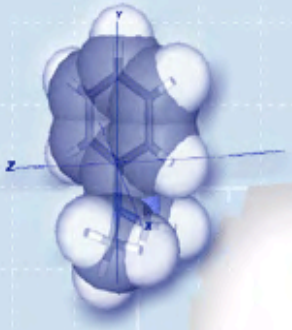


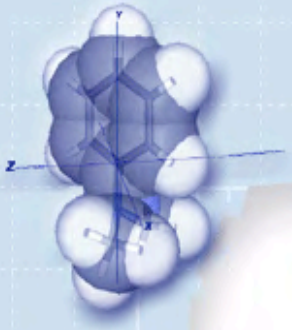
Troubles with Chemoinformatics and Associative Neural Networks

Igor V. Tetko

GSF -- Institute for Bioinformatics (MIPS), Neuherberg, Germany and Institute of Bioorganic & Petrochemistry, Kyiv, Ukraine

Fraunhofer FIRST, Berlin

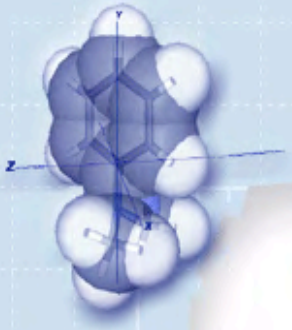
15 November 2006



Layout of presentation

- ✓ Problems with chemoinformatics models developed by Academia
- Introduction to the Associative Neural Network
- Properties of the ASNN: bias estimation and correction
- Properties of the ASNN: applicability domain of models
- Properties of the ASNN: secure sharing of data

gsf



Challenges for development of Chemoinformatics models

Public datasets

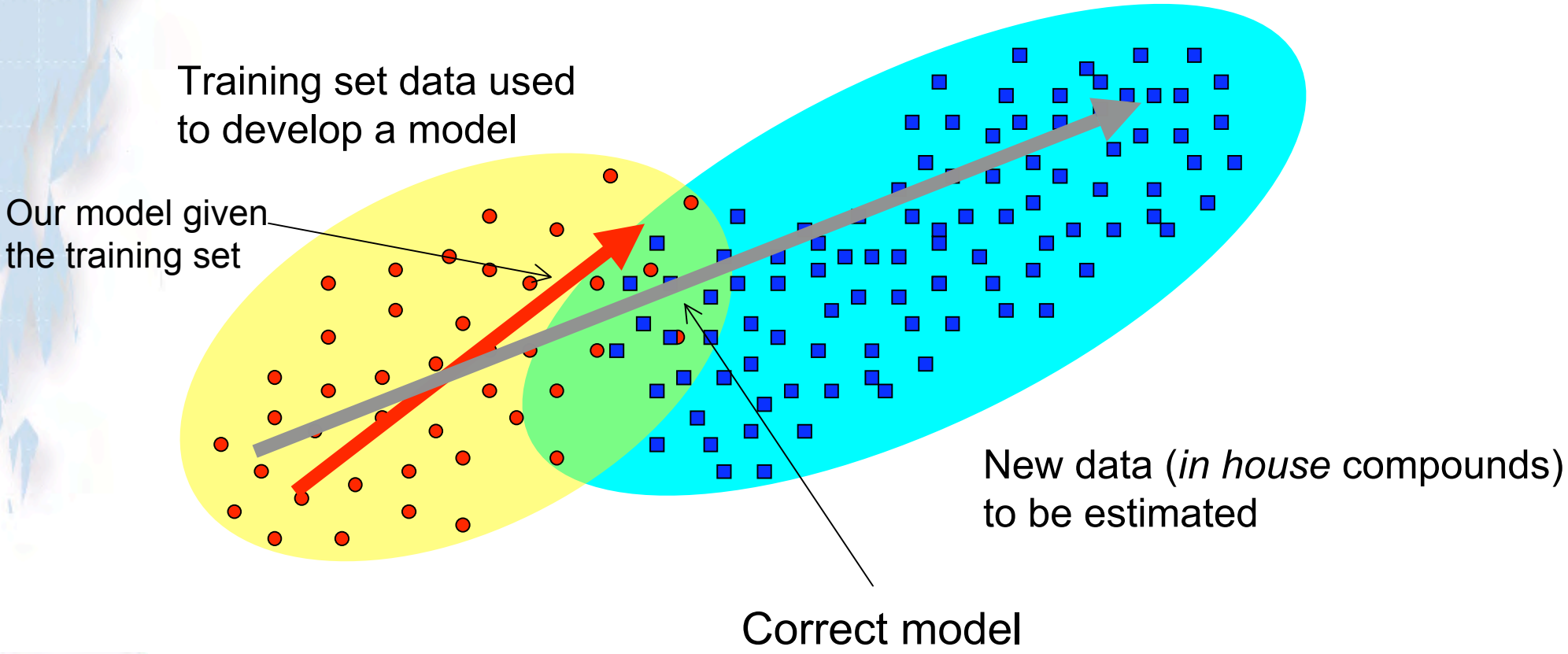
- Limited data availability, data are expensive
- Data are noisy, inconsistent between laboratories (not all conditions controlled/specified)
- No new measurements can be easily made
- Data are heterogenous

In house datasets

- Data are consistent (at least within each company site)
- Data are homogeneous and may correspond just to few series of interest
- New data can be easily and incrementally measured
- Accuracy of data may vary depending on used experimental protocol
- Not available for public development

gsf

"Public" model can fail due to different chemical diversity in training & test sets



"One can not embrace the unembraceable."

