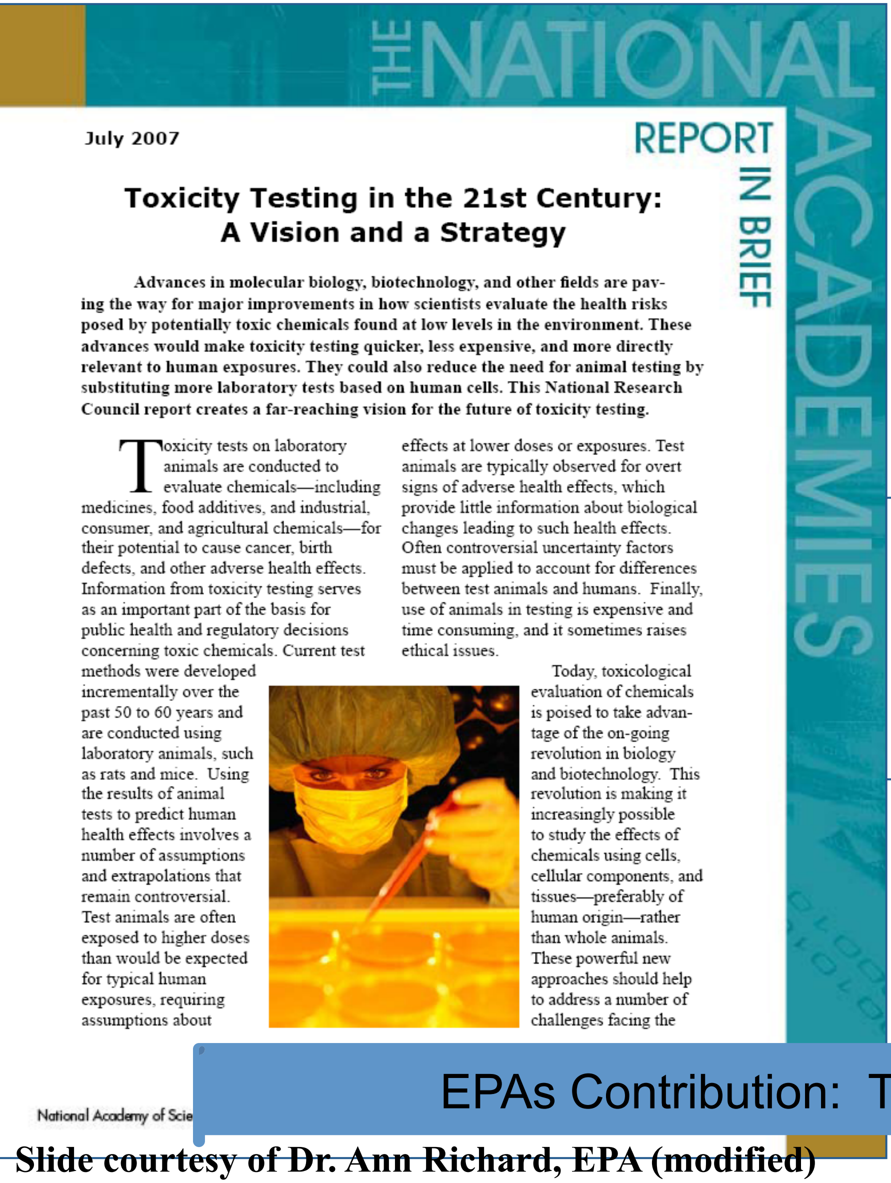
On the use of biological descriptors of chemical compounds to enrich traditional cheminformatics applications

Alexander Tropsha
Laboratory for Molecular Modeling and
Carolina Center for Computational Toxicology
UNC Eshelman School of Pharmacy
UNC-Chapel Hill

QSAR and Chemical Toxicity Testing in the 21 Century



THE NATIONAL ACADEMIES

REPORT IN BRIEF

July 2007

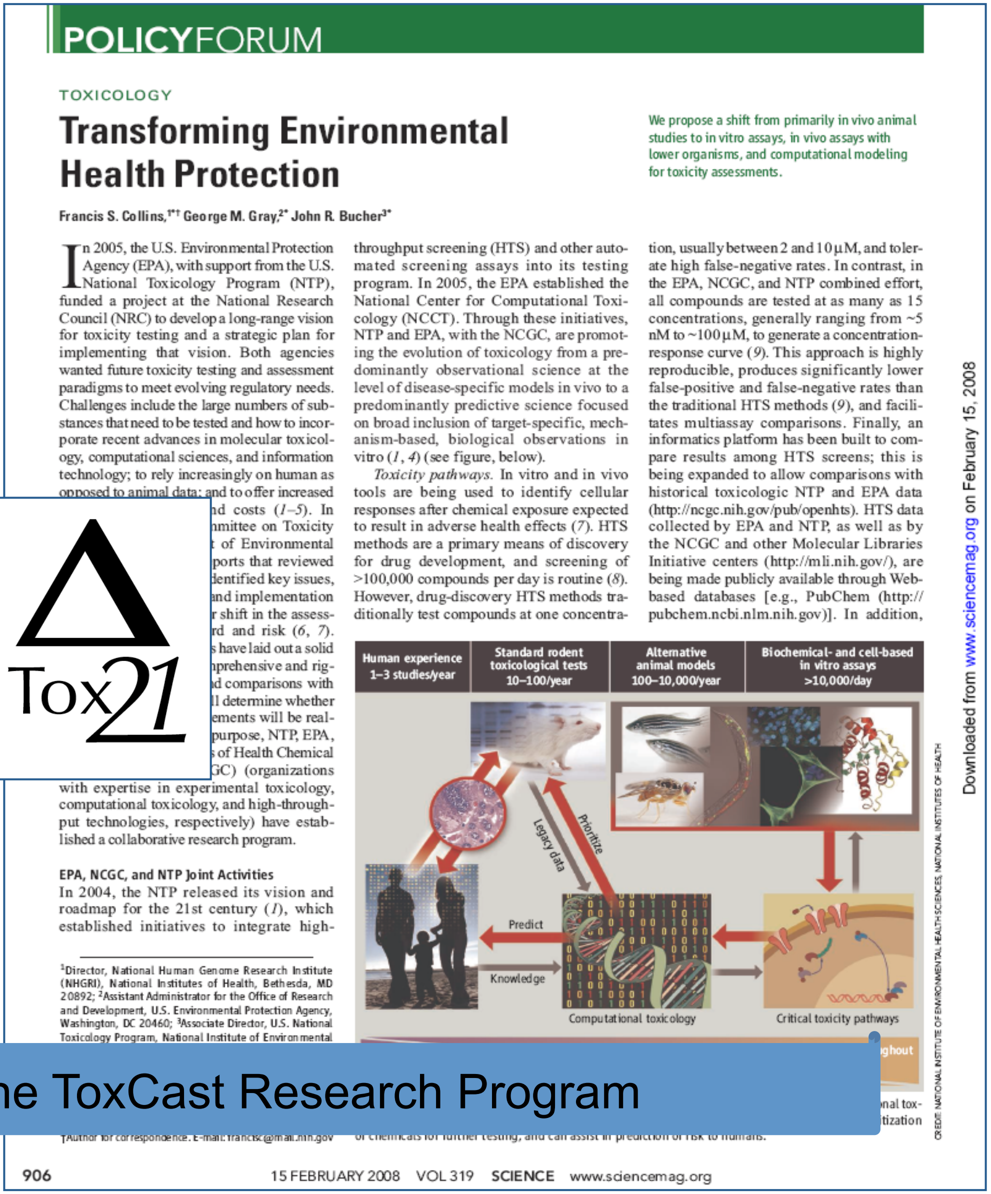
Toxicity Testing in the 21st Century: A Vision and a Strategy

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
nd costs (1–5). In
nmittee on Toxicity
t of Environmental
ports that reviewed
dentified key issues,
and implementation
r shift in the assess-
rd and risk (6, 7).
s have laid out a solid
nprehensive and rig-
d comparisons with
ll determine whether
ements will be real-
purpose, NTP, EPA,
s 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,

nal tox-
tization

or chemicals for further testing, and can assist in prediction of risk to humans.

[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

†Author for correspondence. E-mail: francisc@mail.nih.gov

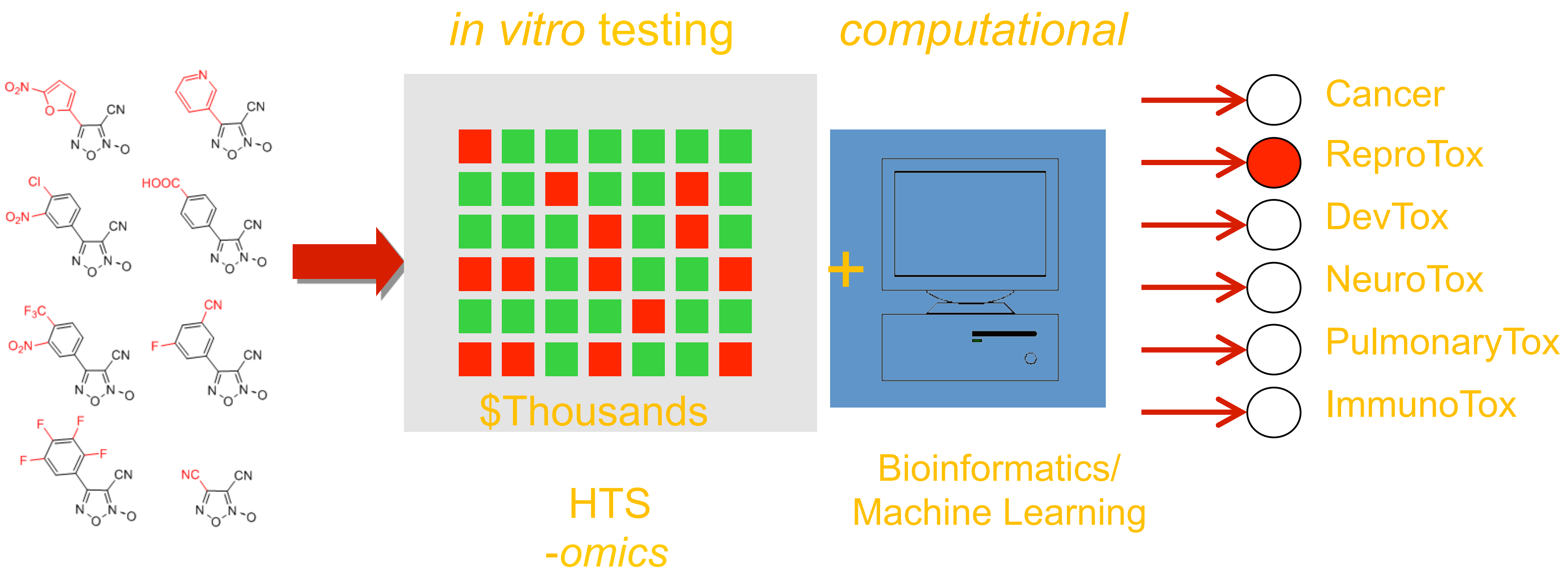
CREDIT: NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES, NATIONAL INSTITUTES OF HEALTH

Downloaded from www.sciencemag.org on February 15, 2008

EPAs Contribution: The ToxCast Research Program

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

QSAR and Chemical Toxicity Testing in the 21 Century



in vitro testing

🟥	🟩	🟩	🟩	🟩	🟩	🟩
🟩	🟩	🟥	🟩	🟩	🟥	🟩
🟩	🟩	🟩	🟥	🟩	🟥	🟩
🟥	🟥	🟩	🟥	🟩	🟩	🟥
🟩	🟩	🟩	🟩	🟥	🟩	🟩
🟥	🟥	🟩	🟥	🟩	🟩	🟥

$Thousands

HTS

-omics

+

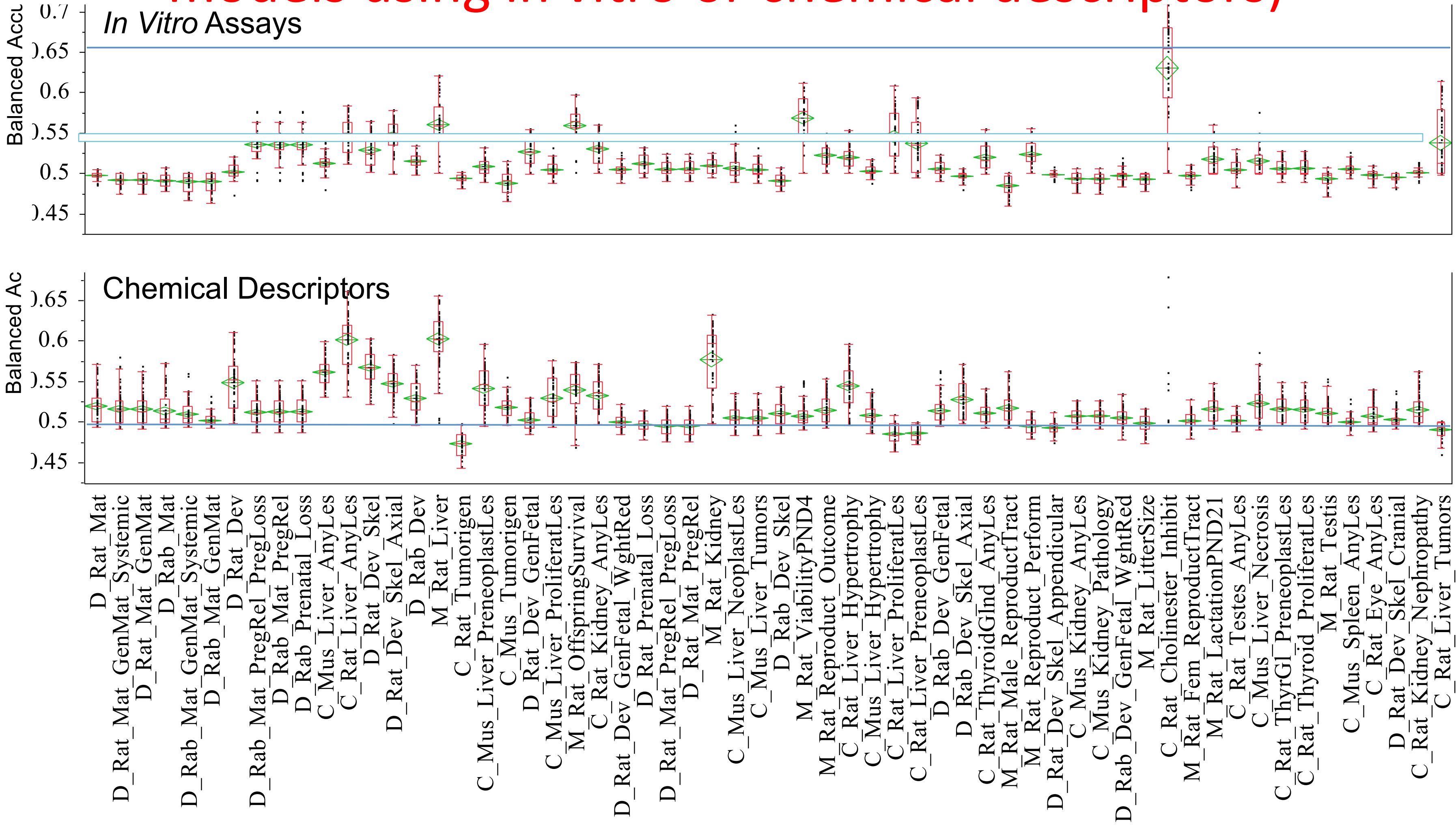
computational

Bioinformatics/
Machine Learning

- ○ Cancer
- ● ReproTox
- ○ DevTox
- ○ NeuroTox
- ○ PulmonaryTox
- ○ ImmunoTox

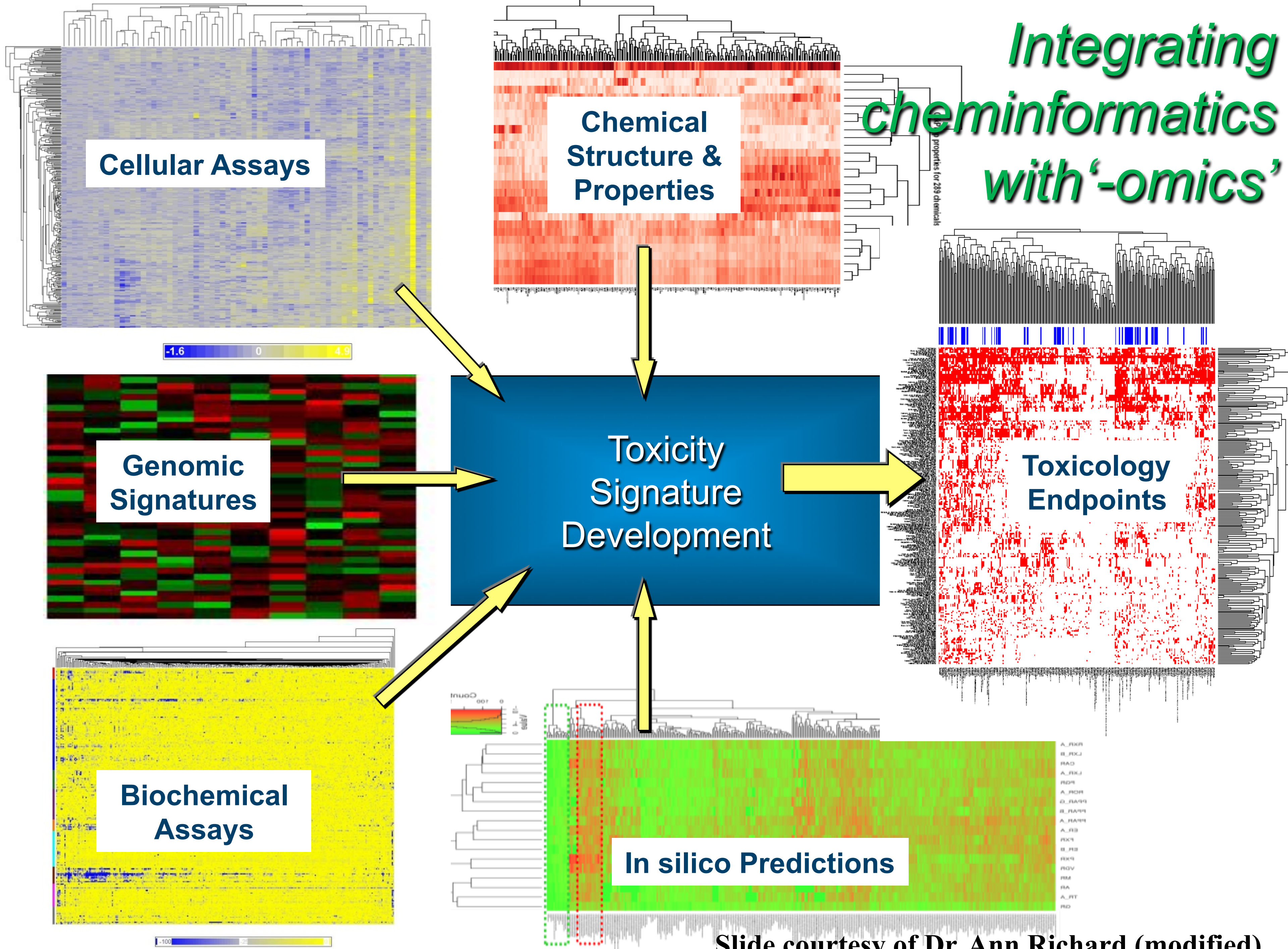
Slide courtesy of Dr. Ann Richard, EPA (modified)

