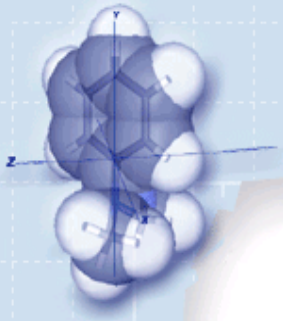
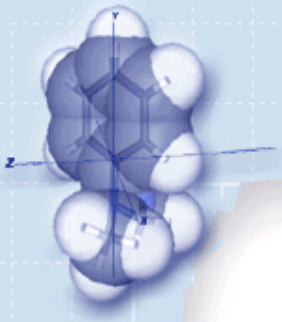


ALOGPS is a free on-line program to predict lipophilicity and aqueous solubility of chemical compounds

Igor V. Tetko & Vsevolod Yu. Tanchuk

IBPC, Ukrainian Academy of Sciences, Kyiv, Ukraine and Institute for Bioinformatics, Munich, Germany

March 15th, ACS

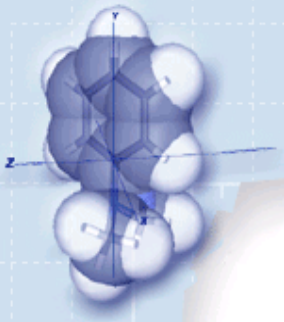


ALOGPS 2.1

•LogP: 75 input variables corresponding to electronic and topological properties of atoms (E-state indices), 12908 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.35, MAE=0.26, n=76 outliers (>1.5 log units)

•LogS: 33 input E-state indices, 1291 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.49, MAE=0.35, n=18 outliers (>1.5 log units)

- **Tetko, Tanchuk & Villa, JCICS, 2001, 41, 1407-1421.**
- **Tetko, Tanchuk, Kasheva & Villa, JCICS, 2001, 41, 1488-1493.**
- **Tetko & Tanchuk, JCICS, 2002, 42, 1136-1145.**



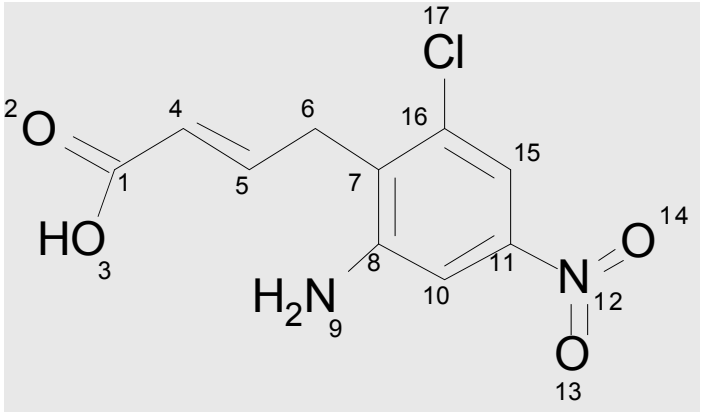
Representation of molecules

SMILES (no stereoisomers)

NH, NA -- number of hydrogen and non-hydrogen atoms

E-state indexes developed by Kier & Hall

- Basic atom-type E-state indexes
- Extended atom-type
- Bond-type

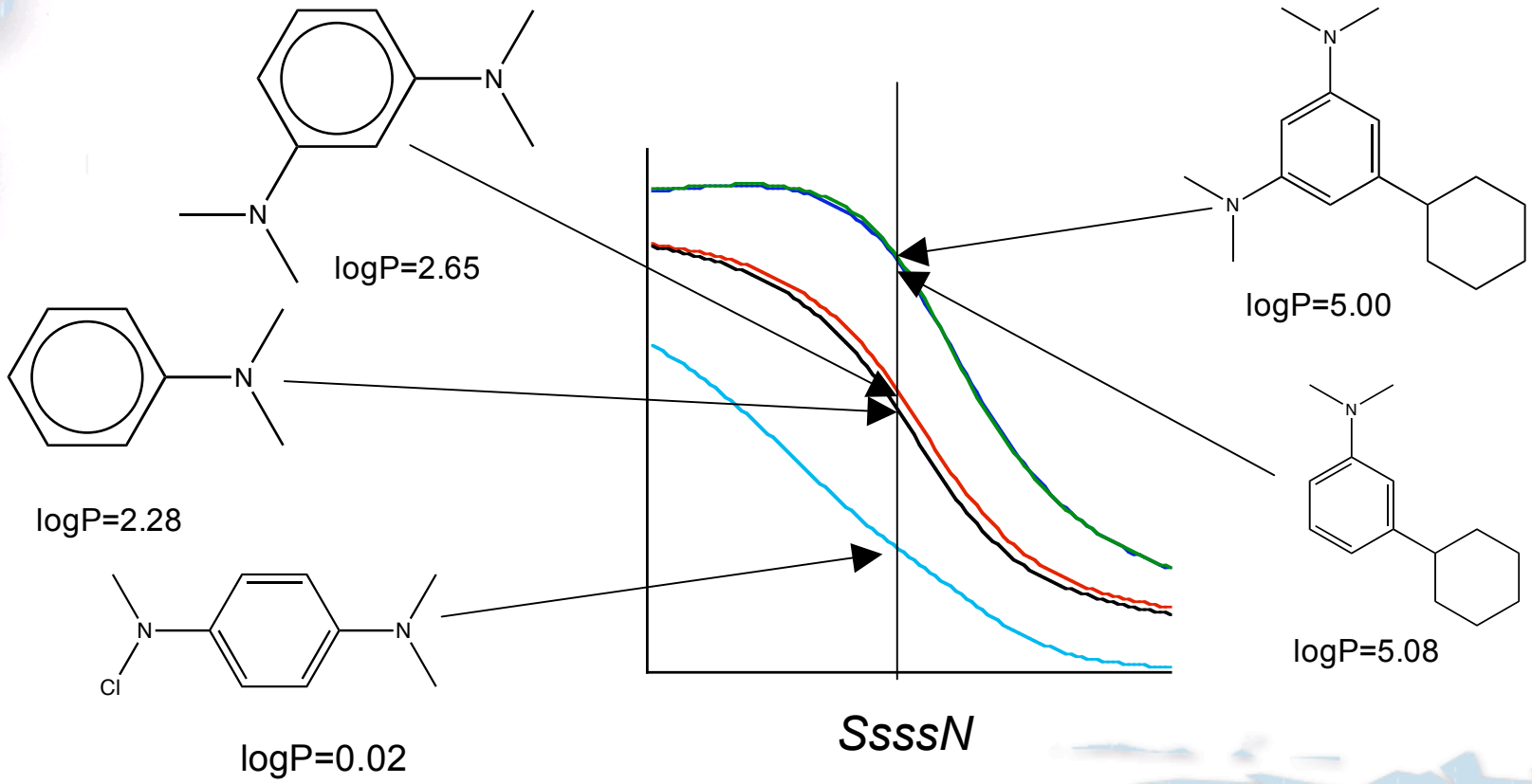


Advantages to use atom-type E-state indices:
No missed fragments!
==> non-linear dependencies indices/property

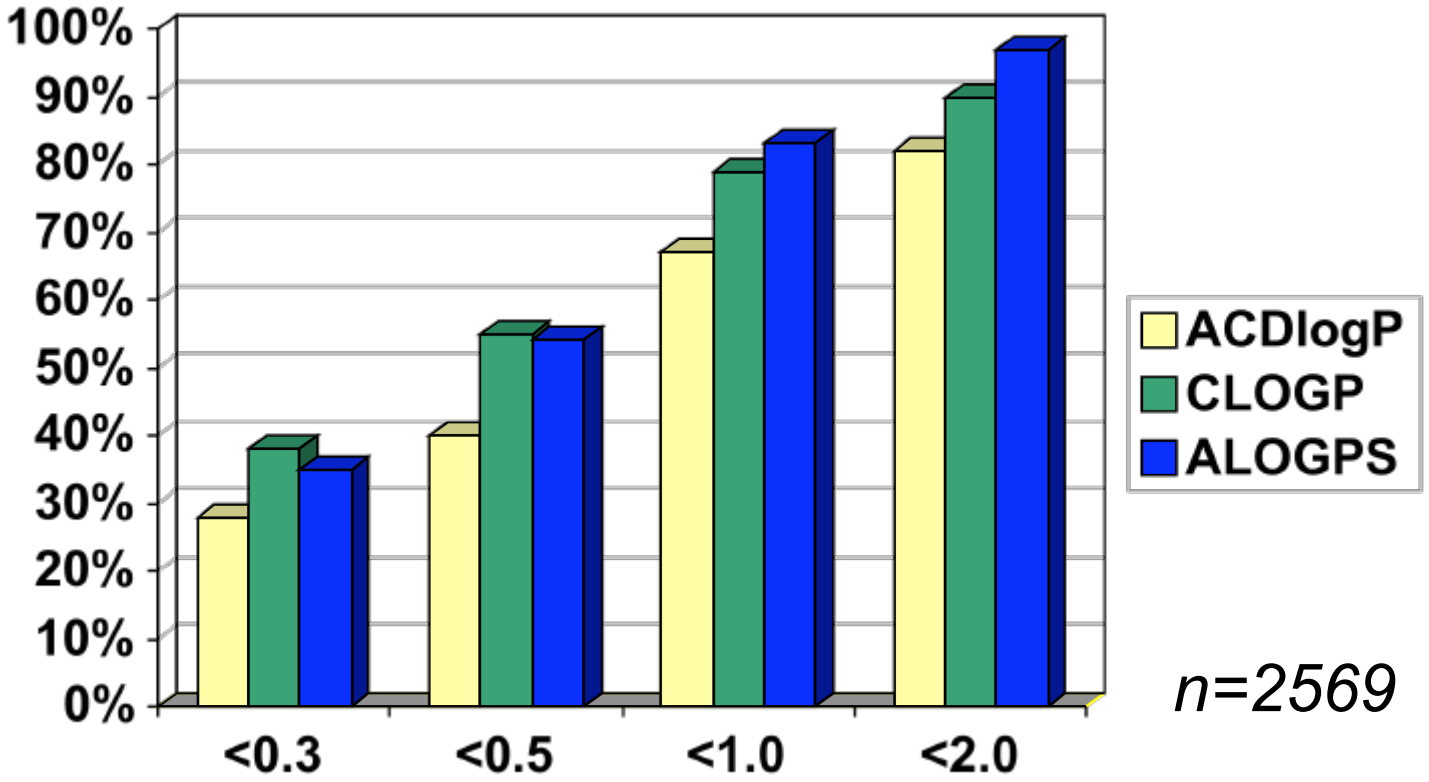
Atom type E-state: SdO, SsOH,SsCl,…
Extended: SdO(nitro), SdO(acid), …
Bond-type: e2NO2, eaC3C3aa,e1C3Cla,...

