The Challenges in Predictive QSPR Modeling

Alexander Tropsha, Hao Zhu, Kun Wang

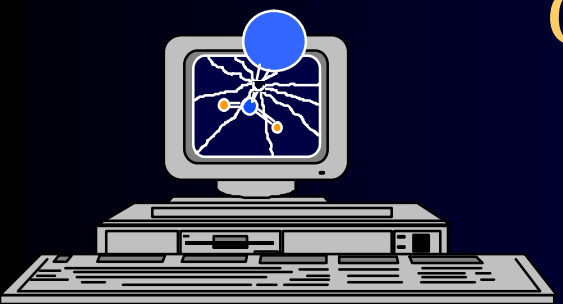
Laboratory for Molecular Modeling

School of Pharmacy

and

Carolina Center for Exploratory Cheminformatics Research

UNC-Chapel Hill



OUTLINE

- The need for developing validated models
  - OECD programme
  - NIH's Molecular Library Initiative and PubChem
  - Possible insufficiency of HTS and -omics data for predictive toxicology
- Predictive QSAR Modeling Workflow
- Examples of the Workflow applications
  - Ames genotoxicity
  - HTS/NTP dataset
  - Carcinogenicity modeling (HTS/NTP/CPDB and CPDB)
- Discussion

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

*Slide is a courtesy of Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)

From Setubal to OECD Principles

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

- be associated with a defined endpoint
- take the form of an unambiguous and easily applicable algorithm;
- ideally, have a mechanistic basis;
- be accompanied by a definition of domain of applicability

-

- be associated with a measure of goodness-of fit (internal validation);

- be assessed in terms of its predictive power by using data not used in the development of the model (external validation).

*Slide is a courtesy of Prof. Paola Gramatica - QSAR Research Unit - DBSF - University of Insubria - Varese (Italy)

NIH Roadmap

ACCELERATING MEDICAL DISCOVERY TO IMPROVE HEALTH

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

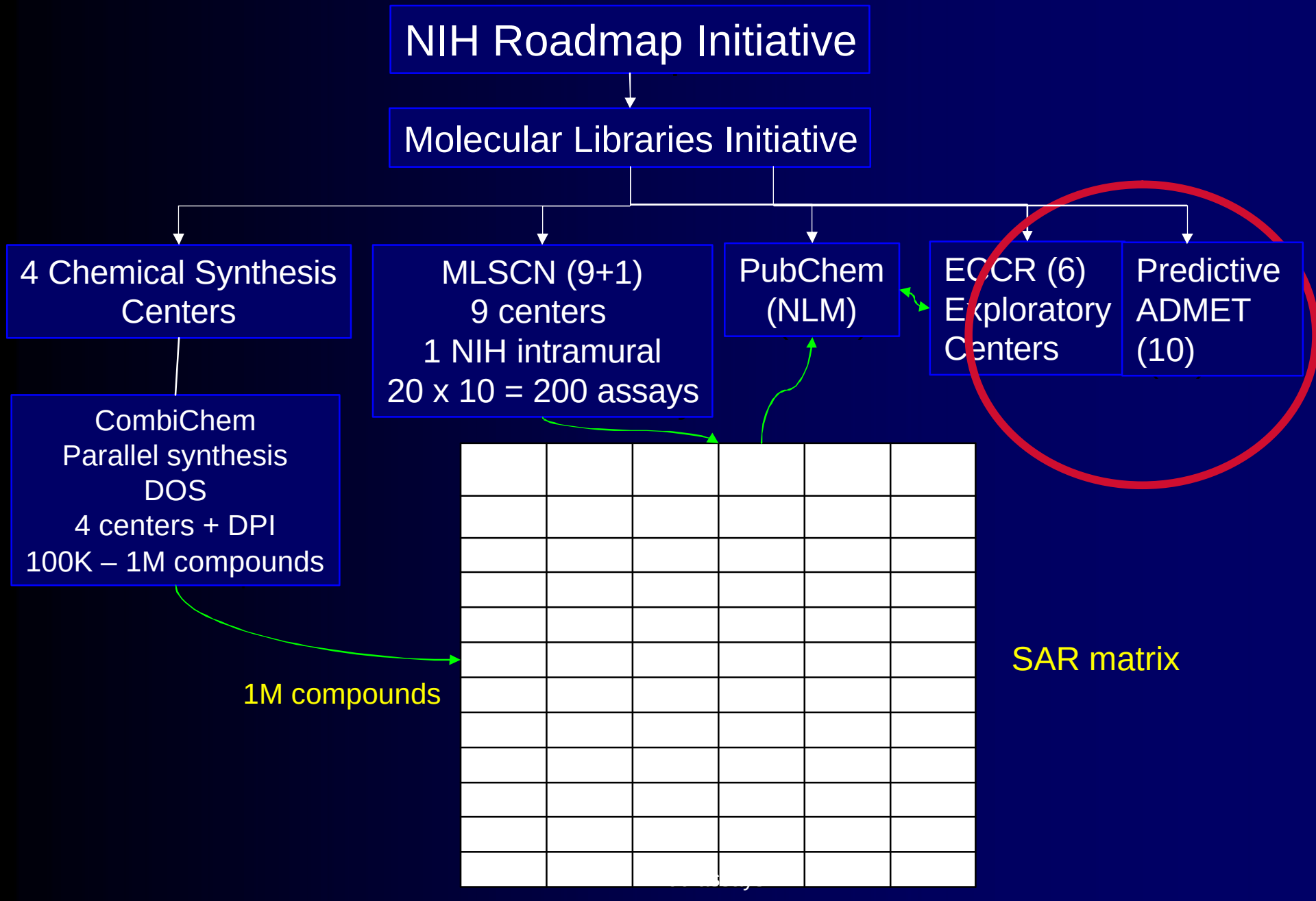
Re-engineering the Clinical Research Enterprise

- 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



Recent MLSCN Screening Results in PubChem

Screen	Center	# of Actives	# Screened	% Active
Prx2	NCGC	0	65535	0
sOGT	NCGC	0	70158	0
Cell Viability (SK-N-SH)	NCGC	92	1408	0.065341
Cell Viability (MRC5)	NCGC	51	1408	0.036222
Cell Viability (HepG2)	NCGC	53	1408	0.037642
Cell Viability (Hek293)	NCGC	80	1408	0.056818
Cell Viability (Jurkat)	NCGC	142	1408	0.100852
Cell Viability (BJ)	NCGC	52	1408	0.036932
IkB Signalling	NCGC	37	69826	0.00053
MKP-1	PMLSC	100	65239	0.001533
FPRL1	NMMLSC	23	9993	0.002302
FPR	NMMLSC	51	9993	0.005104
Pantothenate Synthetase	SRMLSC	2	10011	0.0002
A549 Cell Growth	SRMLSC	278	3317	0.083811
Cell Viability (HPDE-C7K)	SDCCG	215	9984	0.021534
Cell Viability (HPDE-C7)	SDCCG	194	9984	0.019431
Thallium flux through GIRK	VUMLSC	49	8536	0.00574
M4	VUMLSC	72	12369	0.005821

Subset of PubChem relevant to this presentation: NTP-HTS Content Summary of 1408 Compounds

