

CAN PRIMARY HIGH-THROUGHPUT SCREENING DATA BE ANALYZED IN A MEANINGFUL WAY?

Alexander Tropsha

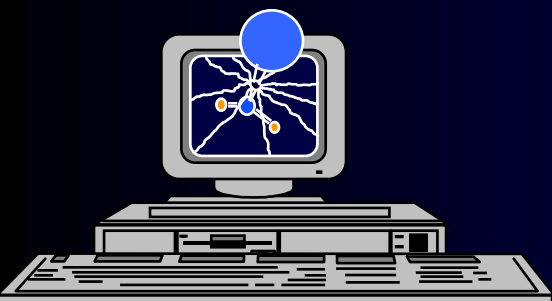
Laboratory for Molecular Modeling

and

Carolina Center for Exploratory
Cheminformatics Research

School of Pharmacy

UNC-Chapel Hill



Meaningful data analysis?

- Why Bother?

- we can not screen every important chemical entity
- screening hit rate is very low

- Who Cares?

- experimentalists need tools for data handling
- there is a need in accurate predictive models of (historic) data (QSAR analysis) that guide the (future) experiment

- And So What?

- modern QSAR modeling can serve as a decision support tool
- rigorous models afford biological data imputation

EU-WHITE PAPER on the Strategy for a Future Chemicals Policy (2001)*

Art. 3.2 ... "to keep animal testing to a minimum"

"in the interest of time- and cost-effectiveness"...

"particular research efforts are needed for development and validation of modelling (e.g. QSAR) and screening methods for assessing the potential adverse effects of chemicals"

➤ "Regulatory acceptance of QSAR models":

- Workshop ICCA/CEFIC (2002):



Setubal Principles



New Pathways to Discovery

- ▶ [Building Blocks, Biological Pathways, and Networks](#)
- ▶ [Molecular Libraries and Imaging](#)
- ▶ [Structural Biology](#)
- ▶ [Bioinformatics and Computational Biology](#)
- ▶ [Nanomedicine](#)

Research Teams of the Future

- ▶ [High-Risk Research](#)
 - [NIH Director's Pioneer Award](#)
- ▶ [Interdisciplinary Research](#)
- ▶ [Public-Private Partnerships](#)

Re-engineering the Clinical Research Enterprise

- ▶ [Re-engineering the Clinical Research Enterprise](#)

- <http://nihroadmap.nih.gov/>

- *Molecular Library Screening Center Network (MLSCN)*

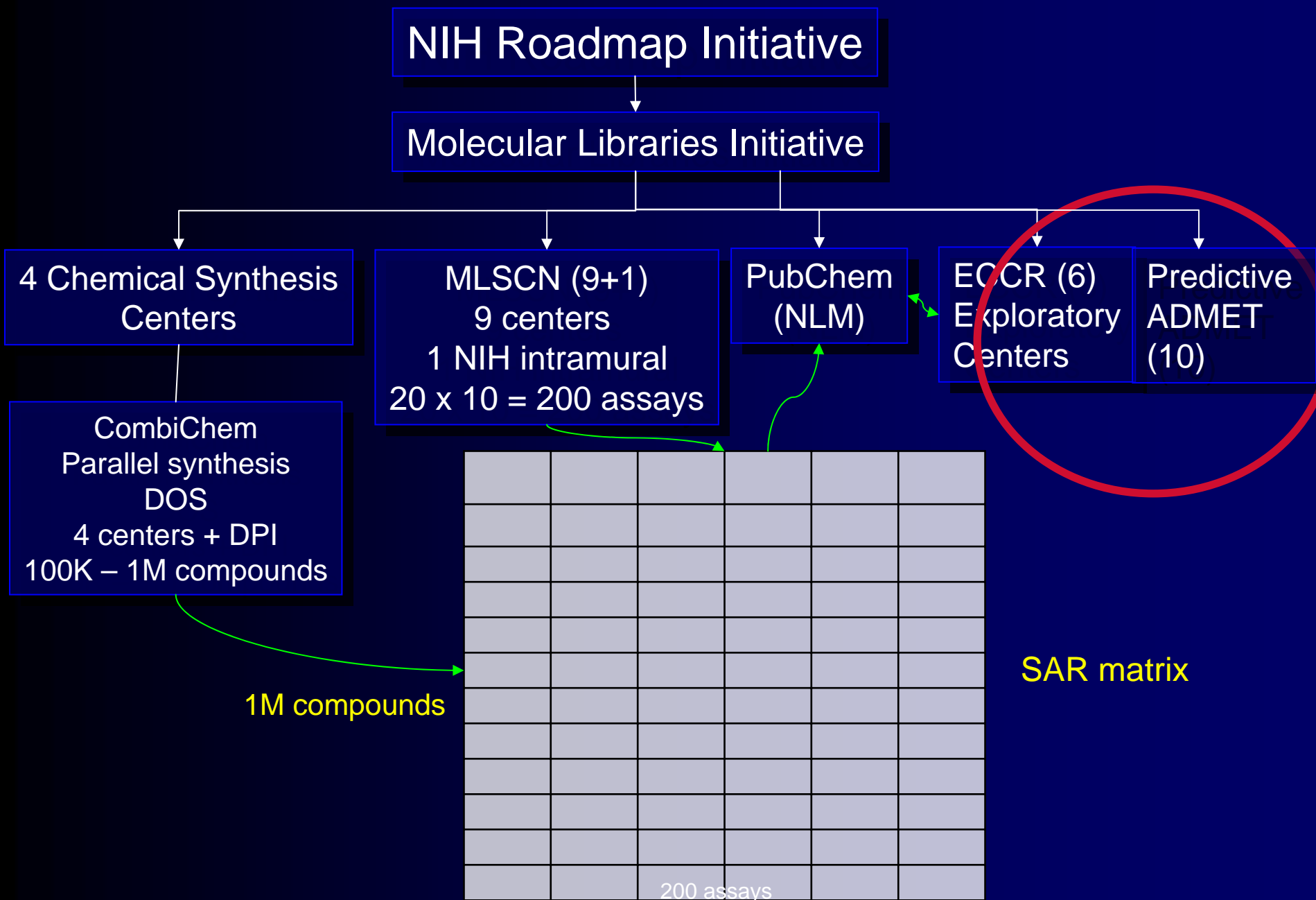
- *Screening Centers*
 - *Admin by NHGRI & NIMH*
- *Compound Repository-Contract*
- *PubChem-NLM*

- *Cheminformatics*

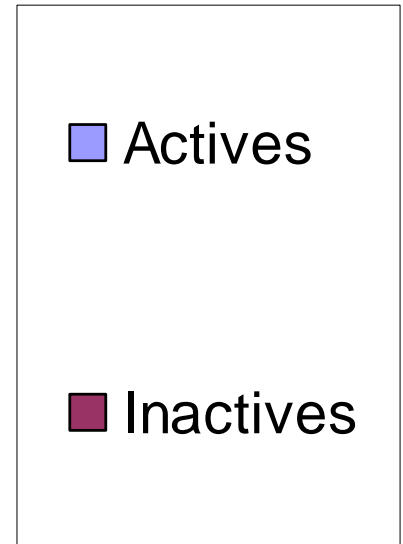
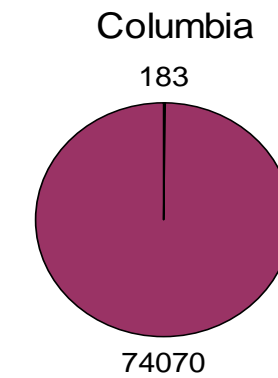
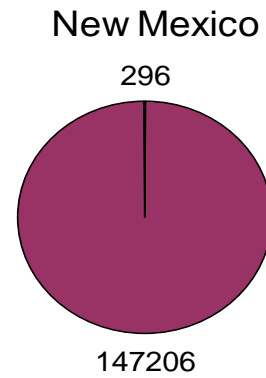
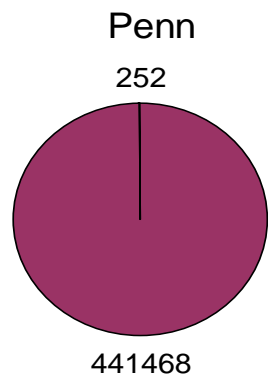
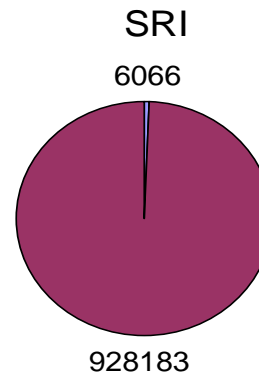
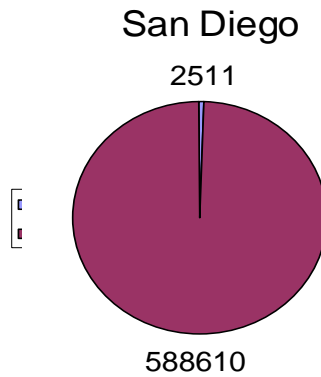
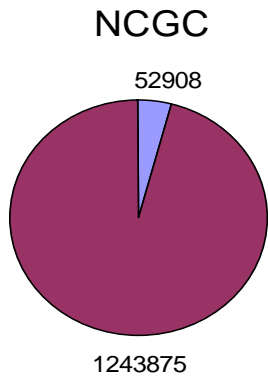
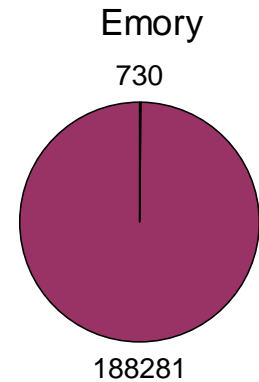
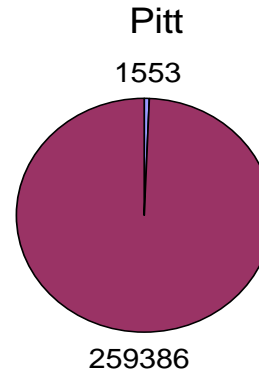
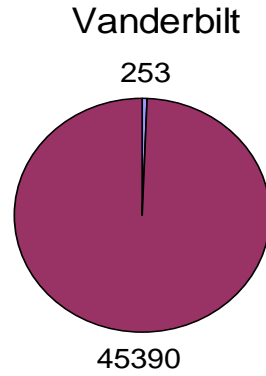
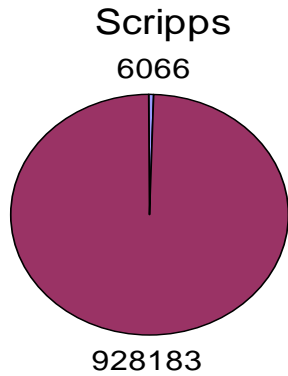
- *Technology Development*

- *Chemical Diversity*
- *Assay Diversity*
 - *Funded research examples* →
- *Instrumentation*

