A Comprehensive Statistical Analysis of Predicting In Vivo Hazard Using High-Throughput In Vitro Screening

Russell S. Thomas,*† Michael B. Black,* Lili Li,+ Eric Healy,* Tzu-Ming Chu,† Wenjun Bao,† Melvin E. Andersen,* and Russell D. Wolfinger†

*The Hamner Institutes for Health Sciences, Research Triangle Park, North Carolina 27709; †SAS Institute Inc., Cary, North Carolina 27513

†To whom correspondence should be addressed at The Hamner Institutes for Health Sciences, 6 Davis Drive, Research Triangle Park, NC 27709. Fax: (919) 558-1300. E-mail: rthomas@thehamner.org.

Received February 24, 2012; accepted April 23, 2012

Over the past 5 years, increased attention has been focused on using high-throughput in vitro screening for identifying chemical hazards and prioritizing chemicals for additional in vivo testing. The U.S. Environmental Protection Agency’s ToxCast program has generated a significant amount of high-throughput screening data allowing a broad-based assessment of the utility of these assays for predicting in vivo responses. In this study, a comprehensive cross-validation model comparison was performed to evaluate the predictive performance of the more than 600 in vitro assays from the ToxCast phase I screening effort across 60 in vivo endpoints using 84 different statistical classification methods. The predictive performance of the in vitro assays was compared and combined with that from chemical structure descriptors. With the exception of chronic in vivo cholinesterase inhibition, the overall predictive power of both the in vitro assays and the chemical descriptors was relatively low. The predictive power of the in vitro assays was not significantly different from that of the chemical descriptors and aggregating the assays based on genes reduced predictive performance. Prefiltering the in vitro assay data outside the cross-validation loop, as done in some previous studies, significantly biased estimates of model performance. The results suggest that the current ToxCast phase I assays and chemicals have limited applicability for predicting in vivo chemical hazards using standard statistical classification methods. However, if viewed as a survey of potential molecular initiating events and interpreted as risk factors for toxicity, the assays may still be useful for chemical prioritization.

Key Words: predictive toxicology; QSAR; chemical structure; in vitro assays; high throughput screening; chronic toxicity; alternatives to animal testing; mode-of-action; ToxCast; developmental toxicity; reproductive toxicity.

Over the past 5 years, there has been increasing attention focused on the use of in vitro assays as alternatives to the traditional high-dose animal studies. This attention has been brought about by numerous factors including the European Cosmetics Directive and the release in the United States of the National Research Council’s report Toxicity Testing in the 21st Century (NRC, 2007). The attention has spawned multiple research efforts using in vitro data to identify chemical hazards. In the United States, the Environmental Protection Agency’s (EPA) ToxCast project and the Tox21 consortium have used high-throughput screening to characterize the in vitro biological activity of chemicals across multiple cellular pathways and biochemical targets with the intent to prioritize compounds for conventional toxicity testing (Dix et al., 2007; Kavlock et al., 2009). In Europe, multiple research consortia such as Sens-it-iv, ACuteTox, and carcinoGENOM-ICS have also examined the ability of in vitro systems for predicting in vivo responses (dos Santos et al., 2009; Jennen et al., 2010; Sjostrom et al., 2008). To date, the research efforts have produced mixed results.

The ToxCast project was developed as a phased approach with phase I including the in vitro testing of 309 unique chemicals across more than 600 in vitro assays. The 309 unique chemicals were primarily pesticides and high–production volume chemicals for which significant in vivo animal testing has been performed (Judson et al., 2009; Martin et al., 2009a). The ToxCast phase I efforts at the EPA have progressed beyond data collection, and the results have been published in a series of articles describing the general activity of the chemicals within specific in vitro assay types or focused on a specific set of endpoints (Houck et al., 2009; Judson et al., 2010; Knudsen et al., 2011; Martin et al., 2010; Reif et al., 2010; Rotroff et al., 2010). Apart from general descriptions of the results, a series of articles have assessed the ability of the in vitro assays for predicting selected in vivo endpoints (Judson et al., 2010; Martin et al., 2010; Martin et al., 2011; Shah et al., 2011; Sipes et al., 2011) and identifying signaling pathways associated with developmental processes (Kleinstreuer et al., 2011). The previous studies have used the results from the individual in vitro assays in the predictive modeling and aggregated assays based on genes and functional similarities (Kleinstreuer et al., 2011; Martin et al., 2011; Sipes et al., 2011). Nonetheless, a comprehensive analysis of the in vitro-to-in vivo predictive capability of the ToxCast high-throughput screening effort has yet to be presented.

Quantitative structure activity relationships (QSAR) have been used to predict potential in vivo toxicological effects for...
PREDICTING IN VIVO EFFECTS WITH IN VITRO ASSAYS

over a decade (Benigni, 1991; Moudgal et al., 2003; Young et al., 2004). The underlying assumption of QSAR is that common chemical structural features lead to similar effects, but the complexity of toxicological responses has led to mixed success. Recent QSAR models have shown promise for predicting general developmental toxicity (Cassano et al., 2010). Based on leave-several-out cross-validation, the predictive accuracy of the models was between 0.72 and 0.77 (Cassano et al., 2010), but the domain of applicability of these models may be relatively restricted (Hewitt et al., 2010). For non-organ-specific chemical carcinogenicity, recent QSAR models had a predictive accuracy between 0.62 and 0.66 based on five-fold cross-validation and a predictive accuracy of 0.60 using an external validation set (Fjodorova et al., 2010). For reproductive toxicity, separate QSAR models have been generated for seven general reproductive toxicity categories including male and female reproductive toxicity, fetal dysmorphogenesis, functional toxicity, mortality, growth, and newborn behavioral toxicity (Matthews et al., 2007). The performance of the models based on 10-fold cross-validation ranged from 0.54 for fetal death to 0.87 for female reproductive toxicity (Matthews et al., 2007). However, the QSAR models for the reproductive endpoints generally had a low sensitivity and high specificity.