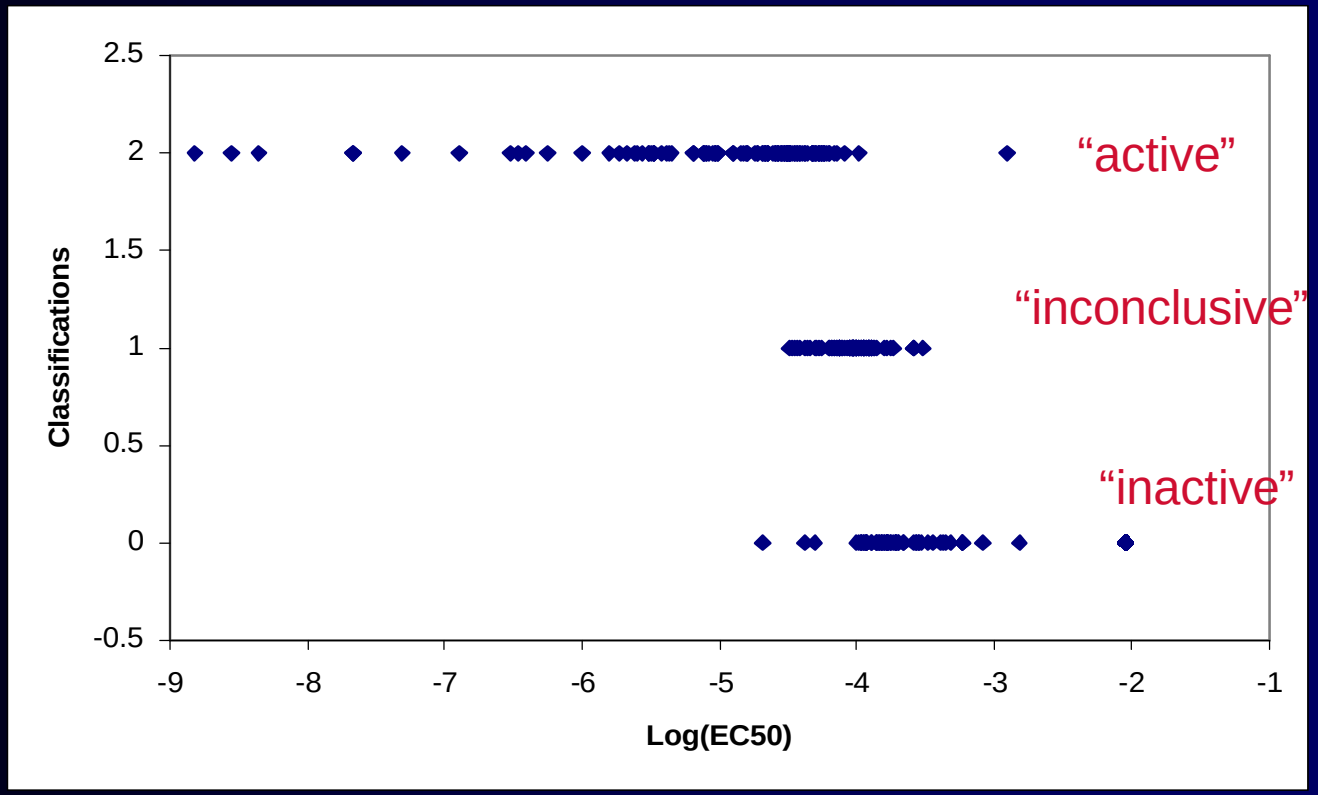
- Chemical Structure Types:
  - Organic: 1,348
  - Inorganic: 27
  - Organometallic: 19
  - No structure: 14
- 1348 Organic compounds contain:
  - Unique: 1,279
  - Complex: 51
  - Salt: 20
  - Duplicates: 53

- Curated subset: 1,289 unique organic compounds

HTS Screening Data (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

Data interpretation: How was the Activity Classified?



The relationship between IC50 and classification of compounds in Jurkat cell line test.

Summary of the experimental data for 1,289 compounds

- 141 compounds are active in at least one test.
- 230 compounds are at least “active” or “inconclusive” in at least one test.
- 1,059 compounds are inactive in all 6 tests

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.

Overview of carcinogenic responses of the 383 compounds in rats and mice

- 229 compounds show positive response in at least one organ of one or more species.
- 92 compounds show negative results in all tests.
- 62 compounds show negative response in all tests but the tests are not complete.

Are HTS results indicative of carcinogenicity?

93 compounds were tested in HTS, 57 of them are or likely to be human carcinogens, 36 of them are not human carcinogens.

	HTS-Actives	HTS-Inconclusives	HTS-Inactives
Human Carcinogens	5	5	47
Non Human Carcinogens	1	2	33

Results based on the IRIS database (EPA 1986, 1996, 1999 carcinogen risk assessment)

Can the explicit use of chemical structure help with the end point prediction: QSPR Modeling

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

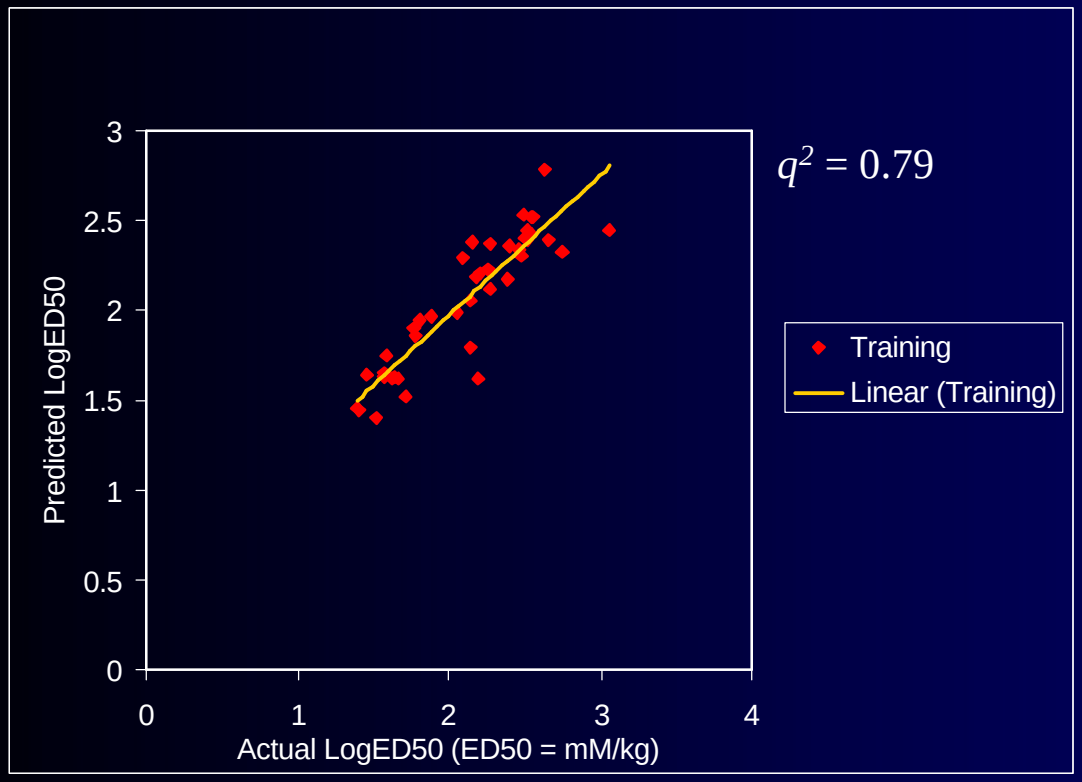
Chemistry	Biology (IC50, Kd...)	Cheminformatics (Molecular Descriptors)			
Comp.1	Value1	D1	D2	D3	D4
Comp.2	Value2	"	"	"	"
Comp.3	Value3	"	"	"	"
- - - - - - - - - - - - - -					
Comp.N	ValueN	"	"	"	"

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

BA (e.g., IC50) = F(D)

Typical QSPR modeling result: Comparison between Actual and Predicted Activity…

…makes everyone happy



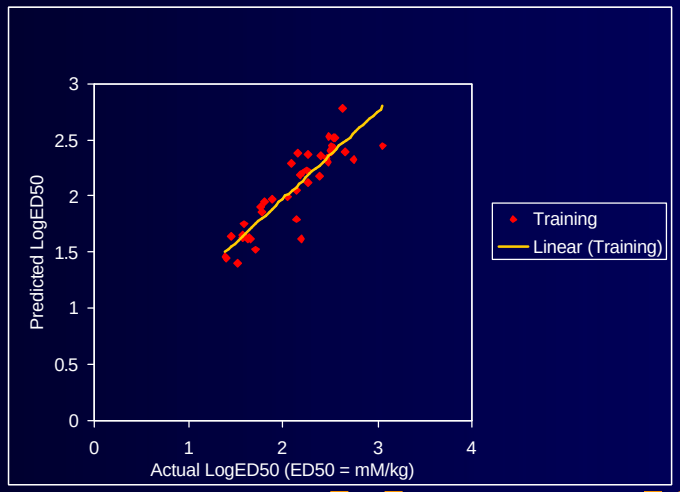
An example of “mechanistic” model of mutagenicity

$$\log \mathrm{TA100} = -12.61592 - 4.58430\ \mathrm{LUMO} - 3.66205\ \mathrm{MR} + 72.46140\ \text{C-carb} + 2.55239\ \log P + 13.09442\ \text{C-}\beta \quad n = 17;\ r^2 = 0.84;\ q^2 = 0.40 \quad (3)$$

- Possible remedies (per authors)
  - retesting some of the compounds;
  - testing further new compounds;
  - checking (if necessary) the use of additional chemical descriptors.

