Fundamentals of QSAR modeling: basic concepts and applications

Alexander Tropsha
University of North Carolina, Chapel Hill, USA
Key points

• Basic concepts and best practices of QSAR modeling
• Data curation
• Case study and model interpretation: alerts about alerts
• Emerging approaches: Hybrid (chemical-biological) QSAR modeling and Chemical Biological Read Across (CBRA)
• Summary of QSAR as (regulatory) decision support tool
The growing appreciation of molecular modeling and informatics

Next RSC president predicts that in 15 years no chemist will do bench experiments without computer-modelling them first

The newly-appointed President-Elect of the Royal Society of Chemistry today forecast the impact of advances in modelling and computational informatics on chemistry.
The chief utility of computational models: Hit identification in external libraries
QSAR Modeling

Quantitative Structure Activity Relationship

Slide credit: UNC MML
Structure representation
Graphs are widely used to represent and differentiate chemical structures, where atoms are vertices and bonds are expressed as edges connecting these vertices.

Molecular graphs allow the computation of numerous indices to compare them quantitatively.

Molecular descriptors
Datasets are represented by a matrix of molecular descriptors

<table>
<thead>
<tr>
<th>Samples (Compounds)</th>
<th>Variables (descriptors)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_1$</td>
</tr>
<tr>
<td>1</td>
<td>$X_{11}$</td>
</tr>
<tr>
<td>2</td>
<td>$X_{21}$</td>
</tr>
<tr>
<td>...</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>n</td>
<td>$X_{n1}$</td>
</tr>
</tbody>
</table>
Compounds represented by vectors in a multidimensional descriptor space

\[ \mathbf{R}_M = \mathbf{R}_M (x_{M1}, x_{M2}, \ldots, x_{MK}) \]
Molecules may form clusters in chemical space.

Cluster 2
Cluster 3
Cluster 4
Cluster 1

Molecules are considered as vectors in the space of descriptors (« chemical » space).

Dimensions of this space correspond to the number of descriptors.

Clustering methods are employed to analyze distances between compounds and identify clusters.
**QSAR Modeling**

Establish **quantitative relationships** between descriptors and the target property capable of predicting activities of novel compounds.

<table>
<thead>
<tr>
<th>Chemistry</th>
<th>Bioactivity (IC50, Kd...)</th>
<th>Cheminformatics (Molecular Descriptors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
<td>Value1</td>
<td>D₁ D₂ D₃</td>
</tr>
<tr>
<td>Comp.2</td>
<td>Value2</td>
<td>&quot; &quot; &quot;</td>
</tr>
<tr>
<td>Comp.3</td>
<td>Value3</td>
<td>&quot; &quot; &quot;</td>
</tr>
<tr>
<td>Comp.N</td>
<td>ValueN</td>
<td>&quot; &quot; &quot;</td>
</tr>
</tbody>
</table>

**BA = F(D)** (linear, e.g., -LogIC50 = k₁D₁ + k₂D₂ + ... + kₙDₙ) or non-linear, e.g. k nearest neighbors

![Graph showing predicted vs actual LogED50](image.png)
QSAR Modeling Workflow: the importance of rigorous validation

Experimental confirmation

Virtual screening (with AD threshold)

Evaluation of external performance

An ensemble of QSAR Models

Internal validation

Model selection

Combi-QSAR modeling

Modeling methods

K-Nearest Neighbors (kNN)

Random Forest (RF)

Support Vector Machines (SVM)

Descriptors

Dragon

MOE


Fully implemented on CHEMBENCH.MML.UNC.EDU
Data dependency and data quality are critical issues in QSAR modeling

Believe it or not: how much can we rely on published data on potential drug targets?

**CORRESPONDENCE**

results that are published are hard to reproduce. However, there is an imbalance between this apparently widespread impression and public recognition (for example, see REF 2) and the surprisingly few scientific publications dealing with this topic. Indeed, to our knowledge, so far there has been no published in-depth, systematic analysis that compares...

Believe it or not: how much do we actually rely on published data for drug targets?

Are the Chemical Structures in Your QSAR Correct?

Douglas Younga, Todd Martinb, Raghuraman Venkatapathyb, and Paul Hartena

a US Environmental Protection Agency, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA; E-mail: young.douglas@epa.gov
b Pegasus Technical Services, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA

Keywords: Databases, N-octanol/water partition coefficient, Quantitative structure-activity relationships, SMILES

Received: June 26, 2008; Revised: August 13, 2008; Accepted: August 21, 2008

DOI: 10.1002/qsar.200810084
Data dependency and data quality are critical issues in QSAR modeling.