The EPA’s ToxCast phase I data set provides an opportunity to broadly evaluate the ability of a specific group of high-throughput in vitro screening assays to predict in vivo toxic responses. The goals of this study were threefold: (1) to compare the performance of the 600 in vitro assays for predicting in vivo toxicity with the performance of chemical structure descriptors, (2) to evaluate the effects of aggregating the in vitro assays based on related genes and biological processes, and (3) to assess the impact of prefiltering the in vitro assays using a univariate statistical test prior to cross-validation. To address these goals, a comprehensive statistical classification analysis was performed to predict responses across 60 in vivo endpoints using a suite of 84 different predictive modeling methods. This comprehensive statistical analysis allows a broad-based evaluation of the predictive performance and the ability to make general conclusions about the expected future performance of the approaches investigated.

MATERIALS AND METHODS

ToxCast chemicals. The ToxCast phase I chemical library consisted of 309 unique compounds (Houck et al., 2009; Sipes et al., 2011). The selection of these chemicals was primarily due to the availability of extensive in vivo toxicity data, but the chemicals were also filtered based on solubility in DMSO, molecular weight, and purity. A complete list of the ToxCast phase I chemicals and associated purity and quality control data is available on the ToxCast web site (http://www.epa.gov/ncct/toxcast/chemicals.html).

In vitro assay data. The in vitro assay data used in this study was derived from the high-throughput screening of the ToxCast phase I chemical library. The data were downloaded from the ToxCast web site (http://www.epa.gov/ncct/toxcast/data.html; version 20110110). In total, the data include the activity of 309 compounds in more than 600 in vitro assays. Nine separate technologies were used, including receptor-binding and enzyme activity assays, cell-based protein and RNA expression assays, real-time growth measured by electronic impedance, and fluorescent cellular imaging. Each chemical-assay combination was run in dose-response, and an AC50 (concentration at 50% of maximum activity) or LEC (lowest effective concentration) value was calculated depending on the range of the dose-response data. In the download files, the reported AC50 and LEC values were in the units of micromolar response. Compounds not having an effect in a specific assay or endpoint were assigned a value of 106 (Judson et al., 2010). A detailed description of the assays and associated data have been provided in earlier publications (Houck et al., 2009; Huang et al., 2011; Judson et al., 2010; Knight et al., 2009; Knudsen et al., 2011; Martin et al., 2010; Rotroff et al., 2010). The assay definitions can be found in the ToxCast_Phase_1_Assays_20110110.txt file downloaded from the ToxCast web site. The AC50 and LEC values were log10 transformed. Compounds not having an effect in a specific assay or endpoint were assigned a value of 6 based on a log10 transformation of the 106 value assigned for inactive compounds (Judson et al., 2010). An alternative data transformation used in a previous analysis (Martin et al., 2011) (i.e., log10([AC50/1000])) was also evaluated but did not have a significant effect on the performance of the statistical classification models (see Supplementary file 1).

The complete set of transformed in vitro assay data used in the statistical classification analysis is provided as supplementary material (Supplementary file 2).

Previous studies have aggregated the more than 600 in vitro assays into different groups based on genes and functional similarities prior to the predictive modeling (Kleinstreuer et al., 2011; Martin et al., 2011; Sipes et al., 2011). For the GO biological processes, the summary value was the average AC50/LEL value of all genes or proteins in that GO biological process that were affected by the chemical (i.e., not assigned a value of 106 as inactive) (Sipes et al., 2011). The complete set of summary data for the genes and GO biological processes is provided as supplemental material (Supplementary files 3 and 4).

Chemical structural descriptors. Chemical structural descriptors were calculated for 297 of the 309 ToxCast phase I chemicals using Dragon (v.5.5; Talete SRL, Milan, Italy). The conventional chemical descriptors employed in this analysis are applicable to single organic molecules, and a total of 1224 chemical structural descriptors were calculated for each chemical. The 12 chemicals without structural descriptors were boric acid (CAS No. 10043-35-3), c dys 4-codacety 9901 55 69 4-91, 90 fenin (CAS No. 76-87-9), ifusilazole (CAS No. 85590-19-9), mancozeb (CAS No. 8018-01-7), maneb (CAS No. 12427-38-2), metiram-zinc (CAS No. 9006-42-2), milbemectin (CAS No. 5159611-3), and probenecide-calcium (CAS No. 127277-53-6). All 12 compounds belong to chemical classes (inorganic, mixture, organometallic, salts, or large macrocycle) in which conventional chemical descriptors could not be computed. A complete list of the structural descriptors is provided as supplementary material (Supplementary file 5).

In vivo toxicity data. The in vivo toxicity testing data were derived from the ToxRef database, which contains the results of guideline animal studies on the ToxCast phase I chemicals (Knudsen et al., 2009; Martin et al., 2009a,b). The in vivo testing data were downloaded from the ToxCast web site (http://www.epa.gov/ncct/toxcast/data.html; version 20110110). The data include a total of 463 in vivo endpoints from chronic testing in mice and rats, developmental testing in rabbits and rats, and multigenerational studies in rats.
In the downloaded file, the *in vitro* effects of each chemical were reported either as low effect levels (LELs) in the units of mg/kg/day or as binary active and inactive designations with active compounds noted as “1.” Inactive chemicals not having an effect in a specific endpoint were assigned a value of 10^6, and chemicals not tested in a particular type of study (e.g., multigenerational study) were assigned a value of NA. For the endpoints that were reported as continuous LEL values, the continuous data were converted into binary values such that a chemical producing an effect at any dose was considered positive and assigned a value of 1 and chemicals that did not have an effect at any dose (those assigned a value of 10^6 in the original data file) were considered negative and assigned a value of 0. The endpoints were then sorted based on the number of positive chemicals. Endpoints with ≥ 35 positive chemicals among the 297 chemicals with both chemical structural descriptors and *in vitro* assay data were selected for classification analysis. In addition, despite the low number of positive chemicals for the chronic rat liver tumor endpoint (21 positive chemicals), this endpoint was also selected for classification analysis due to a previous publication evaluating the ability of the *in vitro* assays for predicting the endpoint (Judson et al., 2010). The selected endpoints and the corresponding names in ToxRefDB are provided as supplementary material (Supplementary file 6).

