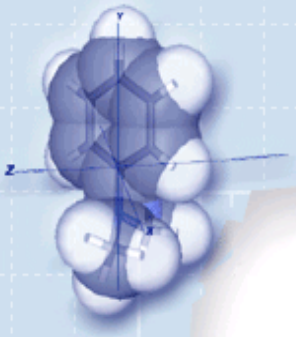


How quality of ADMET property prediction may affect early drug discovery process

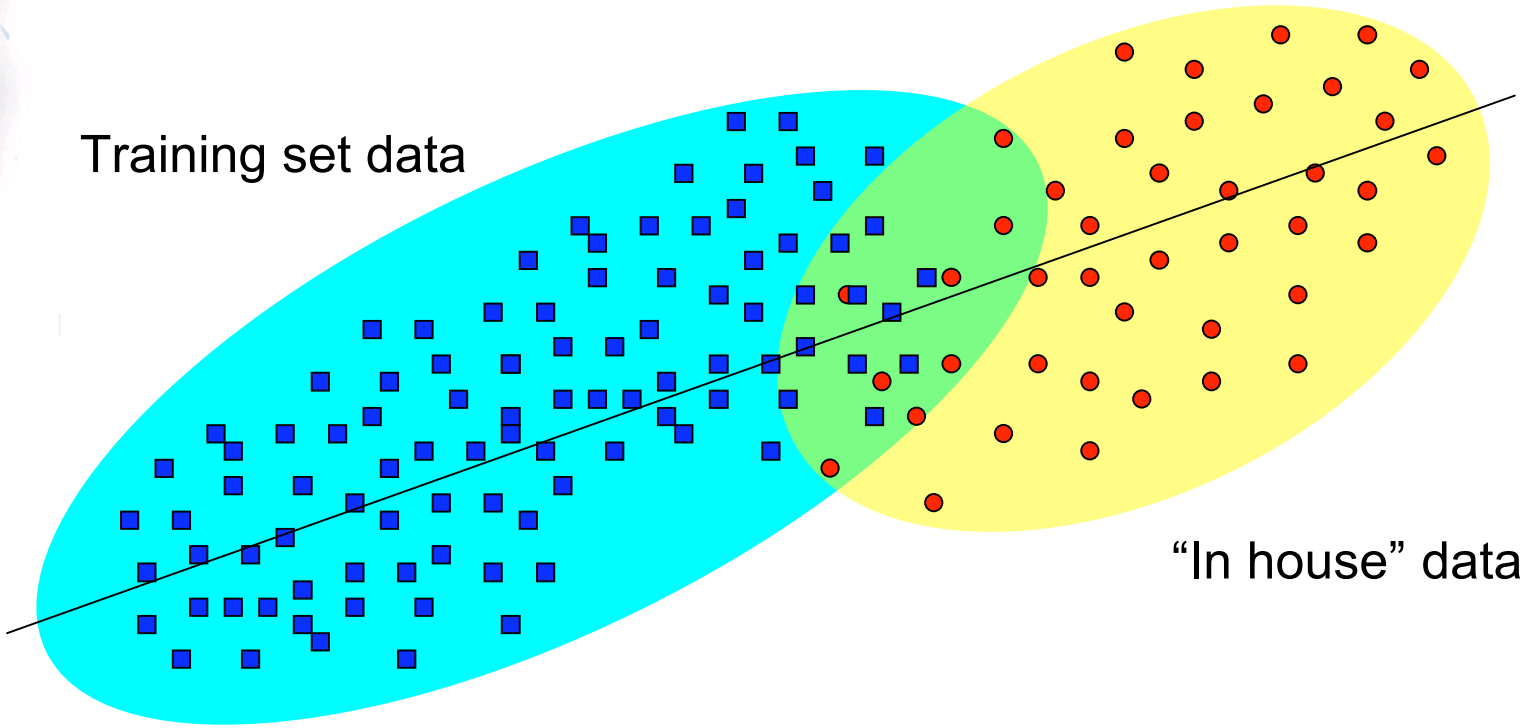
Igor V. Tetko

GSF -- Institute for Bioinformatics (MIPS), Neuherberg, Germany and

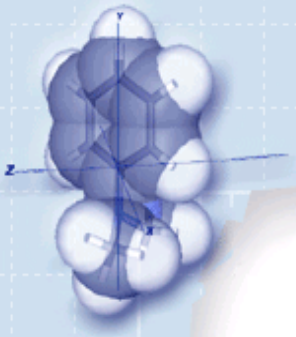
Institute of Bioorganic & Petrochemistry, Kyiv, Ukraine

May 12th, QSAR2006, Lyon, France

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



= Applicability domain



Applicability Domain Methods

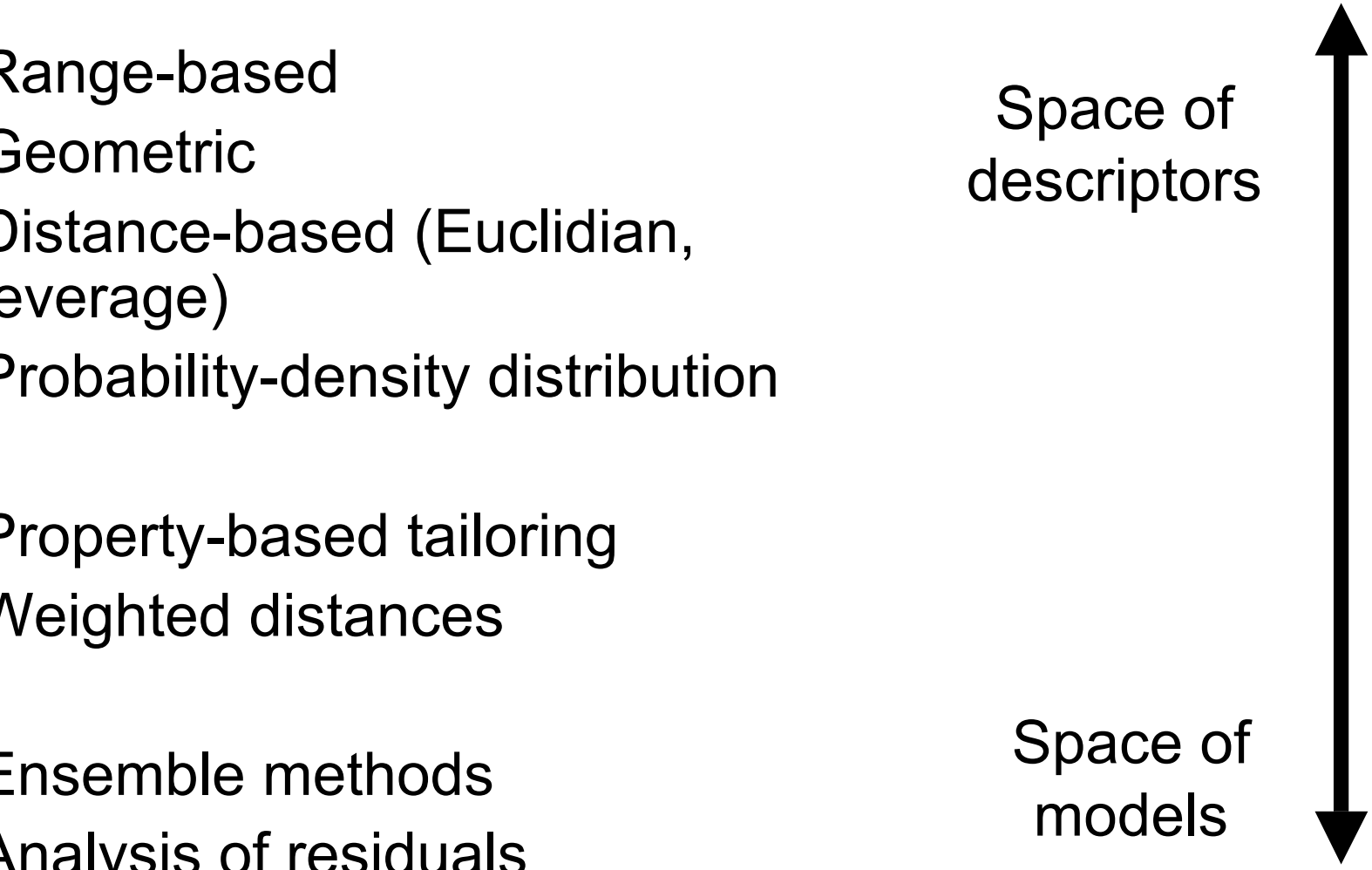
- Range-based
- Geometric
- Distance-based (Euclidian, leverage)
- Probability-density distribution

