The Use of Hybrid Chemical/Biological Descriptors in QSAR Modeling Improves the Accuracy of *In Vivo* Chemical Toxicity Prediction

Hao Zhu
Research Assistant Professor
School of Pharmacy
University of North Carolina at Chapel Hill

Outline

- Request of the computational toxicity predictors
- QSAR modeling approaches
- Applications
  - Hybrid QSAR modeling of rodent toxicity
  - Hierarchical QSAR modeling of rodent toxicity

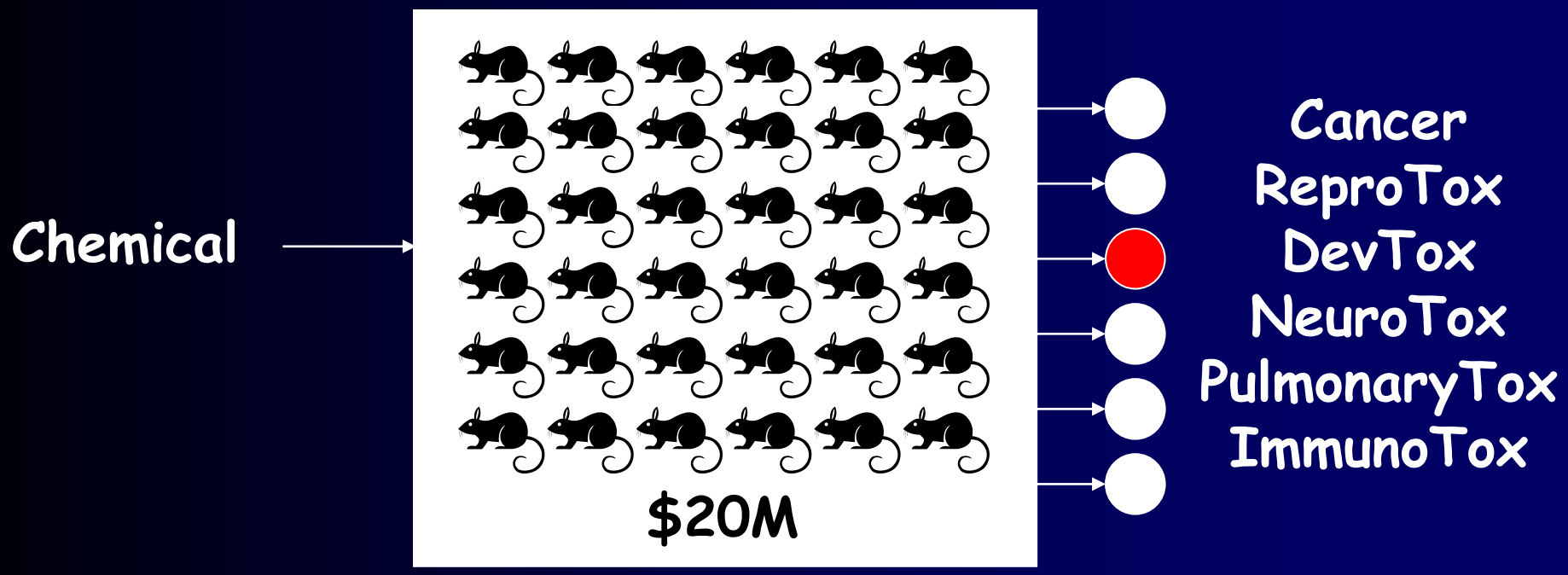
The Compounds Need to be Screened

- High Production Volume (HPV)
- Pesticides
- Drinking Water (DW) components

Around 15,000 compounds total need to be tested.

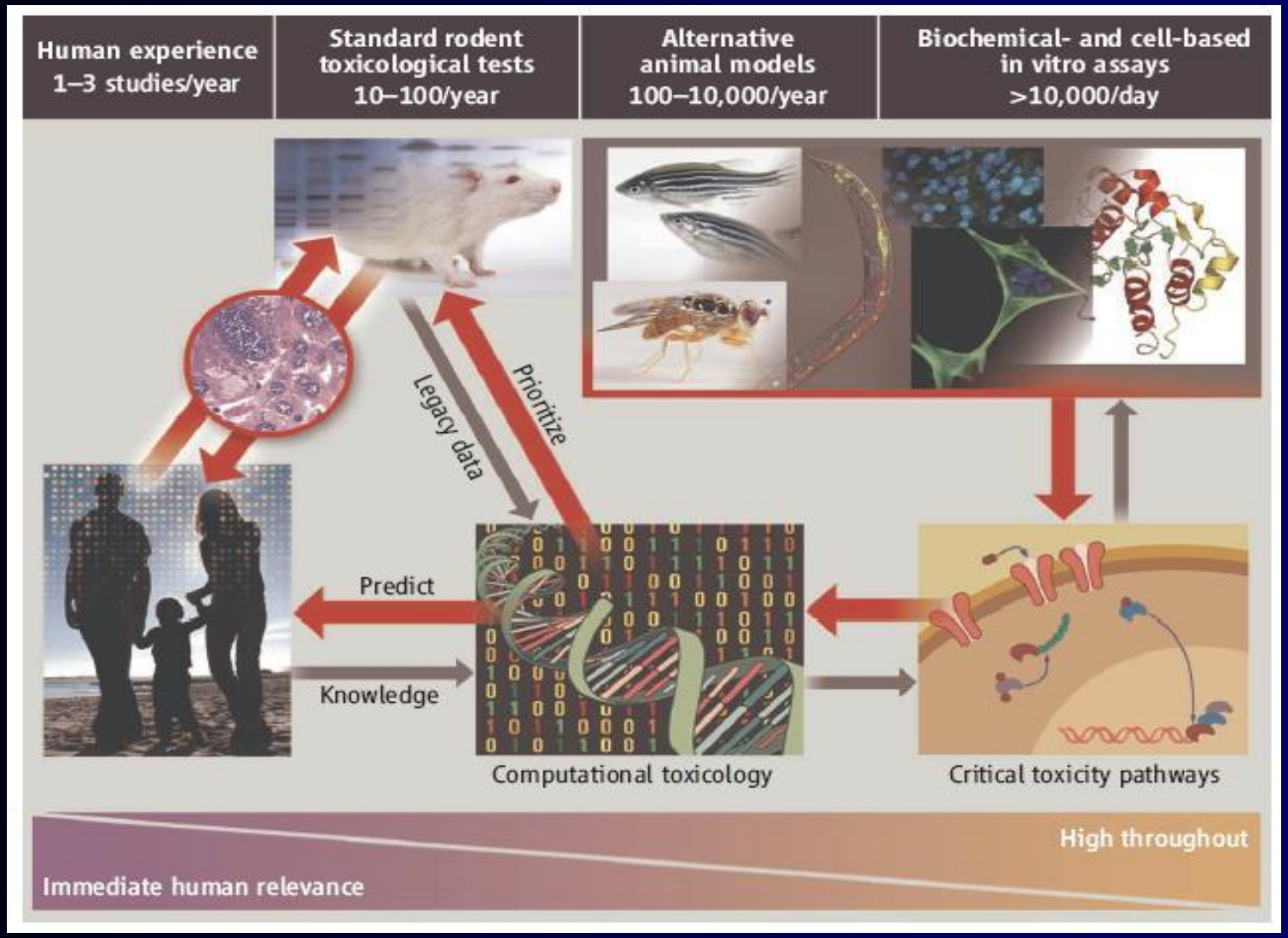
What we have known are less than 500.

Toxicity Prediction Today



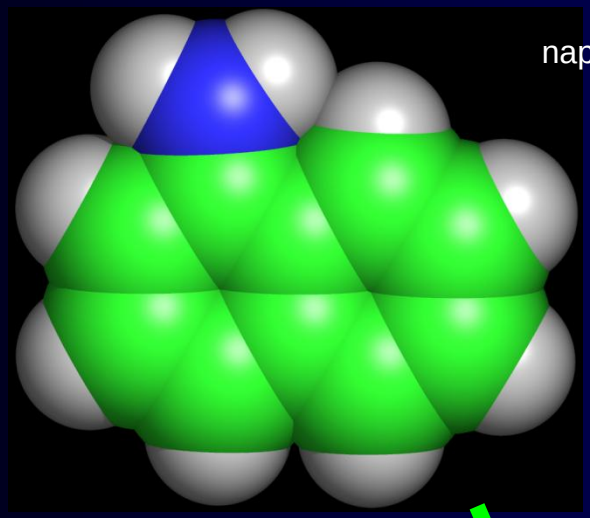
Slide courtesy of Dr. Richard Judson, EPA

