sample 281			1		
Azithromycin	FALSE	0	-16.25	0.80	Sample 105		1			
Buspirone	FALSE	0	-3.12	0.29	Sample 258		1			
Diclofenac	TRUE	1	5.69	0.56	Sample 173	1				

Nefazodone	TRUE	1	3.73	0.67	Sample 283	1			
Pioglitazone	FALSE	0	-5.14	0.40	Sample 27		1		
Propranolol	FALSE	0	-9.58	0.64	Sample 232		1		
Rosiglitazone	FALSE	0	-1.96	0.35	Sample 27		1		
Telithromycin	FALSE	1	-6.25	0.39	Sample 105				1
Troglitazone	FALSE	1	-2.52	0.37	Sample 27				1
Valsartan	FALSE	0	-2.26	0.32	Sample 250		1		
Sudoxicam	TRUE	1	2.33	0.35	Sample 45	1			
Meloxicam	TRUE	0	10.07	0.64	Sample 233			1	
Olmesartan	FALSE	0	-12.75	0.32	Sample 270		1		
Celecoxib	TRUE	1	6.15	0.31	Sample 160	1			
Lapatinib	TRUE	1	5.78	0.31	Sample 138	1			
Gefitinib	TRUE	1	5.51	0.37	Sample 18	1			
					Sum	13	10	5	8

Drug Metabolism and Disposition

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)

Supplemental Table 3. Compounds greater than 70% similar to training set. False =DILI negative (0), True = DILI Positive (1).

Compound Name	DILI Bayesian Prediction	Human DILI Bin (1=positive; 0=negative)	DILI Bayesian score	DILI Bayesian Closest Similarity	DILI Bayesian Closest Sample	True Positive	True Negative	False Positive	False Negative
Streptomycin	FALSE	0	-45.49	0.97	Sample 264		1		
Quinine Sulfate	TRUE	1	30.57	0.96	Sample 251	1			
Clarithromycin	FALSE	1	-13.31	0.91	Sample 105				1
Tobramycin	FALSE	1	-38.58	0.91	Sample 154				1
Glutethimide	FALSE	1	-6.73	0.81	Sample 16				1
Azithromycin	FALSE	0	-16.25	0.80	Sample 105		1		

Nafcillin Sodium	TRUE	0	8.21	0.80	Sample 182			1	
Tolazamide	FALSE	1	-2.78	0.80	Sample 139				1
Ampicillin Sodium	TRUE	0	2.95	0.79	Sample 220			1	
Ifosfamide	TRUE	1	3.09	0.79	Sample 78	1			
Terbutaline Sulfate	TRUE	1	1.20	0.79	Sample 180	1			
Adriamycin	FALSE	0	-25.50	0.78	Sample 146		1		
Doxorubicin HCl	FALSE	1	-25.50	0.78	Sample 146				1
Calcifediol	FALSE	1	-18.49	0.78	Sample 103				1
Econazole Nitrate	FALSE	0	-13.38	0.78	Sample 191		1		
Sulconazole Nitrate	FALSE	0	-9.79	0.78	Sample 191		1		
Fexofenadine	FALSE	0	-30.23	0.78	Sample 272		1		
Methyldopa	FALSE	1	-8.21	0.78	Sample 162				1

Acenocoumarol	FALSE	1	-7.29	0.77	Sample 290				1
Epirubicin	FALSE	0	-25.70	0.76	Sample 146		1		
Fialuridine	TRUE	1	3.44	0.76	Sample 120	1			
Grepafloxacin	TRUE	1	3.23	0.74	Sample 64	1			
Lactose	FALSE	0	-14.06	0.73	Sample 90		1		
Ethinyl estradiol	TRUE	1	7.39	0.72	Sample 108	1			
Prochlorperazine Maleate	TRUE	1	6.07	0.72	Sample 58	1			
Atomoxetine	FALSE	1	-9.79	0.71	Sample 125				1
Iproniazid	TRUE	1	3.58	0.71	Sample 150	1			
Minocycline HCl	TRUE	1	12.49	0.70	Sample 86	1			
					sum	9	8	2	9

Drug Metabolism and Disposition

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)

Supplemental Table 4. Mean physicochemical properties for Recently approved drugs from Prous database

Descriptor	Recently approved drugs from Prous database (N = 77)
ALogP	2.09 ± 3.49
Apol	16315.18 ± 9937.28
LogD	1.42 ± 3.52
MW	427.05 ± 280.31
Number of rotatable bonds	7.05 ± 7.56
Number of rings	3.44 ± 1.70
Number of aromatic rings	2.02 ± 1.21
Number of H bond acceptors	6.01 ± 6.73

Number of H bond donors	2.37 ± 3.28
Molecular surface area	413.89 ± 264.25
Molecular polar surface area	110.85 ± 133.18
Wiener Index	5843.43 ± 17813.73
Zagreb Index	158.23 ± 97.50