Space of descriptors

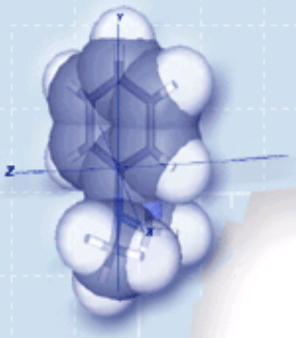
- Property-based tailoring
- Weighted distances

- Ensemble methods
- Analysis of residuals

Space of models



Netzeva et al, ATLA, 2005, 33(2), 155-173.



Why property-based space?

In space of descriptors:

- Detection of correct neighborhood relations depends on selection, pre-processing (e.g., PCA) and normalization of descriptors
- Dependencies in the input space are static and do not change with analyzed properties

But...

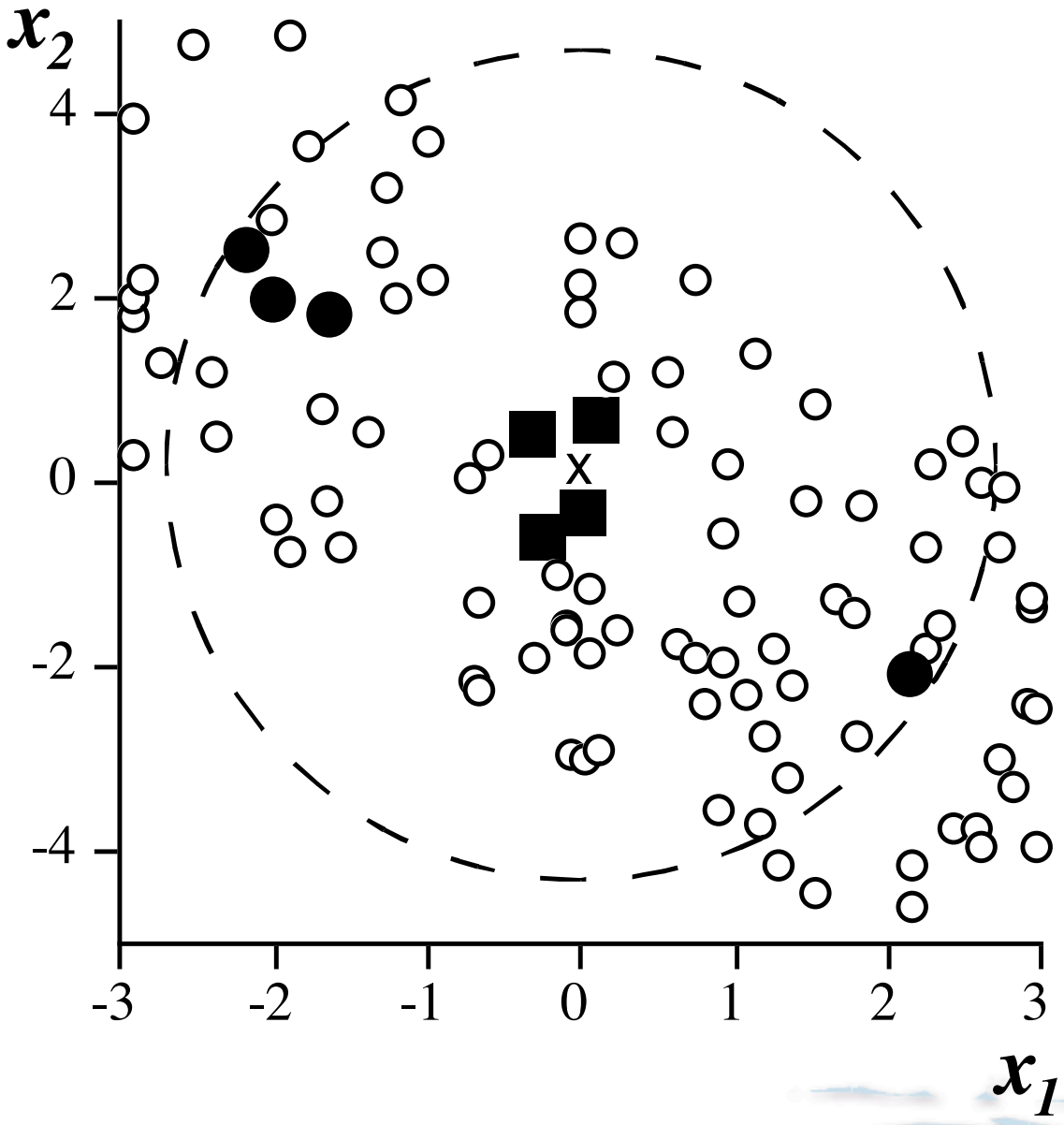
- Supervised learning method select the best combination of descriptors
- Provide their normalization (and non-linear transformations)

Thus

- We should profit from the supervised methods and use the supervised models to define the molecular similarity, **the property-based molecular similarity.**

