



UNC
ESHELMAN
SCHOOL OF PHARMACY



MML
UNC.EDU

Prediction of animal toxicity endpoints of
ToxCast Phase I compounds using a
combination of chemical and biological
in vitro descriptors

Alexander Tropsha

Carolina Center for Computational Toxicology
and

Carolina Center for Environmental Bioinformatics

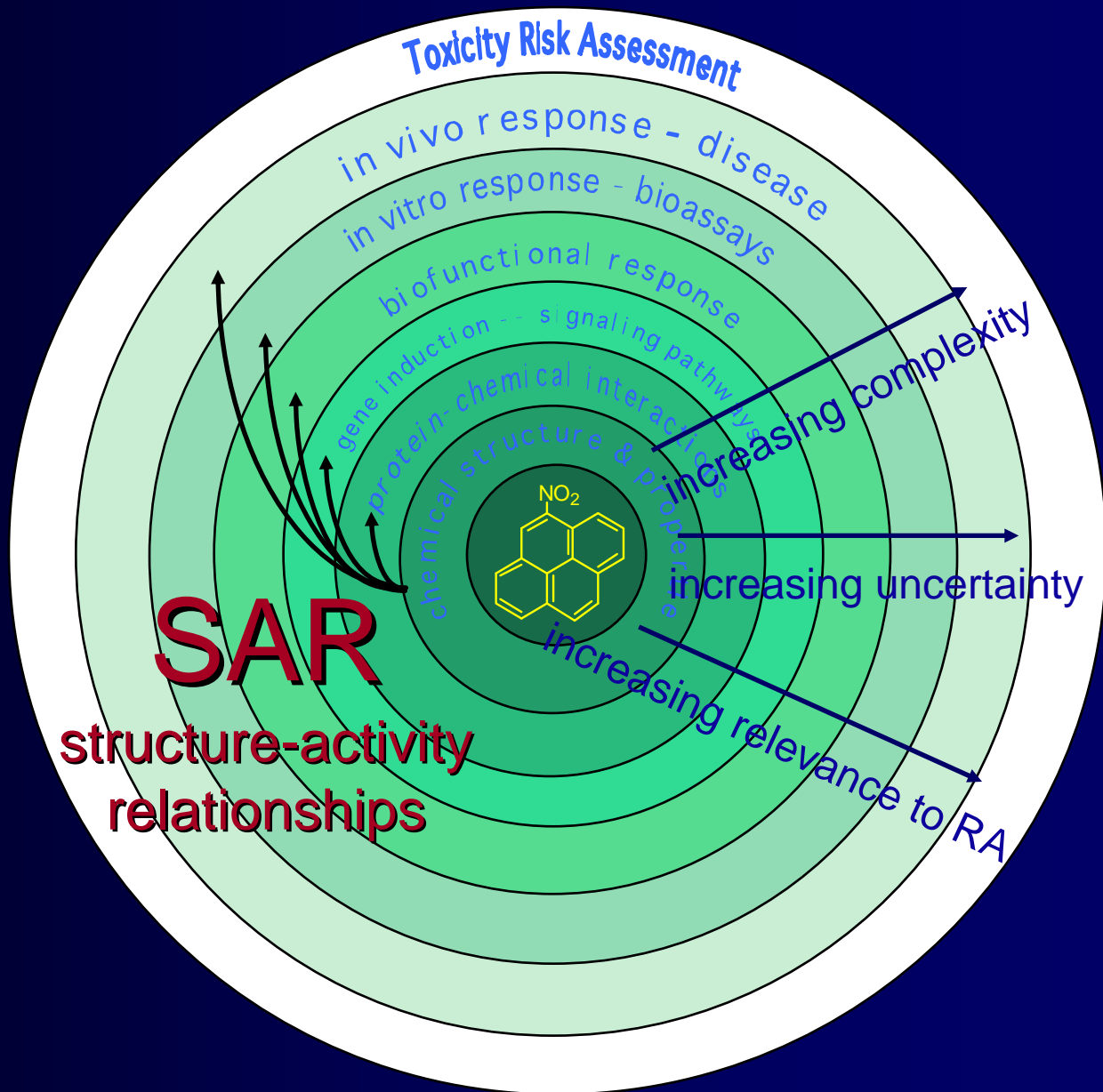
Laboratory for Molecular Modeling

School of Pharmacy, UNC-Chapel Hill

Outline

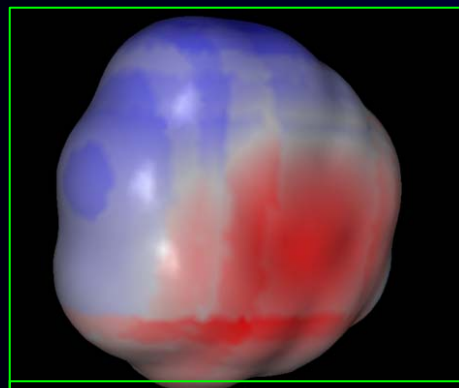
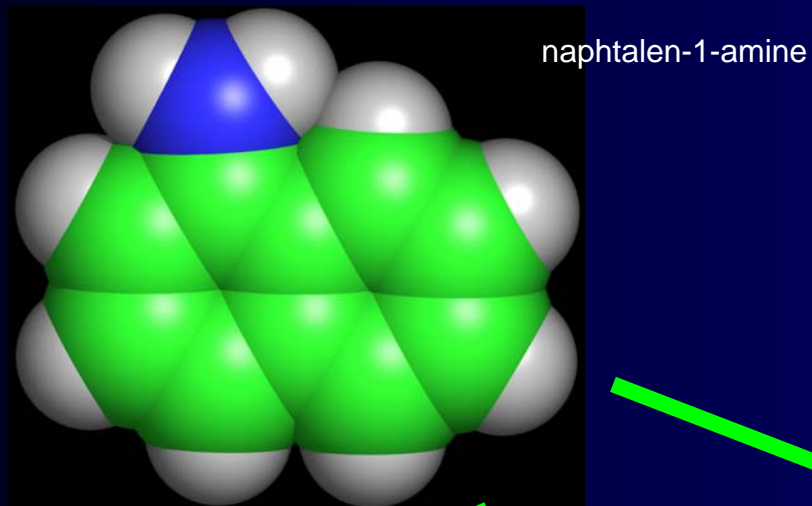
- Overall project vision: exploiting the entire structure – *in vitro* – *in vivo* continuum
- (briefly) Predictive QSAR Modeling Workflow
- Applications
 - novel data partitioning approach based on *in vitro* – *in vivo* correlations: Hierarchical QSAR modeling of rodent toxicity
 - preliminary analysis 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*

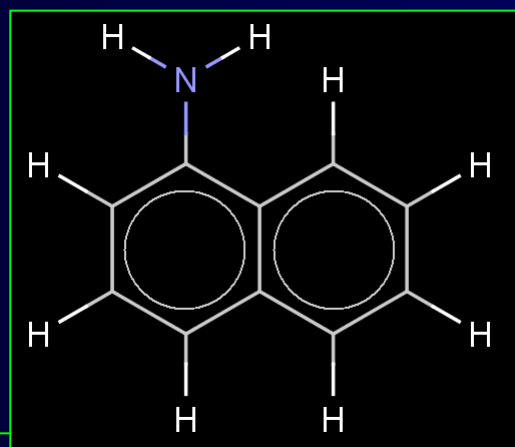


Slide courtesy of Dr. Ann Richard (EPA)

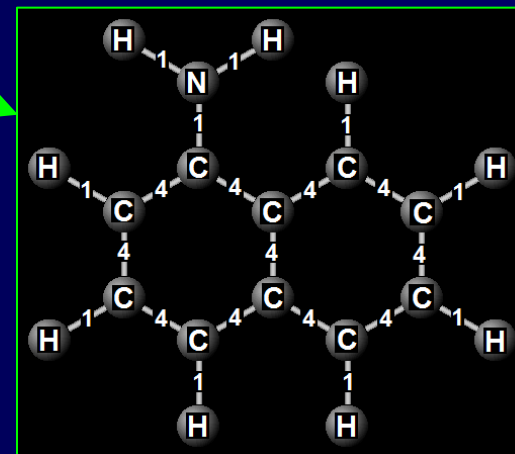
STRUCTURE REPRESENTATION



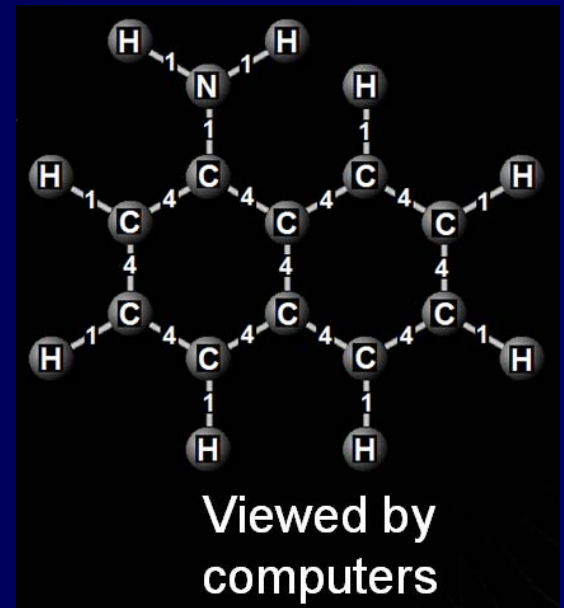
Viewed by another molecule

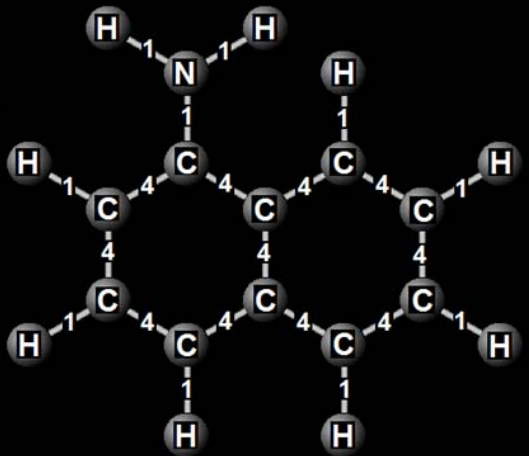


Viewed by chemists



Viewed by computers





Viewed by
computers

Graphs are widely used to represent and differentiate chemical structures, where **atoms are vertices** and **bonds are expressed as edges**

