A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury

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

Collaborations in Chemistry, Jenkintown, Pennsylvania (S.E.); Department of Pharmaceutical Sciences, University of Maryland, College Park, Maryland (S.E.); Department of Pharmacology, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey (S.E.); Royal Society of Chemistry, Wake Forest, North Carolina (A.J.W.); and Merck & Co., Inc., Rahway, New Jersey (J.J.X.)

Received June 22, 2010; accepted September 15, 2010

ABSTRACT:

Drug-induced liver injury (DILI) is one of the most important reasons for drug development failure at both preapproval 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, a 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 results for internal validation. The Bayesian model with extended connectivity functional class fingerprints of maximum diameter 6 (ECFC_6) 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 Smiles Arbitrary Target Specification (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, such as the Abbott alerts, which captures thiol traps and other compounds, may be of use 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 cost-effective selection criteria before in vitro or in vivo experimental studies.

Introduction

Pharmaceutical researchers 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, excretion, and toxicity 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, and 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, 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 drugs are not approved and also the reason 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). Because 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 maximum therapeutic concentration (C_{max}), 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 hepato-
toxic (Xu et al., 2008). The overall concordance of the in vitro human
hepatocyte imaging assay technology (HIAT), when applied to approx-
imately 300 drugs and chemicals, is approximately 75% with regard to
clinical hepatotoxicity, with very few false-positive results (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
preclinical testing paradigm to select drug candidates with improved
therapeutic indexes for clinical hepatotoxicity.

The use of in vitro approaches still comes at a cost. First, the
compound has to physically be made and be available for
testing. Second, the screening system is relatively low throughput
compared with any primary screens and, as a result, whole compound
or vendor libraries cannot be cost effectively screened for prioriza-
tion. Third, 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 $C_{\text{max}}$ is often
imprecise before clinical testing in actual patients. A potential alter-
avative may be to use the historic DILI data to create a computa-
tional 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 before actual in vitro testing.

There have been many examples in which computational quantita-
tive structure-activity relationship 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, artificial neural networks, and machine learning algorithms
(OneR) (Cruz-Montagudo et al., 2008). These modeling techniques
were found to produce models with satisfactory internal cross-valida-
tion statistics (accuracy, sensitivity, and specificity more than 84, 78,
and 90%, respectively). These models were then tested on very small
sets of compounds (6 and 13 compounds, respectively) with more than
80% accuracy. A second study compiled a dataset of compounds
reported to produce a wide range of effects in the liver in different
species and then used binary quantitative structure-activity relation-
ship models (248 active and 283 inactive) to predict whether a
compound would be expected to produce liver effects in humans. The
resultant support vector machine 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., 2010b). 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 results for 626 Pfizer compounds (sen-
itivity of 46%, specificity of 73%, and concordance of 56% for the
latest version) (Greene et al., 2010).

In the current study, we have used a training set of 295 com-
ounds and a test set of 237 molecules. In contrast with earlier
studies, we have used a Bayesian classification approach (Xia et
al., 2004; Bender, 2005) with simple, interpretable molecular de-
scriptors as well as extended connectivity functional class finger-
prints 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 of the
532 molecules in the test and training set, because many pharma-
aceutical companies use Smiles Arbitrary Target Specification
(SMARTS) queries, which specify substructures of interest (http://
www.daylight.com/dayhtml/doc/theory/theory.smarts.html). Com-

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 means (±S.D.) of the calculated values are 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 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 to 2010 of 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 data). 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 (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, and 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 ROC area under the curve (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 DILI (Fig. 1A) and those that are not present in compounds causing DILI (Fig. 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 to generate the cross-validated 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 of >57%, specificity of >61%, and sensitivity of >52% did not seem to be different 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 property that was statistically different among 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 among 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 2.5.5 to select molecules with a common substructure and analyze those that cause DILI compared with those that do not. As demonstrated in Fig. 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 training set of 295 compounds (Supplemental Table 1). The concordance of ~60%, specificity of 67%, and sensitivity of 56% were comparable (Table 3) with results from internal validation (Table 1). A subset of 37 compounds (Supplemental Table 2) of most interest clinically (including similar compounds that were either DILI-causing or not) showed similar testing values with a concordance of greater than 63% (Table 2). The compounds of most interest can be defined as well hepatotoxic drugs [e.g., those hepatotoxic drugs cited elsewhere (U.S. Food and Drug Administration Guidance for Industry “Drug-Induced Liver Injury: Premarketing Clinical Evaluation,” 2009, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072278.pdf], 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 with the hepatotoxic ones (e.g., zolpidem versus alpidem and ibuprofen versus benoxaprofen). 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 sets covered overlapping or similar chemical space (Fig. 2A). However, there were some distinct compounds such as retinyl palmitate that were outside the training set (Fig. 2B). Therefore, focusing on compounds with a Tanimoto similarity greater than 0.7 left 28 compounds (Supplemental Table 3) whose Matthews correlation coefficient and concordance were similar to those of the complete test set. The specificity increased to 80% 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. Of the four sets of independent filters tested, the Abbott alerts had the highest concordance and sensitivity, whereas 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 those observed in Table 3 for the test sets with the Bayesian model.