**Statistical classification analysis.** Statistical classification analysis was performed using a suite of 84 different models and feature selection algorithms in JMP Genomics software (version 5.0; SAS Inc., Cary, NC). The 84 models included 8 discriminant analysis, 8 distance scoring, 12 general linear, 8 k-nearest neighbor, 13 logistic regression, 15 partial least squares (PLS), 16 partition tree, and 4 radial basis machine models. The radial basis machine models are types of support vector machine models. Detailed descriptions and parameters for each model are provided in supplementary material (Supplementary file 7). A total of 60 *in vivo* endpoint variables were used as the class-dependent variables (i.e., what is being predicted).

In order to compare the performance of the *in vitro* assays with that of chemical structure descriptors, evaluate the effects of aggregating the *in vitro* assays based on genes and biological processes, and assess the impact of prefiltering the *in vitro* assays prior to the predictive modeling, the 84 statistical classification models were run for each of the 60 *in vivo* endpoints in seven different scenarios:

1. *In vitro* assay data for the 297 chemicals
2. Chemical structural descriptors for the 297 chemicals
3. Combined *in vitro* assay data plus the chemical structural descriptors for the 297 chemicals
4. *In vitro* assay data for the full set of 309 chemicals without prefiltering
5. *In vitro* assay data for the full set of 309 chemicals without prefiltering aggregated based on gene
6. *In vitro* assay data for the full set of 309 chemicals without prefiltering aggregated based on GO biological process
7. *In vitro* assay data for the full set of 309 chemicals with prefiltering

The performance of the statistical classification models was assessed using 10 iterations of fivefold cross-validation for each of the seven scenarios. Fivefold cross-validation means that both the *in vivo* responses for a particular endpoint and the associated *in vitro* assays or chemical structure descriptors were randomly divided into five equal groups (Fig. 1). The endpoint frequencies in the groups (i.e., the percentage of positive and negative chemicals) were balanced (i.e., balanced stratification). One of the groups was then set aside as the test set. The remaining four groups were then used as the training set. With the training set, the most discriminating features were identified using different feature selection approaches, and a machine learning model was constructed. The model was then used to predict the chemicals in the test set, and the ability of the model to predict each chemical was recorded. This process was repeated until each of the five groups was left out. The fivefold cross-validation process was repeated 10 times (i.e., 10 iterations) to provide a more accurate measure of the predictive performance of the model for the specific endpoint.

The term “prefiltering” as performed on the *in vitro* assay data in the last scenario was defined as employing a univariate t-test prior to and outside the cross-validation loop. Only the *in vitro* assays that were statistically significant (p < 0.05) were used in the predictive modeling.

The performance metrics of the statistical classification models included balanced accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Balanced accuracy is the simple average of sensitivity and specificity. This performance metric is preferable over the normal calculation of accuracy when the number of positive and negative chemicals is unbalanced as is the case in this study. For example, if only 30 chemicals were positive out of 300 chemicals and the predictive model predicted all 300 chemicals to be negative, the normal calculation of accuracy would be 90%. However, the sensitivity would be 0% and specificity would be 100%, which would result in a balanced accuracy of 50%. The calculation of sensitivity, specificity, PPV, and NPV based on a standard 2 x 2 table is shown in Figure 2A for ease of reference. The AUC of the ROC curve is a single summary value for the performance of a statistical classification model that combines both sensitivity and specificity and effectively measures sorting efficiency (Fawcett, 2006). The AUC of the ROC ranges from 0 to 1, with a value of > 0.5 representing a predictive model that is better than chance (Fig. 2B).

A statistical comparison of the AUC values among the seven different analysis runs was performed using a Wilcoxon signed rank test, which is the nonparametric equivalent to the paired t-test (JMP version 9.0, SAS). The statistical test was performed using paired values across all the 60 *in vivo* endpoints. The paired values were the median AUC of all 84 statistical classification models under each analysis condition (e.g., *in vitro* assay data vs. chemical structure descriptors). A p value of < 0.05 was considered significant.

**Hierarchical clustering of chemical and endpoint predictability.** The average root mean square error (RMSE) across all of the 84 statistical classification models was calculated for each chemical and *in vivo* endpoint pair. The RMSE is a quadratic scoring rule that measures the average magnitude of the error of the prediction model. The RMSE values were obtained from the statistical classification analysis using only the *in vitro* assays. The log of the reciprocal of the mean RMSE (denoted log [1/RMSE]) was used as a surrogate for how well the chemical and *in vivo* endpoint pair was predicted. In addition, the sign of this statistic was changed to reflect either a positive true value (i.e., a value of 1 for the chemical-endpoint pair) or a negative true value (i.e., a 0 for the chemical-endpoint pair). The values were two dimensionally clustered using Ward’s minimum variance method with centered rows (JMP Genomics software, version 5.0, SAS). The results were displayed as a heat map with dendrograms for both rows and columns. A dendrogram is a branching diagram representing a hierarchy of categories based on degree of similarity. To further explore the predictability, the matrix of log (1/RMSE) values was subjected to a PLS analysis, using them as y-variables and the matrix of chemical structure descriptors as x-variables.

**Odds ratio analysis.** For the *in vivo* effects, the responses were converted into binary values as described above. For the *in vitro* assays, the AC₅₀ and LEL values were also converted into binary values such that a chemical producing an effect at any dose was considered positive and assigned a value of 1, and chemicals that did not have an effect at any dose (those assigned a value of 10^6 in the original data file) were considered negative and assigned a value of 0. For each *in vivo* endpoint, the positive and negative *in vivo* responses and positive and negative *in vitro* assay responses were compiled into a 2 x 2 contingency table. A standard odds ratio was calculated based on the 2 x 2 contingency table, and a two-tailed p value was calculated using Fisher’s exact test. A p value of < 0.01 was considered significant.

**RESULTS**

**Prevalence of Chemicals Producing a Positive In Vivo Response**

The ToxCast phase I chemicals are primarily pesticides and high-production volume chemicals with a significant amount of *in vivo* animal testing data developed through guideline toxicity
FIG. 1. Flow chart depicting the statistical classification analysis performed in the study. Two different data types were used for input into the statistical classification models—chemical structure and in vitro high-throughput screening assays. The data from the in vitro high-throughput assays were processed either with no aggregation, aggregation based on genes, or aggregation based on GO biological process categories. A total of 84 different statistical classification models were run with each input data type and used to predict 60 different in vivo toxic responses from the ToxRef database. Model performance was assessed using five-fold cross-validation with random partition and 10 iterations. The range and central tendency of the in vivo predictive performance of each input data type across all the 84 statistical models are depicted as a box and whiskers plot.