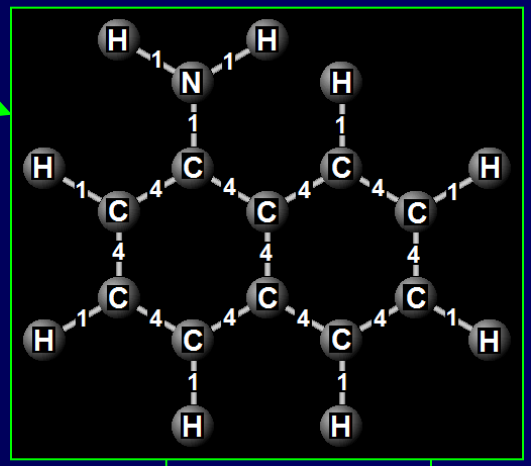
Toxicity Evaluation System



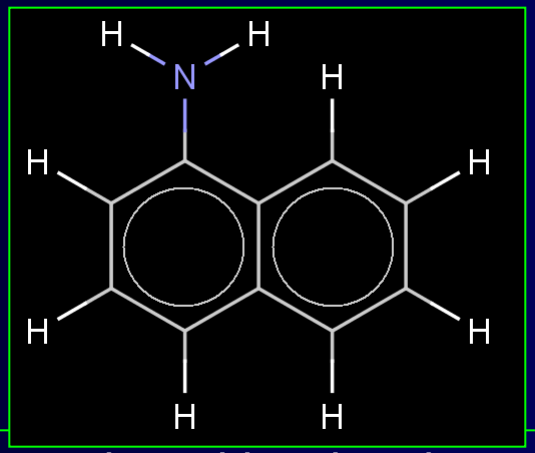
Collins, F. S., Gray, G. M. and Bucher J. R. *Science*, 2008, **319,** 906-907

STRUCTURE REPRESENTATION

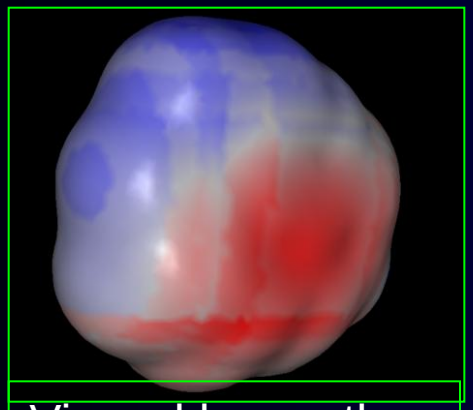




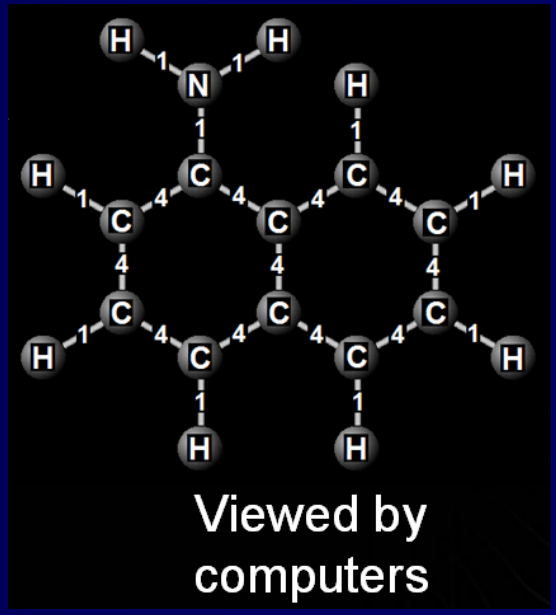
Viewed by computers

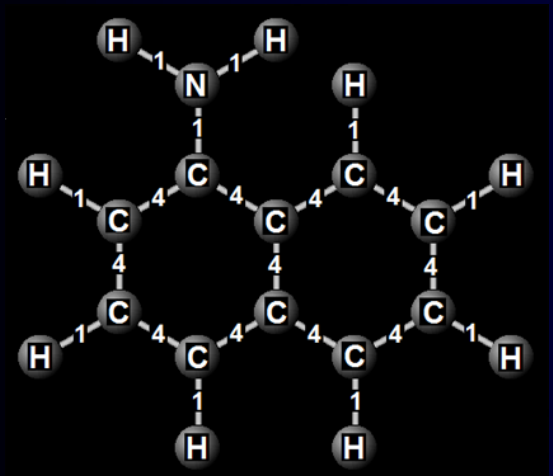


Viewed by chemists



Viewed by another molecule


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

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

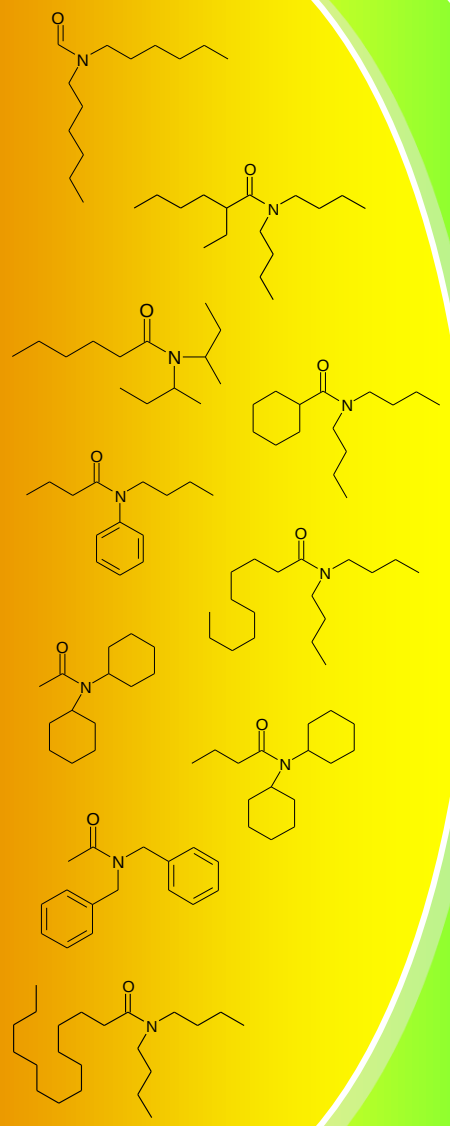
(connectivity table, label-types of bonds)

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

Molecular descriptors

Principles of QSAR/QSPR modeling

COMPOUNDS



DESCRIPTORS

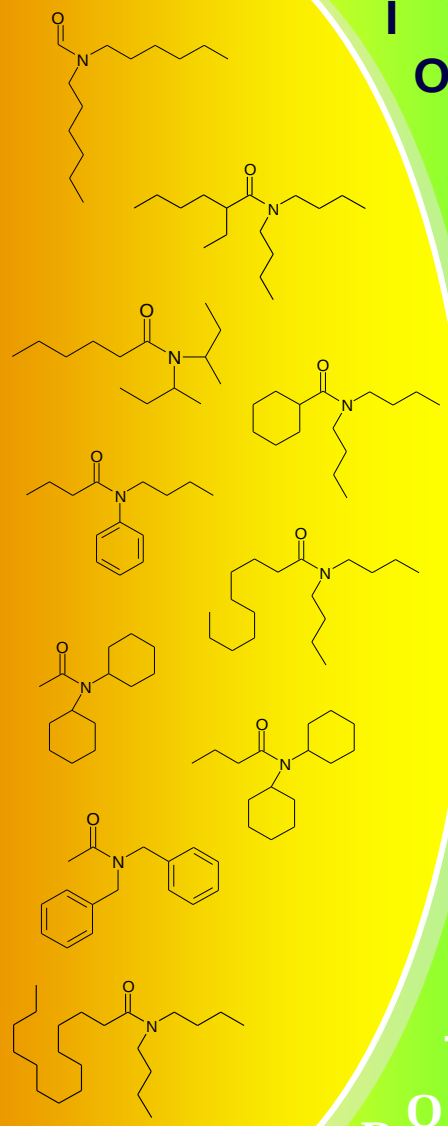
Quantitative
Structure
Property
Relationships

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

Principle of QSAR/QSPR modeling

COMPOUNDS



BIOPROFILES + DESCRIPTORS

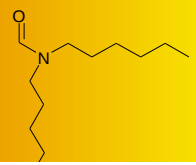
Quantitative
Structure
Property
Relationships