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

Molecular descriptors

MOL File

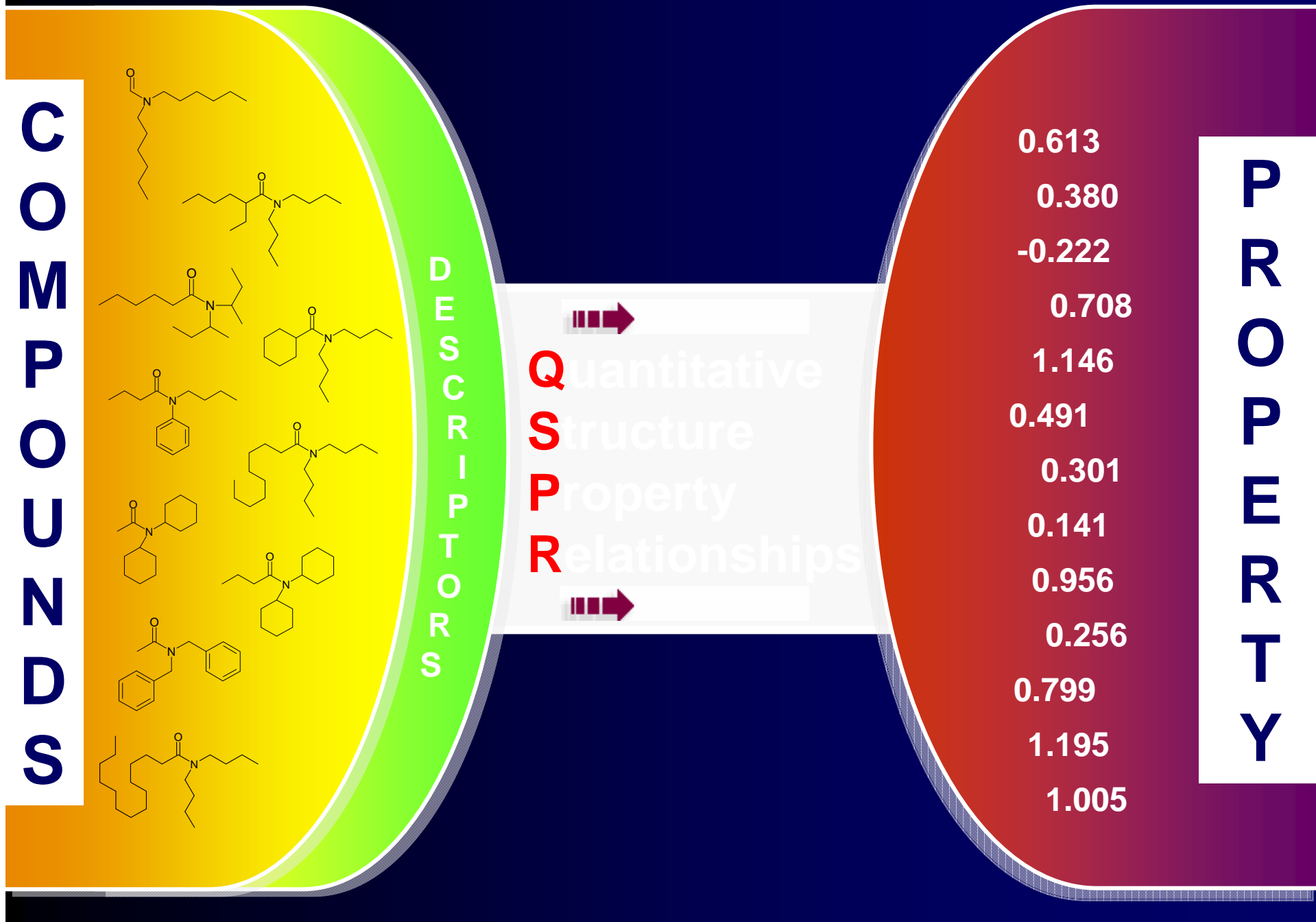
```
-ISIS- 11240515102D
16 15 0 0 0 0 0 0 0 0999 V2000
-0.1958 -2.9667 0.0000 C 0 0 0
0.5167 -2.5500 0.0000 C 0 0 0
0.5125 -1.7250 0.0000 O 0 0 0
1.2292 -2.9625 0.0000 N 0 0 3
1.9417 -2.5458 0.0000 C 0 0 0
2.6542 -2.9583 0.0000 C 0 0 0
3.3667 -2.5417 0.0000 C 0 0 0
4.0792 -2.9542 0.0000 C 0 0 0
4.7917 -2.5375 0.0000 C 0 0 0
5.5042 -2.9500 0.0000 C 0 0 0
1.2250 -3.7875 0.0000 C 0 0 0
0.8083 -4.5000 0.0000 C 0 0 0
1.3917 -5.0833 0.0000 C 0 0 0
0.9750 -5.7958 0.0000 C 0 0 0
1.5583 -6.3792 0.0000 C 0 0 0
0.9708 -6.9625 0.0000 C 0 0 0
8 9 1 0 0 0 0
4 5 1 0 0 0 0
9 10 1 0 0 0 0
2 3 2 0 0 0 0
4 11 1 0 0 0 0
5 6 1 0 0 0 0
11 12 1 0 0 0 0
1 2 1 0 0 0 0
12 13 1 0 0 0 0
6 7 1 0 0 0 0
13 14 1 0 0 0 0
2 4 1 0 0 0 0
14 15 1 0 0 0 0
7 8 1 0 0 0 0
15 16 1 0 0 0 0
M END
```

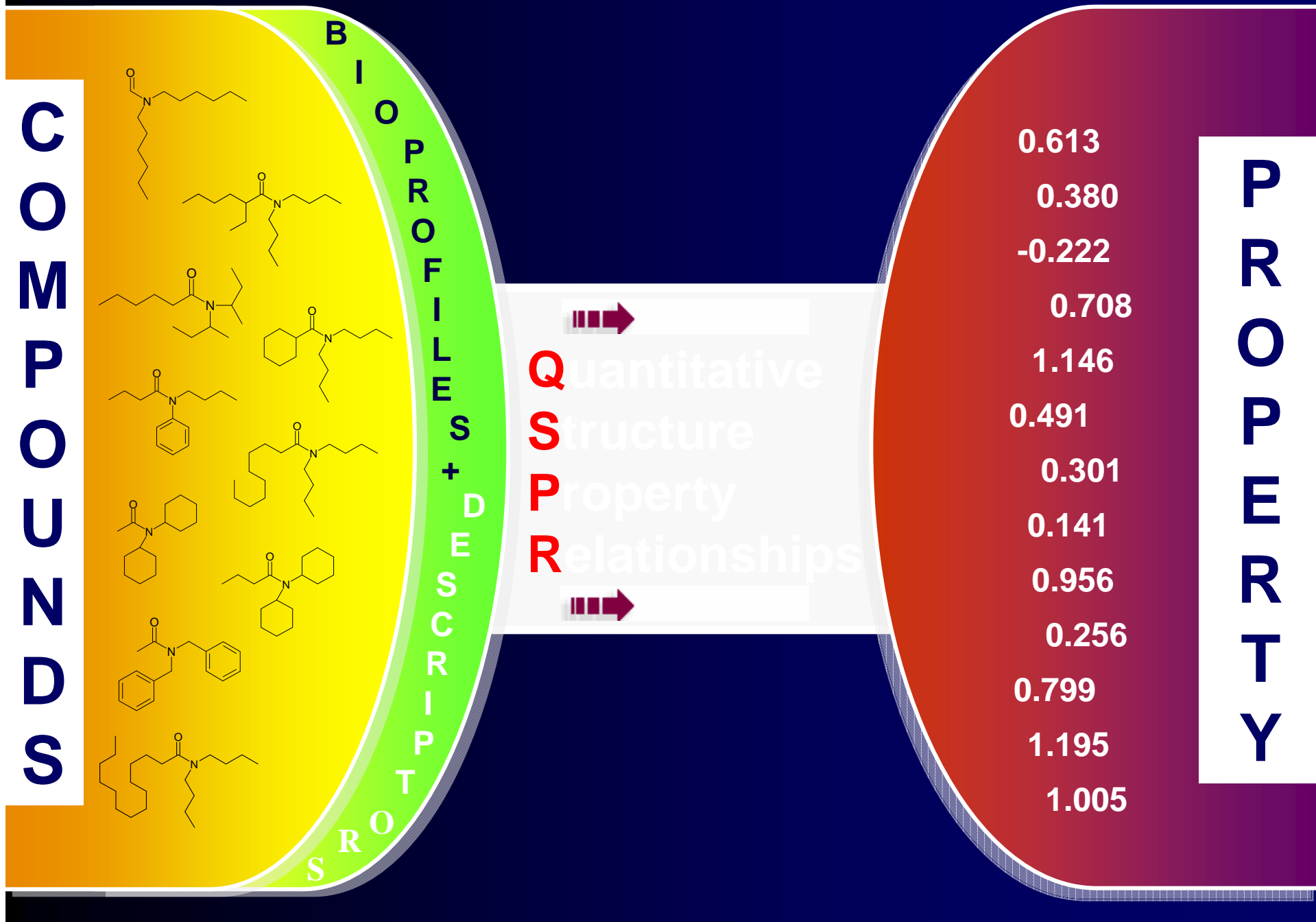
Vertices

(atomic type,
coordinates etc.)

