Can we estimate the accuracy of ADMET predictions?

Igor V. Tetko[1], Pierre Bruneau[2], Hans-Werner Mewes[1],
Douglas Rohrer[3], and Gennadiy Poda[3]

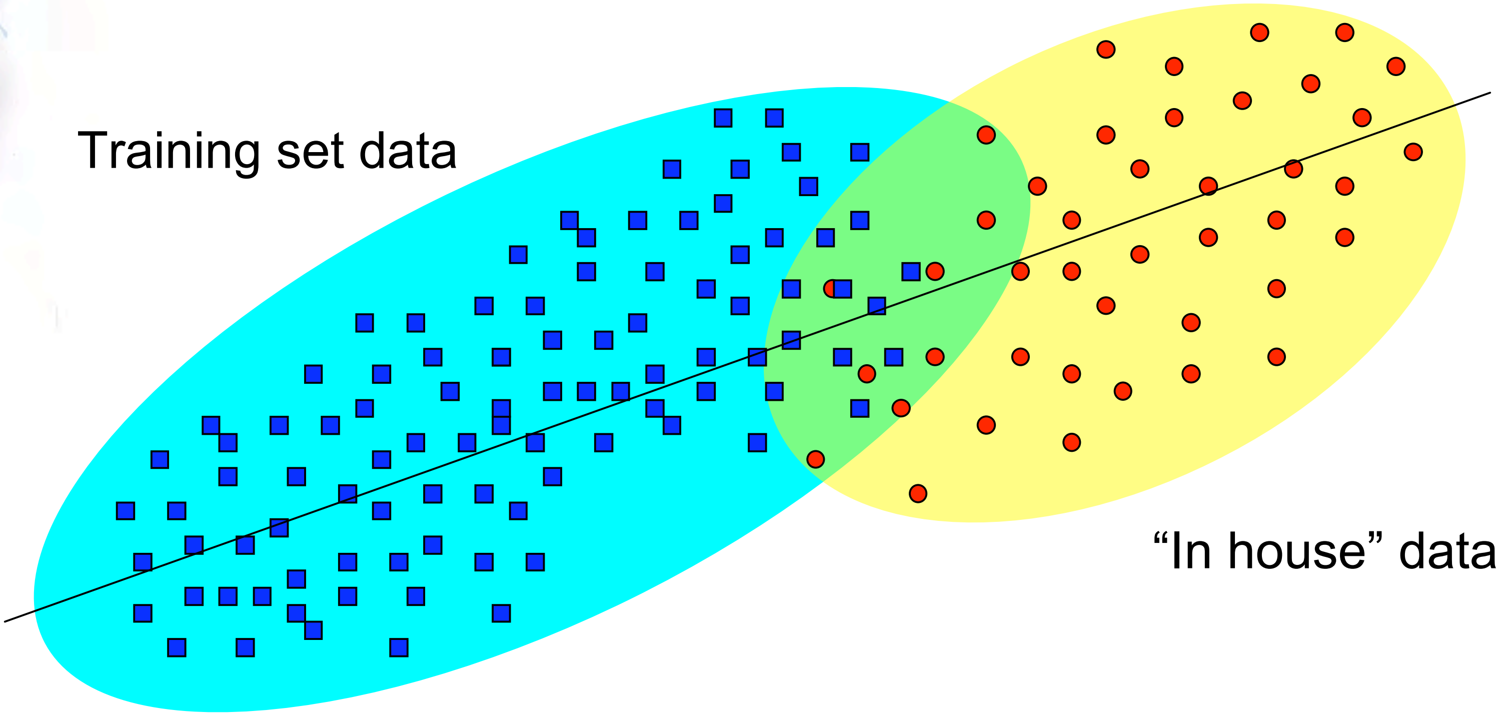
(1) GSF - National Centre for Environment and Health, Institute for Bioinformatics, Ingolstaedter Landstrasse 1, Neuherberg, 85764, Germany,

(2) Centre de Recherche, AstraZeneca, Parc Industriel Pompelle, BP 1050, Reims, France,

(3) Structural & Computational Chemistry, Pfizer Global R & D, 700 Chesterfield Parkway West, Mail Zone BB4G, Chesterfield, MO 63017

Tuesday, 12 September 2006 Moscone Center, 232th ACS meeting, San Francisco

Prediction Space of the model does not cover the "in house" compounds



Training set data

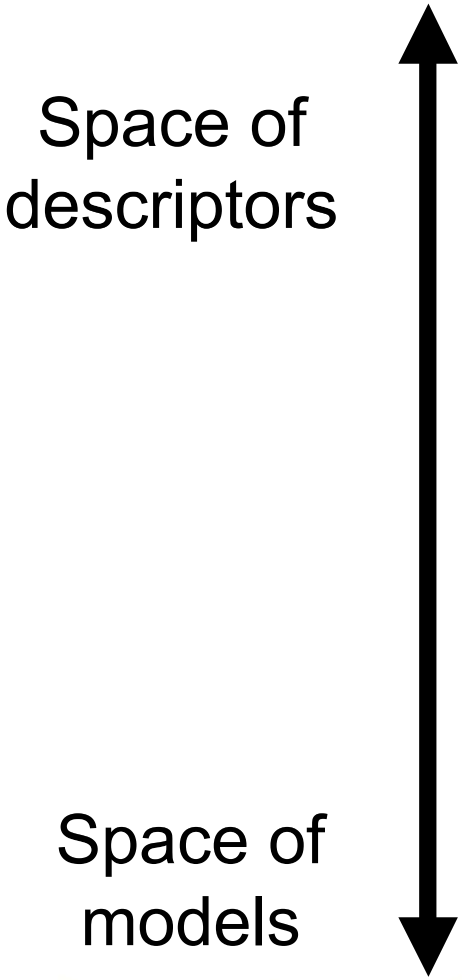
"In house" data

Applicability Domain Methods

- Range-based methods
- Geometric methods
- Distance-based methods
- Probability-density distribution

