

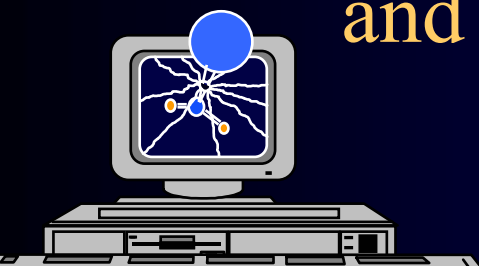
QUIZZING QSAR MODELS: TRUTH OR DARE?

Alexander Tropsha

Laboratory for Molecular Modeling

School of Pharmacy

and Carolina Center for Genome
Sciences



UNC-Chapel Hill

QSAR Modeling

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

Structure	Activity (IC50, Kd...)	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)

Components of QSPR 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

Is QSAR Models Predictive? BEWARE OF q^2 !!!

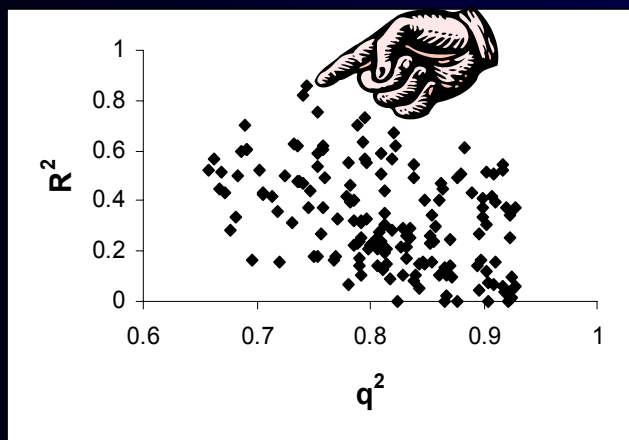
(Golbraikh & Tropsha, *J. Mol. Graphics Mod.* 2002, 20, 269-276.)

31 Cramer steroids [1] (Benchmark to investigate novel QSAR methods [2])

Predictive R^2 versus cross-validated $R^2(q^2)$ for QSAR models with $q^2 > 0.5$. using common definition (e.g., [3]) of training and test sets.

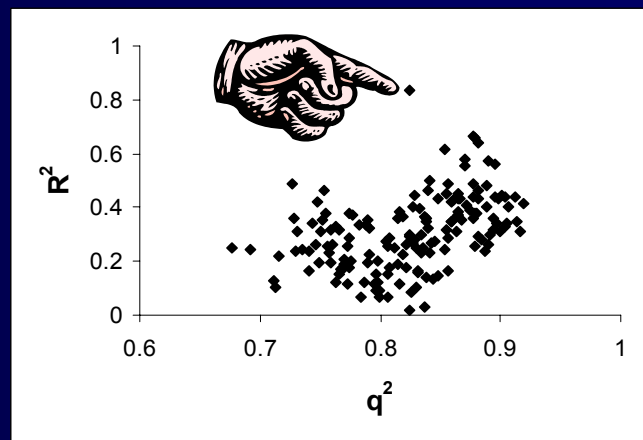
Training set: compounds 1-21

Test set: compounds 22-31



Training set: compounds 1-12 and 23-31

Test set: compounds 13-22



1. Cramer, R.D. III, Patterson, D.E., Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am.Chem.Soc.* **1988**, 110, 5959-5967
2. Coats, E.A. *The CoMFA steroids as a benchmark data set for development of 3D QSAR methods. In 3D QSAR in Drug Design. V.3.* Kubinyi, H., Folkers, G., Martin, Y.C., Eds. Kluwer/ESCOM:Dordrecht, 1998, pp 199-213.
3. Kubinyi, H.; Hamprecht, F.A. & Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D-QSAR) for SEAL Similarity Metric. *J. Mol. Chem.* **1999**, 41, 2553-2564

Competing View of a Statistician:

An excerpt from a recent publication...[1]

“The only motivation to rely on the holdout sample rather than cross-validation would be if there was reason to think the cross-validation not trustworthy - biased or highly variable. But neither **theoretical** results nor the **empiric** results sketched here give any reason to disbelieve the cross-validation results.”

“This conclusion is opposite to the recommendations of Golbraikh and Tropsha [2]. ... We believe therefore that their advice is wrong.”

1. D.M.Hawkins, S.C.Basak, D.Mills. Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci. 2003, 43, 579-586.

2. A.Golbraikh, A.Tropsha. Beware of q^2 ! J.Mol.Graph.Mod. 2002, 20, 269-276.

THE EXPERIMENT OF HAWKINS et al.

- **300 analyses**
- **Each analysis:**
 - ✓ select randomly 100 out of 469 compounds as the training set
 - ✓ select randomly 10, 20 and 50 out of 369 compounds as test sets (holdouts)
 - ✓ use 319 remaining compounds to calculate “true” R^2
 - ✓ select randomly 5 (75 times), 10 (75 times), 20 (75 times), 50 (75 times) out of 232 descriptors

Some results of Hawkins et al

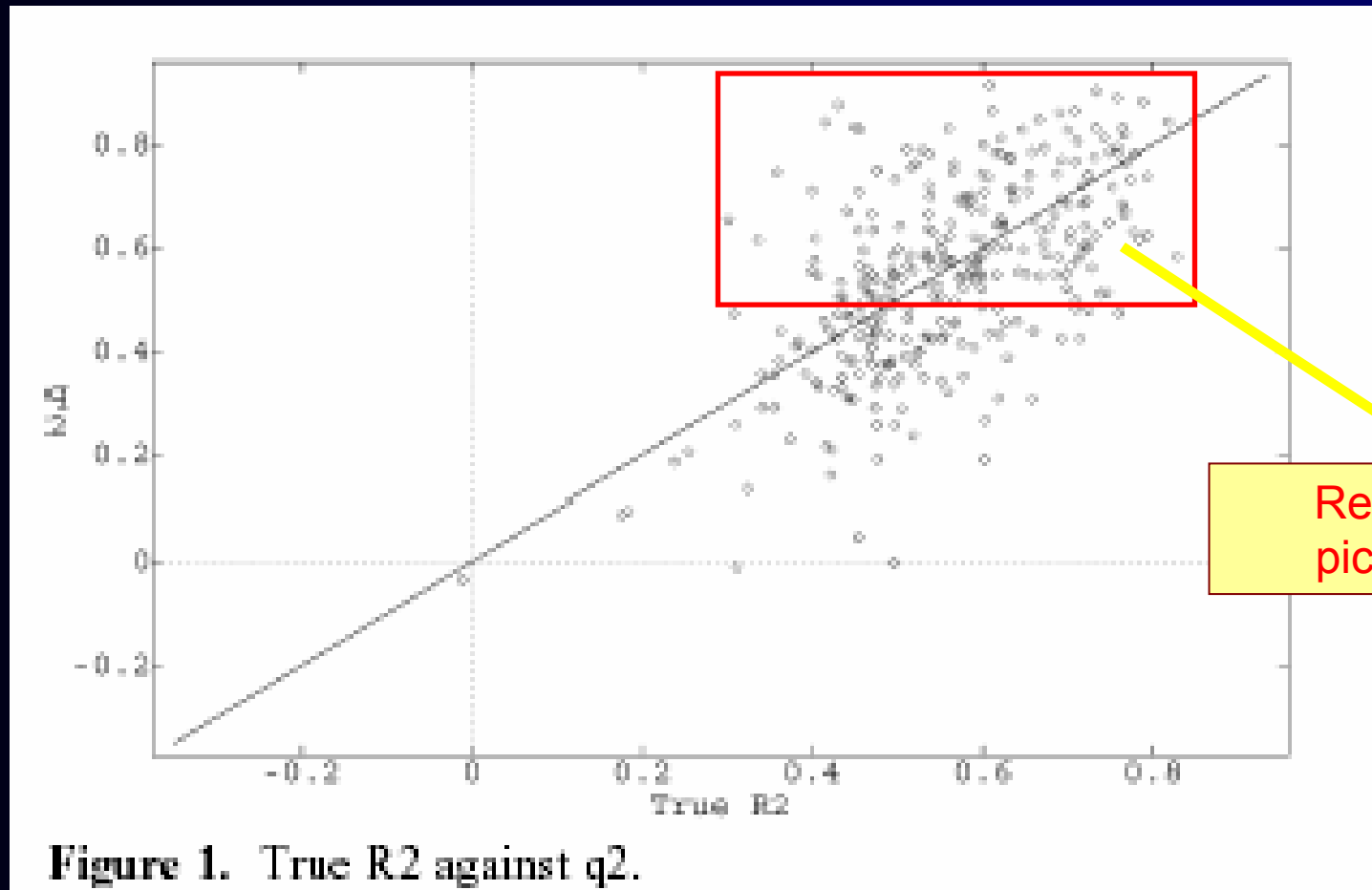


Figure 1. True R^2 against q^2 .

CONCLUSION: “The cross-validation q^2 seems to give a reliable picture of the true R^2 . There is no apparent systematic over- or underestimation.”

DEVISING PREDICTIVE QSAR WORKFLOW

- QSAR as an empirical data modeling exercise
 - Combinatorial QSAR
 - Extensive model validation
 - Applicability domain
- Examples of the Workflow Applications, including application of predictive QSAR models to virtual screening
- Application of QSAR methodologies to docking/scoring
 - Scoring (EnTess)
 - Docking in chemometric space (CoLIBRI)
- Final thoughts

