The good, the bad, and the ugly…practices of QSAR modeling

Alexander Tropsha

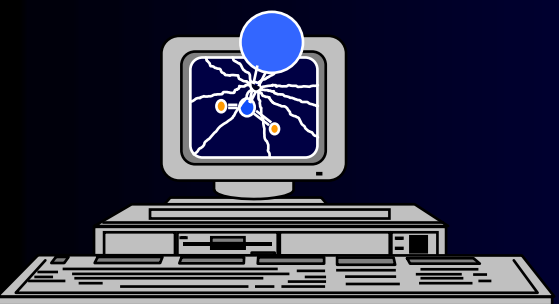
Laboratory for Molecular Modeling

and

Carolina Center for Exploratory Cheminformatics Research

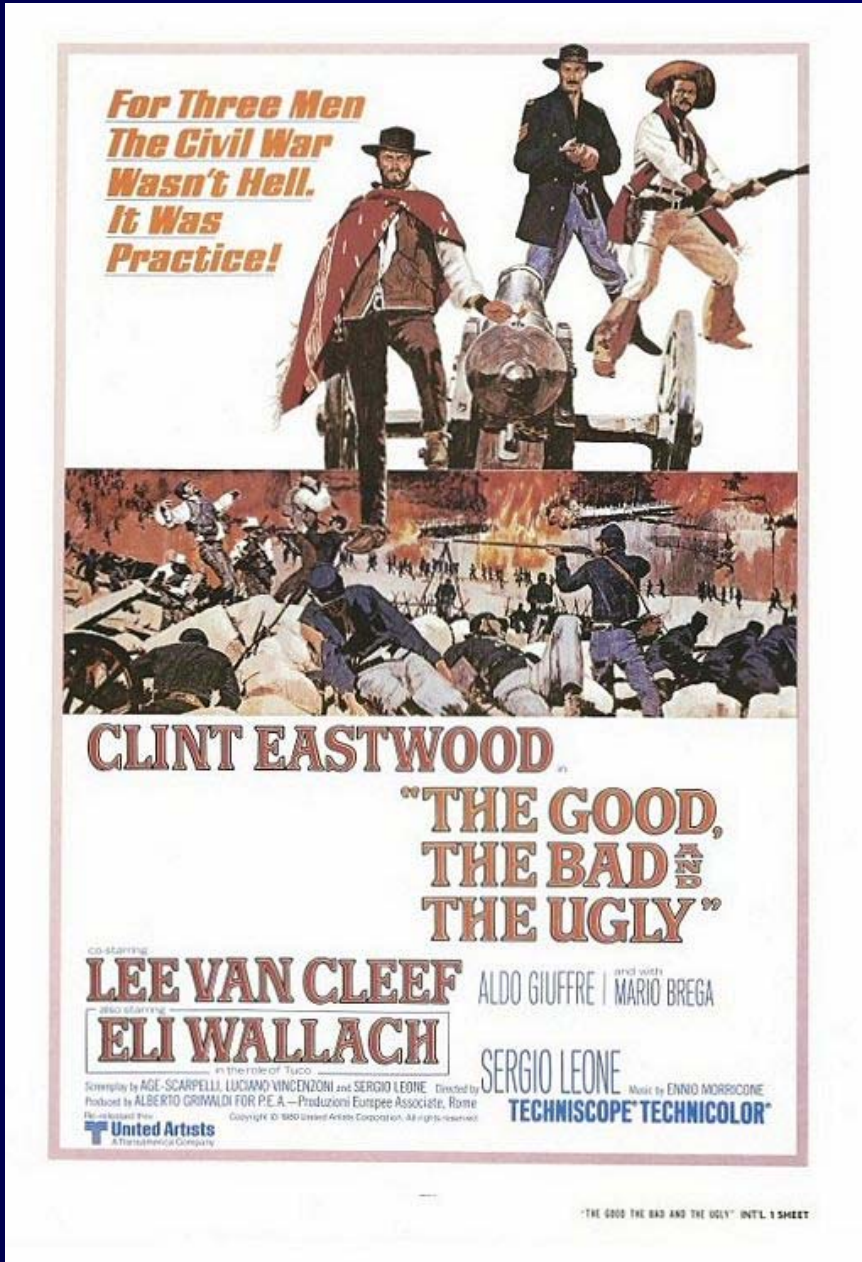
School of Pharmacy

UNC-Chapel Hill



Some definitions

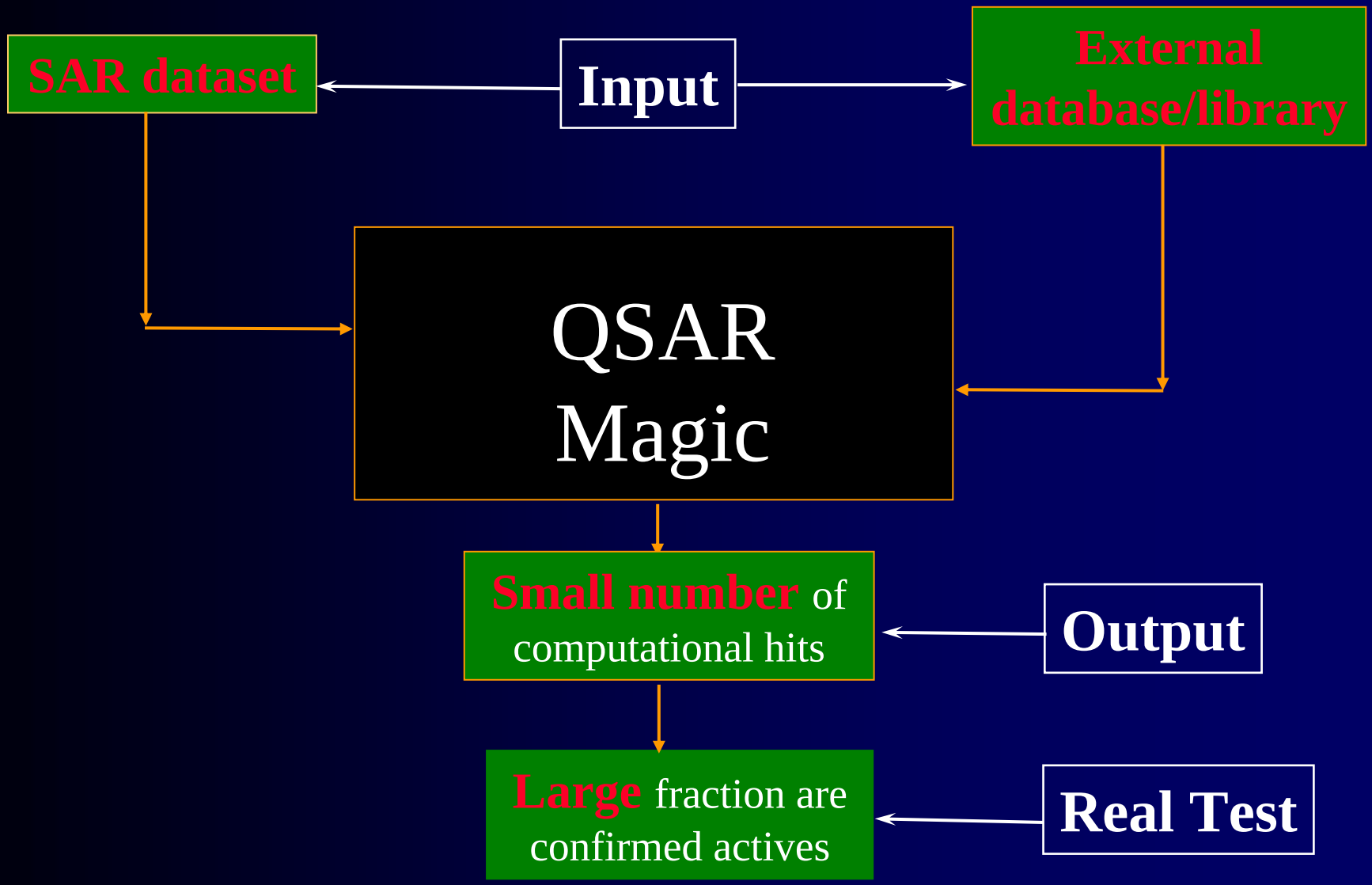
- Clint Eastwood is the "good" (slow to anger, but quick on the trigger)
- Lee Van Cleef is the bad (an elegant exemplar of absolute evil) and
- Eli Wallach is the "ugly" (a menacingly funny, totally amoral *bandido* whose relationship with the Eastwood character consists largely of betrayals).



OUTLINE

- Introduction: The need for developing externally validated → predictive models of biological data
- Why do models fail (bad practices)
- Predictive QSAR Modeling Workflow (good practices)
- Examples of the Workflow applications
  - QSAR based virtual screening and hit identification
  - Consensus QSAR modeling of chemical toxicity
- Conclusions: “best” QSAR modeling is a decision support science → focus on accurate predictions

Key point: Focus on Externally Validated Predictions



QSAR Modeling is easy…

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

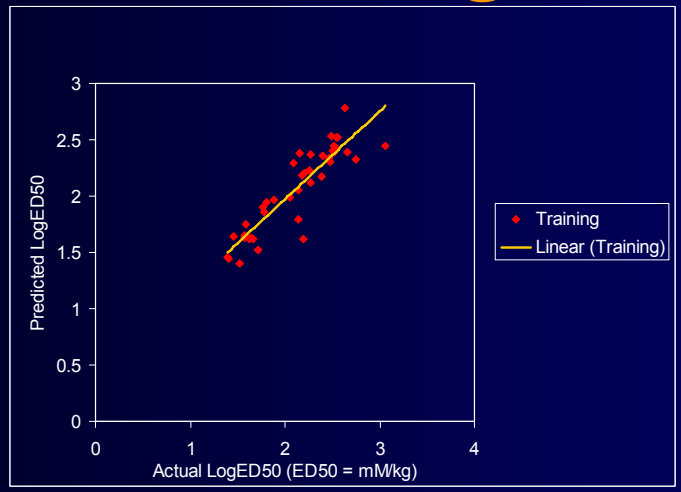
Chemistry	Biology (IC50, Kd...)	Cheminformatics (Molecular Descriptors)				
Comp.1	Value1	D_1	D_2	D_3	|	D_n
Comp.2	Value2	"	"	"	|	"
Comp.3	Value3	"	"	"	|	"
Comp.N	ValueN	"	"	"	|	"

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

BA = F(D) {e.g., …}

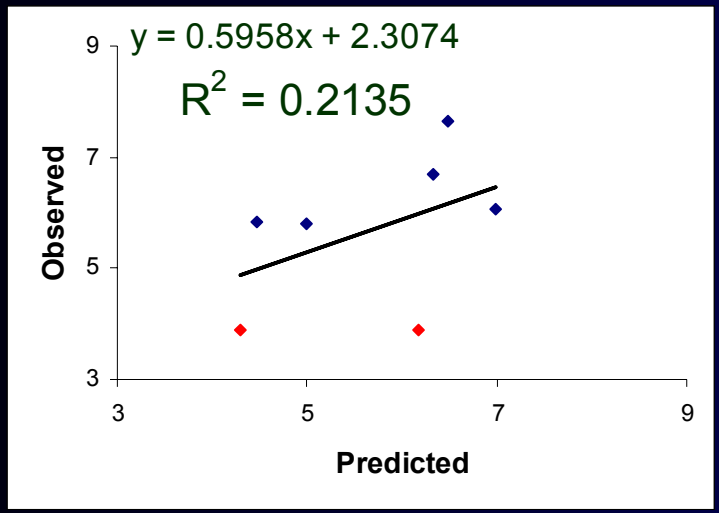
(e.g., -LogIC50 = $k_1D_1+k_2D_2+\ldots+k_nD_n$)

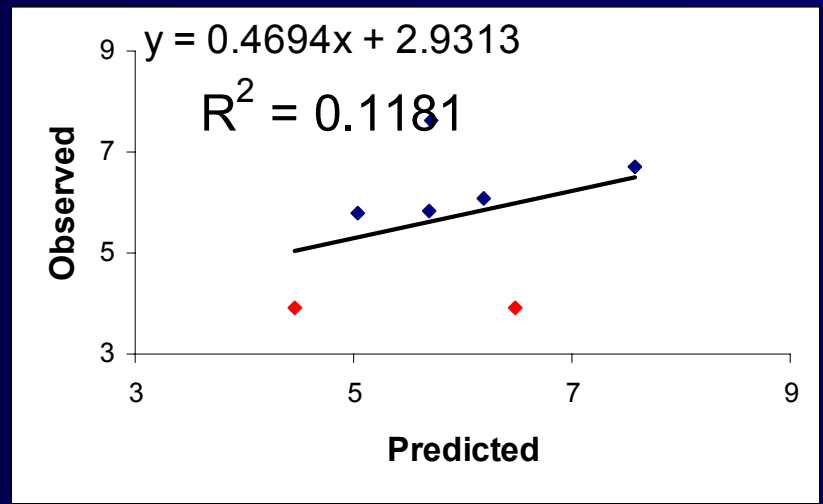
But … the unbearable lightness of model building for training sets…



…leads to unacceptable prediction accuracy.

EXTERNAL TEST SET PREDICTIONS





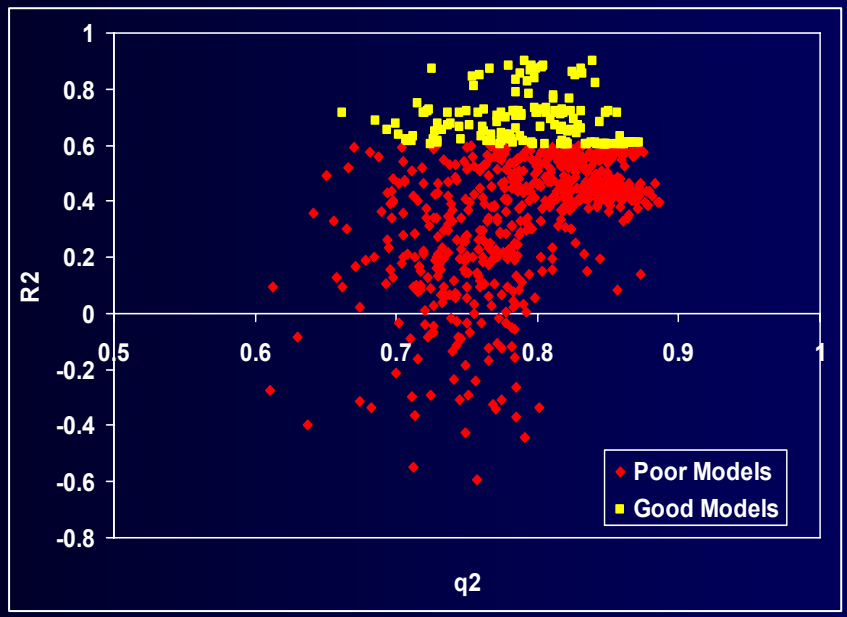
Choices and Practices

- 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

BEWARE OF q^2!!!



•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.