MLRA vs Neural Network

method	indices	R^2	RMSE
MLRA	75 E-state indices, including NA, NH	0.89	0.61
ASNN	the same	0.95	0.35



Prediction of AstraZeneca logP set

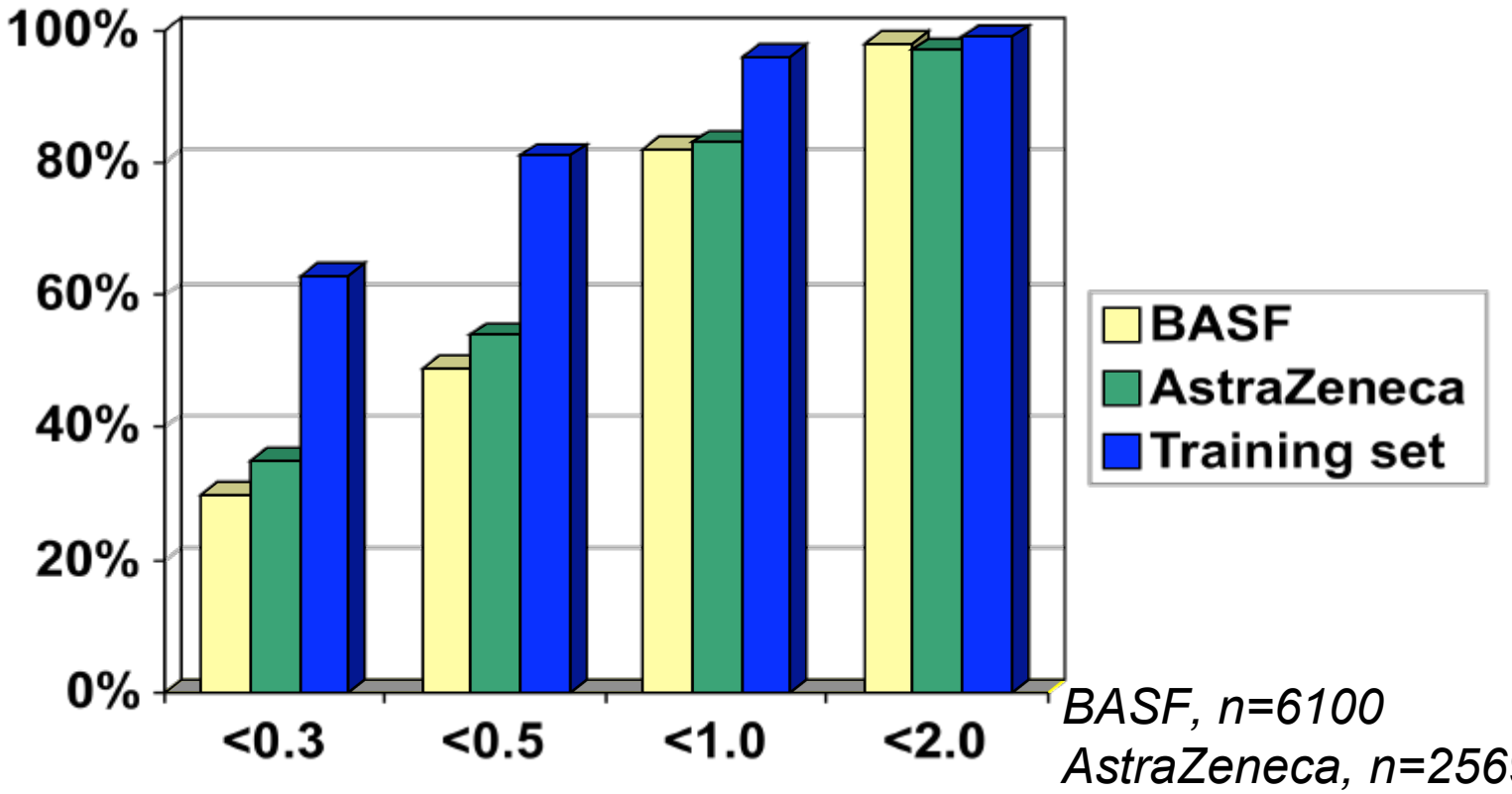


n=2569

ACDlogP (v. 7.0): *MAE = 0.86, RMSE=1.20*
CLOGP (v. 4.71): *MAE = 0.71, RMSE=1.07*
ALOGPS: *MAE = 0.60, RMSE=0.84*

Tetko & Bruneau, J. Pharm. Sci., 2004, 94, 3103-3110.

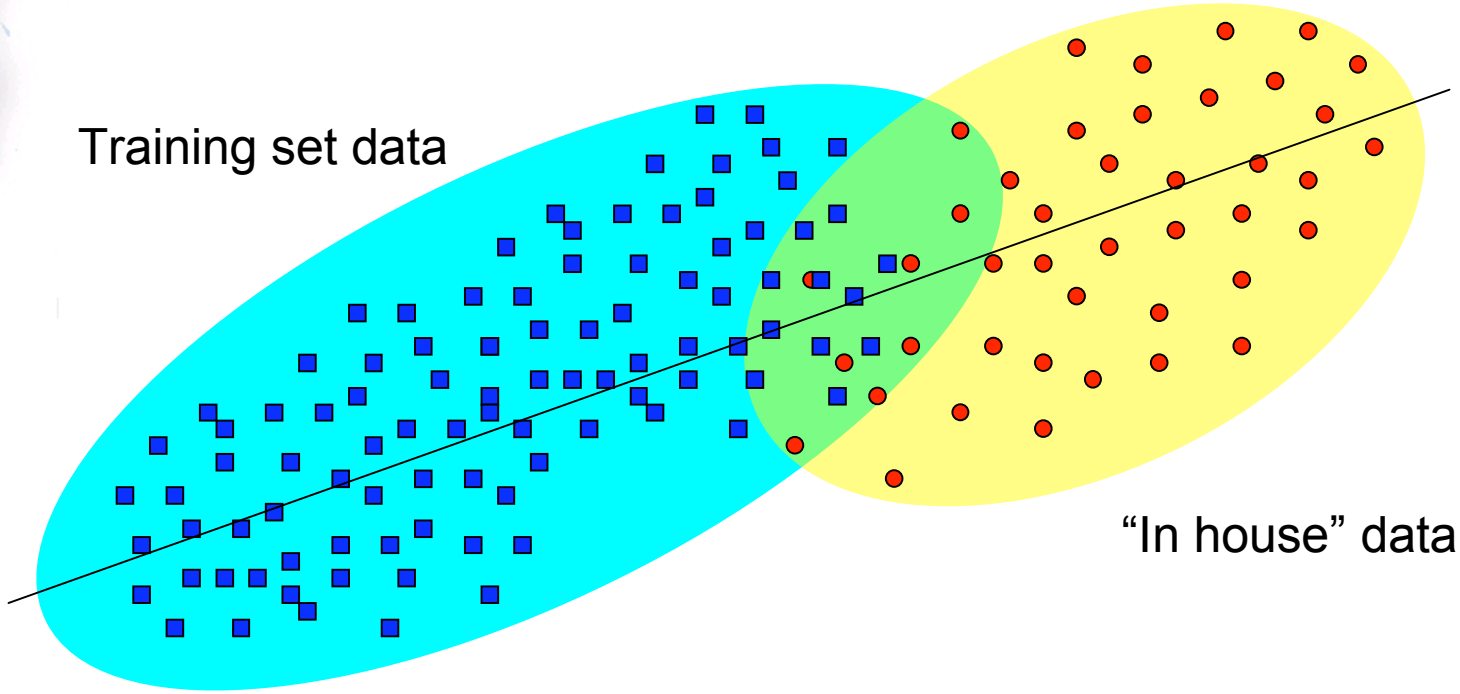
ALOGPS: Extrapolation vs Interpolation

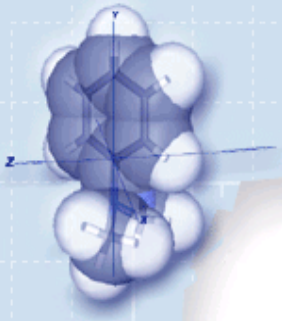


BASF, n=6100
AstraZeneca, n=2569

Tetko, JCICS, 2002, 42, 717-742.
Tetko & Bruneau, J. Pharm. Sci., 2004, 94, 3103-3110.

Prediction Space of the model does not cover the “in house” compounds



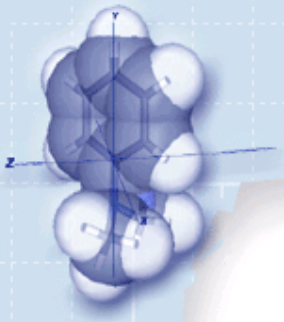


Possible strategies

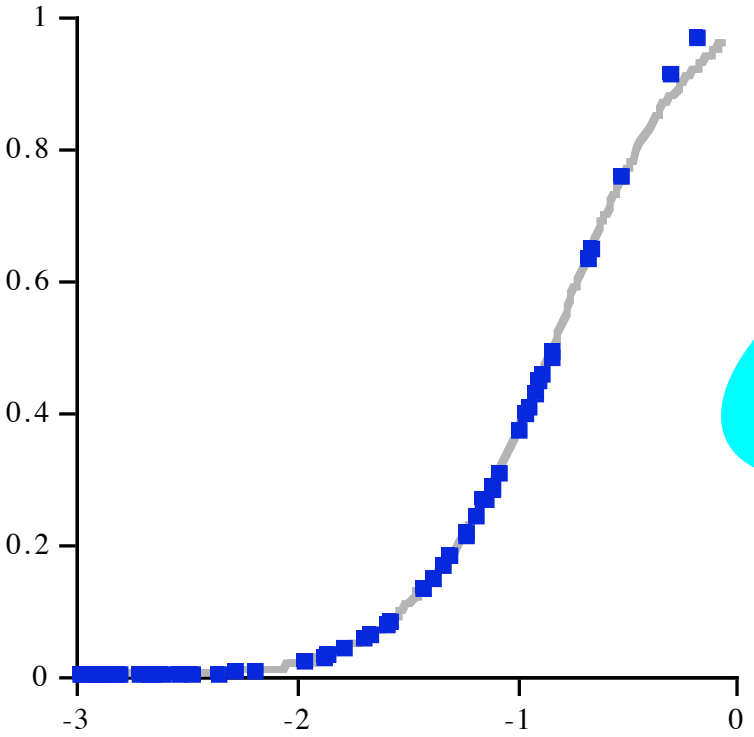
- Generate new indices and build a new model --method is used by fragment-based approaches (ACDLabs, PharmaAlgorithms)

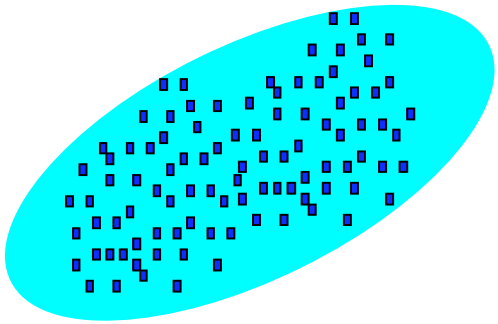
  provides an improvement but may have danger of overfitting that can lower prediction ability

- Do not generate new indices/model but to extend the model into the uncovered space and correct the model using kNN-- no danger of overfitting (Associative Neural Network)