Drug Metabolism and Disposition

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)

Supplemental Text

Each filter has a minimum and maximum number of times that it is allowed to map (in parentheses). The following SMARTS were used with default settings: Sulfonyl halide (0-1), Primary alkyl halide (0-1), Epoxide or aziridine (0-1), Sulfonate ester (0-1), Phosphonate ester (0-1), Long aliphatic chain (0-1), Peroxide (0-1), 1-2 Dicarbonyl (0-1), Acid halide (0-1), Non-Hydrogen atoms (2-35), Carbons (1-30), N-O-S (0-9),Sulfonyl halides (0-0), Acid halides (0-0), Alkyl halides (0-0), Acid anhydrides (0-0), Isocyanates or Isothiocyanates (0-0), Thiocyanates (0-0), Carbodiimides (0-0), Sulfonates (0-0), Acylhydrazides (0-0), Isonitriles (0-1), Imines (0-0), Acrylonitriles (0-0), Propenals (0-0), Macrocycles (0-0), Phosphorus 3 (0-0), Hexanes (0-0), 5 rotatable bonds (0-0), Aliphatic alcohols (0-3), Perchlorates (0-0), Fluorines (0-7), Cl-Br-I (0-3), P halides (0-0), Cyanohydrines (0-0), Sulfate esters (0-0), Pentafluorophenyl esters (0-0), Paranitrophenyl esters (0-0), HOBt esters (0-0), Lawesson's reagents (0-0), Phosphoramides (0-0), Aromatic azides (0-0), Quaternary C-Cl-I-P-S (0-0), Beta carbonyl quaternary N (0-0), Acyl cyanides (0-0), Sulfonyl cyanides (0-0), Thioepoxides (0-0), Benzylic quaternary N (0-0), Di or Triphosphates (0-0), Aminooxy-oxo (0-0), Nitros (0-1), N-halides (0-0), Aldehyde (0-1), Cyano (0-1), Acid halides (0-0), Carbazides (0-0), Sulfate esters (0-0), Sulfonates (0-0), Acid anhydrides (0-0), Peroxides (0-0), Pentafluorophenyl esters (0-0), Paranitrophenyl esters (0-0), Esters of HOBT (0-0), Isocyanates and Isothiocyanates (0-0), Triflates (0-0), Lawesson reagent and derivatives (0-0), Phosphoramides (0-0), Aromatic azides (0-0), Beta carbonyl quaternary Nitrogen (0-0), Acylhydrazide (0-0), Quaternary C or C1 or I or P or S (0-0), Phosphoranes (0-0), Nitroso (0-0), P or S Halides (0-0), Carbodiimide (0-0), Isonitrile (0-0), Triacyloximes (0-0), Cyanohydrins (0-0), Acyl cyanides (0-0), Sulfonyl cyanides (0-0),

Cyanophosphonates (0-0), Azocyanamides (0-0), Azoalkanals (0-0), Aliphatic methylene chains of 7 carbons or more in length (0-0), Compounds with 4 or more acidic groups (0-0), Crown ethers (0-0), Disulfides (0-0), Thiols (0-0), Epoxides or Thioepoxides or Aziridines (0-0), 2-4-5 trihydroxyphenyl (0-0), 2-3-4 trihydroxyphenyl (0-0), Hydrazothiourea (0-0), Thiocyanate (0-0), Benzylic quaternary Nitrogen (0-0), Thioesters (0-0), Cyanamides (0-0), Four numbered Lactones (0-0), Di and Triphosphates (0-0), Betalactams (0-0), Quinones (0-0), Polyenes (0-0), Saponin derivatives (0-0), Cytochalasin derivatives (0-0), Cycloheximide derivatives (0-0), Monensin derivatives (0-0), Cyanidin derivatives (0-0) and Squalestatin derivatives (0-0). A molecule must match this filter or it will be classed as failing the filter.

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

Ekins, Antony J. Williams and Jinghai J. Xu

Drug Metabolism and Disposition

Supplemental Table 1

Training set

Chemical Name	Formula	PubChem_CID	DILI_Bins_Binary *
1,2,3,4,5,6-hexachlorocyclohexane	C6H6Cl6	727	1
3-Acetamidophenol (AMAP)	C8H9NO2	12,124	0
Acetaminophen	C8H9NO2	1,983	1
Acetazolamide	C4H6N4O3S2	1,986	1
Acetohexamide	C15H20N2O4S	1,989	1
Acetylcholine Chloride	C7H16ClNO2	187	0
Acitretin	C21H26O3	5,284,513	1
acivicin	C5H7ClN2O3	294,641	0
Adenosine	C10H13N5O4	60,961	0
Albendazole	C12H15N3O2S	2,082	1
Allopurinol	C5H4N4O	2,094	1
Alloxan Hydrate	C4H4N2O5	312,231	0
Amantadine HCl	C10H18ClN	2,130	0
Amiloride HCl	C6H9Cl2N7O	16,231	0
Aminobenzoate Potassium	C7H6KNO2	978	0
Aminoglutethimide	C13H16N2O2	2,145	0
Aminosalicylic Acid	C7H7NO3	4,649	1
amiodarone	C25H29I2NO3	2,157	1
Amitriptyline HCl	C20H24ClN	2,160	0
ammonium chloride	H4ClN	22,985	0
amoxapine	C17H16ClN3O	2,170	0
Amrinone	C10H9N3O	3,698	1
Amsacrine HCl	C21H20ClN3O3S	2,179	1
Ascorbate	C6H8O6	235	0
Aspirin	C9H8O4	2,244	0
Astemizole	C28H31FN4O	2,247	0