Some reasons as to why QSAR fails

- 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
- Not all statistical criteria are used to estimate predictive power of a model
- No applicability domain
- Incorrectly defined applicability domain
- No Y-randomization
- Leverage (structure) and activity outliers are not removed
- Modeling set is too small

No external validation

- **It is still a problem, particularly in toxicity studies.**

  A typical paper:

  A small dataset of congeneric compounds: n~10-20

  QSAR as a linear regression in the form:

  $\log(1/EC_{50}) = a*\log P + b$

  (sometimes, additional 1-2 descriptors like E_{LOMO} or the number of H-bond donors in a molecule are included)

  The only validation method used: Leave-group-out cross-validation

  Relatively high q^2 (not always) and R^2:

  Typical model acceptance criteria: $q^2>0.5$, $r^2-q^2>0.3$

  (some use R instead of R^2, because $R>R^2$)

  No true validation using compounds not included in the training set

  (in some cases the model is tested on just 2-3 compounds)

Artificial Improvement of Predictive Ability of a QSAR Model

- Johnson, S.R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008,** *48,* 25-26:

  "The common practice has been to select the model with the best fitness function score and predict a small group of observations that were withheld at the beginning. All too often, the model development process stops here, or, worse, the validation set is poorly predicted and models are iteratively tested until one predicts this set of compounds well."

A typical example:

A dataset is divided into a training and test set

Multiple QSAR models with high q^2 values are built using training set

QSAR model with the highest R^2 for the test set is selected

Selected model might have poor predictive ability for other compounds

Another EXTERNAL EVALUATION SETS are necessary

Incorrect division of a dataset into training and test sets

- **Typical division of a dataset into training and test sets**: random
  - **Undesired outcome**:
    - some compounds of the test set can be out of the applicability domain of the training set
    - large gaps of activity in the training or the test set and activity outliers in them
- **Requirements for training and test sets**:
  - Compounds with maximum and minimum activities of the dataset should be included into the training set (important for methods that cannot extrapolate).
  - Large gaps of activities is not allowed neither in training nor in test set.
  - Compounds of the training set should be distributed within the entire area of the descriptor space occupied by the dataset.
  - Each compound of the test set should be close to at least one compound of the training set.

Classification QSAR for Biased Datasets: Incorrect Target Function

- **A typical example of the target function**:

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

  A dataset:

  Class 1: 80 compounds

  Class 2: 20 compounds

  Model: assign all compounds to Class 1.

  Target function: CCR=0.8

  The model has high classification accuracy

- **Better target function**:

  CCR=0.5(Sensitivity+Specificity)

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

Not all statistical parameters are used to estimate predictive power of a model

- Sometimes, cross-validation q^2 for the training set and R^2 for the test set are the only criteria used There are other important criteria which are not always used such as coefficients of determination and slopes for regressions through the origin (predicted vs. observed and observed vs. predicted activities), etc.
- Standard error of prediction alone is not a good statistical parameter, if it is not compared with the standard deviation of activities.

No Applicability Domain 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 as outlier

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

Applicability Domain is Too Large

- **Typical AD**:

  Rectangular hyper-parallelepiped in the descriptor space with edges equal to intervals of change for each descriptor.

  This hyper-parallelepiped can be mostly empty with points concentrated only along certain directions (it is particularly true, if descriptors are linearly dependent).

  Too large AD can lead to the same consequences as not having it at all.

  AD should be defined as a union of relatively small areas around all points of the training set. For example, currently we define it as

  $$D_{cutoff} = \langle d_{nn} \rangle + Z\sigma_{nn},$$

  where $\langle d_{nn} \rangle$ and σ_{nn} are the average of distances between *K* nearest neighbors in the training set and their standard deviation, and Z is a user-defined value (default is 0.5), which can be adjusted.

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

Outliers detection and removal

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

- **Two types of outliers**

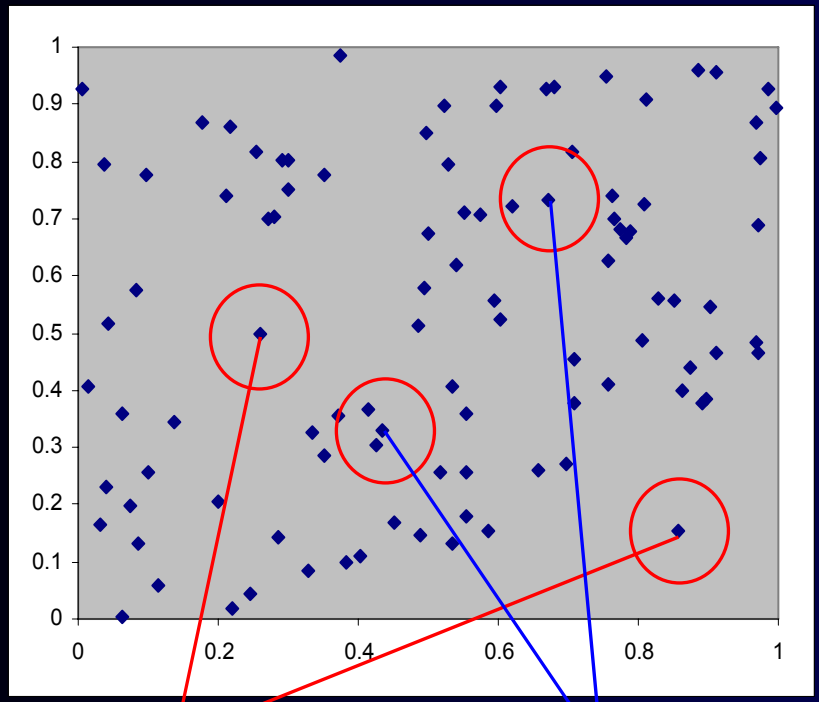
  Leverage outliers: compounds dissimilar from all other compounds in a dataset.

  Activity outliers: compounds similar to some other compounds in the dataset, but their activities are quite different from those of their nearest neighbors.

Detection of Leverage Outliers

- Calculate distance/similarity matrix in the entire descriptor space.
- For each compound, find its nearest neighbor/most similar compound.
- Find compounds which are out of the cutoff distance from their nearest neighbors or have similarity to the most similar compound lower than a predefined threshold. These compounds are leverage outliers for this predefined distance or threshold.

Detection of Leverage Outliers using a Sphere-Exclusion algorithm



PROBE SPHERE RADIUS R:

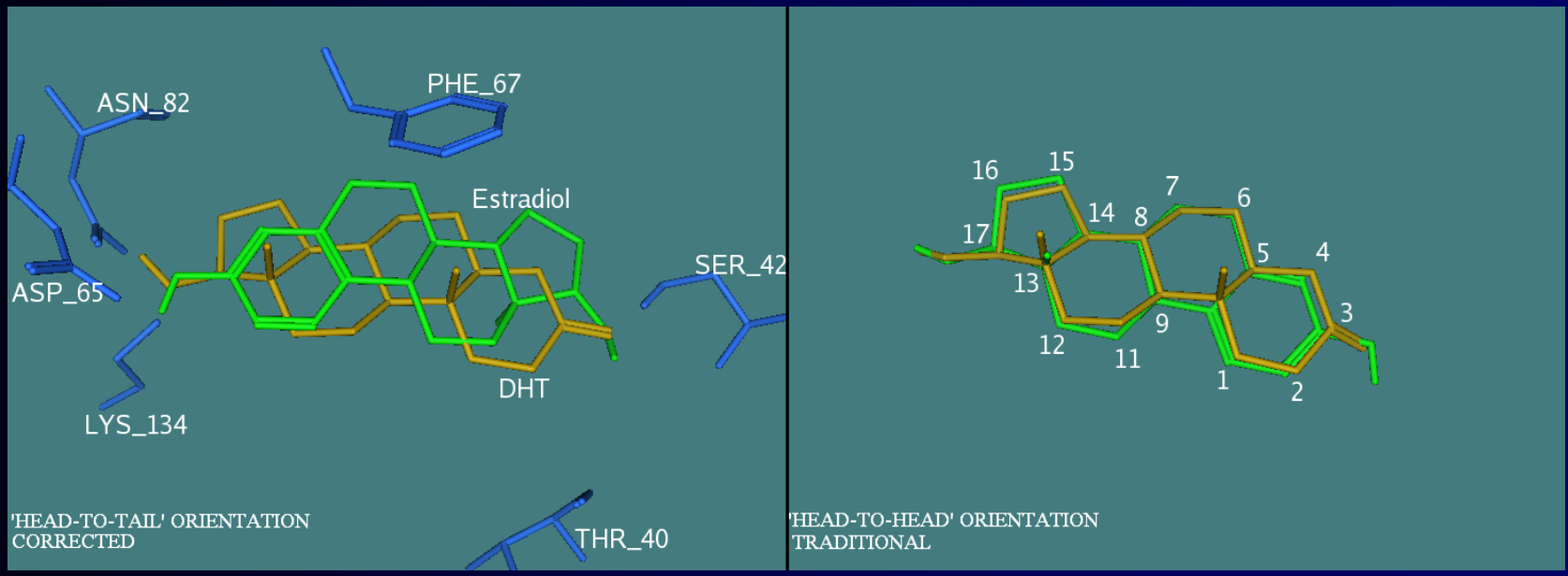
- **Calculate distances between nearest neighbors.**
- **Find mean $\overline{D}$ and standard deviation σ of these distances.**
- **R= $\overline{D}$ + Zσ,**

where Z is a user defined parameter Z-cutoff.

- **Calculations with multiple Z-cutoff values can be carried out.**

Example of “misinformation”: Optimal (left panel) and traditional (right panel) orientations of androgen (DHT shown in gold) and estrogen (estradiol shown in green) within the human SHBG steroid-binding site.

Identical q^2 (CoMFA*) of 0.53



*CoMFA – Completely Misleading Famous Aberration

A. Cherkasov, *JMC*, in press

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?????

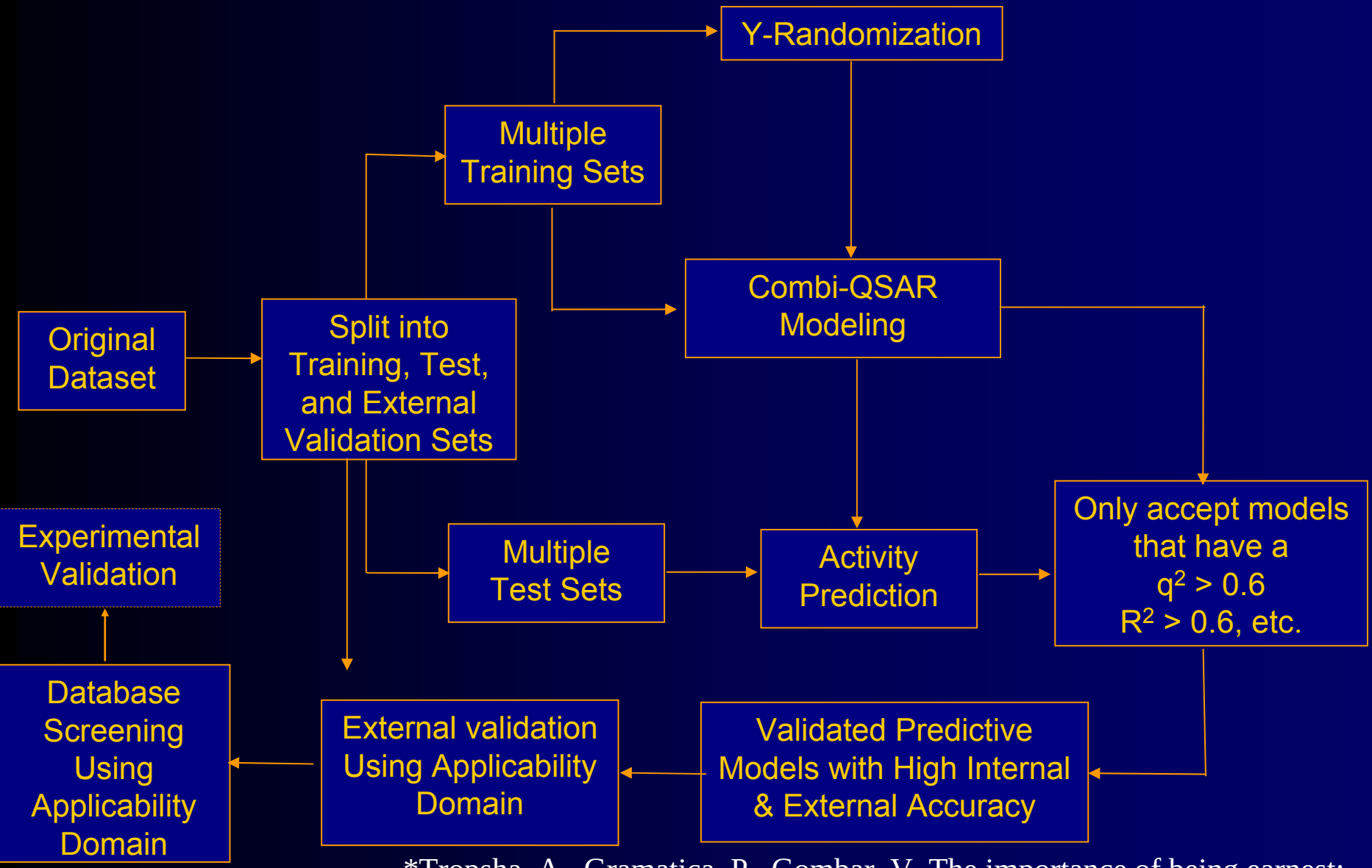
- Change the success criteria! Leave behind the phase of "narcissistic" modeling and focus on external predictivity and experimental validation.
- Recognize QSAR as an empirical data modeling approach: just do it any (all) way you like but VALIDATE on independent datasets!

