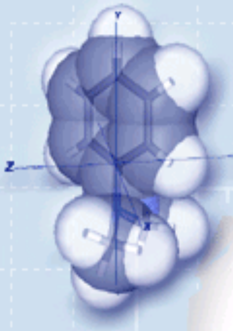


Estimation of the accuracy of ADMET predictions and secure sharing of information are two sides of the same coin

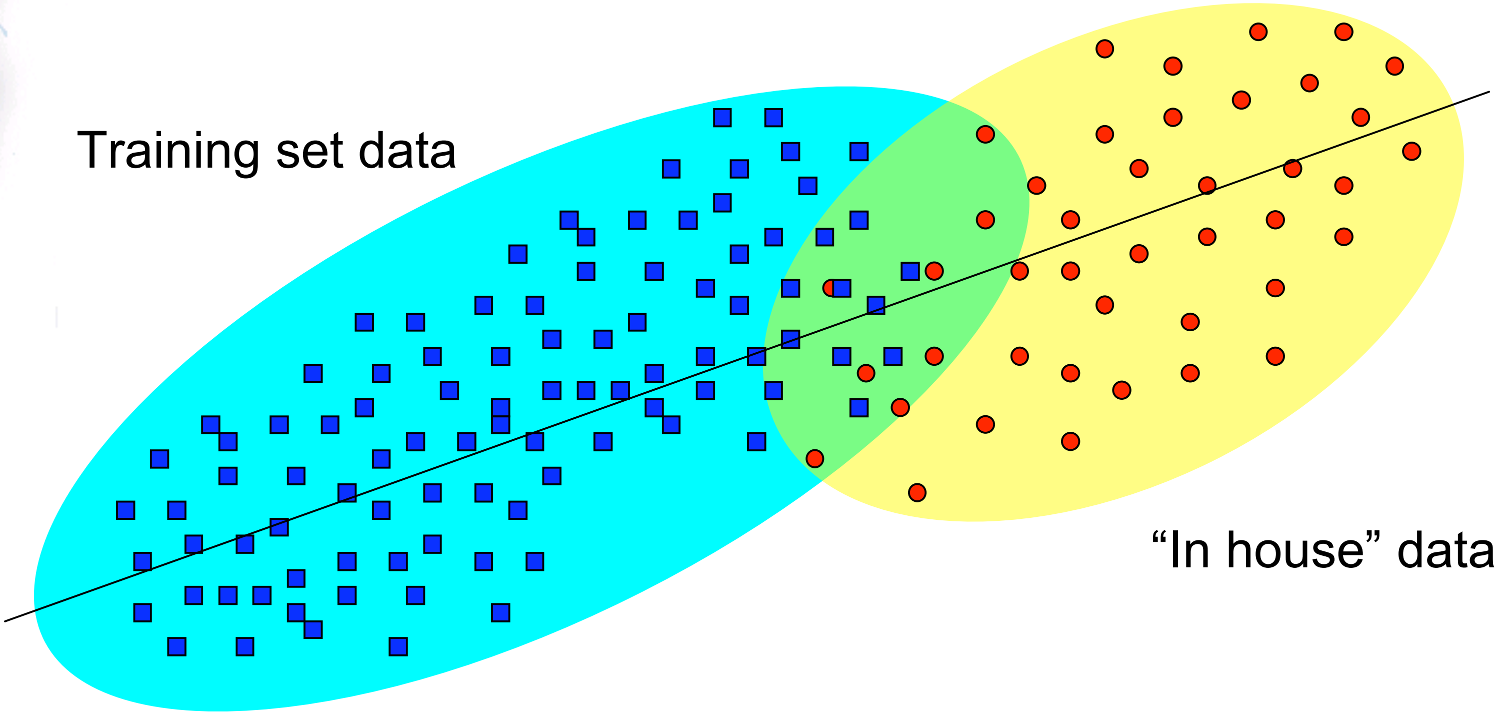
Igor V. Tetko

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

Institute of Bioorganic & Petrochemistry, Kyiv, Ukraine

30 May 2006, Chemoinformatics in Europe, Obernai, France

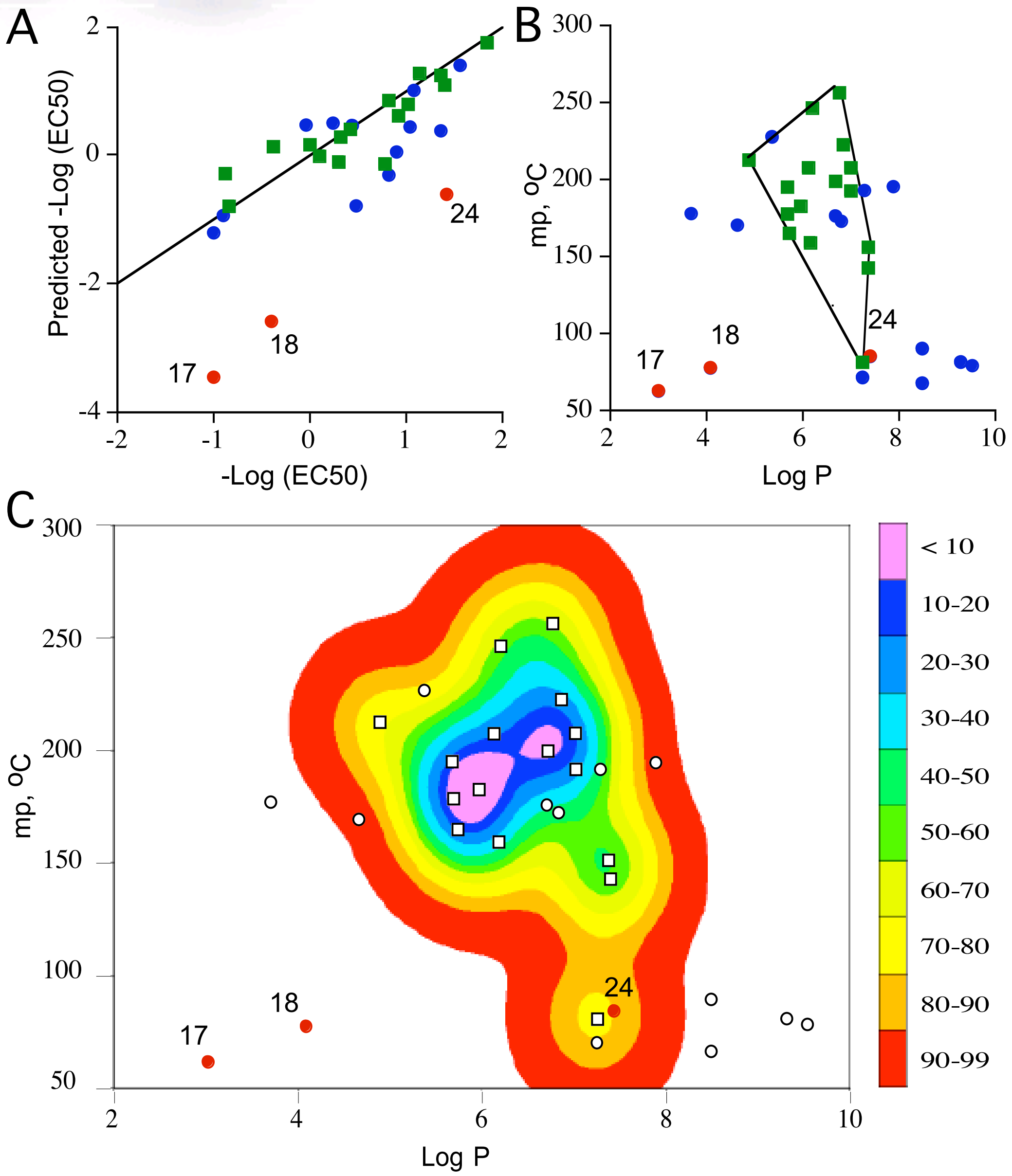
Prediction Space of the model does not cover the “in house” compounds



