UNC
ESHELMAN
SCHOOL OF PHARMACY

MML
UNC.EDU

Fundamentals of QSAR modeling: basic concepts and applications

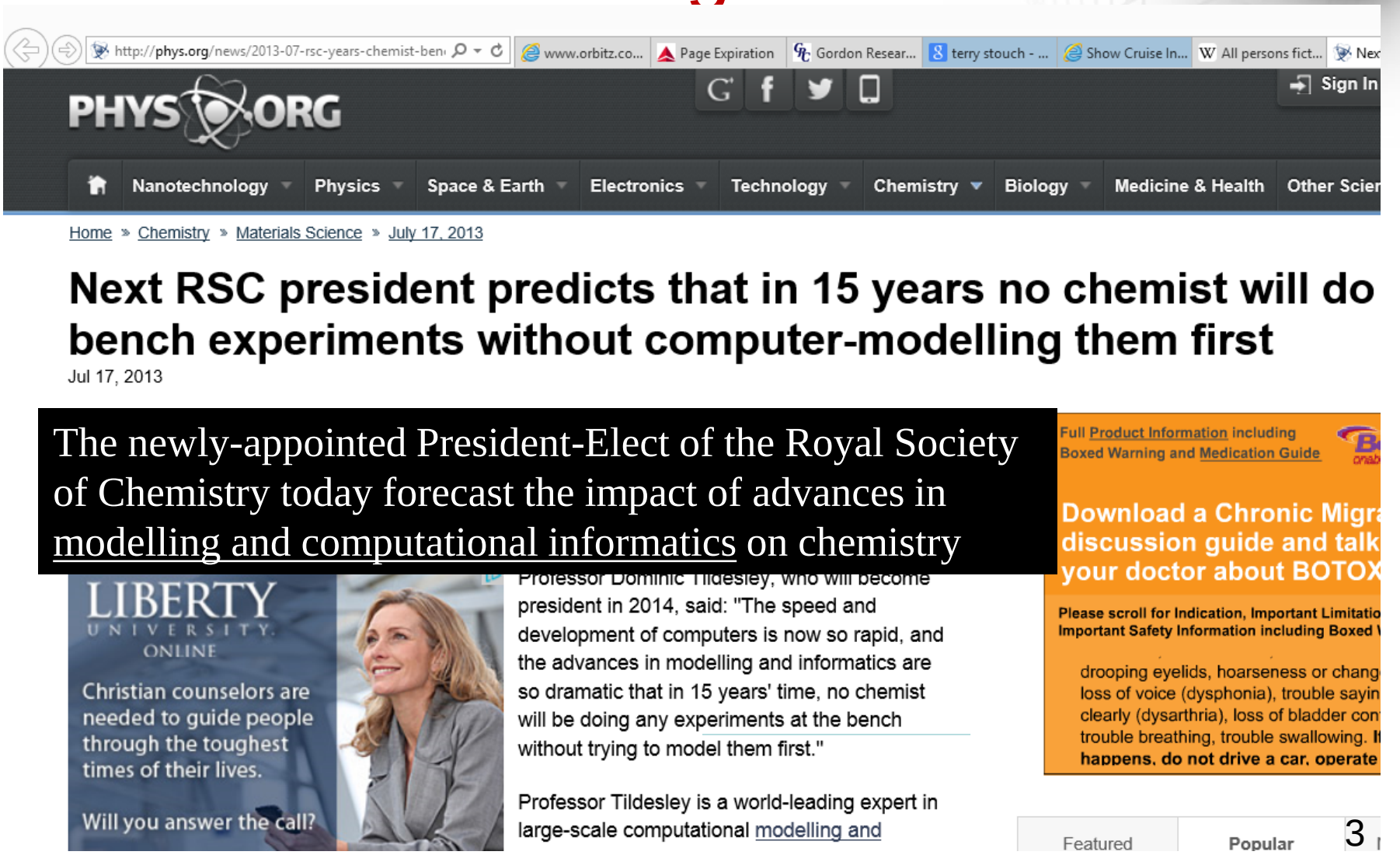
Alexander Tropsha

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



http://phys.org/news/2013-07-rsc-years-chemist-ben

PHYS.ORG

Sign In

Nanotechnology | Physics | Space & Earth | Electronics | Technology | Chemistry | Biology | Medicine & Health | Other Scier

Home » Chemistry » Materials Science » July 17, 2013

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

Jul 17, 2013

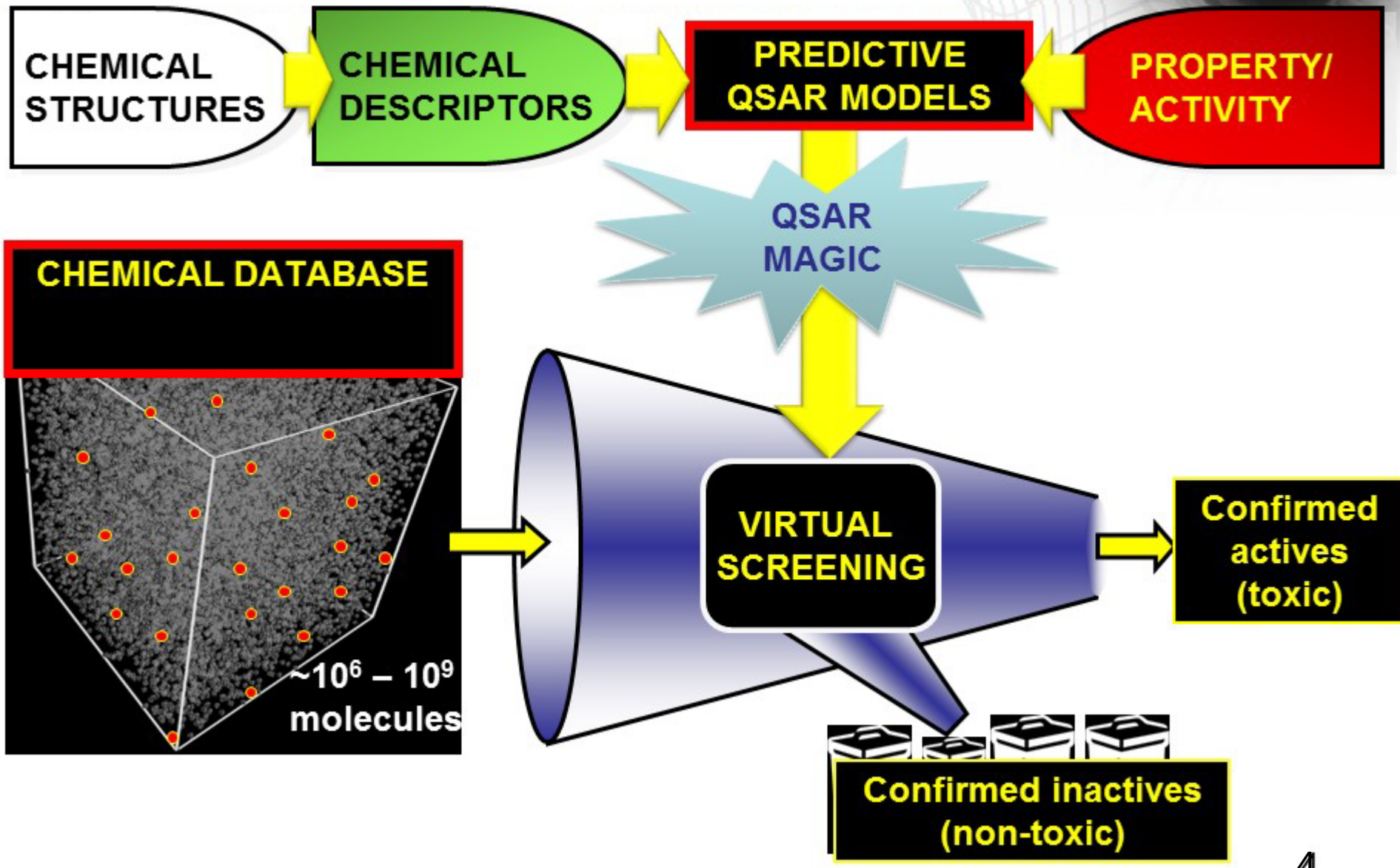
The newly-appointed President-Elect of the Royal Society of Chemistry today forecast the impact of advances in modelling and computational informatics on chemistry

Professor Dominic Tildesley, who will become president in 2014, said: "The speed and development of computers is now so rapid, and the advances in modelling and informatics are so dramatic that in 15 years' time, no chemist will be doing any experiments at the bench without trying to model them first."

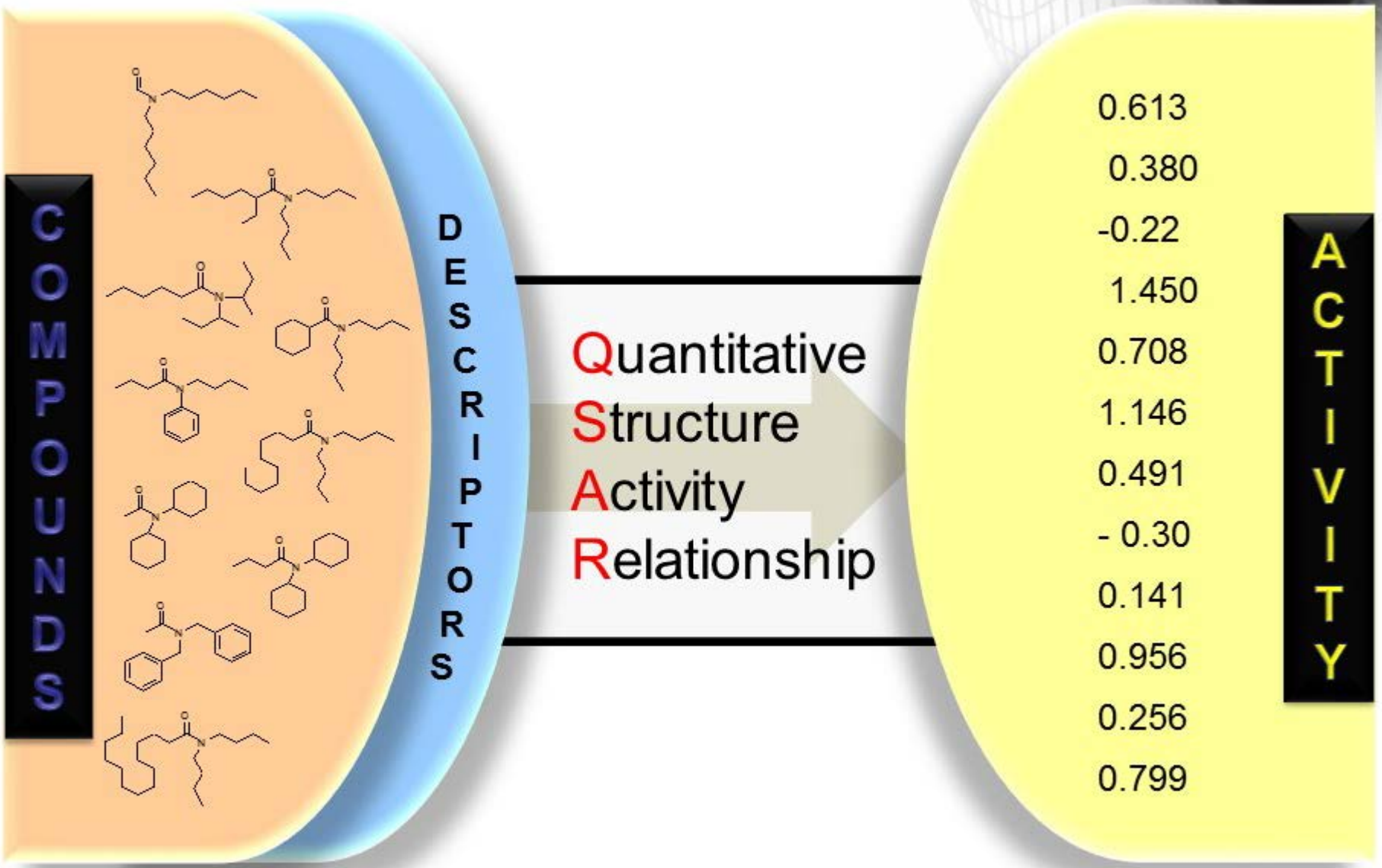
Professor Tildesley is a world-leading expert in large-scale computational modelling and

Featured | Popular

The chief utility of computational models: Hit identification in external libraries