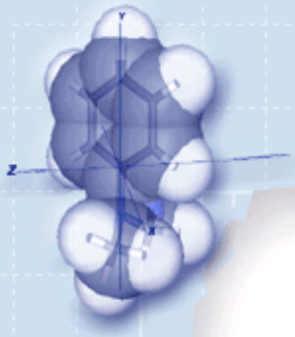
gsf

Nearest neighbors in the input space

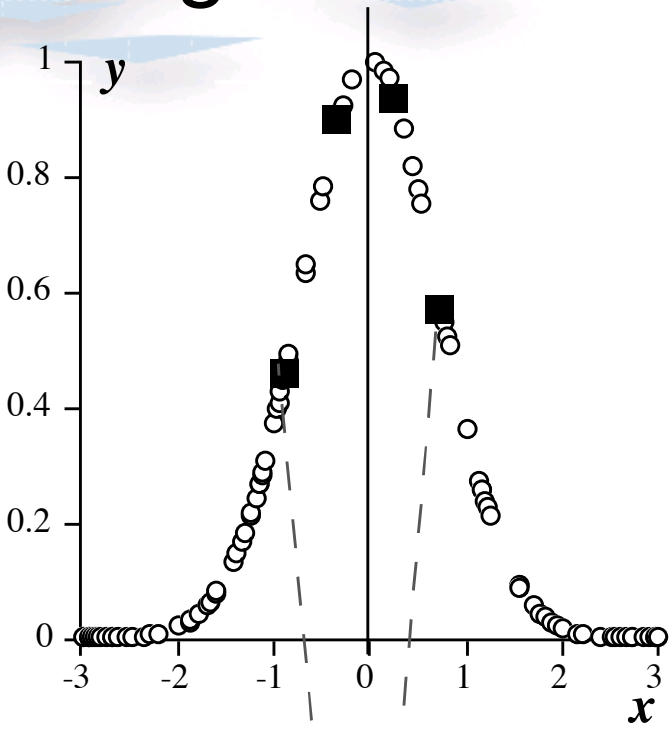


gsf

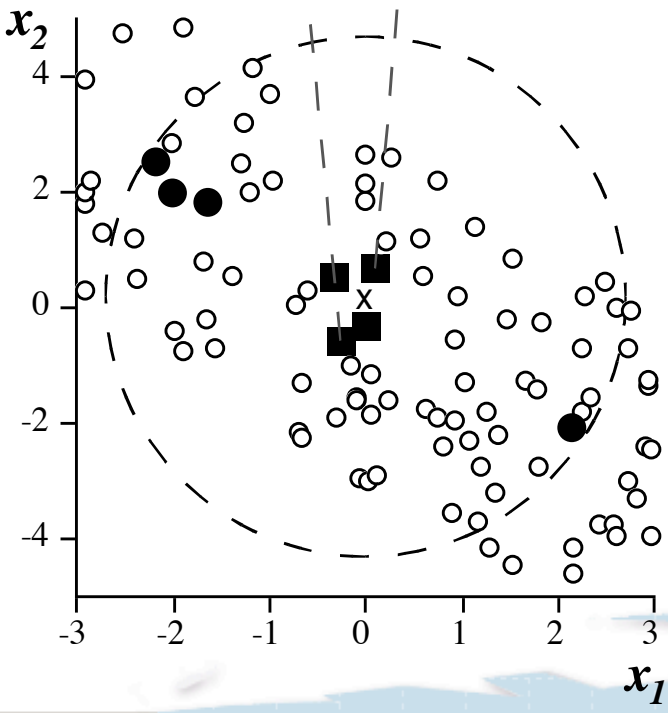
Nearest neighbors and activity



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



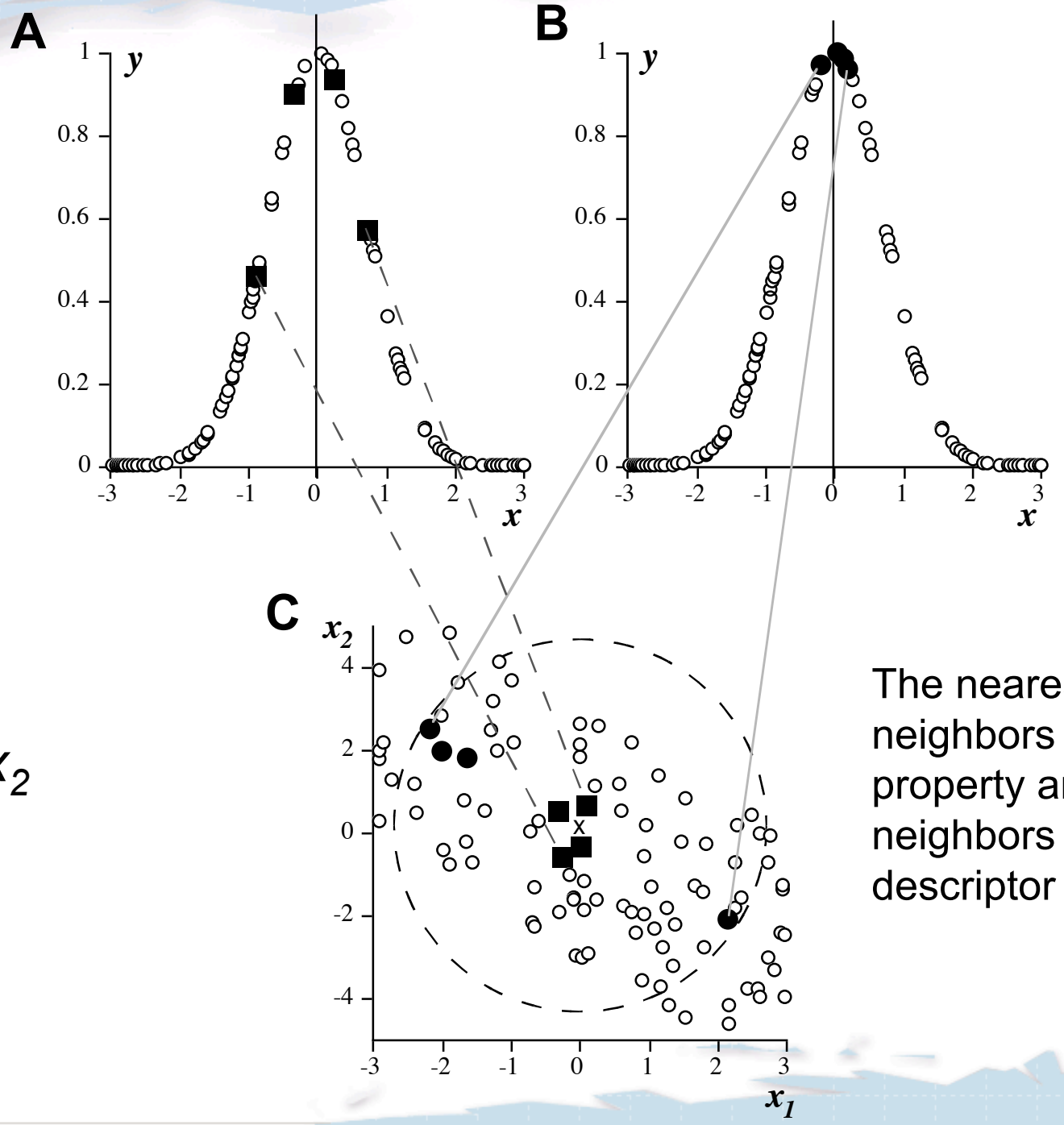
$x=x_1+x_2$!