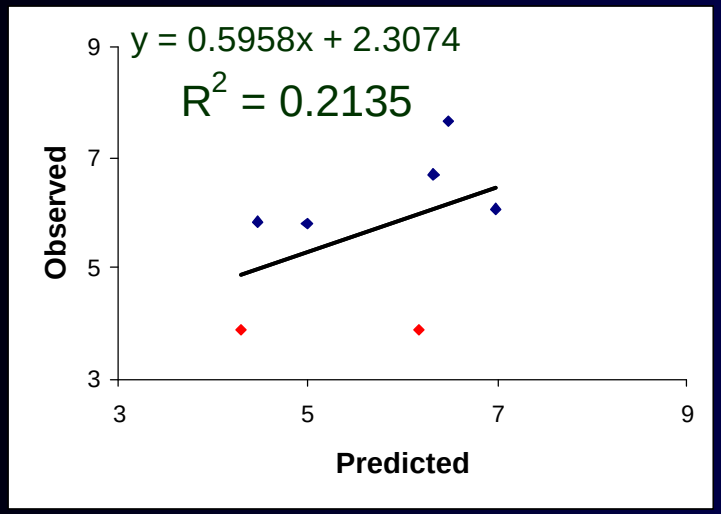
Beningni et al. *Env. Mol. Mutagenesis* 46:268-280 (2005)

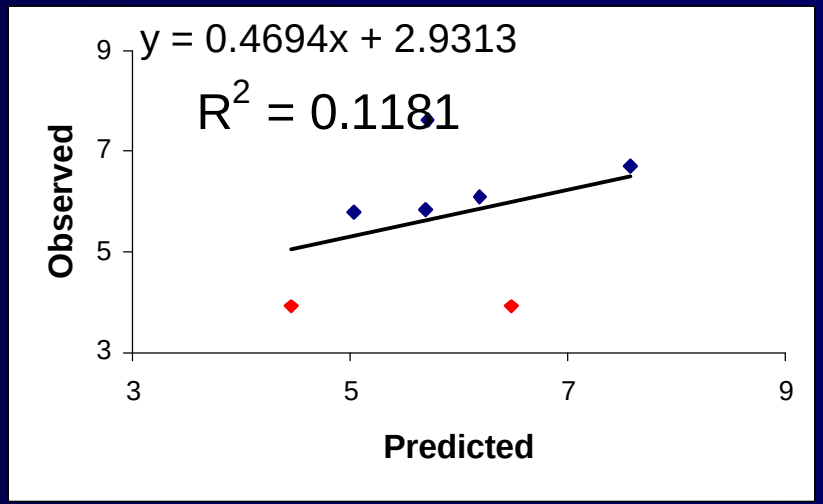
The unbearable lightness of modeling (in this case, CoMFA)



…leads to unacceptable prediction accuracy.

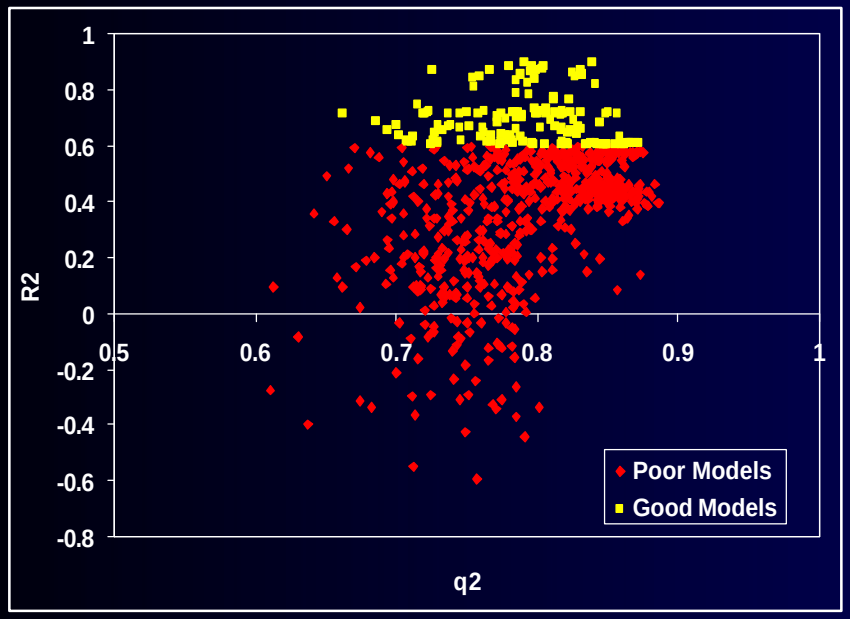
EXTERNAL TEST SET PREDICTIONS

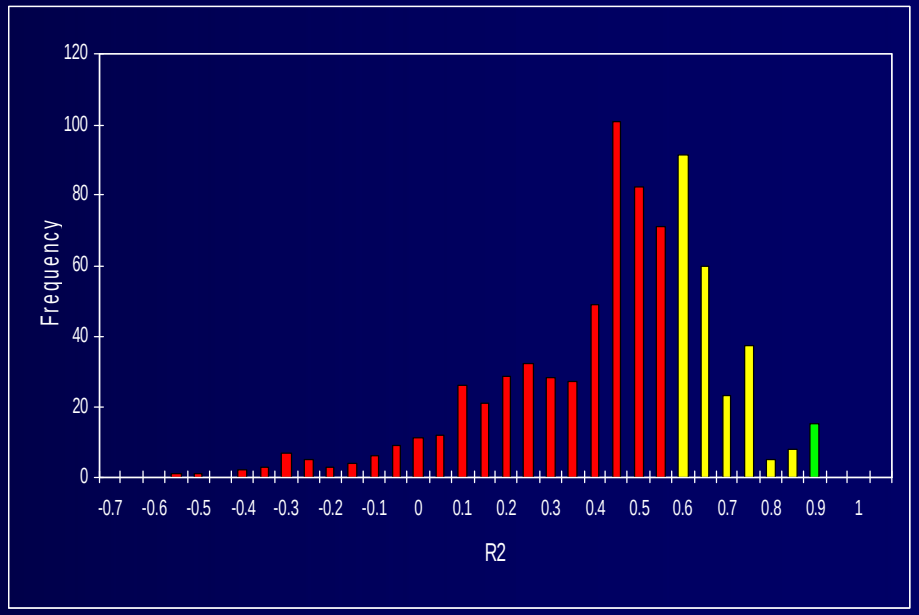




BEWARE OF q^2!!!

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



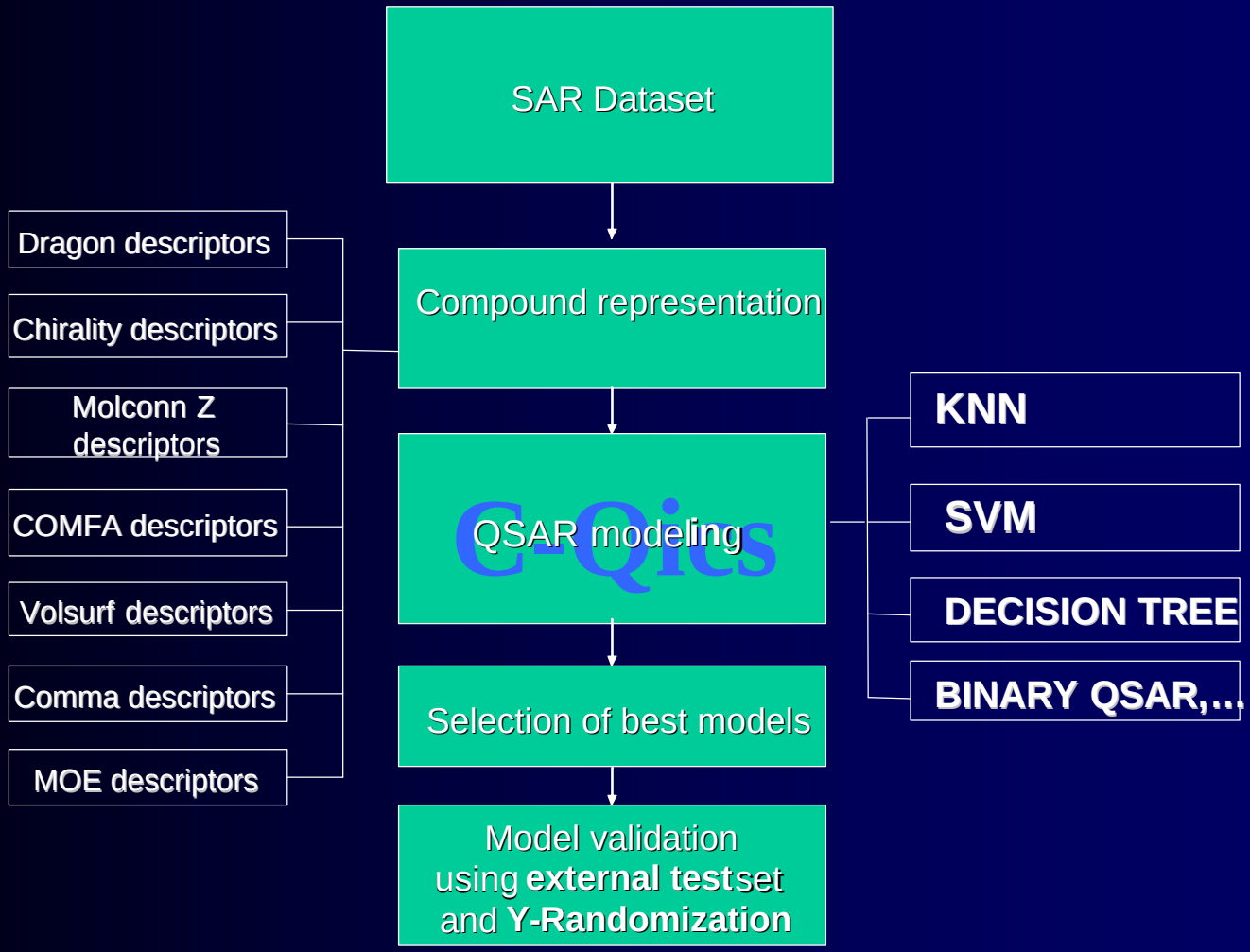


COMPONENTS OF THE PREDICTIVE QSPR MODELING WORKFLOW*

- Model Building: Combination of various descriptor sets and variable selection data modeling methods (Combi-QSAR)
- Model Validation
  - Y-randomization
  - Training and test set selection
  - Applicability domain
  - Evaluation of external predictive power
- Virtual screening