**Discussion**

Pharmaceutical companies are eager 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 process 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 because of its importance in first-pass metabolism) (Ekins et al., 2003; O’Brien and de Groot, 2005). Idiosyncratic liver injury or drug-induced liver injury is 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...
FIG. 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. B, 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.
using either linear discriminant analysis, artificial neural networks, and machine learning algorithms (OneR) (Cruz-Monteagudo et al., 2008), support vector machine (Fourches et al., 2010a), 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 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, whereas 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).

Computational Bayesian models were recently developed for time-dependent inhibition of CYP3A4 using more than 2000 molecules for filtering of compounds that must be screened in vitro because of 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 curve scores of >0.85 and specificity of >61%, concordance of >57%, and sensitivity of >52% (Table 1). Using the ECFC_6 descriptors, we found that many of the fingerprints with high Bayesian scores that are present in many DILI compounds appeared to be reactive in nature, which could cause time-dependent inhibition of cytochromes P450, 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 (Fig. 1A) other than being generally hydrophobic and perhaps enabling increased accumulation. It is possible they may be hydroxylated and then form other metabolites that are in turn reactive. Further analysis of simple molecular descriptors calculated

### Table 1

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DILI-Negative (n = 137)</td>
<td>DILI-Positive (n = 158)</td>
</tr>
<tr>
<td><strong>External ROC Score</strong></td>
<td>0.62 ± 0.08</td>
<td>0.86 ± 0.01</td>
</tr>
<tr>
<td><strong>Internal ROC Score</strong></td>
<td>0.86 ± 0.01</td>
<td>0.84 ± 0.01</td>
</tr>
<tr>
<td><strong>Concordance</strong></td>
<td>58.48 ± 8.31</td>
<td>65.45 ± 15.22</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>65.15 ± 9.18</td>
<td>54.21 ± 9.69</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>61.81 ± 10.57</td>
<td>54.20 ± 9.83</td>
</tr>
</tbody>
</table>

* is test: P < 0.05.
** is test: P < 0.01.

### Table 2

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

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DILI-Negative (n = 137)</td>
<td>DILI-Positive (n = 158)</td>
</tr>
<tr>
<td>ALogP</td>
<td>1.31 ± 3.24</td>
<td>1.89 ± 2.47*</td>
</tr>
<tr>
<td>Apol</td>
<td>12.644.0 ± 6478.29</td>
<td>12.178.1 ± 6061.78</td>
</tr>
<tr>
<td>LogD</td>
<td>0.65 ± 3.43</td>
<td>1.23 ± 2.45</td>
</tr>
<tr>
<td>Molecular weight</td>
<td>355.67 ± 186.93</td>
<td>184.83 ± 184.83</td>
</tr>
<tr>
<td>No. of rotatable bonds</td>
<td>5.17 ± 4.35</td>
<td>4.47 ± 4.04</td>
</tr>
<tr>
<td>No. of rings</td>
<td>2.63 ± 1.51</td>
<td>2.51 ± 1.53</td>
</tr>
<tr>
<td>No. of aromatic rings</td>
<td>1.27 ± 1.04</td>
<td>1.36 ± 1.00</td>
</tr>
<tr>
<td>No. of H bond acceptors</td>
<td>5.20 ± 4.06</td>
<td>4.97 ± 3.61</td>
</tr>
<tr>
<td>Molecular surface area</td>
<td>352.68 ± 180.92</td>
<td>332.88 ± 183.78</td>
</tr>
<tr>
<td>No. of H bond donors</td>
<td>2.51 ± 2.92</td>
<td>2.09 ± 2.38</td>
</tr>
<tr>
<td>Molecular polar surface area</td>
<td>102.17 ± 92.83</td>
<td>96.47 ± 74.51</td>
</tr>
<tr>
<td>Wiener index</td>
<td>2383.90 ± 6919.65</td>
<td>1919.01 ± 5230.99</td>
</tr>
<tr>
<td>Zagreb index</td>
<td>122.38 ± 69.64</td>
<td>115.48 ± 64.32</td>
</tr>
</tbody>
</table>