The nearest neighbors in descriptor space are not neighbors in property!

gsf

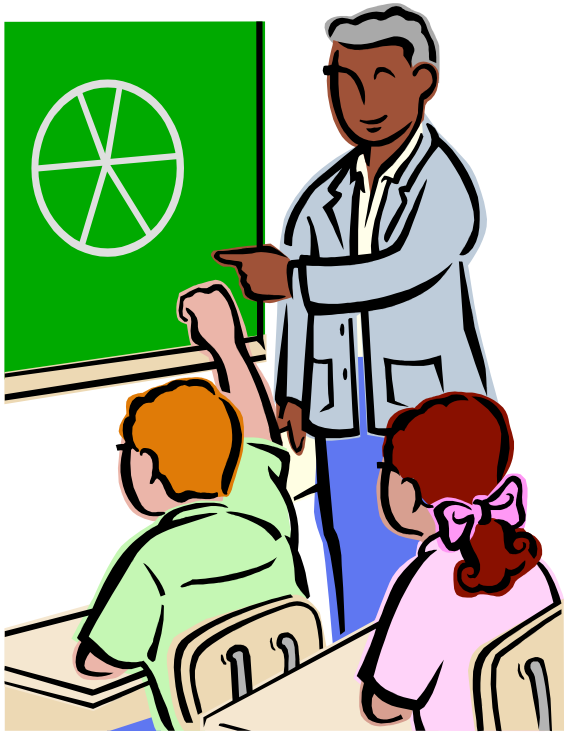
Nearest neighbors and activity



$x=x_1+x_2$

The nearest neighbors in property are not neighbors in descriptor space!

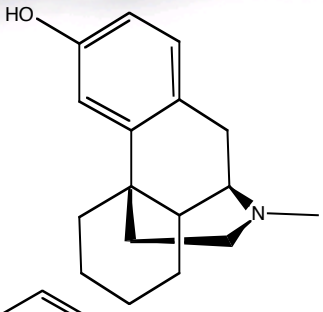
Ensemble methods



Hansen, L.K.; Salamon, P. *IEEE Trans. Pattern. Anal. Mach. Learn.,* 1990, 12, 993.
Tetko, I. V.; Luik, A. I.; Poda, G. I. *J. Med. Chem*., 1993, 36, 811.
Tetko, I.V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* 1995, *35(5)*, 826.