A.

<table>
<thead>
<tr>
<th>Prediction Based on In Vivo Assays of Chemical Structure</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Sensitivity = TP / (TP + FN)  
Specificity = TN / (FP + TN)

PPV = TP / (TP + FP)  
NPV = TN / (FN + TN)

FIG. 2. Overview of metrics used to evaluate the performance of the statistical classification models for predicting the in vivo responses based on the in vitro high-throughput screening assays and the chemical structure descriptors. (A) Calculation of sensitivity, specificity, PPV, and NPV based on a 2×2 table. (B) Calculation of the AUC of the ROC curve. An AUC of the ROC ranges from 0 to 1 with a value of >0.5, representing a predictive model that is better than chance (Fawcett, 2006).
studies. The results of the in vivo animal studies have been collated and summarized in the ToxRef database and published in a series of articles on the chronic in vivo testing in mice and rats (Martin et al., 2009a), developmental testing in rabbits and rats (Knudsen et al., 2009), and multigenerational studies in rats (Martin et al., 2009b). To develop a statistical classification model, an adequate number of positive chemicals are necessary and, optimally, there should be similar numbers of positive and negative chemicals. Of the 463 endpoints in the ToxRef database, only 59 endpoints had ≥35 chemicals producing a positive response in the 297 chemicals with both chemical structural descriptors and in vitro assay data. Only 6 of the in vivo endpoints had a balance between positive and negative chemicals, defined as between 40 and 60% (Fig. 3). Despite the imbalance between positive and negative chemicals for the majority of endpoints, all 59 were retained for further statistical classification analysis. In addition, the chronic rat liver tumor endpoint (21 positive chemicals) was also selected for statistical classification analysis due to a previous publication evaluating the ability of the ToxCast in vitro assays for predicting this endpoint (Judson et al., 2010). Therefore, a total of 60 endpoints were selected for the statistical classification analysis.

Comparison of the Predictive Performance of the In Vitro ToxCast Assays and Chemical Structure Descriptors

Of the 309 ToxCast phase I chemicals, structural descriptors could be derived for 297 chemicals. In order to compare the predictive performance of the in vitro assays with the performance of chemical structure descriptors, the performance of the statistical classification models was compared on the set of 297 chemicals. The AUC of the ROC curve was used as a common metric for predictive performance because it combines both sensitivity and specificity (Fawcett, 2006). For the in vitro assays (Scenario no. 1 described in Materials and Methods section), the majority of the 60 endpoints had ranges of AUC values across the 84 models that were greater than 0.5, suggesting that the majority of the endpoints were predicted better than chance (Fig. 4). The highest median AUC value (0.706) was for predicting rat cholinesterase inhibition (C_Rat_Cholinester_Inhibit). The balanced accuracy showed that 56 of the 60 endpoints were predicted at a level less than 0.55 (Fig. 5). The relatively low balanced accuracy was due to the poor sensitivity for predicting most endpoints (Fig. 6) with an associated high specificity (Fig. 7). For the few endpoints that

FIG. 3. Bar graphs of the number of chemicals showing a positive response in each of the 60 in vivo endpoints selected for statistical classification analysis. Black bars are the number of positive chemicals among the 297 ToxCast phase I chemicals that have both in vitro assay data and Dragon chemical structure descriptors. Gray bars are the number of positive chemicals among the complete set of 309 ToxCast phase I chemicals. The horizontal dashed lines bracket the 40 to 60% range in the number of positive chemicals.
showed reasonable sensitivity, a poor specificity was observed. The PPV and NPV showed similar trends (Figs. 8 and 9). The sensitivity of the various endpoints showed a high degree of positive correlation with the prevalence of positive chemicals, whereas the specificity was inversely correlated. A complete list of model performance metrics for each in vivo endpoint is provided as supplementary material (Supplementary file 8).

Among the six endpoints that had a nearly balanced number of positive and negative chemicals, more moderate sensitivities and specificities were observed. For the gross developmental endpoints in the rat (D_Rat_Dev), the sensitivity and specificity were 0.732 and 0.267, respectively. For maternal pregnancy–related endpoints in the rabbit (D_Rab_Mat_PregRel_PregLoss, D_Rab_Mat_PregRel, and D_Rab_Prenatal_Loss), similar sensitivities (0.717–0.725) and specificities (0.363) were observed. For the general category of mouse liver lesions (C_Mus_Liver_AnyLes), sensitivity was 0.543 and specificity was 0.480. Finally, for rat liver lesions (C_Rat_Liver_AnyLes), the sensitivity and specificity were 0.520 and 0.582, respectively.

To evaluate whether consensus predictions of the statistical classification models improved the ability to predict in vivo responses, the predicted probabilities across all 84 statistical classification models were averaged to get a consensus probability.
for each chemical and in vivo endpoint. The AUC of the ROC curve for each endpoint was then recomputed to evaluate whether the consensus predictions improved the predictive performance.

The consensus predictions of the 84 statistical classification models did not significantly improve the predictive performance of the in vitro assays (Supplementary file 9).

For the chemical structure descriptors (Scenario no. 2 described in Materials and Methods section), the majority of endpoints also had ranges of AUC values across the 84 models that were greater than 0.5 with the highest median AUC value (0.899) for rat cholinesterase inhibition. The balanced accuracy showed that 52 of the 60 endpoints were predicted at a level less than 0.55 (Fig. 5). Similar to the trends observed for the in vitro assays, the relatively low balanced accuracy was due to the poor sensitivity for predicting most endpoints (Fig. 6) with an associated high specificity. For the few endpoints that showed reasonable sensitivity, a poor specificity was observed.

A complete list of model performance metrics for each in vivo endpoint is provided as supplementary material (Supplementary file 8).