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

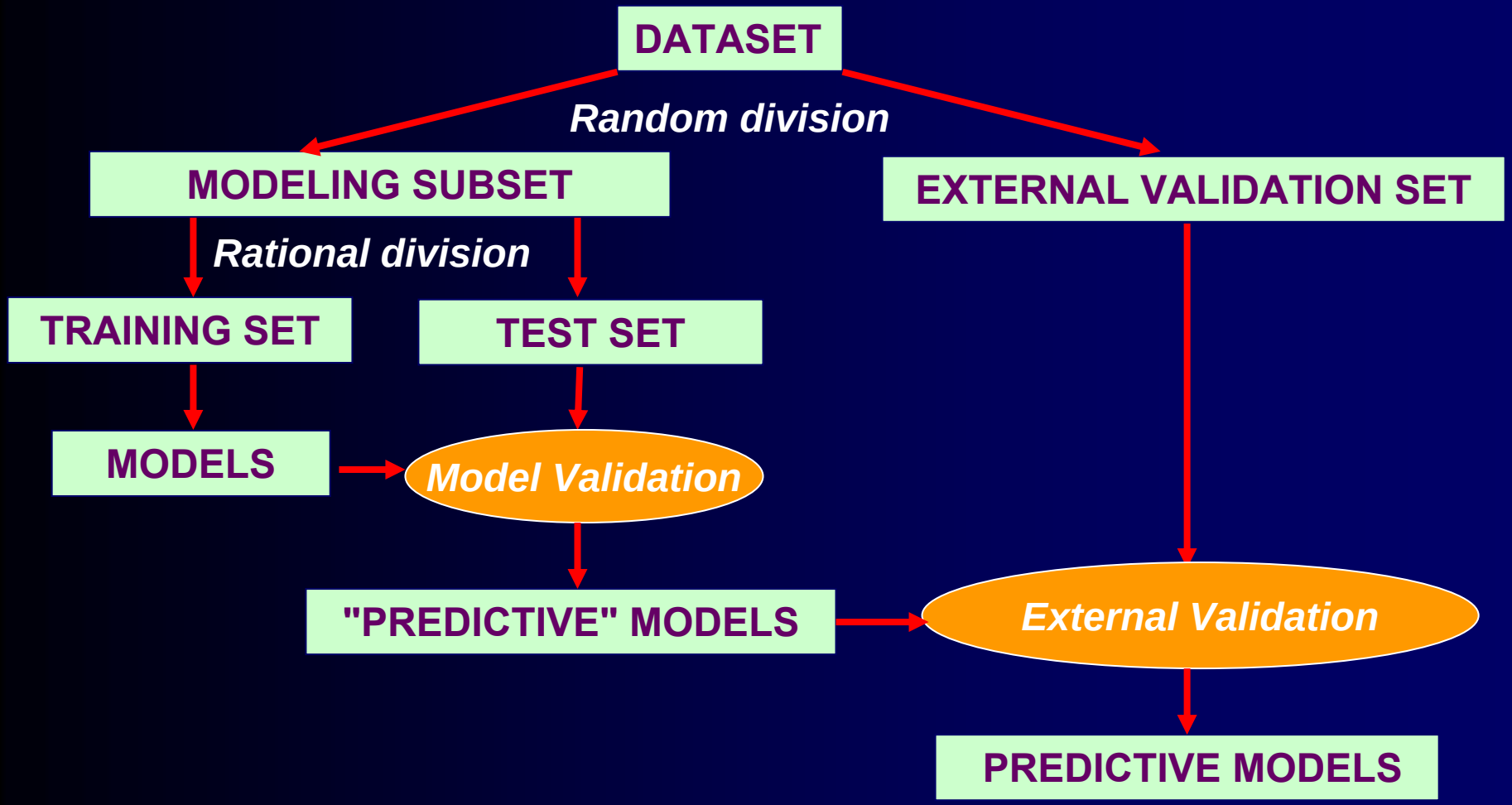
*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:… *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, 22, 69-77.

Predictive QSAR Workflow*

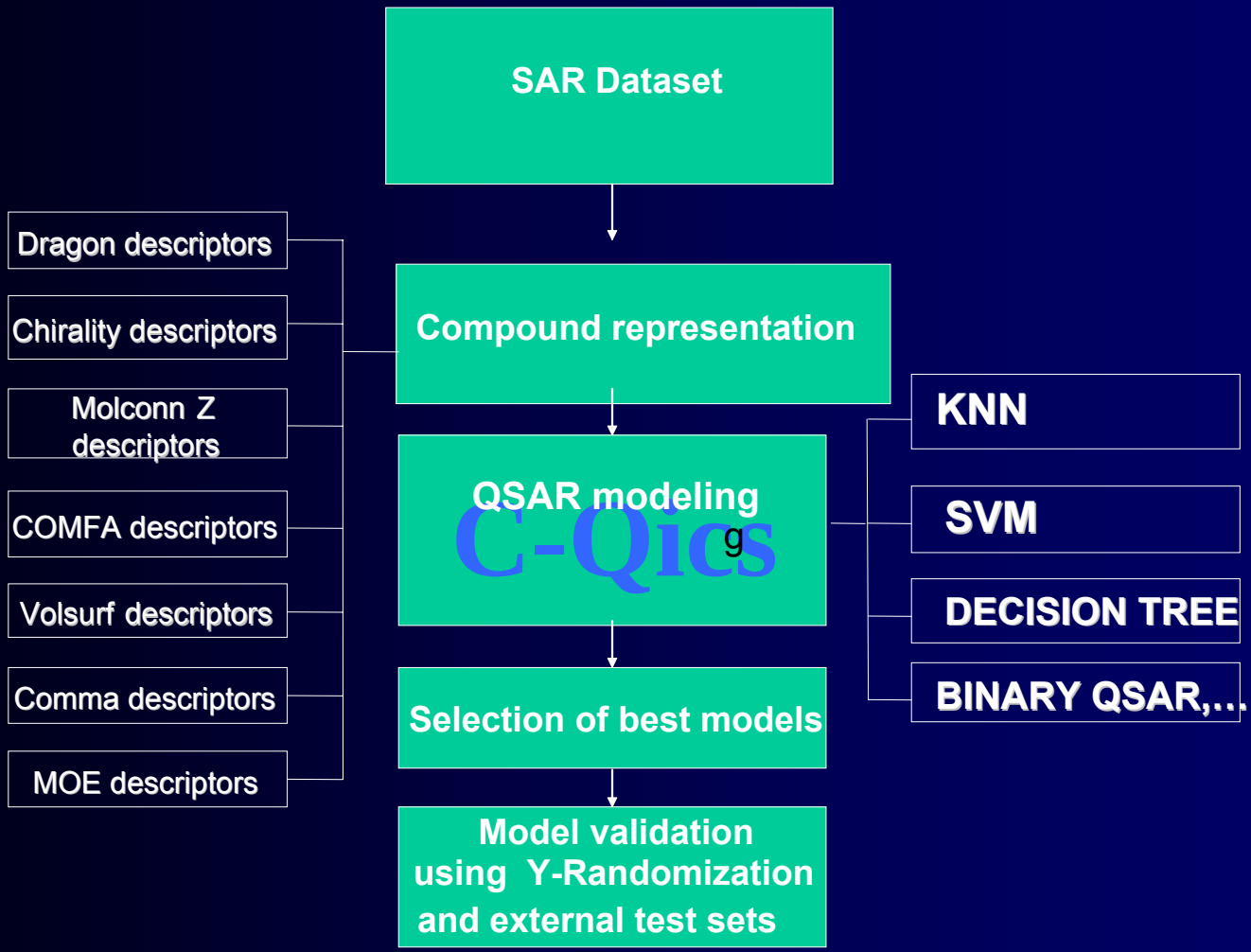


*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:… *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, 22, 69-77.

DIVISION OF A DATASET IN THREE SUBSETS AND EXTERNAL VALIDATION



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

DEFINING THE APPLICABILITY DOMAIN

Training set: 60 compounds
Test set: 35 compounds

MODEL:

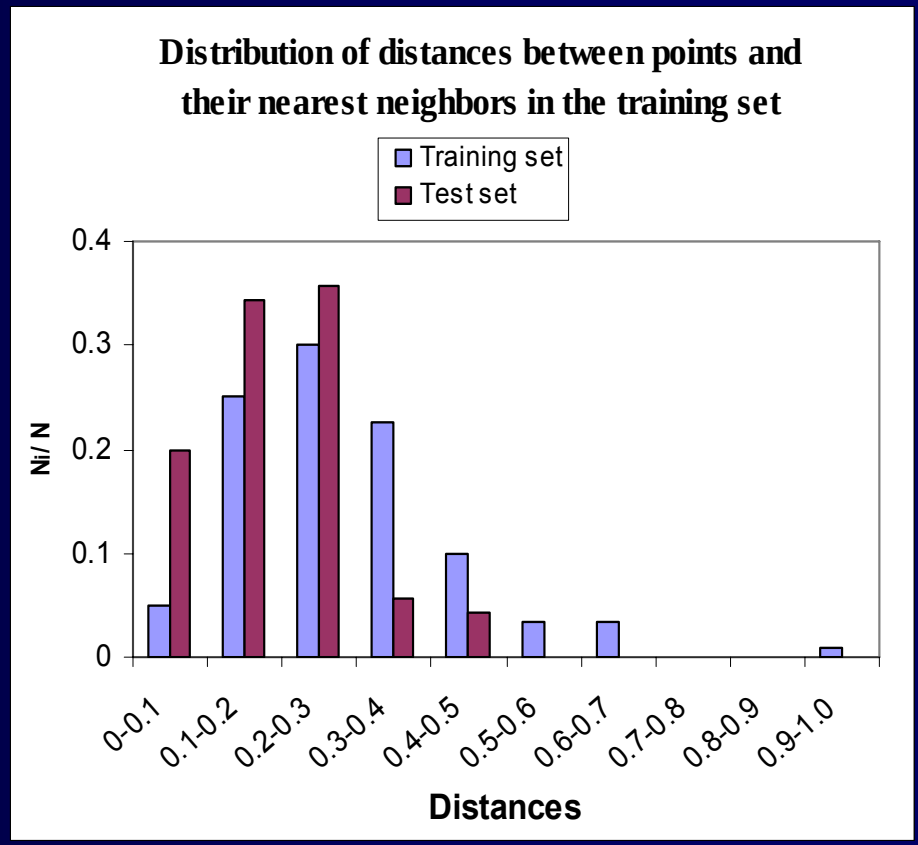
Two nearest neighbors
The number of descriptors: 8
$Q^2(CV)=0.57$ $R^2=0.67$

DISTANCES:

$\langle D\rangle_{train}=0.287$
$StDev(D)_{train}=\sigma=0.149$

Closest nearest neighbors of test set compounds:

$D_{test} \le \langle D\rangle_{train} + \sigma \times Z_{CutOff}$
($Z_{CutOff}=0.5$)

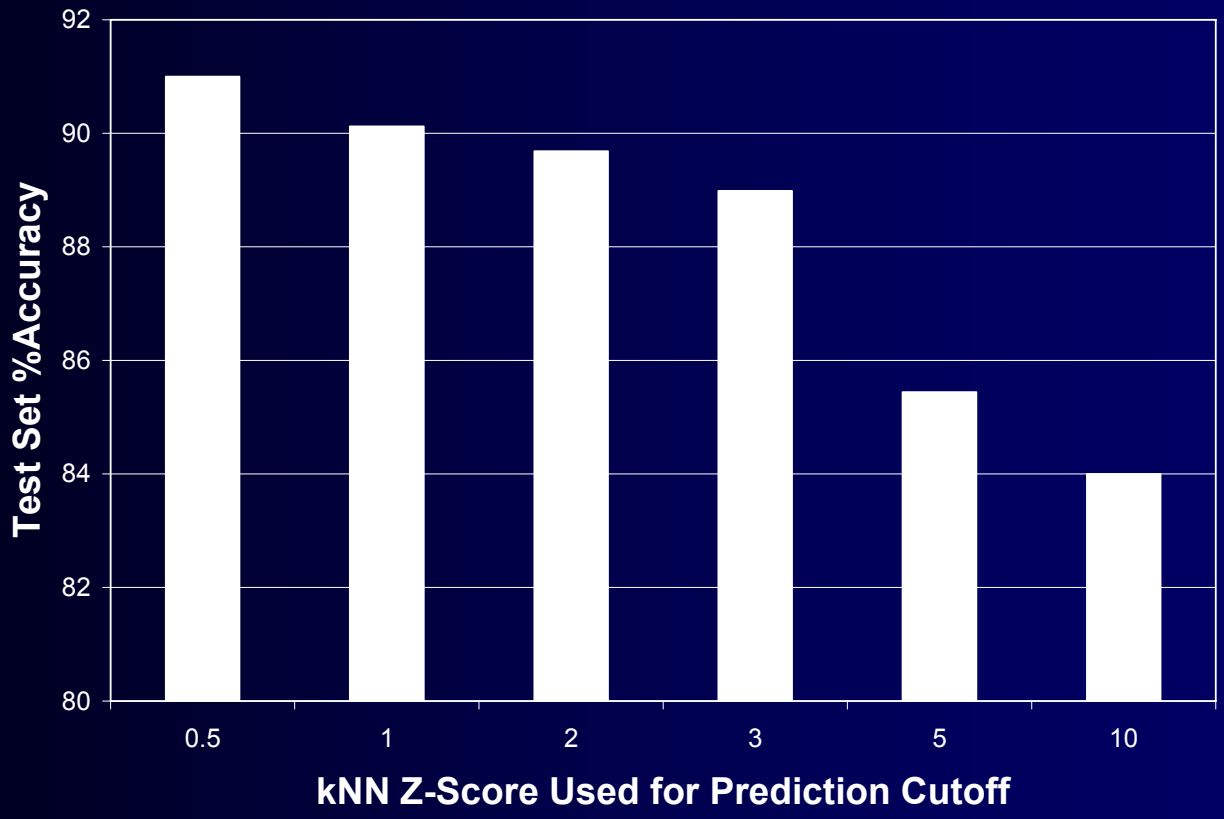


N is the total number of distances
($N_{train}=60$ 2=120; $N_{test}=70$)

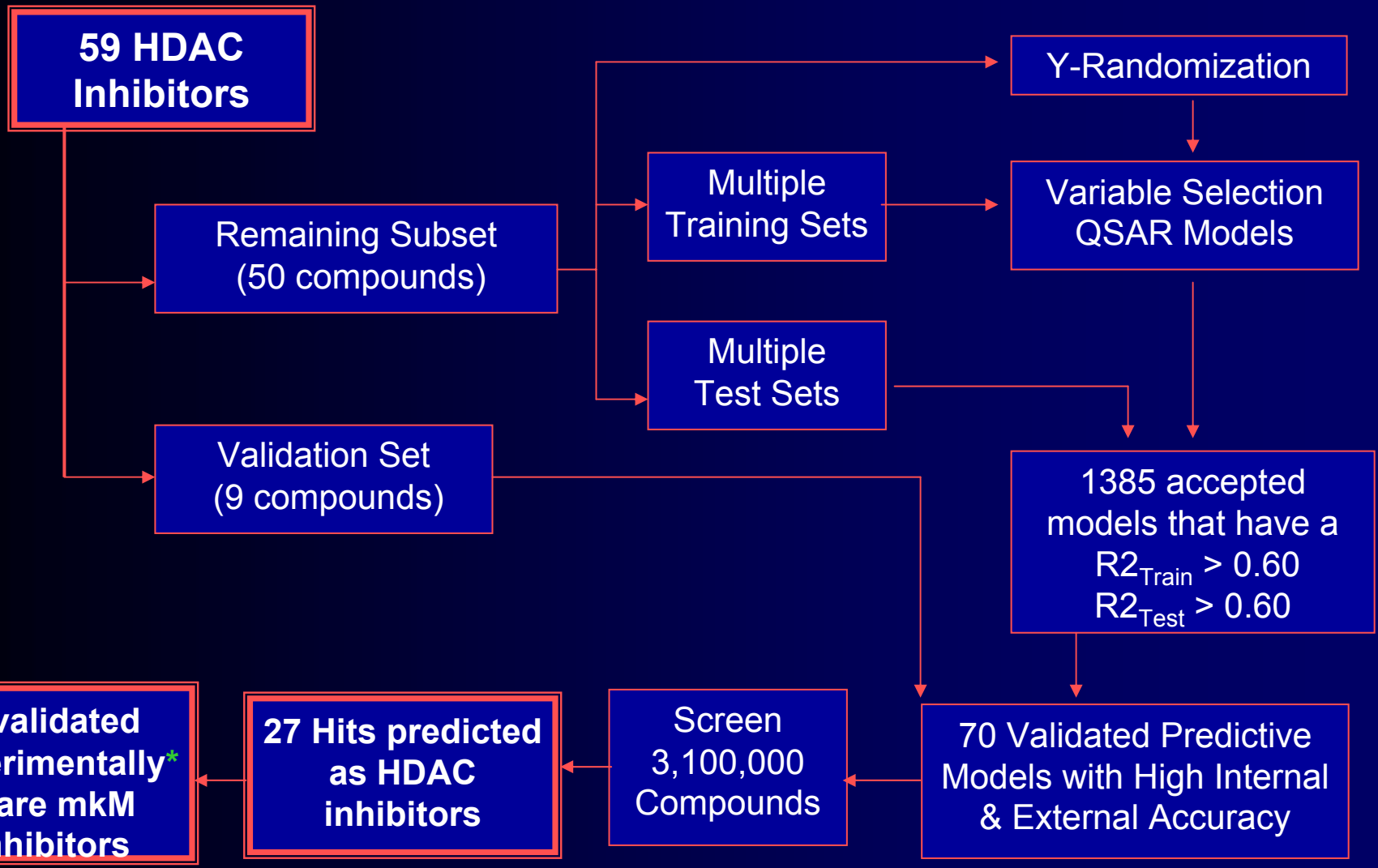
N_i is the number of distances in each category (bin)