= Applicability domain

QSAR of antifilarial antimycin analogues*

in vitro activity
$-\log(EC_{50}) = 0.016$ mp + 0.56 logP - 6.14



*Selwood et al, 1990, *J. Med. Chem*, 33, 136.
Tetko et al, *DDT*, 2006, in press

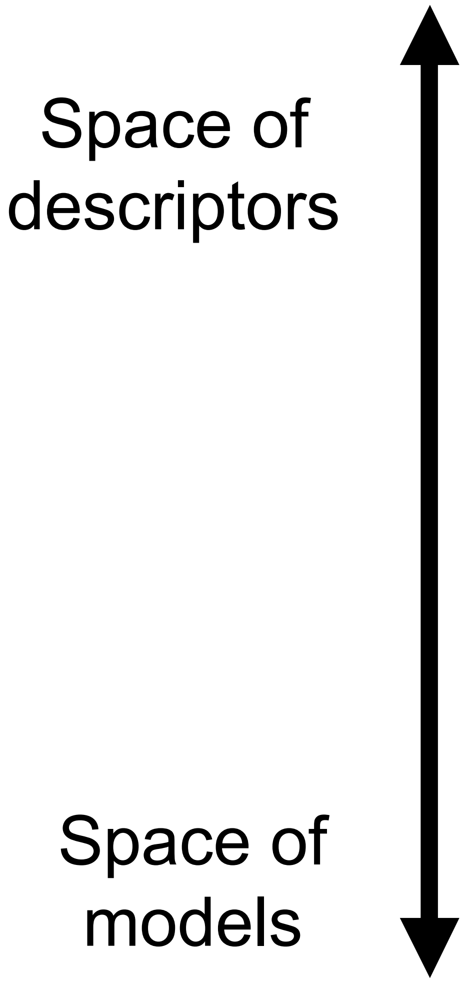
gsf

Applicability Domain Methods

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

- Property-based tailoring
- Weighted distances

- Ensemble methods
- Analysis of residuals


Space of descriptors

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

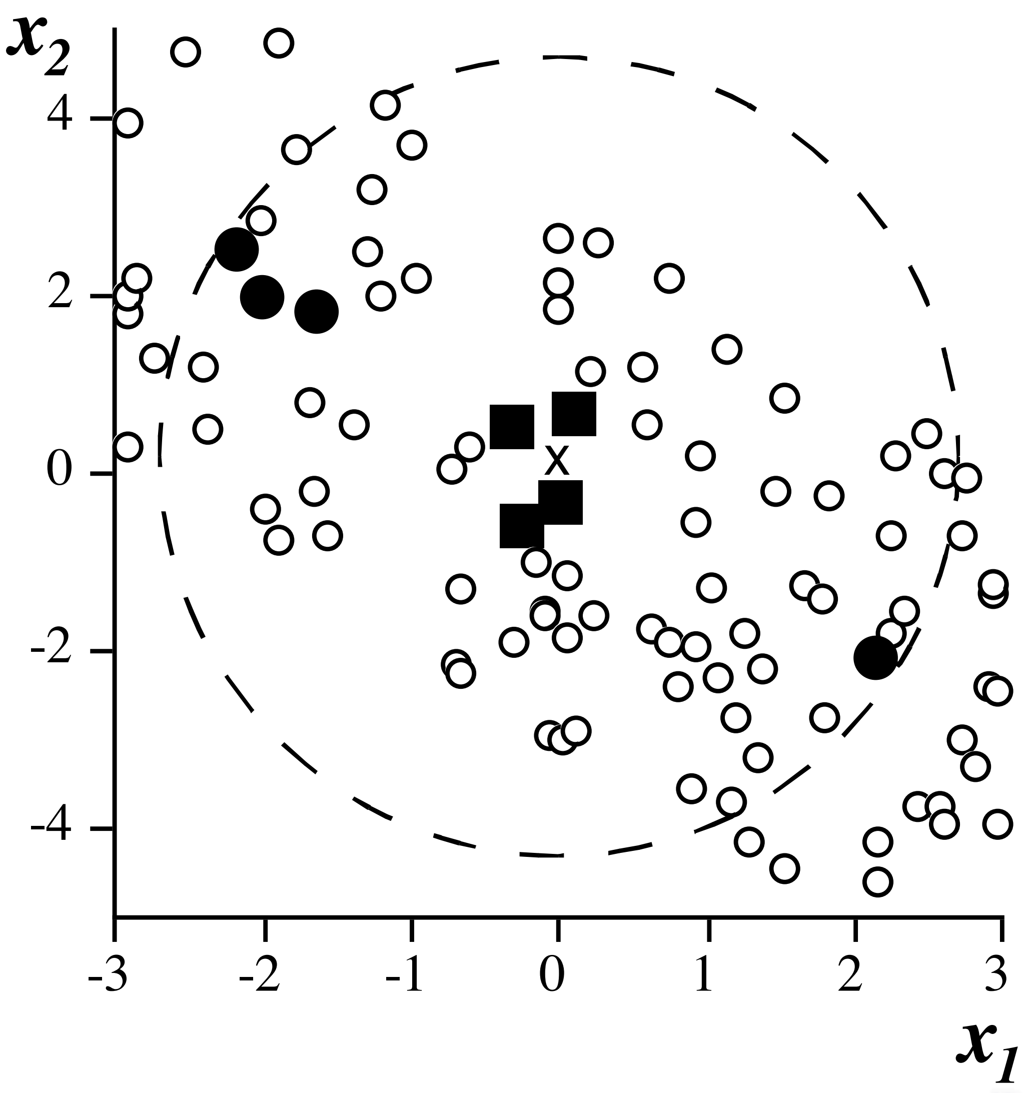
But...