Among the six endpoints that had a nearly balanced number of positive and negative chemicals, more moderate sensitivities and specificities were observed. For the gross developmental endpoints in the rat (D_Rat_Dev), the sensitivity and specificity...
were 0.701 and 0.422, respectively. For maternal pregnancy-related endpoints in the rabbit (D_Rab_Mat_PregRel, D_Rab_Mat_PregRel, and D_Rab_Prenatal_Loss), a sensitivity of 0.703 and a specificity of 0.323 were observed. For the general category of mouse liver lesions (C_Mus_Liver_AnyLes), sensitivity was 0.602 and specificity was 0.530. Finally, for rat liver lesions (C_Rat_Liver_AnyLes), the sensitivity and specificity were 0.666 and 0.562, respectively.

A statistical comparison of the AUC values from the in vitro assays and chemical structure descriptors across all 60 in vivo endpoints showed that the predictive performance of the in vitro assays were not significantly different than the chemical structure descriptors ($p = 0.0709$). The median ratio of the in vitro assay AUC values relative to the chemical structure AUC values was 0.999 with upper and lower quartiles of 1.026 and 0.936, respectively (Fig. 10A).

To assess whether the combination of in vitro assays and chemical structure descriptors increased the predictive performance over either one alone, the statistical classification models were rerun with the combined set of descriptors (Scenario no. 3 described in Materials and Methods section). The AUC values of the combined set of descriptors did not
show significant improvement in performance over either the in vitro assays alone ($p = 0.0746$) or the chemical structure descriptors alone ($p = 0.2829$) (Figs. 10B and C).

To assess potential performance benefits for specific endpoints, the median AUC values for the in vitro assays and chemical structure descriptors were visualized as a scatter plot (Fig. 11). The majority of endpoints showed little difference in the median AUC values for the in vitro assays compared with the chemical structure descriptors. In setting an arbitrary cut-off for the difference in median AUC values of 0.1, no in vivo endpoints showed increased performance from the in vitro assays compared with the chemical structure descriptors. In contrast, six in vivo endpoints had a median AUC values for the chemical structure descriptors that were more than 0.1 higher than those for the in vitro assays. Three of the seven endpoints were related to maternal toxicity in developmental studies (D_Rat_Mat, D_Rat_Mat_Gen_Mat, and D_Rat_Mat_GenMat_Systemic), one endpoint related to rat kidney toxicity in the multigenerational studies (M_Rat_Kidney), one endpoint related to rat liver hypertrophy in chronic studies (C_Rat_Liver_Hypertrophy), and one endpoint related to rat cholinesterase inhibition in chronic studies (C_Rat_Cholinester_Inhibit).
Comparison of the Predictive Performance of the Aggregated and Unaggregated In Vitro ToxCast Assays

In previous studies, the more than 600 *in vitro* assays were aggregated based on genes and GO biological processes prior to the predictive modeling (Kleinstreuer et al., 2011; Sipes et al., 2011). To assess whether aggregating the *in vitro* assays has any effect on predictive performance, the *in vitro* assays were aggregated based on gene and GO biological processes prior to the statistical classification modeling. The aggregated descriptors were used as input for the 84 statistical classification models in each of the 60 *in vivo* endpoints (Scenario nos. 5 and 6 described in Materials and Methods section). The full set of 309 ToxCast phase 1 chemicals was used to compare the effects of *in vitro* assay aggregation. When the *in vitro* assays were aggregated by gene, the aggregated data resulted in significantly lower median AUC values ($p = 0.0006$). The median AUC ratio for the non-aggregated *in vitro* assays relative to the assays aggregated based on genes was 1.017 with upper and lower quartiles of 1.037 and 0.993, respectively (Fig. 12A). When the *in vitro* assays were aggregated by GO biological processes, the aggregated data did not result in a significant change in the median AUC values ($p = 0.3936$). The median
Comparison of the Predictive Performance of the Prefiltered and Non-prefiltered In Vitro ToxCast Assays

In previous studies, more than 600 in vitro assays were prefiltered using univariate t-tests, correlation coefficients, and chi-squared tests based on the in vivo endpoint being examined (Martin et al., 2011; Sipes et al., 2011). The prefiltering in these studies was done outside the cross-validation loop, which usually inflates model performance. To assess whether prefiltering the in vitro assays using univariate statistical tests outside the cross-validation loop has any effect on predictive performance, the in vitro assays were prefiltered using a t-test for each of the 60 in vivo endpoints. The assays showing a significant association with an in vivo endpoint \( (p < 0.05) \) were then used in each of the 84 statistical classification models. A statistical comparison of the AUC values from the
non-prefiltered and prefiltered in vitro assays across all 60 in vivo endpoints showed that prefiltering the assays highly inflated the AUC values ($p < 0.0001$). The median AUC ratio for the non-prefiltered in vitro assays relative to the prefiltered assays was 0.937 with upper and lower quartiles of 0.962 and 0.914, respectively (Fig. 12C). A complete list of model performance metrics for each in vivo endpoint is provided as supplementary material (Supplementary file 8).

**Hierarchical Clustering of Chemical and Endpoint Predictability**

To evaluate whether model performance could be used to identify chemical-endpoint pairs that are consistently correctly or incorrectly predicted and inform chemical prioritization, signed log2 (1/RMSE) values were hierarchically clustered. Hierarchical clustering of the y-axis grouped chemicals based on similar performance patterns, whereas clustering of the x-axis grouped the in vivo endpoints (Fig. 13A). If a chemical is consistently predicted incorrectly across the 84 statistical classification models for a specific endpoint, the color of the block in the heat map will be gray. Conversely, if a chemical is consistently predicted correctly across the 84 statistical classification models for a specific endpoint, the color will be either a deeper shade of blue or red depending on whether it was a positive correct prediction or a negative correct prediction.

The in vivo endpoints on the far right side of the x-axis had chemicals that were generally positive (red cells) with a low RMSE indicating a high degree of confidence in the prediction. Conversely, several in vivo endpoints such as rat liver tumors (C_Rat_Liver_Tumors) and rat kidney nephropathy (C_Rat_Kidney_Nephropathy) had large numbers of chemicals that were negative (blue cells) with a low RMSE also indicating a high degree of confidence in the prediction. These endpoints were those that had a high percentage of negative chemicals (Fig. 3). Because of the skewing of these
results by selection of almost all positive or almost all negative chemicals in the endpoints, they have questionable significance either for guiding prioritization or identifying chemicals that are consistently correctly or incorrectly predicted.