- Property-based tailoring
- Weighted distances

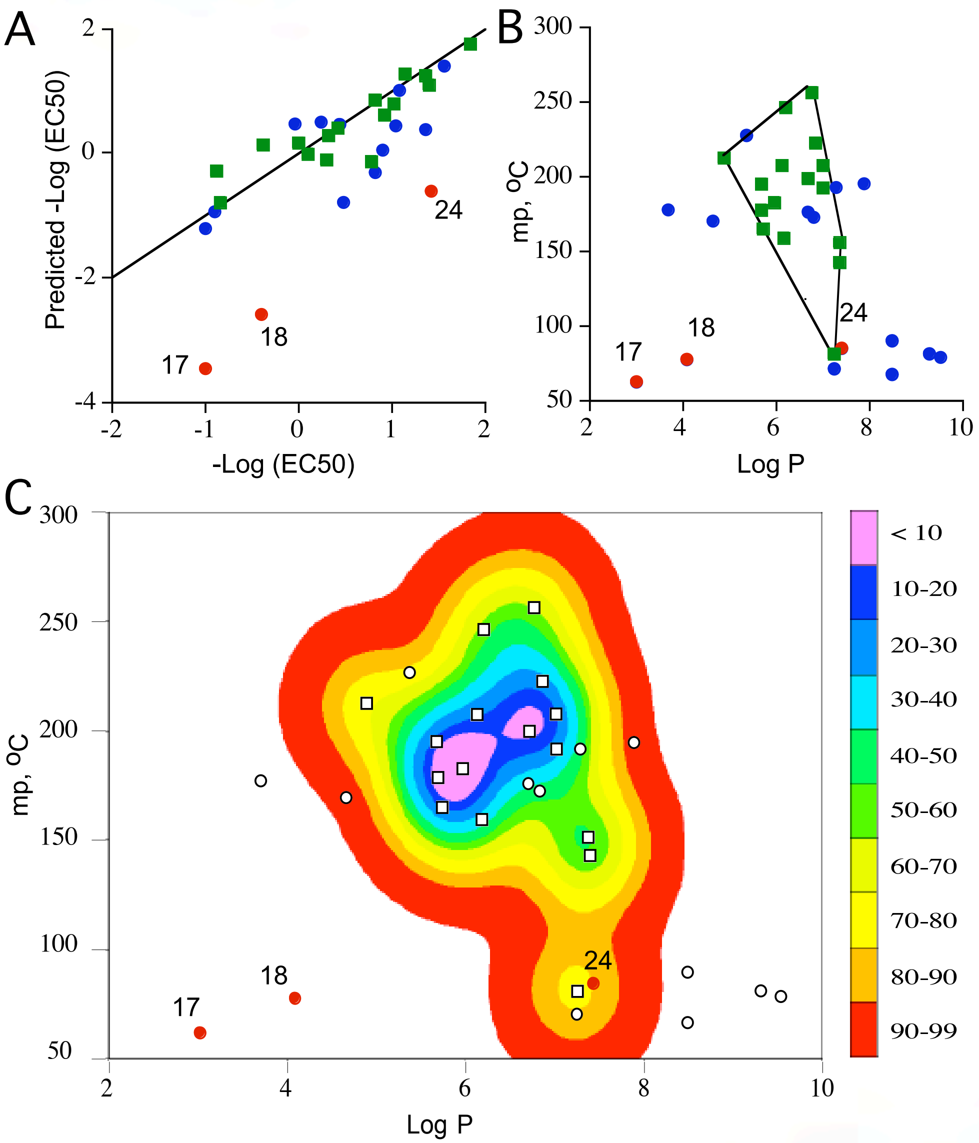
- Ensemble methods
- Analysis of residuals



Space of descriptors

Space of models

Analysis of the antifilarial antimycin analogues (Selwood dataset)



Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

Why property-based space?

In space of descriptors:

- Detection of correct neighborhood relations depends on selection and normalization of descriptors
- Dependencies in the input space are static and do not change with analyzed properties

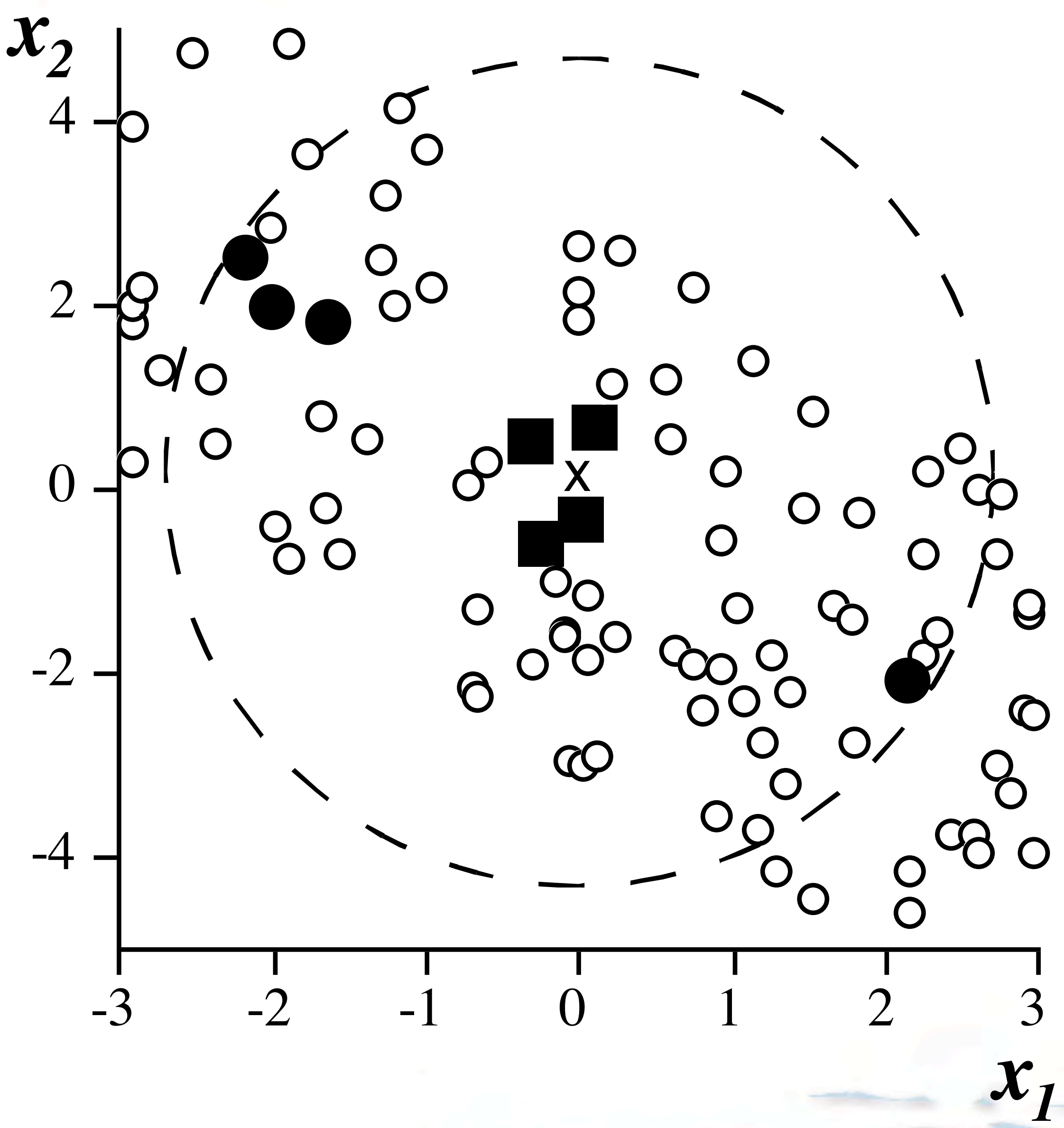
But...

