UNC
ESHELMAN
SCHOOL OF PHARMACY

MML
UNC.EDU

State of the art in QSAR modeling

Alexander Tropsha

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

OUTLINE

MML UNC.EDU

1. Introductory notes on QSAR and modern data streams: Chemical Structure – *in vitro* – *in vivo* data continuum

2. Novel workflows for hybrid and integrative QFAR Modeling (with examples of applications):
   -Hierarchical modeling based on in vitro – in vivo Correlations
   - Hybrid chemical-biological descriptors
   - Consensus between chemical and biological neighbors

3. Conclusions and outlook: QSAR modeling workflows and the use of QSAR for decision support

QSAR modeling progression

MML UNC.EDU

- Experimental Data = pain
  - Structure
  - Activity
- Validated models of data
  - Descriptors
  - Statistical/machine learning techniques
- Data inputation and experimental confirmation
- Reliable models to enable decision support (both in research and regulations) = gain

Current Problems and Challenges in QSAR modeling

- (partially) solved problems:
  - Descriptors
  - Modeling techniques (statistical/machine learning)
  - Model validation approaches
  - Virtual screening and experimental validation of predictions (low bar)
- Challenges
  - Chemical and biological data curation and correction
  - New workflows integrating QSAR with available experimental data
  - Utility of QSAR models outside of research labs (especially, regulatory acceptance)
  - New application areas (e.g., materials informatics; nanotoxicology; integration with PK/PD modeling)

QSAR and Chemical Toxicity Testing in 21 Century

MML UNC.EDU

THE NATIONAL

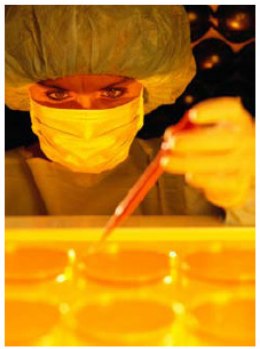
July 2007

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 Academ

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.

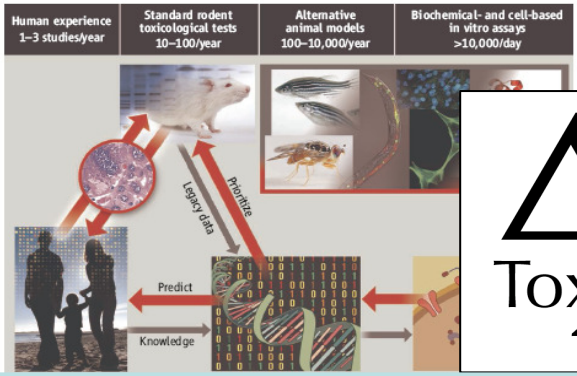
Francis S. Collins,[1*†] George M. Gray,[2*] John R. Bucher[3*]

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 efficiency in design and costs (*1*–*5*). In response, the NRC Committee on Toxicity Testing and Assessment of Environmental Agents produced two reports that reviewed current toxicity testing, identified key issues, and developed a vision and implementation strategy to create a major shift in the assessment of chemical hazard and risk (*6*, *7*). Although the NRC reports have laid out a solid theoretical rationale, comprehensive and rigorously gathered data (and comparisons with historical animal data) will determine whether the hypothesized improvements will be realized in practice. For this purpose, NTP, EPA, and the National Institutes of Health Chemical Genomics Center (NCGC) (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,



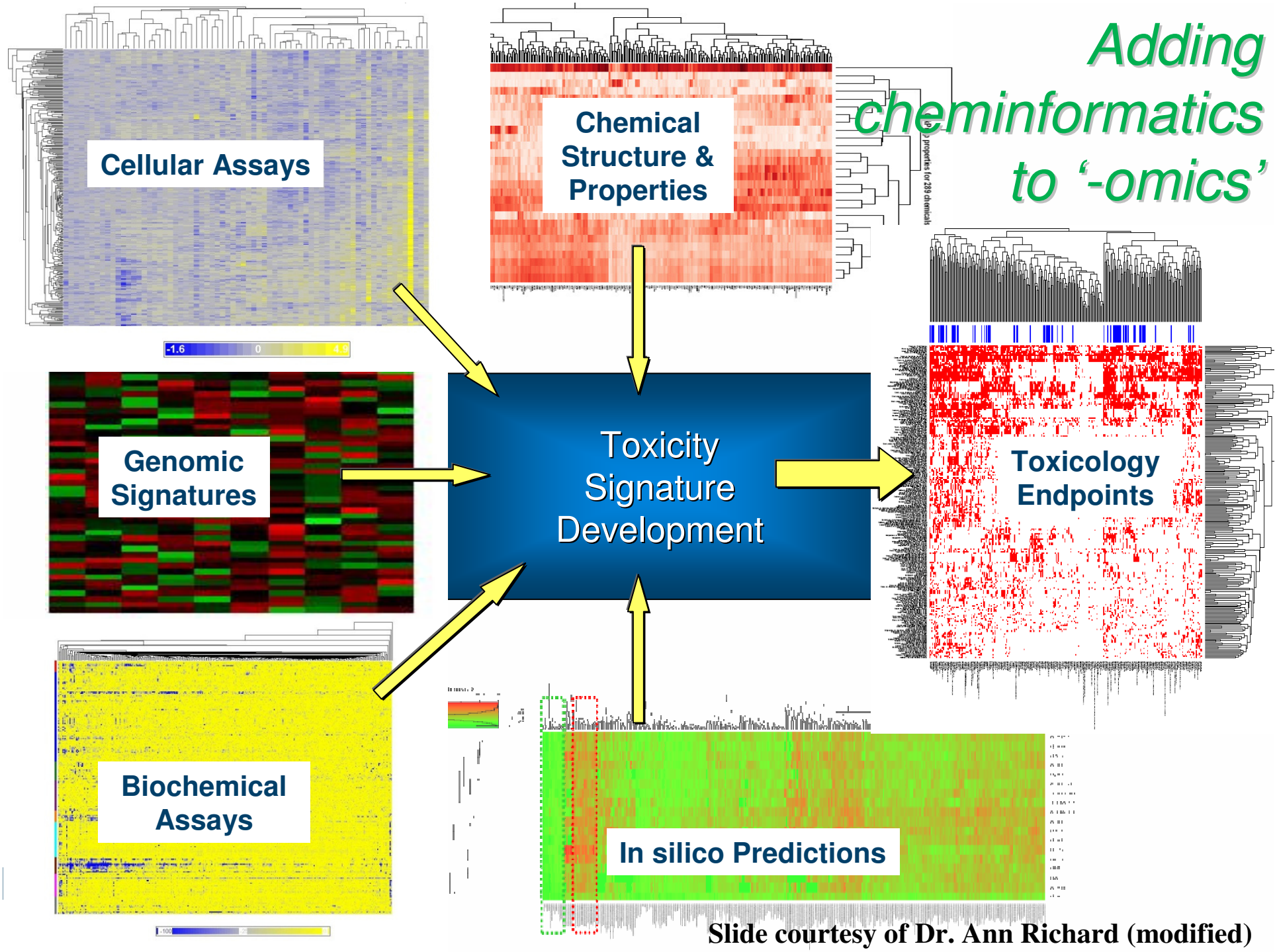
[1]Director, National Human Genome Research Institute (NHGRI), National Institutes of Health, Bethesda, MD 20892; [2]Assistant Administrator for the Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC 20460; [3]Associate Director, National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709. *The views expressed in this Policy Forum are those of the authors and do not necessarily reflect the views and policies of their respective agencies.

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

Transforming toxicology. The studies we propose will test whether high-throughput and computational toxicology approaches can yield data predictive of results from animal toxicity studies, will allow prioritization of chemicals for further testing, and can assist in prediction of risk to humans.

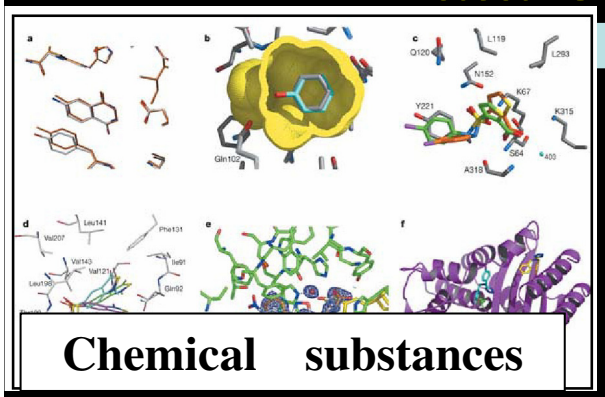
Tox21

EPAs Contribution: The ToxCast Research Program

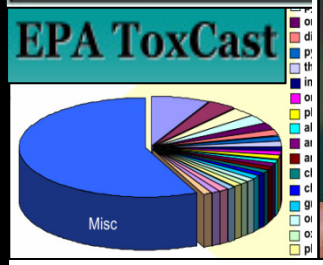

Adding cheminformatics to '-omics'
Cellular Assays
Chemical Structure & Properties
Genomic Signatures
Toxicity Signature Development
Toxicology Endpoints
Biochemical Assays
In silico Predictions
Slide courtesy of Dr. Ann Richard (modified)

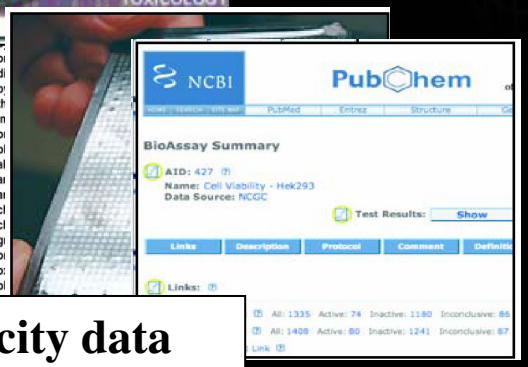
Data Continuum/Modeling workflow

MML UNC.EDU



Chemical substances

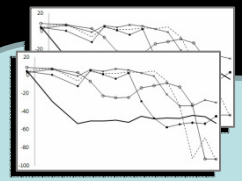




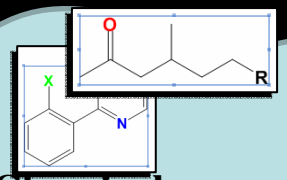
Public toxicity data

Structural data

Data collection and curation



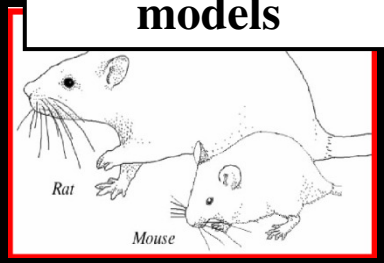
Biological descriptors



Chemical descriptors

Combi-QSAR modeling

In vivo toxicity models



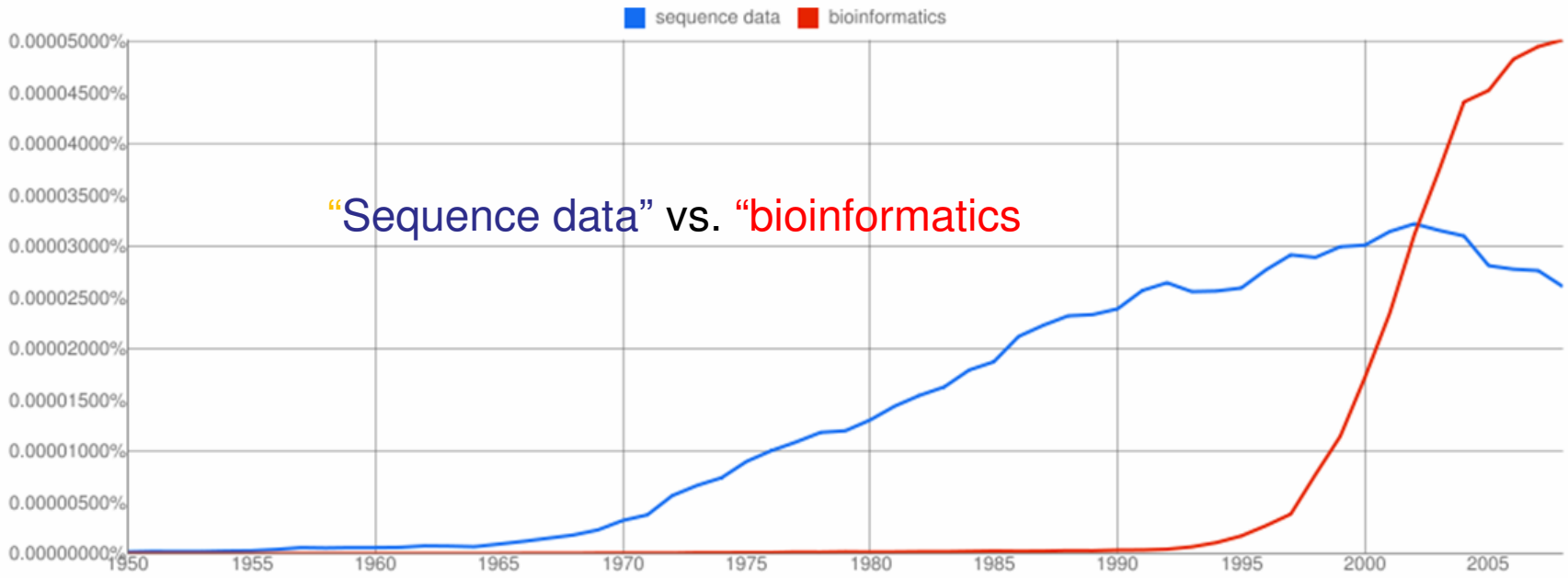
ChemBench portal

The importance of DATA to enable any informatics-dependent discipline: *bioinformatics example*

