Improved quantitative models of chemical toxicity based on combined application of chemical and biological molecular descriptors

Hao Zhu and Alexander Tropsha

Carolina Center for Computational Toxicology

and

Carolina Center for Environmental Bioinformatics

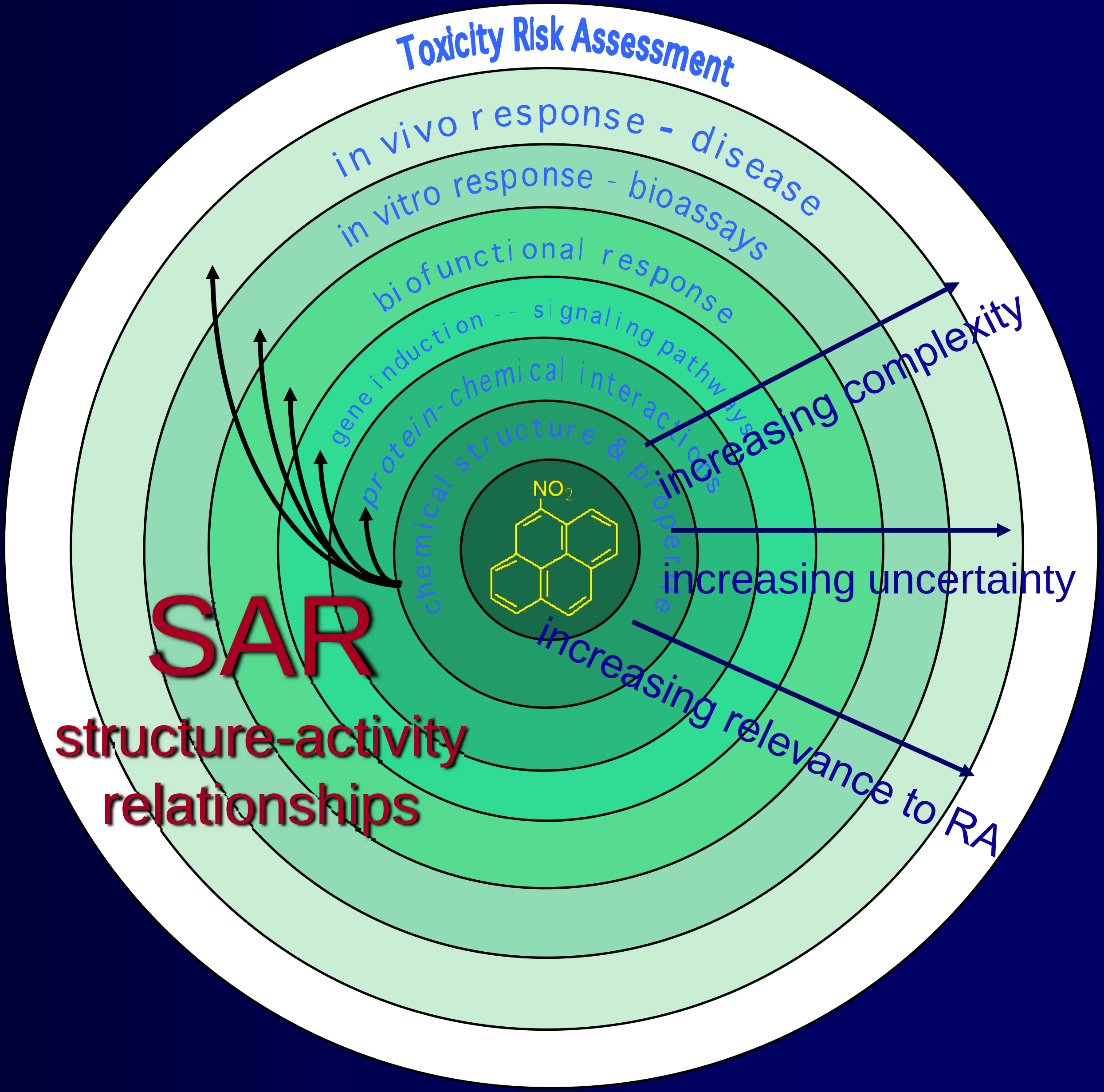
Laboratory for Molecular Modeling

School of Pharmacy, UNC-Chapel Hill

Outline

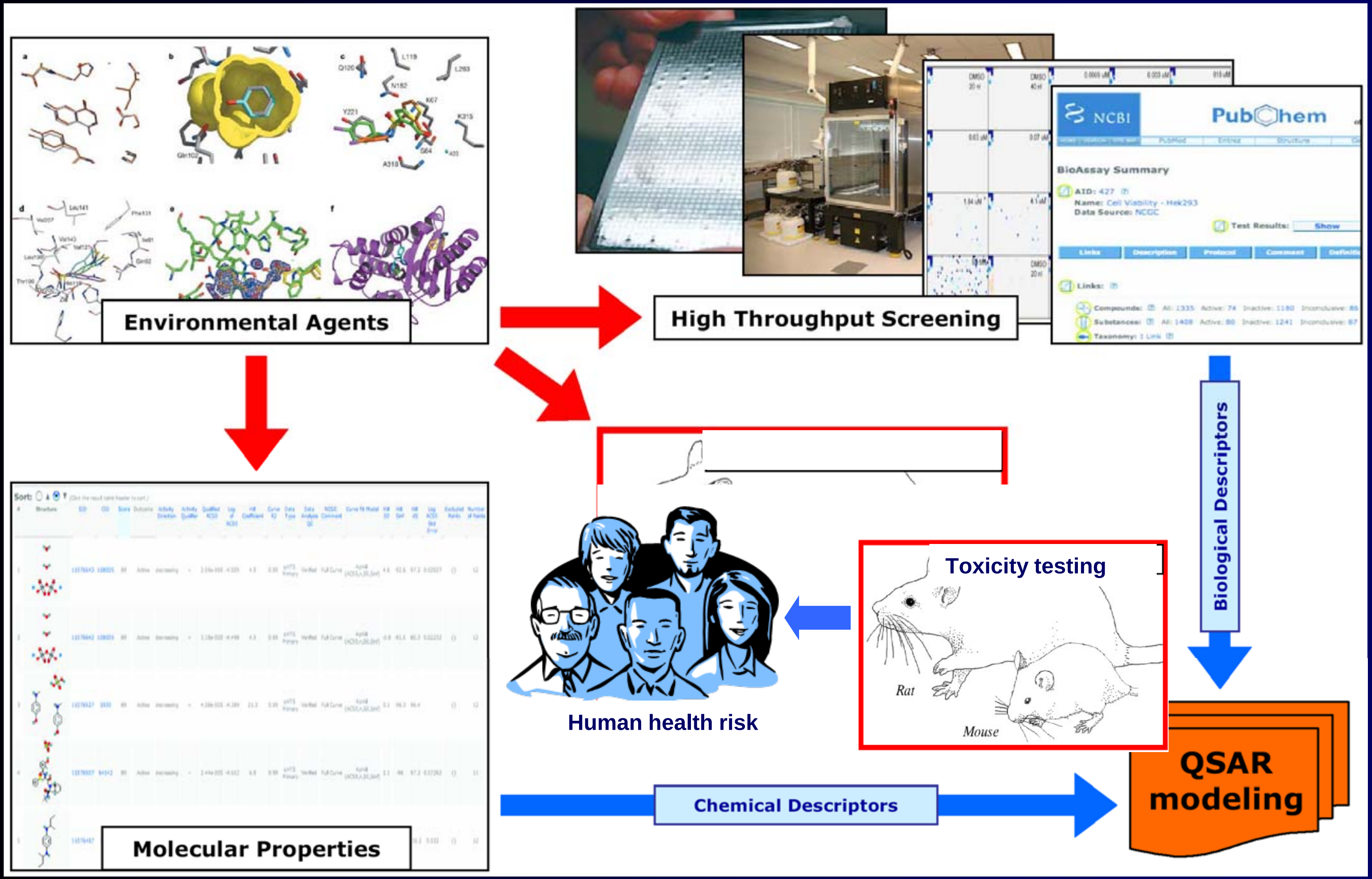
- Overall project vision: exploiting the entire structure – *in vitro* – *in vivo* continuum
- (briefly) Predictive QSAR Modeling Workflow
- Applications
  - novel data partitioning approach based on in vitro – in vivo correlations: Hierarchical QSAR modeling of rodent toxicity
  - analysis of ToxCAST data
    - Modeling of ToxRefDB endpoints using chemical descriptors only
    - Modeling selected *in vivo* end points using hierarchical QSAR modeling

Chemocentric view of biological data



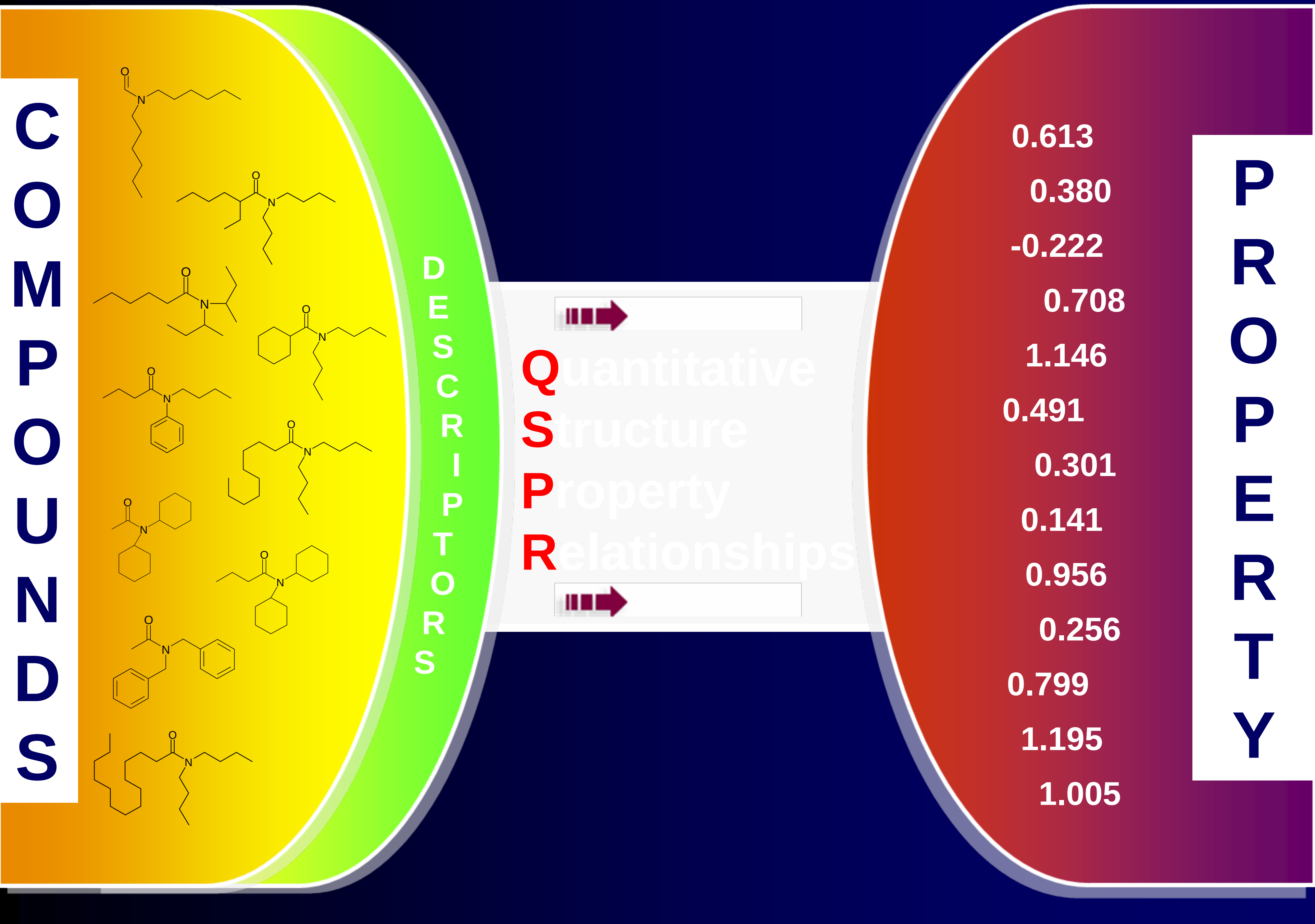
Slide courtesy of Dr. Ann Richard (EPA)

Chemical Structure – *in vitro* – *in vivo* Toxicity Data Continuum.