Believe rely on drug ta

Dispensing Processes Impact Apparent Biological Activity as Determined by Computational and Statistical Analyses

Sean Ekins, Joe Olechno, Antony J. Williams

Abstract

Dispensing and dilution processes may profoundly influence estimates of biological activity of compounds. Published data show Ephrin type-B receptor 4 IC₅₀ values obtained via tip-based serial dilution and dispensing versus acoustic dispensing with direct dilution differ by orders of magnitude with no correlation or ranking of datasets. We generated computational 3D pharmacophores based on data derived by both acoustic and tip-based transfer. The computed pharmacophores differ significantly depending upon dispensing and dilution methods. The acoustic dispensing-derived pharmacophore correctly identified active compounds in a subsequent test set where the tip-based method failed. Data from acoustic dispensing generates a pharmacophore containing two hydrophobic features, one hydrogen bond donor and one hydrogen bond acceptor. This is consistent with X-ray crystallography studies of ligand-protein interactions and automatically generated pharmacophores derived from this structural data. In contrast, the tip-based data suggest a pharmacophore with two hydrogen bond acceptors, one hydrogen bond donor and no hydrophobic features. This pharmacophore is inconsistent with the X-ray crystallographic studies and automatically generated pharmacophores. In short, traditional dispensing processes are another important source of error in high-throughput screening that impacts computational and statistical analyses. These findings have far-reaching implications in biological research.
Data dependency and data quality are critical issues in QSAR modeling.


A Disturbing Note in a Recent SI File

August 6th, 2013

A recently published ASAP article in the journal Organometallics is sure to raise some eyebrows in the chemical community. While the paper itself is a straightforward study of palladium and platinum bis-sulfoxide complexes, page 12 of the corresponding Supporting Information file contains what appears to be an editorial note that was inadvertently left in the published document:

Emma, please insert NMR data here! Where are they? And for this compound, just make up an elemental analysis...

This statement goes beyond a simple embarrassing failure to properly edit the manuscript, as it appears the first author is being instructed to fabricate data. Elemental analyses would be very easy to fabricate, and long-time readers of this blog will recall how fake elemental analyses were pivotal to Bengu Sezen’s campaign of fraud in the work she published from 2002 to 2005 out of Dalibor Sames’ lab at Columbia.

The compound labeled 14 (an acac complex) in the main paper does not appear to correspond to compound 14 in the SI. In fact, the bridged-dichloride compound appears to be listed as an unlabeled intermediate in Scheme 5, which should raise more eyebrows. Did the authors unlist the compound in order to avoid having to provide robust characterization for it?

ChemBark is contacting the corresponding author for comment, and his response will be posted in full when we receive it.
In the Pipeline


April 11, 2014

Biology Maybe Right, Chemistry Ridiculously Wrong

As my correspondent (a chemist himself) mentions, a close look at Figure 2 of the paper raises some real questions. Take a look at that cyclohexadiene enamine - can that really be drawn correctly, or isn't it just N-phenylbenzylamine? The problem is, that compound (drawn correctly) shows up elsewhere in Figure 2, hitting a completely different pathway. These two tautomers are not going to have different biological effects, partly because the first one would exist for about two molecular vibrations before it converted to the second. But how could both of them appear on the same figure?

And look at what they're calling "cyclohexa-2,4-dien-1-one". No such compound exists as such in the real world - we call it phenol, and we draw it as an aromatic ring with an OH coming from it. Thiazolidinedione is listed as "thiazolidine-2,4-quinone". Both of these would lead to red "X" marks on an undergraduate exam paper. It is clear that no chemist, not even someone who's been through second-year organic class, was involved in this work (or at the very least, involved in the preparation of Figure 2). Why not? Who reviewed this, anyway?

DOI: 10.100
Data dependency and data quality are critical issues in QSAR modeling.


Policy: NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

27 January 2014

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

Subject terms: Biological techniques • Lab life • Peer review • Research management

And such conclusions lead to the possible scenario that we may never see the end of the Madden era when we receive it.
Dealing with Irreproducibility

Researchers discuss the growing pressures that are driving increases in retraction rates at AACR.

By Jef Akst | April 8, 2014

Recent years have seen increasing numbers of retractions, higher rates of misconduct and fraud, and general problems of data irreproducibility, spurring the National Institutes of Health (NIH) and others to launch initiatives to improve the quality of research results. Yesterday (April 7), at this year’s American Association for Cancer Research (AACR) meeting, researchers gathered in San Diego, California, to discuss why these problems come to a head—and how to fix them.

“We really have to change our culture and that will not be easy,” said Lee Ellis from the University of...
QSAR modeling with non-curated datasets

- Presence of ERRONEOUS OR WRONG STRUCTURES
- Presence of MISPRINTS AND WRONG NAMES
- ERRORS in the calculation of DESCRIPTORS
- Presence of DUPLICATES
- Presence of MIXTURES
- Presence of SALTS

QSAR MODELS ???