QSPR modeling process revisited



Pharmaco-

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

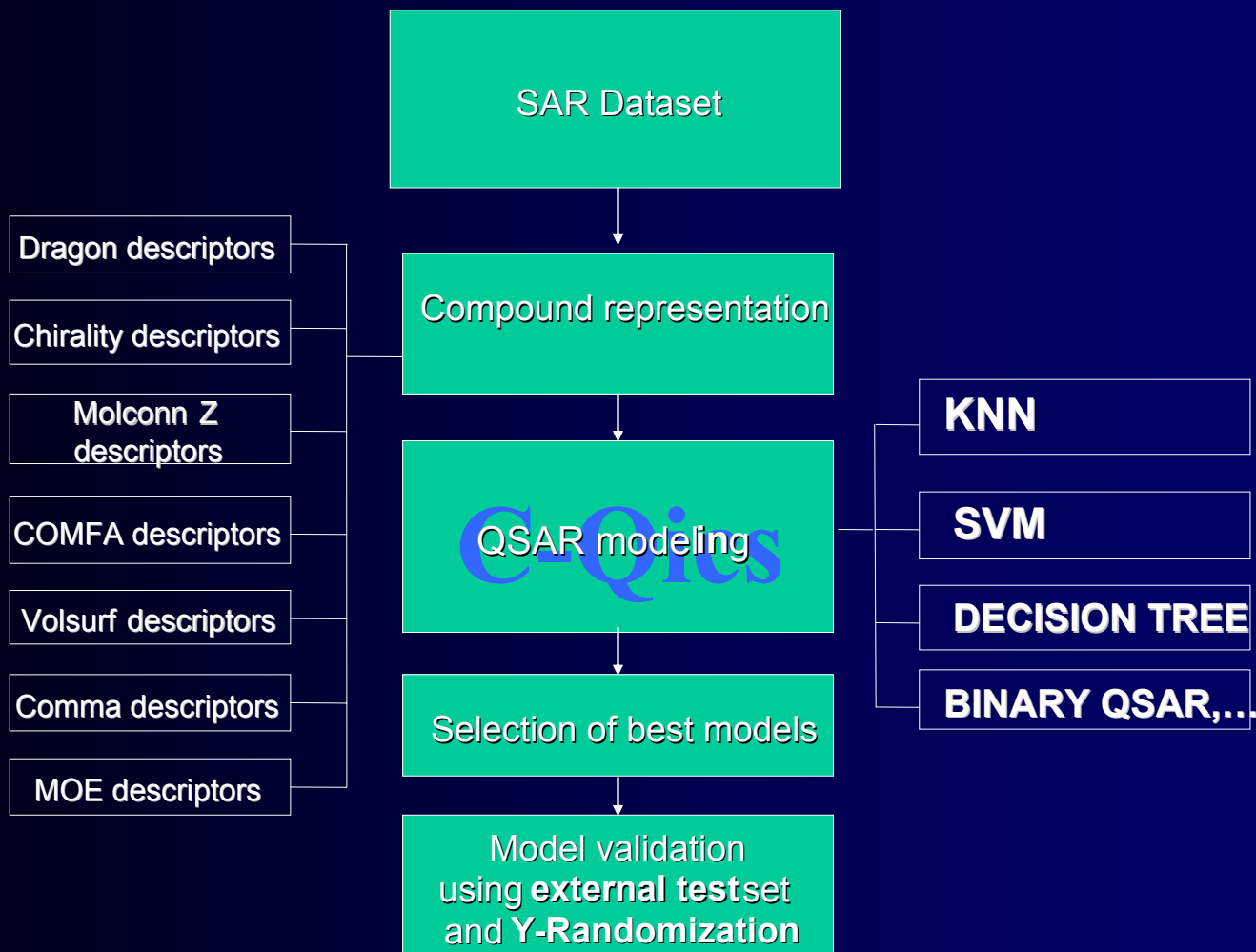
-ICS

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

COMBINATORIAL QSARomics, or

C-Qics

COMBINATORIAL QSAR



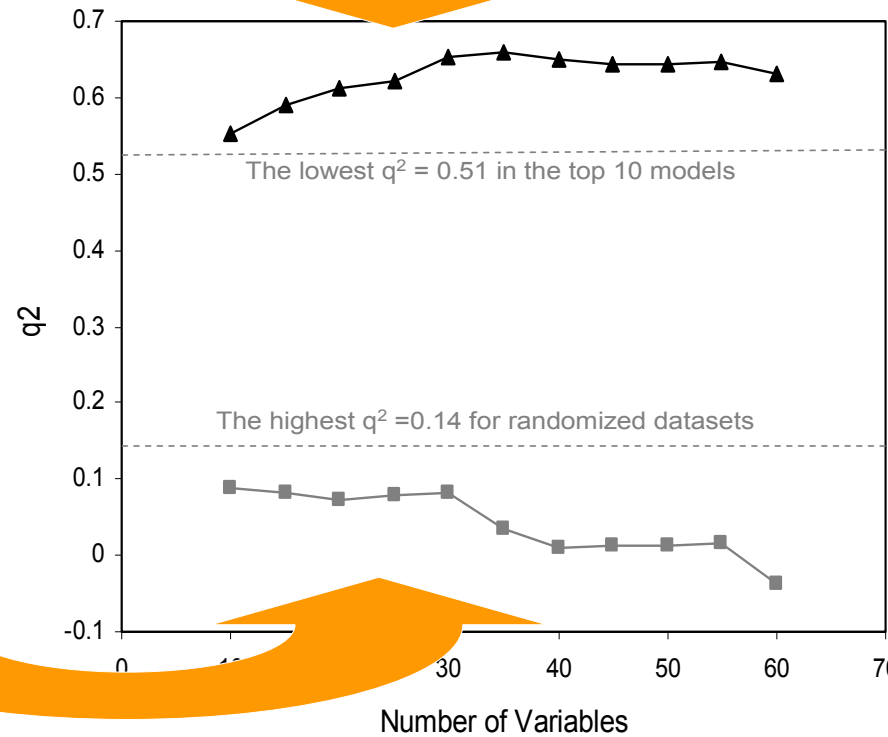
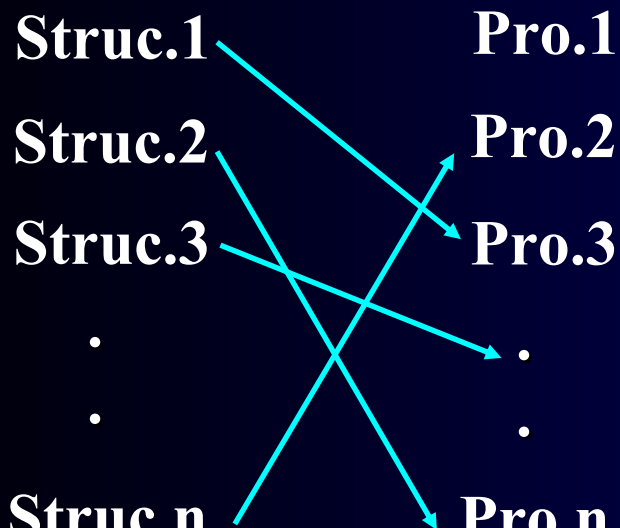
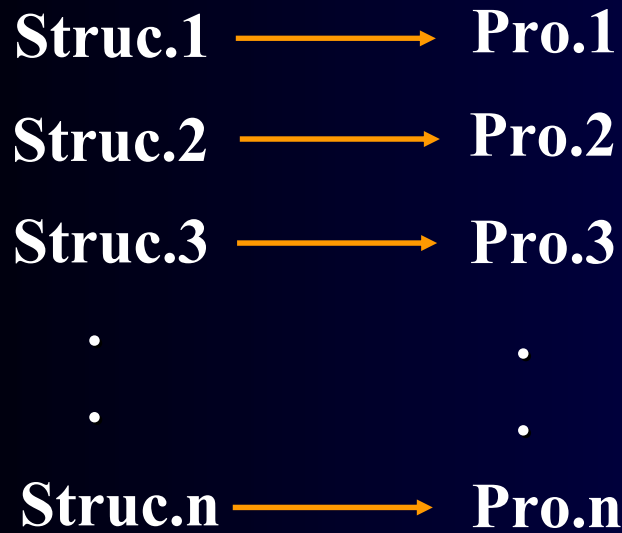
COMPONENTS OF PREDICTIVE QSAR MODELING WORKFLOW*

- 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

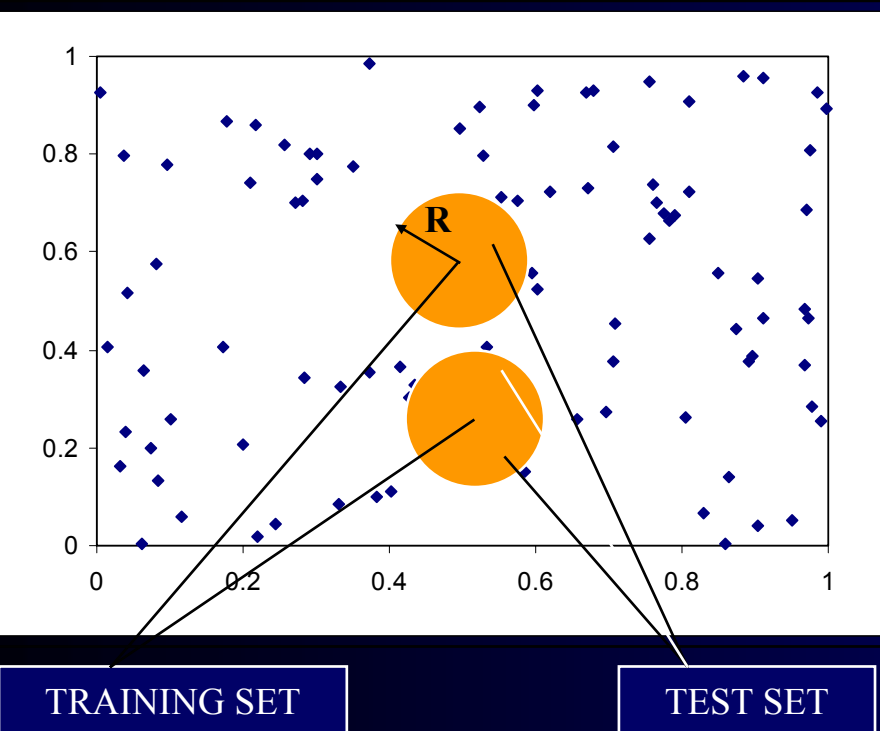
*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:...

Quant. Struct. Act. Relat. Comb. Sci. 2002, 22, 60-77

Activity randomization



RATIONAL SELECTION OF TRAINING AND TEST SETS BASED ON DIVERSITY SAMPLING



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} , $i=1, \dots, n$, $j=1, \dots, m$ - 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.

$$V_p = 1/N \quad R = cV_p^{1/K}$$

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

DEFINING THE APPLICABILITY DOMAIN

Training set: 60 compounds
Test set: 35 compounds

MODEL:

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

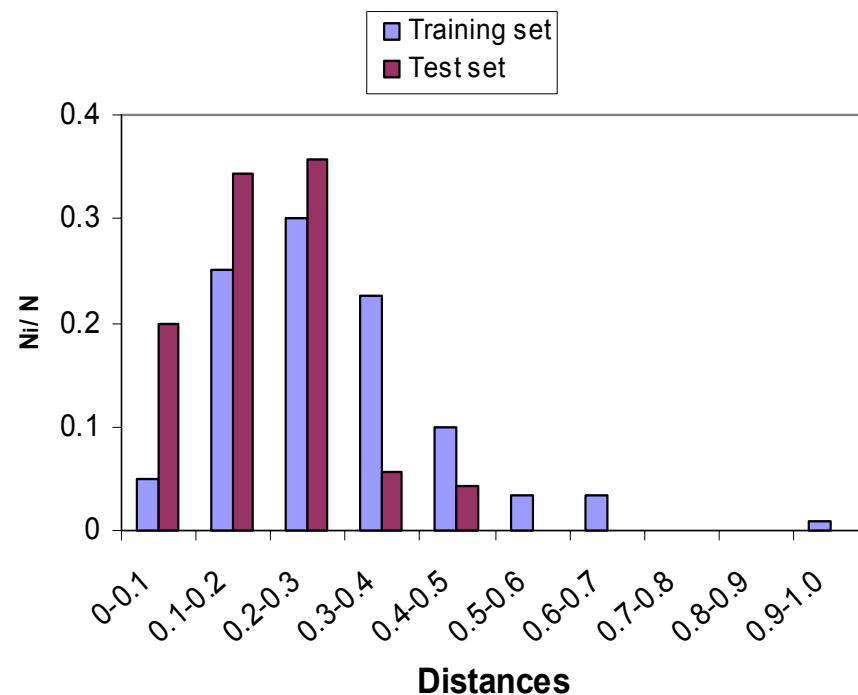
DISTANCES:

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

Closest nearest neighbors of
test set compounds:

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

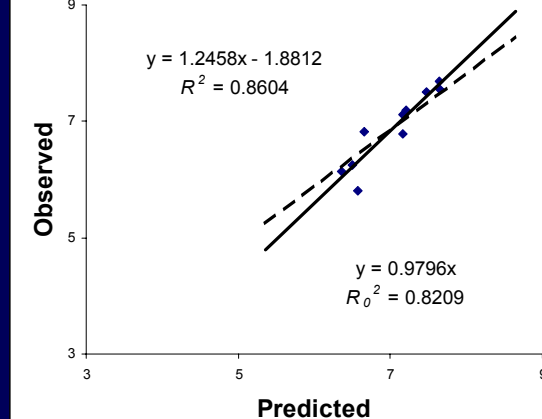
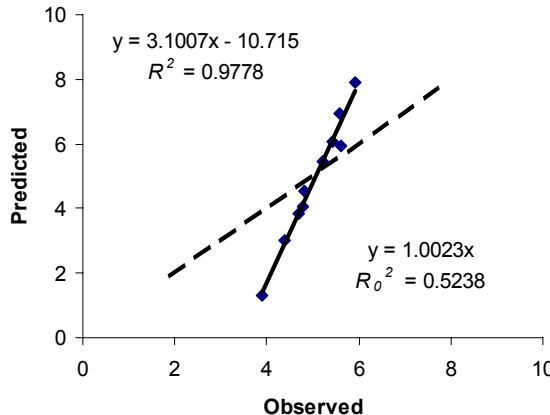
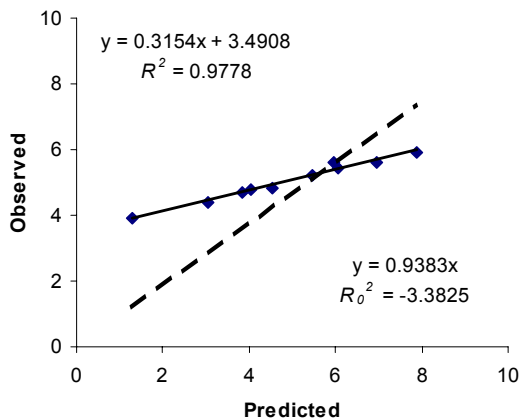
Distribution of distances between points and their nearest neighbors in the training set



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

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

Criteria for Predictive QSAR Model.



Regression

Correlation coefficient

Regression through the origin

Coefficients of determination

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

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

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

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

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

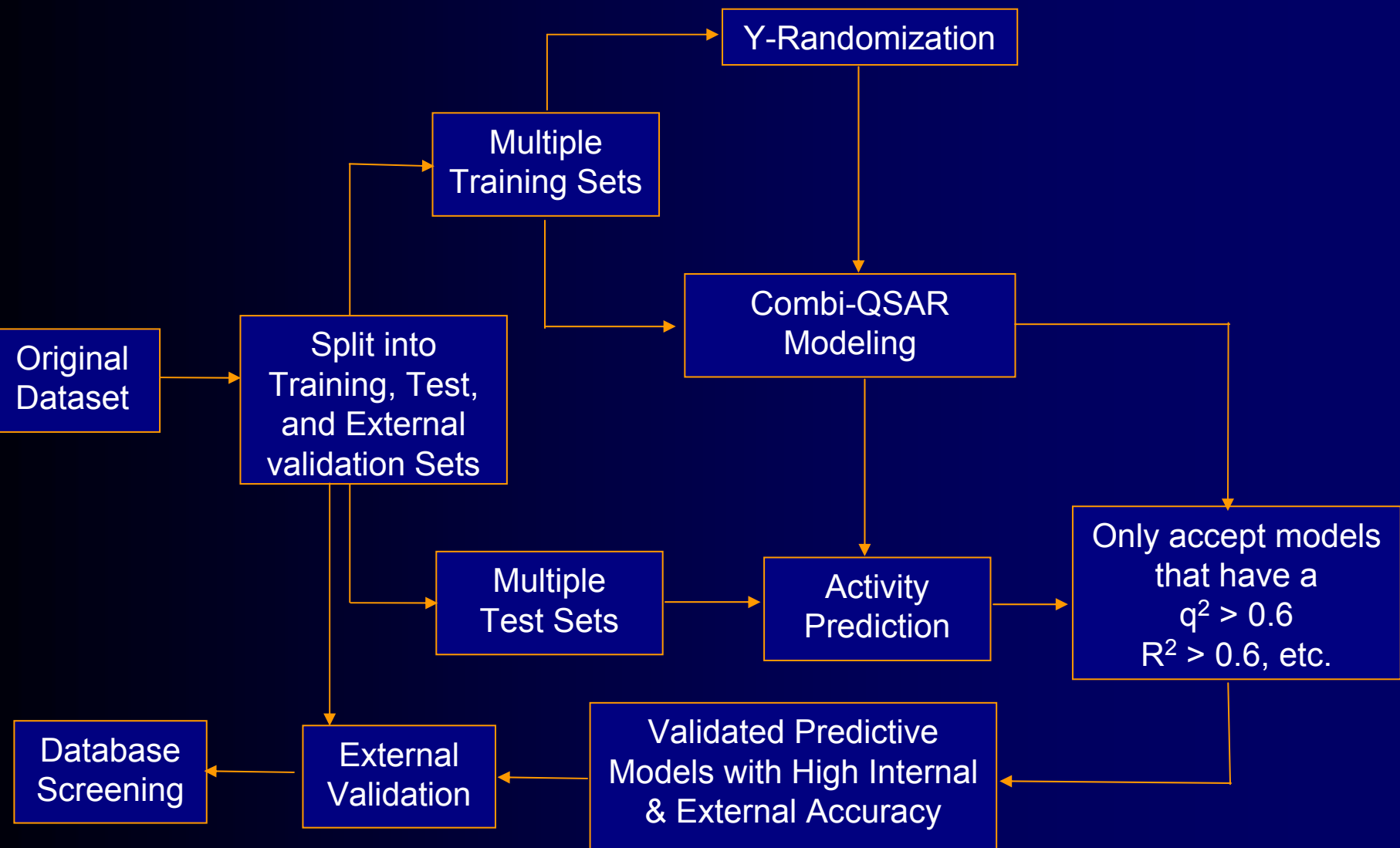
$$R'^2_0 = 1 - \frac{\sum (y_i - \tilde{y}_i^{r_0})^2}{\sum (y_i - \bar{y})^2}$$