Possible: 10^{60} - 10^{100} molecules theoretically exist (> 10^{80} atoms in the Universe)

Achievable: 10^{20} - 10^{24} can be synthesized now (weight of the Moon is ca 10^{23} kg)

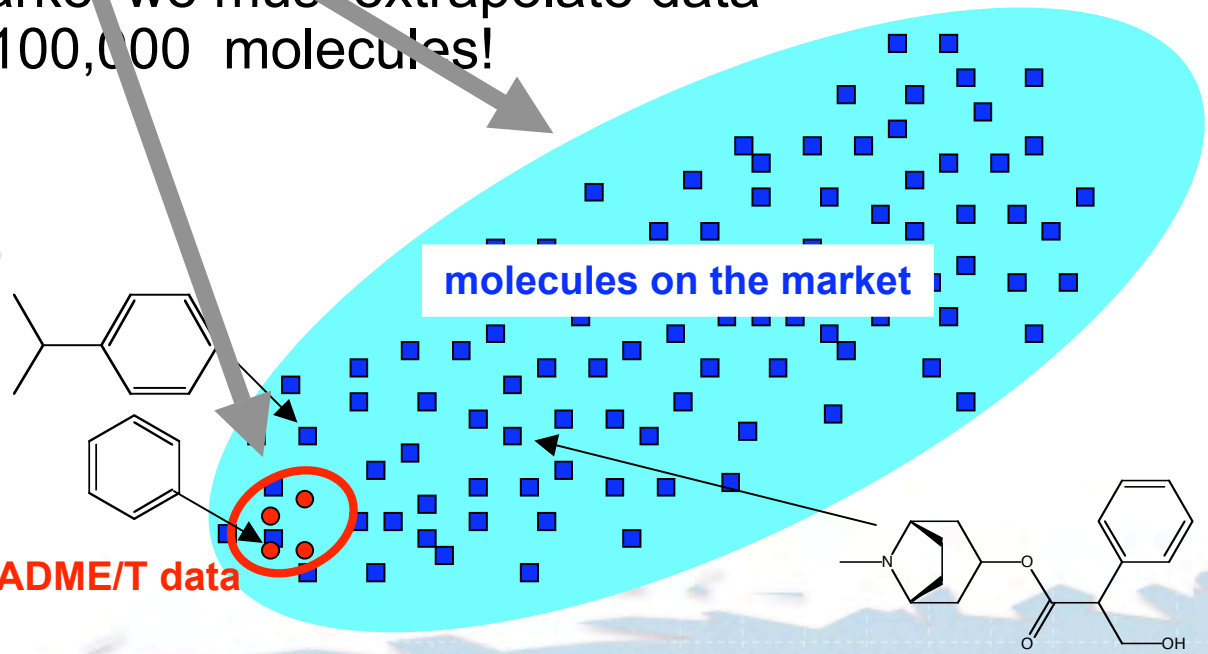
Available: $2*10^{7}$ molecules are on the market

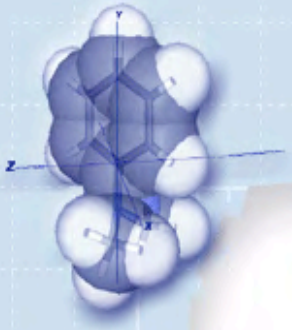
Measured: 10^{2} - 10^{4} molecules with ADME/T data

Problem: To predict ADME/T properties of just molecules on the market we must extrapolate data from one to 1,000 - 100,000 molecules!

Kozma Prutkov

We need methods which can estimate the accuracy of predictions!

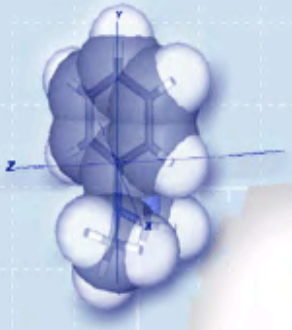




Troubles with models

- Model developed using “public data”
  - May have low prediction accuracy due to limited chemical diversity of both sets (problem of accuracy)
- Model developed using “public” + “in house” data
  - Difficult to receive “in house” data
  - “Public data” may not be really public (problem of IP)
  - May require significant computing time/expertise of the end user
  - Public/in-house data can be incompatible
- Model developed using “public” data and then “*adjusted*” for molecules from “*in house*” data
  - Associative Neural Network

gsf



Layout of presentation

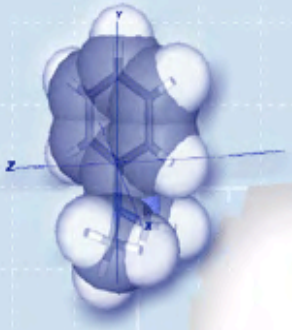
- Problems with chemoinformatics models developed by Academia
- ✓ Introduction to the Associative Neural Network
- Properties of the ASNN: bias estimation and correction
- Properties of the ASNN: applicability domain of models
- Properties of the ASNN: secure sharing of data

gsf

Associative Neural Network (ASNN)

- ✓ Some software tools rely just on one “best” model.
- ✓ Other software tools rely on the ensemble average (“panel of experts”).
- ✓ ASNN explores disagreement of individual models in the ensemble to improve its accuracy and to derive a confidence score.

gsf



Bias-variance decomposition

$$PE(f_i) = (f_i - Y)^2 = (f^* - Y)^2 + (f^* - \bar{f})^2 + (f_i - \bar{f})^2$$

f_i - *prediction of* i^{th} *model;*

Y - *experimental values*

f^* - *ideal function*

$\bar{f}$ - *ensemble average*

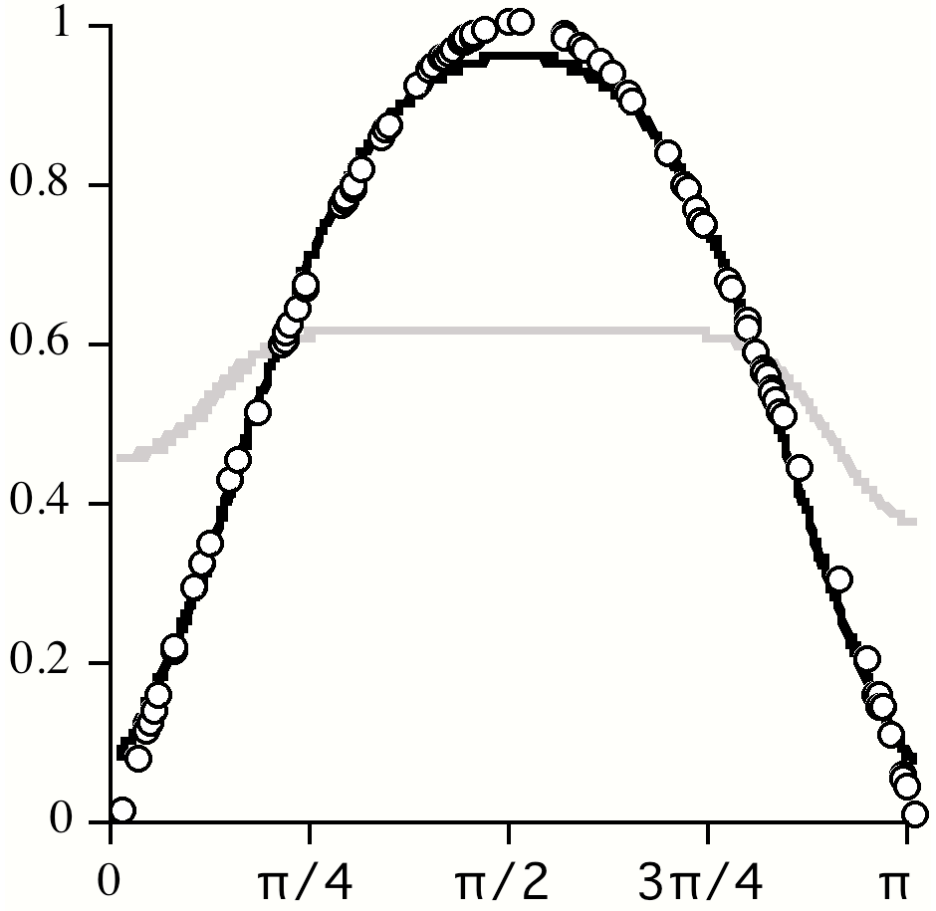
Variance: can be decreased using a large ensemble of models