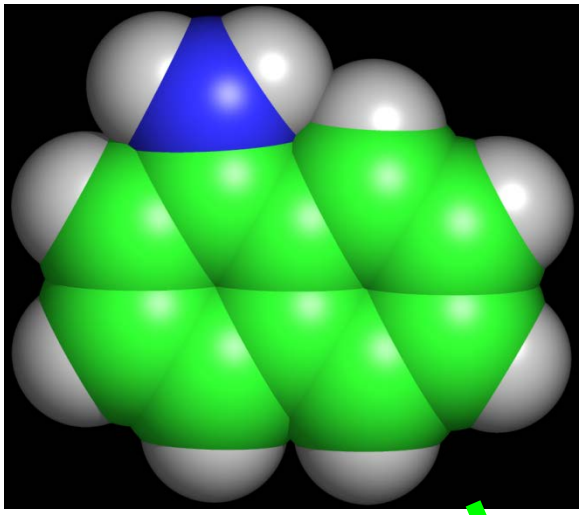


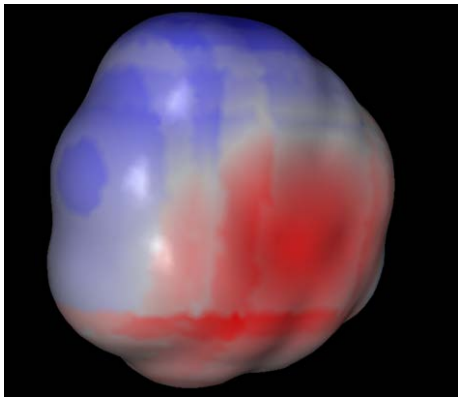
QSAR Modeling

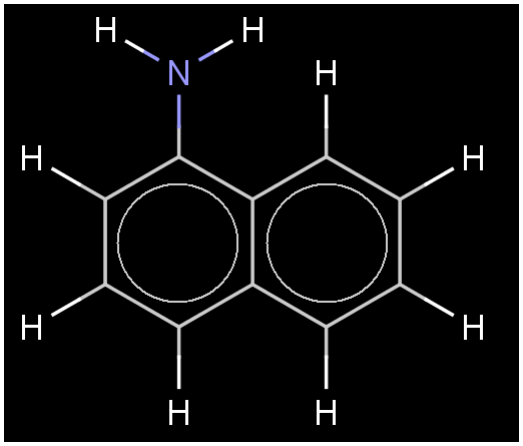


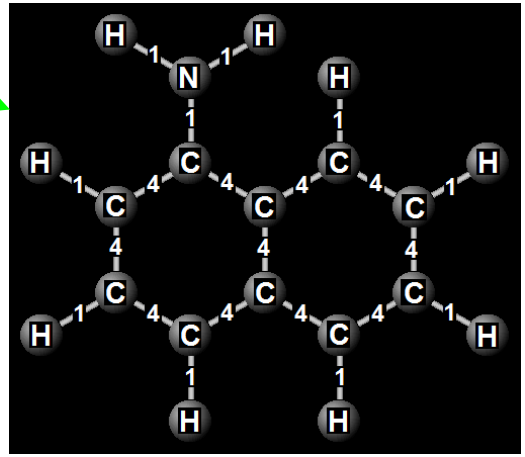
Slide credit: UNC MML

Structure representation

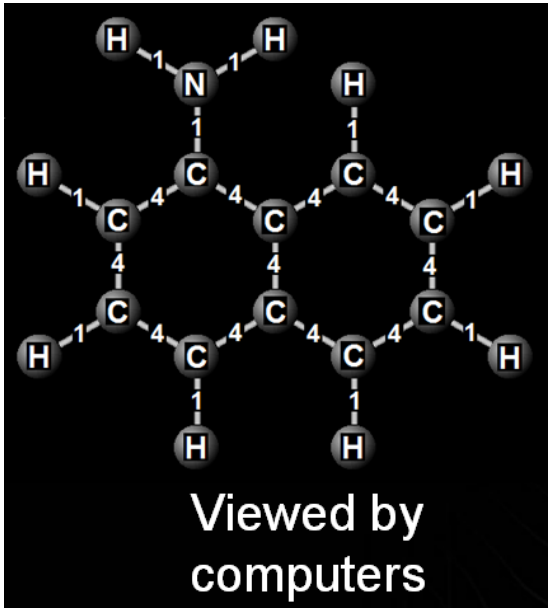




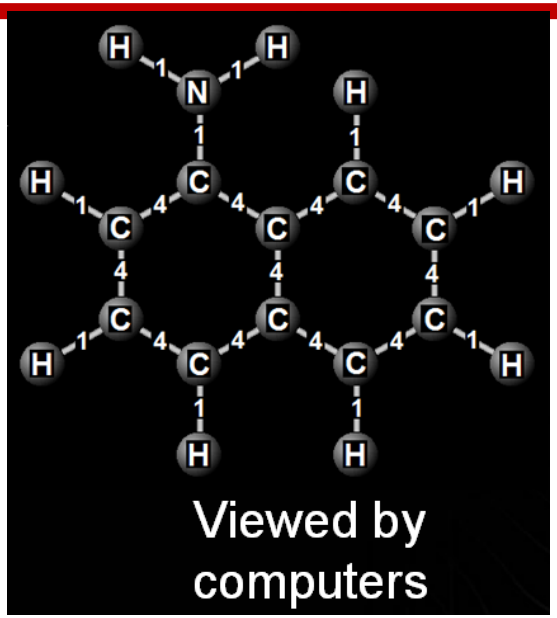




MML
UNC.EDU



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

MOL File

```
 16 15  0  0  0  0  0  0  0  0999 V2000
   -0.1958   -2.9667    0.0000 C   0  0  0
    0.5167   -2.5500    0.0000 C   0  0  0
    0.5125   -1.7250    0.0000 O   0  0  0
    1.2292   -2.9625    0.0000 N   0  0  3
    1.9417   -2.5458    0.0000 C   0  0  0
    2.6542   -2.9583    0.0000 C   0  0  0
    3.3667   -2.5417    0.0000 C   0  0  0
    4.0792   -2.9542    0.0000 C   0  0  0
    4.7917   -2.5375    0.0000 C   0  0  0
    5.5042   -2.9500    0.0000 C   0  0  0
    1.2250   -3.7875    0.0000 C   0  0  0
    0.8083   -4.5000    0.0000 C   0  0  0
    1.3917   -5.0833    0.0000 C   0  0  0
    0.9750   -5.7958    0.0000 C   0  0  0
    1.5583   -6.3792    0.0000 C   0  0  0
    0.9708   -6.9625    0.0000 C   0  0  0
  8  9  1  0  0  0  0
  4  5  1  0  0  0  0
  9 10  1  0  0  0  0
  2  3  2  0  0  0  0
  4 11  1  0  0  0  0
  5  6  1  0  0  0  0
 11 12  1  0  0  0  0
  1  2  1  0  0  0  0
 12 13  1  0  0  0  0
  6  7  1  0  0  0  0
 13 14  1  0  0  0  0
  2  4  1  0  0  0  0
 14 15  1  0  0  0  0
  7  8  1  0  0  0  0
 15 16  1  0  0  0  0
M  END
```

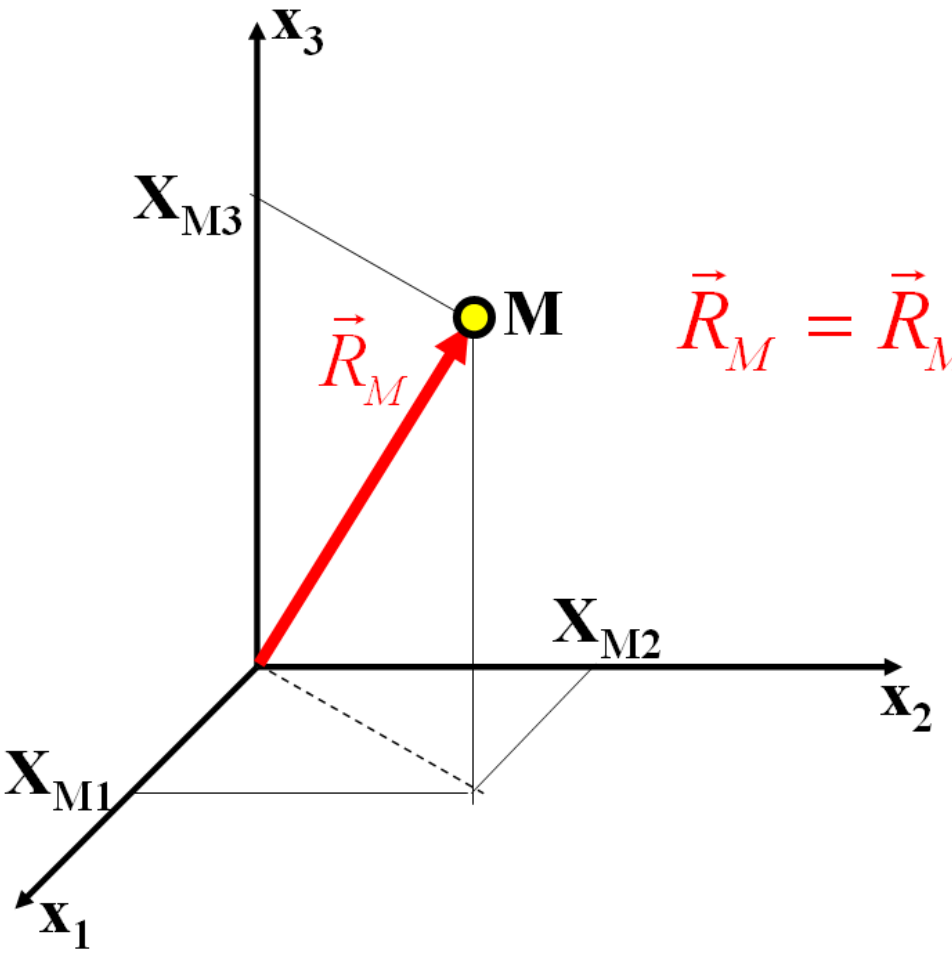
Vertices
(atomic type, coordinates etc.)

Edges
(connectivity table, label-types of bonds)

Datasets are represented by a matrix of molecular descriptors

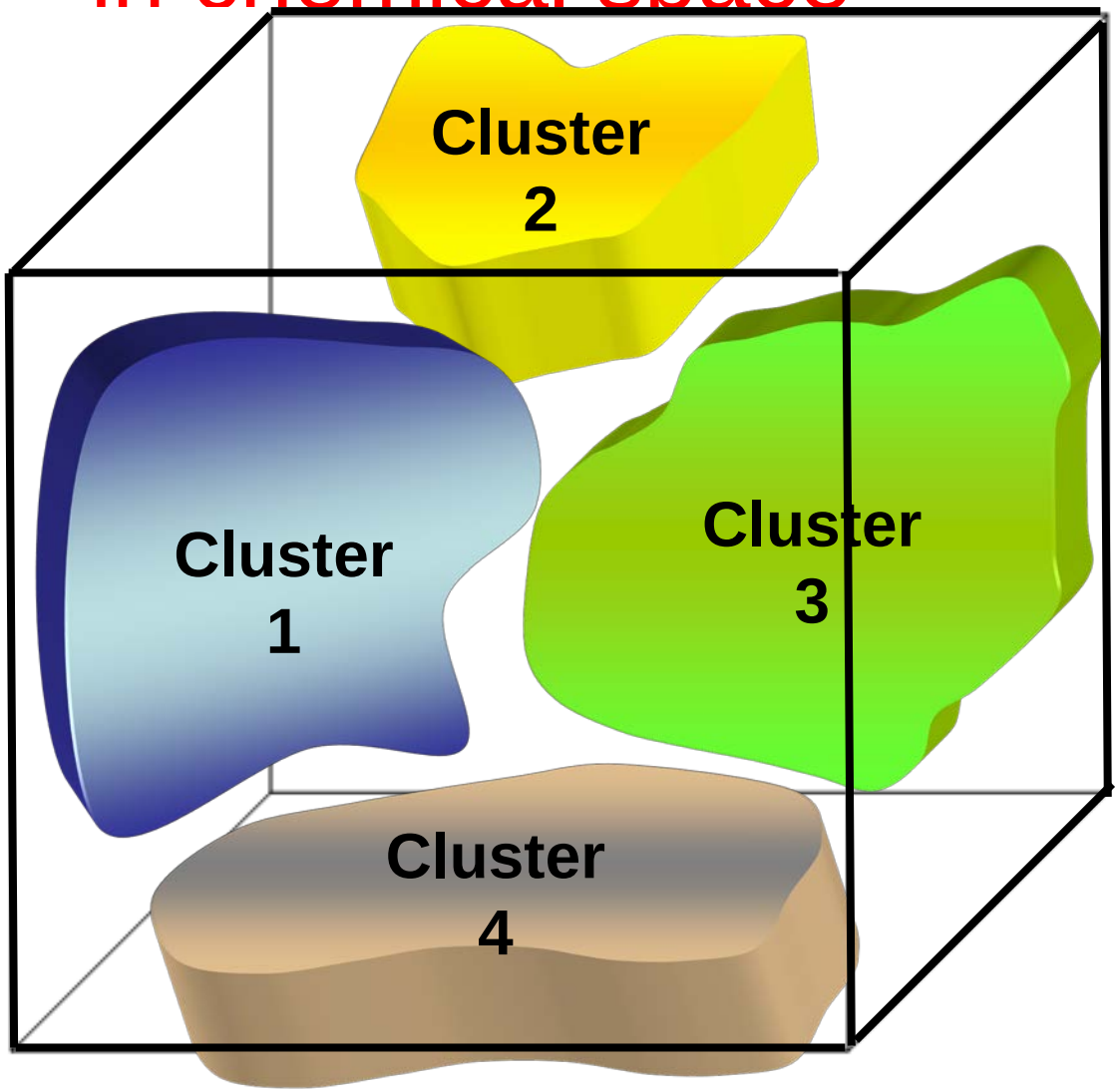
Samples (Compounds)	Variables (descriptors)			
	X_1	X_2	...	X_m
1	X_{11}	X_{12}	...	X_{1m}
2	X_{21}	X_{22}	...	X_{2m}
...	...	...	...	...
n	X_{n1}	X_{n2}	...	X_{nm}

Compounds represented by vectors in a multidimensional descriptor space



$$\vec{R}_M = \vec{R}_M(x_{M1}, x_{M2}, \ldots, x_{MK})$$

Molecules may form clusters in chemical space



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

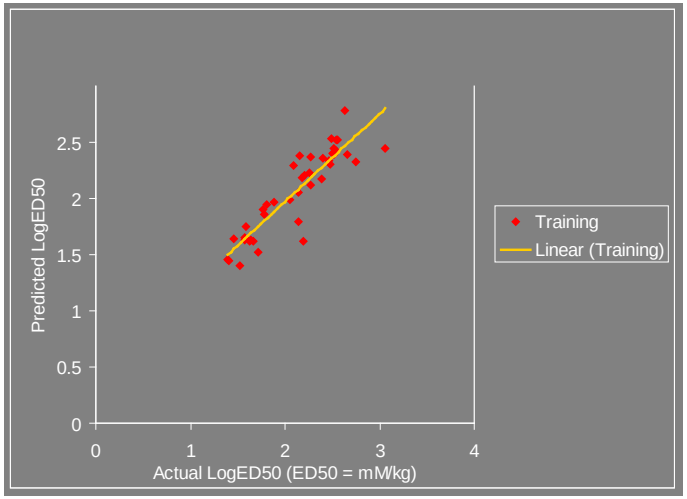
Establish quantitative relationships between descriptors and the target property capable of predicting activities of novel compounds.

Chemistry	Bioactivity (IC50, Kd...)	Cheminformatics (Molecular Descriptors)				
Comp.1	Value1	D_1	D_2	D_3	|	D_n
Comp.2	Value2	"	"	"	|	"
Comp.3	Value3	"	"	"	|	"
Comp.N	ValueN	"				

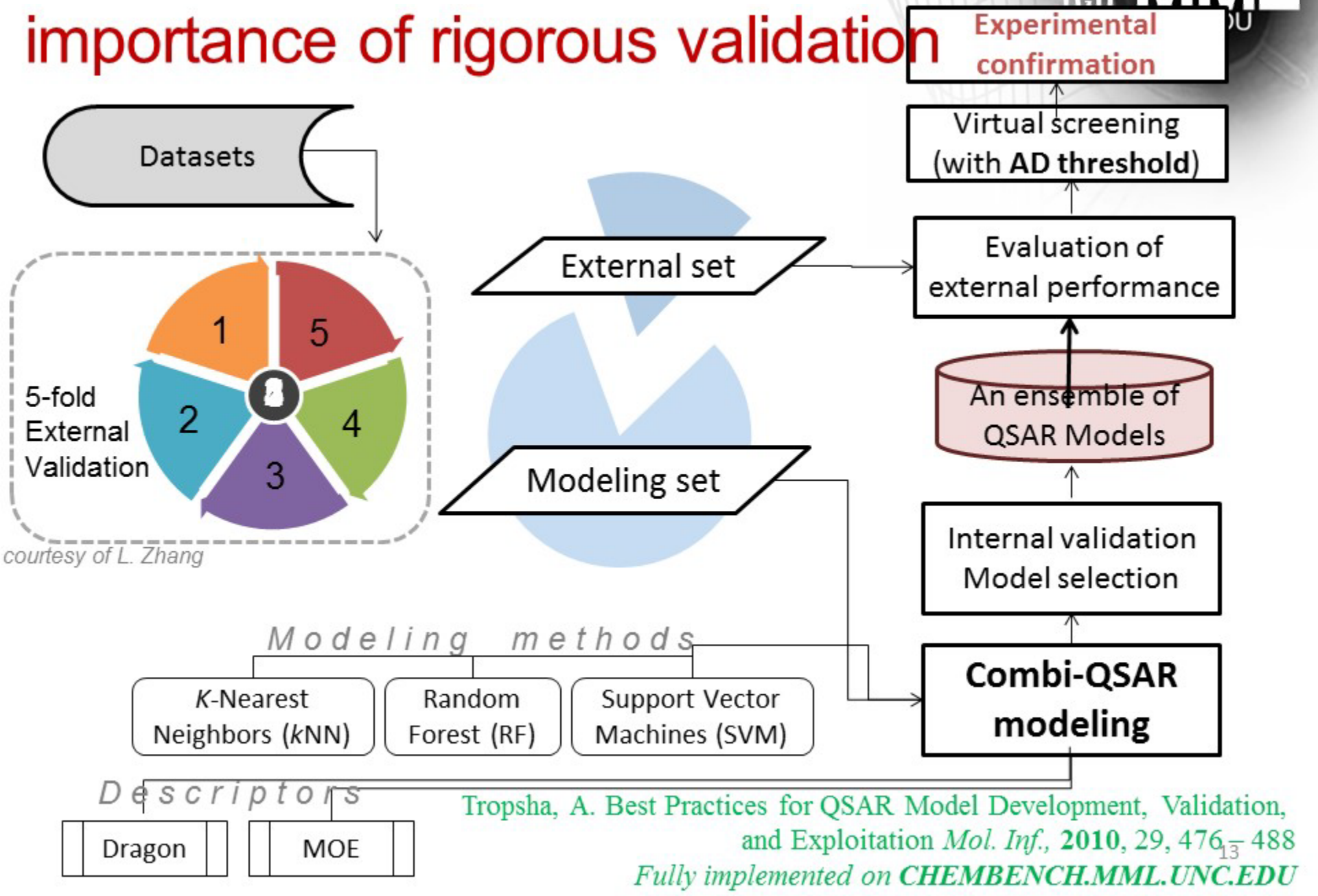
BA = F(D) (linear,

e.g., -LogIC50 = $k_1D_1+k_2D_2+\ldots+k_nD_n$)

or non-linear, e.g. k nearest neighbors



QSAR Modeling Workflow: the importance of rigorous validation



Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation *Mol. Inf.*, **2010**, 29, 476 – 488

