Nonanimal Models for Acute Toxicity Evaluations: Applying Data-Driven Profiling and Read-Across

Daniel P. Russo,1 Judy Strickland,2 Agnes L. Karmaus,2 Wenyi Wang,1 Sunil Shende,1,3 Thomas Hartung,4,5 Lauren M. Aleksunes,6 and Hao Zhu1,7

1Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, USA
2Integrated Laboratory Systems (ILS), Research Triangle Park, North Carolina, USA
3Department of Computer Science, Rutgers University, Camden, New Jersey, USA
4Johns Hopkins Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, Maryland, USA
5University of Konstanz, CAAT-Europe, Konstanz, Germany
6Department of Pharmacology and Toxicology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, New Jersey, USA
7Department of Chemistry, Rutgers University, Camden, New Jersey, USA

BACKGROUND: Low-cost, high-throughput in vitro bioassays have potential as alternatives to animal models for toxicity testing. However, incorporating in vitro bioassays into chemical toxicity evaluations such as read-across requires significant data curation and analysis based on knowledge of relevant toxicity mechanisms, lowering the enthusiasm of using the massive amount of unstructured public data.

OBJECTIVE: We aimed to develop a computational method to automatically extract useful bioassay data from a public repository (i.e., PubChem) and assess its ability to predict animal toxicity using a novel bioprofile-based read-across approach.

METHODS: A training database containing 7,385 compounds with diverse rat acute oral toxicity data was searched against PubChem to establish in vitro bioprofiles. Using a novel subspace clustering algorithm, bioassay groups that may inform on relevant toxicity mechanisms underlying acute oral toxicity were identified. These bioassays were used to predict animal acute oral toxicity using read-across through a cross-validation process. Finally, an external test set of over 600 new compounds was used to validate the resulting model predictivity.

RESULTS: Several bioassay clusters showed high predictivity for acute oral toxicity (positive prediction rates range from 62–100%) through cross-validation. After incorporating individual clusters into an ensemble model, chemical toxicants in the external test set were evaluated for putative acute toxicity (positive prediction rate equal to 76%). Additionally, chemical fragment–in vitro–in vivo relationships were identified to illustrate new animal toxicity mechanisms.

CONCLUSIONS: The in vitro bioassay data-driven profiling strategy developed in this study meets the urgent needs of computational toxicology in the current big data era and can be extended to develop predictive models for other complex toxicity end points. https://doi.org/10.1289/EHP3614

Introduction

There are currently over 100,000 chemicals available on the market that lack toxicity information, comprising roughly 90% of the 140,000 consumer products in use (Hartung and Rovida 2009; Judson et al. 2009). Traditional toxicology evaluations require the use of animal models for testing new compounds. However, these animal models are costly and time-consuming, and they raise ethical concerns regarding the well-being of animals (Hartung 2017). Under this paradigm, generating substantial toxicity data for a limited number of compounds could take years, and it would be financially impossible to test all the available compounds using animal testing protocols (Hartung 2016). In 2007, the National Research Council Committee on Toxicity Testing and Assessment of Environmental Agents addressed this issue by proposing a new framework to accurately and more quickly evaluate the health risks due to environmental chemical exposures (National Research Council 2007). This federal effort stressed the importance of integrating/establishing the use of computational and in vitro–based alternative methods for chemical risk evaluation. One such alternative, called read-across, relies on using toxicity information from structurally similar compounds to estimate the toxicity of untested compounds (Patlewick et al. 2014; Wang et al. 2012). This strategy can be used to fill toxicity data gaps for untested chemicals and has been implemented by various regulatory agencies (Hartung 2016). Previous read-across studies relied solely on chemical structure similarity searching (Enoch et al. 2008; Hewitt et al. 2010; Koleva et al. 2008; Luechtfeld et al. 2016; Wu et al. 2013). However, this type of read-across is not applicable for compounds with unique chemical structures and can be confounded by “activity cliffs” (i.e., structurally similar compounds with distinctly different toxicity characteristics) (Maggora 2006). More recently, efforts to include biological information as a basis for similarity in read-across approaches have started (Zhu et al. 2016). Previous studies using biological data for chemical toxicity evaluations were mostly based on in-house biological data and were limited to specific mechanisms (Judson et al. 2015; Kleinstreuer et al. 2017). This paper addresses the challenge of identifying and integrating biological data from various resources into read-across modeling.

In vitro high-throughput screening (HTS) is capable of rapidly testing large numbers of chemicals to study their effects on molecular targets using whole-cell and cell-free assays. Because of their relatively low cost and high-throughput, efforts such as the Toxicity Testing in the 21st Century (Tox21) program have focused on the application of HTS techniques as the basis for chemical hazard assessment (Attene-Ramos et al. 2013). The direct result of these efforts is a rapidly growing amount of in vitro bioassay data being generated for thousands of chemicals and stored in databases accessible to public users, allowing for new statistical and computational techniques to be developed. The impact of such large publicly available databases for chemical toxicity evaluation is profound, with several projects having successfully used HTS data to better evaluate chemicals for potential hazards (Brown et al. 2015; Hartung 2016; Kim et al. 2016; Kleinstreuer et al. 2017).
However, rapidly changing public data sources represent a dynamic data landscape, and integrating such data to chemical information for toxicity evaluation is an area that remains largely unexplored. Development of automated computational methods to deeply exploit this rich and dynamic data landscape to establish predictive nonanimal toxicity models is needed.

Acute oral toxicity testing is conducted to determine the immediate health effects of an orally administered chemical substance and is expressed in terms of the lethal dosage that kills 50% of the population (LD50) of animals tested (Strickland et al. 2018). Acute oral toxicity data are used by a number of regulatory agencies for hazard classification and labeling of products to alert handlers and consumers of potential toxicity hazards, to determine acceptable human exposure limits and personal protective equipment needed for handling, and determine countermeasures that should be employed in the event of toxic exposures (Corvaro et al. 2016; Strickland et al. 2018; Walum 1998). In some cases, acute oral toxicity data may be used to establish doses for longer-term studies, identify target organs for toxicity, and assess the hazard of accidental ingestions of chemical contaminants in food (Strickland et al. 2018). To date, there are no in vitro tests accepted by regulatory agencies as stand-alone replacements for acute oral animal tests (Kinsner-Ovaskainen et al. 2009; Strickland et al. 2018).