Poor structure – in vivo or in vitro-in vivo correlations for Toxcast data (ca. 80 models using in vitro or chemical descriptors)*



***Thomas et al., Toxicol Sci. 2012 Aug;128(2):398-417.**

Integrating cheminformatics with‘-omics’



Cellular Assays

Chemical Structure & Properties

Genomic Signatures

Biochemical Assays

In silico Predictions

Toxicity Signature Development

Toxicology Endpoints

Slide courtesy of Dr. Ann Richard (modified)

40 CENTS

THE Chemoinformatics Manifesto

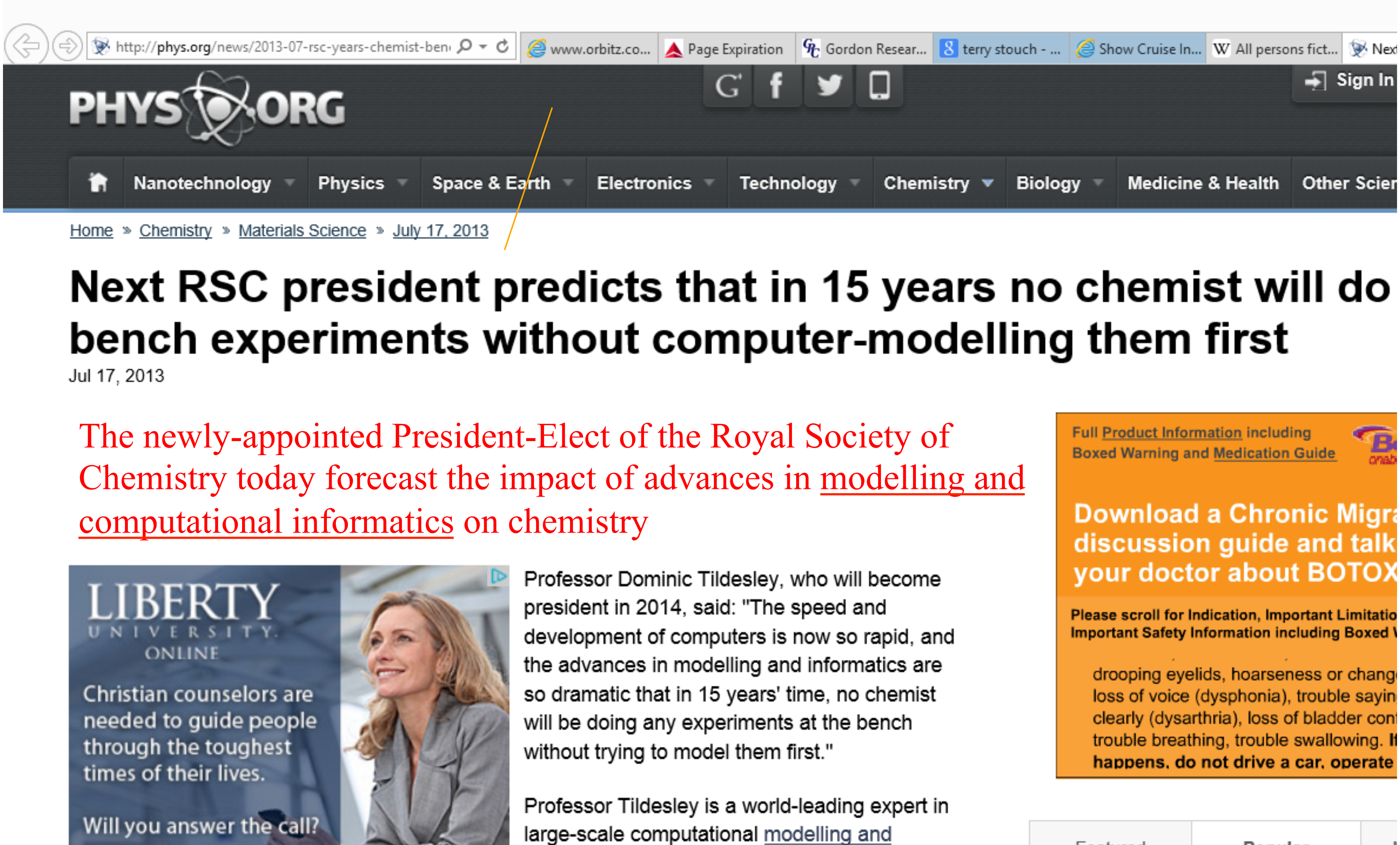
INTERNATIONAL PUBLISHERS

40 CENTS

THE Chemoinformatics Manifesto

A spectre is haunting Europe -- the spectre of [chemoinformatics]. [Chemoinformatics] is already acknowledged by all European powers to be itself a power. It is high time that [Chemoinformaticians] should openly, in the face of the whole world, publish their views, their aims, their tendencies, and meet this nursery tale of the spectre of [chemoinformatics] with a manifesto of the party itself.

The importance of modeling is acknowledged and appreciated



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

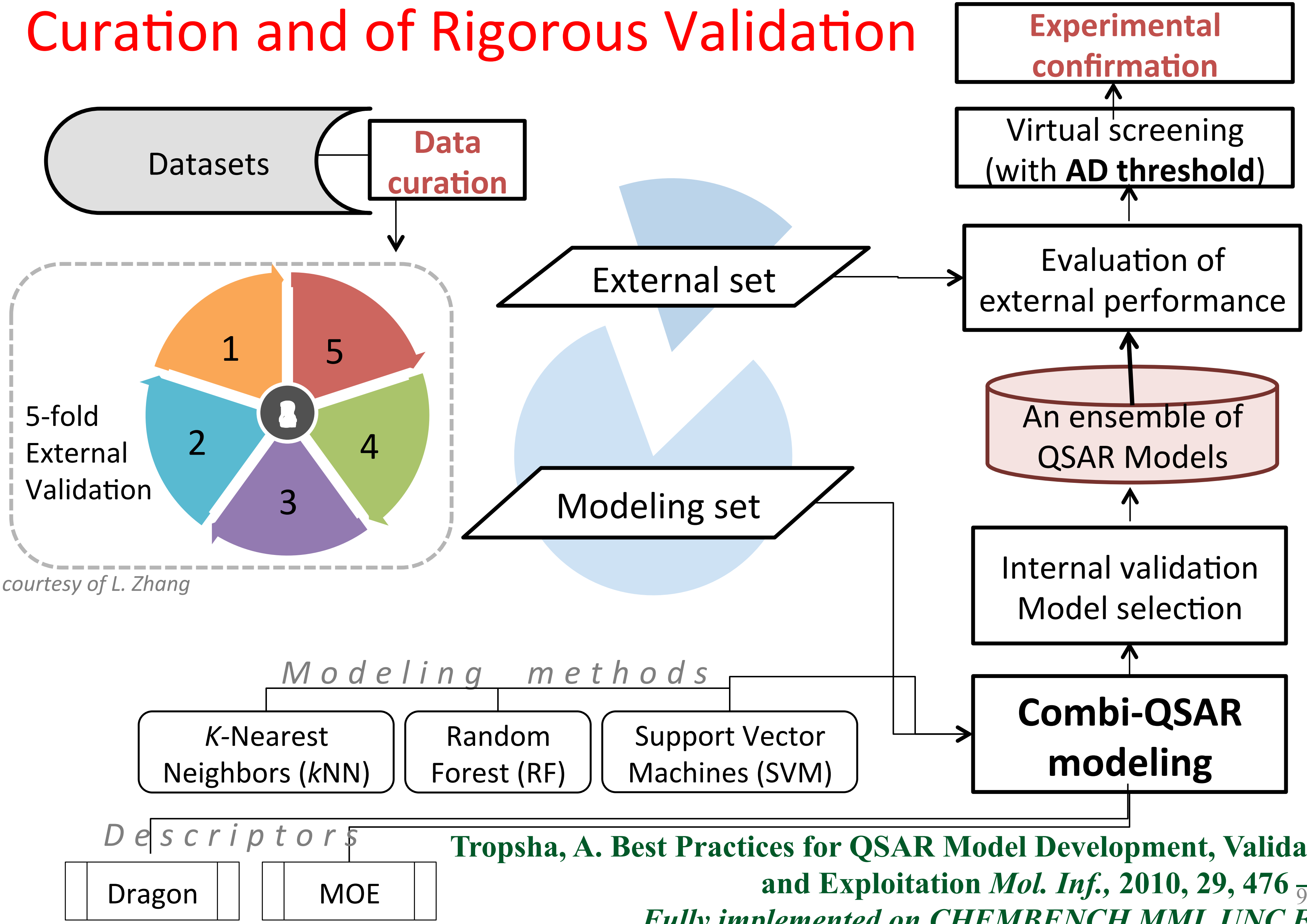
Jul 17, 2013

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

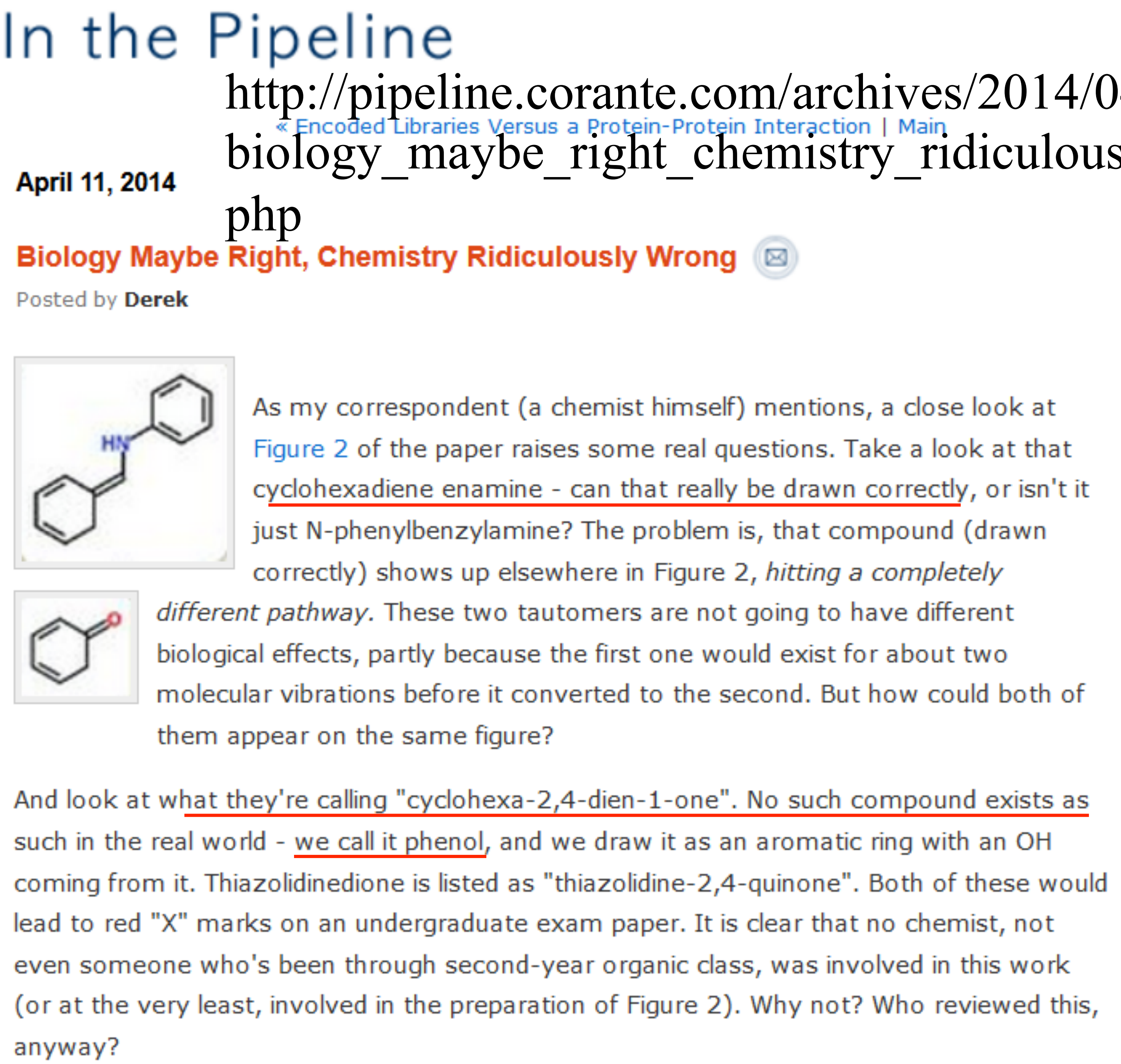
QSAR Modeling Workflow: the Importance of Data Curation and of Rigorous Validation


Experimental confirmation
Virtual screening (with **AD threshold**)
Evaluation of external performance
Datasets
Data curation
External set
5-fold External Validation
1
2
3
4
5
Modeling set
An ensemble of QSAR Models
Internal validation Model selection
courtesy of L. Zhang
Modeling methods
K-Nearest Neighbors (*k*NN)
Random Forest (RF)
Support Vector Machines (SVM)
Combi-QSAR modeling
Descriptors
Dragon
MOE

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

Fully implemented on CHEMBENCH.MML.UNC.EDU

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



In the Pipeline

« Encoded Libraries Versus a Protein-Protein Interaction | Main

April 11, 2014

Biology Maybe Right, Chemistry Ridiculously Wrong

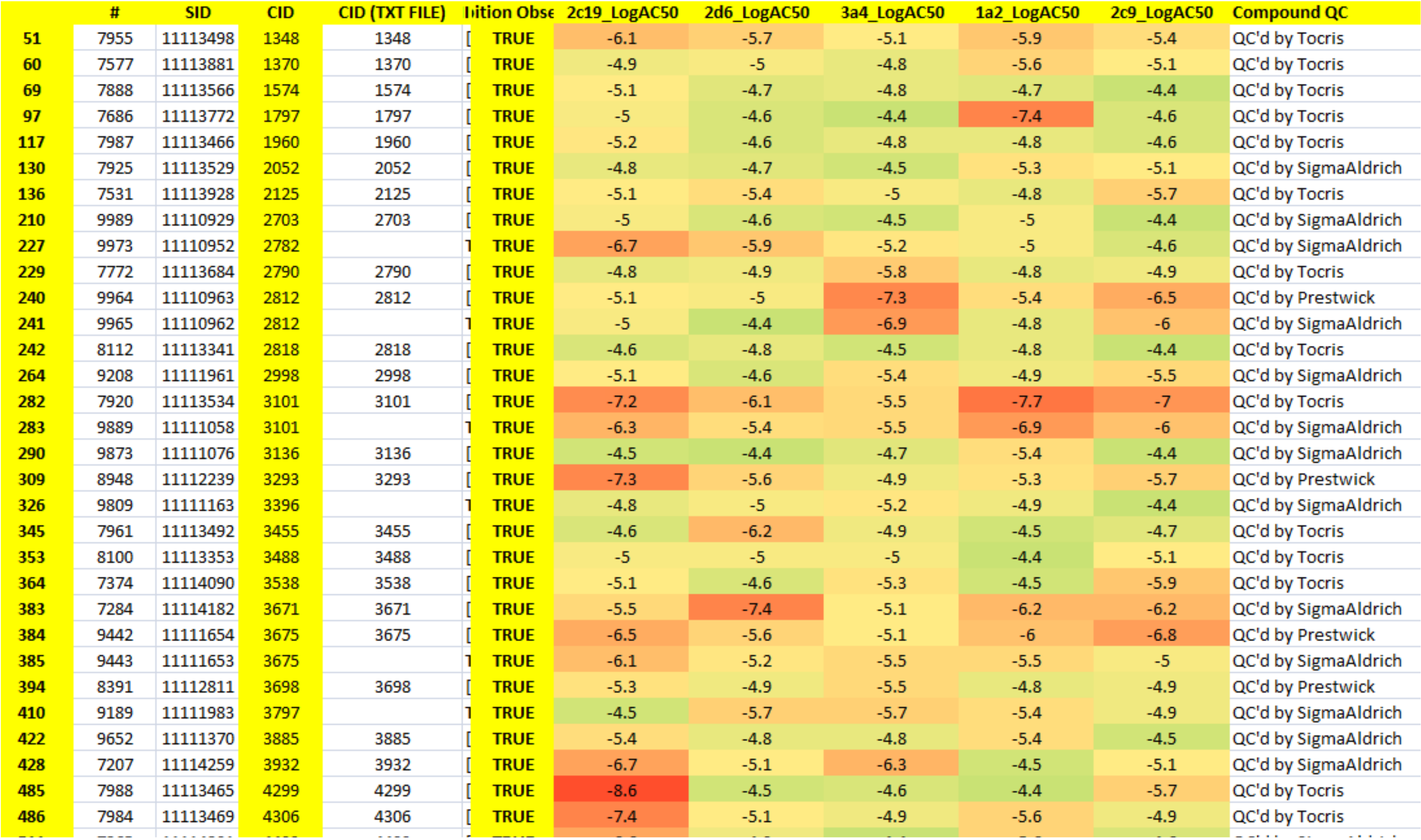
Posted by **Derek**

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?