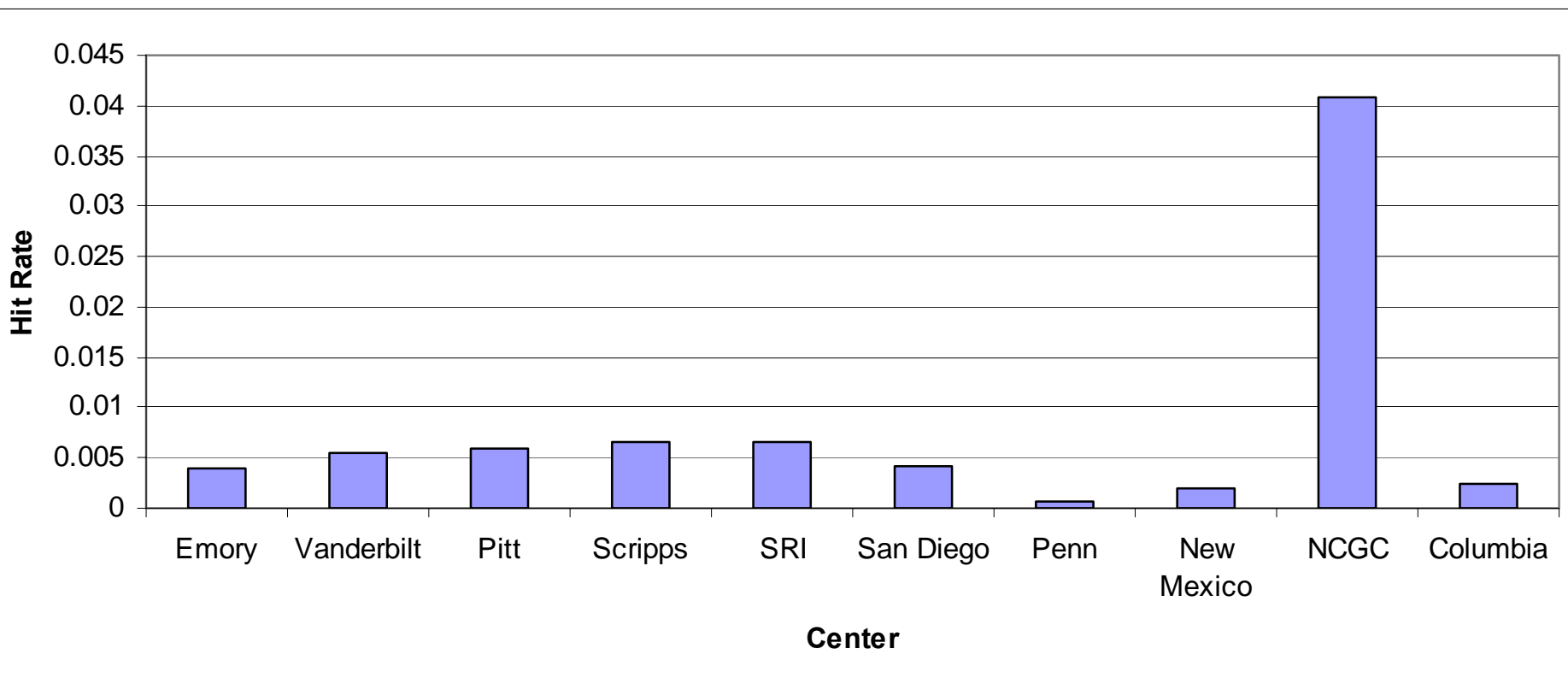
NIH's Molecular Libraries Initiative in Numbers



MLSCN Assay Results (by the Center)



Average MLSCN Hit Rates

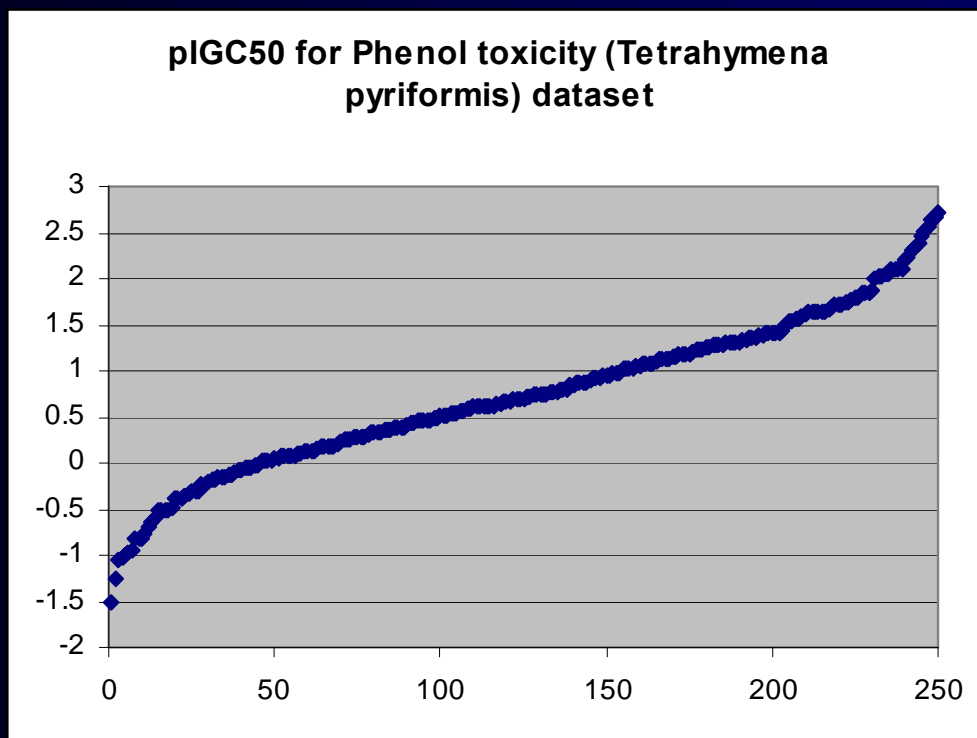


Note: 4 of 35 Assays for NCGC account for 92% of their hits. Without these screens the hit rate for NCGC = 0.004

Modern Experimental Datasets Present Multiple Challenges for their Analysis

	FIVE - TEN YEARS AGO	NOW
Availability (!)	Poor	GREAT (e.g., PubChem)
Size	30-300	30-5000 and more
Diversity	Usually relatively low diversity (several groups of similar compounds)	Usually very high
Number of Activities	Usually one	Sometimes Multiple
Type of Activity	Usually continuous	In many cases categorical
Distribution of activities for continuous QSAR	Usually uniform or close to normal	In many cases non-uniform and non-normal
Distribution between classes for category QSAR	Usually more or less balanced	In many cases highly unbalanced

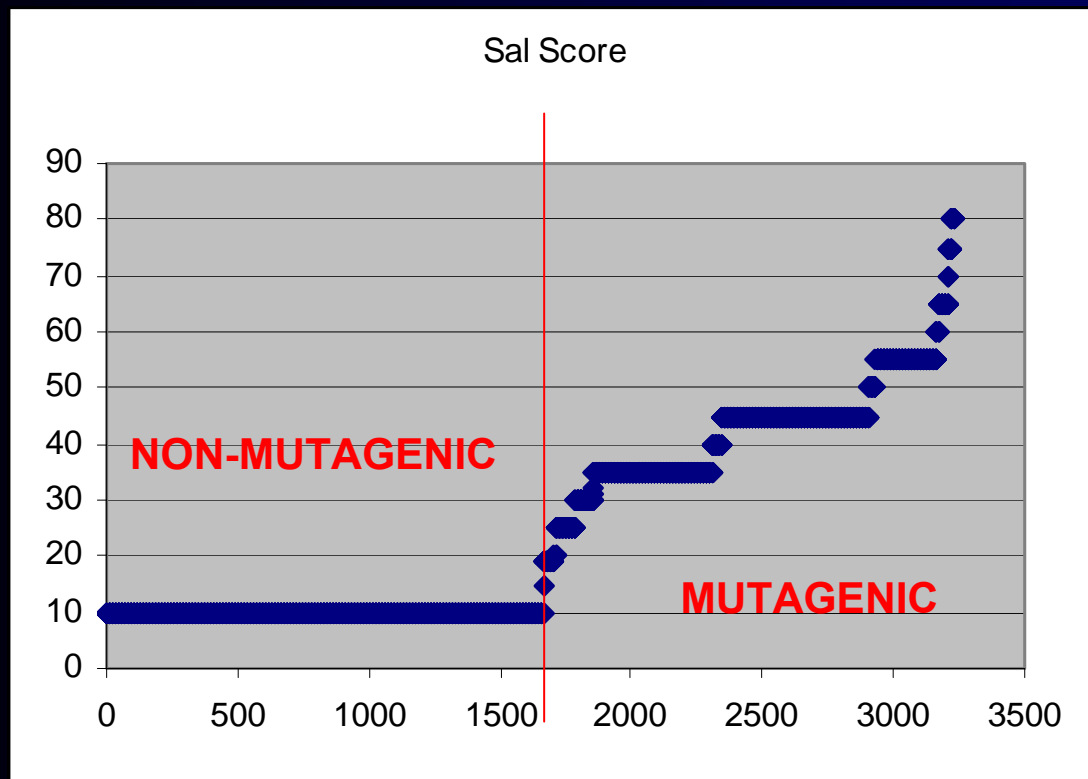
Typical activity distribution for a small or average size dataset for QSAR studies (for older datasets)



Cronin MTD et al. Chemosphere 2002, 49, 1201-1221.

Transformation of Modern Datasets from Continuous to Categorical

- 3229 Salmonella mutagenicity dataset compounds (after cleaning). Can be divided into mutagenic (1442 comp) and non-mutagenic (1787 comp)



**ACTIVITY (Sal Score):
degree of mutagenicity**

10 – non-mutagenic

80 – very mutagenic

CURRENT VERSION:

AMES dataset:¹

4337 compounds

2401 mutagenic

1936 nonmutagenic

¹ Kazius, etc. J. Med. Chem.
2005, 48 (1), 312 - 320

QSAR as the Primary Data Modeling Approach

- **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

Typical modern datasets

- 1948 hERG Potassium channel openers-closers:
 - 58 openers
 - 191 closer
 - 1700 inactive
- 1408 NTP-HTS dataset compounds (Data for six cell lines)
 - overlap with CPDB compounds: 320
 - 270 among them positive for more than one cell line
 - 199 – in more than one organ
 - 190 – in more than one species
 - 249 – for both sexes.