Atenolol	C14H22N2O3	2,249	0
atractyloside	C30H44K2O16S2	5,702,200	1
Azaserine	C5H7N3O4	5,284,344	1
Azathioprine	C9H7N7O2S	2,265	1
Aztreonam (Z-isomer)	C13H17N5O8S2	5,742,832	0
Bacitracin	C66H103N17O16S	6,474,109	0
Bambuterol	C18H29N3O5	54,766	0
Beclomethasone Dipropionate	C28H37ClO7	20,469	0
Benazepril	C24H28N2O5	5,362,124	0
Benzbromarone	C17H12Br2O3	2,333	1
Bepridil HCl	C24H35ClN2O	2,351	1
Betahistine DiHCl	C8H14Cl2N2	2,366	1
Betamethasone	C22H29FO5	9,782	0
Bezafibrate	C19H20ClNO4	39,042	0
Bicalutamide	C18H14F4N2O4S	56,069	1
Biotin	C10H16N2O3S	171,548	0
Brompheniramine Maleate	C20H23BrN2O4	6,834	0
Bumetanide	C17H20N2O5S	2,471	1
Bupivacaine	C18H31ClN2O2	2,474	0
Bupropion HCl	C13H19Cl2NO	444	0
Busulphan	C6H14O6S2	2,478	1
cadmium chloride	CdCl2	176	1
Calcium Pantothenate	C18H32CaN2O10	6,109	0
Captopril	C9H15NO3S	44,093	1
carbendazim	C9H9N3O2	25,429	1
Carbenoxolone Disodium	C34H48Na2O7	6,419,769	0
Carbidopa	C10H14N2O4	34,359	0
Cefoperazone Dihydrate	C25H31N9O10S2	6,420,003	0
Chenodiol	C24H40O4	10,133	1
Chloramphenicol Palmitate	C27H42Cl2N2O6	5,959	1
Chlorpheniramine Maleate	C20H23ClN2O4	2,725	1
chlorpromazine	C17H19ClN2S	2,726	1
Chlortetracycline HCl	C22H24Cl2N2O8	5,280,963	1
Chlorzoxazone	C7H4ClNO2	2,733	1
Ciclopirox	C12H17NO2	38,911	0
Cimetidine	C10H16N6S	2,756	0
Ciprofibrate	C13H14Cl2O3	2,763	1
Ciprofloxacin HCl	C17H19ClFN3O3	2,764	1

Flupenthixol	C23H25F3N2OS	5,281,881	0
Cisapride	C23H29ClFN3O4	2,769	0
citalopram	C20H21FN2O	2,771	0
Cladribine	C10H12ClN5O3	20,279	0
clinafloxacin	C17H17ClFN3O3	60,063	0
Clomiphene Citrate	C32H36ClNO8	3,033,832	1
Clomipramine	C19H23ClN2	2,801	1
Clonidine HCl	C9H10Cl3N3	2,803	1
Clotrimazole	C22H17ClN2	2,812	0
Clozapine	C18H19ClN4	2,818	1
Colchicine	C22H25NO6	6,167	0
Cromolyn	C23H16O11	2,882	0
Cyanocobalamin	C63H88CoN14O14P	5,460,135	0
Cyclophosphamide	C7H15Cl2N2O2P	2,907	1
cyclosporin A	C62H111N11O12	5,284,373	1
Cyproheptadine HCl	C21H22ClN	2,913	1
Cyproterone acetate	C24H29ClO4	5,284,537	1
Danazol	C22H27NO2	28,417	1
Dantrolene Sodium	C14H9N4NaO5	9,568,637	1
Dapsone	C12H12N2O2S	2,955	1
Deferoxamine Mesylate	C26H52N6O11S	2,973	1
Demeclocycline HCl	C21H22Cl2N2O8	5,281,008	1
desipramine	C18H22N2	2,995	0
Dexamethasone	C22H29FO5	5,743	0
Dextromethorphan HBr	C18H26BrNO	5,360,696	0
d-galactosamine	C6H13NO5	24,154	1
Didanosine	C10H12N4O3	50,599	1
Diethylcarbamazine	C10H21N3O	3,052	1
Diflunisal	C13H8F2O3	3,059	1
Digoxin	C41H64O14	30,322	0
dilthiazem HCl	C22H27ClN2O4S	62,920	0
Diphenhydramine HCl	C17H22ClNO	3,100	0
Disopyramide Phosphate	C21H32N3O5P	107,858	1
Dobutamine HCl	C18H24ClNO3	36,811	0
Donepezil	C24H29NO3	3,152	0
Dopamine	C8H11NO2	681	0
Doxycycline hyclate	C46H58Cl2N4O18	5,281,011	1
Edrophonium Chloride	C10H16ClNO	3,202	0