0.613
0.380
-0.222
7.08
1.146
0.491
0.301
0.141
0.956
0.256
0.799
1.195
1.005

Etc.
Chemical Structure Curation

Chemical structures should be cleaned and standardized (duplicates removed, salts stripped, neutral form, canonical tautomer, etc.) to enable rigorous model development.

**Quinine sulfate dihydrate**

**Pyridostigmine Bromide**

**Fenoprofen Sodium**

QSAR modeling of nitro-aromatic toxicants

- Case Study 1: 28 compounds tested in rats, log(LD50), mmol/kg.
- Case Study 2: 95 compounds tested against Tetrahymena pyriformis, log(IGC50), mmol/ml.

- Five different representations of nitro groups.
  - Case Study 1: after the normalization of nitro groups $R^2_{\text{ext}} \approx 0.45$ increased to $R^2_{\text{ext}} \approx 0.9$.
  - Case Study 2: after the normalization of nitro groups $R^2_{\text{ext}} \approx 0$ increased to $R^2_{\text{ext}} \approx 0.5$

Even small differences in structure representation can lead to significant errors in prediction accuracy of models

Artemenko, Muratov et al. SAR QSAR 2011, 22 (5-6), 1-27.
QSAR modeling of nitro-aromatic toxicants

- Case Study 1: 28 compounds tested in rats, log(LD50), mmol/kg.
- Case Study 2: 95 compounds tested against *Tetrahymena pyriformis*, log(IGC50), mmol/ml.

Data curation affects the accuracy (up or down!) of QSAR models.

Even small differences in structure representation can lead to significant errors in prediction accuracy of models.

Artemenko, Muratov et al. SAR QSAR 2011, 22 (5-6), 1-27.
Curation of Bioactivity: Case study

Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data

Hongmao Sun,*,† Henrike Veith,† Menghang Xia,† Christopher P. Austin,† and Ruili Huang†

†National Institutes of Health (NIH) Chemical Genomics Center, NIH,

ABSTRACT: The human cytochrome P450 (CYP450) isozymes are the most important enzymes in the body to metabolize many endogenous and exogenous substances including environmental toxins and therapeutic drugs. Any unnecessary interactions between a small molecule and CYP450 isozymes may raise a potential to disarm the integrity of the protection. Accurately predicting the potential interactions between a small molecule and CYP450 isozymes is highly desirable for assessing the metabolic stability and toxicity of the molecule. The National Institutes of Health Chemical Genomics Center (NCGC) has screened a collection of over 17,000 compounds against the five major isozymes of CYP450 (1A2, 2C9, 2C19, 2D6, and 3A4) in a quantitative high throughput screening (qHTS) format. In this study, we developed support vector classification (SVC) models for these five isozymes using a set of customized generic atom types. The CYP450 data sets were randomly split into equal-sized training and test sets. The optimized SVC models exhibited high predictive power against the test sets for all five CYP450 isozymes with accuracies of 0.93, 0.89, 0.89, 0.85, and 0.87 for 1A2, 2C9, 2C19, 2D6, and 3A4, respectively, as measured by the area under the receiver operating characteristic (ROC) curves. The important atom types and features extracted from the five models are consistent with the structural preferences for different CYP450 substrates reported in the literature. We also identified novel features with significant discerning power to separate CYP450 actives from inactives. These models can be useful in prioritizing compounds in a drug discovery pipeline or recognizing the toxic potential of environmental chemicals.

CONCLUSION

SVM classification models have been built for the five most important isozymes of CYP450 (1A2, 2C9, 2C19, 2D6, and 3A4) based on a large qHTS data set with over 6000 compounds available for both model training and testing. The five CV optimized SVC models built by using the atom typing molecular descriptors exhibited consistently high predictive power when applied to the equally populated test sets with accuracies between 0.85 and 0.93, as measured by the AUC of ROC plots. The results indicated that the atom typing descriptors generated from a large, high quality data set were capable of feeding information rich learning materials to the SVM learner. Useful information of structural features was derived from feature importance analysis for each isozyme of CYP450. The privileged structural features that could result in inhibitory and stimulatory activity against different CYP450 isozymes can serve as valuable guidelines in the drug discovery process.
Dataset Curation summary

17143 compounds
- Removal of inorganics and mixtures
- Structural conversion, cleaning of salts
- Normalization of specific chemotypes
- Treatment of tautomeric forms
- Removal of duplicates
- Manual inspection

17121 compounds
- 17121 compounds
- 17121 compounds
- 17121 compounds
- 16142 compounds
- 16142 compounds

NCGC dataset: analysis of duplicates

- Out of 1280 duplicate couples:
  - 406 had no discrepancies—no values or no values for comparison
  - 874 had biological profile differences