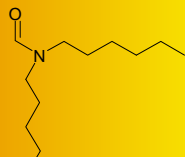
Edges

(connectivity table,
label-types of bonds)



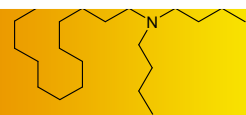
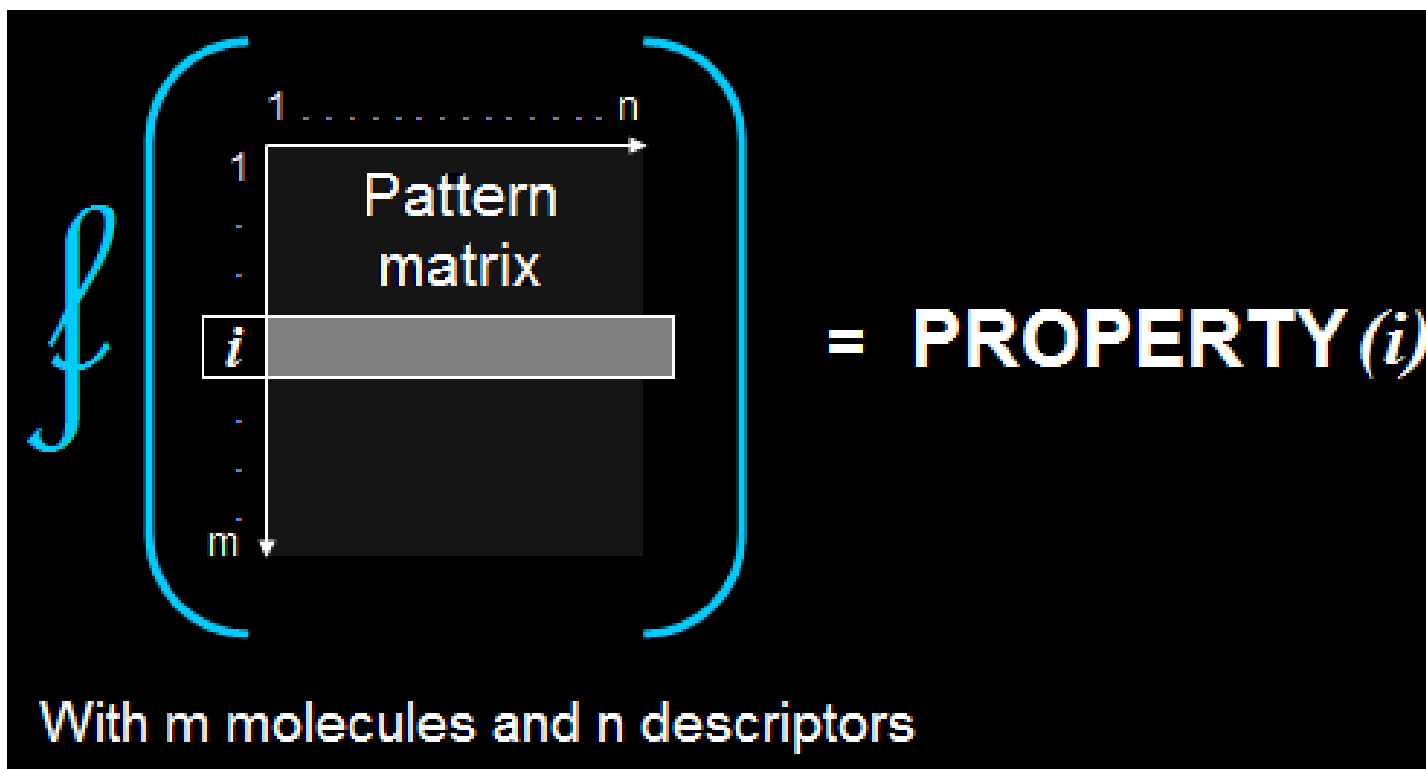