- Supervised learning selects the best combination of descriptors
- Provides their normalization (and non-linear transformations)

Thus

- We should profit from the supervised method and use the models to determine the similarity!

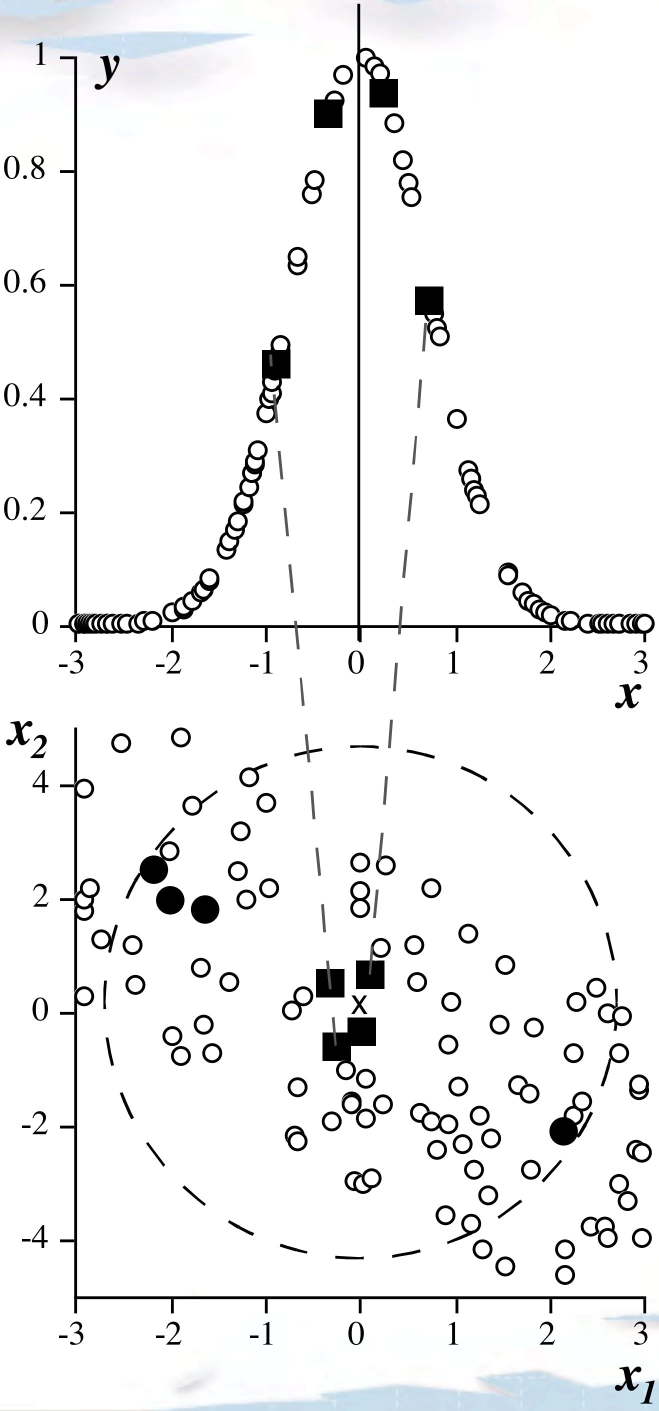
Nearest neighbors in the input space



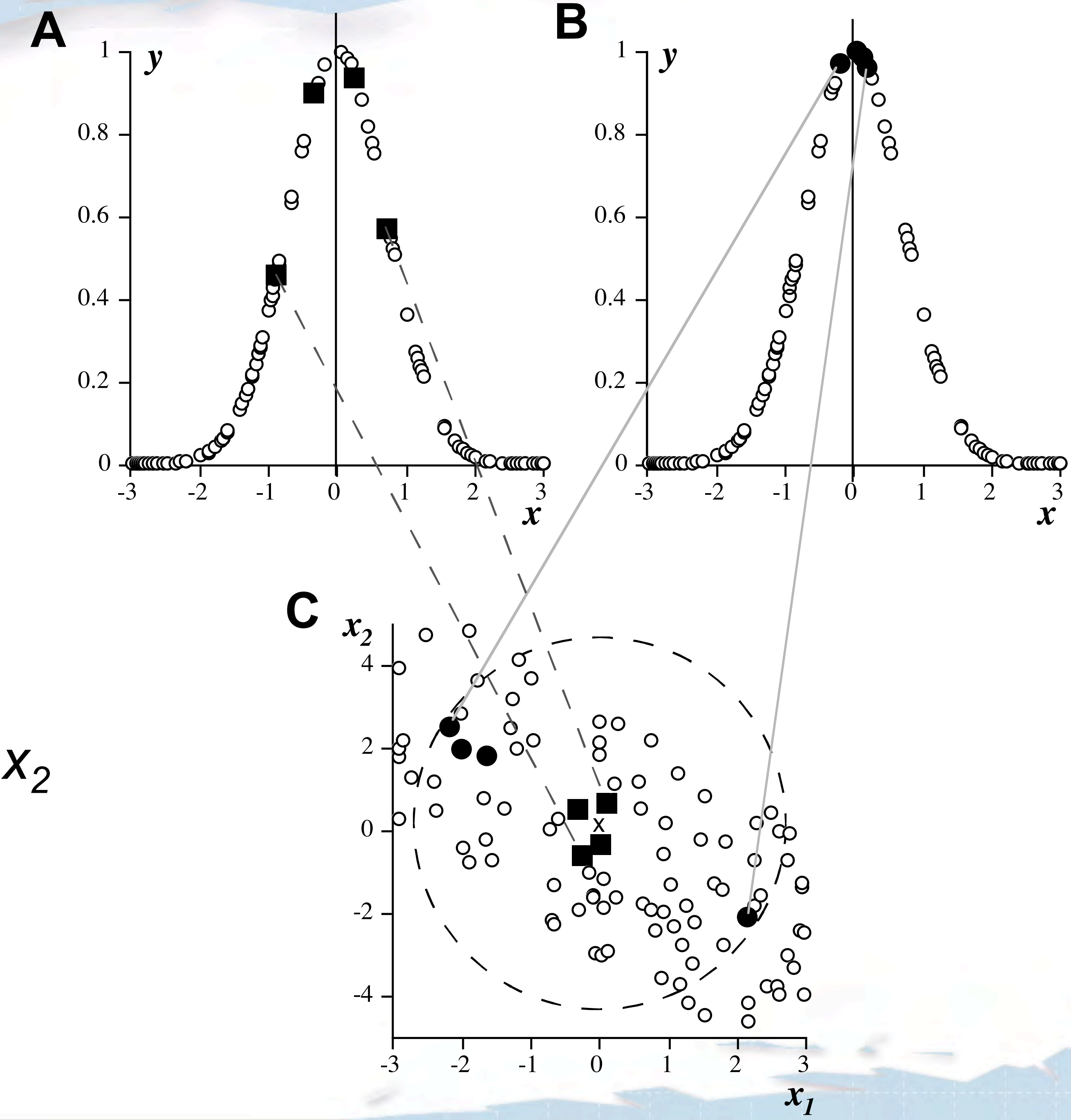
Nearest neighbors and the property

$y=exp(-(x_1+x_2)^2)$

$x=x_1+x_2$!



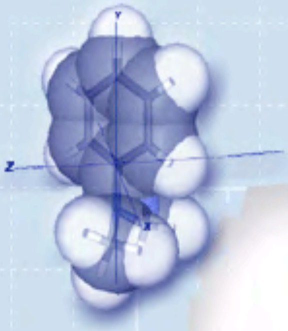