If one removes the endpoints at both extremes, chemical clusters were identified that were generally grouped together based on similar predictivity profiles. For example, one cluster contained predominantly the conazole fungicides (Fig. 13B). The conazoles generally showed a low RMSE (i.e., large values for 1/RMSE) and a positive response for mouse and rat liver toxicity (C_Mus_Liver_AnyLes and C_Rat_Liver_AnyLes). The conazoles are well known for their hepatotoxicity in both rat and mouse (Allen et al., 2006; Juberg et al., 2006; Peffer et al., 2007; Wolf et al., 2006). Two of the conazoles, hexaconazole and tetraconazole, had RMSE values near 1 for the prediction of hepatotoxicity in the mouse, suggesting these were consistently incorrectly predicted. A complete list of RMSE values for each chemical-endpoint pair is provided as supplementary material (Supplementary file 10).

To explore if the predictability profiles from the in vitro assays could be specifically related to any of the chemical structure descriptors, a PLS analysis was performed with the y-variables containing the signed log2(1/RMSE) values for the endpoints displayed in Figure 13A, the x-variables equal to the chemical structure descriptors, and each row corresponding to a unique chemical. The first PLS factor explained 1.4% of the y-variability and 26.1% of the x-variability. The weights assigned to each of the Dragon descriptors in the PLS component were inspected, and the highest weights were assigned to the Dragon descriptors JGI3-JGI10, all of which refer to the mean topological charge index of order 3 through 10. In general, these parameters describe the charge transfers between pairs of atoms and the global charge transfers in the molecule (Galvez et al., 1994). Exploring connections between the topological molecular charge and the biological activity of the studied compounds is beyond the scope of this study, but these results illustrate how predictability modeling can generate scientific hypotheses of potential promise and aid in...
prioritizing chemicals for further investigation according to a specific structural property like topological charge.

**Odds ratio analysis.** In a previous study, the in vitro assays were proposed to measure molecular initiating events and represent risk factors for in vivo toxicity (Judson et al., 2010). Odds ratios were calculated for all 600 in vitro assays relative to each of the 60 in vivo endpoints (Fig. 14). Both the in vitro assays and in vivo responses were converted into binary values (1 = positive at any dose or concentration; 0 = no response), and for each in vivo endpoint, the positive and negative in vivo responses and positive and negative in vitro assay responses were compiled into a 2×2 contingency table. A standard odds ratio was then calculated with a two-tailed p-value using Fisher’s exact test. A total of 262 in vitro assay:in vivo endpoint combinations showed significant odds ratios (p < 0.01) (Fig. 14). Of the significant odds ratios, 197 showed increased risk (positive odds ratios) from a positive assay response and 65 showed decreased risk (negative odds ratio) from a positive assay response. A complete list of the significant odds ratios is provided in Supplementary file 11.

**DISCUSSION**

The value of high-throughput in vitro assays for identifying chemical hazards and prioritizing chemicals for additional testing is currently being evaluated in both the United States and Europe. A significant amount of high-throughput screening data has already been generated allowing a broad-based assessment of the performance of these assays for predicting in vivo responses. In this study, a suite of 84 different statistical classification methods were used to evaluate the ability of the ToxCast phase I in vitro screening assays to predict 60 in vivo endpoints. The predictive performance of the in vitro assays...
FIG. 13. Hierarchical clustering of the predictability of each chemical and \textit{in vivo} endpoint combination. (A) Heat map of the \( \log_2 \) of the reciprocal of the RMSE across all of the 84 statistical classification models for each chemical and \textit{in vivo} endpoint pair. The sign of the statistic was changed to reflect either
FIG. 14. Analysis of the ToxCast in vitro assays as risk factors for in vivo toxic responses. Distribution of the significant log transformed odds ratios (p < 0.01) for all 600 in vitro assays relative to each of the 60 in vivo endpoints.

was benchmarked against chemical structure descriptors. The effects of in vitro assay aggregation based on genes and biological function and prefiltering were also evaluated because these have been previously applied in other predictive modeling studies using the ToxCast data.

Overall, the current in vitro ToxCast assays and chemicals were poorly predictive of the 60 in vivo responses using standard statistical classification methods with the possible exception of predicting cholinesterase inhibition in chronic rat studies. The overall median balanced accuracy for the in vitro assays across the 60 in vivo endpoints was 0.504 with a median sensitivity and specificity of 0.130 and 0.921, respectively. For most of the endpoints in this study, the imbalance in the number of positive and negative chemicals resulted in either a low sensitivity and high specificity or a high sensitivity and low specificity depending on the nature of the imbalance. Among the endpoints with a nearly equal number of positive and negative chemicals, the median balanced accuracy was 0.540 with a median sensitivity and specificity of 0.717 and 0.363.

Other analyses of the ToxCast phase I data for rat reproductive toxicity and rat developmental toxicity showed balanced accuracies of 0.74 and 0.71, respectively (Martin et al., 2011; Sipes et al., 2011). In our study, rat reproductive toxicity (M_Rat_Reproductive_Outcome) showed a median balanced accuracy, sensitivity, and specificity of 0.525, 0.136, and 0.892 across all the 84 models, respectively. For rat developmental toxicity (D_Rat_Dev), a median balanced accuracy, sensitivity, and specificity of 0.504, 0.719, and 0.265 were obtained across all the 84 models.

A direct comparison of our results with these studies is difficult because significant assay aggregation and prefiltering were performed outside the cross-validation loop in the previous studies, and in vivo endpoints were aggregated differently. An alternative point of comparison for predicting developmental toxicity is the frog embryo teratogenesis assay (FETAX). Using single decision criteria, the accuracy of the FETAX assay ranges between 0.54 and 0.63 with a range of sensitivity between 0.40 and 0.78 and a range in specificity between 0.45 and 0.71 (NICEATM, 2000). Using the FETAX as a basis for comparison, the ToxCast phase I assays showed a similar accuracy and sensitivity but a lower specificity for predicting developmental toxicity.