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

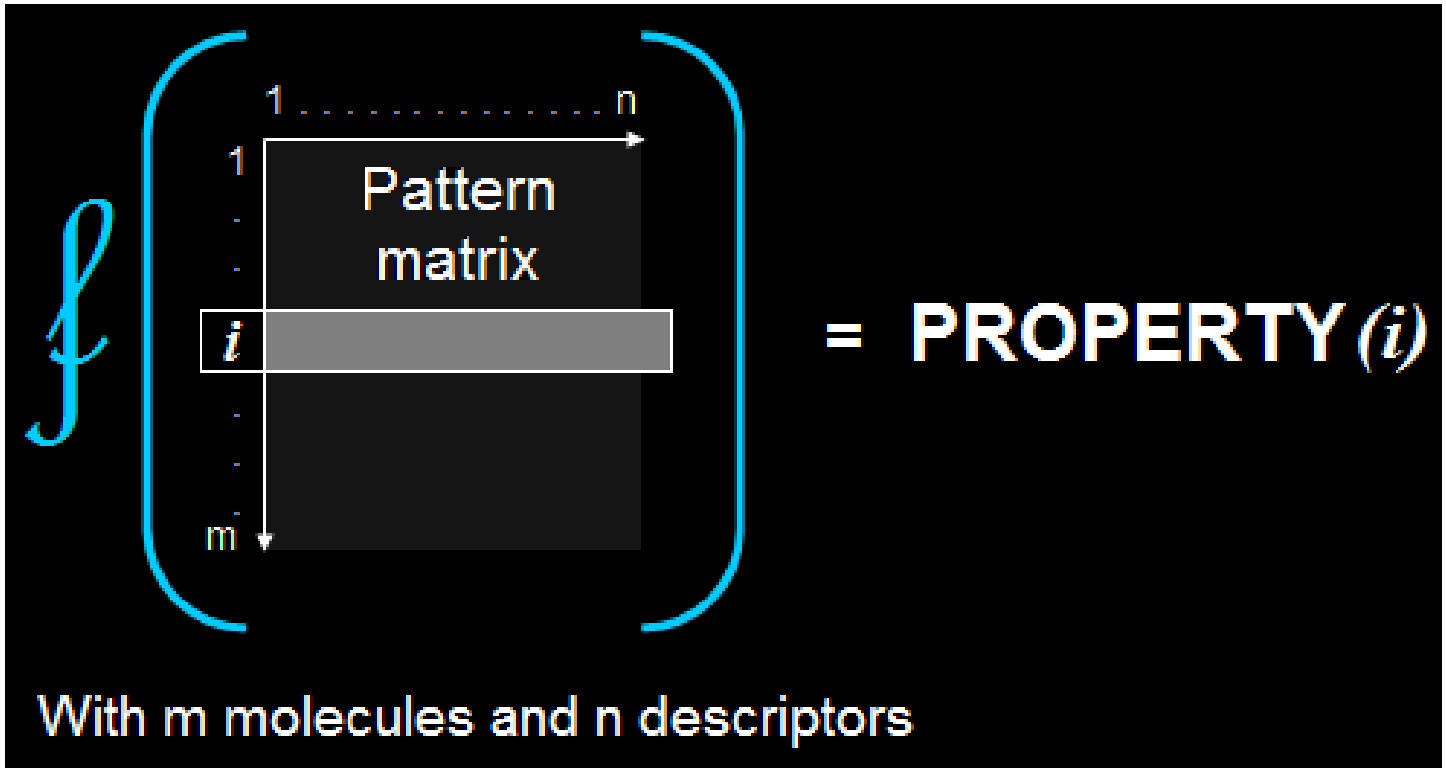
PROPERTY

Principles of QSAR/QSPR modeling

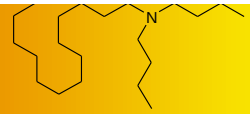
COMPOUNDS



0.613

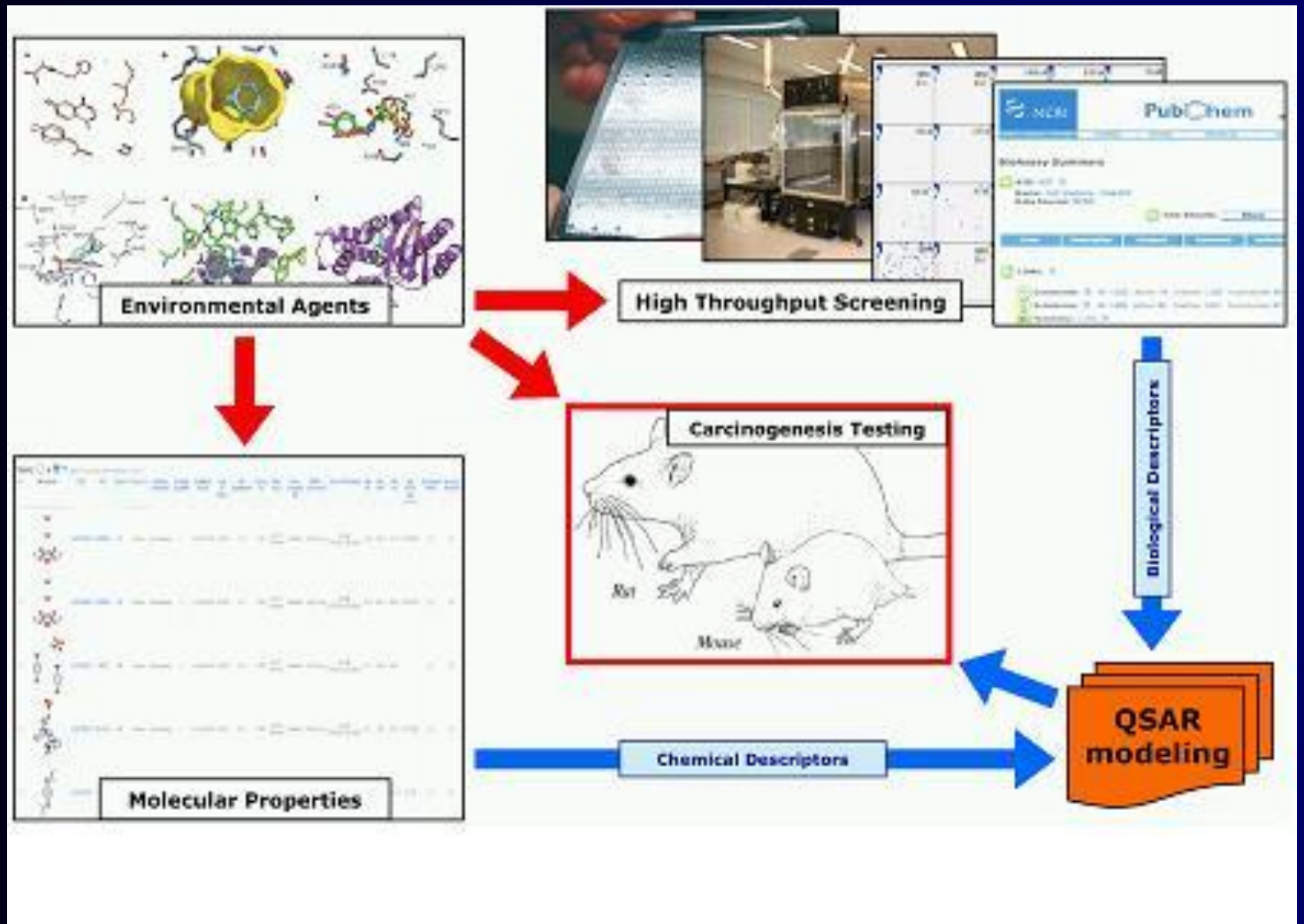


PROPERTY

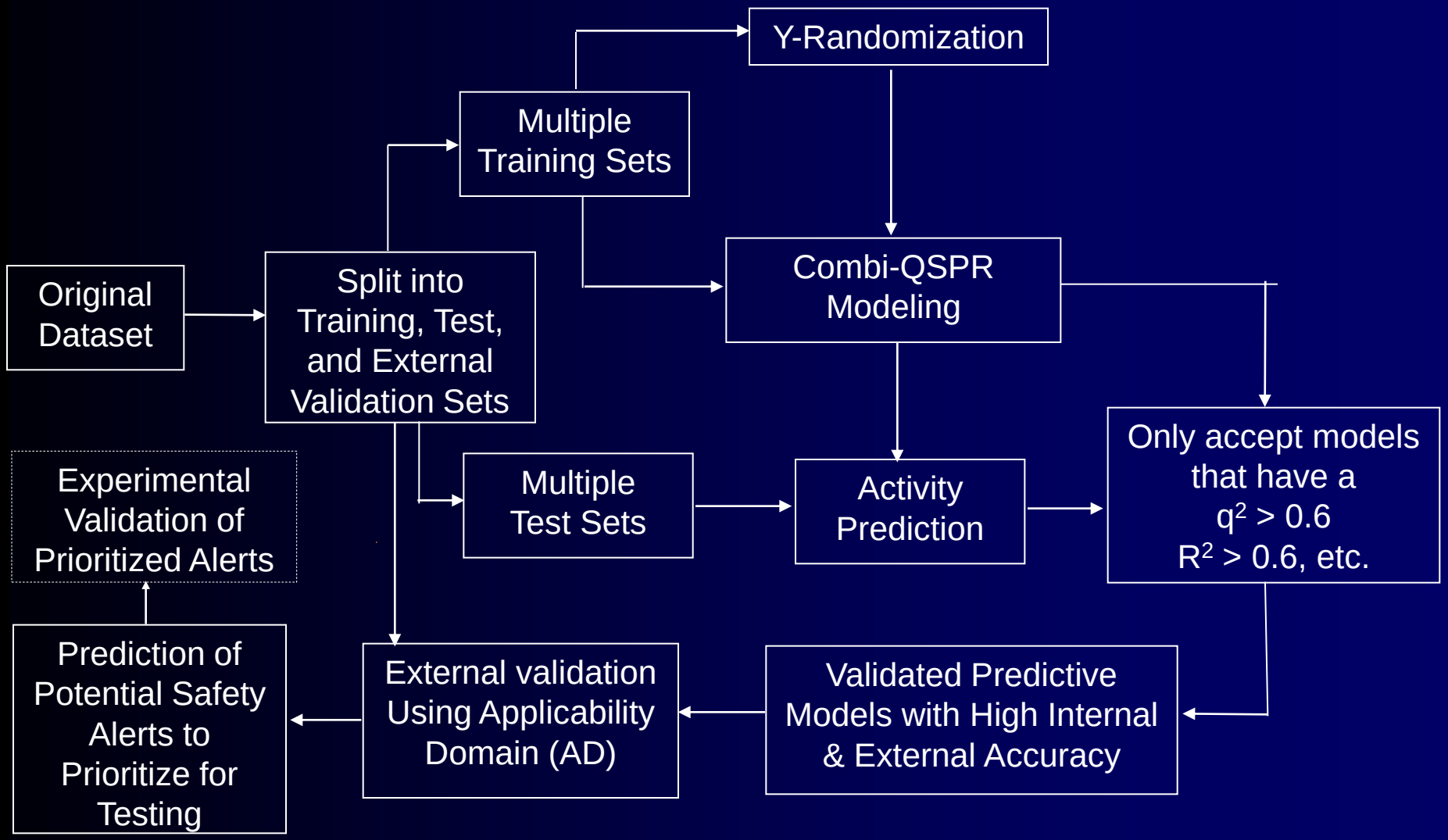


1.005

Using *In Vitro* Assay Results as Descriptors in QSAR Modeling of *In Vivo* Endpoints