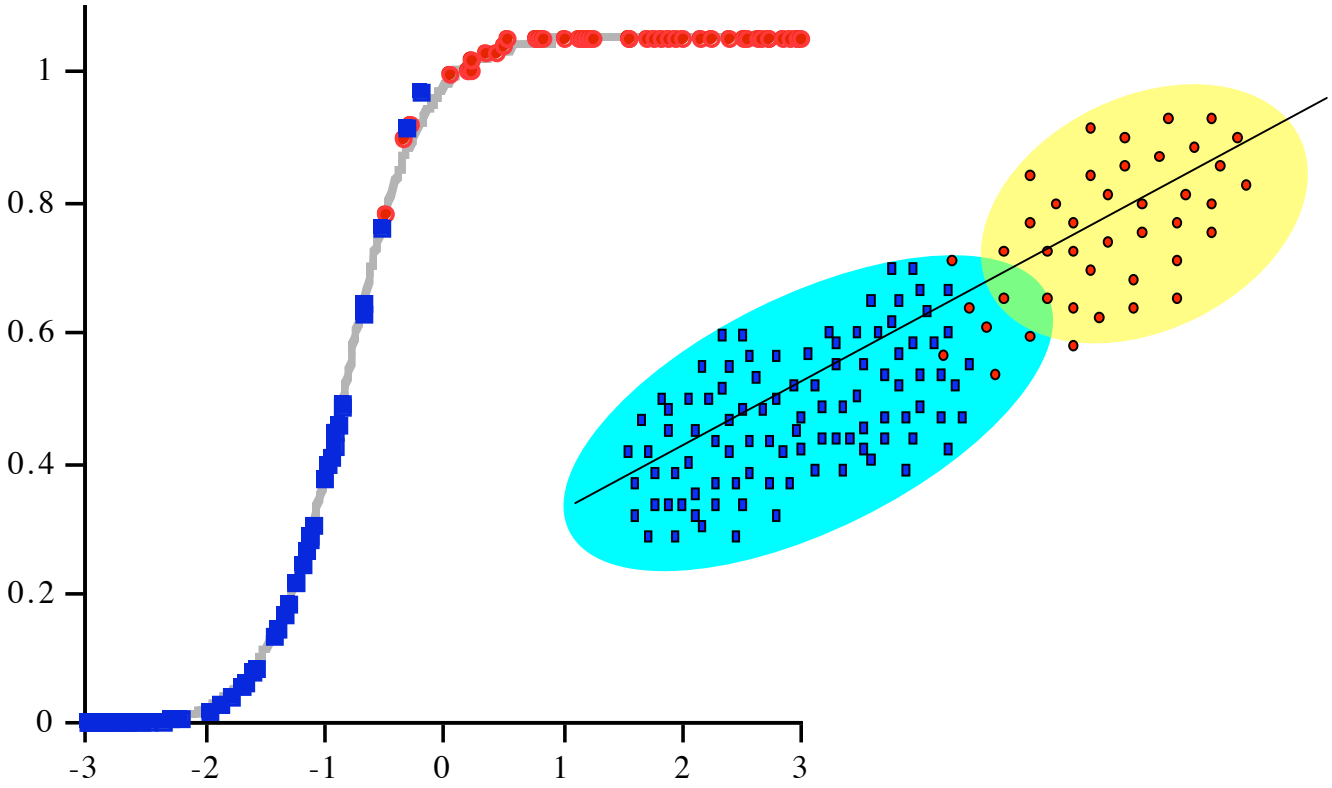
Training set model





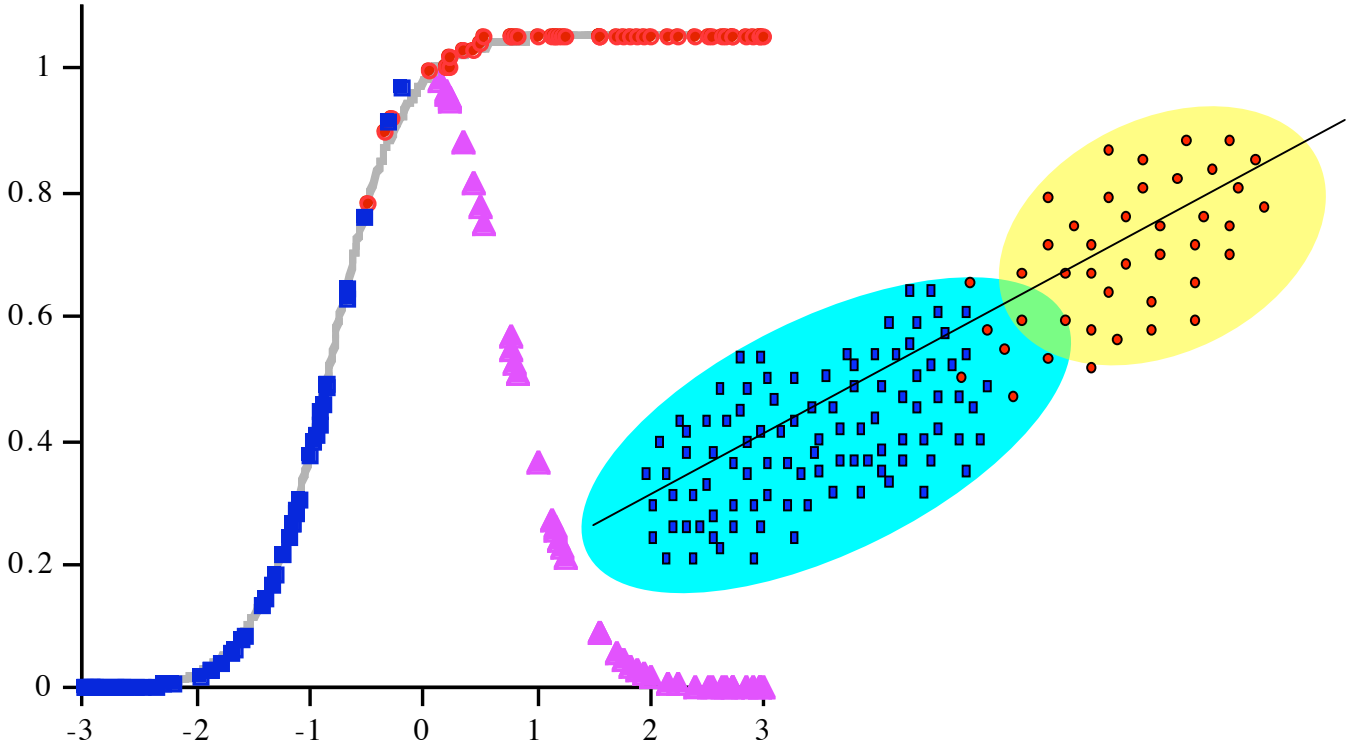
Interpolation

Prediction of the test compounds

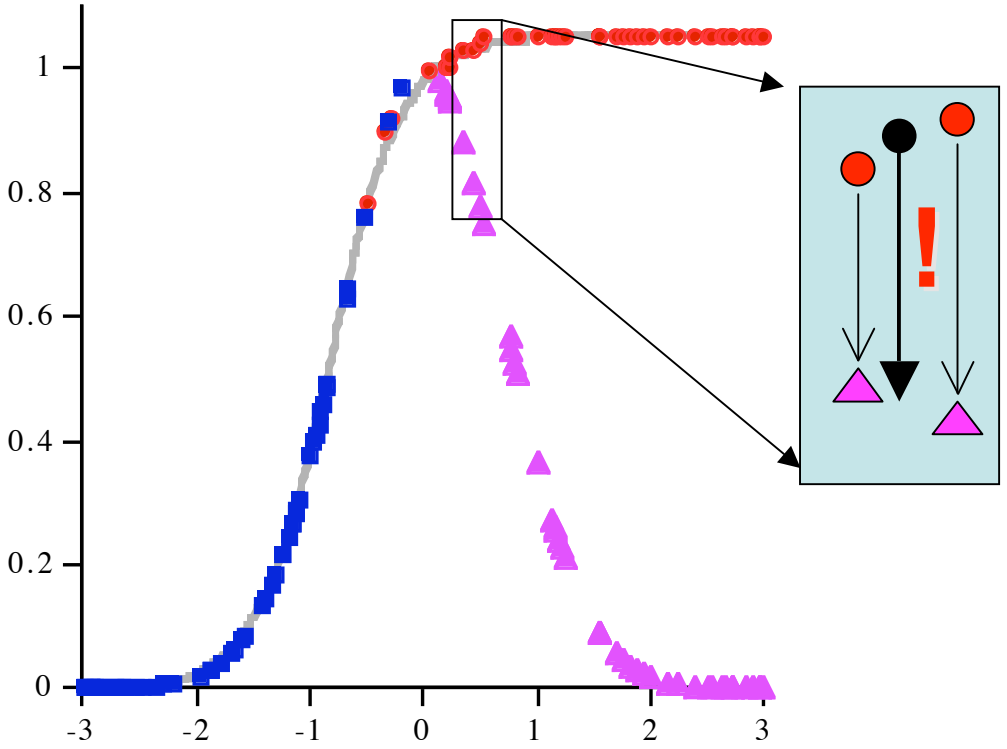


extrapolation

New data prediction

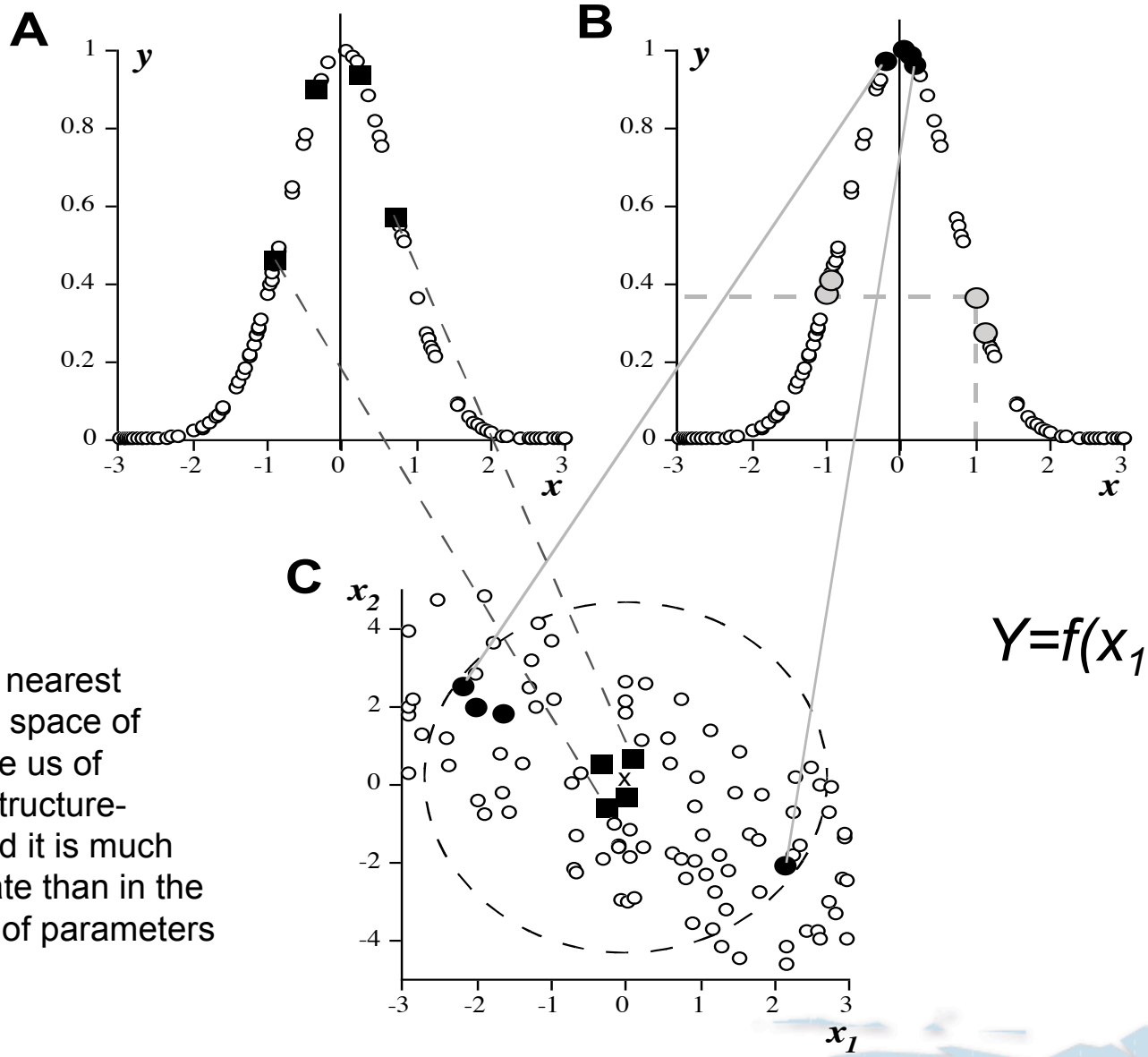


k-NN correction



- We can make an adjustment of the predicted value by identifying the nearest neighbors (NN) of the analyzed data case
- How to detect the nearest neighbors?
- Why we need for this ensemble of models?