MML UNC.EDU

Google labs Books Ngram Viewer

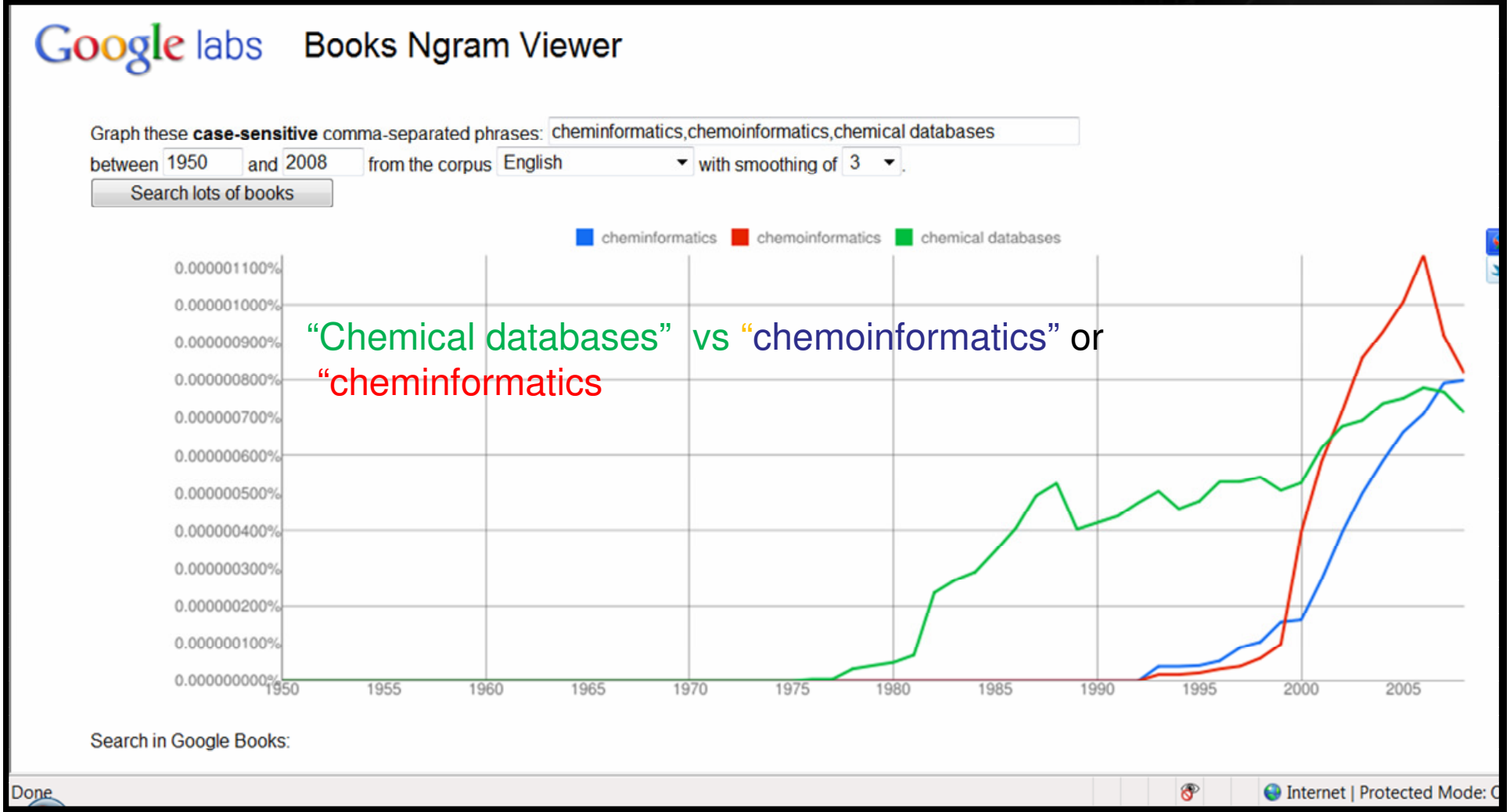
Graph these **case-sensitive** comma-separated phrases: sequence data,bioinformatics
between 1950 and 2008 from the corpus English with smoothing of 3.
Search lots of books



Search in Google Books:

The importance of DATA to enable any informatics-dependent discipline: *cheminformatics example*

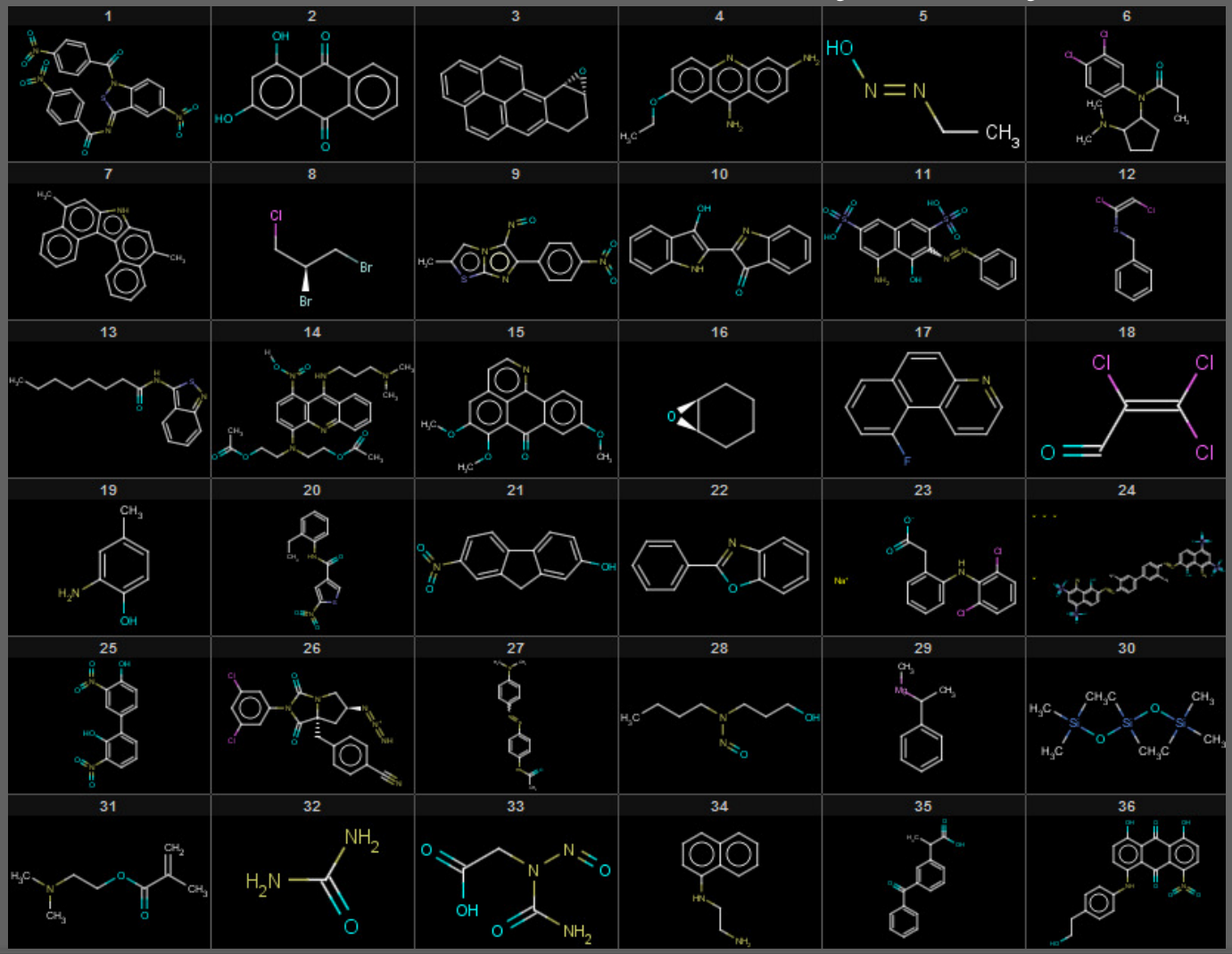
MML UNC.EDU



Data dependency and data quality are critical issues in QSAR modeling

- Cheminformaticians are at the mercy of data providers. Prediction performance of (Q)SAR models could depend strongly on the quality of input data (both structures and activities).
- Both chemical and biological data in a dataset may be inaccurate and in need of thorough curation
- The number of published QSAR models that were poor or not too successful due to data quality issue is unknown but possibly large
- Often considered trivial, the basic steps to curate a dataset of compounds are not so obvious especially for beginners.

