UNC
ESHELMAN
SCHOOL OF PHARMACY

MML
UNC.EDU

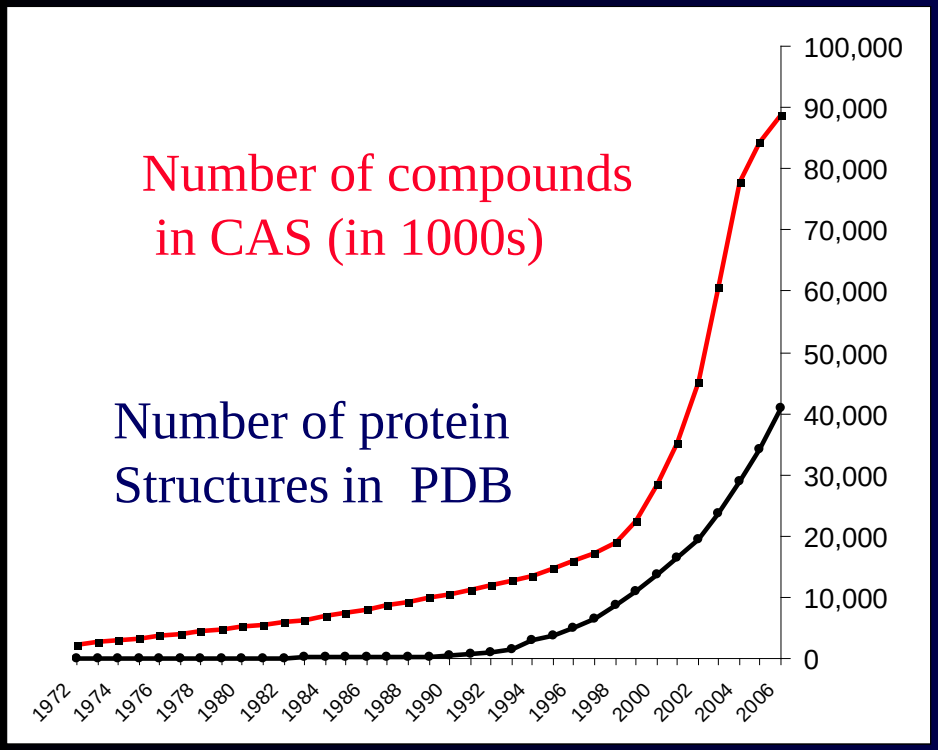
Best practices for developing predictive QSAR models

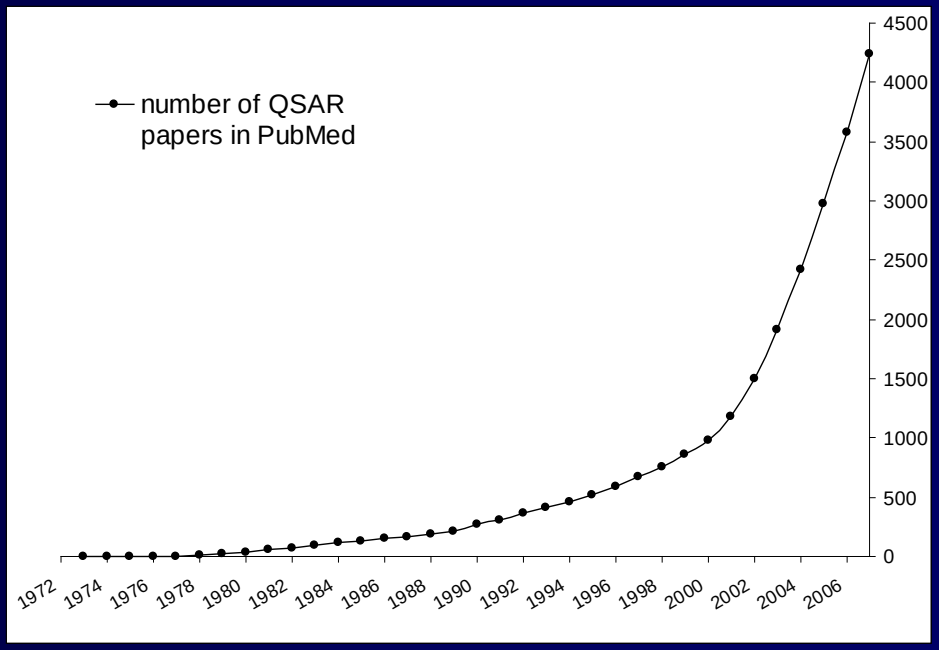
Alexander Tropsha
Laboratory for Molecular Modeling
and
Carolina Center for Exploratory Cheminformatics Research
School of Pharmacy
UNC-Chapel Hill

OUTLINE

- Introduction: Brief outline of the QSAR approach
- Why models fail (bad practices)
- Good practices.
  - Predictive QSAR Modeling Workflow
  - Examples of the Workflow applications
  - Emerging applications of QSAR: chemocentric informatics
- Conclusions: QSAR modeling is a decision support

The rumors of QSAR demise have been greatly exaggerated

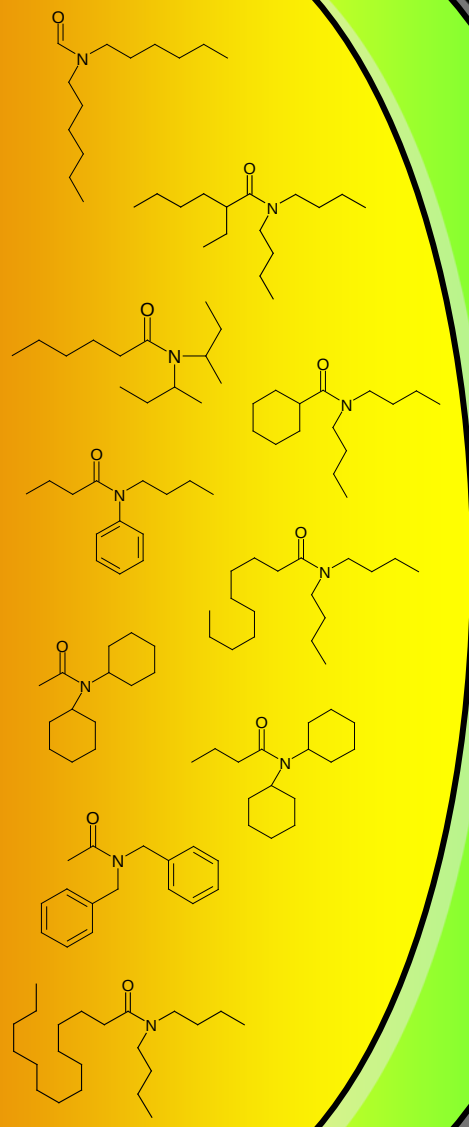




Graphs are courtesy of Prof. A. Cherkasov

Principles of QSAR modeling

COMPOUNDS



DESCRIPTORS

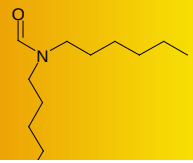
Quantitative Structure Activity Relationships

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

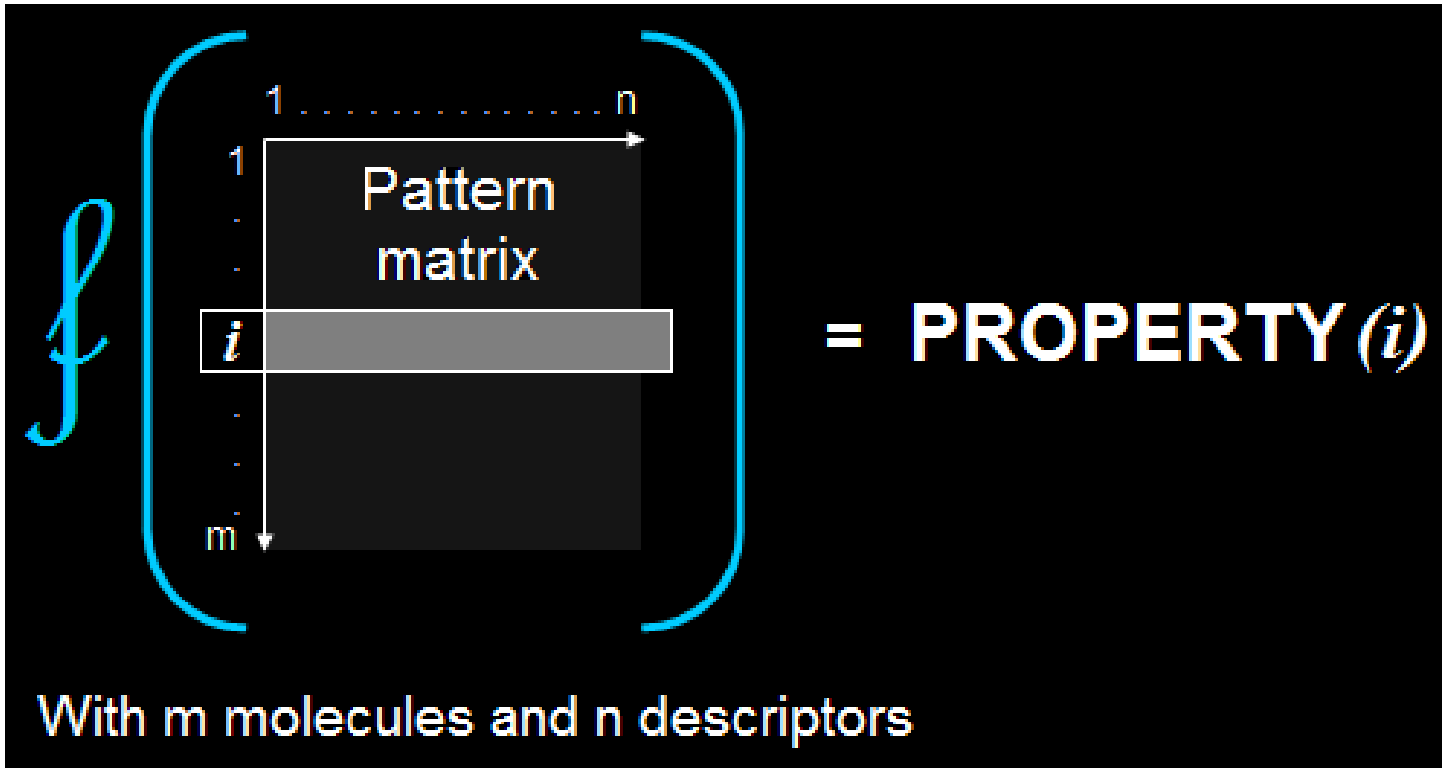
ACTIVITY

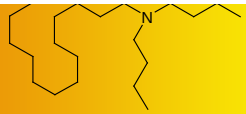
Principles of QSAR/QSPR modeling

COMPOUNDS



0.613





1.005

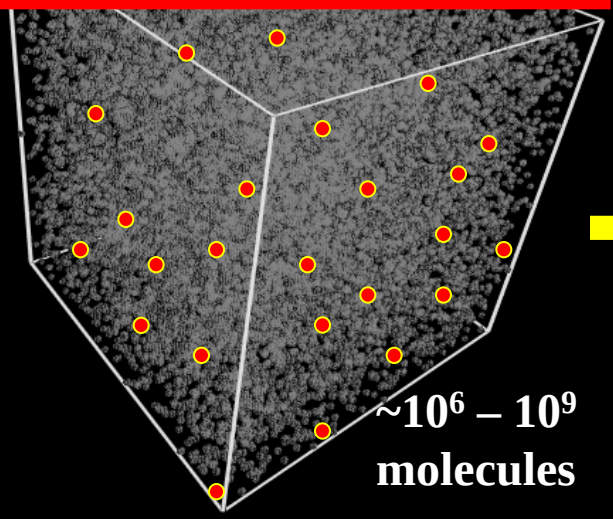
PROPERTY

The utility of QSAR models

MML
UNC.EDU

CHEMICAL STRUCTURES → CHEMICAL DESCRIPTORS → PREDICTIVE QSAR MODELS ← PROPERTY/ ACTIVITY

CHEMICAL DATABASE



~10^6 – 10^9 molecules

VIRTUAL SCREENING

HITS

INACTIVES

QSAR Modeling appears easy…

Goal: Establish correlations between descriptors and the target property capable of predicting activities of novel compounds

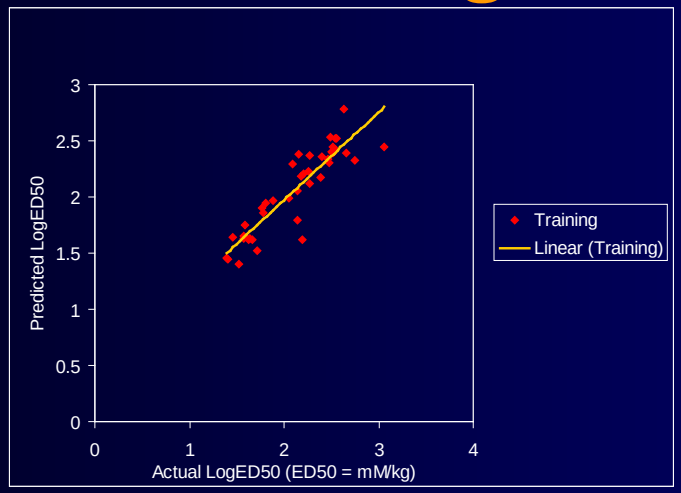
Chemistry	Biology (IC50, Kd...)	Cheminformatics (Molecular Descriptors)				
Comp.1	Value1	D_1	D_2	D_3	|	D_n
Comp.2	Value2	"	"	"	|	"
Comp.3	Value3	"	"	"	|	"
Comp.N	ValueN	"	"	"	|	"

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\bar{y} - y_i)^2}$$

BA = F(D) {e.g., …}

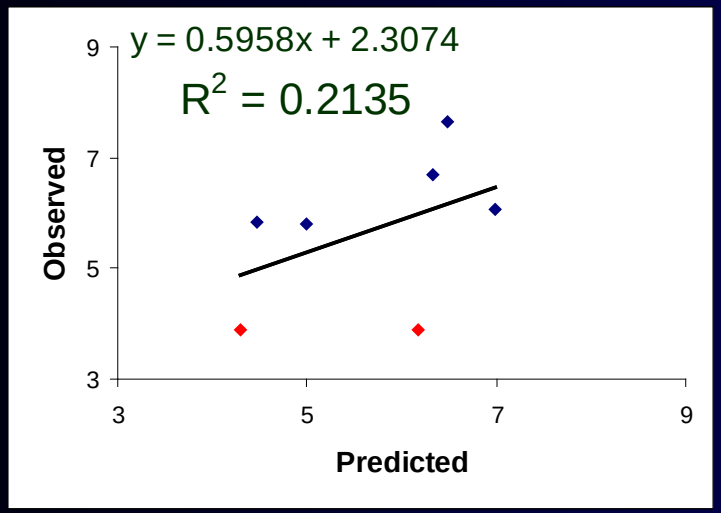
(e.g., -LogIC50 = $k_1D_1+k_2D_2+\ldots+k_nD_n$)

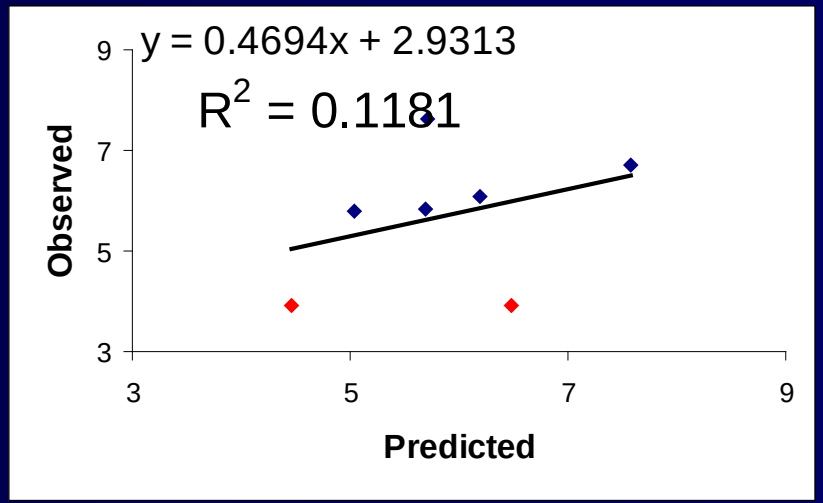
But … the unbearable lightness of model building for training sets…



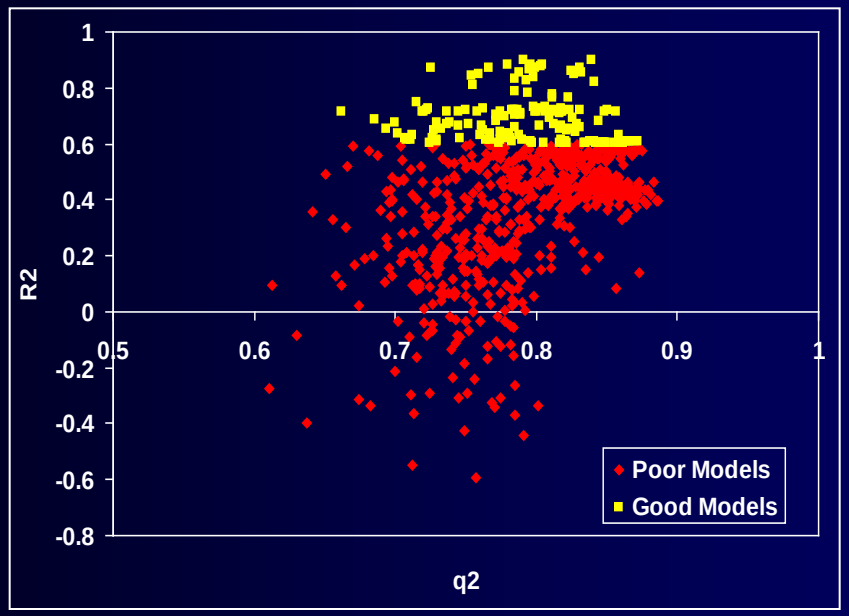
…leads to unacceptable prediction accuracy.

EXTERNAL TEST SET PREDICTIONS





BEWARE OF q^2 (Kubinyi paradox)!!!



•Only a small fraction of "predictive" training set models with LOO $q^2 > 0.6$ is capable of making accurate predictions ($r^2 > 0.6$) for the test sets.

Golbraikh & Tropsha, *J. Mol. Graphics Mod*. 2002, 20, 269-276.

Major components of QSAR modeling

- **Target properties (dependent variable)**
  - Continuous (e.g., IC50)
  - Categorical unrelated (e.g., different pharmacological classes)
  - Categorical related (e.g., subranges described as classes)
- **Descriptors (or independent variables)**
  - Continuous (allows distance based similarity)
  - Categorical related (allows distance based similarity)
  - Categorical unrelated (require special similarity metrics)
- **Correlation methods (with and w/o variable selection)**
  - Linear (e.g., LR, MLR, PCR, PLS)
  - Non-linear (e.g., kNN, RP, ANN, SVM)
- **Validation and prediction**
  - Internal (training set) vs. external (test set) vs. independent evaluation set
- **Examples of applications and pitfalls**

QSAR Pill

Complexity of QSAR modeling: Choices and Practices

- Descriptors (thousands and counting)
- Data-analytical methods (dozens and counting)
- Validation approaches (unfortunately (!) only a handful but counting)
- Experimental validation as part of model building (very rare)

BUT

- We typically use one (or at best very few) modeling techniques
- Publish successes only
- Compete but (mostly) indirectly

Why models may fail

- Incorrect data (structures and activities) in the dataset
- Modeling set is too small
- No external validation
- Incorrect selection of an external test set
- Incorrect division of a dataset into training and test sets
- Incorrect measure of prediction accuracy
- Insufficient statistical criteria to estimate predictive power of models
- Lack or incorrect definition of applicability domain
- No Y-randomization test (overfitness)
- Presence of leverage (structure) and activity outliers

Also, see Dearden JC, Cronin MT, Kaiser KL. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ Res. 2009;20(3-4):241-66

Some reasons why QSAR models may fail: using incorrect target function in classification QSAR for biased datasets:

QSAR Pill

- **A typical target function (Classification Rate)**:

  CR=N(classified correctly)/N(total)

  A dataset:

  Class 1: 80 compounds

  Class 2: 20 compounds

  Model: assign all compounds to Class 1.