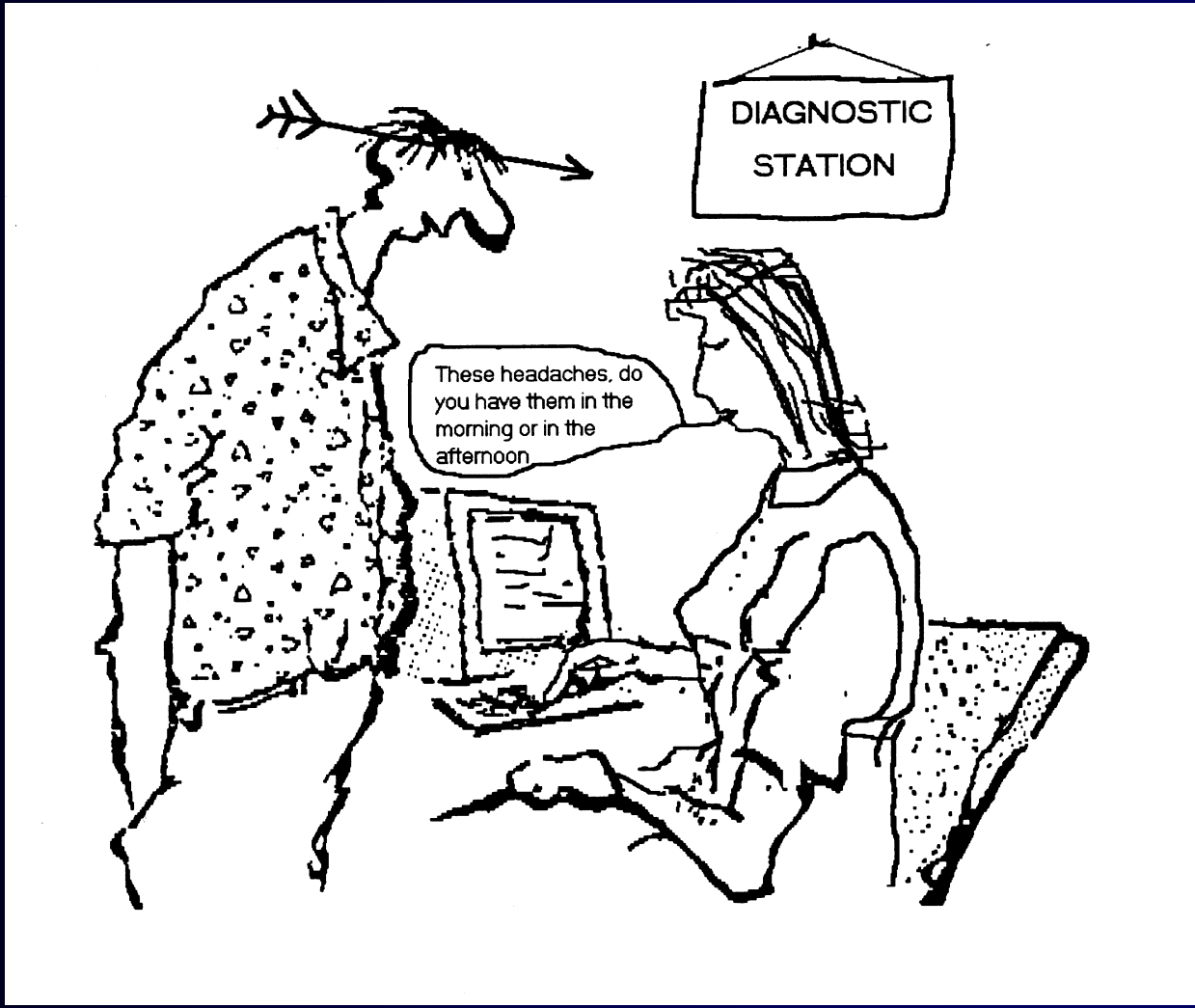


Predictive QSAR Workflow*



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

Applicability Domain in QSAR Studies



Slide courtesy of Dr. Gilles Klopman, MultiCASE Inc.

Experimental Study I. Using Full High-Throughput Screening Dose Response Curves as Biological Fingerprints of Organic Compounds in QSAR Studies*

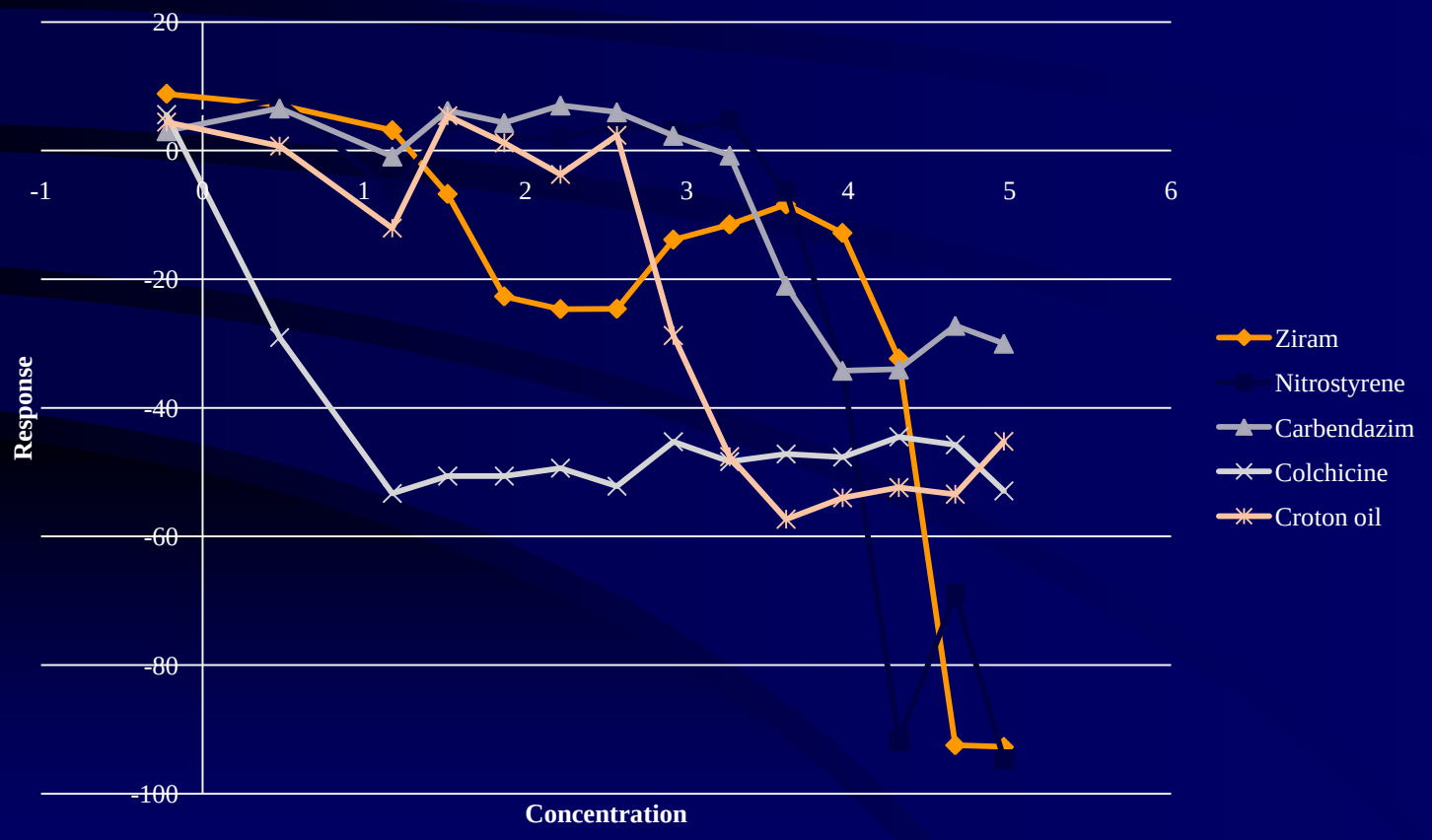
Zhu, Sedykh, et al, in preparation; EPA Collaborator: Ann Richard

***Zhu H, Rusyn I, Richard A, Tropsha A.* Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ Health Perspect* 2008; (116): 506-513**

NCGC HTS Dataset

- 1408 chemicals, 1353 unique compounds
- 13 cell lines: 9 human species, 2 rat species, 2 mouse species
- Cell viability assay
- All datasets available via PubChem