Fully implemented on ***CHEMBENCH.MML.UNC.EDU***

Data dependency and data quality are critical issues in QSAR modeling

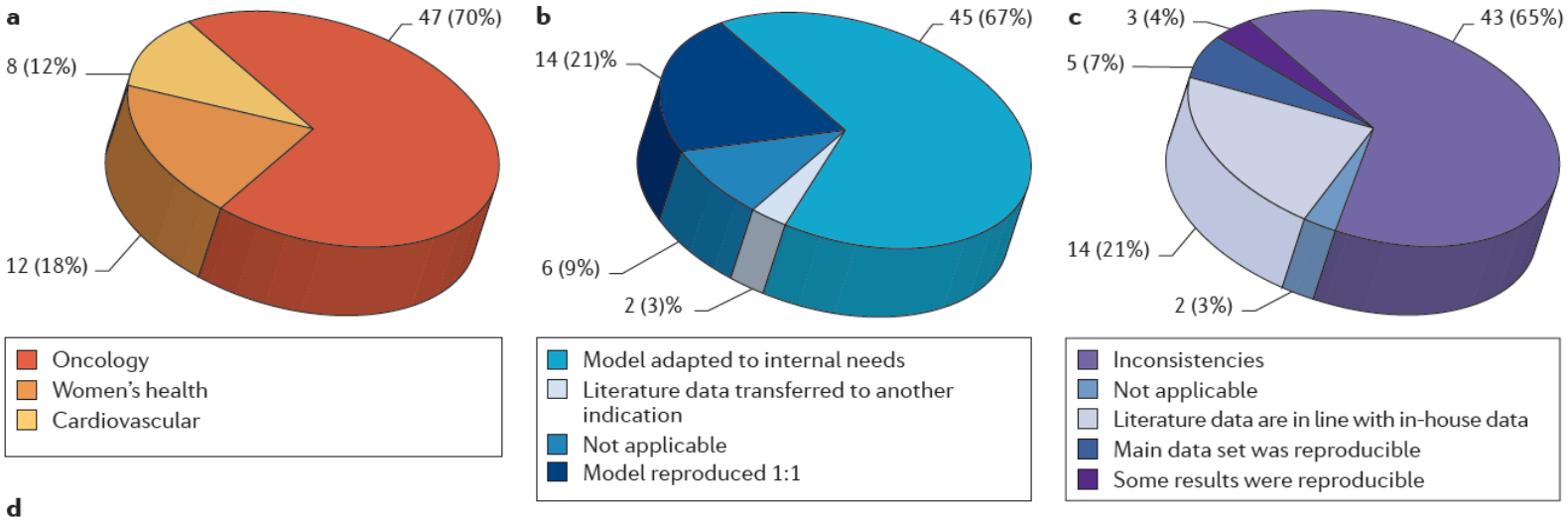
CORRESPONDENCE

LINK TO ORIGINAL ARTICL

Florian Prinz, Thomas Schlange and Khusru Asadullah. Nature Rev. Drug Disc. Sep 2011

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

results that are published are hard to repr
duce. However, there is an imbalance betwe
this apparently widespread impression and i
public recognition (for example, see REFS 2,3
and the surprisingly few scientific public
tions dealing with this topic. Indeed, to ou
knowledge, so far there has been no publishe
in-depth, systematic analysis that compar



d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

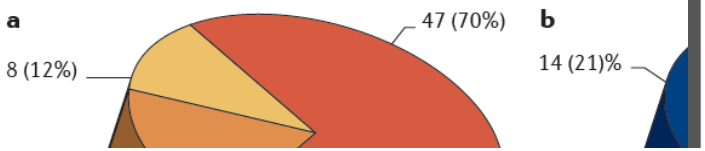
Data dependency and data quality are critical issues in QSAR modeling

CORRESPONDENCE

Florian Prinz, Thomas Schlange

Disc. Sep 2011

Believe it or not: how m
rely on published data
drug targets?



Drug Discovery Today • Volume 16, Numbers 17/18 • September 2011 EDITORIAL

ELSEVIER

editorial

Antony J. Williams

medicine and now drug repositioning or repurposing efforts. Their utility depends on the quality of the underlying molecular structures used. Unfortunately, the quality of much of the chemical structure-based data introduced to the public domain is poor. As an example we describe some of the errors found in the recently released NIH Chemical Genomics Center 'NPC browser' e as an example. There is an urgent need ernment funded data curation to improve lity of internet chemistry and to limit the ation of errors and wasted efforts.

Full Papers

QSAR & Combinatorial Science

QCS

Are the Chemical Structures in Your QSAR Correct?

Douglas Young[a]*, Todd Martin[a], Raghuraman Venkatapathy[b], and Paul Harten[a]

[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

ling agencies have been investing in the development of main chemistry platforms with the primary attention n to the informatics platform itself rather the quality of content. This is clearly exemplified by the recently PC browser from the NIH Chemical Genomics Center 1]. Public online databases such as PubChem, ChemID- d the EPA's ACToR [3], to name just a few, have rapidly usted valuable resources which researchers rely on for lable chemical structures and associated data. While emistry databases can certainly be of value, we feel the ould be immediately alerted to consider issues of data hen using these resources and we call into question both s and the trust we place in them. To our knowledge the raise, using the example of a recently released database, been described elsewhere and the user community, and gencies, should not ignore them any longer. The devel- f cheminformatics platforms without due care given to quality they contain, is a poor strategy for long term

In the last ten years, public online databases have

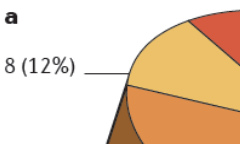
In the last decade numerous attempts have been made to

Data dependency and data quality are criti

MML UNC.EDU

CORRESP
Florian Pri
Disc. Sep

Believe
rely on
drug ta



Full Papers

Are the Chemi

Douglas Young[a]*, Todd Ma

[a] US Environmental Protectio
E-mail: young.douglas@epa.
[b] Pegasus Technical Services,

Keywords: Databases, *N*-octan
relationships, SMILES

Received: June 26, 2008; Revis

DOI: 10.1002/qsar.200810084

MyUNC : UNC Campus | Grade Roster | PLOS ONE: Dispensing Pro

v.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0062325

OPEN ACCESS PEER-REVIEWED

4,666 VIEWS

RESEARCH ARTICLE

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 IC_{50} 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.

In the last ten years, public online databases have

In the last decade numerous attempts have been made to

Data dependency and data quality are

MML UNC.EDU

ChemBark

News, Analysis, and Commentary for the World of Chemistry & Chemical Research

« Hacks for Septa

Organometallics Responds to the Dorta Situation »

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...

ChemBark

Investigates

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 an as 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?

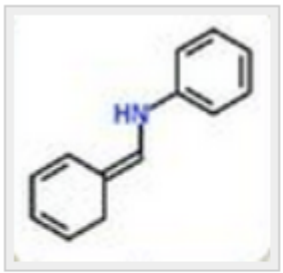
ChemBark is contacting the corresponding author for comment, and his response will be posted in full when we receive it.

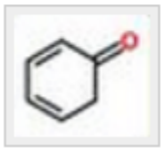
In the Pipeline

http://pipeline.corante.com/archives/2014/04/11/biology_maybe_right_chemistry_ridiculously_wrong.php

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

when we receive it.

NATURE | COMMENT

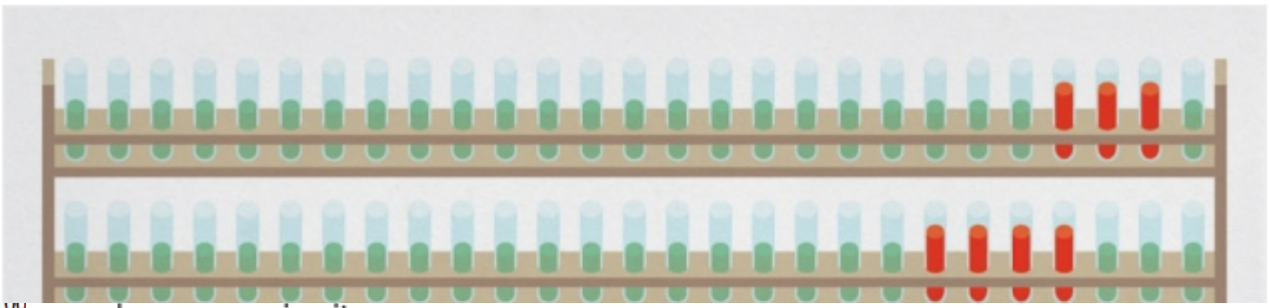
Policy: NIH plans to enhance reproducibility

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



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

FLICKR, UNIVERSITY OF EXETER

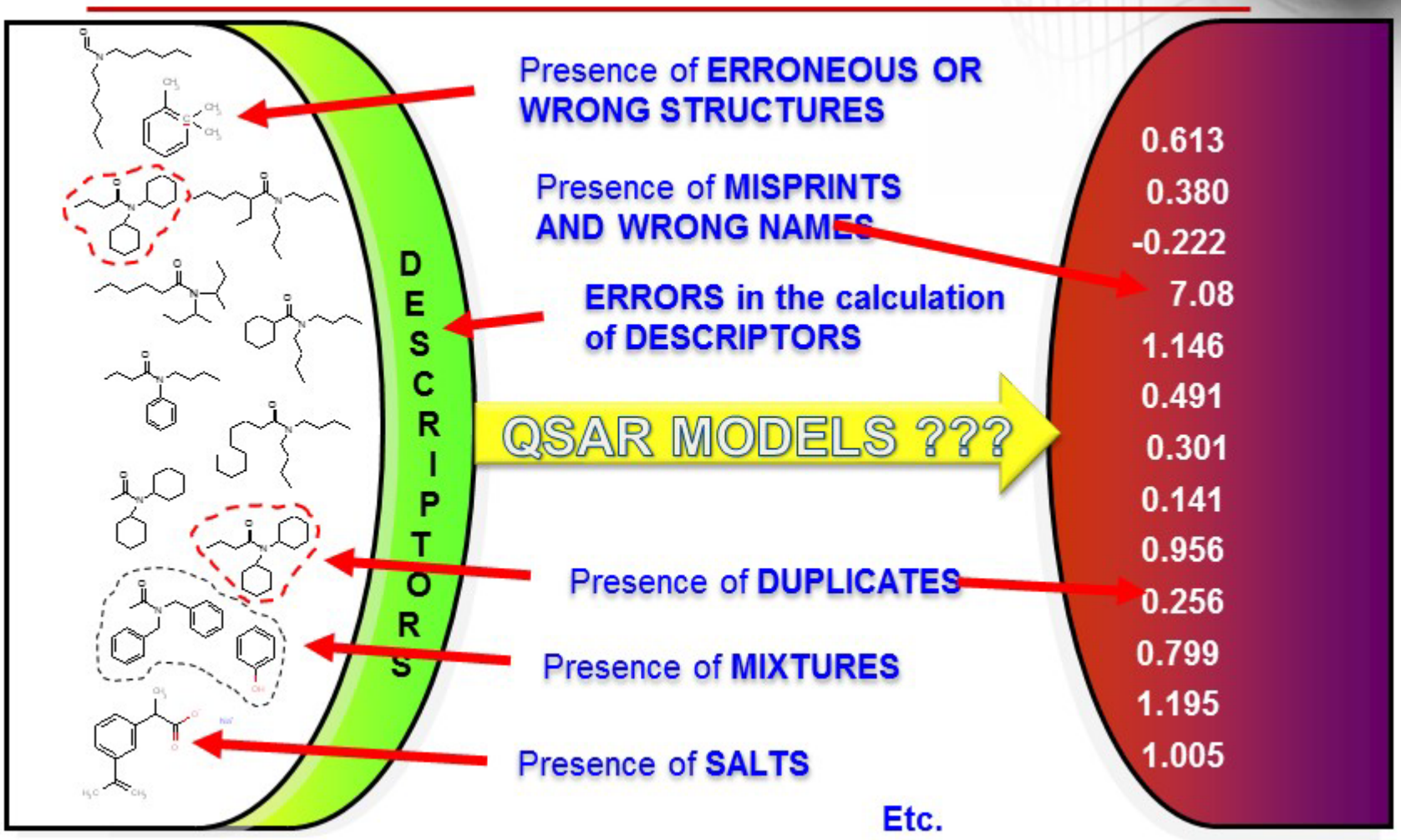
Recent years have seen increasing numbers of retractions, higher rates of misconduct and fraud, and general problems of data irreproducibility, spurring the National Institutes of Health (NIH) and others to launch initiatives to improve the quality of research results. Yesterday (April 7), at this year's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

when we receive it.