*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)

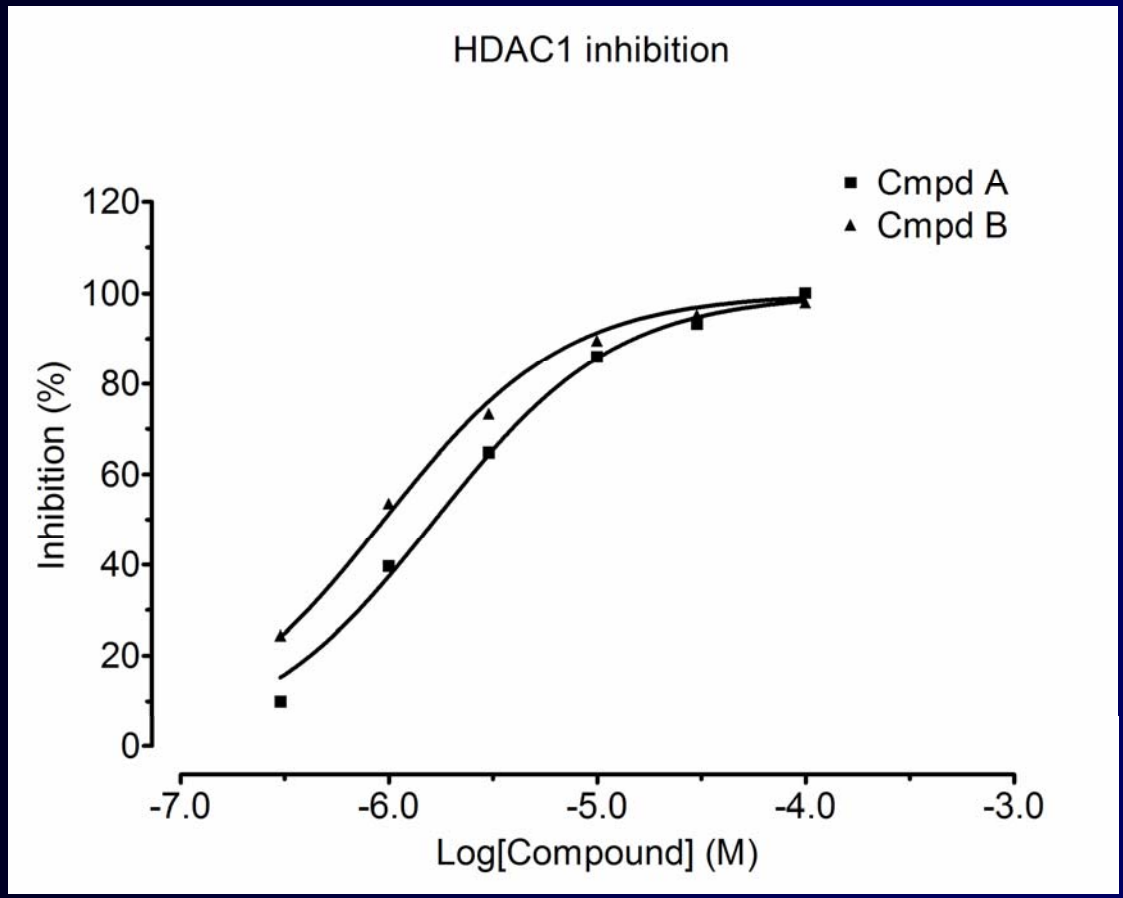


Application of the Predictive QSAR Workflow to HDAC Inhibitors*



*collaboration with Bryan Roth, UNC

Experimental validation of HDAC computational hits (data from Bryan Roth's lab, UNC-Pharmacology))



Application to GGTase-I Inhibitors

Training:

48 GGTI compounds (from -log 3.8 to 7.6)

274 descriptors (MolconnZ)

611 models generated

best q^2 = 0.77 best r^2 = 0.94

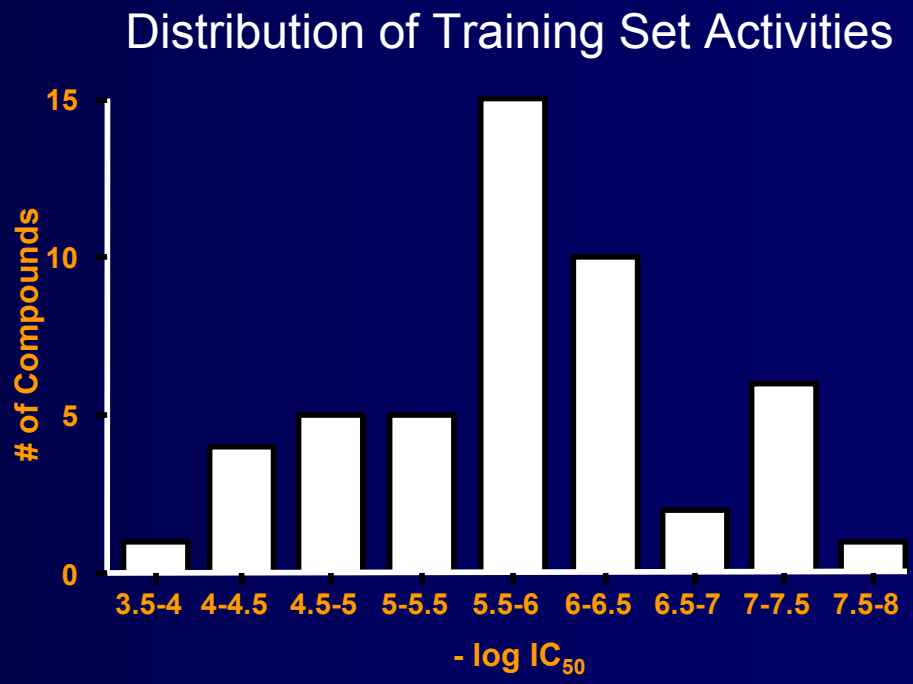
104 models pass cutoff (0.6 for q^2&r^2)

Database Search:

~9 million small molecules (ZINC + Raw Database)

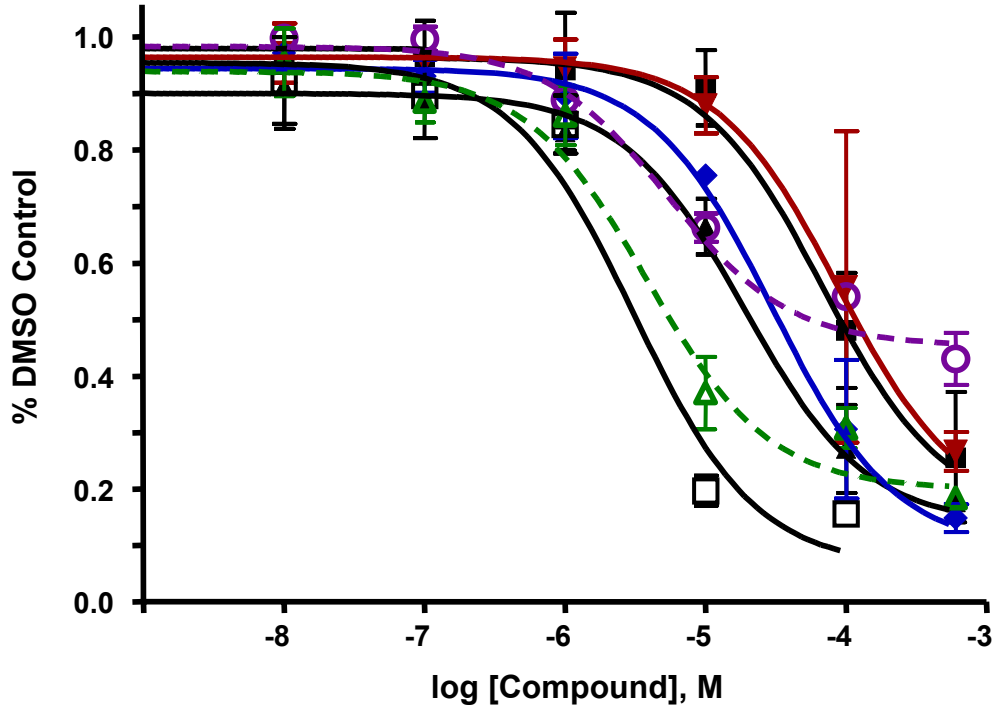
79 predicted actives using linear consensus kNN (0.5 cutoff)
range = 4.51 to 5.96 (1.45 log)
56 cmpds > 5.5 predicted activity

7 selected for further analysis based on high predicted activity, uniqueness of structure (divergent from training set), & availability



*collaboration with Y. Peterson & P. Casey, Duke

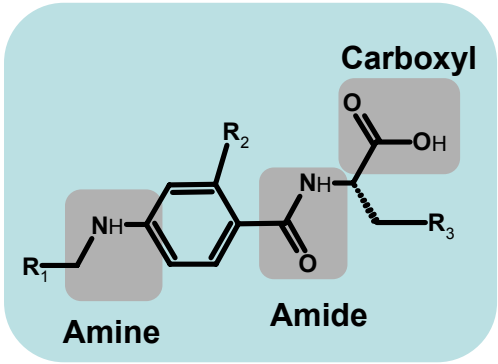
GGTI QSAR Hits – GGTase-I in vitro Activity Assay



	IC_{50} (μM)	log IC_{50}
Sig 1	139.6 +/- 76.69	-3.855
En1	122.6 +/-76.4	-3.912
As1	68.55 +/- 26.54	-4.16
En2	30.118+/- 8.473	-4.521
As2	18.91 +/- 6.84	-4.723
Sig3	3.85 +/'- 1.12	-5.415
Sig2	3.12 +/- 1.61	-5.506