CRITERIA

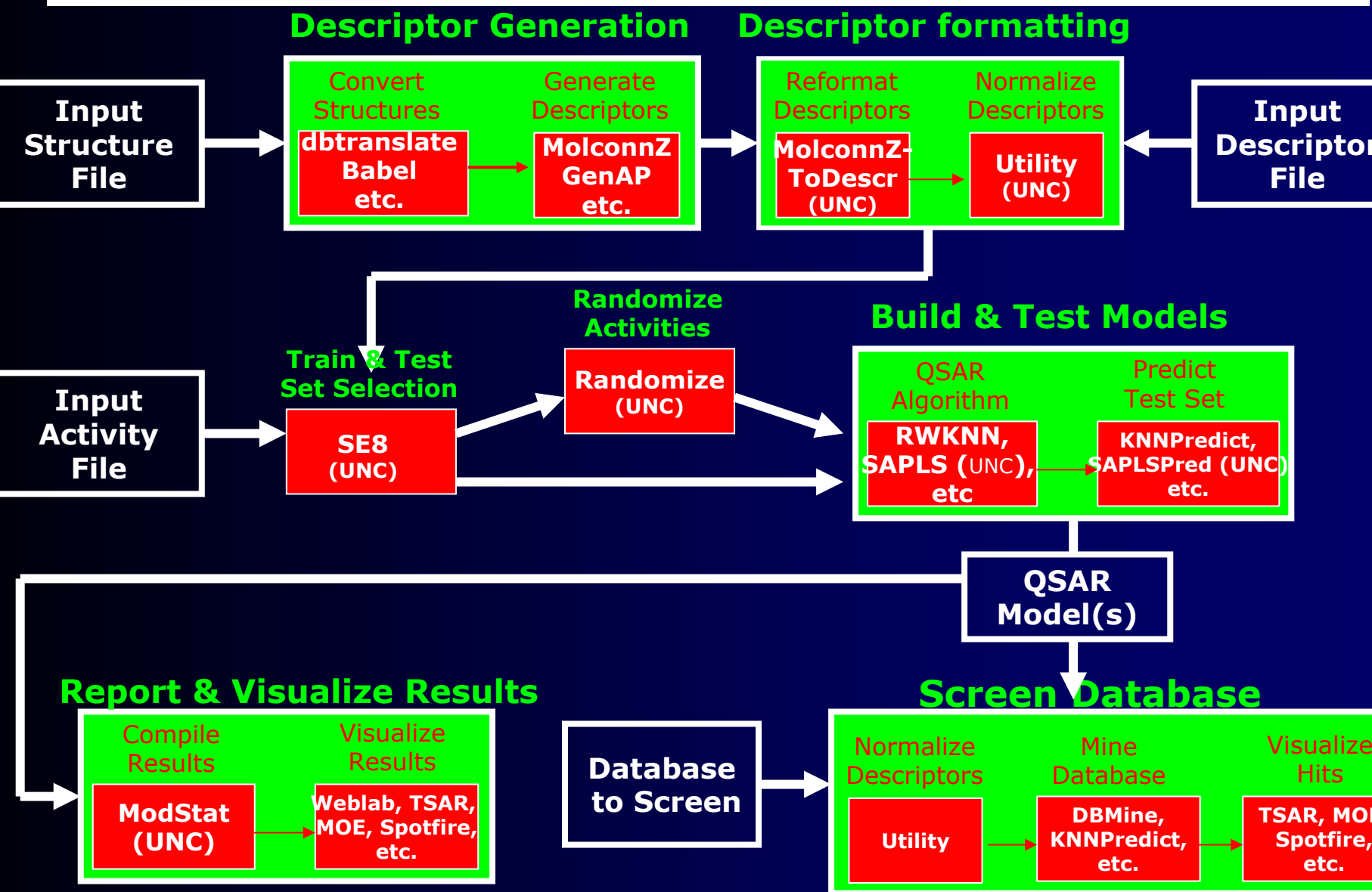
$$q^2 > 0.5; R^2 > 0.6;$$

$$k \text{ or } k' \approx 1.0; R_0^2 \text{ or } R'^2_0 \approx R^2$$

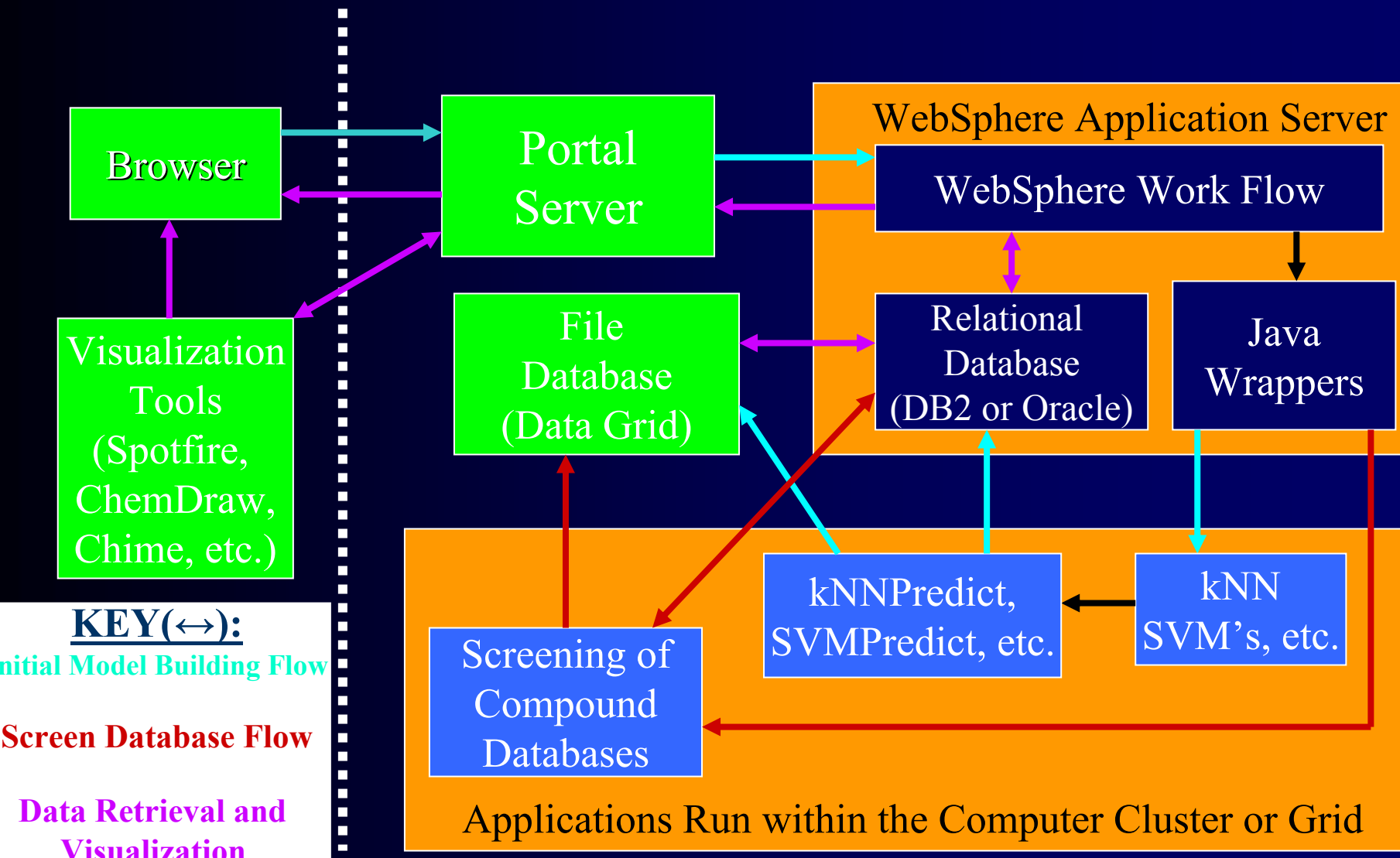
Predictive QSAR Workflow



Structure-Activity Relationships for the Design of Molecules (STARDOM™): WORKFLOW



STARDOM™ as an Automated Web Based, Grid Enabled Application

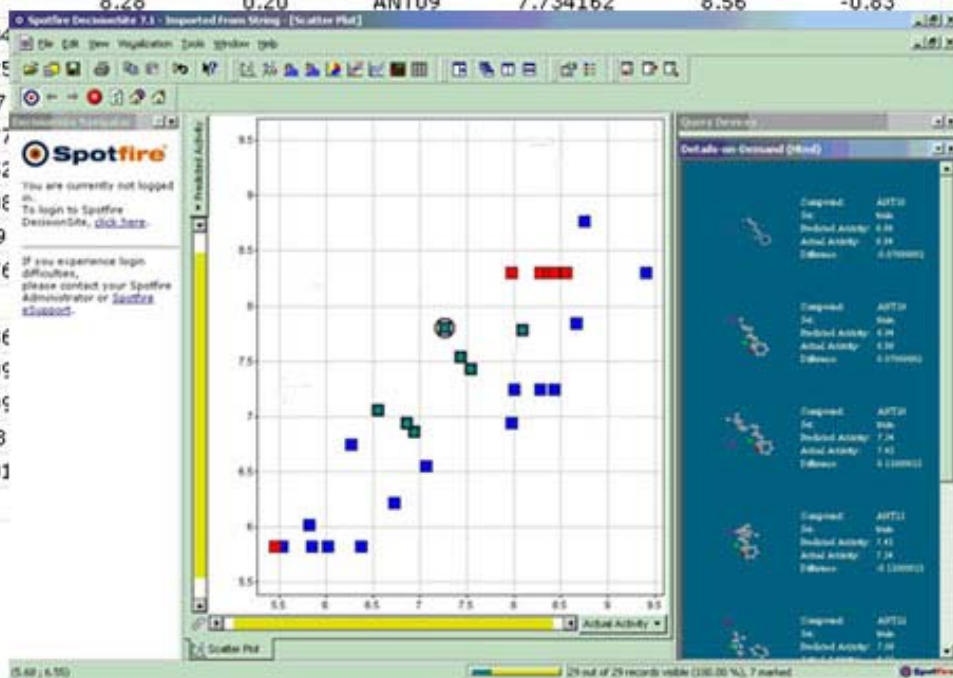


THE MODEL DETAILS.