Nearest neighbors and the property



$x=x_1+x_2$

Ensemble methods

Tetko, I. V.; Luik, A. I.; Poda, G. I. Applications of neural networks in structure-activity relationships of a small number of molecules, *J. Med. Chem*., 1993, 36, 811-4.

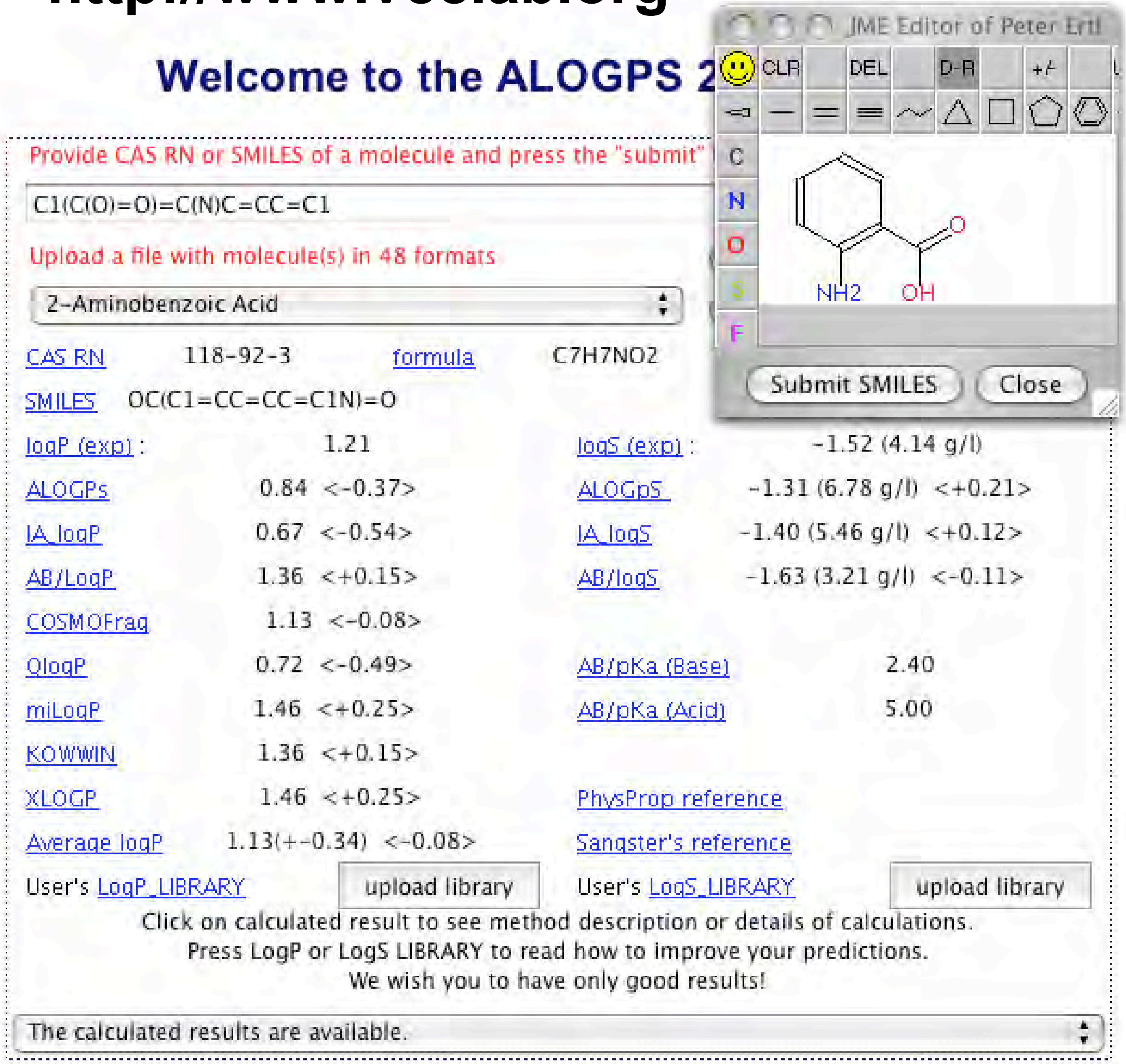


Virtual Computational Chemistry Laboratory

http://www.vcclab.org

ALOGPS 2.1

- LogP: 75 variables, 12908 molecules, RMSE=0.35, MAE=0.26
- LogS: 33 variables, 1291 molecules, RMSE=0.49, MAE=0.35