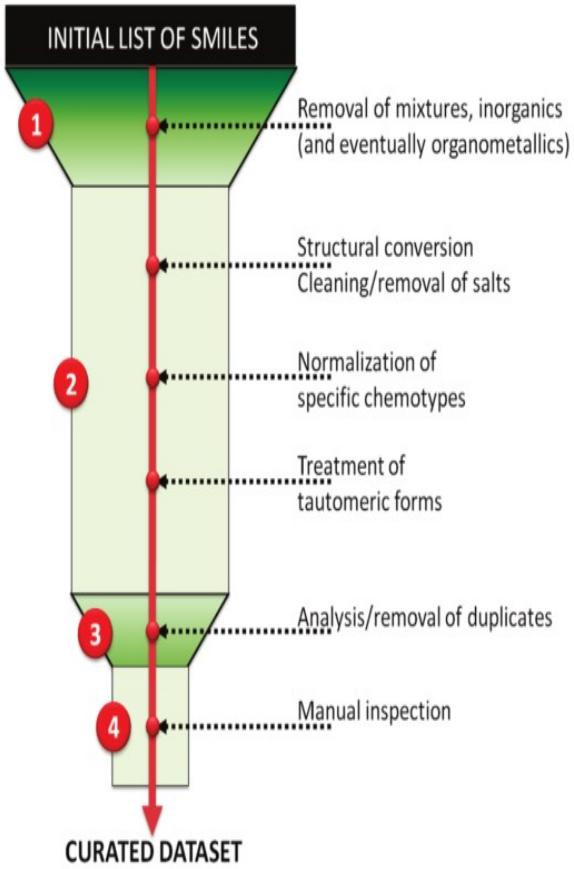
DOI: 10.100

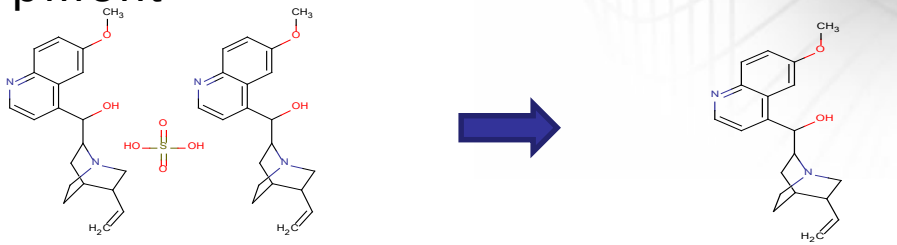
QSAR modeling with non-curated datasets



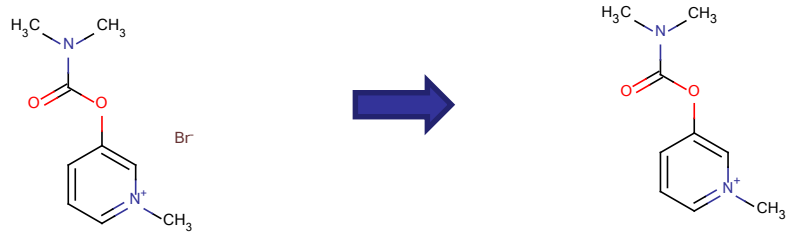
Chemical Structure Curation

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

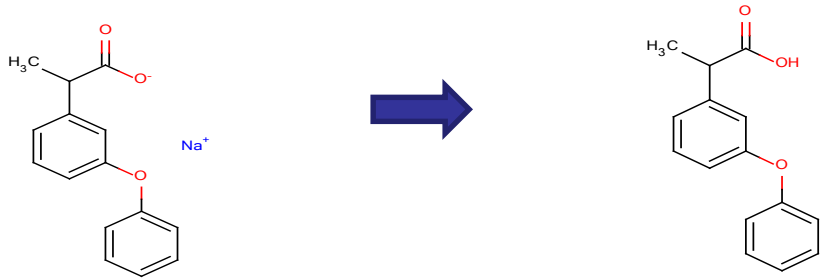




- Quinine sulfate dihydrate



- Pyridostigmine Bromide

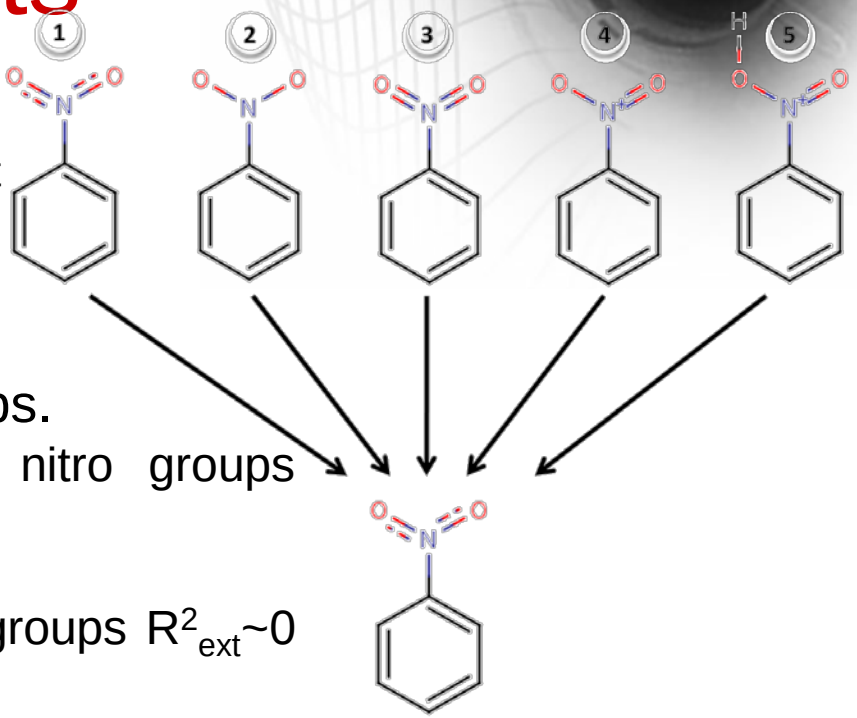


- Fenoprofen Sodium

Muratov, Fourches, Tropsha. Trust but verify. *JC J. Chem. Inf. Model*. **2010**, 50, 1189-1204.

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_{ext}~0.45 increased to R^2_{ext}~0.9.

-Case Study 2: after the normalization of nitro groups R^2_{ext}~0 increased to R^2_{ext}~0.5

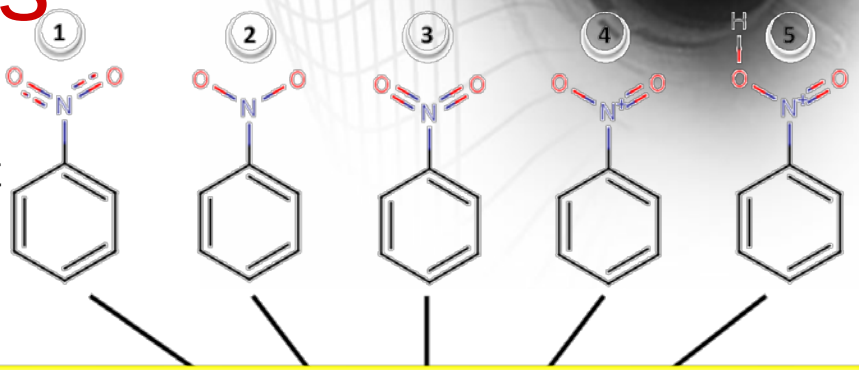
Even small differences in structure representation can lead to significant errors in prediction accuracy of models

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,

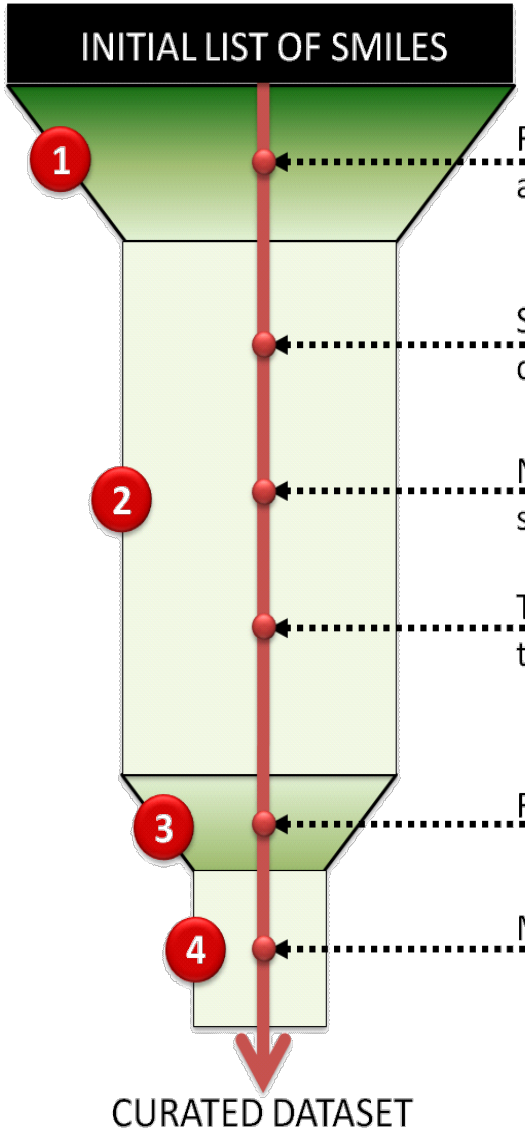
ABSTRACT: The human cytochrome P450 (CYP450) isozymes are the most important enzymes in the body to metabolize many endogenous and exogenous substances including environmental toxins and therapeutic drugs. Any unnecessary interactions between a small molecule and CYP450 isozymes may raise a potential 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 isoforms 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.

J. Chem. Inf. Model. 2011, 51, 2474–2481

Dataset Curation summary



17143 compounds

Removal of inorganics and mixtures — 17121 compounds

Structural conversion, cleaning of salts — 17121 compounds

Normalization of specific chemotypes — 17121 compounds

Treatment of tautomeric forms — 17121 compounds

Removal of duplicates — 16142 compounds

Manual inspection — **16142** compounds

Fourches D, et al. J Chem Inf Model. 2010 50(7):1189-204.

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

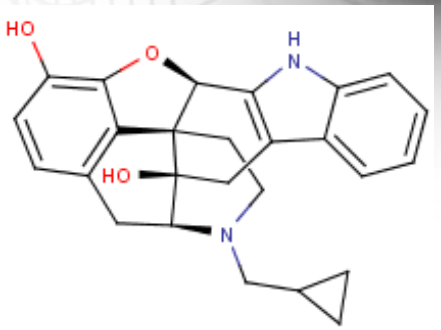
- A total of 1535 discrepancies were found in the

	CYP2C9	CYP1A2	CYP3A4	CYP2D6	CYP2C19
# of discrepancies	154	363	426	422	170

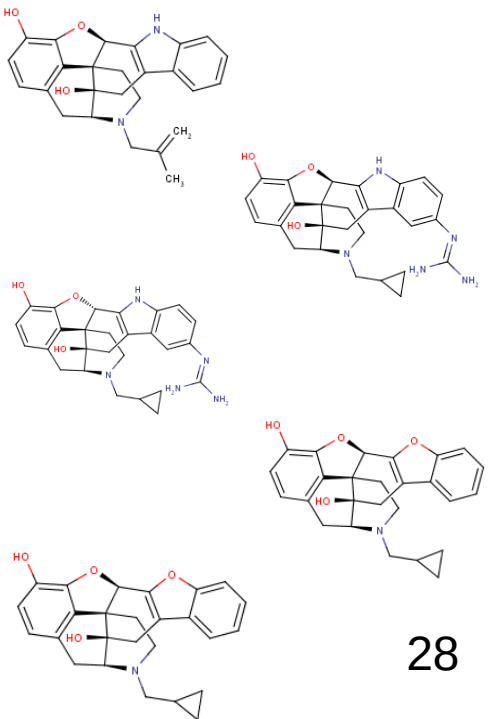
Neighborhood Analysis for Duplicates

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

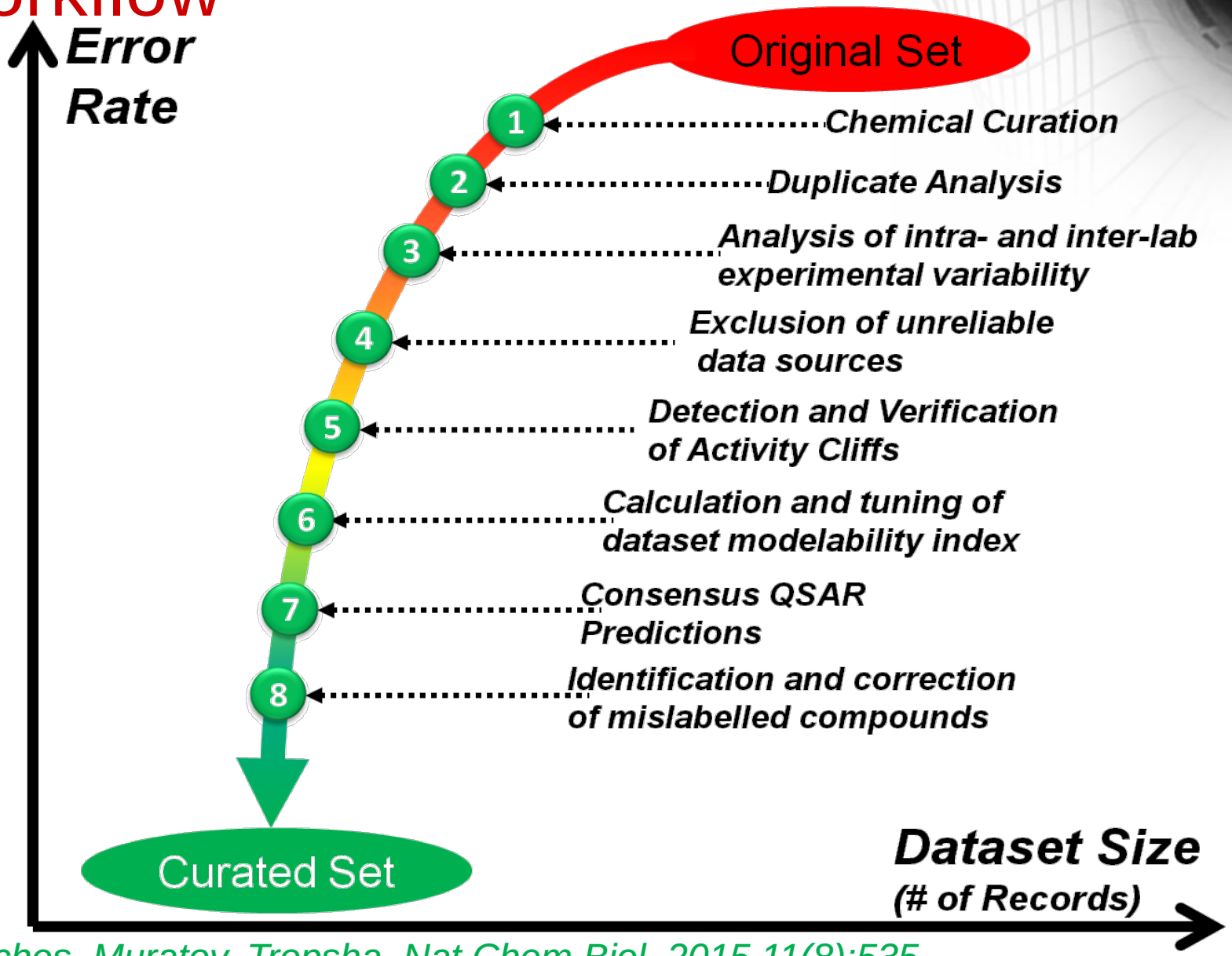
Tocris-0740	SID	Supplier	2C9	1A2	3A4	2D6	2C19
CID_6603937	11113673	Tocris	-4.6	-4.4	**-4.6**	-6.2	-4.5
CID_6603937	11111504	Sigma Aldrich	-4.4		**-8**	-5.6	-5



5 Nearest neighbors	Tanimoto Similarity	SID	Supplier	2C9	1A2	3A4	2D6	2C19
6604862	**0.98**	11114071	Tocris			**-4.5**		-5.5
6604106	**0.98**	11112029	Sigma Aldrich			**-5.1**		
6604846	**0.98**	11114012	Tocris					
6604136	**0.95**	11112054	Sigma Aldrich			**-4.8**	-5.9	
6604137	**0.95**	11113764	Tocris		-4.4	**-4.7**	-4.5	



Chemical/Biological data curation workflow



Fourches, Muratov, Tropsha. Nat Chem Biol. 2015,11(8):535.

