What is a property-based similarity?

Igor V. Tetko

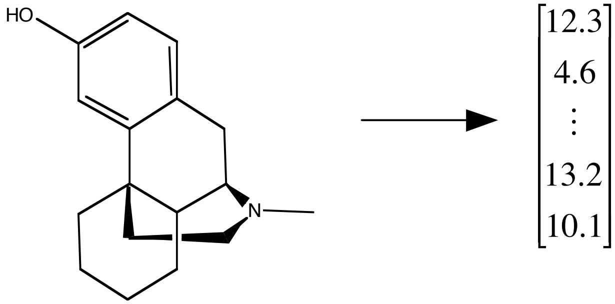
(1) GSF - National Centre for Environment and Health, Institute for Bioinformatics, Ingolstaedter Landstrasse 1, Neuherberg, 85764, Germany,

(2) Institute of Bioorganic & Petrochemistry, Ukrainian Academy of Sciences, Kyiv, Ukraine

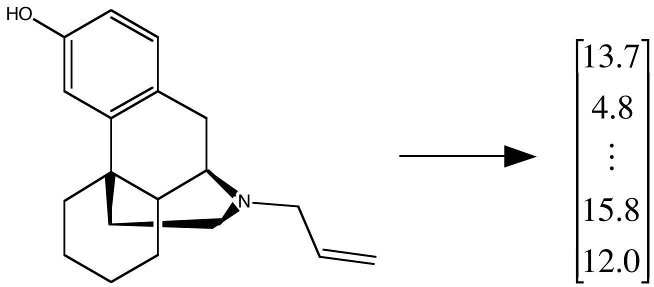
Monday, 11 September 2006 Moscone Center, 232th ACS meeting, San Francisco

Similarity of Molecules

- Usually we describe a molecule with a set of descriptors (topological 2D, 3D, etc.).
- This set of descriptors can be used for similarity search (Tanimoto, Euclidina distance, etc.).
- Problem is how to select and normalize them to better relate to the target property?

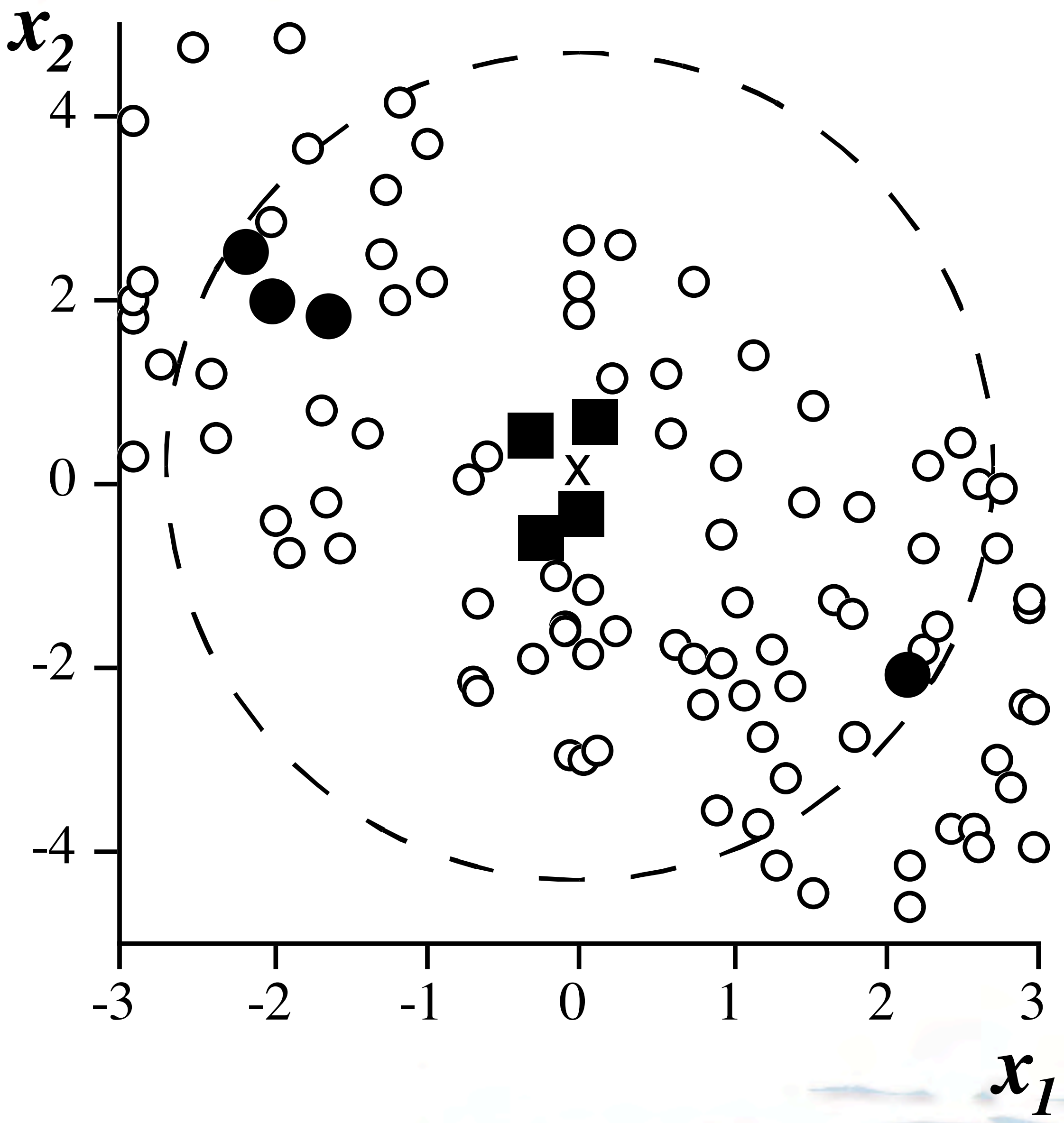


$$\rightarrow \begin{bmatrix} 12.3 \\ 4.6 \\ \vdots \\ 13.2 \\ 10.1 \end{bmatrix}$$



$$\rightarrow \begin{bmatrix} 13.7 \\ 4.8 \\ \vdots \\ 15.8 \\ 12.0 \end{bmatrix}$$