Here, we present a new computational technique to automatically extract pertinent data from PubChem, which is updated daily, and develop a predictive ensemble model for estimating acute oral toxicity (outlined in Figure 1). In this work, a large dataset consisting of compounds with broad acute oral toxicity values was profiled by their PubChem in vitro bioassay data to generate bioprofiles. Using these initial bioprofiles, we characterized and clustered mechanistically similar PubChem bioassays using fragments of the chemicals tested within them. The resulting chemical fragment—in vitro—in vivo relationships were the foundation for bioprofile-based read-across studies conducted using clusters of PubChem bioassays considered to inform on similar toxicity mechanisms.

**Methods**

**Acute Oral Toxicity Datasets**

An in-house rat acute oral toxicity database was previously collected and curated from ChemIDplus (https://chem.nlm.nih.gov/chemidplus; Zhu et al. 2009). This in-house rat acute oral toxicity database was used as the training set and consisted of 7,385 compounds with their most conservative LD50 values (i.e., lowest recorded LD50 value for a single compound) ranging from 0.02 log kg to 389.39 g/kg (Figure 2A). For modeling purposes, the LD50 values were normalized by the following logistic function:

\[
f(x) = \frac{e^{x_0/s}}{1 + e^{x_0/s}},
\]

where \(x\) represents a log10-transformed LD50 (logLD50) value, \(x_0\) represents the midpoint of the curve, and \(s\) is a number to control the shape of the curve. Figure 2B shows the relationship between the logistic function outputs and the logLD50 values through various values of \(s\). The outputs of the logistic function are all between 0 and 1, so we chose the threshold of 0.5 to distinguish toxic (\(f \geq 0.5\)) and nontoxic (\(f < 0.5\)) compounds. To obtain a

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Overview of the data-driven read-across approach developed in this study. (A) Profiling compounds via PubChem portal, (B) clustering PubChem assays based on chemical fragment in vitro relationships, (C) identifying bioassays capable of predicting acute oral toxicity, and (D) predicting new toxicants by read-across and mechanism illustrations.
balanced dataset and easily interpretable cutoff, we set the midpoint \((x_0)\) at 3 (which is equivalent to 1,000 mg/kg), yielding 3,791 toxic and 3,594 nontoxic compounds.

A separate dataset consisting of 3,852 compounds with rat acute oral LD\(_{50}\) values was collected from a variety of sources, including ChemIDplus (https://chem.nlm.nih.gov/chemidplus/), the Hazardous Substances Data Bank (https://toxnet.nlm.nih.gov/newtoxnet/hsdb.htm), European Chemical Agency (https://echa.europa.eu/information-on-chemicals), and the U.S. Environmental Protection Agency (EPA) (U.S. EPA 2016). This dataset served as an external test set to evaluate the generated models. We refined this dataset by excluding compounds already included in the training set, standardizing chemical structures, and removing compounds with an LD\(_{50}\) value reported as a range (e.g., LD\(_{50} > 500\) mg/kg). Thus, the curated external test set ultimately contained 639 compounds with LD\(_{50}\) values ranging from 0.012 mg/kg to 63.97 g/kg. The LD\(_{50}\) values were also converted to classifications using the logistic function (including the same threshold and midpoint) as described above.

**Bioprofiling and Subspace Clustering of PubChem Data**

To form a training set of compounds for modeling, public in vitro bioassay data for 7,385 compounds were extracted from PubChem (https://pubchem.ncbi.nlm.nih.gov/) using an automatic data mining portal (http://ciipro.rutgers.edu/) (Russo et al. 2017; Zhang et al. 2014a). To establish meaningful relationships between chemical fragment descriptors and PubChem bioassays, a feature reduction approach was applied. Briefly, those bioassays with very limited data across the chemicals in our training set were removed to avoid overfitting by training the model using minimal signal. Thus, only bioassays with at least five active responses among the training set compounds were included in the modeling procedure. This effort resulted in a sufficiently large dataset containing 3,543 training set compounds with in vitro data from 1,077 PubChem bioassays.

The bioactivity values of the 3,543 training set compounds across each of the 1,077 PubChem assays comprised the initial bio-profile. To identify potential toxicity mechanisms and further optimize the initial bio-profile, the 1,077 PubChem bioassays were clustered based on shared chemical fragments relevant to bioassay responses. To achieve this, we used the established ToxPrint fingerprints, a set of 729 chemical fragments relevant to toxicity reported in a previous study (Yang et al. 2015); fingerprints were generated using ChemoTyper (Molecular Networks GmbH, Erlangen, Germany) software (version 1.0). Then, the relationship between these ToxPrint fragments and each PubChem bioassay were determined using Fisher’s exact test. Fisher’s exact test requires constructing a \(2 \times 2\) contingency matrix, which we define below:

\[
\begin{array}{cc}
\text{Compounds with fragment} & \text{Compounds without fragment} \\
\text{Compounds active in assay} & a & b \\
\text{Compounds inactive in assay} & c & d \\
\end{array}
\]

where \(a\) is the number of compounds with an active response in this assay and contain this fragment, \(b\) is the number of compounds with an active response and do not contain the fragment, \(c\) is the number of compounds with an inactive response and contain the fragment, and \(d\) is the number of compounds with an inactive response and do not contain the fragment.