Ergocalciferol	C28H44O	5,280,793	0
Ergonovine Maleate	C23H27N3O6	6,437,065	0
erythromycin	C37H67NO13	12,560	0
Erythromycin estolate (1)	C52H97NO18S	12,560	1
Eserine	C15H21N3O2	5,983	0
estradiol	C18H24O2	5,991	0
estradiol 17b glucuronide	C24H32O8	66,424	1
Estrone	C18H22O2	9,919	1
Ethynodiol Diacetate	C24H32O4	6,432,306	1
Etoposide	C29H32O13	36,462	1
Famotidine	C8H15N7O2S3	3,325	0
FCCP	C10H5F3N4O	3,330	1
Felbamate	C11H14N2O4	3,331	1
Fenofibrate	C20H21ClO4	3,339	1
Fenoprofen Sodium	C15H13NaO3	3,342	1
Fenoterol HBr	C17H22BrNO4	3,343	1
Flecainide Acetate	C19H24F6N2O5	41,022	1
Floxuridine	C9H11FN2O5	5,790	1
Fluconazole	C13H12F2N6O	3,365	1
Flucytosine	C4H4FN3O	3,366	1
Fludrocortisone Acetate	C23H31FO6	5,875	0
Flumazenil	C15H14FN3O3	3,373	0
fluoxetine	C17H18F3NO	3,386	0
Flurbiprofen	C15H13FO2	3,394	1
Flutamide	C11H11F3N2O3	3,397	1
Fluvastatin	C24H26FNO4	5,281,101	0
fluvoxamine	C15H21F3N2O2	5,324,346	1
Folate	C19H17N7O6	6,037	0
Furazolidone	C8H7N3O5	5,323,714	1
Furosemide	C12H11ClN2O5S	3,440	1
Gabapentin	C9H17NO2	3,446	0
Gallamine Triethiodide	C30H60I3N3O3	6,172	0
Gallium Nitrate Hydrate	H2GaN3O10	61,635	0
gatifloxacin	C19H22FN3O4	5,379	0
Gemfibrozil	C15H22O3	3,463	1
Glafenine	C19H17ClN2O4	3,474	1
Gliclazide	C15H21N3O3S	3,475	0
Glimepiride	C24H34N4O5S	3,476	0

Griseofulvin	C17H17ClO6	441,140	1
Hycanthone	C20H24N2O2S	3,634	1
Hydrochlorothiazide	C7H8ClN3O4S2	3,639	1
Hydrocortisone	C21H30O5	26,133	0
Hydroxyurea	CH4N2O2	3,657	1
Idarubicin HCl	C26H28ClNO9	107,865	0
Idoxuridine	C9H11IN2O5	5,905	1
Imipramine HCl	C19H25ClN2	3,696	1
indomethacin	C19H16ClNO4	3,715	1
isoniazid	C6H7N3O	3,767	1
Isoproterenol HCl	C11H18ClNO3	3,779	0
Isotretinoin	C20H28O2	444,795	1
Isoxsuprine HCl	C18H24ClNO3	3,783	0
Kanamycin Sulfate	C18H38N4O15S	441,374	0
Ketorolac Tromethamine	C19H24N2O6	3,826	1
Ketotifen	C19H19NOS	3,827	0
Labetalol	C19H24N2O3	3,869	1
Lamivudine	C8H11N3O3S	60,825	0
L-arginine	C6H14N4O2	6,322	0
Leflunomide	C12H9F3N2O2	3,899	1
L-Ethionine	C6H13NO2S	25,674	1
Levodopa	C9H11NO4	6,047	0
levofloxacin	C18H20FN3O4	149,096	0
Lidocaine	C14H22N2O	3,676	0
Lisinopril	C21H31N3O5	5,362,119	1
Lithocholic acid	C24H40O3	9,903	1
Lomefloxacin HCl	C17H20ClF2N3O3	3,948	1
Loperamide HCl	C29H34Cl2N2O2	3,955	0
Lovastatin	C24H36O5	53,232	0
maleic acid	C4H4O4	21,954	1
Maprotiline	C20H23N	4,011	0
Mebendazole	C16H13N3O3	4,030	1
Meclofenamate Sodium	C14H12Cl2NNaO3	4,037	1
Medroxyprogesterone Acetate	C24H34O4	5,702,080	0
Mefenamic Acid	C15H15NO2	4,044	1
Melatonin	C13H16N2O2	896	0
Memantine	C12H21N	4,054	0
Mercaptopurine	C5H4N4S	667,490	1