242 chemical records + 1 binary activity

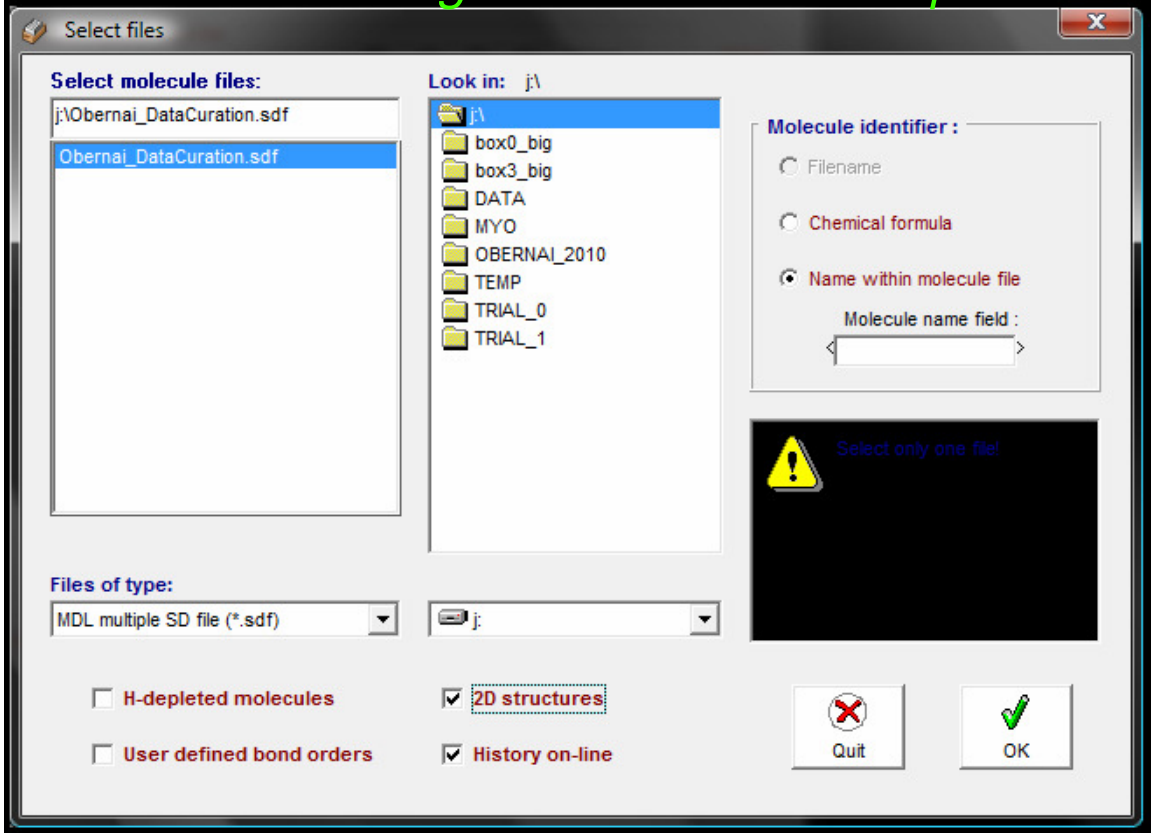


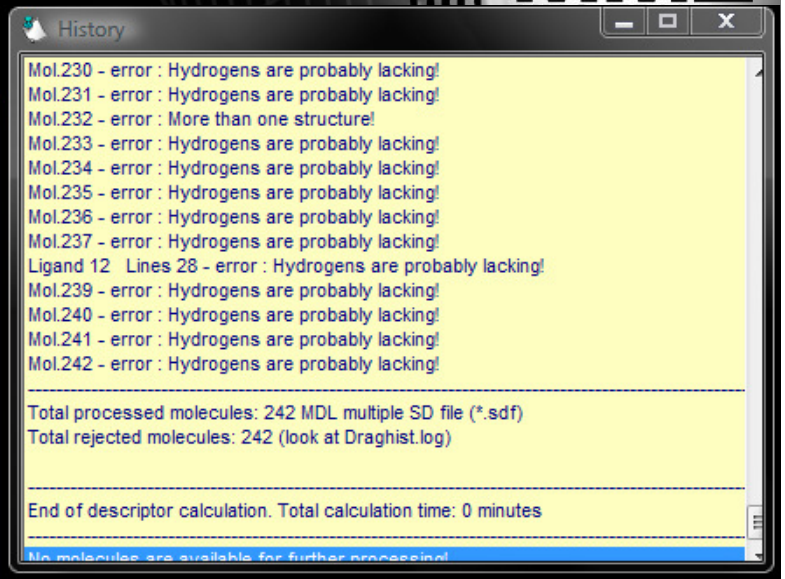
MML
UNC.EDU

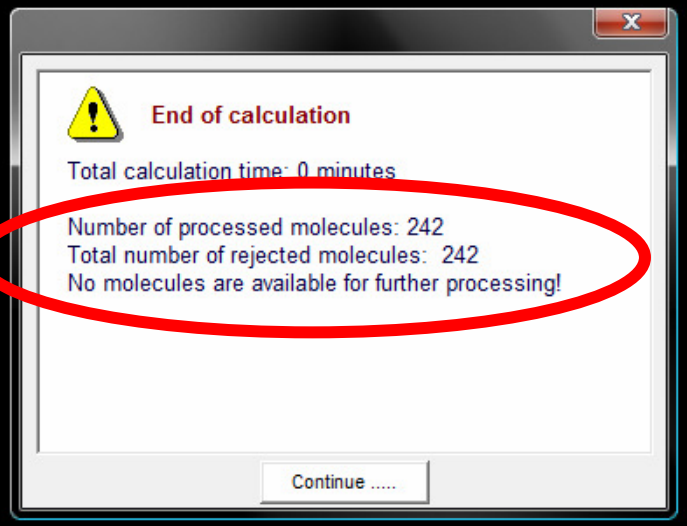
Looks clean …

Looks clean … but …

Calculation of Dragon molecular descriptors





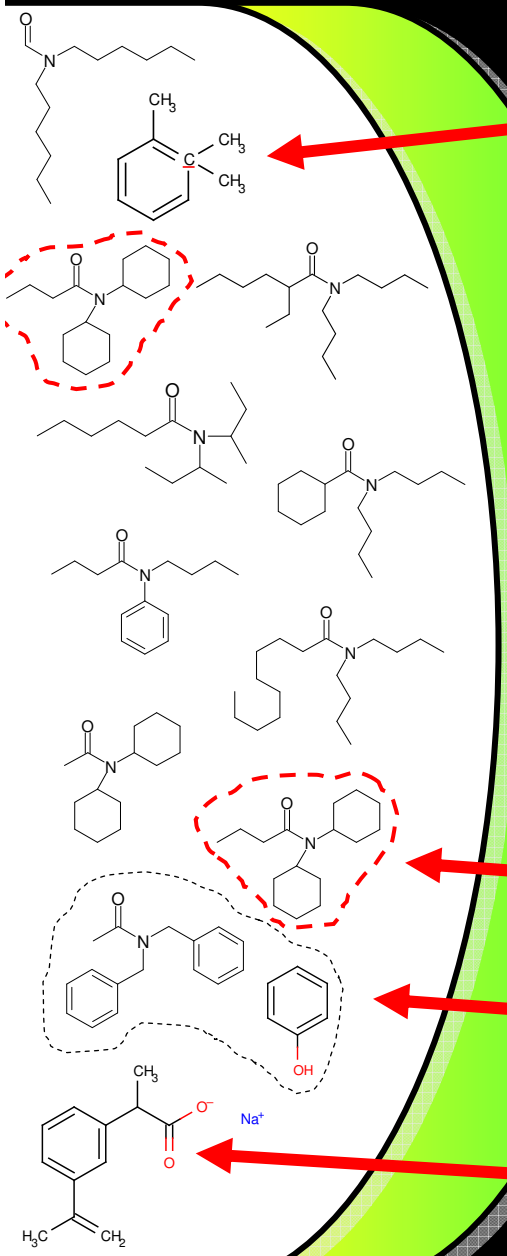


All compounds are in fact incorrect

(presence of inorganics, salts, organometallics, duplicates; certain hydrogens are lacking; wrong standardization; etc.)

http://chembench.mml.unc.edu

QSAR modeling with non-curated datasets



DESCRIPTORS

Presence of **ERRONEOUS AND/OR WRONG STRUCTURES**

Presence of **MISPRINTS AND WRONG NAMES**

ERRORS in the calculation of DESCRIPTORS

QSAR MODELS ???

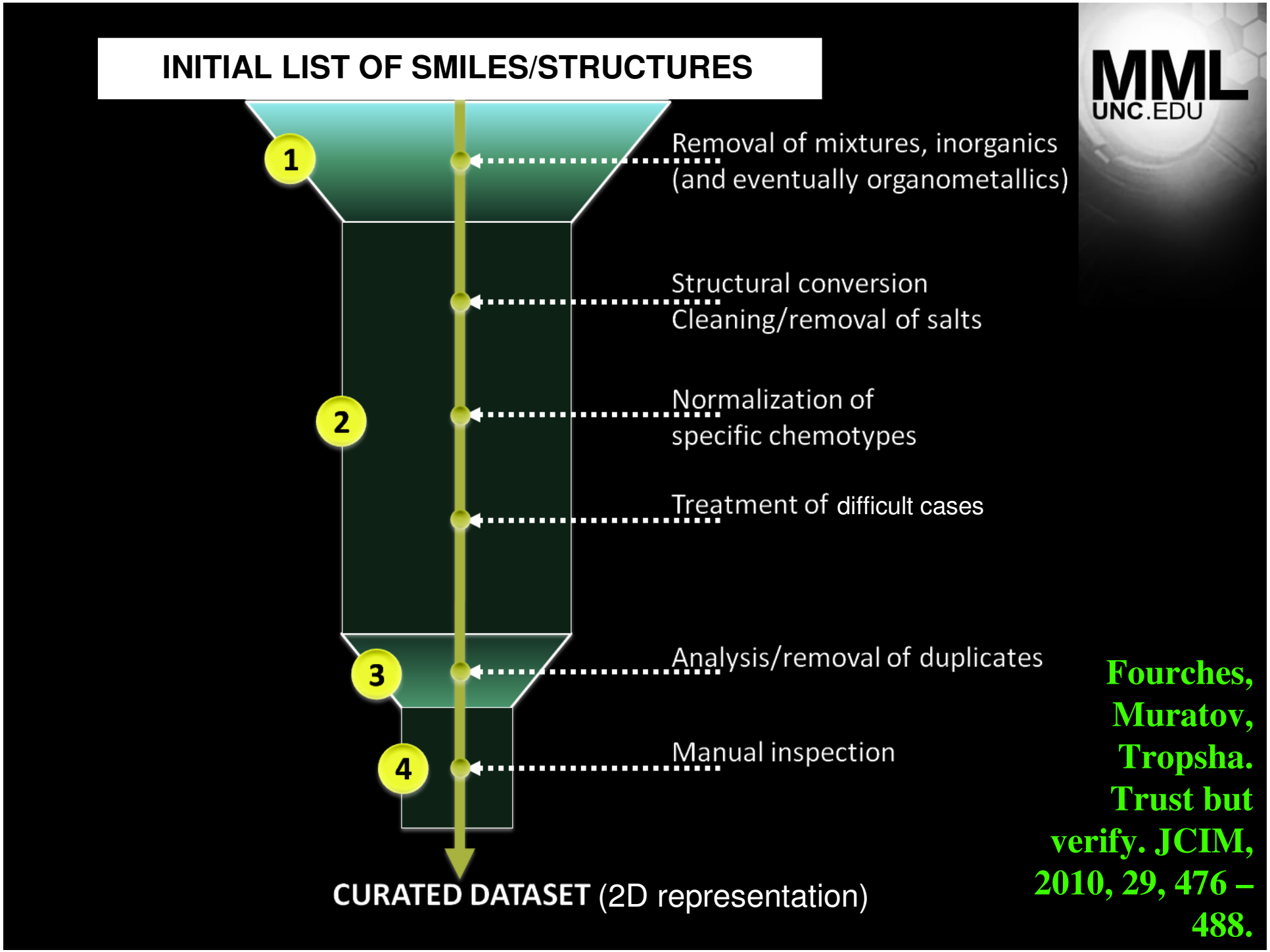
Presence of **DUPLICATES**

Presence of **MIXTURES**

Presence of **SALTS**

Etc.

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


INITIAL LIST OF SMILES/STRUCTURES
MML
UNC.EDU
1
Removal of mixtures, inorganics
(and eventually organometallics)
Structural conversion
Cleaning/removal of salts
2
Normalization of
specific chemotypes
Treatment of difficult cases
3
Analysis/removal of duplicates
4
Manual inspection
CURATED DATASET (2D representation)
Fourches, Muratov, Tropsha. Trust but verify. JCIM, 2010, 29, 476 – 488.

Procedures	Software	Availability
Inorganics Removal	ChemAxon/JChem	Free for Academia
	OpenEye/Filter	Free for Academia
Structure Normalization (fragment removal, structural curation, salt neutralization)	ChemAxon/Standardizer	Free for Academia
	OpenBabel	Free
	Molecular Networks/ CHECK,TAUTOMER	Commercial
Duplicate Removal	ISIDA/Duplicates	Free for Academia
	HiT QSAR	Free for Academia
	CCG/MOE	Commercial
SDF management/ viewer File format converter	ISIDA/EdiSDF	Free
	Hyleos/ChemFileBrowser	Free
	OpenBabel	Free
	ChemAxon/MarwinView	Free for Academia
	CambridgeSoft/ChemOffice	Commercial
	Schrodinger/Canvas	Commercial
	ACD/ChemFolder	Commercial
	Symyx Cheminformatics	Commercial
	CCG/MOE	Commercial
	Accelrys/Accord	Commercial
	Tripos/Benchware Pantheon	Commercial

MML UNC.EDU

Summary of major procedures and corresponding relevant software for every step of the data curation process

The effect of curation on dataset size.

Dataset	Number of compounds	
	Original set	Curated set
Liver toxicants (DILI)	1061	951 (90%)
Nitroaromatics (rats)	28	28 (100%)
Nitroaromatics (*T. pyriformis*)	95	95 (100%)
ToxRefDB	320	292 (91%)
Ames mutagenicity	7090	6542 (92%)
Bioavailability (UCSD)	805	734 (91%)

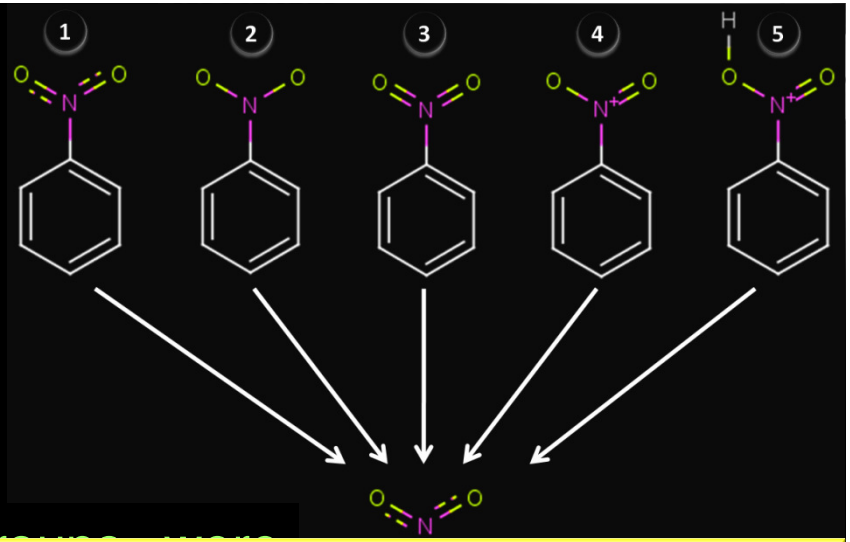
Statistical parameters of QSAR models before and after curation.