The output of this test is a \(p\)-value denoting the statistical significance of the relationship between the fragment and bioassay activity. When considering multivariate comparisons, it is necessary to ensure that the resulting correlations are not spurious. However, strict multiple testing corrections, such as the Bonferroni correction, greatly reduced the number of significant relationships (<90% and a loss of ~60% of the biological data) and were conservative for inclusion purposes. Alternatively, the number of false positives was approximated in this analysis by comparing \(p\)-values to those obtained from a permutation test (Figure S1). The standard \(p\)-value

---

**Figure 2.** Activity distribution for compounds in this study. (A) Histogram for the 7,385 log-transformed LD\(_{50}\) values from compounds in the original dataset (hashed bars in background) and the final 3,543 compounds with biological data ultimately used for modeling (solid bars in foreground). The parameter, \(x_0\), is equal to the midpoint of the logistic curve displayed in (B) and was used as the cutoff for determining compound toxicity. (B) The parameter \(s\) controls the shape of the curve. Lines represent the logistic function at various values for \(s\) (0.1 through 0.9 at steps of 0.1). The final curve, with \(s = 0.5\), is bold. The thin areas on the line indicate log-transformed LD\(_{50}\) values <2.5 or >3.5 and their corresponding values in \(f\) space. The thick line indicate log-transformed LD\(_{50}\) values \(>3\) and their corresponding values in \(f\) space. The thick line indicate log-transformed LD\(_{50}\) values \(\geq 5\).
The biosimilarity between two compounds in a cluster biosimilar compound in the modeling set, based on sharing the Jaccard dissimilarity ($J_d$) between each bioassay was calculated using the fragment profile, defined as:

$$J_d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

where $A$ and $B$ represent the sets of significant fragments for PubChem bioassays $A$ and $B$, respectively. Calculating dissimilarity ($J_d$) between bioassays allows for the representation of potential relationships among assays as a network graph, where nodes represent bioassays, and edges are the $J_d$ values between two bioassays. The network graph was created and manipulated using the software package Gephi (https://gephi.org/) (version 0.9.1). Clusters of PubChem assays were determined by using the Louvain modularity algorithm available within the Gephi software package, using the “resolution” parameter to determine the bioassays that belong to the same cluster (Blondel et al. 2008). A larger resolution value allows more bioassays to form larger clusters, and a smaller resolution value restricts bioassays with higher similarity to small clusters. A resolution value of 0.25 was used in this study, with the maximum number of bioassays in the resulting clusters set to 60.

Bioassay-Based Read-Across for Acute Oral Toxicity Classifications

We delineated clusters of bioassays that were highly predictive of acute toxicity to identify relevant bioassays and unveil potential toxicity mechanisms. Read-across studies were then performed using the bioassay responses within a cluster to predict acute oral toxicity classification; prediction results were evaluated by five-fold cross-validation. If bioassays showed highly correlated responses with at least 10 mutual test responses ($R^2 > 0.9$), one of them was randomly selected and incorporated into model building, and the other highly correlated assays were removed from any further analysis. Training set compounds with at least one in vitro bioassay response in a cluster were randomly split into five equivalent subsets. For each of five iterations, one subset acted as a pseudotest set, while the remaining compounds served as a pseudotraining set. The acute oral toxicity classification of each validation set compound was predicted by the most biosimilar compound in the modeling set, based on sharing the most similar PubChem bioassay responses within the cluster. The biosimilarity between two compounds in a cluster $c$ can be calculated by:

$$\text{Biosimilarity} (A, B)_c = \frac{|A_a \cap B_c| + |A_i \cap B_c| \cdot w}{|A_a \cap B_c| + |A_i \cap B_c| \cdot w + |A_a \cap B_i| + |A_i \cap B_i|},$$

where $A_a$ and $B_c$ represent the sets of active responses in PubChem bioassays within a cluster $c$ for compounds $A$ and $B$, respectively. Conversely, the terms $A_i$ and $B_i$ represent the sets of inactive responses. The term $w$ weights the inactive responses less than active responses since the proportion of active data, which indicates more significant biological phenomena, is much lower than inactive data. In this study, $w$ was calculated as the ratio of total active responses to total inactive responses for each cluster.

Public bioassay data is inherently sparse, and therefore, calculating biosimilarity alone can offer misleading results due to missing data. For example, when two compounds are both only tested in one assay within a cluster, their biosimilarity result is less reliable than two other compounds that share responses in multiple assays. For this reason, the relative cluster confidence ($rcc$) of a biosimilarity calculation was also computed using the equation below:

$$rcc(A, B)_c = \frac{|A_a \cap B_c| + |A_i \cap B_c| \cdot w + |A_a \cap B_i| + |A_i \cap B_i|}{N},$$

where $N$ is the number of noncorrelated assays in cluster $c$ (i.e., total number of bioassays used in the model). Here, a high $rcc$ is indicative of the two compounds being tested in many of the same assays within a cluster.

The underlying mechanisms responsible for toxicity in animal acute oral studies are vast, and thus, it is unlikely to expect a few bioassays to explain all the various toxic phenomena. Therefore, when using bioassays as models for acute oral toxicity, it is reasonable to expect a relatively high false negative rate (compounds inactive in bioassay response yet active in acute oral animal toxicity test). In these cases, toxicity may be elicited via mechanisms that are not represented by the biochemical coverage of the in vitro assays mined. Therefore, to focus on the ability to identify toxic compounds, the positive predictive value ($ppv$) was used to evaluate the model performance and is defined as:

$$ppv = \frac{TP}{TP + FP},$$

where $TP$ represents the number of true positives (toxic compounds correctly predicted as toxic), and $FP$ represents the number of false positives (nontoxic compounds incorrectly predicted as toxic).

Quantitative Structure–Activity Relationship Models

Missing data severely limits the identification of relationships in this study. Simple data imputation methods, such as random sampling, are not sufficiently robust and may create further issues. In order to correctly impute biological data, especially for the toxicants with missing data, more reasonable data imputation methods were necessary. To realize this, Random Forest and Naïve Bayes, both implemented using the Python library scikit-learn (version 0.18.1), were used to develop quantitative structure–activity relationship (QSAR) models (Pedregosa et al. 2011). These QSAR approaches were implemented for PubChem bioassays to fill the missing bioassay data, allowing for sufficient data points to evaluate chemical fragment—in vitro—in vivo relationships (i.e., when prediction performance is increased by the presence of a toxiphore). More specifically, of the original 7,385 chemicals identified from the in-house rat acute oral toxicity database, only 3,543 had sufficient bioassay data for initial analyses. To increase scope for later chemical fragment—in vitro—in vivo relationship analyses, the remaining 3,842 chemicals were subjected to the QSAR workflow.