  Target function: CR=0.8

  The model appears to have high classification accuracy

- **Better target function**:

  CCR=0.5(Sensitivity+Specificity)

  In the above example, CCR = 0.5

- **General formula**:

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

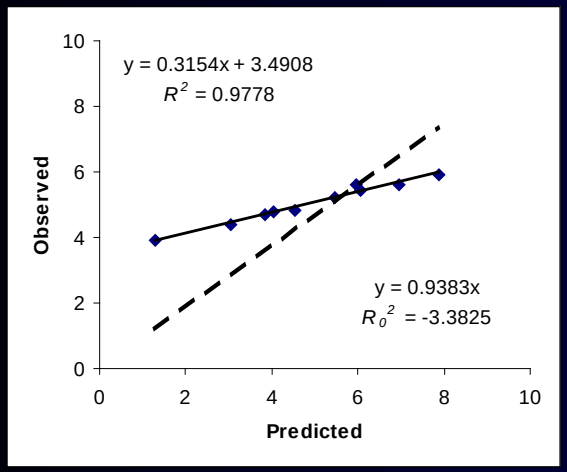
K – the number of classes

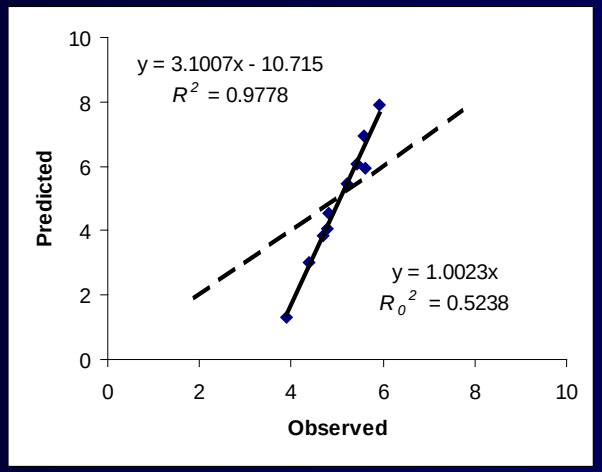
N_k^{corr} – the number of compounds of class k assigned to class k

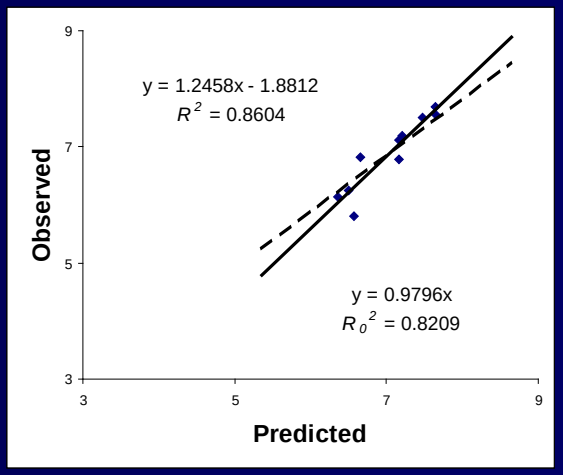
N_k^{total} – total number of compounds of class k

- **For categorical response variable, target functions can depend also on the absolute errors (differences between predicted and observed classes).**

HOW TO DEFINE A PREDICTIVE QSAR MODEL







Regression

$$\tilde{y}^{r} = a' y + b'$$

$$a' = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum (y_i - \bar{y})^2}$$

$$b' = \bar{\tilde{y}} - a' \bar{y}$$

Correlation coefficient

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}}$$

Regression through the origin

$$\tilde{y}^{r_0} = k' y$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}$$

Coefficients of determination

$$R_0^2 = 1 - \frac{\sum (\tilde{y}_i - y_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2}$$

$$R'^2_0 = 1 - \frac{\sum (y_i - \tilde{y}_i^{r_0})^2}{\sum (y_i - \bar{y})^2}$$

CRITERIA

$$q^2 > 0.5; R^2 > 0.6;$$

$$k \text{ or } k' \approx 1.0; R_0^2 \text{ or } R_0'^2 \approx R^2$$

Some reasons why QSAR models may fail: No Applicability Domain is defined for the Model

- **Compounds which are highly dissimilar from all compounds of the training set (according to the set of descriptors selected) cannot be predicted reliably**

  Lack of the AD:

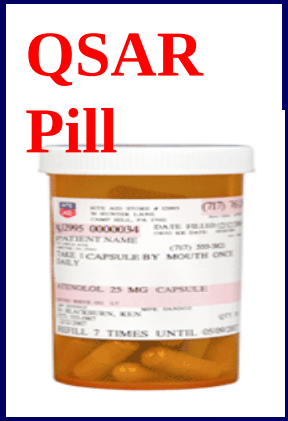
  unjustified extrapolation

  wrong prediction

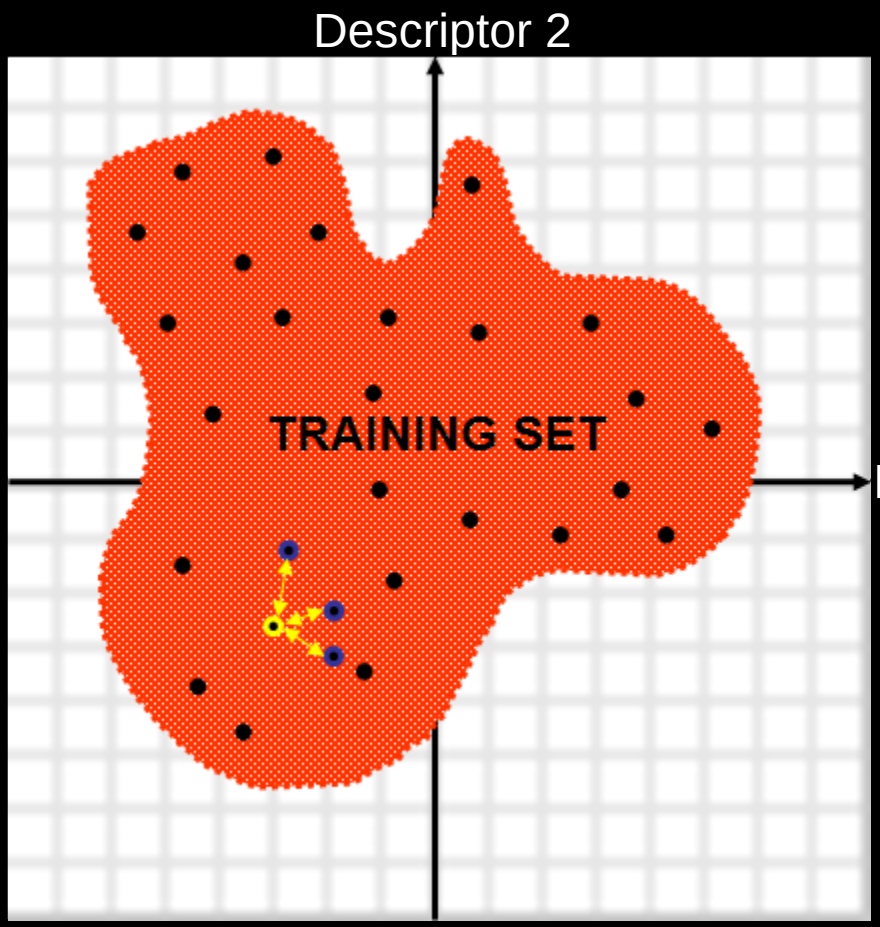
  Typical situation:

  a compound of the test set for which error of prediction is high is considered an outlier

  HOWEVER: a compound of the test set dissimilar from all compounds of the training set can be by chance predicted accurately



Applicability domain of QSAR models

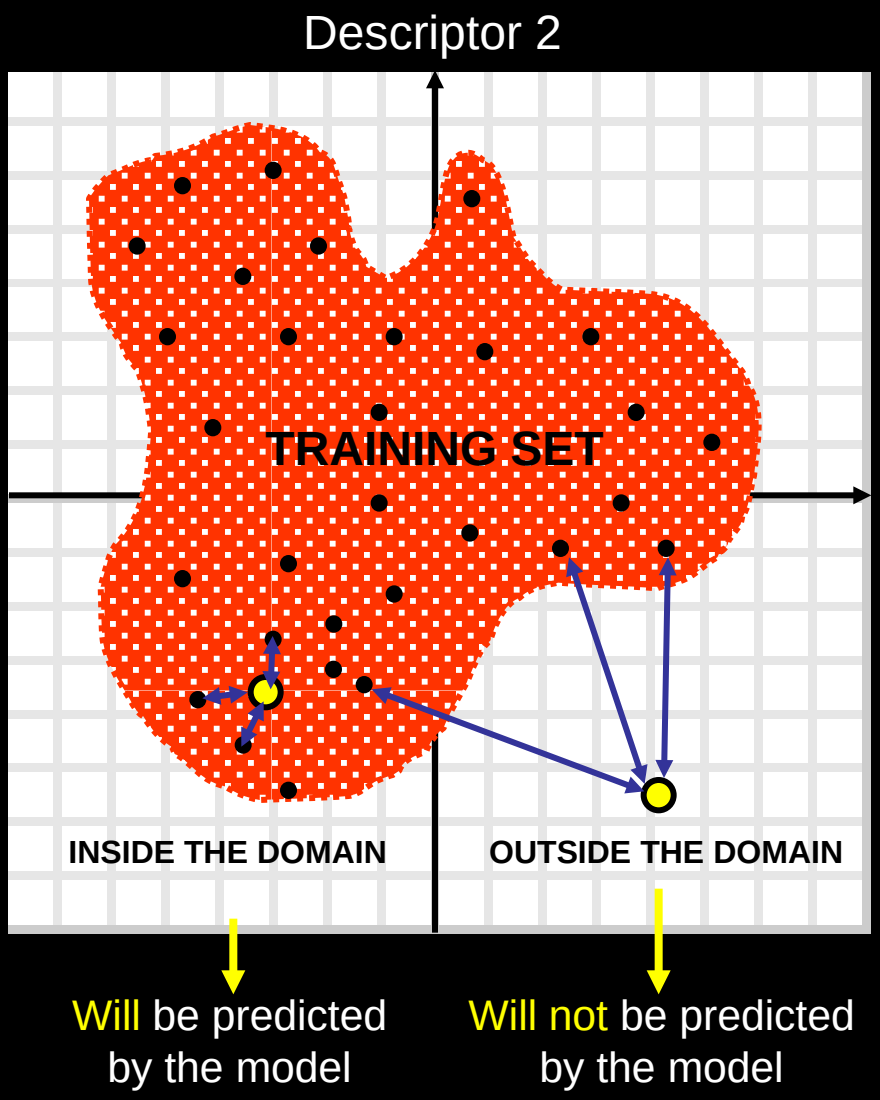


For a given model, two parameters are calculated:

- $\langle D_k \rangle$: average Euclidian distance between each compound of the training set and its *k* nearest neighbors in the descriptors space;

- s_k : standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.

Applicability domain of QSAR models



For a given model, two parameters are calculated:

- $\langle D_k \rangle$: average euclidian distance between each compound of the training set and its *k* nearest neighbors in the descriptors space;

- s_k : standard deviation of the distances between each compound of the training set and its k nearest neighbors in the descriptors space.

● = NEW COMPOUND

For each test compound *i*, the distance D_i is calculated as the average of the distances between i and its *k* nearest neighbors in the training set.

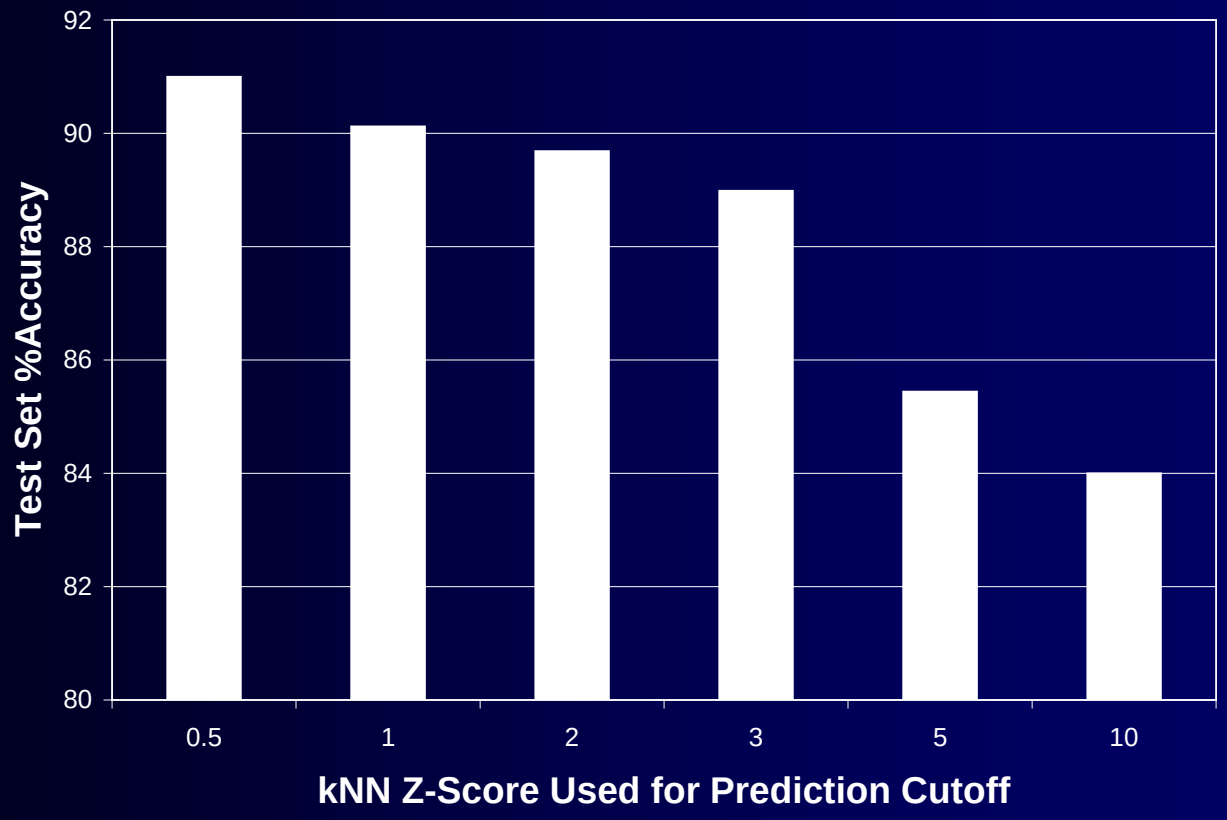
The new compound will be predicted by the model, only if :

$$D_i \leq \langle D_k \rangle + Z \times s_k$$

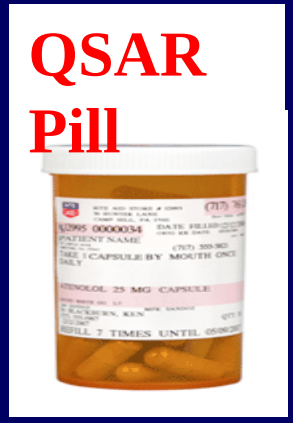
with Z, an empirical parameter (0.5 by default)

*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:…
Quant. Struct. Act. Relat. Comb. Sci. **2003**, 22, 69-77.

Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)



Some reasons why QSAR models may fail: Y-randomization test is not carried out



- **Y-randomization test**:
  - Scramble activities of the training set
  - Build models and get model statistics.
  - If statistics are comparable to those obtained for models built with real activities of the training set, the last are unreliable and should be discarded.

Frequently, Y-randomization test is not carried out.

Y-randomization test is of particular importance, if there is:

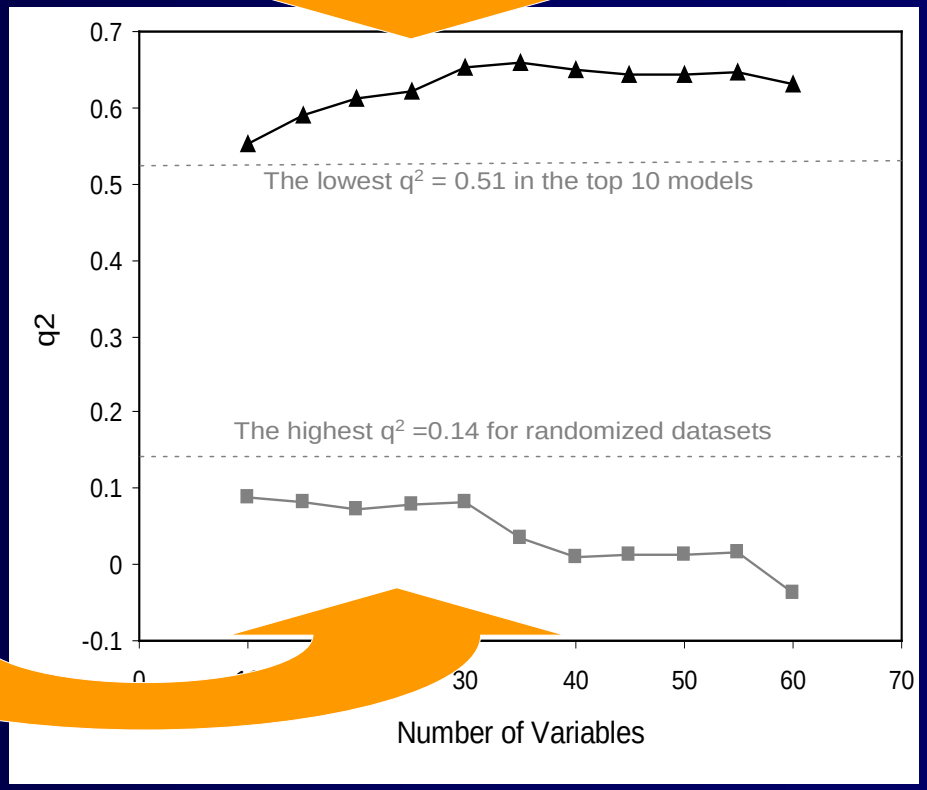
- a small number of compounds in the training or test set

- response variable is categorical

Activity randomization: model robustness

Struc.1 → Pro.1
Struc.2 → Pro.2
Struc.3 → Pro.3
.
.
Struc.n → Pro.n

Struc.1 Pro.1
Struc.2 Pro.2
Struc.3 Pro.3
. .
. .
Struc.n Pro.n



Training set with real property values is expected to produce much higher q^2 values than the same set with randomized property values.

Some reasons why QSAR models may fail: outliers

- Many potential outliers can be detected in the dataset prior to QSAR studies, but typically this is not done.

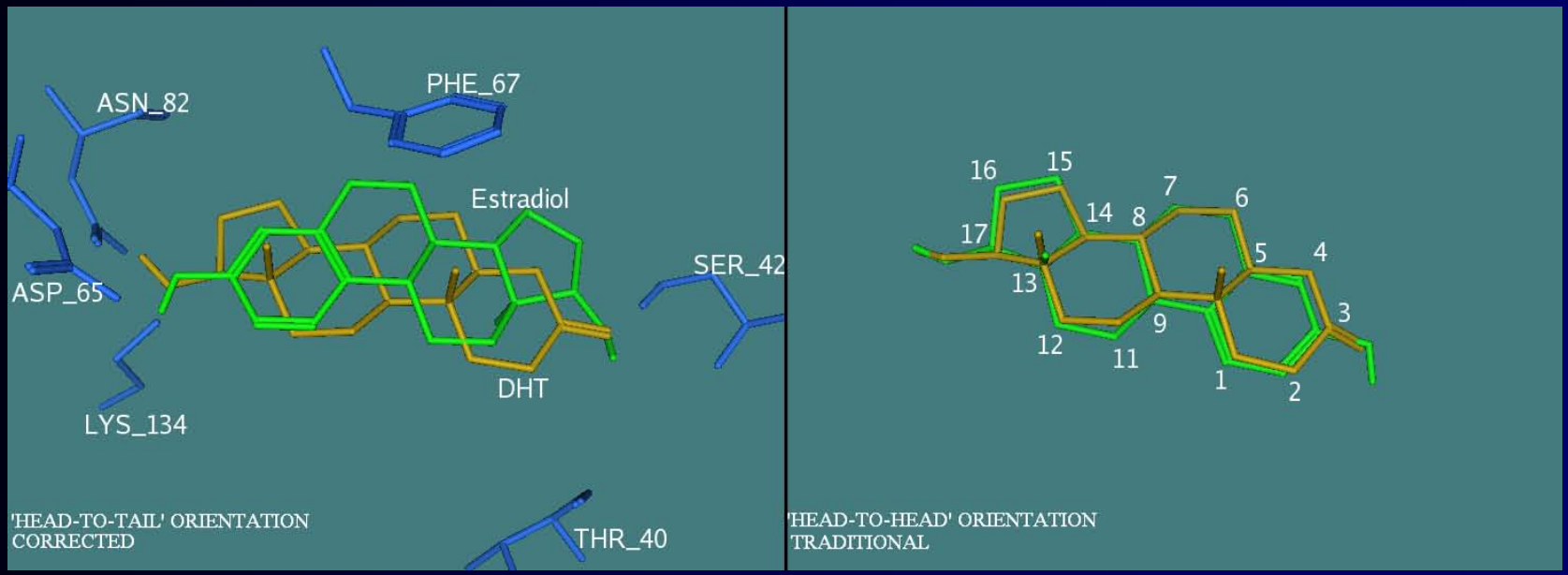
- **Two types of outliers**

  - **Leverage outliers:** compounds dissimilar from all other compounds in a dataset in the chemistry space.

  -**Activity outliers:** compounds similar to some other compounds in the dataset, but with activities quite different from those of their nearest neighbors (activity cliffs) .

Why QSAR models may fail: insensitive descriptors. [Example: Optimal (left panel) and traditional (right panel) orientations of androgen (DHT shown in gold) and estrogen (estradiol shown in green) within human SHBG steroid-binding site].

Identical q^2 (CoMFA*) of 0.53



*CoMFA – Completely Misleading Famous Aberration

A. Cherkasov, *JMC*, 2008

Full Papers

QSAR & Combinatorial Science QCS

Are the Chemical Structures in Your QSAR Correct?

Douglas Young[a]*, Todd Martin[a], Raghuraman Venkatapathy[b], and Paul Harten[a]

[a] US Environmental Protection Agency, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA; E-mail: young.douglas@epa.gov

[b] Pegasus Technical Services, 26 West Martin Luther King Drive, Cincinnati, OH 45268, USA

Keywords: Databases, *N*-octanol/water partition coefficient, Quantitative structure-activity relationships, SMILES

Received: June 26, 2008; Revised: August 13, 2008; Accepted: August 21, 2008

DOI: 10.1002/qsar.200810084 QSAR Comb. Sci. *27*, **2008**, No. 11-12, 1337 – 1345

- Recently, D.Young et al. pointed out the **importance of cleaning data, especially, in the context of QSAR modeling.**

- They investigated several public and commercial databases to calculate their **error rates**: the latter were ranging from **0.1 to 3.4%** depending on the database.

- Their main conclusions were that **small structural errors within dataset could lead to significant loss of predictive abilities for the QSAR models** which have been built using those erroneous input data.

Why can't we get it Right? Have not we tried enough?

- Descriptors? No, we have plenty (e.g., 1000's in Dragon)
- Datamining methods? No, we also have plenty (e.g., SAS)
- Training set statistics? NO, it does not work
- Test set statistics? Maybe, but it is still insufficient

So…what else can we do?????