ID	Name	R^2	Q^2	R^2_{EF}	S_{ws}	S_{cv}	S_{EF}	R^2_{EVS}	$R^2_{EVS(NM)}$
1	Rat	0.96	0.84-0.93	0.89-0.92	0.11-0.13	0.16-0.24	0.20-0.26	–	–
2	$Rat_{(NM)}$	0.91-0.97	0.89-0.95	0.45-0.88	0.10-0.18	0.14-0.28	0.28-0.58	–	–
3	TP	0.83	–	0.76	0.33	–	0.38	0.54	-0.58
4	$TP_{(NM)}$	0.85	–	0.54	0.31	–	0.54	0.49	0.44
5	Biowisdom non-curated	No modeling was possible							
6	Biowisdom	Modeling Set 5-fold external CV Accuracy = 62-68% External sets Accuracy = 56-73%							
7*	[59] Ames non-curated	$Sensitivity_{RF}$=83%; $Sensitivity_{SVM}$=87%; $Specificity_{RF}$=$Specificity_{SVM}$=75% AUC_{GP}=88%; AUC_{SVM}=89%; AUC_{RF}=83%							
8*	[60] Ames curated	$Sensitivity_{RF}$=$Sensitivity_{SVM}$=79%; $Specificity_{RF}$=$Specificity_{SVM}$=81% AUC_{GP}=86%; AUC_{SVM}=84%; AUC_{RF}=83%							

Where:
TP – *Tetrahymena pyriformis* dataset, $_{(NM)}$ – modeling with various representations of nitro groups
R^2 - determination coefficient, Q^2 - cross validation determination coefficient
R^2_{EF}- determination coefficient for external folds extracted from the modeling set
S_{ws} - standard error of a prediction for work set
S_{cv} - standard error of prediction for work set in cross validation terms
S_{ts} - standard error of a prediction for external folds extracted from the modeling set
A - number of PLS latent variables, D - number of descriptors, M - number of molecules in the work set
R^2_{EVS} - determination coefficient for external validation set
$R^2_{EVS(NM)}$ - determination coefficient for external validation set with shuffled nitro groups
AUC – Area Under Curve statistical parameter
RF – Random Forest, SVM – Supporting Vector Machine, GP – Gaussian Processes, * Prediction performances are reported for external validation set.

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 were

Data curation improves the true accuracy (up or down!) of QSAR models

significantly modified the overall prediction accuracy obtained by models trained with standardized (R^2_{ext}~0.5) vs. non-standardized (R^2_{ext}<0) compounds.

Results show that even small differences in structure representation can lead to significant errors, and even robust and inherently predictive models can fail on non-curated external validation sets.

Artemenko, Muratov et al. J. SAR QSAR 2011, Accepted.

Case Study I: A Two-step Hierarchical QSAR Modeling Workflow for Predicting *in vivo* Chemical Toxicity*

*Zhu, Richard, Rusyn, Wright, et al, EHP, 2009, 117(8):1257-64

Focusing on a small subset of ToxCast™ data: Chronic Mouse Toxicity

- Continuity (overlaps with previous ToxRefDB data)
- Manageable (has only 7 *in-vivo* assays)
- 3 assays with the highest fraction of actives chosen for initial studies

  CHR_Mouse_LiverProliferativeLesions (87 actives)

  CHR_Mouse_LiverTumors (68 actives)

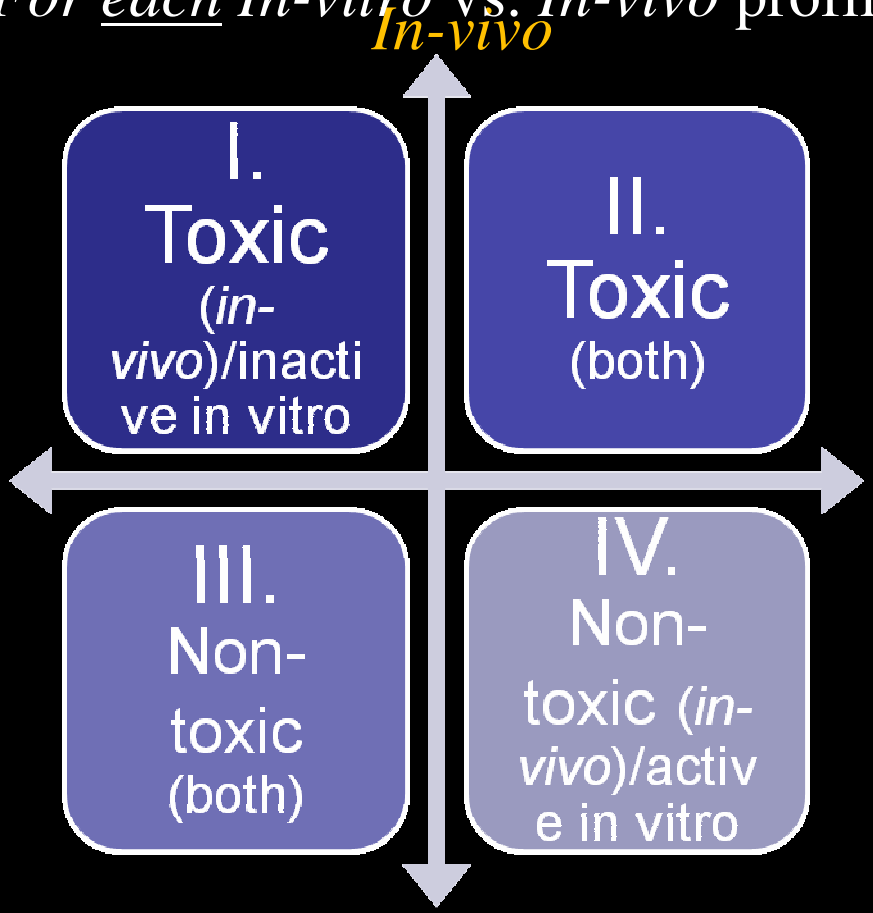
  CHR_Mouse_Tumorigen (88 actives)

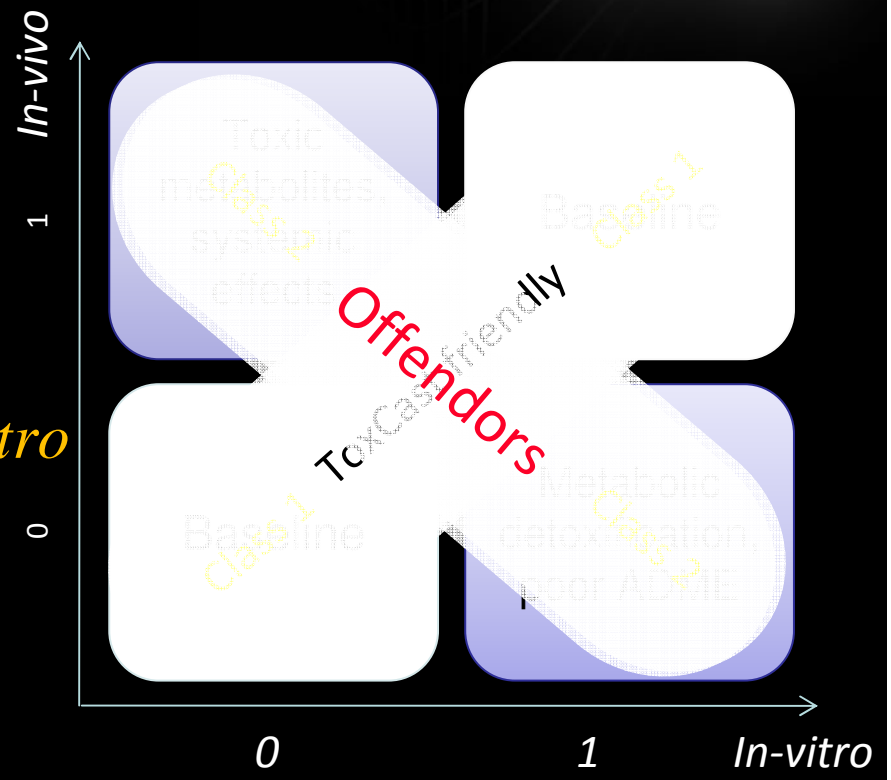
- 1 composite endpoint:

  CHR_Mouse_Liver_tox (110 actives)

Data partitioning based on *in vitro-in vivo* correlations as part of the QSAR Modeling workflow

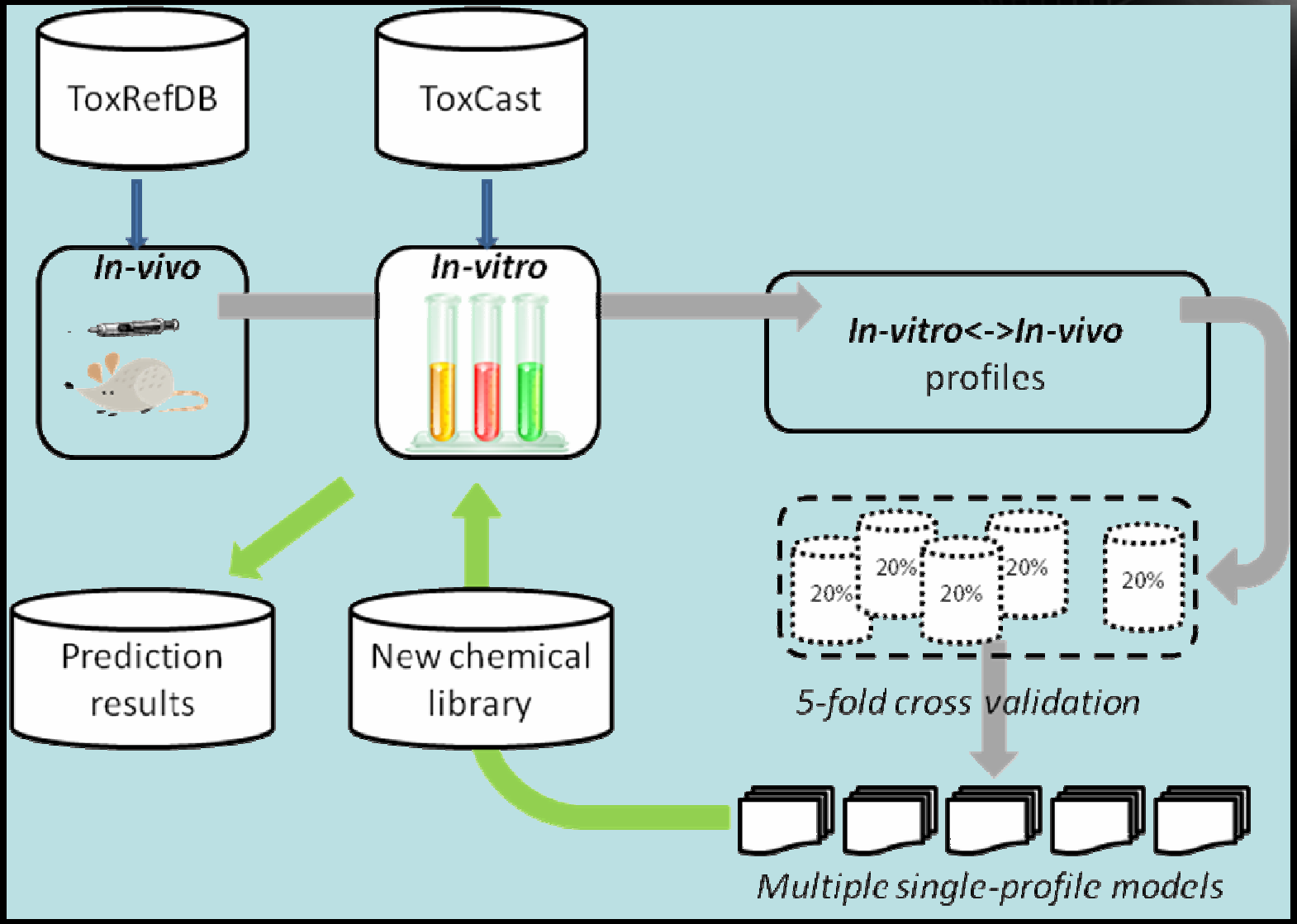
For <u>each</u> In-vitro vs. *In-vivo* profile (3 x 353 = 1059 combinations):