Random Forest is an ensemble algorithm consisting of constructing many decision trees and then making a prediction by combining the output among the trees (Breiman 2001). Naïve Bayes is an algorithm that predicts by estimating the probability of membership to a certain class using Bayes theorem (Friedman et al. 1997). The QSAR model development herein followed the workflow used in our previous studies and briefly outlined below (Kim et al. 2014; Solimeo et al. 2012; Sprague et al. 2014; Wang et al. 2015; Zhao et al. 2017). The training data for QSAR model development was retrieved from PubChem’s PUG-REST web service by using a PubChem
Assay Identifier (AID) as the query (Kim et al. 2015). The information retrieved for a single bioassay included chemical structures and activity classifications (active/inactive/inconclusive) for all tested compounds in the bioassay. First, inconclusive results were eliminated. Then, the active/inactive ratio was balanced by randomly selecting and eliminating compounds until an equal ratio was achieved. Training set data was set to not exceed 10,000 compounds. The rdkit implementation of extended-connectivity fingerprints was used as chemical features for the remaining compounds in QSAR model training (Rogers and Hahn 2010).

A fivefold cross-validation procedure available within the scikit-learn package was used to evaluate the resulting models. This procedure searched and stored the models based on fivefold cross-validation prediction accuracy (acc) as defined as:

\[
acc = \frac{TP + TN}{TP + FP + TN + FN},
\]

where true positive (TP) is the number active compounds correctly predicted as active, false positive (FP) is the number of inactive compounds incorrectly predicted as active, false negative (FN) is the number of active compounds incorrectly predicted as inactive, and true negative (TN) is the number of inactive compounds correctly predicted as inactive. Only models with a cutoff of acc > 0.6 were used to fill data gaps for substances missing in vitro bioassay results. If both Random Forest and Naïve Bayes models had an acc > 0.6 for the same PubChem assay, the model with higher acc value was used.

**Mechanism-Driven Toxicity Pathway Analysis**

By integrating QSAR predictions into bioprofiles to address missing data (i.e., for chemicals that were not tested), chemical fragment—*in vitro*—*in vivo* relationship analyses can be performed by examining the chemical fragments of the compounds with predicted bioassay results. If compounds contain a specific chemical fragment that produces superior prediction accuracy within a cluster, this chemical fragment can be selected as a potential toxicophore, which is the principal chemical feature to induce a toxicity pathway (Allen et al. 2014). In this way, new toxicity mechanisms can be revealed by linking chemical fragments, relevant *in vitro* bioprofiles, and *in vivo* acute oral toxicity.

**Results**

**Acute Oral Toxicity Classifications**

A logistic function (f) was used in this study to convert experimental log-transformed LD50 (logLD50) values into classifications (toxic/nontoxic), applying a threshold of f ≥ 0.5 to define toxic compounds based on an established LD50 cutoff of 1,000 mg/kg. This threshold balances the toxic/nontoxic ratio (see “Methods” section) and is close to the EPA’s current criterion to classify acute oral toxicants as Category II (50 < LD50 ≤ 500 mg/kg) or Category III (500 < LD50 ≤ 5,000 mg/kg) (U.S. EPA 2012).

Many compounds have logLD50 values close to the threshold (i.e., a logLD50 of 3 ± 0.5), making these compounds difficult to classify as either toxic or nontoxic. Changing the parameter s has the effect of shrinking or extending the range of the involved values of compounds close to this threshold (bold line in Figure 2B). By selecting an s value of 0.5, the outputted f value of these compounds would fall between 0.75 and 0.25 (e.g., logLD50 value as 2.5 with the associated f output as 0.75). As shown in Figure 2A,B, many compounds in the database can fall within these two areas.

**Subspace Clustering of PubChem Assays**

Among the 1,077 PubChem bioassays in the original bioprofile, 707 had at least one ToxPrint chemical fragment that was significantly correlated with activity, resulting in a total of 15,064 significant chemical fragment—*in vitro* bioassay activity relationships (p-value < 0.05). These relationships were used to cluster the PubChem assays by calculating the dissimilarity (Jd) between every pair of bioassays. As shown in Figure 3, the edges between two nodes (bioassays) indicate dissimilarity values less than 0.75. Sixty-seven bioassays are unique compared to others (no Jd values less than 0.75) and are not shown in this figure. Within the remaining 640 bioassays, the Louvain modularity algorithm (Blondel et al. 2008) identified 45 unique clusters with 2 to 60 bioassays per cluster (Excel Table S1). The clusters result in grouping bioassays that are potentially related in their ability to inform on biological mechanism as it pertains to acute oral toxicity.

The 640 bioassays came from different source depositories, which are summarized in Table 1. Bioassays from different sources coexisted within clusters, with 37 out of 45 clusters containing bioassays from ≥ two sources. The biological targets of bioassays varied, and the majority could be classified as overt toxicity (e.g., cytotoxicity assays), biomarkers of cellular responses (e.g., mitochondrial membrane potential assays), or specific protein targets (e.g., agonists of the androgen receptor). Some clusters showed a clear preponderance of one bioassay type (e.g., Cluster 1 consisted of only cytotoxicity assays).