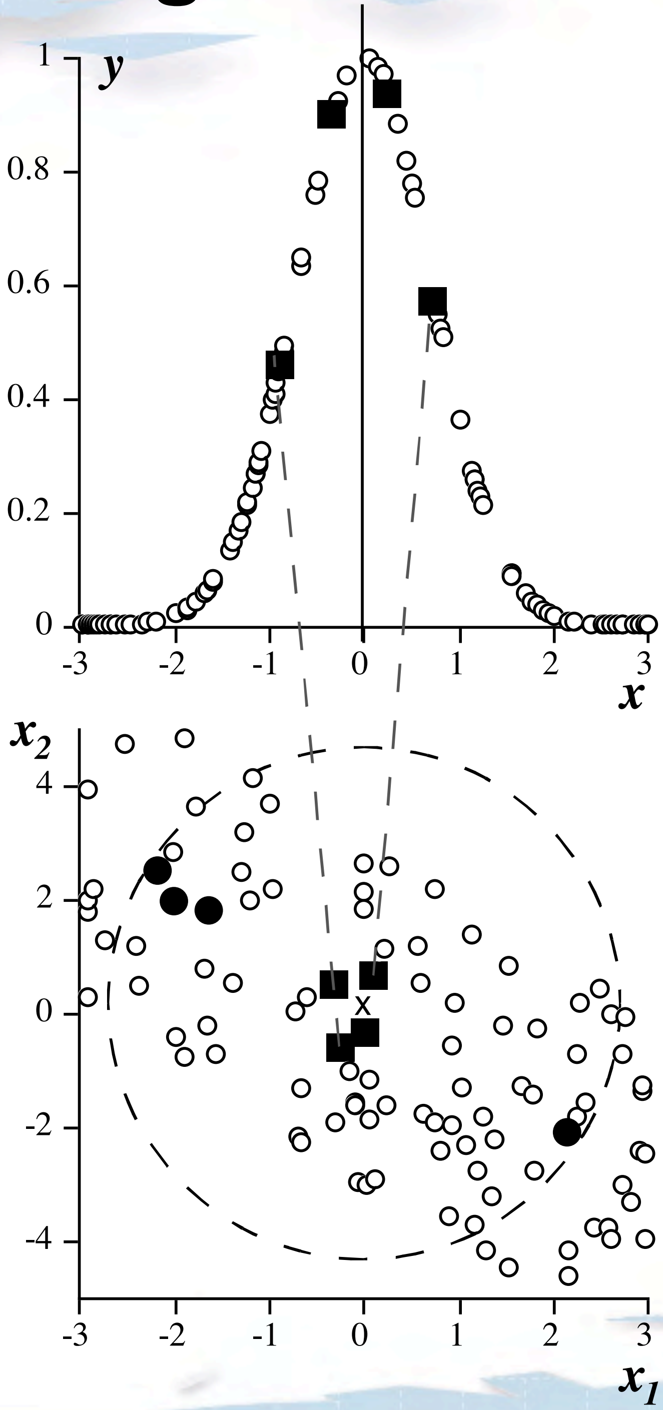
Nearest neighbors in the input space



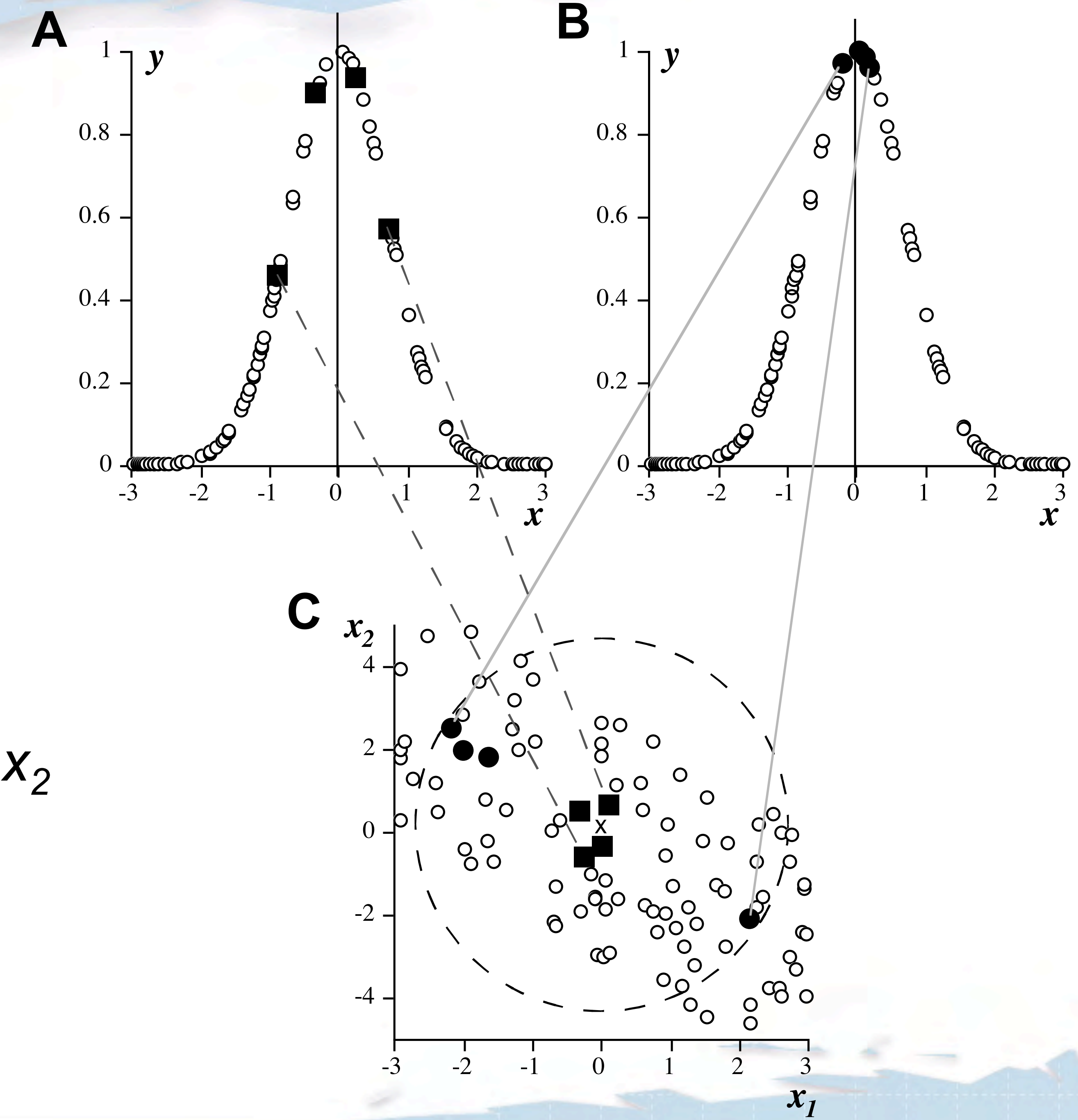
Nearest neighbors and activity

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

$x=x_1+x_2$!



Nearest neighbors and activity



$x=x_1+x_2$

Ensemble methods

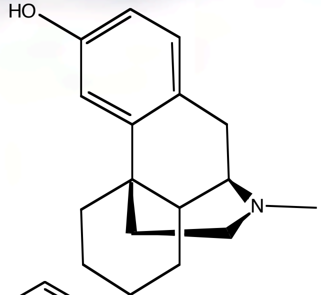
✓Some methods rely just on one “best” model.

✓Other methods rely on the ensemble average (“panel of experts”).

✓ We explore disagreement of individual models in the ensemble to derive a similarity score and improve the ensemble accuracy and to estimate the reliability score.