Bias: can be partially decreased using more flexible models (large number of hidden neurons), early stopping, AdaBoost algorithm, etc. (indirect methods)

Question: Can we directly estimate (and correct) the bias of the ensemble?

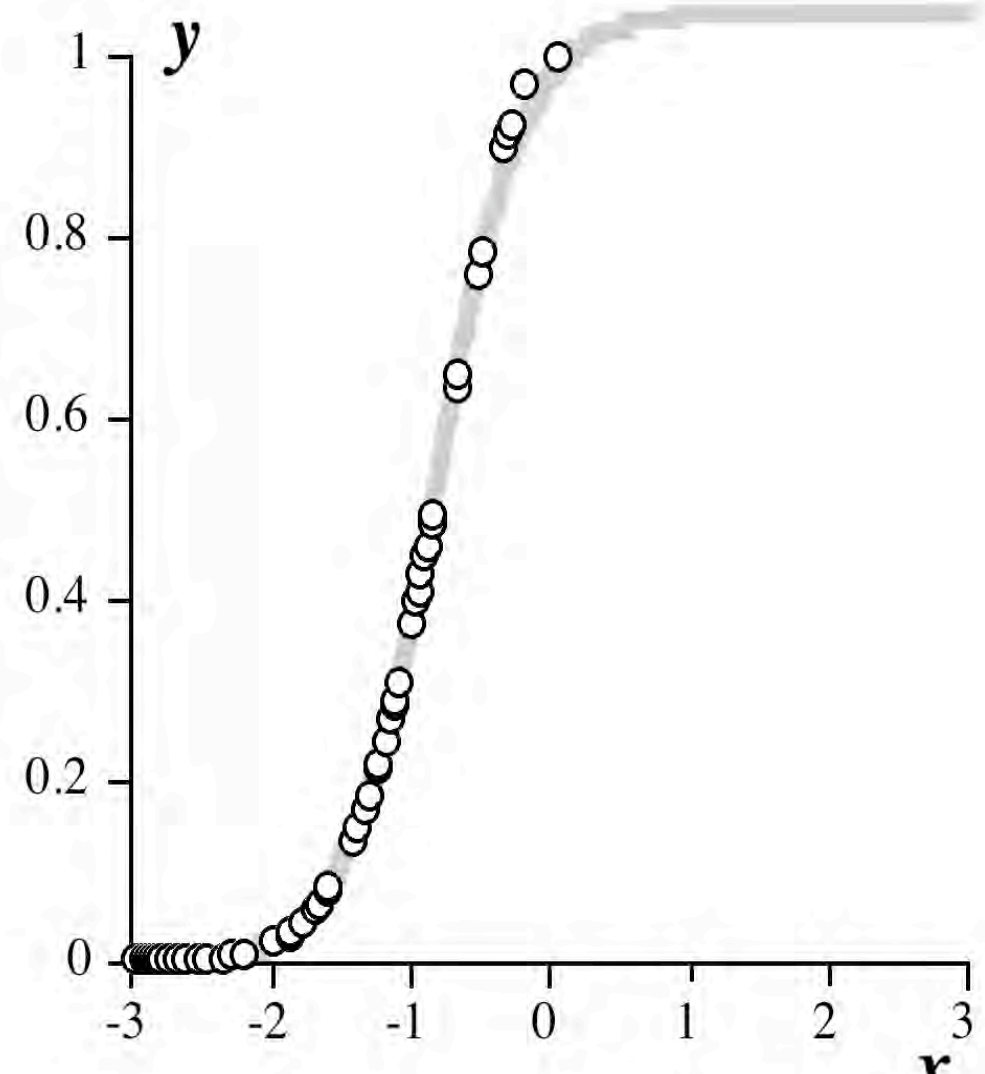
Answer: Yes, if we can correctly detect nearest neighbors of the target data case in "*functional space*"!

Sources of bias: underfitting

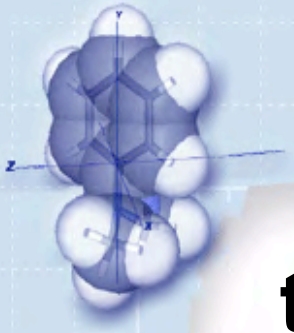


Sine function approximation by neural networks with one and two hidden neurons ($x=x_1+x_2$)