QSAR Modeling

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	"	"	"		"

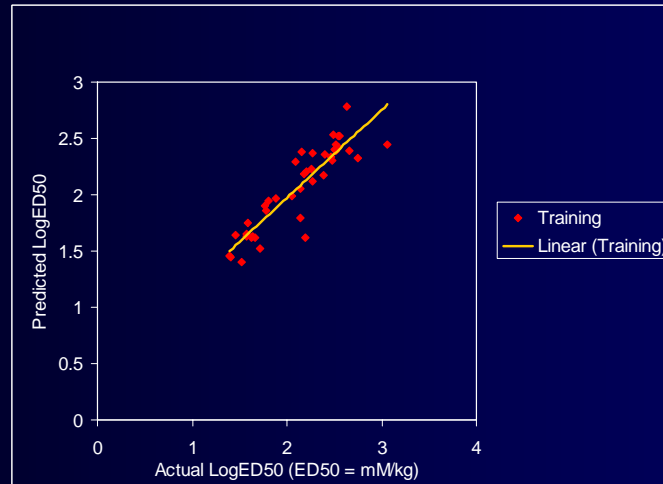


$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\bar{y} - y_i)^2}$$

BA = F(D) {e.g., ...}

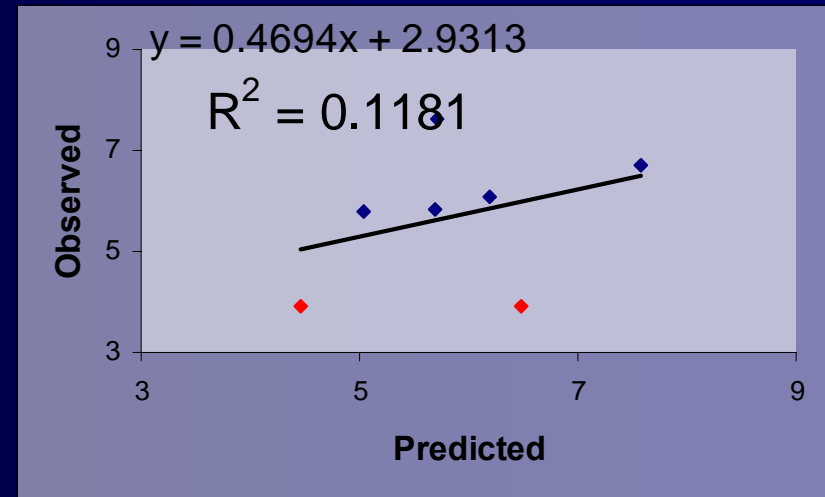
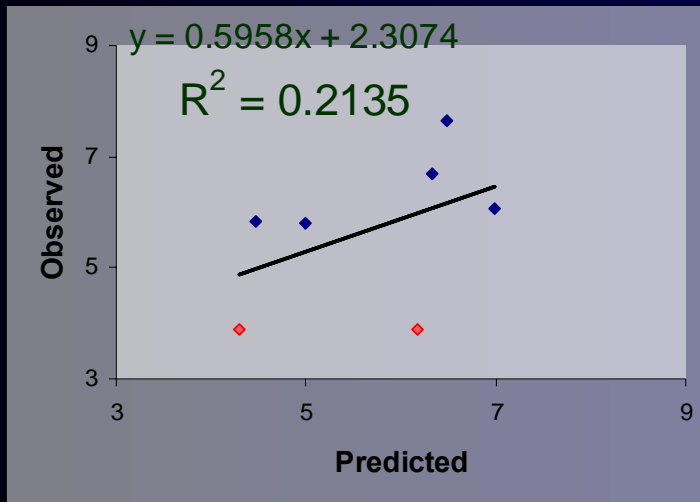
(e.g., $-\text{LogIC50} = k_1D_1 + k_2D_2 + \dots + k_nD_n$)

The unbearable lightness of model building...

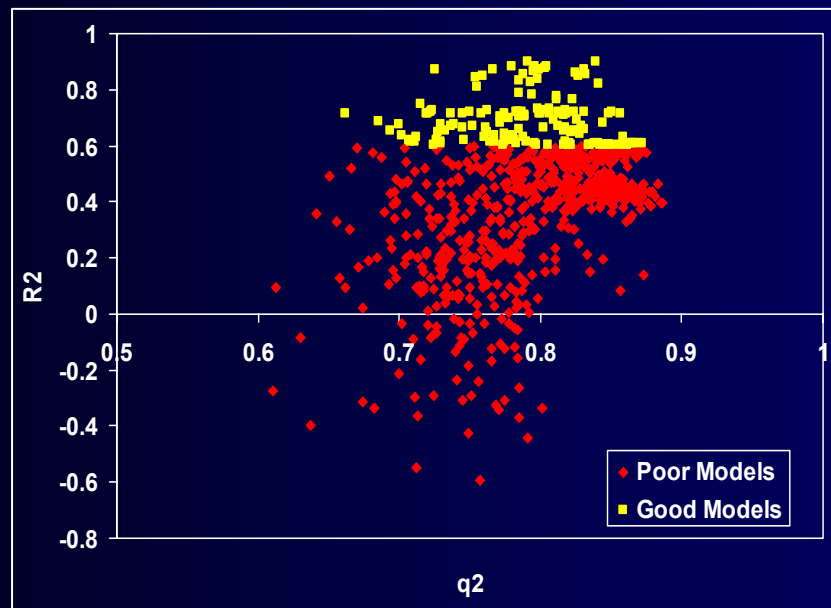


...leads to unacceptable prediction accuracy.

EXTERNAL TEST SET PREDICTIONS



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.

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

- Descriptors? No, we have plenty (e.g., Dragon)
- Datamining methods? No, we also have plenty
- 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 is an empirical data modeling approach: just do it any (all) way you like but **VALIDATE** on independent datasets!

QSPR modeling process revisited



Pharmaco-

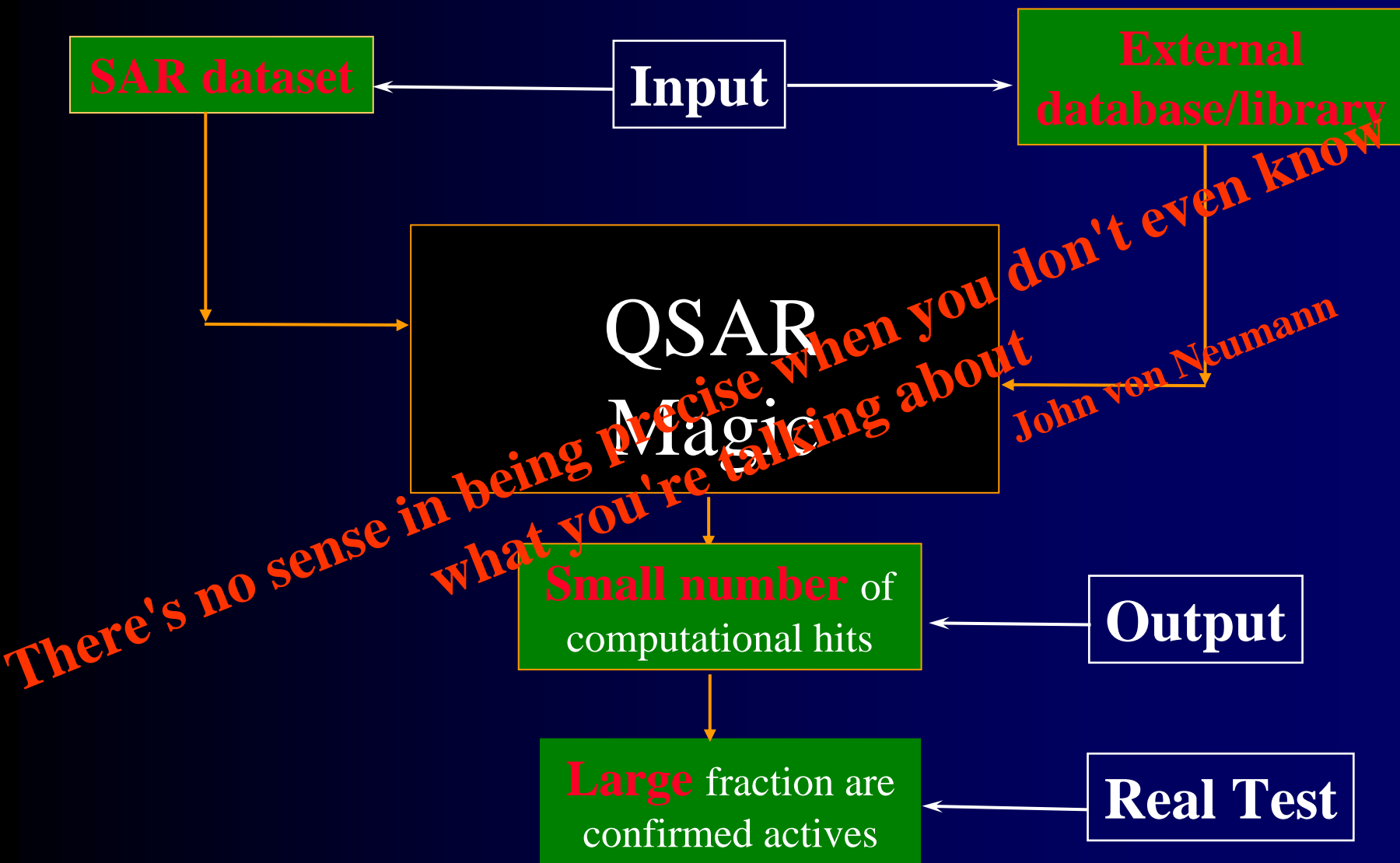
GENET-
GENOM-
PROTEOM-
BIOINFORMAT-
MEDINFORMAT-
CHEMOGENOM-
CHEMOINFORMAT-
PROTEOCHEMOMETR-

-ICS

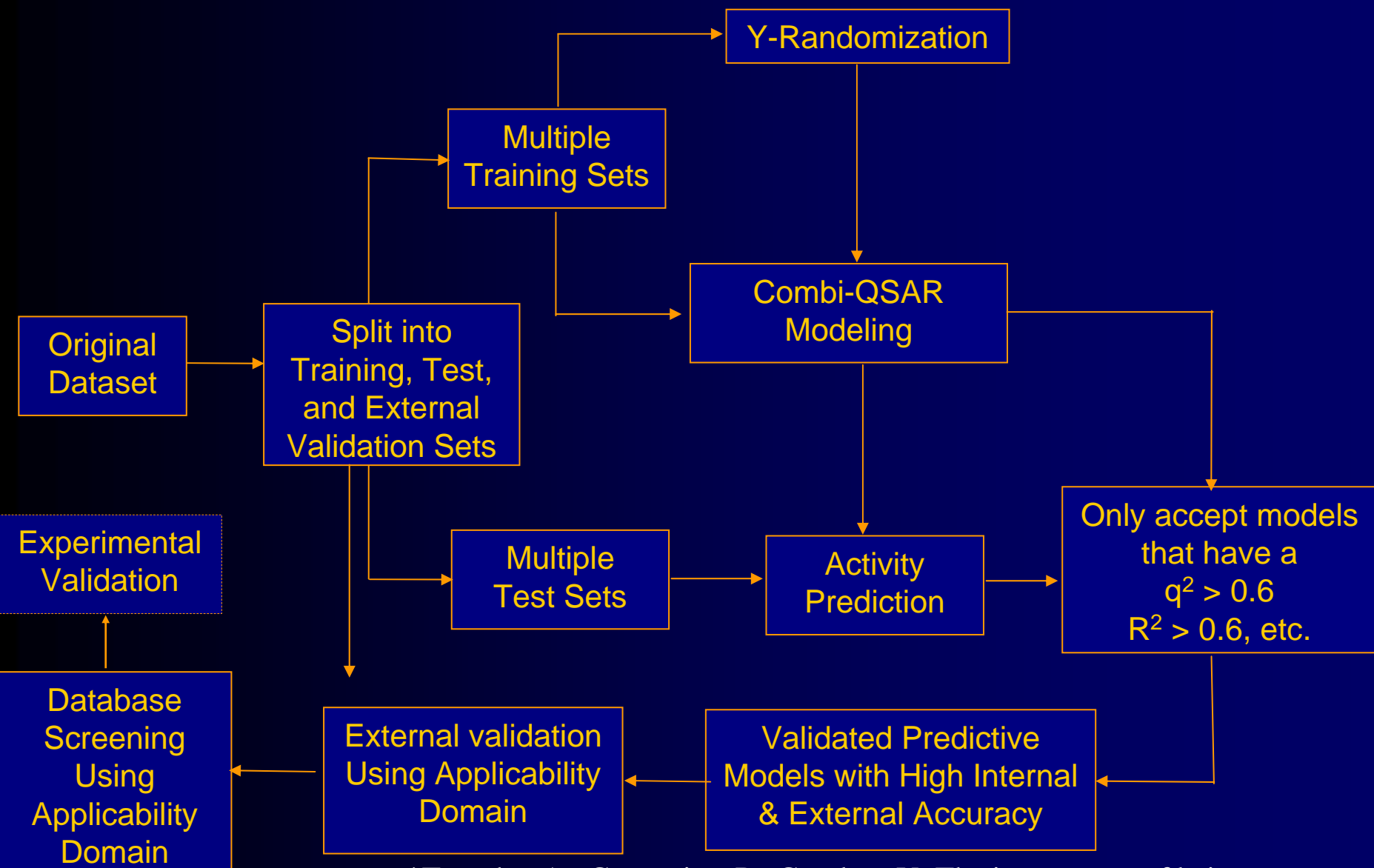
“-ics” – an old Latin suffix that means “way too much”