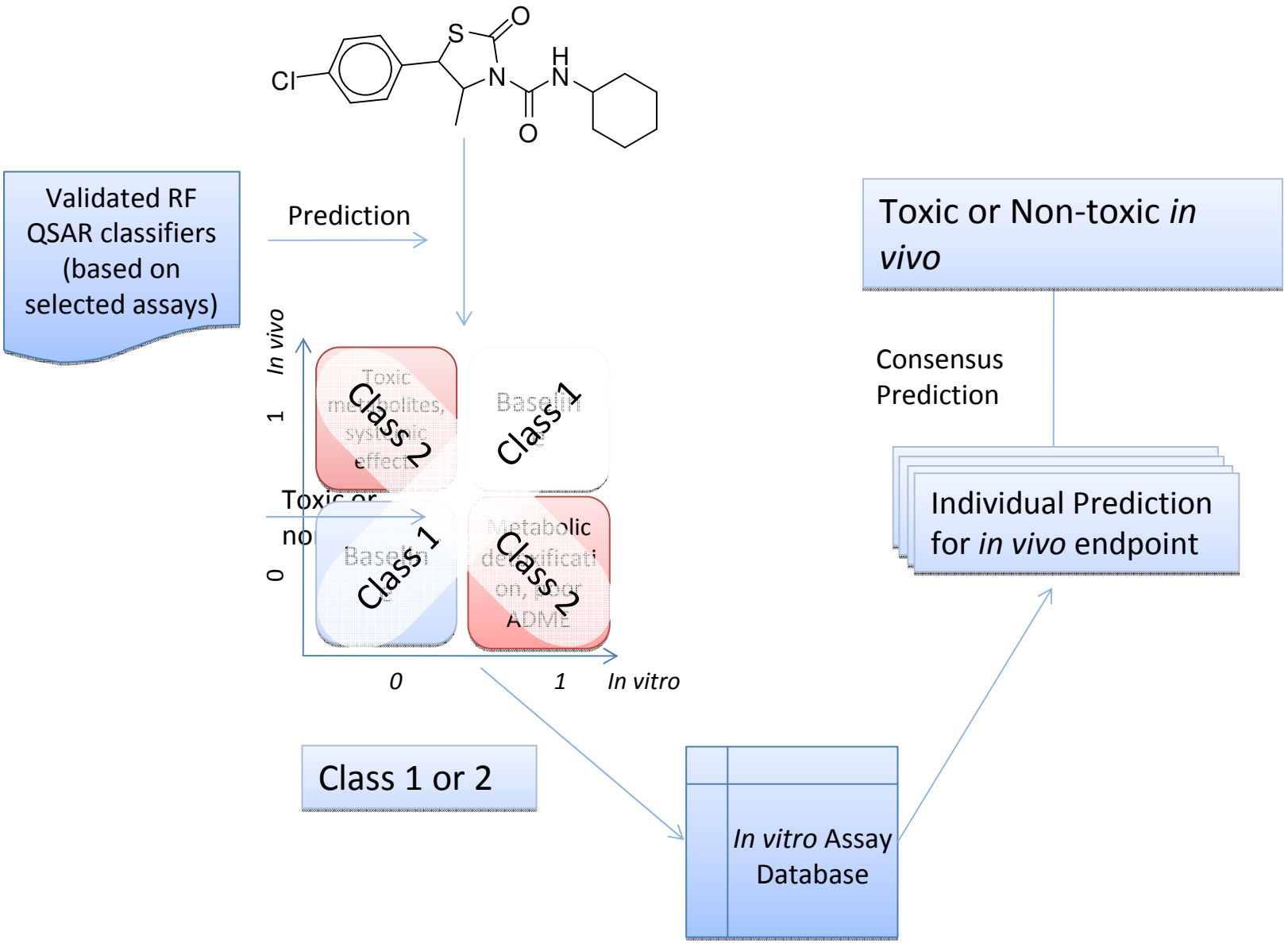


- Binary classification QSAR for “baseline” (II & III) vs. off-line (I & IV) using chemical descriptors only

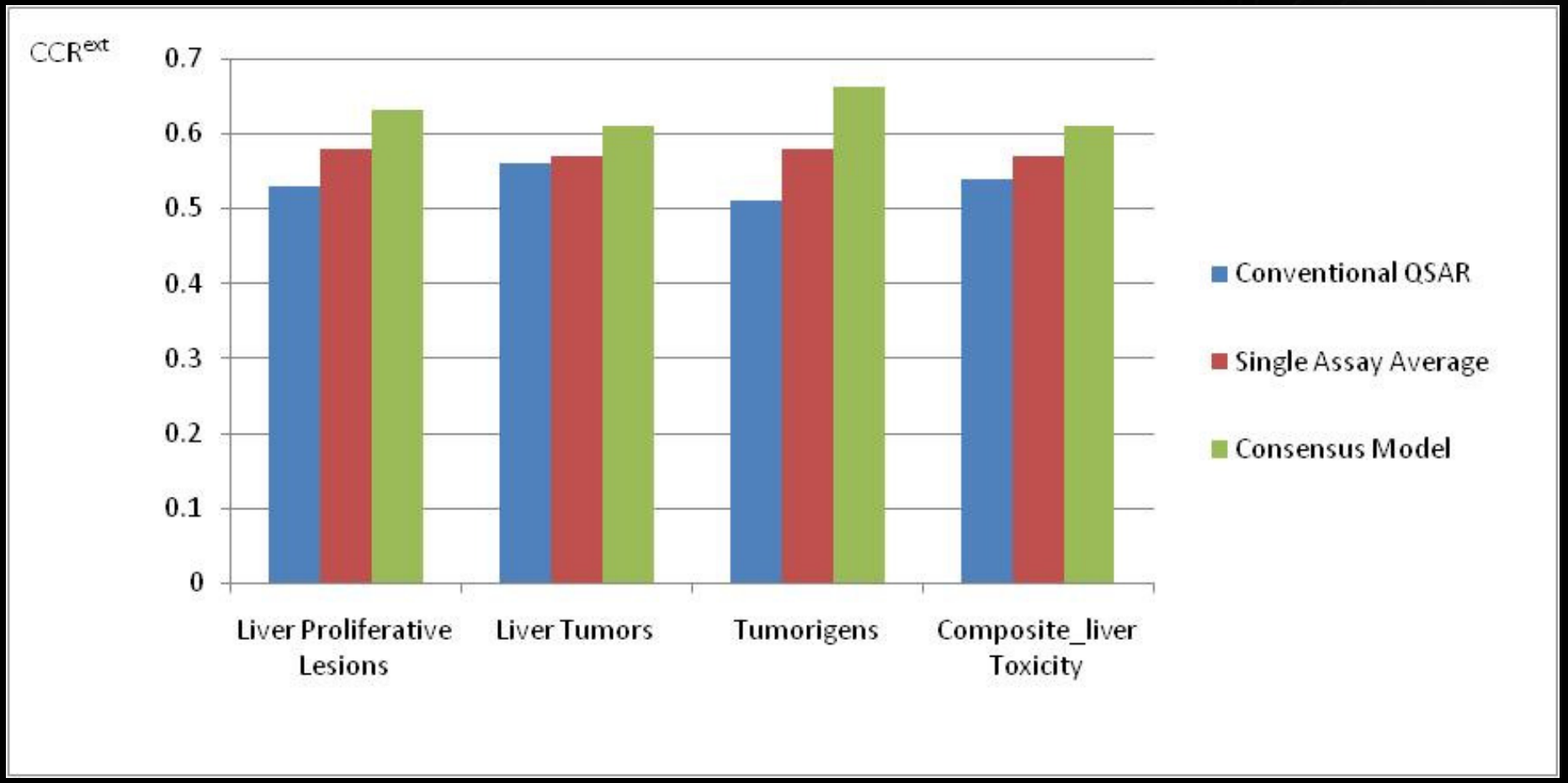
Modeling Workflow for ~1400 *in vitro – in vivo* Series Using Various QSAR Approaches and Dragon descriptors



External prediction workflow



Comparison of External Prediction Accuracy of conventional vs. hierarchical QSAR Models

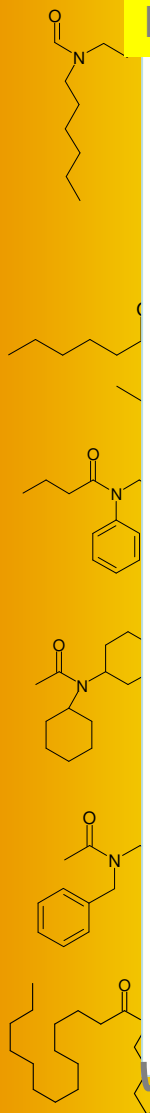


Case Study 2. Hybrid chemical-biological (short-term assays) descriptors*

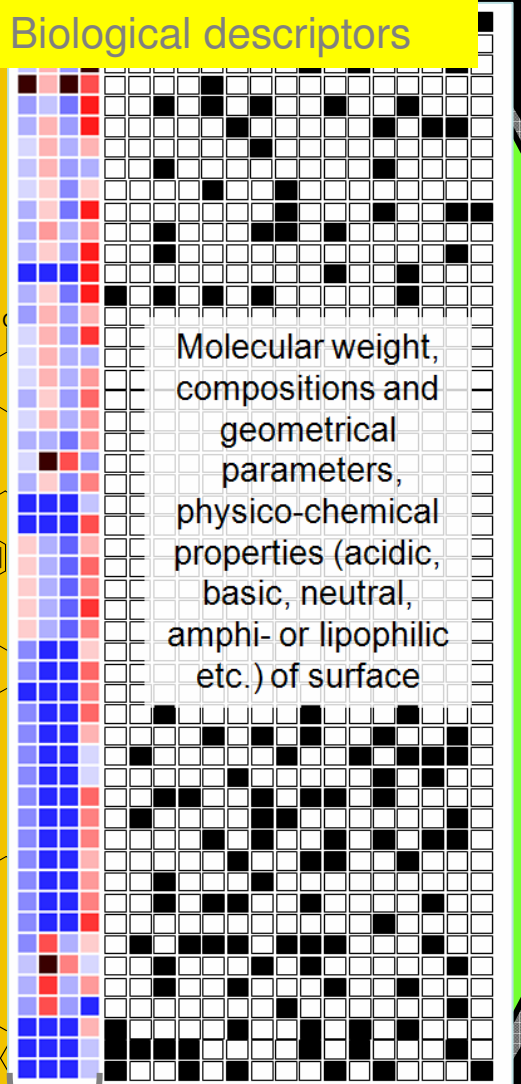
Zhu et al, EHP, 2008, (116): 506-513
Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. EHP, 2011, 119(3):364-70.

QFAR modeling using hybrid chemical-biological descriptors

COMPOUNDS



Combined chemical/
Biological descriptors



In vitro assays

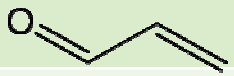
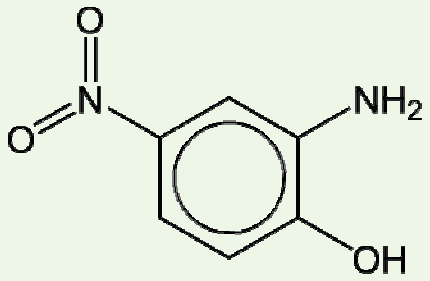
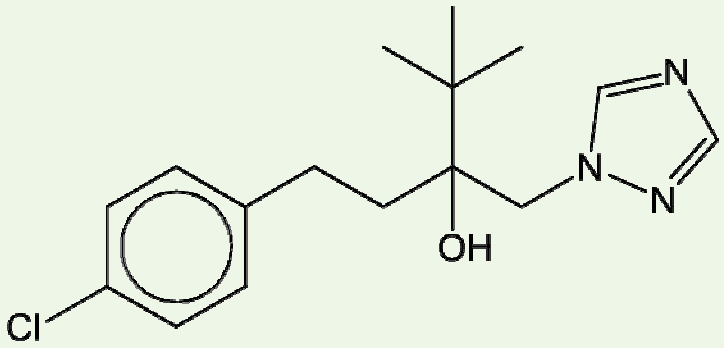
Quantitative
Structure
Property
Relationships

0.613
0.380
-0.222
0.708
1.146
0.491
0.301
0.141
0.956
0.256
0.799
1.195
1.005

PROPERTY

QSAR TABLE – CHEMICAL DESCRIPTORS

Descriptor #: 1 2 … 201

ID	Name	Structure	MW	arom. atoms	…	-C- sp2
1	Acrolein		56.0	0	…	1
2	2-Amino-4-nitrophenol		154.0	6	…	6
…	…	…	…	…	…	…
369	Tebuco-nazole		307.1	11	…	5

IN VITRO DOSE-RESPONSE DATA IMPROVE THE PREDICTIVE POWER OF QSAR MODELS

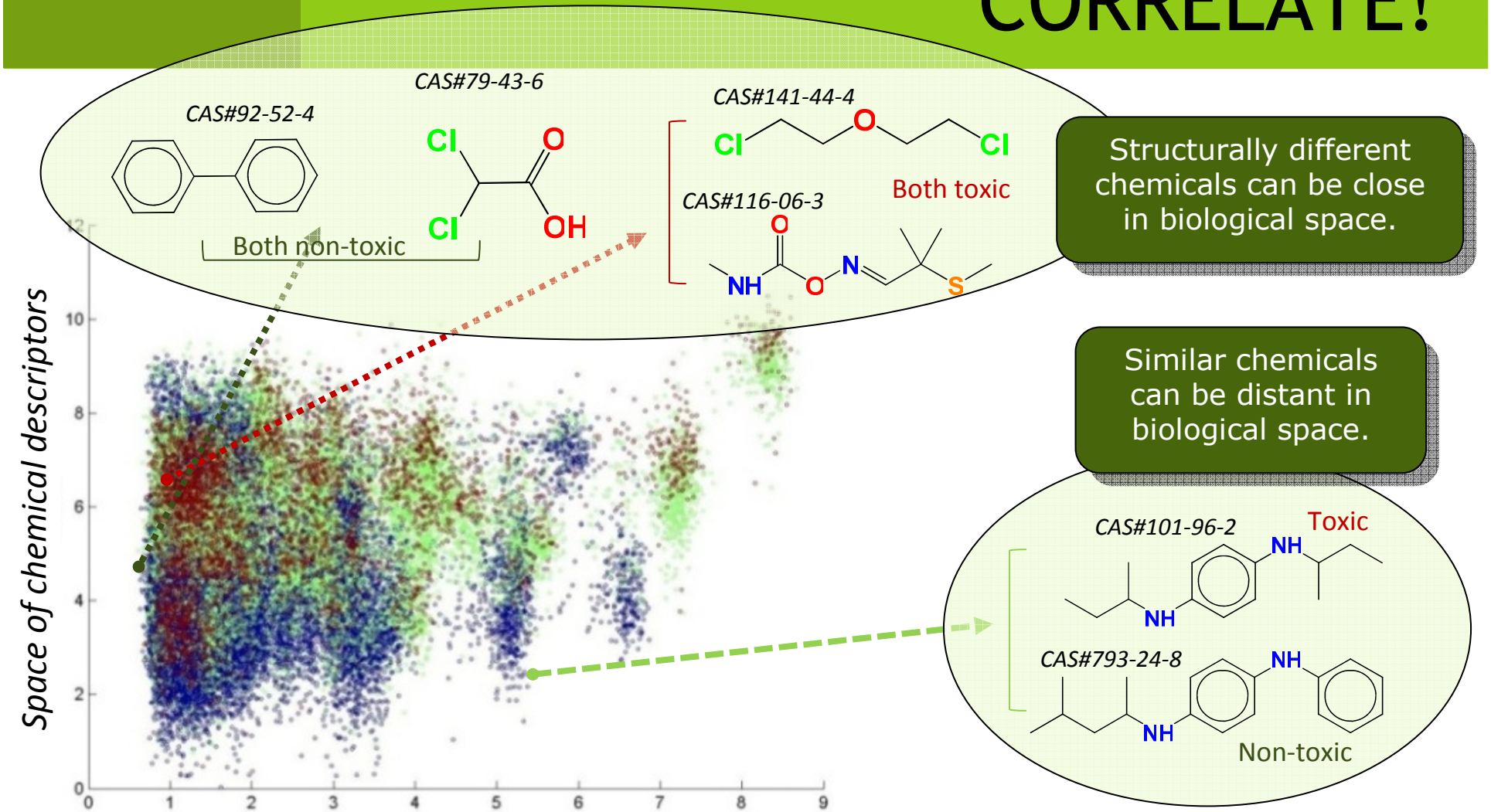
Case Study: prediction of *in vivo* toxicity (rat LD_{50})