Example of logP prediction in ALOGPS

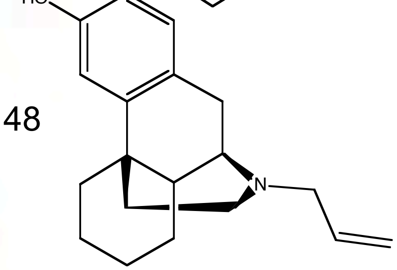
logP=3.11



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

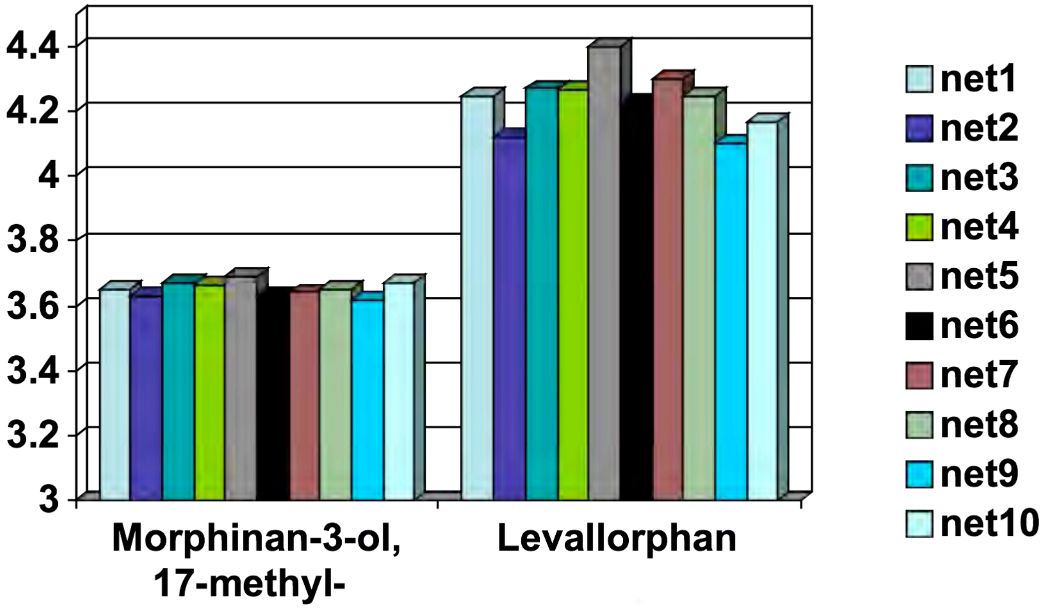
Morphinan-3-ol, 17-methyl-

logP=3.48



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

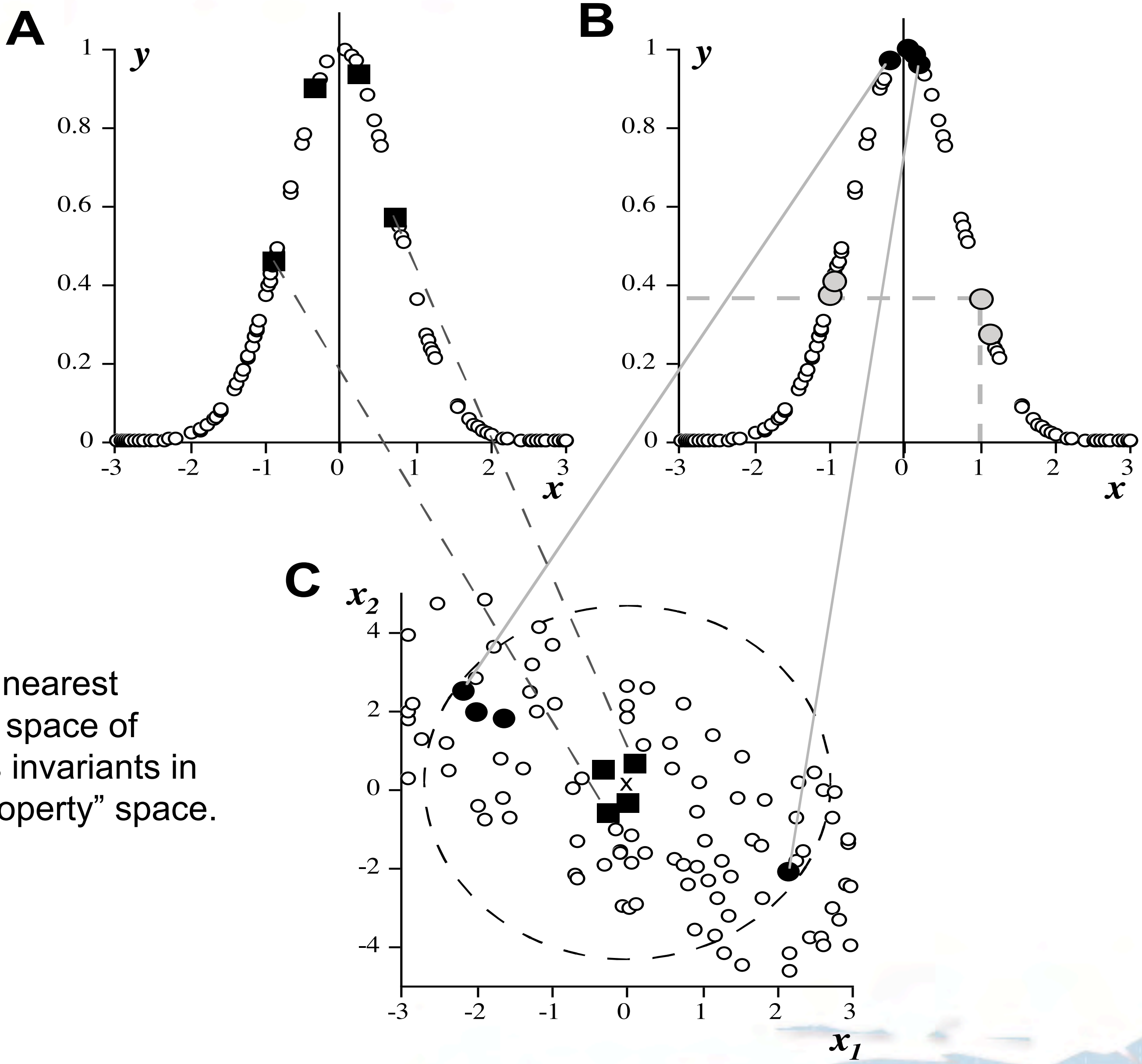
Levallorphan



R^2 of ensemble residuals = the property-based similarity of molecules

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

Nearest neighbors for Gauss function

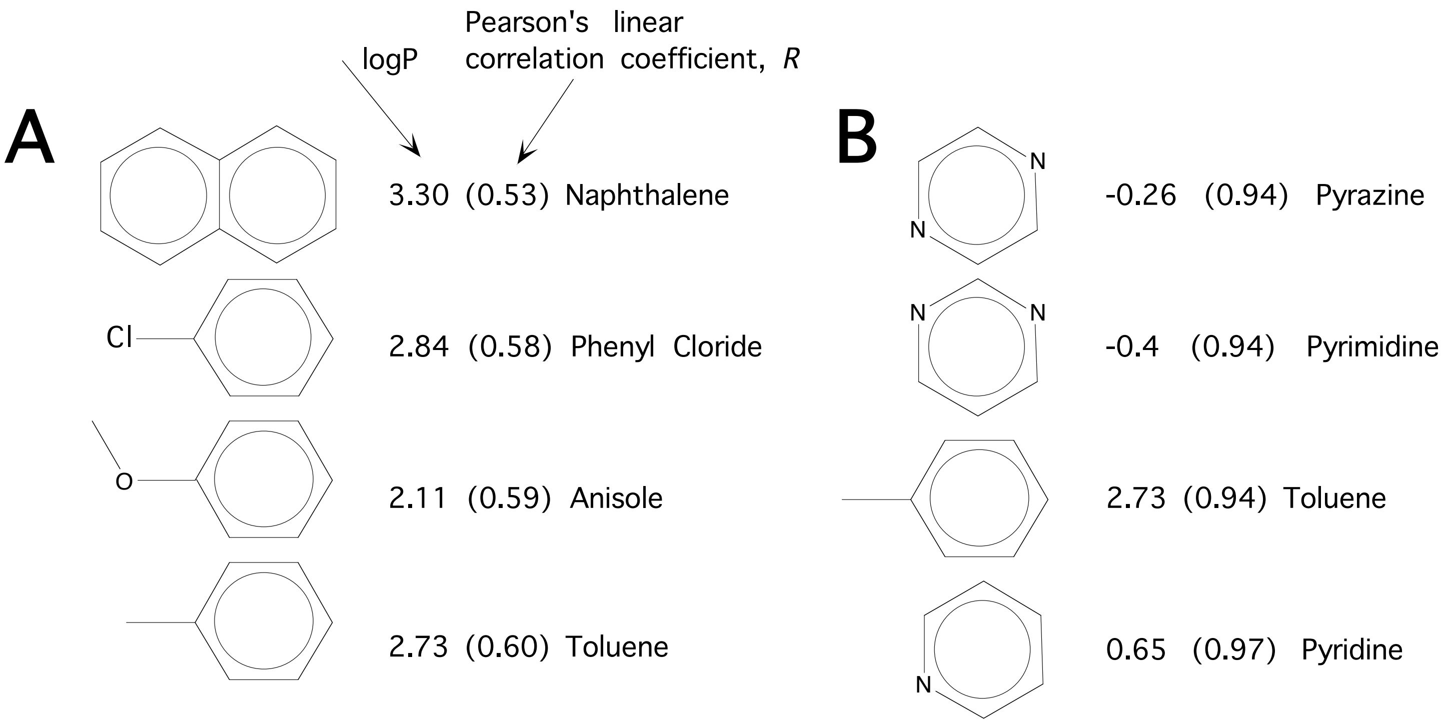


Detection of nearest neighbors in space of models uses invariants in “structure-property” space.