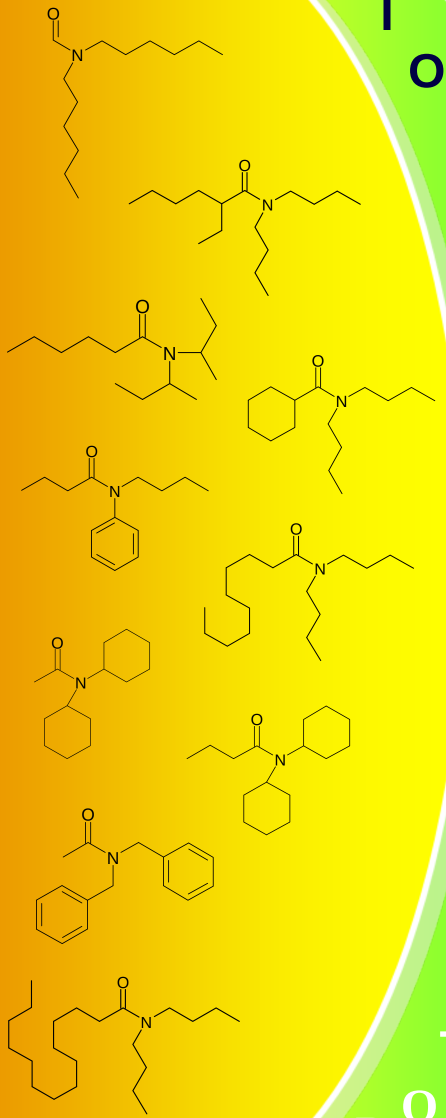
Slide is courtesy of Dr. Ivan Rusyn (UNC)

Principles of QSAR/QSPR modeling



Principle of QSAR/QSPR modeling

COMPOUNDS



BIOPROFILES + DESCRIPTORS

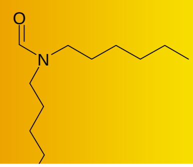
Quantitative **S**tructure **P**roperty **R**elationships

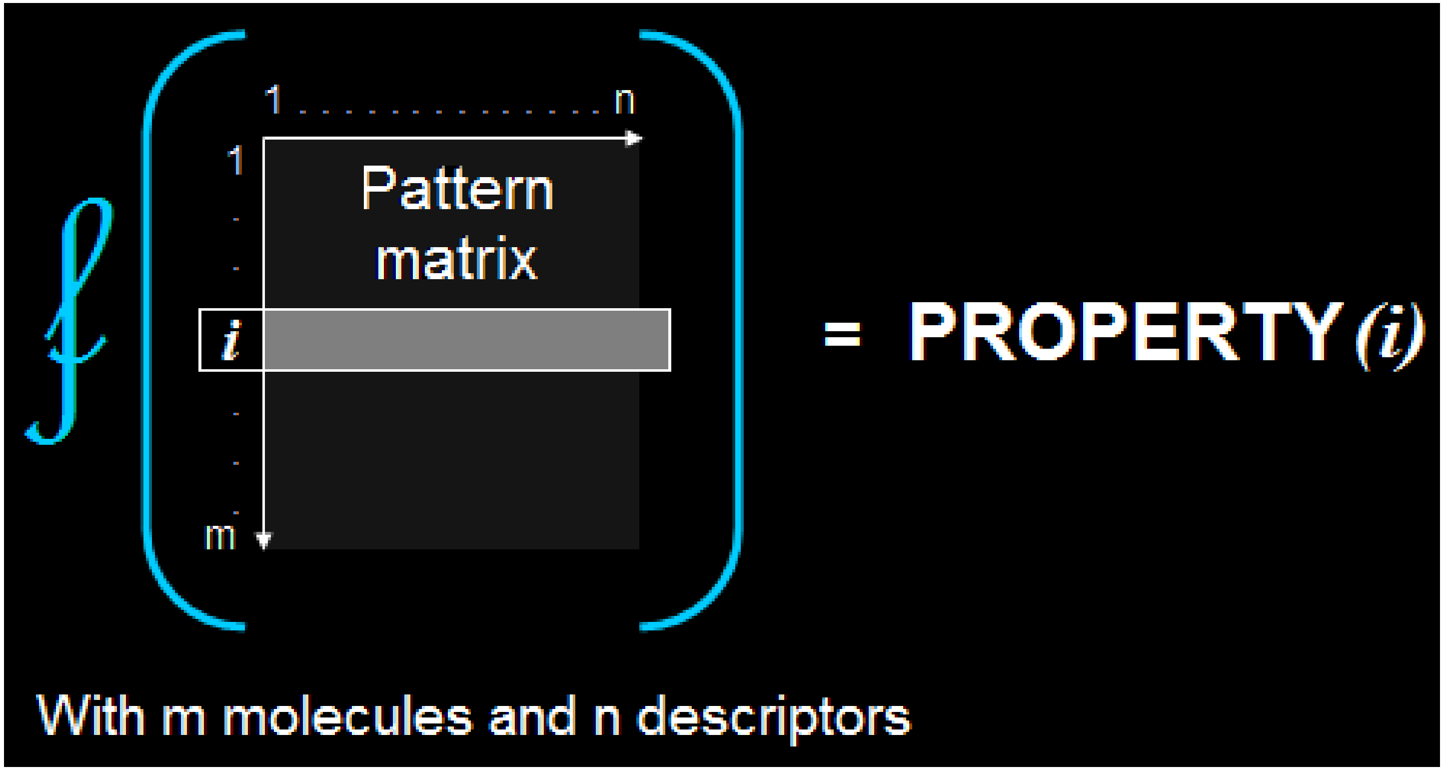
0.613
0.380
-0.222
0.708
1.146
0.491
0.301
0.141
0.956
0.256
0.799
1.195
1.005

PROPERTY

Principles of QSAR/QSPR modeling

C O M P O U N D S



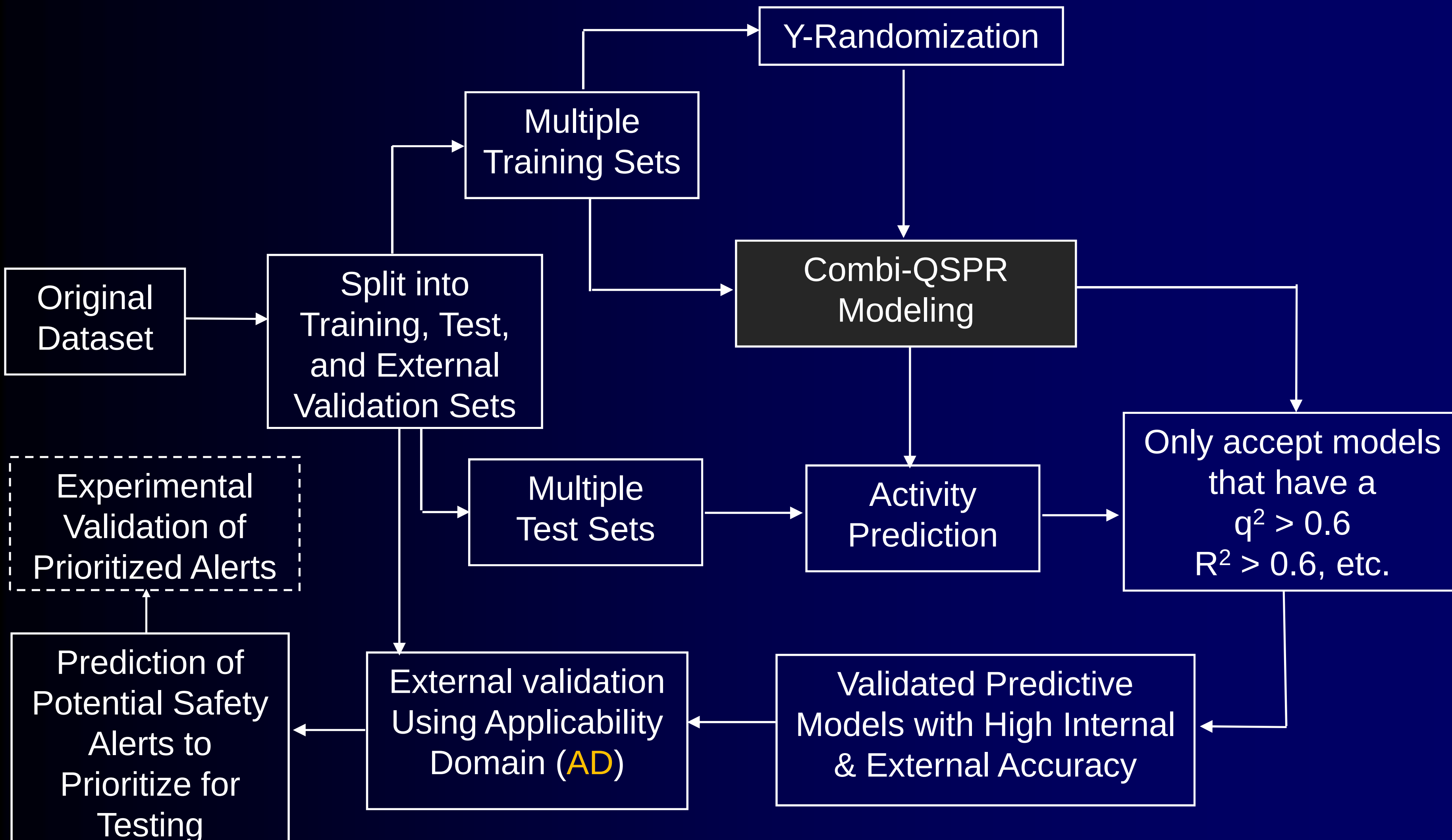


0.613

1.005

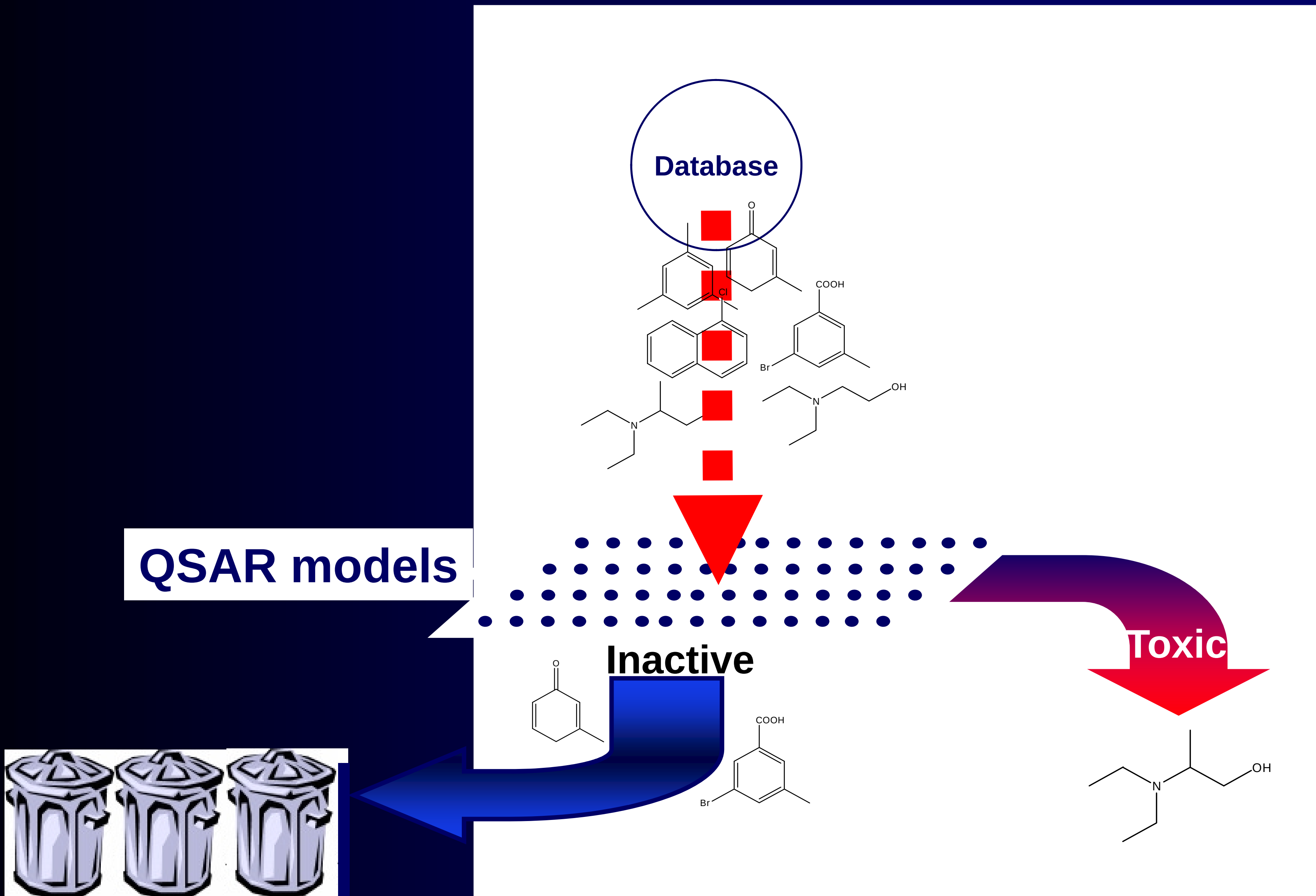
P R O P E R T Y

Predictive QSAR Workflow*



Tropsha, A., Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. Curr. Pharm. Des., 2007, 13, 3494-3504.

Compound prioritization using QSAR models



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

*Zhu, Rusyn, Wright, et al, EHP, 2009(8), 1257-64; in collaboration with Ann Richard, NCCT, US EPA