**Acute Oral Toxicity Model Selection and Ensemble Modeling**

For modeling purposes, the 45 PubChem bioassay clusters shown in Figure 3 were evaluated for their ability to predict acute oral toxicity. Bioprofile-based read-across studies were performed within each cluster and assessed using a fivefold cross-validation procedure. Different thresholds of biosimilarity (ranging from 0.5 to 0.9) and rcc values (ranging from 0 to 100%) were used to determine the most similar compound in the training set to a test compound. The best ppv for each cluster was recorded, with cross-validated ppv values from 0 to 100% across the bioassay clusters (Figure 4). The clusters with a ppv above 60% were selected as viable potential models of acute oral toxicity. To ensure the models are statistically significant and interpretable, clusters with <five bioassays were removed. This procedure resulted in 19 PubChem bioassay models that were potentially applicable for acute oral toxicity prediction. In order to leverage predictions across multiple models (and thus multiple potential biological mechanisms), an ensemble model was created by averaging the predictions across all the cluster-specific models.

**PubChem Bioassays Selected for Acute Toxicity Classifications**

The PubChem bioassays identified by our approach for toxicity predictions came from a variety of screening programs. These programs have different research goals and vary in scope and size, and many were not designed to be used for acute oral toxicity evaluation purposes. However, the analysis of the top-ranked cluster-based models unveiled several mechanisms clearly relevant to acute oral toxicity, including cytotoxicity/growth inhibition in cells of various origins including tumor cell lines, viral growth inhibition, and assays measuring protein interaction and/or function.

Cluster 1 had the highest accuracy in the cross-validation procedure (100% ppv). It consisted entirely of data deposited from the Developmental Therapeutics Program at the National Cancer Institute (NCI). The 17 NCI bioassays comprising Cluster 1 measure cytotoxicity/growth inhibition against tumor cell lines and
have been shown to be highly correlated with acute toxicity in an earlier study (Zhang et al. 2014a). Cluster 2 (96% ppv in the cross-validation procedure) included not only cytotoxicity assays from NCI but also a significant number of bioassays with protein targets regulating cell growth (e.g., steroid receptor coactivators 1 and 3, paternally expressed gene 3, hypoxia-inducible factor 2 alpha). Cluster 3 (87% ppv in the cross-validation procedure) contained 60 PubChem in vitro bioassays and was the largest among all

Table 1. Source depositors of the PubChem 640 assays used in this study.

| Source institution                                      | Number of assays | Cluster membership
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NCGC</td>
<td>200</td>
<td>2, 3, 4, 8, 9, 10, 12, 13, 14, 16, 17, 19, 20, 21, 23, 24, 25, 26, 27, 29, 30, 32, 35, 36, 37, 39, 40, 41, 42, 44, 45</td>
</tr>
<tr>
<td>Tox21</td>
<td>114</td>
<td>3, 9, 10, 11, 12, 15, 17, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 38, 44</td>
</tr>
<tr>
<td>DTP/NCI</td>
<td>96</td>
<td>1, 2, 5, 6, 7, 8, 10, 18, 19, 20, 21, 27</td>
</tr>
<tr>
<td>Scripps Research Institute Molecular Screening Center</td>
<td>67</td>
<td>2, 3, 4, 5, 8, 10, 14, 16, 18, 19, 20, 21, 27, 35, 37, 39, 40</td>
</tr>
<tr>
<td>Sanford-Burnham Center for Chemical Genomics</td>
<td>41</td>
<td>2, 3, 4, 5, 8, 10, 14, 15, 18, 20, 21, 27, 35, 37, 43, 45</td>
</tr>
<tr>
<td>Broad Institute</td>
<td>31</td>
<td>2, 3, 8, 10, 12, 14, 16, 18, 20, 21, 27, 34, 35</td>
</tr>
<tr>
<td>Cheminformatics and Chemogenomics Research Group</td>
<td>16</td>
<td>3, 15, 17, 20, 24, 27, 29, 32, 36</td>
</tr>
<tr>
<td>EPA DSSTox</td>
<td>11</td>
<td>3, 6, 7, 37, 38</td>
</tr>
<tr>
<td>Johns Hopkins Ion Channel Center</td>
<td>11</td>
<td>3, 10, 13, 17, 20</td>
</tr>
<tr>
<td>Southern Research Institute</td>
<td>10</td>
<td>3, 14, 15, 17, 20, 21, 24</td>
</tr>
<tr>
<td>Southern Research Specialized Biocontainment Screening Center</td>
<td>10</td>
<td>3, 12, 27, 37</td>
</tr>
<tr>
<td>Emory University Molecular Libraries Screening Center</td>
<td>9</td>
<td>3, 14, 21, 28</td>
</tr>
<tr>
<td>University of New Mexico</td>
<td>8</td>
<td>14, 18, 20, 21, 41</td>
</tr>
<tr>
<td>ICCB-Longwood/NSRB Screening Facility, Harvard Medical School</td>
<td>4</td>
<td>3, 12, 14</td>
</tr>
<tr>
<td>Vanderbilt High Throughput Screening Facility</td>
<td>3</td>
<td>3, 17, 20</td>
</tr>
<tr>
<td>Columbia University Molecular Screening Center</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>University of Pittsburgh Molecular Library Screening Center</td>
<td>2</td>
<td>3, 17</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>2</td>
<td>14, 27</td>
</tr>
<tr>
<td>Milwaukee Institute for Drug Discovery</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Institute for Research in Immunology and Cancer</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Psychoactive Drug Screening Program</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

Note: ChEMBL, European Molecular Biology Laboratory chemistry database; DTP/NCI, Developmental Therapeutics Program/National Cancer Institute; EPA DSSTox, Environmental Protection Agency Distributed Structure-Searchable Toxicity Database; GPCR, G protein–coupled receptors; ICCB, Institute of Chemistry and Cell Biology; NCGC, National Center for Advancing Translational Sciences Chemical Genomics Center; NSRB, National Screening Laboratory for the Regional Centers of Excellence for Biodefence and Emerging Infectious Diseases; Tox21, Toxicity Testing in the 21st Century.