- 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

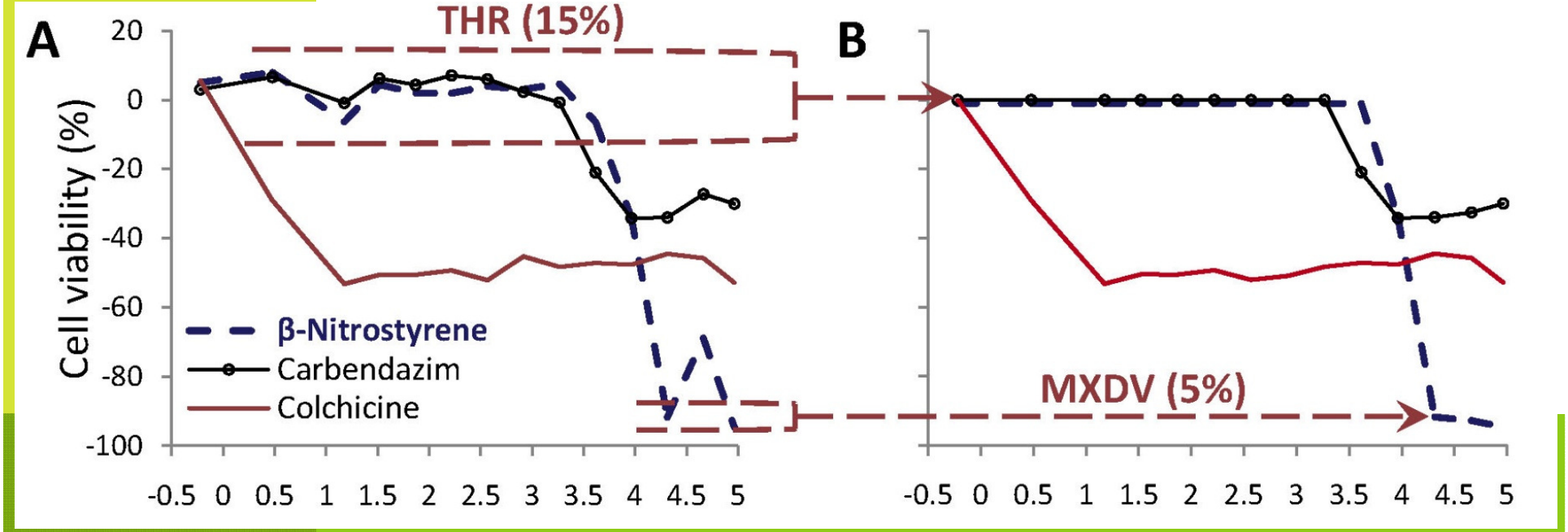
CHEMICAL AND BIOLOGICAL SIMILARITIES DO NOT CORRELATE!



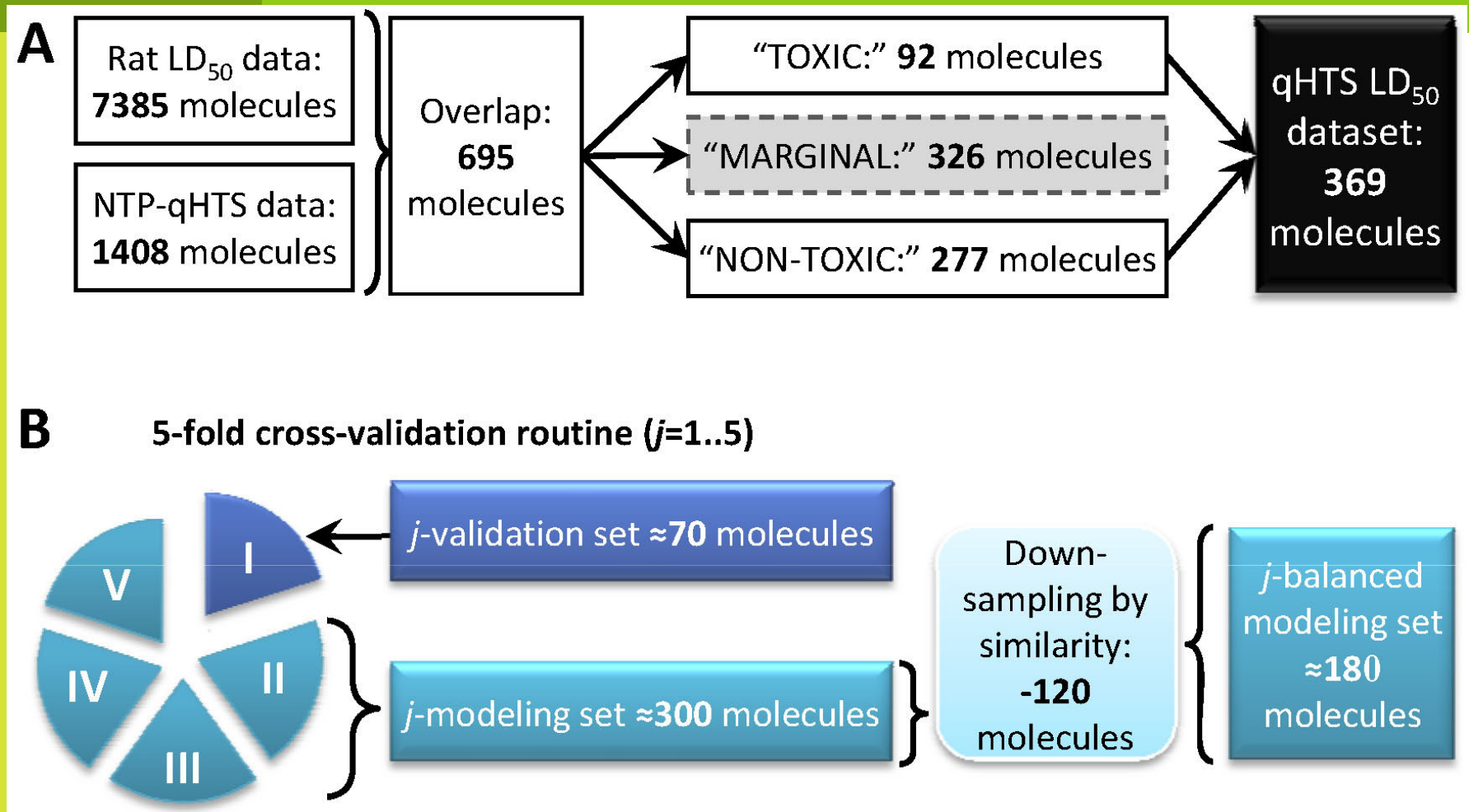
Pairwise distances of modeling compounds in space of chemical (*y*-axis) and biological (*x*-axis) descriptors:

● Between "toxic" compounds; ● Between "non-toxic" compounds; ● Between "toxic" and "non-toxic" compounds

EXAMPLES OF QHTS CONCENTRATION-RESPONSE CURVES AND THEIR NOISE-FILTERING TRANSFORMATIONS



MODELING WORKFLOW

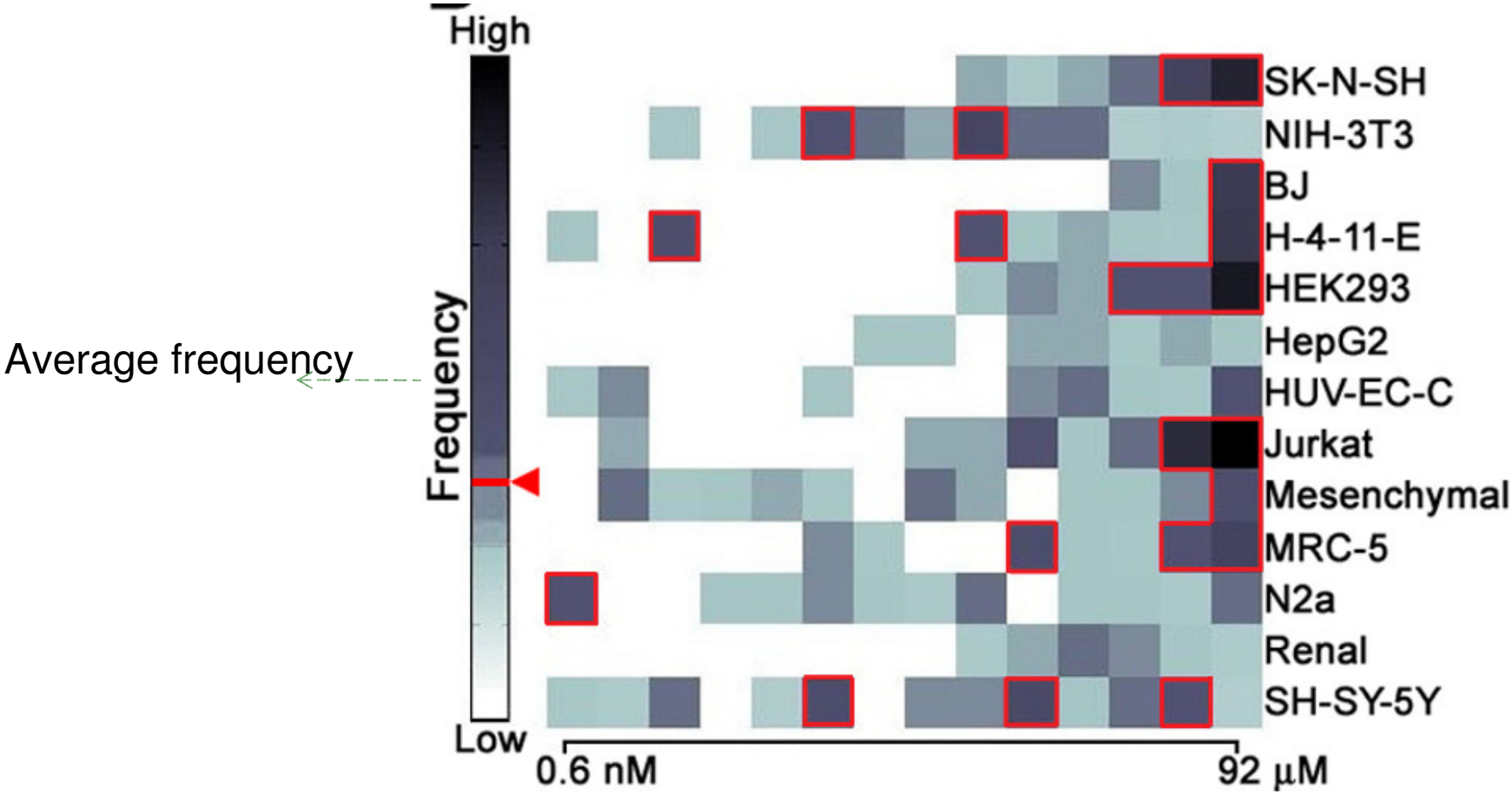


NOISE-REDUCTION APPLIED TO CONCENTRATION-RESPONSE DATA IMPROVES PREDICTION ACCURACY OF HYBRID MODELS.

	%	Chemical descriptors only	Hybrid descriptors (Original, binary)	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**

BIOLOGICAL DESCRIPTOR ANALYSIS

Relative contribution of qHTS dose-response descriptors to QSAR models varies between cell-lines



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).

Case Study 3: QSAR + Toxicogenomics*

Rationale

- Hybrid models of chemical and biological descriptors (in vitro assays, dose response curves) improved prediction of toxicity (Zhu 2008, Sedykh 2010)
- Toxicogenomics has proven predictive and interpretative value

Hypothesis

- Adding toxicogenomics data as biological descriptors to hybrid QSAR models may improve predictivity.