An example ensemble analysis

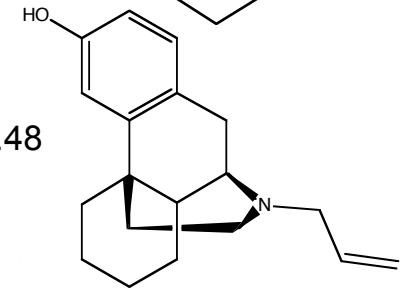
logP=3.11



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

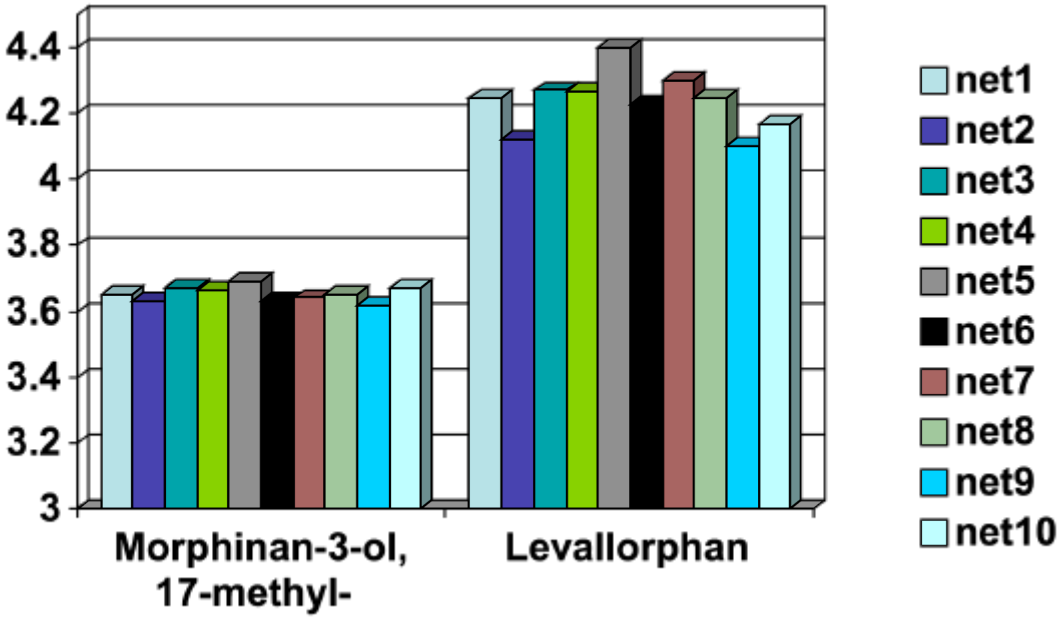
Morphinan-3-ol, 17-methyl-

logP=3.48



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

Levallorphan

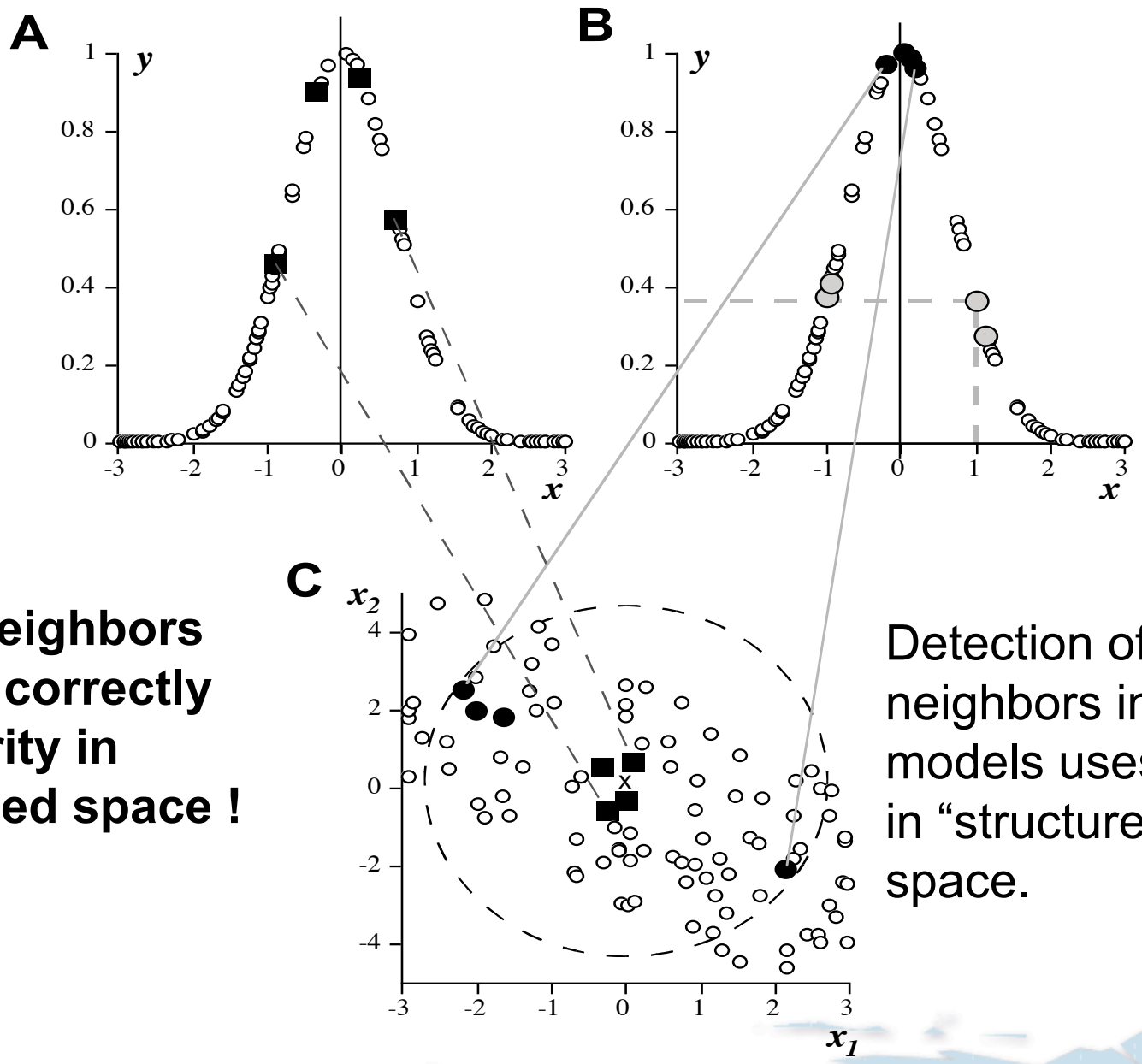


-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals amid >12,000 molecules!

R is the similarity in space of models.

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

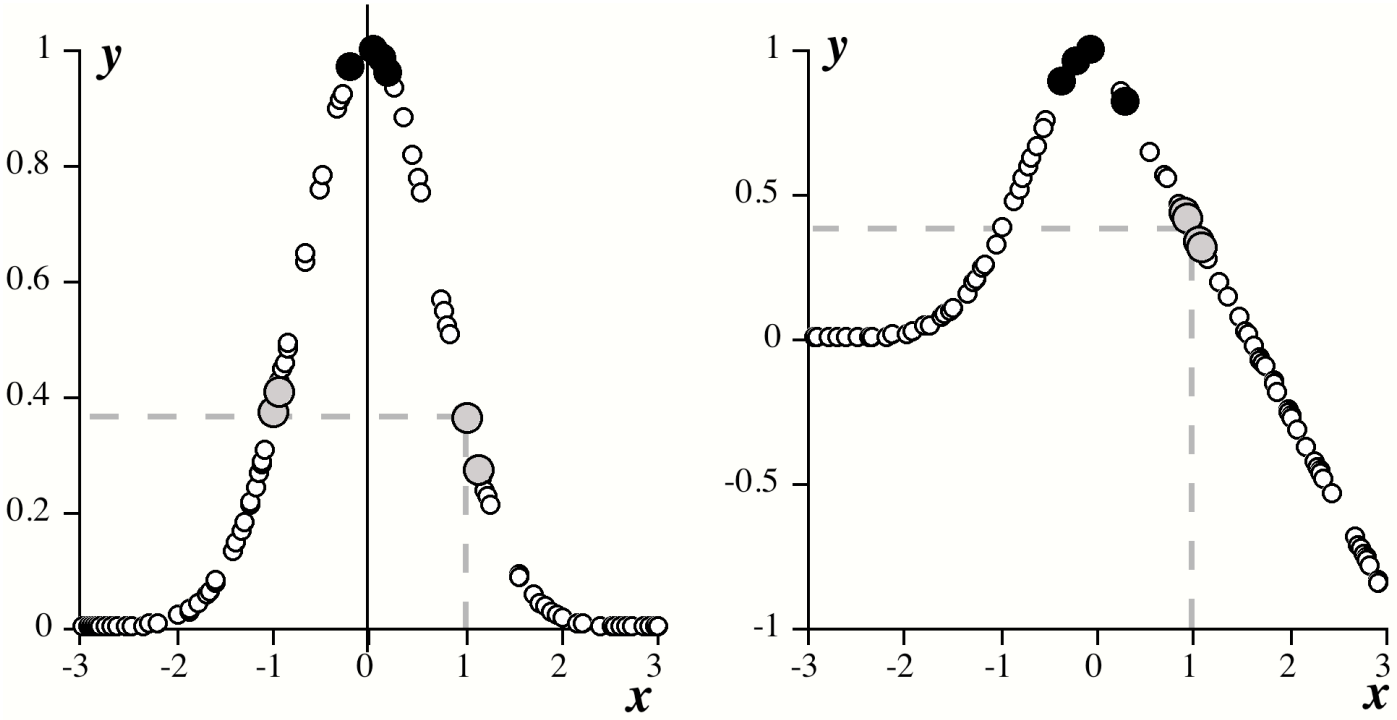
Nearest neighbors for Gauss function



All nearest neighbors are detected correctly using similarity in property-based space !

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

Nearest neighbors for different functions



Molecules with the same response values are not necessary nearest neighbors in the property-based space!

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***

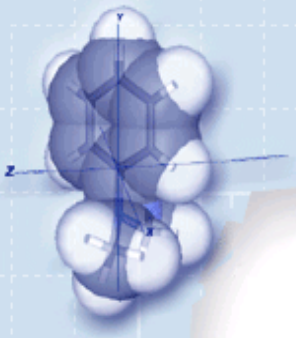
$$\bar{z}_i' = \bar{z}_i + \frac{1}{k} \sum_{j \in N_k(\mathbf{x}_i)} (y_j - \bar{z}_j)$$

<<= ASNN bias correction

ASNN is actually kNN in space of model residuals => allows instance learning of new data via kNN (LIBRARY mode or memory correction).

1) Tetko, I.V. *JCICS*, **2002**, 42, 717.
2) Tetko, I.V. *Neural Proc. Lett.*, **2002**, 16(2), 187-199.