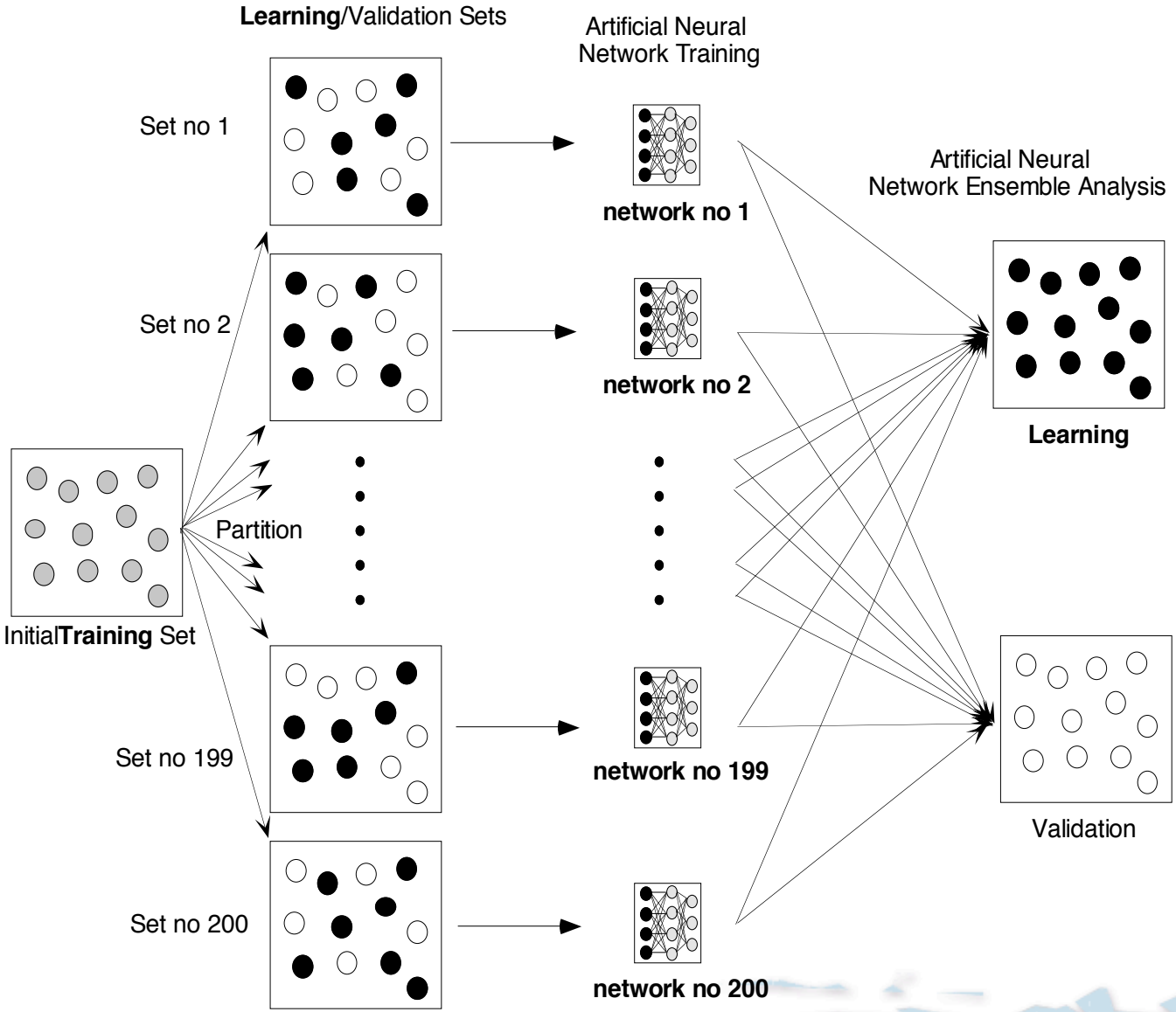
Nearest neighbors for Gauss function



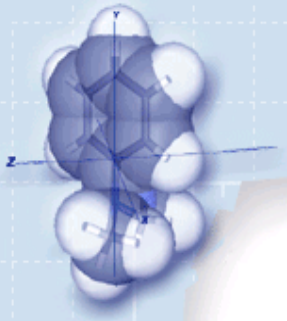
Detection of nearest neighbors in space of models make us of invariants “structure-property” and it is much more accurate than in the initial space of parameters

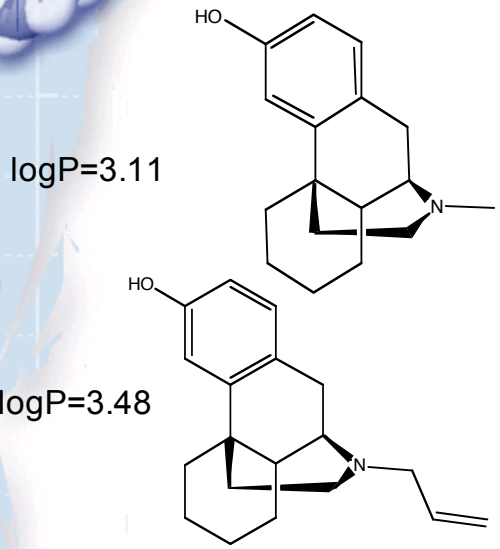
$Y=f(x_1+x_2)$

Early Stopping Over Ensemble (ESE)



ASNN: an example correction





logP=3.11

logP=3.48

$$\begin{bmatrix} 12.3 \\ 4.6 \\ \vdots \\ 13.2 \\ 10.1 \end{bmatrix} \rightarrow \begin{bmatrix} net\ 1 \\ net\ 2 \\ \vdots \\ net\ 63 \\ net\ 64 \end{bmatrix}$$

$$\begin{bmatrix} 13.7 \\ 4.8 \\ \vdots \\ 15.8 \\ 12.0 \end{bmatrix} \rightarrow \begin{bmatrix} net\ 1 \\ net\ 2 \\ \vdots \\ net\ 63 \\ net\ 64 \end{bmatrix}$$

1-kNN correction

Morphinan-3-ol, 17-methyl-

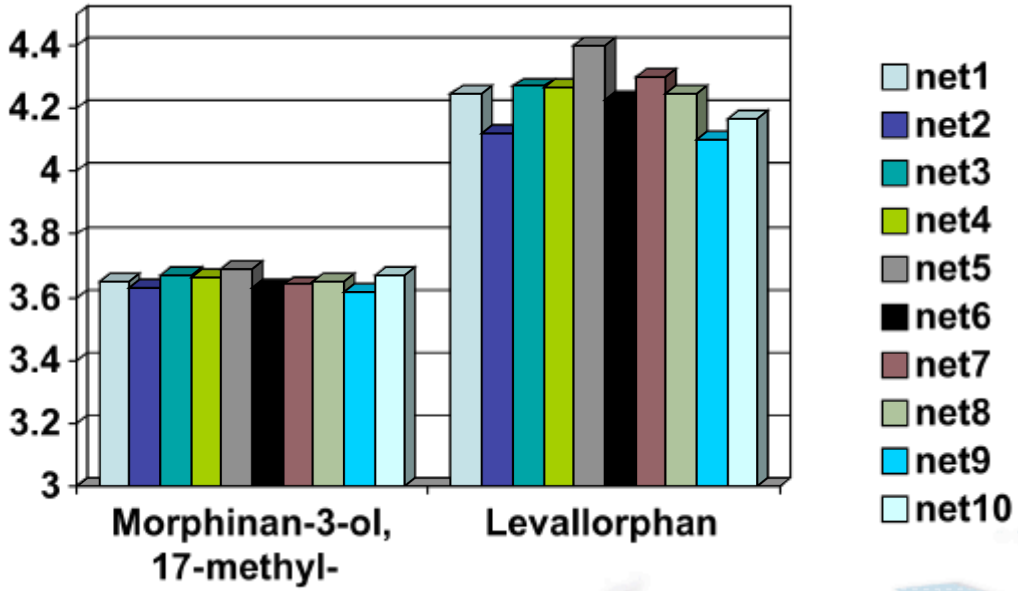
Calculated logP=3.65, δ=+0.54

--> 3.65-0.76=2.89 (δ=+0.22)

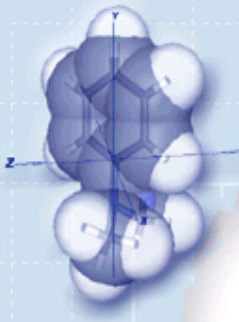
Levallorphan

Calculated logP=4.24, δ=+0.76

--> 4.24-0.54=3.70 (δ=+0.22)



-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals!



Associative Neural Network (ASNN)

A prediction of case i: $[\mathbf{x}_i] \bullet [\mathbf{ANNE}]_M = [\mathbf{z}_i] = \begin{bmatrix} z_1^i \\ \vdots \\ z_k^i \\ \vdots \\ z_M^i \end{bmatrix}$

Ensemble approach:

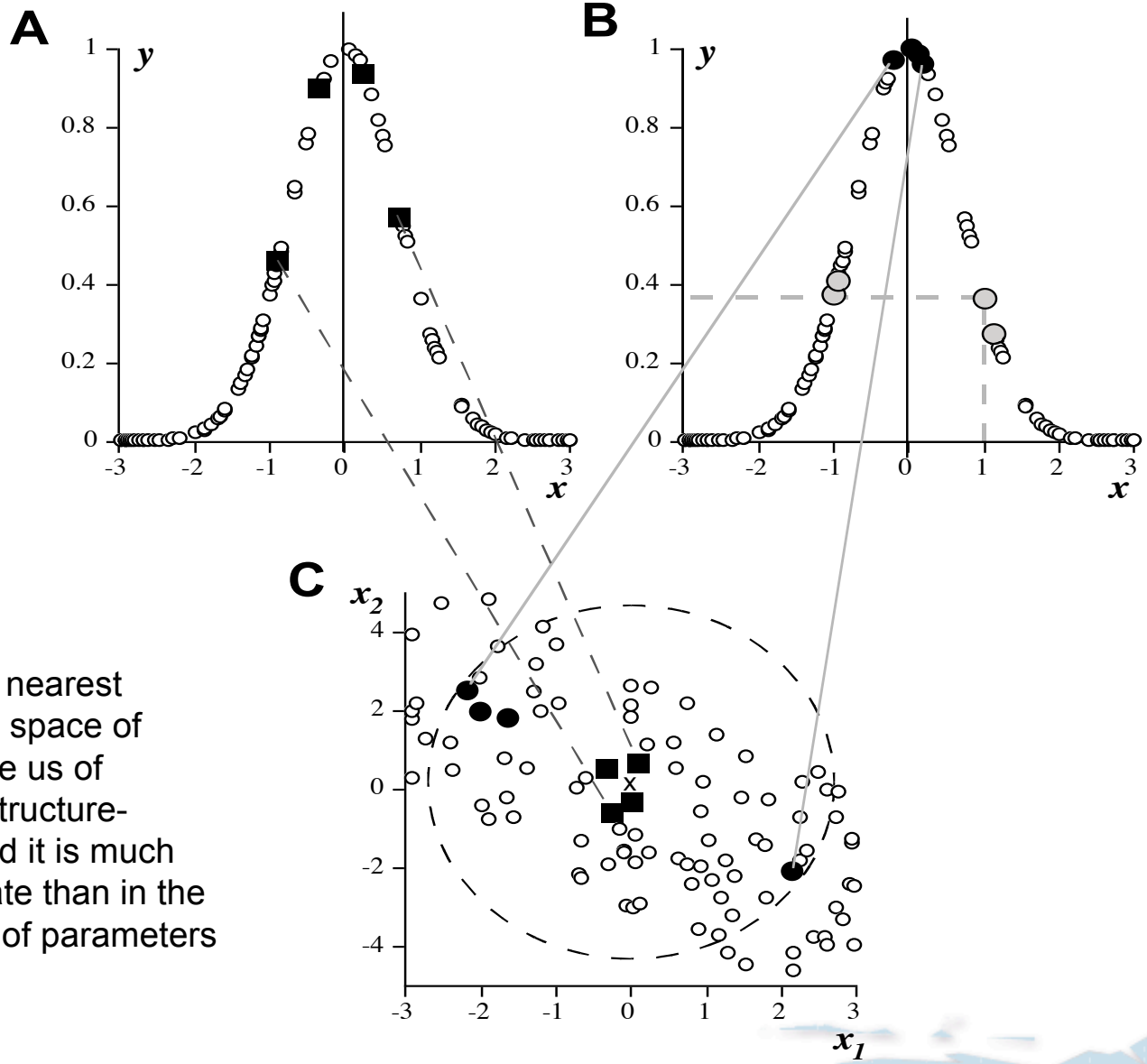
$$\bar{z}_i = \frac{1}{M} \sum_{k=1,M} z_k^i$$