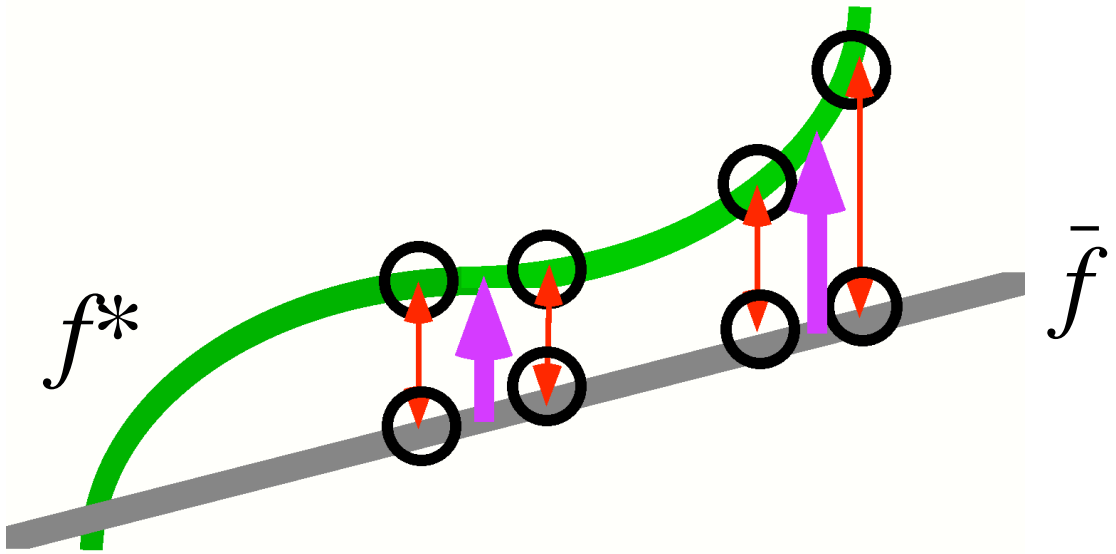
Sources of bias: incomplete data (extrapolation, applicability domain)



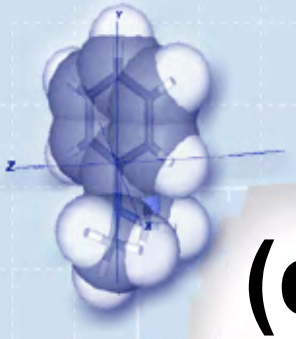
Gauss function extrapolation ($x=x_1+x_2$)



Correction of model bias by the error of the nearest neighbors of the target point

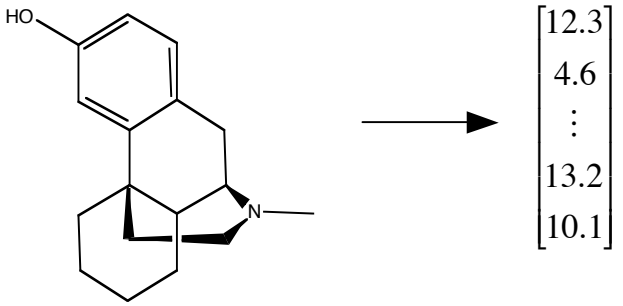


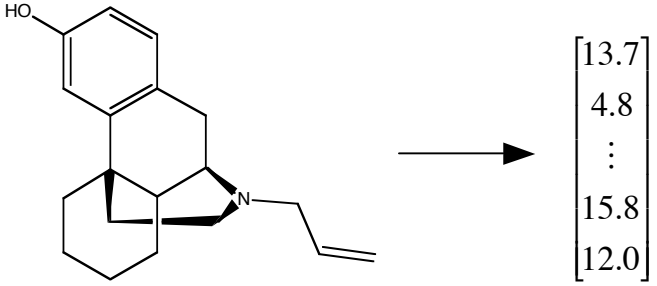
gsf



Representation of molecules (data cases) for machine learning

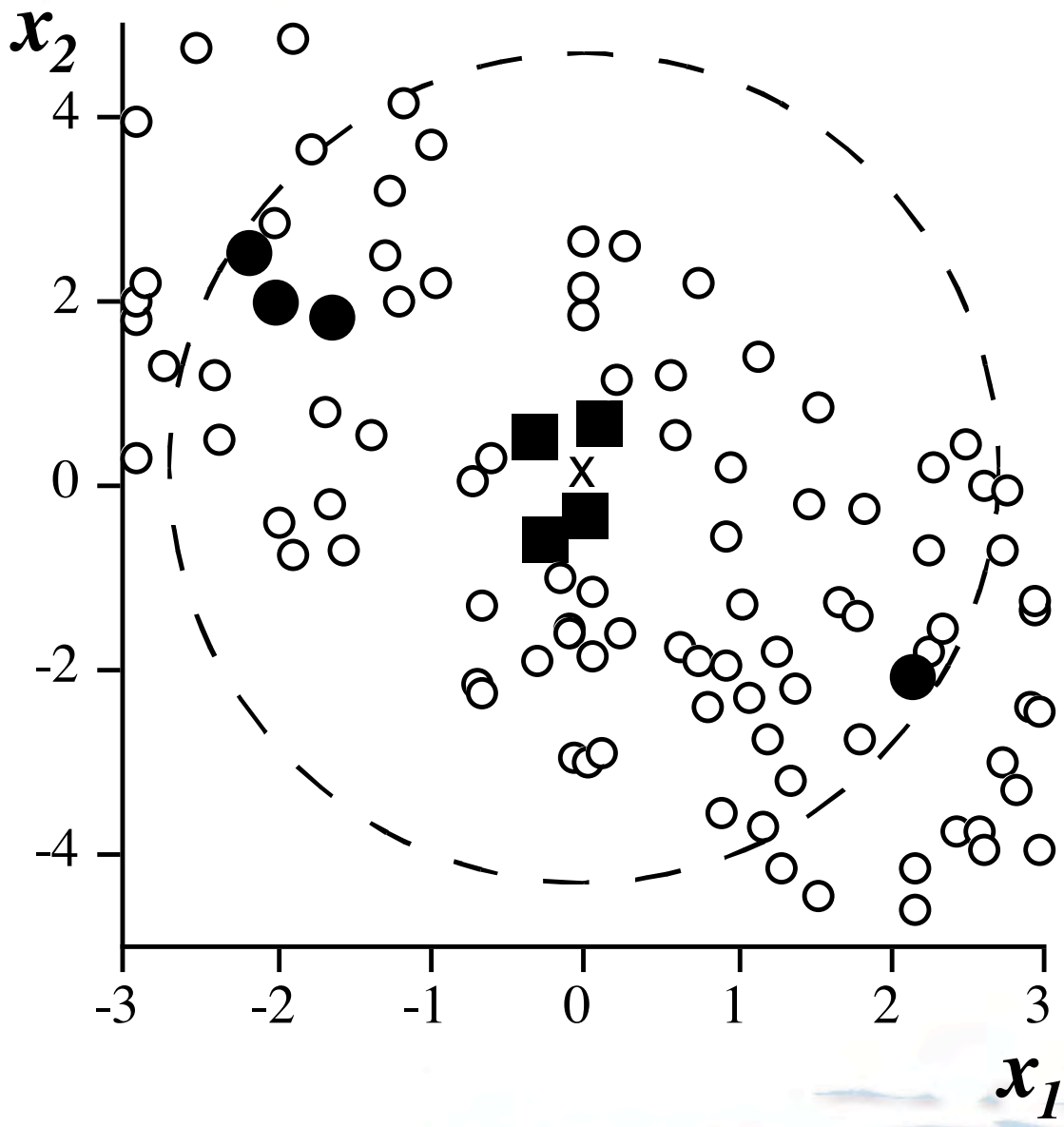
- Can be defined with calculated properties (logP, quantum-chemical parameters, etc.)
- Can be defined with a set of structural descriptors (topological 2D, 3D, etc.).
- In general: any set of descriptors