- 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.**

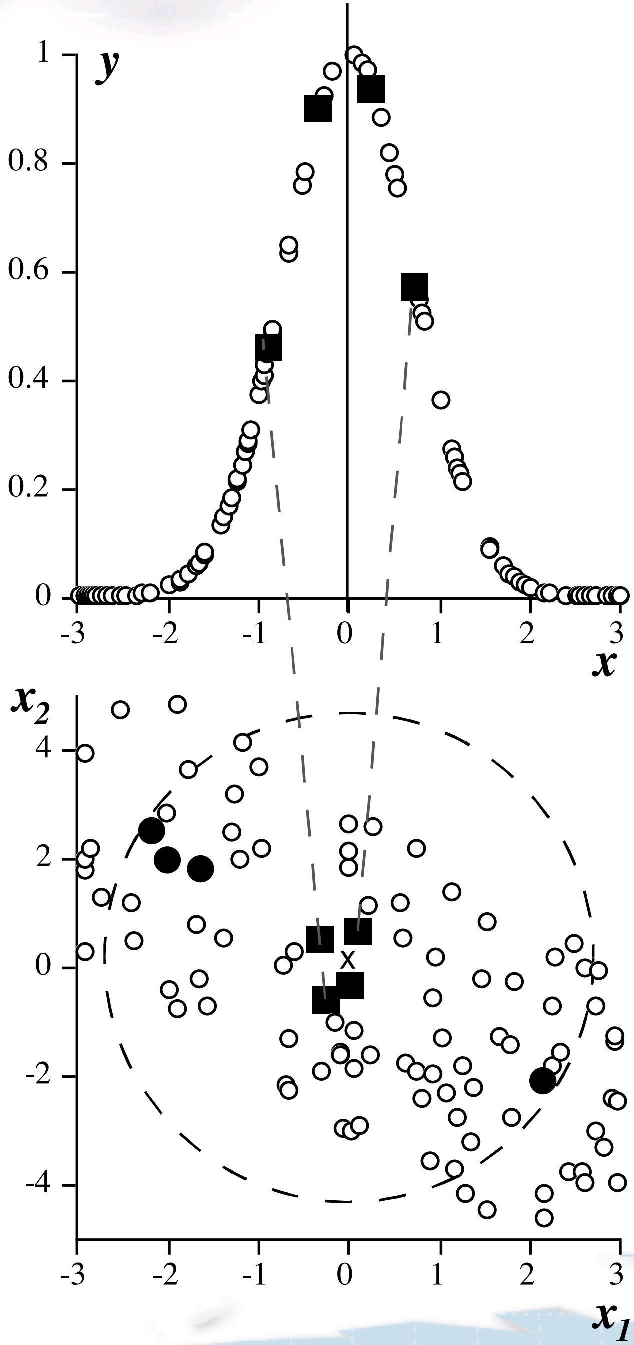
Nearest neighbors in the input space



Nearest neighbors and activity

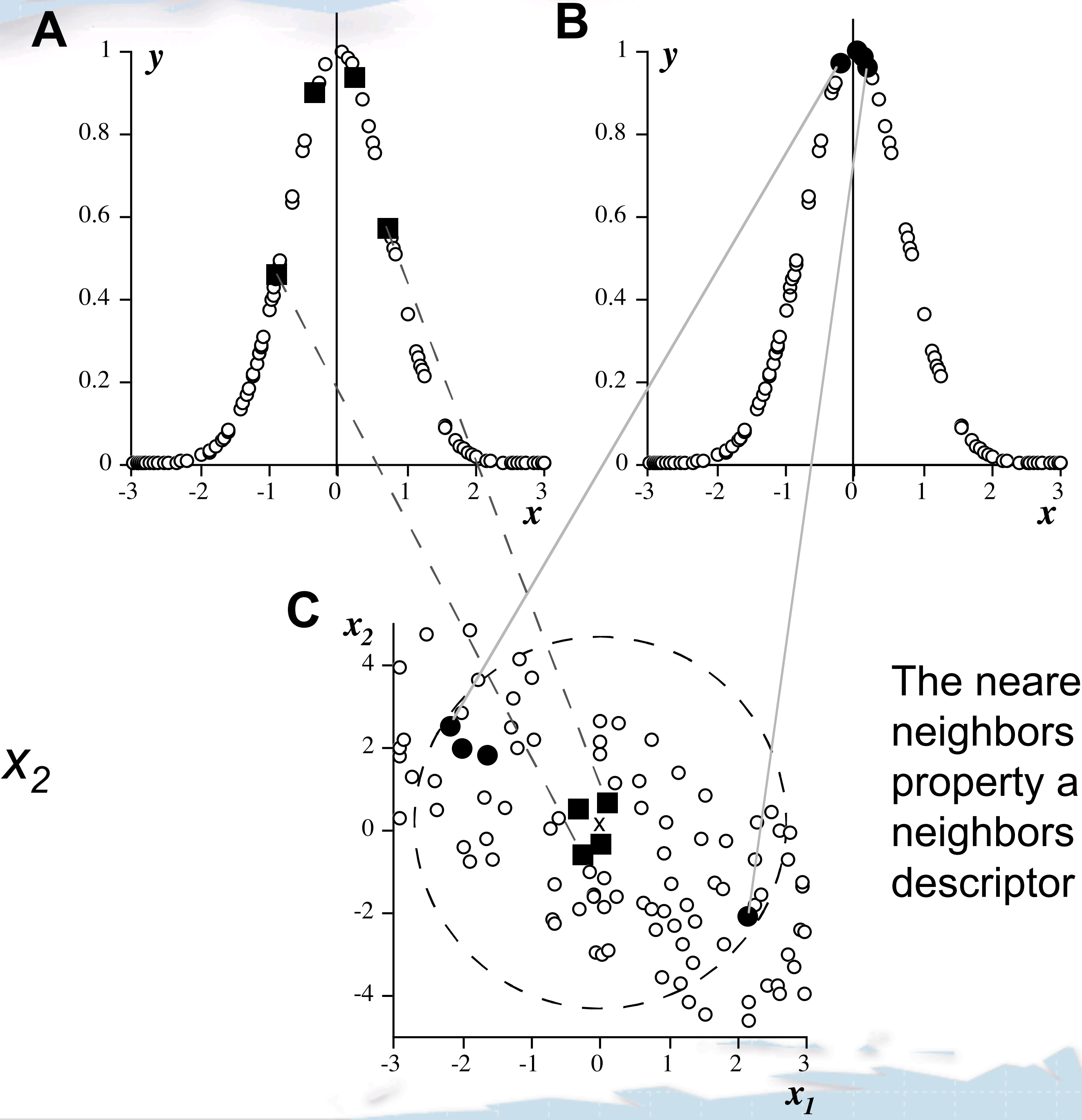
$y=exp(-(x_1+x_2)^2)$

$x=x_1+x_2$!



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

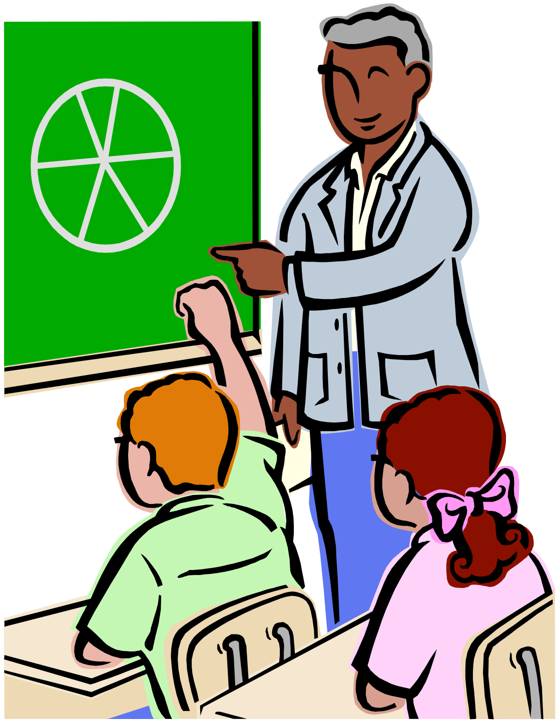
Nearest neighbors and activity



$x=x_1+x_2$

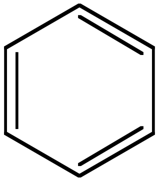
The nearest neighbors in property are not neighbors in descriptor space!

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.