*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

Example of application in a Combi-QSAR Study

Percent Classification Accuracy for the PGP Dataset*

Method / Descriptors	kNN		DECISION TREE		SVM	
	Training	Test	Training	Test	Training	Test
MOLCONNZ	92	78	88	67	90	67
ATOM PAIR	87	80	83	76	94	80
MOE	89	53	88	69	84	62
VOLSURF	83	76	86	78	88	80

Percent Classification Accuracy for the Fragrance Dataset**

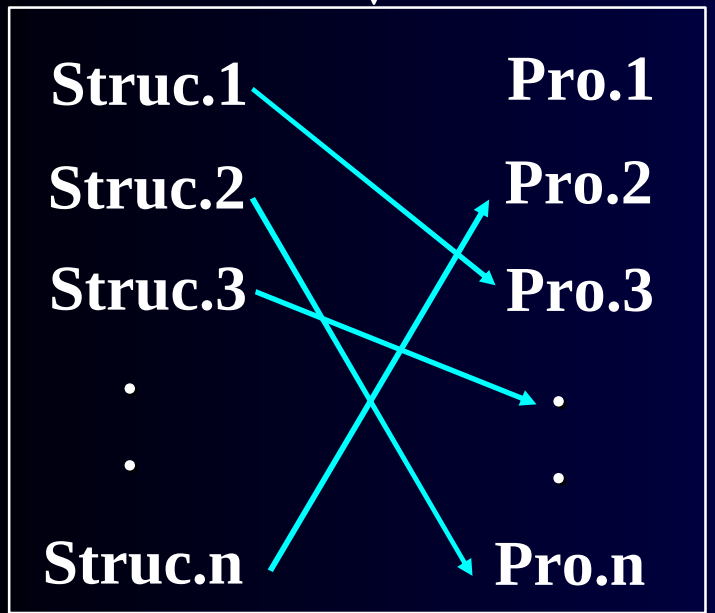
Method / Descriptors	kNN		BINARY QSAR		DECISION TREE		SVM	
	Training	Test	Training	Test	Training	Test	Training	Test
DRAGON	70	86	72	76	70	78	83	68
CMTD	72	65	76	50	67	74	81	58
CMTD/MOLCONNZ	67	60	85	47	62	53	87	53
CoMFA	76	89	71	65	75	62	83	75
VOLSURF	78	85	74	70	77	60	94	53
MOE	77	65	74	86	74	71	77	65
COMMA/MOE	77	75	73	70	74	72	73	69
COMMA/MOE	77	75	73	70	74	72	73	69

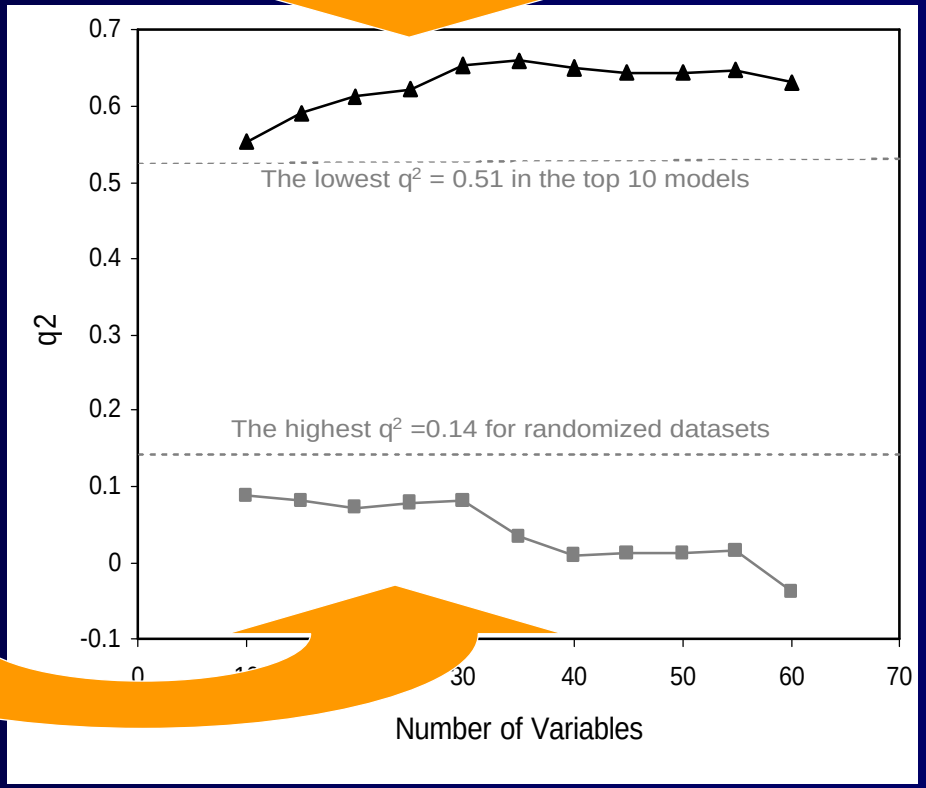
*Lima, et al. JCIM, 2006, in press.

Kovatcheva, Golbraikh, **Oloff, et al. JCICS, 44: 582-595, 2004.

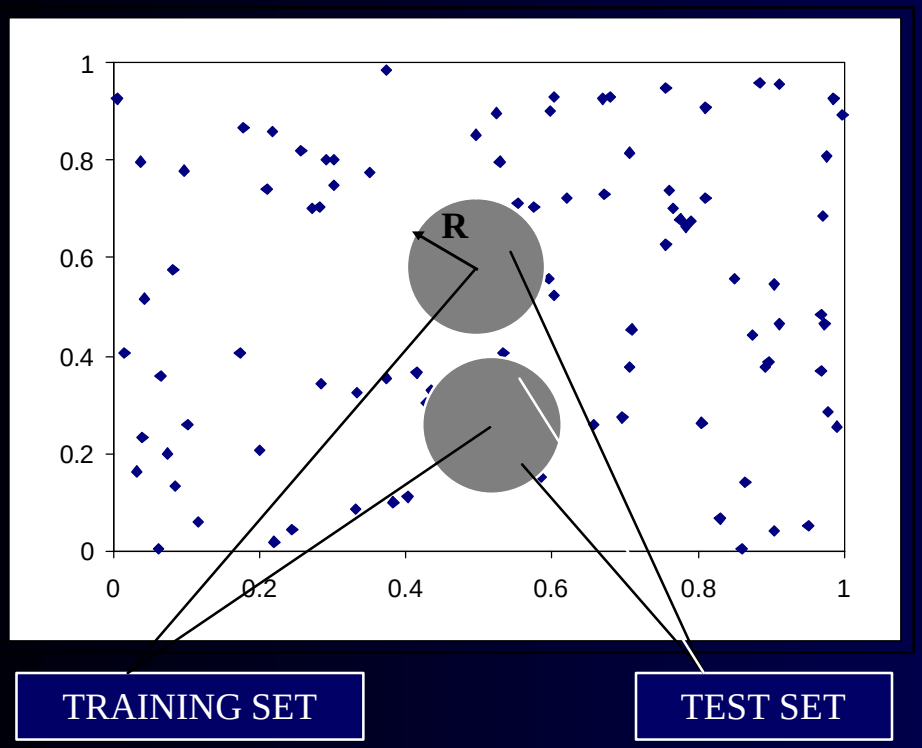
Activity randomization: model robustness

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





RATIONAL SELECTION OF TRAINING AND TEST SETS BASED ON DIVERSITY SAMPLING



$V_p = V / N$

$R = cV_p^{1/K}$

N – number of points

V_p- volume corresponding to one point

V - the occupied volume in the descriptor space

c - dissimilarity level

K - dimensionality of the descriptor space

ALGORITHMS 1 to 3

1. Volume corresponding to one point is 1/N.
2. Select a compound with the highest activity.
3. Include this compound into the training set.
4. Construct a sphere with the center in the representative point of this compound with radius $R = c(V/N)^{1/K}$.
5. Include compounds within this sphere except for the center in the test set.
6. Exclude all points within this sphere. For algorithm 1, select randomly a compound and go to 3. If no compounds left, go to 10.
7. n - the number of remaining compounds. m - the number of spheres already constructed. d_{ij}, i=1,…,n, j=1,…,m – distances of compounds left to the sphere surfaces.
8. Select a compound with the smallest (algorithm 2) or largest (algorithm 3) d_{ij}.
9. Go to step 3.
10. Stop.