- A total of 1535 discrepancies were found in the

<table>
<thead>
<tr>
<th></th>
<th>CYP2C9</th>
<th>CYP1A2</th>
<th>CYP3A4</th>
<th>CYP2D6</th>
<th>CYP2C19</th>
</tr>
</thead>
<tbody>
<tr>
<td># of discrepancies</td>
<td>154</td>
<td>363</td>
<td>426</td>
<td>422</td>
<td>170</td>
</tr>
</tbody>
</table>
Neighborhood Analysis for Duplicates

17,000 compounds screened against five major CYP450 isozymes. 1,280 pairs of duplicates couples were found (874 had different bioprofiles)

<table>
<thead>
<tr>
<th>Tocris-0740</th>
<th>SID</th>
<th>Supplier</th>
<th>2C9</th>
<th>1A2</th>
<th>3A4</th>
<th>2D6</th>
<th>2C19</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID_6603937</td>
<td>11113673</td>
<td>Tocris</td>
<td>-4.6</td>
<td>-4.4</td>
<td>-4.6</td>
<td>-6.2</td>
<td>-4.5</td>
</tr>
<tr>
<td>CID_6603937</td>
<td>11111504</td>
<td>Sigma Aldrich</td>
<td>-4.4</td>
<td>-5.6</td>
<td>-5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 Nearest neighbors

<table>
<thead>
<tr>
<th>Tanimoto Similarity</th>
<th>SID</th>
<th>Supplier</th>
<th>2C9</th>
<th>1A2</th>
<th>3A4</th>
<th>2D6</th>
<th>2C19</th>
</tr>
</thead>
<tbody>
<tr>
<td>6604862</td>
<td>0.98</td>
<td>11114071</td>
<td>Tocris</td>
<td>-4.5</td>
<td></td>
<td>-5.5</td>
<td></td>
</tr>
<tr>
<td>6604106</td>
<td>0.98</td>
<td>11112029</td>
<td>Sigma Aldrich</td>
<td>-5.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6604846</td>
<td>0.98</td>
<td>11114012</td>
<td>Tocris</td>
<td>-4.5</td>
<td></td>
<td>-5.9</td>
<td></td>
</tr>
<tr>
<td>6604136</td>
<td>0.95</td>
<td>11112054</td>
<td>Sigma Aldrich</td>
<td>-4.8</td>
<td>-5.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6604137</td>
<td>0.95</td>
<td>11113764</td>
<td>Tocris</td>
<td>-4.4</td>
<td>-4.7</td>
<td>-4.5</td>
<td></td>
</tr>
</tbody>
</table>
Chemical/Biological data curation workflow

1. Chemical Curation
2. Duplicate Analysis
3. Analysis of intra- and inter-lab experimental variability
4. Exclusion of unreliable data sources
5. Detection and Verification of Activity Cliffs
6. Calculation and tuning of dataset modelability index
7. Consensus QSAR Predictions
8. Identification and correction of mislabelled compounds

Original Set

Curated Set

Error Rate

Dataset Size (# of Records)

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm;
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation if possible
- Should be added: data used for modeling should be carefully curated
## Table 1. Types of error in QSAR/QSPR development and use.

<table>
<thead>
<tr>
<th>No.</th>
<th>Type of error</th>
<th>Relevant OECD principle(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Failure to take account of data heterogeneity</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Use of inappropriate endpoint data</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Use of collinear descriptors</td>
<td>2, 4, 5</td>
</tr>
<tr>
<td>4</td>
<td>Use of incomprehensible descriptors</td>
<td>2, 5</td>
</tr>
<tr>
<td>5</td>
<td>Error in descriptor values</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Poor transferability of QSAR/QSPR</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>Inadequate/undefined applicability domain</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Unacknowledged omission of data points</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Use of inadequate data</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>Replication of compounds in dataset</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Too narrow a range of endpoint values</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>Over-fitting of data</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>Use of excessive numbers of descriptors in a QSAR/QSPR</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>Lack of/inadequate statistics</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>Incorrect calculation</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>Lack of descriptor auto-scaling</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>Misuse/misinterpretation of statistics</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>No consideration of distribution of residuals</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Inadequate training/test set selection</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>Failure to validate a QSAR/QSPR correctly</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>Lack of mechanistic interpretation</td>
<td>5</td>
</tr>
</tbody>
</table>
Model accuracy and interpretation: Case studies (modeling of skin sensitization and Ames genotoxicity)

- The Local Lymph Node Assay (LLNA) is generally regarded as the preferred test for evaluating skin sensitization.\(^1\)

- Although LLNA has a good correlation with human skin sensitization, it has been shown that LLNA fails in several cases to predict human skin sensitization.\(^2\)

- Ca. 3.89% (39,090) of the 1,004,873 animals used for safety testing in Europe are used in skin sensitization/irritation tests\(^2\); this creates a strong need to evaluate skin sensitization potential for a chemical without expensive and time-consuming animal testing.