Cheminformatics Analysis of (inaccuracy of) qHTS Data

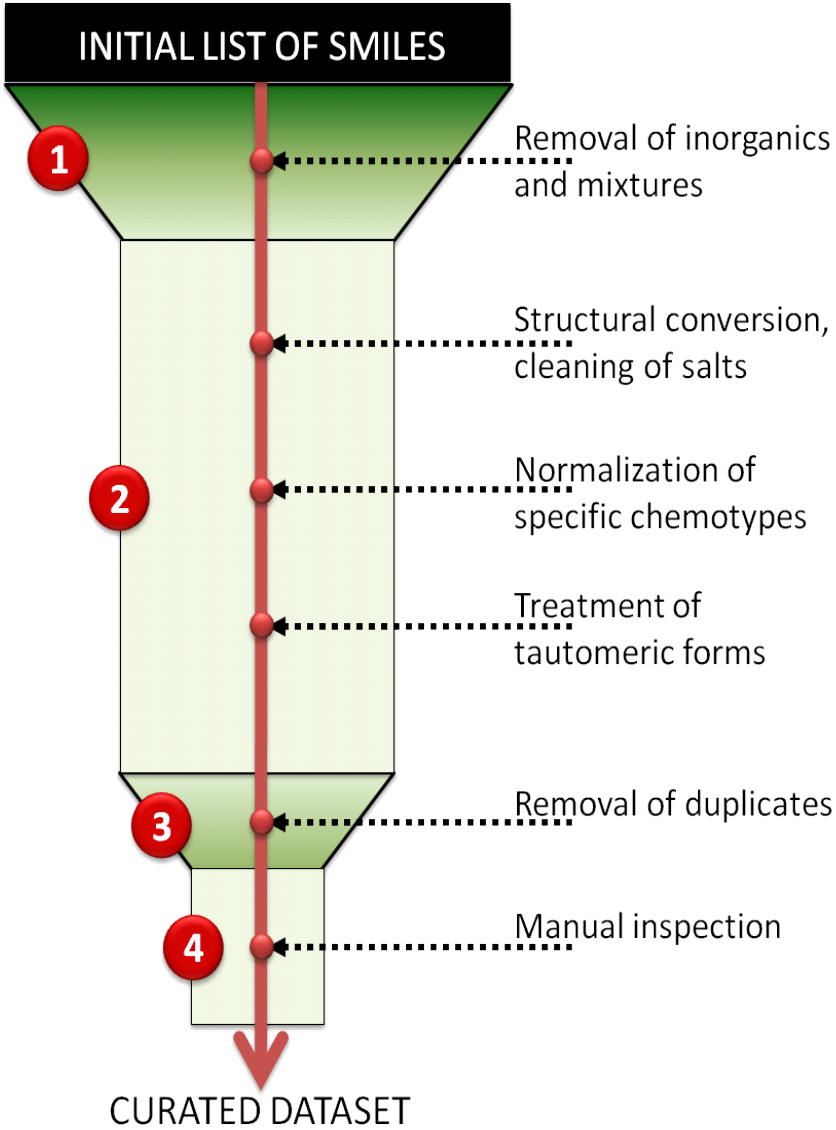
over 17,000 compounds screened against five major CYP isozymes using In Vitro bioluminescent qHTS assay



	#	SID	CID	CID (TXT FILE)	lition	Obse	2c19_LogAC50	2d6_LogAC50	3a4_LogAC50	1a2_LogAC50	2c9_LogAC50	Compound QC
51	7955	11113498	1348	1348	[	TRUE	-6.1	-5.7	-5.1	-5.9	-5.4	QC'd by Tocris
60	7577	11113881	1370	1370	[	TRUE	-4.9	-5	-4.8	-5.6	-5.1	QC'd by Tocris
69	7888	11113566	1574	1574	[	TRUE	-5.1	-4.7	-4.8	-4.7	-4.4	QC'd by Tocris
97	7686	11113772	1797	1797	[	TRUE	-5	-4.6	-4.4	-7.4	-4.6	QC'd by Tocris
117	7987	11113466	1960	1960	[	TRUE	-5.2	-4.6	-4.8	-4.8	-4.6	QC'd by Tocris
130	7925	11113529	2052	2052	[	TRUE	-4.8	-4.7	-4.5	-5.3	-5.1	QC'd by SigmaAldrich
136	7531	11113928	2125	2125	[	TRUE	-5.1	-5.4	-5	-4.8	-5.7	QC'd by Tocris
210	9989	11110929	2703	2703	[	TRUE	-5	-4.6	-4.5	-5	-4.4	QC'd by SigmaAldrich
227	9973	11110952	2782		T	TRUE	-6.7	-5.9	-5.2	-5	-4.6	QC'd by SigmaAldrich
229	7772	11113684	2790	2790	[	TRUE	-4.8	-4.9	-5.8	-4.8	-4.9	QC'd by Tocris
240	9964	11110963	2812	2812	[	TRUE	-5.1	-5	-7.3	-5.4	-6.5	QC'd by Prestwick
241	9965	11110962	2812		T	TRUE	-5	-4.4	-6.9	-4.8	-6	QC'd by SigmaAldrich
242	8112	11113341	2818	2818	[	TRUE	-4.6	-4.8	-4.5	-4.8	-4.4	QC'd by Tocris
264	9208	11111961	2998	2998	[	TRUE	-5.1	-4.6	-5.4	-4.9	-5.5	QC'd by SigmaAldrich
282	7920	11113534	3101	3101	[	TRUE	-7.2	-6.1	-5.5	-7.7	-7	QC'd by Tocris
283	9889	11111058	3101		T	TRUE	-6.3	-5.4	-5.5	-6.9	-6	QC'd by SigmaAldrich
290	9873	11111076	3136	3136	[	TRUE	-4.5	-4.4	-4.7	-5.4	-4.4	QC'd by SigmaAldrich
309	8948	11112239	3293	3293	[	TRUE	-7.3	-5.6	-4.9	-5.3	-5.7	QC'd by Prestwick
326	9809	11111163	3396		T	TRUE	-4.8	-5	-5.2	-4.9	-4.4	QC'd by SigmaAldrich
345	7961	11113492	3455	3455	[	TRUE	-4.6	-6.2	-4.9	-4.5	-4.7	QC'd by Tocris
353	8100	11113353	3488	3488	[	TRUE	-5	-5	-5	-4.4	-5.1	QC'd by Tocris
364	7374	11114090	3538	3538	[	TRUE	-5.1	-4.6	-5.3	-4.5	-5.9	QC'd by Tocris
383	7284	11114182	3671	3671	[	TRUE	-5.5	-7.4	-5.1	-6.2	-6.2	QC'd by SigmaAldrich
384	9442	11111654	3675	3675	[	TRUE	-6.5	-5.6	-5.1	-6	-6.8	QC'd by Prestwick
385	9443	11111653	3675		T	TRUE	-6.1	-5.2	-5.5	-5.5	-5	QC'd by SigmaAldrich
394	8391	11112811	3698	3698	[	TRUE	-5.3	-4.9	-5.5	-4.8	-4.9	QC'd by Prestwick
410	9189	11111983	3797		T	TRUE	-4.5	-5.7	-5.7	-5.4	-4.9	QC'd by SigmaAldrich
422	9652	11111370	3885	3885	[	TRUE	-5.4	-4.8	-4.8	-5.4	-4.5	QC'd by SigmaAldrich
428	7207	11114259	3932	3932	[	TRUE	-6.7	-5.1	-6.3	-4.5	-5.1	QC'd by SigmaAldrich
485	7988	11113465	4299	4299	[	TRUE	-8.6	-4.5	-4.6	-4.4	-5.7	QC'd by Tocris
486	7984	11113469	4306	4306	[	TRUE	-7.4	-5.1	-4.9	-5.6	-4.9	QC'd by Tocris

Veith et al., Nature Biotechnology, 2009, 27:1050-5
Sun et al., J. Chem. Inf. Model., 2011, 51:2474-81

Dataset Curation summary



INITIAL LIST OF SMILES	**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
CURATED DATASET	

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

Chemical Duplicate Analysis

- Carried out by ISIDA/Duplicates program
- 1,280 duplicate couples were found
  - 406 had a complete matching profile
  - 874 had profile differences
  - A total of 1,535 discrepancies were found in the 874 duplicates couples CYP annotation:

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

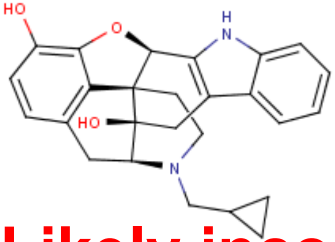
PROBLEM: CYP bioprofiles for some duplicates are dramatically different

➡ Need biological curation!

Neighborhood analysis helps to choose correct value

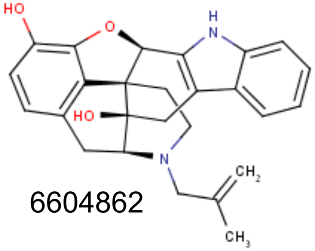
Case Study: structural duplicates found in NCGC CYP450 qHTS data

Tocris-0740	SID	Supplier	Cytochrome P450 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	INA	**-8**	-5.6	-5

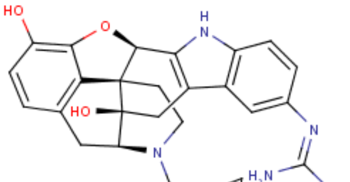


Likely incorrect!

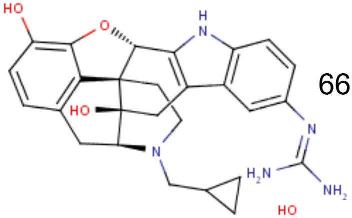
5 Nearest neighbors	Tanimoto Similarity	SID	Supplier	Cytochrome P450 2C9	 1A2	 3A4	 2D6	 2C19
6604862	**0.98**	11114071	Tocris	INA	INA	**4.5-**	INA	5.5-
6604106	**0.98**	11112029	Sigma Aldrich	INA	INA	**5.1-**	INA	INA
6604846	**0.98**	11114012	Tocris	INA	INA	**INA**	INA	INA
6604136	**0.95**	11112054	Sigma Aldrich	INA	INA	**4.8-**	5.9-	INA
6604137	**0.95**	11113764	Tocris	INA	4.4-	**4.7-**	4.5-	INA



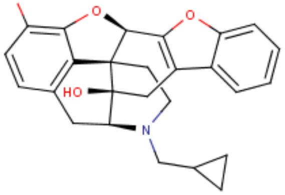
6604862



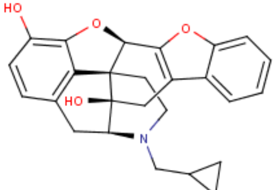
6604106



6604846

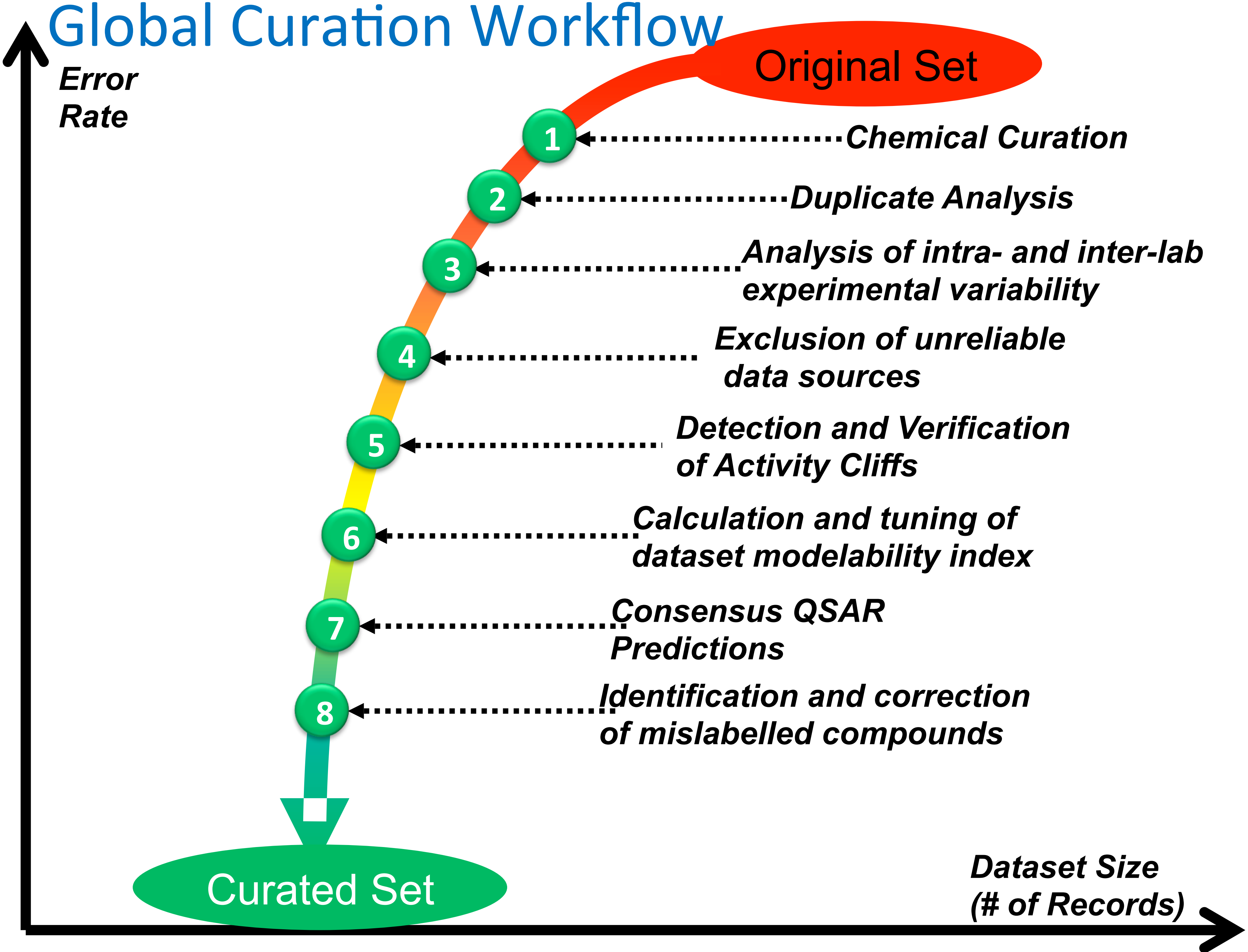


6604136



6604137

Global Curation Workflow



Original Set

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

Curated Set

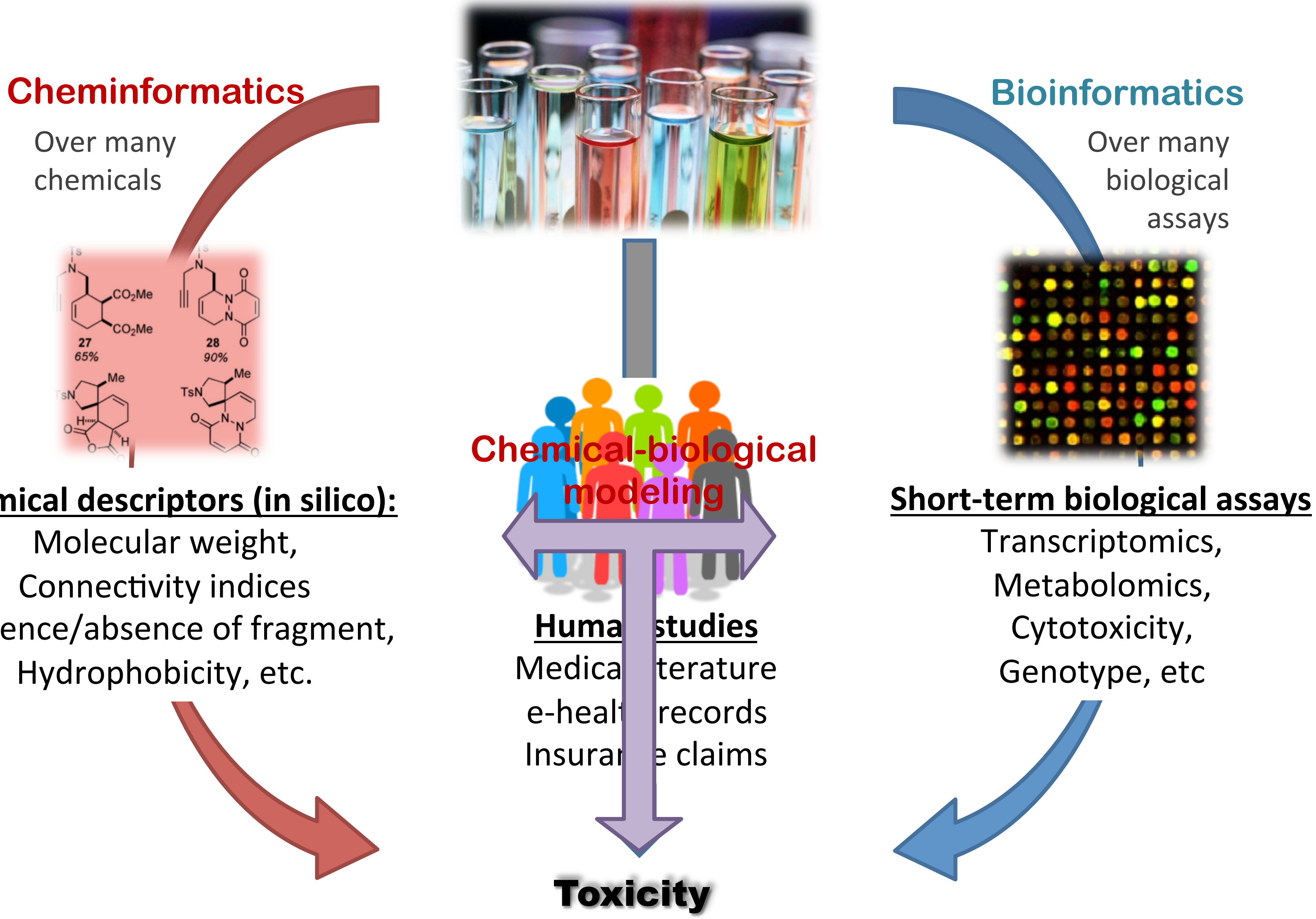
Notes on the importance of data curation

- The curation of chemical data is critical prior to any cheminformatics analysis and modeling. Difficult cases require human interventions and cannot be fully automated.

- Prediction outliers may be due to structural outliers, real activity cliffs or mislabeled compounds. Many of them can still be detected and removed prior to modeling studies boosting the reliability of QSAR model.

- Rigorously developed QSAR models can be even used to correct erroneous biological data associated with certain compounds.