Example of property-based similarity

A: lipophilicity prediction

B: molecular weight prediction



logP

Pearson's linear correlation coefficient, *R*

A

3.30 (0.53) Naphthalene

2.84 (0.58) Phenyl Cloride

2.11 (0.59) Anisole

2.73 (0.60) Toluene

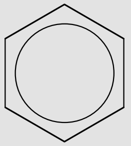
B

-0.26 (0.94) Pyrazine

-0.4 (0.94) Pyrimidine

2.73 (0.94) Toluene

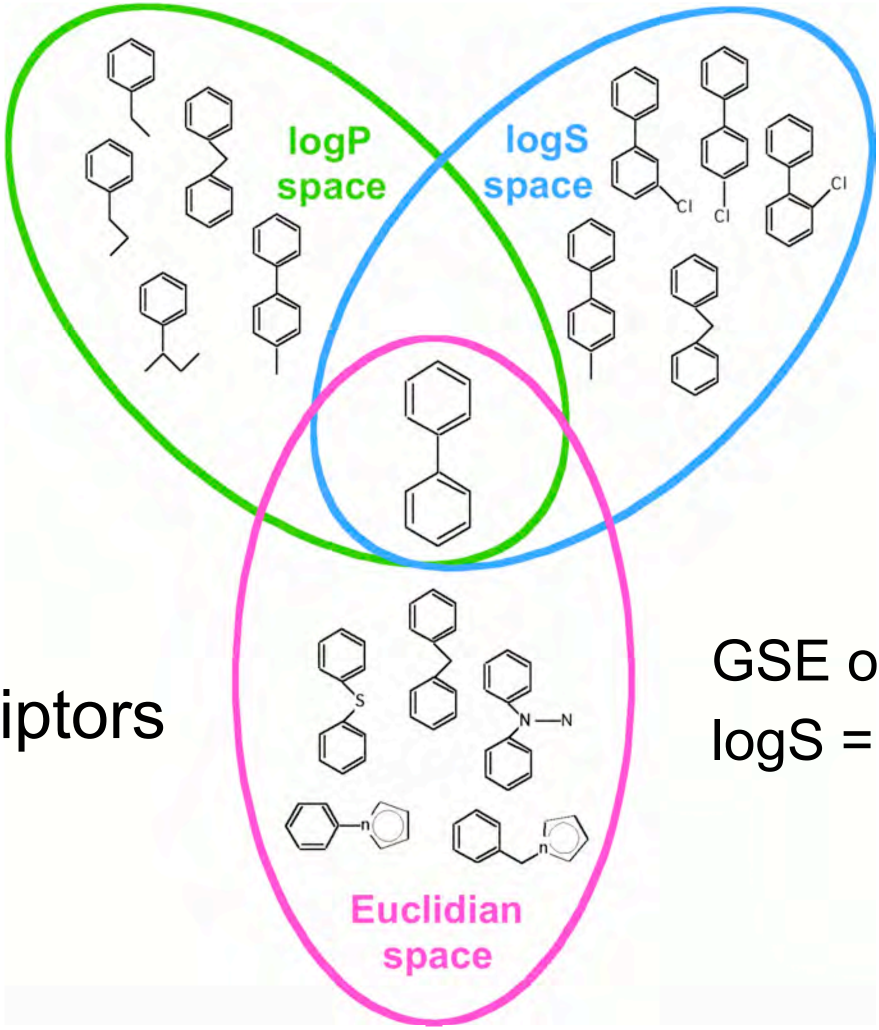
0.65 (0.97) Pyridine



2.13 Benzene

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