Mesoridazine Besylate	C27H32N2O4S3	4,078	1
Metaproterenol Sulfate	C22H36N2O10S	4,086	1
Methacycline HCl	C22H23ClN2O8	5,281,092	1
Methicillin Sodium	C17H19N2NaO6S	23,689,098	1
Methimazole	C4H6N2S	1,349,907	1
Methotrexate	C20H22N8O5	126,941	1
Methoxamine HCl	C11H18ClNO3	6,082	0
Methylergonovine Maleate	C24H29N3O6	8,226	0
Methysergide Maleate	C21H27N3O2	9,681	0
Metoclopramide HCl	C14H23Cl2N3O2	4,168	1
Metronidazole	C6H9N3O3	4,173	1
Mexiletine HCl	C11H18ClNO	21,467	1
Miconazole	C18H14Cl4N2O	4,189	0
Mitoxantrone diHCl	C22H30Cl2N4O6	4,212	1
Molindone HCl	C16H25ClN2O2	23,897	1
Monocrotaline	C16H23NO6	104,764	1
montelukast	C35H36ClNO3S	5,281,040	0
Nadolol	C17H27NO4	39,147	0
Nalidixic Acid	C12H12N2O3	4,421	1
Nalmefene	C21H25NO3	5,284,594	0
Naltrexone	C20H23NO4	5,360,515	1
Niacin	C6H5NO2	938	1
Nicardipine HCl	C26H30ClN3O6	4,474	1
Nifedipine	C17H18N2O6	4,485	1
Nimesulide	C13H12N2O5S	4,495	1
Nimodipine	C21H26N2O7	4,497	1
Nisoldipine	C20H24N2O6	4,499	0
Nocodazole	C14H11N3O3S	4,122	0
nomifensine	C16H18N2	4,528	1
Norethindrone	C20H26O2	6,230	1
Norgestrel	C21H28O2	5,991	1
Nortriptyline HCl	C19H22ClN	4,543	1
Novobiocin	C31H36N2O11	4,546	1
Orphenadrine Citrate	C24H31NO8	4,601	0
oxybendazole	C12H15N3O3	4,622	1
Oxybutynin HCl	C22H32ClNO3	4,634	0
Oxyphenonium	C21H34NO3	5,749	0
Pamidronate	C3H9NNa2O7P2	73,351	0

Paromomycin Sulfate	C23H47N5O18S	165,580	0
Paroxetine	C19H20FNO3	43,815	0
p-bromophenol	C6H5BrO	7,808	1
Penicillin G Sodium	C16H17N2NaO4S	2,349	1
Perhexilene	C19H35N	4,746	1
Phenacetin	C10H13NO2	4,754	1
Phenazopyridine HCl	C11H12ClN5	4,756	1
phenelzine	C8H12N2	61,100	0
Phenoxybenzamine HCl	C18H23Cl2NO	4,768	0
Phentolamine Mesylate	C18H23N3O4S	91,430	1
Phenylbutazone	C19H20N2O2	4,781	1
Phenylpropanolamine HCl	C9H14ClNO	26,934	0
phenytoin	C15H12N2O2	1,775	1
Pilocarpine	C11H16N2O2	5,910	0
Pinacidil	C13H21N5O	4,826	0
Pindolol	C14H20N2O2	4,828	0
Piroxicam	C15H13N3O4S	5,280,452	1
potassium dichromate	Cr2K2O7	8,232	1
Praziquantel	C19H24N2O2	4,891	0
Prednisone	C21H26O5	5,865	0
Primaquine Phosphate	C15H27N3O9P2	4,908	0
Primidone	C12H14N2O2	4,909	0
Procarbazine HCl	C12H20ClN3O	4,915	1
Progesterone	C21H30O2	12,419	1
Promazine HCl	C17H21ClN2S	4,926	0
Promethazine HCl	C17H21ClN2S	4,927	0
Propafenone HCl	C21H28ClNO3	4,932	0
Pseudoephedrine HCl	C10H16ClNO	7,028	0
puromycin	C22H29N7O5	4,984	1
Pyrazinamide	C5H5N3O	1,046	1
Pyridostigmine Bromide	C9H13BrN2O2	4,991	0
Pyridoxine	C8H11NO3	1,054	0
Pyrimethamine	C12H13ClN4	4,993	1
Quinapril	C25H30N2O5	54,892	1
quinidine	C20H24N2O2	11,069	1
quinine	C20H24N2O2	8,549	1
raloxifene	C28H27NO4S	5,035	0
ranitidine	C13H22N4O3S	3,001,055	0