*In silico* methods are highly recommended for time and cost saving of skin-related research.\(^4\)

---


\(^3\) European Commission. Seventh report on the statistics on the number of animals used for experimental and other scientific purposes in the member states of the 2013.

\(^4\) European Commission. On the animal testing and marketing ban and on the state of play in relation to alternative methods in the field of cosmetics 2013.
Model accuracy and interpretation: Case studies


• QSAR models of human data could replace mLLNA test for predicting human skin sensitization potential of chemicals (Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. *In preparation*).
Skin Sensitization Dataset (mLLNA)

**Source**

ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods) report 2009

<table>
<thead>
<tr>
<th>Vehicle type</th>
<th>Non-sensitizer</th>
<th>Sensitizer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>14</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>AOO</td>
<td>51</td>
<td>178</td>
<td>229</td>
</tr>
<tr>
<td>dH2O</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>DMF</td>
<td>40</td>
<td>27</td>
<td>67</td>
</tr>
<tr>
<td>DMSO</td>
<td>16</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>PG</td>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Pluronic L92 (1%)</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Others</td>
<td>4</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>135</strong></td>
<td><strong>273</strong></td>
<td><strong>408</strong></td>
</tr>
</tbody>
</table>

Abbreviations: AOO, acetone&olive oil (4:1 by volume); ACE, acetone; DMF, dimethyl formamide; DMSO, dimethyl sulfoxide; PG, propylene glycol.

254 compounds were retained for QSAR modeling:
127 non-sensitizers + 127 sensitizers
133 remaining sensitizers were used for additional external validation
QSAR models of skin sensitization (mLLNA)

Statistical characteristics of the models

254 compounds (127 sensitizers + 127 non-sensitizers)

Fair comparison with QSAR Toolbox

Showing results for 153 compounds
Not present in QSAR Toolbox DB

Models were built using Random Forest approach – 5-fold External CV results
### ALERTS vs. QSAR: ACTIVATED PYRIDINE/PYRIMIDINE

<table>
<thead>
<tr>
<th>Chemical Structure</th>
<th>QSAR ToolBox</th>
<th>QSAR</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethyl 2,6-dichloro-5-fluoro-b-oxo-3-pyridinepropanoate</td>
<td><img src="#" alt="Contains Activated Pyridine" /></td>
<td>Non Sensitizer</td>
<td>Non Sensitizer</td>
</tr>
<tr>
<td>N-(2-Chloro-4-pyrimidinyl)-N,2,3-trimethyl-2H-indazol-6-amine</td>
<td><img src="#" alt="Contains Activated Pyridine" /></td>
<td>Non Sensitizer</td>
<td>Non Sensitizer</td>
</tr>
<tr>
<td>N-(2-Chloro-4-pyrimidinyl)-2,3-dimethyl-2H-indazol-6-amine</td>
<td><img src="#" alt="Contains Activated Pyridine" /></td>
<td>Non Sensitizer</td>
<td>Non Sensitizer</td>
</tr>
</tbody>
</table>
### ALERTS vs. QSAR: NO PROTEIN BINDING ALERTS

<table>
<thead>
<tr>
<th>QSAR Toolbox</th>
<th>QSAR</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Chemical Structure" /></td>
<td>Non Sensitizer</td>
<td>Non Sensitizer</td>
</tr>
<tr>
<td><img src="image2.png" alt="Chemical Structure" /></td>
<td>Non Sensitizer</td>
<td>Non Sensitizer</td>
</tr>
<tr>
<td><img src="image3.png" alt="Chemical Structure" /></td>
<td>Sensitizer</td>
<td>Sensitizer</td>
</tr>
</tbody>
</table>

**1-[3,5-Bis(trifluoromethyl)phenyl]-N-methylethanamine**

**1-[3-(Cyclopentyloxy)-4-methoxy-phenyl]-4-oxocyclohexane carbonitrile**

**3-Aminomethyl-3,5,5-trimethylcyclohexyl amine**
Chemical Alerts (rules) of Toxicity: are they truly reliable?

Toxtree

Skin sensitisation reactivity domains

Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach.

Available since ToxTree 2.1.0 (under name "Skin sensitisation alerts" and "Skin sensitisation alerts (M.Cronin)"). The name is changed to "Skin sensitisation reactivity domain" by P&G team suggestion in order to reflect the fact the alerts provide grouping into reactivity mode of action and do not predict skin sensitisation potential.