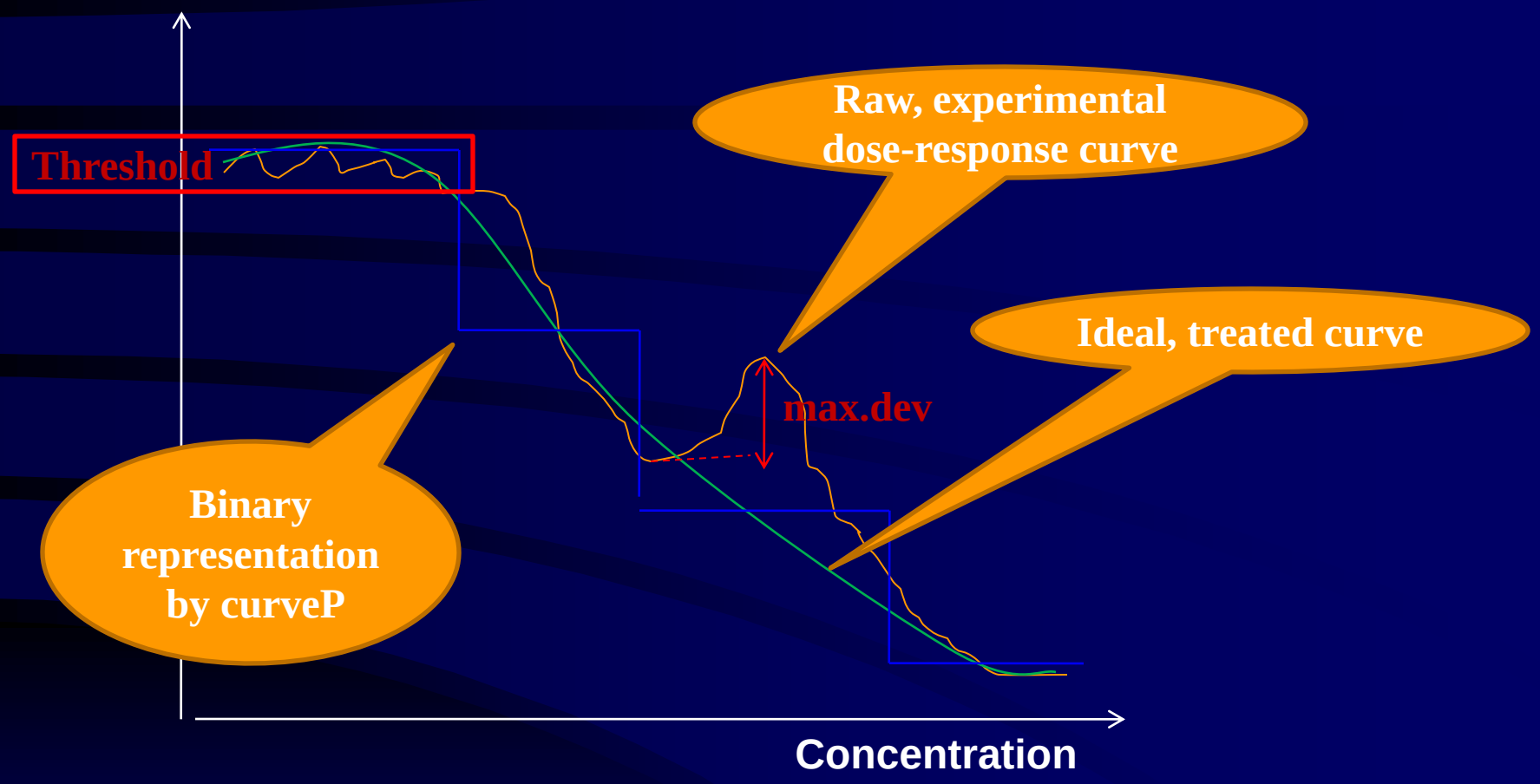
How to use the full dose response curve?



Objective : Remove the Noise in the Dose Response Curves

- 1. Deviation treatment: Compare two adjacent points: if the difference between them are less than a certain value, they will be treated as the same.
- 2. Threshold treatment: Give a special treatment of the low dose response (data near the baseline): if the change compared to the previous response is less than a certain value, the change will be erased.

HTS Raw Data treatment



Threshold – to control data deviation near baseline (“noise”)
Max.dev – to control deviation from monotonous behaviour

QSAR Tools

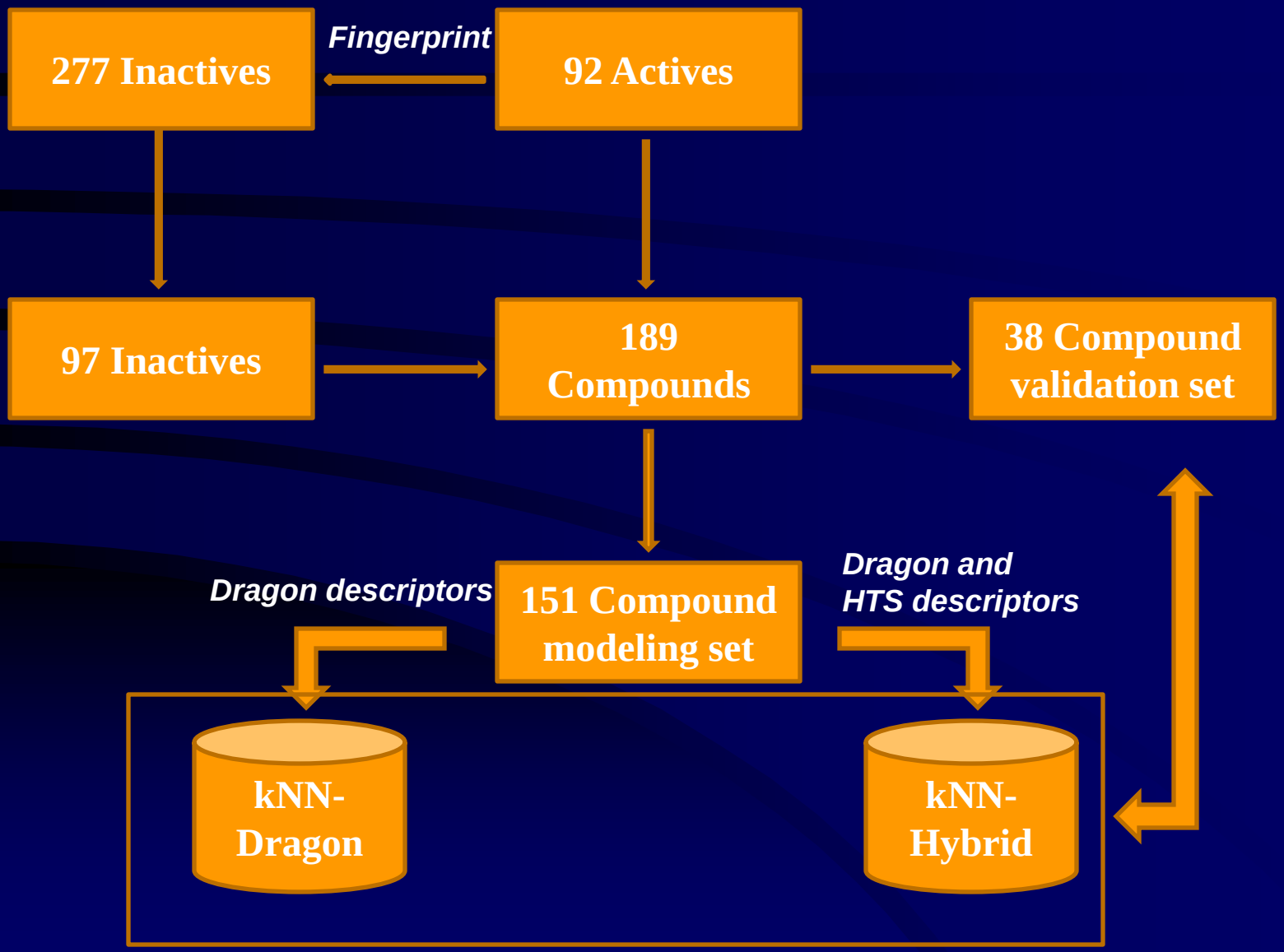
- k Nearest Neighbor (kNN)
- Random Forest (RF)
- Applicability Domain

Using HTS Dose Response Curve to Assist QSAR Modeling of Rat Acute Toxicity

- Three types of descriptors:

  Chemical calculated by Dragon software (300+); Biological (150+); Hybrid (400+)
- LD50 data: 690 unique organic compounds, 92 actives, 321 marginally actives and 277 inactives.

HTS vs. Rat LD50s



Prediction of the 38 compound External Validation Set by kNN LD50 models

	kNN-Dragon	kNN-Hybrid
Sensitivity	59%	72%
Specificity	94%	98%
CCR	77%	85%
Coverage	76%	93%

The data are the average values after repeating the experiments 5 times.

Prediction of the 38 compounds in the External Validation Set by RF LD50 models

	RF-Dragon	RF-Hybrid
Sensitivity	71%	88%
Specificity	85%	92%
CCR	78%	90%
Coverage	100%	100%

The data are the average values after repeating experiments 5 times.