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



Cheminformatics

Over many chemicals

Bioinformatics

Over many biological assays

Chemical-biological modeling

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

Human studies
Medical literature
e-health records
Insurance claims

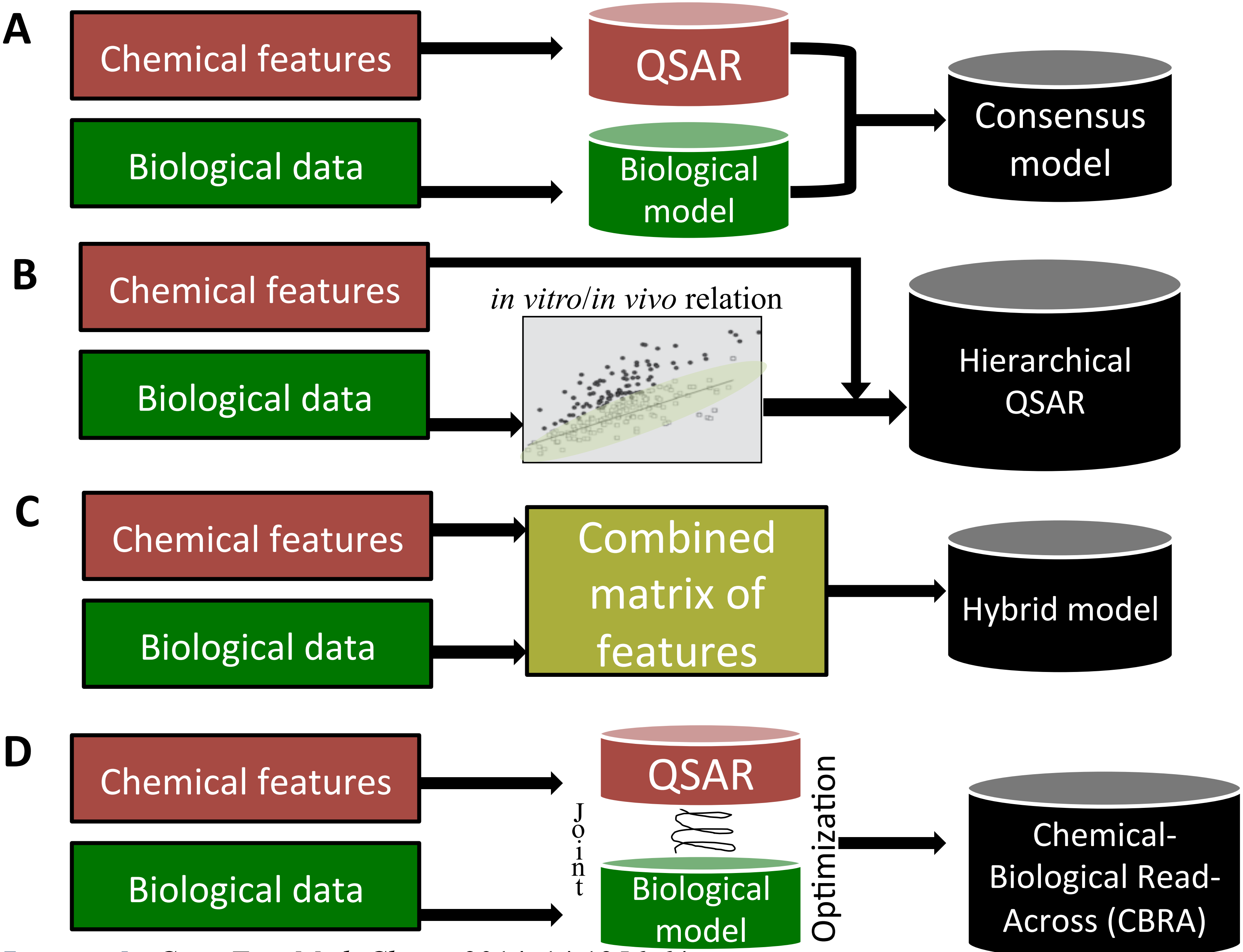
Short-term biological assays
Transcriptomics,
Metabolomics,
Cytotoxicity,
Genotype, etc

Toxicity

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

Approaches to Integrative QSAR Modeling



A

Chemical features → QSAR

Biological data → Biological model

QSAR + Biological model → Consensus model

B

Chemical features → Hierarchical QSAR

Biological data → *in vitro/in vivo* relation → Hierarchical QSAR

C

Chemical features → Combined matrix of features

Biological data → Combined matrix of features

Combined matrix of features → Hybrid model

D

Chemical features → QSAR

Biological data → Biological model

Joint

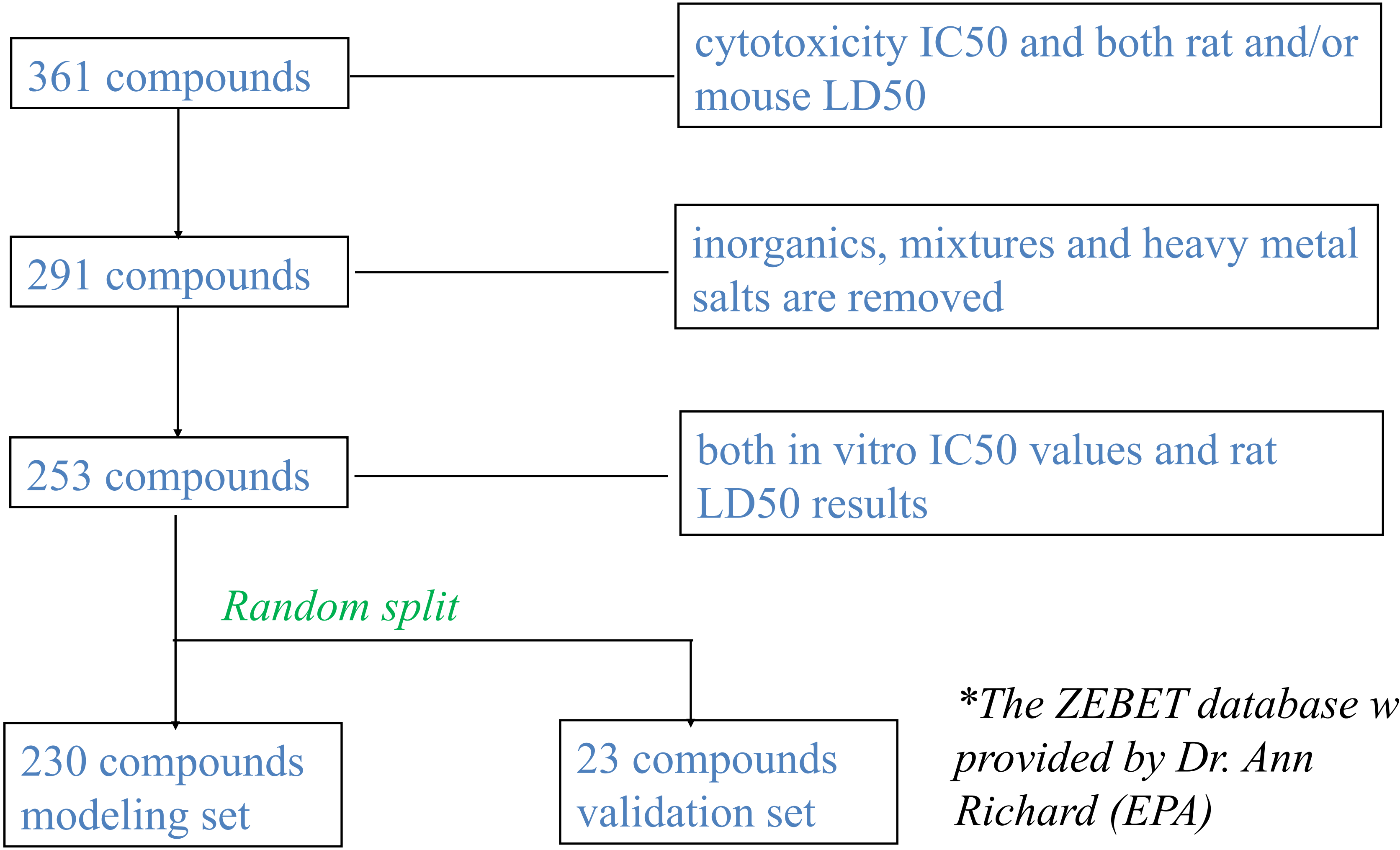
Optimization

→ Chemical-Biological Read-Across (CBRA)

Low et al,. *Curr. Top. Med. Chem.*, **2014**, 14:1356-64

Hierarchical QSAR: Using *in vitro* IC50 data to develop improved QSAR models for *in vivo* Rat Oral LD50.

ZEBET Database* and Data Preparation



361 compounds

cytotoxicity IC50 and both rat and/or mouse LD50

291 compounds

inorganics, mixtures and heavy metal salts are removed

253 compounds

both in vitro IC50 values and rat LD50 results

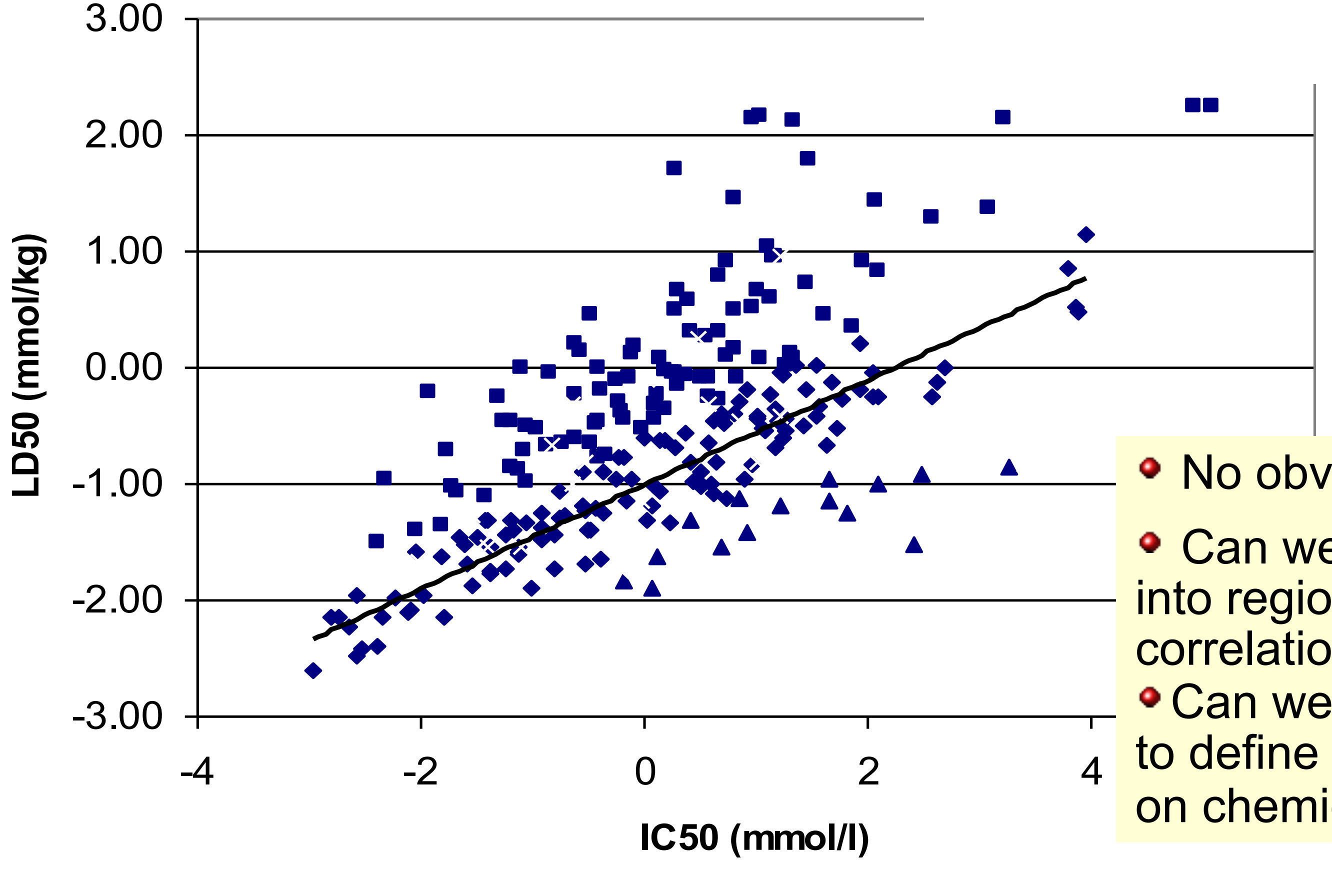
Random split

230 compounds modeling set

23 compounds validation set

**The ZEBET database was provided by Dr. Ann Richard (EPA)*

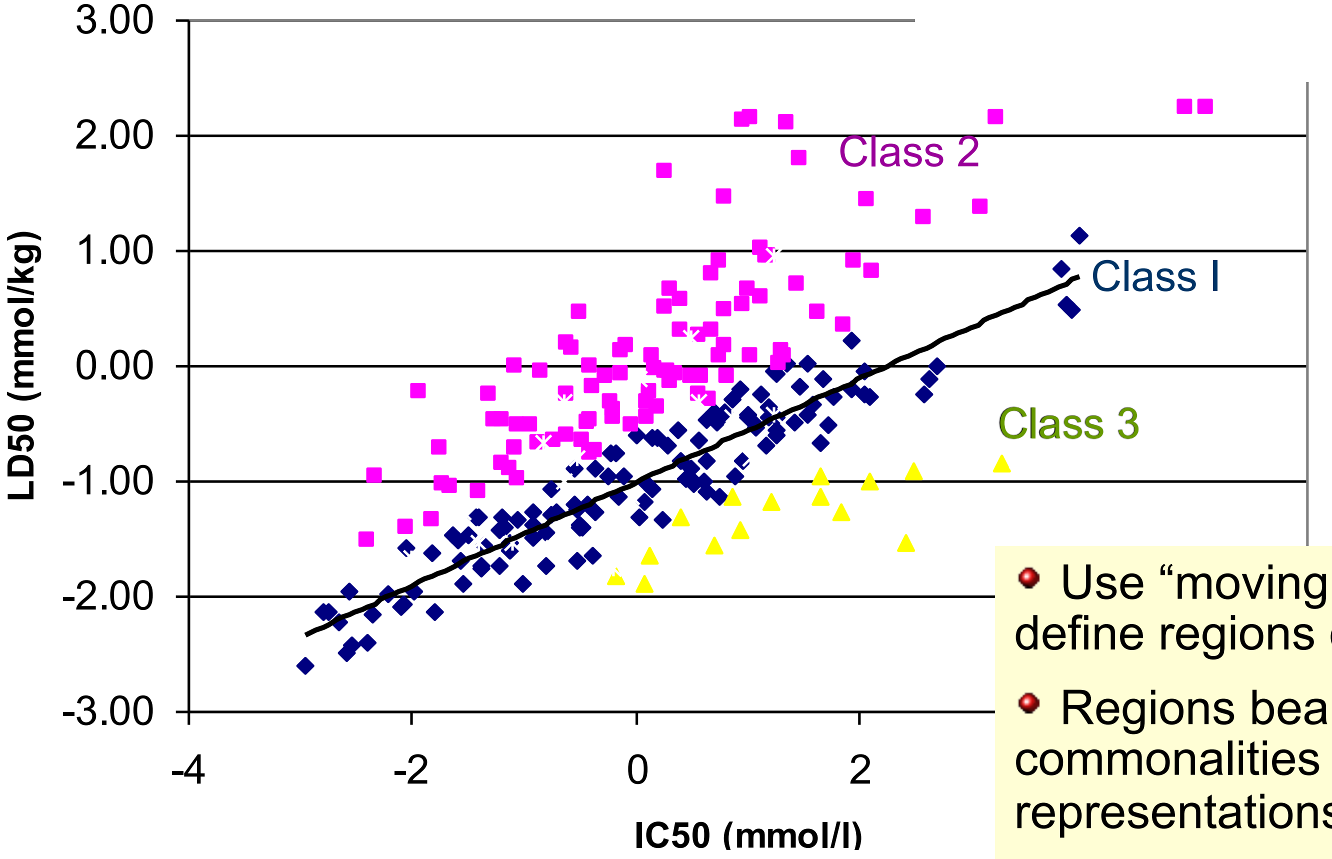
Relatively poor correlation between *in vitro* IC50 data and *in vivo* Rat Oral LD50



- No obvious correlation
- Can we break the problem into regions of higher correlation?
- Can we use QSAR methods to define those regions based on chemical structure alone?

Zhu H, Ye L, Richard A, Golbraikh A, Rusyn I, Tropsha A. (2009) EHP 117:1257-1264.