Developed by IdeaConsult Ltd., (Sofia, Bulgaria), with collaboration with and support from Procter and Gamble © 2010
Chemical Alerts (rules) of Toxicity: are they truly reliable?
Model interpretation: identifying statistically important fragments as complex alerts

**Ames data set**
- 5,439 compounds
- 2,121 mutagenic
- 3,318 non-mutagenic

### Results from 5-fold external cross validation

<table>
<thead>
<tr>
<th></th>
<th>Full model (967 fragments)</th>
<th>Reduced model (76 fragments)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specificity</strong></td>
<td>0.92 ±0.009</td>
<td>0.92 ±0.009</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>0.78 ±0.005</td>
<td>0.81 ±0.005</td>
</tr>
<tr>
<td><strong>Balanced Accuracy</strong></td>
<td>0.85 ±0.005</td>
<td>0.87 ±0.005</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td>0.91 ±0.004</td>
<td>0.94 ±0.003</td>
</tr>
</tbody>
</table>

Slightly improved
Example of fragment (alert) interaction

Nitro’s mutagenic effect is:
- increased by furan (*synergism*)
- decreased by primary alkanes (*antagonism*)

### Synergistic interaction

\[
\text{N}^\circ\text{O}^\circ
\]

- 100% mutagenic
- 100:0

### Antagonistic interaction

\[
\text{C}^\circ\text{C}^\circ\text{N}^\circ\text{O}^\circ
\]

- 69% mutagenic
- 100:46

\[
\text{C} - \text{C} - \text{C} - \text{H}
\]

- 29% mutagenic
- 785:1884

\[
\text{C} - \text{C} - \text{C} - \text{H}
\]

- 94% mutagenic
- 79:5

\[
\text{O} = \text{N} = \text{O}
\]

- 84% mutagenic ("penetrance")
- 620:118

---

Number of mutagenic : non-mutagenic compounds

- 41
Nitro compounds are active when paired with aromatic rings inactive when paired with primary alkanes

Examples

- 645-12-5 5-nitro-2-furanoate  
  Mutagenic

- 5275-69-4 2-acetyl-5-nitrofuran  
  Mutagenic

- Nitroalkanes (primary)  
  Nitro(prop – hex)ane  
  Non-mutagenic

Mechanism

- aromatic nitro more likely to be bioactivated
- nitro reductase
- nitro radical
- nitroso

- aliphatic nitro less likely to be bioactivated

Benigni 2011 Chem Rev
Helguera 2006 Toxicol
McCalla 1983 Env Mutagen
Marrying SAR and QSAR in CWAS: Deriving alerts from validated QSAR models
Can models replace testing? Skin sensitization modeling of human data

**human** DSA$_{0.05}$ data: induction dose per skin area (DSA) that produces a positive response in 5% of the tested population using human maximization test (HMT) and the human repeat-insult patch test (HRIPT)

---

Comparison of external predictive accuracy for human data: QSAR gives more reliable predictions than mLLNA

Accessed by 5-fold external cross validation; SVM: Support Vector Machine; AD: Applicability Domain.
No. of compounds = 63 sensitizers + 46 non sensitizers
QSAR and toxicity testing in the 21st century

EPAs Contribution: The ToxCast Research Program

Transforming Environmental Health Protection

Francis S. Gilliam,*† George M. Gray,*† John R. Bucher,*†

In 2005, the U.S. Environmental Protection Agency (EPA), with support from the U.S. National Toxicology Program (NTP), funded a project at the National Research Council (NRC) to develop a long-range vision for toxicity testing and a strategic plan for implementing that vision. Both agencies wanted future toxicity testing and assessment paradigms to meet evolving regulatory needs. Challenges include the large numbers of substances that need to be tested and how to incorporate recent advances in molecular toxicology, computational sciences, and information technology to rely increasingly on human as opposed to animal data and to offer increased risk and reduced costs (1-3). In response, the Toxicity Testing Branch of the National Toxicology Program (NTP) reviewed identified key issues and implementation goals in the assessment and risk (4, 5, 6) and developed a suite of technologies that allows for high-throughput screening (HTS) and other automated screening assays into its testing program. In 2005, the EPA established the National Center for Computational Toxicology (NCCT). Through these initiatives, NTP and EPA, with the NCGC, are promoting the evolution of toxicology from a predominantly observational science at the level of disease-specific models in vivo to a predominantly predictive science focused on broad inclusion of target-specific, mechanism-based, biological observations in vivo (1, 4) (see figure below).

Toxicity pathways. In vitro and in vivo tools are being used to identify cellular responses after chemical exposure expected to result in adverse health effects (7). HTS methods are a primary means of discovery for drug development, and screening of >100,000 compounds per day is routine (8). However, drug discovery HTS methods traditionally test compounds at one concentration, usually between 2 and 10 μM, and tolerate high false-negative rates. In contrast, in the EPA, NCCT, and NTP combined effort, all compounds are tested at as many as 15 concentrations, generally ranging from ~5 nM to ~100 μM, to generate a concentration-response curve (9). This approach is highly reproducible, produces significantly lower false-positive and false-negative rates than the traditional HTS methods (9), and facilitates multispecies comparisons. Finally, an informatics platform has been built to compare results among HTS screens; this is being expanded to allow comparisons with historical toxicologic NTP and EPA data (http://narrative.nih.gov/pubchem). HTS data collected by EPA and NTP, as well as by the NCCT and other Molecular Libraries Initiative centers (http://nci.nih.gov), are being made publicly available through Web-based databases (e.g., PubChem (http://pubchem.ncbi.nlm.nih.gov)). In addition, we propose a shift from primarily in vivo animal studies to in vitro assays, in vivo assays with lower organisms, and computational modeling for toxicity assessments.