Database Mining Reveals Unique Chemical Entities

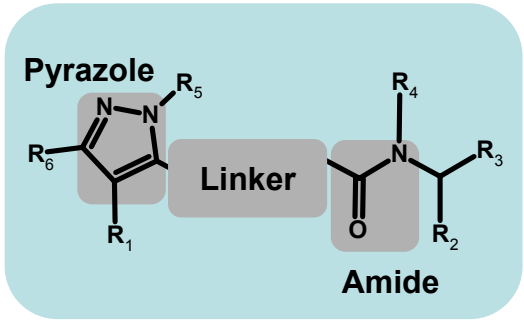
2 Training Set Scaffolds



GGTI–x Series

Peptidomimetics

Hamilton & Sebti

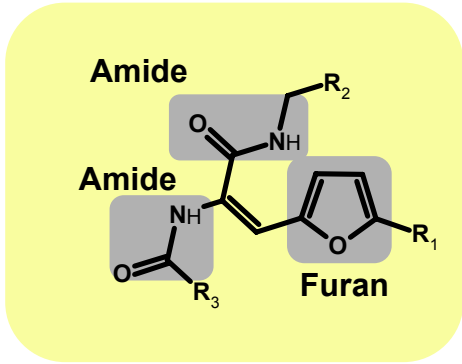


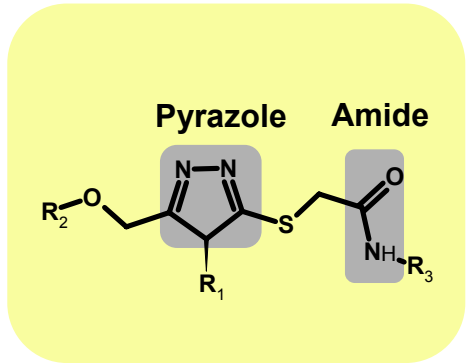
GGTI–DUx Series

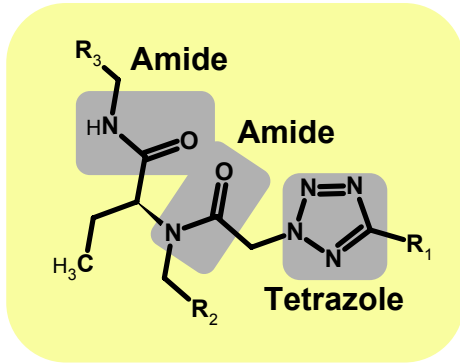
Pyrazoles

Peterson & Casey

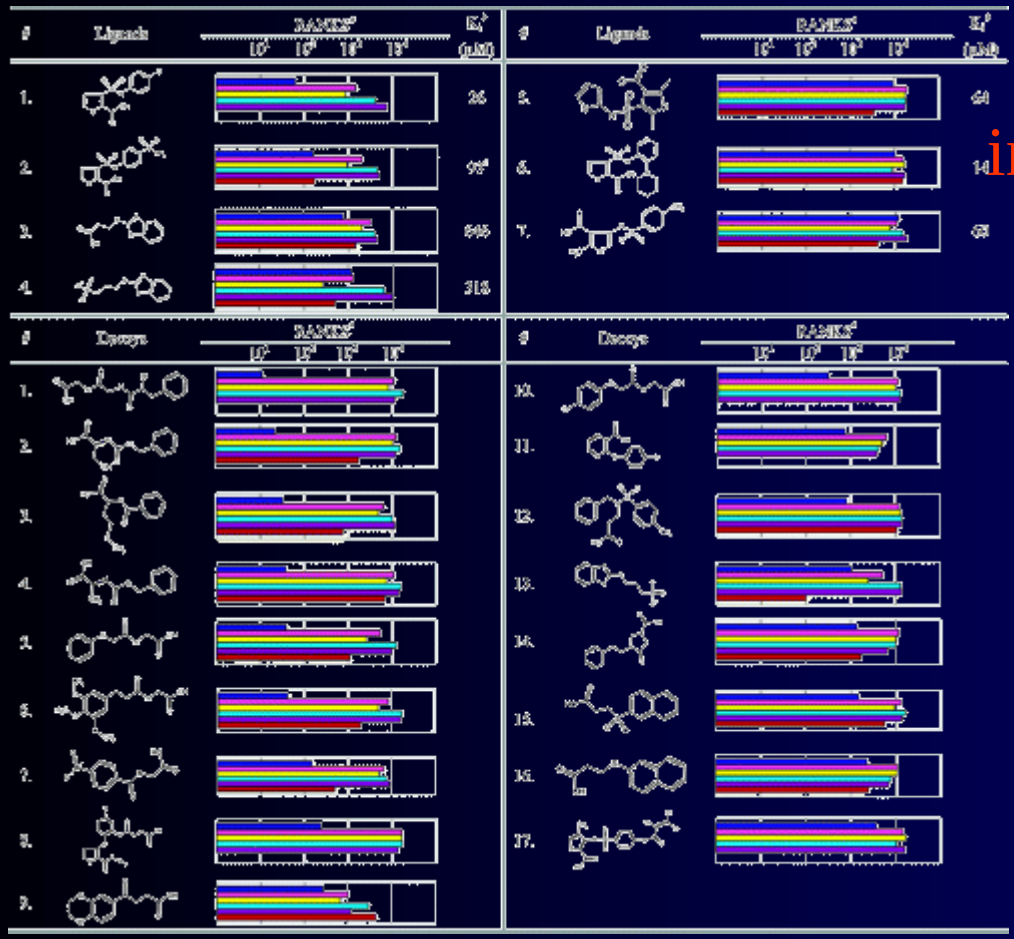
Novel Scaffolds Discovered



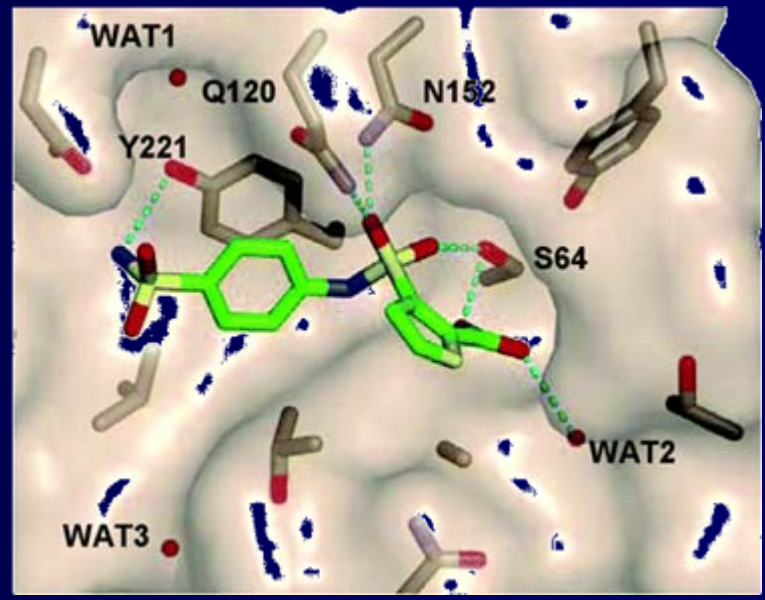




Application of Cheminformatics Approaches to Binding Decoys



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

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.
- Amp inhibitors: Zhang L. et al, *J. Comp. Aid. Molec. Des*., **2008 (ASAP)**
- HDAC inhibitors: Wang, S. *et al*, (unpublished)
- GGT-I inhibitors: Wang, Peterson, et al (provisional patent)

PubChem NCGC AmpC Dataset Reexamined by QSAR and Docking based Virtual Screening

J Comput Aided Mol Des
DOI 10.1007/s10822-008-9199-2

Differentiation of AmpC beta-lactamase binders vs. decoys using classification *k*NN QSAR modeling and application of the QSAR classifier to virtual screening

Jui-Hua Hsieh · Xiang S. Wang · Denise Teotico · Alexander Golbraikh · Alexander Tropsha

ASAP March 13, 2008

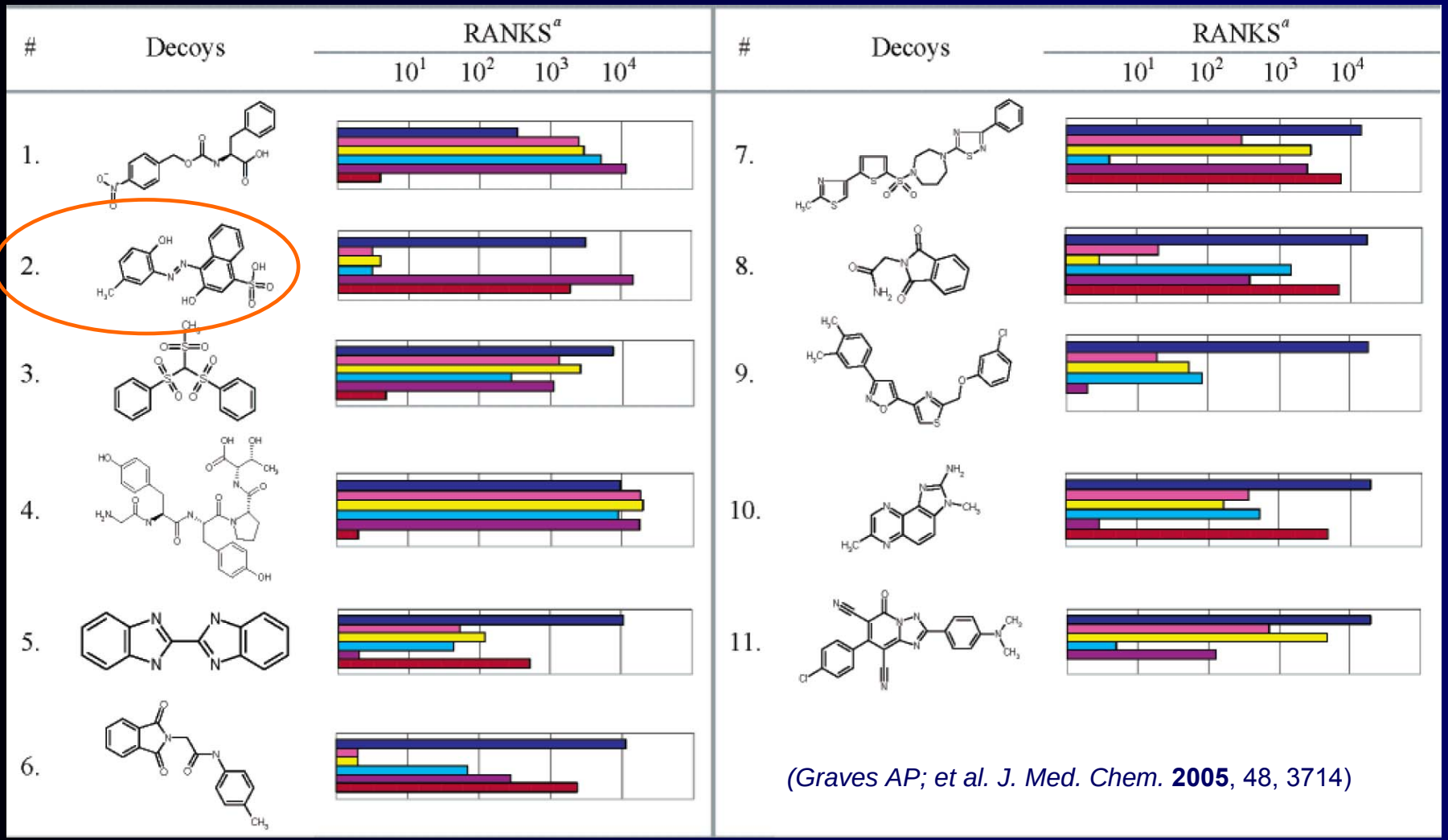
J. Med. Chem. **XXXX,** *xxx,* 000–000

ASAP March 12, 2008

Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against *β*-Lactamase

Kerim Babaoglu,[†] Anton Simeonov,[‡] John J. Irwin,[†] Michael E. Nelson,[‡] Brian Feng,[†] Craig J. Thomas,[‡] Laura Cancian,[||] M. Paola Costi,[||] David A. Maltby,[§] Ajit Jadhav,[‡] James Inglese,[‡] Christopher P. Austin,*[,‡] and Brian K. Shoichet*[,†]

QSAR modeling of binders vs. binding decoys and structure-less virtual screening



(Graves AP; et al. J. Med. Chem. **2005**, 48, 3714)

[blue:DOCK, magenta: ScreenScore, yellow: FlexX, cyan: PLP, purple: PMF, and red: SMoG]

Classification QSAR modeling of binders vs. apparent non-binding decoys.*

	T4 lysozyme L99A	AmpC β-lactamase
Inhibitors	55	21
Non-binders	65	80
Total	120	101

Inhibitors: (AmpC) compounds can inhibit competitively

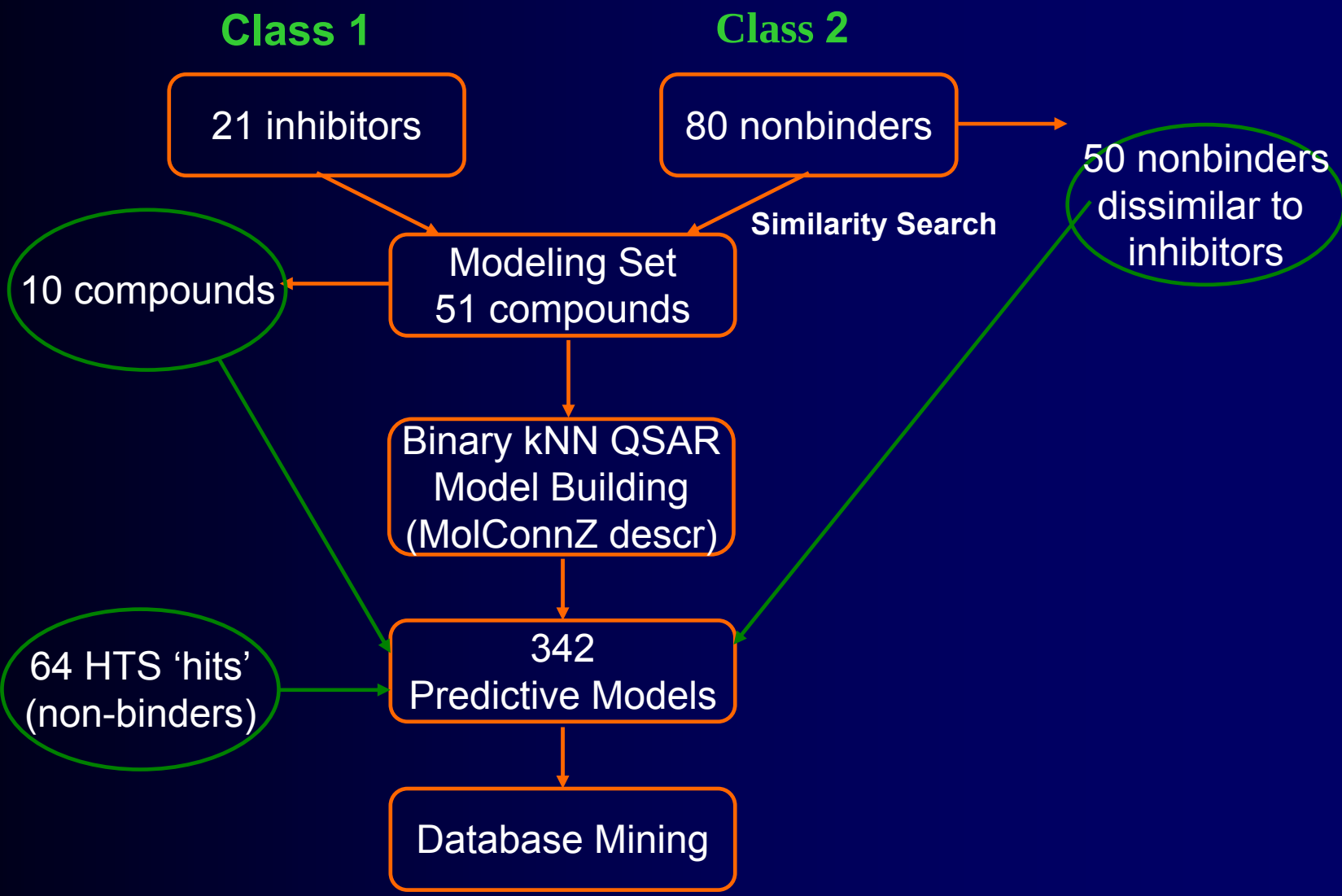
(T4) compounds are known to bind to T4 lysozyme L99A

Non-binders: (AmpC) compounds do not inhibit at 1 mM

(T4) compounds can't detect binding at high concentrations

*http://shoichetlab.compbio.ucsf.edu/take-away.php

Study Design



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

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

External Validation Results

10 randomly excluded compounds

Experimental / Predicted	Inhibitor	Nonbinder	Total
Inhibitor	5	0	5
Nonbinder	0	5	5
Total	5	5	10