Golbraikh et al., *J. Comp. Aid. Mol. Design* 2003, 17, 241–253

DEFINING THE APPLICABILITY DOMAIN

Training set: 60 compounds
Test set: 35 compounds

MODEL:

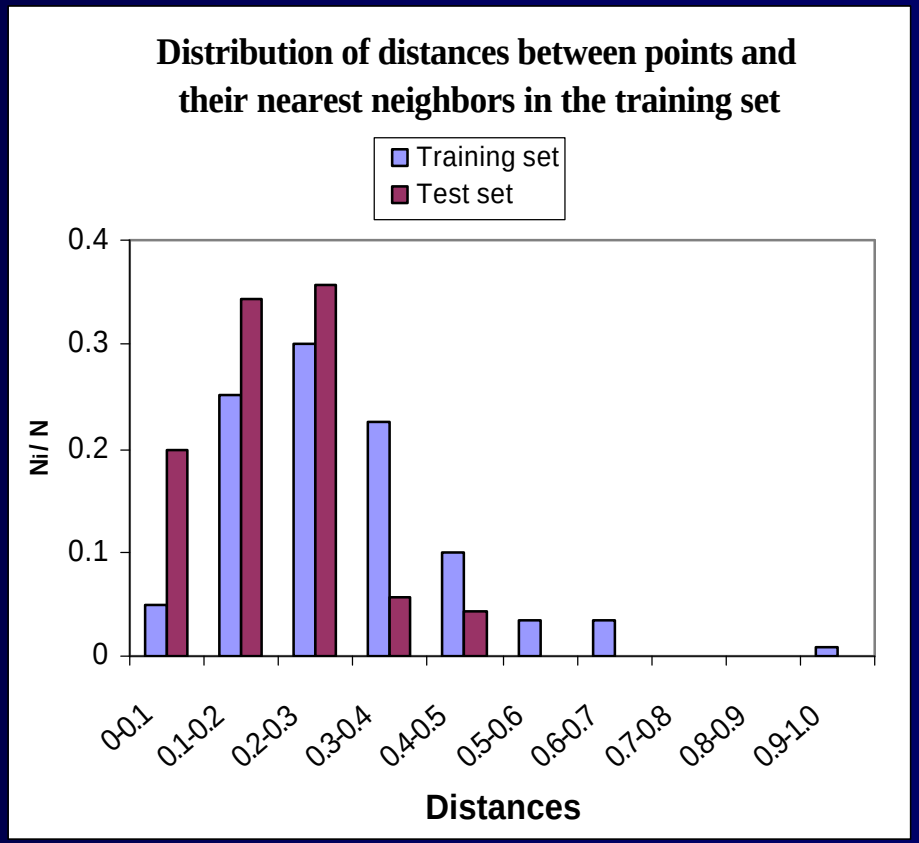
Two nearest neighbors
The number of descriptors: 8
Q^2(CV)=0.57 R^2=0.67

DISTANCES:

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

Closest nearest neighbors of test set compounds:

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

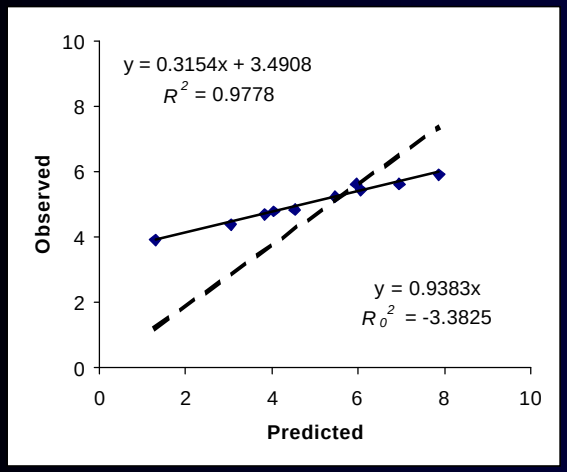


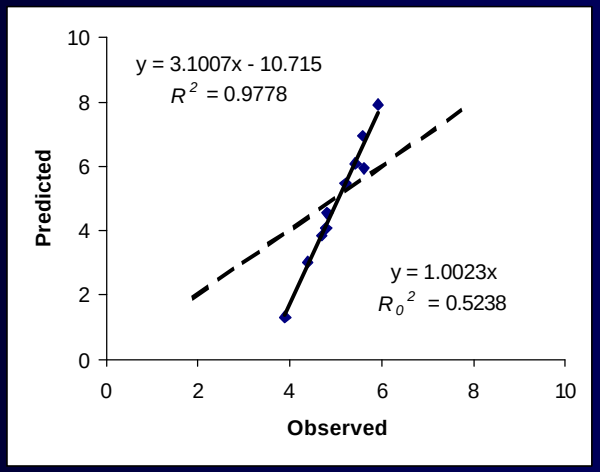
N is the total number of distances
(N_{train}=60 2=120; N_{test}=70)

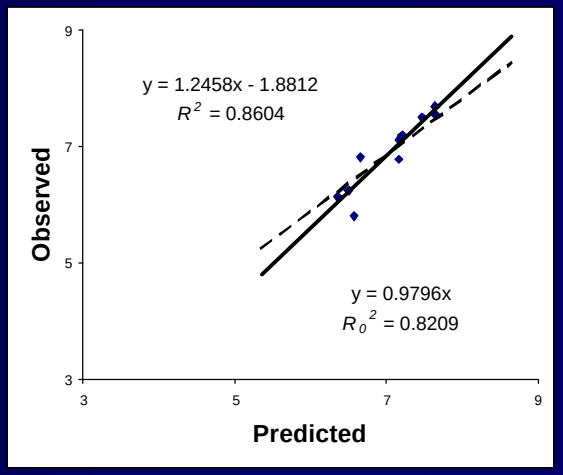
N_i is the number of distances in each category (bin)

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

Criteria for Predictive QSAR Model.