gsf



ALOGPS 2.1

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

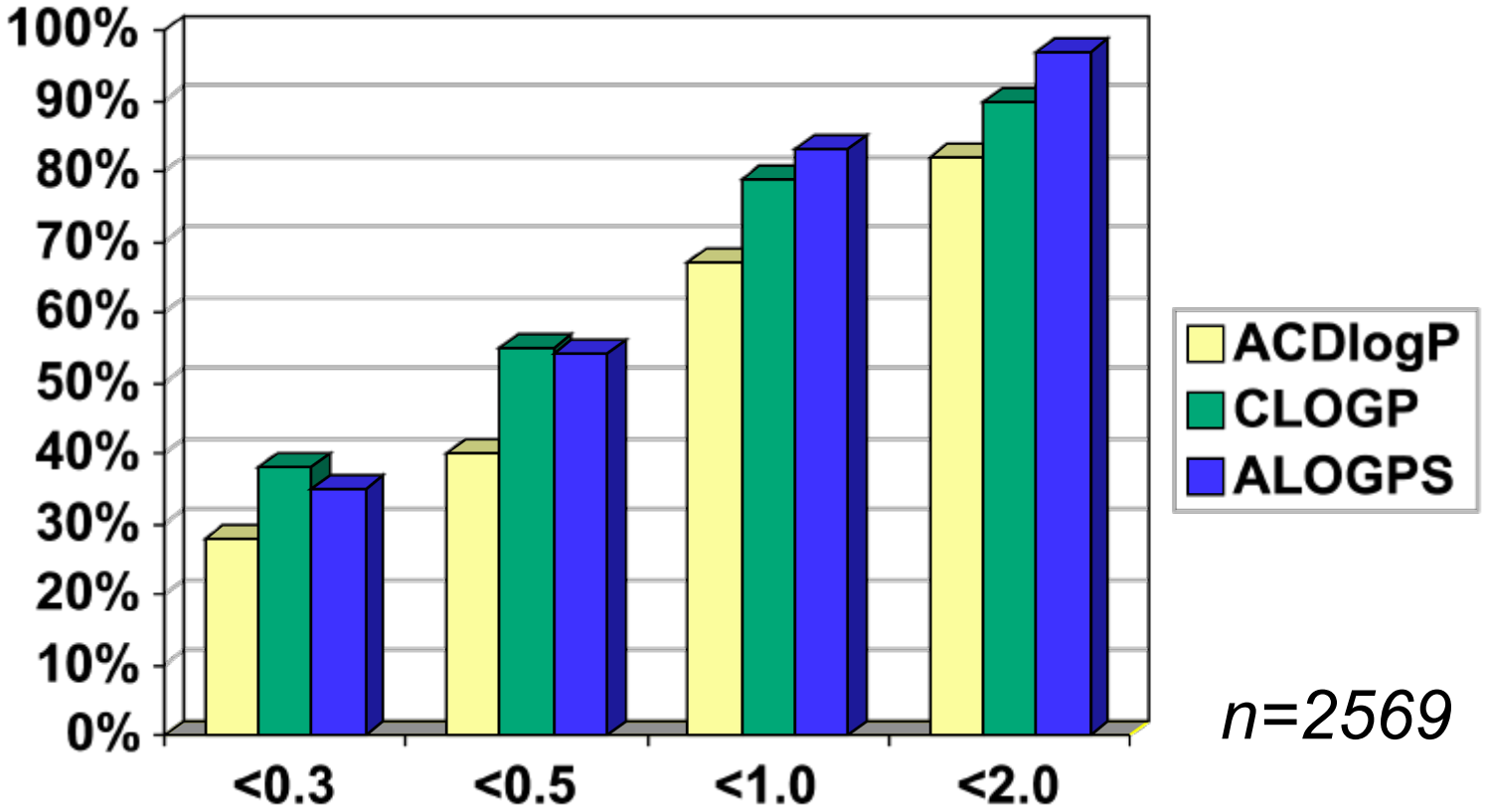
•LogS: 33 input E-state indices, 1291 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.49, MAE=0.35, n=18 outliers (>1.5 log units)

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

Available free at http://www.vcclab.org site.

gsf

Prediction of AstraZeneca logP set



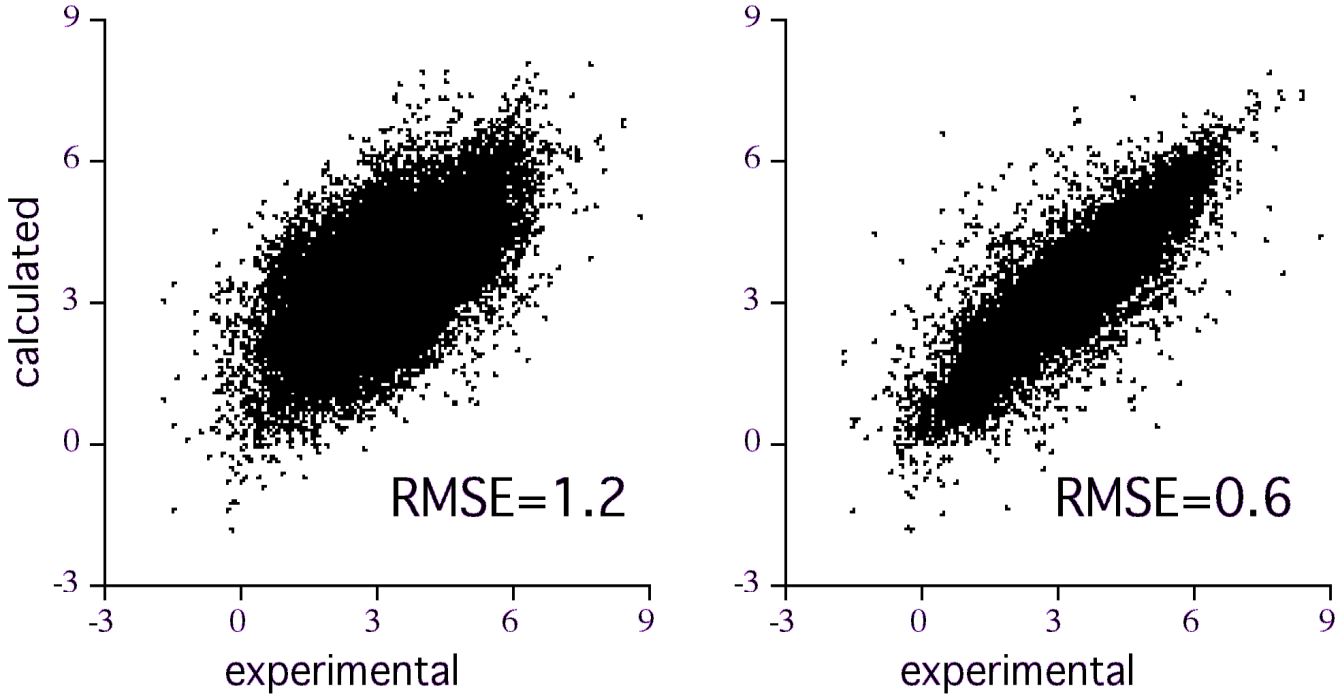
n=2569

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.