Published guidance on model development and validation: The OECD Principles

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

21 “how not to do QSAR” principles

Table 1. Types of error in QSAR/QSPR development and use.

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

Dearden JC et al., **2009,** SAR and QSAR in Environmental Research, Vol. 20, Nos. 3–4, April–June 2009, 241

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]

[1]OECD. Test No. 429: Skin Sensitisation http://iccvam.niehs.nih.gov/SuppDocs/FedDocs/OECD/OECD-TG429-2010.pdf (accessed Jan 23, 2013).
[2]Api, A. M.; Basketter, D.; Lalko, J.; Basketter, D.; Lalko, J. *Cutan. Ocul. Toxicol.* **2014**, *9527*, 1–5.
[2]European Commission. Seventh teport 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 skin sensitization and their application to identify potentially hazardous compounds (Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. *Toxicol Appl Pharmacol*. **2015** 284(2):262-72)
- QSAR models of skin permeability and the relationships between skin permeability and skin sensitization (Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. *Toxicol Appl Pharmacol*. **2015** 284(2):273-80)
- 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)

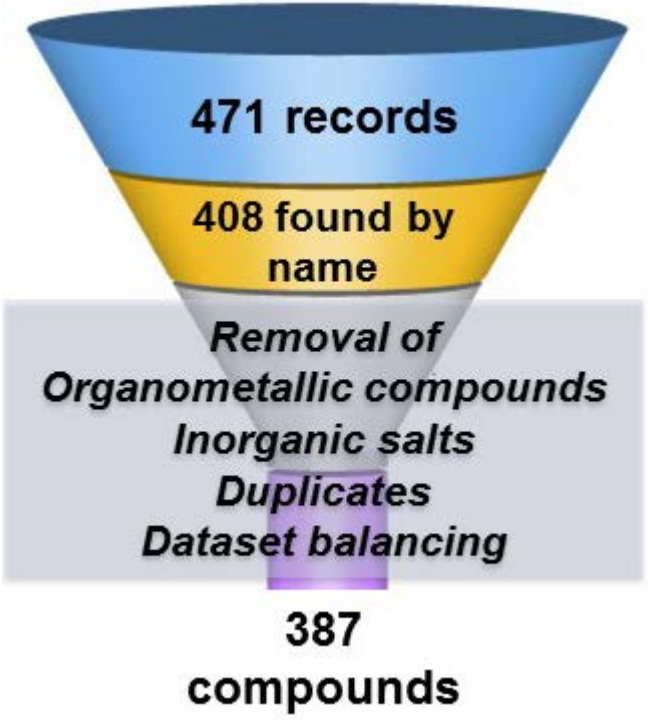
MML UNC.EDU

Source

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

Vehicle type	Non-sensitizer	Sensitizer	Total
ACE	14	31	45
AOO	51	178	229
dH_2O	2	2	4
DMF	40	27	67
DMSO	16	15	31
PG	6	8	14
Pluronic L92 (1%)	2	5	7
Others	4	7	11
Total	**135**	**273**	**408**

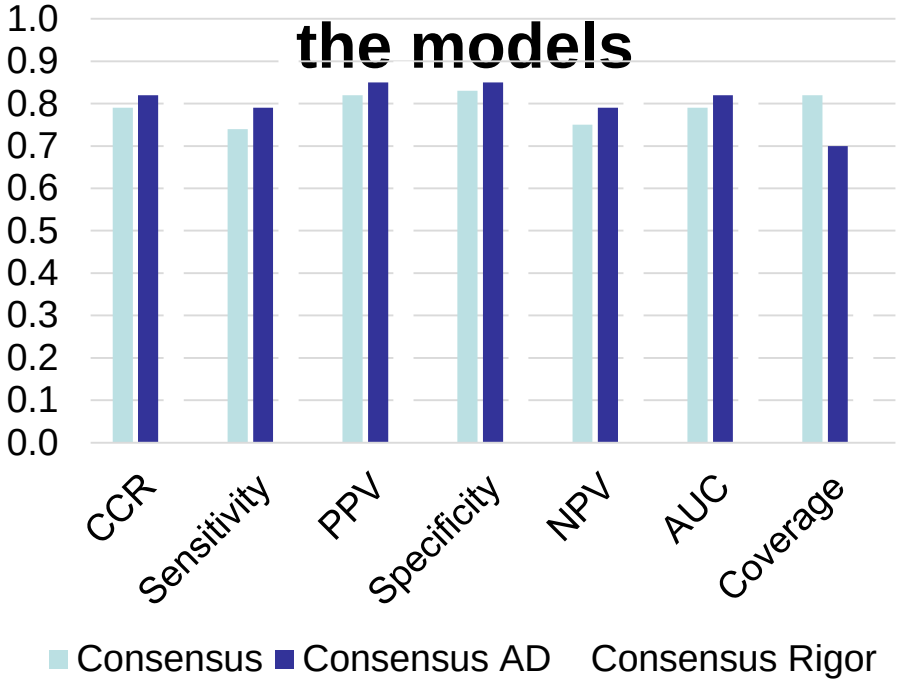
Abbreviations: AOO, acetone&olive oil (4:1 by volume); ACE, acetone; DMF, dimethyl formamide; DMSO, dimethyl sulfoxide; PG, propylene glycol.



254 compounds were retained for QSAR modeling:
127 non-sensitizers + 127 sensitizers
133 remaining sensitizers were used for additional external validation

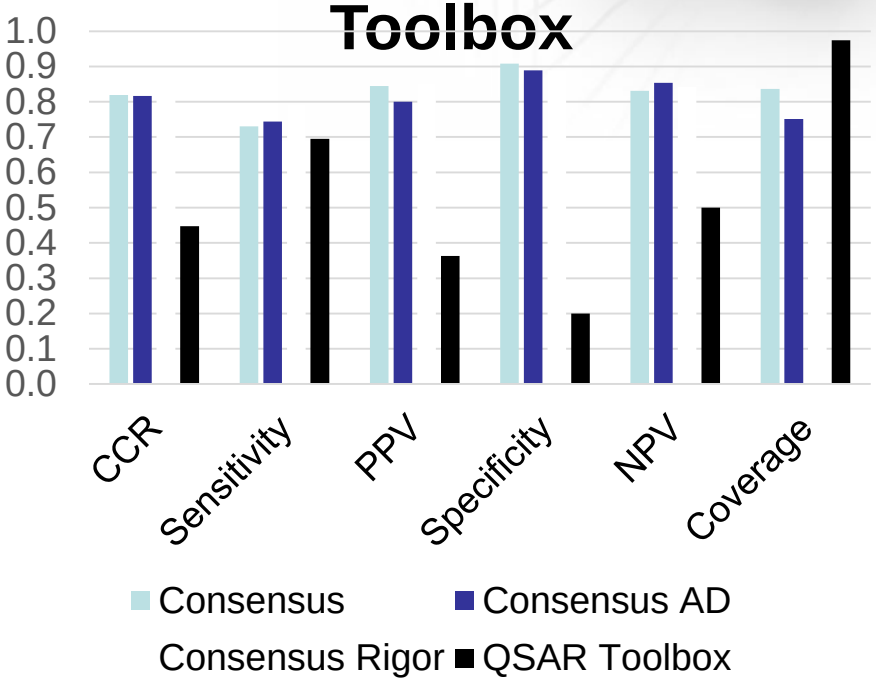
QSAR models of skin sensitization (mLLNA)

Statistical characteristics of the models



254 compounds (127 sensitizers + 127 non-sensitizers)

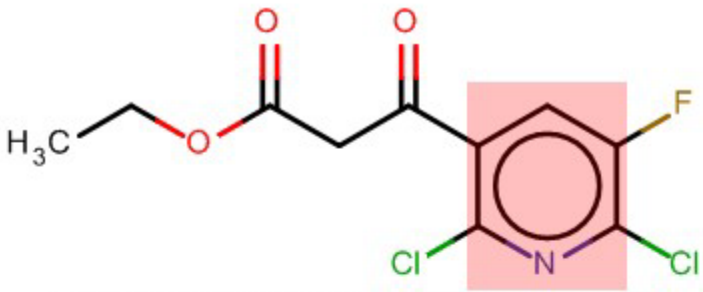
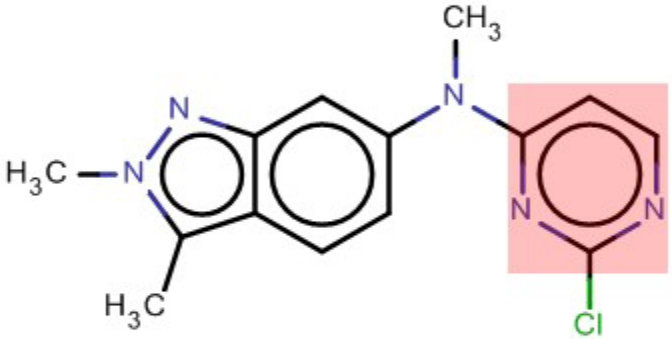
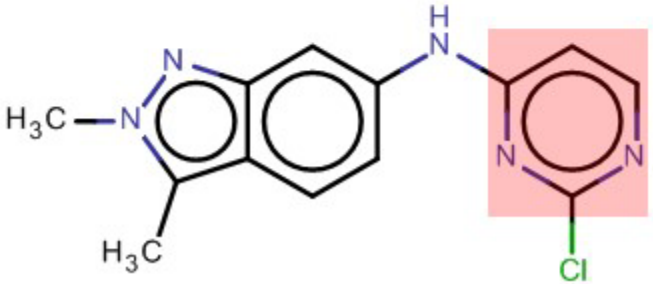
Fair comparison with QSAR Toolbox



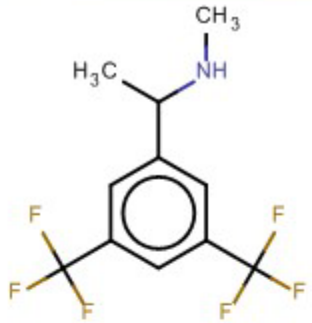
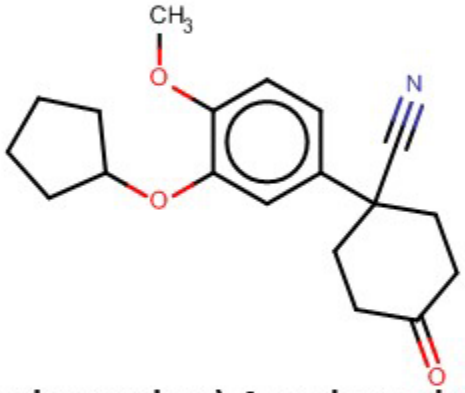
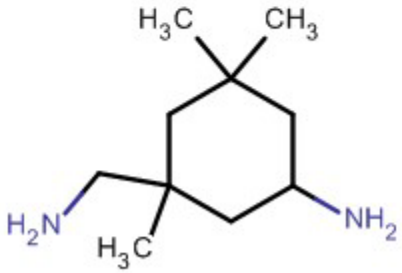
Showing results for 153 compounds Not present in QSAR Toolbox DB

Models were built using **Random Forest** approach – 5-fold External CV results

ALERTS vs. QSAR: ACTIVATED PYRIDINE/PYRIMIDINE

	QSAR Toolbox	QSAR	Experiment
 Ethyl 2,6-dichloro-5-fluoro-b-oxo-3-pyridinepropanoate	✓ Contains Activated Pyridine → Sensitizer	Non Sensitizer	Non Sensitizer
 N-(2-Chloro-4-pyrimidinyl)-N,2,3-trimethyl-2H-indazol-6-amine	✓ Contains Activated Pyridine → Sensitizer	Non Sensitizer	Non Sensitizer
 N-(2-Chloro-4-pyrimidinyl)-2,3-dimethyl-2H-indazol-6-amine	✓ Contains Activated Pyridine → Sensitizer	Non Sensitizer	Non Sensitizer

ALERTS vs. QSAR: NO PROTEIN BINDING ALERTS

	QSAR Toolbox	QSAR	Experiment
 1-[3,5-Bis(trifluoromethyl)phenyl]-N-methylethanamine	No alert → Sensitizer	Non Sensitizer	Non Sensitizer
 1-[3-(Cyclopentyloxy)-4-methoxy-phenyl]-4-oxocyclohexane carbonitrile	No alert → Sensitizer	Non Sensitizer	Non Sensitizer
 3-Aminomethyl- 3,5,5-trimethylcyclohexyl amine	No alert → Non sensitizer	Sensitizer	Sensitizer

Chemical Alerts (rules) of Toxicity: are they truly reliable?

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

967 fragments → 76 fragments

Data preprocessing → Build random forests (RF) → Rank fragments by RF importance → Rebuild RF models with top fragments

Chemical curation
Remove invariant, highly correlated fragments

	Full model (967 fragments)		Reduced model (76 fragments)
Specificity	0.92 ±0.009		0.92 ±0.009
Sensitivity	0.78 ±0.005	Slightly improved	0.81 ±0.005
Balanced Accuracy	0.85 ±0.005		0.87 ±0.005
AUC	0.91 ±0.004		0.94 ±0.003

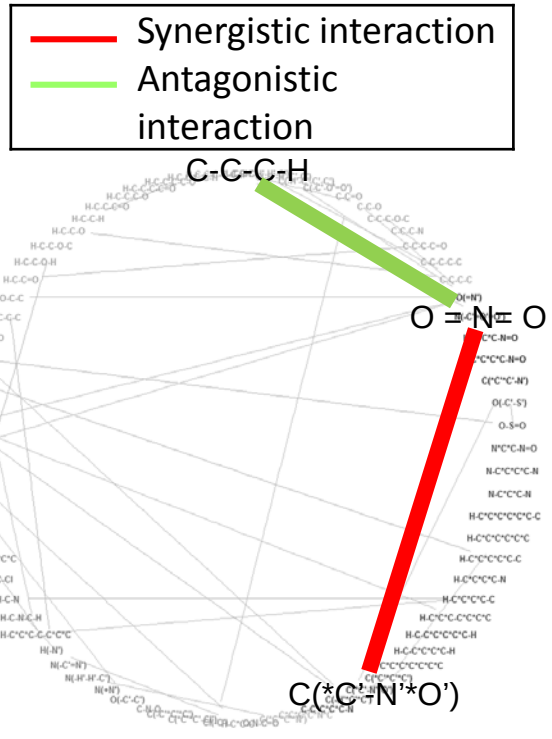
Results from 5-fold external cross validation

Example of fragment (alert) interaction

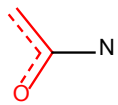
Nitro's mutagenic effect is:

increased by furan (synergism)

decreased by primary alkanes(antagonism)