Welcome to the ALOGPS 2

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

C1(C(O)=O)=C(N)C=CC=C1

Upload a file with molecule(s) in 48 formats

2-Aminobenzoic Acid

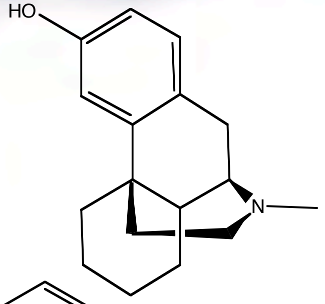
CAS RN	118-92-3	formula	C7H7NO2
SMILES	OC(C1=CC=CC=C1N)=O		
logP (exp) :	1.21	logS (exp) :	-1.52 (4.14 g/l)
ALOGPs	0.84 <-0.37>	ALOGpS	-1.31 (6.78 g/l) <+0.21>
IA_logP	0.67 <-0.54>	IA_logS	-1.40 (5.46 g/l) <+0.12>
AB/LogP	1.36 <+0.15>	AB/logS	-1.63 (3.21 g/l) <-0.11>
COSMOFrag	1.13 <-0.08>		
QlogP	0.72 <-0.49>	AB/pKa (Base)	2.40
miLogP	1.46 <+0.25>	AB/pKa (Acid)	5.00
KOWWIN	1.36 <+0.15>		
XLOGP	1.46 <+0.25>	PhysProp reference	
Average logP	1.13(+-0.34) <-0.08>	Sangster's reference	
User's LogP_LIBRARY	upload library	User's LogS_LIBRARY	upload library

Click on calculated result to see method description or details of calculations.
Press LogP or LogS LIBRARY to read how to improve your predictions.
We wish you to have only good results!

The calculated results are available.