Retinoic Acid	C20H28O2	444,795	1
Ribavirin	C8H12N4O5	37,542	0
Riluzole	C8H5F3N2OS	5,070	1
Risperidone	C23H27FN4O2	5,073	1
Sertraline	C17H17Cl2N	68,617	0
Simvastatin	C25H38O5	54,454	0
Sorbitol	C6H12O6	5,780	0
Spironolactone	C24H32O4S	5,833	1
Stavudine	C10H12N2O4	18,283	1
Streptomycin Sulfate	C21H41N7O16S	19,649	0
Sulfasalazine	C18H14N4O5S	5,353,980	1
Sulindac	C20H17FO3S	1,548,887	1
Sumatriptan	C14H21N3O2S	5,358	0
tacrine	C13H14N2	1,935	0
tamoxifen	C26H29NO	2,733,526	1
telmisartan	C33H30N4O2	65,999	0
Temozolomide	C6H6N6O2	5,394	0
terfenadine	C32H41NO2	5,405	0
Tetracaine HCl	C15H25ClN2O2	8,695	0
tetracycline	C22H24N2O8	5,497,101	1
thioacetamide	C2H5NS	2,723,949	1
Thioguanine	C5H5N5S	2,723,601	1
Thiothixene	C23H29N3O2S2	941,651	0
tianeptine	C21H25ClN2O4S	68,870	1
Ticlopidine	C14H14ClNS	5,472	1
Timolol Maleate	C17H28N4O7S	5,478	1
Tolmetin	C15H15NO3	5,509	1
Tranylcypromine HCl	C9H12ClN	19,493	0
Trazodone HCl	C19H23Cl2N5O	5,533	1
trifluoperazine	C21H24F3N3S	5,566	1
Trimethadione	C6H9NO3	5,576	1
Trovafloxacin	C20H15F3N4O3	483,952	1
Ursodeoxycholic acid	C24H40O4	31,401	0
Vancomycin	C66H75Cl2N9O24	444,193	1
Vidarabine	C10H15N5O5	21,704	0
Warfarin	C19H16O4	6,691	0
Zafirlukast	C31H33N3O6S	5,717	1
Zalcitabine	C9H13N3O3	24,066	0

Zidovudine	C10H13N5O4	35,370	1
Zileuton	C11H12N2O2S	60,490	1
Zomepirac	C15H14ClNO3	5,733	0

Test set

Chemical Name	**Formula**	**PubChem_CID**	**DILI_Bins_Binary ***
sudoxicam	C13H11N3O4S2		1
meloxicam	C14H13N3O4S2		0
olmesartan	C29H30N6O6		0
celecoxib	C17H14F3N3O2S		1
lapatinib	C29H26ClFN4O4S		1
gefitinib	C22H24ClFN4O3		1
6-mercaptopurine	C5H4N4S		1
Abacavir	C14H18N6O		1
Acenocoumarol	C19H15NO6		1
Acetylcysteine	C5H9NO3S		0
Aclarubicin HCl	C42H54ClNO15		0
adriamycin	C27H30ClNO11		0
aflatoxin B1	C17H12O6		1
Allyl alcohol	C3H6O		1
allyl formate	C4H6O2		1
alpidem	C21H23Cl2N3O		1
Amineptine HCl	C22H28ClNO2		1
Aminocaproic Acid	C6H13NO2		0
Amodiaquine HCl	C20H23Cl2N3O		1
Amoxicillin	C16H19N3O5S		0
Amphotericin B	C47H73NO17		0
Ampicillin Sodium	C16H18N3NaO4S		0
Anileridine HCl	C22H29ClN2O2		0
ANIT (1-Naphthyl isothiocyanate)	C11H7NS		1
Arsenic Trioxide	As2O3		1
atomoxetine	C17H21NO		1
Auranofin	C19H32AuO9PS		1
Aurothioglucose	C6H11AuO5S		1
aurothiolamate	C4H3AuNa2O4S		1
Azacytidine	C8H12N4O5		1
Azatadine Maleate	C24H26N2O4		1
Azlocillin Sodium	C20H22N5NaO6S		0