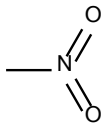


Synergistic influence

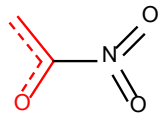


94% mutagenic
79:5

+



84% mutagenic ("penetrance")
620:118



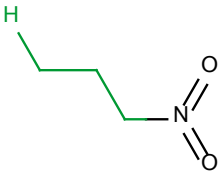
100% mutagenic
79:0

+

C – C – C – H

29% mutagenic
785:1884

Antagonistic influence

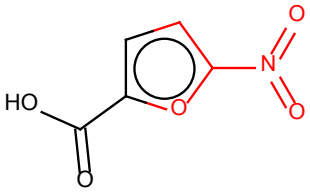


69% mutagenic
100:46

Number of mutagenic compounds : Number of non-mutagenic compounds

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

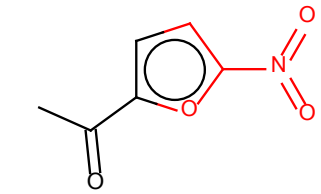
Examples



645-12-5

5-nitro-2-furanoate

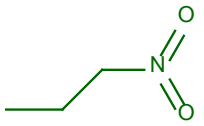
Mutagenic

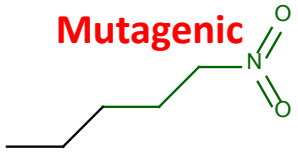


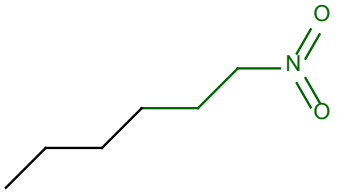
5275-69-4

2-acetyl-5-nitrofuran

Mutagenic







nitroalkanes (primary)

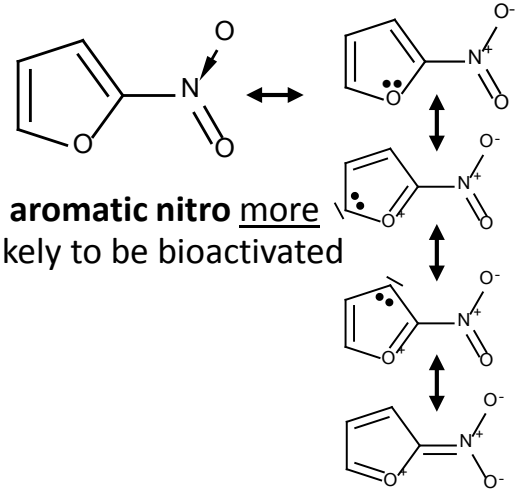
Nitro(prop – hex)ane

Non-mutagenic

Mechanism

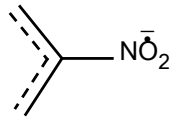
multiple resonance forms likely to be reduced

reactive metabolites

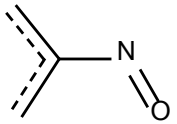


aromatic nitro more likely to be bioactivated

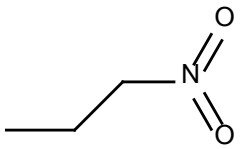
nitro reductase



nitro radical



nitroso



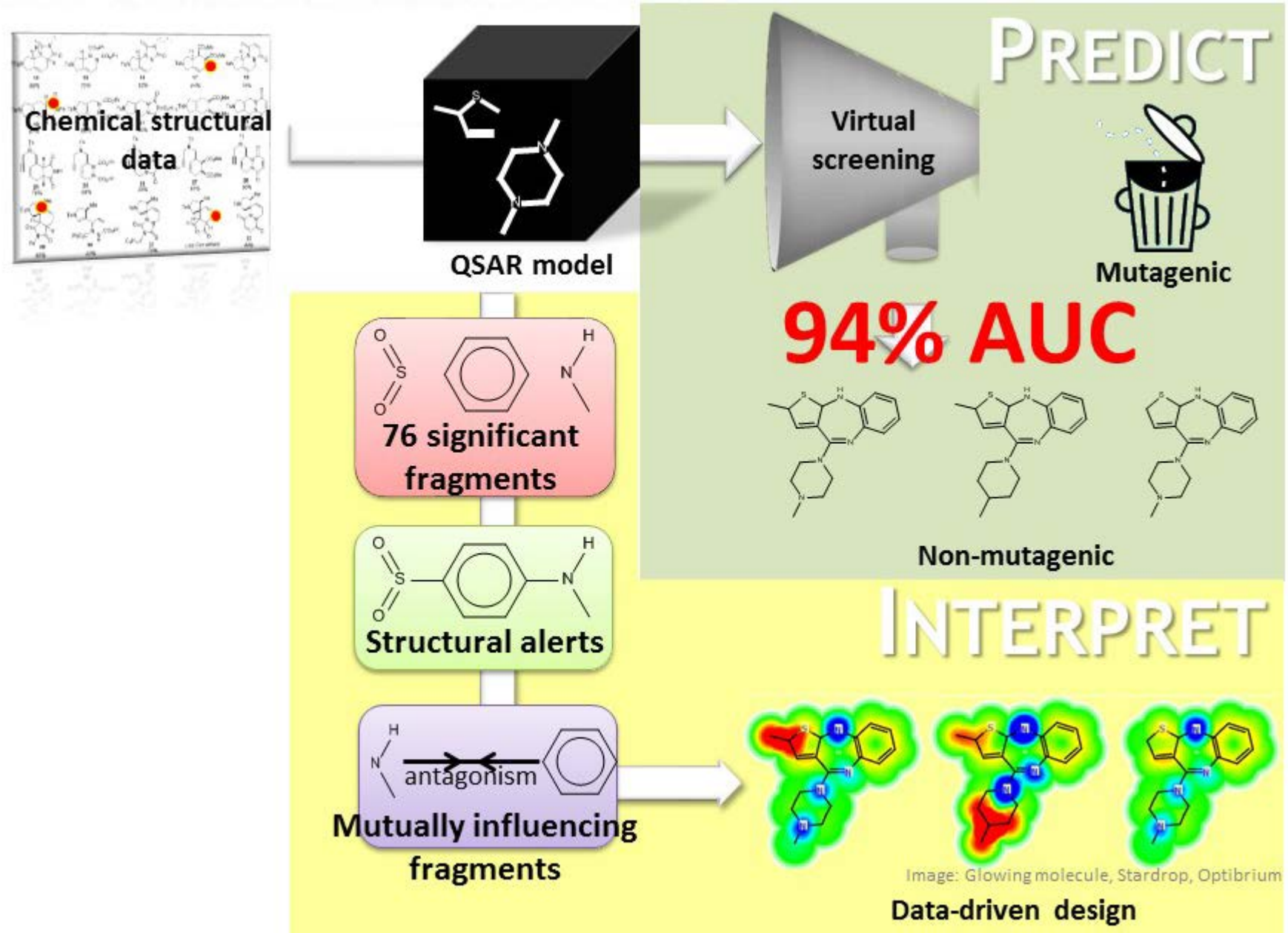
aliphatic nitro less likely to be bioactivated

Benigni 2011 *Chem Rev*

Helguera 2006 *Toxicol*

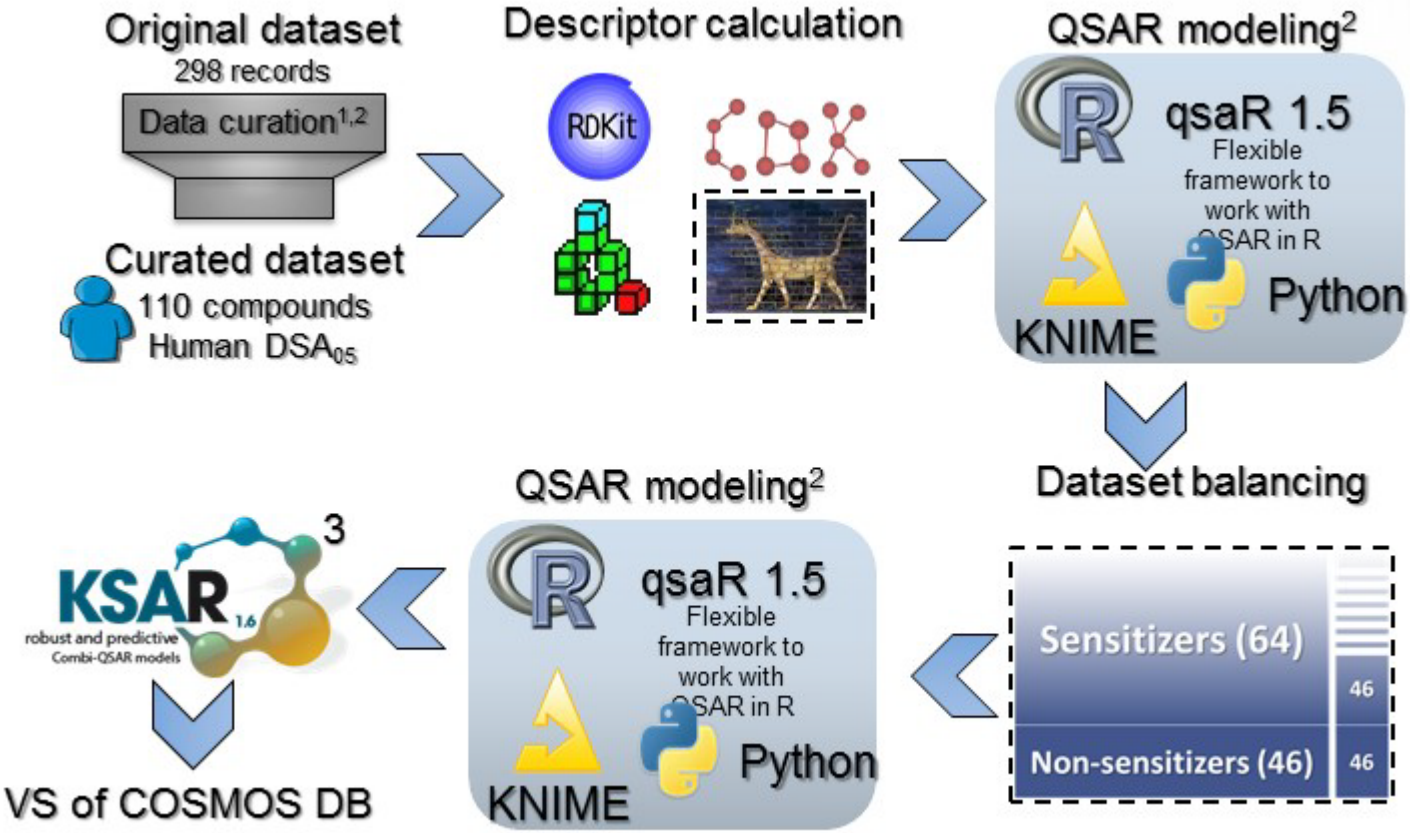
McCalla 1983 *Env Mutagen*

Marrying SAR and QSAR in CWAS: Deriving alerts from validated QSAR models



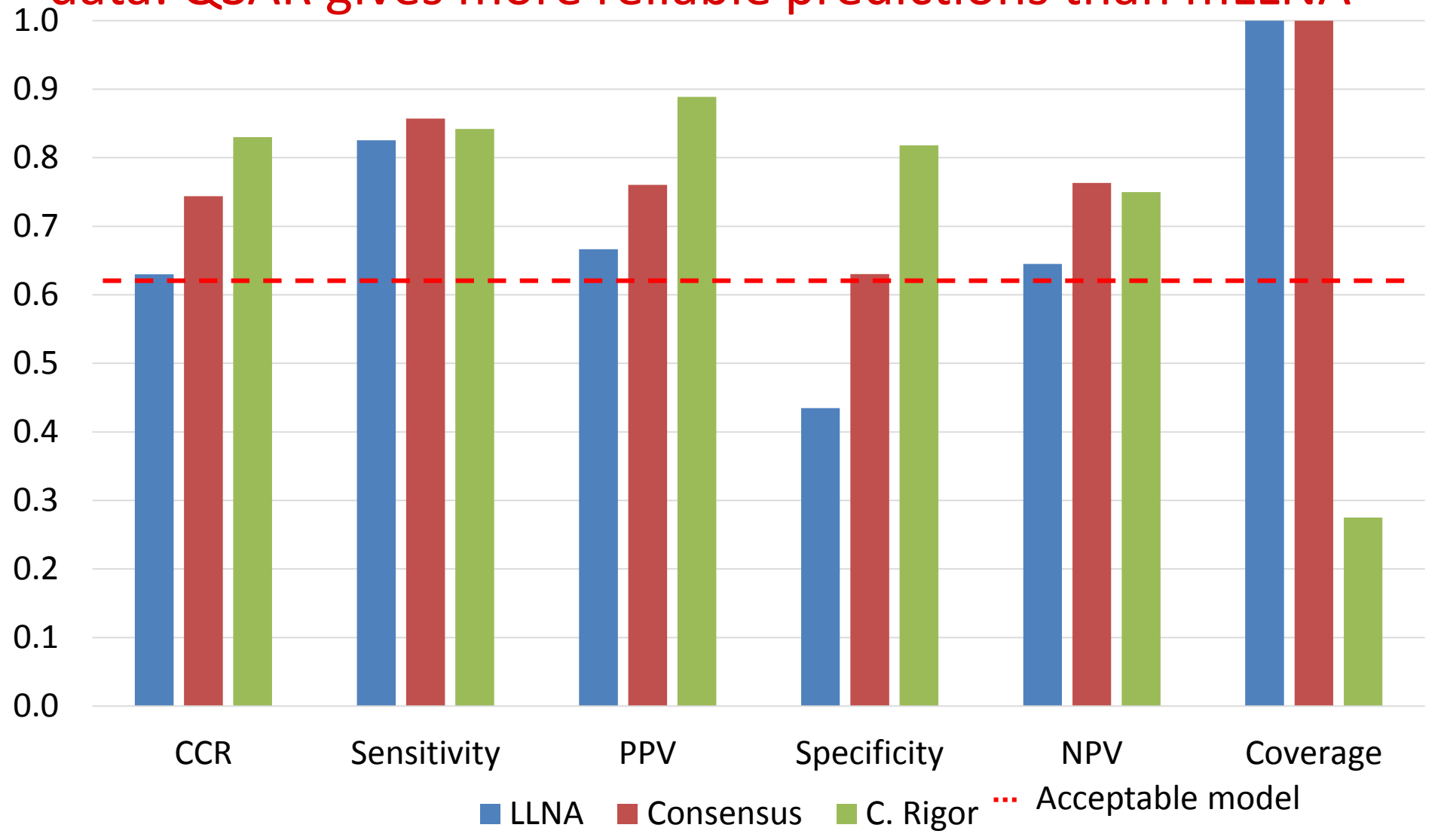
Can models replace testing? Skin sensitization modeling of human data

human DSA_{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)



[1]Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.
[2]Tropsha, A. *Mol. Inform.* **2010**, *29*, 476–488.
[3] Braga, R. C.; Alves, V. M. et al. *Curr. Top. Med. Chem.* **2014**, *14*, 1399–1415.

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

THE NATIONAL ACADEMIES

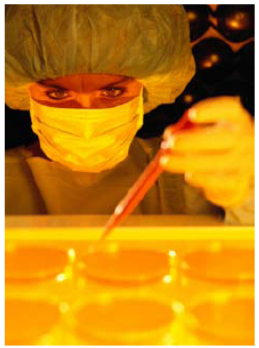
July 2007

REPORT IN BRIEF

Toxicity Testing in the 21st Century: A Vision and a Strategy

Advances in molecular biology, biotechnology, and other fields are paving the way for major improvements in how scientists evaluate the health risks posed by potentially toxic chemicals found at low levels in the environment. These advances would make toxicity testing quicker, less expensive, and more directly relevant to human exposures. They could also reduce the need for animal testing by substituting more laboratory tests based on human cells. This National Research Council report creates a far-reaching vision for the future of toxicity testing.