Pearson's (Spearman) correlation coefficient $r_{ij}=R(z_i,z_j)>0$ ***in space of residuals***

$$\bar{z}_i' = \bar{z}_i + \frac{1}{k} \sum_{j \in N_k(\mathbf{x}_i)} \left(y_j - \bar{z}_j \right)$$ **<<= ASNN bias correction**

The correction of neural network ensemble value is performed using errors (biases) calculated for the neighbor cases of analyzed case $\mathbf{x}_i$ detected in space of neural network models

Nearest neighbors for Gauss function

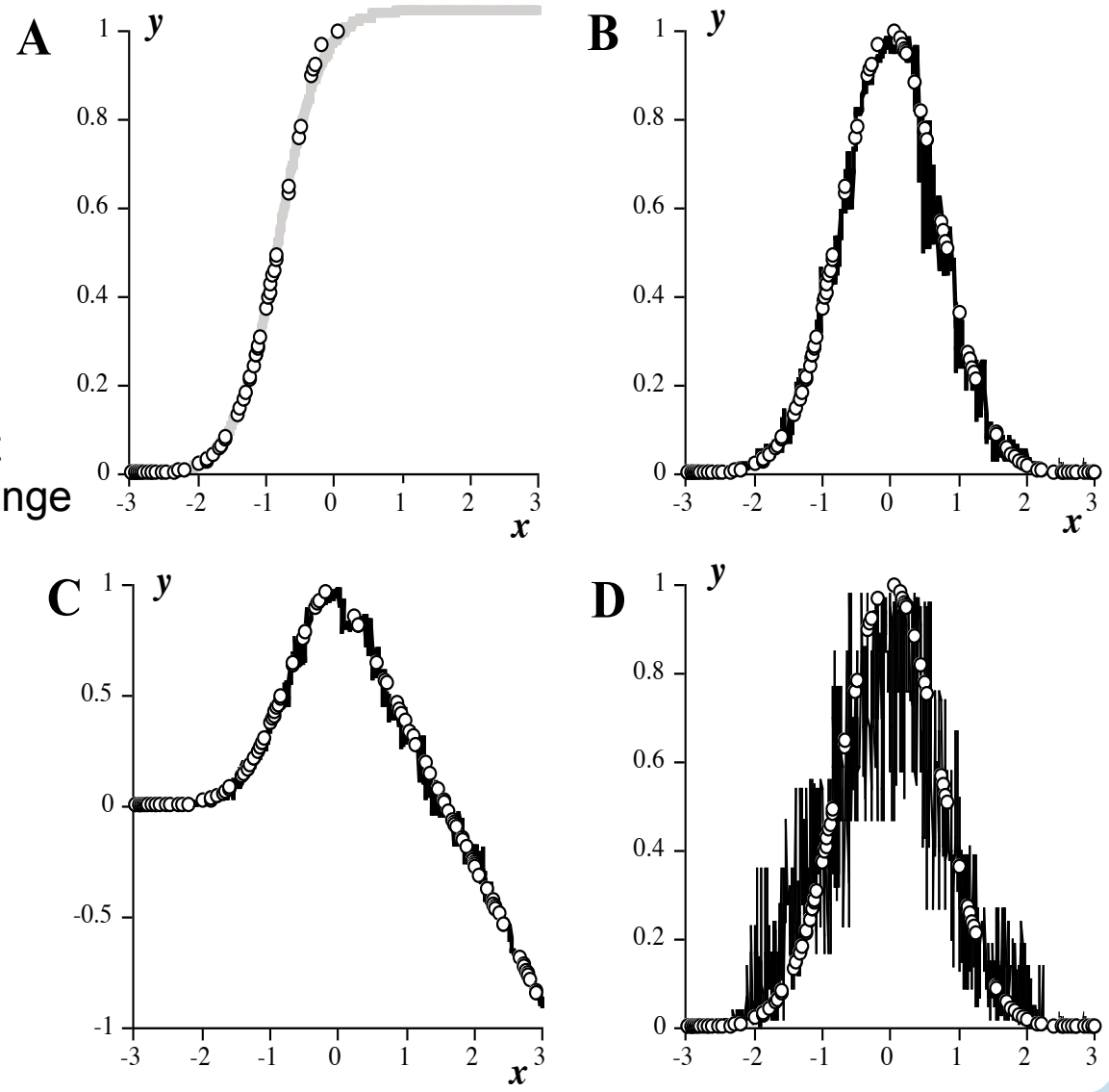


Detection of nearest neighbors in space of models make us of invariants “structure-property” and it is much more accurate than in the initial space of parameters

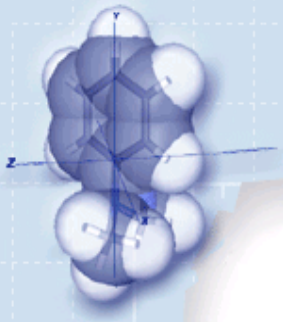
Gauss function extrapolation

Advantages: fast, no neural network retraining; correction is not limited by the range of values in the training set.

Notice: $y=f(x=x_1+x_2)$

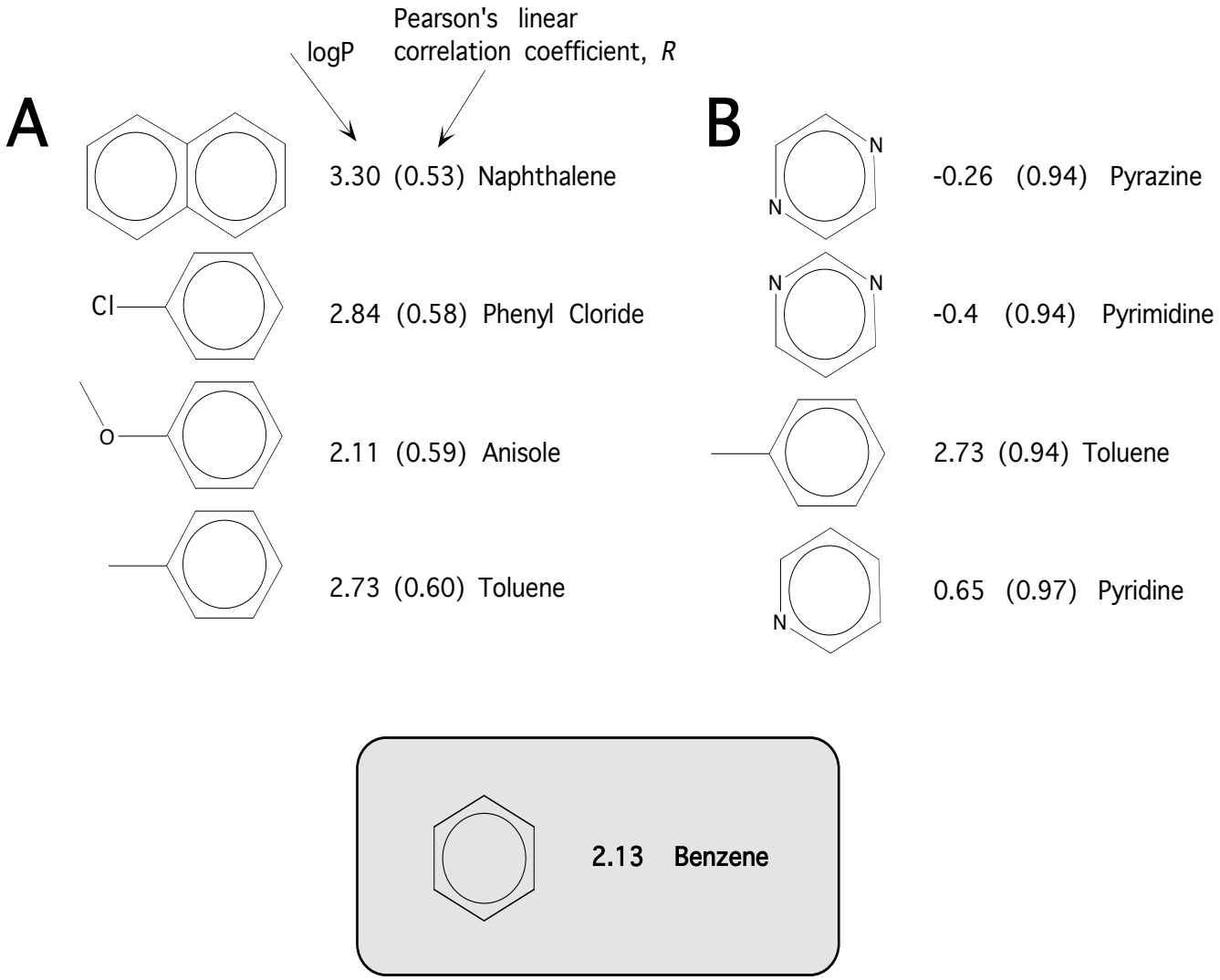


Property-based clustering



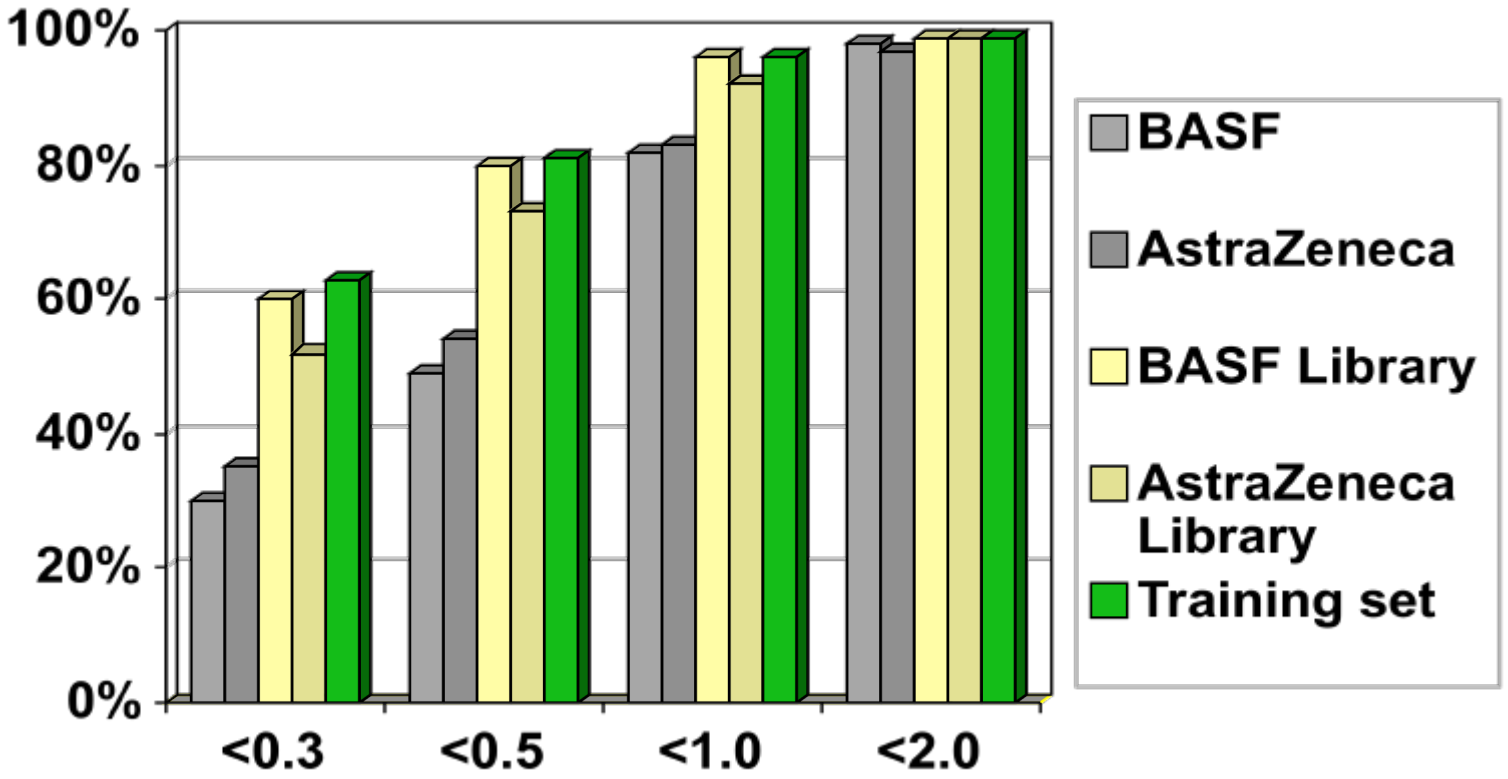
A: lipophilicity prediction

B: molecular weight prediction



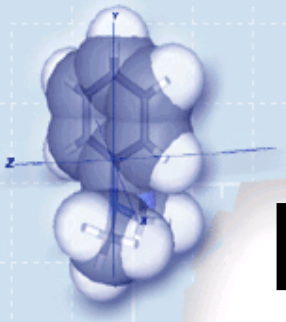
Tetko, JCICS, 2002, 42, 717-728.