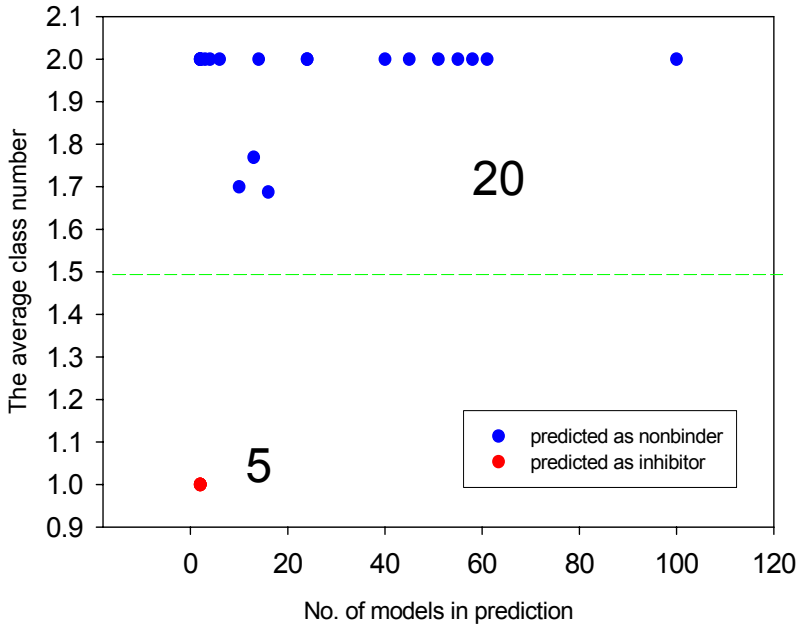
CCR = 0.5 * (5/5 + 5/5) = 1

50 nonbinders dissimilar to inhibitors

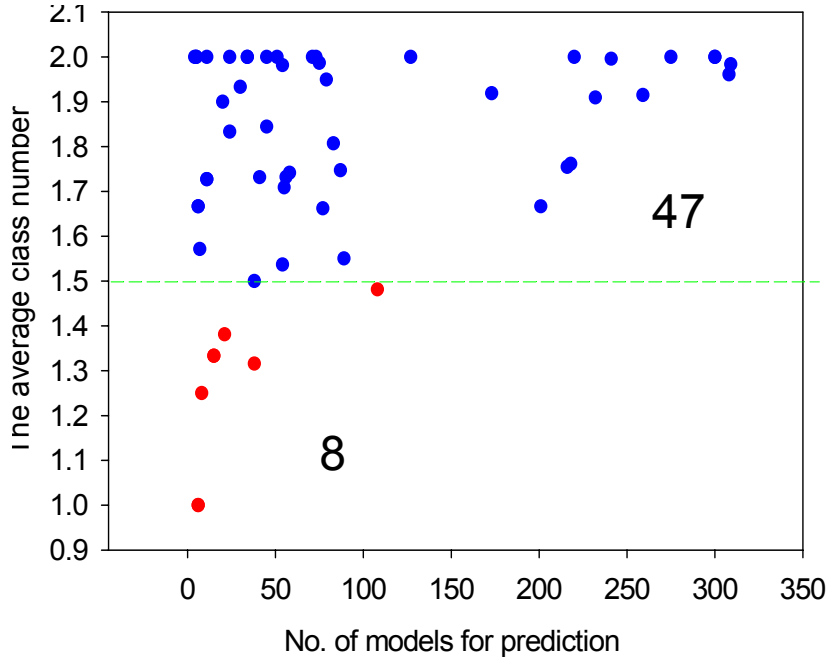
Experimental / Predicted	Inhibitor	Nonbinder	Total
Inhibitor	0	6	6
Nonbinder	0	41	41
Total	0	47	47

Accuracy = 41/47 = 0.87

The QSAR models do not predict the majority of the 64 HTS ‘Hits’ as binders



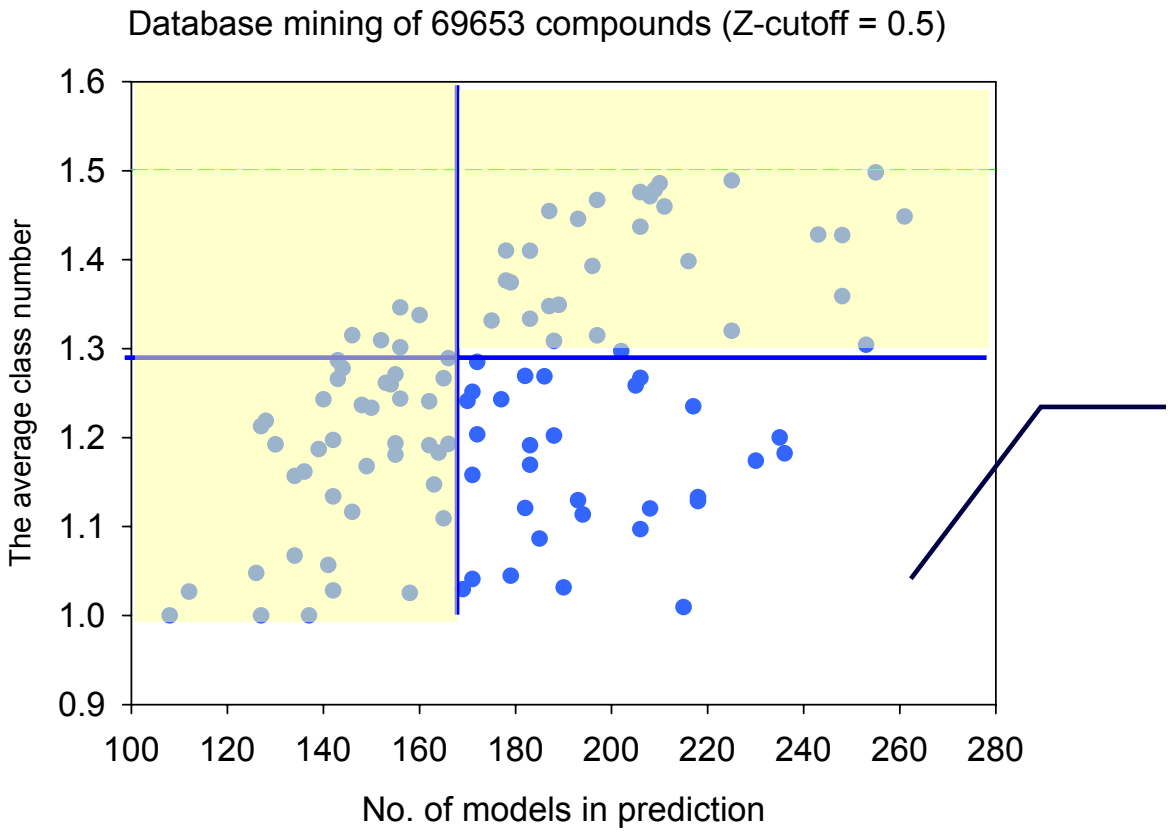
Z = 0.5; **Accuracy = 20/25 = 0.8**



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

- assigned as nonbinder
- assigned as inhibitor

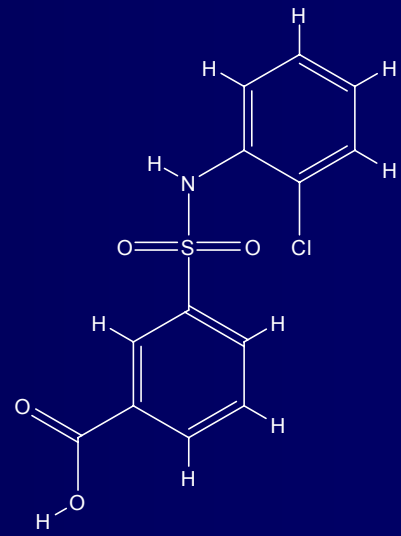
Virtual Screening of a PubChem AMPc HTS dataset of 69,653 Compounds

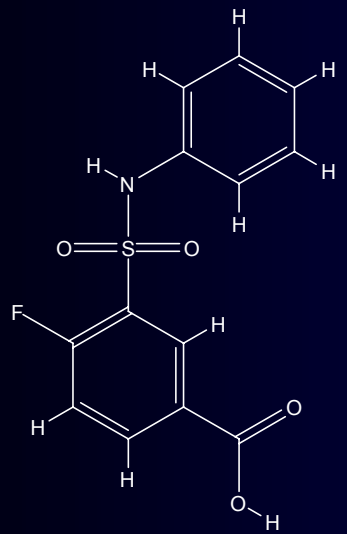


Hit selection criteria:
- Within AD of at least 50% of models
- 80% of those predict a compound as an inhibitor

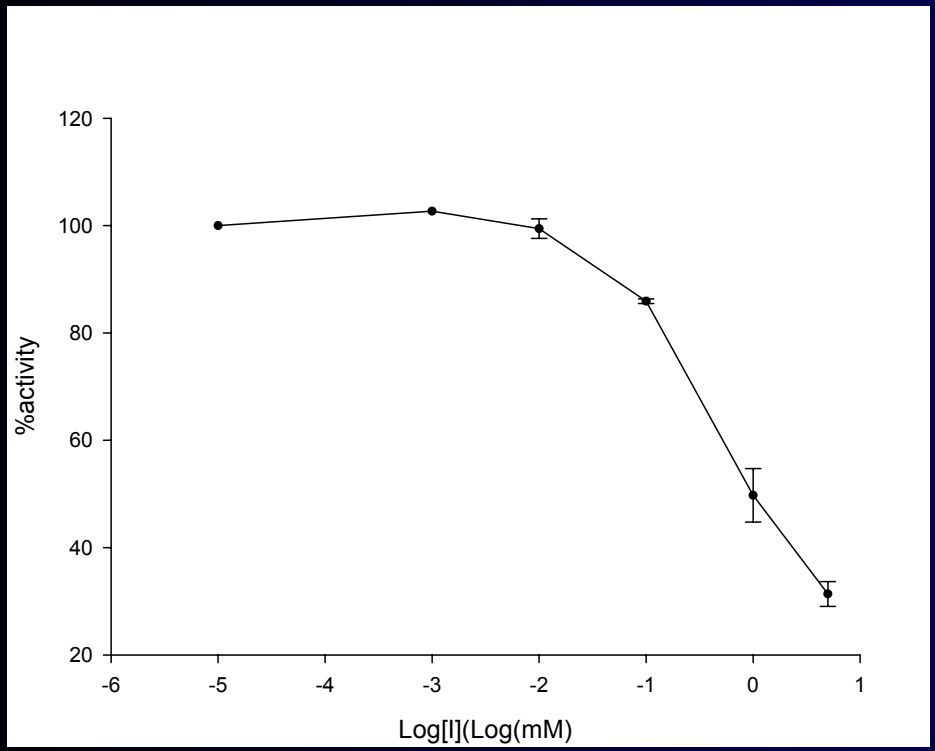
This leads to 15 Hits

5 Compounds Selected for Experimental Testing

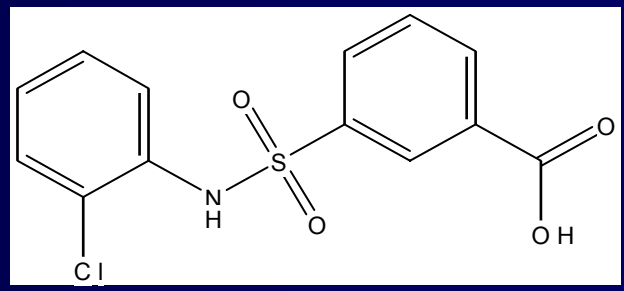




One compound (CID 69951) Shows Micro-molar Inhibitory Activity



CID: 699751

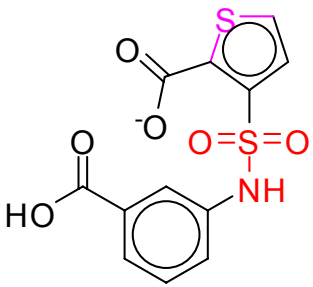


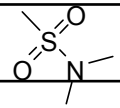
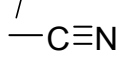
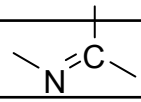
Kd = 18μM, Ki = 135μM

Experiments done by Dr. D. Teotico at UCSF

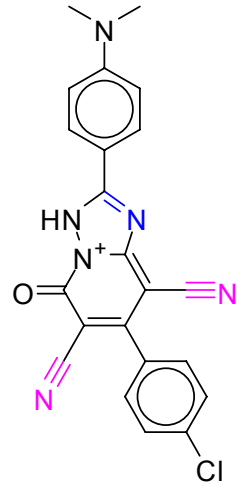
Descriptor Interpretation

inhibitor



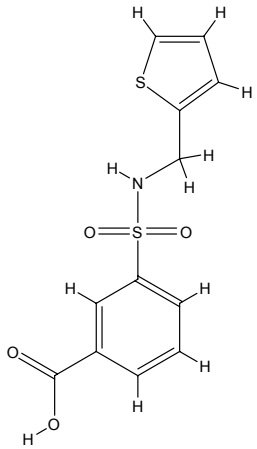
Rank	Descriptor ID	Frequency	Interpretation
1	**nHCsatu**	32.2	CHn (unsatuaed)
2	**Hsulfonamide**	28.4	
3	**nnitrile**	27.5	
4	**Hmin**	27.2	
5	**naaO**	26.3	:O:(aromatic)
6	**naaS**	26.3	:S: (aromatic)
7	**SaaCH**	26.0	:CH:
8	**n3Pad24**	26.0	
9	**SssCH2**	26.0	-CH_2-
10	**SHBint5**	25.4	
11	**Xvch5**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