Analysis of Pfizer data

ALOGPS prediction for ElogD set of 17,861 compounds

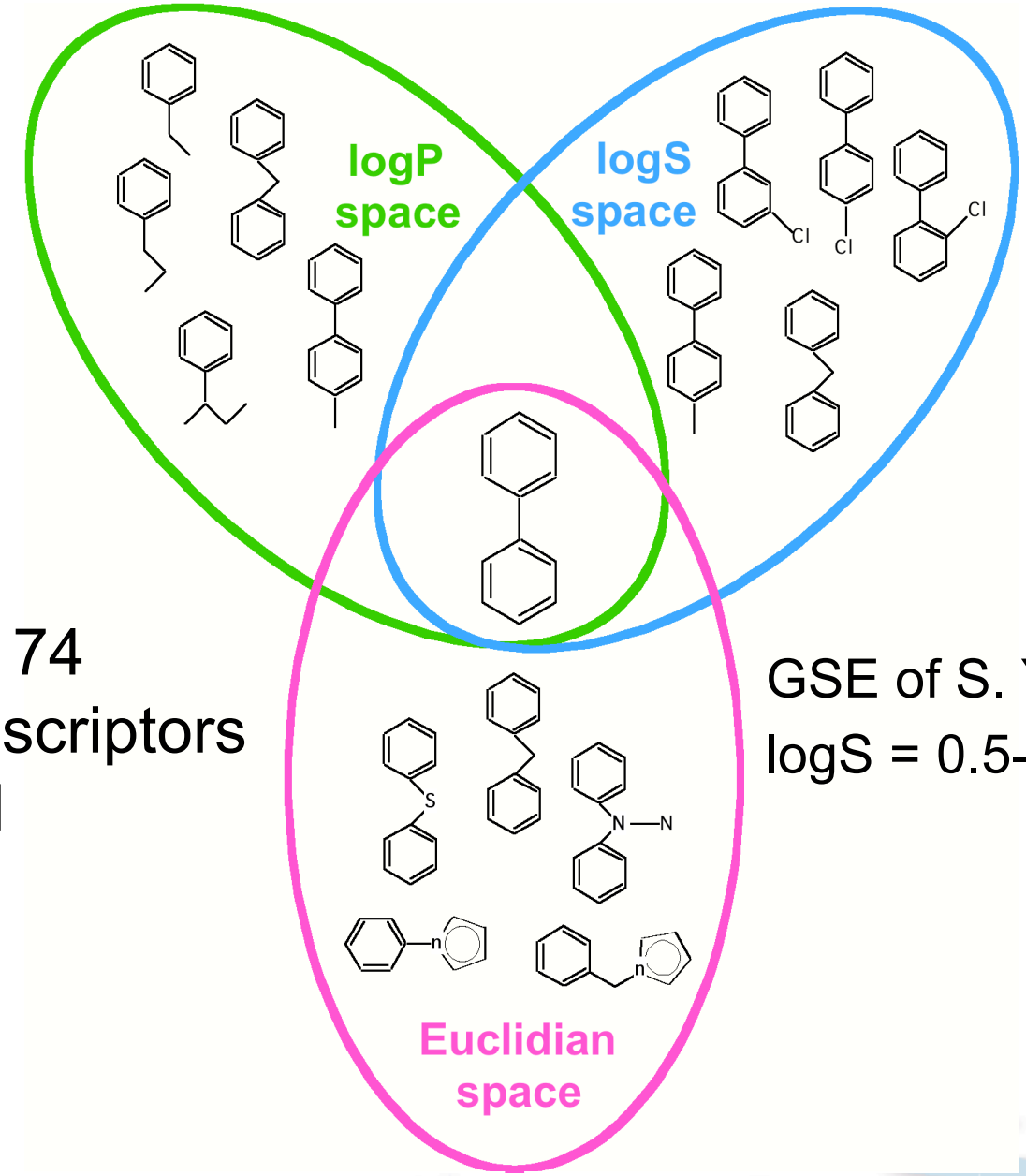


ALOGPS "as is" → ALOGPS LIBRARY

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.

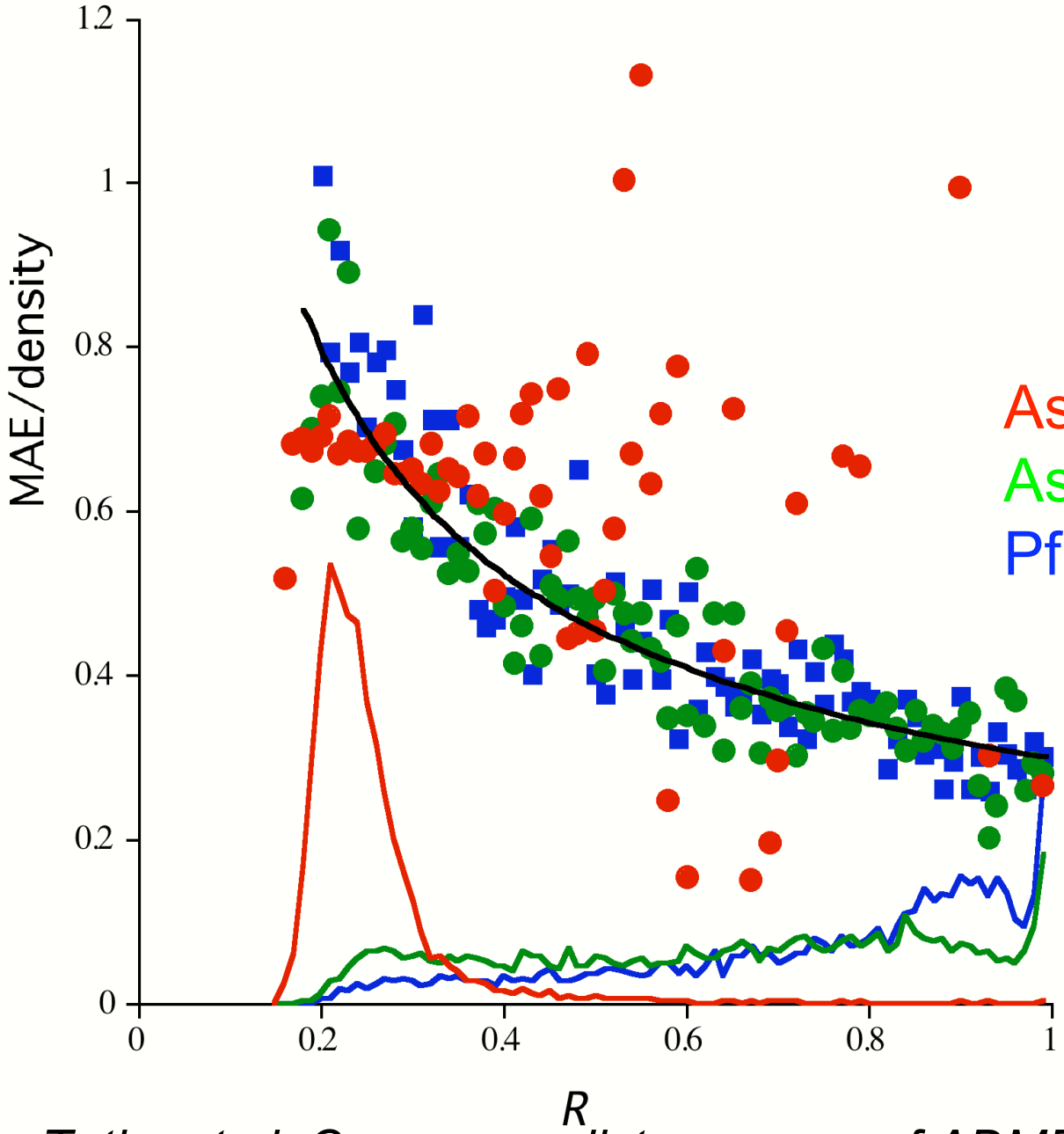
Nearest neighbors in different spaces



The same 74
E-state descriptors
were used

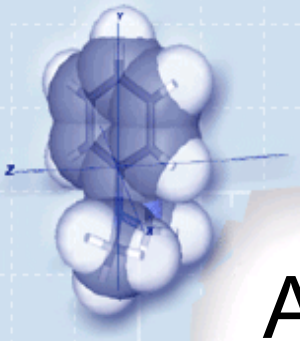
GSE of S. Yalkowsky
logS = 0.5-0.01(MP-25) - logP

Accuracy of logP prediction as function of R



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

Tetko et al, Can we predict accuracy of ADMET? DDT, in press.