ALOGPS: Extrapolation vs Interpolation

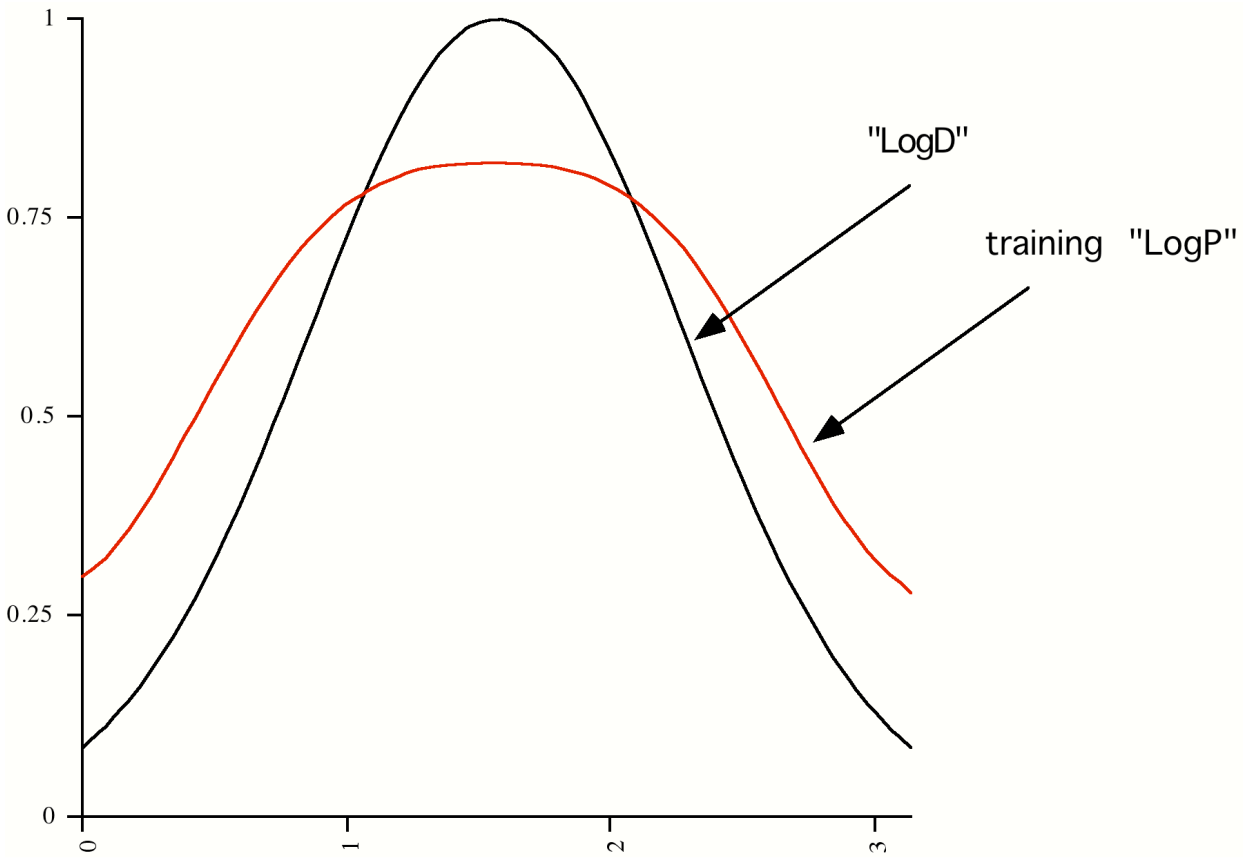


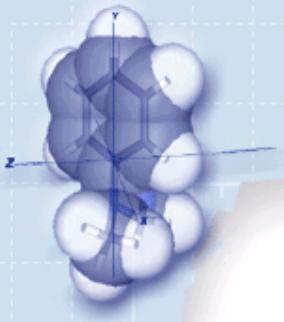
ALOGPS logP (blind) :*MAE = 1.27, RMSE=1.63*
ALOGPS logP (LIBRARY):*MAE = 0.49, RMSE=0.70*

Tetko, JCICS, 2002, 42, 717-742.
Tetko & Bruneau, J. Pharm. Sci., 2004, 94, 3103-3110.