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

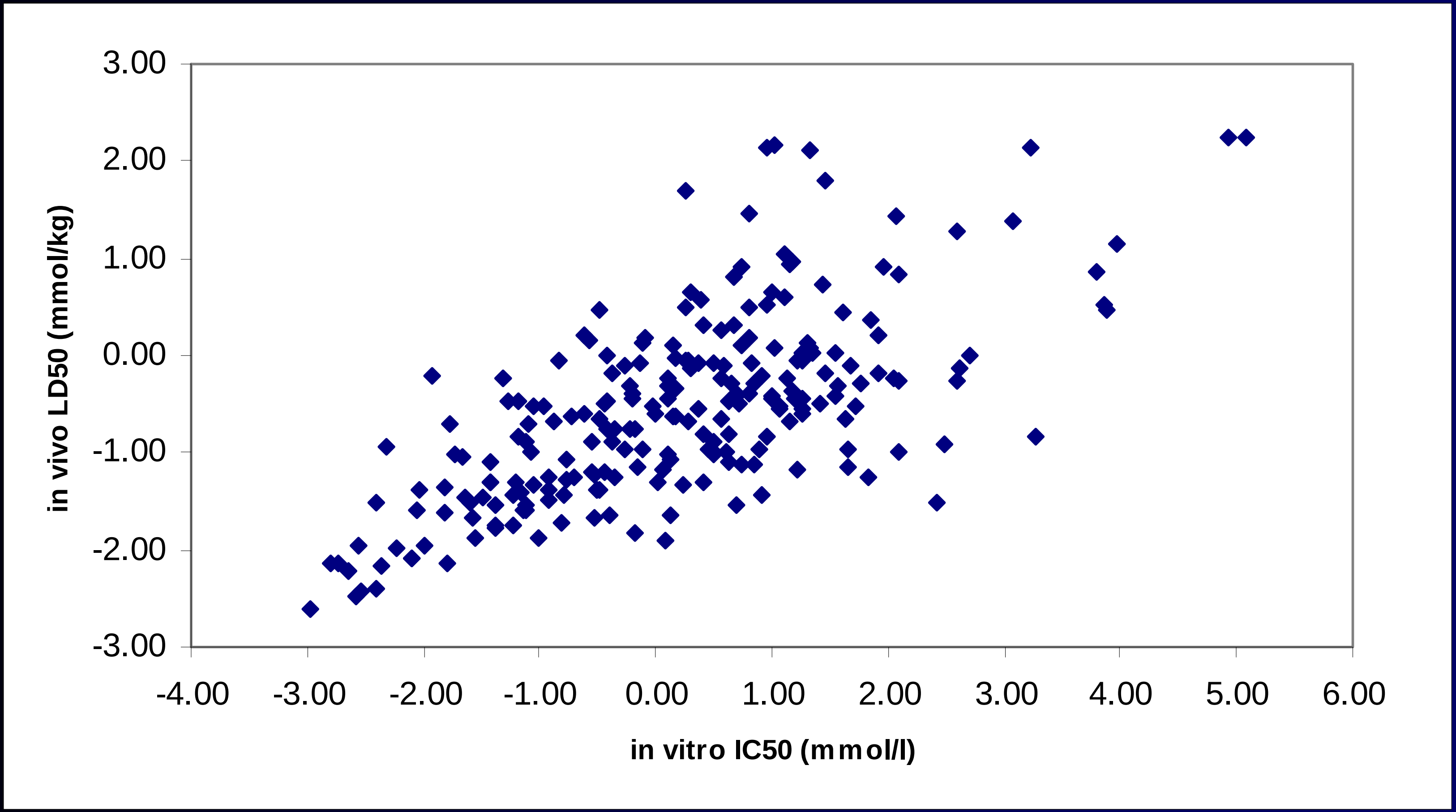
Random split

230 compounds modeling set

23 compounds validation set

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

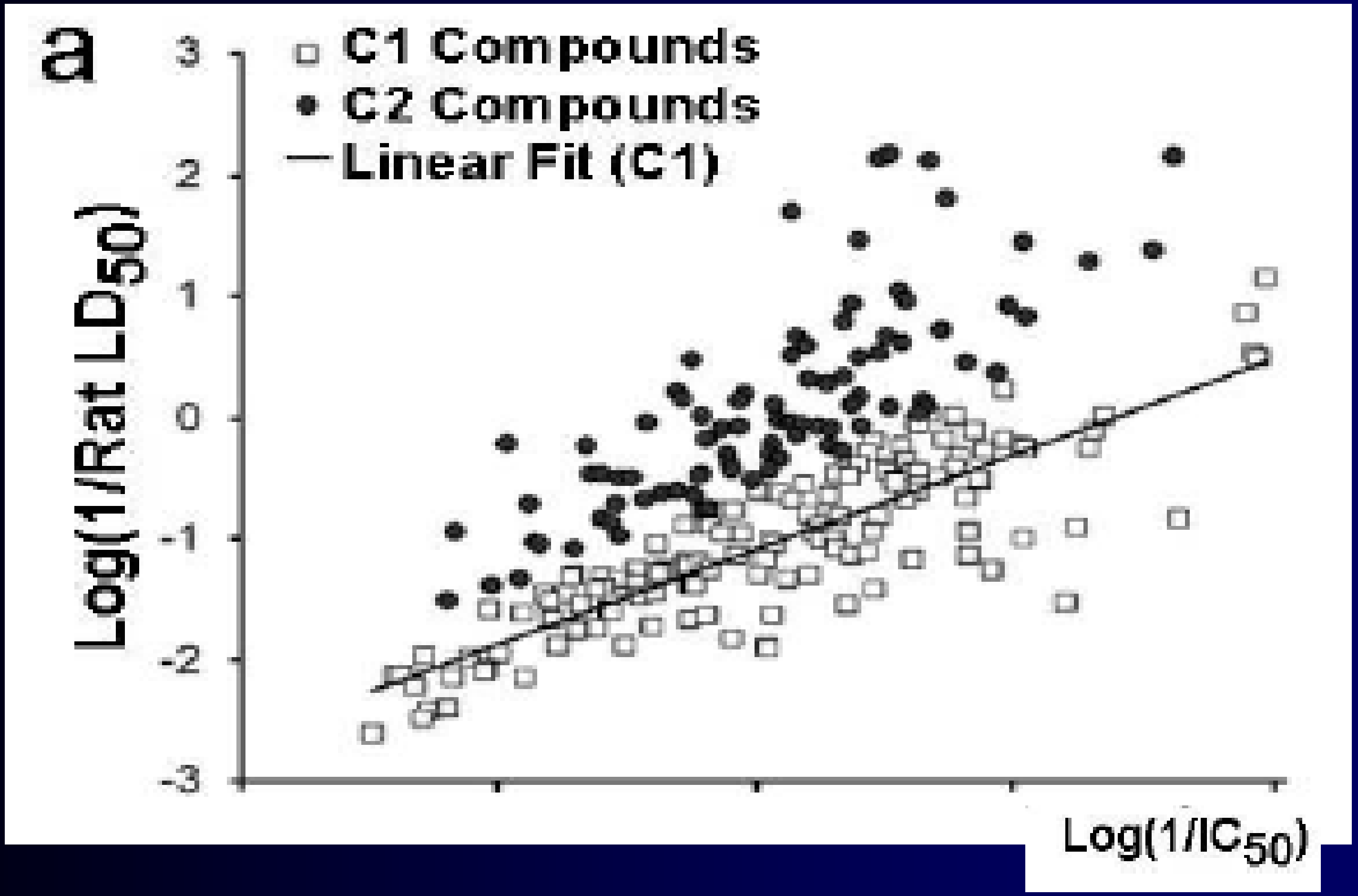
(Rather) poor in vitro-in vivo Correlation Between IC50 and Rat LD50 Values



$R^2=0.46$

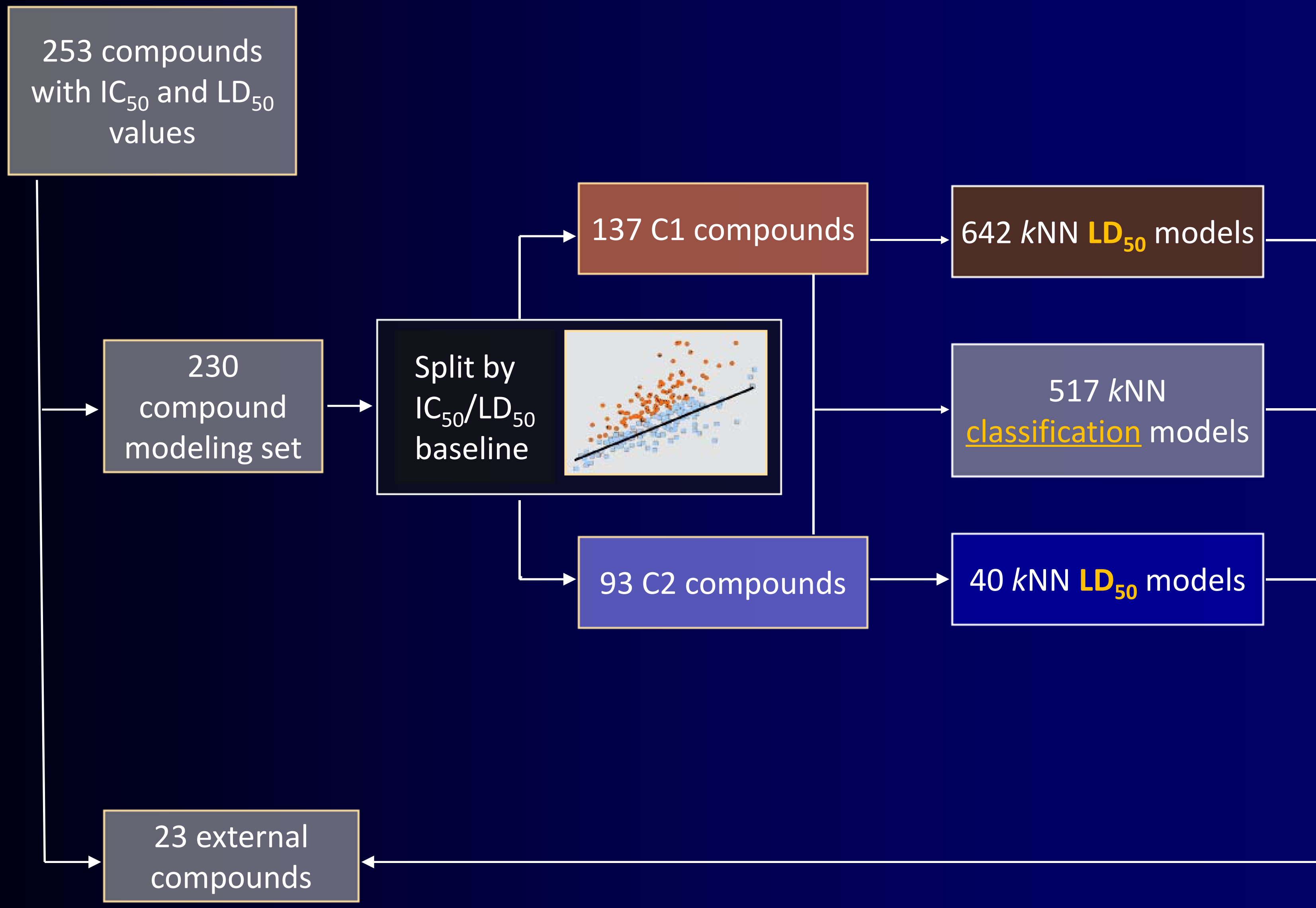
Data partitioning based on the moving regression approach

- IC50 vs. rat LD50 values

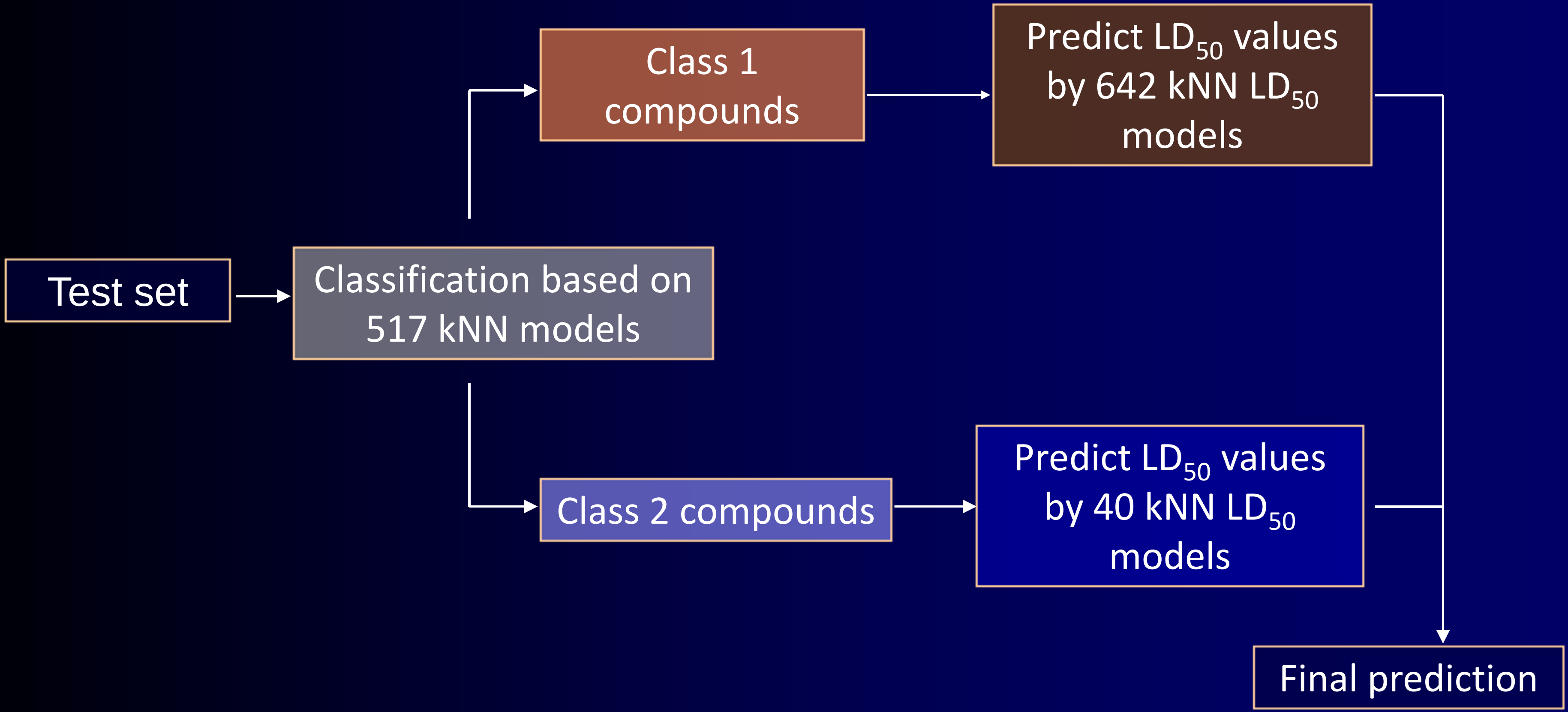


R^2=0.74 for Class 1 compounds

Modeling Workflow



Prediction Workflow



Classification of the Rat LD50 Values for the External Set of 23 Compounds

No AD:
Classification rate = 62%

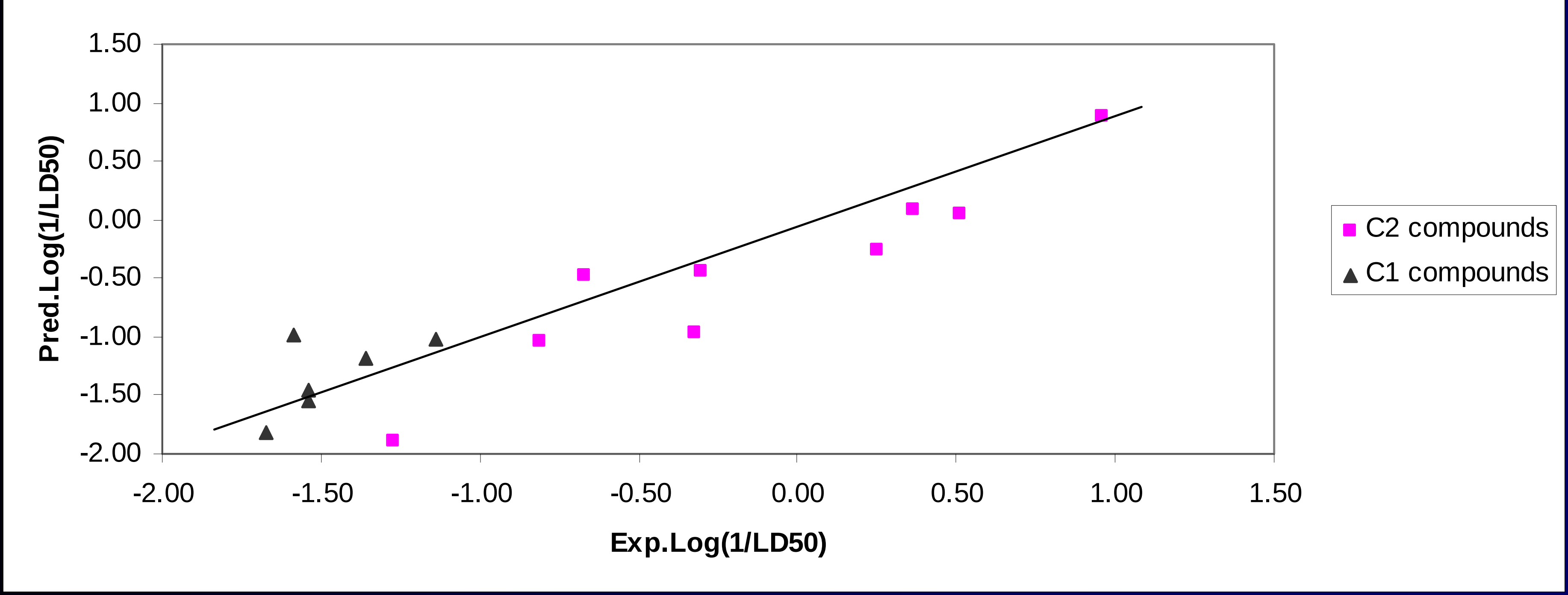
	Pred. C1	Pred. C2
Exp. C1	7	2
Exp. C2	6	5