About the data set

- The Toxicogenomics Project (TGP), Japan collected microarray profiles of 150 common drugs covering various hepatotoxic modes of action http://toxico.nibio.go.jp/

*Low et al. Predicting Drug-induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem. Res. Toxicol.*, **2011**, 24(8):1251-62.

Data available from TGP

http://toxico.nibio.go.jp/

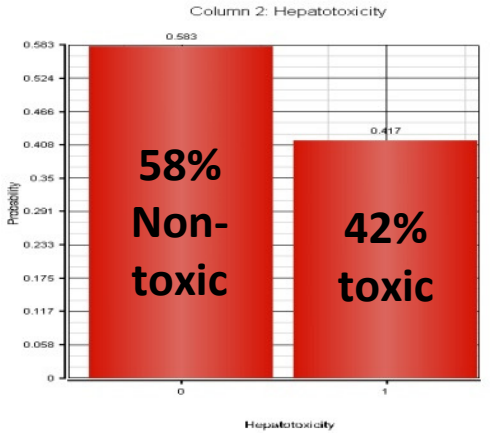
Table 2. The standard study protocol in TGP

In vivo	
Animal	Sprague–Dawley rat (6 wk old, $N = 5$ for each group)
Vehicle	0.5% Methylcellulose or corn oil (oral dose) Saline or 5% glucose solution (intravenous dose)
Dose	Low, middle, and high (mainly 1:3:10)
Route	Oral (intravenous in a few cases)
Sacrifice	3, 6, 9, and 24 h after a single administration 24 h after the last dose of repeated administration for 3, 7, 14 and 28 days
Sampling	Liver, kidney, and plasma
Microarray analysis	Affymetrix GeneChip ($N = 3$ for each group)
Items examined	Histopathology: liver and kidney Body weight, organ weight (liver and kidney), food consumption, hematology, and b[illegible]

Source: Uehara 2010

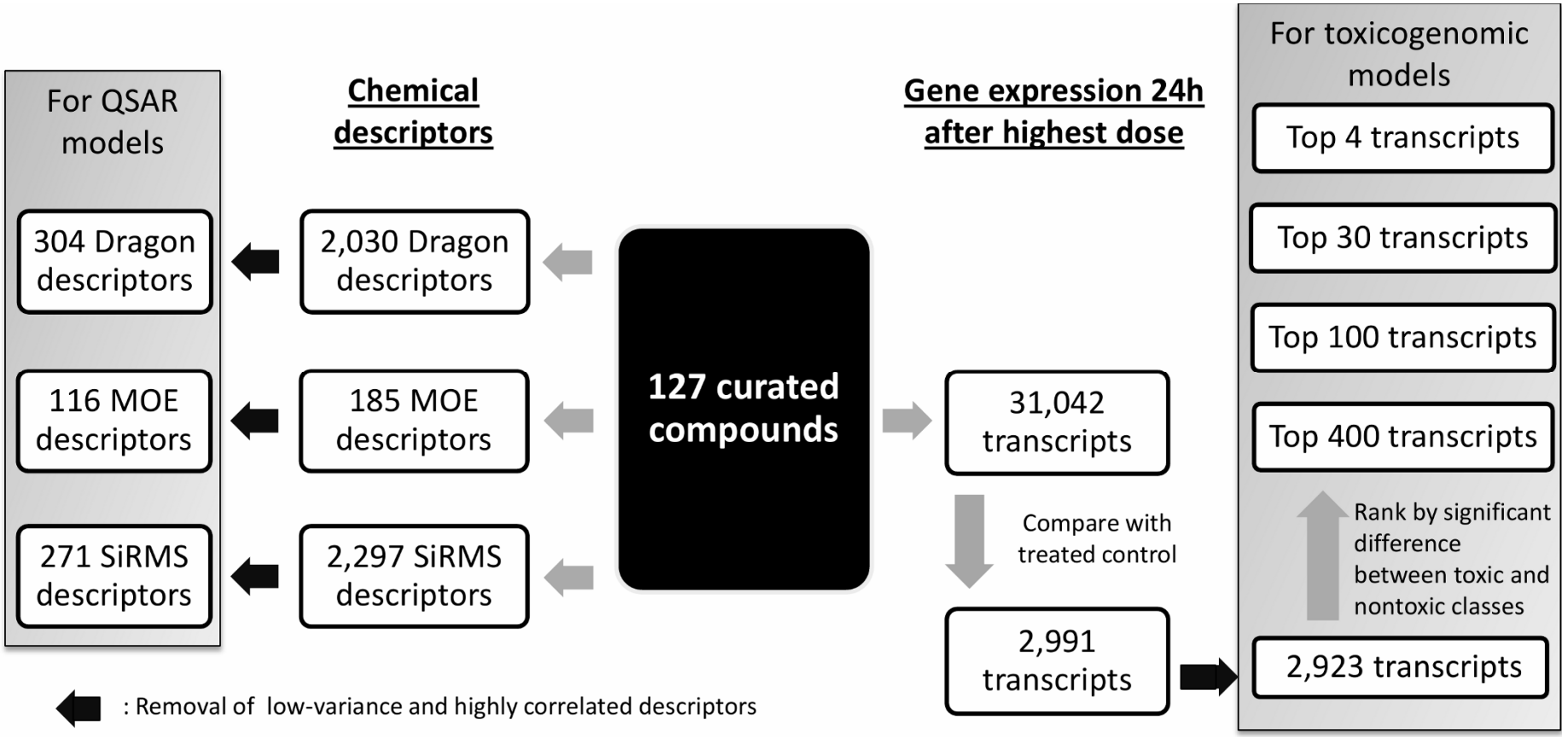
127 drugs

Hepatotoxicity (based on serum chemistry and histopathology)

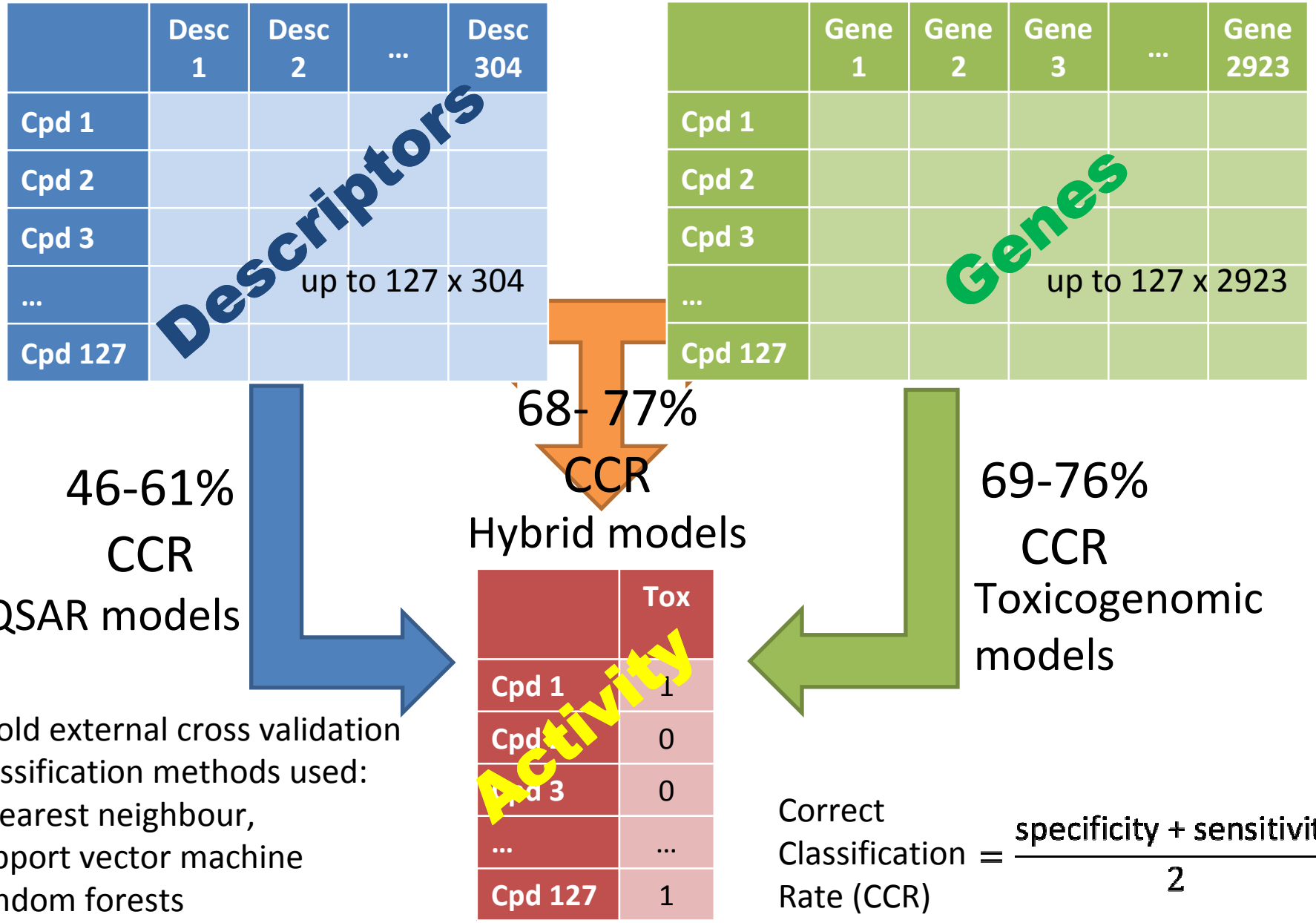


Gene expression data used: taken from rats treated with the highest dose after 24h.

Chemical curation and feature selection

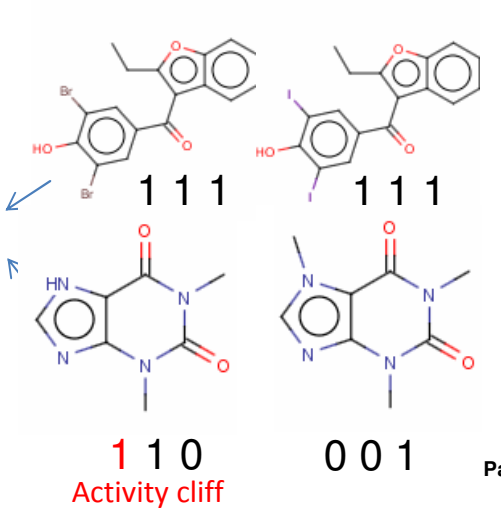


Predicting hepatotoxicity from chemical descriptors and/or gene expression



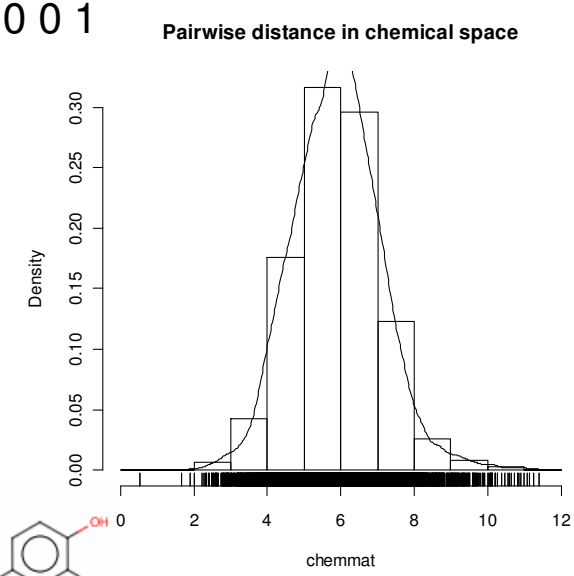
50% of 40 closest pairs in chemical space are activity cliffs