### Table 3

Results of external validation of Bayesian model for DILI

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Matthews Correlation Coefficient</th>
<th>Concordance</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete test set (n = 237)</td>
<td>0.22</td>
<td>59.91</td>
<td>66.67</td>
<td>56</td>
</tr>
<tr>
<td>Subset of most interest (n = 37)</td>
<td>0.28</td>
<td>63.88</td>
<td>66.67</td>
<td>61.9</td>
</tr>
<tr>
<td>Compounds &gt;70% similar to training set (n = 28)</td>
<td>0.29</td>
<td>60.71</td>
<td>80.00</td>
<td>50</td>
</tr>
</tbody>
</table>
for the test and training sets showed only differences in ALogP for the training set, whereas 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 a lower sum of atomic polarizabilities (Apol)] but ALogP was not (Table 2). When we used branching as measured by the Zagreb index and a lower sum of atomic property values (Supplemental Table 4 compared with Supplemental Table 3). These combined analyses suggest that the test and training sets used for the DILI model are representative of current medicinal chemistry efforts.

A further approach we have taken on the basis of 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, compounds with false-positive results, and frequent hitters from their high throughput screening libraries or to filter vendor compounds (Williams et al., 2009). For example, REOS from Vertex (Walters and Murcko, 2002) and filters from Glaxo (Hann et al., 1999), Bristol-Myers Squibb (Pearce et al., 2006), Abbott (Huth et al., 2005, 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 more than 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 finding suggests that these SMARTS filters may be useful as a prescreen to remove potential DILI-causing compounds alongside the Bayesian models, which perform better.

**TABLE 4**

Summary of SMARTS filtering for the combined DILI test and training sets

<table>
<thead>
<tr>
<th>Filters/DILI Class</th>
<th>Molecules Passing Filter</th>
<th>Molecules Failing Filter</th>
<th>Concordance</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blake (Pfizer) total</td>
<td>283</td>
<td>249</td>
<td>50.7</td>
<td>54.7</td>
<td>47.9</td>
</tr>
<tr>
<td>DILI-negative</td>
<td>121</td>
<td>100</td>
<td>52.0</td>
<td>56.5</td>
<td>49.7</td>
</tr>
<tr>
<td>DILI-positive</td>
<td>162</td>
<td>149</td>
<td>47.6</td>
<td>52.0</td>
<td>46.3</td>
</tr>
<tr>
<td>Glaxo total</td>
<td>458</td>
<td>74</td>
<td>44.2</td>
<td>86.4</td>
<td>14.1</td>
</tr>
<tr>
<td>DILI-negative</td>
<td>191</td>
<td>30</td>
<td>63.5</td>
<td>70.0</td>
<td>56.2</td>
</tr>
<tr>
<td>DILI-positive</td>
<td>267</td>
<td>44</td>
<td>52.0</td>
<td>66.6</td>
<td>38.8</td>
</tr>
<tr>
<td>Abbott total</td>
<td>192</td>
<td>340</td>
<td>55.8</td>
<td>40.3</td>
<td>66.9</td>
</tr>
<tr>
<td>DILI-negative</td>
<td>89</td>
<td>132</td>
<td>57.1</td>
<td>55.8</td>
<td>49.7</td>
</tr>
<tr>
<td>DILI-positive</td>
<td>103</td>
<td>208</td>
<td>47.9</td>
<td>49.8</td>
<td>46.6</td>
</tr>
</tbody>
</table>

When we narrowed down the dataset to only those molecules with >70% similarity to the training set (n = 28) on the basis of the Tanimoto similarity (with MDL keys descriptors), the specificity increased to greater than 80% and concordance increased slightly to ~64%. Such an increase in concordance statistics is analogous to that observed with other computational chemistry predictions, because 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 (Fig. 2A), an approach we have used previously (Zientek et al., 2010), which shows that many of the molecules in the test set cover chemical space similar to that of those in the training set, whereas there are some compounds that may be outliers such as retinyl palmitate (Fig. 2B), which in this case 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 to 2010 extracted from the Prous Integrity database (Supplemental Fig. 1). Again we found that these molecules were distributed throughout the combined training and test sets, representative of overlap, which is also suggested from the mean physicochemical property values (Supplemental Table 4 compared with Supplemental Table 2). These combined analyses suggest that the test and training sets used for the DILI model are representative of current medicinal chemistry efforts.
In summary, we present the first large-scale testing of a machine learning model for DILI that uses similarly sized training and test sets. Our model may have use 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 that could be evaluated in future studies. It is also feasible that combinations of such computational approaches may also be of use to identify DILI-causing compounds.

Acknowledgments. S.E. gratefully acknowledges Dr. Maggie A. Z. Hucpey for chemistry discussions, Accelerx for providing Discover Studio 2.1 and 2.5.5., and the reviewers for their excellent suggestions.

References
Address correspondence to: Dr. Sean Ekins, Collaborations in Chemistry, 601 Runnymede Ave., Jenkintown, PA 19046. E-mail: ekinssean@yahoo.com