Regression

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

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

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

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

Golbraikh et al., *J. Comp. Aid. Mol. Design* 2003, 17, 241–253

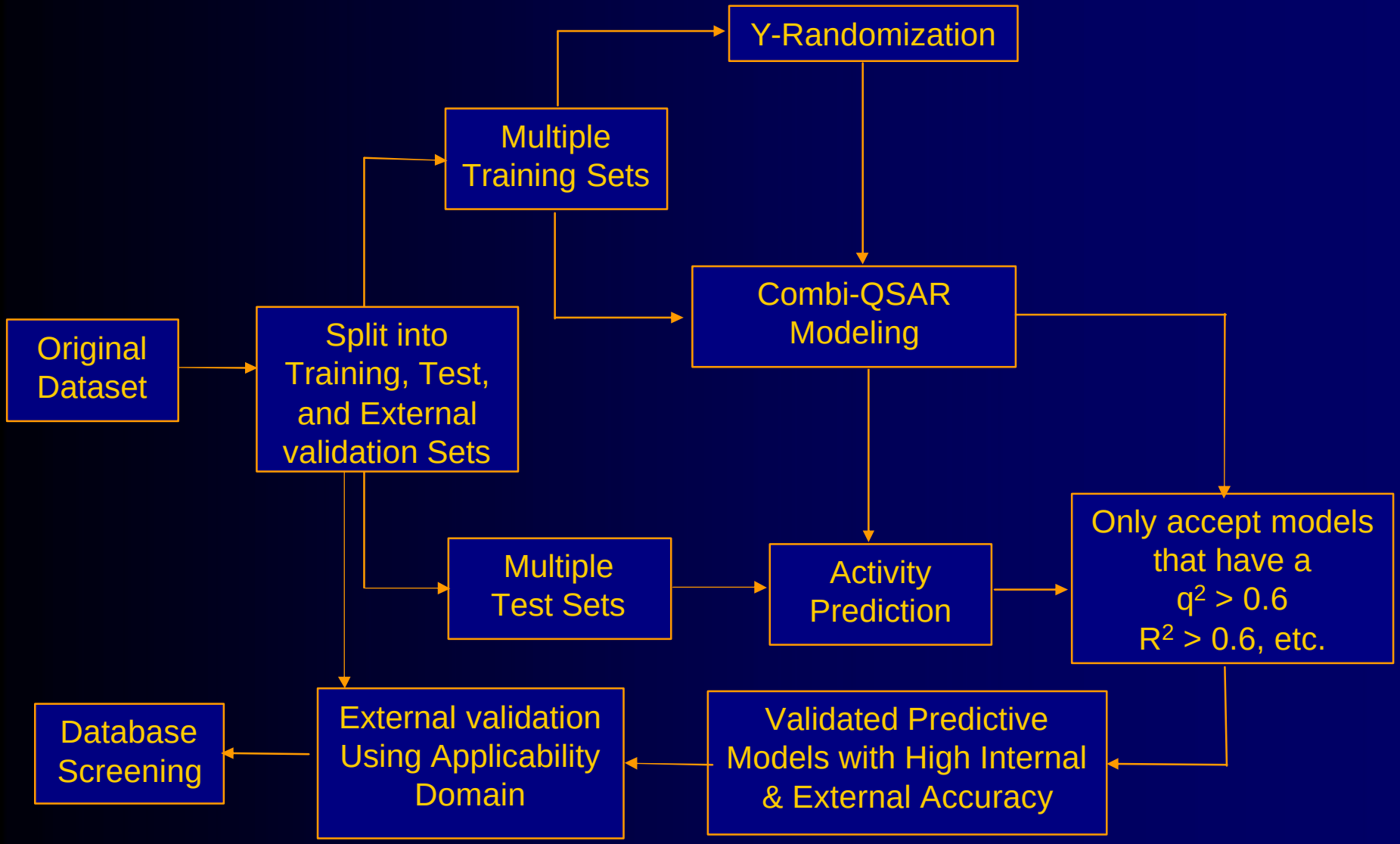
Why can't we get it right? Have not we tried enough?

- Descriptors? No, we have plenty (e.g., Dragon)
- Methods? No, we also have plenty, and still searching (e.g., adapting datamining techniques).
- 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!!!
- QSAR is an empirical data modeling exercise: just do it any way you like but VALIDATE on independent datasets!

Predictive QSPR Workflow*



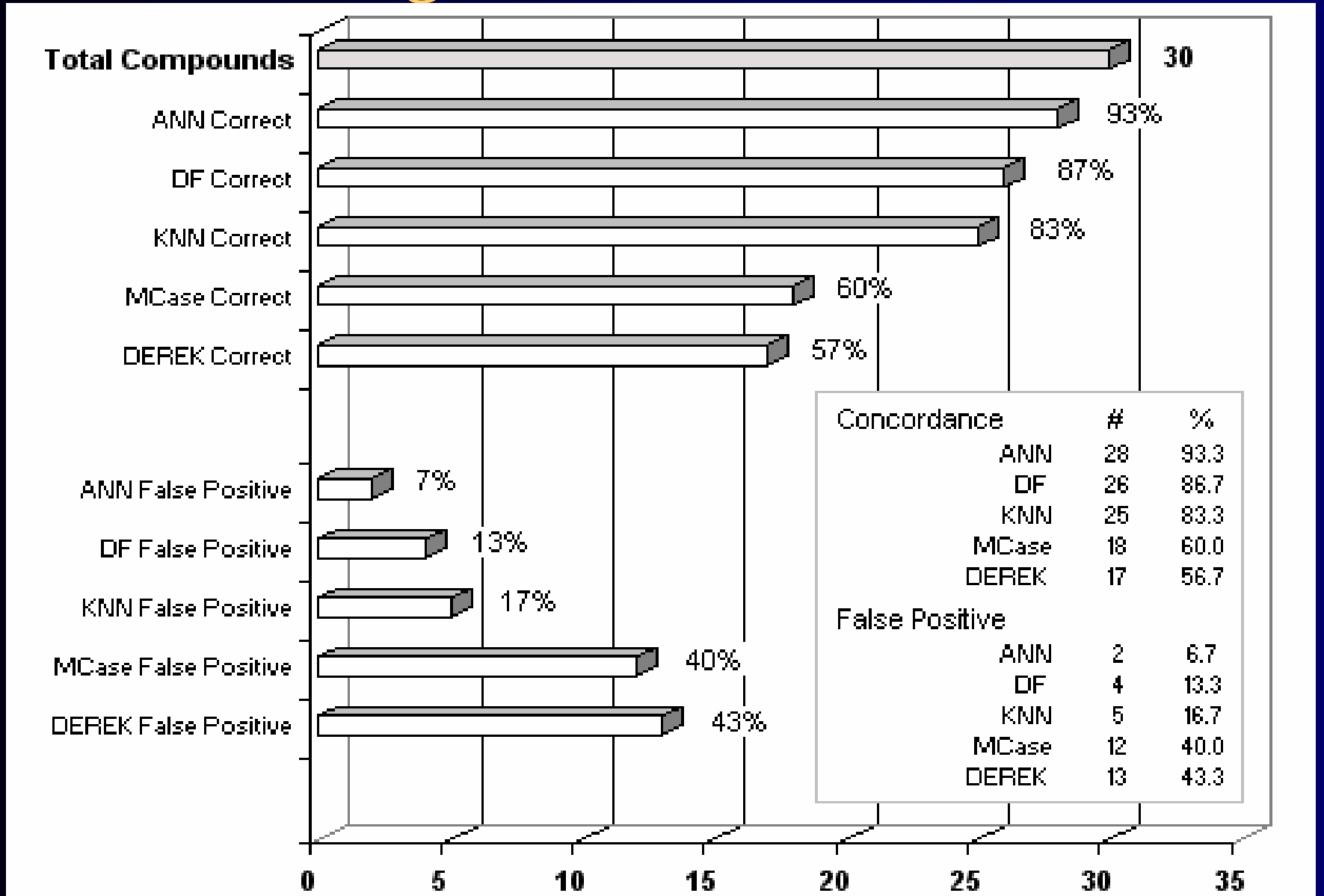
*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:… *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, 22, 69-77.

Example. Consensus QSPR models for the prediction of Ames genotoxicity*

- 3,363 diverse compounds (including >300 drugs) tested for their Ames genotoxicity
  - 60% mutagens, 40% non mutagens
  - 148 initial topological descriptors
  - ANN, kNN, Decision Forest (DF) methods
- 2963 compounds in the training set, 400 compounds (39 drugs) in randomly selected validation set

*Votano JR, Parham M, Hall LH, Kier LB, Oloff S, Tropsha A, Xie Q, Tong W. *Mutagenesis*, **2004**, 19, 365-77.