*Cluster membership displays an exhaustive list of all the clusters to which a particular source has at least one bioassay belong to, as identified by the Louvain modularity algorithm.
19 selected clusters. It consisted of data deposited from the National Center for Advancing Translational Sciences (formerly National Chemical Genomics Center) (25%), the Scripps Research Institute (20%), Tox21 (12%), Emory University Molecular Libraries Screening Center (10%), Johns Hopkins University (8%), and others (25%). Only a few of these bioassays were cytotoxicity assays, with the majority being associated with protein targets, including nuclear receptors, membrane channels, kinases, and various enzymes. Cluster 8 (ppv 80%) consisted of a mixture of cell viability, protein targets, and viral bioassays from five sources.

**Ensemble Predictions for New Compounds**

Prior experimentation has demonstrated that integration of multiple models can show superior predictivity to individual models (Kim et al. 2014; Solimeo et al. 2012; Sprague et al. 2014; Wang et al. 2015; Zhao et al. 2017; Zhu et al. 2009). In this project, one such approach was applied, which involved generating an ensemble model using the 19 models selected (i.e., those having ppv > 60%) to predict acute oral toxicity. For evaluation purposes, the resulted models were used to predict an external test set consisting of 639 compounds that were not present within the original training set. To build the ensemble model, bioprofile-based read-across prediction output values were averaged, per chemical, from all 19 models (Figure 5). Many new compounds contained varying amounts of in vitro biological data among the 19 PubChem bioassay clusters, and the read-across performance was also different among these 19 models. When applying increasing thresholds of confidence for the read-across predictions, some of these models were not able to provide predictions for new compounds due to insufficient data in the associated PubChem bioassays. To quantify this source of uncertainty, confidence values (rcc) were computed and evaluated. Ensemble model predictions made with a low confidence threshold (e.g., rcc values ≤ 10%, indicating that the external test set chemical and training set chemical were tested in few of the same bioassays) resulted in poor prediction of new compounds with ppv of 60%. However, when the confidence threshold was increased to eliminate unreliable predictions (e.g., rcc values > 30%), the ppv value increased significantly (ppv > 76%). This improved ensemble model performance at higher confidence thresholds was similar to the cross-validation results of the most selected models shown in Figure 4.

**Toxicity Pathways for Acute Oral Toxicity**

To identify potential toxicity mechanisms from predictions, chemical fragment–in vitro–in vivo relationship analysis was performed within each individual cluster. Unfortunately, a complete dataset is critical for this analysis. To resolve the missing data issue, QSAR model predictions were used to fill in the missing data for the 3,842 compounds that were initially omitted from the training set due to insufficient bioassay data. The use of QSAR model predictions will add additional uncertainty since the QSAR model predictivity of PubChem bioassay responses for new compounds was generally around 70% (Figure S2).

The population of the bioprofile data by QSAR predictions can provide insights to reveal potential toxicity mechanisms within each cluster. For example, investigation into the chemical scaffolds of the predicted toxic compounds within a cluster can reveal toxicophores related to acute oral toxicity. These toxicophores can be used to explain toxicity mechanisms by integrating bioassay data used for read-across in relevant clusters. Two of the 19 clusters showed exceptional chemical fragment–in vitro–in vivo relationships. Cluster 1, consisting of 17 NCI tumor cell line growth inhibition assays, had 24 predicted toxic compounds at an rcc of 100%, 13 of which were true positives (i.e., toxic in vivo) and 11 of which were false positives (i.e., nontoxic in vivo based on our toxic/nontoxic designations using a 1,000-mg/kg threshold). Examination of the chemical structures of the true positives and false positives (Figure 6; Excel Table S2) revealed distinctly different chemical scaffolds and suggested a potential toxicophore. Within the compounds predicted as toxic by this model, eight compounds had a steroid backbone (Figure 6B), and seven were classified as toxic in acute oral animal studies. Additionally, 10...
Figure 5. Bioprofile-based prediction performance on the external test set of 639 chemicals. The 639 external test set compounds were used to evaluate the predictiveity of 19 clusters and the ensemble model at varying relative cluster confidence (rcc) levels. (A) Positive predictive value (ppv) of the ensemble model at increasing rcc thresholds (same x-axis as grid plot in panel B). (B) Grid plot showing confusion matrices of the ensemble model and the individual clusters at different rcc thresholds. Each confusion matrix contains the number of false positives (upper left corner), true positives (upper right corner), false negatives (lower left corner), and true negatives (lower right corner) for a model at a particular rcc threshold.
of the 17 bioassays in this model had this chemical fragment statistically relevant to their activity by Fisher’s exact test.

Cluster 8 consisted of seven cytotoxicity assays and seven protein and viral targets, the latter of which are listed in Table 2. This model showed a remarkable predictivity (modeling and external test sets had ppr of 80 and 63%, respectively). The associated toxicophore within this cluster is a trifluoromethyl-substituted benzimidazole ring (Figure 7). While this ring structure appears in 129 of

Figure 6. Cluster 1 predictions. (A) Principal component analysis (PCA) of Cluster 1 model predictions (true positives as highlighted circles and false positives as diamonds). (B) Two examples of true positive compounds with their common representative substructure in bold. (C) Two examples of false positive compounds with their common representative substructure in bold.
the predicted chemicals, and all but one was toxic based on our classification criteria, not all were within the domain of this cluster’s model. Yet the 46 toxic compounds that contained this fragment and were within the domain of this cluster’s model were all correctly classified as toxic (Excel Table S3). This highlights the important role of an applicability domain assessment, which we integrated into the evaluation of the model. The compounds containing this toxicophore are pesticides and behave as casein kinase 2 (CK2) inhibitors, (e.g., 4,5,6,7-tetrabromobenzimidazole, CAS 749,234-11-5) (Adamson et al. 1984; Jones and Watson 1965; Pagano et al. 2004).

Discussion
We have recently advocated for the integration of biological information, such as data from PubChem, into read-across (Zhu et al. 2016). In this study, we demonstrated a novel approach to integrate diverse in vitro data from a rapidly evolving public resource to serve as the foundation for bioprofile-based read-across predictions. The resulting bioassay cluster-based predictive models, and the ensemble model generated by combining cluster predictions, show great promise for predicting acute oral toxicity as well as informing on possible mechanisms contributing to toxicity.