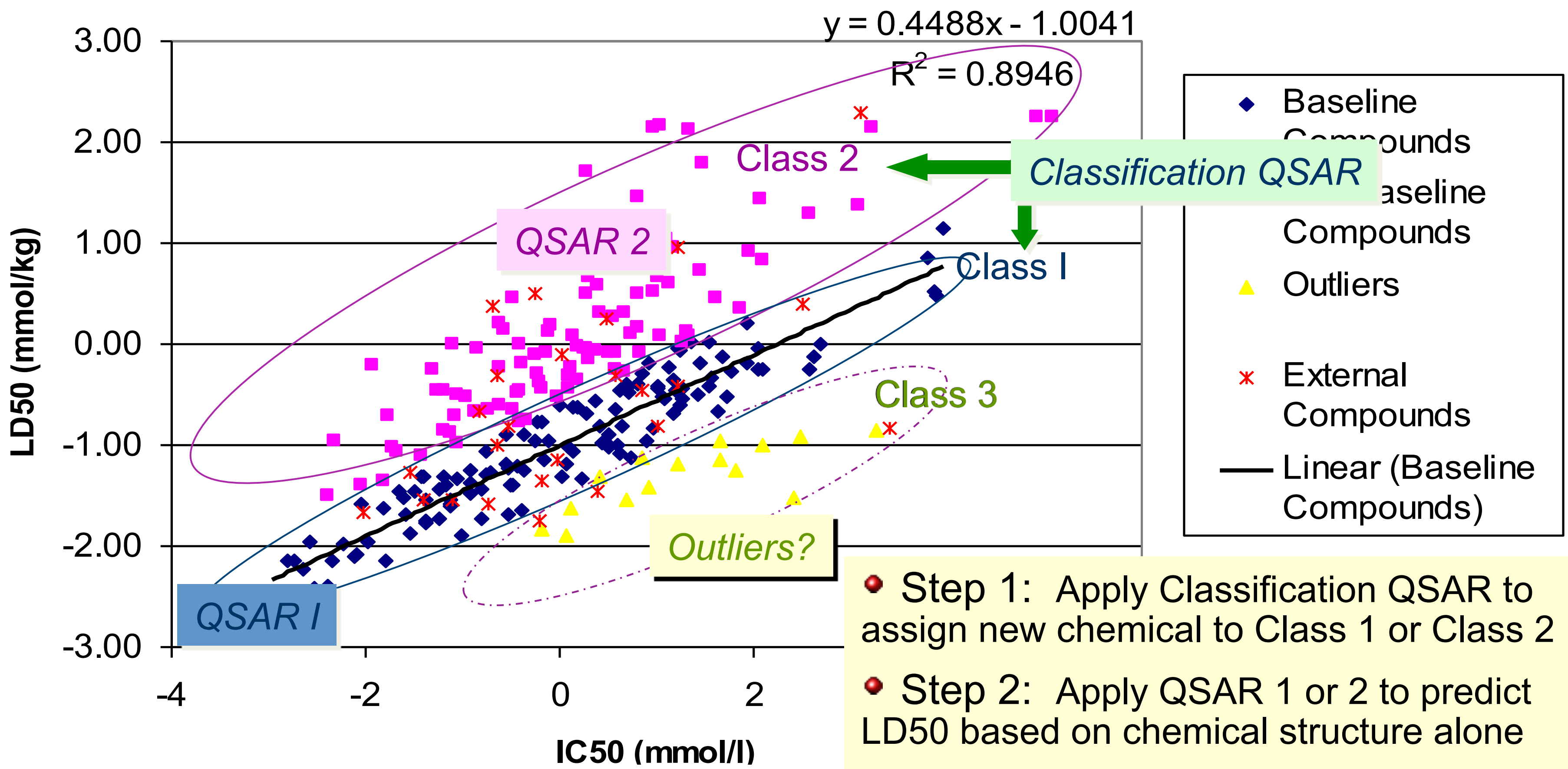
Different regions of *in vitro* IC50 - *in vivo* Rat Oral LD50 relationships



- Use "moving regression" to define regions of higher correlation
- Regions bear some commonalities to "baseline toxicity" representations
- Attempt to distinguish regions based on chemical structure alone

Zhu H, Ye L, Richard A, Golbraikh A, Rusyn I, Tropsha A. (2009) EHP 117:1257-1264.

Hierarchical QSAR modeling



Class 2

Classification QSAR

QSAR 2

Class I

Class 3

Outliers?

QSAR I

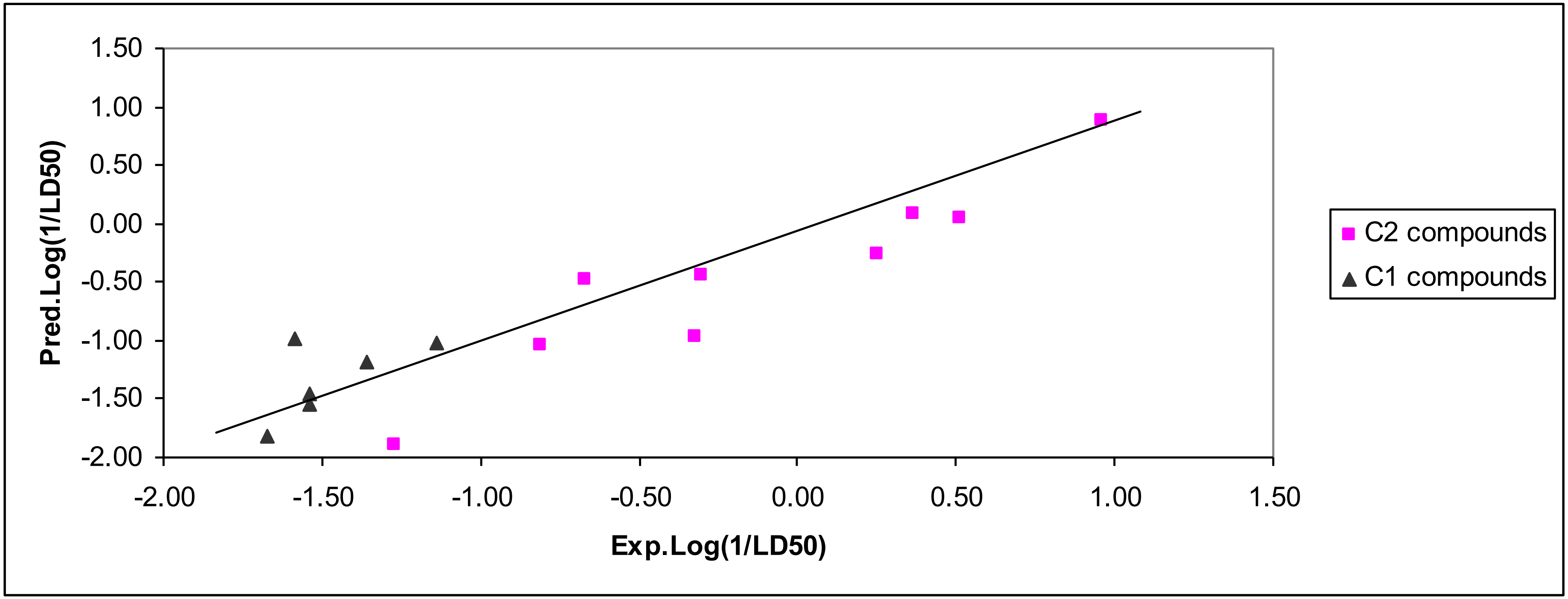
IC50 used to inform construction of QSARs, but not needed for prediction

- Step 1: Apply Classification QSAR to assign new chemical to Class 1 or Class 2
- Step 2: Apply QSAR 1 or 2 to predict LD50 based on chemical structure alone
- Step 3: Validate approach with external data

Zhu H, Ye L, Richard A, Golbraikh A, Rusyn I, Tropsha A. (2009) EHP 117:1257-1264.

Prediction of the Rat LD50 Values for the External set of 23 Compounds

- R^2=0.79, *MAE*=0.37, Coverage=74% (17 out of 23)



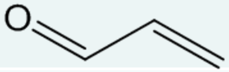
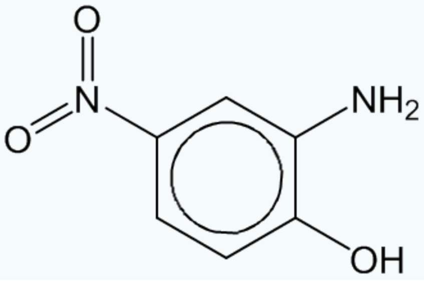
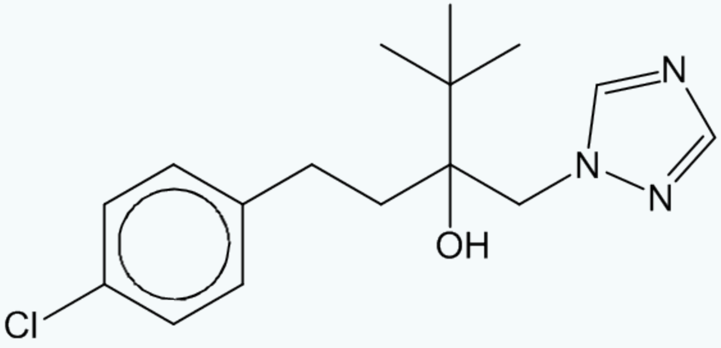
Hybrid QSAR: *In vitro* dose-response data improve the predictive power of QSAR models of in vivo toxicity (rat LD_{50})

- 1408 substances
- 382 chemical structure descriptors (Dragon v5.5)
- 13 *in vitro* NCGC cell viability assays * :
  - ◎ qHTS (quantitative HTS) data
  - ◎ 14 test concentrations: 0.6nm .. 92.2µm

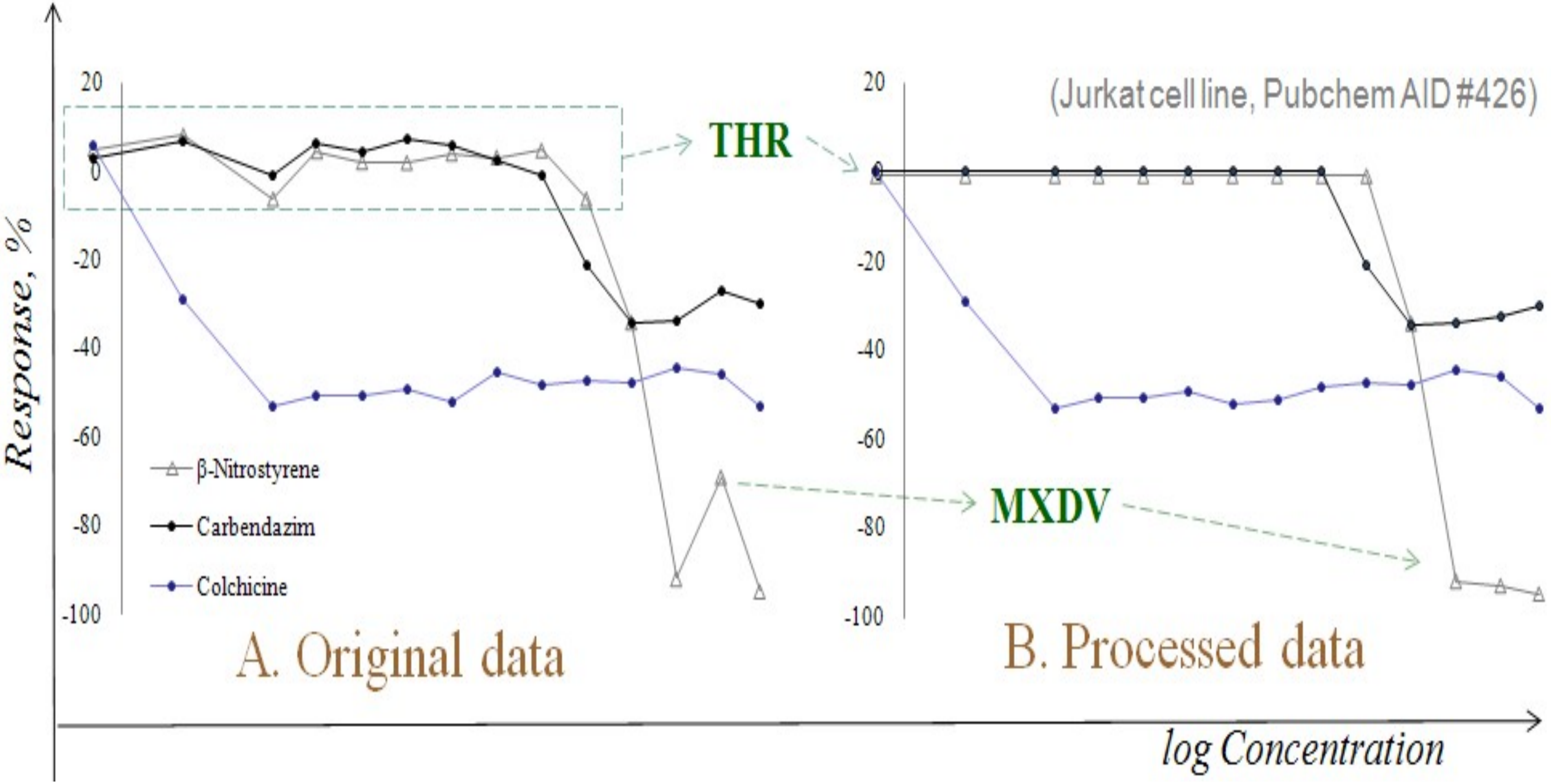
May yield up to 13x14 = 182 *in vitro* qHTS descriptors, but the issue of data noise becomes important.

*Inglese J., Douglas S. A. et al. *PNAS,* **2006**, v103(31), p11473

QSAR-like Table – qHTS descriptors

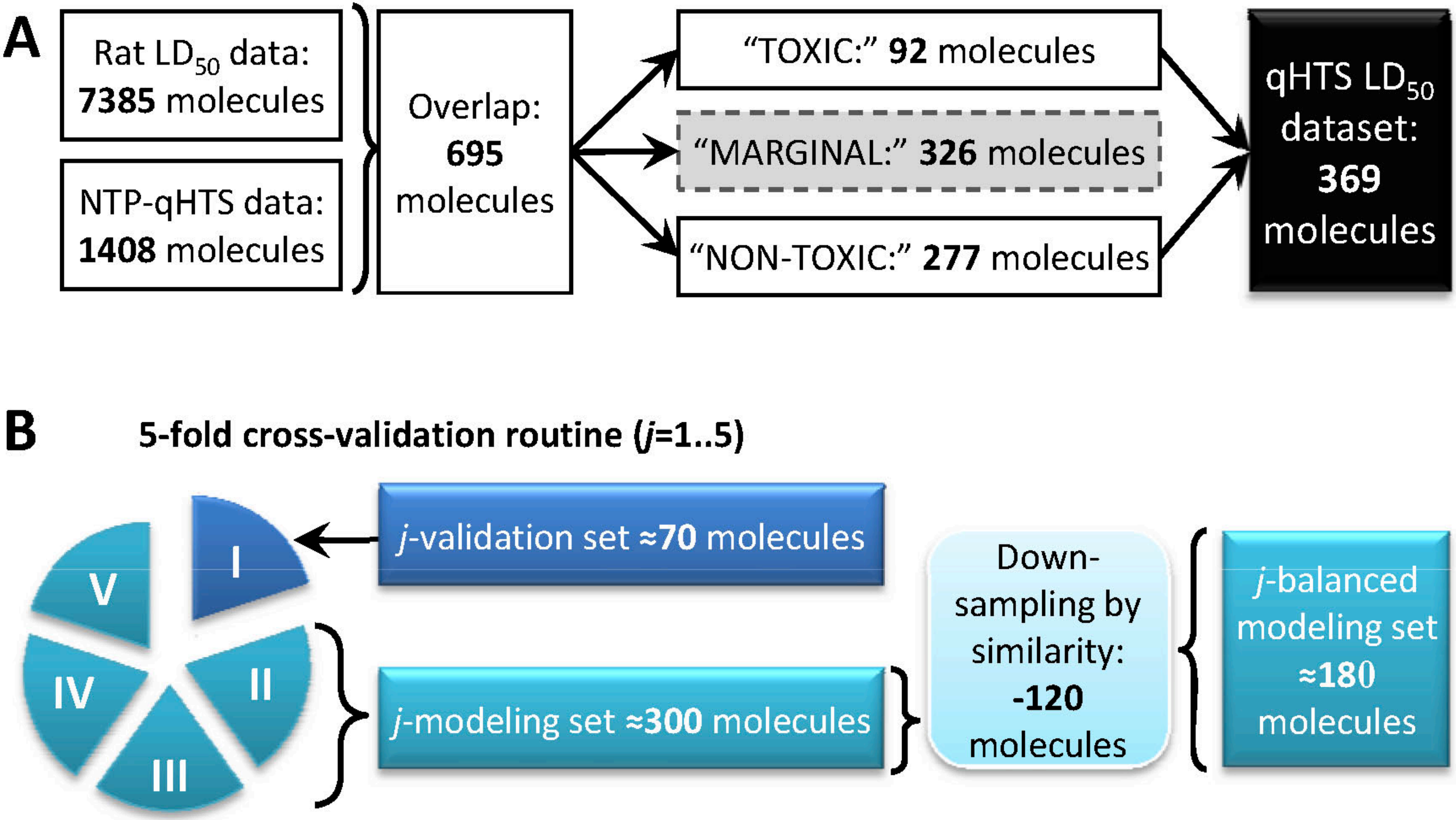
		Descriptor #:	*1*	*2*	*…*	*182*
ID	**Name**	**Structure**	**3T3 9.2mkM**	**3T3 21mkM**	**…**	**SHSY 92mkM**
1	Acrolein		0	0	…	-92
2	2-Amino-4-nitrophenol		0	-22	…	0
…	…	…	…	…	…	…
369	Tebuconazole		-21	-24	…	-18

SMOOTHING CONCENTRATION-RESPONSE CURVES (NOISE SUPPRESSION).



Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. EHP, 2011, 119(3):364-70

Modeling Workflow



A

Rat LD_{50} data: **7385** molecules

NTP-qHTS data: **1408** molecules

Overlap: **695** molecules

"TOXIC:" **92** molecules

"MARGINAL:" **326** molecules

"NON-TOXIC:" **277** molecules

qHTS LD_{50} dataset: **369** molecules

B

5-fold cross-validation routine (*j*=1..5)

j-validation set ≈**70** molecules

j-modeling set ≈**300** molecules

Down-sampling by similarity: **-120** molecules

j-balanced modeling set ≈**180** molecules

Smoothing the concentration-response data improves the prediction accuracy of hybrid models.

	%	Chemical descriptors only	Hybrid descriptors (Original)	Hybrid descriptors (THR=15%)
*k*NN models	*Sensitivity*	68±8	63±9	76±5
	Specificity	85±4	86±4	87±2
	CCR	**76 ±5 ***	**74 ±5**	**82 ±3**
Random Forest (RF) models	*Sensitivity*	74±9	66±8	77±10
	Specificity	82±7	87±4	86±3
	CCR	**78 ±4 ***	**77 ±5**	**82 ±5**

Shown are averaged results of five-fold external validation. *Chemical descriptors only models were significantly different ($p < 0.05$) from all other models of the corresponding group by the permutation test (10,000 times).

Hybrid QSAR models have higher predictive power than commercial software TOPKAT

%	TOPKAT	Chemical descriptors only		Hybrid descriptors (Original)		Hybrid descriptors (THR=15%)	
		kNN	*RF*	*kNN*	*RF*	*kNN*	*RF*
Sensitivity	0.45	0.73	0.73	0.55	0.82	0.91	0.91
Specificity	0.93	0.78	0.80	0.85	0.78	0.85	0.83
CCR	**0.69 ***	**0.75**	**0.77**	**0.70**	**0.80**	**0.88**	**0.87**

Results are shown for 52 compounds in our external validation sets, which were also absent in the TOPKAT training set.

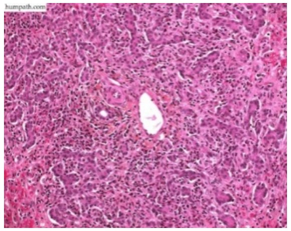
*TOPKAT model was significantly different ($p < 0.05$) from all other models by the permutation test (10,000 times).