Encoding of a molecule as a rank of models

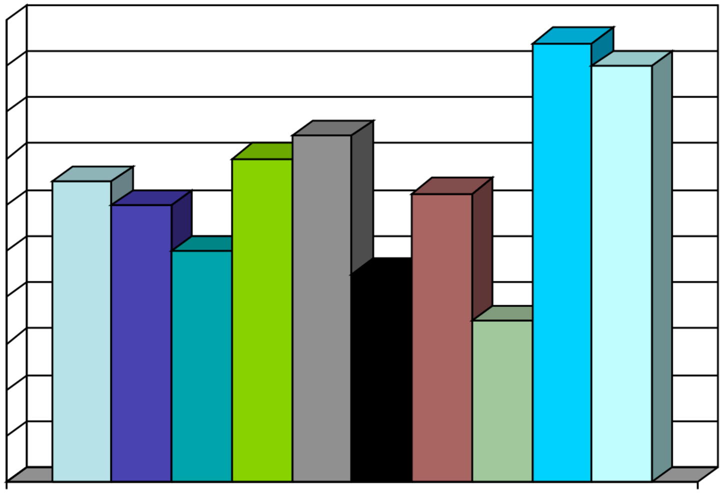


--> C1=CC=CC=C1

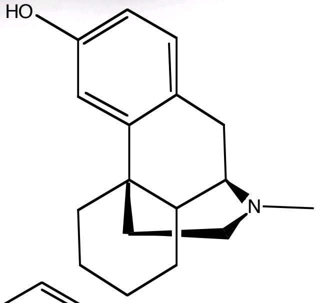
--> 3D, E-state descriptors

--> + property

0.89	0.88	0.86	0.90	0.91	0.85	.885	0.83	0.95	0.94
5	7	8	4	3	9	6	10	1	2



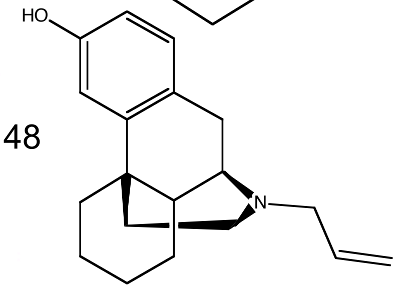
An example of an ensemble analysis



logP=3.11

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

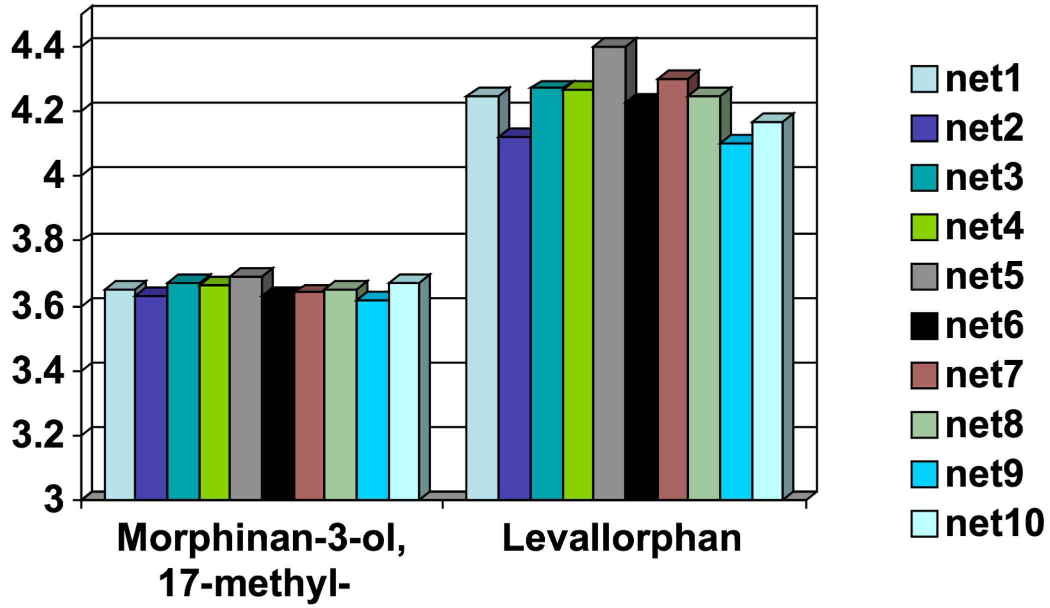
Morphinan-3-ol, 17-methyl-



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

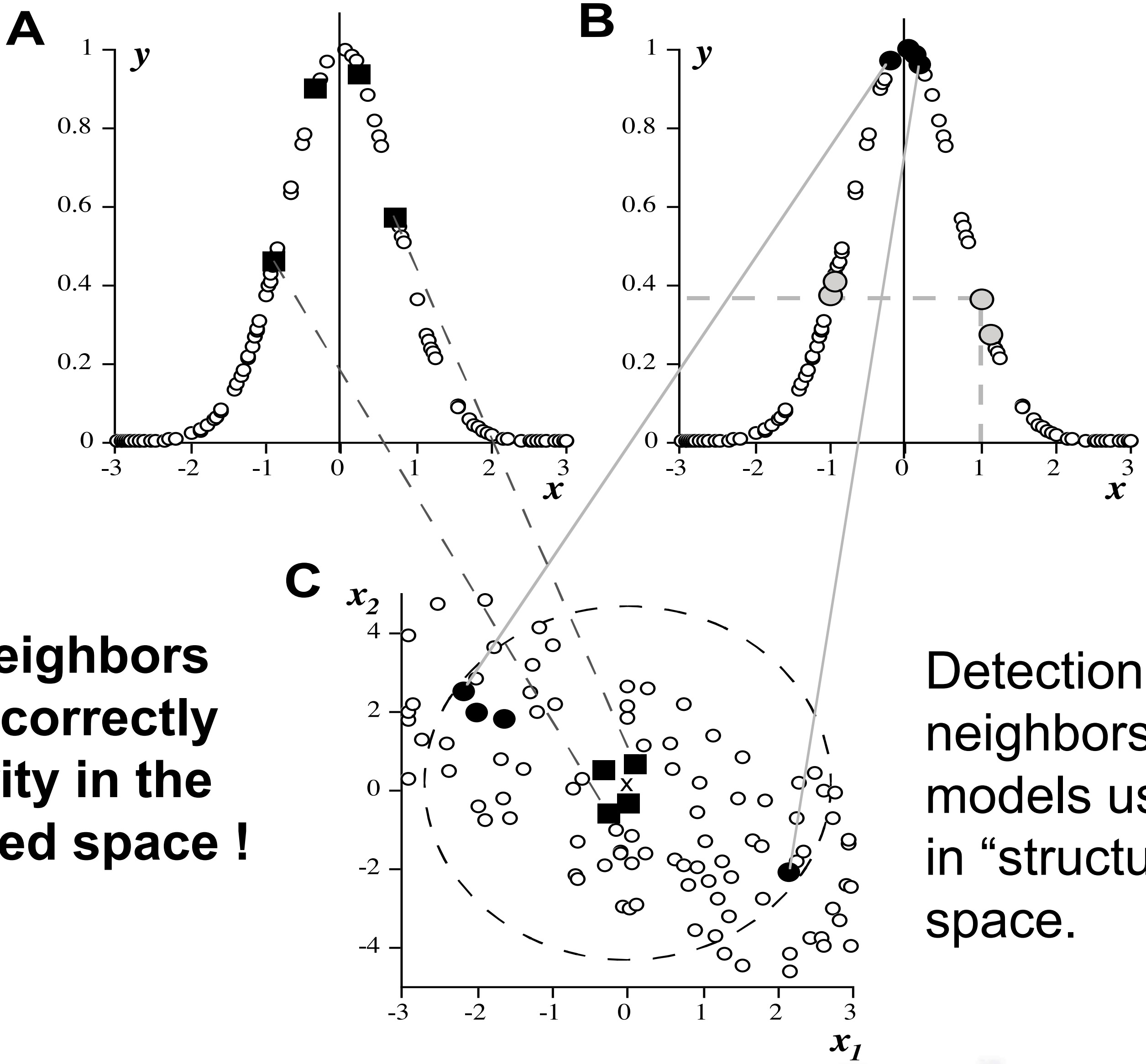


-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals amid >12,000 molecules!