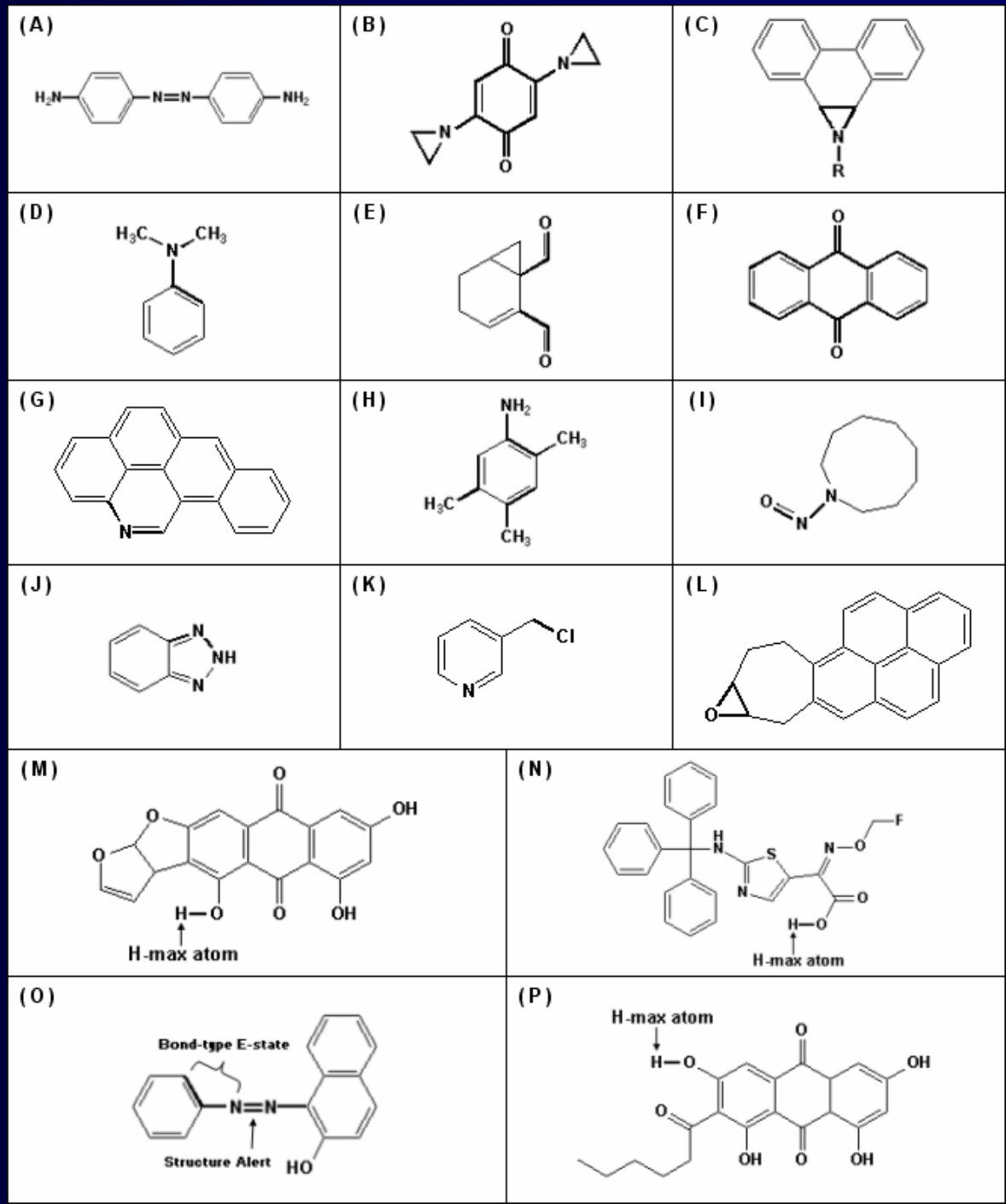
Comparison of GenTox prediction for 30 drugs in the external test set



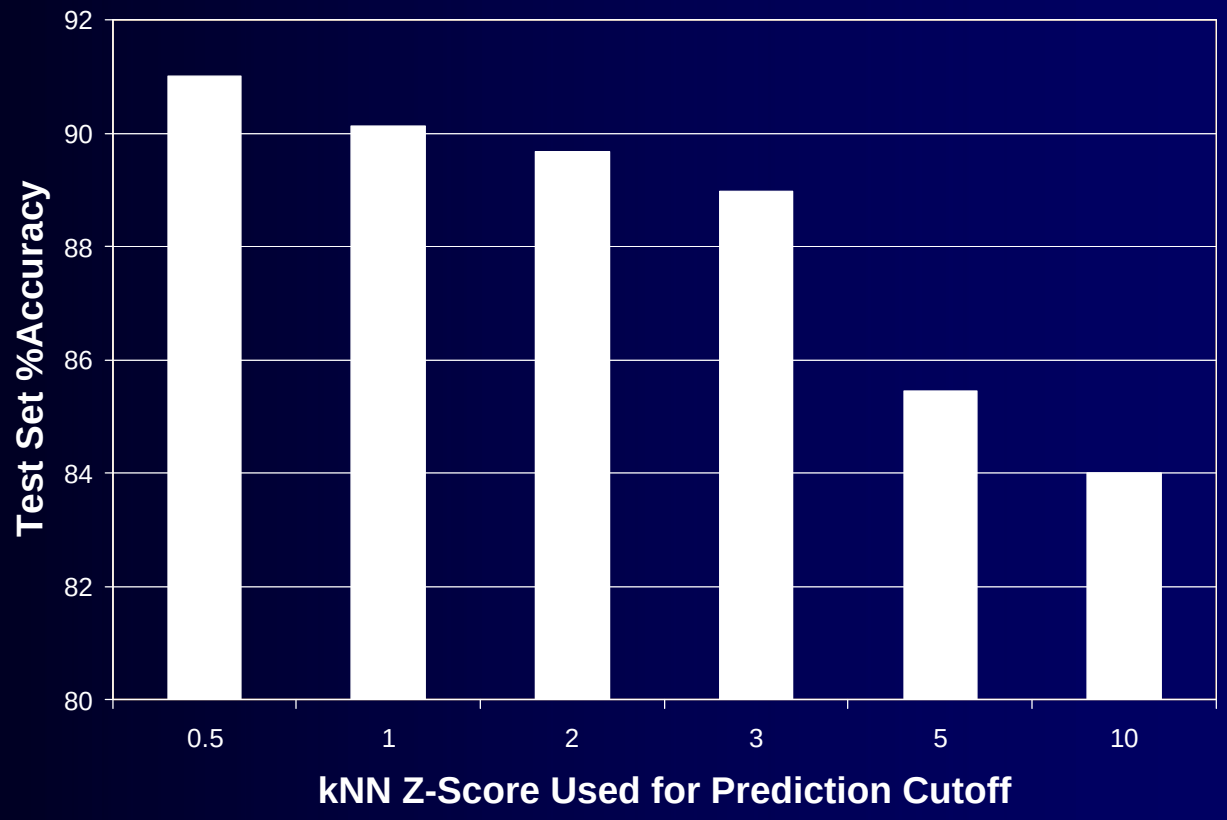
Concordance	#	%
ANN	28	93.3
DF	26	86.7
KNN	25	83.3
MCase	18	60.0
DEREK	17	56.7
False Positive		
ANN	2	6.7
DF	4	13.3
KNN	5	16.7
MCase	12	40.0
DEREK	13	43.3

Frequent MI descriptors map onto (some known) structural alerts

Bold, wide bonds show positions within structures where descriptors indicate a structural alert for Ames mutagenicity as found among most important E-State indices



Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)



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

Prediction of the External Set

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%

Carcinogenicity Model Based on the 187 Compounds

- Modeling set: 167 compounds
- External validation set: 20 compounds
- The number of kNN QSAR models based on modeling set for different cutoff values:

Training/test set predictivity cutoff	Chemical descriptors only	**Combined HTS+chemical descriptors**
0.7/0.7	315	919
0.75/0.75	29	86
0.8/0.8	1	4

Prediction of the 20 External Compounds

	Chemial 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%	

Modeling of the complete carcinogenicity dataset: The Carcinogenic Potency Database (CPDB)

- Lois Swirsky Gold, Ph.D., Director
- Unique and widely used international repository http://potency.berkeley.edu/
- 1485 chemicals
- Species, strain, and sex of test animals
- Target organ, tumor types, and tumor incidence
- Carcinogenic potency (TD50)
- Shape of the dose-response
- Experts's conclusion on carcinogenicity
- Literature citation through1997
- Incorporated in The Distributed Structure-Searchable Toxicity (DSSTox) Database Network. http://www.epa.gov/nheerl/dsstox/sdf_cpdbas.html