With AD:
Classification rate = 78%

	Pred. C1	Pred. C2
Exp. C1	6	0
Exp. C2	4	5

Prediction of the Rat LD50 Values of the External 23 Compounds

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



Experimental Study II:

Analysis of ToxCAST data:

(i) ToxRefDB Overview

(ii) Application of Hierarchical QSAR Approach

ToxRefDB* Summary

I. Effective Dose, in mg/kg/day

Chronic toxicity: CHR_Mouse (7); CHR_Rat (14)

Developmental toxicity: DEV_Rabbit (18); DEV_Rat (18)

Reproductive toxicity: Multigeneration, MGR_Rat (19)

-LTD, HTD - Lowest and Highest Tested Doses (10)

II. Category data:

- Supplementary carcinogenicity data,

lesions/tumors of various organs: Mouse (174); Rat(174)

- OPP Carcinogenic potential (2)

*http://www.epa.gov/NCCT/toxrefdb/

Toxcast Assays (08-07-2009). 320 chemical entries

Source	#	Type	Species	Description	Endpoint	Max. test conc.(μM)	≠0	0	N/A
ACEA	7	*In vitro*	*Human*	Cell-growth and morphology	IC50	100	17%	83%	0%
Attagene	73	*In vitro*	*Human*	Transcription factors	LEL	100	8%	92%	0%
BioSeek	174	*In vitro*	*Human*	BioMap system (pharmac.targets, adverse effects), protein markers	LEL	40	11%	89%	0%
Cellumen	57	*In vitro*	*Human*	Cellular toxicity indicators	IC50	200	11%	89%	0%
CellzDirect	42	*In vitro*	*Human*	Gene expression change (transport proteins, metabolic enzymes, etc)	LEL	40	17%	83%	0%
Gentronix	1	*In vitro*	*Human*	HTS (Genotoxicity)	LEL	200	10%	90%	0%
NCGC	18	*In vitro*	*Human(23); Rat(1)*	HTS (Nuclear receptor, cell viability and p53 assays)	IC50	200	3%	97%	0%
NovaScreen	239	*In vitro (biochem)*	*Human(146); Rat(67); Mouse(2); Rabbit(2); Pig(1); Guinea Pig(10); Sheep(2); Cow(9)*	HTS (ADME-Tox, enzyme, nuclear receptor, GPCR)	IC50	20 & 50	3%	97%	0%
Solidus	4	*In vitro*	*Human*	Cytotoxicity	LC50	960	22%	78%	0%
Genes	315	*Mapped*	*Human(231); Rat(65); Mouse(2); Pig(1); Guinea Pig(6); Sheep(2); Cow(8)*	values = min. LEL (over all the assays relevant to the particular gene)	LEL	-	9%	91%	0%
Pathways	438	*Mapped*	*Human(425); Rat(7); Mouse(6)*	min.LEL if > 5 mapped assays	LEL	-	8%	92%	0%
ToxRefDB	76	*In vivo*	*Rat(51); Mouse(7); Rabbit(18)*	In-vivo animal toxicity (mg/kg/day)	LEL	≈10g/kg/day	11%	68%	21%
ToxRefDB	10	*In vivo*	*Rat(6); Mouse(2); Rabbit(2)*	In-vivo animal toxicity (mg/kg/day)	LTD, HTD	HTD	79%	0%	21%
	348	*In vivo*	*Rat(174); Mouse(174)*	Carcinogenicity data	Binary	-	2%	77%	21%
OPP Carc	2	*In vivo*	*Rodents*	Human carcinogen risk judgement	Score	-	22%	46%	32%

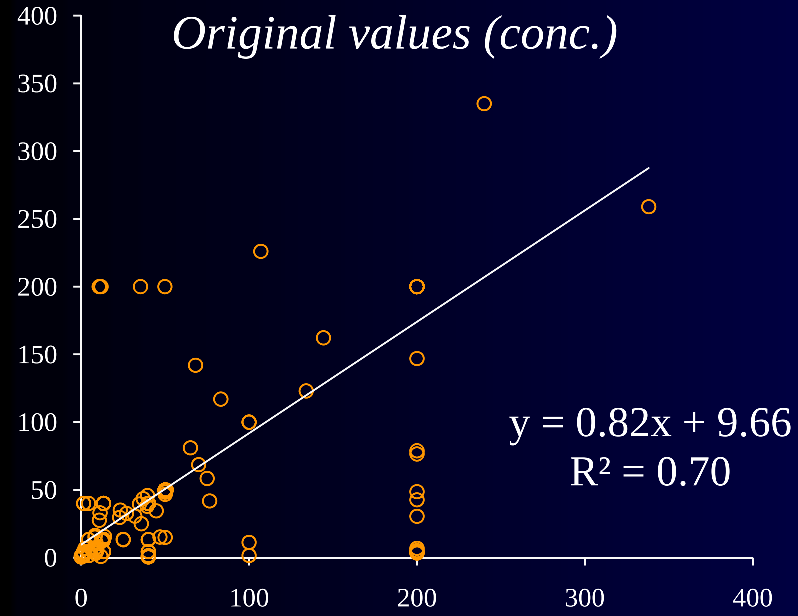
LEL – lowest effective level; LTD, HTD – lowest, highest tested dose

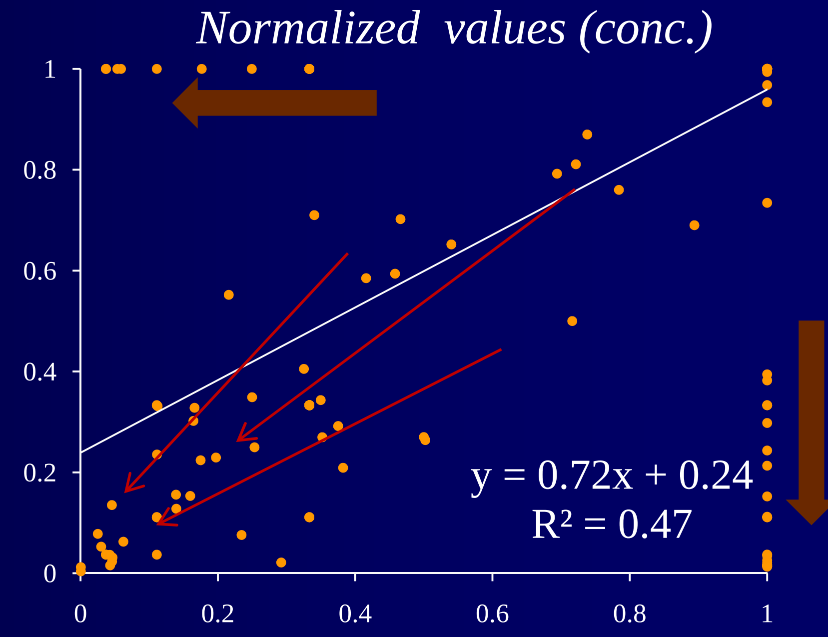
Comparison of the ToxCAST *in vitro* Assay Results for Duplicates/Triplicates (QC)

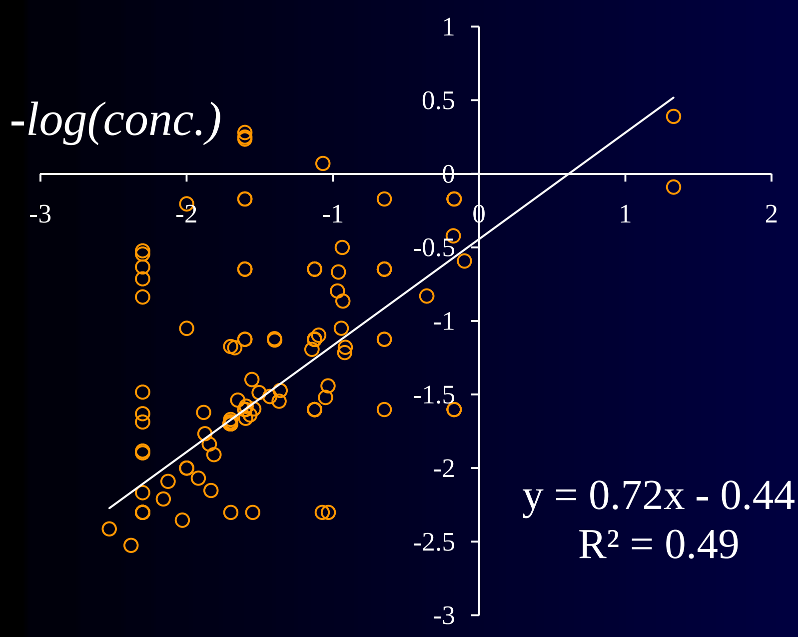
	Assay->	Total	ACEA	ATG	BSK	Cellu	CLZD	NVS	NCGC	Total	ACEA	ATG	BSK	Cellu	CLZD	NVS	NCGC
	n=	*610*	*7*	*73*	*174*	*57*	*42*	*239*	*18*	*610*	*7*	*73*	*174*	*57*	*42*	*239*	*18*
CAS	**Chemical**	R correlation coefficients (-log [Conc.])								# of non-zero signals							
55406-53-6	3-Iodo-2-propynyl-butylcarbamate	0.70	0.91	NA	0.57	0.40	0.44	0.93	NA	90	4	0	48	15	1	17	0
741-58-2	Bensulide	0.87	0.38	0.97	0.16	0.35	0.47	0.91	1.00	83	2	4	14	12	17	30	2
64902-72-3	Chlorsulfuron	0.67	NA	NA	-0.03	NA	0.00	1.00	NA	1	0	0	1	0	0	0	0
84-74-2	Dibutyl phthalate	0.74	NA	0.60	0.02	-0.02	0.02	0.95	NA	22	0	4	12	1	1	4	0
51338-27-3	Diclofop-methyl	0.88	1.00	0.97	0.22	0.45	0.31	0.26	1.00	23	2	4	11	1	3	1	2
759-94-4	EPTC	0.89	NA	NA	-0.01	NA	0.98	0.00	NA	3	0	0	0	0	3	0	0
66441-23-4	Fenoxaprop-ethyl	0.85	NA	0.33	-0.02	0.10	0.00	0.03	NA	21	0	1	17	2	0	1	0
94125-34-5	Prosulfuron	0.86	NA	0.66	-0.02	NA	0.71	0.50	NA	9	0	1	3	0	5	1	0
		R correlation coefficients (binary data)								Accuracy (binary data) = (TP + TN)/(TP+TN+FP+							
55406-53-6	3-Iodo-2-propynyl-butylcarbamate	0.74	0.73	NA	0.70	0.47	0.48	0.89	NA	0.93	0.86	0.99	0.86	0.74	0.93	0.98	1.00
741-58-2	Bensulide	0.76	0.55	0.94	0.56	0.43	0.76	0.94	1.00	0.94	0.79	0.99	0.90	0.75	0.87	0.99	1.00
64902-72-3	Chlorsulfuron	0.07	NA	NA	0.07	NA	NA	NA	NA	0.97	0.71	0.99	0.91	0.98	0.95	1.00	1.00
84-74-2	Dibutyl phthalate	0.49	NA	0.52	0.48	0.49	0.38	0.81	NA	0.94	0.57	0.90	0.89	0.95	0.88	0.99	1.00
51338-27-3	Diclofop-methyl	0.53	1.00	1.00	0.40	0.48	0.45	0.35	1.00	0.95	1.00	1.00	0.86	0.92	0.85	0.99	1.00
759-94-4	EPTC	0.27	NA	NA	-0.04	NA	1.00	NA	NA	0.98	0.86	1.00	0.92	1.00	1.00	1.00	1.00
66441-23-4	Fenoxaprop-ethyl	0.53	NA	0.43	0.65	0.35	NA	0.35	NA	0.95	0.71	0.95	0.90	0.89	0.98	0.98	1.00
94125-34-5	Prosulfuron	0.47	NA	0.70	0.23	NA	0.77	1.00	NA	0.98	1.00	0.99	0.93	0.98	0.94	1.00	1.00

triplicate results are averaged; shaded cells reflect significant # of signals

ToxCast Assays comparison example : 3-Iodo-2-propynylbutylcarbamate

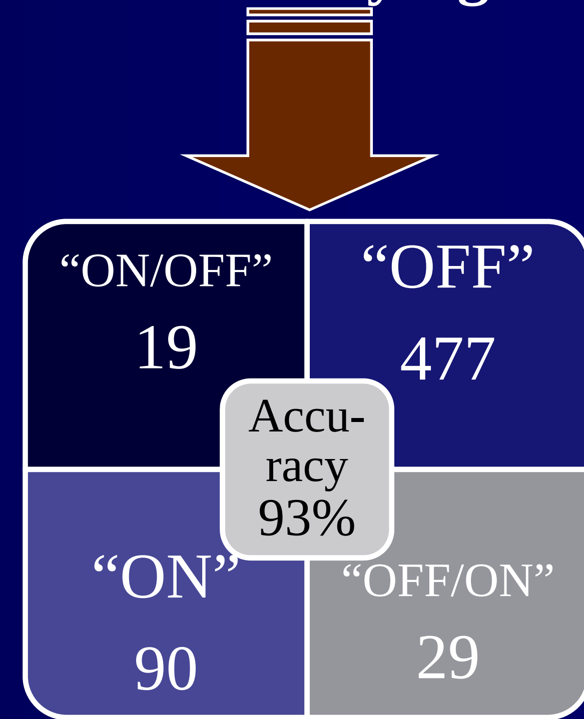






Quantitave data too "noisy"…

Binary signal



Data Curation

- *In-vitro* assays: 615→ 284
  - Remove one of two highly correlated (R^2 >0.95) assays and low-signal (<10 non-zero entries) assays
- Chemicals: 320 → 230~250
  - duplicate structures, mixtures, inorganic compounds, macromolecules were removed
  - Kept only those for which ***in-vivo*** data is available