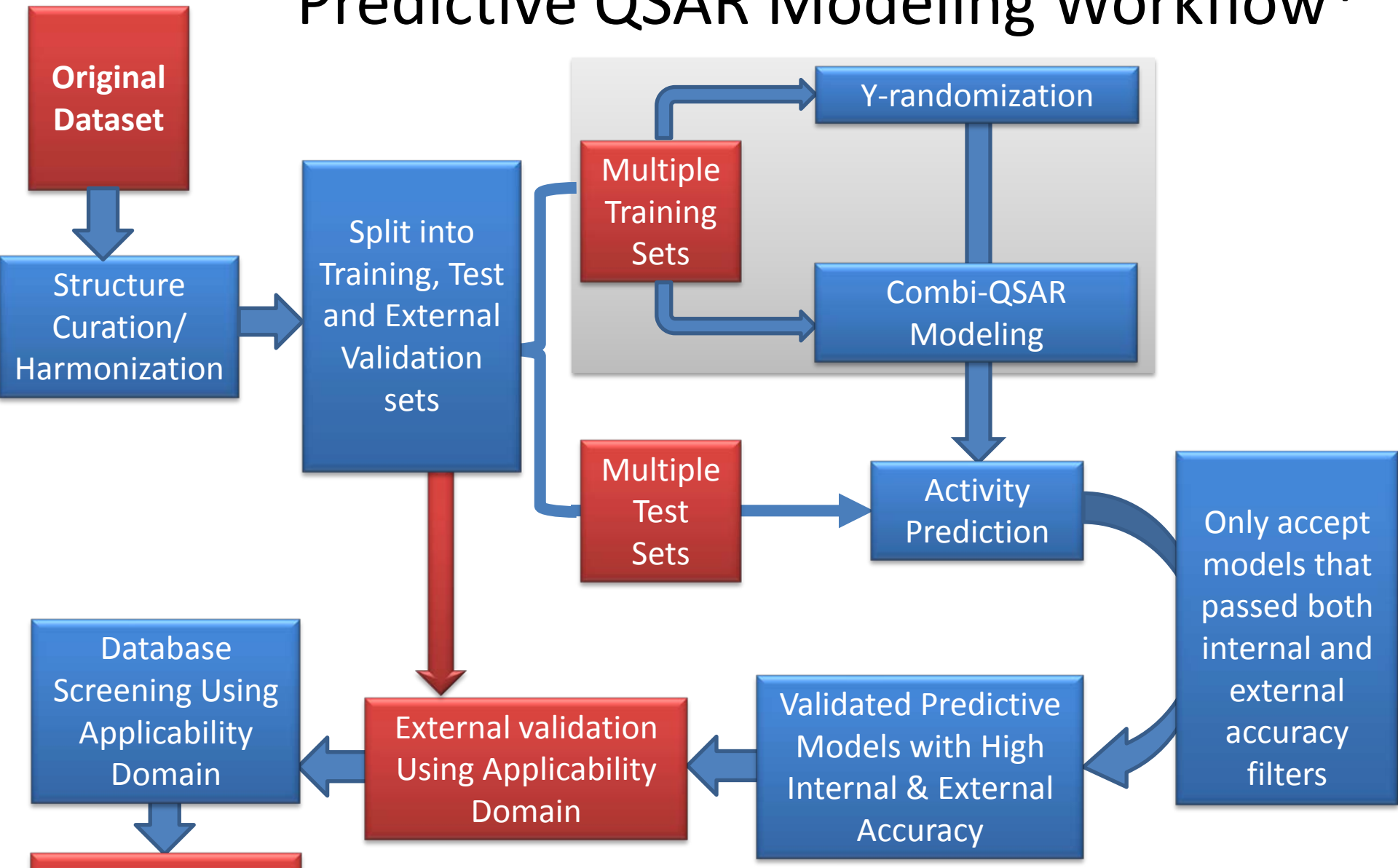
- Change the success criteria! Leave behind the phase of "narcissistic" modeling and focus on external predictivity and experimental validation.
- Recognize QSAR as an empirical data modeling approach: just do it any (all) way you like but VALIDATE on independent datasets!

Revising QSAR Modeling Process : Predictive QSAR Modeling Workflow*

- Model Building: Combination of various descriptor sets and variable selection data modeling methods (Combi-QSAR)
- Model Validation
  - Y-randomization
  - Training, test, AND evaluation set selection
  - Model sampling and selection criteria
  - Applicability domain
- Consensus prediction using multiple models

*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:… *Quant. Struct. Act. Relat. Comb. Sci*. **2003**, 22, 69-77;
Tropsha & Golbraikh, Curr. Pharm. Des., 2007, 13, 3494-3504

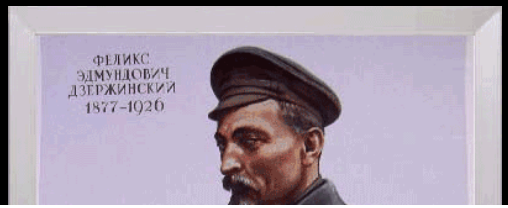
Predictive QSAR Modeling Workflow*



*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:… *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, 22, 69-77.
Tropsha & Golbraikh, Curr. Pharm. Des., 2007, 13, 3494-3504

The importance of data curation: What do these two men have in common?.

MML UNC.EDU



ARMS CONTROL ASSOCIATION

Search

Printer Friendly Page | E-mail to a Friend | Adjus

- Arms Control Today
  - Current Issue
  - Archived Issues
  - Archived Focus Editorials
  - Submissions and Letters
  - Archived Indexes
  - Advertising
  - Permission/Re-Print request
- Subject Resources
- Country Resources

Home

TRANSCRIPT - NEXT STEPS IN U.S.-RUSSIAN NUCLEAR ARMS REDUCTIONS: THE START FOLLOW-ON NEGOTIATIONS AND BEYOND

WELCOME AND MODERATOR:

DARYL KIMBALL,

EXECUTIVE DIRECTOR,

ARMS CONTROL ASSOCIATION

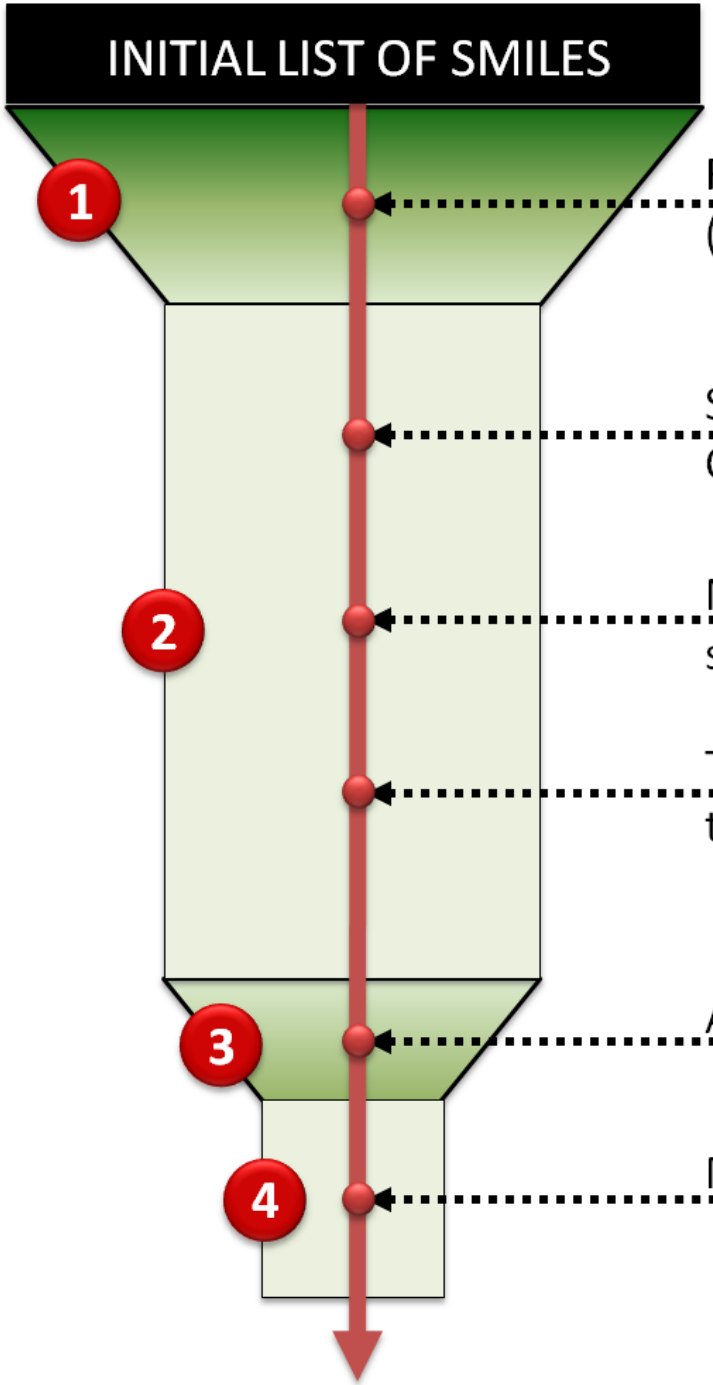
Latest ACA Resources

Russia

Russia Vetoes UN Mission in Georgia

did not think that some of the provisions were necessary because they were dealing with scenarios that were totally unrealistic.

But in order to satisfy the skeptics on the U.S. side that any agreement could be carried out, obviously, one settled for even more stringent verification measures than some people thought were necessary at the time in order to win order skeptics in agreement. We all remember, you know, trust but verify - that Russian proverb that Ronald Reagan and Felix Dzerzhinsky liked so much. But it wasn't trust but verify. It was we don't trust you, and therefore, we have to verify. And we have to verify very rigorously. That was the atmosphere at the time. And, you know, there are visages of that today.



SOFTWARE

Removal of mixtures, inorganics (and eventually organometallics)
ChemAxon - Standardizer
OpenEye - Filter

Structural conversion
Cleaning/removal of salts

Normalization of specific chemotypes

Treatment of tautomeric forms

ChemAxon - Standardizer
OpenBabel
Molecular Networks - CHECK,TAUTOMER

Analysis/removal of duplicates

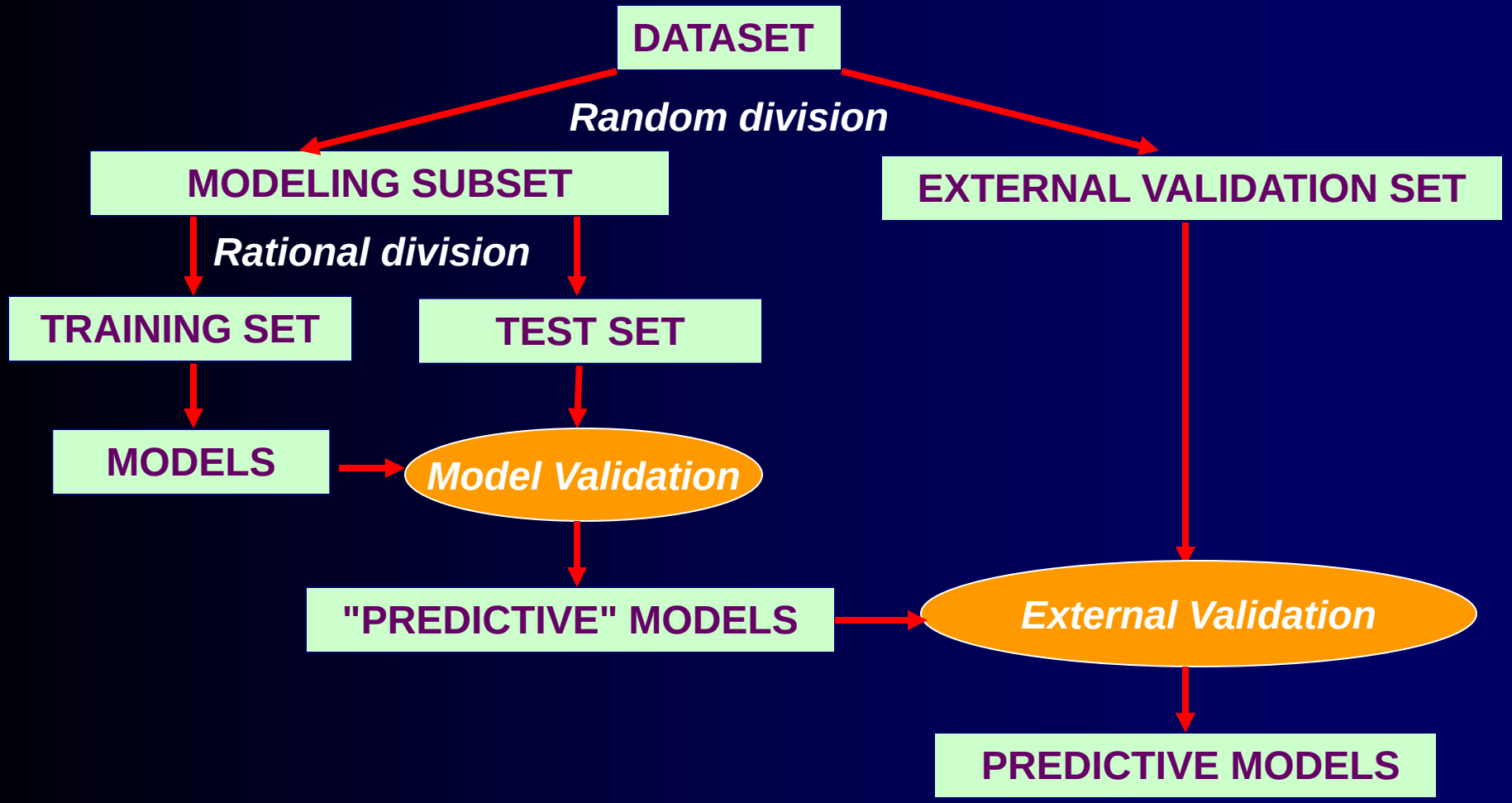
ISIDA - Duplicates
HiT QSAR
CCG - MOE

Manual inspection

ISIDA - EdiSDF
Hyleos - ChemFileBrowser
OpenBabel
ChemAxon - MarwinView

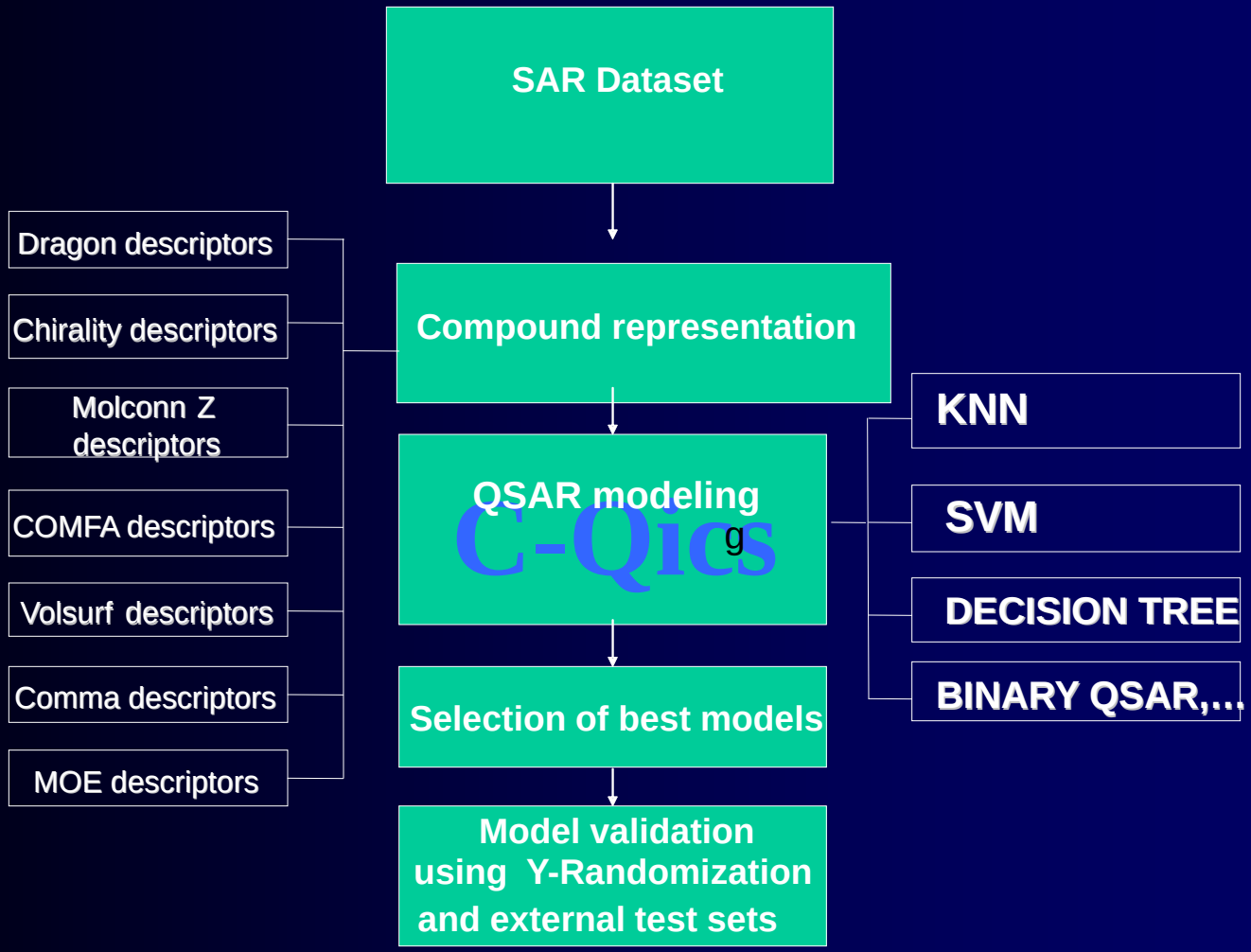
Muratov, Fourches, Tropsha. Trust but verify. JCIM, 2010, in press.

Division of the Dataset into Three Subsets and External Validation



Golbraikh et al., J. Comp. Aid. Mol. Design 2003, 17, 241–253.
Tropsha & Golbraikh, Curr. Pharm. Des., 2007, 13, 3494-3504

COMBINATORIAL QSAR



Lima, P., Golbraikh, A., Oloff, S., Xiao, Y., Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Info. Model.*, **2006** 46, 1245-1254.

Kovatcheva, A., Golbraikh, A., Oloff, S., Xiao, Y., Zheng, W., Wolschann, P., Buchbauer, G., Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J Chem. Inf. Comput. Sci*. **2004**, 44, 582-95

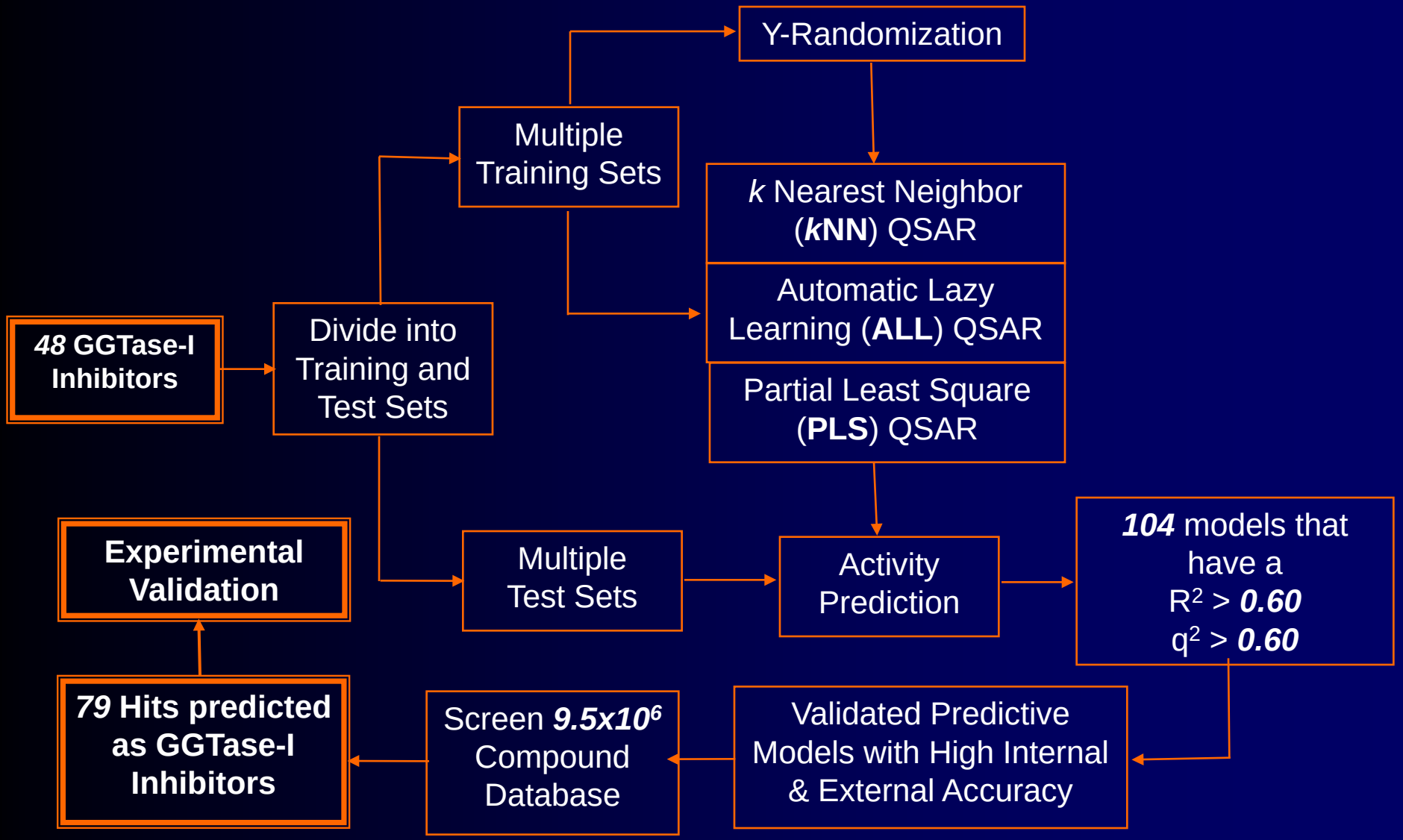
The OECD Principles of model validation

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm;
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation, if possible;

http://www.oecd.org/dataoecd/33/37/37849783.pdf

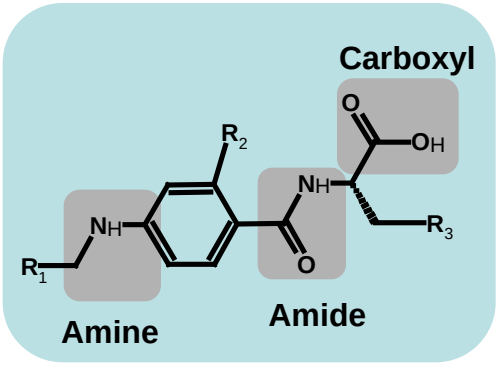
Application of Predictive QSAR Workflow to GGTase-I Inhibitors*