C
O
M
P
O
U
N
D
S



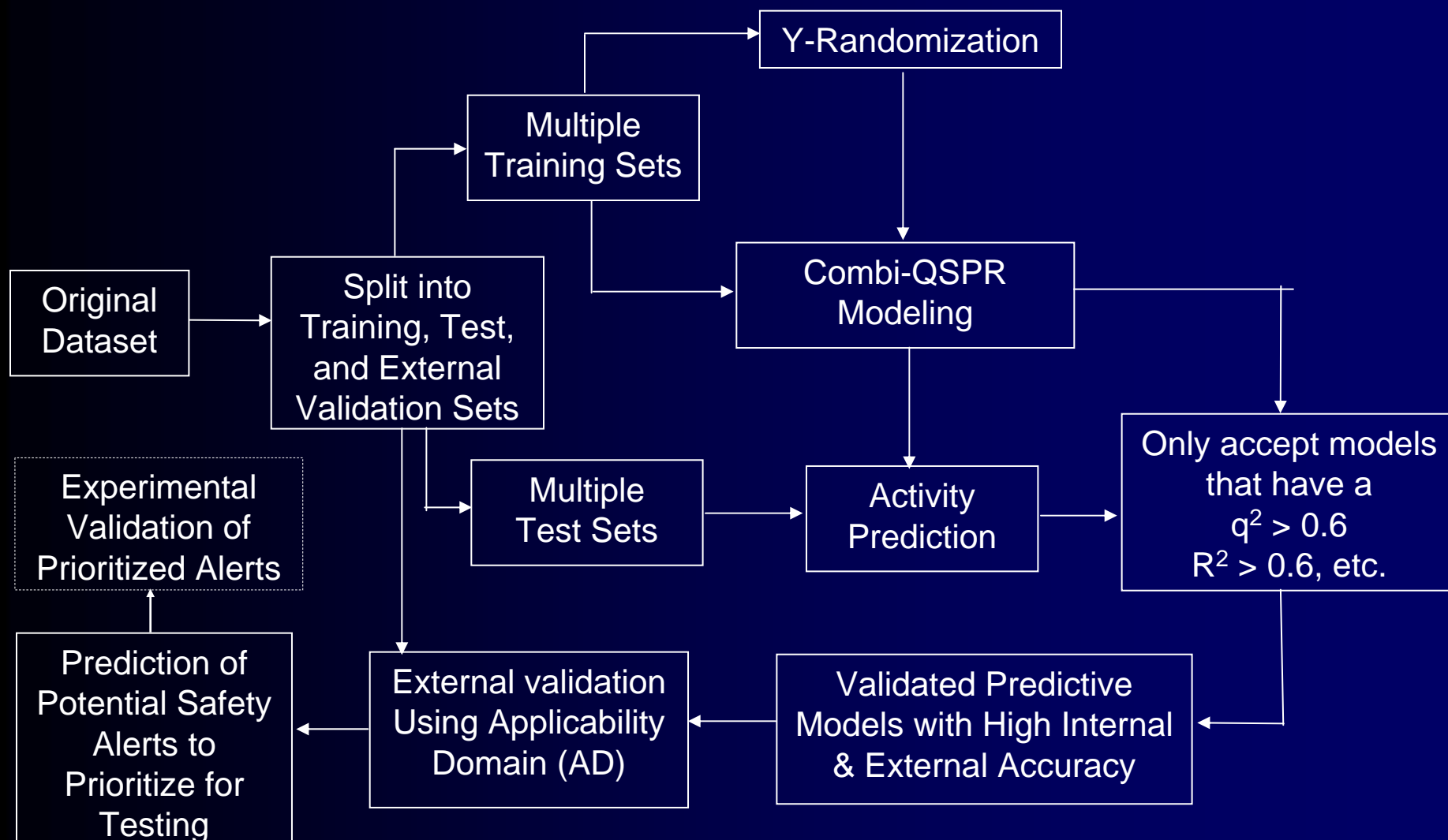
0.613

P
R
O
P
E
R
T
Y



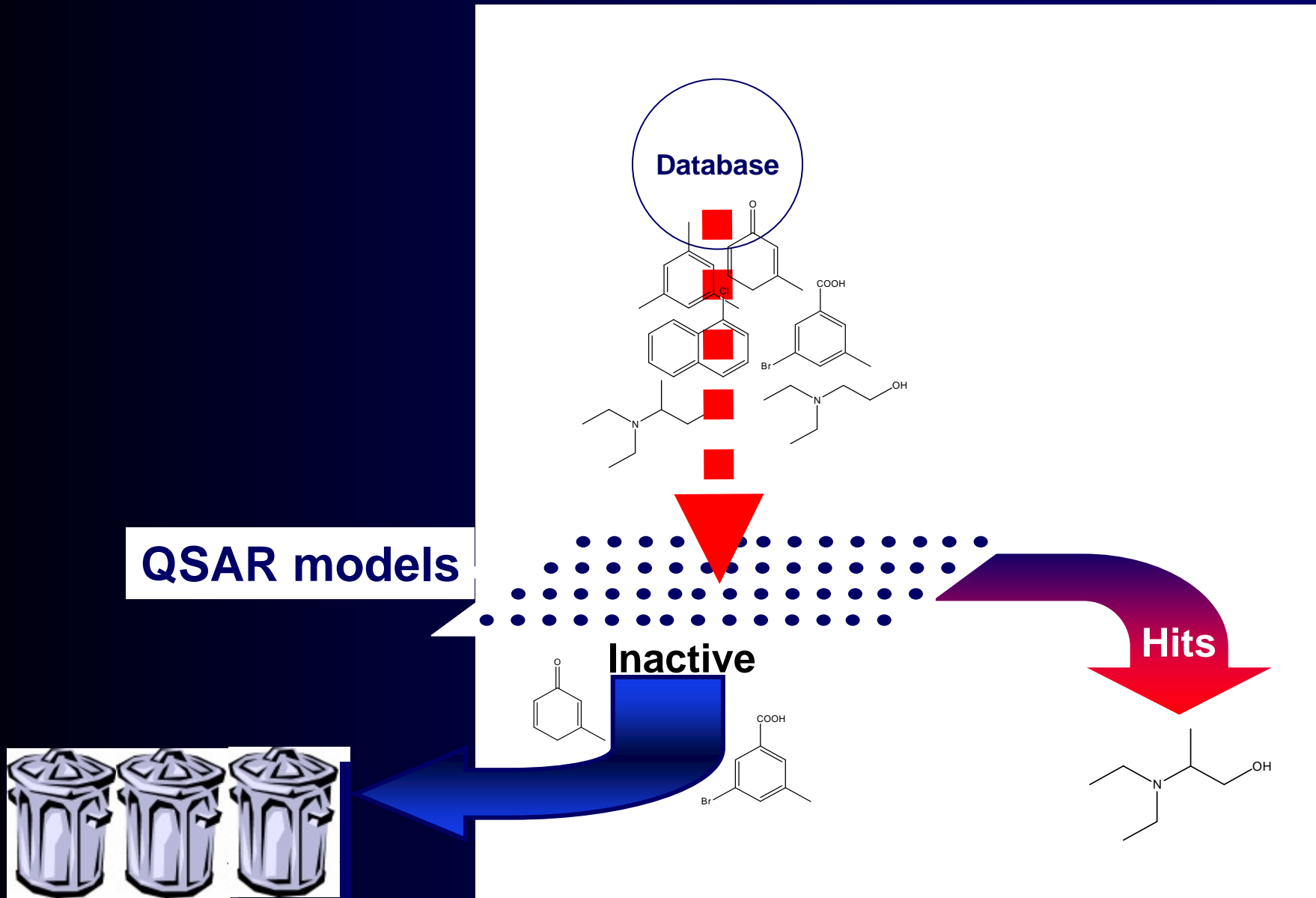
1.005

Predictive QSAR Workflow*



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

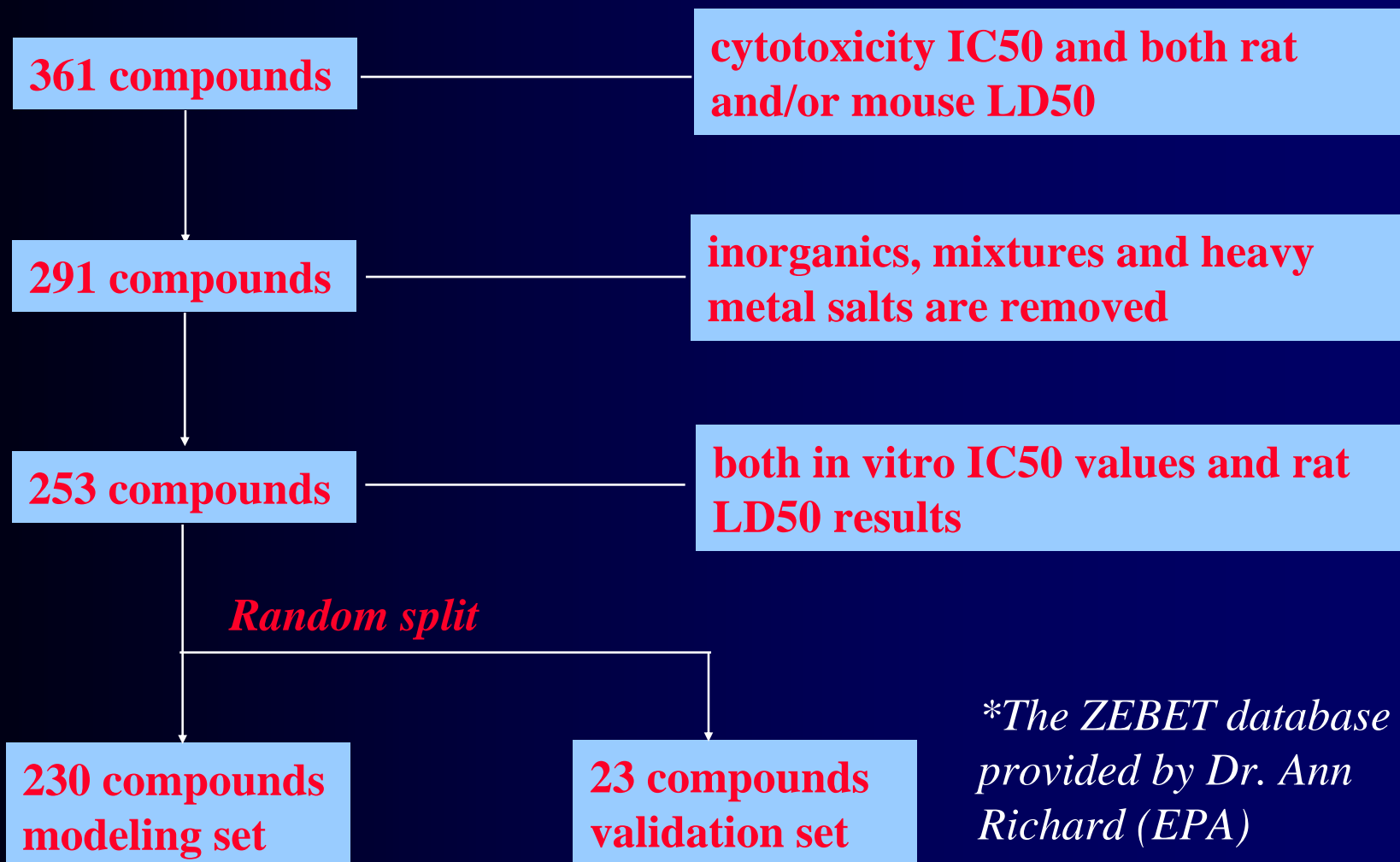
Compound prioritization using QSAR models



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

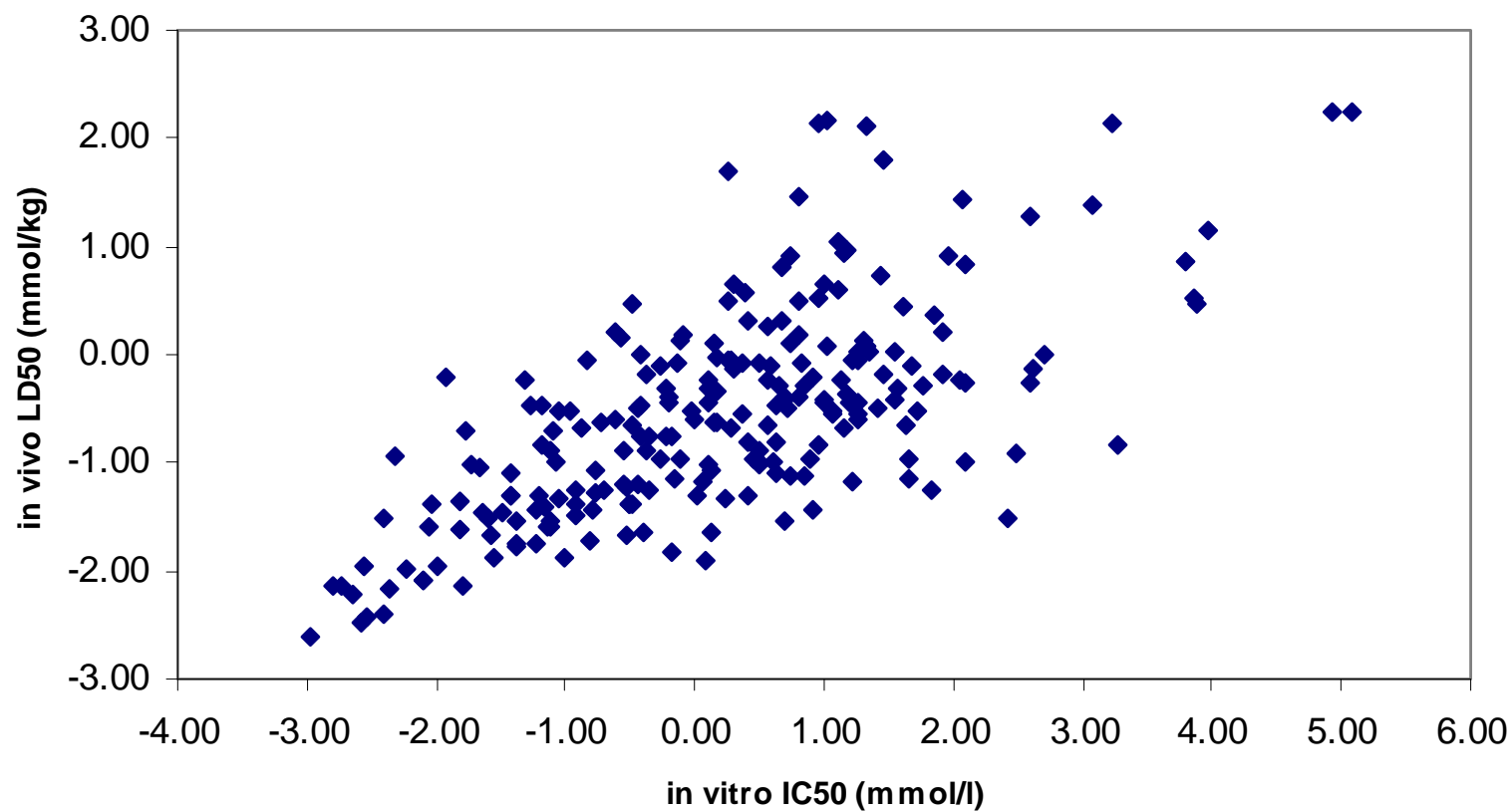
*Zhu, Rusyn, Wright, et al, EHP, 2009, in press; in collaboration with
Ann Richard, NCCT, US EPA

ZEBET Database* and Data Preparation



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

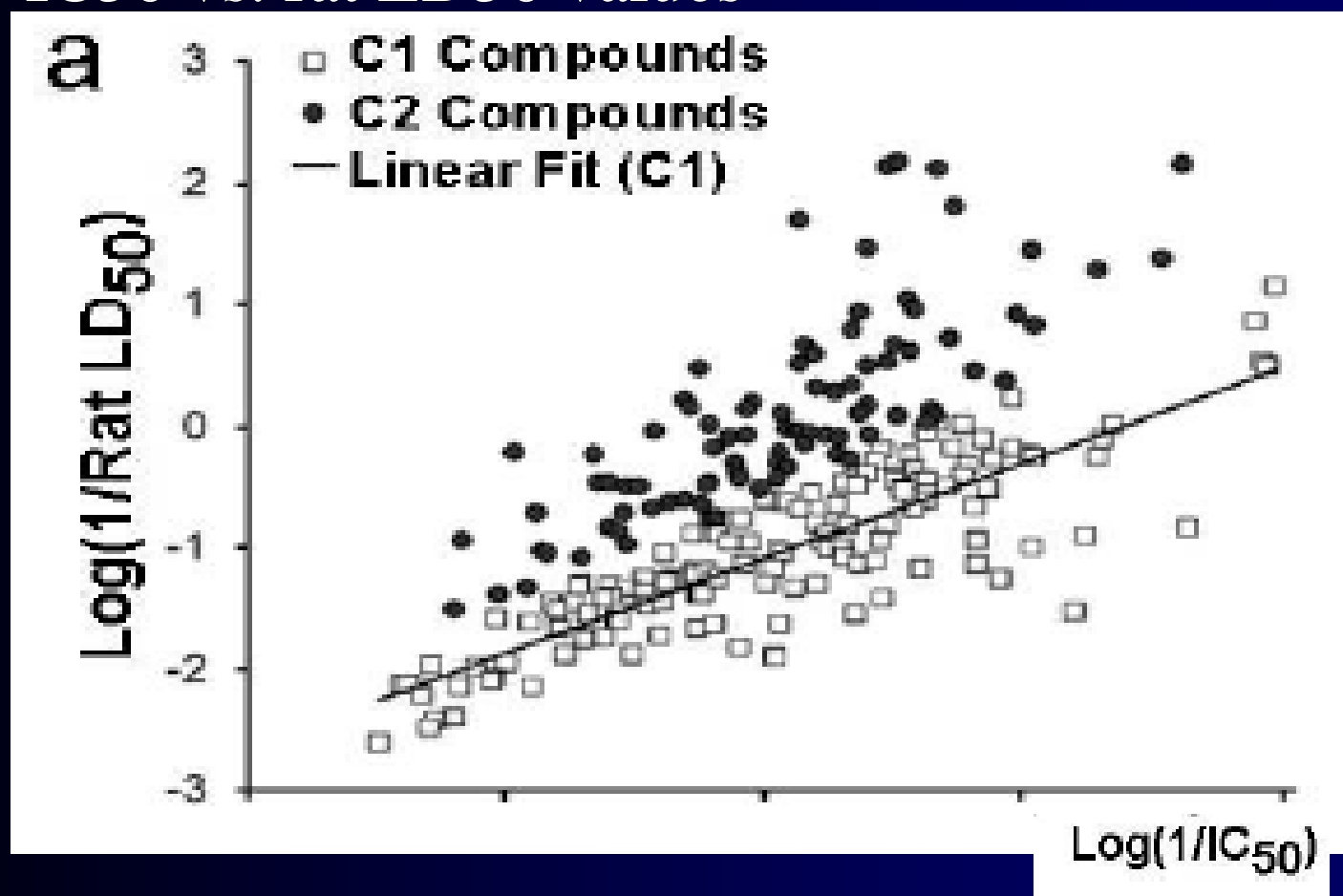
(Rather) poor in vitro-in vivo
Correlation Between IC50 and Rat
LD50 Values