The success of bioprofile-based read-across for the external test set (Figure 5) was different for each model as compared to the cross-validation within training sets (Figure 4). For example, Cluster 1, which mostly consisted of growth inhibition/cytotoxicity assays and yielded the best cross-validation accuracy (Figure 4), had no significant contribution to the external test set predictions, as shown in Figure 5. This is due to the lack of sufficient bioassay data for the external test set compounds in the subset of

Table 2. PubChem Assays involving protein or viral targets within Cluster 8 model.

<table>
<thead>
<tr>
<th>PubChem AID</th>
<th>Name</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>488899</td>
<td>MITF measured in cell-based system using plate reader</td>
<td>Microphthalmia-associated transcription factor (Homo sapiens)</td>
</tr>
<tr>
<td>504444</td>
<td>Nrf2 qHTS screen for inhibitors</td>
<td>Nuclear factor erythroid 2-related factor 2 isoform 2 (H. sapiens)</td>
</tr>
<tr>
<td>540276</td>
<td>qHTS for inhibitors of binding or entry into cells for Marburg virus</td>
<td>Gene 4 small orf (Marburg virus)</td>
</tr>
<tr>
<td>588413</td>
<td>uHTS identification of Gli-Sufu antagonists in a luminescence reporter assay</td>
<td>Glioma-associated oncogene 1 (Mus musculus)</td>
</tr>
<tr>
<td>624169</td>
<td>Luminescence-based cell-based primary high-throughput screening assay to identify agonists of the mouse 5-hydroxytryptamine (serotonin) receptor 2A (HTR2A)</td>
<td>5-Hydroxytryptamine receptor 2A (M. musculus)</td>
</tr>
<tr>
<td>624354</td>
<td>uHTS identification of Caspase-8 TRAIL sensitizers in a luminescence assay</td>
<td>Tumor necrosis factor receptor superfamily member 10B isoform 1 precursor (H. sapiens)</td>
</tr>
<tr>
<td>651820</td>
<td>qHTS assay for inhibitors of hepatitis C virus (HCV)</td>
<td>Hepatitis C virus</td>
</tr>
</tbody>
</table>

Note: MITF, microphthalmia-associated transcription factor; Nrf2, nuclear factor erythroid 2-related factor 2; qHTS, quantitative high-throughput screening; TRAIL, tumor necrosis factor-related apoptosis-inducing ligand; uHTS, ultra-high-throughput screening.

Figure 7. Cluster 8 predictions. (A) Principal component analysis (PCA) of Cluster 8 model predictions (true positives as highlighted circles and false positives as diamonds). (B) Three examples of true positives and the potential toxicophores associated with this cluster in bold are shown.
PubChem bioassays. Although cytotoxicity assays have been proven to be useful in predicting acute oral toxicity (Barile et al. 1994; Garle et al. 1994; Ukelis et al. 2008), these bioassays will not contribute to predicting the toxicity of new compounds if the new compounds lack sufficient data. This issue can be solved by predicting new compound bioactivity in vitro (e.g., by QSAR modeling conducted herein and accepting an additional margin of uncertainty) or by testing new compounds using these bioassays. Consensus models have typically provided superior or equivalent predictive relationships to the individual models in our previous QSAR studies (Kim et al. 2014; Solimeo et al. 2012; Sprague et al. 2014; Wang et al. 2015; Zhao et al. 2017; Zhu et al. 2009). Here we used a similar approach, where an ensemble model was used to partially resolve this issue of insufficient data. Using this approach, the predictivity for new compounds in the external test set showed goodppv (76%) at reasonable prediction confidence (rec of 30%) but with a low coverage (~10%). This reflects the importance of having enough bioassay data to evaluate new compounds for our bioprofile-based prediction approach to succeed.

In both fivefold cross-validation and the external evaluation, cytotoxicity assays appear as strong predictors of acute oral toxicity. The best performing models in fivefold cross-validation contain cytotoxicity assays from NCI (e.g., those in Clusters 1 and 2); however, they lacked sufficient data for evaluation purposes using the external test compounds. On the other hand, Clusters 11 and 23 showed accurate predictivity and contained many cytotoxicity assays from the Tox21 screening program. With reasonable confidence (rec of 30%), they both showed high predictivity for the external test set compounds (ppv of 88% and 68%, respectively).

Unfortunately, the external test set compounds were not numerous enough to establish meaningful chemical fragment–in vitro–in vivo relationships. However, by using the QSAR predictions, enough data could be generated to see relationships between certain chemical fragments and in vitro bioassay activity responses within a cluster. This information can be used to enhance read-across predictions and/or explore toxicity mechanisms. While the use of QSAR undoubtedly introduces uncertainty into the read-across predictions, manual review of the model outputs (namely, chemical scaffolds of predicted compounds) can aid in bolstering confidence in the results. For example, identifying the distinction of true positive vs. false positive predictions within Cluster 1 (Figure 6) was explored. This cluster consisted entirely of cytotoxicity/growth inhibition bioassays in immortalized cell lines. Many of the true positives in these bioassays were steroids, which have been established as lead anticancer agents stemming from their affinity for nuclear receptors (Carvalho et al. 2010; Gupta et al. 2013). Indeed, the literature supports that steroid type toxicants induce acute toxicity because of cytotoxicity (Tantawy et al. 2017; Ur Rahman et al. 2017). On the other hand, 10 of the 11 false positive compounds were antibiotics that contain a large ring structure (greater than six members). Because of this unique chemical structure, antibiotics are likely to be misclassified by the QSAR models and should likely be excluded from the domain of applicability for the Cluster 1 model. These results suggest that cytotoxicity assays are reasonable alternative approaches for screening steroids to inform on acute oral toxicity and highlight the implications of conducting applicability domain assessments. Another consideration worth noting is that these bioassays use human cell lines but are being used to predict acute oral toxicity identified in rats. However, we do not expect a significant interspecies effect with regards to cytotoxicity assays because previous studies have demonstrated good concordance between cytotoxicity results obtained from rat and human cell lines (Clemenson and Ekwall 1999; Clothier et al. 2013).