*Collaboration with P. Casey and Y. Peterson, Duke

Database Mining Reveals Unique Chemical Entities

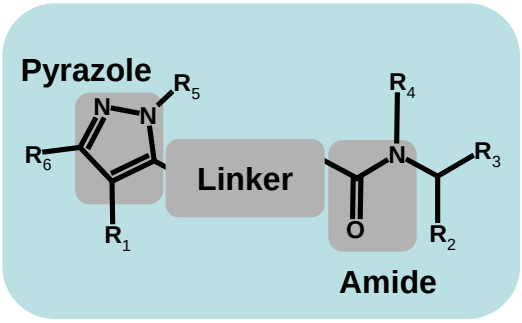
2 Training Set Scaffolds



GGTI 羊x Series

Peptidomimetics

Hamilton & Sebti

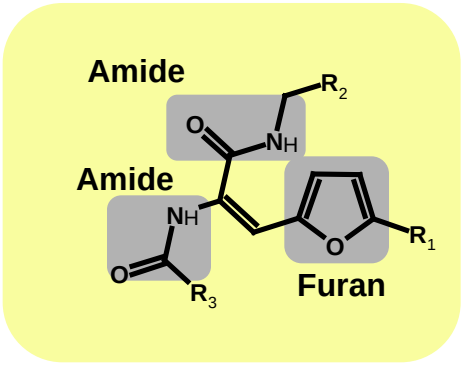


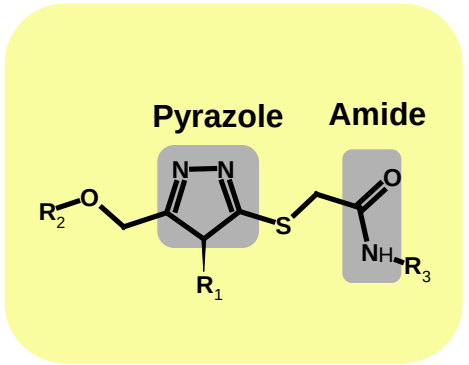
GGTI 羊DUx Series

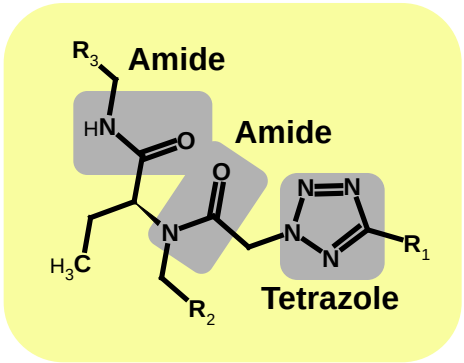
Pyrazoles

Peterson & Casey

Novel Scaffolds Discovered







Database Mining: Similarity Search vs. QSAR Search

A Large Commercial Database of 515,000 Compounds

Similarity Search

- ***Similarity Metric: Tanimoto Coefficient; of every single compound in the training set***
- ***Fingerprint: MACCS Structural Keys***
- ***425 hits obtained for TC=0.80; 2 hits obtained for TC=0.90***

QSAR Database Search

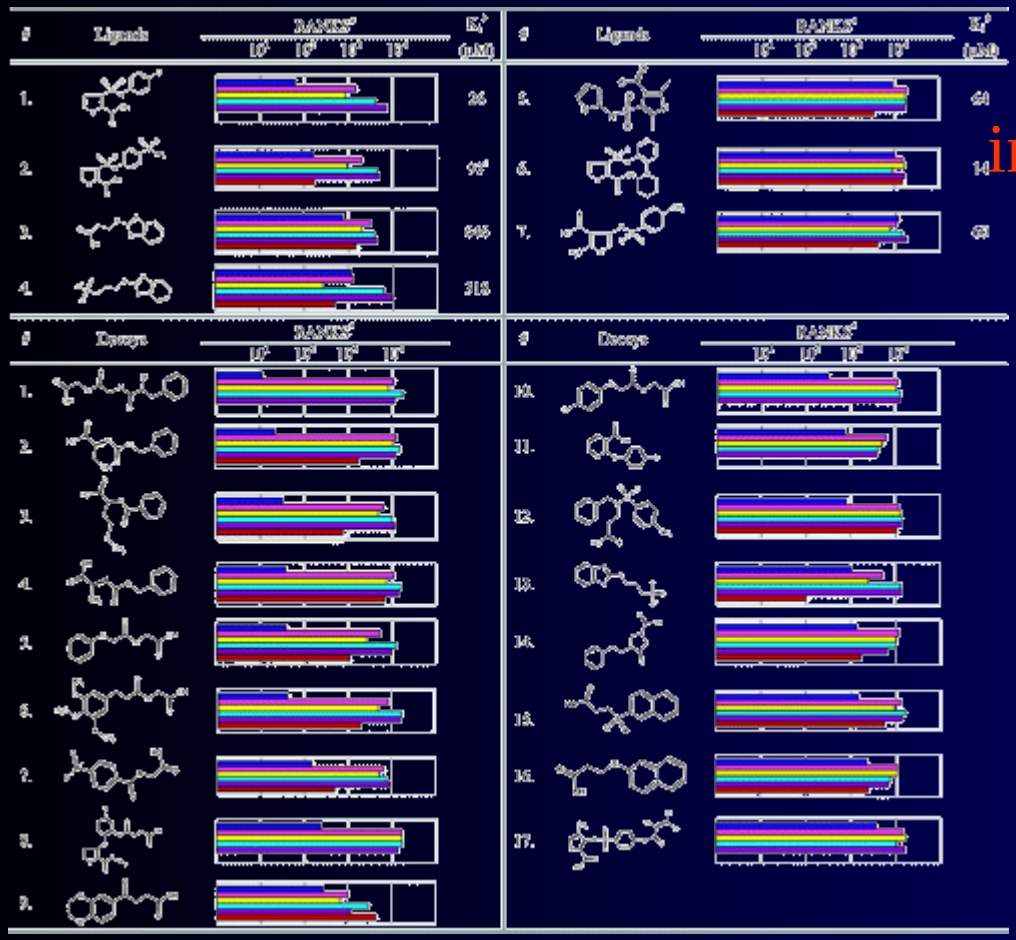
- ***Global search based on the whole chemical space (MZ 4.09 des.) of training set***
- ***12 hits obtained after global search (Z = 0.5) and subjected to consensus predictions***
- ***2 selected for experimental validation based on high predicted activity, uniqueness of structure & availability***

There was NO overlap between the hits from two protocols; All 12 QSAR hits were below TC=0.80 of training set.

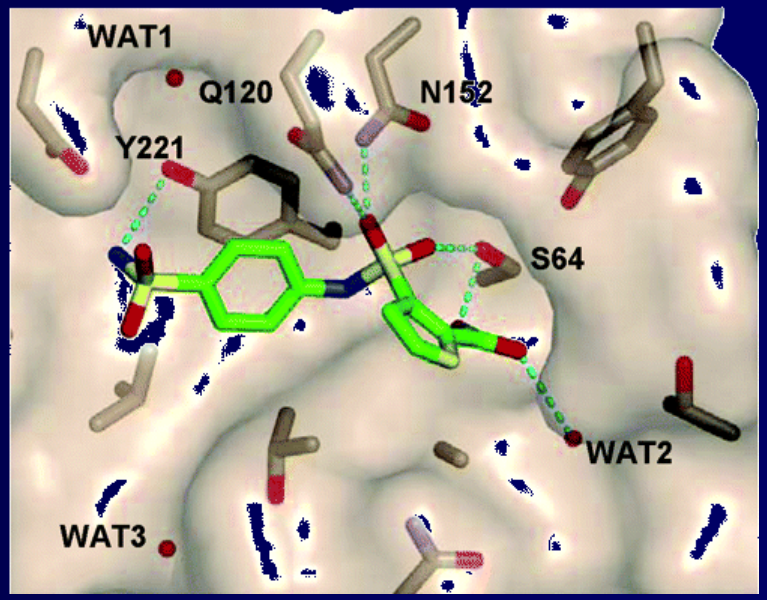
Recent examples of experimentally validated QSAR-based predictions

- Anticonvulsants: Shen, M. *et al*, *J. Med. Chem.* **2004**, 47, 2356-2364.
- HIV-1 reverse transcriptase inhibitors: Medina-Franco, J., *et al*, *J. Comput. Aided. Mol. Des*., **2005,** 19, 229–242
- D1 receptor antagonists: Oloff et al, *J. Med. Chem*., **2005,** 48, 7322-32
- Anticancer agents: Zhang *et al*, *J. Comp. Aid. Molec. Des*., **2007**, 21, 97-112.
- AmpC inhibitors: Hsieh, J.-H.. et al, *J. Comp. Aid. Molec. Des*., **2008,** 22(9):593-609
- HDAC inhibitors: Wang, S. *et al*, (JCIM, 2009, 49, 461-76)
- GGT-I inhibitors: Wang, Peterson, et al (JMC, 2009, 52(14):4210-20; provisional patent)
- 5HT7 binders: Zhao et al (in preparation)

QSAR vs. Docking: Application of QSAR Approaches to the Analysis of Binding Decoys



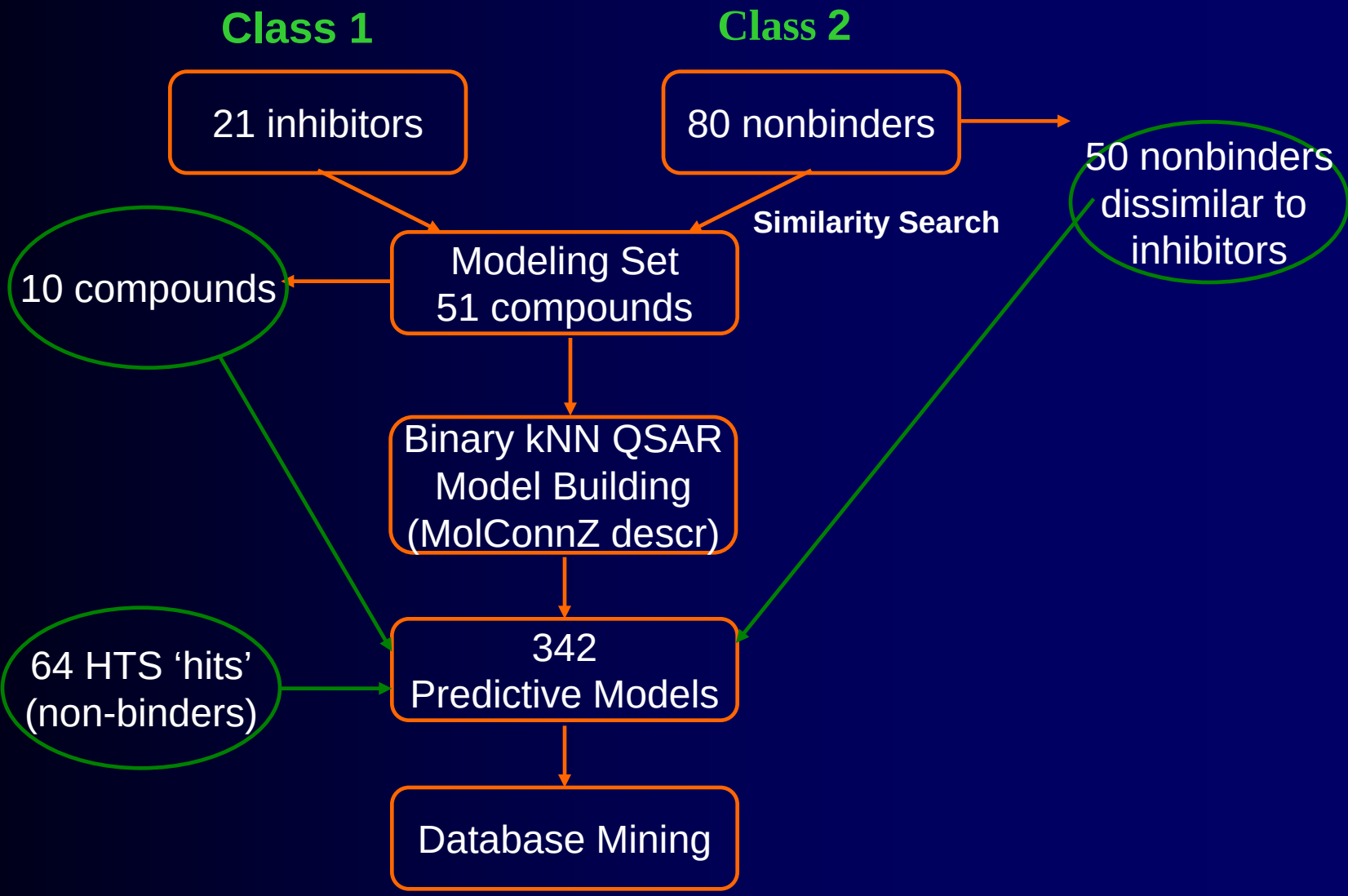
Decoys are frequently indistinguishable from binders using typical SB scoring functions.*



Characteristic AmpC Ligands and Decoys and Their Ranks by Different Scoring Functions. Blue = DOCK, magenta = ScreenScore, yellow = FlexX, cyan = PLP, purple = PMF, and red = SMoG (SMoG ranks are based on a ranking, which does not include halogenated compounds).

J. Med. Chem.* **2005, *48,* 3714-3728

Study Design (AmpC β-lactamase dataset)

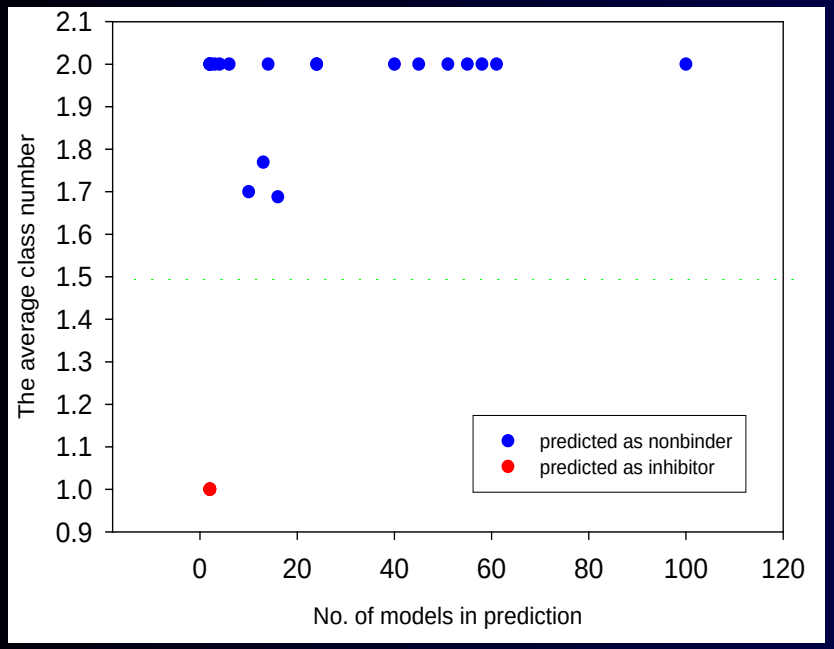


Correct Classification Rate (CCR) = 0.5 * (TP/N_1+TN/N_0)

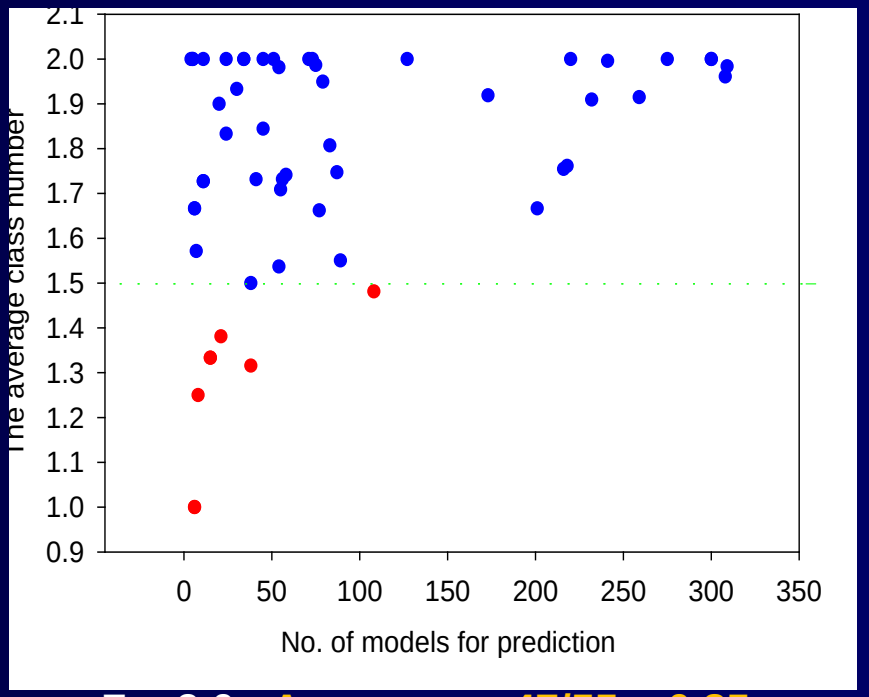
N_1: Total number of inhibitors
N_0: Total number of nonbinders

Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A
J Comput Aided Mol Des. 2008;22(9):593-609

The QSAR models do not predict the majority of the 64 HTS ‘Hits’ as binders in agreement with experimental study by Shoichet group



Z = 0.5; **Accuracy = 20/25 = 0.8**



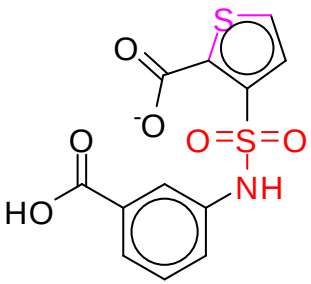
Z = 3.0; **Accuracy = 47/55 = 0.85**

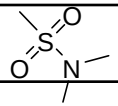
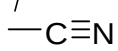
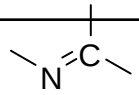
- assigned as nonbinder
- assigned as inhibitor

Babaoglu et al. J Med Chem. 2008 Apr 24;51(8):2502-11

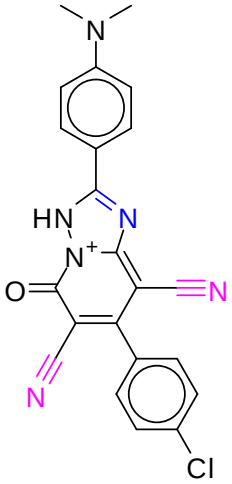
Descriptor Interpretation

inhibitor

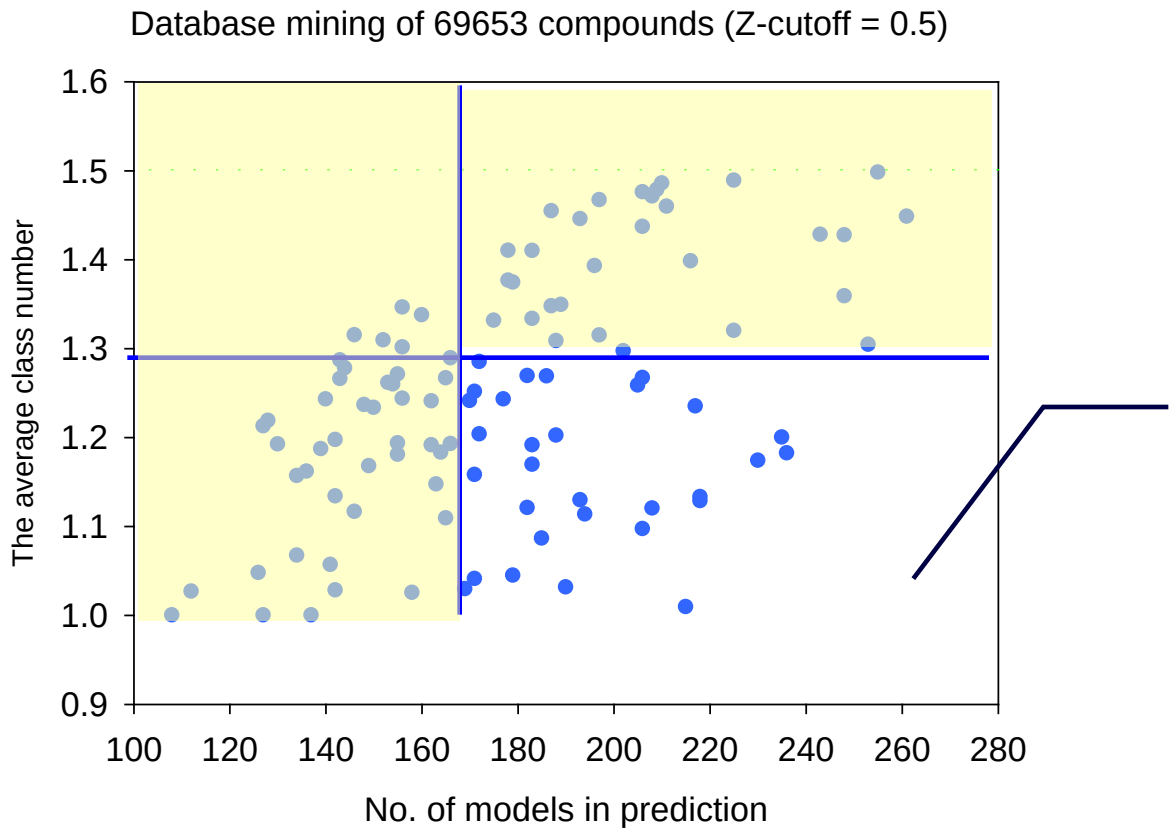


Rank	Descriptor ID	Frequency	Interpretation
1	**nHCsatu**	32.2	CHn (unsatuaed)
2	**Hsulfonamide**	28.4	
3	**nnitrile**	27.5	
4	**Hmin**	27.2	
5	**naaO**	26.3	:O:(aromatic)
6	**naaS**	26.3	:S: (aromatic)
7	**SaaCH**	26.0	:CH:
8	**n3Pad24**	26.0	
9	**SssCH2**	26.0	-CH_2-
10	**SHBint5**	25.4	
11	**Xvch5**	24.3	
12	**n2Pag23**	24.3	
13	**IDW**	24.0	
14	**htets2**	23.7	
15	**nimine**	23.7	

nonbinder



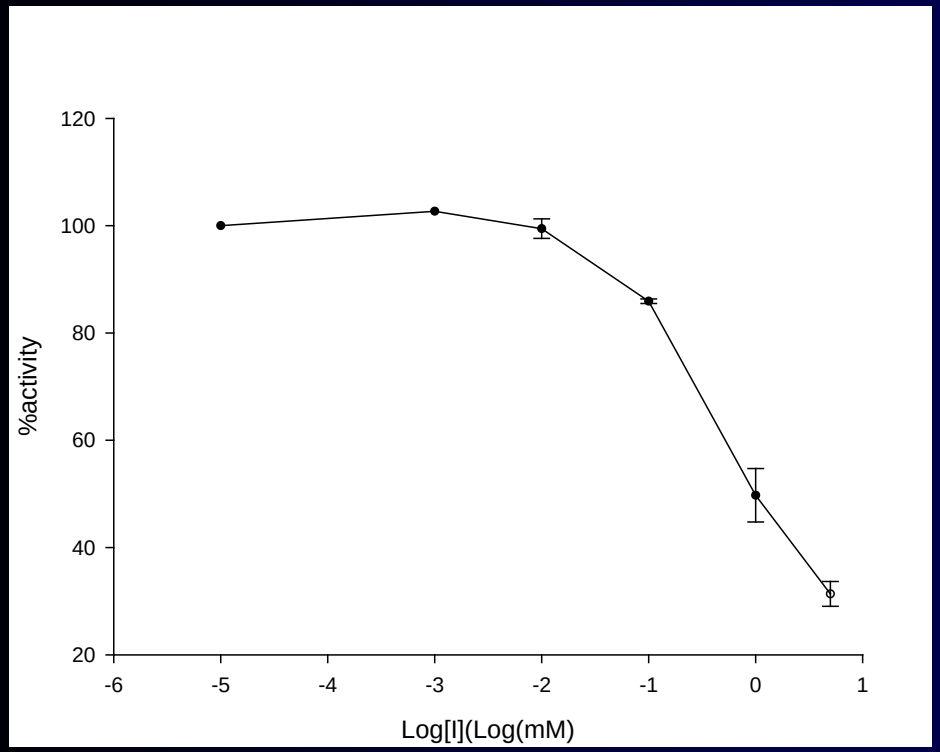
Virtual Screening of a PubChem AMPc HTS dataset of 69,653 Compounds