BEA (bromoethanamine)	C2H6BrN	1
benoxaprofen	C16H12ClNO3	1
Betaine HCl	C5H12ClNO2	0
Bismuth Subsalicylate	C7H5BiO4	0
Bithionol	C12H6Cl4O2S	1
BNIT (2-Naphthyl isothiocyanate)	C11H7NS	0
Bosentan	C27H29N5O6S	1
bromfenac	C15H12BrNO3	1
Bromobenzene	C6H5Br	1
buthinioninesulphoxime	C20H10Br4O10S2	1
Butoconazole Nitrate	C19H18Cl3N3O3S	0
Butylated hydroxytoluene	C15H24O	1
Caffeine	C8H10N4O2	0
Calcifediol	C27H44O2	1
candesartan	C24H20N6O3	0
Capsaicin	C18H27NO3	0
carbamzepine	C15H12N2O	1
carbon tetrachloride (2)	CCl4	1
Carboplatin	C6H12N2O4Pt	1
Cefadroxil	C16H19N3O6S	1
Cefamandole Sodium	C18H17N6NaO5S2	1
Cefotetan	C17H17N7O8S4	0
Cefotiam HCl	C18H24ClN9O4S3	1
Cefoxitin	C16H17N3O7S2	1
Ceftazidime	C22H22N6O7S2	0
Ceftriaxone Sodium E-Isomer	C18H17N8NaO7S3	1
cephaloridine	C19H17N3O4S2	1
Cephalothin Sodium	C16H15N2NaO6S2	1
Cephapirin Sodium	C17H16N3NaO6S2	0
Cephradine	C16H19N3O4S	0
cerivastatin	C26H34FNO5	1
Chloroform	CHCl3	1
Chloroquine Phosphate	C18H32ClN3O8P2	0
Chlorpropamide	C10H13ClN2O3S	1
Ciglitizone	C18H23NO3S	1
Cinchophen	C16H11NO2	1
cisplatin	H6Cl2N2Pt	1
Citicoline	C12H22N4O11P2	0

Clarithromycin	C38H69NO13	1
Clemastine Fumarate	C25H30ClNO5	0
Clindamycin HCl	C18H34Cl2N2O5S	1
Clofibrate	C12H15ClO3	1
Cloxacillin Sodium	C19H19ClN3NaO6S	1
Cyclizine	C18H22N2	1
Dacarbazine	C6H10N6O	1
Dactinomycin	C62H86N12O16	1
DCB (dichlorobenzene)	C6H4Cl2	1
Dextroamphetamine Sulfate	C18H28N2O4S	1
dichloroethylene	C2H4Cl2	1
dichlorophenyl succinimide	C9H10Cl2N2O	1
Dicloxacillin Sodium	C19H16Cl2N3NaO5S	1
Diethylhexylphthalate (phthalate ester)	C24H38O4	1
difluoropentane	C5H10F2	1
Dimercaprol	C3H8OS2	1
dimethylnitrosamine	C2H6N2O	1
dinitrophenol	C6H4N2O5	1
Di-n-pentyl-phthalate	C18H26O4	1
Diphenoxylate HCl	C30H33ClN2O2	0
Dipyridamole	C24H40N8O4	0
Diquat	C12H12Br2N2	1
Divalproex Sodium	C16H31NaO4	1
d-limonene	C10H16	1
Doxorubicin HCl	C27H30ClNO11	1
Econazole Nitrate	C18H16Cl3N3O4	0
entacapone	C14H15N3O5	0
Epirubicin	C27H29NO11	0
eprosartan	C23H24N2O4S	0
Ergotamine l-Tartrate	C37H41N5O11	0
Ethane Dimethane Sulfonate	C14H10Cl4	1
ethinyl estradiol	C20H24O2	1
Ethosuximide	C7H11NO2	1
ethylene glycol	C2H6O2	1
Fexofenadine	C32H39NO4	0
Fialuridine	C9H10FIN2O5	1
Fipexide	C20H21ClN2O4	1
Flufenamic Acid (Flufenamate)	C14H10F3NO2	0

flunoxaprofen	C16H12FNO3	0
Fluorouracil	C4H3FN2O2	1
Fluspirilene	C29H31F2N3O	1
Foscarnet	CH3O5P	0
Glutethimide	C13H15NO2	1
grepafloxacin	C19H22FN3O3	1
Guanethidine Sulfate	C10H24N4O4S	1
Hexachlorophene	C13H6Cl6O2	1
Hydrazine	H4N2	1
Hyoscyamine Sulfate	C17H25NO7S	1
Ibufenac	C12H16O2	1
Ibuprofen	C13H18O2	0
Ifosfamide	C7H15Cl2N2O2P	1
indacrinone	C18H14Cl2O4	1
indinavir sulphate	C36H49N5O8S	0
Iopamidol	C17H22I3N3O8	1
iproniazid	C9H13N3O	1
irbesartan	C25H28N6O	0
Isocarboxazid	C12H13N3O2	0
Isosorbide dinitrate	C6H8N2O8	0
Isoxicam	C14H13N3O5S	0
Ketoconazole	C26H28Cl2N4O4	1
ketoprofen	C16H14O3	0
Lactose	C12H22O11	0
Leucovorin Calcium	C20H21CaN7O7	0
Liothyronine	C15H12I3NO4	0
Loracarbef	C16H18ClN3O5	0
losartan	C22H23ClN6O	1
Lumiracoxib	C15H13ClFNO2	1
Menadione	C11H8O2	1
Mephobarbital	C13H14N2O3	1
Meprobamate	C9H18N2O4	1
Mestranol	C21H26O2	1
Metformin	C4H11N5	0
Methapyrilene	C14H19N3S	1
Methapyrilene	C14H19N3S	1
Methoxyacetic Acid	C3H6O3	1
Methyldopa	C10H13NO4	1