Cluster 8 contained several pesticides that are putative CK2 inhibitors sharing a benzimidazole ring (Figure 7). CK2 is a serine/threonine protein kinase targeting a wide array of proteins involved in several cell processes, including mediating cell cycle and apoptosis (Hamacher et al. 2007; Litchfield 2003; Yamane and Kinsella 2005), and at least two PubChem bioassays within this cluster are relevant to it. The first assay (PubChem AID 504444) screens for inhibitors of nuclear factor erythroid 2–related factor 2 (Nrf2), a transcription factor intimately involved in the cellular response to oxidative stress (Hur and Gray 2011). Nrf2 responds to oxidative stress by translocating to the nucleus, where it binds to the antioxidant response element (ARE) and induces expression of an array of genes encoding antioxidants (Jaiswal 2004). It has been shown that Nrf2 phosphorylation by CK2 is required for translocation to the nucleus and subsequent ARE activation (Apopa et al. 2008; Pi et al. 2007). Another assay (PubChem AID 588413) measures the inhibition of Gli1, a transcription factor involved in the hedgehog (Hh) signaling pathway (Ruiz i Altaba 1999; Villavicencio et al. 2000). The Hh pathway is involved in cell proliferation, cell maintenance, cell differentiation, and embryonic development (Gupta et al. 2010; Mahindroor et al. 2009). In addition to these processes, Hh signaling and Gli1 expression have been delineated as responders to oxidative stress, suggesting they have a role in regulating antioxidant genes (Chen et al. 2017; Yao et al. 2017). Similar to Nrf2, recent work has also shown CK2 to be a positive modulator of Gli1 activation (Jin et al. 2011) and downstream Hh pathway signaling (Zhang et al. 2012, 2014b). Thus, toxic compounds containing the pharmacophore in Figure 7 may initiate a toxicity pathway by inhibition of CK2 and disruption of cell homeostasis in one of the following ways: a) by obstructing Nrf2 translocation to the nucleus, the antioxidant response is diminished and leaves the cell more susceptible to oxidative stress; and b) inhibition of Gli1 could result in silencing of the Hh pathway, disrupted cellular growth, and additional lowered antioxidant response. This example highlights the additional context that could be inferred from in-depth review of relationships between the bioassays within clusters to help inform on potential novel molecular mechanisms underlying acute oral toxicity.

Developing nonanimal models for acute oral toxicity evaluations still poses a significant challenge. While QSAR models built from the same rat acute oral toxicity training set offered an acceptableppv on the same external test set compounds (ppv > 62%, shown in Excel Table S4), the prediction accuracy was lower than that achieved by the presented ensemble model (Figure 5). Furthermore, by leveraging in vitro bioactivity data, the bioprofile-based read-across method presented herein offers a solution to potential activity cliff issues existing in normal QSAR modeling (Maggiora 2006) because the new compound predictions are not limited to the information only obtained from chemical structures. To succeed, the integration of even more in vitro bioassays will likely be necessary to associate toxicity mechanisms for untested potential toxicants. Many of the toxicity mechanisms are obscure, so it is limiting to rely solely on the available bioassay data, especially unstructured public data, for toxicity evaluation purposes. This study shows that toxicity evaluations can be significantly enhanced by using meaningful biological data.

The new computational approach developed in this study can automatically identify useful biological data from public data sources and is capable of performing bioprofile-based read-across for untested compounds based on chemical fragment–in vitro–in vivo relationships. By doing so, chemical toxicity mechanisms can be elucidated. Such promising bioassays can then be used to characterize unknown substances. For example, these bioassays can be incorporated into weight of evidence approaches, such as an integrated testing strategy (Hartung et al. 2013; Rovida et al. 2015). Such strategies enable the incorporation of additional data that

Environmental Health Perspectives
047001-11
127(4) April 2019
may be critical for more relevant risk assessment, including bioassays capable of in vitro metabolism (Jacobs et al. 2013; McKim 2010; Yoon et al. 2012) as well as computational models to predict bioavailability (Bhattacharai et al. 2015; Kim et al. 2014). Read-across studies using biological data strongly depend on the reliability of bioassay testing protocols and the quality of biological data. Potential experimental errors and other relevant issues (e.g., data reproducibility) may affect the confidence of data from bioassays. Although it is beyond the scope of this study, we have recommended potential solutions to improve the quality of public data in previous publications (Sedykh et al. 2011; Zhao et al. 2017).

Conclusions

For complex animal toxicity end points, such as acute oral toxicity, the complete replacement of animal testing is still not feasible. However, efforts to prioritize potentially hazardous chemicals by leveraging reliable and sufficient bioassay data that can be linked to specific toxicity mechanism(s) can significantly reduce the number of animals used, save great resources in chemical toxicology studies, and facilitate hazard assessment of high-priority chemicals. The data-driven profiling strategy presented in this study provides a novel way of extracting pertinent information from a daily updated, unstructured public resource. In contrast to previous studies, our method incorporates both chemical (i.e., chemical structure) and biological (in vitro bioassays) data into the workflow. This approach not only predicts acute oral toxicity classification but also infers biological mechanism information, offering novel insights into mechanisms of acute oral toxicity as well as in vitro bioassays and their utility for predicting in vivo toxicity. Furthermore, this method can easily be expanded to develop nonanimal models to evaluate other complex animal toxicities beyond acute oral systemic toxicity.

Acknowledgments

D.P.R., W.W., and H.Z. were partially supported by the National Institute of Environmental Health Sciences (grant number R15ES023148), the Colgate-Palmolive Grant for Alternative Research, and the Johns Hopkins Center for Alternatives to Animal Testing (CAAT) grant. J.S. and A.K. were supported by NIH R15ES023148), the Colgate-Palmolive Grant for Alternative National Institute of Environmental Health Sciences (grant number.

References


Environmental Health Perspectives

047001-12 127(4) April 2019