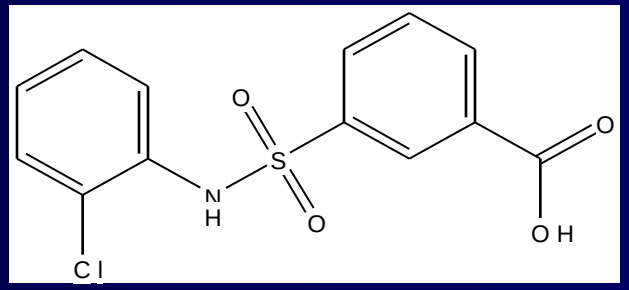
Hit selection criteria:
- Within AD of at least 50% of models
- 80% of those predict a compound as an inhibitor

This leads to 15 Hits

One “inactive” compound (CID 69951) Shows Micro-molar Inhibitory Activity



CID: 699751



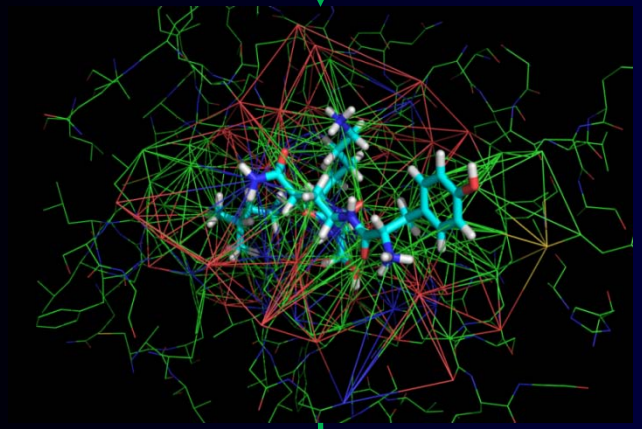
Kd = 18μM, Ki = 135μM

Experiments done by Dr. D. Teotico at UCSF

Pairwise potential (PPL) descriptors are applied to characterize protein-ligand interactions

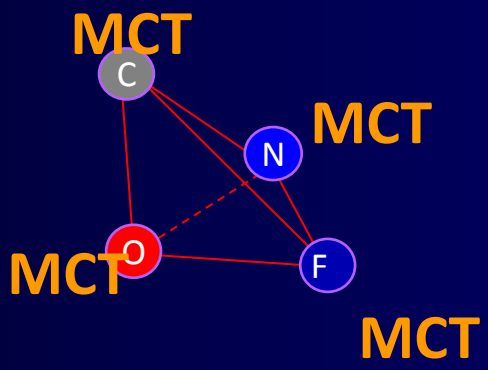
Protein-ligand interfaces

Delaunay Tessellation



$$PPL_m = \sum_{k=1}^{n} \sum_{p}^{1\sim3} \sum_{l}^{1\sim3} (MCT_p \times MCT_l / d_{pl})_k$$

1. **Delaunay tessellation of protein-ligand interface**



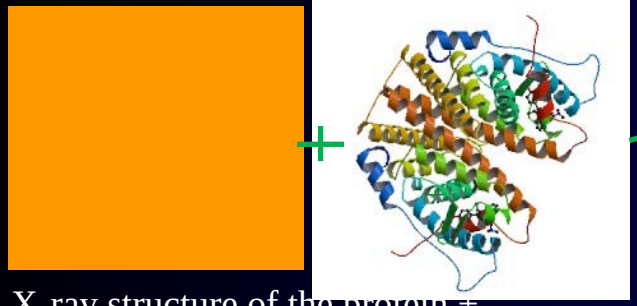
Each tetrahedron is categorized by
a) receptor/ligand atoms
b) Chemical atom type
In total, there are 554 theoretical descriptor types (m) [2].

2. For each atom at the protein-ligand interface, assign the maximal charge transfer (MCT) value calculated by Conceptual Density Functional theory DFT [1].

3. Each descriptor's value is the SUM of protein (p)-ligand (l)pairwise potential for the same tetrahedral type at the interface (n)

[1]: Parr RG, Szentpaly LV, Liu S J. Am. Chem. Soc. **1999**, 121, 1922
[2]: Zhang S. et al. J. Med. Chem., **2006**; 49(9); 2713-2724

Target-specific filter construction



X-ray structure of the protein + its bound native ligand

Docking (e.g. Fred) to generate ~1000 geometric poses

Generate pairwise potential descriptors for each pose

Classify the poses into native-like/decoy based on RMSD cutoff (4Å)

Training/test set

External validation set

Build binary classification models using k-NN or LIBSVM

Select acceptable models (Accuracy > 95%) for consensus prediction

Accuracy > 95%

A proposed screening protocol for structure-based virtual screening

A mix of ligands and decoys from DUD

Docking (e.g. Fred) and save multiple poses for each compound

Target-specific filter /multi-target filter

Exclude poses predicted as decoys

Re-ranking (e.g. MedusaScore*)

Performance evaluation

MedusaScore was developed in Dr. N. Dokholyan's lab in UNC
J. Chem. Inf. Model. ***2008, 48, 1656–1662***

Protein-ligand complex(es)

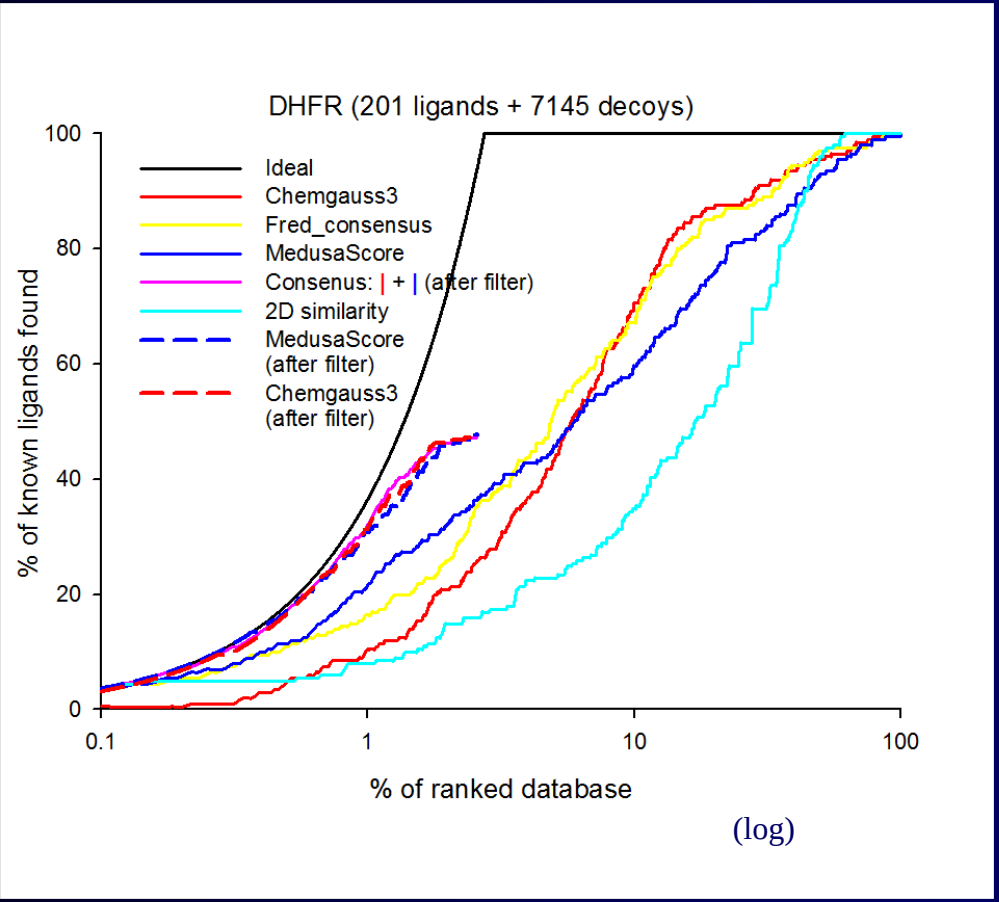
Generate geometric poses

Generate Pairwise potential descriptors

Build QSAR classification model

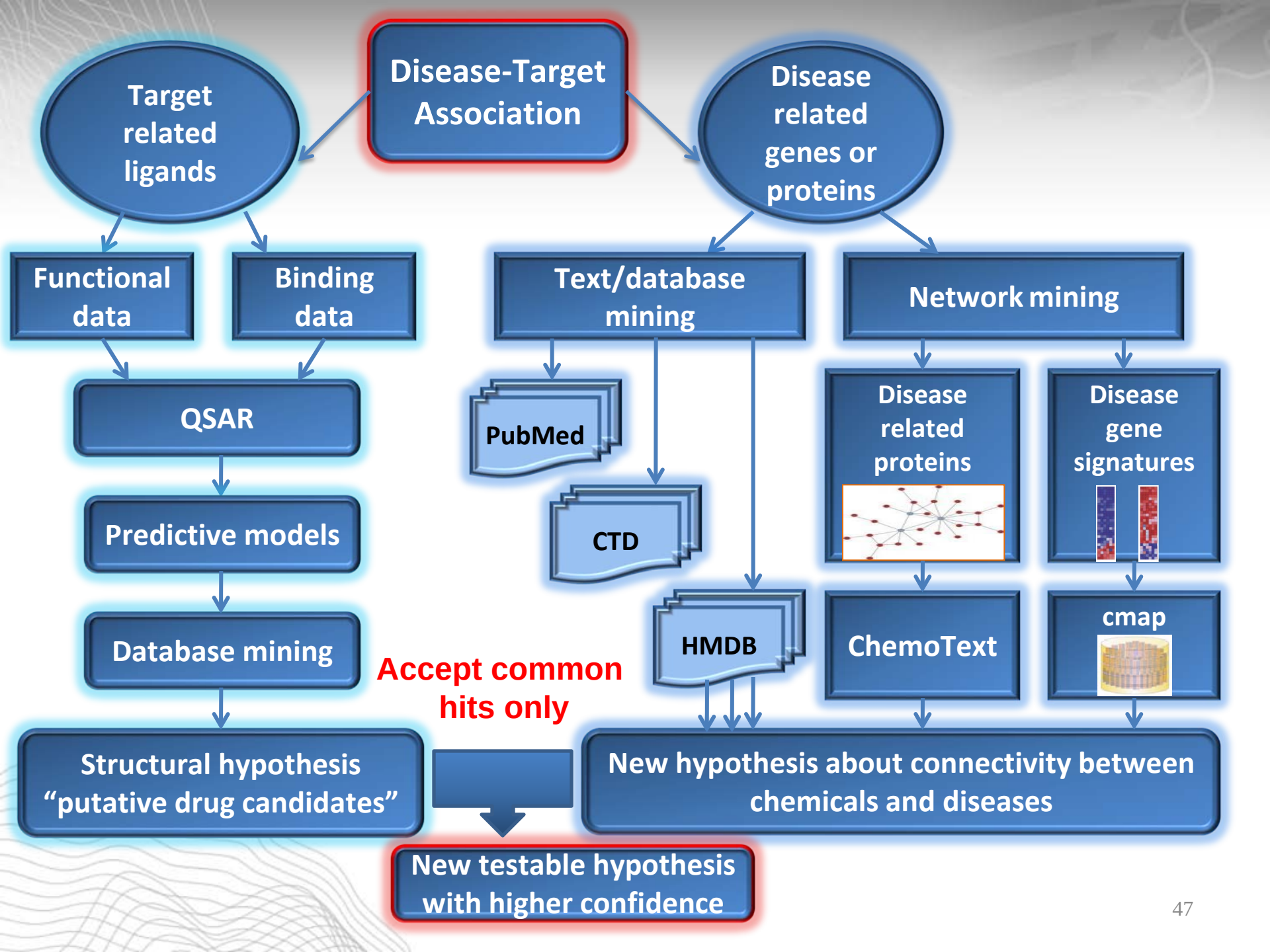
Select acceptable models

Docking enrichment plot for DHFR using DUD



Chemocentric Informatics: Integration of QSAR modeling with other approaches to drug discovery: structural hypothesis fusion.

Application to 5-HT_6 receptor linked to Alzheimer's disease

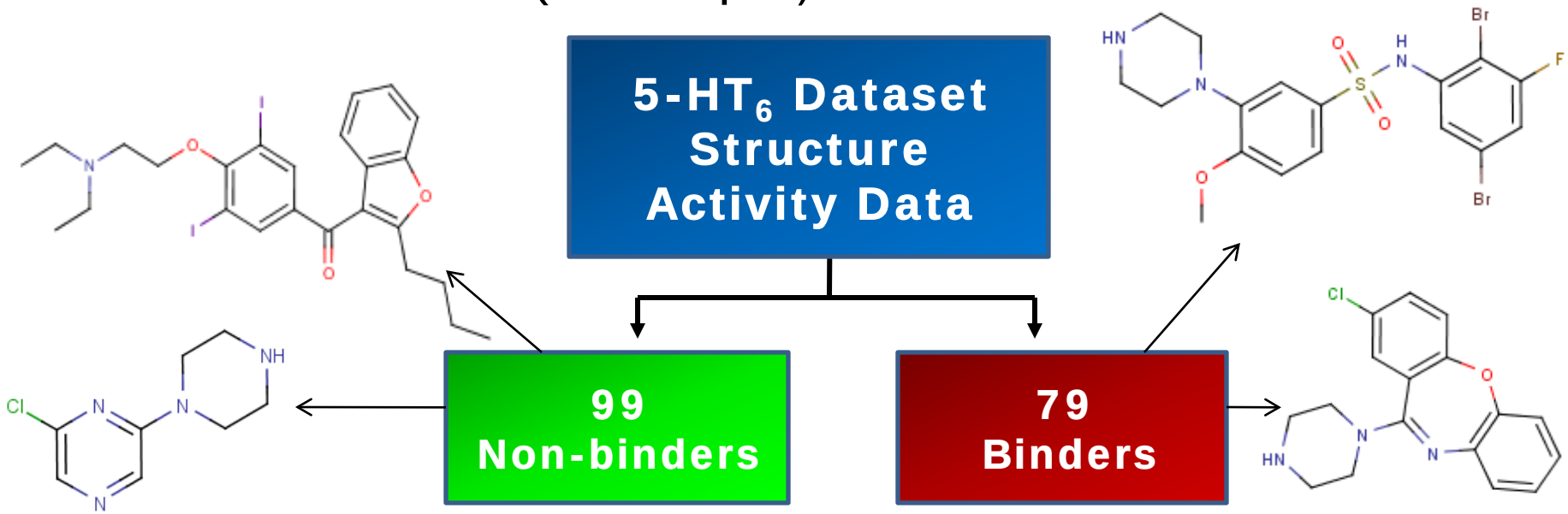

Disease-Target Association
Target related ligands
Disease related genes or proteins
Functional data
Binding data
Text/database mining
Network mining
QSAR
PubMed
Disease related proteins
Disease gene signatures
Predictive models
CTD
Database mining
HMDB
ChemoText
cmap
Accept common hits only
Structural hypothesis "putative drug candidates"
New hypothesis about connectivity between chemicals and diseases
New testable hypothesis with higher confidence

QSAR Modeling of 5-HT_6 Ligands

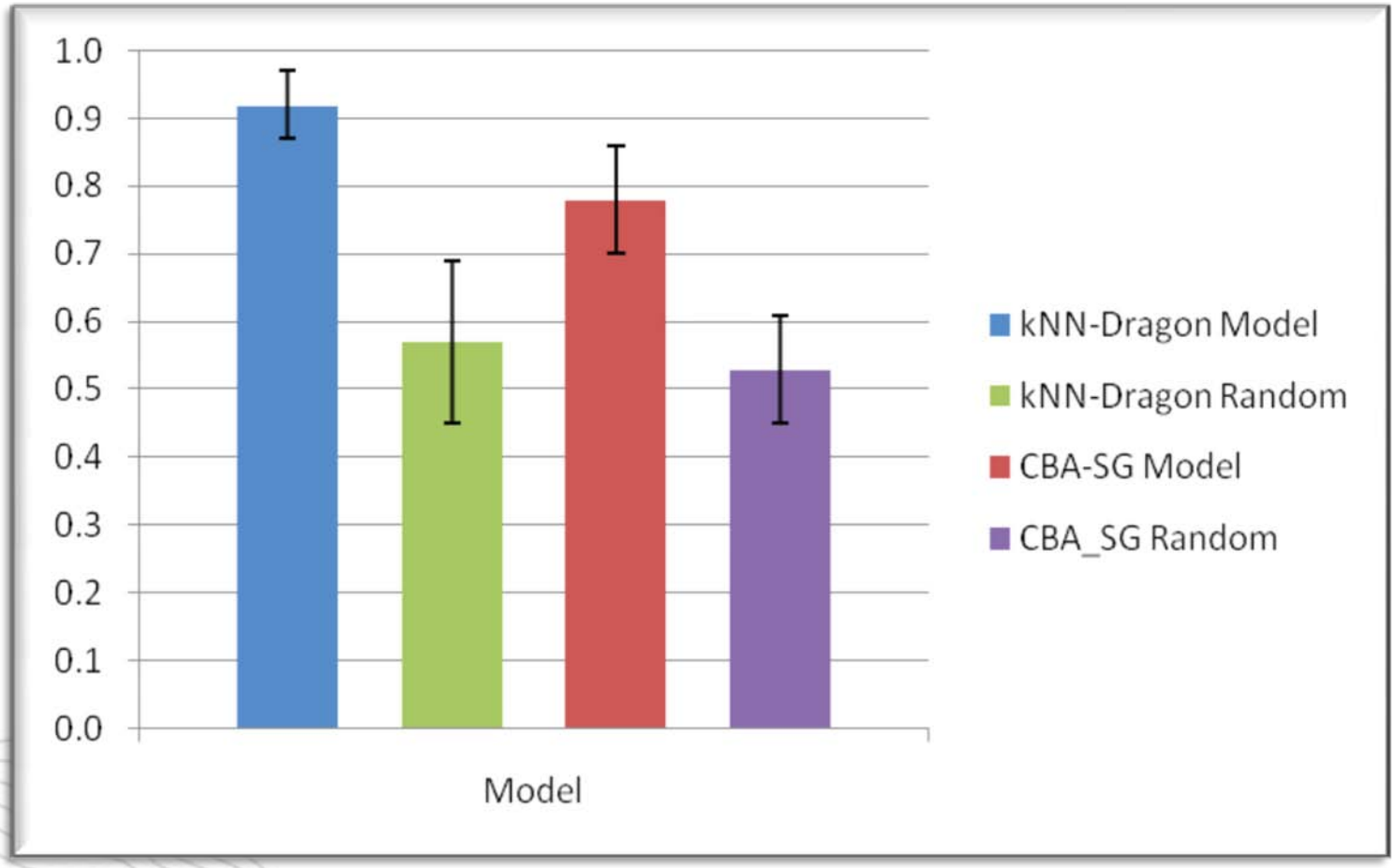
5-HT_6 Dataset:

Source: PDSP Ki Database

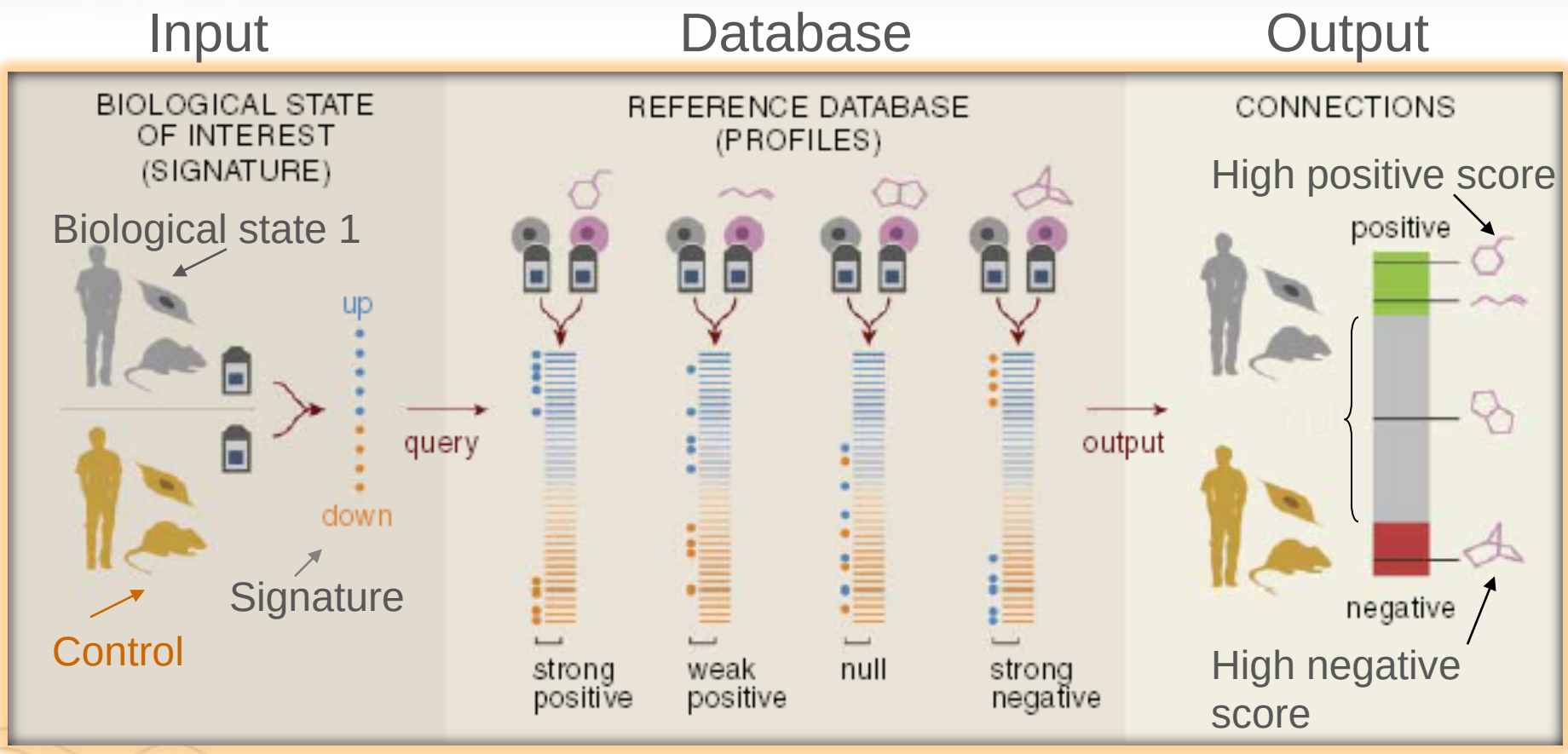
- 79 Binders (Ki < 10 μM),
- 99 Non-binders (Ki ≥ 10μM)