    (may vary for different endpoints)

Focusing on a small subset of data: Multi-Generation Rat Toxicity

- Important: Reproductive Toxicity
- 3 out of 19 assays with the highest fraction of actives chosen for initial studies:

  MGR_Rat_Kidney (78 actives)

  MGR_Rat_Liver (110 actives)

  MGR_Rat_ViabilityPND4 (70 actives)

Conventional QSAR Modeling

- Using chemical descriptors only:

  -1224 Dragon chemical descriptors

- QSAR approaches:

  - Random Forest (RF)

  - SVM linear kernal

  - SVM rbf kernel

Breiman L. Machine Learning 45 (2001): 5-32

5-Fold Cross Validation Result

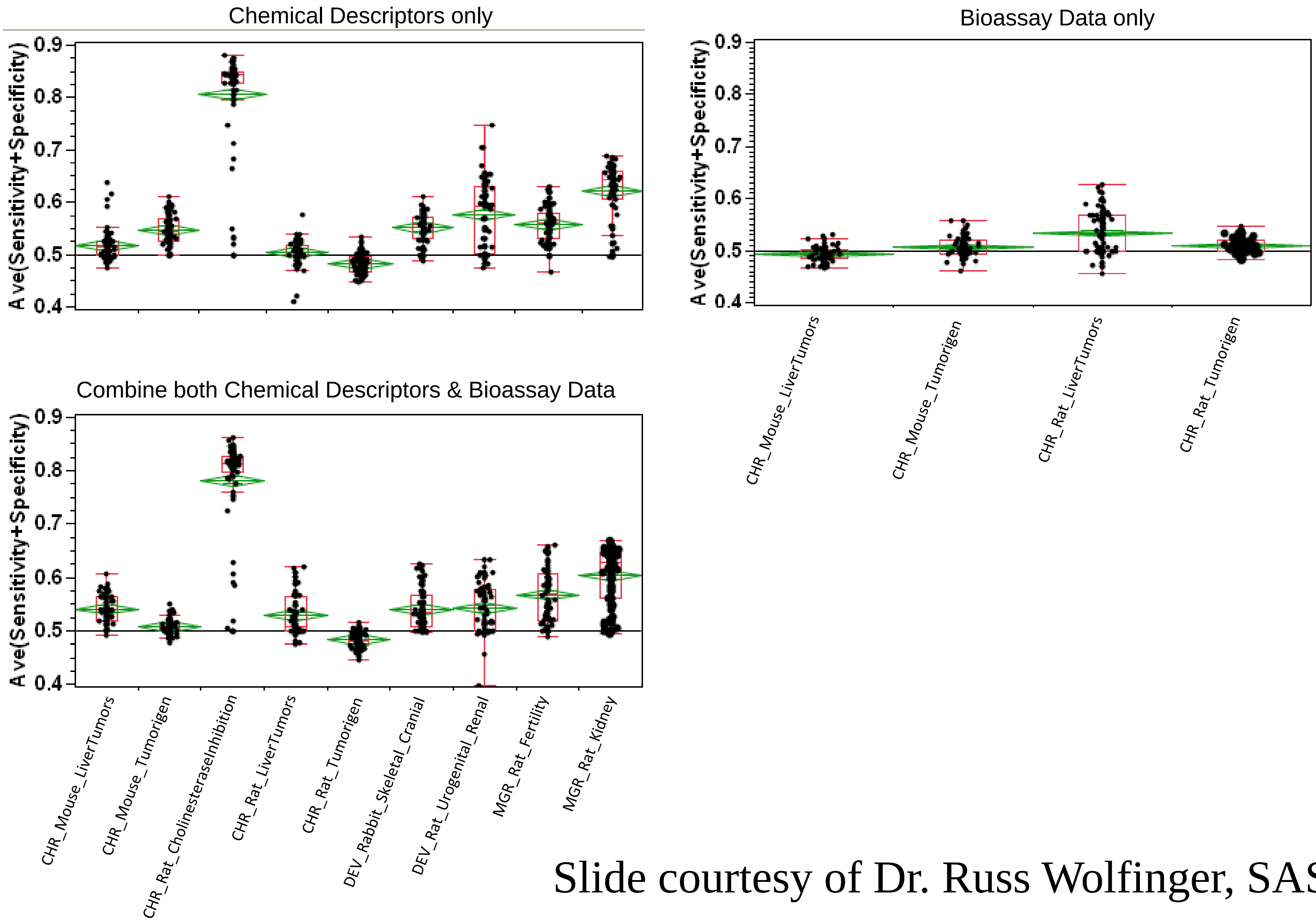
	MGR_Rat Kidney	MGR_Rat Liver	MGR_Rat ViabilityPND4
RF sensitivity	0.22	0.53	0.12
RF specificity	0.90	0.75	0.93
RF CCR	**0.56**	**0.64**	**0.53**
SVM_linear sensitivity	0.35	0.56	0.32
SVM_linear specificity	0.65	0.64	0.73
SVM_linear CCR	**0.50**	**0.60**	**0.53**
SVM_rfb sensitivity	0.37	0.61	0.23
SVM_rfb specificity	0.81	0.69	0.83
SVM_rfb CCR	**0.59**	**0.65**	**0.53**

$$CCR = \frac{1}{K}\sum_{k=1}^{K}\frac{N_k^{corr}}{N_k^{total}}$$

In vivo toxicity prediction using either biological or hybrid (chemical plus biological descriptors)

- Using ToxCAST bioassay results as biological descriptors did not result in any statistically significant models.
- The use of hybrid (biological + chemical descriptors) did not improve the results either.
- These results are similar to the analysis of ToxCAST data by the SAS team.

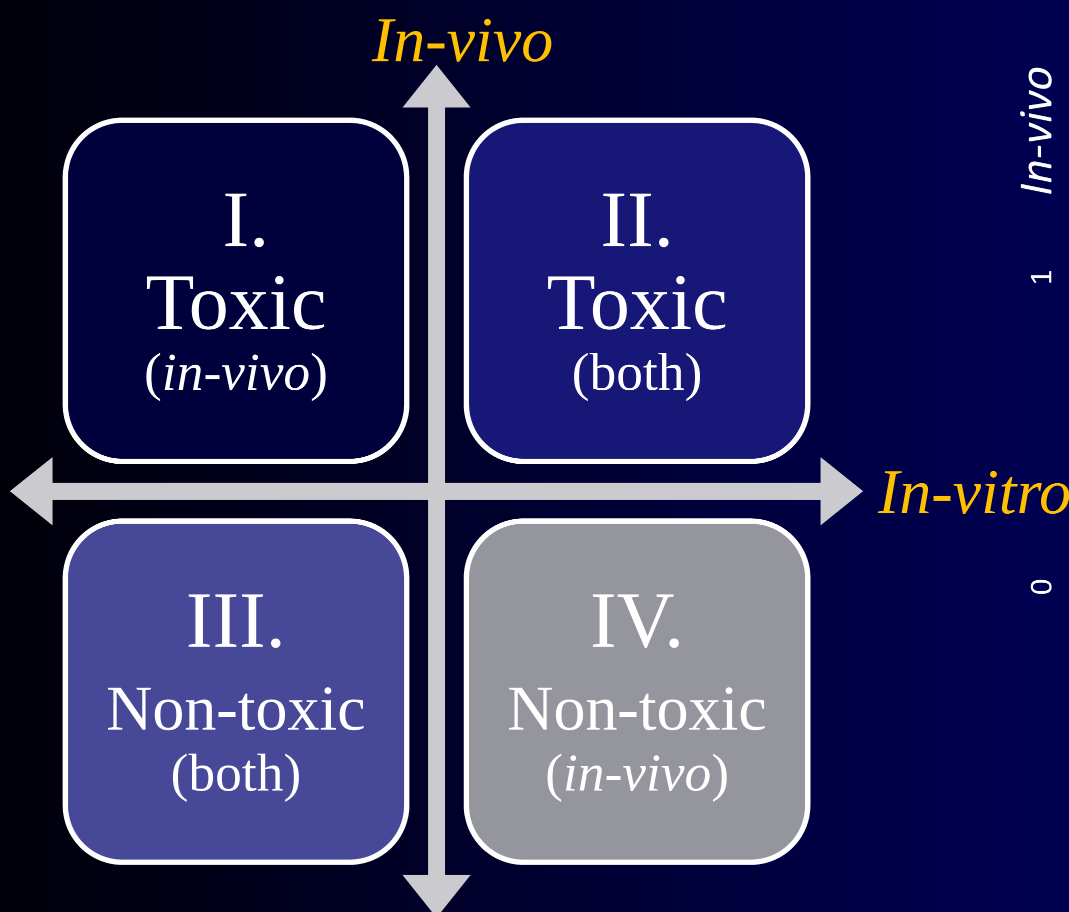
Prediction Comparison Based on Ave (Sensitivity + Specificity)

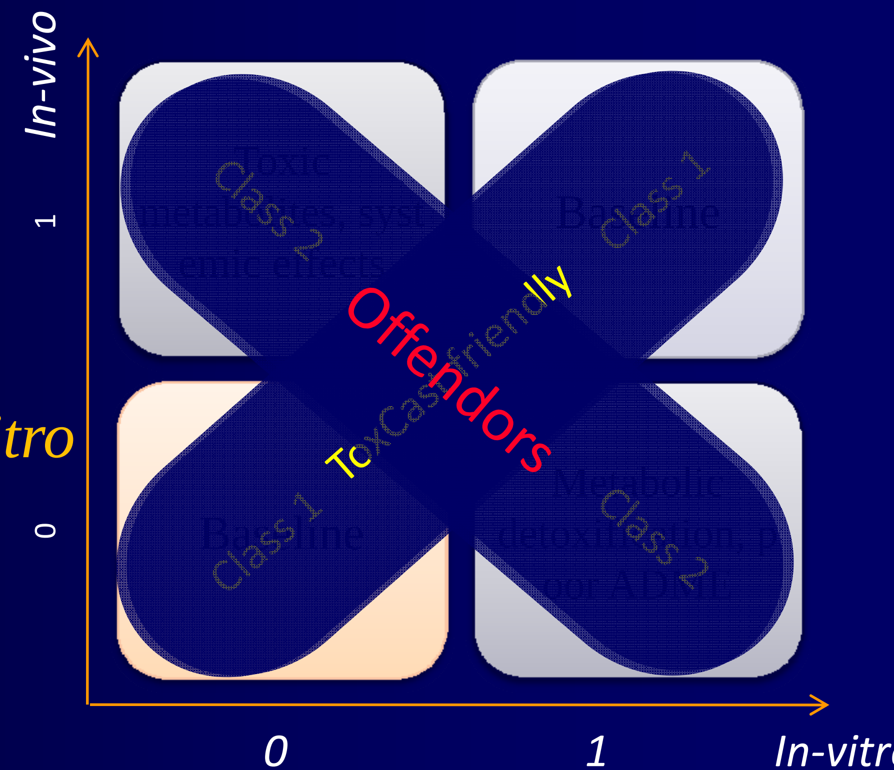


Slide courtesy of Dr. Russ Wolfinger, SAS

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

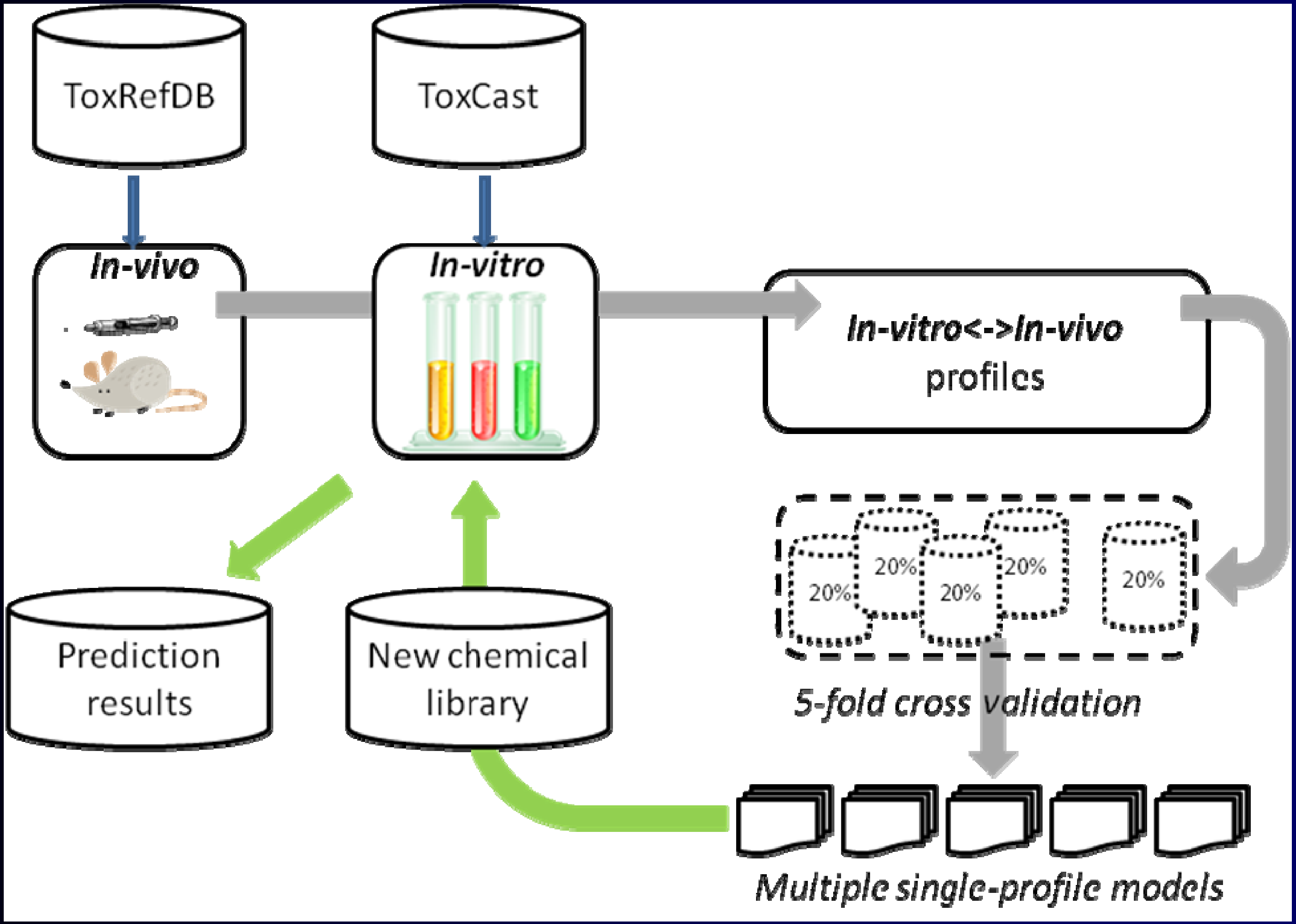
For <u>each</u> In-vitro vs. *In-vivo* profile (~1000 combinations):