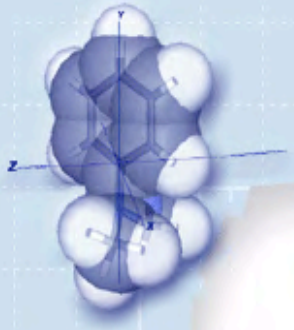


gsf

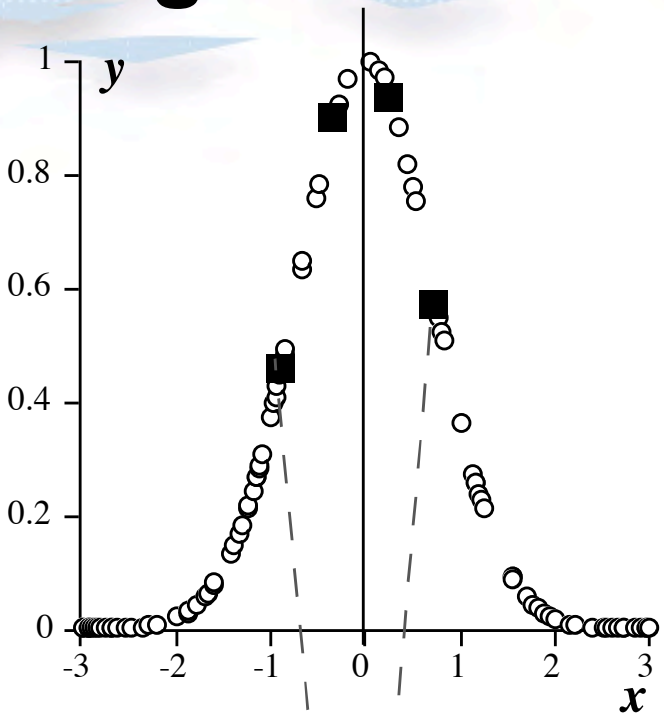
Nearest neighbors in the input space (space of descriptors)



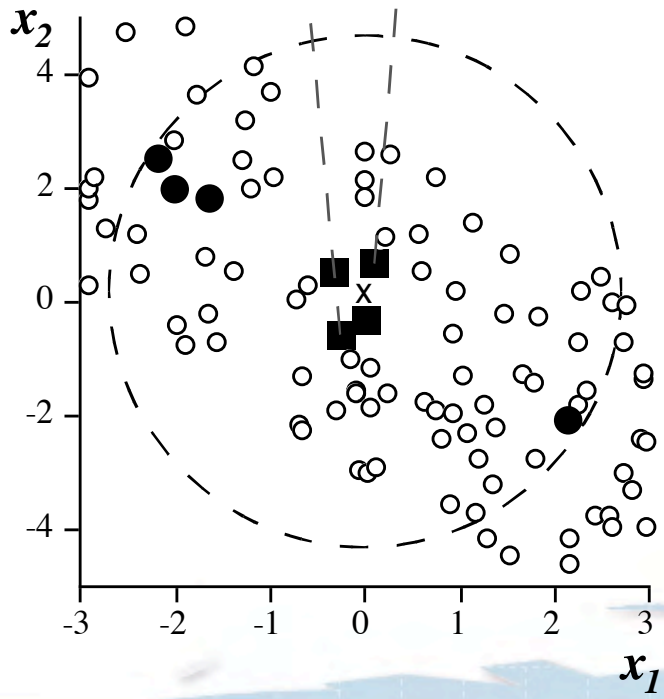
Nearest neighbors and activity



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



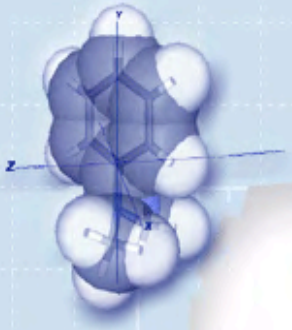
$x=x_1+x_2$!