Comparison of the QSAR Approaches to Classify 5-HT_6 Receptor Ligands



The Connectivity Map



Step1: upload signature

Step2: query the cmap

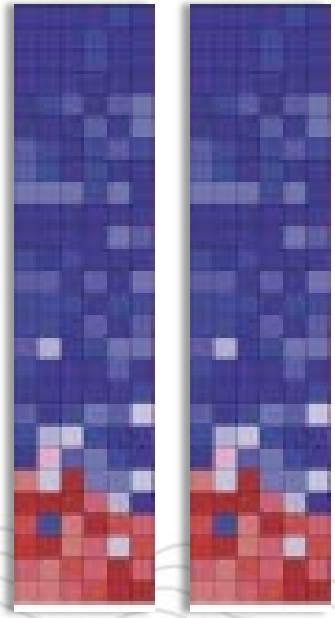
Step3 : list of correlated compounds

Lamb, J. *et al*. Science, 313, 1929-1935 (2006)

Querying the cmap with Alzheimer's Disease Gene Signatures

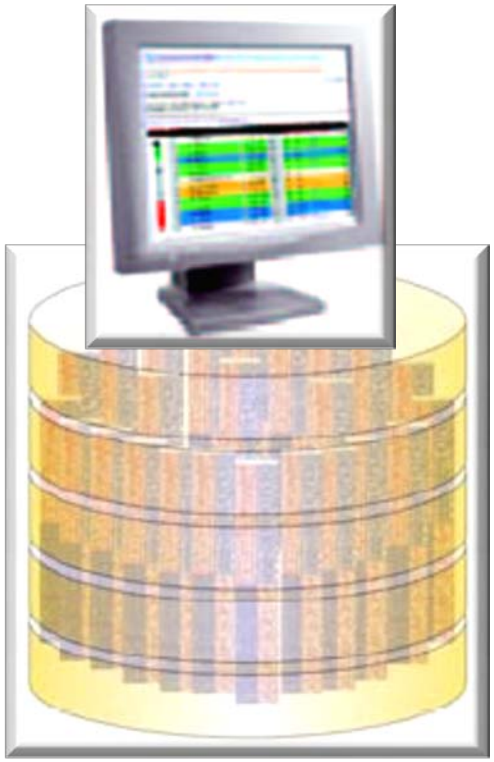
Upload signature

Alzheimer's disease gene signatures "**Two different signatures"** from hippocampus (**S1**) and cerebral cortex (**S2**) from two independent reports



(S1) (S2)

Query the cmap



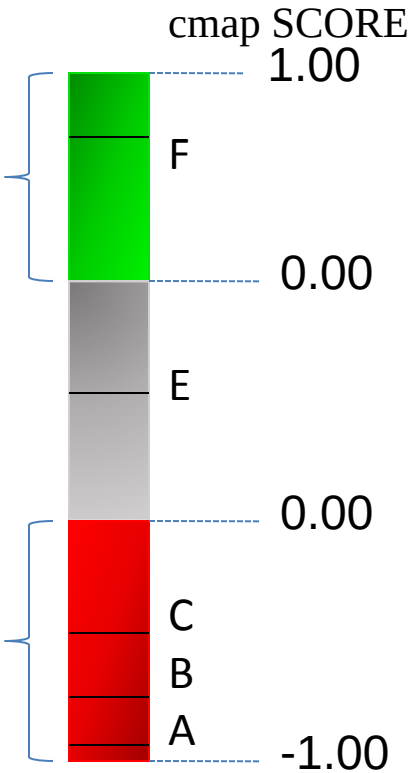
cmap

Signature database "Pattern Matching"

List of compounds

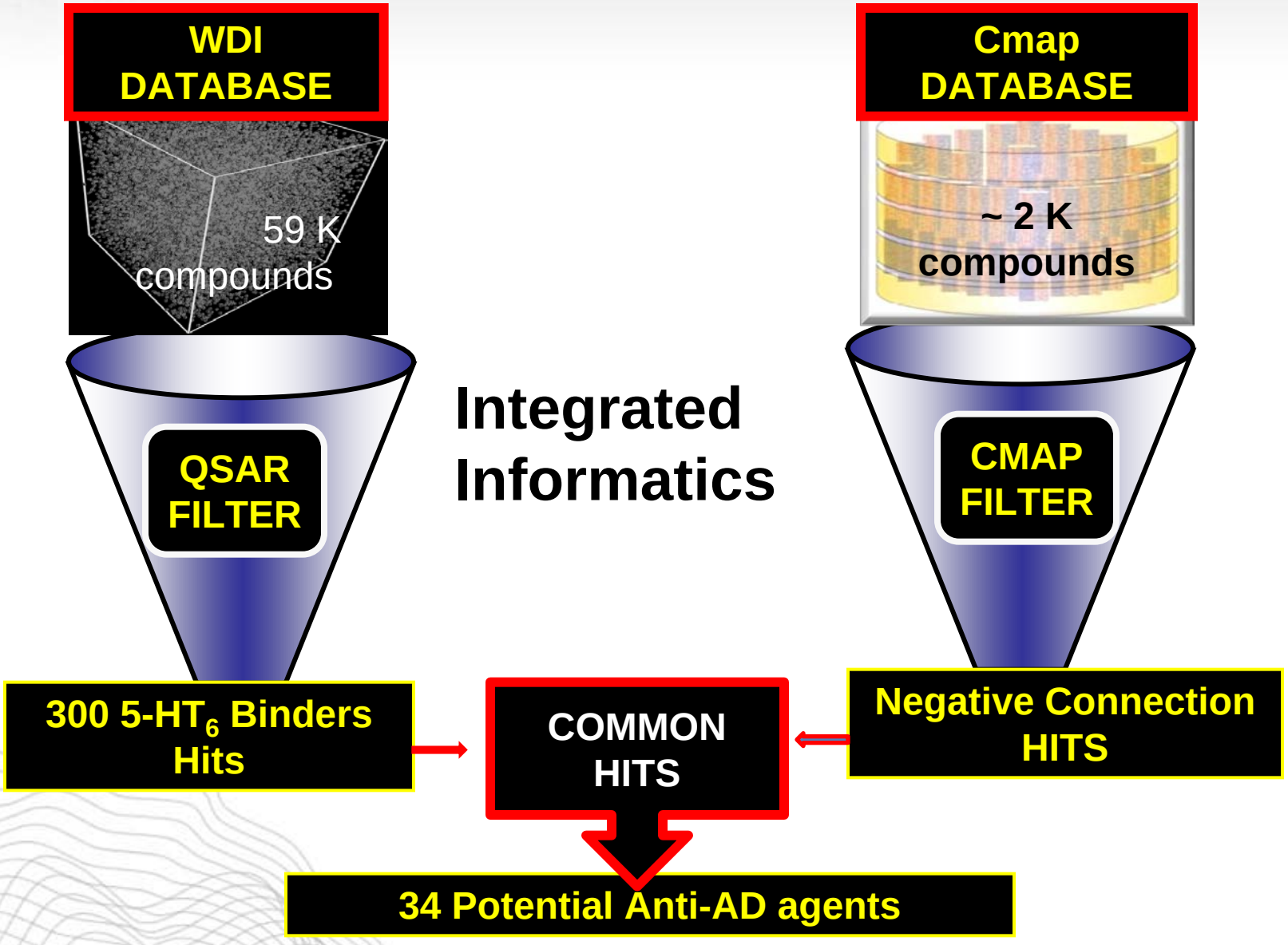
Positive Connectivity "possible causes for disease state"

Negative Connectivity "possible treatments for disease state"

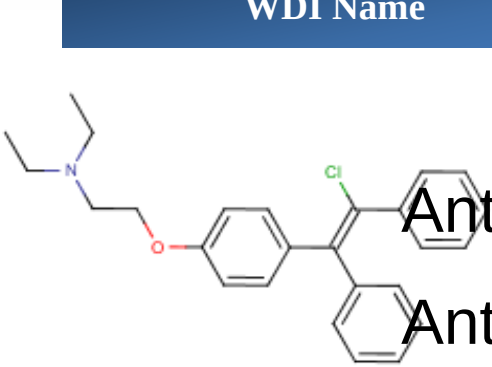


Identification of possible treatments (A,B,C) and causes (F)

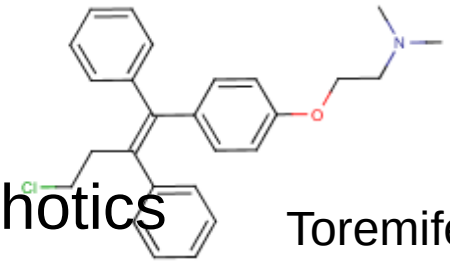
Virtual Screening Results



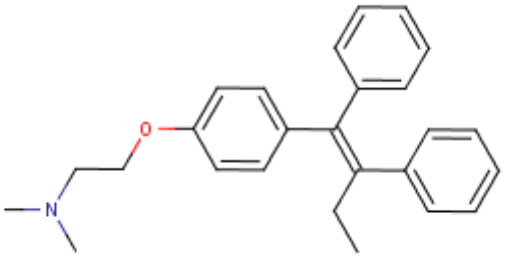
Selected Common Hits from QSAR and the cmap



Clomiphene



Toremifene



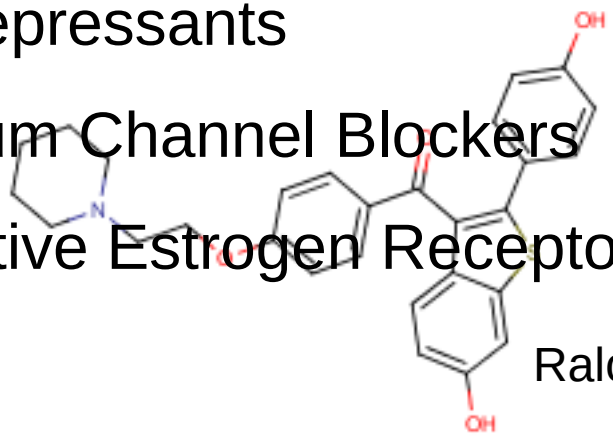
Tamoxifene

Antipsychotics

Antidepressants

Calcium Channel Blockers

Selective Estrogen Receptor Modulators



Raloxifene

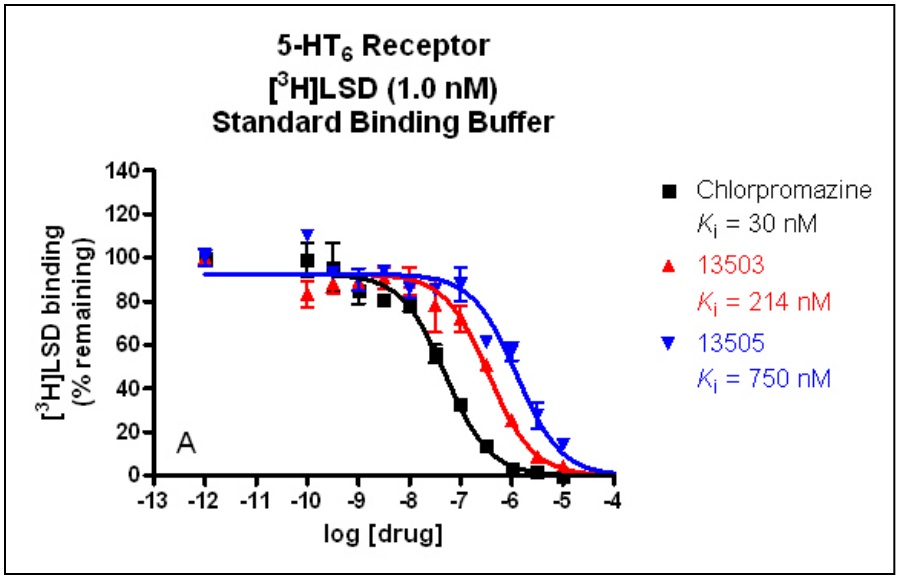
WDI Name		Pred.	cmap	cmap Score
BI-3			-0.768	-0.425
ENCLOMIFENE			-0.414	-0.611
DO-897			-0.741	-0.619
LY-294002			-0.351	-0.303
ACEFYLLINE-PRENYLA			-0.589	-0.457
NISOXETINE			-0.491	-0.408
IFENPRODIL			-0.541	-0.489
FENDILINE			-0.388	-0.683
NAFTIFINE			-0.790	-0.591
RALOXIFENE	miloxifene	809	0.64	-0.378	-0.482
MEBEVERINE	mebeverine	820	0.57	-0.543	-0.798
LOBELANIDINE	lobelanidine	826	0.56	-0.508	-0.488
LOBELINE	lobeline	882	0.55	-0.514	-0.750
AZACYCLONOL	azacyclonol	840	0.54	-0.448	-0.556

Selective Estrogen Receptor Modulators (SERMs) predicted as 5-HT$_6$ receptor ligands and potential therapeutics for AD: A power of the integrated chemogenomic approach

Raloxifene is a 5-HT_6 Binder and Potential Anti-Alzheimer's

A Power of the Integrated Chemogenomic Approach

- Raloxifene binds to 5-HT_6 receptor with a **K_i= 750 nM.**
- Raloxifene given at a dose of 120 mg/day led to reduced risk of cognitive impairment in post-menopausal women.
- Yaffe, K. *et al.,* ***Am J Psychiatry,*** **162, 683–690** (2005) .



Raloxifene (blue triangle) and Chlorpromazine (square) versus [3H] LSD competition binding at 5-HT_6 receptors. Tested by our collaborators at PDSP.

Raloxifene is predicted to bind several receptor families using QSAR-based VS

Classification models used prospectively to predict raloxifene's promiscuity

Receptor family	Type of model	Number of models	Average score	Total number of models	CCR_models (tr, ts &ex)
5-HT	Classification	512	0.57	650	≥0.7
Alpha 2	Classification	1686	0.76	2045	≥0.9
Dopamine	Classification	350	0.70	482	≥0.7
Muscarinic	Classification	444	0.66	500	≥0.7
Sigma	Classification	730	0.69	898	≥0.7

Regression models used retrospectively to predict raloxifene's binding affinity

Receptor	Type of models	Number of models	plogKi	SD	Ki
Alpha 2A	Regression	-	-	-	-
Alpha 2B	Regression	25	6.2	0.00	631
Alpha 2C	Regression	1	6.8	0.42	158

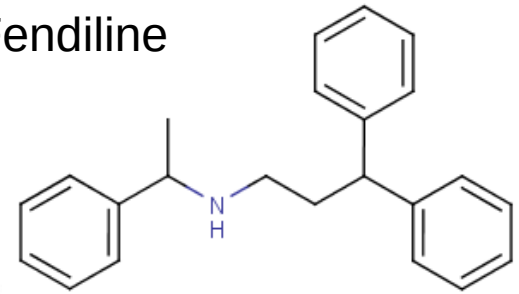
Comprehensive screening results

CMPD	PI	Tier	5ht1a	5ht1b	5ht1d	5ht1e	5ht2a	5ht2b	5ht2c	5ht3	5ht4	5ht5a	5ht6	5ht7
13505	Hajjo	Raloxifene	2,330.00	624	1,222.00	1,868.00	1,049.00	69	1,642.00		5,050.00	1,219.00	750	1,220.00
14821	Hajjo	Fendiline			3,550.00			3,217.00				894		3,085.00
678	PDSP (MUL)	Tamoxifen	3,477.00	1,618.00			2,596.00		4,282.00			2,123.00	931.1	1,077.00
10572	PDSP	Tamoxifen	>10,000	7,857.00			2,720.00	1,952.00	5,787.00			7,821.00	1,698.00	>10,000

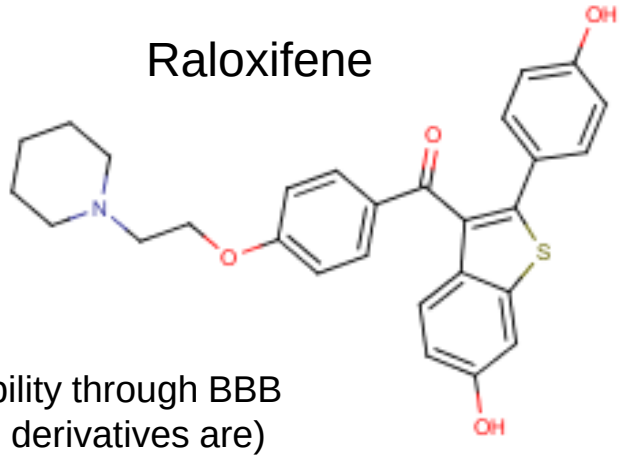
Alpha1A	Alpha1B	Alpha1D	Alpha2A	Alpha2B	Alpha2C	Beta3	D1	D2	D3	D4	D5	DAT	DOR	EP4	GabaA
247.7	534.6	478.2	1,288.20		61	1,104.00	1,626.00	683	>10,000	3,023.00	3,803.00	928	3,158.00		>10,000
3,056.00			592.1	873.3	768.4			9,655.00				6,881.00			
							1,508.00	1,682.00	498	7,817.00	>10000	4,328.00			
			1,211.00				657	5,517.00	1,740.00	>10,000		2,820.00			

H1	H2	H3	H4	KOR	M1	M2	M3	M4	M5	MOR	NET	SERT	Sigma 1	Sigma 2
5,356.00	1,436.00		7,072.00	186	>10,000	2,037.00		1,229.00	8,127.00		1,026.00	5,761.00	247.8	
2,295.00	993				9,603.00	9,073.00	3,201.00	327.8	2,278.00	1,455.00	3,981.00		227.3	471.9
											720			
	1,980.00										5,083.00		481	331

Fendiline



Raloxifene

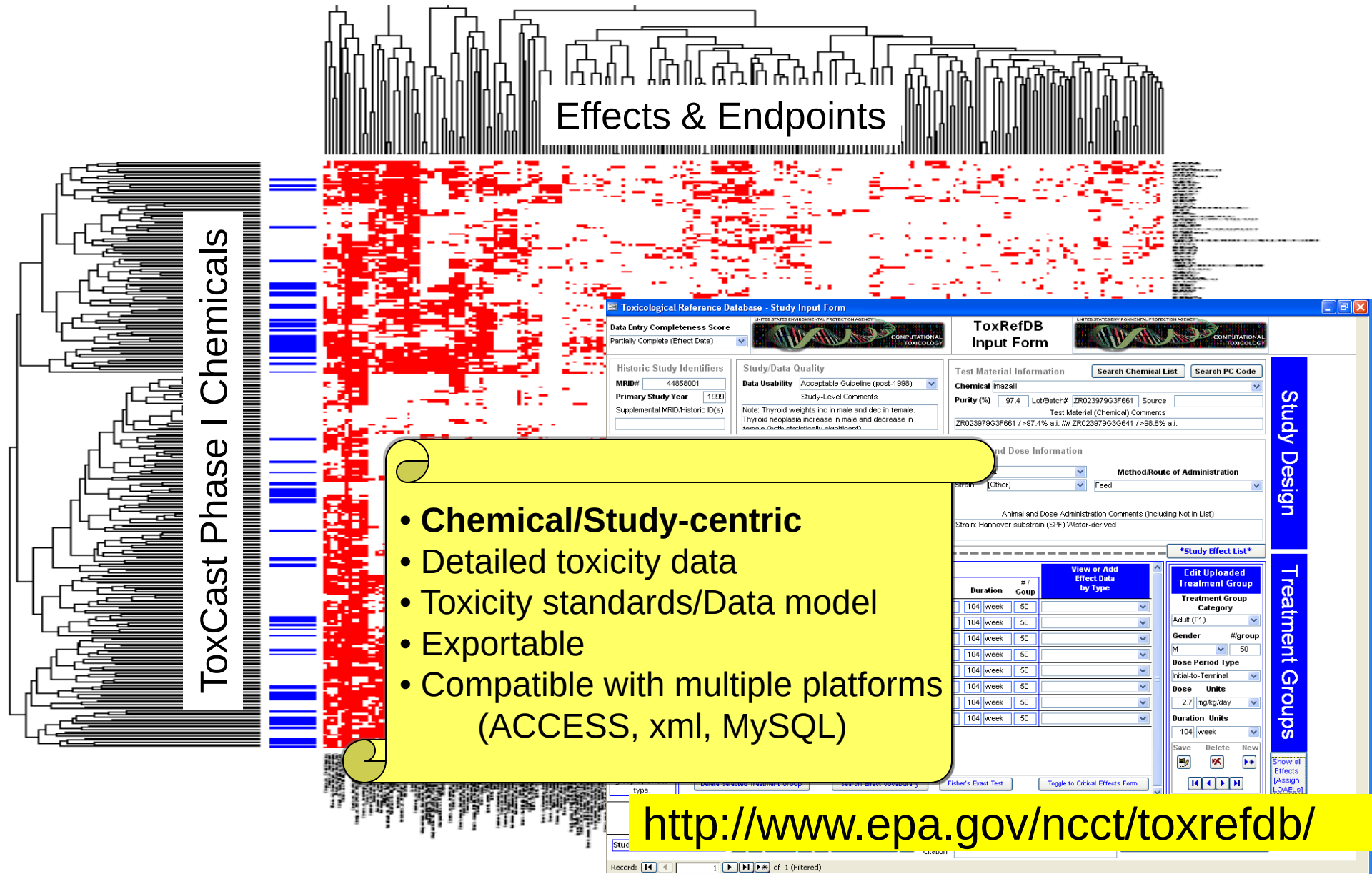


Ca-channel blocker (L-type “as neuronal”) And has high permeability through BBB
Could be a positive **potentiator of $GABA_B$** (to be tested: because derivatives are)
And has high permeability through BBB