Toxicity tests on laboratory animals are conducted to evaluate chemicals—including medicines, food additives, and industrial, consumer, and agricultural chemicals—for their potential to cause cancer, birth defects, and other adverse health effects. Information from toxicity testing serves as an important part of the basis for public health and regulatory decisions concerning toxic chemicals. Current test methods were developed incrementally over the past 50 to 60 years and are conducted using laboratory animals, such as rats and mice. Using the results of animal tests to predict human health effects involves a number of assumptions and extrapolations that remain controversial. Test animals are often exposed to higher doses than would be expected for typical human exposures, requiring assumptions about effects at lower doses or exposures. Test animals are typically observed for overt signs of adverse health effects, which provide little information about biological changes leading to such health effects. Often controversial uncertainty factors must be applied to account for differences between test animals and humans. Finally, use of animals in testing is expensive and time consuming, and it sometimes raises ethical issues.



Today, toxicological evaluation of chemicals is poised to take advantage of the on-going revolution in biology and biotechnology. This revolution is making it increasingly possible to study the effects of chemicals using cells, cellular components, and tissues—preferably of human origin—rather than whole animals. These powerful new approaches should help to address a number of challenges facing the

National Academy of Scie

POLICYFORUM

TOXICOLOGY

Transforming Environmental Health Protection

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.

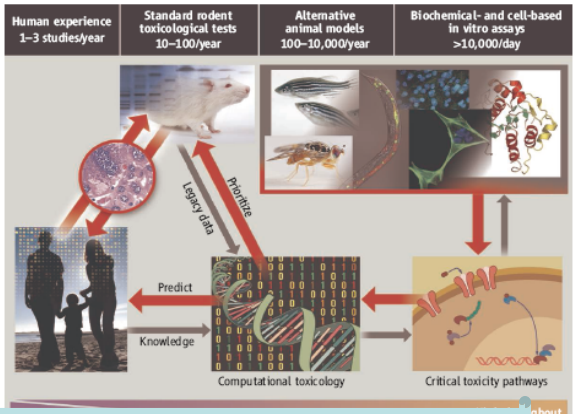
Francis S. Collins,[1*†] George M. Gray,[2*] John R. Bucher[3*]

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 d costs (*1*–*5*). In mittee on Toxicity of Environmental ports that reviewed dentified key issues, nd implementation r shift in the assess- rd and risk (*6*, *7*). s have laid out a solid prehensive and rig- d comparisons with l determine whether ements will be real- purpose, NTP, EPA, of Health Chemical GC) (organizations with expertise in experimental toxicology, computational toxicology, and high-throughput technologies, respectively) have established a collaborative research program.

EPA, NCGC, and NTP Joint Activities

In 2004, the NTP released its vision and roadmap for the 21st century (*1*), which established initiatives to integrate 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 vitro (*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, NCGC, 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 multiassay 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://ncgc.nih.gov/pub/openhts). HTS data collected by EPA and NTP, as well as by the NCGC and other Molecular Libraries Initiative centers (http://mli.nih.gov/), are being made publicly available through Web-based databases [e.g., PubChem (http://pubchem.ncbi.nlm.nih.gov)]. In addition,

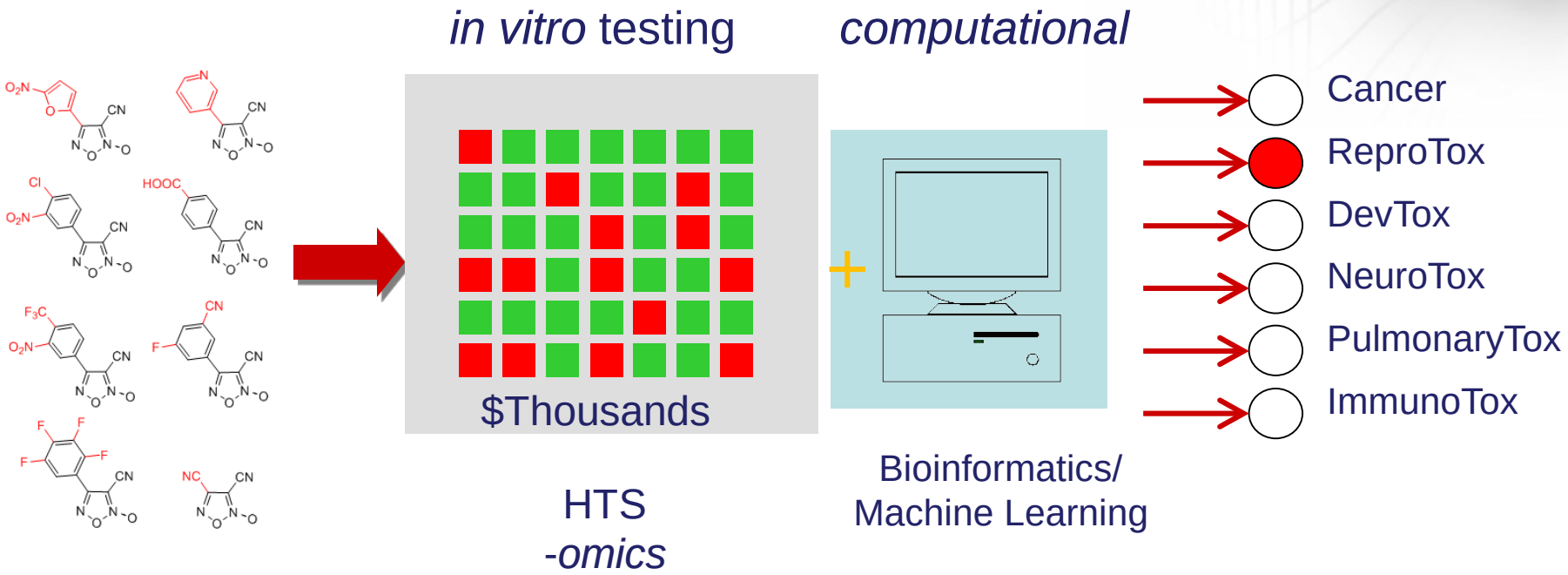


[1]Director, National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, MD 20892; [2]Assistant Administrator for the Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC 20460; [3]Associate Director, U.S. National Toxicology Program, National Institute of Environmental

EPAs Contribution: The ToxCast Research Program

Slide courtesy of Dr. Ann Richard, EPA (modified)

QSAR and Chemical Toxicity Testing in the 21 Century

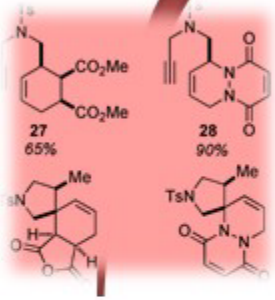


Slide courtesy of Dr. Ann Richard, EPA (modified)

Integration of Diverse Data Streams into QSAR Modeling to Improve Toxicity Prediction

Cheminformatics

Over many chemicals



Chemical descriptors (in silico):

Molecular weight,
Connectivity indices
Presence/absence of fragment,
Hydrophobicity, etc.

Bioinformatics

Over many biological assays

Short-term biological assays

Transcriptomics,
Metabolomics,
Cytotoxicity,
Genotype, etc

Huma studies

Medica terature
e-healt records
Insura e claims

Toxicity

QSAR modeling: chemical descriptors

	x1	x2	...	...	...	...	xz
Chemical 1							
Chemical 2	Chemical descriptors				Bioassay data		
Chemical 3							
...							
Chemical n							

High dimensional data, X

Machine learning

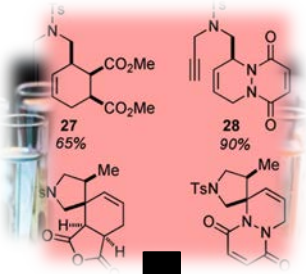
$y=f(X)$

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

Response, *y*

Zhu H et al. (2008) *Environ. Health Perspect.* 116, 506-513;
Low Y et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262;
Sedykh A *et al*. (2011) *Environ. Health Perspect.* (119): 364-370

QSAR modeling: in vitro assay descriptors



	x1	x2	...	...	...	...	xz
Chemical 1							
Chemical 2	Chemical descriptors				Bioassay data		
Chemical 3							
...							
Chemical n							

High dimensional data, X

Machine learning

$y=f(X)$

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

Response, y

Zhu H et al. (2008) *Environ. Health Perspect.* 116, 506-513;
Low Y et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262;
Sedykh A *et al*. (2011) *Environ. Health Perspect.* (119): 364-370

QSAR modeling: hybrid descriptors

	x1	x2	...	...	...	...	xz
Chemical 1							
Chemical 2	Chemical descriptors				Bioassay data		
Chemical 3							
...							
Chemical n							

High dimensional data, X

Machine learning

$y=f(X)$

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

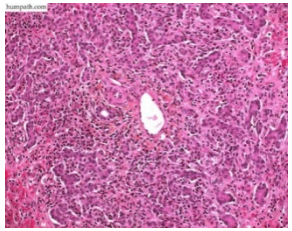
Response, *y*

Zhu H et al. (2008) *Environ. Health Perspect.* 116, 506-513;
Low Y et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262;
Sedykh A *et al*. (2011) *Environ. Health Perspect.* (119): 364-370

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

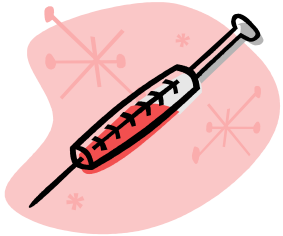
- Zhu, H., *et al*. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *EHP*, **2008**, (116): 506-513
- Sedykh A, *et al*. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *EHP*, **2011**, 119(3):364-70.
- Low *et al*., Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem Res Toxicol*. **2011** Aug 15;24(8):1251-62
- Rusyn *et al*, Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. *Tox. Sci*., **2012**, 127(1):1-9
- Low Y, *et al*. Integrative chemical-biological read-across approach for chemical hazard classification. *Chem Res Toxicol*. **2013**, 26(8):1199-208
- Low, Y, *et al*. Integrative Approaches for Predicting In Vivo Effects of Chemicals from their Structural Descriptors and the Results of Short-Term Biological Assays. *Curr. Top. Med. Chem*., **2014**, 14(11):1356-64
- Low et al, Cheminformatics-Aided Pharmacovigilance: Application to Stevens Johnson Syndrome. JAMIA, 2015 (in press).

Predicting Subchronic Hepatotoxicity from 24h Toxicogenomics Profiles



Liver histopathology

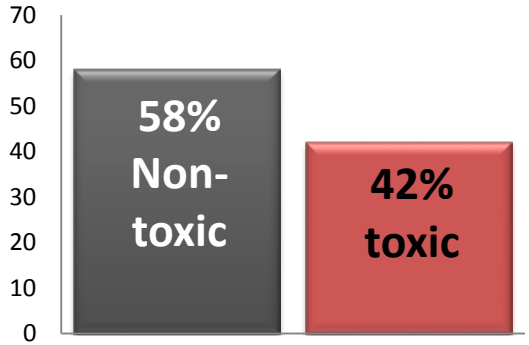
Rats in triplicates
6-8 weeks old
Sprague Dawley



Clinical chemistry

Assigned by pathologist

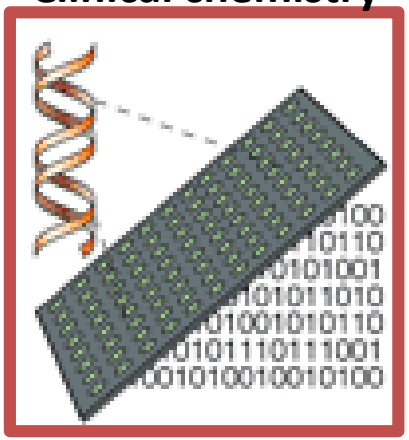
127 compounds in 2 classes



Subchronic 28-day hepatotoxicity

Doses: low, med, high

Time points:
3h, 6h, 9h, 24h,
3, 7, 14 and 28 days



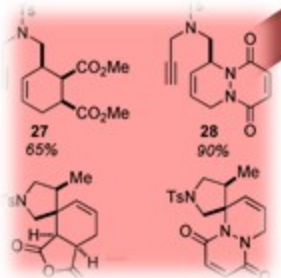
***In vivo* hepatic gene expression**
(24h, high dose)

Predict

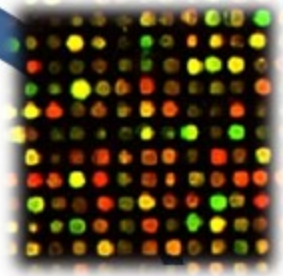
Data source: Open TG-GATEs http://toxico.nibio.go.jp/

QSAR models < Hybrid models < Toxicogenomics models

Data source: TGP2 Toxicogenomics Informatics Project in Japan



Chemical descriptors



Toxicogenomics expression
(24h)