nonbinder

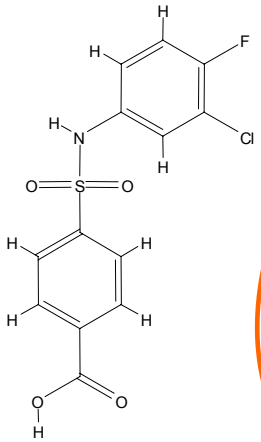


Comparison between QSAR and Docking Hits

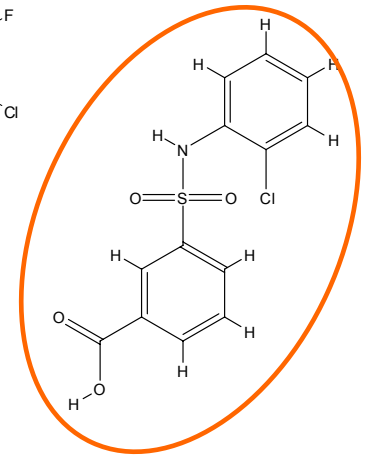
QSAR Hits



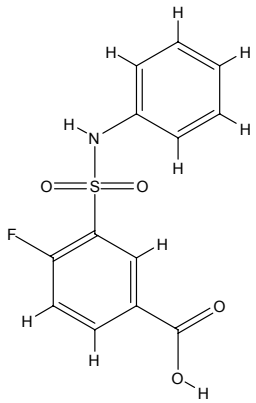
IC_{50} = 3.0mM



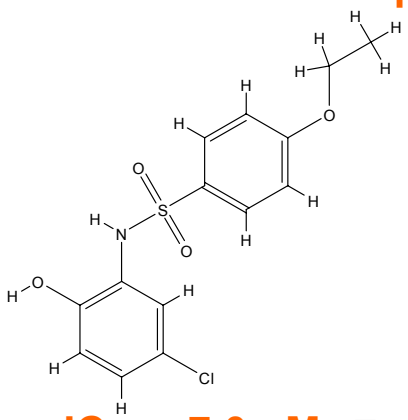
IC_{50} = 9.0mM



IC50 = 0.7mM
Ki = 135µM

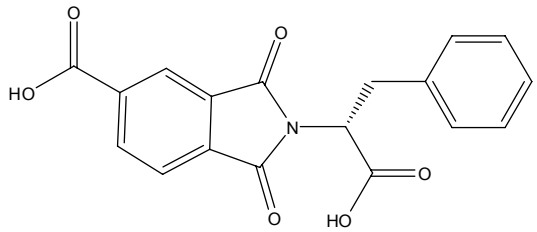


IC_{50} = 1.8mM

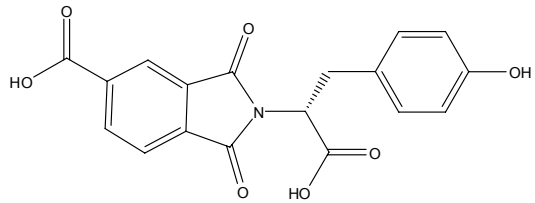


IC_{50} = 7.0mM

Docking Hits



Ki = 55µM

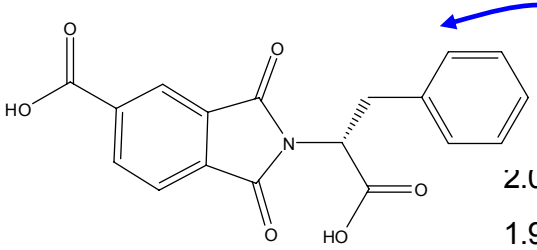


Ki = 37µM

2 true hits out of 16 tested

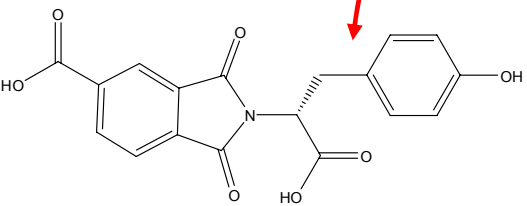
5 true hits out of 5 VS hits tested

QSAR Consensus Prediction of VS Hits

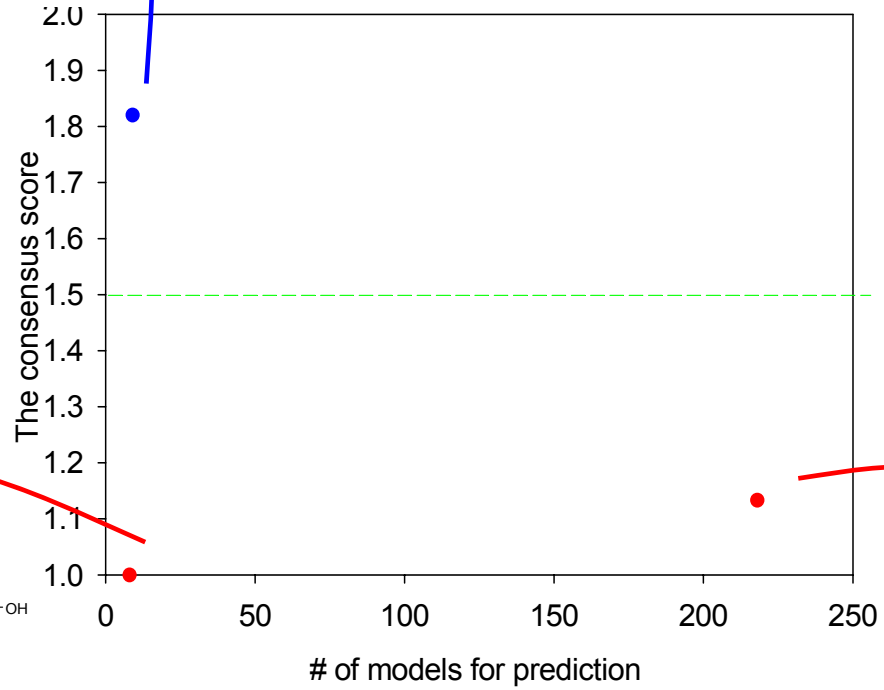


Ki = 55μM

Docking Hits



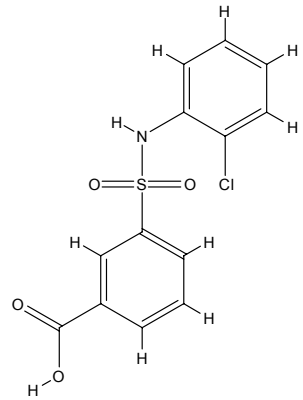
Ki = 37μM



assigned as nonbinder

assigned as inhibitor

QSAR Hits



Ki = 135μM

QSAR Modeling* of the TETRATOX aquatic toxicity endpoint

- 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, 2008, in press

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

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

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

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

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%

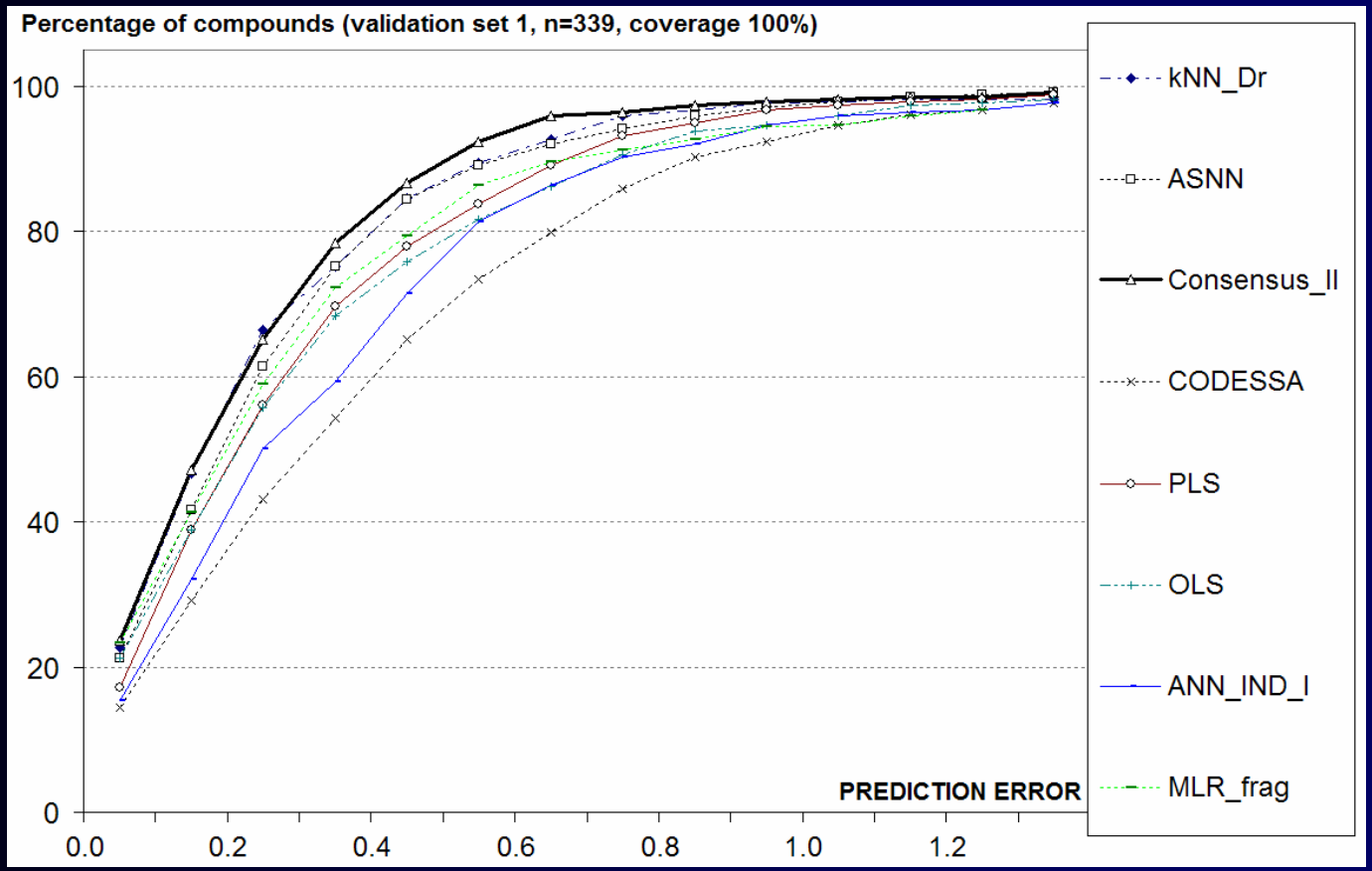
Which model is best?

- 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?

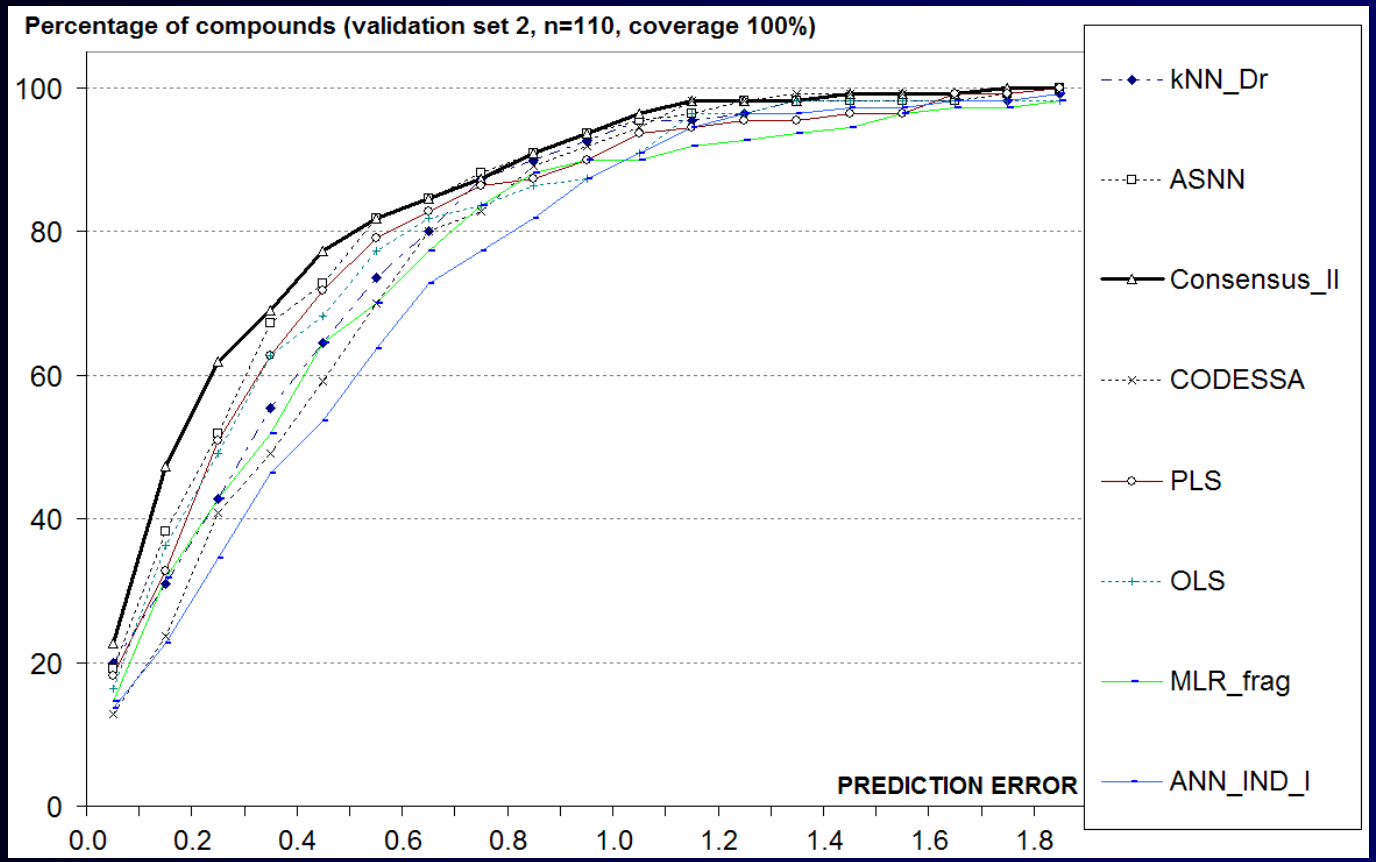
Should we choose?

- 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 all predictions.

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



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



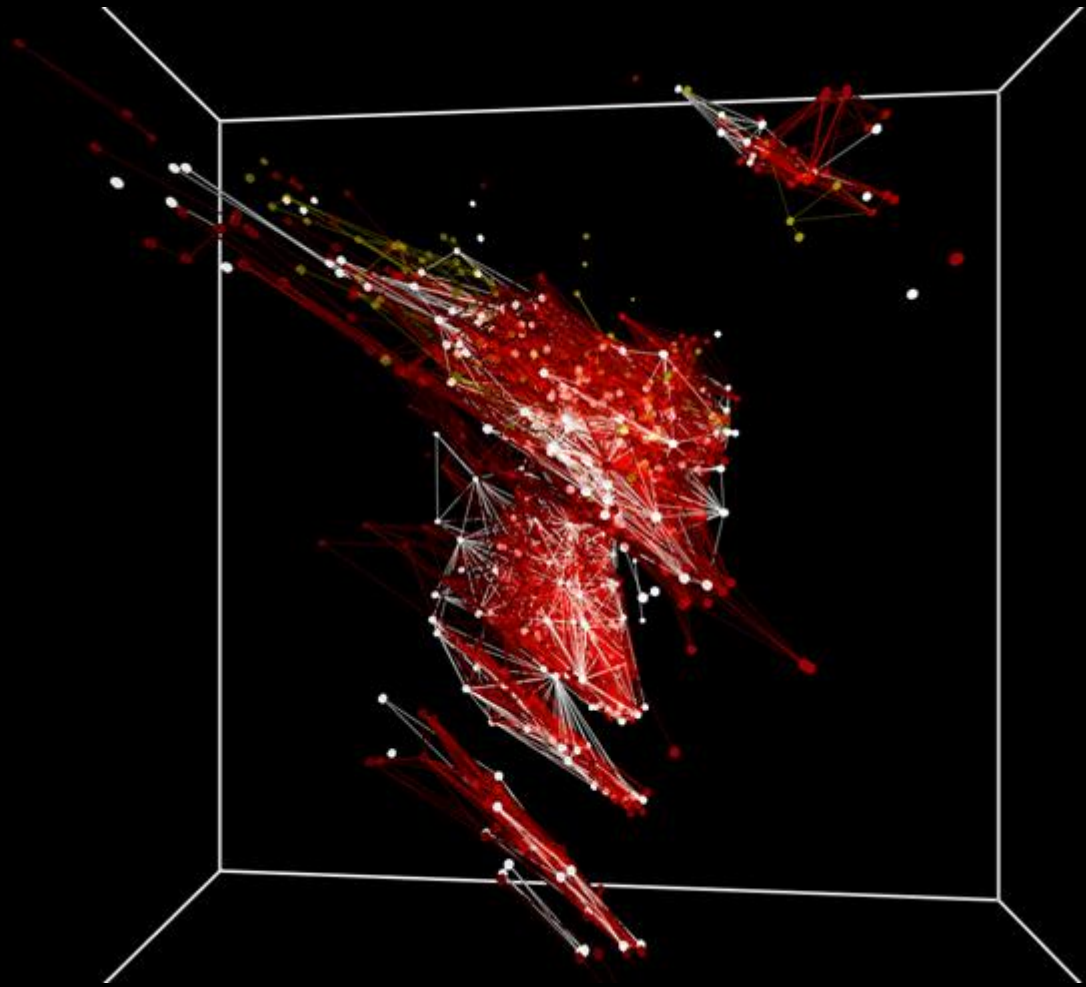
Data visualization (ROI is filed)

Calculations using 74 Dragon descriptors normalized between 0 and 1.

The three first components are visualized.