Toxicology

Transforming Environmental Health Protection

Francis S. Gilliam,*† George M. Gray,*† John R. Bucher,*†

In 2005, the U.S. Environmental Protection Agency (EPA), with support from the U.S. National Toxicology Program (NTP), funded a project at the National Research Council (NRC) to develop a long-range vision for toxicity testing and a strategic plan for implementing that vision. Both agencies wanted future toxicity testing and assessment paradigms to meet evolving regulatory needs. Challenges include the large numbers of substances that need to be tested and how to incorporate recent advances in molecular toxicology, computational sciences, and information technology to rely increasingly on human as opposed to animal data and to offer increased risk and reduced costs (1-3). In response, the Toxicity Testing Branch of the National Toxicology Program (NTP) reviewed identified key issues and implementation goals in the assessment and risk (4, 5, 6) and developed a suite of technologies that allows for high-throughput screening (HTS) and other automated screening assays into its testing program. In 2005, the EPA established the National Center for Computational Toxicology (NCCT). Through these initiatives, NTP and EPA, with the NCGC, are promoting the evolution of toxicology from a predominantly observational science at the level of disease-specific models in vivo to a predominantly predictive science focused on broad inclusion of target-specific, mechanism-based, biological observations in vivo (1, 4) (see figure below).

Toxicity pathways. In vitro and in vivo tools are being used to identify cellular responses after chemical exposure expected to result in adverse health effects (7). HTS methods are a primary means of discovery for drug development, and screening of >100,000 compounds per day is routine (8). However, drug discovery HTS methods traditionally test compounds at one concentration, usually between 2 and 10 μM, and tolerate high false-negative rates. In contrast, in the EPA, NCCT, and NTP combined effort, all compounds are tested at as many as 15 concentrations, generally ranging from ~5 nM to ~100 μM, to generate a concentration-response curve (9). This approach is highly reproducible, produces significantly lower false-positive and false-negative rates than the traditional HTS methods (9), and facilitates multispecies comparisons. Finally, an informatics platform has been built to compare results among HTS screens; this is being expanded to allow comparisons with historical toxicologic NTP and EPA data (http://narrative.nih.gov/pubchem). HTS data collected by EPA and NTP, as well as by the NCCT and other Molecular Libraries Initiative centers (http://nci.nih.gov), are being made publicly available through Web-based databases (e.g., PubChem (http://pubchem.ncbi.nlm.nih.gov)). In addition,
QSAR and Chemical Toxicity Testing in the 21 Century

in vitro testing  computational

$Thousands

HTS -omics

Bioinformatics/ Machine Learning

Cancer
ReproTox
DevTox
NeuroTox
PulmonaryTox
ImmunoTox

Slide courtesy of Dr. Ann Richard, EPA (modified)
Integration of Diverse Data Streams into QSAR Modeling to Improve Toxicity Prediction

**Cheminformatics**
- Over many chemicals

**Bioinformatics**
- Over many biological assays

**Chemical descriptors (in silico):**
- Molecular weight,
- Connectivity indices
- Presence/absence of fragment,
- Hydrophobicity, etc.

**Chemical-biological modeling**

**Human studies**
- Medical literature
- e-Health records
- Insurance claims

**Short-term biological assays**
- Transcriptomics,
- Metabolomics,
- Cytotoxicity,
- Genotype, etc.

**Toxicity**
QSAR modeling: chemical descriptors

High dimensional data, $X$

$$y = f(X)$$

Chemical descriptors

Bioassay data

Toxicity

Chemical 1 1
Chemical 2 0
Chemical 3 0
... ...
Chemical n 1

Response, $y$

**QSAR modeling: in vitro assay descriptors**

Machine learning

\[ y = f(X) \]

<table>
<thead>
<tr>
<th>Chemical 1</th>
<th>Chemical 2</th>
<th>Chemical 3</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>Chemical n</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>x2</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

High dimensional data, \( X \)

**QSAR modeling: hybrid descriptors**

Machine learning

\[ y = f(X) \]

Chemical descriptors

Bioassay data

High dimensional data, \( X \)

<table>
<thead>
<tr>
<th>Chemical 1</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>( x_z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical 1</td>
</tr>
<tr>
<td>Chemical 2</td>
</tr>
<tr>
<td>Chemical 3</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Chemical n</td>
</tr>
</tbody>
</table>