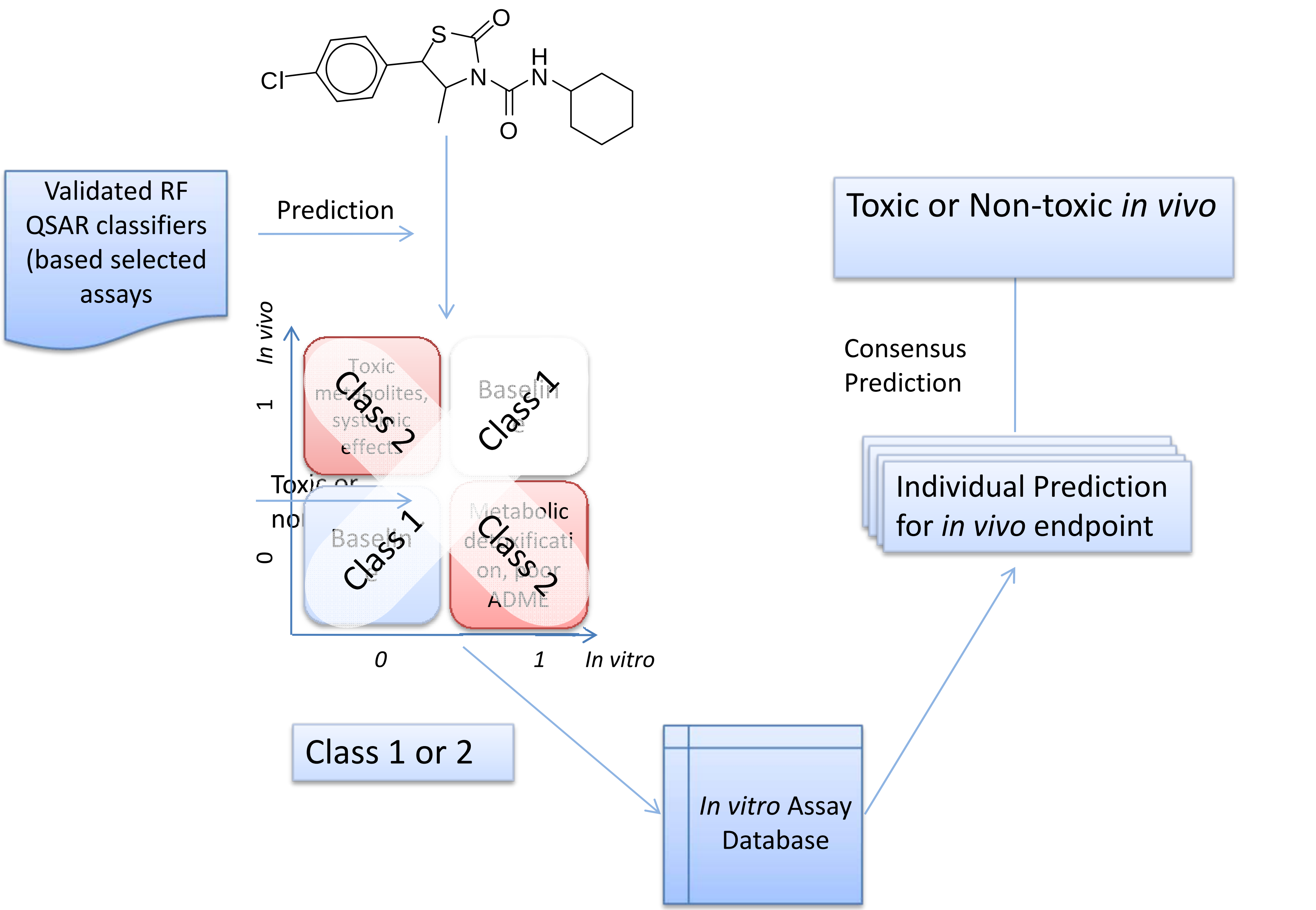


- Binary classification QSAR for "baseline" (II & III) vs. off-line (I & IV) using chemical descriptors only

Modeling Workflow for ~1000 in vitro – in vivo Series Using Various QSAR Approaches and Dragon descriptors



External prediction workflow



General Consensus Prediction Gave the Similar Prediction Accuracy Compared to Conventional QSAR Model

	MGR_Rat Kidney		**MGR_Rat Liver**		**MGR_Rat ViabilityPND4**	
	Hybrid	Chemical	Hybrid	Chemical	Hybrid	Chemical
RF sensitivity	0.25	0.22	0.6	0.53	0.12	0.12
RF specificity	0.88	0.90	0.77	0.75	0.93	0.93
RF CCR	**0.57**	**0.56**	**0.69**	**0.64**	**0.53**	**0.53**
SVM_linear sensitivity	0.37	0.35	0.56	0.56	0.34	0.32
SVM_linear specificity	0.68	0.65	0.67	0.64	0.73	0.73
SVM_linear CCR	**0.53**	**0.50**	**0.61**	**0.60**	**0.53**	**0.53**
SVM_rfb sensitivity	0.35	0.37	0.6	0.61	0.25	0.23
SVM_rfb specificity	0.81	0.81	0.7	0.69	0.83	0.83
SVM_rfb CCR	**0.58**	**0.59**	**0.65**	**0.65**	**0.54**	**0.53**

Possible Reasons for the Lack of Significant Improvement

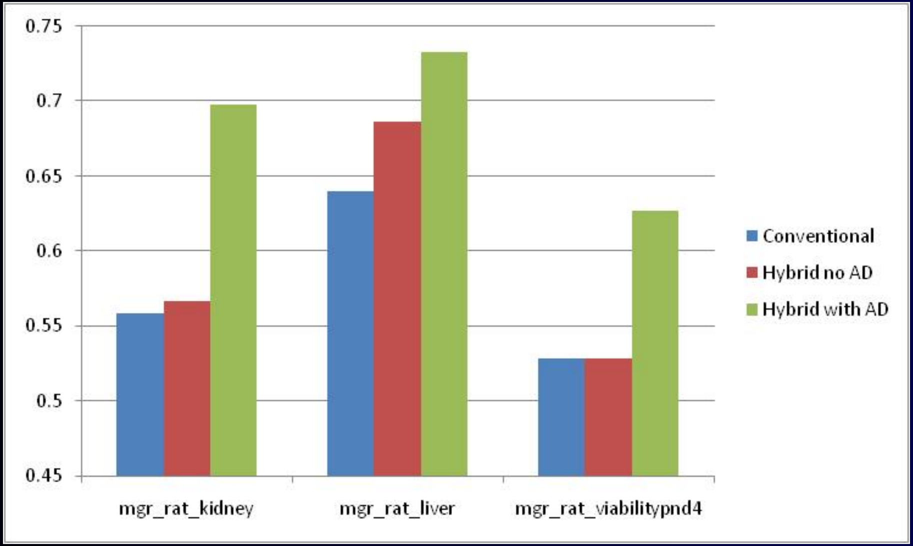
- 1. The quality of the data
- 2. Too many irrelevant bioassays were included in the consensus prediction

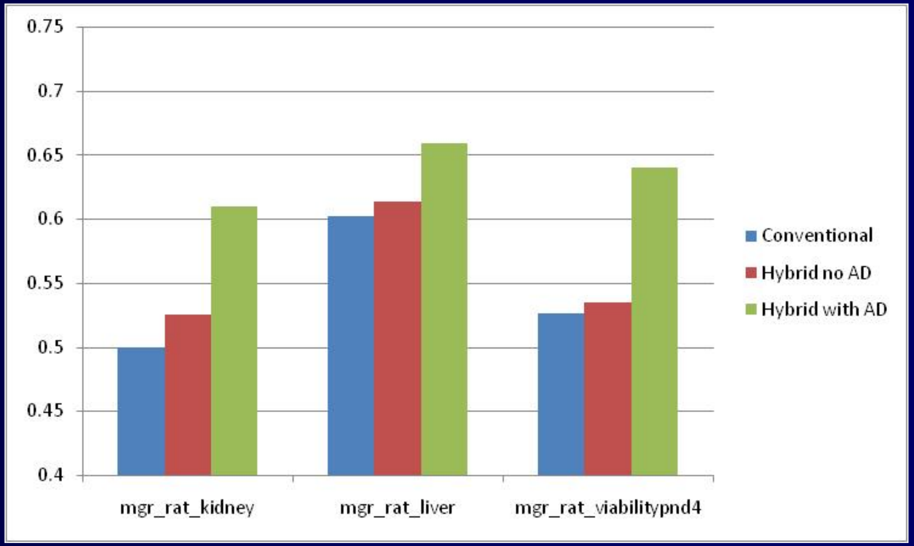
- Solution: the use of more restrictive threshold for active/inactive definitions as a special <u>Applicability Domain</u>
- >0.7 instead of 0.5 was used to define actives; <0.1 instead of 0.5 was used to define inactives.

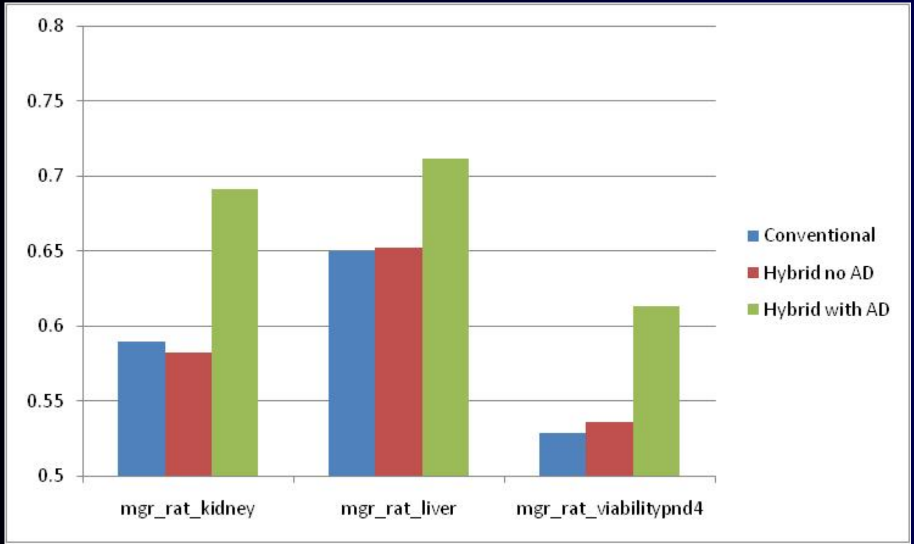
General Consensus Prediction Gave Clear Better Prediction Accuracy with Applicability Domain

	MGR_Rat Kidney		MGR_Rat Liver		MGR_Rat ViabilityPND4	
	Hybrid	Chemical	Hybrid	Chemical	Hybrid	Chemical
RF sensitivity	0.56	0.22	0.77	0.53	0.33	0.12
RF specificity	0.84	0.90	0.70	0.75	0.92	0.93
RF CCR	**0.70**	**0.56**	**0.73**	**0.64**	**0.63**	**0.53**
SVM_linear sensitivity	0.70	0.35	0.86	0.56	0.71	0.32
SVM_linear specificity	0.52	0.65	0.46	0.64	0.57	0.73
SVM_linear CCR	**0.61**	**0.50**	**0.66**	**0.60**	**0.64**	**0.53**
SVM_rfb sensitivity	0.68	0.37	0.86	0.61	0.48	0.23
SVM_rfb specificity	0.71	0.81	0.56	0.69	0.75	0.83
SVM_rfb CCR	**0.69**	**0.59**	**0.71**	**0.65**	**0.61**	**0.53**

Comparison of Prediction Accuracy (CCR) of Conventional and Hybrid QSAR Models







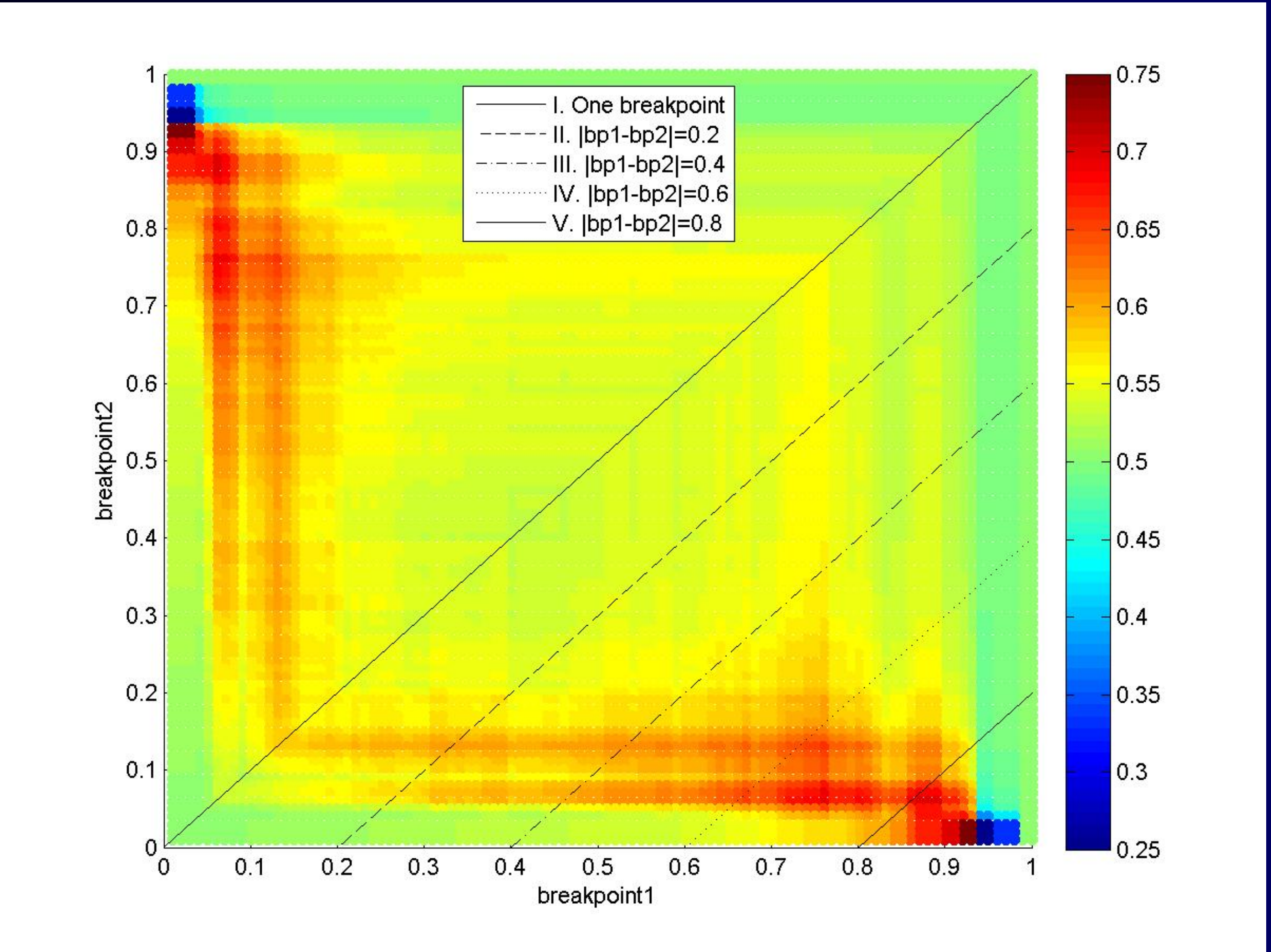
- RF Models
- SVM linear kernel models
- SVM rbf kernel models