Rank correlation of models residuals defines the property based similarity of molecules.

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 the property-based space !

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

Tetko, I.V. *JCICS*, **2002**, 42, 717.

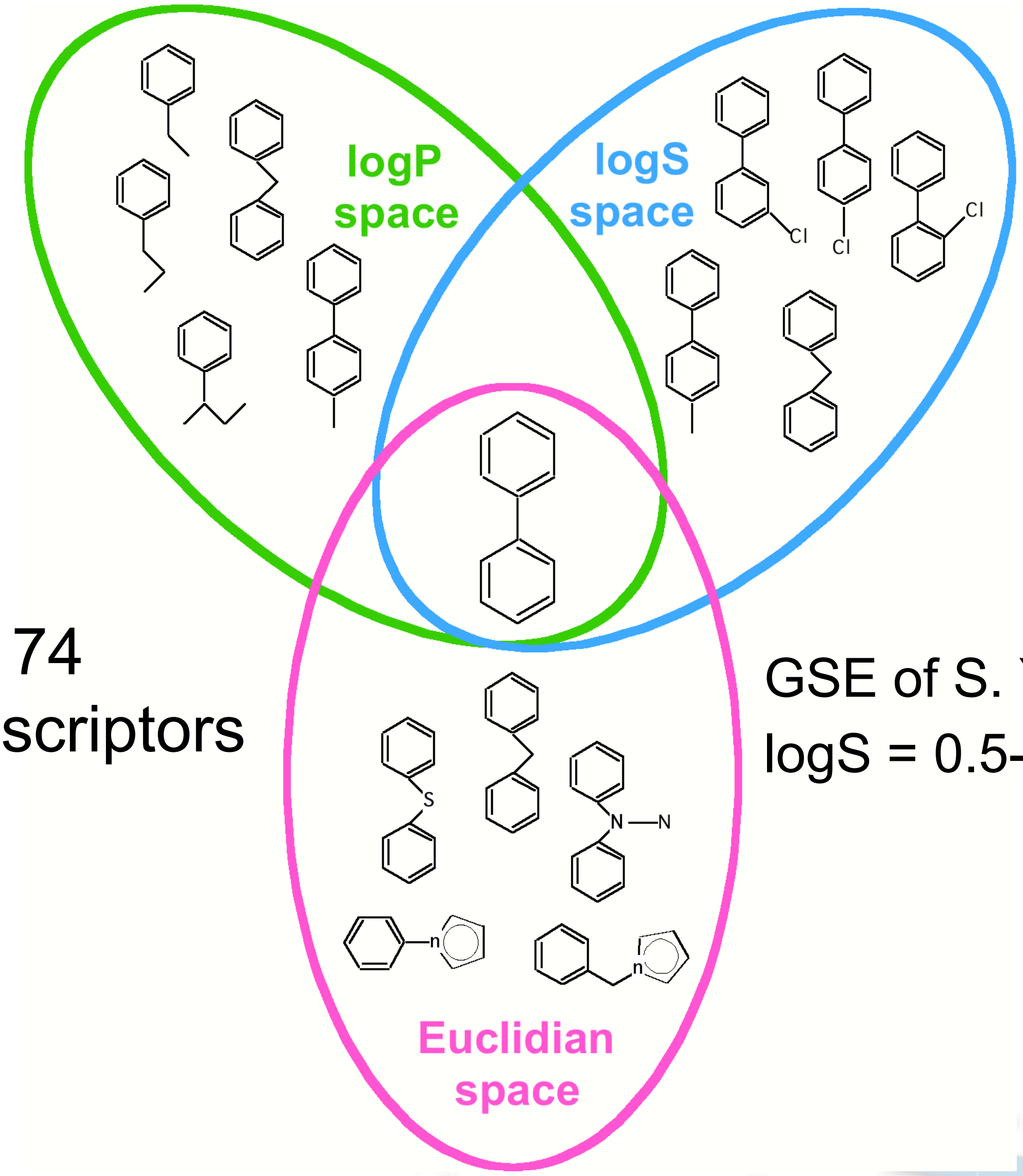
ALOGPS 2.1

- 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)

Both models use property-based similarity for model correction.

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

Nearest neighbors in different spaces



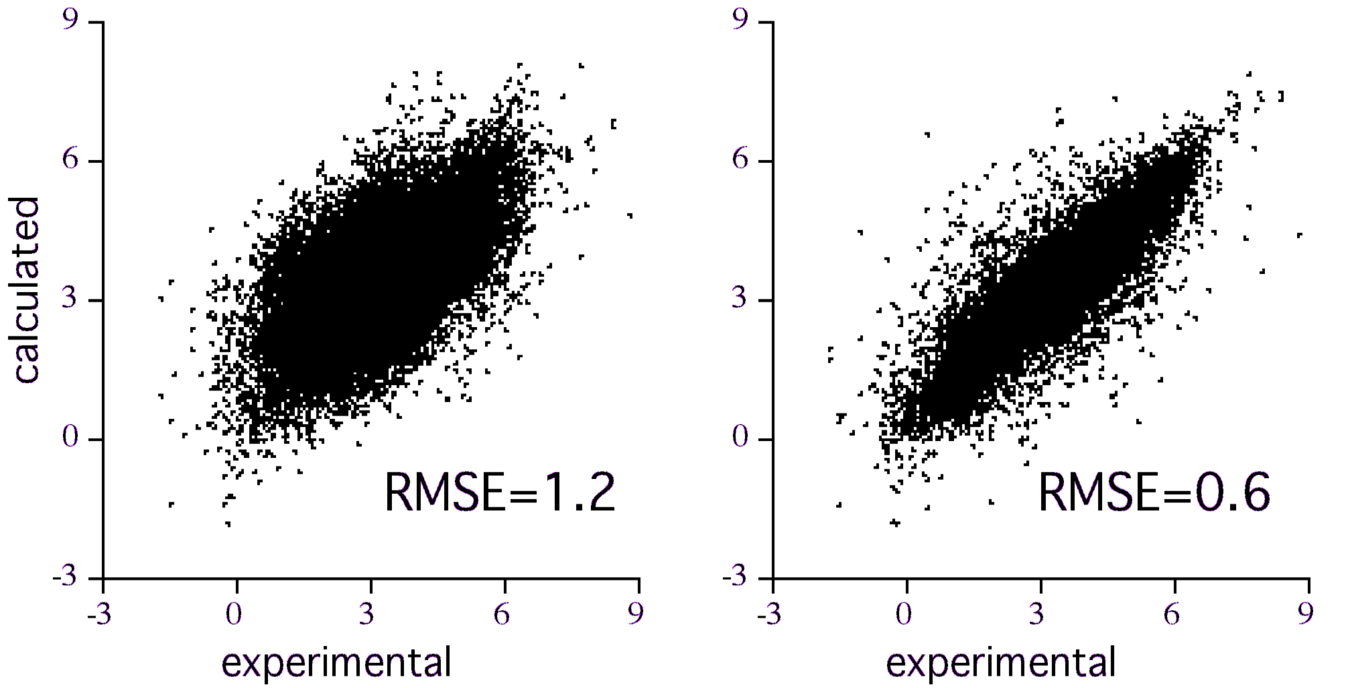
The same 74
E-state descriptors
were used