Nearest neighbors in different spaces



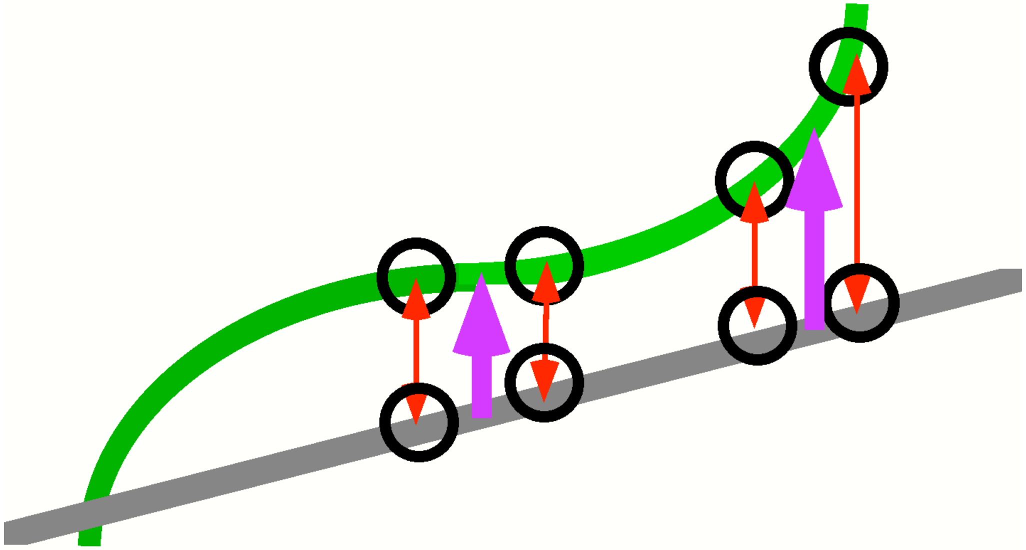
The same 74
E-state descriptors
were used

GSE of S. Yalkowsky

logS = 0.5-0.01(MP-25) - logP

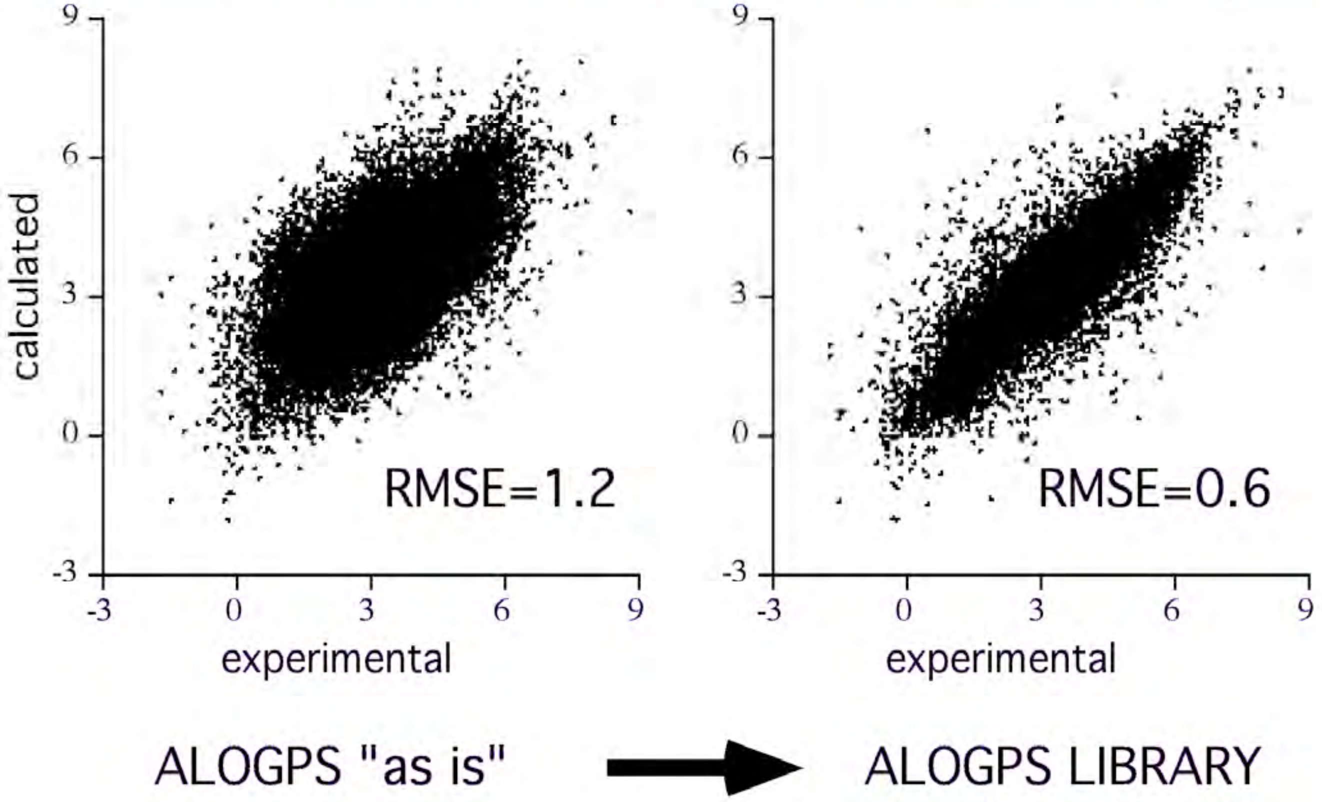
Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions?, Drug Discov. Today, 2006, 11, 700-7

Correction of a model by the nearest neighbors



Prediction of proprietary data

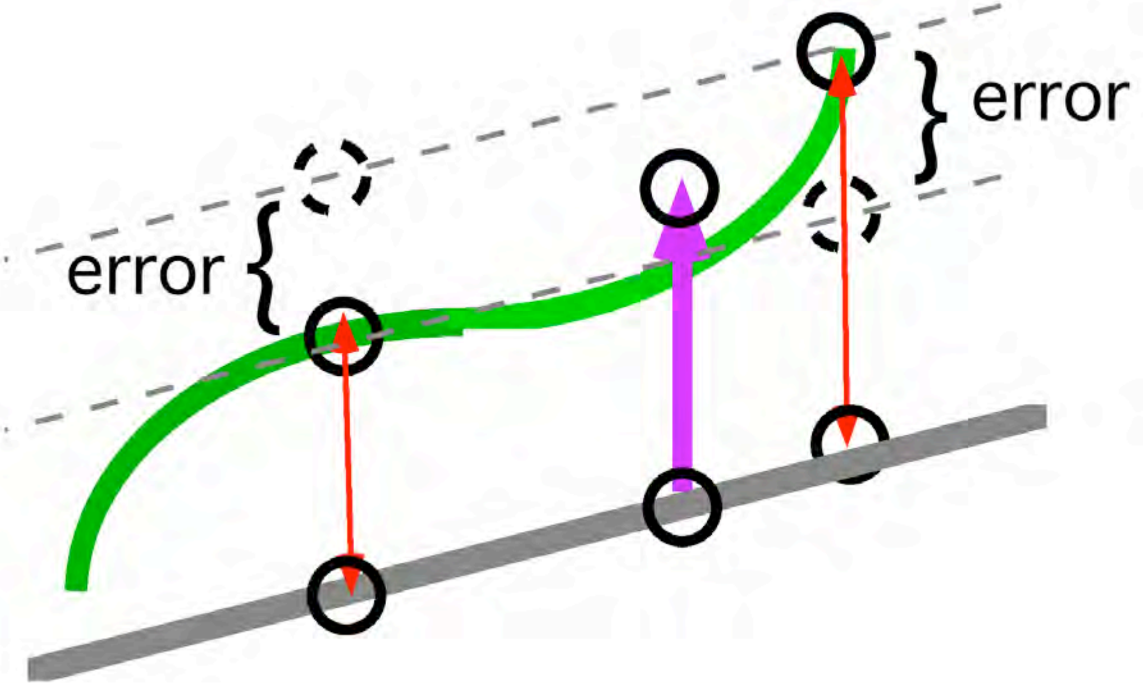
ALOGPS prediction for ElogD set of 17,861 compounds