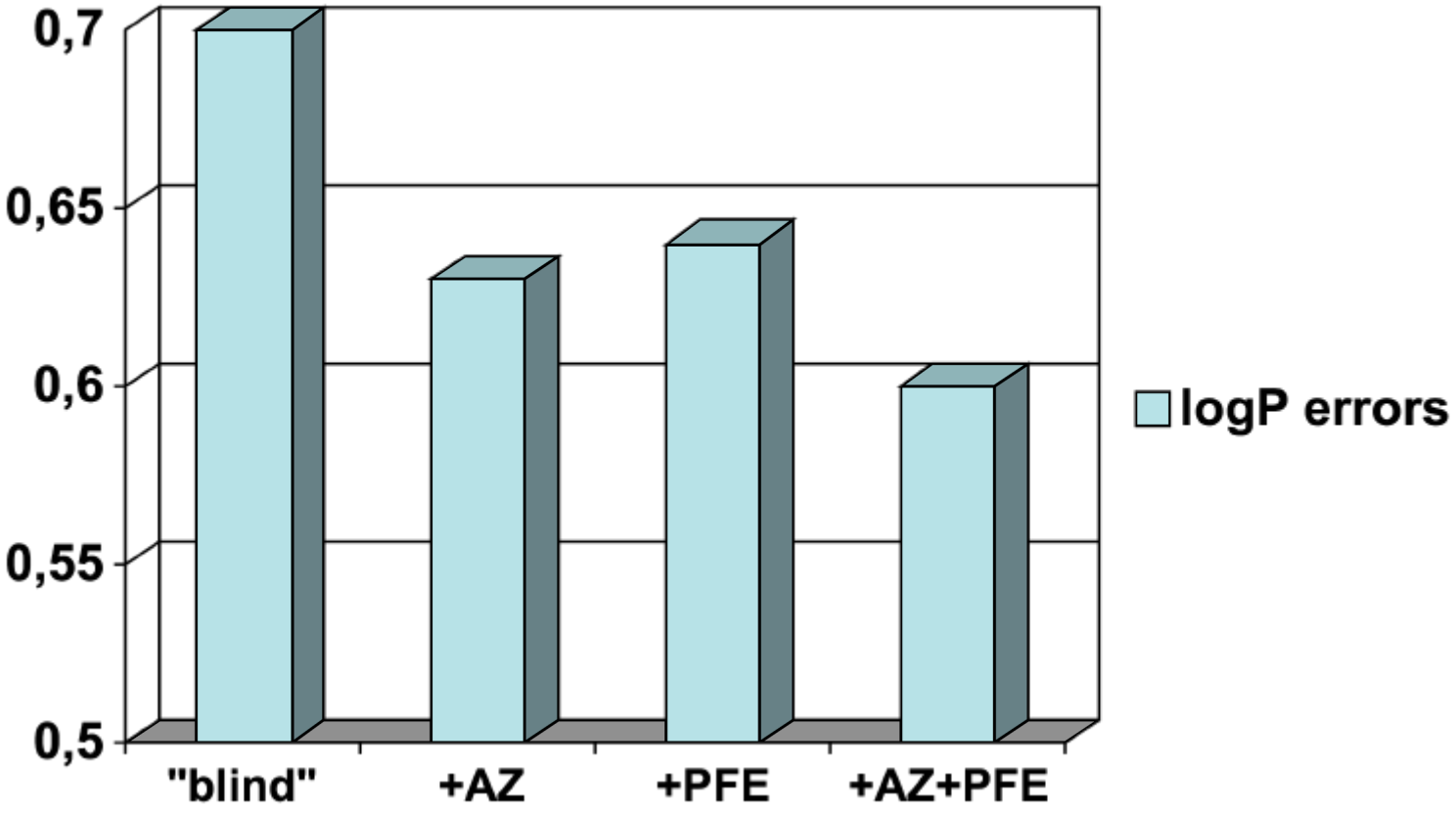


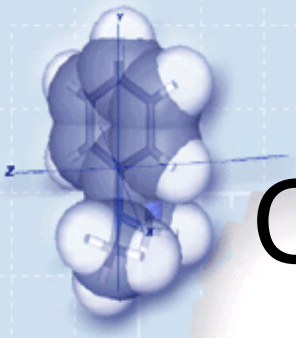
Estimated and calculated MAE for AstraZeneca (AZ) and Pfizer (PFE) sets

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

gsf

Estimated errors for >1,3*10^7 iResearchLibrary molecules

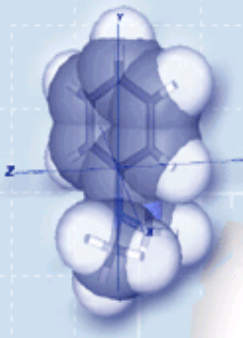




Changes using PFE LIBRARY

- >514,000 compounds having logP> 5 units --> logP<5
- 495,000 compounds changed |logP| > 1 log unit

gsf



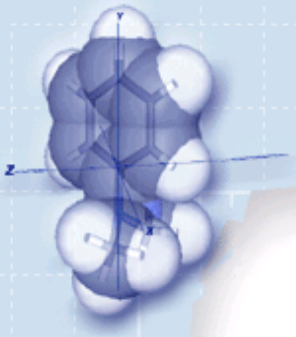
Similarity in property-based space

- is introduced as correlation between vector of residuals of models[1,2]
- is a heart of the Associative Neural Network method[2,3] used in the ALOGPS[2] and 1H NMR[7] prediction programs
- is specific for the target property[3,4]
- detects meaningful nearest neighbors, allows mechanistic interpretation[3,4]
- estimates accuracy of prediction (applicability domain) of programs[5]
- can be used for secure data sharing[6]
- methodology is used in logP LIBRARY builder of TRIDENT (Wavefunction Inc) and (will be) used in ADMET predictor of SimulationPlus Inc.*

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 …*, CRC, **2007**, in press.
5) Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D., Poda, G.I. *DDT*, **2006**, in press.
6) Tetko, I.V.; Abagyan, R.; Oprea, T.I. *J. Comp. Aid. Mol. Des*. **2005**, 19, 749.
7) Da Costa, F. B.; Binev, Y.; Gasteiger, J.; *Tetrahedron Letters* **2004,** 45, (37), 6931.

gsf

*-personal communication from Dr. R. Fraczkiewicz



Acknowledgement

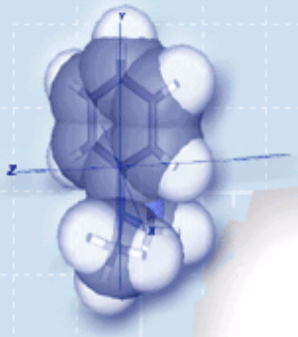
Part of this work was done thanks to
Virtual Computational Chemistry Laboratory
INTAS-INFO 00-0363 project

I thank Pierre Bruneau (AstraZeneca), Gennadiy Poda (Pfizer), Douglas Rohrer (Pfizer), and Hans-Werner Mewes (IBI, GSF) for collaboration in this work and Dr. Scott Hutton for providing compounds from the iResearch Library (ChemNavigator).

Thank you for your attention!

gsf

Free (use/download) at http://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

gsf

Do not remember? Just google “log p” or “logp”

Early Stopping Over Ensemble (ESE)