GSE of S. Yalkowsky

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

Analysis of Pfizer data

ALOGPS prediction for ElogD set of 17,861 compounds

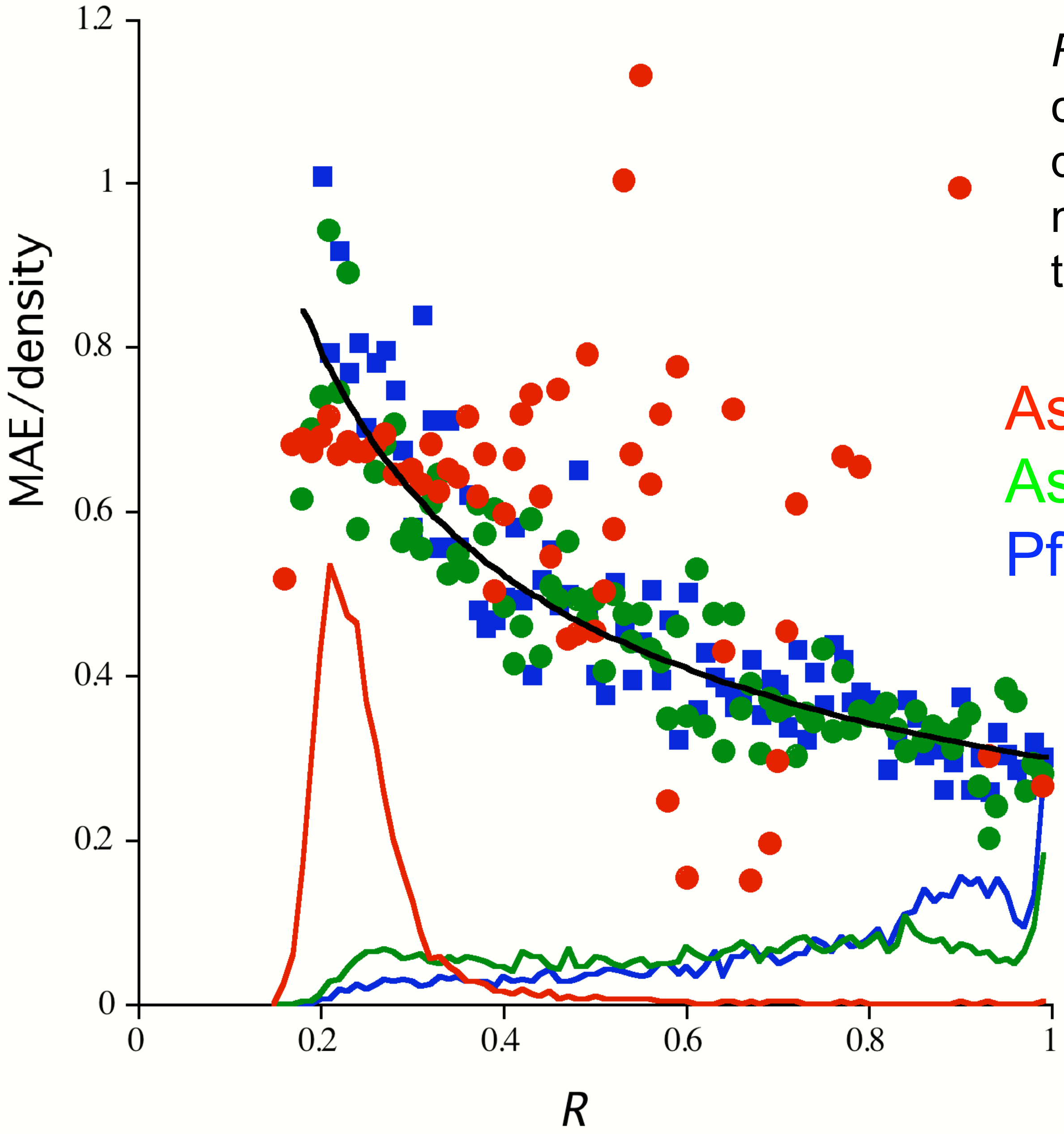


ALOGPS "as is" → ALOGPS LIBRARY

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.

Accuracy of logP prediction as function of R



R is a maximum correlation (r^2) of a query molecule to a molecule in the training set (LIBRARY)