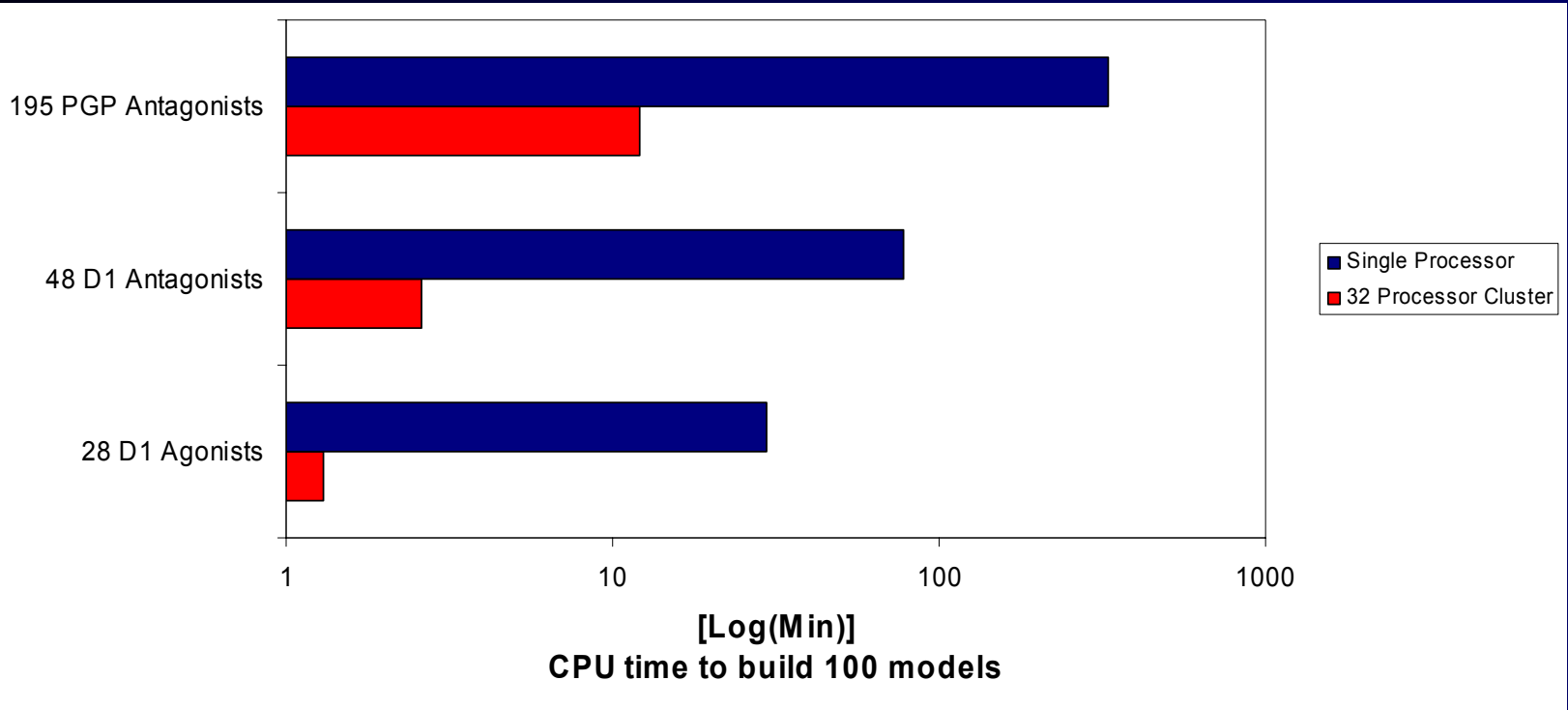
Load: 0

SPOTFIRE

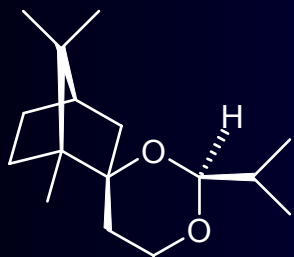
Results		Training Set				Test Set			
Parameter	Value	Compound	Predicted Activity	Actual Activity	Difference	Compound	Predicted Activity	Actual Activity	Difference
Status	Completed	ANT01	8.8	9.0	-0.20	ANT04	7.749543	8.43	-0.68
Risk Min	20	ANT02	7.783544	7.97	-0.19	ANT05	7.732079	8.38	-0.65
Allowed Error	180	ANT03	7.58189	8.28	-0.70	ANT07	7.75548	8.43	-0.67
Model ID	6	ANT06	8.48	8.28	0.20	ANT09	7.734162	8.56	-0.83
C	180.0	ANT08	7.801834						
Epsilon	0.2	ANT10	8.022425						
Training R ²	0.921158	ANT12	6.49057						
Test R ²	0.906579	ANT14	8.620377						
Number of Descriptors	229	ANT16	8.794252						
Constant B	8.767768	ANT17	8.100008						
		ANT18	6.88509						
		ANT19	7.234176						
		ANT20	7.23						
		ANT21	7.655836						
		ANT22	6.749999						
		ANT23	7.259999						
		ANT25	5.72998						
		ANT27	5.820001						
		ANT28	6.02						



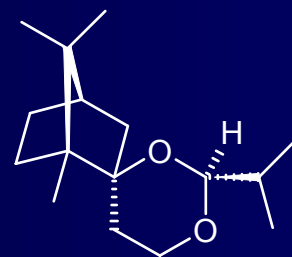
Computational Efficiency of STARDOM™



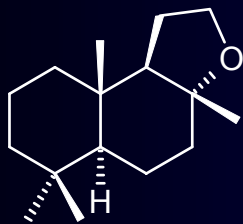
EXAMPLE 1: COMBINATORIAL QSAR OF AMBERGRIS FRAGRANCE COMPOUNDS*



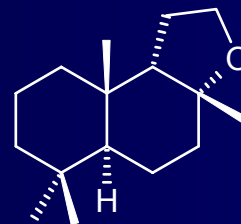
**Amber, woody,
cedarwood,
animal, strong**



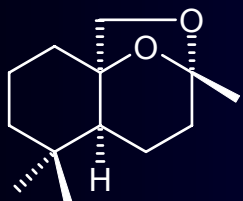
**Amber woody,
camphoraceous,
spicy, weak**



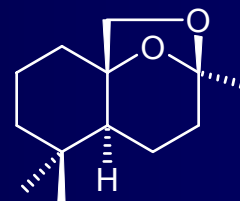
**Amber, exotic
woody, animal**



Strong amber

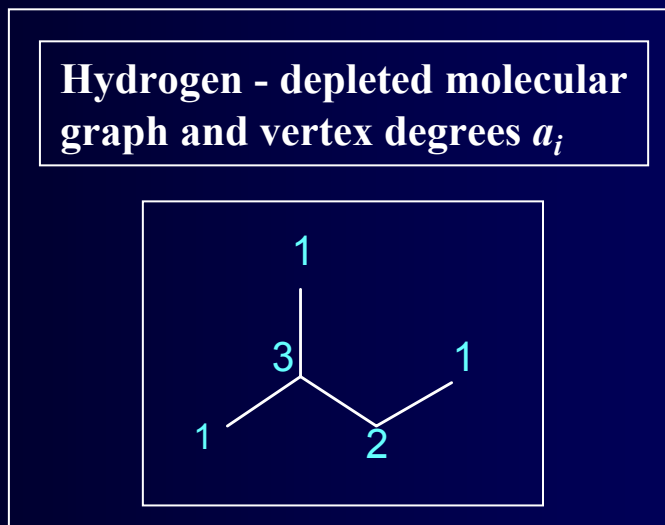


**Amber,
camphoraceous**



**Amber woody,
sea water**

CHIRALITY DESCRIPTORS DERIVED FROM 2D TOPOLOGY*



- Vertex degree a_i is defined as the number of edges originating from this vertex

$$a_i = \sum_{j=1}^N a_{ij}$$

- Correction for an asymmetric atom [D. Bonchev]:

$$a_i \leftarrow a_i + c \quad \text{for R-atom}$$

$$a_i \leftarrow a_i - c \quad \text{for S-atom}$$

Chiral Atom Pair* descriptors

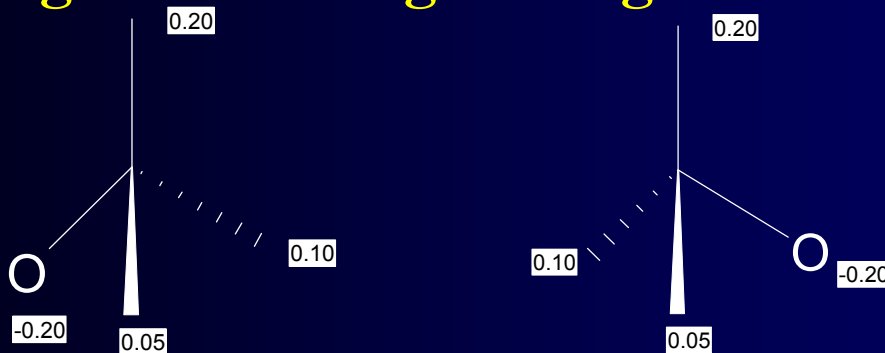
atom type i ----- (distance) ----- atom type j

15 atom types:

negative/positive charge center, HBA/HBD, N, O, S,
Double/triple bond center etc.

2 additional chiral atom types :

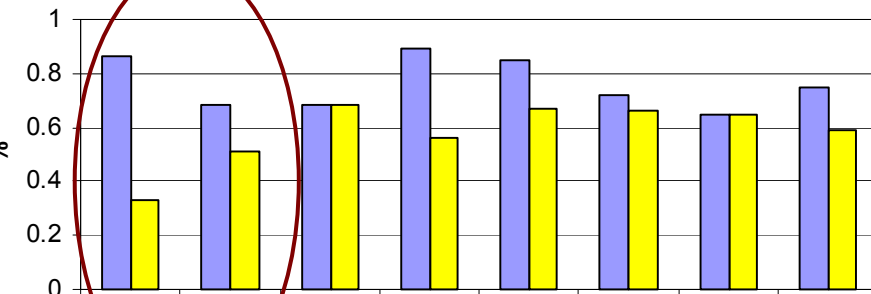
special atom types defined by the partial charge values
of substituents at a chiral atom: substituents with
higher charge values are given higher seniority



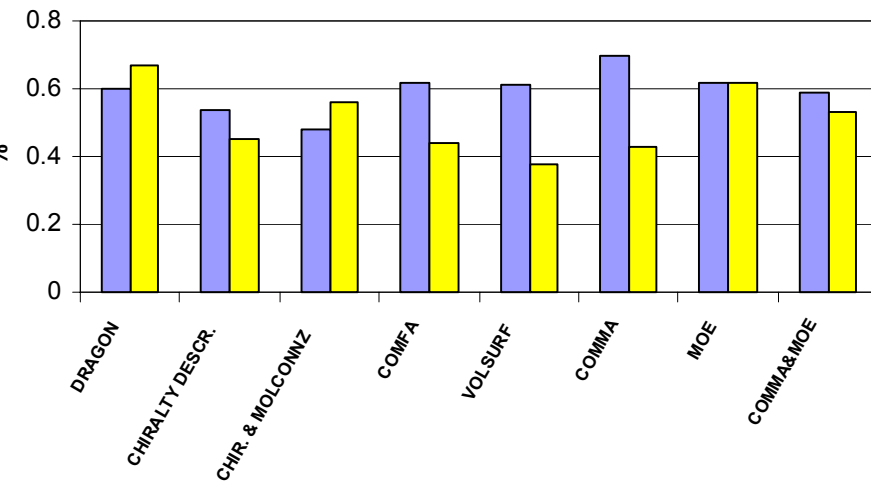
TOTAL PREDICTION ACCURACY FOR THE TEST SET USING BEST ACTUAL & RANDOMIZED MODELS

Test set predicted by real models ■ Test set predicted by randomized models

KNN-CLASSIFICATION

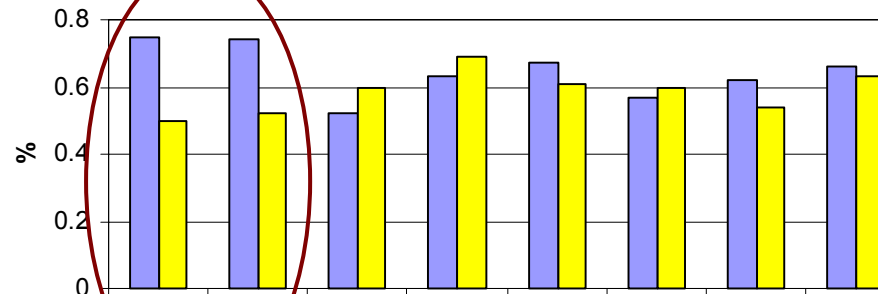


BINARY QSAR

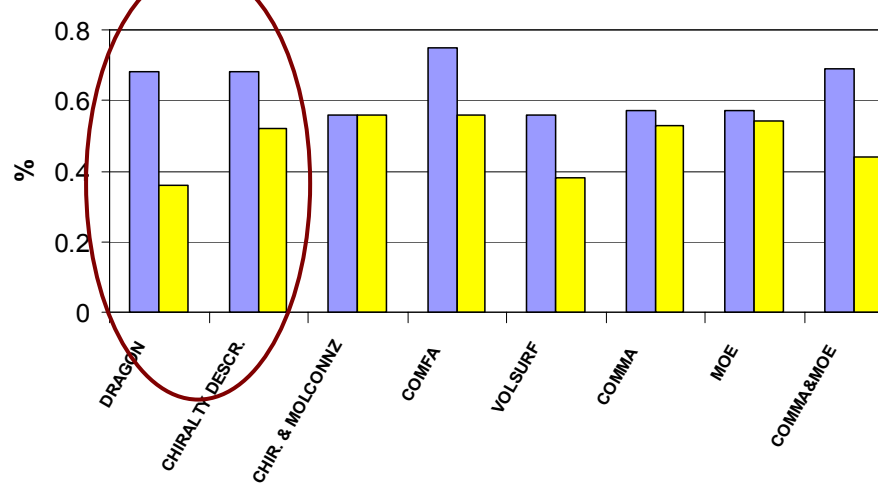


Test set predicted by real models ■ Test set predicted by randomized models

DECISION TREE



SUPPORT VECTOR MACHINES



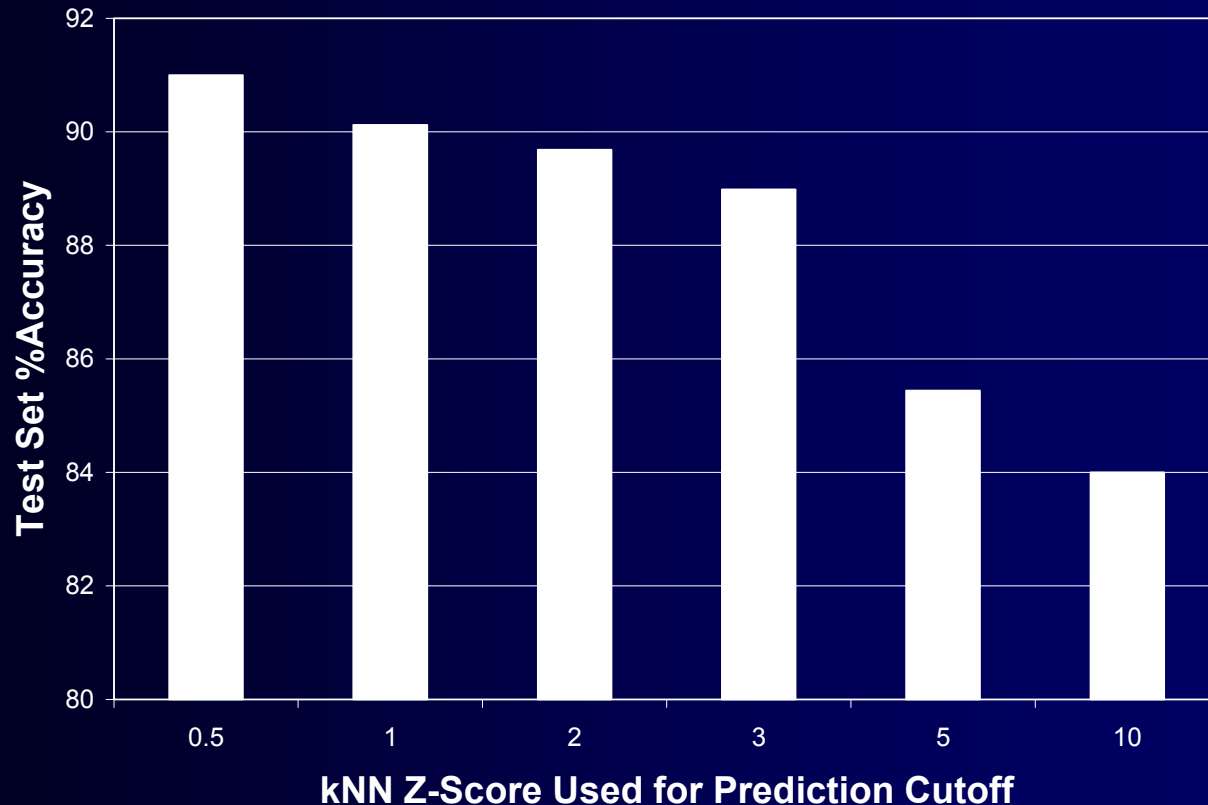
Correct classification rate (%) for internal and external prediction