Experimental Study II: A Two-step Hierarchical Quantitative Structure Activity Relationship Modeling Workflow for Predicting *in vivo* Chemical Toxicity from Molecular Structure*

*Zhu, Rusyn, Wright, et al, EHP, 2009, in press; 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**

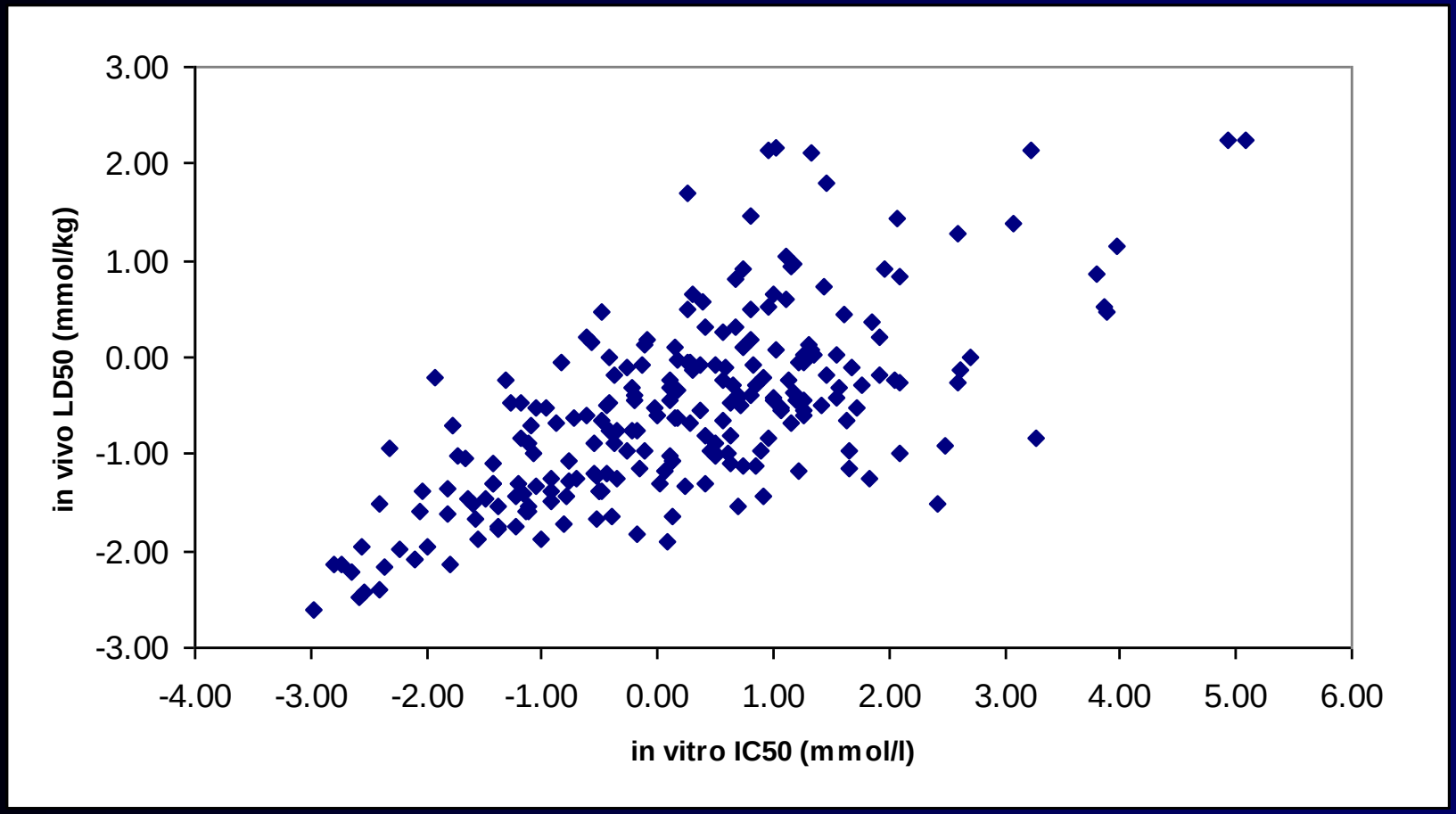
Random split

↓ **230 compounds modeling set**

↓ **23 compounds validation set**

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

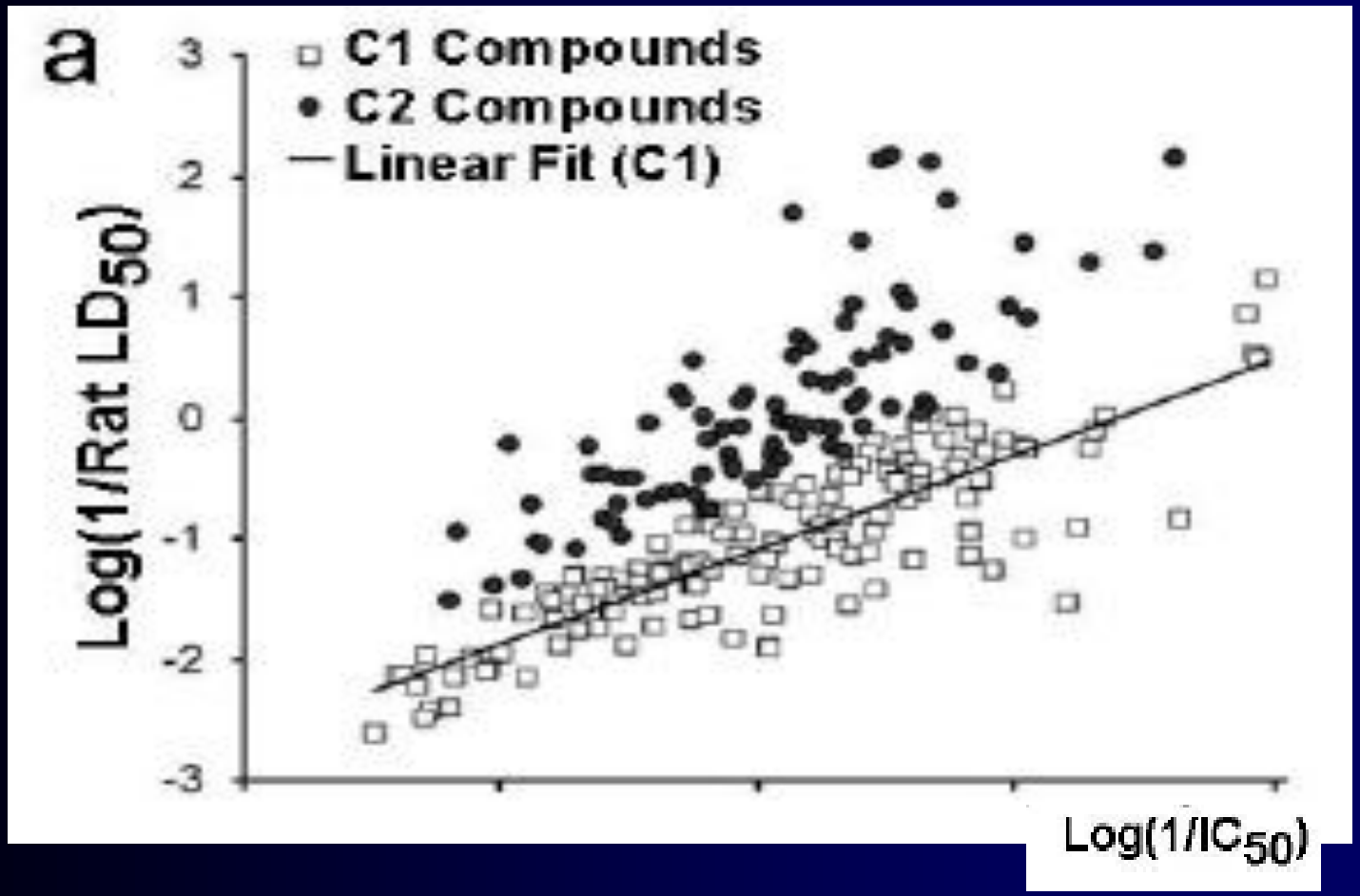
Poor in vitro-in vivo Correlation Between IC50 and Rat LD50 Values



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

Moving Regression for Data Partitioning

$$\eta(x_i, y_i) = \begin{cases} 1, \text{if } \; y_i \in [ax_i + b - d_1, ax_i + b + d_2] \\ 0, \text{otherwise} \end{cases}$$

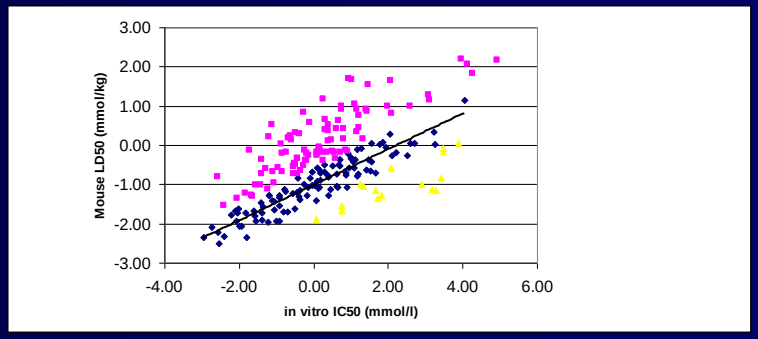
$$F(a,b) = \sum_{i=1}^{n} \eta(x_i, y_i)(y_i - ax_i - b)^2$$

$$\eta(x_i, y_i) \sim \frac{1}{2}\left\{\frac{1}{1+\exp[-P_1(y_i - ax_i - b + d_1)]} + \frac{1}{1+\exp[P_2(y_i - ax_i - b - d_2)]}\right\}$$