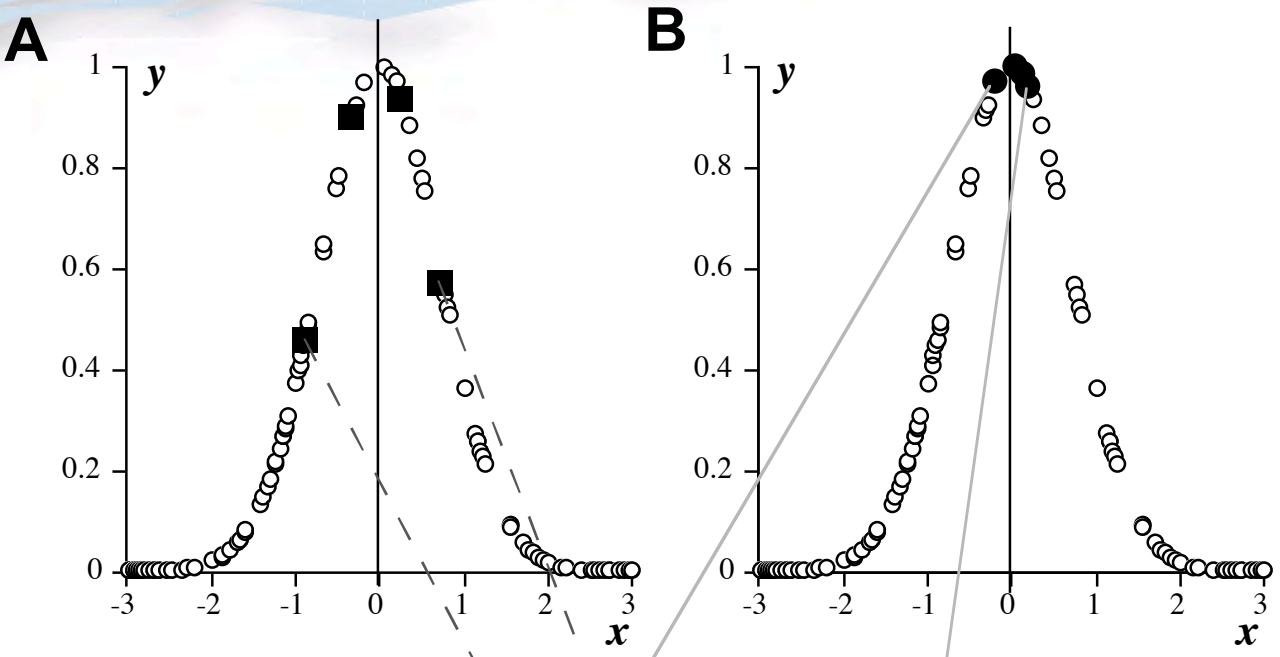


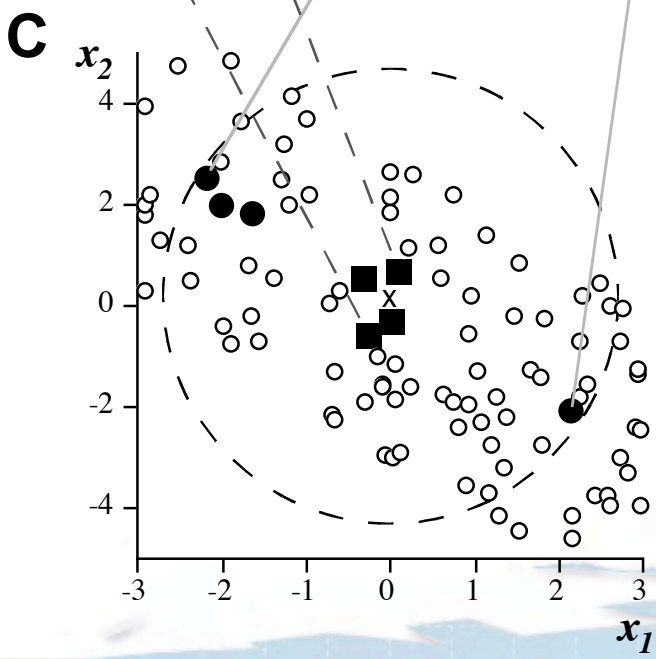
The nearest neighbors in the descriptor space are not necessary neighbors in the property space!