Hybrid QSAR: Predicting subchronic hepatotoxicity using both chemical descriptors and 24h toxicogenomics profiles

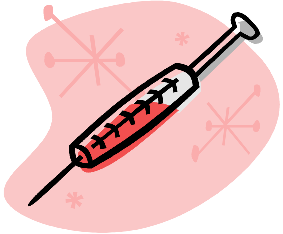
Rats in triplicates
6-8 weeks old
Sprague Dawley

Doses: low, med, high

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



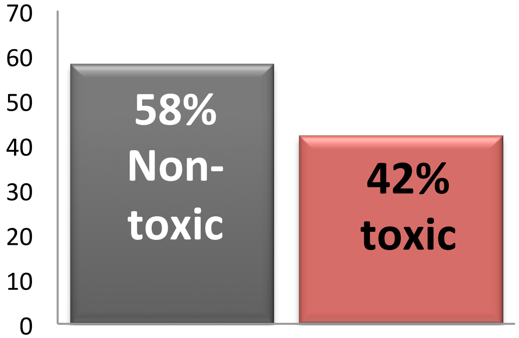
Liver histopathology



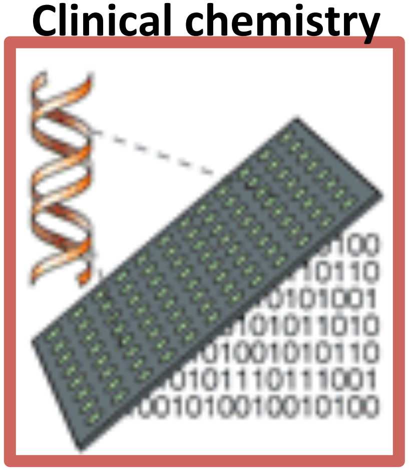
Clinical chemistry

Assigned by pathologists

127 compounds in 2 classes



Subchronic 28-day hepatotoxicity

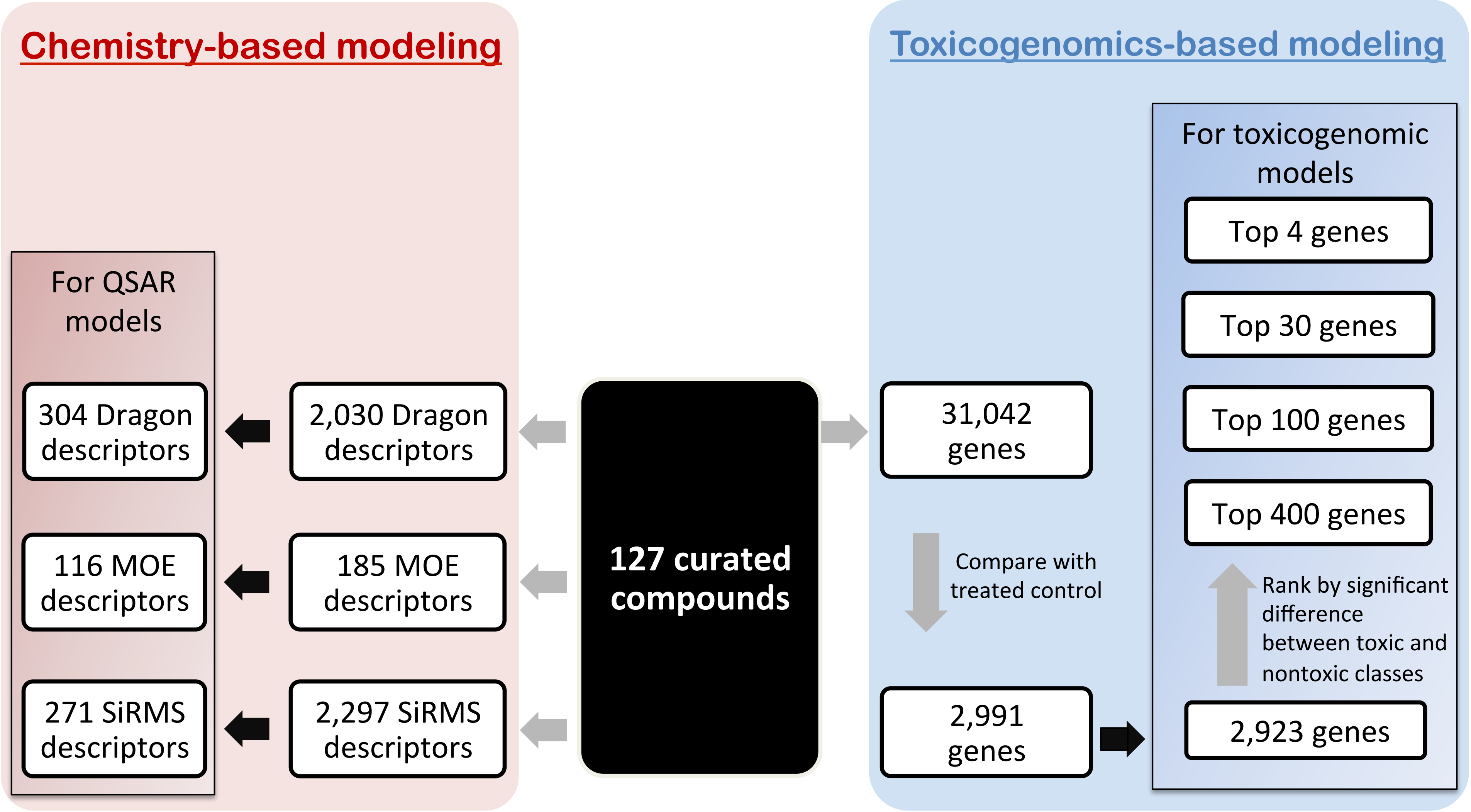


***In vivo* hepatic gene expression**
(24h, high dose)

Predict

Data source: Japanese toxicogenomics project; Open TG-GATEs http://toxico.nibio.go.jp/

Selection of chemical descriptors and transcripts for model building



Chemistry-based modeling

For QSAR models

304 Dragon descriptors

116 MOE descriptors

271 SiRMS descriptors

2,030 Dragon descriptors

185 MOE descriptors

2,297 SiRMS descriptors

127 curated compounds

Toxicogenomics-based modeling

31,042 genes

Compare with treated control

2,991 genes

2,923 genes

Rank by significant difference between toxic and nontoxic classes

For toxicogenomic models

Top 4 genes

Top 30 genes

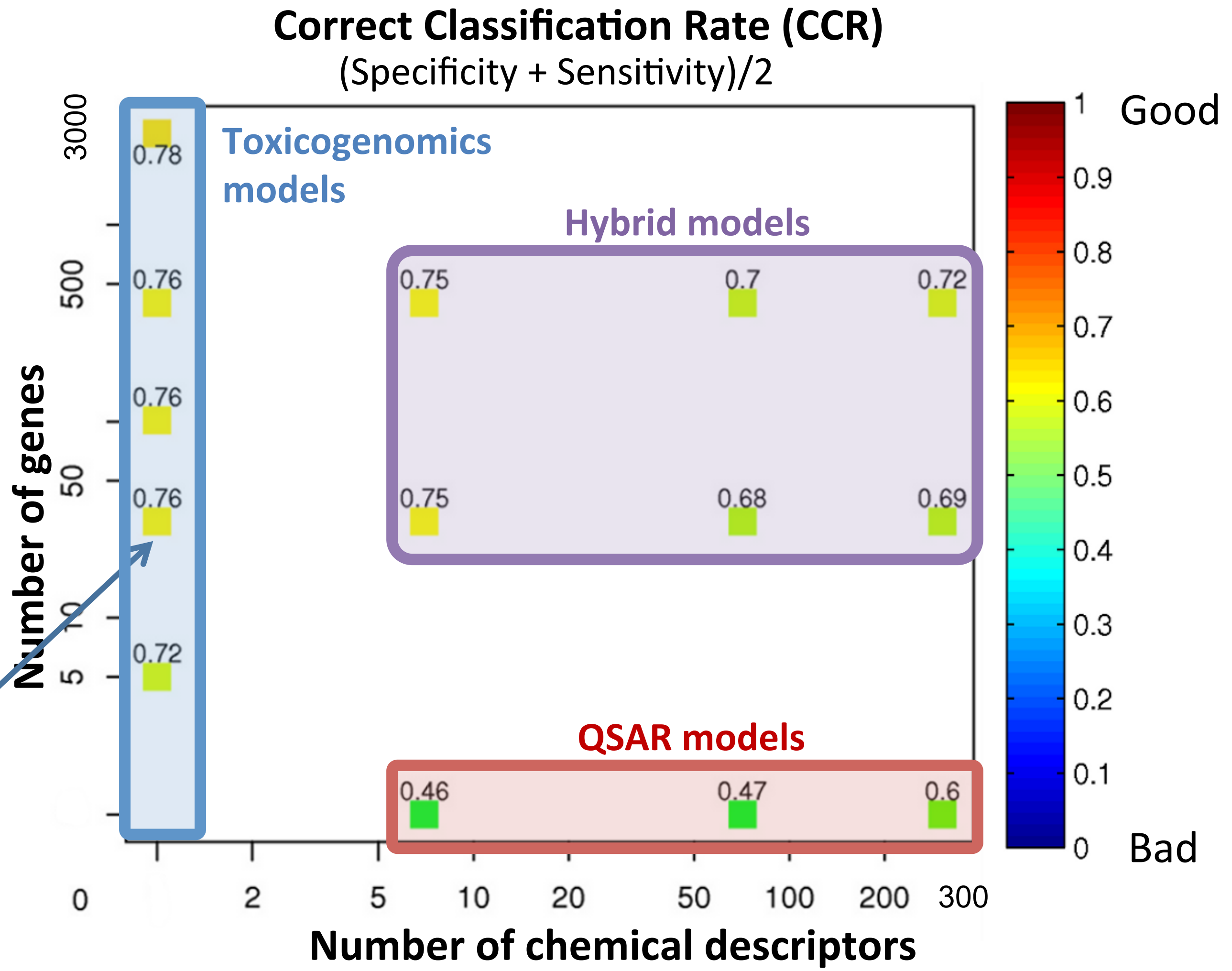
Top 100 genes

Top 400 genes

➡ Removal of low-variance and highly correlated descriptors

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

Comparison of models



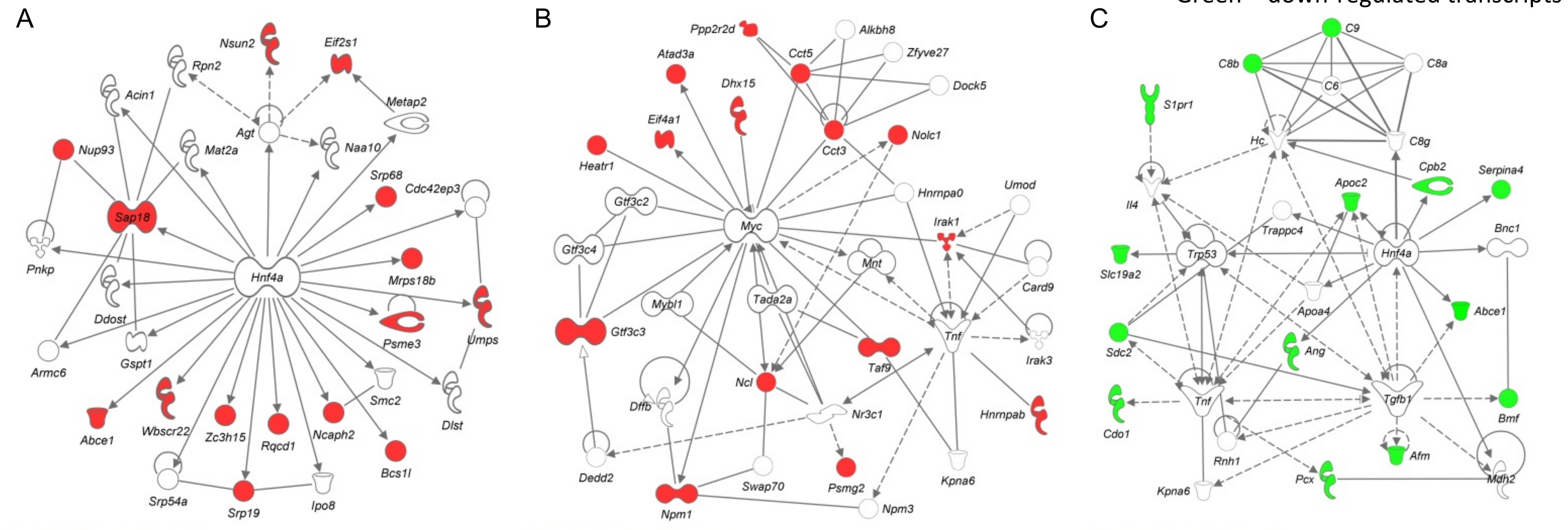
Optimal model with 85 genes

Toxicogenomics models > Hybrid models > QSAR models

Model interpretation (biology): Pathway analysis shows that selected genes are mechanistically relevant

Networks were generated by IPA (Ingenuity)

Red = up-regulated transcripts
Green = down-regulated transcripts



Hnf4a is assoc. with

- Morphological and functional differentiation of hepatocytes
- Liver architecture
- ER stress (Parviz 2003, Watt 2003, Luebke-Wheeler 2008)

Myc is assoc. with

- Cell proliferation
- Cell differentiation
- Apoptosis

(Lin 2009)

Cellular function- and maintenance-related interactomes

Model Interpretation (chemistry) Significant chemical descriptors are interpreted in the form of structural alerts

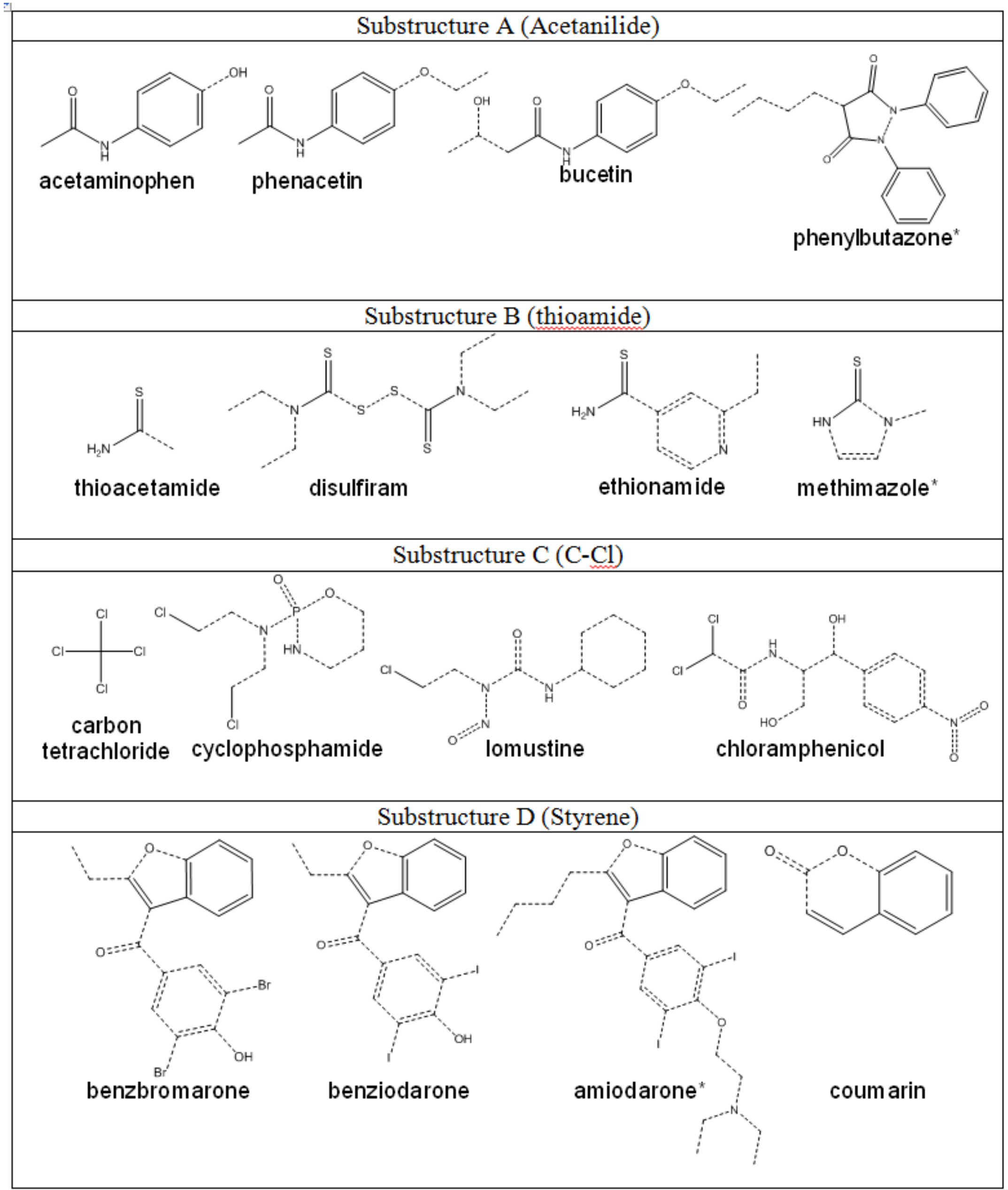
Toxic species

N-hydroxylamines
Nitroso compounds

sulfur species

Alkyl radicals

Epoxides



Substructure A (Acetanilide)

acetaminophen, phenacetin, bucetin, phenylbutazone*

Substructure B (thioamide)

thioacetamide, disulfiram, ethionamide, methimazole*

Substructure C (C-Cl)

carbon tetrachloride, cyclophosphamide, lomustine, chloramphenicol

Substructure D (Styrene)