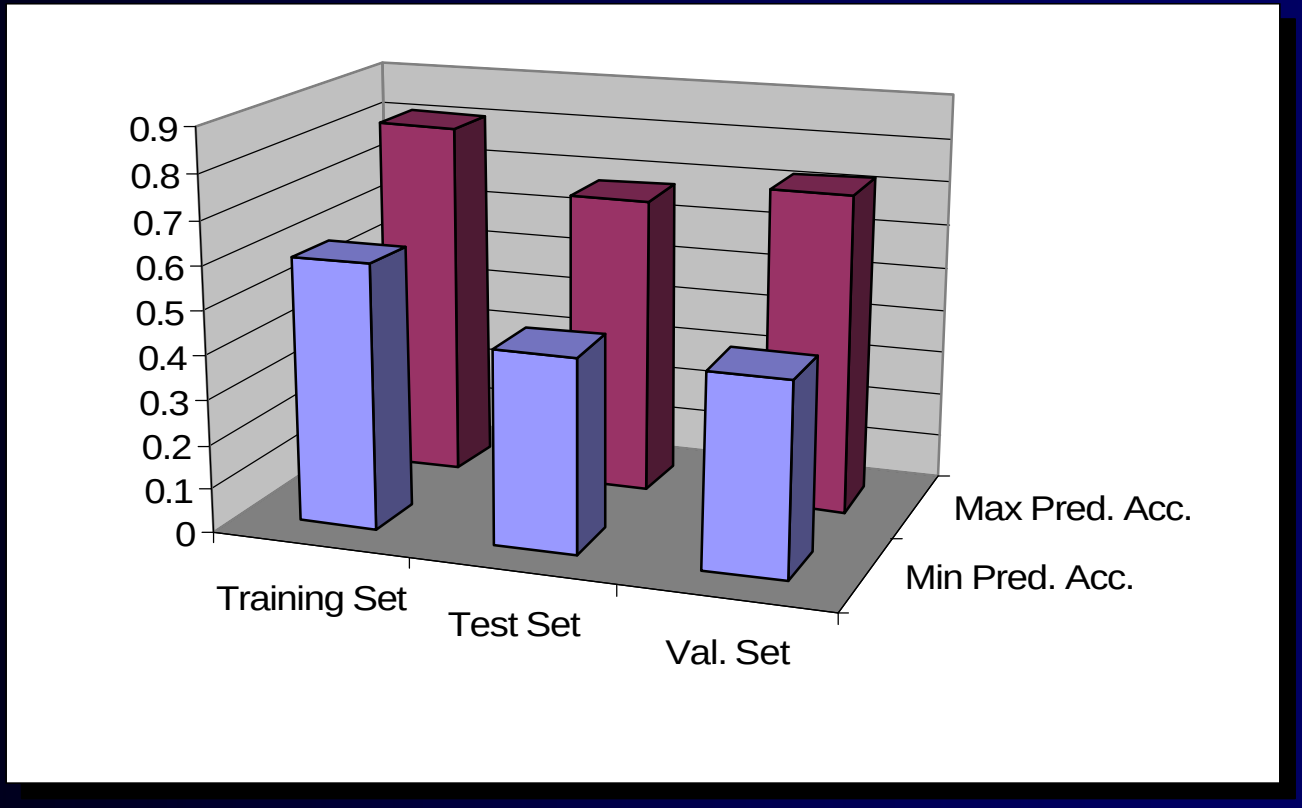
Database Curation

- Total entries: (1481)
- Delete entries with no structure (1444 left)
- Delete entries containing inorganic elements (1244 left)
- Clean duplicates / triplicates and keep one copy (1216 left)
- Delete chiral compounds (1214 left)
- Delete all entries missing mutagenicity data (693 left)

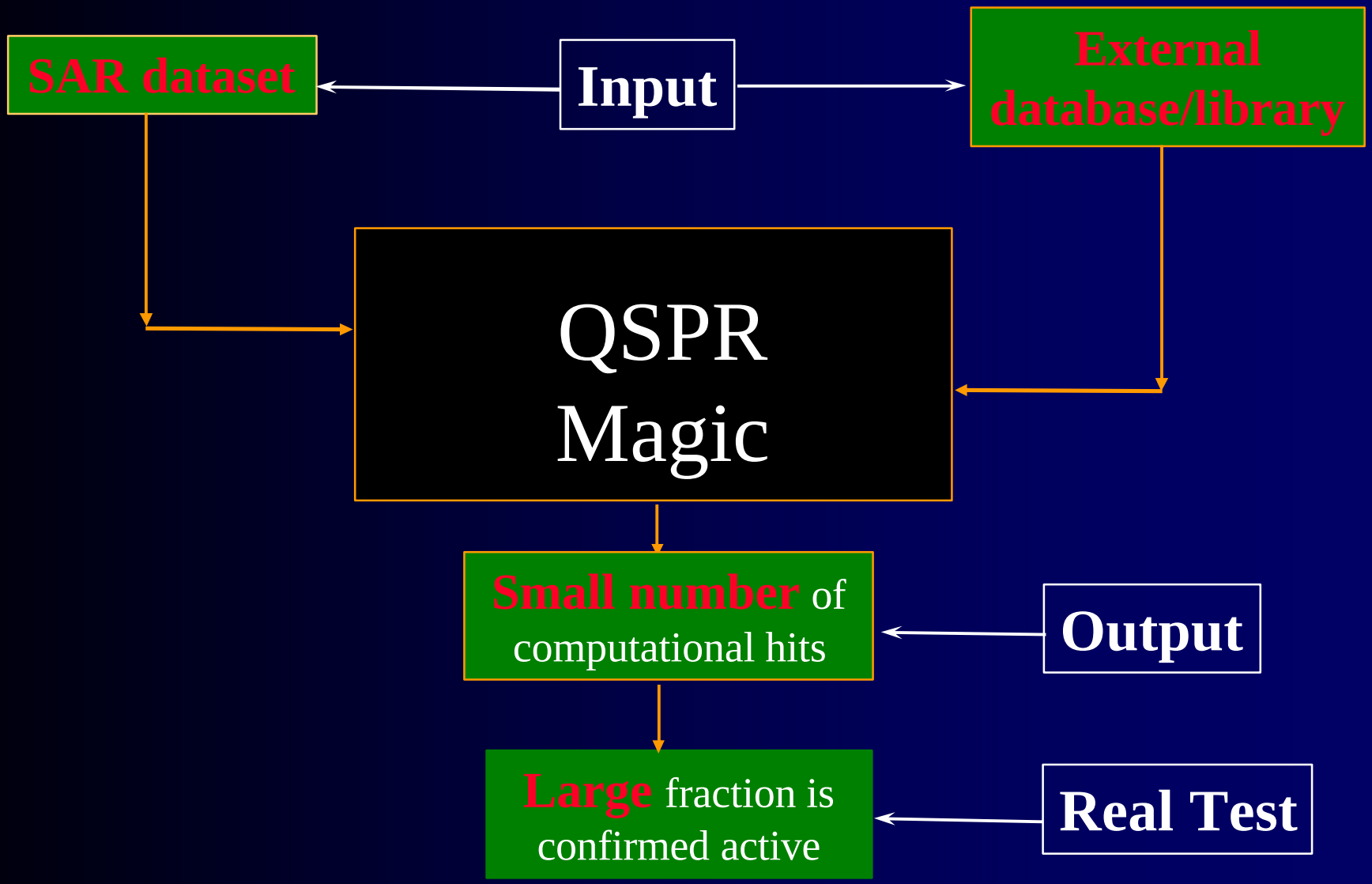
Statistics of a Working Subset for the Animal Carcinogenicity Modeling

	T Train/Test Set	Val. Set	Total
Inactive	210	59	269
Active	343	81	424
Total	553	140	693

Accuracy of kNN QSAR Models of Animal Carcinogenicity



QSPR Workflow: Emphasis on Successful Predictions, not statistics or interpretations



CECCR - ChemBench - Welcome - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://ceccr.unc.edu/chembench Go Links

03/02/2006 - Welcome, Guest ChemBench Help Center Member Login

C-CHEMBENCH v.1.0

ACCELERATING CHEMICAL GENOMICS RESEARCH BY CHEMINFORMATICS

Welcome | Tools | Predictors | ChemSearch

Username:
berkzafer

Password:
••••••••••

Sign in

Not a Member? Register for free.

Forgot your password?

ChemBench News

ChemSearch users, please click here

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Donec vitae ligula. Morbi pellentesque. Maecenas hendrerit orci non mi. Etiam lobortis porttitor massa. Sed urna erat, ultricies id, dictum ac, feugiat ut, diam. Donec risus augue, blandit consequat, fringilla et, dapibus porttitor, leo. Maecenas ut erat. Sed pulvinar nonummy enim. Praesent suscipit libero non urna.

Aenean tristique. Donec feugiat. Quisque lacus velit, sollicitudin facilisis, faucibus ut, mollis at, orci. In hac habitasse platea dictumst. Quisque sit amet ligula ut dui feugiat semper. Quisque ante. Pellentesque laoreet sapien. Morbi pede felis, fermentum sed, tincidunt sit amet, interdum facilisis, nisl. Donec condimentum elementum sapien. Duis vitae arcu. Aenean elementum ipsum a arcu.

ChemBench Project is supported by NIH grant number 1P20HG003898-01

UNC

ceccr.unc.edu/chembench

Internet

Summary and thoughts

The public has an insatiable curiosity to know everything, except what is worth knowing.

Oscar Wilde

- HTS and –omics data may be insufficient to achieve the desired accuracy of the end point property prediction. Should be explored as biodescriptors in conjunction with chemical descriptors
- Predictive QSPR workflow with extensive validation affords statistically significant models that can serve as reliable property predictors
- Mechanistic model interpretation should only be attempted IFF models have been externally validated

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
–Patricia LIMA
–Assia KOVACHEVA
–M. KARTHIKEYAN

Current

•Protein structure group:

– Yetian CHEN
– Tanarat KIETSAKORN
– Berk ZAFER

Cheminformatics group:

- Kun WANG
– Alex GOLBRAIKH
– Raed KHASHAN
– Chris GRULKE
- Hao TANG
- Simon WANG
- Hao ZHU

- Rima HAJJO
- Mei WANG
- Julia GRACE
- Hao HU
- Mihir SHAH
- Jui-Hua Hsieh
- Tong-Ying Wu

- Funding
  - NIH RoadMap
  - EPA (STAR award)
- Collaborators
  - Ann Richard (EPA)
  - Ivan Rusyn (UNC)
  - Tudor Oprea (UNM)