$$R^2=0.46$$

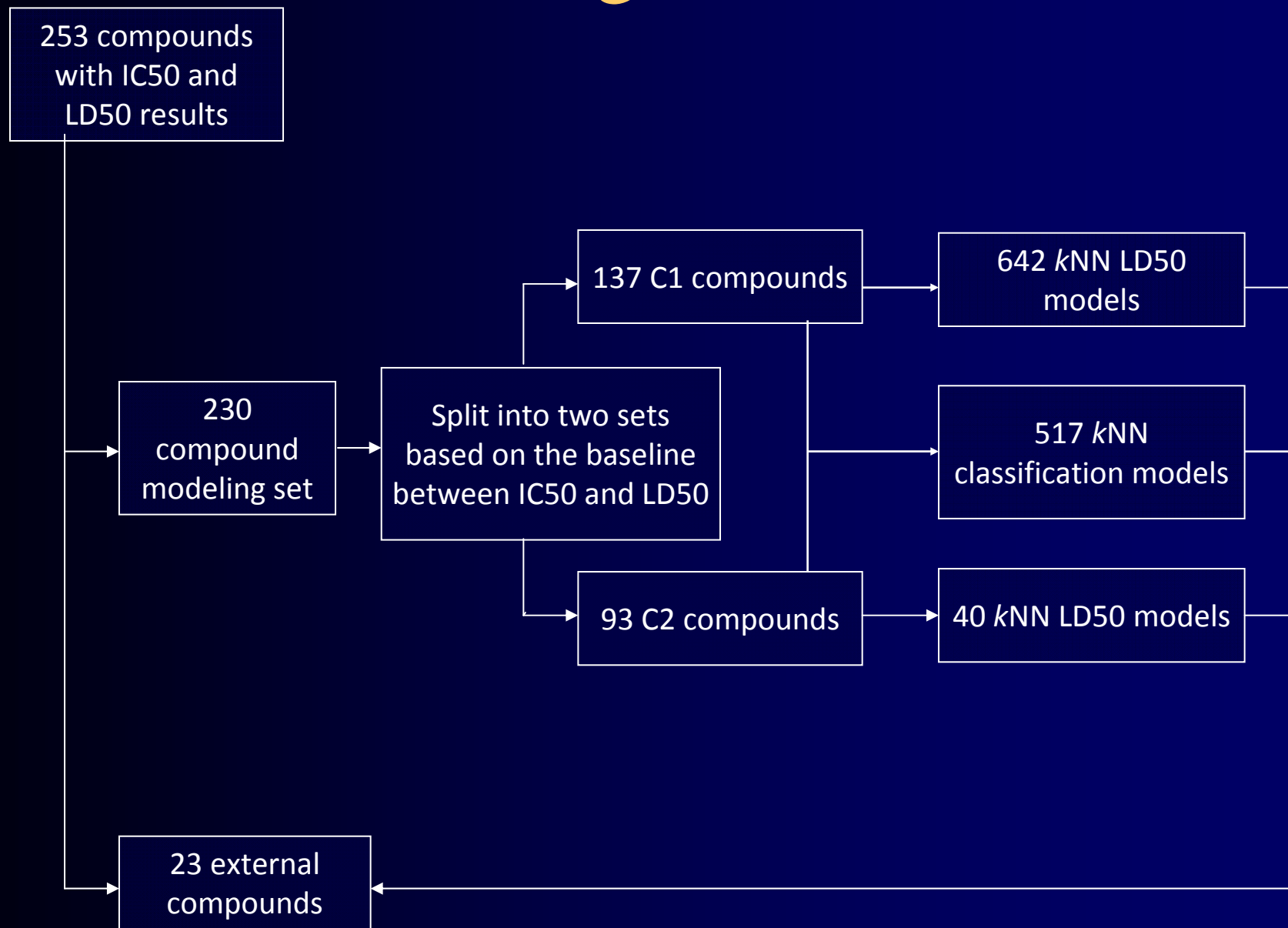
Data partitioning based on the moving regression approach

- IC50 vs. rat LD50 values

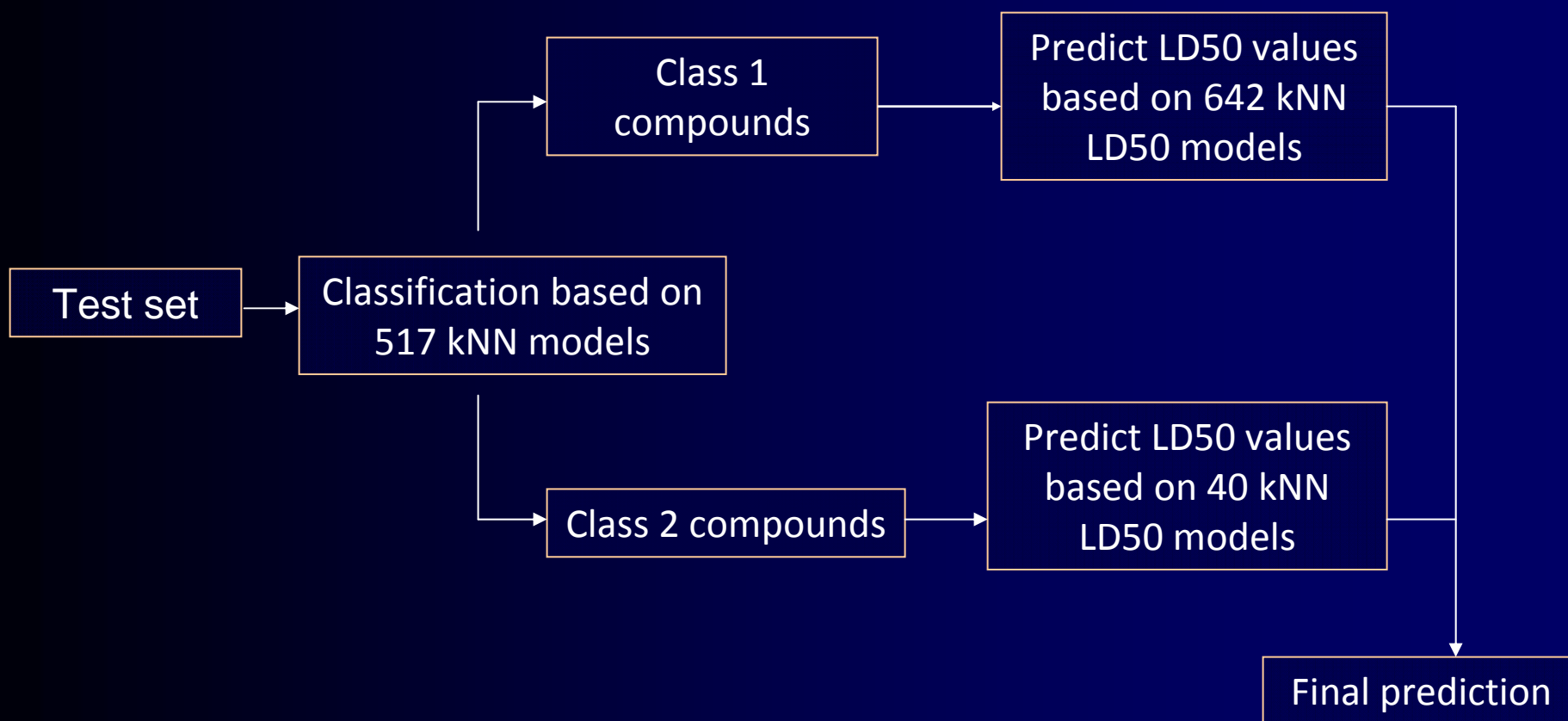


$R^2=0.74$ for Class 1 compounds

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

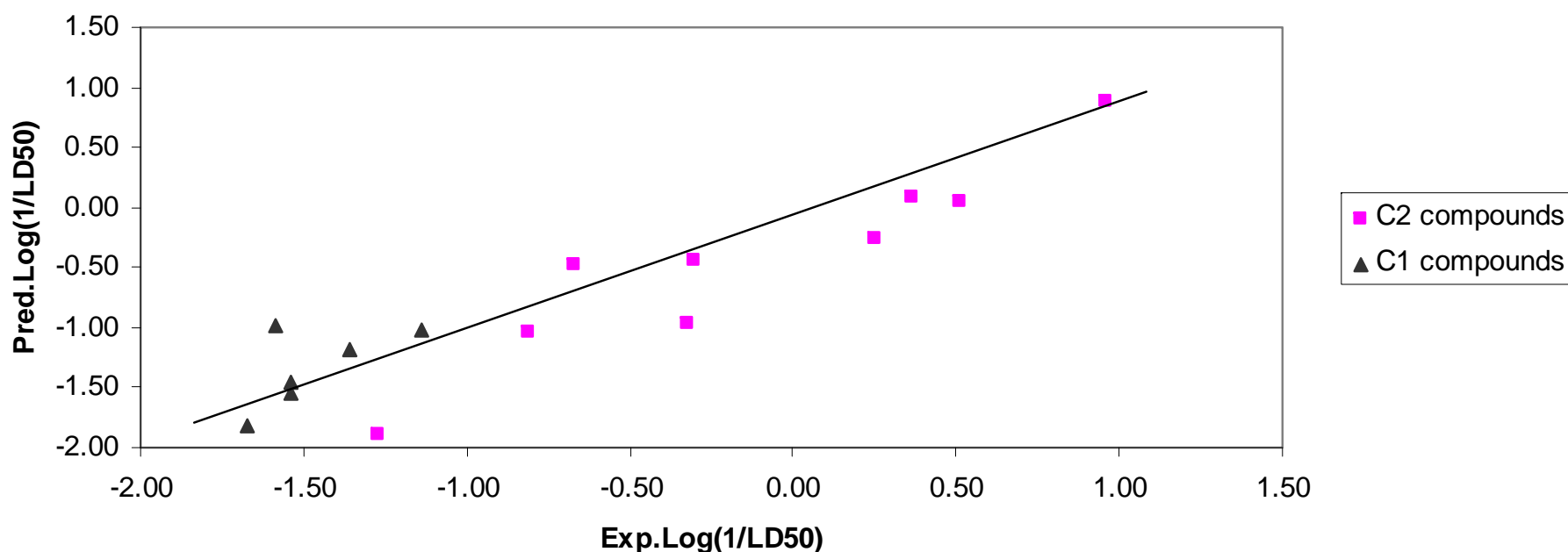
With AD:

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:
Preliminary analysis of
ToxCAST data:**

- (i) ToxRefDB***
- (ii) using hierarchical QSAR
modeling approach**

Endpoints Summary in ToxRefDB (26 endpoints)

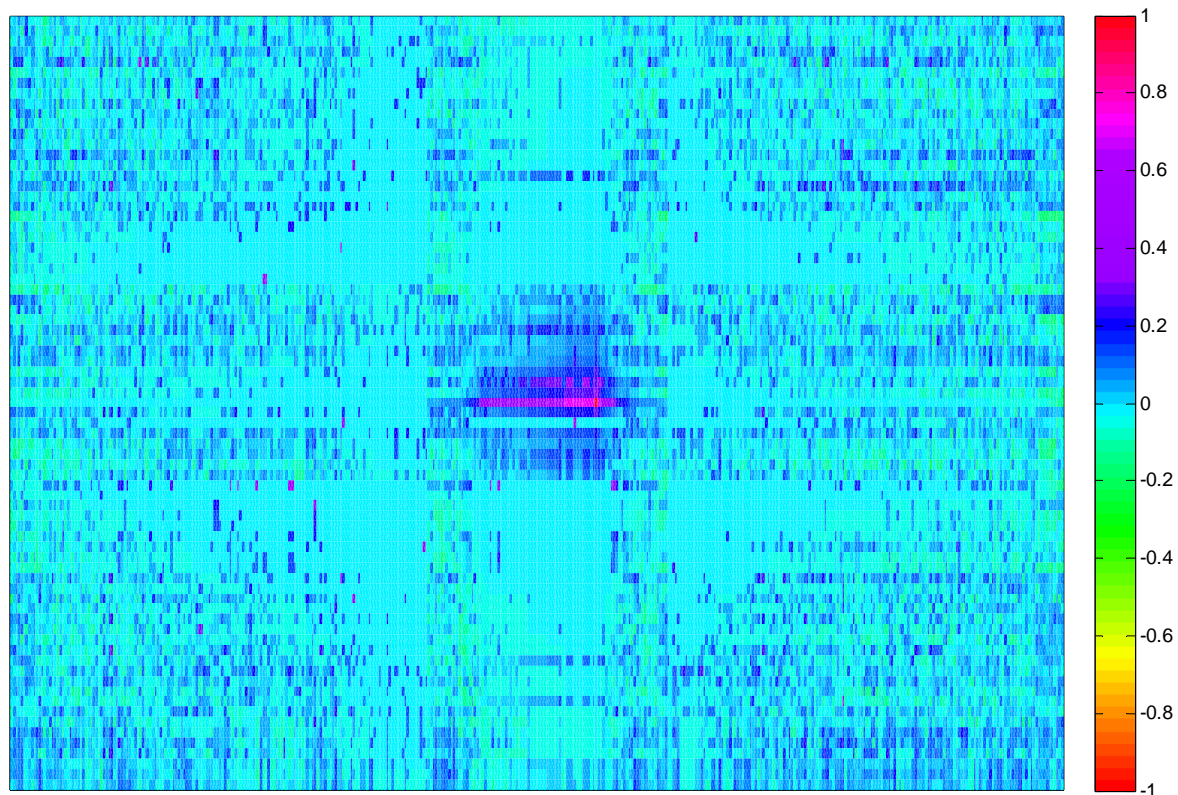
Rat Liver Tumors	----->	All neoplastic liver lesions
Rat Proliferative Liver Lesions	----->	All neoplastic and non-neoplastic proliferative liver lesions
Rat Liver Apoptosis/Necrosis	----->	All effect related to apoptosis and necrosis
Rat Liver Hypertrophy	----->	Liver hypertrophy (individual effect)
Rat Kidney Nephropathy	----->	Kidney nephropathy (individual effect)
Rat Proliferative Kidney Lesions	----->	All neoplastic and non-neoplastic proliferative kidney lesions
Rat Proliferative Thyroid Lesions	----->	All neoplastic and non-neoplastic proliferative thyroid lesions
Rat Thyroid Tumors	----->	All neoplastic thyroid lesions
Rat Thyroid Hyperplasia	----->	Thyroid hyperplasia (individual effect)
Rat Testicular Tumors	----->	All neoplastic testis tumors
Rat Testicular Atrophy	----->	Testis atrophy (individual effect)
Rat Spleen Pathology	----->	All non-neoplastic and neoplastic spleen pathology
Rat Cholinesterase Inhibition	----->	Any cholinesterase inhibition measurement (e.g., brain and erythrocyte)
Rat Tumorigen	----->	All neoplastic lesions (any target)
Rat Multigender Tumorigen	----->	All neoplastic lesions (occurring in male and female)
Rat Multisite Tumorigen	----->	All neoplastic lesions (occurring in multiple target organs)
Mouse Proliferative Liver Lesions	----->	All neoplastic and non-neoplastic proliferative liver lesions
Mouse Liver Apoptosis/Necrosis	----->	All effect related to apoptosis and necrosis
Mouse Liver Hypertrophy	----->	Liver hypertrophy (individual effect)
Mouse Kidney Pathology	----->	All non-neoplastic and neoplastic kidney pathology
Mouse Lung Tumors	----->	All neoplastic lung lesions
Mouse Tumorigen	----->	All neoplastic lesions (any target)
Mouse Multigender Tumorigen	----->	All neoplastic lesions (occurring in male and female)
Mouse Multisite Tumorigen	----->	All neoplastic lesions (occurring in multiple target organs)
Multispecies Tumorigen	----->	All neoplastic lesions (occurring in both rat and mouse)

Initial Results of Global QSAR Modeling for Toxicity Endpoints (5-fold external cross-validation)

Category Name	Endpoints Involved#	Non-toxic Comp #	Toxic Comp #	Descriptor#	Model #	Aver. CCR	Aver. SE	Aver. SP
total	26	52	240	705	73	0.58	0.97	0.19
rat	16	86	206	705	80	0.51	0.84	0.18
mouse	9	143	149	705	408	0.6	0.59	0.61
tumor	7	132	160	705	97	0.51	0.56	0.45
rat liver	4	184	108	705	168	0.58	0.38	0.77
mouse liver	4	170	122	705	116	0.63	0.52	0.73
Cell viability	7	123	98	737	104	0.81	0.84	0.81

Relationships Between *in vivo* and *in vitro* Assays in ToxCast™ (based on Matthew's Correlation Coefficient, MCC*)

75 (49+26) *in vivo* and 409 *in vitro* endpoints



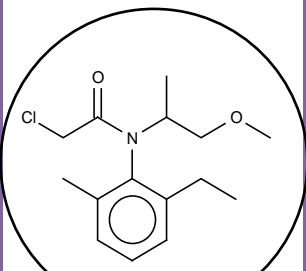
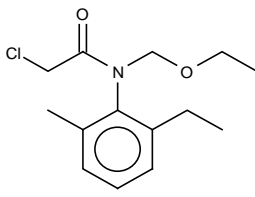
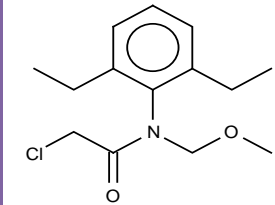
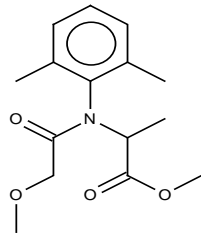
$$*MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

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

Compounds	Total	ACEA	ATG	BSK	Cellumen	NVS	CellzDirect
	500	7	81	87	33	239	48
3-Iodo-2-propynylbutylcarbamate	0.71	0.73	0.18	0.53	0.49	0.89	0.15
Bensulide	0.64	0.09	0.71	0.4	0.69	0.95	0.04
Chlorsulfuron	0.24	N/A	N/A	0.4	N/A	N/A	-0.1
Dibutyl phthalate	0.55	N/A	0.62	0.51	0.7	0.81	-0.1
Diclofop-methyl	0.36	1	0.89	0.15	N/A	-0	-0.1
EPTC	0.13	N/A	N/A	-0.1	N/A	N/A	0.33
Fenoxaprop-ethyl	0.47	N/A	0.56	0.59	0.31	0.35	0.01
Prosulfuron	0.55	N/A	0.68	0.08	N/A	1	0.4

$$*MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Hard cases: chemically similar compounds have different profiles

		IN VITRO ASSAYS						IN VIVO ASSAYS						
		ACEA_IC50	ATG_EGR_CIS	CELLLOSS_24HR	P53ACT_24HR	NVS_ADME_RCYP3A2	NVS_GPCR	SOLIDUS_P450	MOUSE_KIDNEY	RAT_SKELETAL_AXIAL	MGR_RAT_LIVER			MGR_RAT_KIDNEY
METOLACHLOR		0	1	0	0	1	1	0	0	0	0	0		NN1-ACETOCHLOR
N2-ALACHLOR		1	0	1	1	0	0	1	-	0	-	-		NN3-METALAXYL
		0	0	0	0	0	0	0	0	1	1	0		

Richard's Law # 4: it is not rare to find pairs of similar or identical compounds that violate all other Richard's laws

Focusing on a small subset of 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)

Data Curation

- *In-vitro* assays: 524 → 353
 - Remove one of two highly correlated ($R^2 > 0.95$) assays and low-variance (<4 non-zero entries) assays
- Chemicals: 320 → 228
 - duplicate structures, mixtures, inorganic compounds, macromolecules were removed
 - Kept only those for which *in-vivo* data is available (i.e. chronic mouse toxicity)

Conventional QSAR Modeling

- Using chemical descriptors only:
 - 1224 Dragon chemical descriptors
 - Random Forest (RF) approach

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

Endpoints	Sensitivity	Specificity	CCR _{ext}
Liver Proliferative Lesions	0.34	0.72	0.53
Liver Tumors	0.21	0.90	0.56
Tumorigen	0.44	0.57	0.51