An example of logP prediction

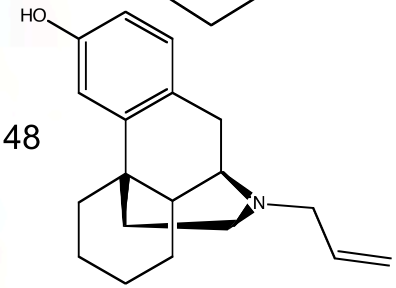
logP=3.11



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

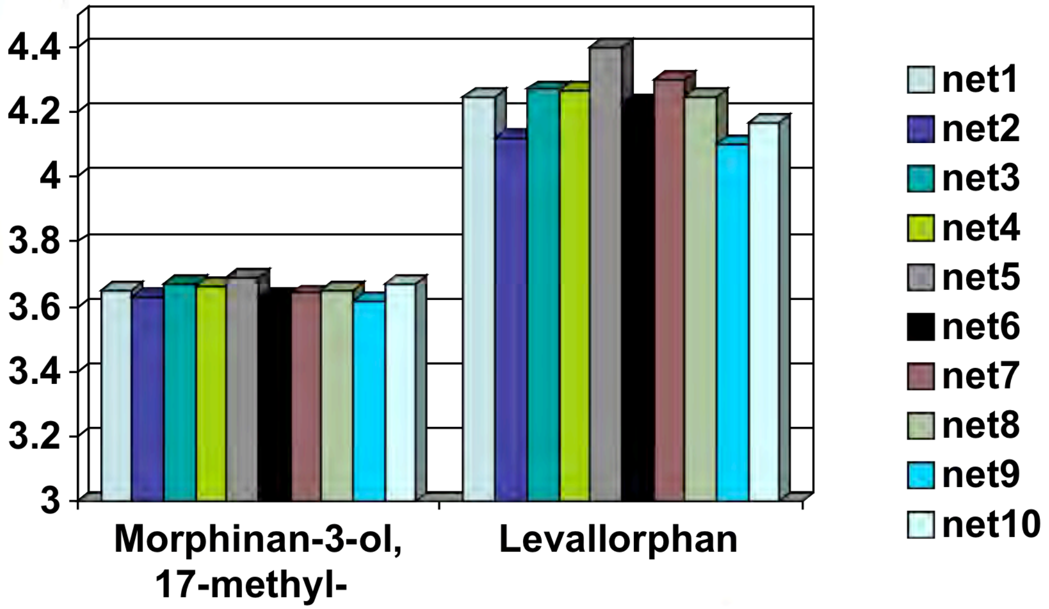
Morphinan-3-ol, 17-methyl-

logP=3.48



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

Levallorphan

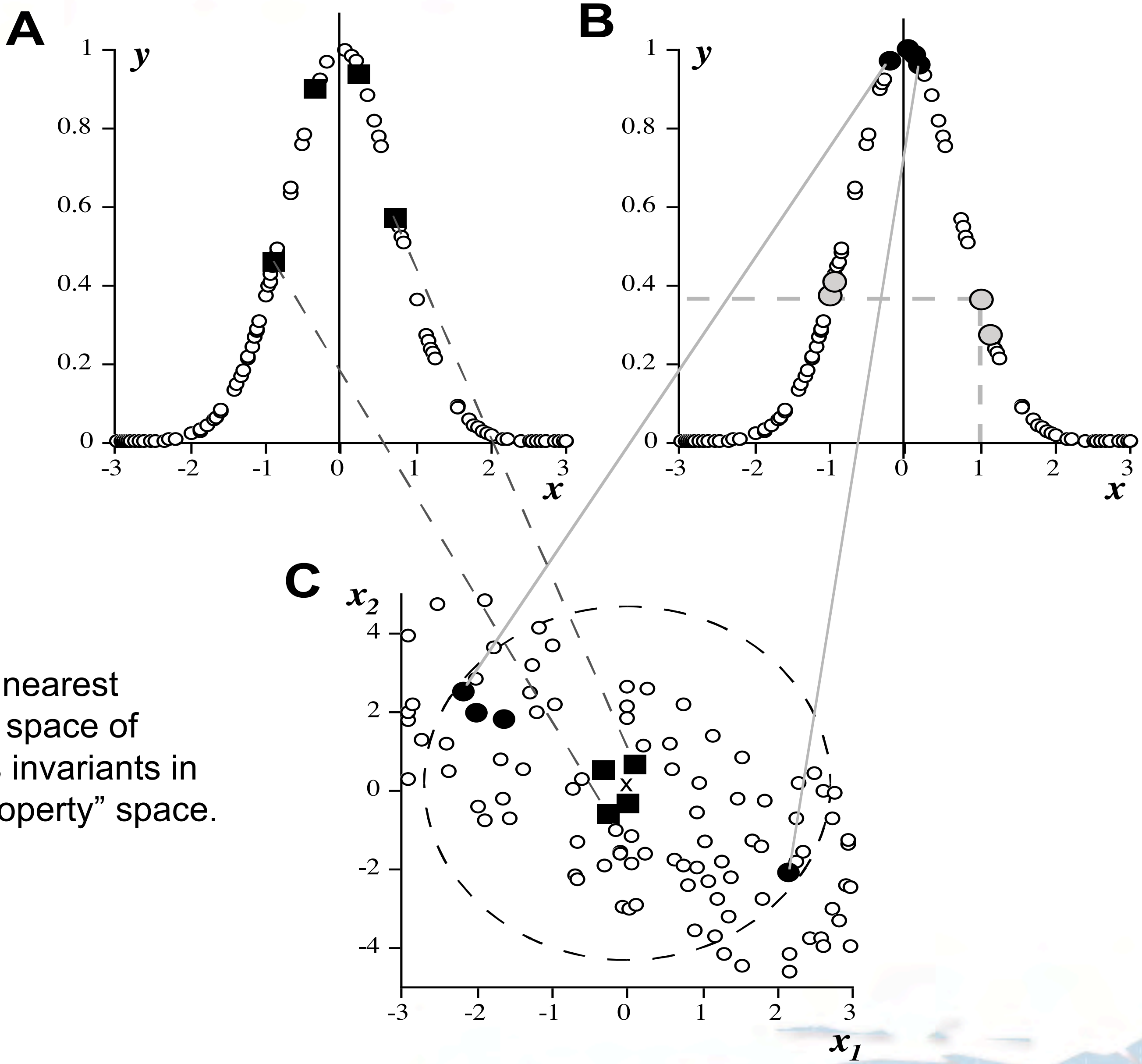


-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals amid >12,000 molecules!

R^2 amid ensemble residuals **IS** the property-based similarity

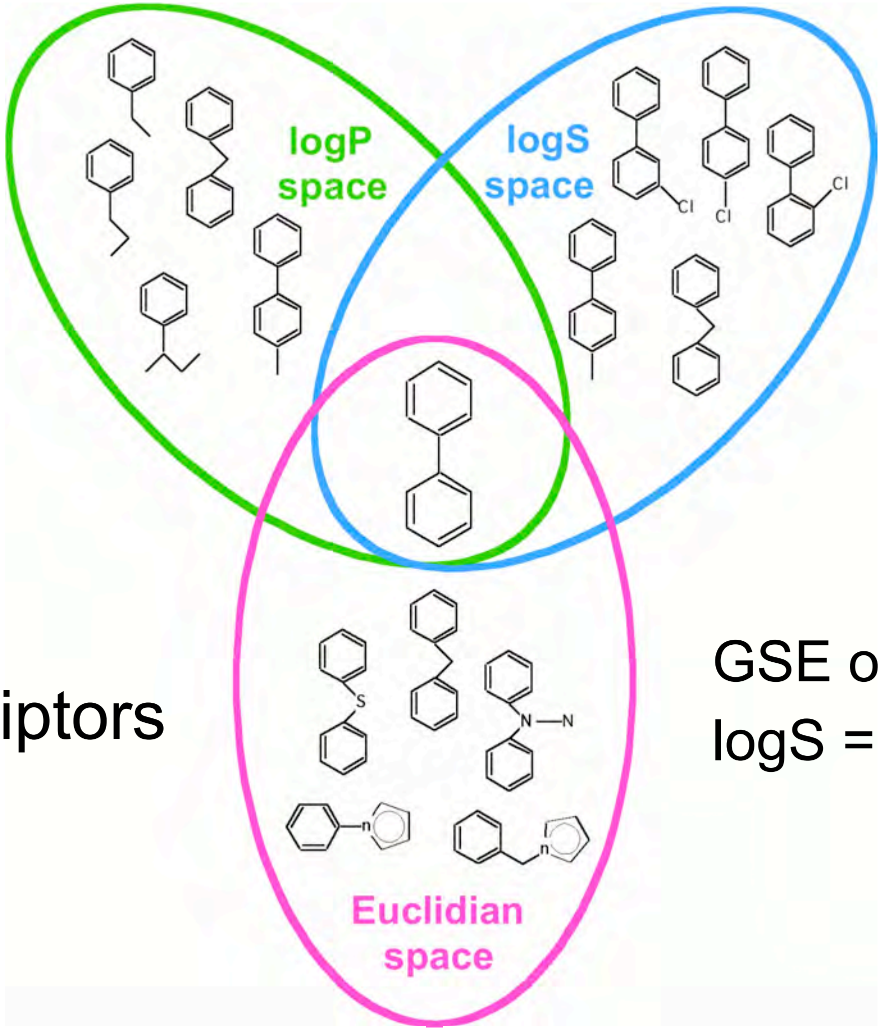
Tetko, I.V.; Villa, A.E.P. Neural Networks, 1997, 10, 1361-1374

Nearest neighbors for Gauss function



Detection of nearest neighbors in space of models uses invariants in “structure-property” space.