methylene dianiline	C16H18ClN3S	1
Methylphenidate	C14H19NO2	1
Metolazone	C16H16ClN3O3S	1
Mezlocillin Sodium	C21H24N5NaO8S2	1
mianserin	C18H20N2	1
Mibefradil	C29H38FN3O3	0
Microcystin-LR	C49H74N10O12	1
Minocycline HCl	C23H28ClN3O7	1
Mometasone Furoate	C27H30Cl2O6	0
Moricizine HCl	C22H26ClN3O4S	1
Moxalactam Disodium	C20H18N6Na2O9S	0
Myo-inositol	C6H12O6	0
N-acetyl cysteine	C5H9NO3S	0
Nafcillin Sodium	C21H21N2NaO5S	0
naproxen	C14H14O3	0
Nitrofurantoin	C8H6N4O5	1
o-bromophenol	C6H5BrO	1
Octyl Methoxycinnamate	C18H26O3	1
Oxaprozin	C18H15NO3	1
Oxyquinoline Sulfate	C9H9NO5S	1
Paclitaxel	C47H51NO14	1
PAP (para-aminophenol)	C18H24NO5PS2	0
paraquat	C14H20N2O8S2	1
Pargyline	C11H13N	0
pemoline	C9H8N2O2	1
Penbutolol Sulfate	C36H60N2O8S	1
Pentanoic Acid	C5H10O2	0
Permethrin	C21H20Cl2O3	0
phenothiazine	C12H9NS	1
phenyl isothiocyanate	C7H5NS	1
Pimozide	C28H29F2N3O	0
Pirprofen	C13H14ClNO2	1
probenecid	C13H19NO4S	0
Probucol	C31H48O2S2	0
Prochlorperazine Maleate	C28H32ClN3O8S	1
Propofol	C12H18O	0
Quinine Sulfate	C40H50N4O8S	1
ragaglitazar	C25H25NO5	0

Retinyl palmitate	C36H58O3	1
Rifabutin	C46H62N4O11	1
Rifampicin	C43H58N4O12	0
ritonavir	C37H48N6O5S2	1
rotenone	C23H22O6	1
Saquinavir base / mesylate	C38H50N6O5	1
Spectinomycin HCl	C14H25ClN2O7	1
Streptomycin	C21H39N7O12	0
Streptozocin	C8H15N3O7	0
Sulconazole Nitrate	C18H16Cl3N3O3S	0
Sulfabenzamide	C13H12N2O3S	0
Sulfamethizole	C9H10N4O2S2	1
Sulfanilamide	C6H8N2O2S	1
Sulfinpyrazone	C23H20N2O3S	1
Tasosartan	C23H21N7O	1
telenzepine	C19H22N4O2S	0
Temafloxacin	C21H18F3N3O3	1
Terbutaline Sulfate	C24H40N2O10S	1
Tetraethythiuran	C10H20N2S4	0
Theophylline	C7H8N4O2	0
Thiamine	C12H17ClN4OS	0
ticrynafen	C13H8Cl2O4S	1
Tobramycin	C18H37N5O9	1
Tolazamide	C14H21N3O3S	1
Tolbutamide	C12H18N2O3S	1
tolcapone	C14H11NO5	1
Tolrestat	C16H14F3NO3S	1
Topiramate	C12H21NO8S	0
trichloroethylene	C2HCl3	1
Trientine HCl	C6H19ClN4	1
Trimethobenzamide HCl	C21H29ClN2O5	1
Trimethoprim	C14H18N4O3	1
Tripelennamine HCl	C16H22ClN3	1
Troleandomycin	C41H67NO15	1
Tromethamine	C4H11NO3	0
TUDC (tauroursodeoxycholic acid)	C26H45NO5S	1
Uracil Mustard	C8H11Cl2N3O2	1
Valproic Acid	C8H16O2	1

Verapamil HCl	C27H39ClN2O4		0
Vinblastine Sulfate	C46H60N4O13S		0
Vincristine Sulfate	C46H58N4O14S		1
Ximelagatran	C24H35N5O5		1
zolpidem	C19H21N3O		0
Azithromycin	C38H72N2O12	447,043	0
buspirone	C21H31N5O2	2,477	0
diclofenac	C14H11Cl2NO2	3,033	1
Nefazodone	C25H32ClN5O2	4,449	1
Pioglitazone	C19H20N2O3S	4,829	0
Propranolol	C16H21NO2	4,946	0
Rosiglitazone	C18H19N3O3S	77,999	0
telithromycin	C43H65N5O10	3,002,190	1
troglitazone	C24H27NO5S	5,591	1
valsartan	C24H29N5O3	60,846	0