COMBINATORIAL QSAR-omics, or
C-Qics

Key point: Emphasis on Successful Predictions, not Statistics or Interpretation

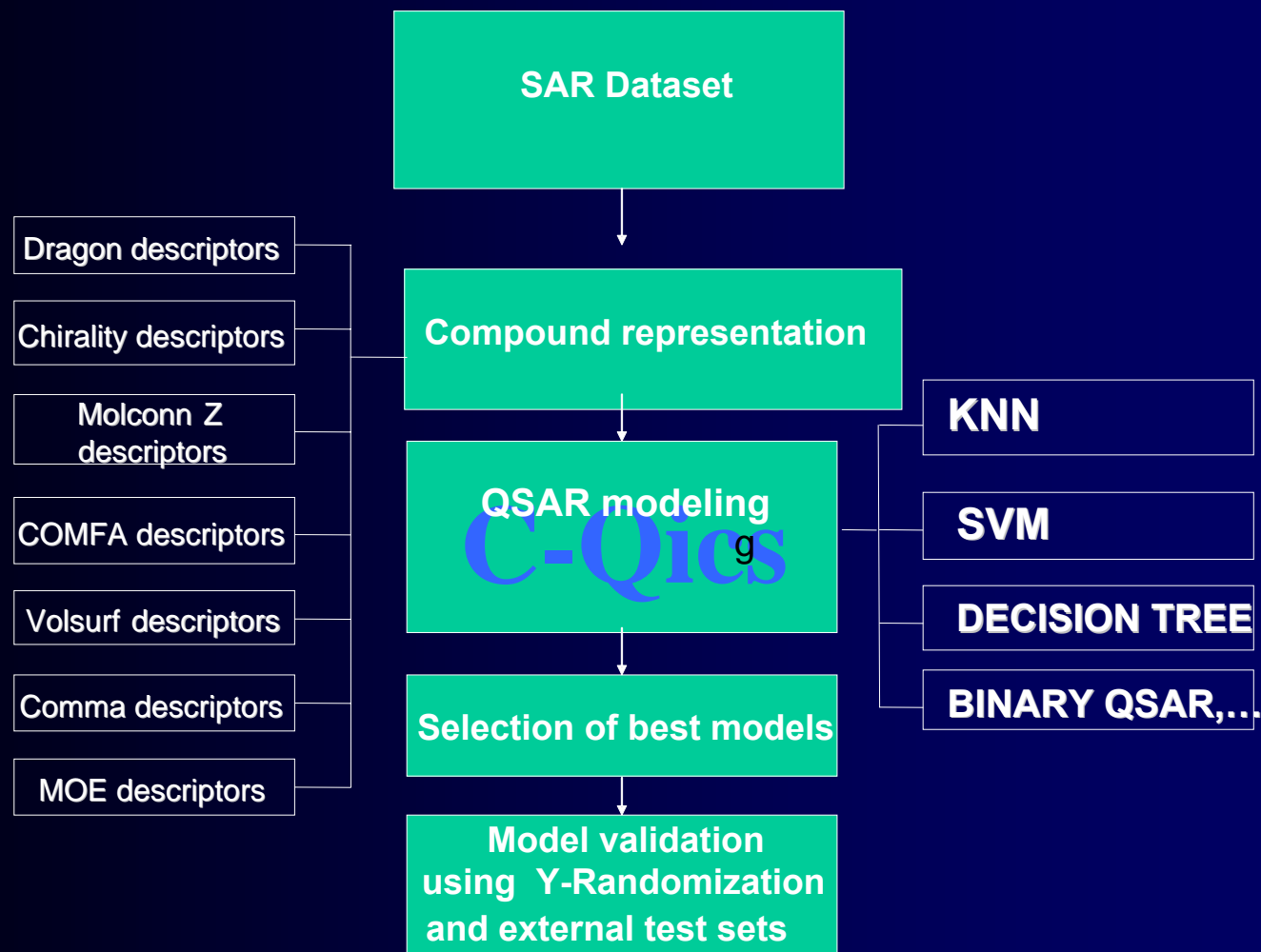


Predictive QSAR Workflow*



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

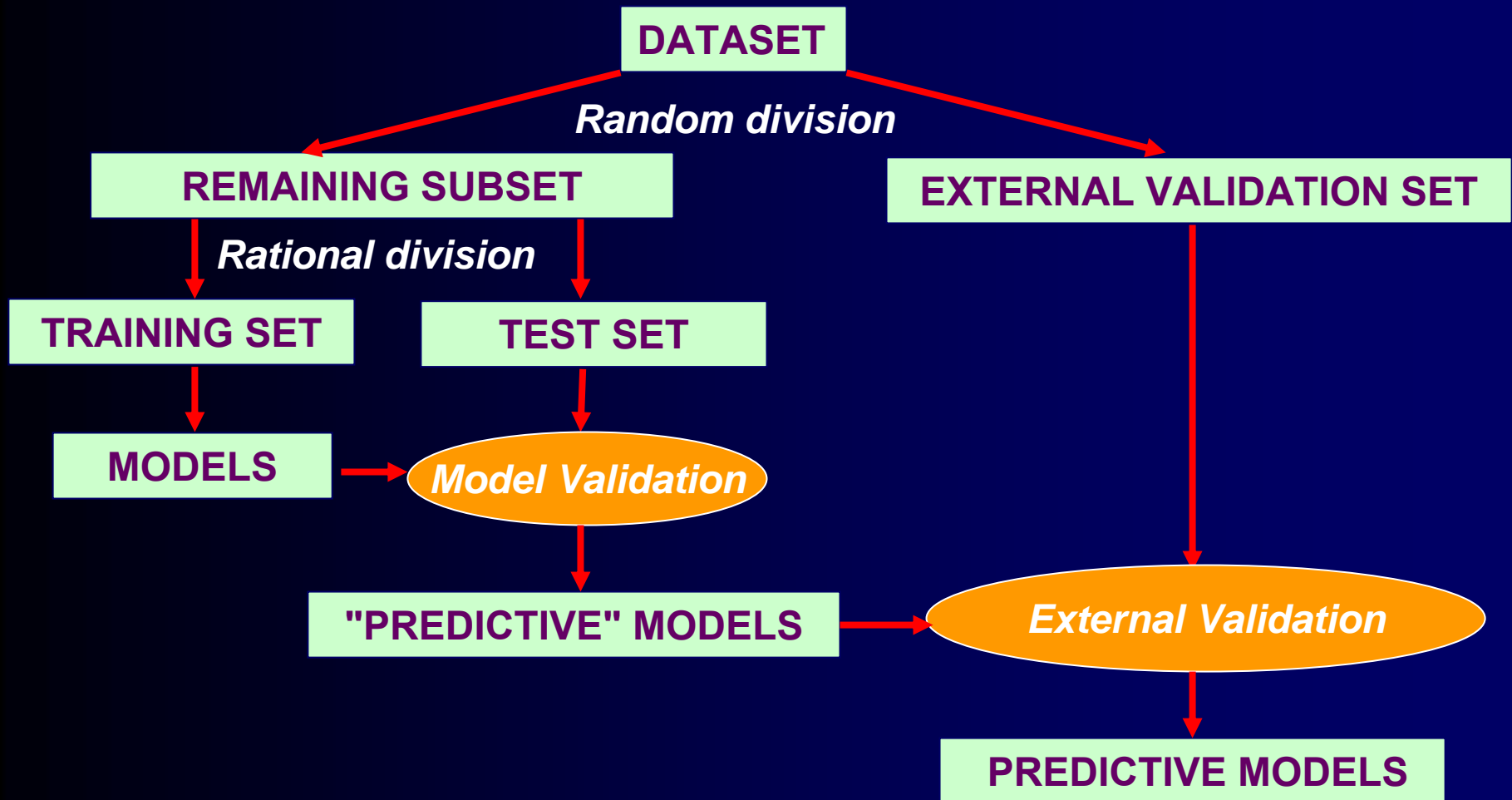
COMBINATORIAL QSAR



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

DIVISION OF A DATASET IN THREE SUBSETS AND EXTERNAL VALIDATION



DEFINING THE APPLICABILITY DOMAIN

Training set: 60 compounds

Test set: 35 compounds

MODEL:

Two nearest neighbors

The number of descriptors: 8

$Q^2(\text{CV})=0.57$ $R^2=0.67$

DISTANCES:

$\langle D \rangle_{\text{train}}=0.287$

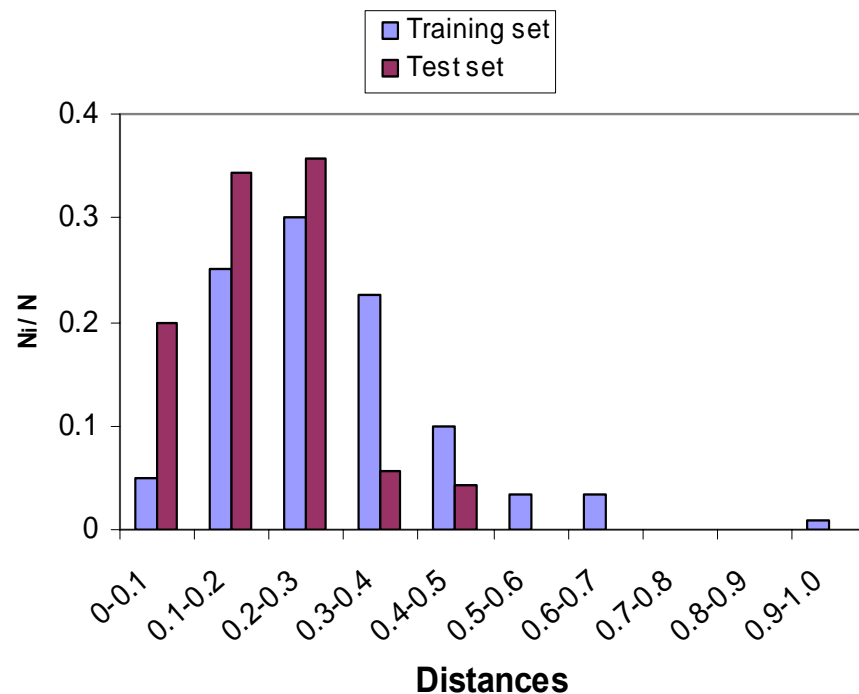
$\text{StDev}(D)_{\text{train}}=\sigma=0.149$

Closest nearest neighbors of test set compounds:

$$D_{\text{test}} \leq \langle D \rangle_{\text{train}} + \sigma \times Z_{\text{CutOff}}$$

($Z_{\text{CutOff}}=0.5$)

Distribution of distances between points and their nearest neighbors in the training set

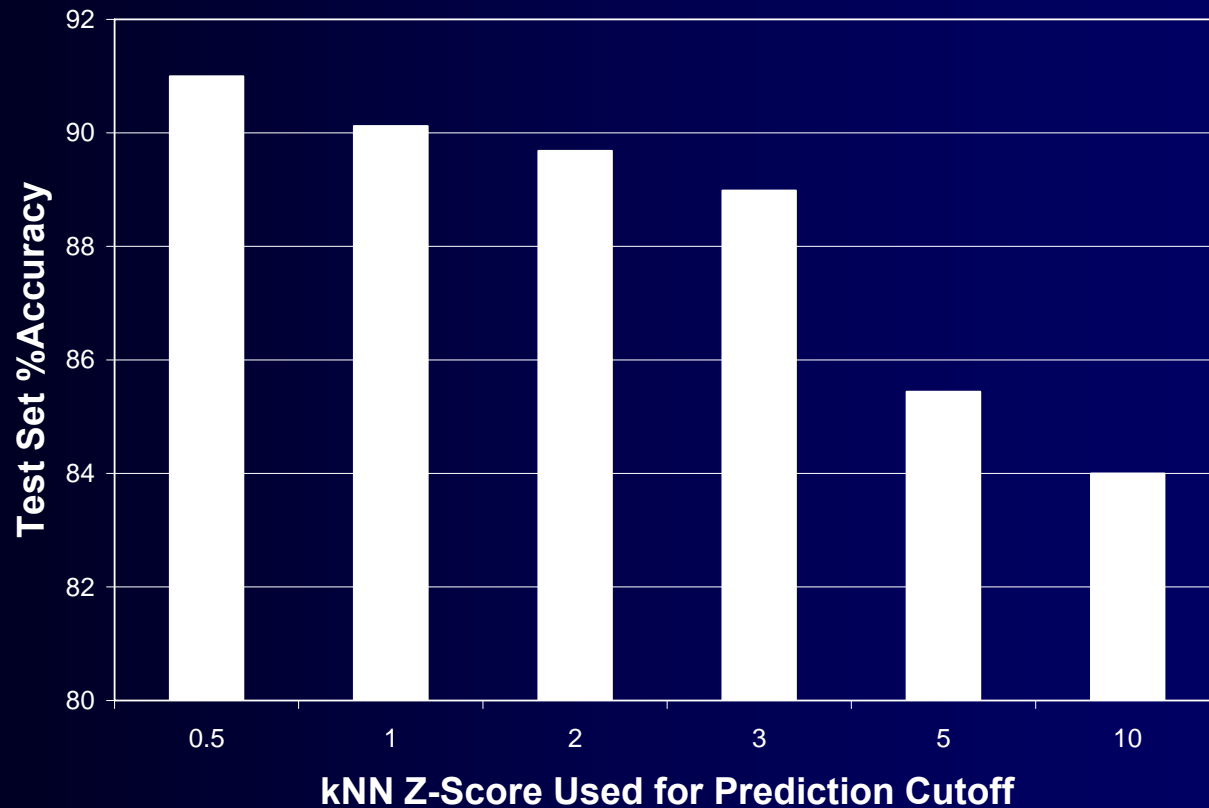


N is the total number of distances

($N_{\text{train}}=60 \quad 2=120$; $N_{\text{test}}=70$)

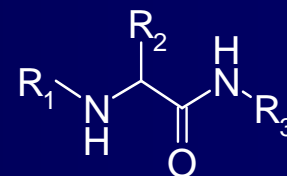
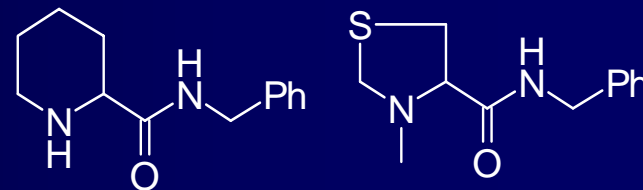
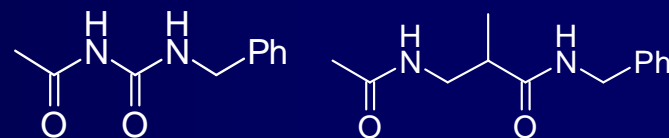
N_i is the number of distances in each category (bin)

Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)



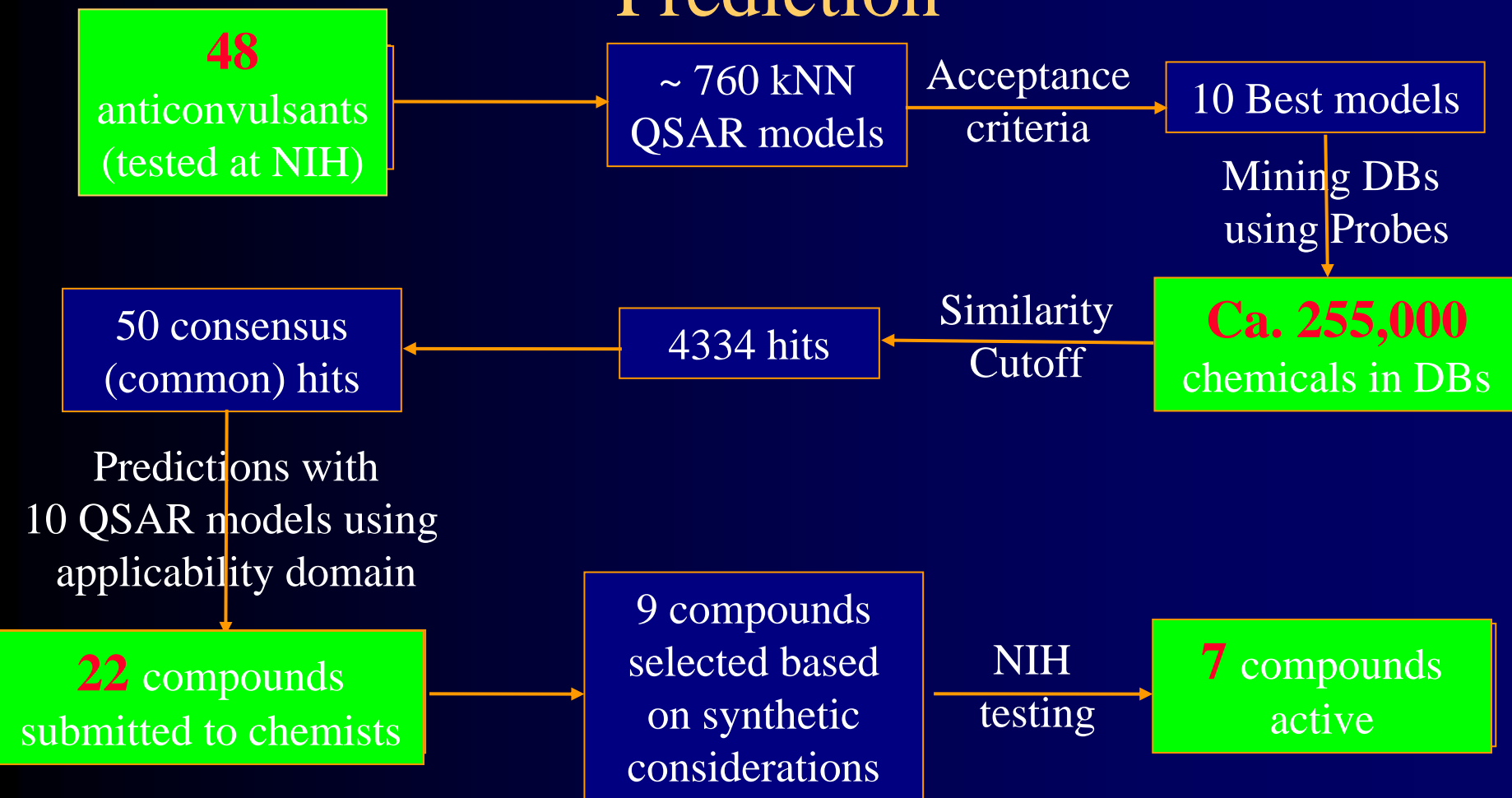
Example. Rational Discovery of Novel Anticonvulsant Agents based on QSAR Modeling and Virtual Screening

- 48 Functionalized amino acid (FAA) anticonvulsant agents
- Experimental activity value:
 - mice ED_{50} (mg/kg)
 - $\log(\text{mM}/\text{kg})$ value: 1.30-3.06
- Five sub-structural classes of FAA