Distance cutoff : 0.7

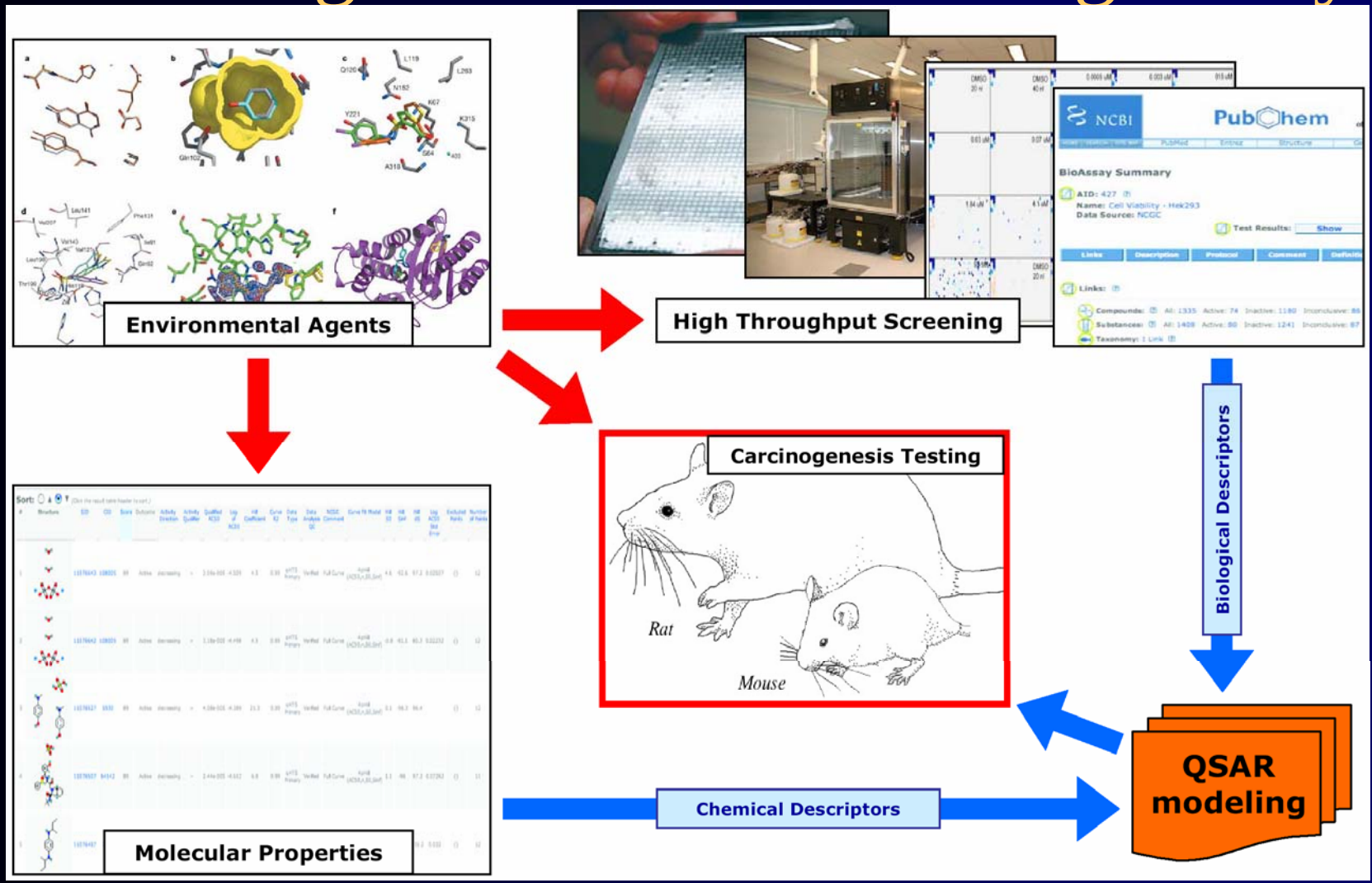
- Training Set
- Test Set 1
- Test Set 2



Conclusions of the aquatic toxicity modeling

- Training set modeling is insufficient to guarantee externally predictive models
- The use of AD is critical to achieve respectable external predictivity of individual models BUT one should keep in mind the balance between predictivity and space coverage
- Consensus prediction
  - affords high predictive power
  - has lowest MAE
  - stable against relatively inefficient individual models
  - avoids the problem of making a choice!!!

Emerging approaches: Combining chemical and biological descriptors in QSAR modeling of chemical carcinogenicity.



NTP-HTS Content Summary of 1408 Compounds

- Chemical Structure Types:
  - Organic: 1,348
  - Inorganic: 27
  - Organometallic: 19
  - No structure: 14
- 1348 Organic compounds contain:
  - Unique: 1,279
  - Complex: 51
  - Salt: 20
  - Duplicates: 53

- Curated subset: 1,289 unique organic compounds

Additional biological data on 1,289 NTP/HTS compounds*

NTP-HTS	NTPBSI	NTPGTZ	HPVCSI	CPDB	IRISSI
1,289	1,153	1,053	423	270	181

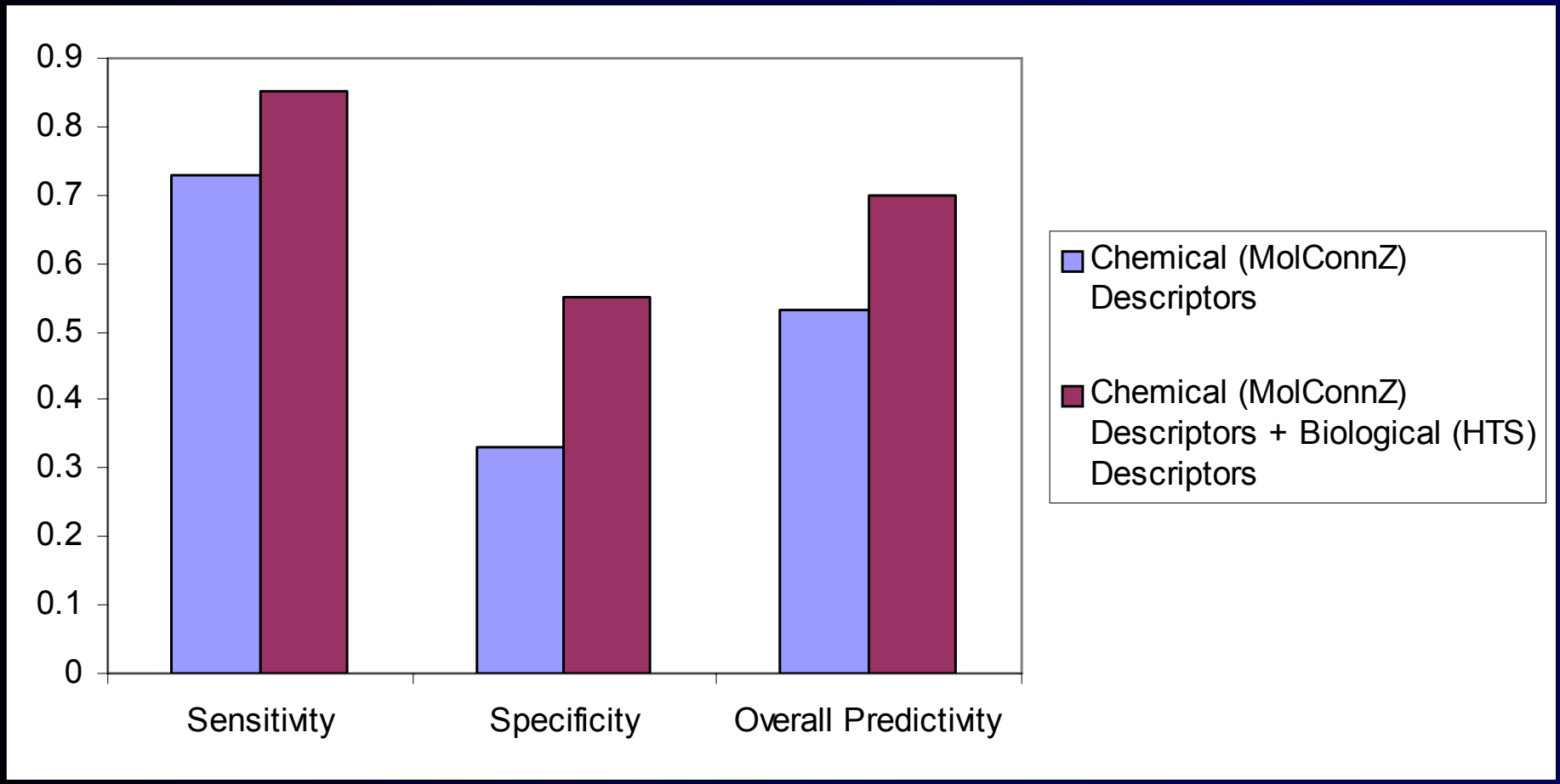
NTPBSI: National Toxicology Program Chemical Structure Index file
NTPGTZ: National Toxicology Program genotoxicity
HPVCSI: High Production Volume Chemicals
CPDB: Carcinogenic Potency Data Base All Species
IRISSI: EPA Integrated Risk Information System

*Based on the DSSTox project of Dr. Ann Richard at EPA.

The relationship between HTS activity and rodent carcinogenicity of 270 compounds

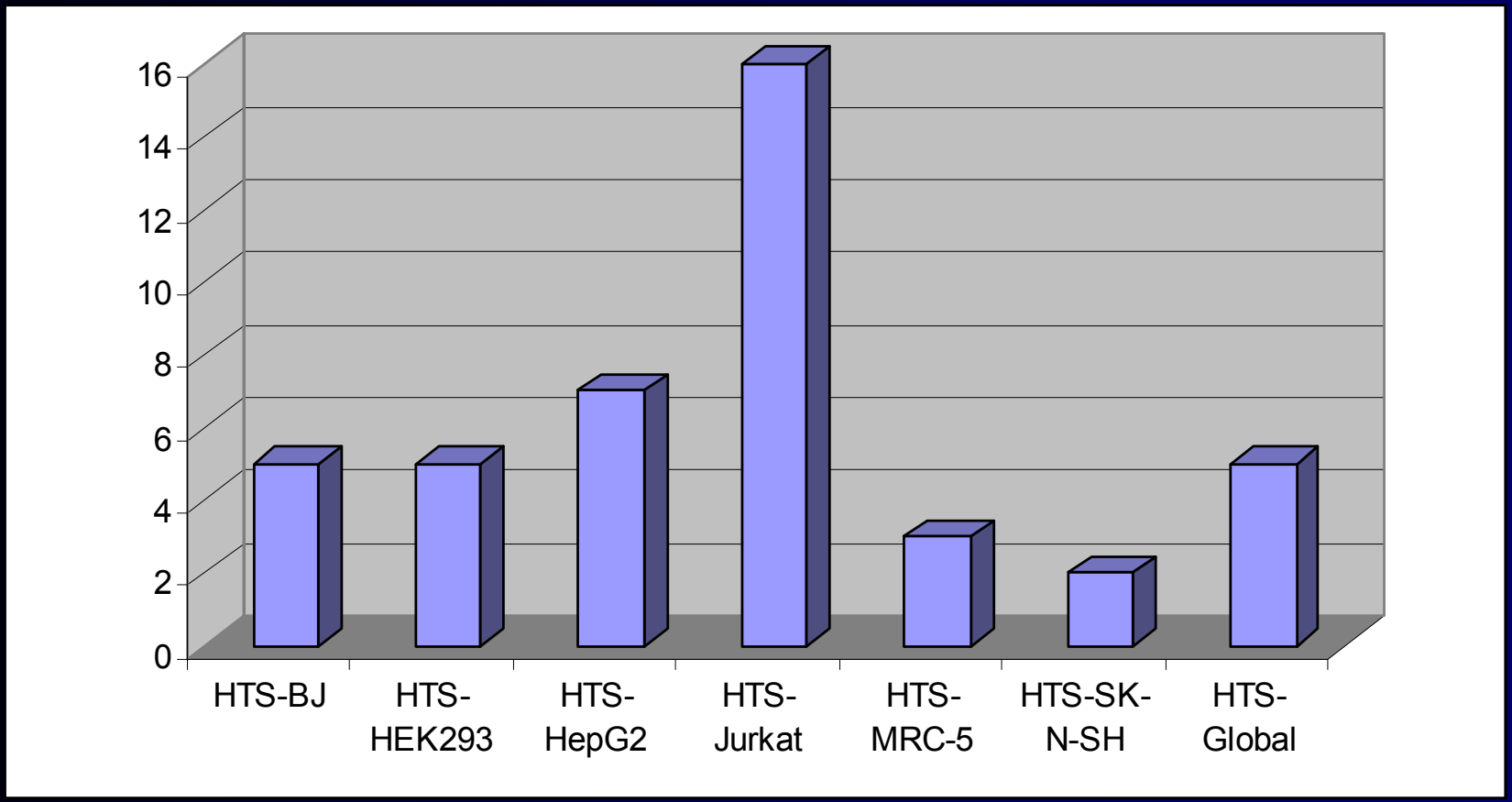
	HTS actives	HTS inconclusives	HTS inactives
CPDB actives	30	12	136
CPDB Inactives	7	11	74
Correlation	81%	-	35%

Comparison between Predictive Power of QSAR Models using Conventional vs. Hybrid Descriptors.



Zhu et al, EHP, 2008, in press

Relative contributions of HTS descriptors to 34 acceptable models



Final Thoughts

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
  - Genotype - phenotype correlations
- Focus on Knowledge Discovery (accurate testable predictions!) in Biomolecular Databases
- Practice best practices! Collaborate!!!

ACKNOWLEDGMENTS

UNC ASSOCIATES

Former:

-Stephen CAMMER
-Sung Jin CHO
-Weifan ZHENG
- Min SHEN
-Bala KRISHNAMOORTHY
-Shuxing ZHANG
–Peter ITSKOWITZ
–Scott OLOFF
-Shuquan ZONG
-Raed KHASHAN

– Jun FENG
– Yun-De XIAO
–Yuanyuan QIAO
–Ruchir SHAH
–Patricia LIMA
–Assia KOVACHEVA
–Julia GRACE
–Hao HU

Current

•Structural bioinformatics group:

– Yetian CHEN
– Tanarat KIETSAKORN
– Berk ZAFER

Cheminformatics group:

- Kun WANG
– Sasha GOLBRAIKH
– Raed KHASHAN
– Chris GRULKE
- Hao TANG
- Simon WANG
- Hao ZHU
- M. KARTHIKEYAN
- Rima HAJJO
- Mei WANG
- Lin YE
- Liying ZHANG
- Mihir SHAH
- Jui-Hua Hsieh
- Tong-Ying Wu

- Collaborators
  - Hal Kohn (UNC)
  - Richard Mailman (UNC)
  - Bryan Roth (UNC)
  - Yuri Peterson (Duke)
  - Diane Pozefsky (UNC)
- Funding
  - NIH
    - P20-HG003898 (RoadMap)
    - R21GM076059 (RoadMap)
    - R01-GM66940
    - GM068665
  - EPA (STAR award)