Method \ Descriptors	KNN		BINARY CLASSIFICATION		DECISION TREE		SVM	
	Training	Test	Training	Test	Training	Test	Training	Test
FRAGON	70	86	72	76	70	78	83	68
MTD	72	65	76	50	67	74	81	58
MTD/ COLCONNZ	67	60	85	47	62	53	87	53
MOE	77	65	74	86	74	71	77	65
COMFA	76*	89*	71	65	75	62	83	75
OLSURF	78	85	74	70	77	60	94	53
COMMA/MOE	77	75	73	70	74	72	73	69
MAP	82	94						

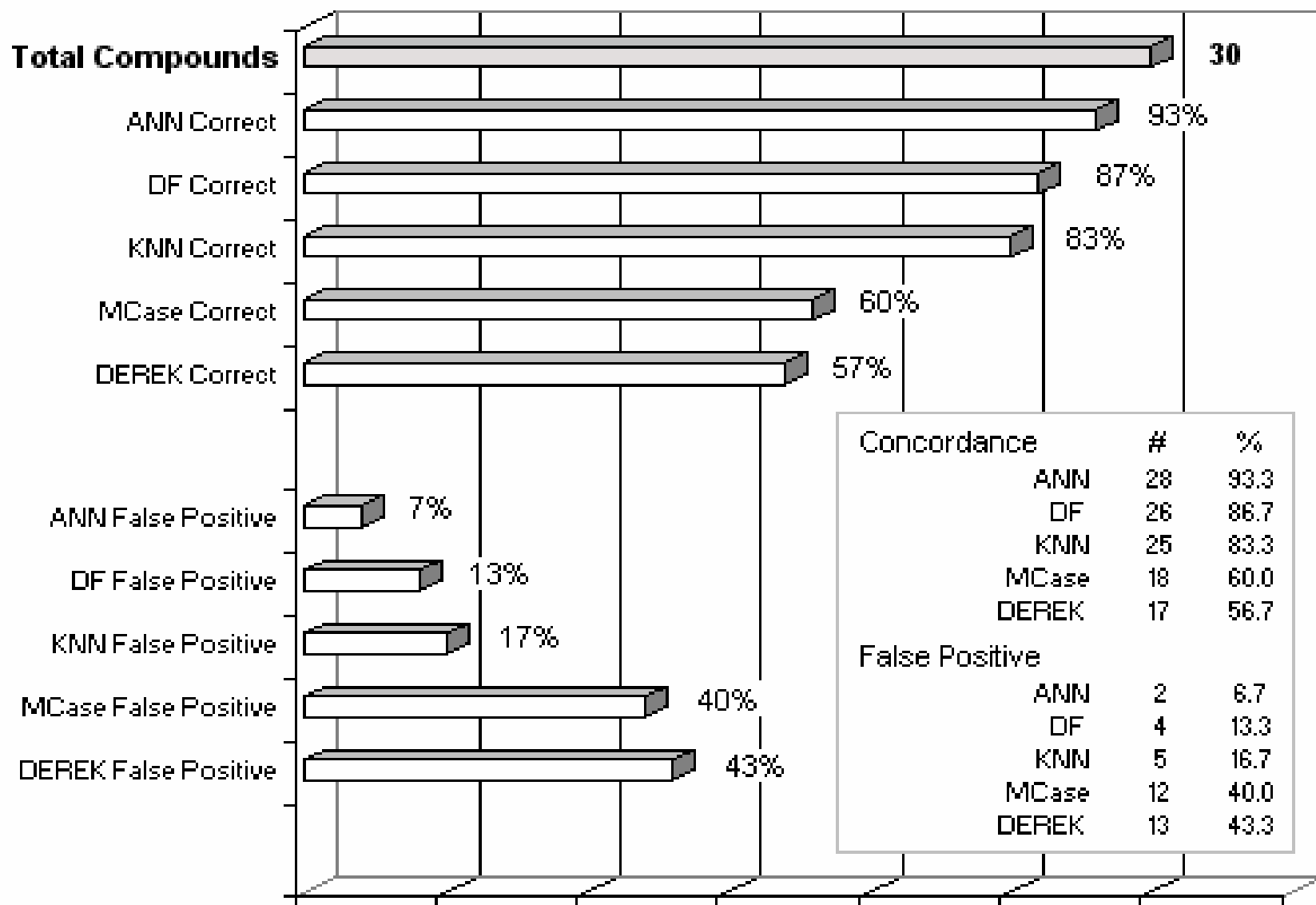
Example 2. Consensus QSAR models for the prediction of Ames genotoxicity*

- 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 test set

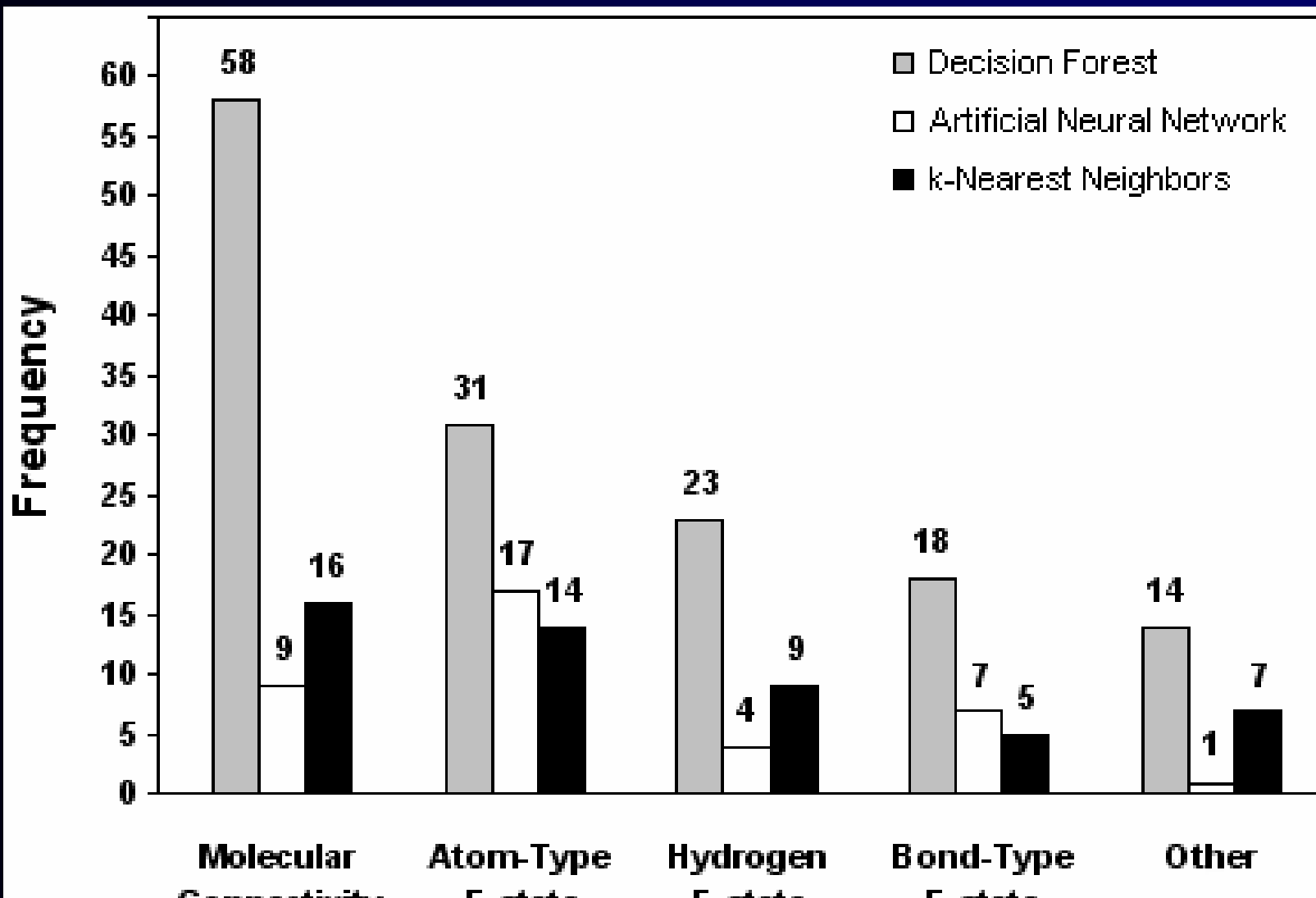
Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)



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

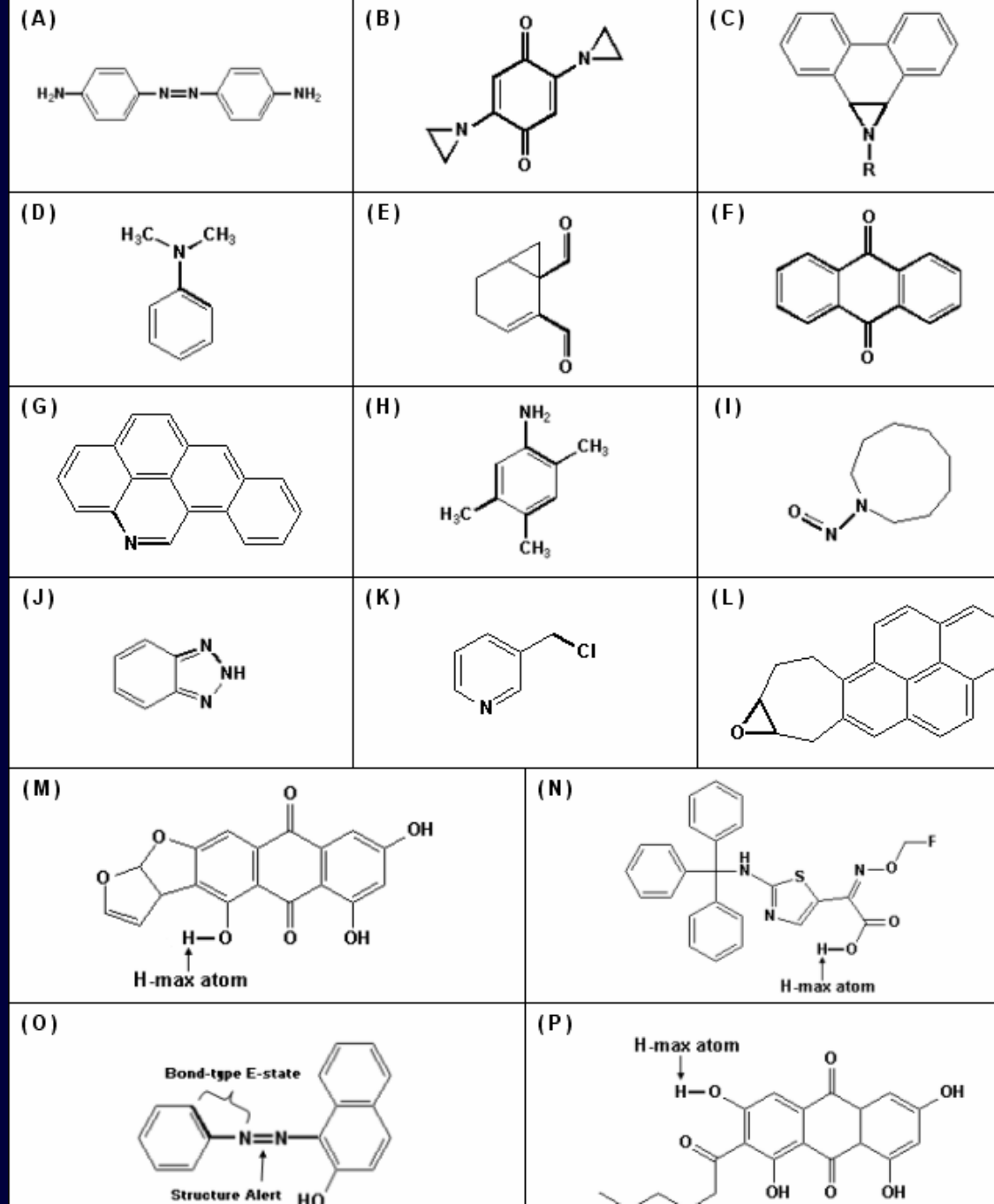


Content-dependent descriptor types identified by different models



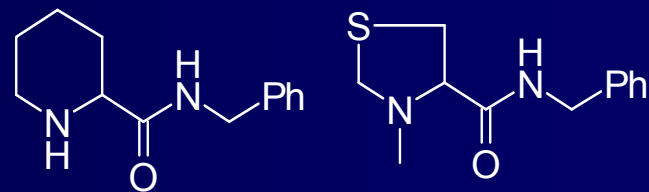
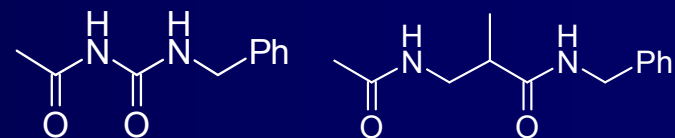
Frequent MI Descriptors map onto (some known) structural alerts