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

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

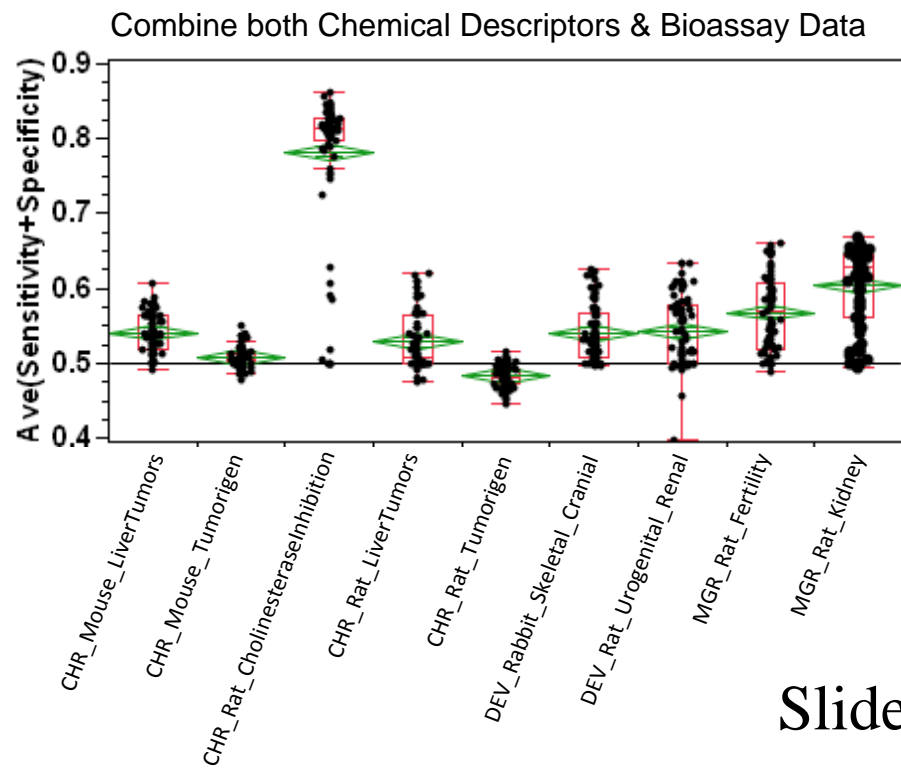
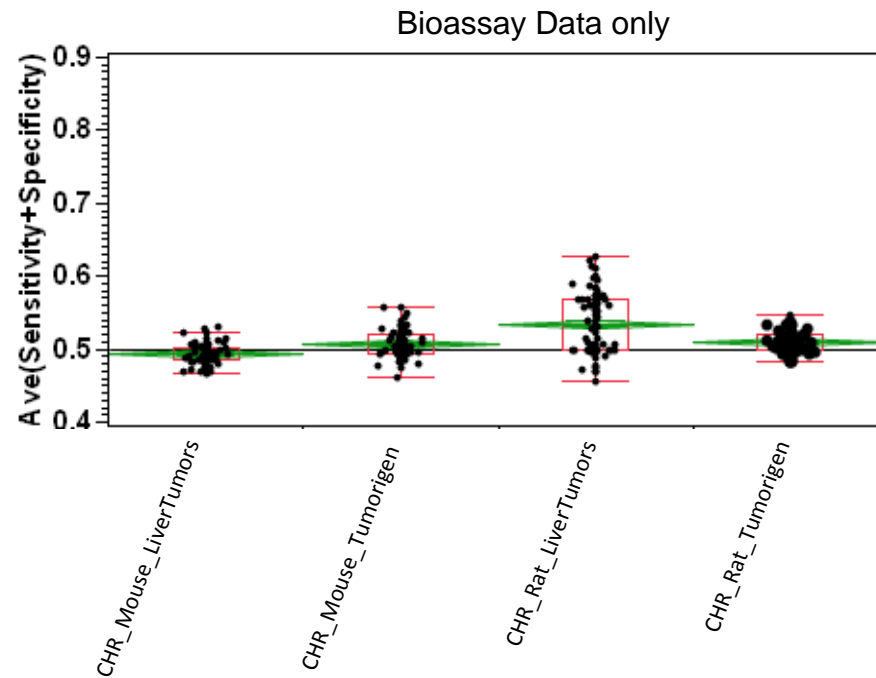
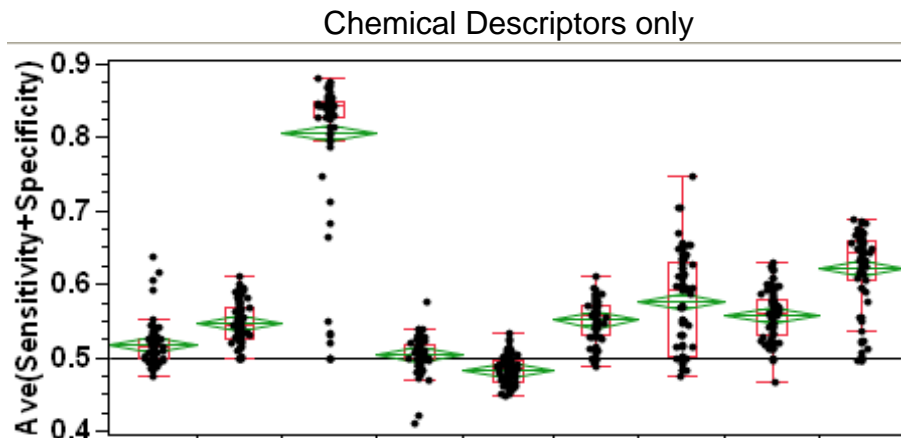
- Using ToxCAST bioassay results as biological descriptors did not result in any statistical significant models.
- The use of hybrid (biological + chemical descriptors) did not improve the results either.
- These results are similar to the analysis of ToxCAST data by the SAS team.

Toxcast Data Prediction

- Fifty endpoints are selected based on the rank of their frequencies
- Total of 1899 predictors from various tables
- Eight classification methods for prediction, 84 total models
 - Discriminant analysis (DA)
 - Distance Scoring (DS)
 - K-Nearest Neighbor (KNN)
 - General linear model selection (GLM)
 - Logistic regression (LR)
 - Partial least square (PLS)
 - Partition tree (PT)
 - Radial basis machine (RBM)
- 5-fold cross-validation, 10 iterations, performance metrics:
 - AUC
 - Accuracy
 - RMSE

Slide courtesy of Dr. Russ Wolfinger, SAS

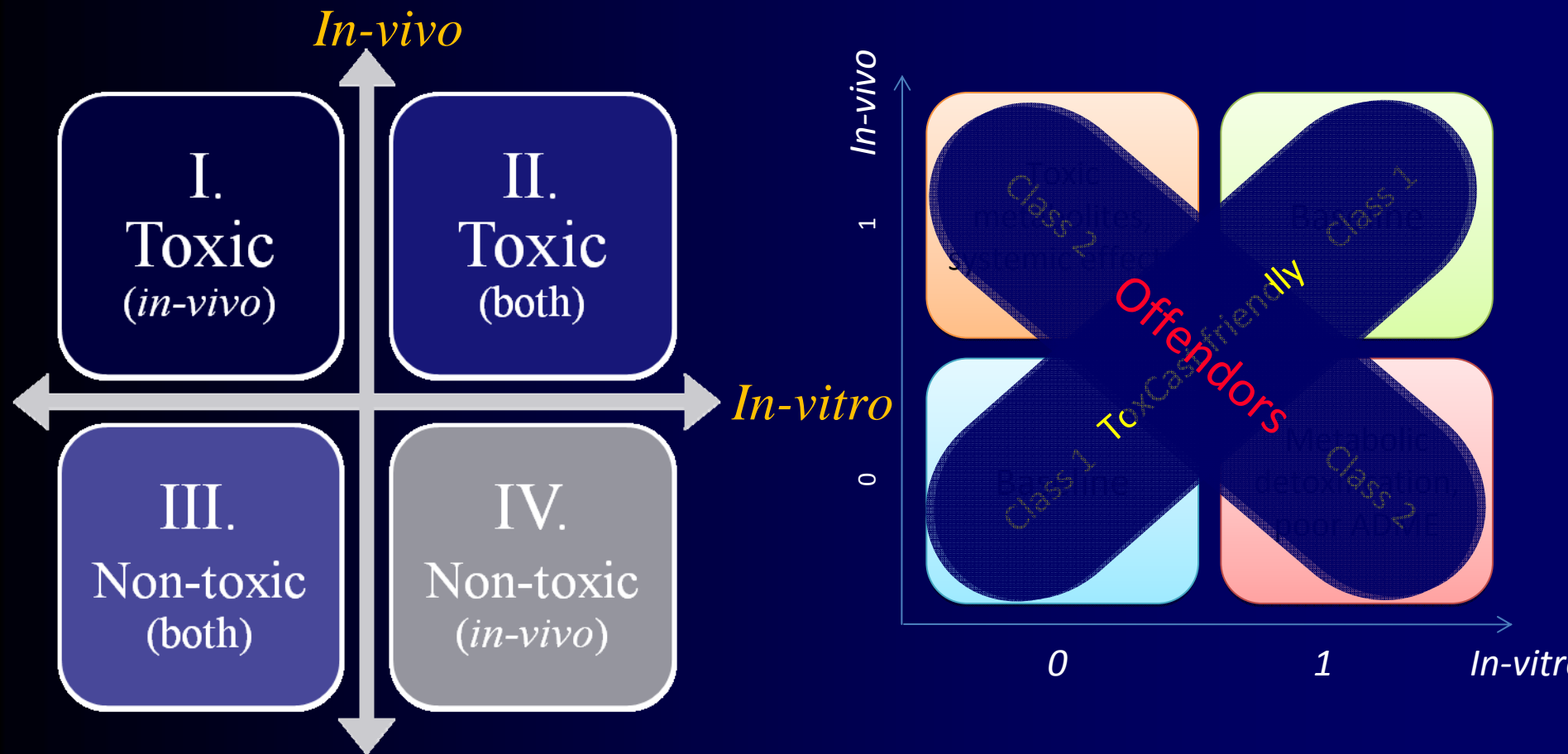
Prediction Comparison Based on Ave (Sensitivity + Specificity)



Slide courtesy of Dr. Russ Wolfinger, SAS

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

For each *In-vitro* vs. *In-vivo* profile ($3 \times 353 = 1059$ combinations):