AstraZeneca blind

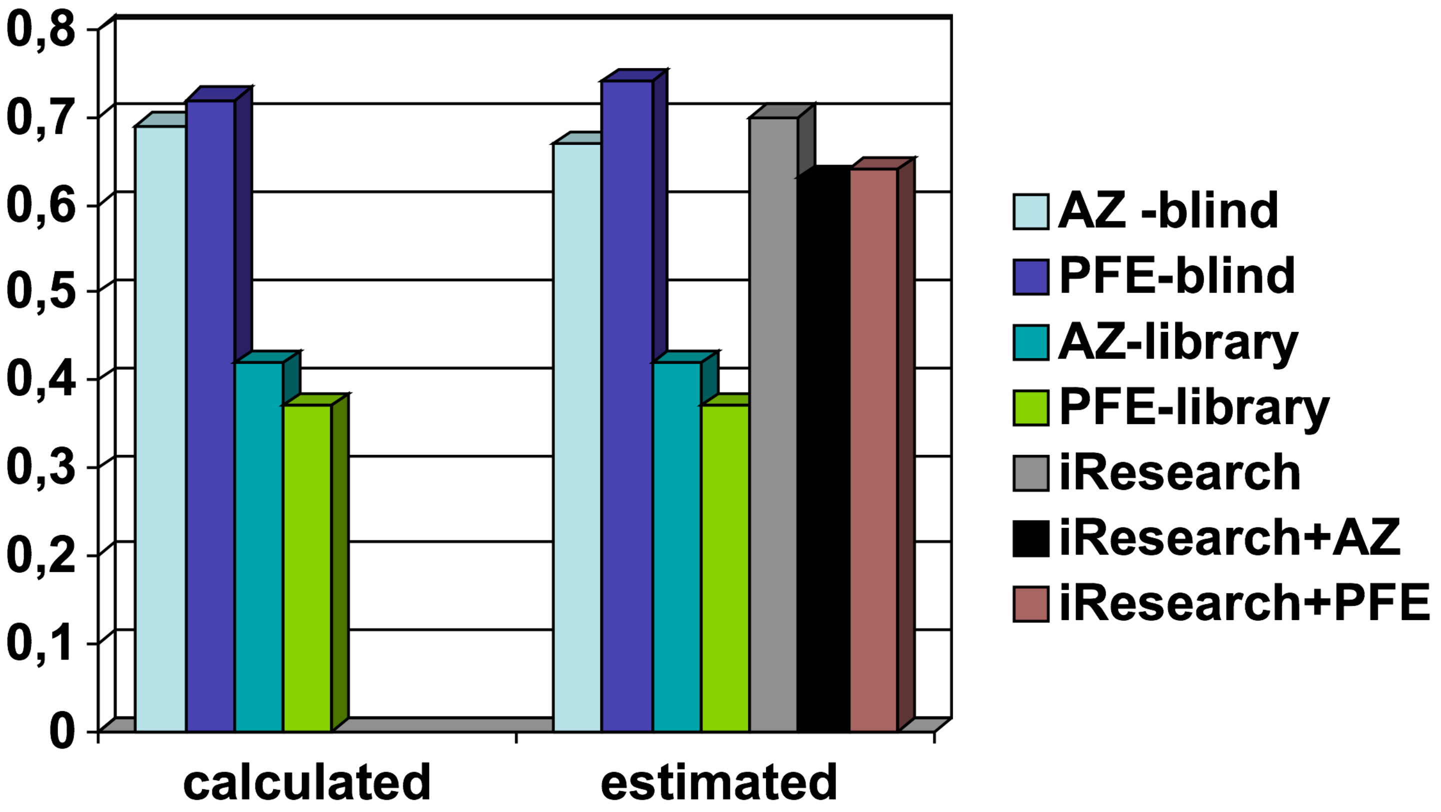
AstraZeneca LIBRARY

Pfizer LIBRARY

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

Tetko et al, Can we predict accuracy of ADMET? DDT, 2006, in press.

Estimated and calculated Mean Absolute Errors for AstraZeneca (AZ), Pfizer (PFE) and iResearch Library sets

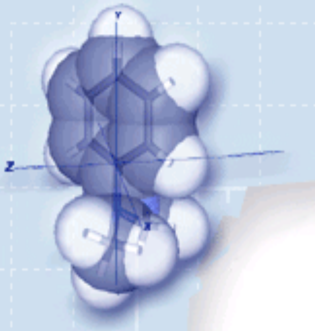


AZ - 7498 molecules
PFE - 8750 molecules
IResearch ChemNavigator Library - 13,333,629 molecules

Prediction of iResearch Library ($13*10^6$ molecules) in blind mode and using PFE LIBRARY

- >514,000 molecules logP> 5 --> logP<5
- >495,000 molecules changed |logP| > 1

PFE dataset contains 8750 molecules



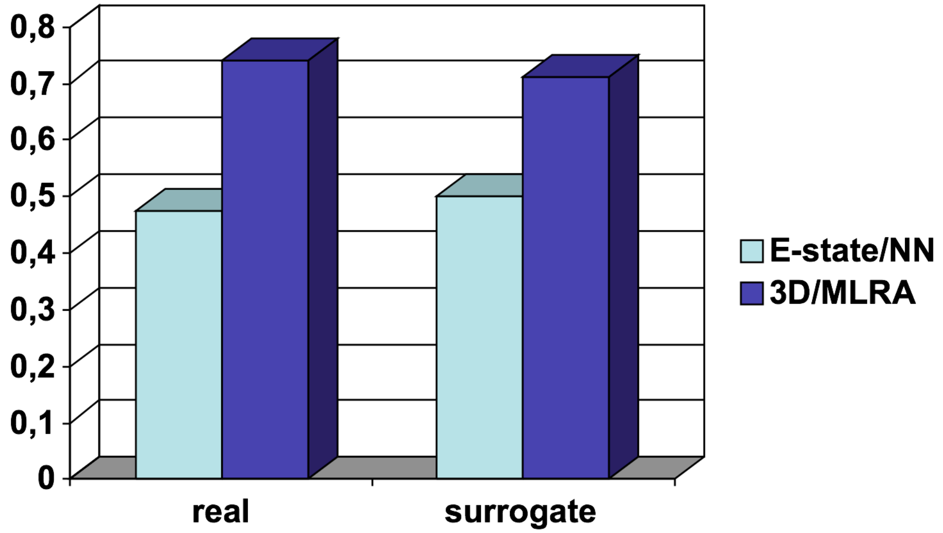
Secure sharing of information but not molecules

- Organized by T. Oprea, 229th ACS, San Diego
- Two dedicated session (CINF, COMP) ca 20 participants
- Too secure sharing makes impossible model development (relevant information is lost)
- Less than 1 bit/atom is required to store molecules in “zip” file (1 float value for molecule with 35 atoms)
- Thus, any proposed method can be secure until they are “hacked”
- Probably sharing molecular descriptors of a target molecule is a quite difficult business
- We can share ranks of models -- limited to the existing model
- But …. let us share reliably predicted molecules!
- These are the molecules with high ***R*** in property space to the target molecule

Real data vs surrogate data model for logP prediction

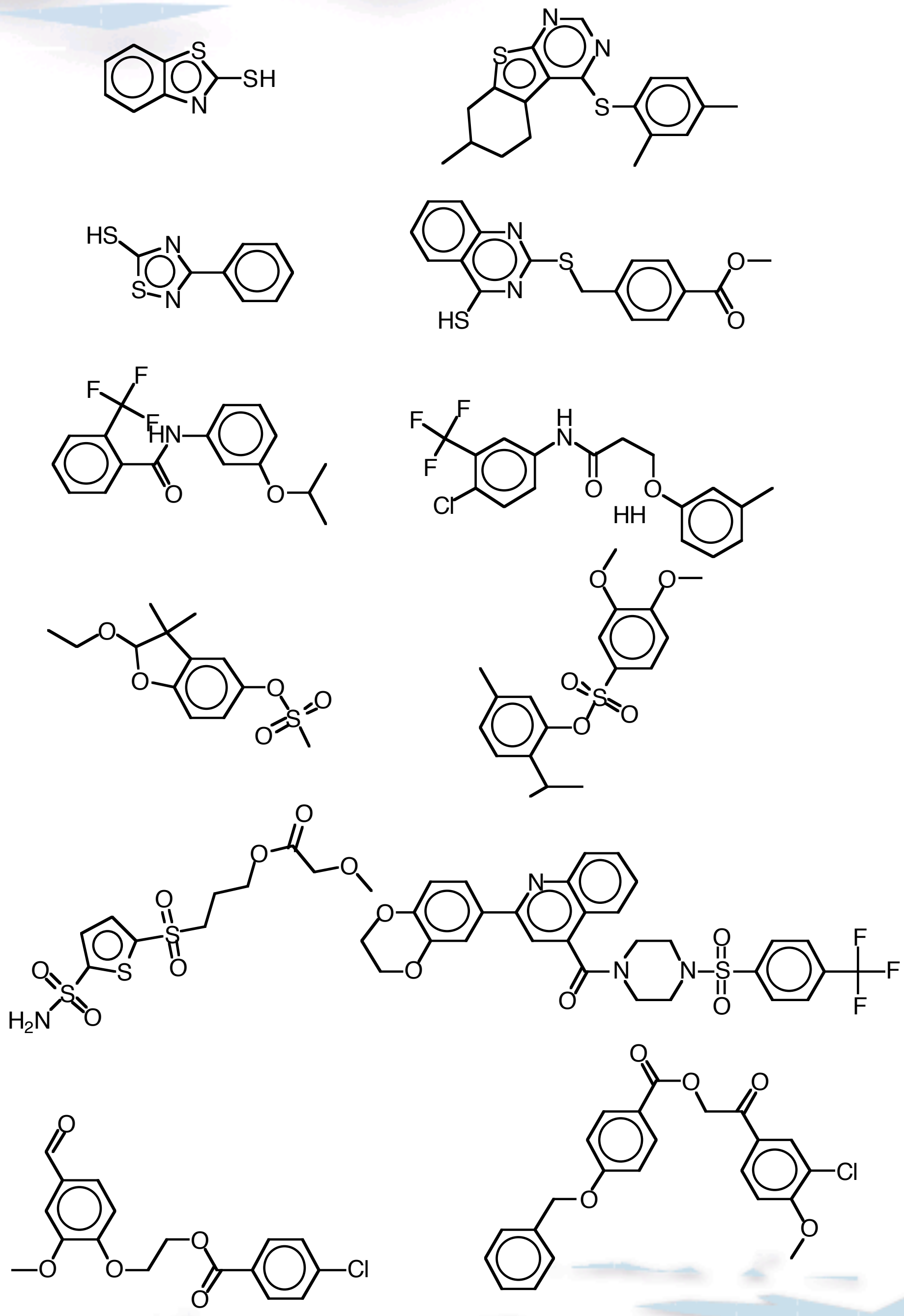
- Take a “real” molecule from PHYSPROP logP dataset
- Find for it a significantly correlated molecule $r^2>0.3$ in the IResearchLibrary (use additional filters to filter structurally similar ones)
- Name it as a “surrogate” molecule, calculate for it logP value --> “surrogate data”
- Use “real” molecules with real logP values and “surrogate data” (dissimilar molecules with predicted logP) to develop models
- Predict all 12908 PHYSPROP molecules using both models