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

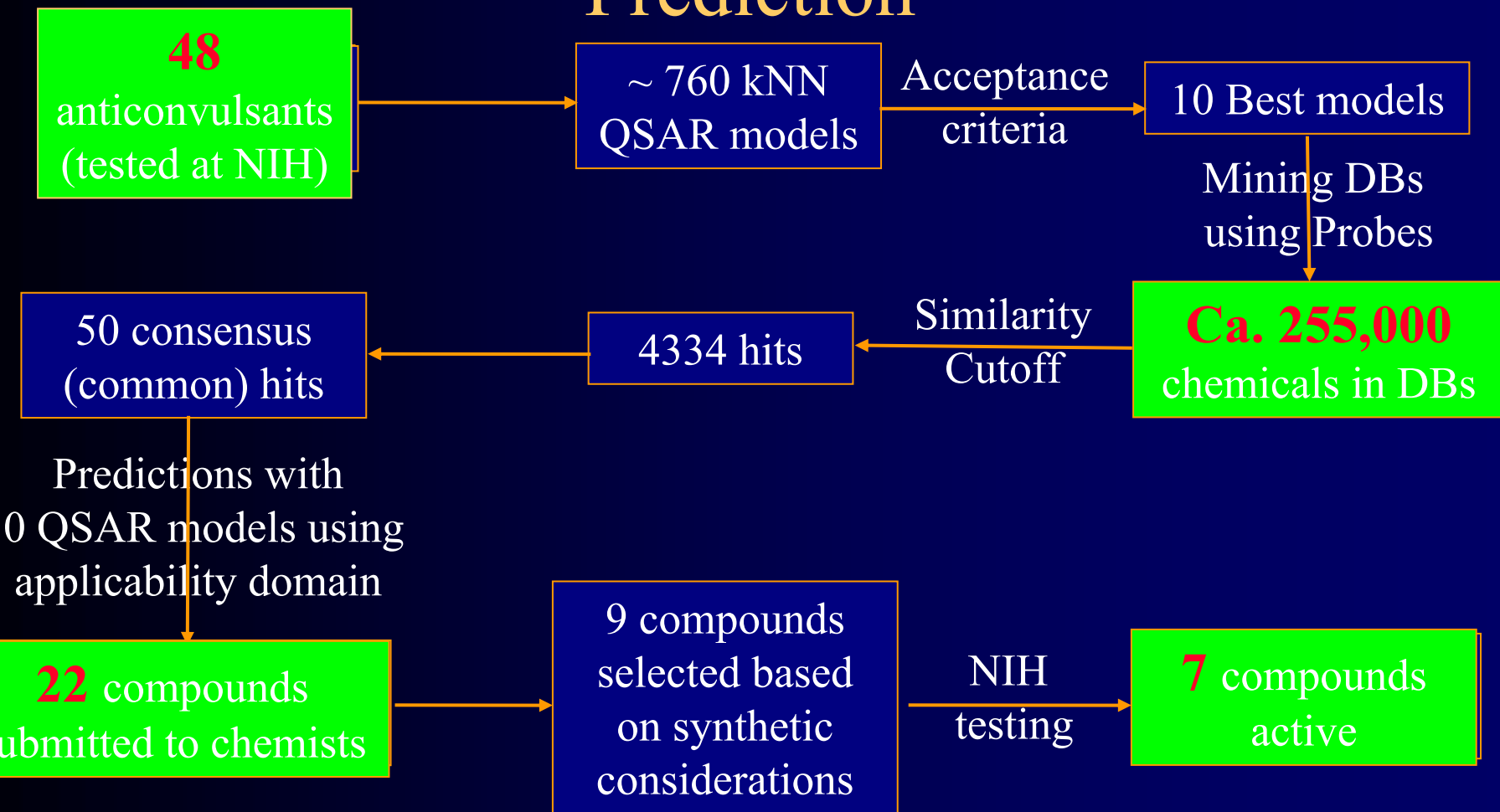


Example 3. Rational Discovery of Novel Anticonvulsant Agents based on QSAR Modeling and Virtual Screening

- 48 Functionalized amino acid (FAA) anticonvulsant agents
- Experimental activity value:
 - mice ED₅₀ (mg/kg)
 - log(mM/kg) value: 1.30-3.06
- Five sub-structural classes of FAA

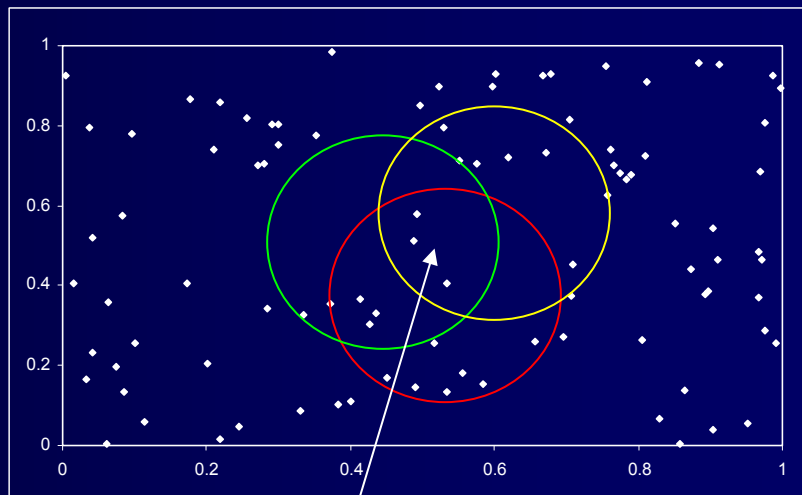


Mining Combined NCI/Maybridge Database with Multiple kNN QSAR models: Consensus Prediction*



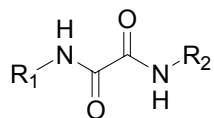
Consensus Database Mining

- Each model captures partial information about SAR
- Multiple models with high predictive power afford consensus database mining

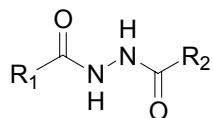


Common hits

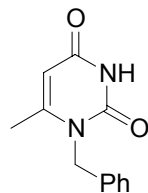
Novel structural classes provided by computational hits



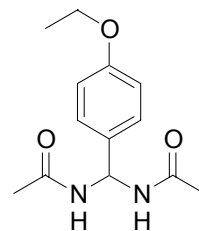
ED₅₀ = 11-51



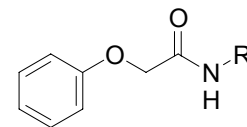
ED₅₀ = 13-50



ED₅₀ = 17



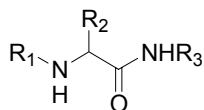
ED₅₀ = 16



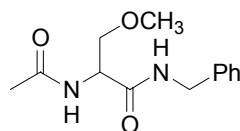
ED₅₀ = 53-70

Compounds are predicted by at least 20 validated QSAR models. ED₅₀ data are in mg/kg.

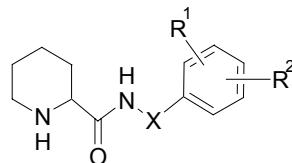
Structural classes of training set compounds



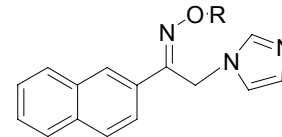
1



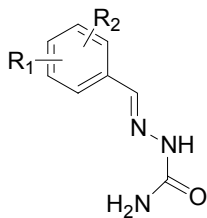
2



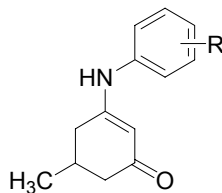
3



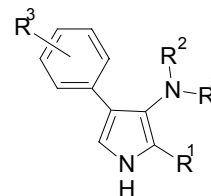
4



5



6



7

ALL-QSAR: A Novel Automated Lazy Learning QSAR Approach

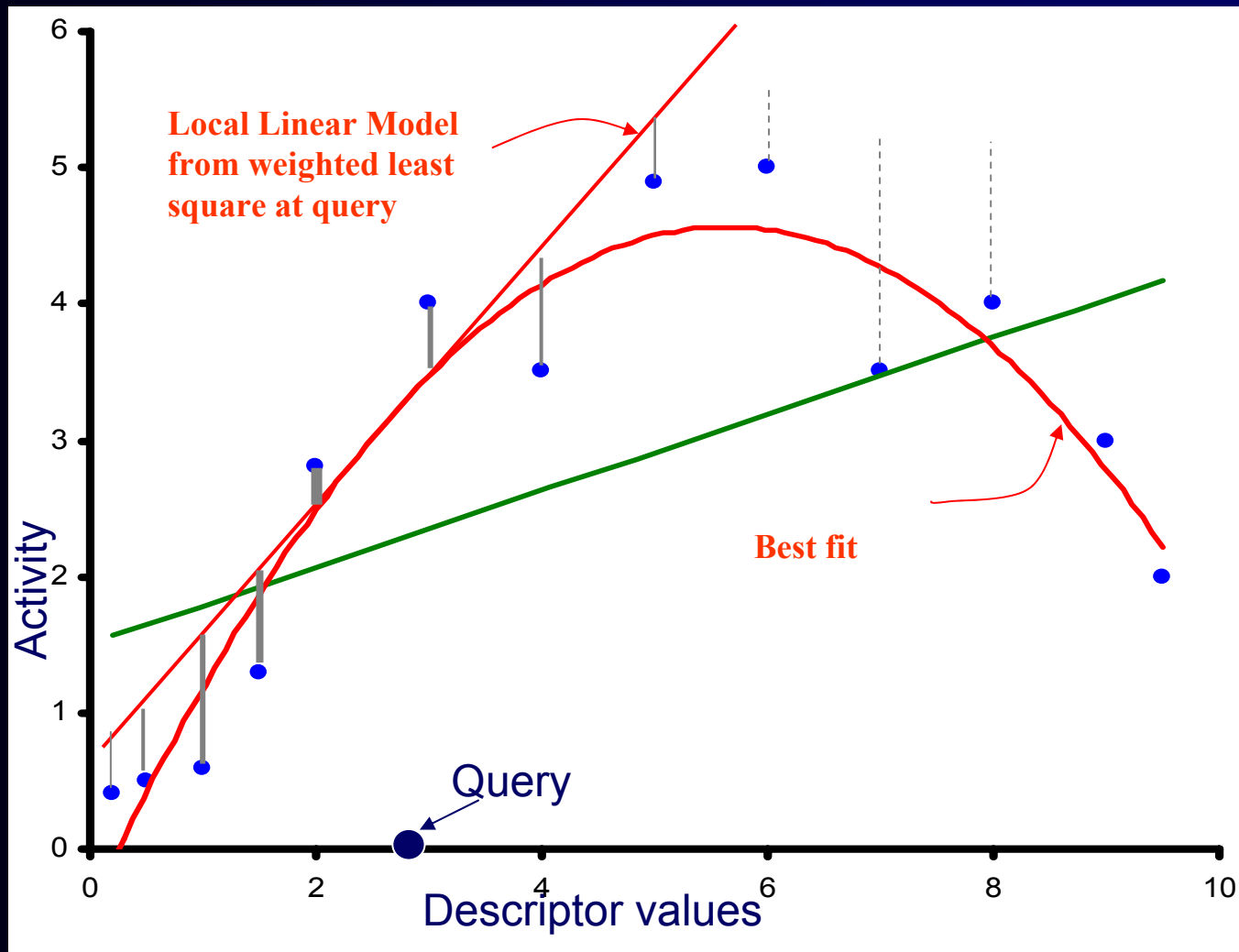
- Lazy learning methods process training data only when a query needs to be answered, so they are query (test set) oriented learning methods. They are also called memory-based or instance-based learning methods.
- We use local smoothing (weighting) linear regression algorithm

Atkeson, C. G.; Moore, A. W.; Schaal, S. Locally weighted learning. *Artif. Intell. Rev.* **1997**, *11*, 11-73.

Wettschereck, D.; Aha, D. W.; Mohri, T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artif. Intell. Rev.* 1997, *11*, 273-314.

Atkeson, C. G. Memory-based approaches to approximating continuous functions. 503-521. 1992. Casdagli and Eubank. Proceedings of a Workshop on Nonlinear Modeling and Forecasting. 9-17-1990.

Locally Weighted Regression



Tuning for Best Fitting

Cost Function

$$C(\mathbf{q}) = \sum_i \left[(f(\mathbf{x}_i, \boldsymbol{\beta}) - y_i)^2 K\left(\frac{d(x_i, q)}{h}\right) \right]$$

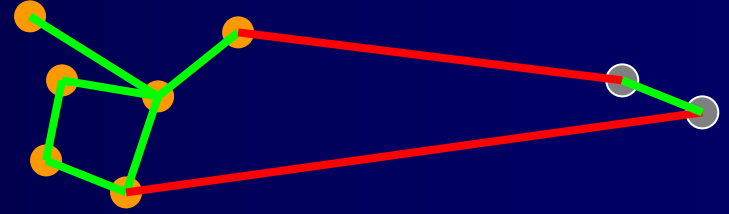
Fitting parameters:

1. Kernel width: h
2. Distance function: $d()$
3. Weighting function: $K()$

Applicability Domain

$$APD = D + Z\sigma$$

1. Used to define the predictable points
2. D is the average of Euclidean distances for those points which have Euclidean distance less than the average of the whole training set
3. Z is an empirical cutoff value
4. σ is the standard deviation of these Euclidean distances



1. Calculate the average of all pairwise distances.
2. Calculate the average (D) of all pairwise distances below the overall average.
3. Calculate the standard deviation (σ) of all pairwise distances below the overall average