304 Dragon descriptors

Hybrid models
68- 75% BAcc

2,923 genes

Rank by differential expression

Top 400 genes

Top 100 genes

Top 30 genes

Top 4 genes

QSAR models
55-61% BAcc

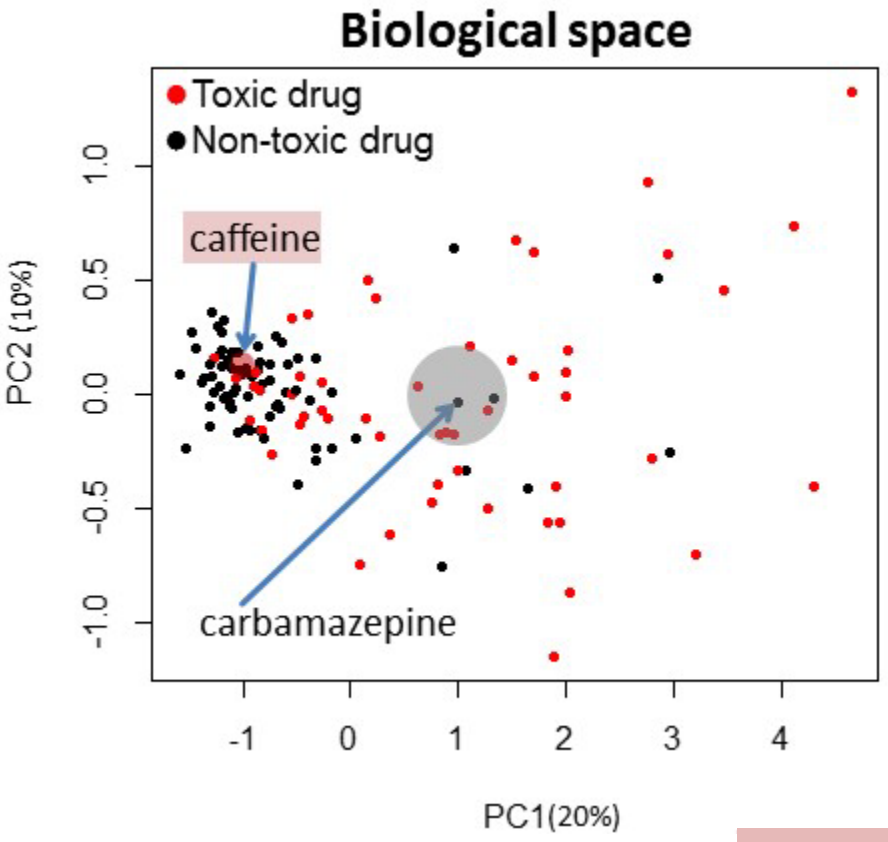
Toxicogenomics models
69-78% BAcc

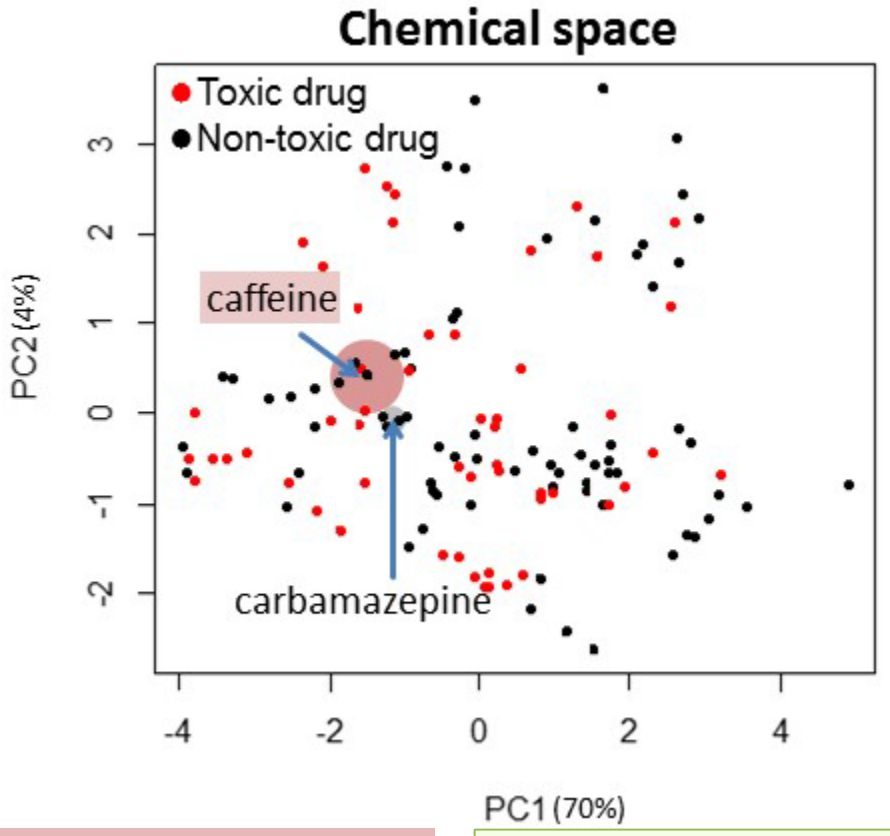
Hepatotoxicity
(28 day)

4 classification methods
(RF, SVM, kNN, DWD)

Low et al. (2011) *Chem. Res. Toxicol.* 24,1251-1262

Conflicting Predictions by QSAR and Toxicogenomics Models





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

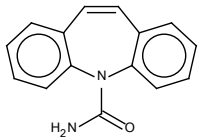
A_{pred}=similarity-weighted average of toxicity values

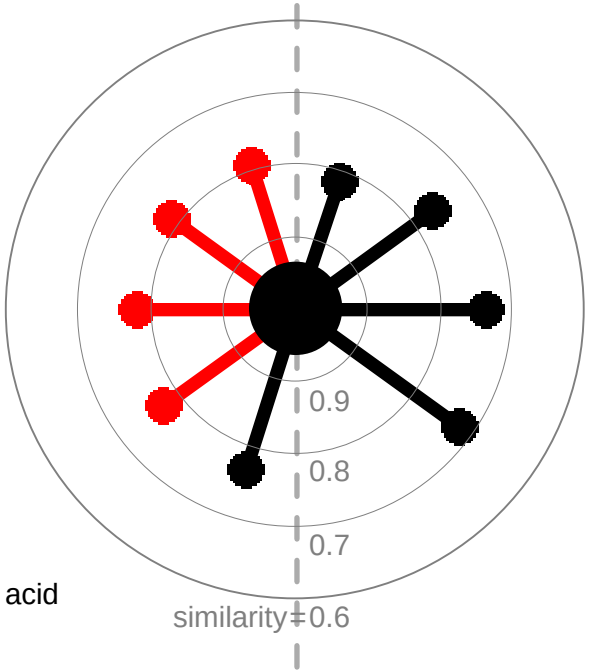
overall correctly predicted as nontoxic

wrongly predicted as toxic

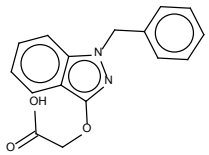
rightly predict as nontoxic

CARBAMAZEPINE
Non-toxic

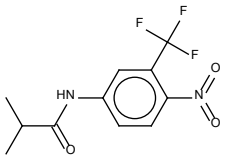




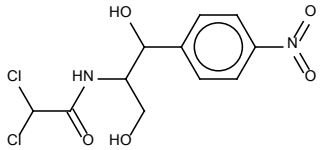
Biological neighbors
(nearest on top)



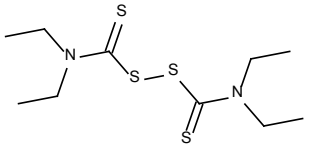
Bendazac
Toxic
0.790



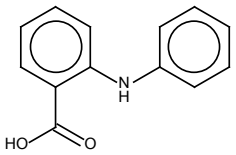
Flutamide
Toxic
0.783



Chloramphenicol
Toxic
0.776



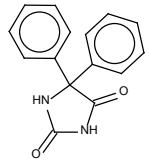
Disulfiram
Toxic
0.770



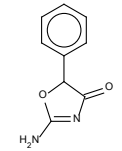
Phenylanthranilic acid
Non-toxic
0.767

Chemical neighbors

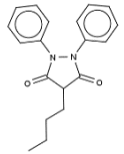
Phenytoin
Non-toxic
0.813



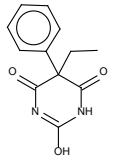
Pemoline
Non-toxic
0.766



Phenylbutazone
Non-toxic
0.737



Phenobarbital
Non-toxic
0.721



Low et al, Chem Res Toxicol. 2013, 26(8):1199-208

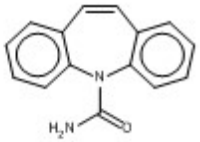
Chemical-biological read-across (CBRA): learning from both sets of neighbors

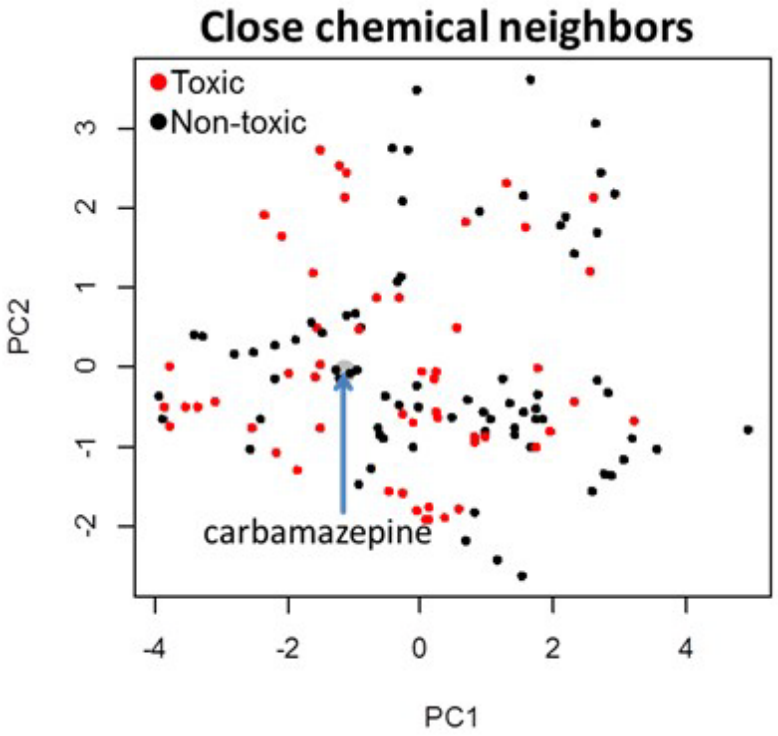
A_{pred}=similarity-weighted average of toxicity values

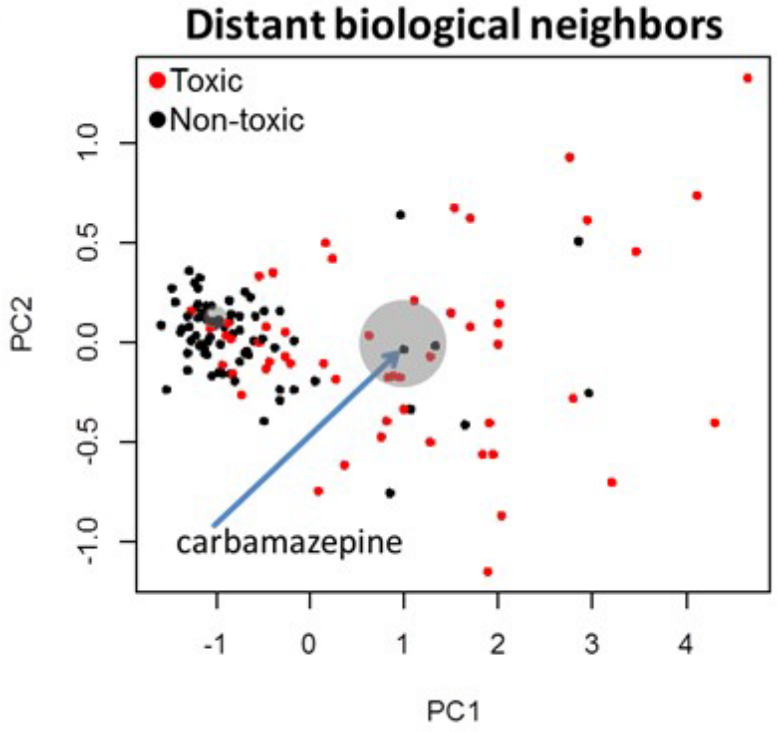
overall correctly predicted as nontoxic

CARBAMAZEPINE

Non-toxic







Low et al, Chem Res Toxicol. 2013, 26(8):1199-208

CBRA outperforms other models

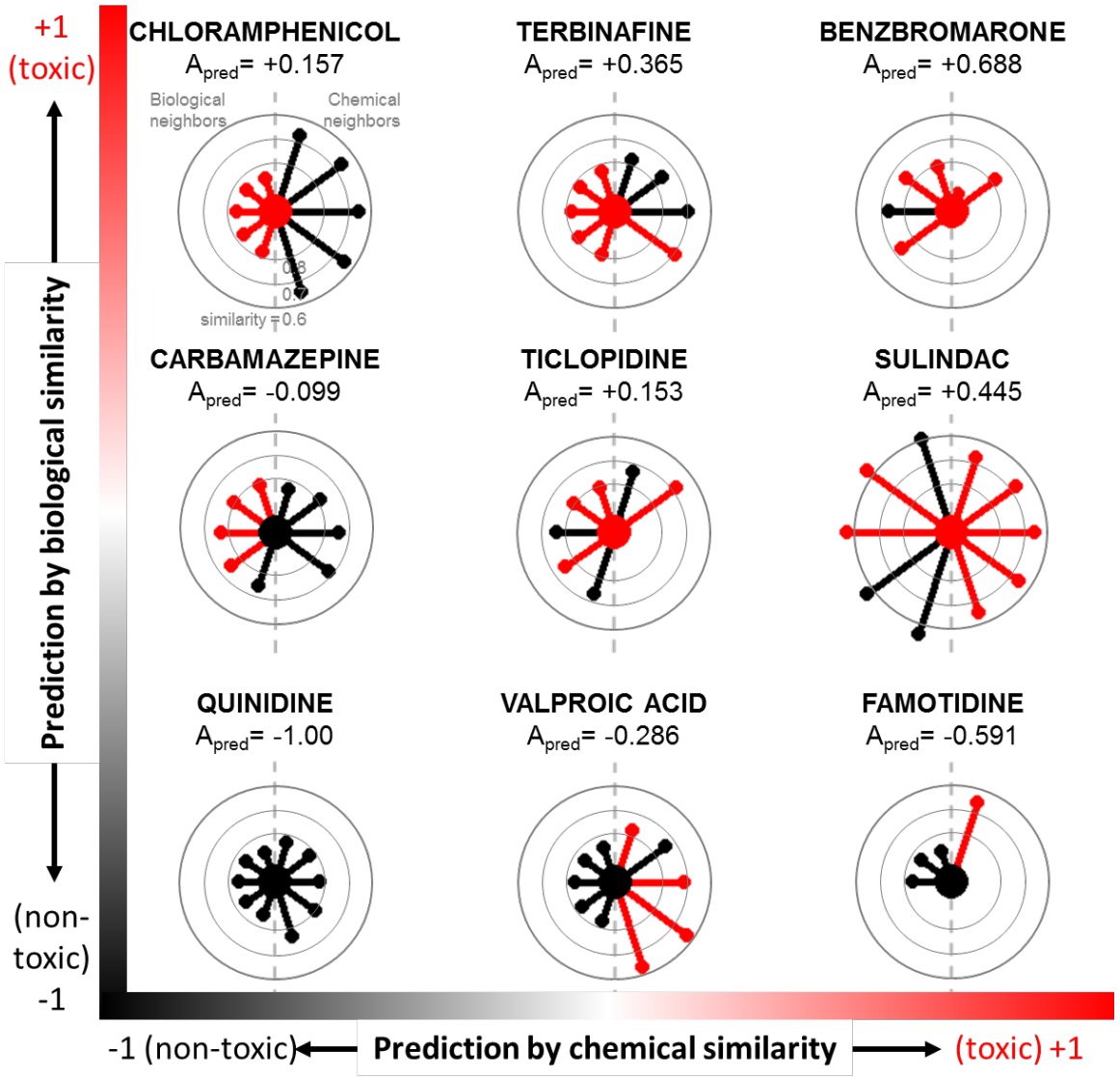
Model	Specificity	Sensitivity	Balanced accuracy (CCR)
Chemical read-across	0.73 ± 0.07	0.34 ± 0.05	0.53 ± 0.04
Biological read-across	0.85 ± 0.07	0.66 ± 0.04	0.76 ± 0.04
Hybrid read-across	0.85 ± 0.07	0.58 ± 0.04	0.72 ± 0.04
Multi-space read-across	0.89 ± 0.07	0.66 ± 0.04	**0.78** ± 0.04

Results of 5-fold external cross-validation

- 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

Low et al, Chem Res Toxicol. 2013, 26(8):1199-208

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



Low et al, Chem Res Toxicol. 2013, 26(8):1199-208

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
  - Pubic web portals (e.g., Chembench, OCHEM)

Acknowledgments

Principal Investigator
Alexander Tropsha

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

Postdoctoral Fellows
Olexander 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)

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

MAJOR FUNDING

NIH
- R01-GM66940
- R01-GM068665

NSF
- ABI 9179-1165

EPA (STAR awards)
- RD832720
- RD833825
- RD834999

ONR