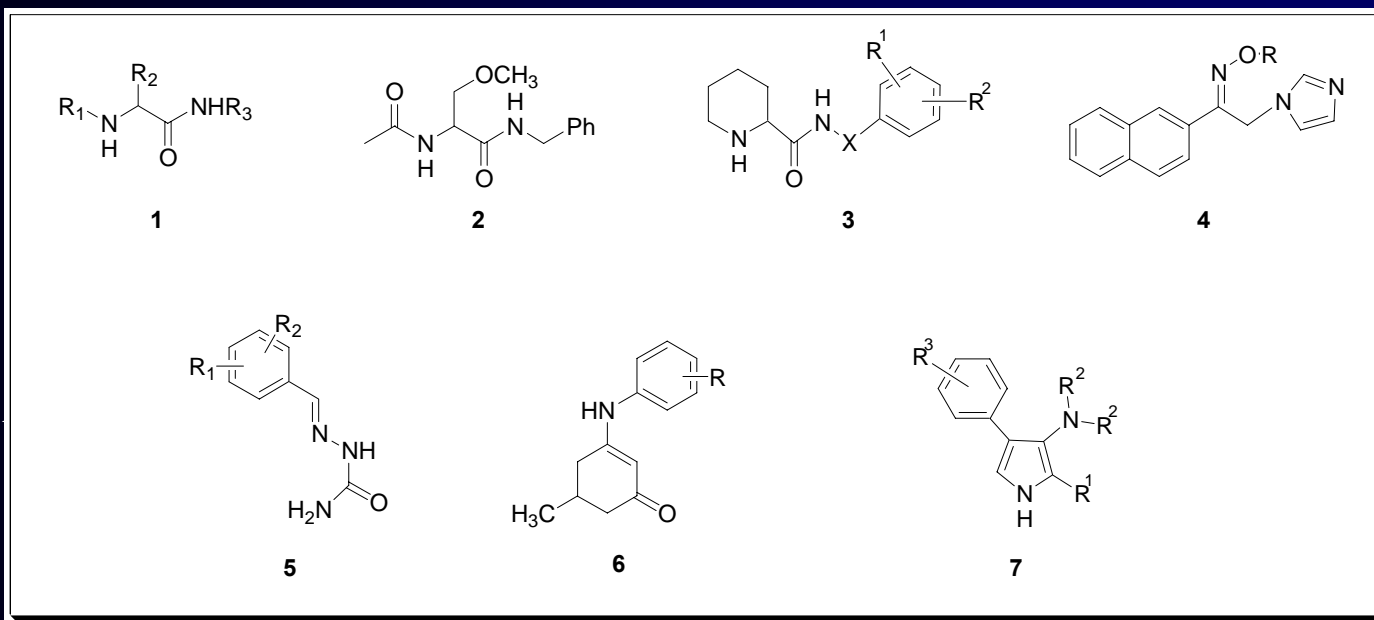


Mining Combined NCI/Maybridge Database with Multiple kNN QSAR models: Consensus Prediction*

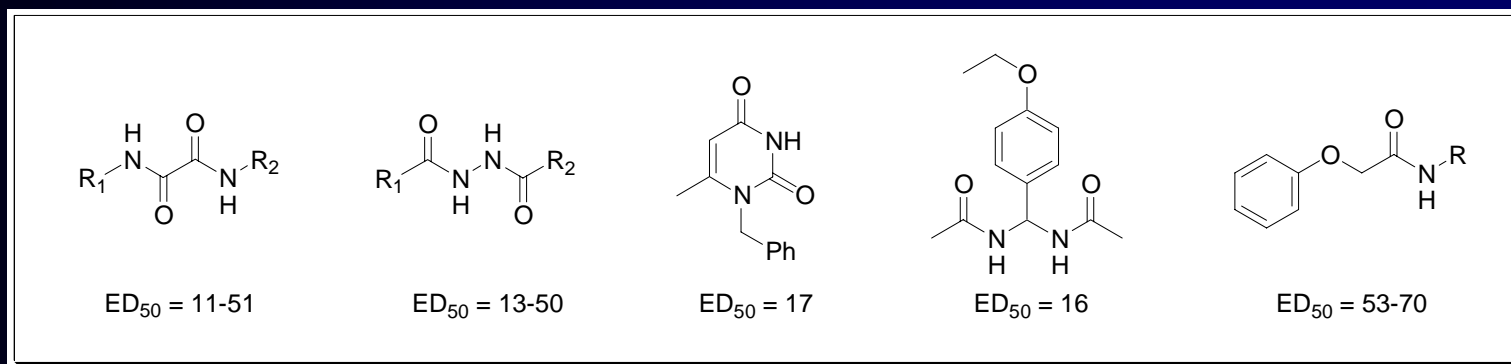
Prediction*



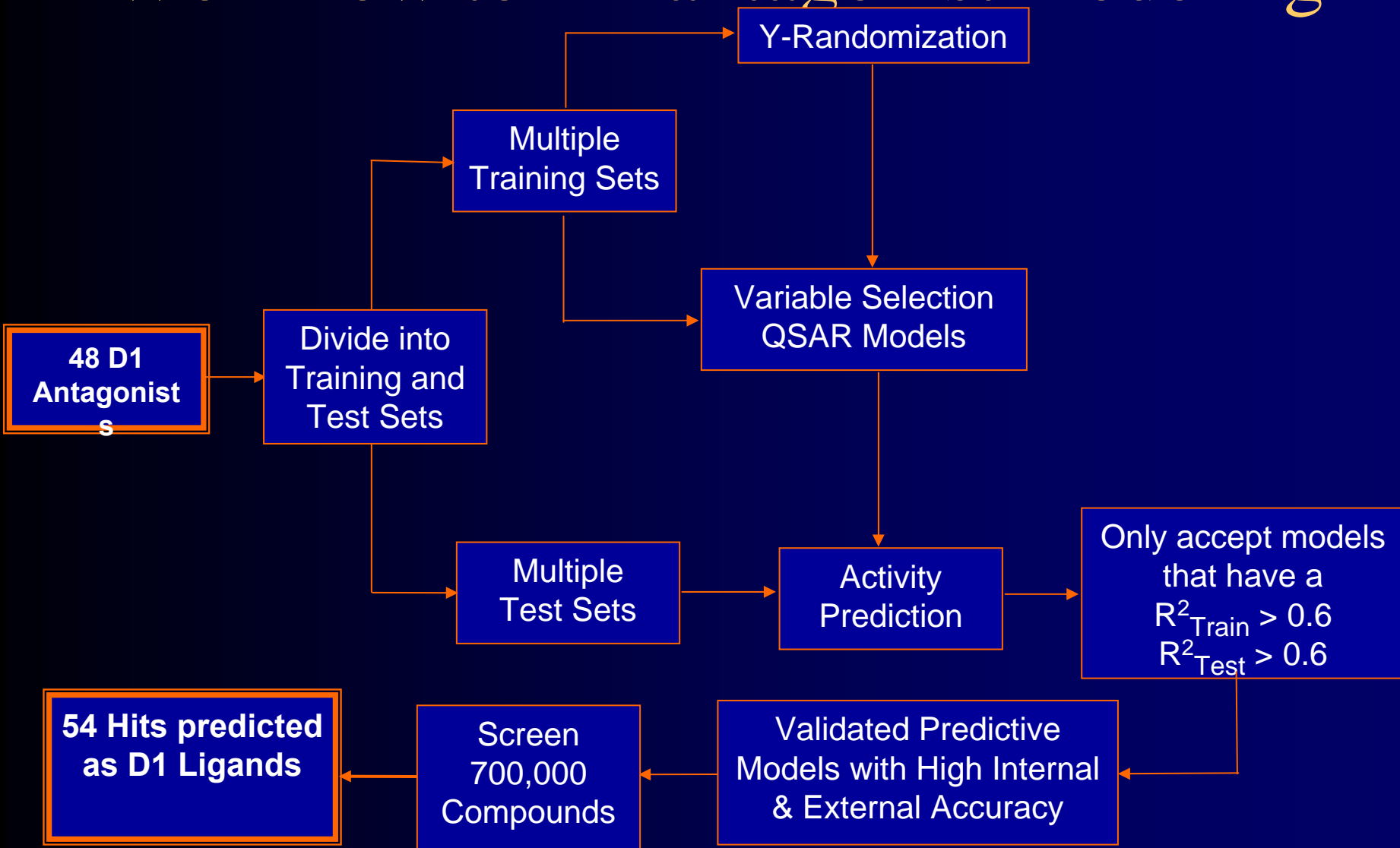
Structural classes of training set compounds



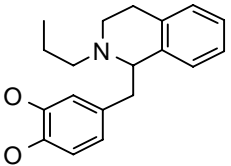
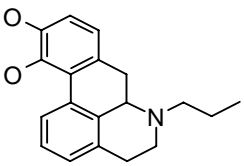
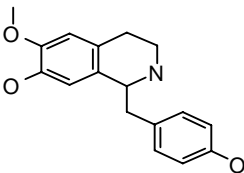
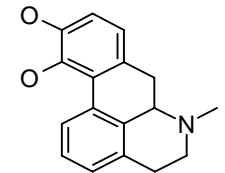
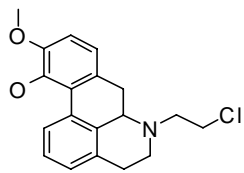
Novel structural scaffolds found in computational hits



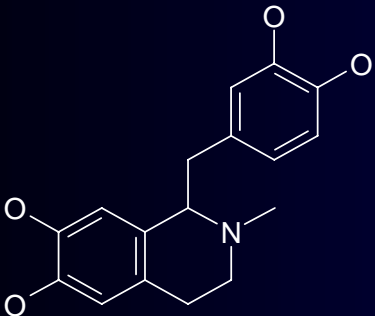
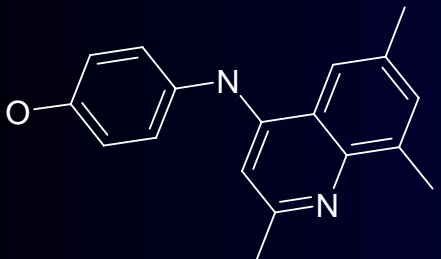
Application of the Predictive QSAR Workflow to D1 antagonist modeling



Compounds identified via virtual screening of databases that were previously characterized as D1 ligands

Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted Log(K0.5)	Std. Dev. of Predictions	Published -Log(K0.5)
	87590-49-6	SVM	84.5	6.5	0.1	5.6
		kNN	86.5	6.7	0.6	
	18426-20-5	SVM	98.6	6.5	0.2	6.5
		kNN	81.8	6.9	1.1	
	2196-60-3	kNN	78.8	6.7	0.6	6.9
	314-19-2	kNN	56.5	7.2	1.1	8.0
	73378-11-7	SVM	85.2	7.5	0.4	6.1
		kNN	72.4	8.4	0.8	

Experimental Validation of Database Screening Predictions

	Prediction Algorithm	% Models Predicted	Predicted $-\text{Log}(K_{0.5})$	Std. Dev. of Prediction	Actual $-\text{Log}(K_{0.5})$
	SVM	51	7.5	0.84	5.5
	kNN	34	8.1	0.2	6.0

Analysis of qHTS Screening Data (from NCGC) for 1,289 NTP Compounds

	BJ	Jurkat	Hek293	HepG2	MRC5	SK-N-SH
Actives	42	121	63	41	37	74
Inconclusives	44	89	79	47	44	54
Inactives	1,203	1,079	1,147	1,201	1,208	1,161

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

NTP- HTS	NTPBSI	NTPGTZ	HPVCSI	CPDB	IRISSI
1,289	1,153	1,053	423	383*	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

*15 of 383 compounds in CPDB database are "technique class".