The Coverage is Reasonable for the Resulting Models

	MGR_Rat Kidney	MGR_Rat Liver	MGR_Rat ViabilityPND4
RF	0.41	0.56	0.40
SVM_linear	0.61	0.66	0.64
SVM_rbf	0.43	0.53	0.38

Using Different Thresholds Could Result in Different Prediction Accuracy

SVM_rbf MGR_Rat ViabilityPND4 results:

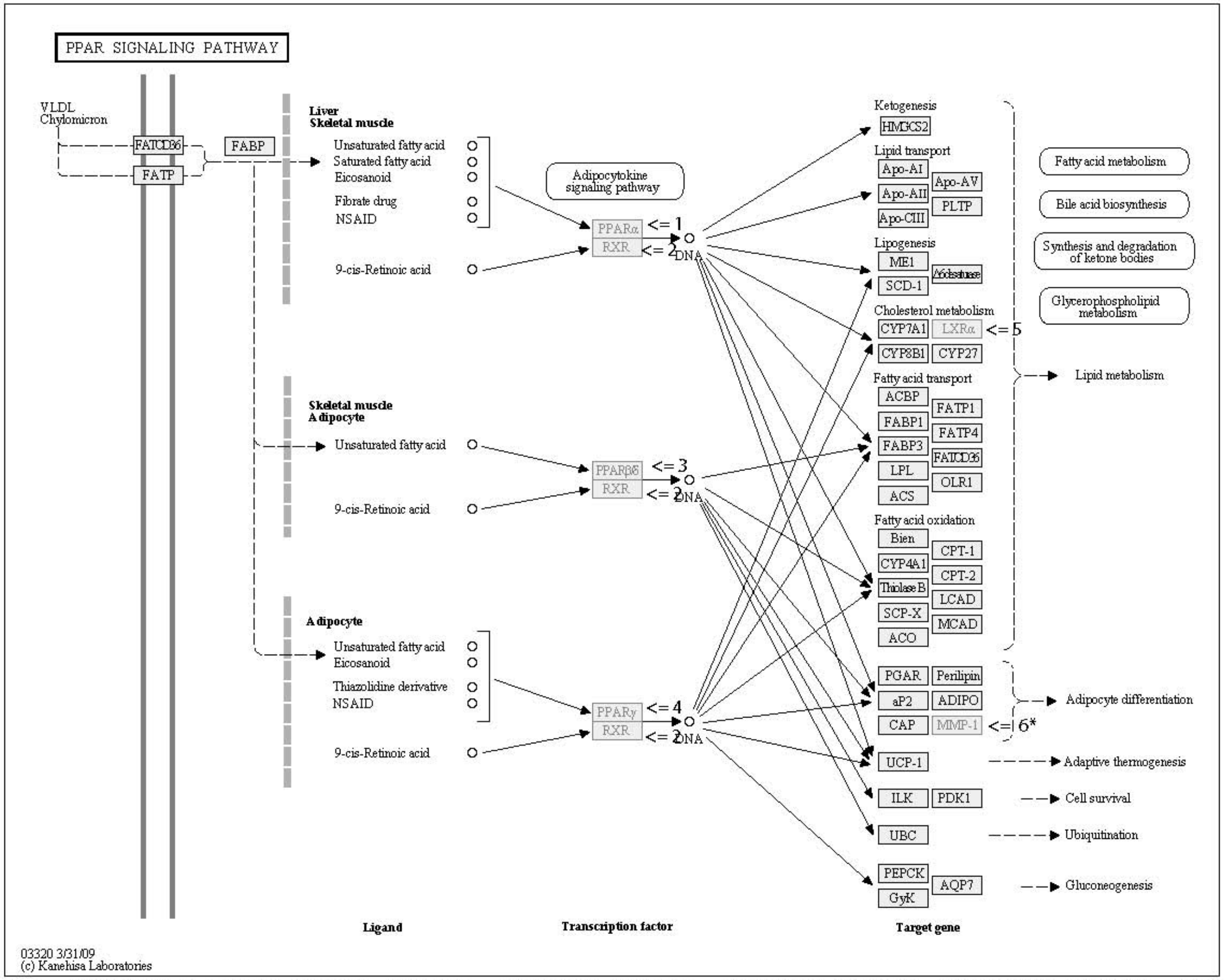


The Use of Toxicological Pathway Information to Improve the Model*

- There are three groups of data could be relevant to toxicological pathways:
  - Attagene: 73 assays;
  - BioSeek: 174 assays;
  - NovaScreen: 239 assays.

*All pathway information was provided by Dr. Shawn Gomez

Example of Some ToxCast Assays That Could be Mapped onto Pathways



1) PPARA hsa:5465 K07294 ATG_PPARa_TRANS & ATG_PPRE_CIS & NVS_NR_hPPARa

2) ATG_RXRa_TRANS & ATG_RXRb_TRANS RXRA hsa:6256 K08524 & RXRB hsa:6257 K08525

3) PPARD hsa:5467 K04504 ATG_PPARd_TRANS

4) PPARG hsa:5468 K08530 ATG_PPARg_TRANS & NVS_NR_hPPARg

5) NR1H3 hsa:10062 k08536 ATG_DR4_LXR_CIS & ATG_LXRa_TRANS

*6) MMP1 hsa:4312 k01388 BSK_BE3C_MMP1 & BSK_hDFCGF_MMP1 & NVS_ENZ_hMMP1

Obtained from Dr. Shawn Gomez

Overview of the ToxCast Assays that have been mapped to pathways (so far)

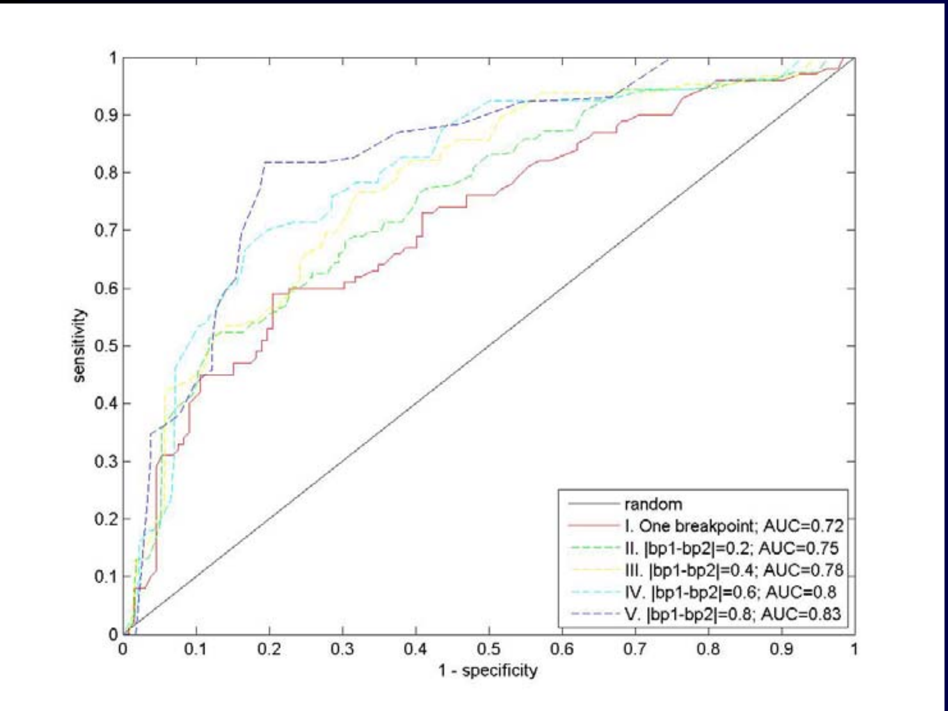
	Nm. of assays	**Nm. of assays (no low active %)**
JAK-STAT Signaling Pathway	8	0
PPAR Signaling Pathway	13	7
Focal adhesion	14	5
T cell receptor signaling pathway	13	4
TGF-beta signaling pathway	10	6
Toll-like receptor signaling pathway	25	20
Apoptosis	15	9
Wnt signaling pathway	10	2
Adipocytokine signaling pathway	12	4
Leukocyte transendothelial migration	11	9
MAPK signaling pathway	24	14
ErbB signaling pathway	11	3
Natural killer cell mediated cytotoxicity	13	4
Cell cycle	9	3
p53 signaling pathway	7	4
BMPR2 interaction with PPARG	8	6
All	95	54

Using 54 ToxCast Pathway-mapped Assays for Consensus Prediction

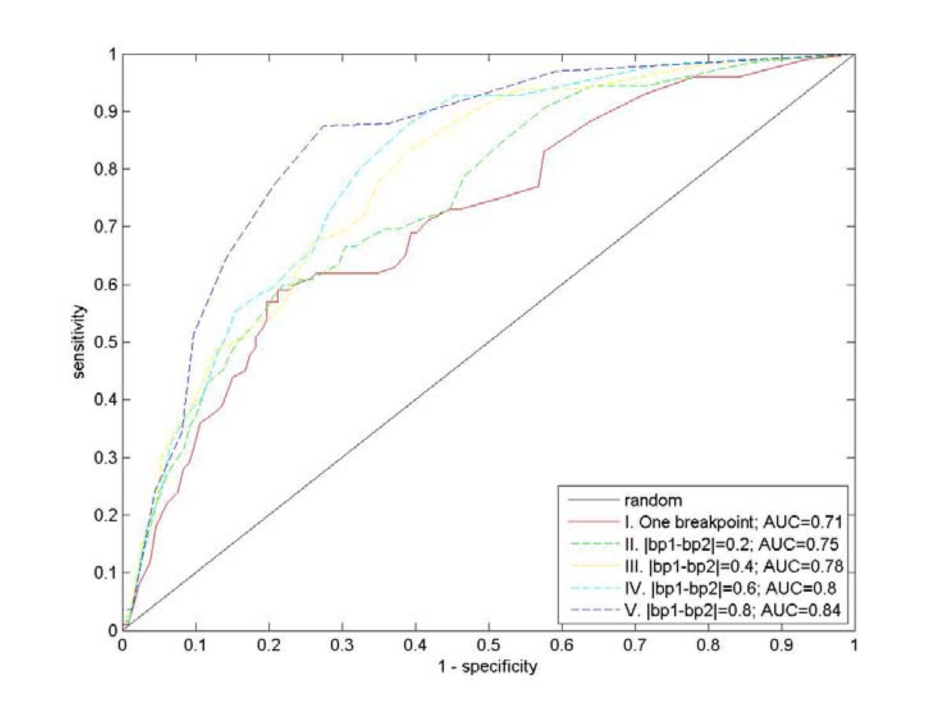
	MGR_Rat Kidney		**MGR_Rat Liver**		**MGR_Rat ViabilityPND4**	
	CCR	Coverage	CCR	Coverage	CCR	Coverage
RF all assays	0.70	0.41	0.73	0.56	0.63	0.4
54 pathway assays	**0.67**	**0.48**	**0.72**	**0.6**	**0.60**	**0.46**
23 pathway assays	0.65	0.44	0.73	0.55	0.58	0.4
10 pathway assays	0.62	0.43	0.71	0.5	0.54	0.36
SVM_linear all assays	0.61	0.38	0.66	0.42	0.64	0.32
54 pathway assays	**0.64**	**0.4**	**0.63**	**0.42**	**0.64**	0.29
23 pathway assays	0.63	0.28	0.57	0.31	0.64	0.22
10 pathway assays	0.6	0.28	0.54	0.24	0.61	0.2
SVM_rfb all assays	0.69	0.43	0.71	0.53	0.61	0.38
54 pathway assays	**0.69**	**0.45**	**0.71**	**0.53**	**0.68**	**0.36**
23 pathway assays	0.67	0.38	0.69	0.4	0.64	0.29
10 pathway assays	0.64	0.28	0.67	0.32	0.53	0.2