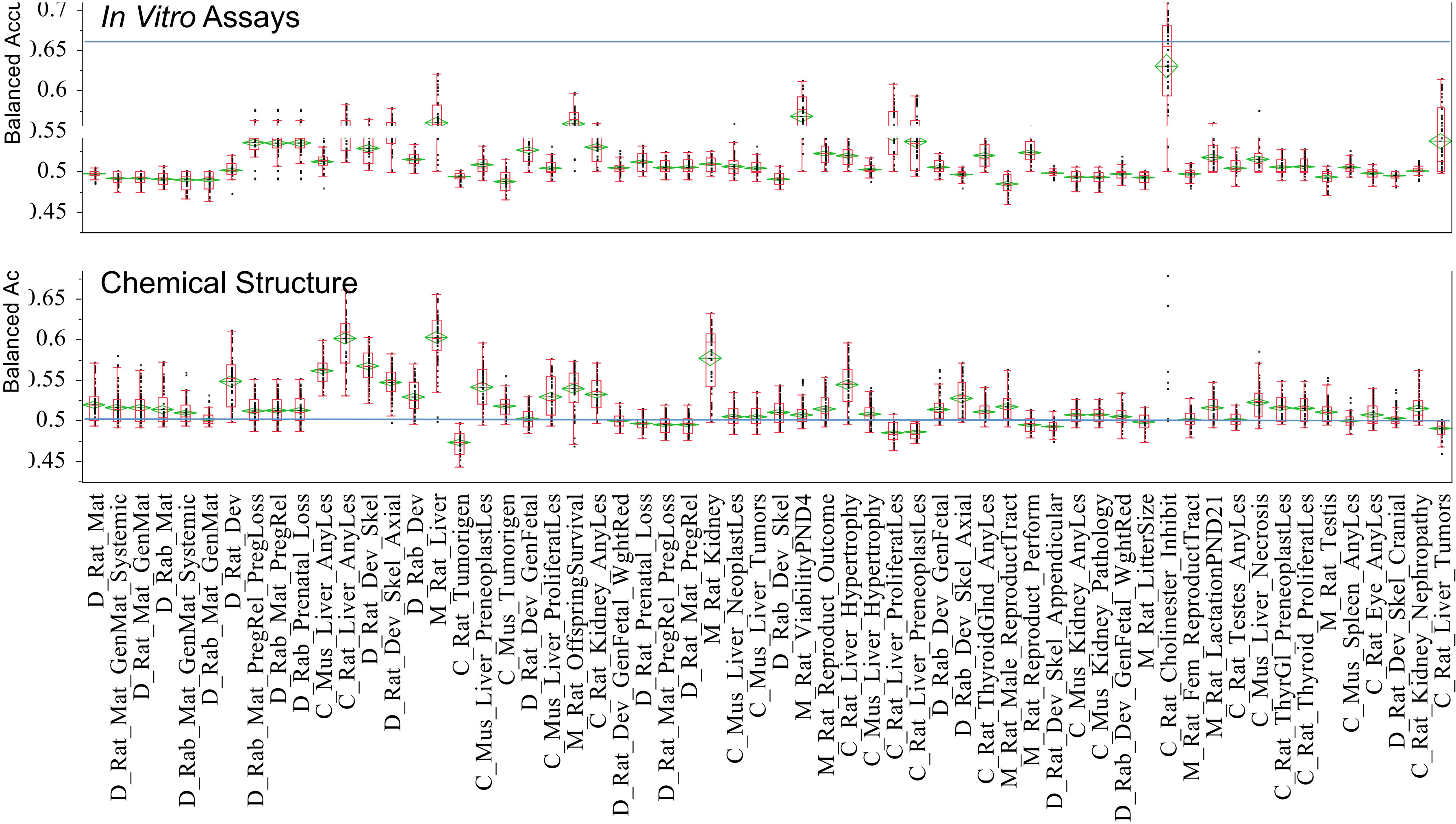
benzbromarone, benziodarone, amiodarone*, coumarin

Why is gene expression more predictive than chemical descriptors?

- Small and chemically diverse data set
  - Too few congeneric compounds is a challenge for QSAR
- Effect of activity cliffs
  - 50% of top 40 nearest neighbor pairs in chemistry space are activity cliffs
  - 33% of top 40 nearest neighbor pairs in biology space are activity cliffs

Dataset Modelability: does it make sense to model any SAR data?

Example: Poor *structure – in vivo* or *in vitro-in vivo* correlations for Toxcast data*



Toxicol Sci. 2012 Aug;128(2):398-417.

The Concept of Modelability

- We often fail to build a predictive QSAR model. However, it may be possible to evaluate ***modelability*** of the dataset prior to QSAR study.
- MODI-index: Balanced accuracy (BA) of a kNN model with K=1 (the activity class of each compound is predicted to be the same as that of its nearest chemical neighbor)

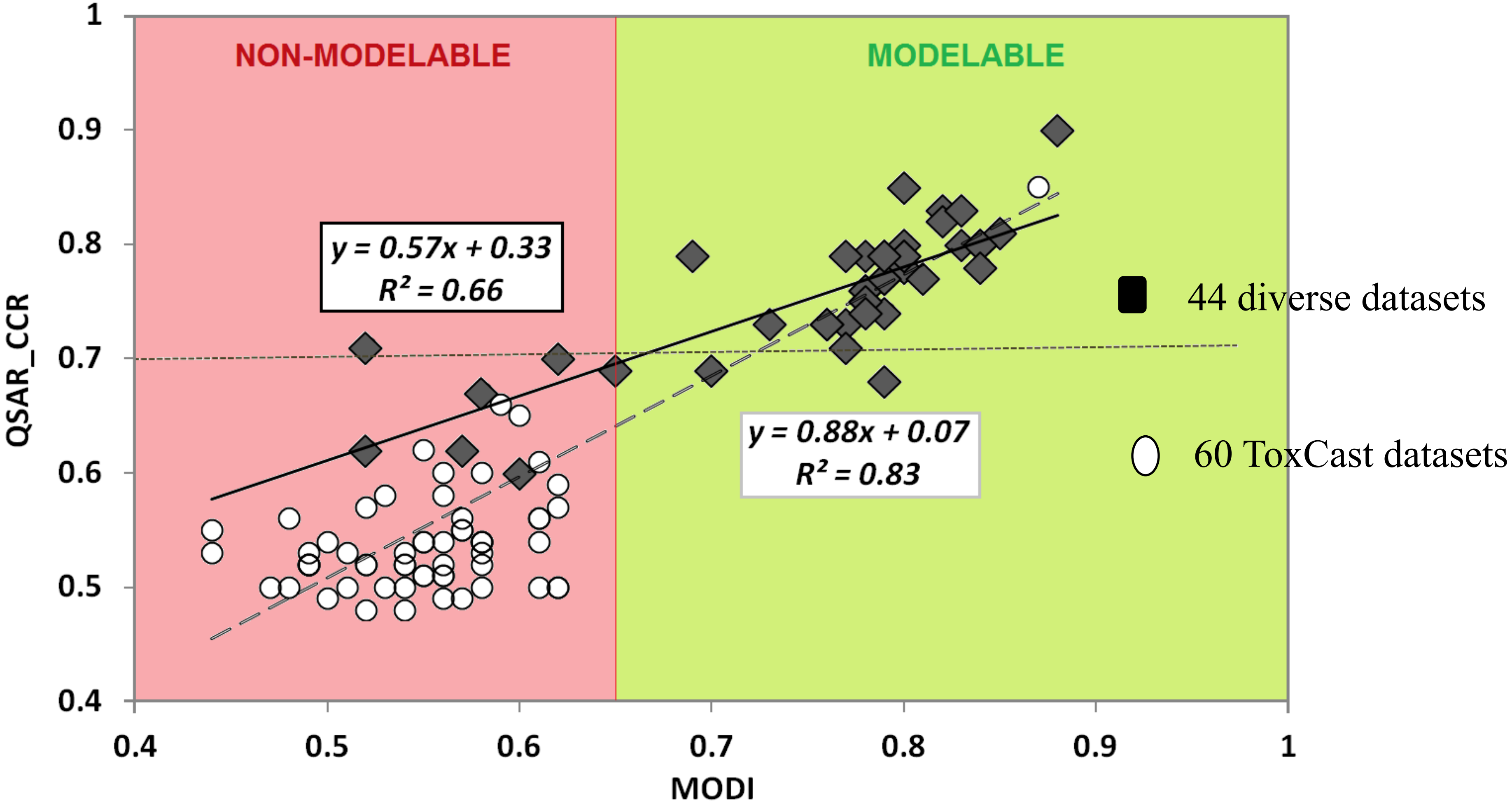
$$SE = N_{00}/N_0$$
$$SP = N_{11}/N_1$$

$$BA = \tfrac{1}{2}\,(SE + SP)$$

CONFUSION MATRIX

PREDICTED	OBSERVED CLASS 0	OBSERVED CLASS 1	TOTAL
CLASS 0	N_{00}	N_{10}	$N_{.0}$
CLASS 1	N_{01}	N_{11}	$N_{.1}$
TOTAL	$N_{0.}=N_0$	$N_{1.}=N_1$	$N_{..}=N$

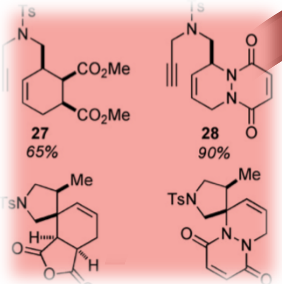
Prediction of Dataset Modelability



Golbraikh A, et al. Data Set Modelability by QSAR. *J Chem Inf Model*. **2014,** 54(1):1-4

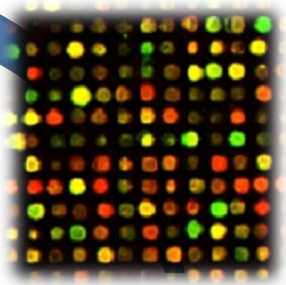
QSAR models < Toxicogenomics models

Data source: TGP2 Toxicogenomics Informatics Project in Japan



Chemical descriptors

127 drugs



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

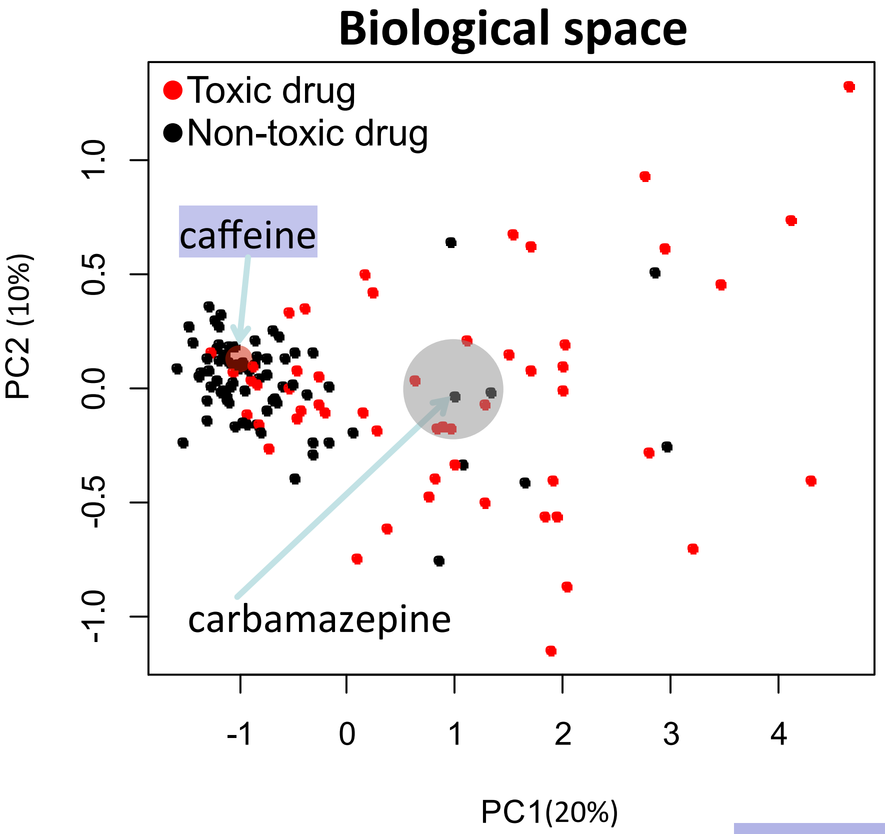
69-78% BAcc

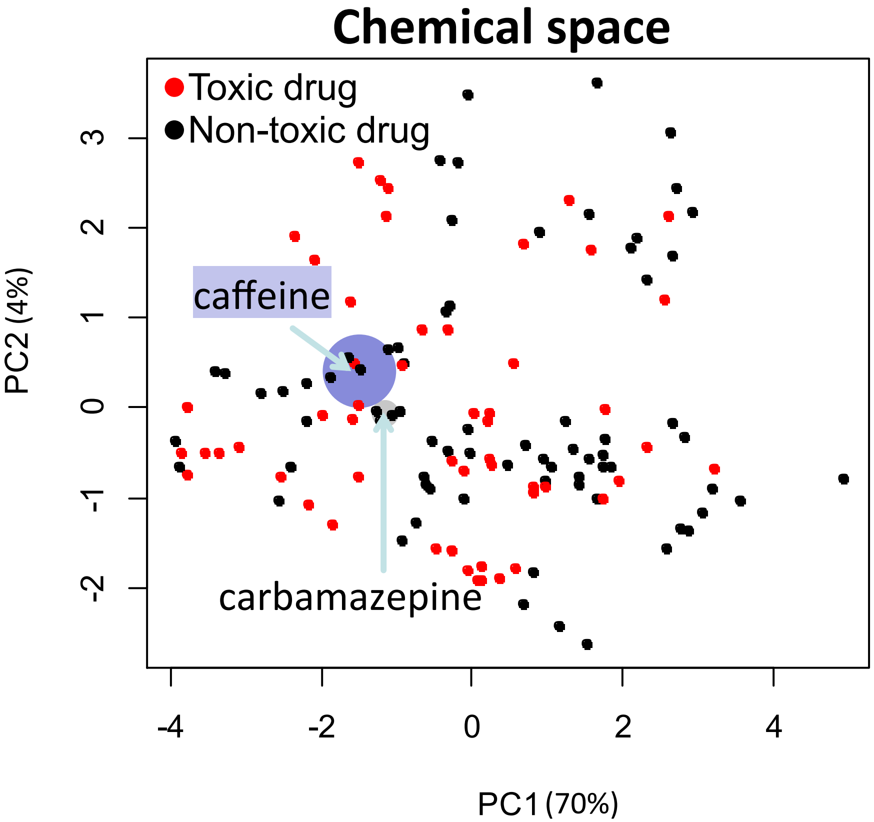
Hepatotoxicity (28 day)

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

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

Conflicting Predictions by QSAR and Toxicogenomics Models





<u>Carbamazepine</u>

☒Distant biological neighbors

☑Close chemical neighbors

=> Chemical similarity works better

<u>Caffeine</u>

☑Close biological neighbors

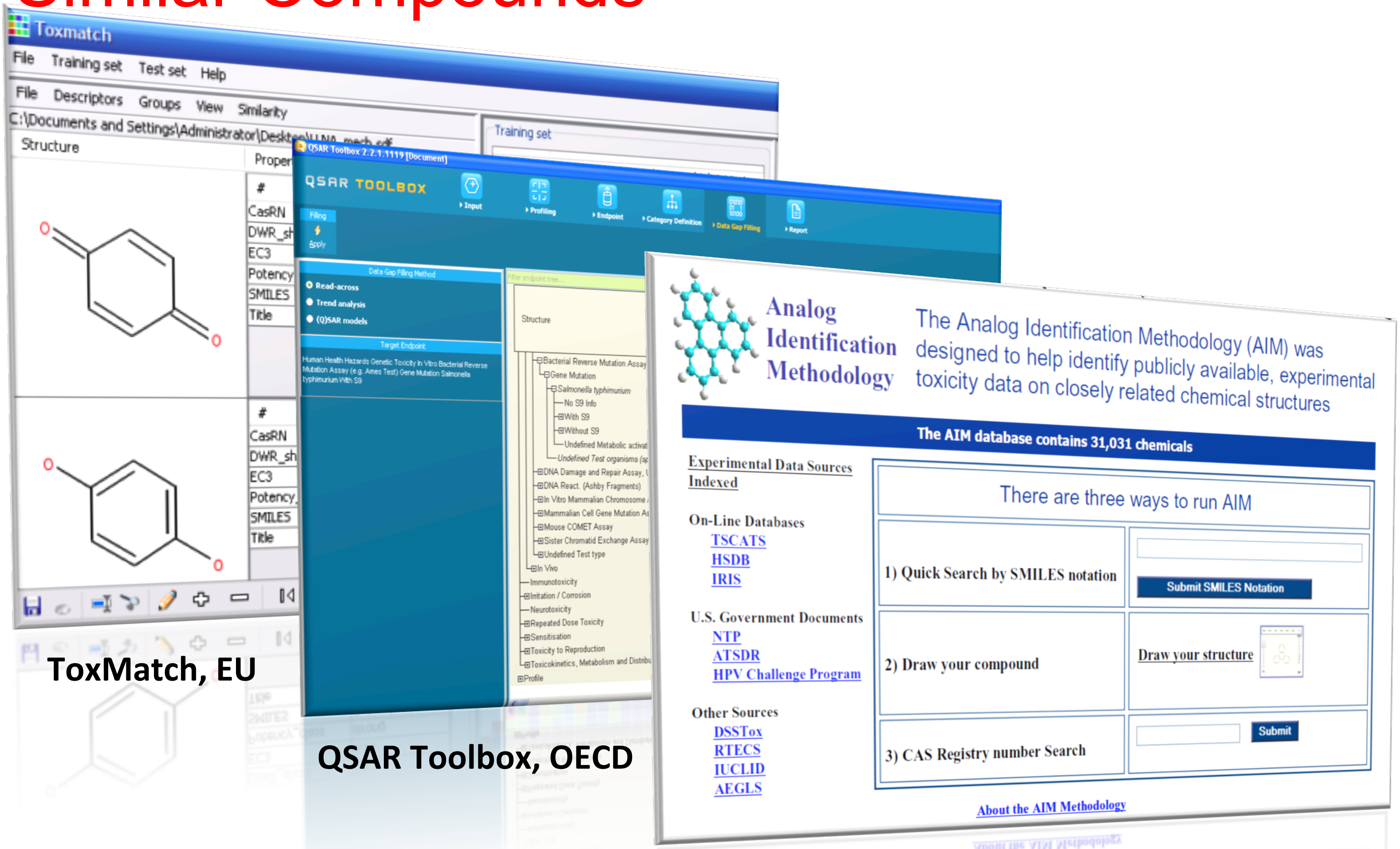
☒Distant chemical neighbors

=> TGx similarity works better

<u>Improved predicion:</u>

Learn from both sets of neighbors

Chemical Read-Across: Learning from Similar Compounds



Human Health Hazards Genetic Toxicity In Vitro Bacterial Reverse Mutation Assay (e.g. Ames Test) Gene Mutation Salmonella typhimurium With S9

Analog Identification Methodology

The Analog Identification Methodology (AIM) was designed to help identify publicly available, experimental toxicity data on closely related chemical structures