Nearest neighbors in different spaces



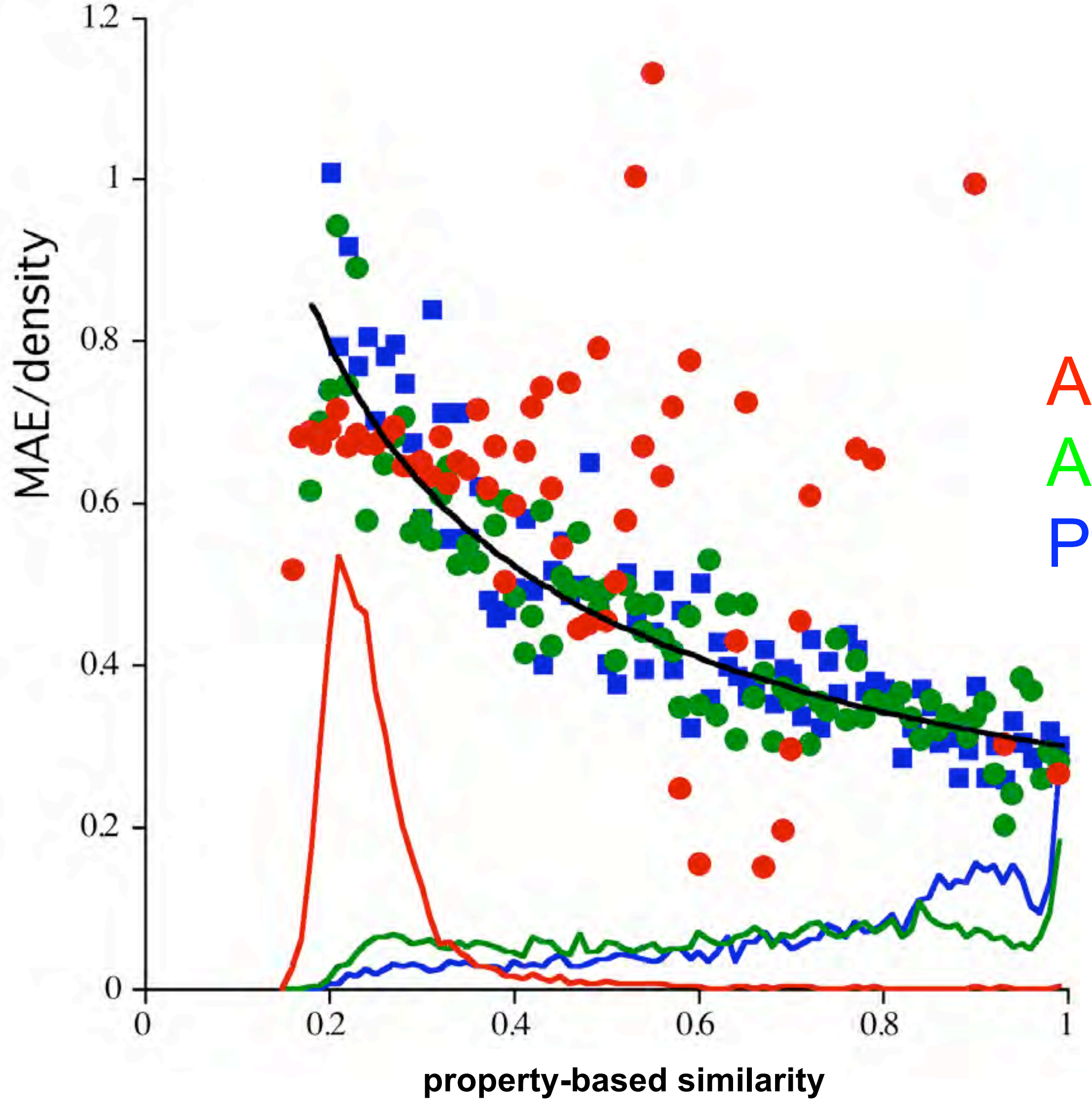
The same 74
E-state descriptors
were used

GSE of S. Yalkowsky

$\log S = 0.5 - 0.01(MP-25) - \log P$

Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

Accuracy of logP prediction as a function of property-based similarity



AstraZeneca blind

AstraZeneca LIBRARY

Pfizer LIBRARY

$$MAE_{pred}=0.302*R^{-0.6}$$

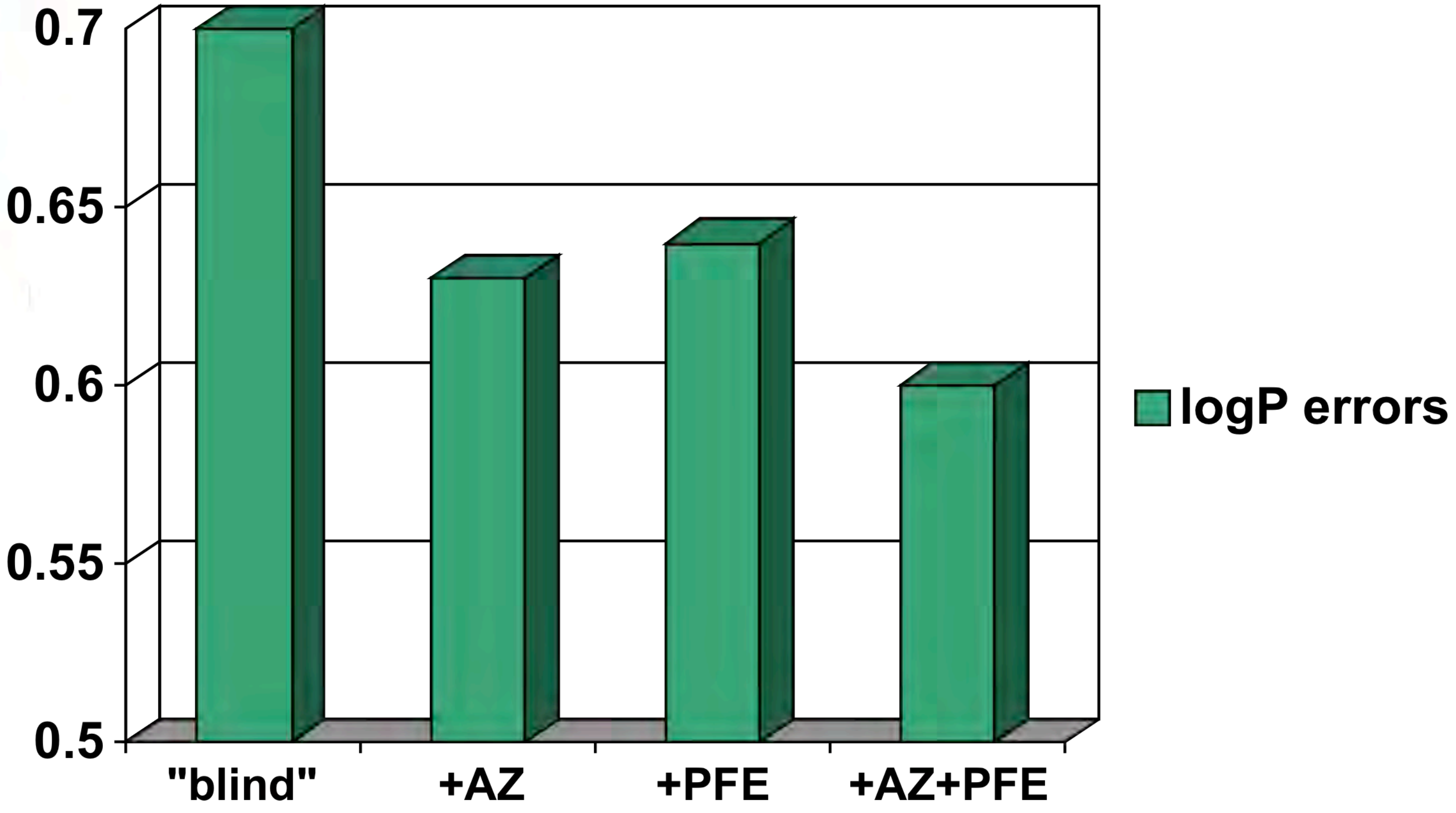
Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