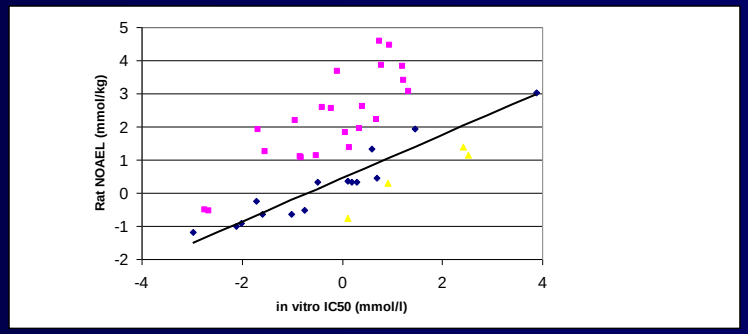
$$F(a,b) = \sum_{i=1}^{n} \frac{1}{2}\left\{\frac{1}{1+\exp[-P_1(y_i - ax_i - b + d_1)]} + \frac{1}{1+\exp[P_2(y_i - ax_i - b - d_2)]}\right\}(y_i - ax_i - b)^2$$

Cytotoxicity IC50 Values vs. in vivo Toxicity Measures

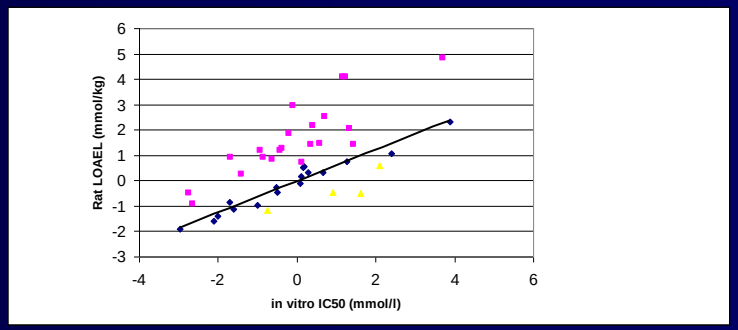
- IC50 vs. mouse LD50 values



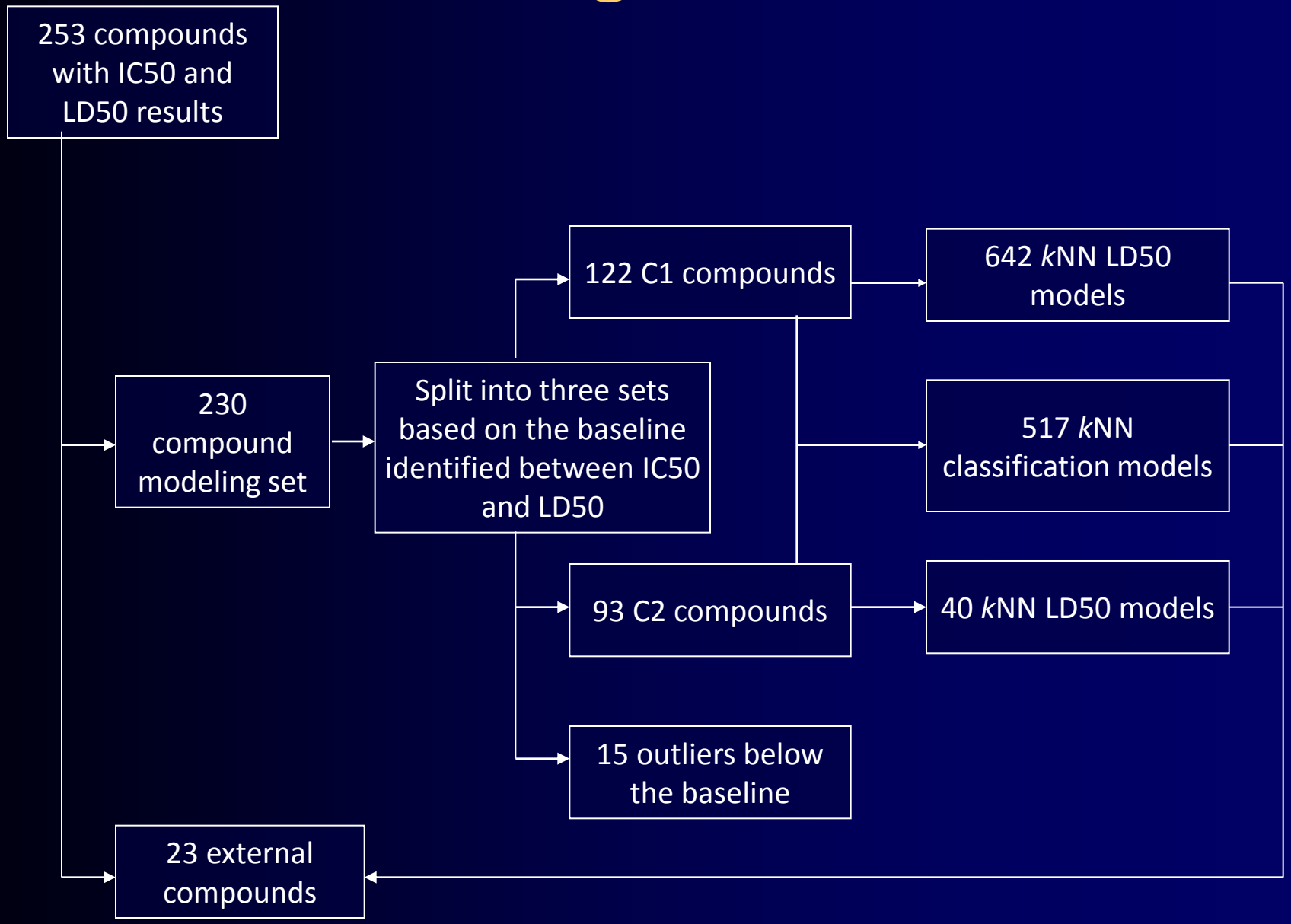
- IC50 vs. rat NOAEL values



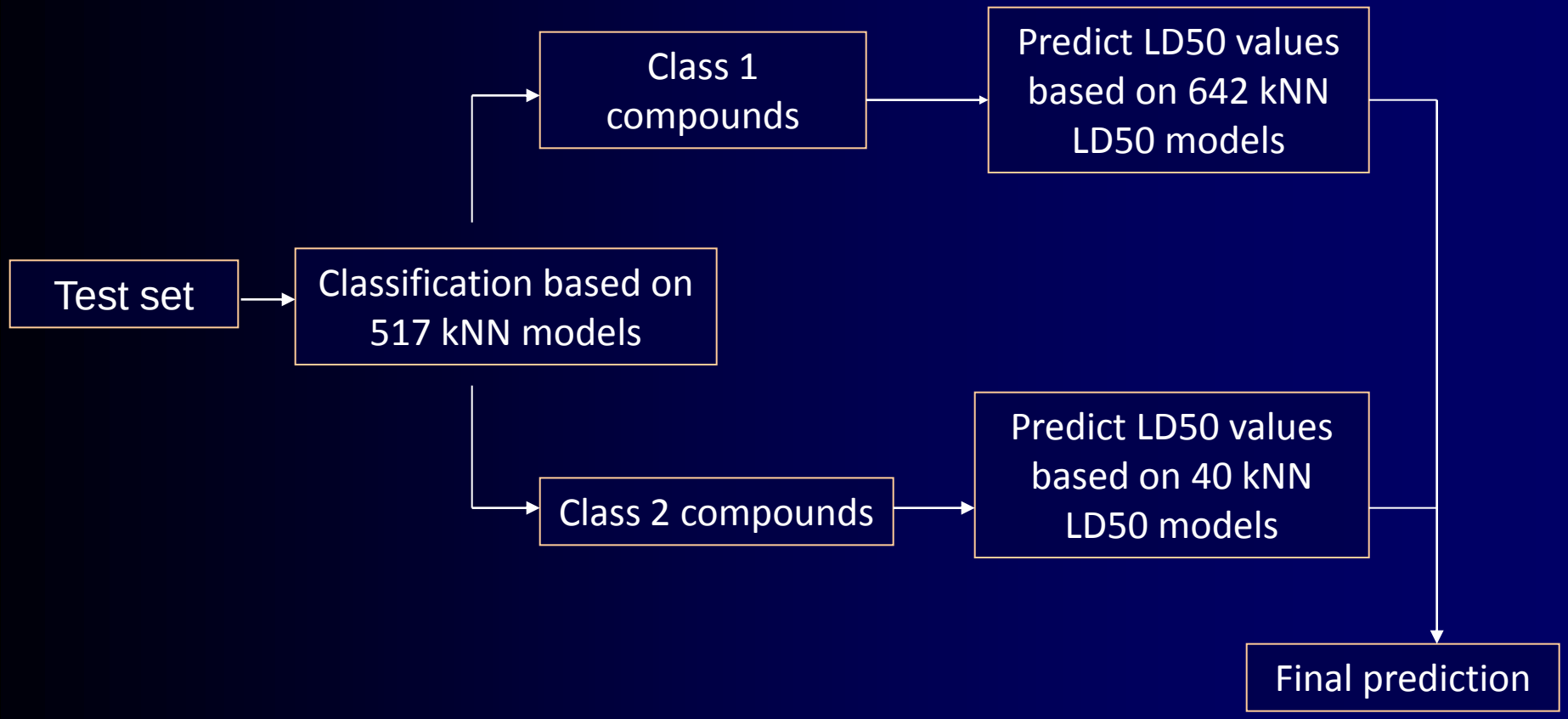
- IC50 vs. rat LOAEL values



Modeling Workflow



Prediction Workflow



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

No AD:
Classification rate = 62%

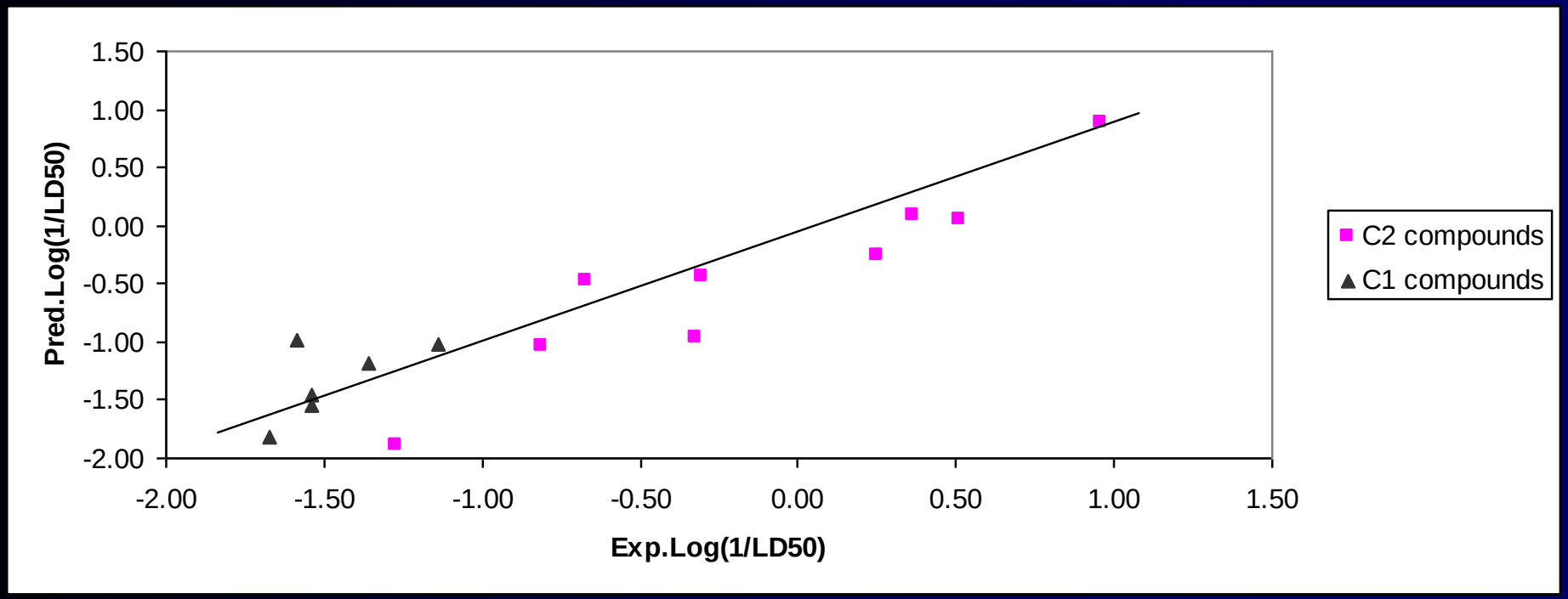
	Pred. C1	Pred. C2
Exp. C1	7	2
Exp. C2	6	5

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)



Prediction of New ZEBET Compounds

- Additional 115 ZEBET compounds with rat LD50 testing results obtained from Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM).

- R^2, *MAE* and prediction coverage of 0.60, 0.46, and 62%

Comparison Between Our Model and Toxicity Prediction by Komputer Assisted Technology (TOPKAT) LD50 Predictor

- 27 out of the 115 new ZEBET compounds do not exist in the TOPKAT LD50 training set (version 6.1).
- Prediction of 27 new ZEBET compounds

	This model		TOPKAT	
	No AD	With AD	No AD	With AD
R^2	0.69	0.73	0.16	0.50
MAE	0.42	0.34	0.78	0.46
Coverage	100%	70%	100%	70%

Conclusions

- Focus on accurate prediction of external datasets is much more critical than accurate fitting of existing data: validate, then interpret!
  - validation!!!
  - applicability domain
  - 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

Current Project

ToxCast™ Program | National Center for Computational Toxicology | US EPA - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

http://www.epa.gov/ncct/toxcast/

Toxcast

Most Visited Getting Started Latest Headlines Customize Links Free Hotmail My eBay Windows Marketplace Windows Media Windows

Search Web Shopping Health Bookmarks My Yahoo! Yahoo! Mail Finance News

U.S. ENVIRONMENTAL PROTECTION AGENCY

National Center for Computational Toxicology

Share

Contact Us Search: All EPA This Area Go

You are here: EPA Home » National Center for Computational Toxicology » ToxCast™ Program »

- Home
- Basic Information
- Organization
  - Post Doc Profiles
- Framework
- Research Activities
  - ACToR
  - DSSTox
  - ToxCast™
  - ToxRefDB
  - v-Liver™
  - v-Embryo™
- Conferences and Seminars
- Products
  - Journal Articles and Book Chapters
  - Presentations
- BOSC Information
- EPA Communities of Practice
- Jobs and Opportunities
- Related Information

ToxCast™ Program

Predicting Hazard, Characterizing Toxicity Pathways, and Prioritizing the Toxicity Testing of Environmental Chemicals

STAY UP TO DATE WITH E-MAIL ALERTS

Introduction

In 2007, EPA launched ToxCast™ in order to develop a cost-effective approach for prioritizing the toxicity testing of large numbers of chemicals in a short period of time. Using data from state-of-the-art high throughput screening (HTS) bioassays developed in the pharmaceutical industry, ToxCast™ is building computational models to forecast the potential human toxicity of chemicals. These hazard predictions will provide EPA regulatory programs with science-based information helpful in prioritizing chemicals for more detailed toxicological evaluations, and lead to more efficient use of animal testing.