gsf

Nearest neighbors and activity



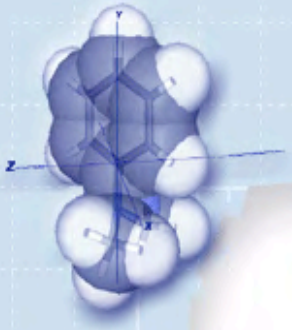


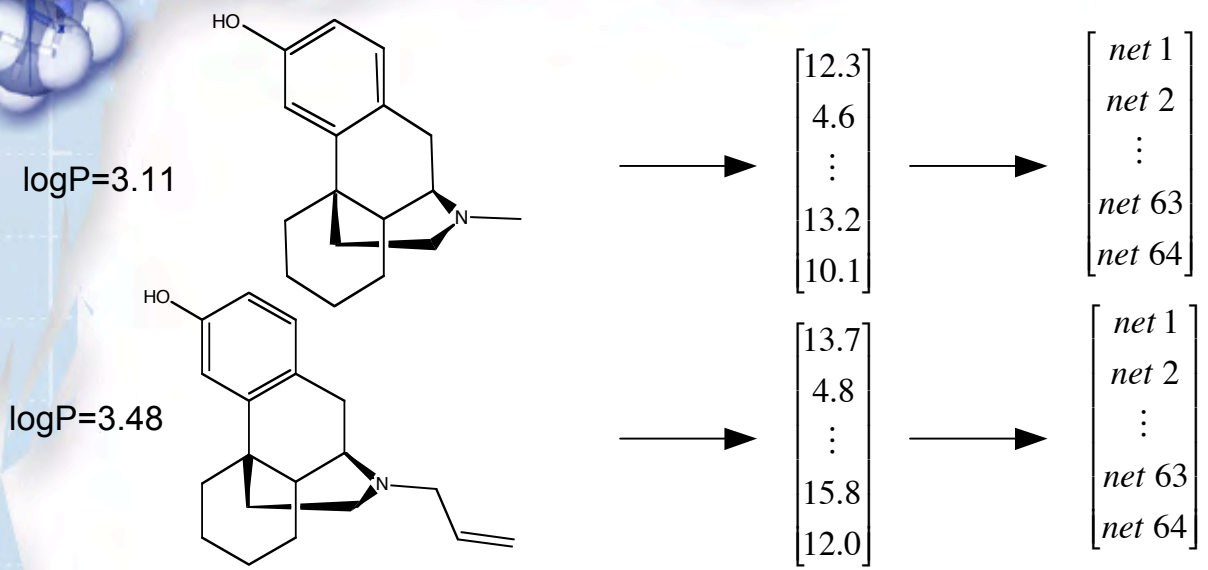


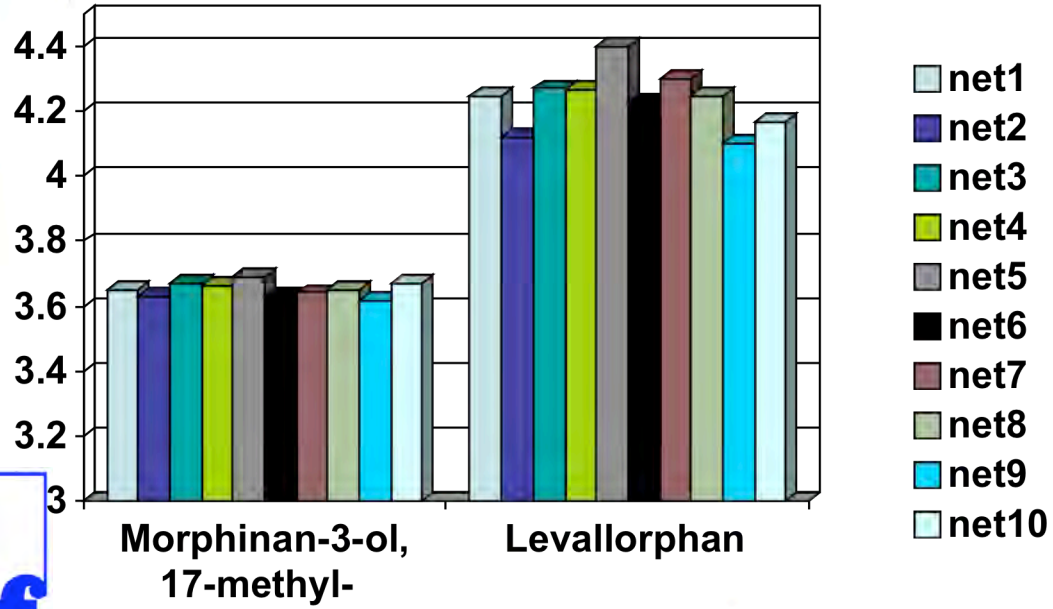
$x=x_1+x_2$

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

A property-based similarity





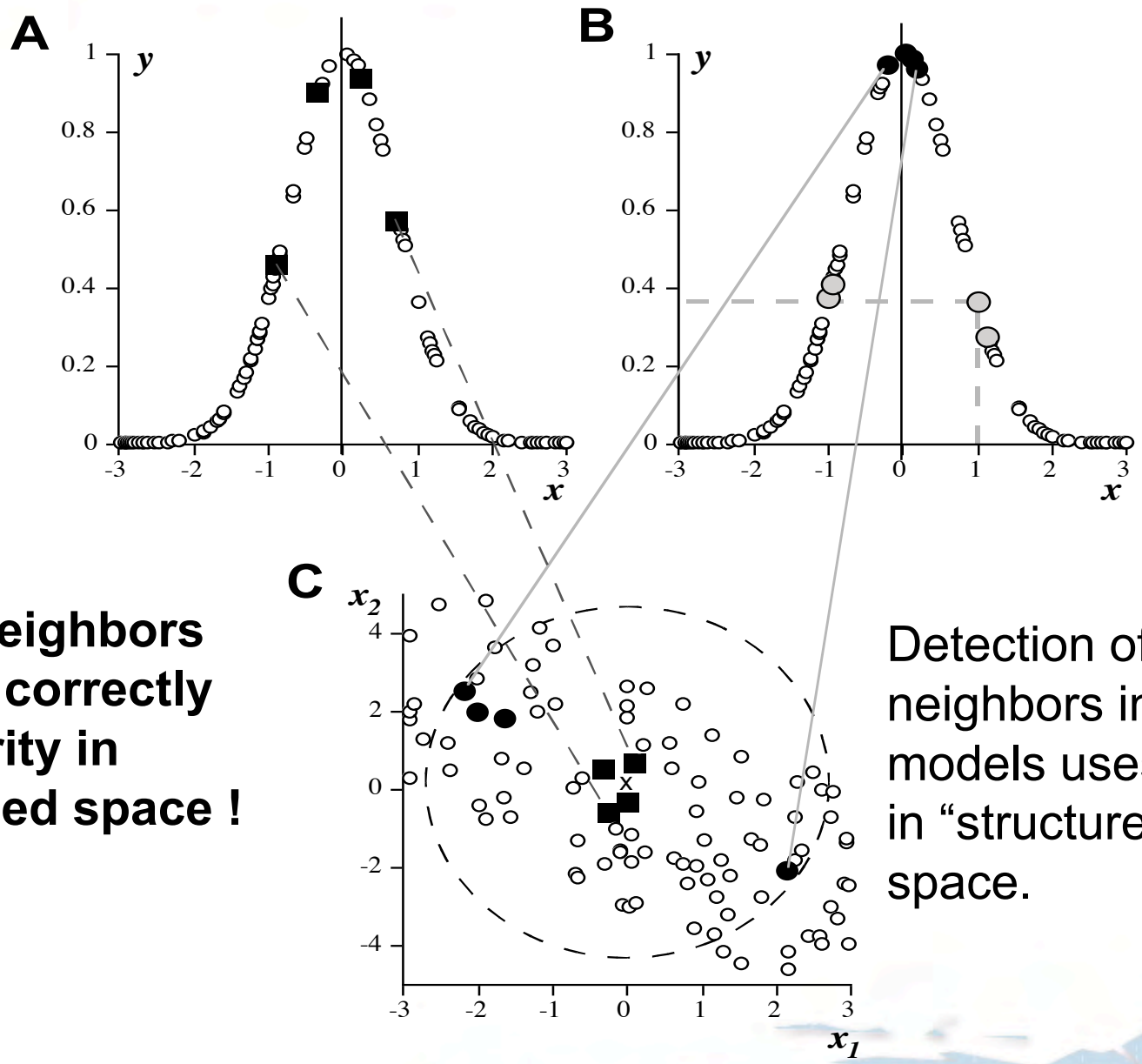


-- both molecules are the nearest neighbors, r^2=0.47, in space of residuals amid >12,000 molecules!

The **property-based similarity*** is defined as *correlation of ensemble residuals*

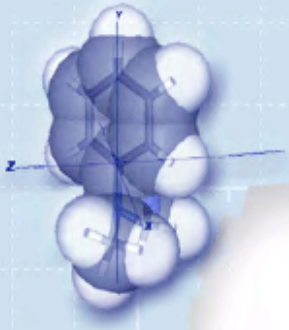
*Tetko, I.V.; Villa, A.E.P. Neural Networks, 1997, 10, 1361-1374

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.