Estimated and calculated MAE for AstraZeneca (AZ) and Pfizer (PFE) sets

dataset	size	training set	estimated	calculated
AZ	7498	PHYSPROP	0.69	0.67
AZ	7498	PHYSPROP+AZ	0.42	0.42
PFE	8750	PHYSPROP	0.72	0.74
PFE	8750	PHYSPROP+PFE	0.37	0.37

Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

Estimated errors for >13,000,000 iResearchLibrary molecules

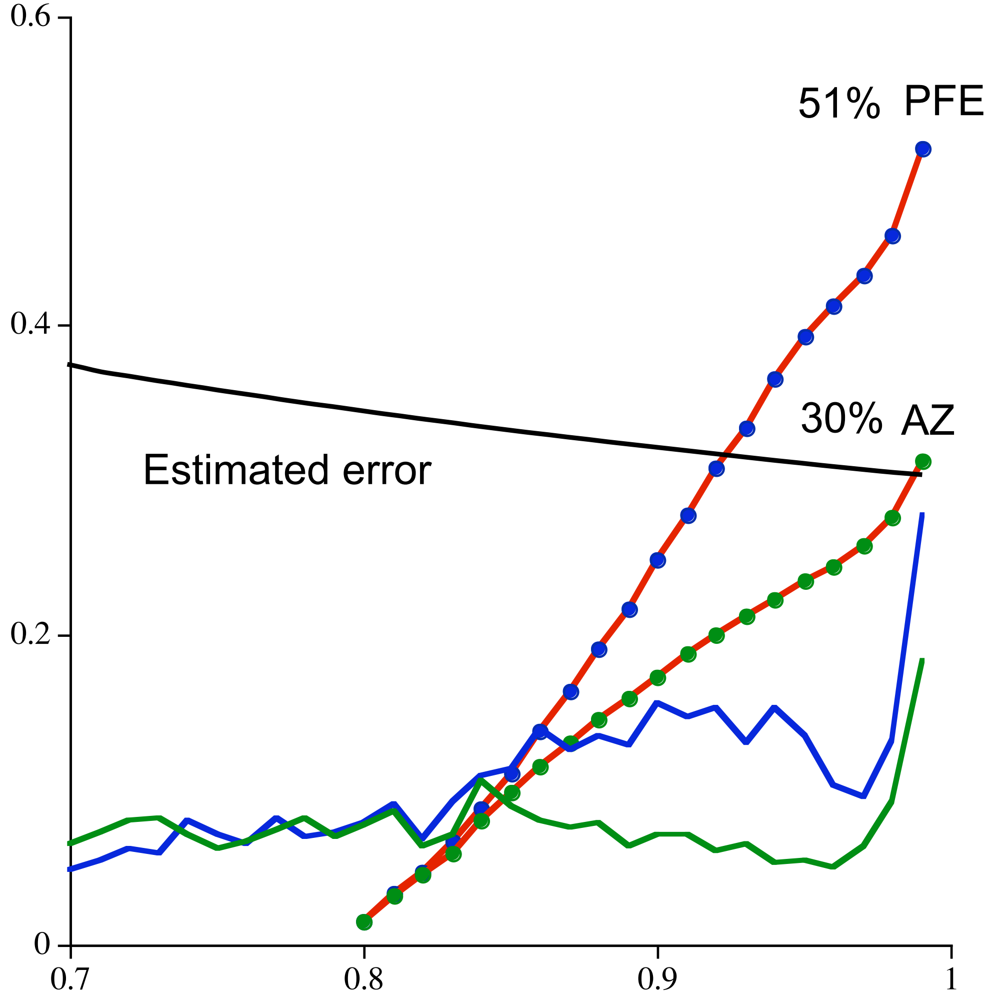


Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

The advantages to use the “in house” data (LIBRARY mode) and error estimations

- 514,000 compounds with logP> 5 for blind prediction were predicted with logP<5 using *PFE in house* data
- 495,000 compounds changed |logP| > 1 log unit in LIBRARY compared to the blind prediction
- Some scaffolds of compounds have incorrect predicted values in “blind prediction” but are correctly predicted in LIBRARY mode using program enriched with *PFE in-house* data
- Some scaffolds are still incorrectly predicted even using *PFE in-house* data.
- But, all these scaffolds with non-reliable predictions can be identified, marked and measured or excluded!

Redundant measurements: R>0.8, MAE < 0.35

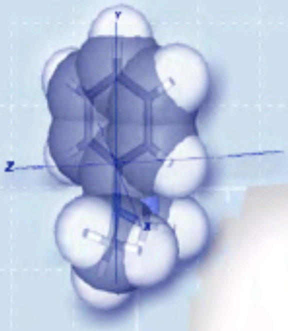


The estimated accuracy allows to avoid measurement of the accurately predicted compounds (30% and 50% for AZ and PFE sets, respectively). The experimental resources can be used to measure the problematic scaffolds.

Similarity in property-based space

- is introduced as correlation between vector of residuals of models[1,2]
- is a heart of the Associative Neural Network method[2,3]
- is specific for the target property[3,4]
- detects meaningful nearest neighbors[3,4]
- estimates accuracy of prediction (applicability domain) of programs[5]
- can be used for secure data sharing[6]

1) Tetko, I.V.; Villa, A.E.P. *Neural Networks*, 1997, 10, 1361.
2) Tetko, I.V.; Tanchuk, V. Yu. *JCICS*, 2002, 42, 1136.
3) Tetko, I.V. *JCICS*, 2002, 42, 717.
4) Tetko, I.V. in D.J. Livingstone, *Neural Networks: Methods and Applications*, CRC, in press.
5) Tetko, I.V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. *DDT*, 2006, 11, 700-7.
6) Tetko, I.V.; Abagyan, R.; Oprea, T.I. *J. Comp. Aid. Mol. Des*. 2005, 19, 749.



Acknowledgement

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

Thank you for your attention!