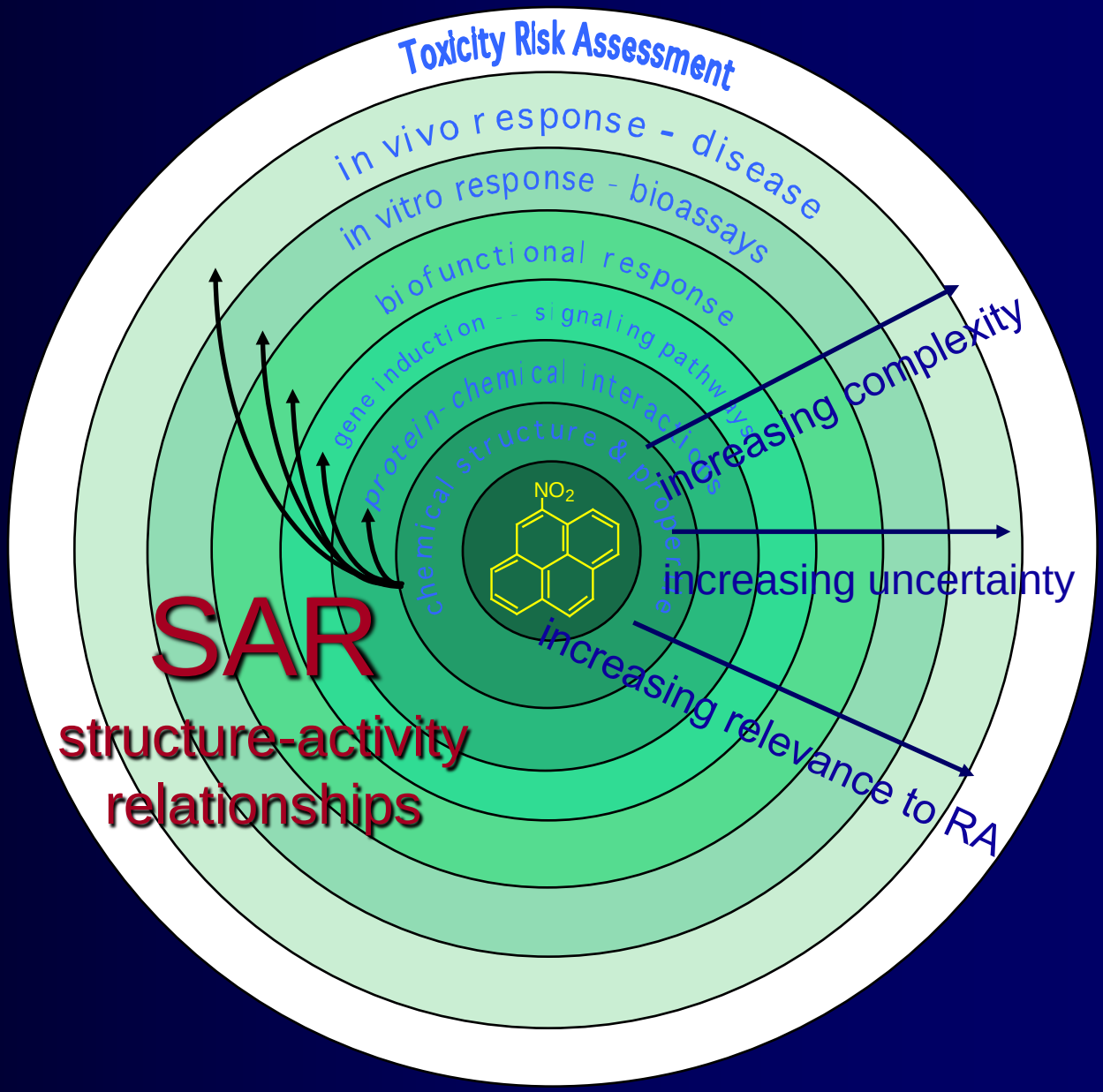
Importance of hypothesis fusion

WDI Name	cmap Name	No. of Models	Av. Pred. Value	cmap Score S1	cmap Score S2
CLOZAPINE	clozapine	900	1.00	-0.398	-0.366
TAMOXIFEN	tamoxifen	910	0.99	0.358	-0.507
FLUSPIRILENE	fluspirilene	854	0.99	-0.493	-0.551
ZUCLOPENTHIXOL	zuclopenthixol	883	0.98	-0.609	-0.746
BI-2	imipramine	898	0.98	-0.503	-0.415
CIDOXEPIN	doxepin	908	0.97	-0.463	-0.777
NORTRIPTYLINE	nortriptyline	883	0.96	-0.555	-0.410
BI-3	clomipramine	893	0.95	-0.768	-0.425
ENCLOMIFENE	clomifene	899	0.91	-0.414	-0.611
DO-897	Prestwick-559	858	0.79	-0.741	-0.619
LY-294002	LY-294002	866	0.70	-0.351	-0.303
ACEFYLLINE-PRENYLAMINE	prenylamine	679	0.69	-0.589	-0.457
NISOXETINE	nisoxetine	899	0.68	-0.491	-0.408
IFENPRODIL	ifenprodil	900	0.66	-0.541	-0.489
FENDILINE	fendiline	765	0.66	-0.388	-0.683
NAFTIFINE	naftifine	724	0.66	-0.790	-0.591
RALOXIFENE	raloxifene	809	0.64	-0.378	-0.482
MEBEVERINE	mebeverine	820	0.57	-0.543	-0.798
LOBELANIDINE	lobelanidine	826	0.56	-0.508	-0.488
LOBELINE	lobeline	882	0.55	-0.514	-0.750
AZACYCLONOL	azacyclonol	840	0.54	-0.448	-0.556

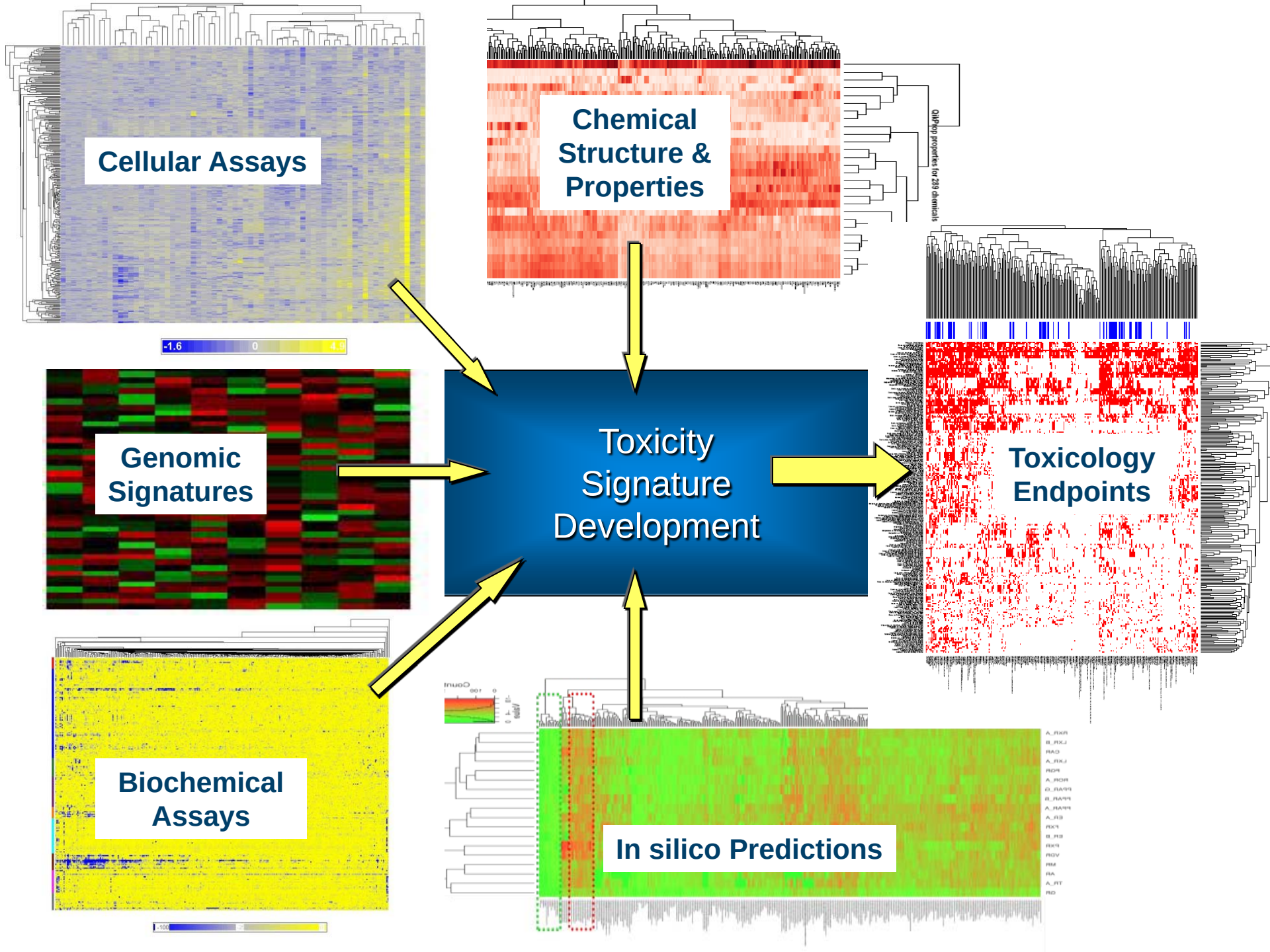
ToxRefDB: >$1Billion Million Dollars Worth of *In Vivo* Chronic/Cancer Bioassay Effects and Endpoints



Chemocentric view of biological data

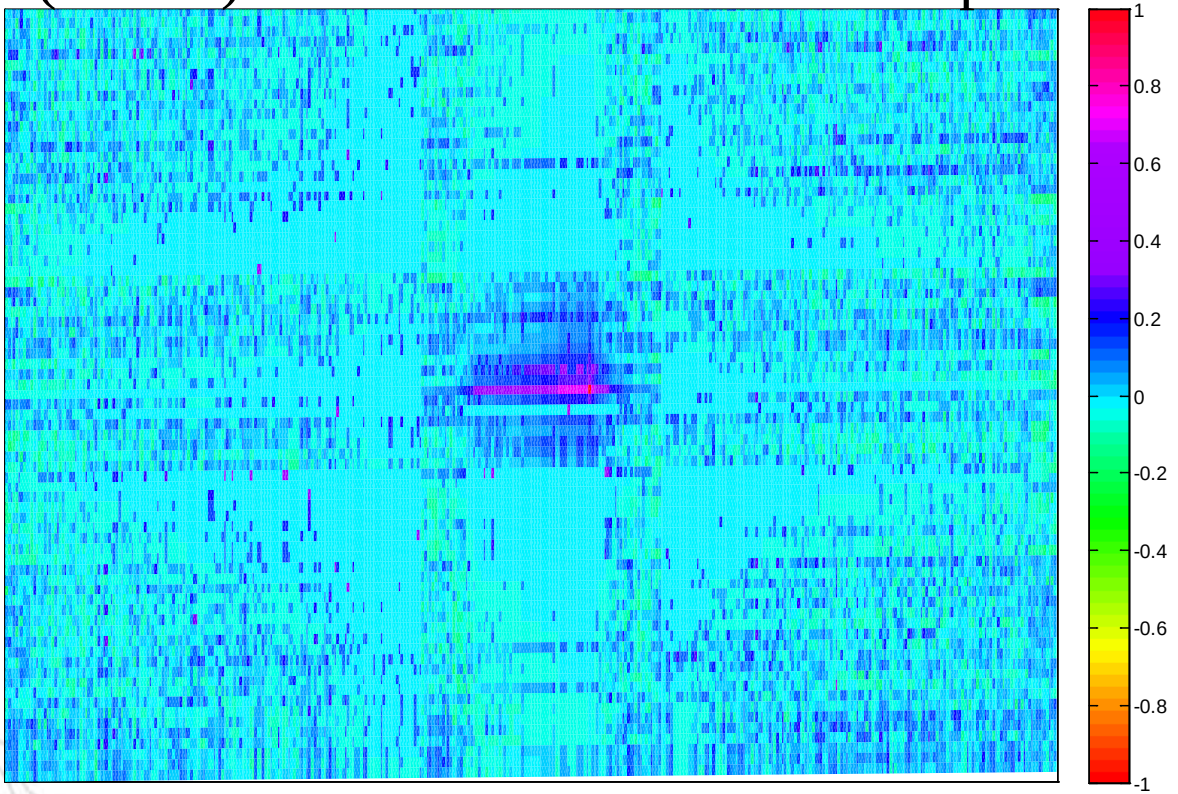


Slide courtesy of Dr. Ann Richard (EPA)


Cellular Assays
Chemical Structure & Properties
Genomic Signatures
Toxicity Signature Development
Toxicology Endpoints
Biochemical Assays
In silico Predictions

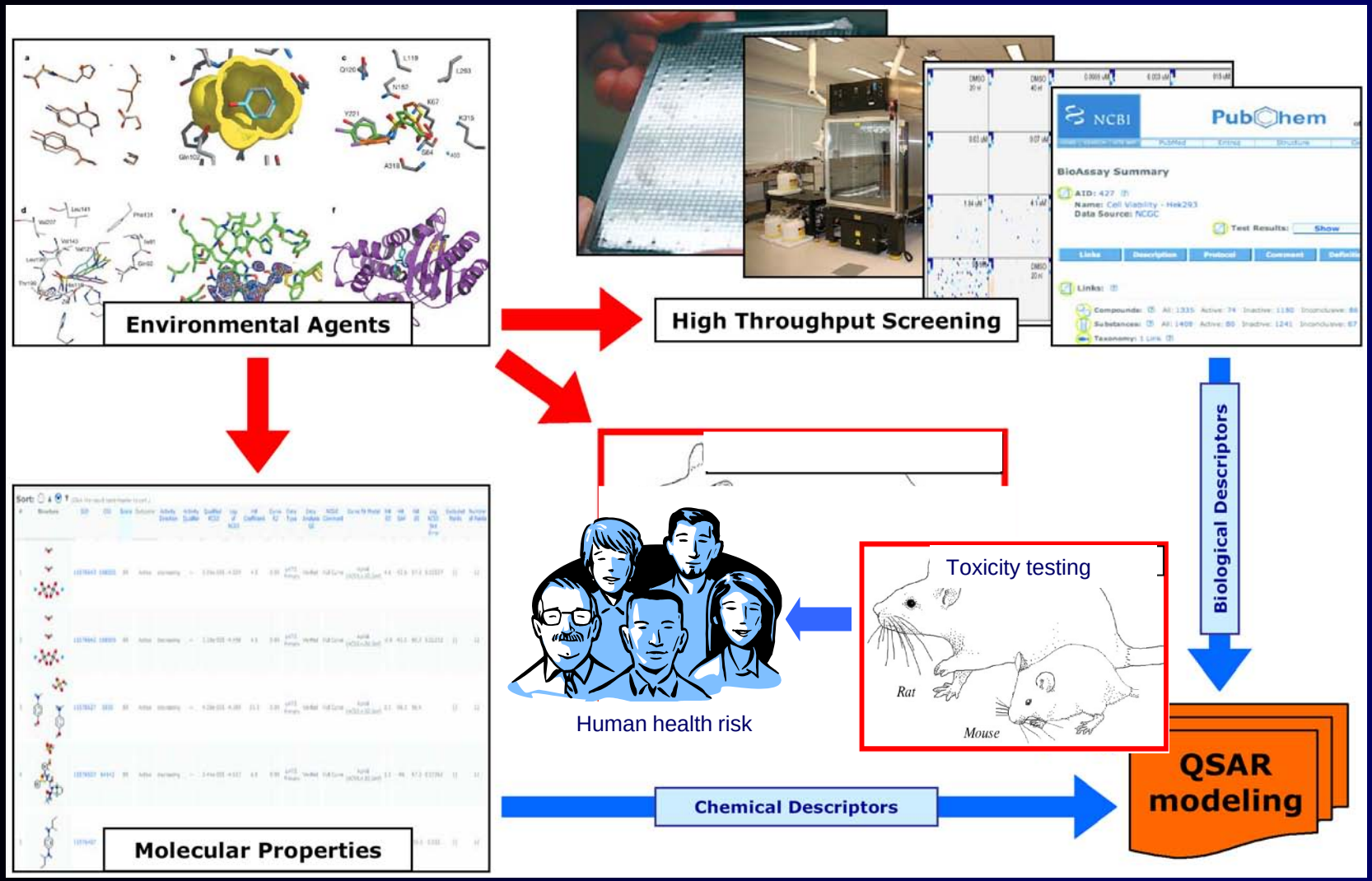
Poor global relationships between in vivo and in vitro assays in ToxCast™ (based on Matthew's Correlation Coefficient, MCC*)

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



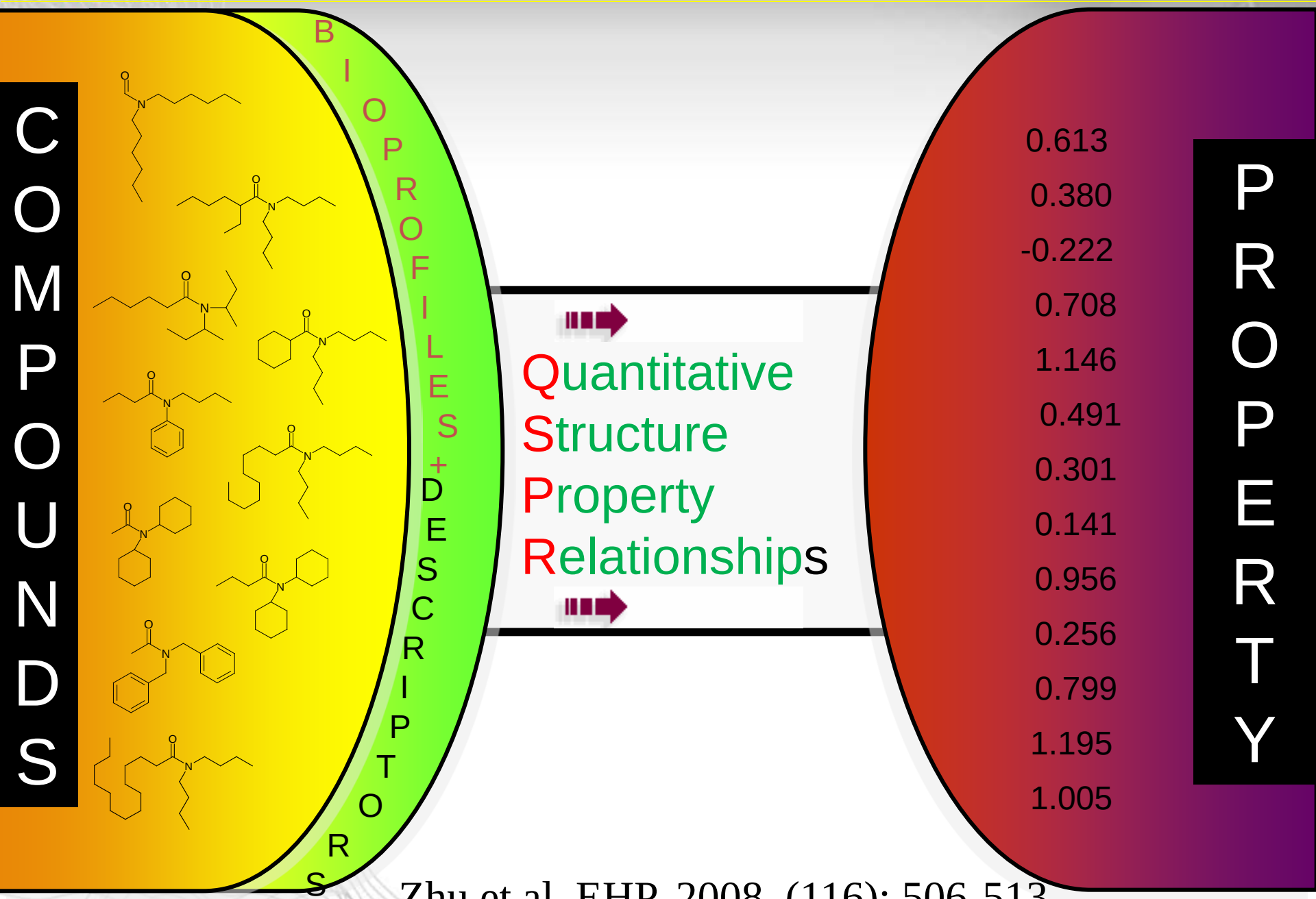
$$^{*}MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

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



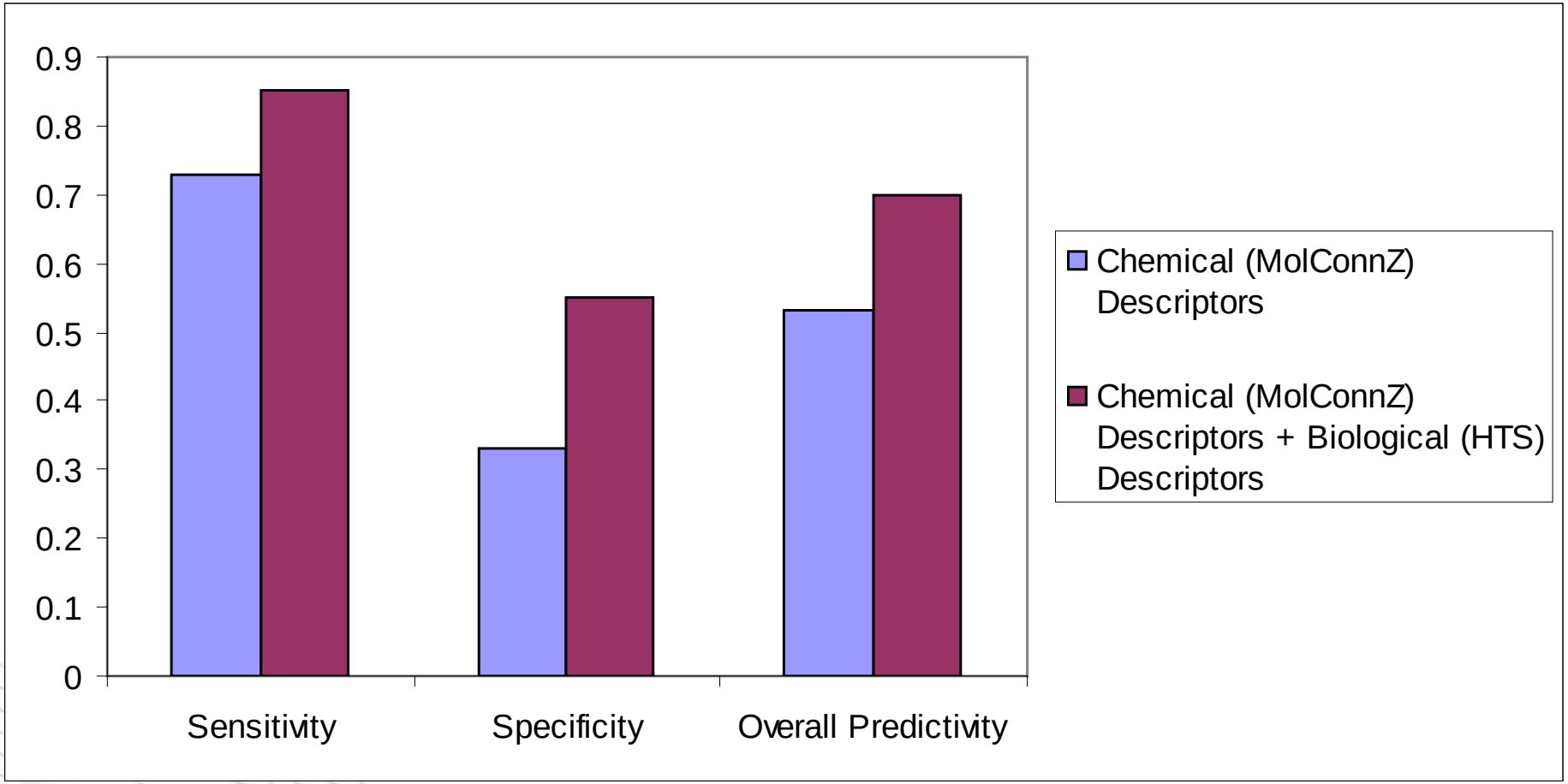
Slide is courtesy of Dr. Ivan Rusyn (UNC)