The AIM database contains 31,031 chemicals

Experimental Data Sources Indexed

On-Line Databases
TSCATS
HSDB
IRIS

U.S. Government Documents
NTP
ATSDR
HPV Challenge Program

Other Sources
DSSTox
RTECS
IUCLID
AEGLS

There are three ways to run AIM

1) Quick Search by SMILES notation — Submit SMILES Notation

2) Draw your compound — Draw your structure

3) CAS Registry number Search — Submit

About the AIM Methodology

ToxMatch, EU

QSAR Toolbox, OECD

AIM, US EPA/OPPT

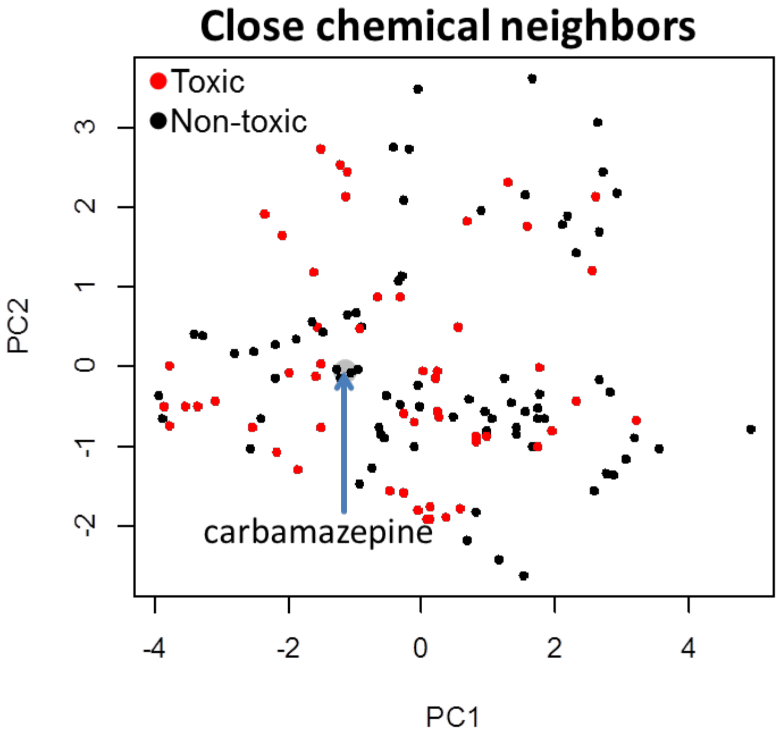
Chemical-biological read-across (CBRA): learning from both sets of neighbors

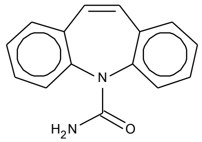
A_{pred}=similarity-weighted average of toxicity values

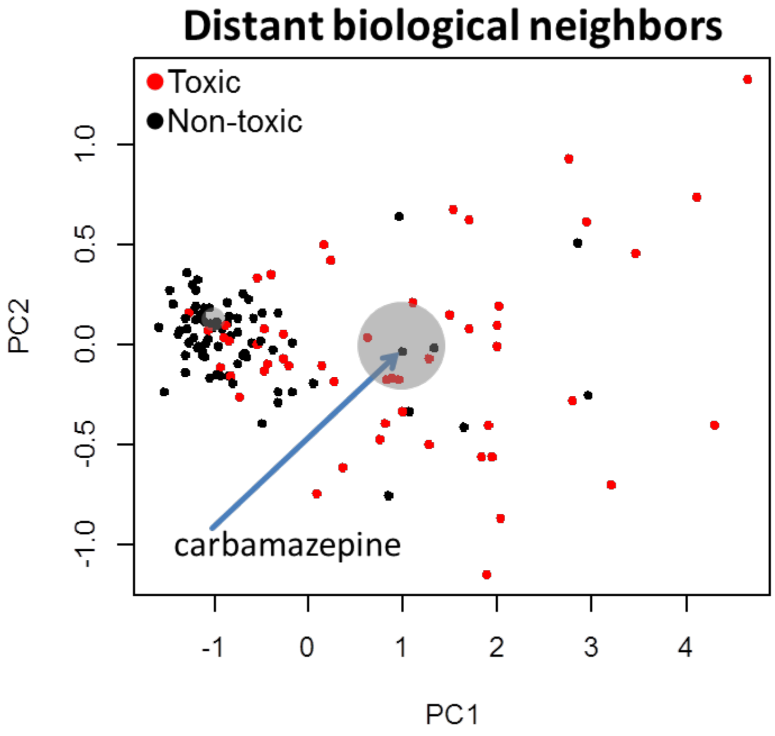
overall correctly predicted as nontoxic

CARBAMAZEPINE

Non-toxic


Close chemical neighbors




Distant biological neighbors

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

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

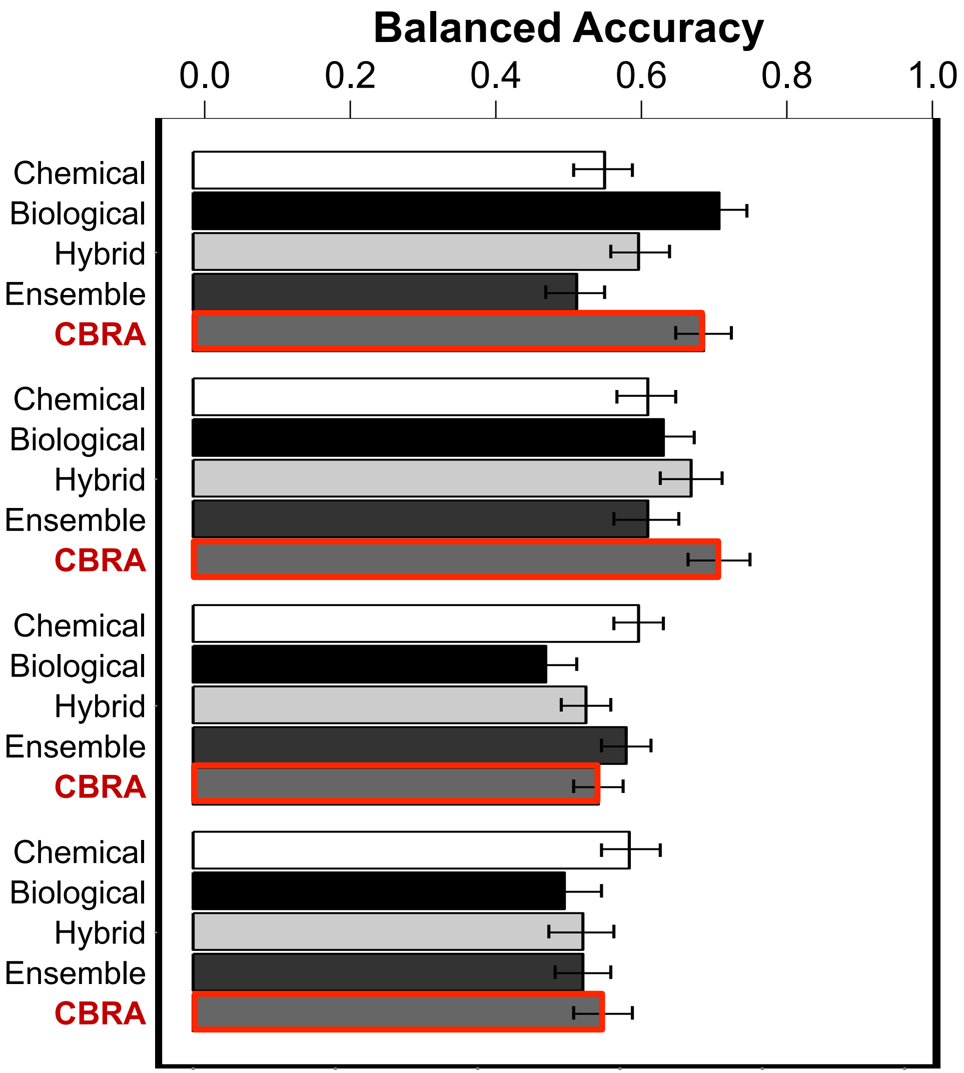
CBRA Shows Consistently Top Performance for Four Benchmark Data Sets

Rat Hepatotoxicity
127 compounds
85 genes

Rat Hepatocarcinogenicity
132 compounds
200 genes

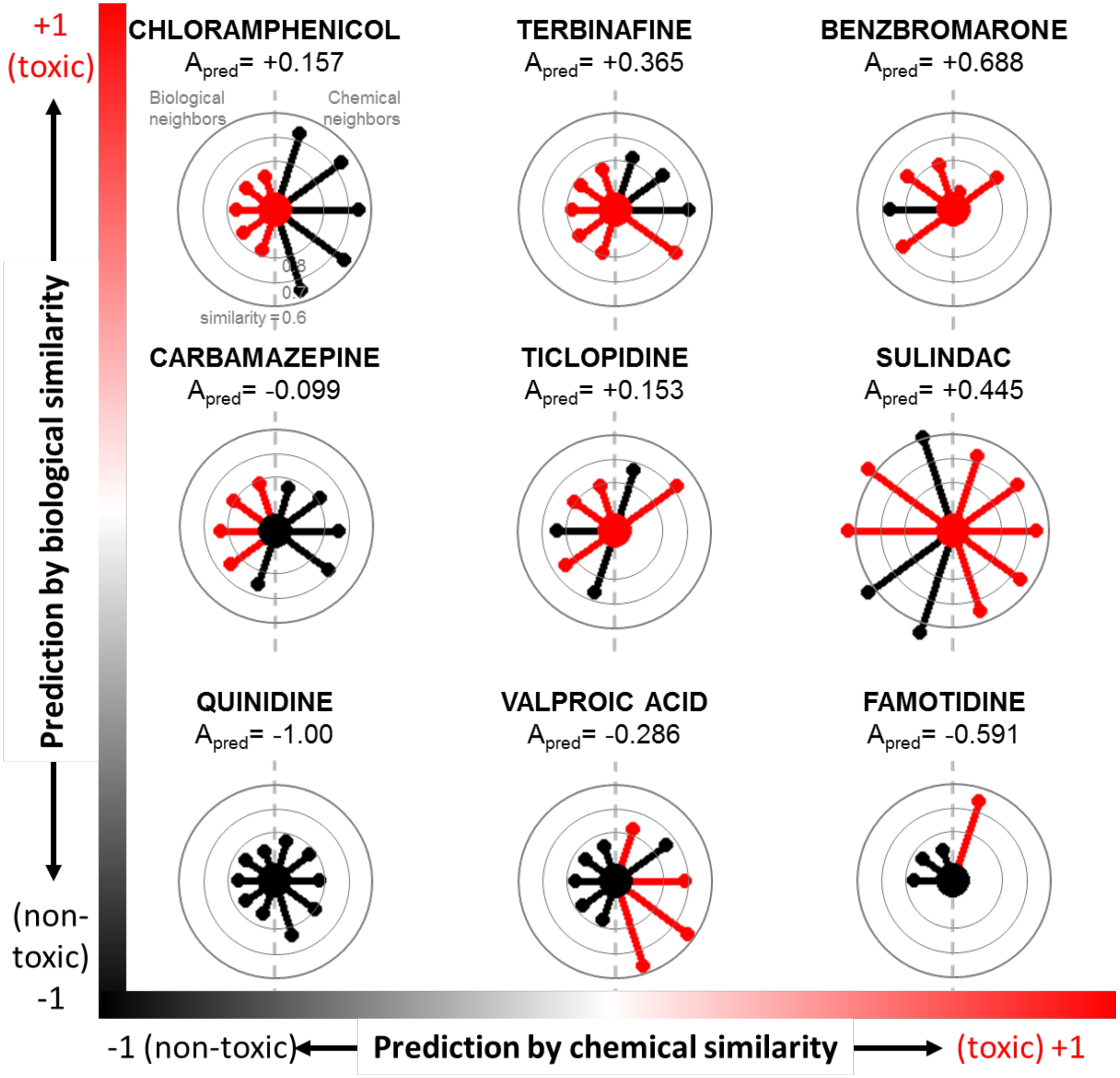
Mutagenicity (Ames Test)
185 compounds
148 cytotoxicity assays

Rat Acute Toxicity (Oral LD_{50})
122 compounds
148 cytotoxicity assays



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

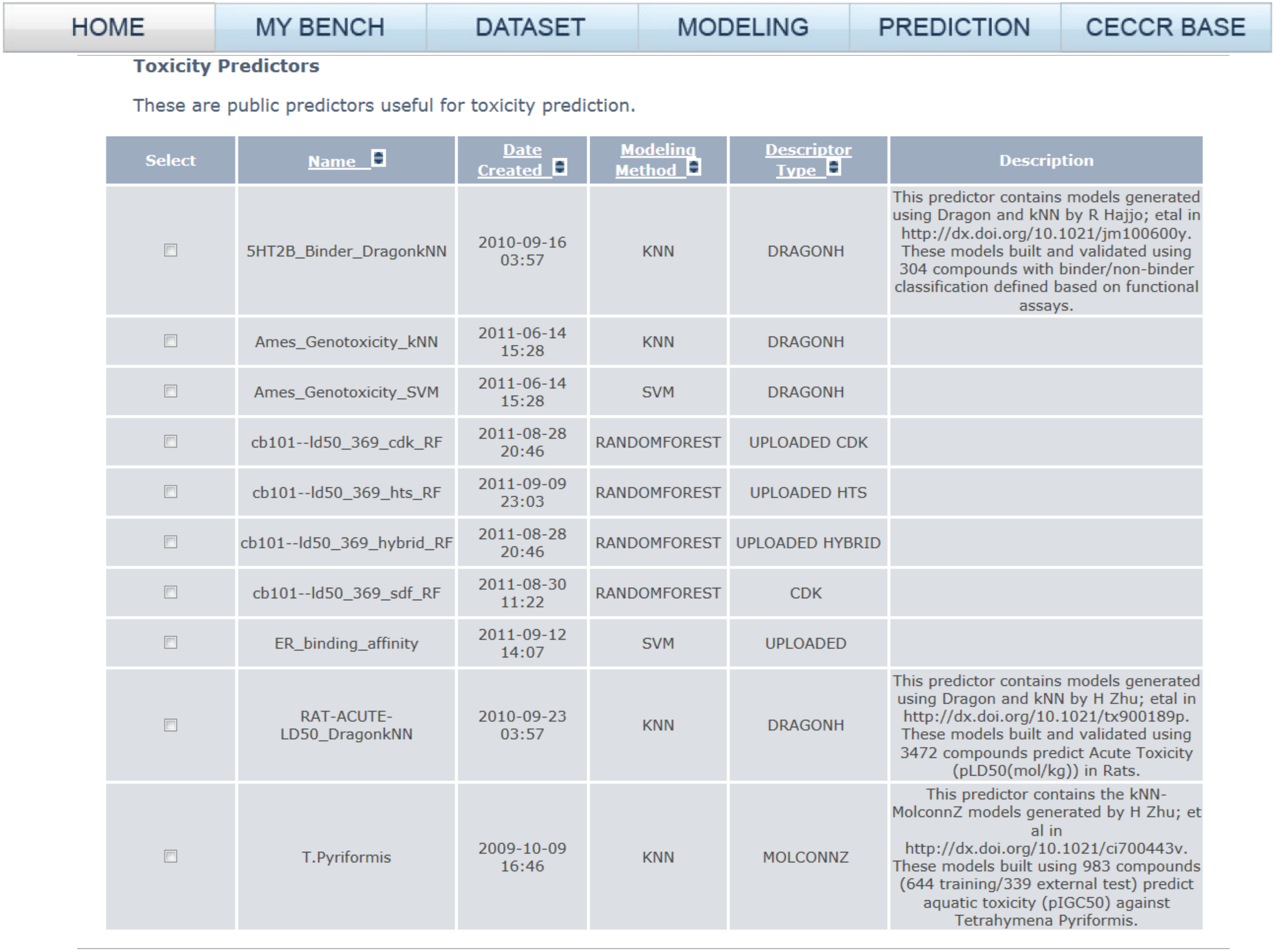


Low et al, Chem Res Toxicol. 2013, 26(8):1199-208

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
- Tool and data sharing
  - Pubic web portals (e.g., Chembench, OCHEM)

CHEM BENCH **http://chembench.mml.unc.edu**



Toxicity Predictors

These are public predictors useful for toxicity prediction.

Select	Name	Date Created	Modeling Method	Descriptor Type	Description
☐	5HT2B_Binder_DragonkNN	2010-09-16 03:57	KNN	DRAGONH	This predictor contains models generated using Dragon and kNN by R Hajjo; etal in http://dx.doi.org/10.1021/jm100600y. These models built and validated using 304 compounds with binder/non-binder classification defined based on functional assays.
☐	Ames_Genotoxicity_kNN	2011-06-14 15:28	KNN	DRAGONH	
☐	Ames_Genotoxicity_SVM	2011-06-14 15:28	SVM	DRAGONH	
☐	cb101--ld50_369_cdk_RF	2011-08-28 20:46	RANDOMFOREST	UPLOADED CDK	
☐	cb101--ld50_369_hts_RF	2011-09-09 23:03	RANDOMFOREST	UPLOADED HTS	
☐	cb101--ld50_369_hybrid_RF	2011-08-28 20:46	RANDOMFOREST	UPLOADED HYBRID	
☐	cb101--ld50_369_sdf_RF	2011-08-30 11:22	RANDOMFOREST	CDK	
☐	ER_binding_affinity	2011-09-12 14:07	SVM	UPLOADED	
☐	RAT-ACUTE-LD50_DragonkNN	2010-09-23 03:57	KNN	DRAGONH	This predictor contains models generated using Dragon and kNN by H Zhu; etal in http://dx.doi.org/10.1021/tx900189p. These models built and validated using 3472 compounds predict Acute Toxicity (pLD50(mol/kg)) in Rats.
☐	T.Pyriformis	2009-10-09 16:46	KNN	MOLCONNZ	This predictor contains the kNN-MolconnZ models generated by H Zhu; et al in http://dx.doi.org/10.1021/ci700443v. These models built using 983 compounds (644 training/339 external test) predict aquatic toxicity (pIGC50) against Tetrahymena Pyriformis.

Acknowledgements

Research Professors

Alexander Golbraikh, Denis Fourches, Eugene Muratov

Postdoctoral Fellows

Olexander Isayev,
Regina Politi

Collaborators

Ivan Rusyn
Diane Pozefsky

Adjunct Members

Weifan Zheng, Shubin Liu

Graduate students

Andrew Fant,
Yen Low
Mary La

MAJOR FUNDING

NIH

- R01-GM66940
- R01-GM068665

NSF

- ABI 9179-1165

EPA (STAR awards)

- RD832720
- RD833825
- RD834999

ONR