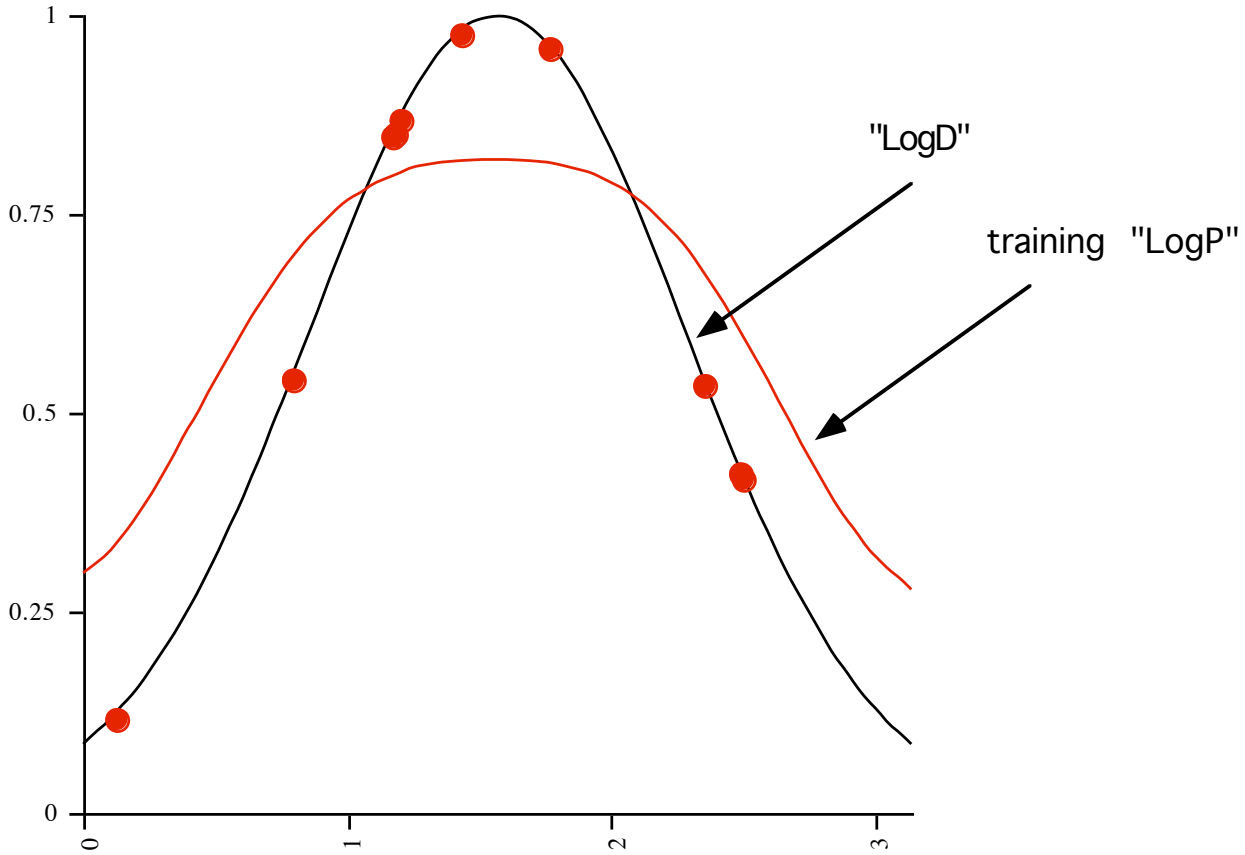


Function training and real

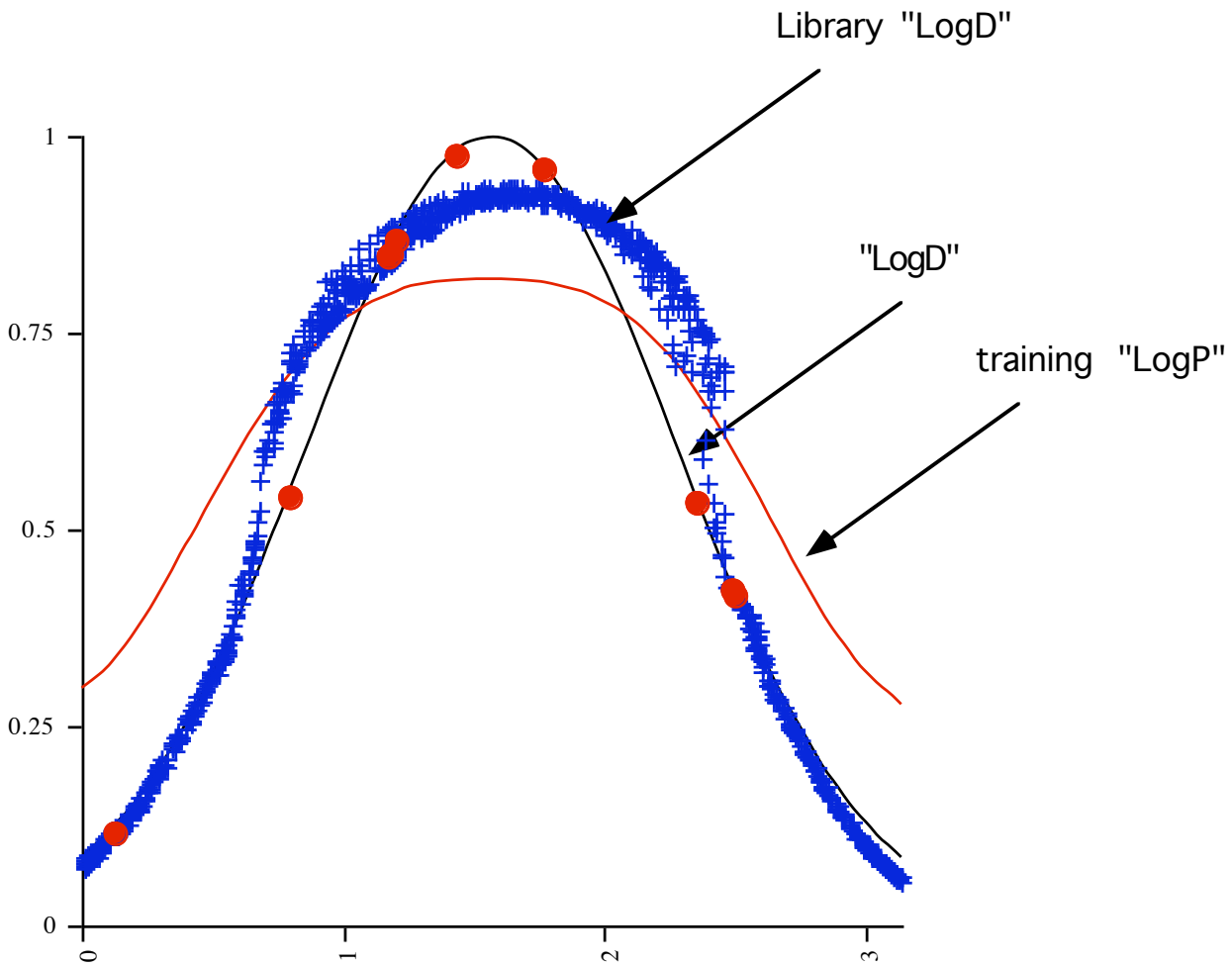




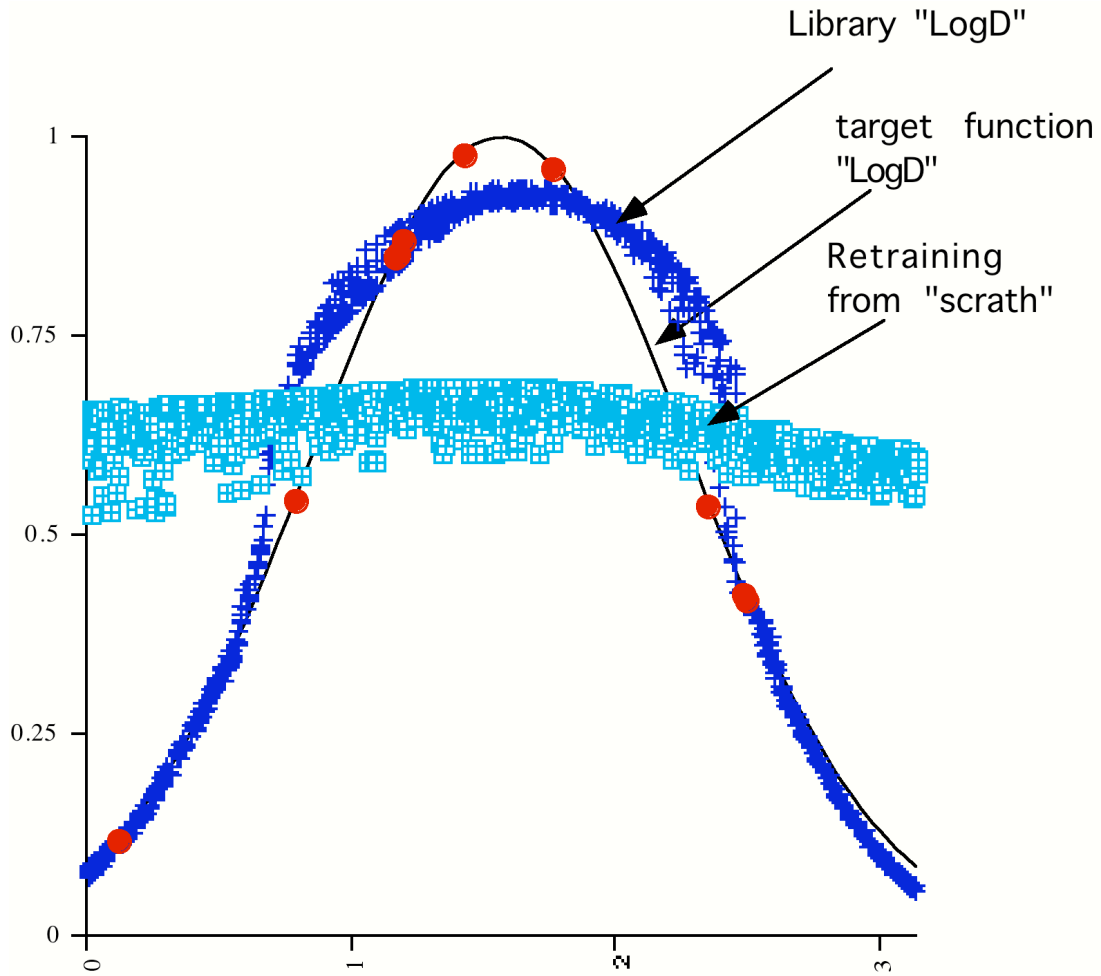
10 new points are measured



Library mode (no retraining)

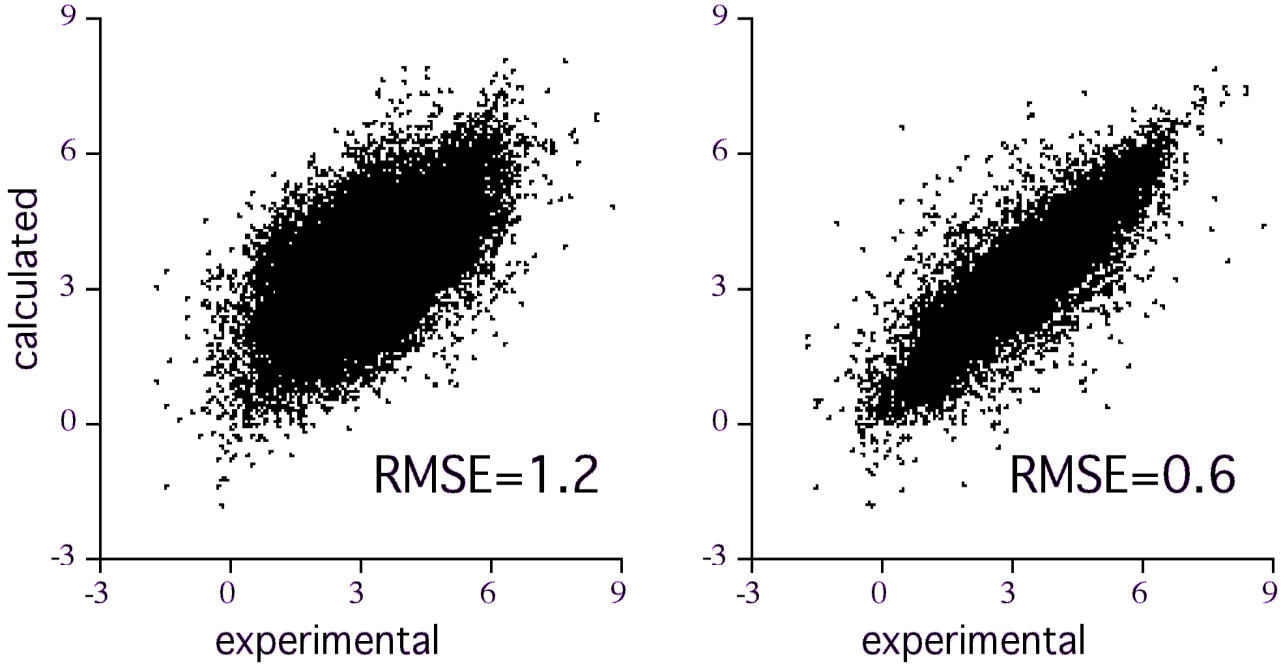


Library mode vs training using 10 cases



Analysis of Pfizer data

ALOGPS prediction for ElogD set of 17,861 compounds



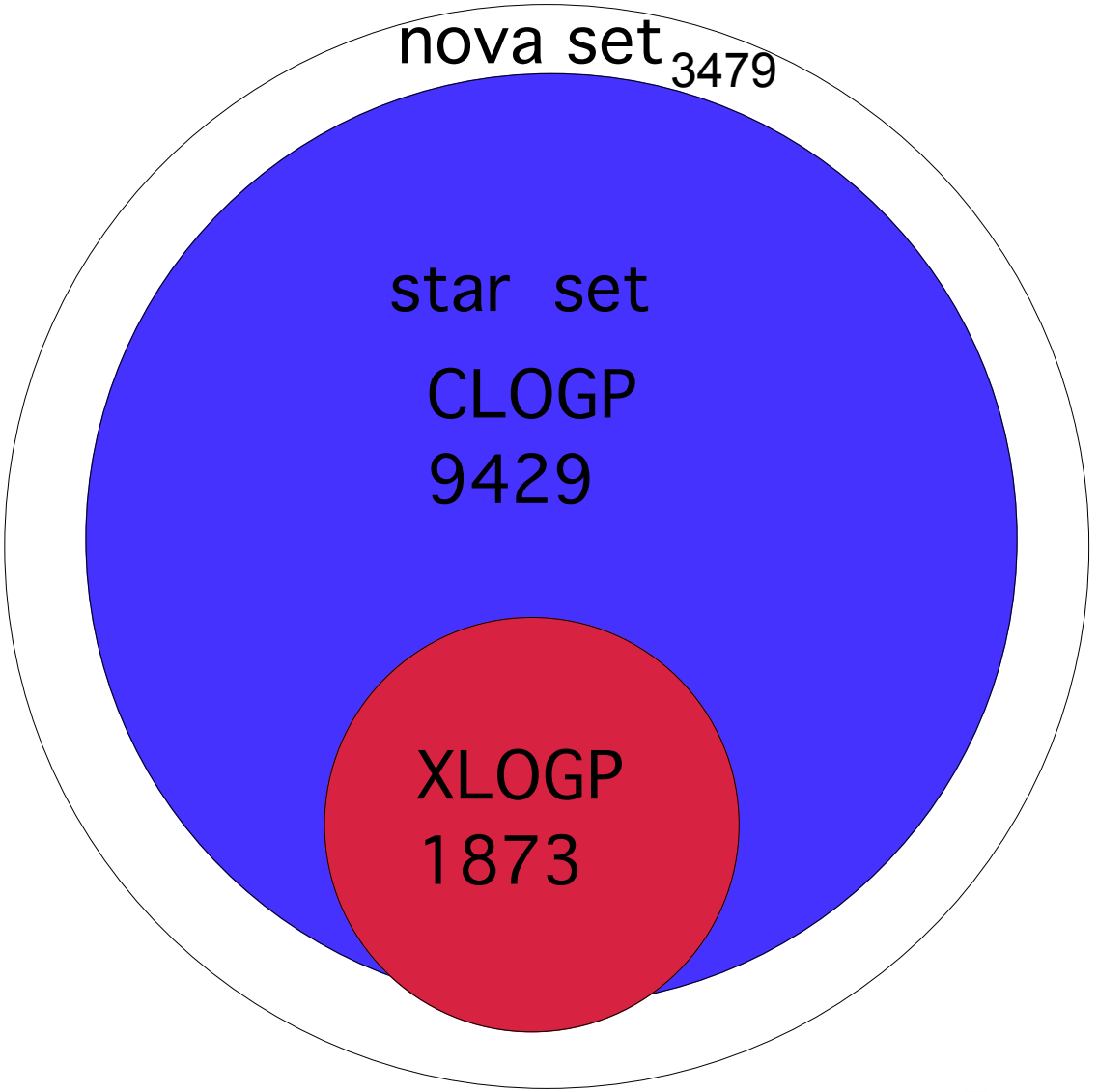
ALOGPS "as is" → ALOGPS LIBRARY

Pallas PrologD :	*MAE = 1.06, RMSE=1.41*
ACDlogD (v. 7.19):	*MAE = 0.97, RMSE=1.32*
ALOGPS:	*MAE = 0.92, RMSE=1.17*
ALOGPS LIBRARY:	*MAE = 0.43, RMSE=0.64*

Tetko & Poda, J. Med. Chem., 2004, 94, 5601-5604.