Although high-throughput in vitro toxicity screening is meant to be a cost-effective alternative to traditional in vivo toxicity studies, the in vitro assays must provide a significant predictive performance benefit or mechanistic insights that cannot be obtained using in silico methods such as QSAR. In this study, the predictive performance of the in vitro assays was compared with that of chemical structure descriptors to assess whether significant performance benefits could be achieved. When analyzed as a whole, the in vitro assays did not provide a significant difference in predictive performance compared with the chemical structure descriptors. Previous studies have also benchmarked biological descriptors against chemical structure in other contexts (Liu et al., 2011).

The predictive performance of the in vitro assays was lower than that reported for previously published QSAR models for carcinogenicity and developmental toxicity but similar to that reported for reproductive toxicity. QSAR models predicted rodent carcinogenicity with balanced accuracy of 0.615 with a sensitivity and specificity of 0.640 and 0.589 (Fjodorova et al., 2010). In our study, the prediction of rat carcinogenicity (C_Rat_Tumorigen) showed a median balanced accuracy, sensitivity, and specificity 0.487, 0.161, and 0.801 across all the 84 models, respectively. For developmental toxicity, QSAR models predicted with a balanced accuracy of 0.86 and a sensitivity and specificity of 0.90 and 0.82, respectively (Cassano et al., 2010). In our study of the ToxCast phase I assays, a median balanced accuracy, sensitivity, and specificity of 0.504, 0.719, and 0.265 were obtained for rat developmental toxicity (D_Rat_Dev) across all the 84 models, respectively. For reproductive toxicity, the QSAR model for male rat reproductive toxicity had a balanced accuracy of 0.59 with a sensitivity of 0.280 and a specificity of 0.90 (Matthews et al., 2007). By comparison, the ToxCast phase I assays showed a median balanced accuracy, sensitivity, and specificity of 0.525,
Thus, selective aggregation of the assays may provide performance (see Supplementary file 12). However, for some descriptors used in their study did not show improved predictive models across all 60 assays analyzed in this study. In previous studies, mixed success was achieved by combining chemical structure descriptors with experimental measurements of biological responses. For example, chemical structure descriptors combined with in vitro cytotoxicity measurements were more successful for predicting acute toxicity (Sedykh et al., 2011) and carcinogenicity (Zhu et al., 2008) than either single data type. In carcinogenicity, the QSAR model predicted with 0.623 accuracy, which increased to 0.727 when the chemical structure descriptors were combined with the results from a suite of in vitro cytotoxicity assays (Zhu et al., 2008). For other toxicological endpoints such as drug-induced liver injury, the combination of transcriptomic data and chemical descriptors was less predictive than transcriptomic data alone (Low et al., 2011). In our study, the combination of in vitro assays and chemical structure descriptors did not increase the predictive performance above that observed with either the in vitro assays or the chemical structure descriptors alone.

The differences in the predictive performance observed in our study compared with that reported in previous analyses of the ToxCast phase I data for rat reproductive toxicity and rat developmental toxicity may be due to the assay aggregation and prefiltering that were performed outside the cross-validation loop in these studies (Martin et al., 2011; Sipes et al., 2011). To understand the potential influence of these data processing steps on model performance, each step was evaluated separately. The more than 600 in vitro assays were aggregated based on genes and GO biological process categories. These aggregated descriptors were used as input into the 84 statistical classification models for predicting the 60 in vivo endpoints. When compared with the unaggregated in vitro assays, the gene aggregation reduced predictive performance, whereas aggregation by GO biological process had no significant effect.

The differences in the predictive performance observed in our study compared with that reported in previous analyses of the ToxCast phase I data for rat reproductive toxicity and rat developmental toxicity may be due to the assay aggregation and prefiltering that were performed outside the cross-validation loop in these studies (Martin et al., 2011; Sipes et al., 2011). To understand the potential influence of these data processing steps on model performance, each step was evaluated separately. The more than 600 in vitro assays were aggregated based on genes and GO biological process categories. These aggregated descriptors were used as input into the 84 statistical classification models for predicting the 60 in vivo endpoints. When compared with the unaggregated in vitro assays, the gene aggregation reduced predictive performance, whereas aggregation by GO biological process had no significant effect.

In contrast to our results on aggregation, an earlier study suggested that combining assays yielded a more balanced and stable model (Martin et al., 2011). However, it is difficult to make exact comparisons because the previous analysis used an aggregation strategy that was significantly different from that tested in our study. In the earlier study, 36 assays were combined into eight aggregated descriptors based on genes, protein function (e.g., G-protein coupled receptors and cytochrome P-450s), and a lumped category called “other” (Martin et al., 2011). When analyzed using the 84 statistical classification models across all 60 in vivo endpoints, the eight aggregated descriptors used in their study did not show improved predictive performance (see Supplementary file 12). However, for some endpoints within reproductive toxicity, the aggregated assays identified in the earlier study increased predictive performance. Thus, selective aggregation of the assays may provide performance benefits for specific endpoints, but this approach will not be universally successful.

In contrast to assay aggregation, prefiltering the in vitro assays using a univariate t-test outside the cross-validation loop significantly improved predictive performance. This result was expected because prefiltering the in vitro assay data biases the assays available to the classification model toward those assays that discriminate between positive and negative chemicals among the full set of chemicals. Although univariate analyses may be useful for identifying in vitro assays for external validation studies with independent sets of chemicals, the use of prefiltering outside the cross-validation loop is inadvisable when evaluating the predictive performance of a classification model. External validation has been performed on a previous ToxCast study that used prefiltering to select in vitro assays to predict reproductive toxicity (Martin et al., 2011).

There are several possible reasons for the poor predictivity of in vivo toxicity using the ToxCast phase I in vitro assays. First, the current in vitro assays may not capture the biochemical and cellular responses in the in vivo tissues with adequate fidelity. This lack of fidelity may include cellular properties such as metabolic competence and transporter expression or more complex cell-cell interactions such as the interactions between immune cells and epithelial cells following cytotoxicity. Second, the current in vitro assays may not provide the proper biological context necessary to predict toxicological responses. For example, activation of the NFκB pathway can be pro- or antiapoptotic depending on the context in which it is activated (Lin et al., 1999). The current suite of assays may not reliably capture context-specific outcomes. Third, the current set of assays may not provide sufficient coverage of pathways, protein targets, and cell types. Fourth, there were significant cross-species differences between the in vitro assays and in vivo endpoints being predicted. The current suite of in vitro assays used predominantly human-derived cells, cell lines, and proteins, whereas the in vivo results were based on studies in rats, mice, and rabbits. Finally, the number of positive chemicals present for each endpoint may not be sufficient to redundantly capture the broad array of mechanisms leading to in vivo toxicity. In the cross-validation process, 20% of the chemicals were randomly left out during model training. If multiple chemicals were not present for each mechanism, then the relevant assays may not be selected prior to prediction of the test set. Realistically, it is likely that some combination of these reasons, and potentially additional ones, led to the lack of predictive performance in the ToxCast phase I data set. Additional studies that explore the relative contributions of these various factors to in vivo predictivity will be necessary for developing confidence in the use of high-throughput in vitro assays for prioritization and hazard prediction.