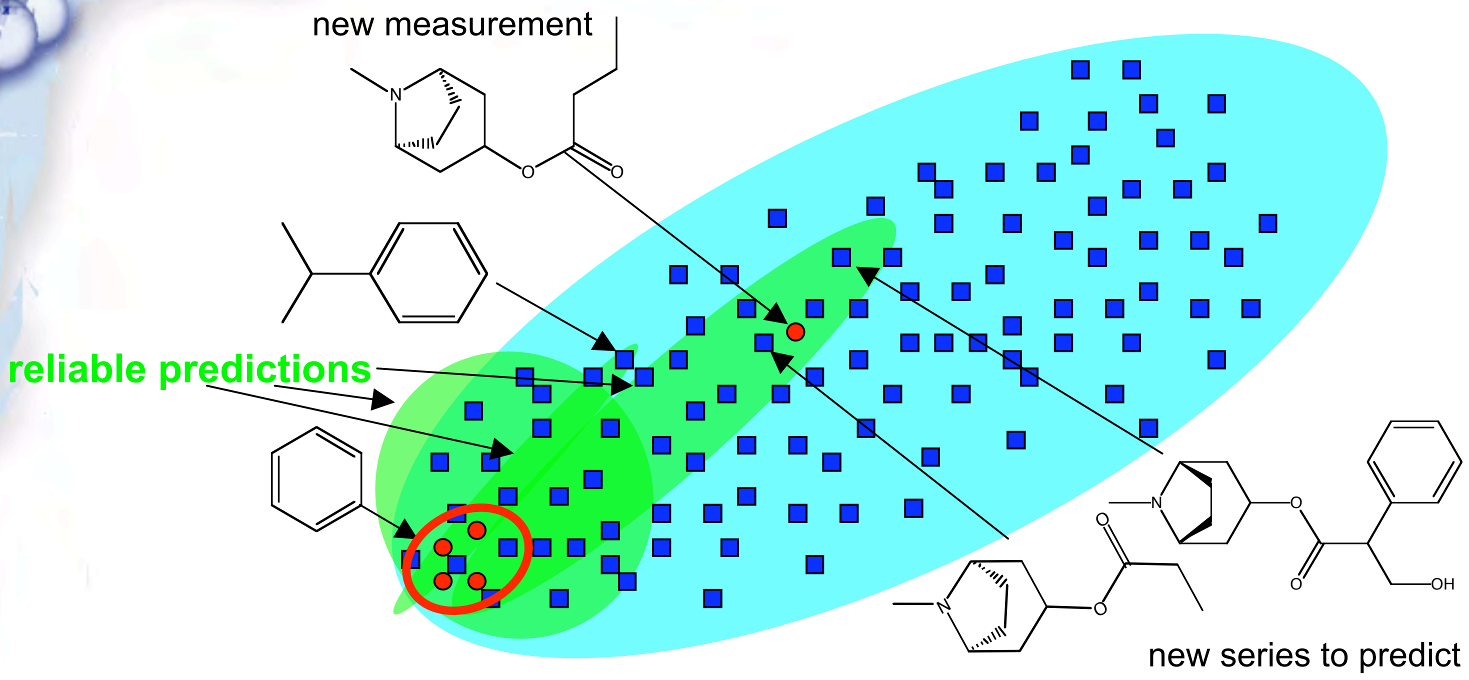
Pallas PrologD :	*MAE = 1.06, RMSE=1.41*
ACDlogD (v. 7.19):	*MAE = 0.97, RMSE=1.32*
ALOGPS:	*MAE = 0.92, RMSE=1.17*
ALOGPS LIBRARY:	*MAE = 0.43, RMSE=0.64*