Real = surrogate = 1949 molecules

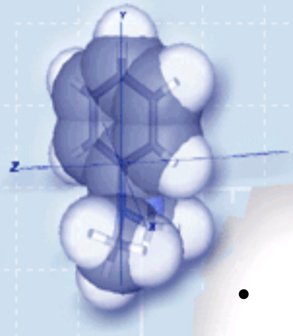


Att: It is a property-specific data sharing!!!

Real and surrogate molecules for logP



Tetko, Abagyan,Oprea
J. Comp. Aid. Mol. Des.
2005,19, 749.



Conclusions

- Residuals of an ensemble provide a new, target-activity-specific, representation of molecules -- they are not a noise but a very valuable information!
- Similarity in property-based space can be introduced as a distance (e.g., rank correlation) between vector of residuals[1,2] that is very specific for the target property[3,4]
- This similarity is a heart of the Associative Neural Network method[2,3] used in the ALOGPS[2] and 1H NMR[7] prediction programs
- It detects meaningful nearest neighbors, allows mechanistic interpretation[3,4]
- It can be used to estimate accuracy of prediction of models[5] **-- YES**
- It can be used for secure data sharing[6] and it is used in 1H NMR program* - **YES**
- The 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: Methods and Applications*, 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.; Aires-De-Sousa, J. *Tetrahedron Letters* **2004,** 45, (37), 6931.

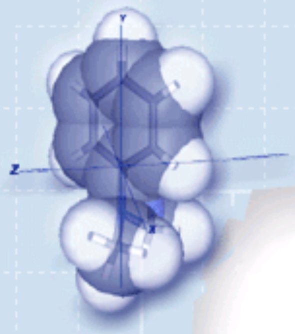
*-personal communication from Prof. J. Aires-De-Sousa
**-personal communication from Dr. R. Fraczkiewicz

Acknowledgement

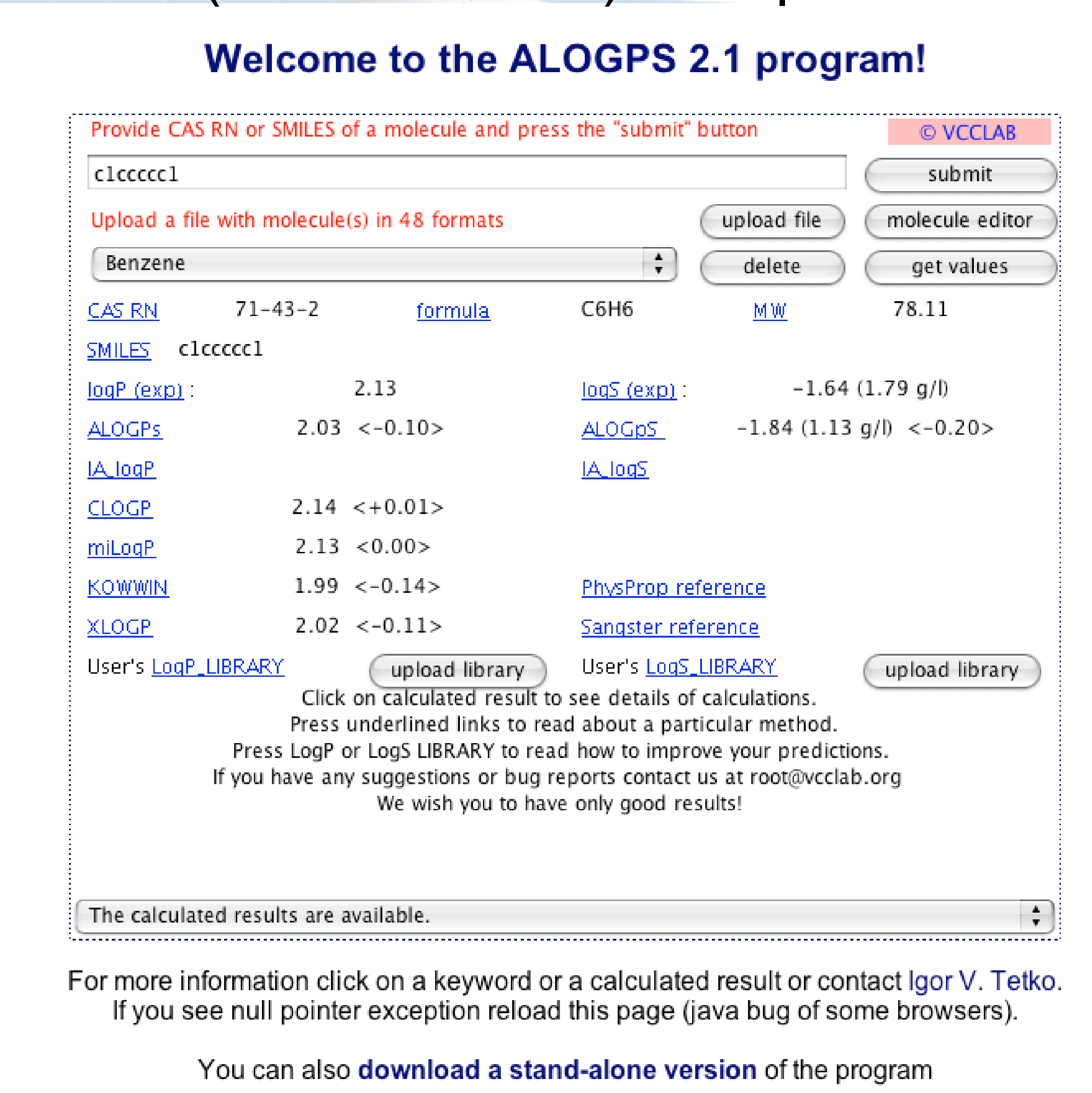
Part of this work was done thanks to
Virtual Computational Chemistry Laboratory
INTAS-INFO 00-0363 project

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

Thank you for your attention!



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

c1ccccc1

Upload a file with molecule(s) in 48 formats

Benzene

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

See also VCCLAB poster!