The Comparison between General Consensus and Pathway Consensus Results

RF MGR_Rat Liver all 284 models

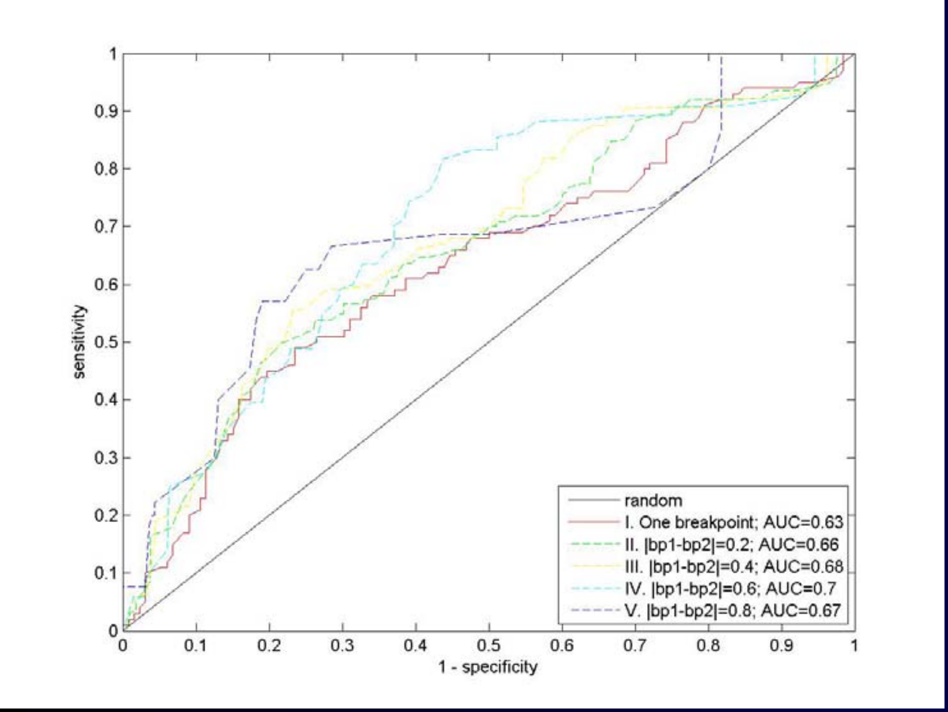


RF MGR_Rat Liver 54 pathway models

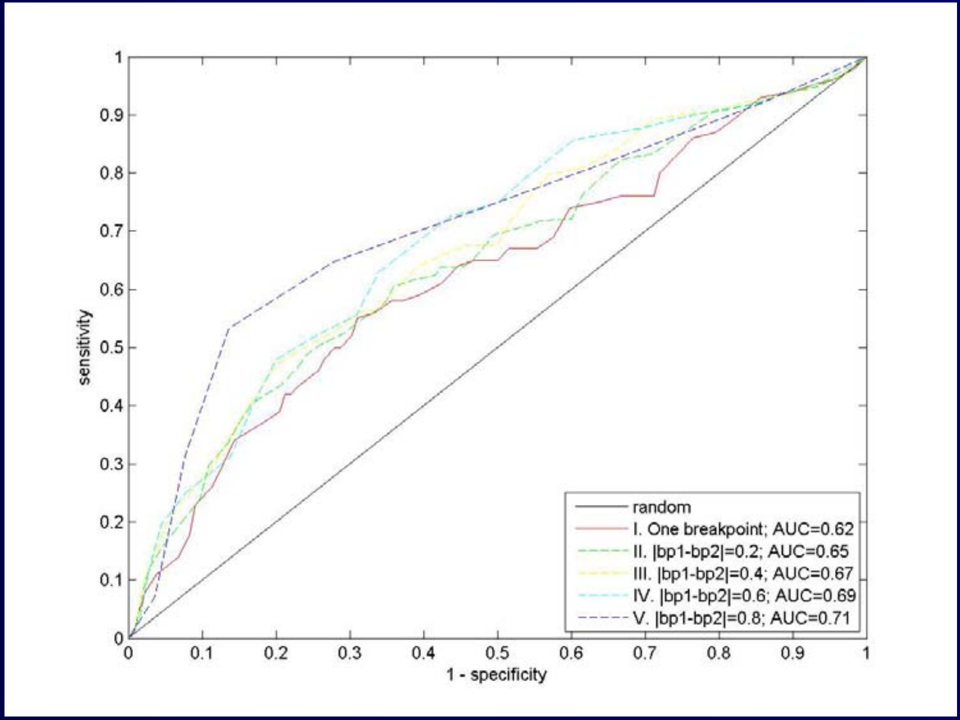


The Comparison between General Consensus and Pathway Consensus Results II

SVM_linear MGR_Rat Liver all 284 models

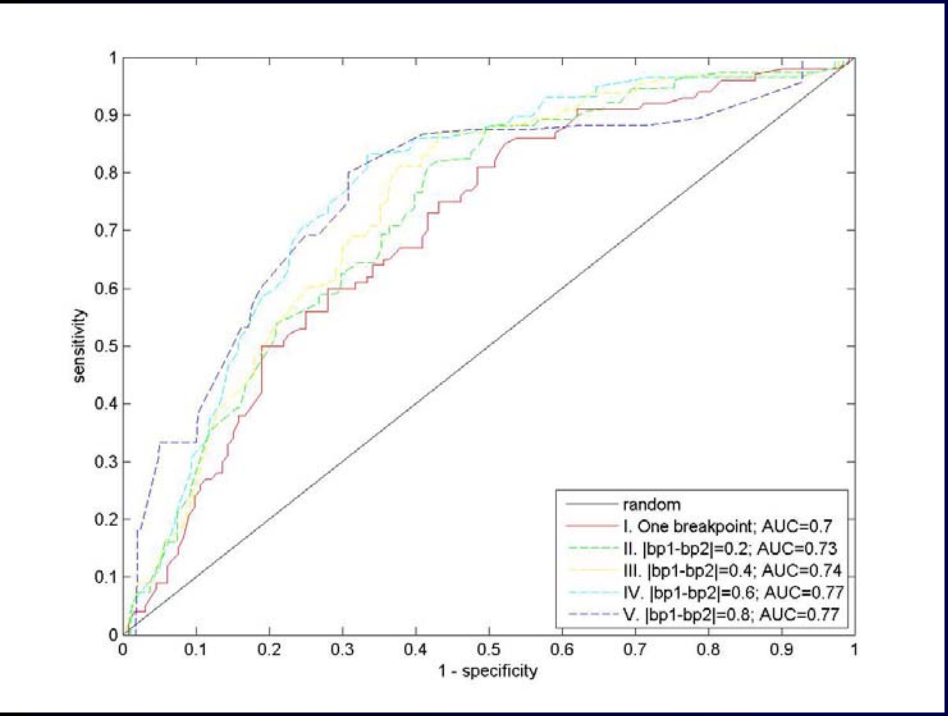


SVM_linear MGR_Rat Liver 54 pathway models

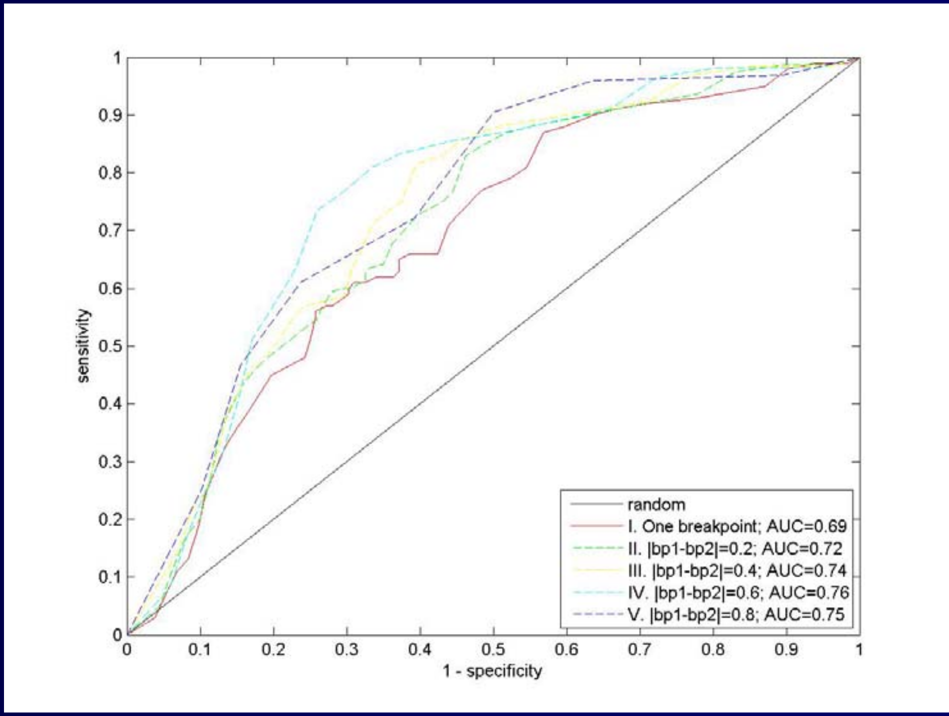


The Comparison between General Consensus and Pathway Consensus Results III

SVM_rbf MGR_Rat Liver all 284 models



SVM_rbf MGR_Rat Liver 54 pathway models



Using Individual Pathways for MGR_Rat Liver Modeling

	RF		**SVM_linear**		**SVM_rbf**	
	CCR	Coverage	CCR	Coverage	CCR	Coverage
All 54	0.73	0.56	0.66	0.42	0.71	0.53
PPAR (7)	**0.77**	**0.47**	0.54	0.3	0.61	0.35
Focal adhesion (5)	**0.7**	**0.57**	**0.7**	**0.46**	**0.72**	**0.52**
T cell (4)	0.69	0.7	0.61	0.4	0.64	0.49
TGF-beta (6)	**0.73**	**0.52**	0.57	0.54	**0.72**	**0.58**
Toll-like (20)	**0.71**	**0.61**	0.59	0.39	**0.69**	**0.43**
Apoptosis (9)	0.69	0.30	0.60	0.32	0.67	0.35
Wnt (2)	0.66	0.74	0.59	0.53	0.66	0.58
Adipocytokine (4)	**0.70**	**0.62**	0.60	0.4	0.58	0.47
Leukocyte (9)	**0.71**	**0.59**	0.59	0.45	**0.72**	**0.54**
MAPK (14)	**0.72**	**0.63**	0.61	0.41	**0.71**	**0.49**
ErbB (3)	**0.70**	**0.56**	**0.68**	**0.35**	0.66	0.41
Natural killer cell (4)	0.66	0.76	**0.65**	**0.62**	**0.69**	**0.64**
Cell cycle (3)	0.69	0.65	0.62	0.68	**0.69**	**0.75**
p53 (4)	0.64	0.66	0.62	0.52	0.66	0.59
BMPR2 (6)	**0.72**	**0.4**	**0.63**	**0.24**	**0.70**	**0.29**

Using Individual Pathways for MGR_Rat Kidney Modeling

	RF		**SVM_linear**		**SVM_rbf**	
	CCR	Coverage	CCR	Coverage	CCR	Coverage
All 54	0.67	0.48	0.64	0.4	0.69	0.45
PPAR (7)	**0.71**	**0.3**	0.51	0.5	0.6	0.29
Focal adhesion (5)	0.62	0.49	0.56	0.53	0.65	0.5
T cell (4)	0.61	0.63	**0.61**	**0.45**	0.58	0.47
TGF-beta (6)	0.61	0.49	0.58	0.53	0.64	0.59
Toll-like (20)	**0.64**	**0.51**	**0.64**	**0.33**	**0.66**	**0.38**
Apoptosis (9)	0.61	0.49	**0.62**	**0.29**	**0.66**	**0.30**
Wnt (2)	0.57	0.75	0.56	0.62	0.58	0.65
Adipocytokine (4)	0.63	0.51	0.57	0.44	0.60	0.45
Leukocyte (9)	**0.65**	**0.53**	**0.61**	**0.41**	**0.67**	**0.48**
MAPK (14)	0.61	0.62	**0.65**	**0.44**	0.64	0.49
ErbB (3)	0.61	0.59	0.59	0.49	0.59	0.52
Natural killer cell (4)	0.61	0.71	**0.61**	**0.56**	0.63	0.6
Cell cycle (3)	0.56	0.72	0.56	0.69	0.62	0.77
p53 (4)	0.60	0.63	0.53	0.5	0.61	0.53
BMPR2 (6)	**0.70**	**0.31**	0.57	0.2	0.63	0.28

Using Individual Pathways for MGR_Rat ViabilityPND4 Modeling

	RF		**SVM_linear**		**SVM_rbf**	
	CCR	Coverage	CCR	Coverage	CCR	Coverage
All 54	0.63	0.4	0.64	0.32	0.61	0.38
PPAR (7)	**0.63**	**0.3**	0.52	0.25	0.5	0.26
Focal adhesion (5)	0.56	0.43	0.59	0.39	0.53	0.41
T cell (4)	0.51	0.62	0.54	0.43	0.54	0.46
TGF-beta (6)	0.53	0.4	**0.61**	**0.39**	**0.60**	**0.5**
Toll-like (20)	0.58	0.47	**0.61**	**0.27**	**0.65**	**0.30**
Apoptosis (9)	**0.60**	**0.46**	**0.63**	**0.23**	**0.60**	**0.27**
Wnt (2)	0.53	0.76	0.48	0.60	0.52	0.63
Adipocytokine (4)	0.51	0.52	0.56	0.39	0.5	0.37
Leukocyte (9)	**0.63**	**0.46**	0.58	0.37	0.55	0.44
MAPK (14)	0.53	0.57	**0.61**	**0.35**	0.56	0.63
ErbB (3)	0.54	0.56	0.56	0.41	0.51	0.46
Natural killer cell (4)	0.54	0.72	**0.61**	**0.54**	**0.63**	**0.57**
Cell cycle (3)	0.48	0.66	0.52	0.65	0.53	0.77
p53 (4)	0.56	0.63	0.54	0.61	0.57	0.54
BMPR2 (6)	0.57	0.3	0.55	0.22	**0.59**	**0.26**

Future Studies

- Go deeply to each pathway assay model
  - Significant tests
  - Any chemical scaffold could be identified?
- Combine different pathway models to achieve predictions with higher accuracy
- Model analysis to identify significant assay-chemical combinations that are predictive of *in vivo* outcomes
- Apply model prospectively to prioritise new compounds for ToxCast testing.

Conclusions and plans

- Focus on accurate prediction of <u>external</u> datasets is much more critical than accurate fitting of existing data: validate, <u>then</u> interpret!
  - validation!!!
  - applicability domain
  - consensus prediction using all acceptable models
  - Ideally, experimental validation of a <u>small</u> number of computational hits
  - Outcome: decision support tools in selecting future experimental screening sets
- HTS and –omics data may be insufficient to achieve the desired accuracy of the end point property prediction BUT should be explored as biodescriptors in combination with chemical descriptors
  - New computational approaches (e.g., hierarchical QSAR)
  - Understanding of <u>both</u> chemistry and biology

Acknowledgements

Principal Investigator
Alexander Tropsha

Research Professors
Clark Jeffries, Alexander Golbraikh, Hao Zhu, Simon Wang

Postdoctoral Fellows
Georgiy Abramochkin, Lin Ye, Denis Fourches

Visiting Research Scientist
Aleks Sedykh

Adjunct Members
Weifan Zheng, Shubin Liu

MML
UNC.EDU

MAJOR FUNDING
NIH
- P20-HG003898
- R21GM076059
- R01-GM66940
- R0-GM068665

EPA (STAR awards)
- RD832720
- RD833825

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

Graduate Research Assistants
Christopher Grulke, Nancy Baker, Kun Wang, **Hao Tang**, Jui-Hua Hsieh, Rima Hajjo, Tanarat Kietsakorn, Tong Ying Wu, **Liying Zhang**, Melody Luo, Guiyu Zhao, Andrew Fant

Research Programmer
Theo Walker

System Administrator
Mihir Shah