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

Estimation of the model accuracy by the nearest neighbors



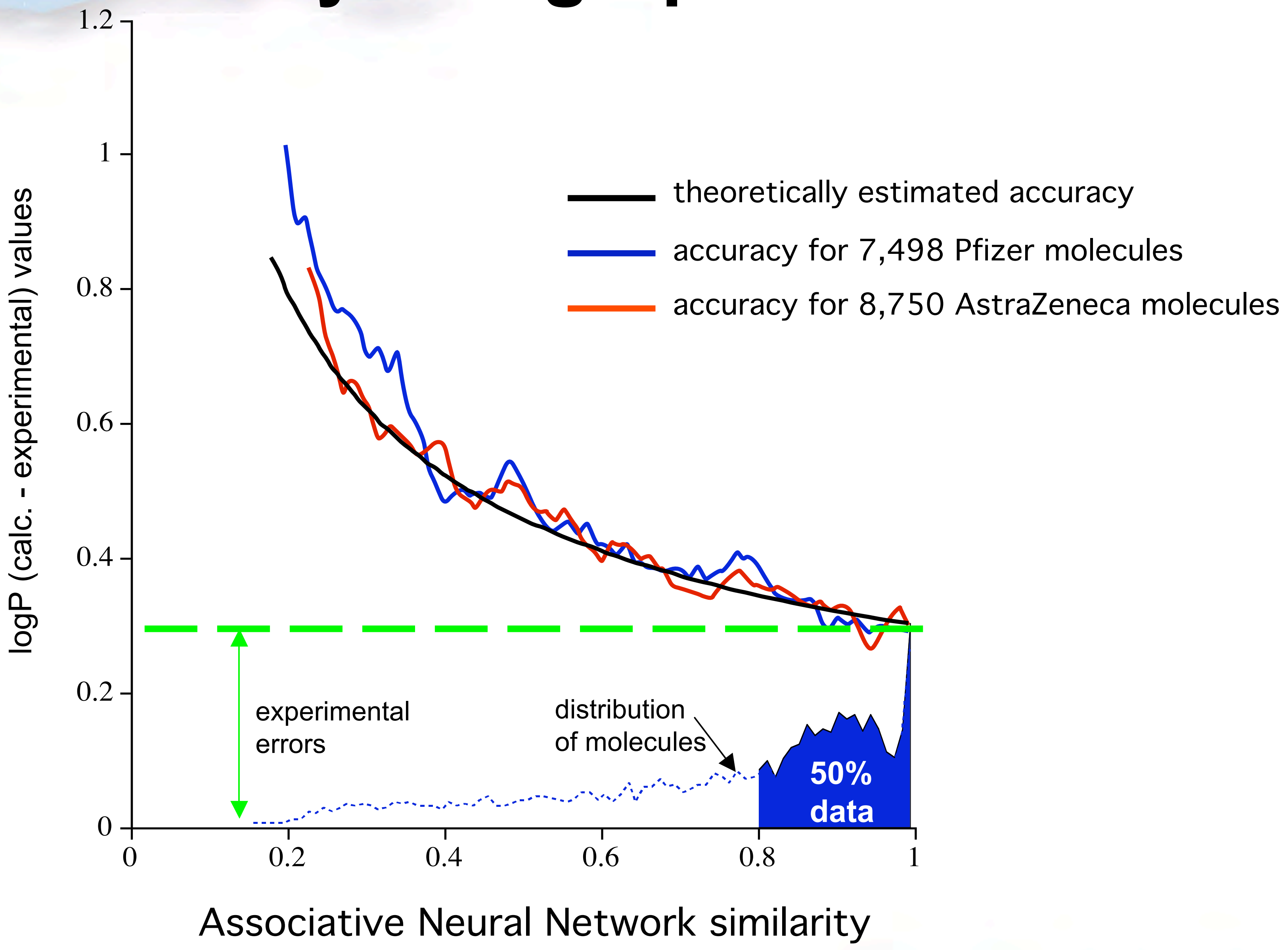
Challenges and solutions



The analysis in the property-based space allows estimation of the accuracy of predictions.

- ✓ Allows to estimate which compounds can/can’t be reliably predicted.
- ✓ Allows to develop targeted models to cover specific series.
- ✓ Allows an experimental design to minimize costs for new measurements.

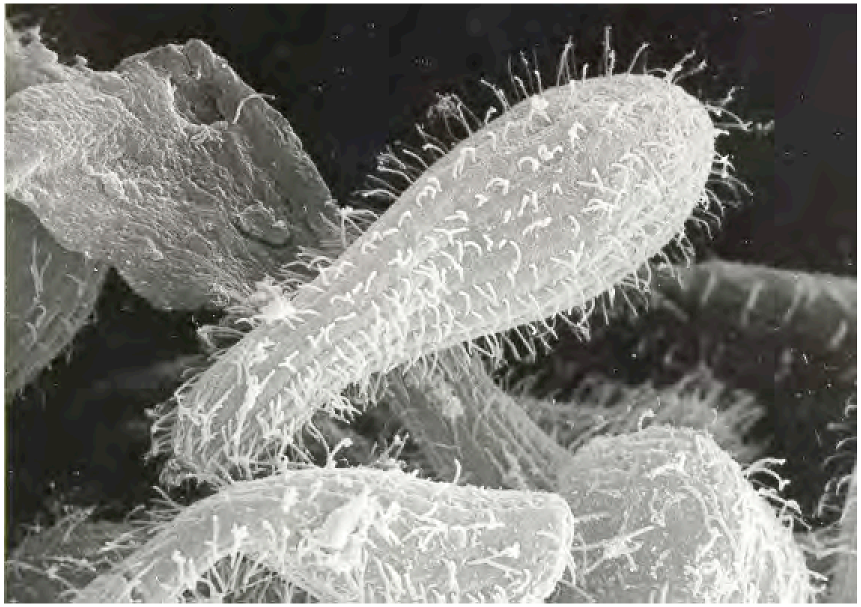
Accuracy of logP prediction



Tetko, I.V. et al, Drug Discovery Today, 2006.

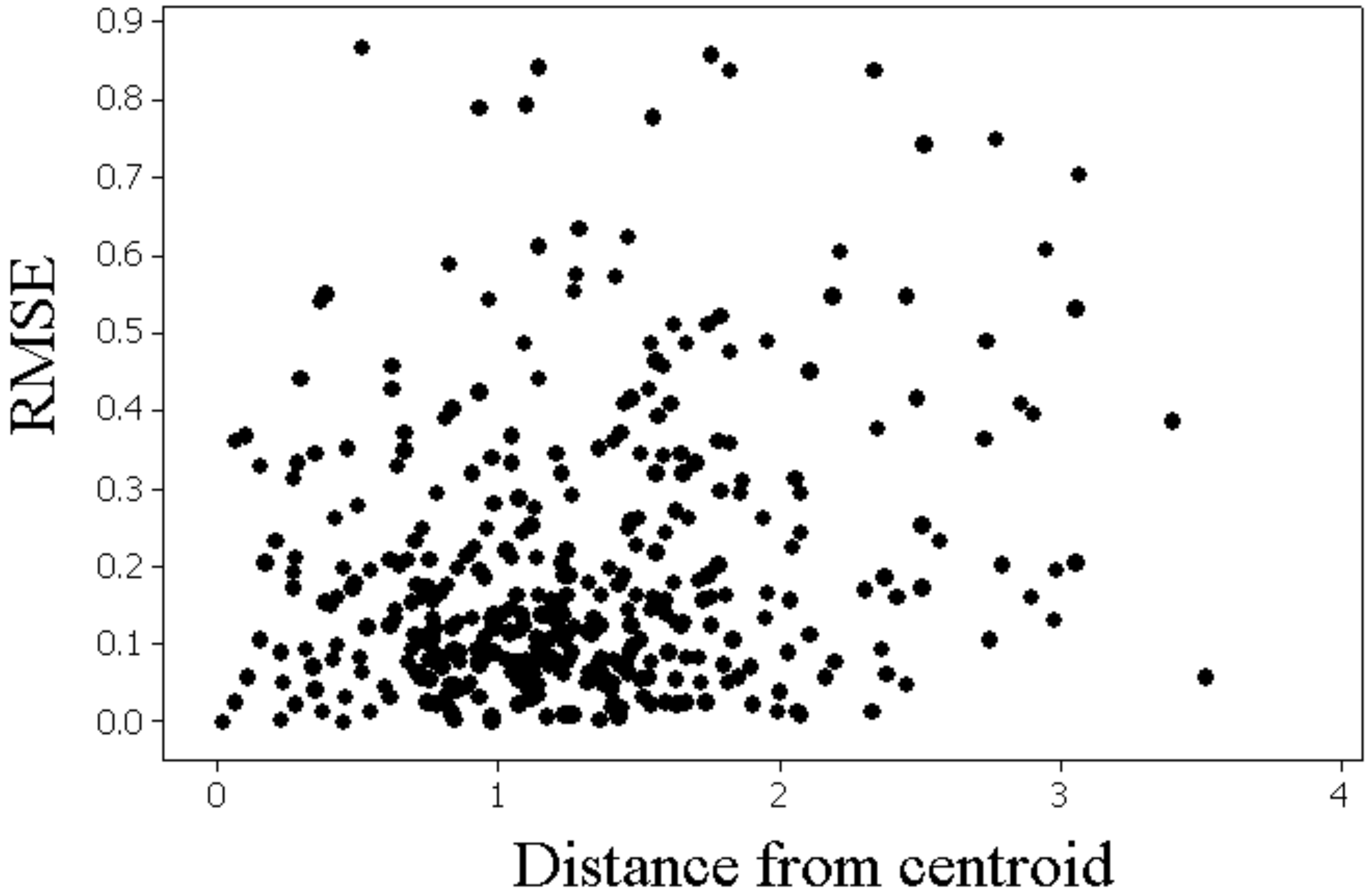
Estimation toxicity of *T. pyriformis*

- Toxicity of 384 aromatic compounds to *Tetrahymena pyriformis*
- *Model organism to estimate toxicity*
- *Very good data (measured in one lab during ca 20 years)*
- *Log/(IGC50-1)= 0.54logP+16.2*A_{max}*-5.9*
- *What is the applicability domain of the model?*