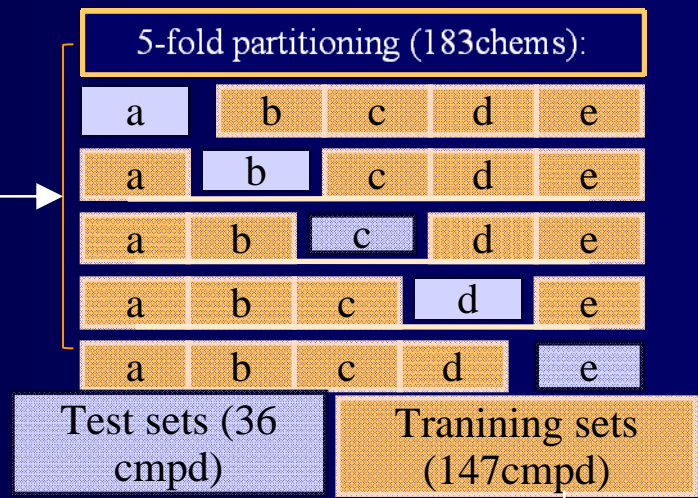
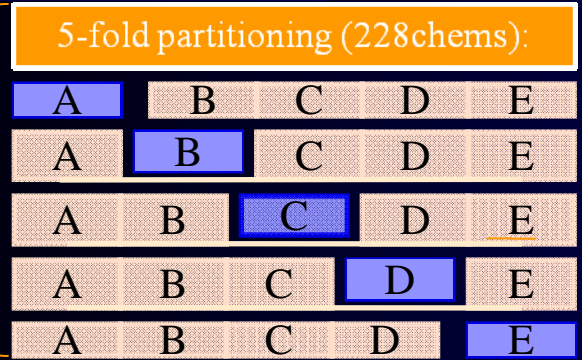


- Binary classification QSAR for “baseline” (II & III) vs. off-line (I & IV) using chemical descriptors only

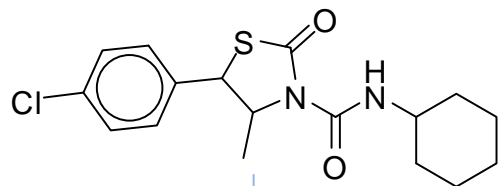
Modeling Workflow for each of the 1059 in vitro – in vivo series using RF and Dragon descriptors

228 chemicals
1224 chemical descriptors
3 in-vivo endpoints
353 in-vitro assays

Evaluation sets (55 cmpd) Modeling sets (183 cmpd)

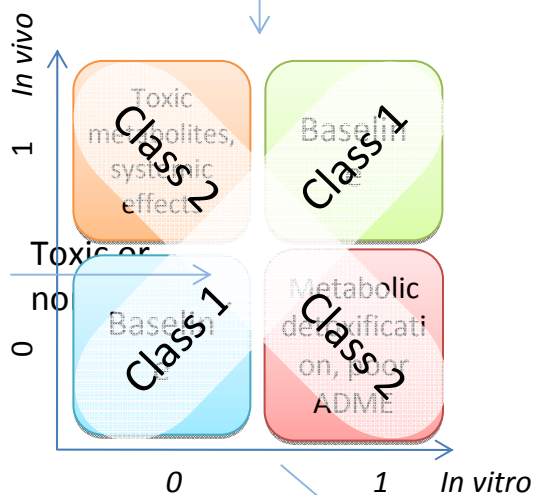


External prediction workflow



Validated RF
QSAR classifiers
(based selected
assays)

Prediction



Toxic or Non-toxic *in vivo*

Consensus Prediction

Individual Prediction for *in vivo* endpoint

Class 1 or 2

In vitro Assay Database

Top ToxCAST Bioassays (selected based on the mean of the 5-Fold External Cross-validation CCR)

For mouse liver proliferative lesions:

Bioassays	CCR1	CCR2	CCR3	CCR4	CCR5	Aver. CCR
NVS_GPCR_h5HT5A	0.63	0.58	0.63	0.56	0.62	0.60
BSK_3C_ICAM1	0.61	0.61	0.65	0.51	0.61	0.60
NVS_TR_rSERT	0.63	0.58	0.61	0.58	0.60	0.60
NVS_GPCR_hNPY1	0.64	0.57	0.59	0.55	0.63	0.59
NVS_GPCR_rAdrRa2						
NonSelective	0.60	0.60	0.59	0.59	0.58	0.59

Top ToxCAST Bioassays (selected based on the mean of the 5-Fold External Cross-validation CCR)

For mouse liver tumors:

Bioassays	CCR1	CCR2	CCR3	CCR4	CCR5	Aver. CCR
ABCB1_48	0.57	0.59	0.64	0.62	0.59	0.60
ATG_p53_CIS	0.64	0.53	0.61	0.56	0.63	0.59
BSK_3C_Proliferation	0.64	0.55	0.60	0.61	0.56	0.59
BSK_3C_uPAR	0.52	0.55	0.62	0.62	0.64	0.59
ATG_LXRb_TRANS	0.58	0.57	0.66	0.62	0.49	0.58

Top ToxCAST Bioassays (selected based on the mean of the 5-Fold External Cross-validation CCR

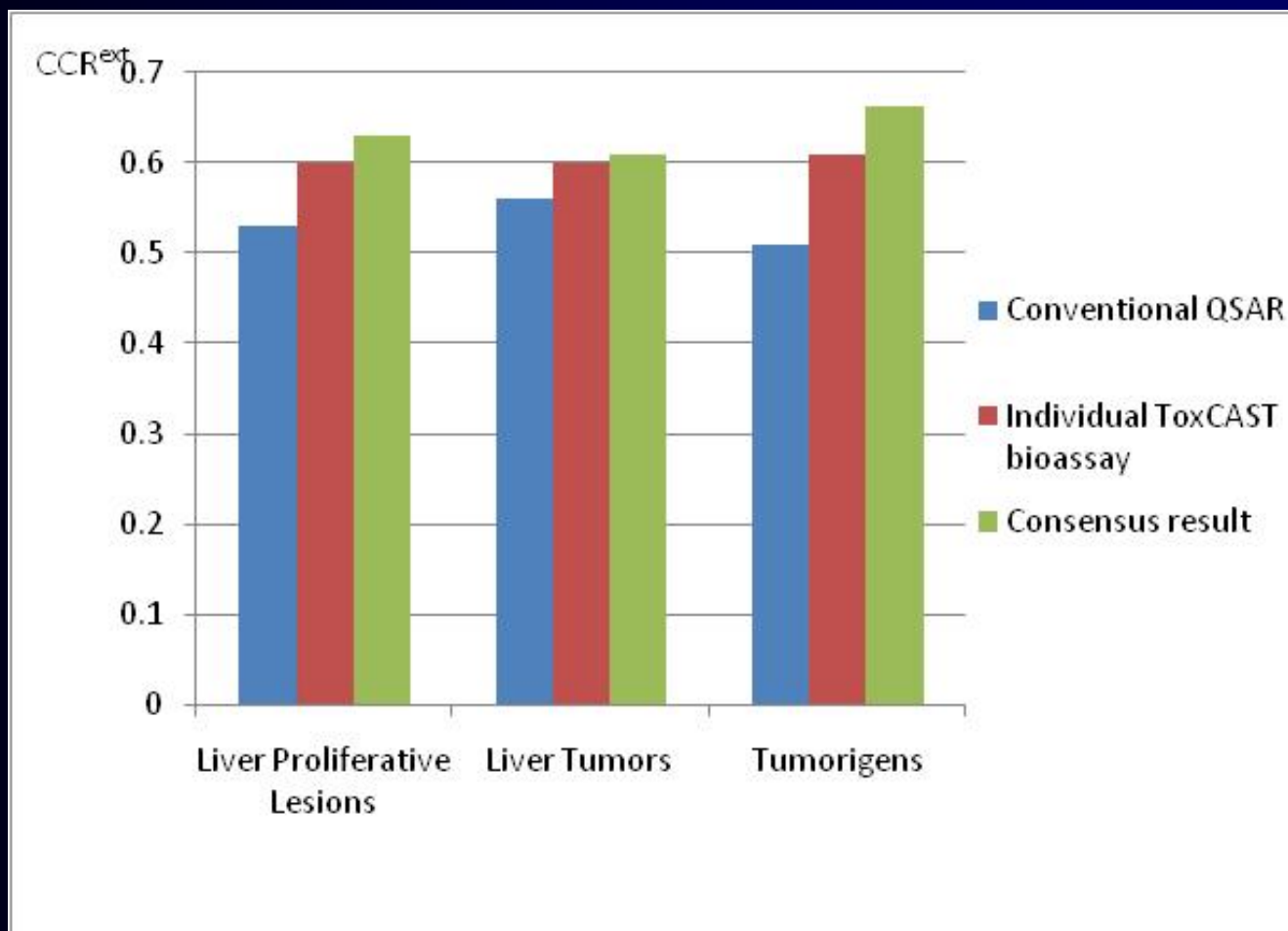
For mouse tumorigens:

Bioassays	CCR1	CCR2	CCR3	CCR4	CCR5	Aver. CCR
BSK_hDFCGF_SRB	0.61	0.62	0.64	0.59	0.61	0.61
Solidus_NoEnzyme	0.63	0.57	0.62	0.62	0.59	0.61
Solidus_PhaseII	0.64	0.60	0.59	0.62	0.58	0.60
MitoticArrest_1hr	0.61	0.60	0.59	0.61	0.61	0.60
ATG_E_Box_CIS	0.59	0.58	0.57	0.63	0.64	0.60

Consensus Modeling by Using the Top ToxCAST Bioassays for each *in vivo* Toxicity Endpoint

- 353 ToxCAST bioassays were ranked based on the average CCR of the 5 fold external validation for each *in vivo* end point
- The top 10-30 bioassays (judged by CCR) were selected to develop the consensus model.

The Consensus Prediction Generally Improves the Predictivity of QSAR Models



Conclusions and plans

- Focus on accurate prediction of external datasets is much more critical than accurate fitting of existing data: validate, then interpret!
 - validation!!!
 - applicability domain
 - consensus prediction using all acceptable models
 - Ideally, experimental validation of a small 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 both chemistry and biology
 - Integration of cheminformatics and bioinformatics: interpretation of significant descriptors in terms of pathways between Molecular Initiating Events (MIE) and Adverse Effects Outcome (AOE) (T. Schultz and OECD QSAR Toolbox paradigm)

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

Collaborators:

UNC: I. Rusyn, F. Wright

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, Liyang Zhang, Melody Luo, Guiyu Zhao, Andrew Fant

Research Programmer

Theo Walker

System Administrator

Mihir Shah



MAJOR FUNDING

NIH

- P20-HG003898 (RoadMap)
- R21GM076059 (RoadMap)
- R01-GM66940
- R0-GM068665

EPA (STAR awards)

- RD832720
- RD833825