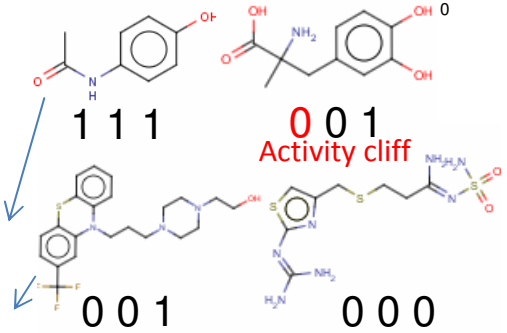
CID	Compounds	Distance	Toxicity
74,118	mefenamic acid, phenylanthranilic acid	0.54	0, 0
40,53	methyltestosterone, ethinylestradiol	1.66	1, 1
63,64	imipramine, amitriptyline	1.91	0, 0
20,110	benzbromarone, benziodarone	2.03	1, 1
83,84	theophylline, caffeine	2.21	1, 0
72,127	erythromycin ethylsuccinate, doxorubicin	2.30	0, 0
5,14	valproic acid, aspirin	2.39	0, 1
24,94	methapyrilene, promethazine	2.48	1, 1
11,25	phenylbutazone, phenytoin	2.50	0, 0
16,115	thioacetamide, ethanol	2.53	1, 0
116,117	phenacetin, bucetin	2.56	1, 1
30,66	gemfibrozil, ibuprofen	2.62	1, 0
52,67	tamoxifen, quinidine	2.63	0, 0
44,105	perhexiline, terbinafine	2.67	0, 1
17,49	carbamazepine, pemoline	2.75	0, 0
19,62	nitrofurantoin, nitrofurazone	2.80	0, 0
10,112	allyl alcohol, bromoethanamine	2.81	1, 0
8,46	rifampicin, tetracycline	2.87	0, 0
85,95	papaverine, colchicine	2.88	0, 1
77,103	chlorpheniramine, clomipramine	2.94	0, 1
99,111	simvastatin, etoposide	3.00	1, 0
38,122	adapin, acetamidofluorene	3.03	0, 1
2,60	isoniazid, iproniazid	3.03	0, 0
39,82	labetalol, enalapril	3.13	0, 0
104,123	trimethadione, nitrosodiethylamine	3.46	1, 1
58,90	tacrine, mexiletine	3.48	0, 1
9,113	naphthyl isothiocyanate, ethionamide	3.53	1, 1
56,78	monocrotaline, nifedipine	3.59	1, 0
69,107	fenofibrate, chlormadinone	3.65	1, 0
1,54	acetaminophen, methyldopa	3.69	1, 0
…	…	…	
36,75	fluphenazine, famotidine	6.56	0,0



Key:

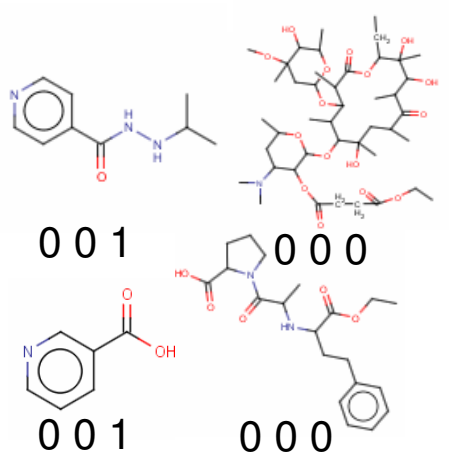
Y_{obs} $Y_{pred,gene}$ $Y_{pred,chem}$





33% of 40 closest pairs in gene expression space are activity cliffs

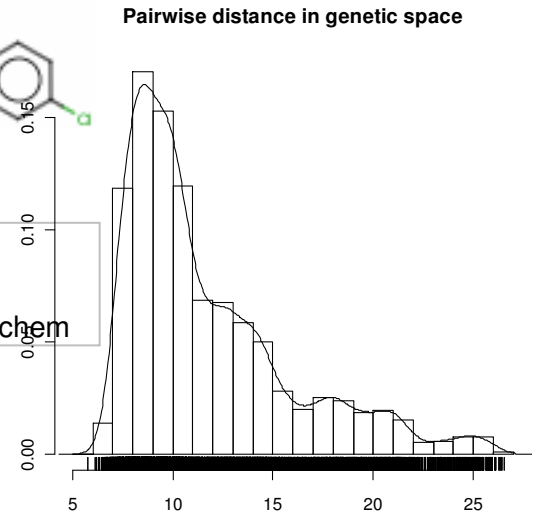
CID	Compounds	Distance	Toxicity
60,72	iproniazid, erythromycin ethylsuccinate	5.75	0, 0
42,45	griseofulvin, ketoconazole	6.18	0, 0
34,39	cimetidine, labetalol	6.30	0, 0
71,82	nicotinic acid, enalapril	6.50	0, 0
70,125	chlorpropamide, gentamicin	6.56	1, 0
76,81	ranitidine, captopril	6.62	0, 0
35,36	haloperidol, fluphenazine	6.64	0, 0
18,115	diclofenac, ethanol	6.67	0, 0
46,48	tetracycline, ciprofloxacin	6.69	0, 0
9,88	naphthyl isothiocyanate, triamterene	6.72	1, 0
25,65	phenytoin, hydroxyzine	6.75	0, 0
51,64	metformin, amitriptyline	6.75	0, 0
7,91	methotrexate, tiopronin	6.77	0, 0
102,107	triazolam, chlormadinone	6.81	0, 0
55,73	methimazole, ethambutol	7.00	0, 1
66,74	ibuprofen, mefenamic acid	7.20	0, 0
97,103	sulpiride, clomipramine	7.34	0, 1
2,37	isoniazid, thioridazine	7.43	0, 0
58,63	tacrine, imipramine	7.46	0, 0
61,105	chloramphenicol, terbinafine	7.58	1, 1
47,98	lomustine, acarbose	7.68	1, 0
27,126	allopurinol, vancomycin	7.73	0, 0
44,68	perhexiline, furosemide	7.85	0, 0
19,22	nitrofurantoin, diazepam	7.85	0, 1
6,110	clofibrate, benziodarone	7.91	0, 1
40,104	methyltestosterone, trimethadione	7.92	1, 1
…	…	…	
3,120	carbon tetrachloride, cyclosporine A	8.56	1,1
…	…	…	
87,113	sulindac, ethionamide	20.49	1,1



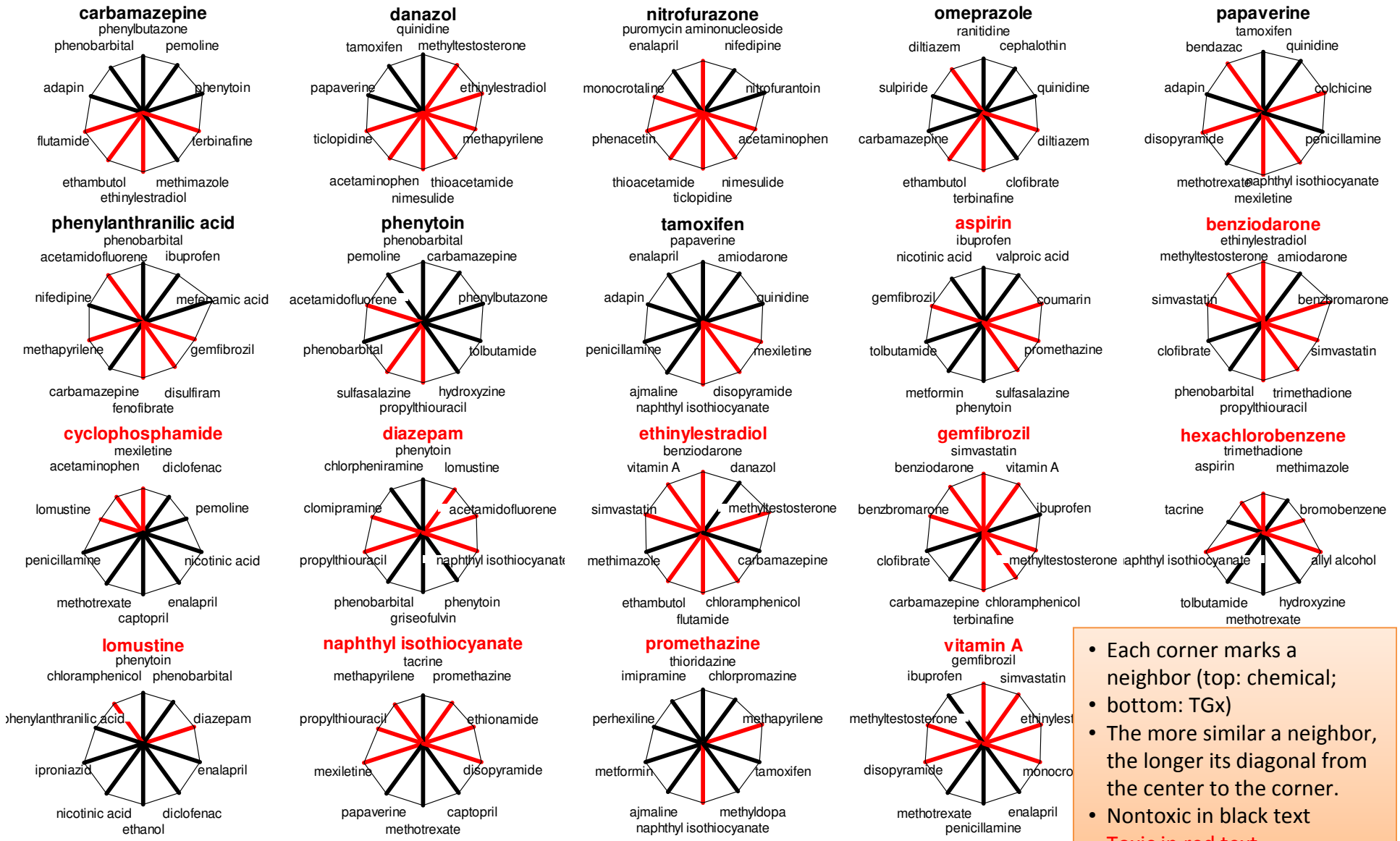
Key:

Y_{obs} $Y_{pred,gene}$ $Y_{pred,chem}$

1 1 1 1 1 0

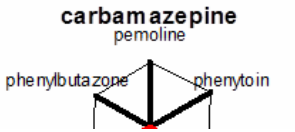


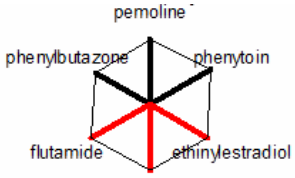
A novel consensus kNN approach: learning from nearest chemical and toxicogenomic neighbors (k=5)



- Each corner marks a neighbor (top: chemical; bottom: TGx)
- The more similar a neighbor, the longer its diagonal from the center to the corner.
- Nontoxic in black text
  Toxic in red text

Consensus vs. Hybrid Descriptors kNN





Neighbors	Space defined by	k	Sens	Spec	CCR
Chemical neighbors only	304 Dragon desc	7	0.56	0.65	0.59
Toxicogenomic neighbors only	85 transcripts	5	0.75	0.79	0.74
Hybrid neighbors	304 Dragon desc 85 transcripts TOGETHER	6	0.69	0.77	0.71
Chemical neighbors & toxicogenomic neighbors	304 Dragon desc, 85 transcripts SEPARATELY	5	0.76	0.80	**0.78**

Results after 5-fold ext CV

5-fold ext CV used to select optimal k

Conclusions

- Methodology:
  - Data curation is critical!
  - consensus (collaborative!) prediction using all acceptable models
  - outcome: decision support tools in selecting future experimental screening sets
- The highest accuracy is achieved by models that employ both chemical and biological descriptors of compounds
  - Integration of cheminformatcs and bioinformatics: predictive model s of selected endpoints using integrated short term biological profiles (biodescriptors) and chemical descriptors for compound subsets
  - New computational approaches (e.g., hybrid and hierarchical QSAR)
  - Interpretation of significant chemical and biological descriptors

Cheminformatics for masses!

MML UNC.EDU

CHEM BENCH

http://chembench.mml.unc.edu

HOME | MY BENCH | DATASET | MODELING | PREDICTION | CECCR BASE

ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

- -

Chembench is a free portal that enables researchers to mine available chemical and biological data. Chembench can help researchers rationally design or select new compounds or compound libraries with significantly enhanced hit rates in screening experiments.

It provides cheminformatics research support to molecular modelers, medicinal chemists and quantitative biologists by integrating robust model builders, property and activity predictors, virtual libraries of available chemicals with predicted biological and drug-like properties, and special tools for chemical library design. Chembench was initially developed to support researchers in the Molecular Libraries Probe Production Centers Network (MLPCN) and the Chemical Synthesis Centers.

Please cite this website using the following URL: http://chembench.mml.unc.edu

- -

The Carolina Cheminformatics Workbench (Chembench) is developed by the Carolina Exploratory Center for Cheminformatics Research (CECCR) with the support of the National Institutes of Health (grants P20HG003898 and R01GM066940) and the Environmental Protection

Please login

username

login

Or, login as a guest

New Users

Please register here

Help & Links

Chembench Overview
Chembench Workflows & Methodology
Links to More Cheminformatics Tools

We thank the following commercial vendors:

CHEMICAL COMPUTING GROUP
Scalable Software. Scalable Science.

TALETE srl

Acknowledgements

Research Professors
Clark Jeffries, Alexander Golbraikh, Hao Zhu, Denis Fourches

Postdoctoral Fellows
Aleks Sedykh, Ashutosh Tripathy

Visiting Research Scientist
Eugene Muratov

Adjunct Members
Weifan Zheng, Shubin Liu

MML UNC.EDU

MAJOR FUNDING
NIH
- R01-GM66940
- R01-GM068665
EPA (STAR awards)
- RD832720
- RD833825
- RD834999

Collaborators:
UNC: I. Rusyn, F. Wright
EPA: T. Martin, D. Young
A. Richard, R. Judson, D. Dix, R. Kavlock

Graduate Research Assistants
Nancy Baker, Hao Tang, Jui-Hua Hsieh, Tanarat Kietsakorn, Tong Ying Wu, Liying Zhang, Guiyu Zhao, Andrew Fant, Stephen Bush, Yen Low

Research Programmer
Theo Walker

System Administrator
Mihir Shah