Use of hybrid descriptors for structure - in vitro – in vivo modeling



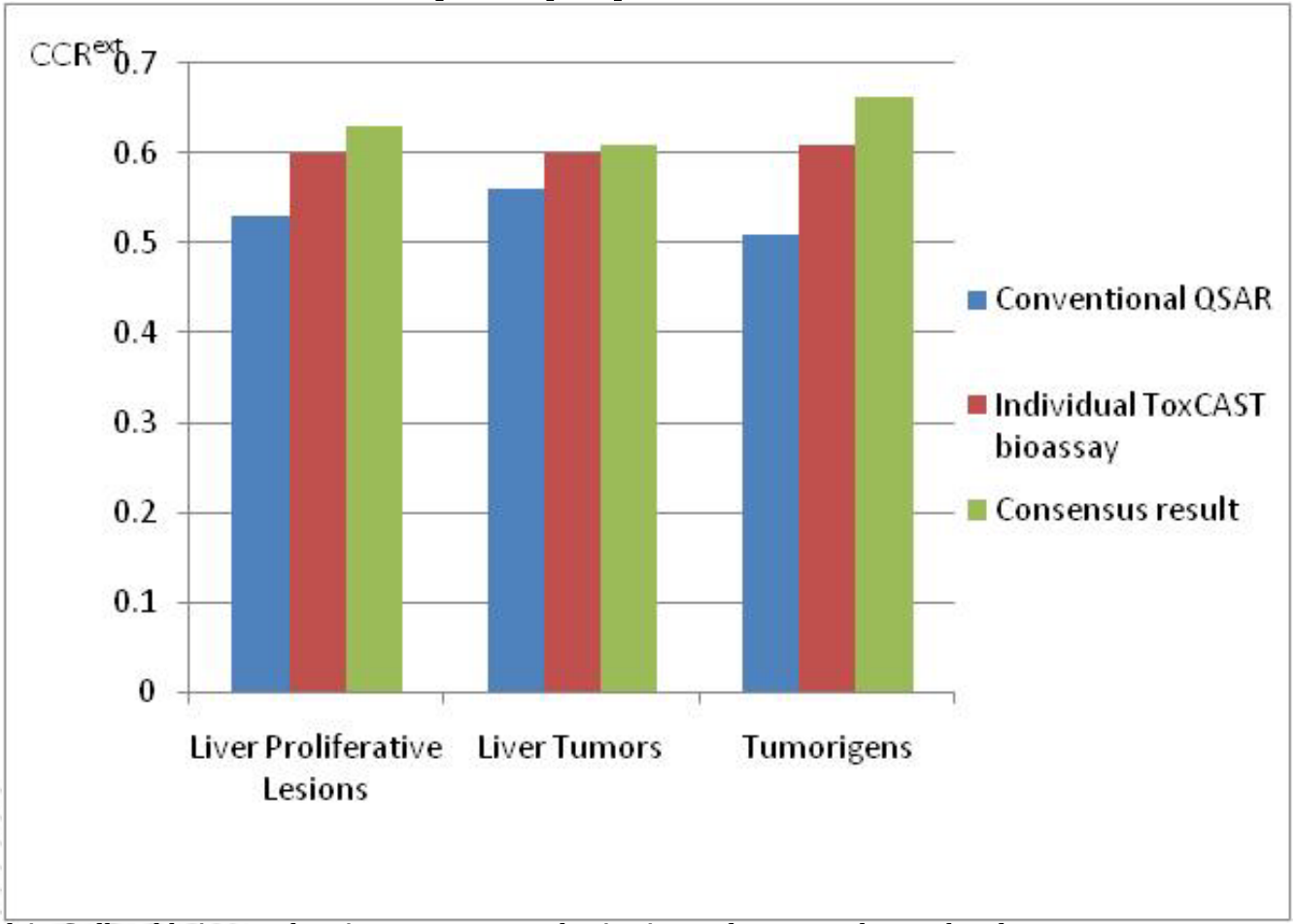
Zhu et al, EHP, 2008, (116): 506-513

Use of HTS based biological descriptors improves predictive power of QSAR Models of chemical carcinogenicity*



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

A new "hierarchical QSAR" approach* relying on the relationship between in vitro and in vivo ToxCast assays results affords highly predictive models

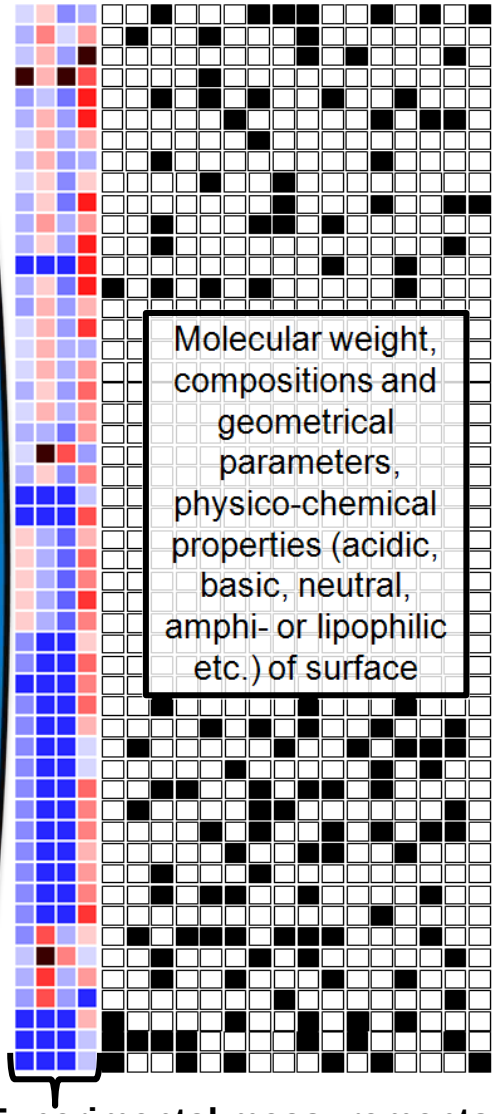


Zhu H, Ye L, Richard A, Golbraikh A, Wright FA, Rusyn I, Tropsha A. A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. Environ Health Perspect. 2009, 117(8):1257-64

Introducing QNTR modeling

NANOMATERIALS

Nanoparticle fingerprints



Molecular weight, compositions and geometrical parameters, physico-chemical properties (acidic, basic, neutral, amphi- or lipophilic etc.) of surface

DESCRIPTORS

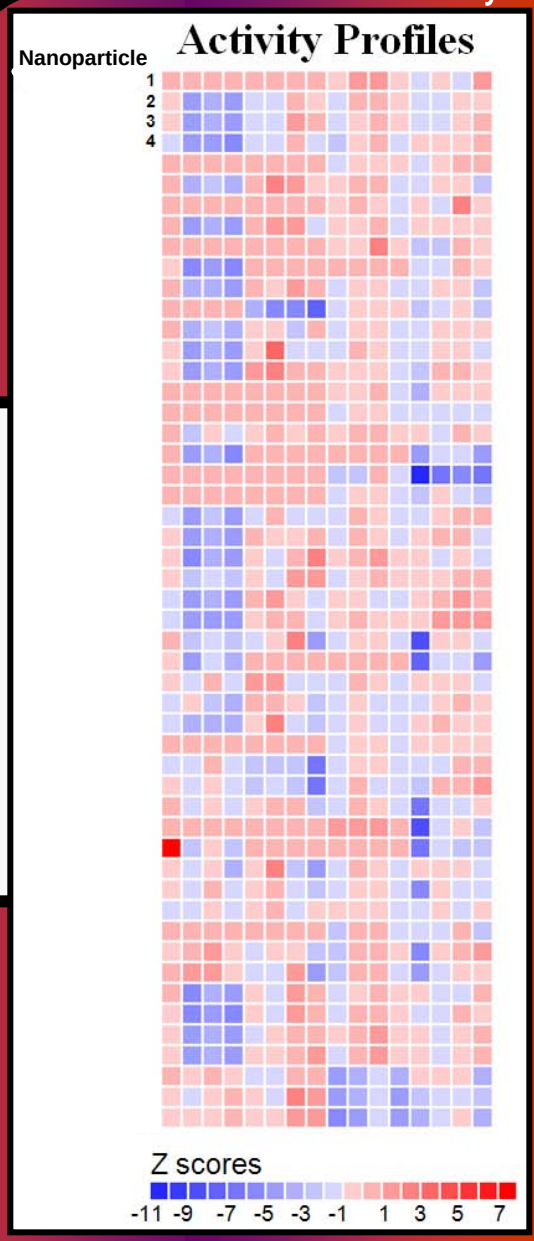
Experimental measurements
(size, relaxivities, zeta potential etc.)

Quantitative Nanostructure Toxicity Relationships

- **Building of models** using machine learning methods (NN, SVM etc.);

- **Validation of models** according to numerous statistical procedures, and their applicability domains.

High-throughput cellular-based assays



Case Study 1

Recently[1], 51 diverse NPs were tested *in-vitro* against 4 cell lines in 4 different assays at 4 different concentrations (→ 51x64 data matrix).

NANOPARTICLES

- → cross-linked iron oxide (**CLIO**)-based (23 NPs)
- → pseudocaged nanoparticle (**PNP**)-based (19 NPs)
- → monocrystalline iron oxide nanoparticle (**MION**)-based (4 NPs)
- → **quantum dot**-based with a CdSe core, a ZnS shell, and a polymer coating (3 NPs)
- → two other iron-based MNPs: Feridex IV (approved for in vivo imaging) and Ferrum Hausmann (approved for iron supplementation)

[1] **Shaw et al. Perturbational profiling of nanomaterial biologic activity. PNAS, 2008, 105, 7387-7392**

CS1. QNTR matrix

	Effect	Size		Zeta pot.		Relaxitivities	
NP-01	High	0.4865	0.5278	0.2941	0.3986		
NP-02	Low	0.4054	0.7222	0.4837	0.6476		
NP-03	High	0.4324	0.5833	0.3529	1.0000		
NP-04	Low	1.0000	0.5833	1.0000	0.7991		
NP-05	High	0.3649	0.4722	0.2353	0.9403		
NP-06	High	0.3919	0.6111	0.3333	0.9079		
NP-07	High	0.5135	0.5833	0.4052	0.6270		

For 44 NPs, size, zeta potential and relaxitivities were available, and then normalized between 0 and 1, to form the QNTR matrix.

Is it possible to predict whether a given particle will induce low or high biological effects using QNTR models?

CS1. QNTR modeling results of 44 diverse NPs using MML-WinSVM and a 5 fold external cross-validation

Fold	MODELING SETS				EXTERNAL SETS				
	n	# models	% accuracy internal 5-fold CV	% accuracy	*n*	% accuracy	% CCR[a]	% Sensitivity	% Specificity
1	35	11	51.4 – 60.0	71.4 – 82.9	9	**78**	83	67	100
2	35	13	51.4 – 60.0	71.4 – 77.1	9	**78**	75	50	100
3	35	16	57.1 – 62.9	74.3 – 82.9	9	**78**	78	80	75
4	35	11	60.0 – 62.9	77.1 – 88.6	9	**56**	55	50	60
5	36	4	66.7	83.3 – 86.1	8	**75**	67	33	100
					44	**73**	**73**	**60**	**86**

[a]CCR – Correct Classification Rate.

Prediction performances are surprisingly good : the overall prediction accuracy for those 44 NPs is equal to 73 %

QSAR and toxicity prediction: QSAR Modeling* of the TETRATOX aquatic toxicity

- Schultz, T.W. TETRATOX: Tetrahymena pyriformis population growth impairment endpoint-A surrogate for fish lethality. Toxicol. Methods (1997) 7: 289-309
- A short-term, static protocol using the common freshwater ciliate *Tetrahymena pyriformis* (strain GL-C) to test aquatic toxicity.
- The 50% impairment growth concentration (IGC50) is the recorded endpoint.
- Website: http://www.vet.utk.edu/TETRATOX/

*Zhu et al, JCIM, J Chem Inf Model 2008; (48): 766-784

International Virtual Collaboratory* of Computational Chemical Toxicology

- USA: UNC-Chapel Hill (UNC) - H. Zhu and A. Tropsha
- France: University of Louis Pasteur (ULP) – D. FOURCHES and A. VARNEK
- Italy: University of Insubria (UI) – E. PAPA and P. GRAMATICA
- Sweden: University of Kalmar (UK) – T. ÖBERG
- Germany: Munich Information Center for Protein Sequences/Virtual Computational Chemistry Laboratory (VCCLAB)– I. TETKO
- Canada: University of British Columbia (UBC) – A. CHERKASOV

*a new networked organizational form that also includes social processes; collaboration techniques; formal and informal communication; and agreement on norms, principles, values, and rules

Different countries, different groups, different tools – shared basic principles

- Explore and combine various QSAR approaches
- Use extensive model validation and applicability domains
- Consider external prediction accuracy as the ultimate criteria of model quality

$$Q^2_{abs} = 1 - \sum_Y (Y_{\exp} - Y_{LOO})^2 \Big/ \sum_Y (Y_{\exp} - \langle Y \rangle_{\exp})^2 \qquad (1)$$

$$R^2_{abs} = 1 - \sum_Y (Y_{\exp} - Y_{pred})^2 \Big/ \sum_Y (Y_{\exp} - \langle Y \rangle_{\exp})^2 \qquad (2)$$

$$MAE = \sum_Y |Y - Y_{pred}| \Big/ n \qquad (3)$$

Overview of the Approaches (15 methodologies total)

Group ID	Modeling Techniques	Descriptor Type	Applicability Domain
UNC	*k*NN, SVM	MolConnZ, Dragon	Euclidean distance threshold between a test compound and compounds in the modeling set
ULP	MLR, *k*NN, SVM	Fragments	Euclidean distance threshold between a compound and compounds in the modeling set; bounding box
UI	OLS	Dragon	Leverage approach
UK	PLS	Dragon	Residual standard deviation and leverage within the PLSR model
MIPS	ASNN	E-state	Maximal correlation coefficient of the test molecule to the training set molecules in the space of models
UBC	MLR, ANN, SVM, PLS	IND_I	Descriptor variability

Individual vs. Consensus Models for the Modeling Set

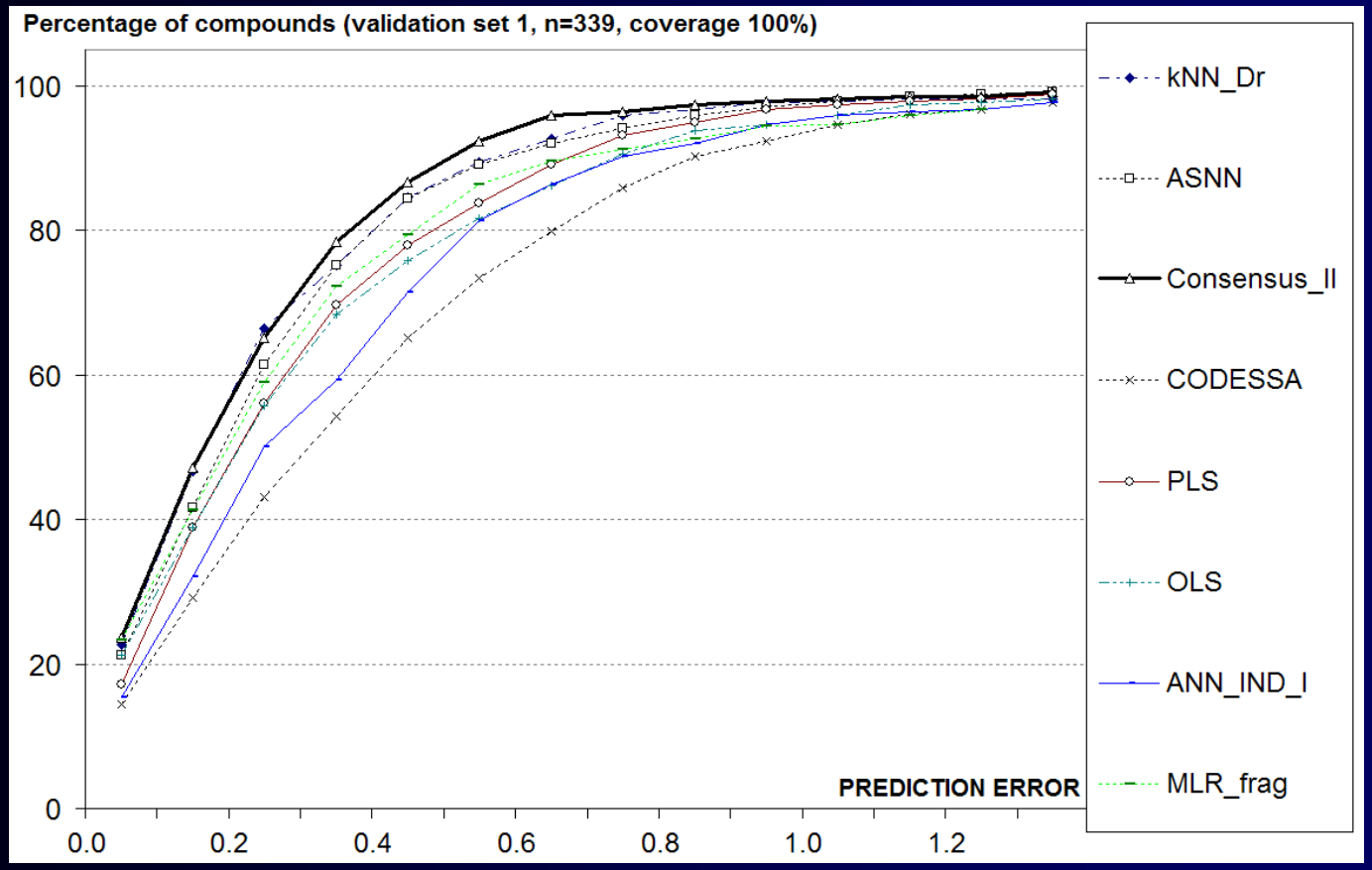
Model	Group ID	Modeling Set (n=644)		
		q^2	SE	Coverage
*k*NN-Dragon	UNC	0.93	0.23	100%
*k*NN-MolconnZ	UNC	0.92	0.26	99.8%
SVM-Dragon	UNC	0.93	0.26	100%
SVM-MolconnZ	UNC	0.89	0.33	100%
*k*NN-Fragmental	ULP	0.77	0.44	100%
SVM-Fragmental	ULP	0.95	0.23	100%
MLR	ULP	0.94	0.25	100%
MLR-CODESSA	ULP	0.72	0.47	100%
OLS	UI	0.86	0.35	92.1%
PLS	UK	0.88	0.34	97.7%
ASNN	MISP	0.92	0.27	83.9%
PLS-IND_I	UBC	0.76	0.39	100%
MLR-IND_I	UBC	0.77	0.39	100%
ANN-IND_I	UBC	0.77	0.39	100%
SVM-IND_I	UBC	0.79	0.31	100%
Consensus Model	-	0.92	0.22	100%

Which model is best?

- Observation: Models that afford most accurate predictions for the validation sets are not necessarily ranked as top models for the modeling set.
- Back to choices and practices: So how do we choose “the best” models?

  Should we choose!?
- Consensus Prediction
  - Only predict compounds within the applicability domain of most models
  - For each compound, exclude predictions that have high deviations from the mean value
  - Final predicted value is the average over all predictions.

Consensus Model gives the lowest MAE of prediction (Validation Set)



Principles of "Safe" QSAR modeling

- Establish an SAR database (target property, descriptor set).
- Rationally divide the dataset into training and test sets
- Develop training set models and characterize them with internal validation parameters.
- Validate training set models using external test set and calculate the external validation parameters
- Finally, explore and exploit validated QSPR models for possible mechanistic interpretation and prediction.*

*Tropsha, Gramatica, Gombar. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. Quant. Struct. Act. Relat. Comb. Sci. 2003, 22, 69-77.)

Important vs. Less Important Directions in QSAR modeling (it is about PREDICTIONS)

- Less important (model development)
  - Descriptor development and/or integration (some exceptions)
  - “novel” data analytical techniques
  - Training set statistics
  - (Harmonizing) definitions (SAR, QSAR, etc.)
  - Mechanistic interpretation (except for validated models)
- More important (model validation)
  - Quality and representation of biological data
  - Analysis of common descriptors and most successful combinations (of descriptors and data modeling techniques) that increase the experimental hit rate
  - Training vs. test vs. evaluation set selection (three-way)
  - Outlier analysis (experimental accuracy or descriptor incapability)
  - Applicability domain (in the context of modeling technique AND TEST SET STATISTICS)
  - The real power of QSAR models is in their ability to design novel active compounds or identify such compounds in databases or virtual libraries
- Independent model evaluation in competitive fashion: CoErPA (similar to CASP) and benchmark dataset depository

Final Word

Nothing that worth knowing can be taught.

Oscar Wilde

- Best time ever to be a cheminformatics scholar
  - Growth of databases
  - Tool development
  - Collaborations with computational and experimental scientists
- Extending cheminformatics approaches to new areas
  - Structure based virtual screening
  - "-omics" data analysis
  - Structure – in vitro – in vivo correlations
  - Toxico-cheminformatics
- Focus on Knowledge Discovery (accurate testable predictions!) in Chemical Databases