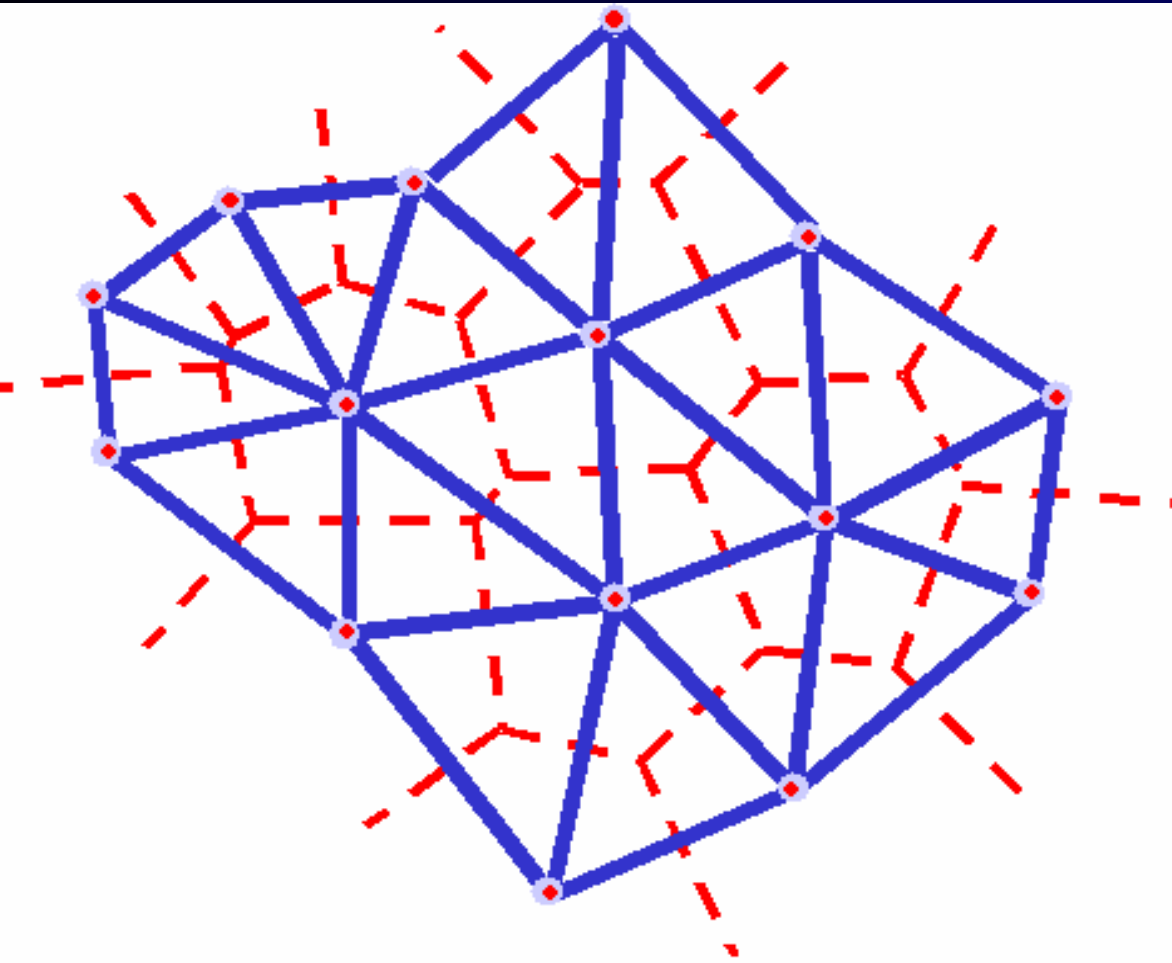
Comparison w/ Other QSAR Approaches

Methods	Training Set Size	Test Set Size	R ² for Test Sets	Consensus R ² for External (10 Best Models)
48 Anti-convulsant Compounds				
<i>k</i> NN	39	9	0.72	-
kNN	38	10	0.67	-
SA-PLS	40	8	<0.67	-
ALL-QSAR	39	9	0.90	-
ALL-QSAR	34	14	0.76	-
48 D1 Antagonist Compounds				
kNN	40	8	0.76	-
SVM	35	13	0.80	-
SA-PLS	40	8	0.63	-
CoMFA	40	8	0.45	-
ALL-QSAR	36	12	0.81	-
Cronin's 250 Phenol Compounds				
Cronin's Methods	200	50	0.66 ~ 0.82	-
<i>k</i> NN	207	43	0.79	-
ALL-QSAR	100-160	40-100	0.71 ~ 0.90	0.86

Example 4: Chemometric Approaches to Virtual Screening

- Majority are structure based
 - Rely on coordinates of active sites
 - Use complex scoring functions (energy based or statistical)
 - Most docking algorithms employ rigid binding pockets; flexible docking is very slow (by orders of magnitude)
 - **Very computationally intensive!!!**
- **Goal:** Develop (perhaps) approximate but **VERY** efficient screening/scoring algorithms using chemometric (chemical descriptors, similarity searching) approaches

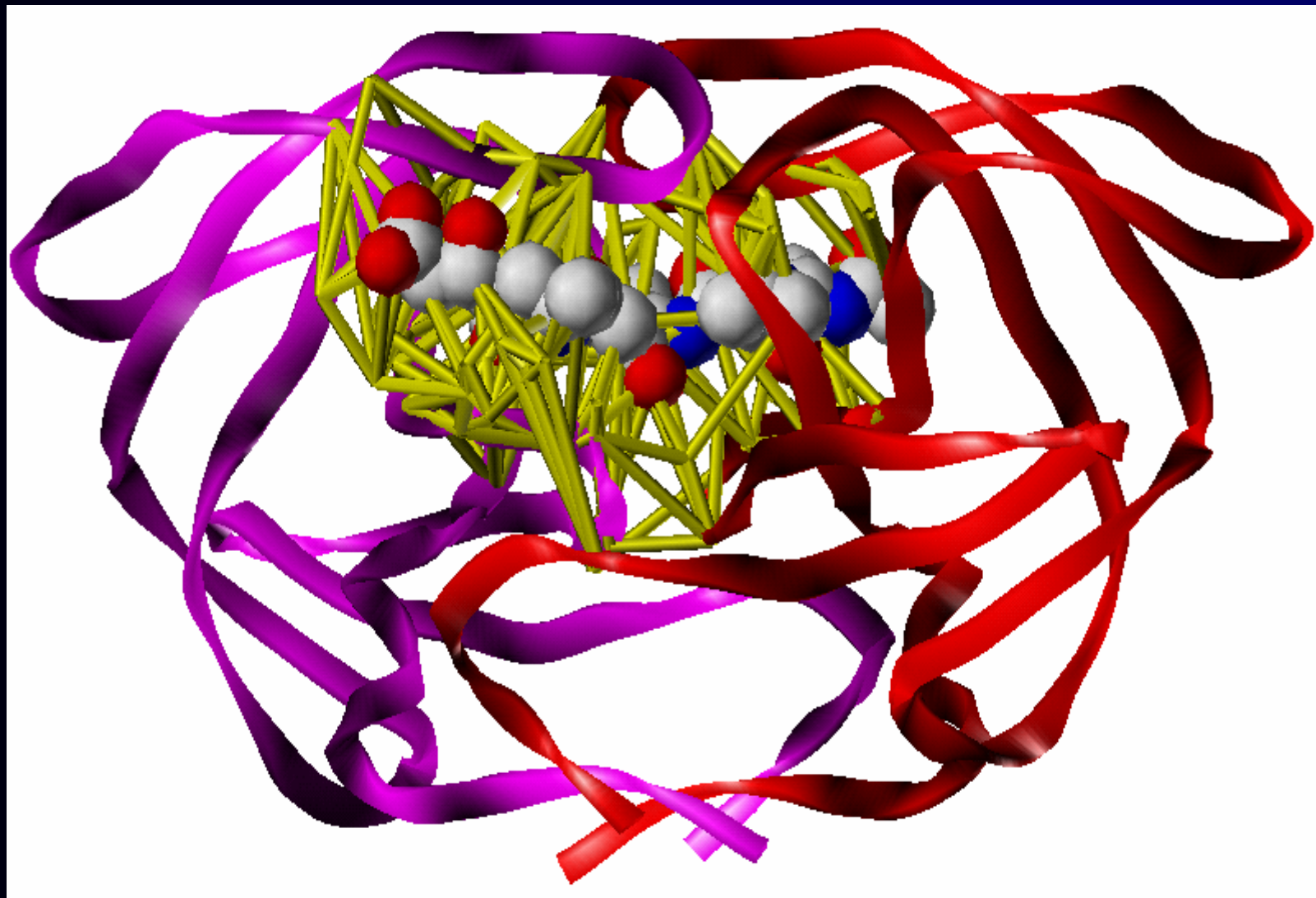
Voronoi/Delaunay Tessellation in 2D and 3D



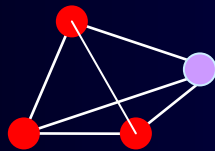
Delaunay Triangulation
(in blue) and Voronoi
Tessellation (in red)

Nearest neighbors are
unambiguously defined in
sets of three (in 2D) or
four (in 3D) vertices.

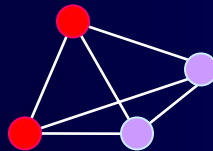
Full Atom-Based Delaunay Tessellation of Protein-Ligand Interface (5HVP)



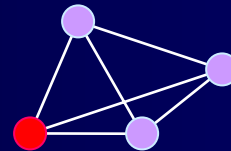
Three Types of Tetrahedra at Protein-Ligand Interfaces



RLLL



RRLL



RRRL

RRRL: Formed by 3 receptor atoms and 1 ligand atoms

RRLL: Formed by 2 receptor atoms and 2 ligand atoms

RLLL: Formed by 1 receptor atoms and 3 ligand atoms

Each of the above tetrahedral types is further discriminated by atom types of the vertices

Atom Type Definition Based on EN Values

Some Statistics: (from 250, 000 and 30, 000 database respectively)
O, C, N, S: >5000 && >1500
F (EN=4.0) and P (EN = 2.19): <1500 && <500

Ligand Atom Types

O EN = 3.4

N EN = 3.0

C EN = 2.5

S EN = 2.4

X P and Halogens, EN = 2.0 ~ 2.4, 4.0

M Metal and all other unexpected atom types, EN = 0.6 ~ 1.6

Receptor Atom Types

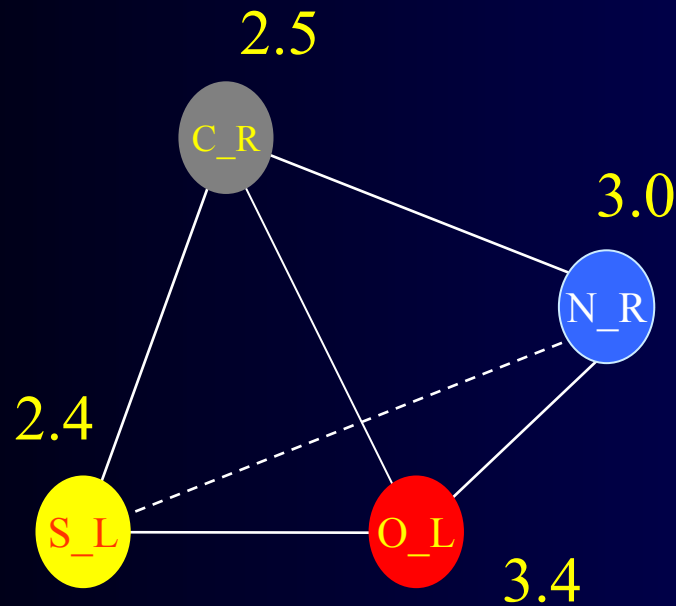
O EN = 3.4

N EN = 3.0

C EN = 2.5

S EN = 2.4

Descriptor Calculation



$$EN_m = \sum_{i=1}^n \sum_j^4 EN_{ij}$$

m : m -th tetrahedral composition type

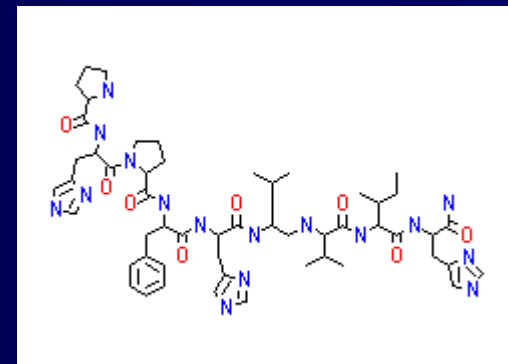
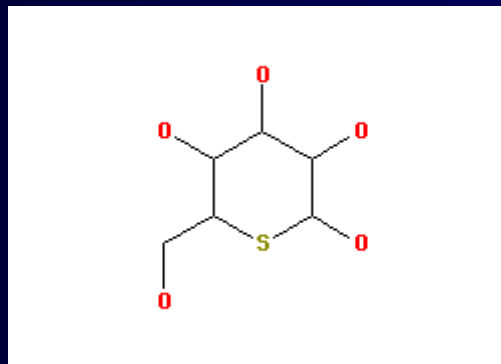
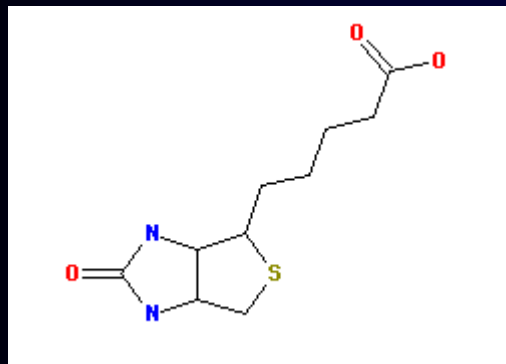
j : Vertex of a tetrahedron