PHYSPROP data set

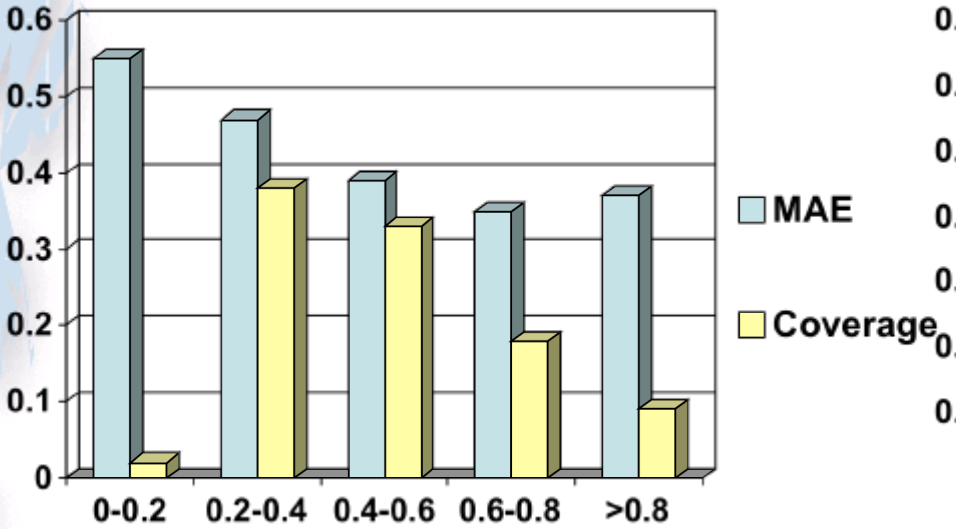
Total: 12908



training
“nova” -->
prediction
star set

Prediction performance as function of similarity in space of "star" set models

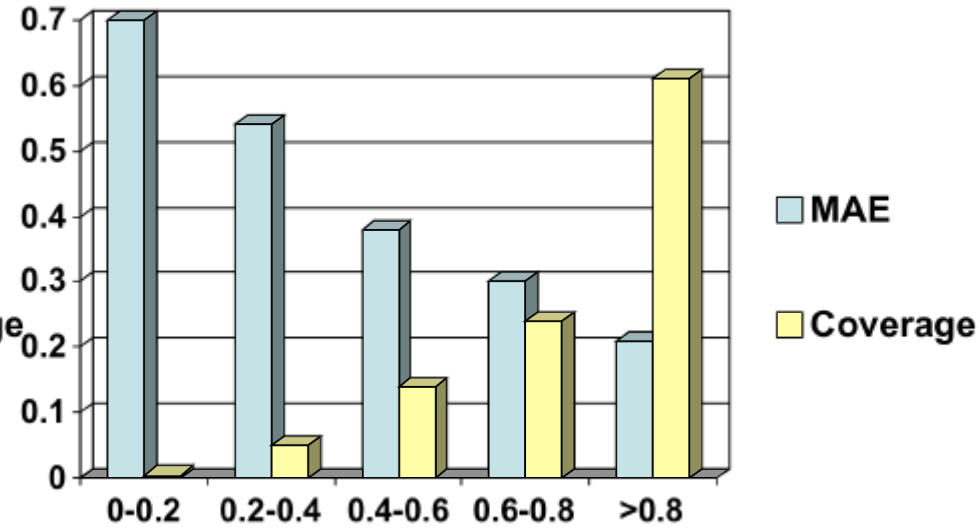
Blind prediction



max correlation coefficient of a test compound to training set compounds

MAE=0.43

LIBRARY mode

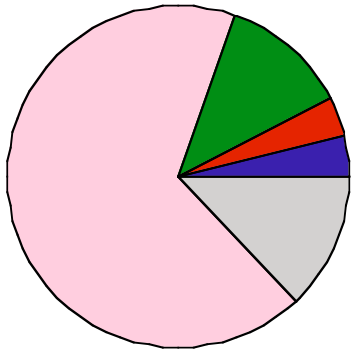


max correlation coefficient of a test compound to LIBRARY compounds

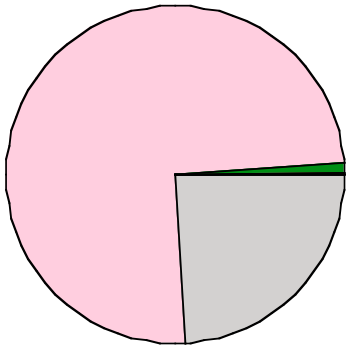
MAE=0.28 (0.26)

Reliability of new compound predictions

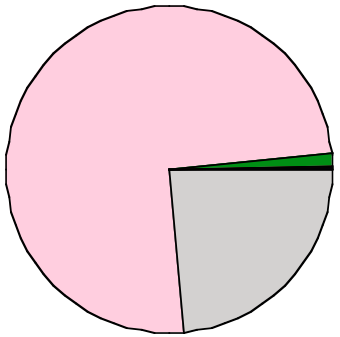
r^2	error
0-0.2	>0.7
0.2-0.4	~0.6
0.4-0.6	~0.5
0.6-0.8	~0.4
0.8-1	<0.3



NCI,
250,000



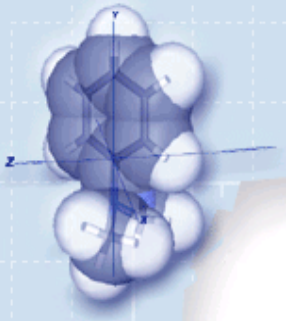
http://asinex.com
120,000



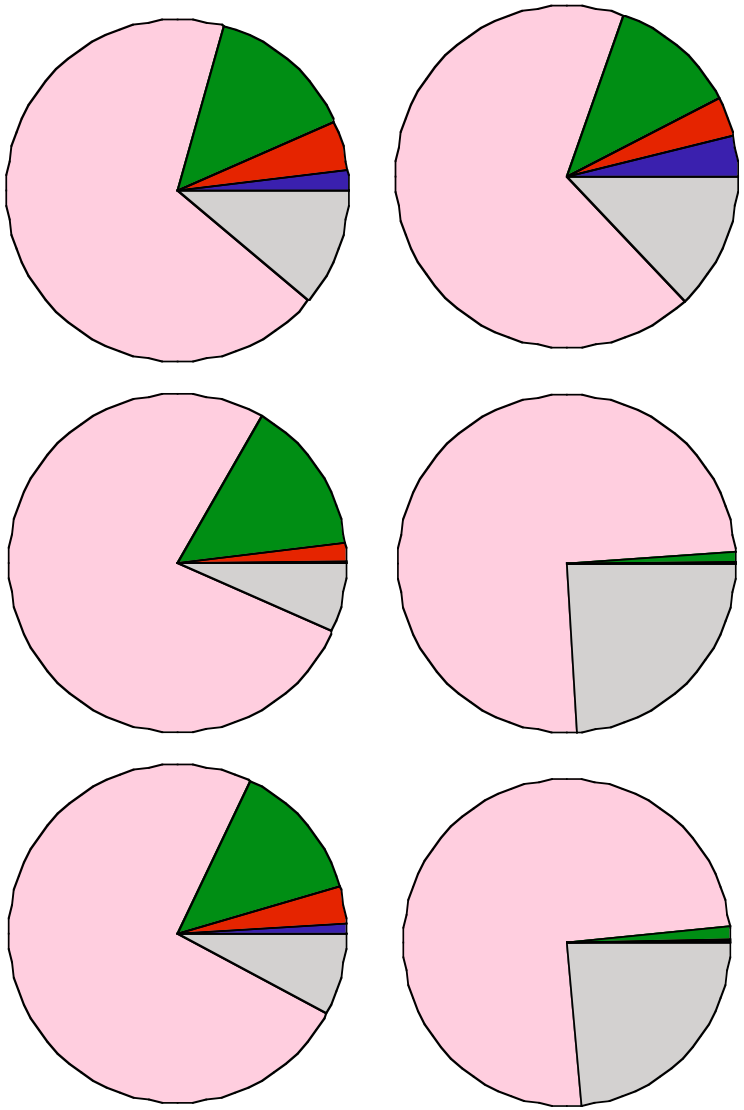
http://ambinter.com
650,000

PHYSPROP

Reliability of new compound predictions



r^2 error	
0-0.2	>0.7
0.2-0.4	~0.6
0.4-0.6	~0.5
0.6-0.8	~0.4
0.8-1	<0.3

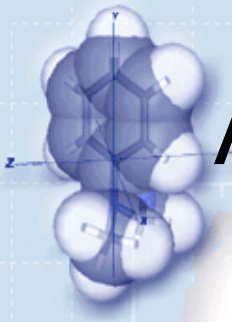


Aurora data PHYSPROP

NCI, 250,000

http://asinex.com 120,000

http://ambinter.com 650,000

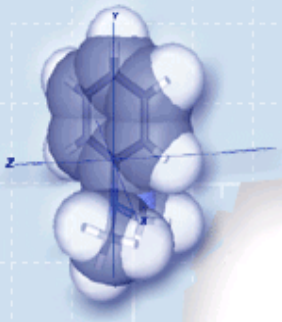


Aqueous solubility / logP prediction for Pfizer data

Towards Predictive ADME Profiling of Drug Candidates: Lipophilicity and Solubility

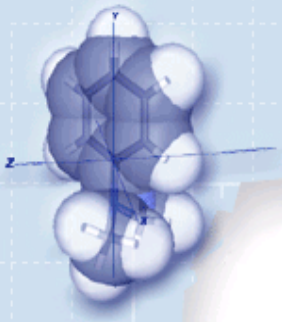
Gennadiy Poda, Igor Tetko and Douglas C. Rohrer

MEDI -- 514
Wednesday, March 16th
18 -- 20



Conclusion

- The LIBRARY mode significantly improves prediction for “in house” logP/S and logD data sets
- The LIBRARY mode can be used with very small number of compounds, i.e one. This number will not be adequate to create new model from “scratch”
- The improvement in this mode due to presence of “invariants” conserved both for training and test sets
- The LIBRARY mode can be used for non-stationary and contradiction data
- An apparent success of logD prediction suggest that similar indices dominates in logP and pKa properties



Acknowledgement

Part of this presentation was done thanks to Virtual Computational Chemistry Laboratory INTAS-INFO 00-0363 project (http://www.vcclab.org).

I thank Prof Hugo Kubinyyi, Drs Pierre Bruneau and Gennadiy Poda for collaboration and Prof. Tudor Oprea for inviting me to participate in this conference.

Thank you for your attention!

Free on-line/batch analysis on http://www.vcclab.org

Welcome to the ALOGPS 2.1 program!

Provide CAS RN or SMILES of a molecule and press the "submit" button © VCCLAB

c1ccccc1 submit

Upload a file with molecule(s) in 48 formats upload file molecule editor

Benzene delete get values

CAS RN	71-43-2	formula	C6H6	MW	78.11
SMILES	c1ccccc1				

logP (exp) :	2.13	logS (exp) :	-1.64 (1.79 g/l)
ALOGPs	2.03 <-0.10>	ALOGpS	-1.84 (1.13 g/l) <-0.20>
IA_logP		IA_logS	
CLOGP	2.14 <+0.01>		
miLogP	2.13 <0.00>		
KOWWIN	1.99 <-0.14>	PhysProp reference	
XLOGP	2.02 <-0.11>	Sangster reference	
User's LogP_LIBRARY	upload library	User's LogS_LIBRARY	upload library

Click on calculated result to see details of calculations.
Press underlined links to read about a particular method.
Press LogP or LogS LIBRARY to read how to improve your predictions.
If you have any suggestions or bug reports contact us at root@vcclab.org
We wish you to have only good results!

The calculated results are available.

For more information click on a keyword or a calculated result or contact Igor V. Tetko.
If you see null pointer exception reload this page (java bug of some browsers).

You can also **download a stand-alone version** of the program