The Use of Biological Screening Data as Additional Biological Descriptors Improves the Prediction Accuracy of Conventional QSAR Models of Chemical Toxicity

Predicting Subchronic Hepatotoxicity from 24h Toxicogenomics Profiles

- Rats in triplicates, 6-8 weeks old, Sprague Dawley
- **Doses**: low, med, high
- **Time points**: 3h, 6h, 9h, 24h, 3, 7, 14 and 28 days
- Liver histopathology
- Clinical chemistry
- *In vivo* hepatic gene expression (24h, high dose)

127 compounds in 2 classes

- 58% Non-toxic
- 42% Toxic

Subchronic 28-day hepatotoxicity

Data source: Open TG-GATEs [http://toxico.nibio.go.jp/]
Conflicting Predictions by QSAR and Toxicogenomics Models

![Biological space and Chemical space plots]

**Carbamazepine**
- Distant biological neighbors
- Close chemical neighbors

=> Chemical similarity works better

**Caffeine**
- Close biological neighbors
- Distant chemical neighbors

=> TGx similarity works better

**Improved prediction:** Learn from both sets of neighbors
Chemical-biological read-across (CBRA): learning from both sets of neighbors

\[ A_{\text{pred}} = \text{similarity-weighted average of toxicity values} \]

overall correctly predicted as nontoxic

wrongly predicted as toxic

Biological neighbors (nearest on top)

Chemical neighbors

CARBAMAZEPINE
Non-toxic

Phenytoin
Non-toxic
0.813

Pemoline
Non-toxic
0.766

Phenylbutazone
Non-toxic
0.737

Phenobarbital
Non-toxic
0.721

Chemical-biological read-across (CBRA): learning from both sets of neighbors

\[ A_{\text{pred}} = \text{similarity-weighted average of toxicity values} \]

overall correctly predicted as nontoxic

CARBAMAZEPINE
Non-toxic

CBRA outperforms other models

- Single space approaches replicated previous results: TGx > hybrid > QSAR
- Multi-space kNN read-across, using both chemical and toxicogenomic neighbors, had the highest predictive power

<table>
<thead>
<tr>
<th>Model</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Balanced accuracy (CCR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical read-across</td>
<td>0.73 ± 0.07</td>
<td>0.34 ± 0.05</td>
<td>0.53 ± 0.04</td>
</tr>
<tr>
<td>Biological read-across</td>
<td>0.85 ± 0.07</td>
<td>0.66 ± 0.04</td>
<td>0.76 ± 0.04</td>
</tr>
<tr>
<td>Hybrid read-across</td>
<td>0.85 ± 0.07</td>
<td>0.58 ± 0.04</td>
<td>0.72 ± 0.04</td>
</tr>
<tr>
<td>Multi-space read-across</td>
<td>0.89 ± 0.07</td>
<td>0.66 ± 0.04</td>
<td><strong>0.78 ± 0.04</strong></td>
</tr>
</tbody>
</table>

Results of 5-fold external cross-validation

Radial Plots Visualize both Chemical and Biological Similarity to Help Forming the Read-across Argument

Conclusions and Outlook

• Rapid accumulation of large biomolecular datasets (especially, in public domain):
  – Strong need for both chemical and biological data curation
  – Cheminformatics approaches support biological data curation

• Novel approaches towards Integration of inherent chemical properties with short term biological profiles (biological descriptors)
  – improve the outcome of structure – in vitro – in vivo extrapolation

• Interpretation of significant chemical and biological descriptors emerging from externally validated models
  – inform the selection or design of effective and safe chemicals and focus the selection of assays/interpretation in terms of MoA

• Tool and data sharing
  – Public web portals (e.g., Chembench, OCHEM)
## Acknowledgments

**Principal Investigator**
Alexander Tropsha

**Research Professors**
Alexander Golbraikh, Denis Fourches (now at NCSU), Eugene Muratov

**Graduate students**
Yen Low (former, now at Netflix)
Vinicius Alves (UNC and UFG, Brazil)
Sherif Farag
Stephen Capuzzi

**Postdoctoral Fellows**
Alexander Isayev, Regina Politi

**Adjunct Members**
Weifan Zheng, Shubin Liu

**Collaborators**
Ivan Rusyn (UNC->Texas A&M)
Diane Pozefsky (UNC)
Judith Strickland (NIEHS/ILS)
Nicole Kleinstruer (NIEHS/ILS)
Carolina Andrade (UFG, Brazil)

## MAJOR FUNDING

**NIH**
- R01-GM66940
- R01-GM068665

**NSF**
- ABI 9179-1165

**EPA (STAR awards)**
- RD832720
- RD833825
- RD834999

ONR