n : Number of m -th composition type

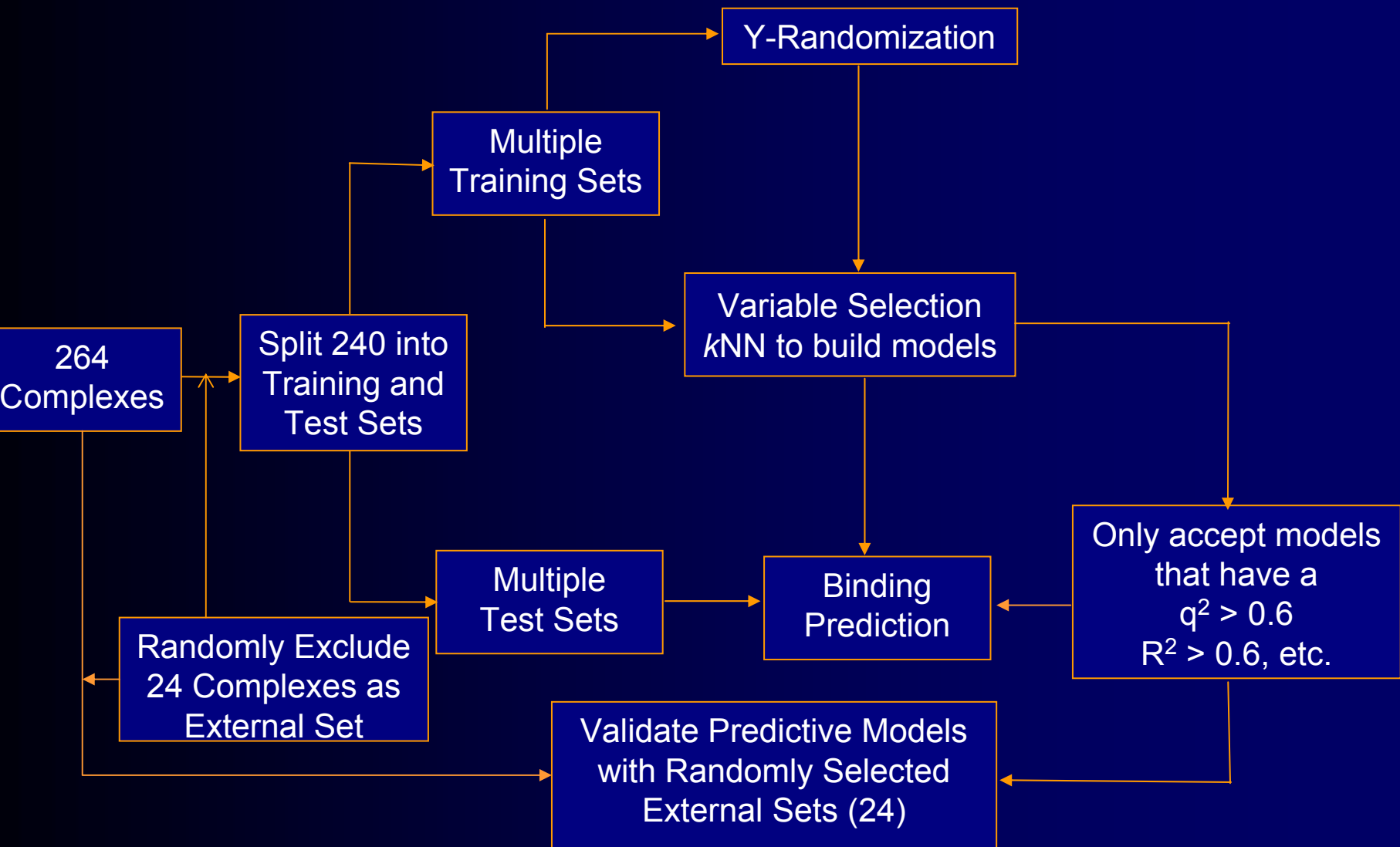
Thus there are 100 descriptors for each protein-ligand complex

Dataset with High Structural Diversity

- 264 protein-ligand complexes
- Belong to more than 80 functional families (some families have only one member)
- Even within one family, the ligand structures are different
- Wide range of binding affinity (PK_d): 1.48 (2XLI) ~ 13.4 (1STP)



Data Modeling Workflow



Comparison with Other Scoring Functions

Methods	Training Set Size	Test Set Size	R ²	R ² for External
BLEEP	351	90	0.53	N/A
PMF	697	77	0.61	N/A
SMoG96	120	46	0.42	N/A
SMoG2001	725	111	0.436	N/A
DT2001	319	67	0.71	N/A
DT2002	319	107	0.54	N/A
ENTESS1	200	40	0.83	0.77
ENTESS2	175	65	0.73	0.85
ENTESS3	122	118	0.63	0.81

Example 3. Life is still difficult. Structure-toxicity relationships for aliphatic electrophiles*

COMPOUNDS: 250

DESCRIPTORS: 160

TRAINING SET: 107

TEST SET: 43

SECOND TEST SET: 100

Zcutoff=1.0

DESCRIPTORS SELECTED:

6 TO 46 STEP 2

MODELS FOR
DESCRIPTORS 10

MODELS 21X10=210

ACCEPTABLE

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

$(R^2 - R_0^2) / R^2 < 0.1$ $0.85 < K < 1.15$

$(R^2 - R_0'^2) / R^2 < 0.1$ $0.85 < K' < 1.15$

$|R_0'^2 - R_0^2| < 0.3$

NETZEVA DATASET: DIVISION INTO THREE SUBSETS

OUT OF 180 MODELS ($q^2 > 0.7$)	SECOND TEST SET WAS ACCEPTABLY PREDICTED BY ... MODELS
OUT OF ... "ACCEPTABLE" MODELS	SECOND TEST SET WAS ACCEPTABLY PREDICTED BY ... MODELS

ENRICHMENT

180	84	180	10	180	10	180	17	180	30
80	49	22	7	7	4	10	3	29	16

1.31

5.73

10.29

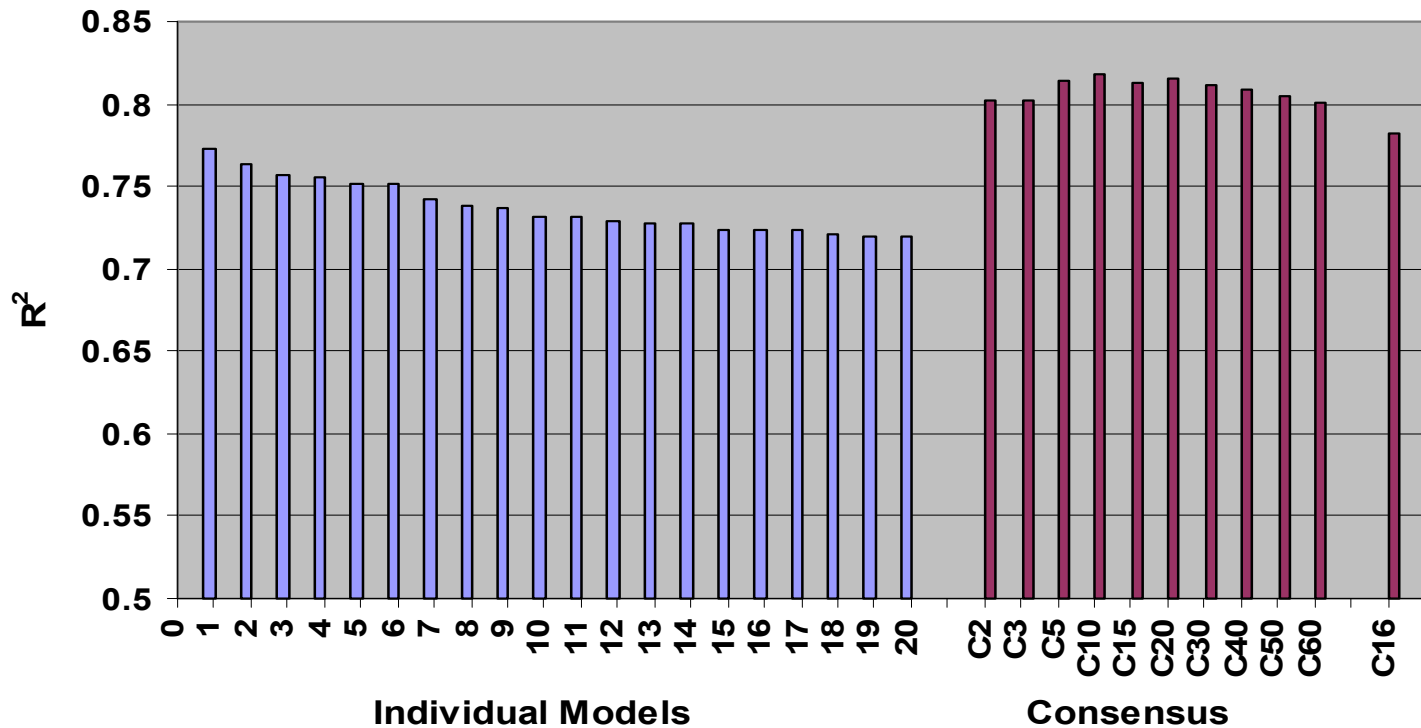
3.15

3.31

CONCLUSION: SET OF MODELS CONTAINS HIGHER FRACTION OF ACCURATELY PREDICTING TEST SETS THAN ALL MODELS WITH HIGH q^2 .

THE SITUATION IS STILL NOT AS SIMPLE... AS IT MAY APPEAR

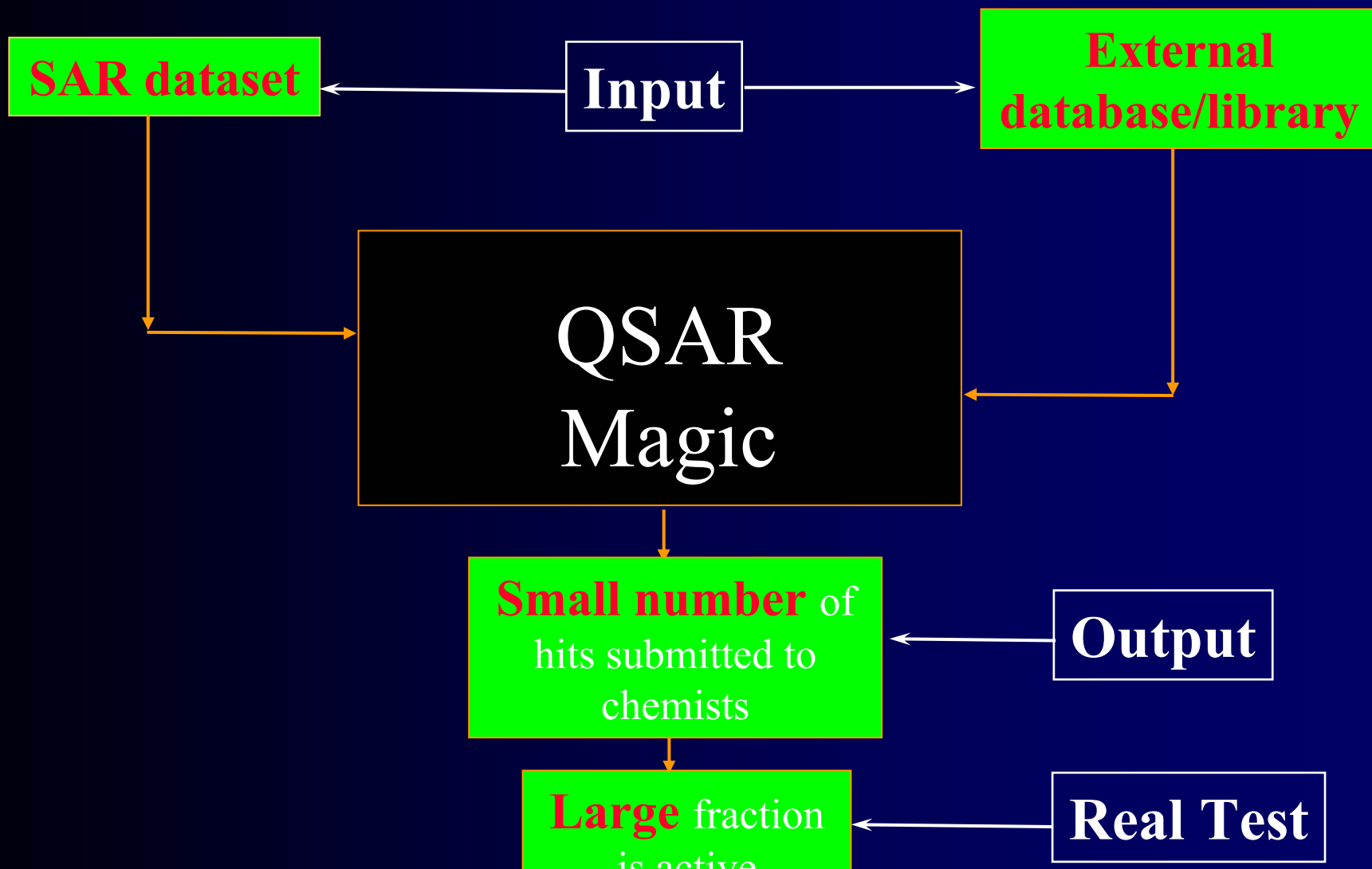
CONSENSUS PREDICTION VS PREDICTION BY INDIVIDUAL MODELS



Future Directions

- Integration of descriptors and data-analytical methods (Combi-QSAR)
- Depository for target specific validated QSAR models
- Analysis of common descriptors and most successful combinations (of descriptors and data modeling techniques) to increase hit rate
- Independent model evaluation in competitive fashion: CAAP (similar to CASP).

Emphasis on Successful Predictions, not models



Summary: The Importance of Being Earnest

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

Oscar Wilde

- Extensive model validation is a must!
- Predictive QSAR workflow affords statistically significant models which can be used directly for consensus database mining
- Automation and web-based accessibility
- QSAR methodologies are applicable to a wide range of problems in chemo-, bio-, and medinformatics.

ACKNOWLEDGMENTS

UNC ASSOCIATES

Former:

Stephen CAMMER

Sung Jin CHO

Weifan ZHENG

Min SHEN

Bala KRISHNAMOORTHY

– Jun FENG

– Yun-De XIAO

–Yuanyuan QIAO

–Patricia LIMA

–Assia KOVACHEVA

–M. KARTHIKEYAN

Current

Protein structure group:

John GRIER

Luke HUAN

Ruchir SHAH

Shuxing ZHANG

Shuquan ZONG

Peter ITSKOWITZ

QSAR group:

– Weifan ZHENG

– Alex GOLBRAIKH

– Raed KHASHAN

– Scott OLOFF

– Kun WANG

– Mei WANG

Chris GRUKE

• Funding

– NIH

– NSF

– NCI-BSF

– Berlex, IBM,
MCNC, GSK,
Inspire,
Millennium,
Ortho-McNeil