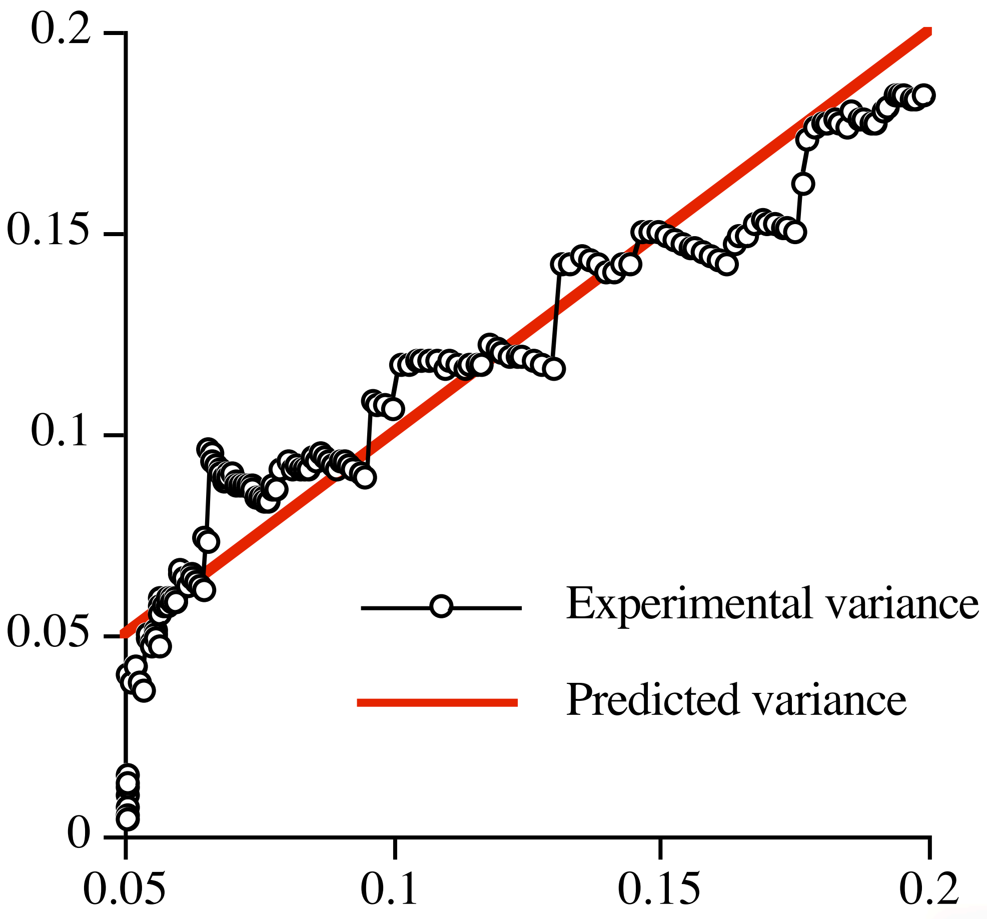
Schultz et al, QSAR Comb Sci, in press.

No Relationship Between RMSE and Distance from Descriptor Space Centroid



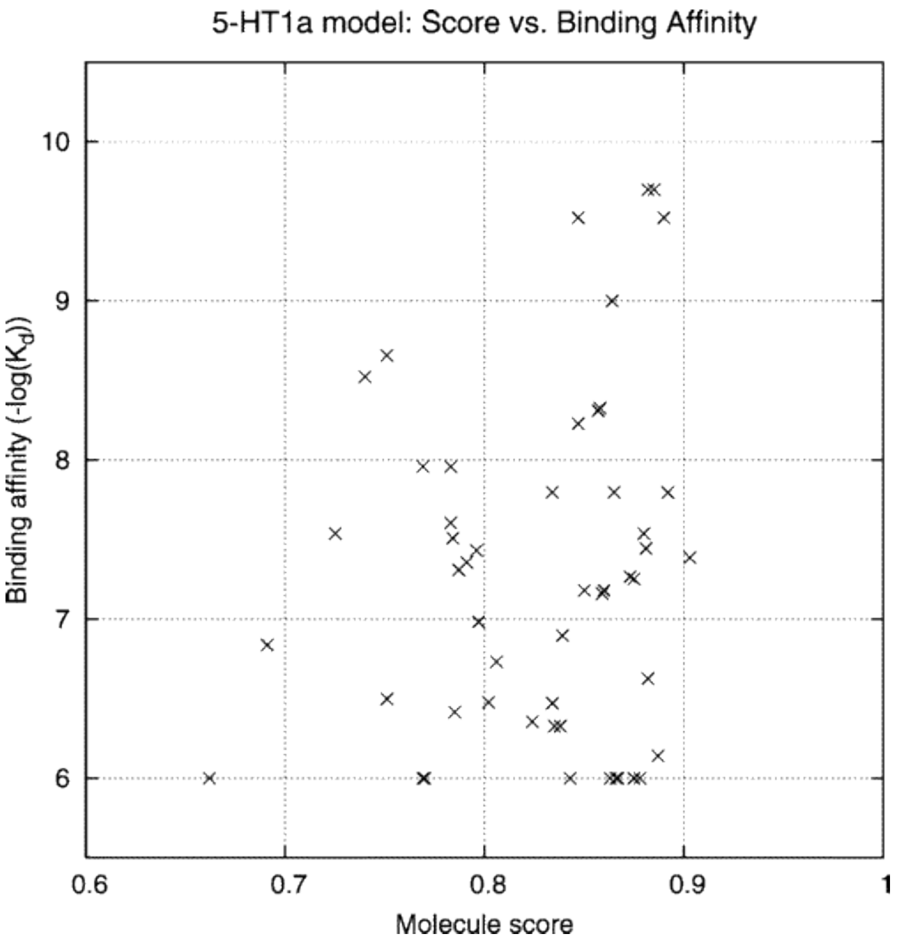
Schultz et al, QSAR Comb Sci, in press.

Experimental vs predicted error for *T. pyriformis*



What is about the biological activity?

- Current screening models usually provide only qualitative activity prediction
- Multiple models can be built and used to estimate quantitatively binding activity (in case if 1-20 molecules with activities are available)
- The descriptors can be, e.g. molecular imprints (Cleves & Jain, J. Med. Chem., 2006)
- The descriptors should be, of course, relevant to the problem!



Jain, A.N., *J. Med. Chem.*, 2004.

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!