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

QSAR modeling of the NTP/NCGC/HTS data only

	Modeling set	Validation set
Actives	103	37
Inconclusives	67	23
Inactives	230*	97*
Total	400	157

*Inactives most similar to actives are selected

The best k-NN models based on the modeling set:

Nm	Pred. Train.	Pred. Test	NNN
1	78.8%	72.8%	2
2	78.8%	79.4%	2
3	78.1%	74.1%	2

Use of Applicability Domain Improves Accuracy of External Prediction

No applicability domain.
Accuracy 75.8%

	Actives	Inactives
Pred. Actives	23	11
Pred. Inactives	13	86
Pred. Accuracy	63.9%	88.7%

Applicability domain filter applied.
Accuracy 83.6%, Coverage 82.8%

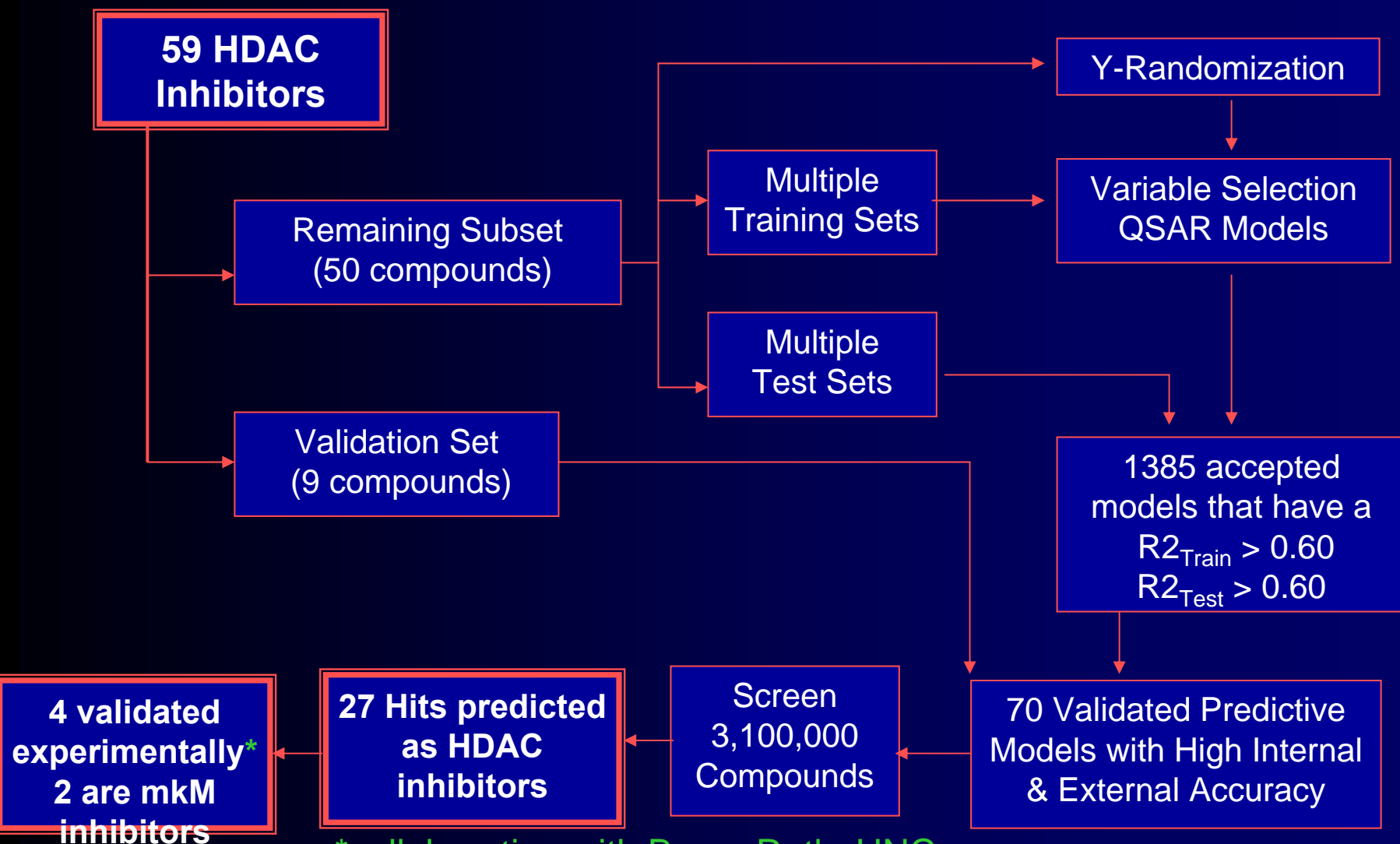
	Actives	Inactives
Pred. Actives	16	7
Pred. Inactives	5	82
Pred. Accuracy	76.2%	92.1%

Use of HTS Data as Biodescriptors improves External Predictive Power of QSAR Models of Animal Carcinogenicity

(modeling set: 167 compounds, External Validation Set: 20 compounds)

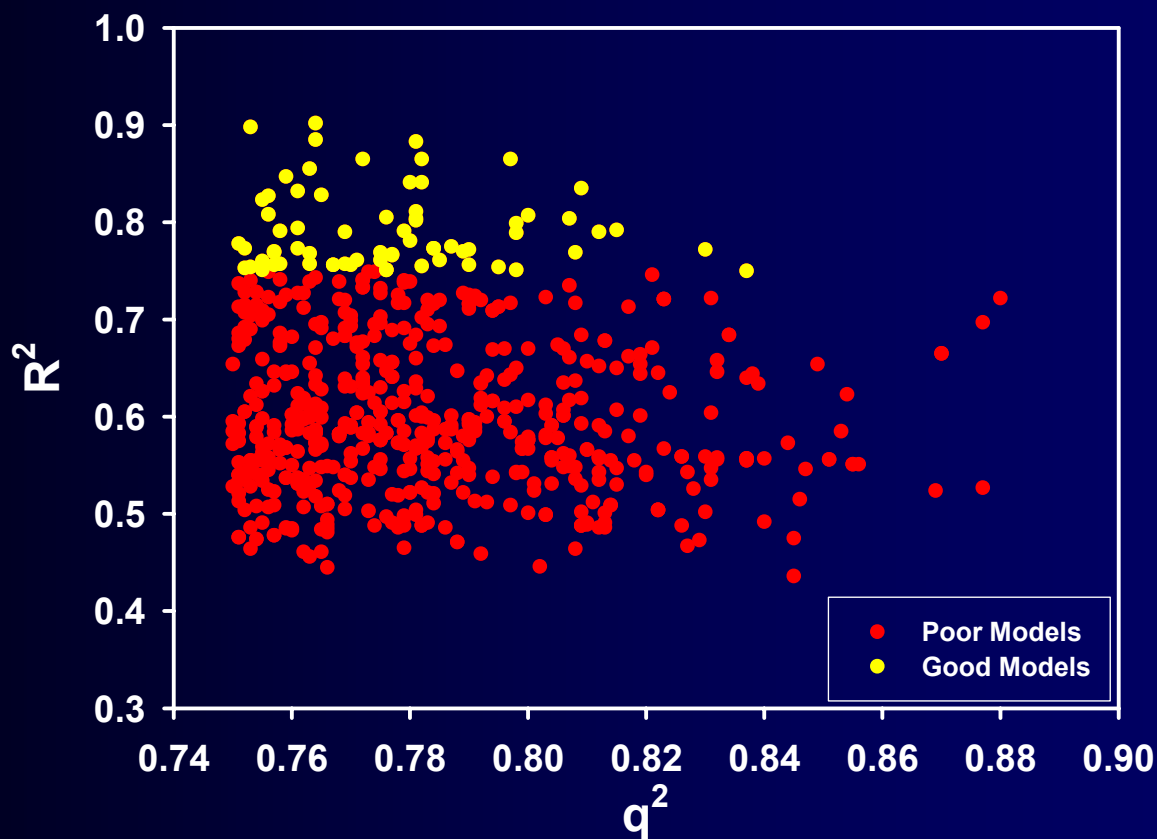
	Chemical descriptors only		Combined HTS+chemical descriptors	
	Exp. Actives	Exp. Inactives	Exp. Actives	Exp. Inactives
Pred. actives	5	1	8	0
Pred. inactives	5	4	3	5
Accuracy	50.0%	80.0%	72.7%	100%
Overall Accuracy	65.0%		86.4%	

Application of Predictive QSAR Workflow to HDAC Inhibitors



*collaboration with Bryan Roth, UNC

HDAC Inhibitor kNN_MZ QSAR Models



- A fraction of training set models with LOO $q^2 > 0.75$ is capable of making accurate predictions ($R^2 > 0.75$) for the test sets.

Comparison of Actual vs Predicted HDAC Inhibitor Activity based on kNN_MZ Models

COMPOUNDS: 59

Training set: 35 Test set: 15

$q^2=0.82$

number of descriptors: 12

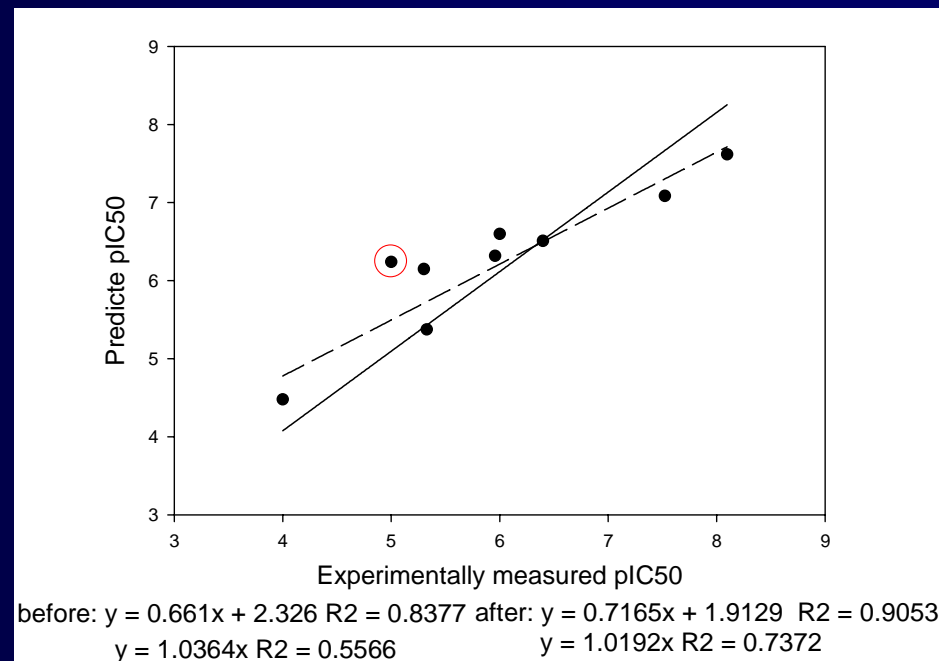
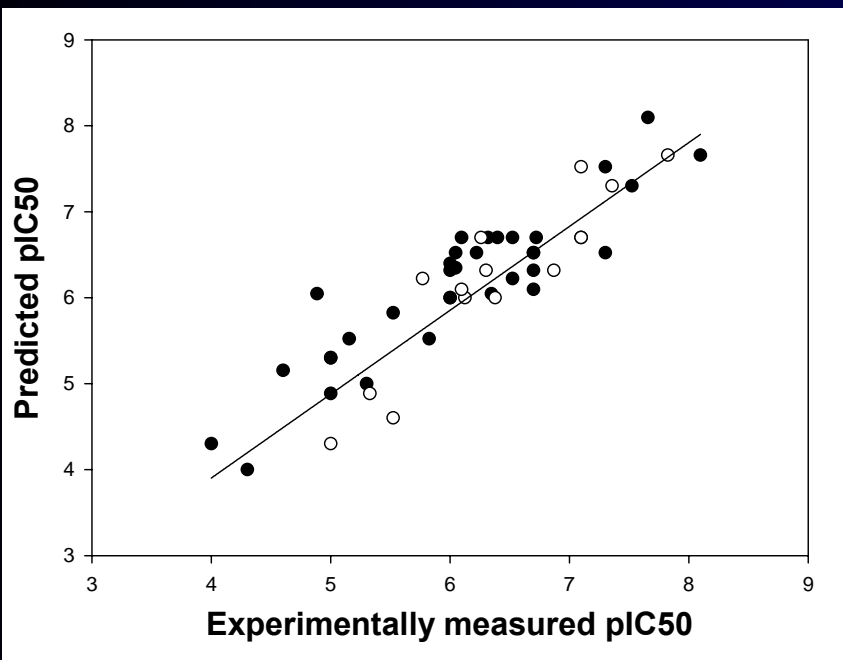
number of nearest neighbors: 1