The relatively poor predictive performance of the high-throughput in vitro assays in this study does not mean that they are not useful in prioritization and toxicity testing nor does it necessarily cast negative light on other in vitro testing efforts. In many respects, the ToxCast in vitro assays analyzed in this study
PREDICTING IN VIVO EFFECTS WITH IN VITRO ASSAYS

In vivo Chronic study, mouse

Chronic study, rat

initiation of hepatocellular proliferation and liver regeneration (Jang et al., 2007b), serotonin plays a role in nonalcoholic liver disease (Nocito et al., 2007a, 2007b).

platelets (Lesurtel, 2006). Previous studies have shown that serotonin is produced in the intestinal enterochromaffin cells with approximately 95% of circulating serotonin stored in platelets (Lesurtel et al., 2006). Previous studies have shown that serotonin plays a role in nonalcoholic liver disease (Nocito et al., 2007a), repair after ischemic liver injury (Nocito et al., 2007b), and cholestatic liver injury (Jang et al., 2012).

in vitro assays with large odds ratios are provided in Table 1. Assays may be viewed as negative risk factors. To demonstrate how these could be used, odds ratios were calculated for all 600 in vitro assays relative to each of the 60 in vivo endpoints and filtered to include only statistically significant odds ratios ($p < 0.01$) (Fig. 14). Examples of two in vivo endpoints that have in vitro assays with large odds ratios are provided in Table 1.

For cholinesterase inhibition in chronic rat studies, the greatest risk factors are associated with binding to acetyl- and butyrylcholinesterase in biochemical assays. As molecular initiating events, binding to these enzymes is biologically consistent with the mechanism. Curiously, binding to the bovine progesterone receptor also had a significant odds ratio for cholinesterase inhibition in chronic rat studies, which may be due to off-target interactions with this class of pesticides.

For mouse liver tumors in chronic studies, the greatest risk factors were the activation of the human peroxisome proliferator activated receptor alpha (PPARA) in a cell-based reporter assay; binding to the guinea pig opioid receptor, kappa 1 (OPRK1), in a biochemical assay; and binding to the human serotonin transporter (SLC6A4) in a biochemical assay. Again, as a molecular initiating event, the activation of PPARA makes biological sense given the association of the receptor with rodent liver tumors (Peters et al., 1997). The other two risk factors are less obvious but are also consistent with the known biological roles of the genes identified.

Serotonin is a neurotransmitter. In the periphery, serotonin is produced in the intestinal enterochromaffin cells with approximately 95% of circulating serotonin stored in platelets (Lesurtel et al., 2006). Previous studies have shown that serotonin plays a role in nonalcoholic liver disease (Nocito et al., 2007a), repair after ischemic liver injury (Nocito et al., 2007b), and cholestatic liver injury (Jang et al., 2012). In addition, the initiation of hepatocellular proliferation and liver regeneration involves the release of platelet-derived serotonin (Lesurtel et al., 2006). The κ-opioid receptor is a G protein–coupled receptor (Feng et al., 2012). Apart from pain modulation and addiction, opioid receptors are involved in various physiological and pathophysiological activities (Feng et al., 2012). The κ-opioid receptor plays a key role in hepatic lipid metabolism (Czzyk et al., 2010) and vascular development (Yamamizu et al., 2011). Both processes are important in liver tumorigenesis.

In viewing the in vitro assays as risk factors for in vivo toxic responses, a combined net risk among the assays that are positive for a particular chemical could then be used in prioritizing chemicals for additional in vivo testing. This more nuanced view of the value of high-throughput in vitro screening transitions away from using high-throughput in vitro screening results for predicting in vivo hazard as a binary yes/no classification toward their application to understanding mode of action. A redirection of this kind will likely provide a better conceptual framework for applying these in vitro assays until future studies can identify the underlying reasons for their poor predictivity.

### TABLE 1

<table>
<thead>
<tr>
<th>In vivo endpoint</th>
<th>In vitro assay name</th>
<th>Description</th>
<th>Odds ratio</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic study, rat</td>
<td>NVS_ENZ_rAChE</td>
<td>Biochemical, rat acetylcholinesterase binding</td>
<td>87.0</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>NVS_ENZ_hAChE</td>
<td>Biochemical, human acetylcholinesterase binding</td>
<td>60.6</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>NVS_ENZ_bAChE</td>
<td>Biochemical, human Butyrylcholinesterase binding</td>
<td>12.8</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>NVS_NR_bPBR</td>
<td>Biochemical, bovine progesterone receptor binding</td>
<td>9.6</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>Chronic study, mouse</td>
<td>ATG_PPARa_TRANS</td>
<td>Cellular, human PPARA reporter</td>
<td>27.8</td>
<td>0.0021</td>
</tr>
<tr>
<td>NVS_GPCR_gOpiateK</td>
<td>Biochemical, guinea pig opioid receptor, kappa 1 binding</td>
<td>22.4</td>
<td>0.0074</td>
<td></td>
</tr>
<tr>
<td>NVS_TR_hSERT</td>
<td>Biochemical, human serotonin transporter binding</td>
<td>22.4</td>
<td>0.0074</td>
<td></td>
</tr>
</tbody>
</table>

The in vitro assays for each in vivo endpoint were filtered to remove those with odds ratios < 5 and p values > 0.01.

### SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci.oxfordjournals.org/.

### FUNDING

American Chemistry Council’s Long-Range Research Initiative.

### REFERENCES