In its first phase, ToxCast™ is profiling over 300 well-characterized chemicals (primarily pesticides) in over 400 HTS endpoints. These endpoints include biochemical assays of protein function, cell-based transcriptional reporter assays, multi-cell interaction assays, transcriptomics on primary cell cultures, and developmental assays in zebrafish embryos. Almost all of the compounds being examined in Phase 1 of ToxCast™ have been tested in traditional toxicology tests, including developmental toxicity, multi-generation studies, and sub-chronic and chronic rodent bioassays. ToxRefDB, a relational database being created to house this information, will contain nearly $1B worth of toxicity studies in animals when completed. ToxRefDB is integrated into a more comprehensive data management system developed by NCCT called ACToR (Aggregated Computational Toxicology Resource), that manages the large-scale datasets of ToxCast™. ACToR is comprised of several independent data repositories linked to a common database of chemical structures and properties, and to tools for development of predictive HTS and genomic bioactivity signatures that strongly correlate with specific toxicity endpoints from ToxRefDB. These ToxCast™ signatures will be defined and evaluated by their ability to predict outcomes from existing

ToxCast™ Navigation

- Introduction
- Chemicals
- Assays
- Information Management
- Partnerships
- Contractors
- Presentations
- Publications
- News
- Data Analysis Summit

Done

ToxCAST Data Overview

320 substances with *in vitro* and *in vivo* experimental results:

Source	ACEA	Attagene	BioSeek	Cellumen	Gentronix	NovaScreen	Solidus	CellzDirect	NCGC	ToxRefDB
#Assays (600)	7	81	87	33	1	239	4	48	24	76
Experiment	In vitro (Cell)	In vitro (Cell)	In vitro (Cell)	In vitro (Cell)	In vitro (Cell)	In vitro (Biochemical)	In vitro (Cell)	In vitro (Cell)	In vitro (Cell)	in vivo
Species	Human	Human	Human	Human	Human	Human(146); Rat(67); Mouse(2); Rabbit(2); Pig(1); Guinea Pig(10); Sheep(2); Cow(9);	Human	Human	Human(23); Rat(1);	Rat(51); Mouse(7); Rabbit(18);
Description	Cell-growth dynamics	Transcription factors	Pharm. targets, adverse effects, protein markers	Cellular toxicity indicators	HTS genotoxicity	HTS: ADME-Tox, enzyme, nuclear receptor, GPCR	Cytotoxicity and metabolism	Gene expression for transport proteins, metabolic enzymes	HTS: nuclear receptor, cell viability and p53 assays	In-vivo animal toxicity, mg/kg/day
Endpoint	IC_{50}	LEL	LEL	IC_{50}	LEL	IC_{50}	LC_{50}	LEL	IC_{50}	LEL
Max. tested conc. (μM)	100	100	40	200	200	20 & 50	960	40	200	2500(?) mg/kg/day
Data richness	17%	22%	23%	13%	10%	3%	22%	12%	2%	11% *[N/A is 21%]*

Acknowledgements

Principal Investigator
Alexander Tropsha

Research Professors
Clark Jeffries, Alexander Golbraikh, 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, **Liying Zhang**, Melody Luo, Guiyu Zhao, Andrew Fant

Research Programmer
Theo Walker

System Administrator
Mihir Shah

MML
UNC.EDU

MAJOR FUNDING
NIH
- P20-HG003898 (RoadMap)
- R21GM076059 (RoadMap)
- R01-GM66940
- R0-GM068665

EPA (STAR awards)
- RD832720
- RD833825