$R^2=0.84$, $R^2_{02}=0.82$,

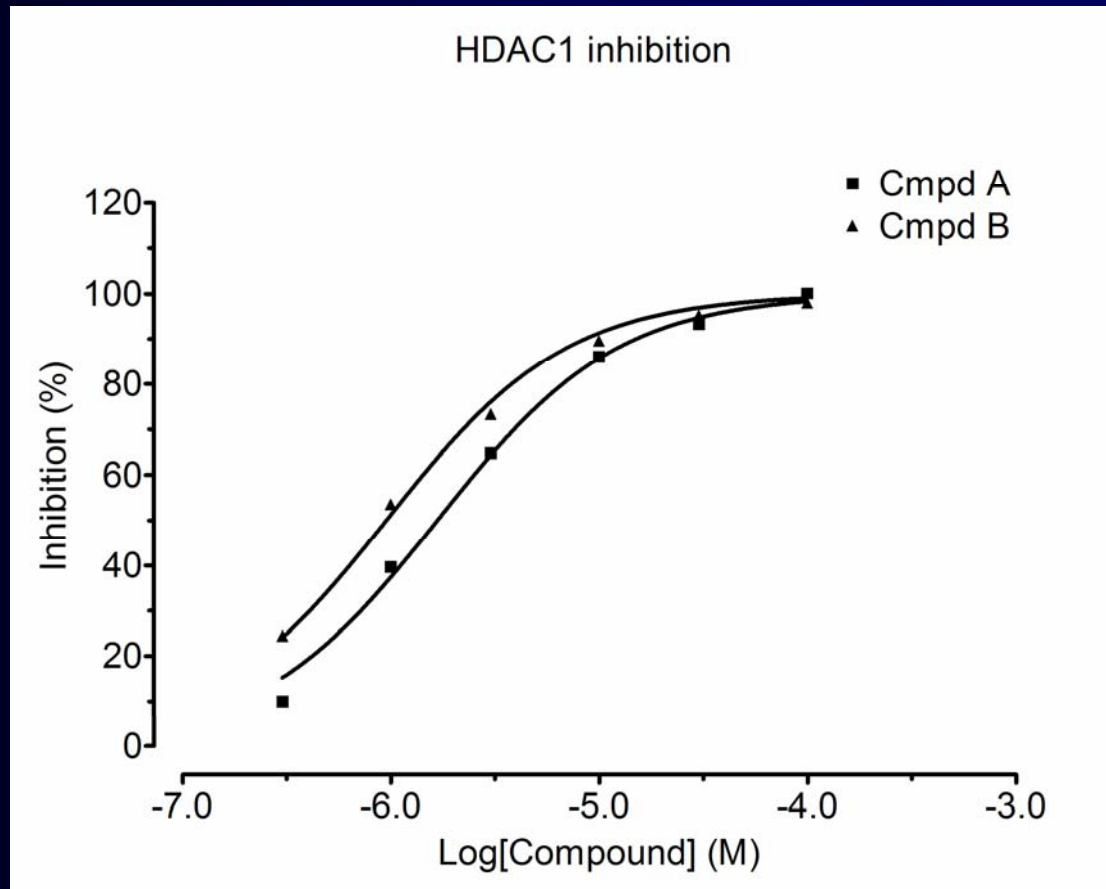
CONSENSUS EXTERNAL VALIDATION:

External validation set: 9

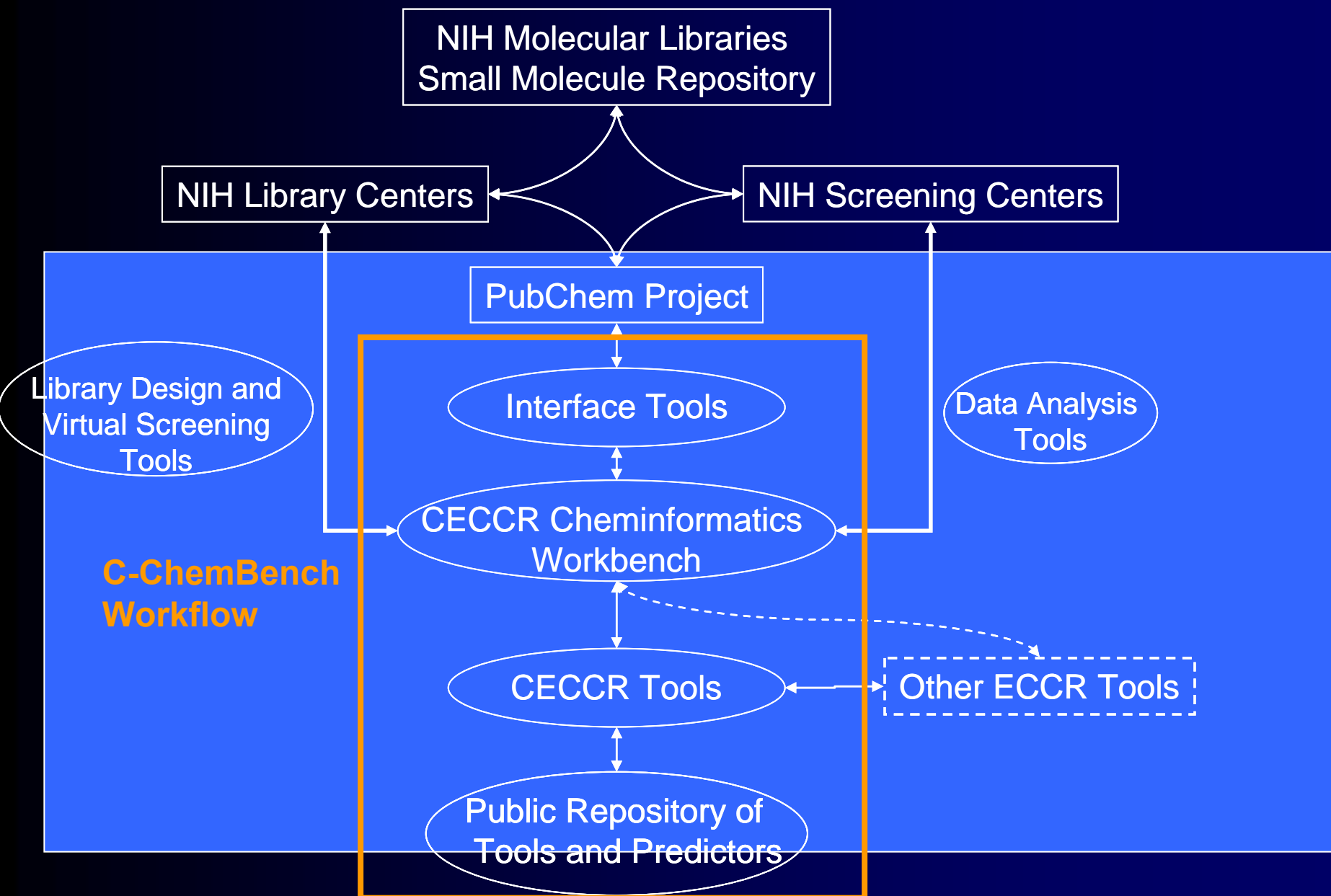
$R^2=0.90$, $R^2_{02}=0.74$



Experimental validation of HDAC computational hits (data from Bryan Roth's lab)



MLI and Carolina ECCCR



C-ChemBench: Web Based QSAR Modeling and Virtual Screening System

C-CHEMBENCH UNC > MODEL BUILDERS - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://ceccr.cs.unc.edu/CECCR-QSAR/submitQsarWorkflow.do

Getting Started Latest Headlines Directory of Departm... post to del.icio.us my del.icio.us

Welcome, test [Logout](#)

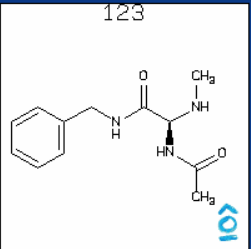
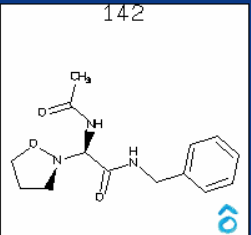
C-CHEMBENCH
ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

WELCOME MODEL BUILDERS PREDICTORS LIBRARY DESIGN VIRTUAL PURCHEN

Of the 1 models generated, 1 passed the training set criteria and 1 pass both training and test set criteria. The top ten models are displayed below.

nnn	q ²	n	r ²	r01 ²	r02 ²	k1	k2
2	0.490	36	0.141	-0.198	-0.281	0.964	0.999

External Validation Results:

Comp_ID	Structure	Actual Value	Predicted Value	# of Models
123	<p>123</p> 	2.27676577	2.421000	1
142	<p>142</p> 	2.053945324	2.326000	1

Done

McAfee SiteAdvisor

Summary

Nothing that is worth knowing can be taught.

Oscar Wilde

- Focus on accurate prediction of external datasets is more critical than accurate fitting of existing data
 - validation!!!
 - applicability domain
 - consensus prediction using all acceptable models
 - experimental validation of a small number of computational hits
- Predictive QSAR workflow with extensive validation affords statistically significant models
 - reliable property predictors
 - decision support tools in selecting experimental screening sets
 - biological data imputation (data imputation is the substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits.)
- HTS and –omics data may be insufficient to achieve the desired accuracy of the end point property prediction BUT should be explored as biodescriptors in combination with chemical descriptors

ACKNOWLEDGMENTS

UNC ASSOCIATES

Former:

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

— Jun FENG
— Yun-De XIAO
—Yuanyuan QIAO
—Ruchir SHAH
-Patricia LIMA
-Assia KOVACHEVA

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

•Protein structure group:

Current

– Yetian CHEN
– Tanarat KIETSAKORN
– Berk ZAFER

Cheminformatics group:

- Kun WANG
– Alex GOLBRAIKH
– Raed KHASHAN
– Chris GRULKE
- Hao TANG
- Simon WANG
- Hao ZHU
- M. KARTHIKEYAN
- Rima HAJJO
- Mei WANG
- Julia GRACE
- Hao HU
- Mihir SHAH
- Jui-Hua Hsieh
- Tong-Ying Wu