Associative Neural Network (ASNN)

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

Traditional ensemble:

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

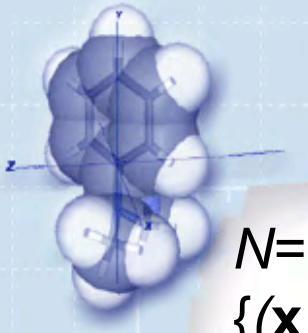
Pearson's (Spearman) correlation coefficient $r_{ij}=R(z_i,z_j)>0$

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

<<= 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 (neural network associations of the given model)

Illustrative example



N=3, three cases $\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),(\mathbf{x}_3,y_3)\}$ in the training set:

M=10 (ten models) in the ensemble $[\mathbf{x}_i]\bullet[\mathbf{ANNE}]_0 = \begin{bmatrix}0.5\\0.3\\0.4\\0.5\\0.6\\0.7\\0.5\\0.4\\0.5\\0.6\end{bmatrix}, \begin{bmatrix}0.9\\0.8\\0.7\\0.9\\0.8\\0.9\\0.8\\0.7\\0.5\\0.4\end{bmatrix}, \begin{bmatrix}0.2\\0.1\\0.3\\0.1\\0.1\\0.3\\0.2\\0.3\\0.2\\0.1\end{bmatrix}$

$\bar{z}_1 = 0.5; \bar{z}_2 = 0.74; \bar{z}_3 = 0.19$ – *ensemble average*

$y_1 = 0.4; y_2 = 0.6; y_3 = 0.17$ – *experimental values*

Test case: $[\mathbf{x}_t]\bullet[\mathbf{ANNE}]_0 = \begin{bmatrix}0.7\\0.4\\0.4\\0.6\\0.7\\0.8\\0.9\\0.7\\0.4\\0.6\end{bmatrix}$

$\bar{z}_t = 0.62$ – – *ensemble prediction*

ASNN result:

$r(x_1,x_t) = 0.55$

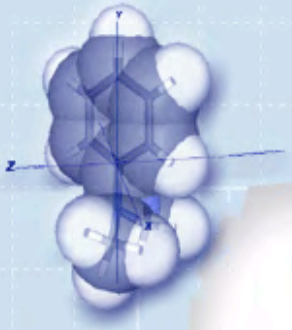
$r(x_2,x_t) = 0.42,\ k = 2$

$r(x_3,x_t) = 0.16$

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

$$=> \bar{z}_t' = 0.62 + \frac{1}{2}\left(\left(0.4-0.5\right)+\left(0.6-0.74\right)\right)$$

$$= 0.62 - 0.12 = 0.5$$

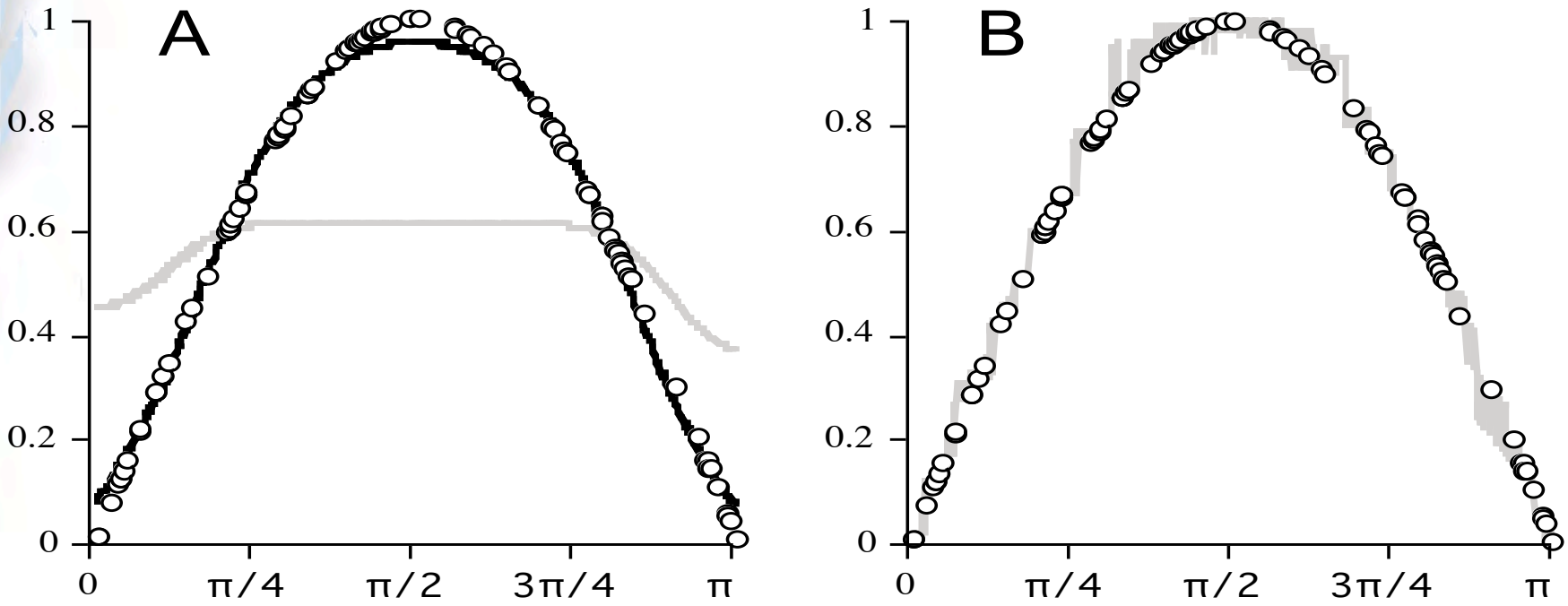


Layout of presentation

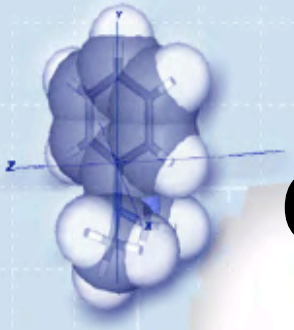
- Problems with chemoinformatics models developed by Academia
- Introduction to the Associative Neural Network
- ✓ Properties of the ASNN: bias estimation and correction
- Properties of the ASNN: applicability domain of models
- Properties of the ASNN: secure sharing of data

gsf

ASNN sine function approximation (correction of the underfitting bias)



A) ANN trained with two (black) and one (grey) hidden neurons. B) ASNN results using one hidden neuron using the same networks from A).



Classification of UCI data sets

dataset	training set	test set	MLP architecture[1]	Boosted results[1]	ANN 50 CC[2]	ASNN[2]
Letter	16000	4000	16-70-50-26	1.5%	4.1%	1.8%
Satellite	4435	2000	35-30-15-6	8.1%	8.3%	7.8%

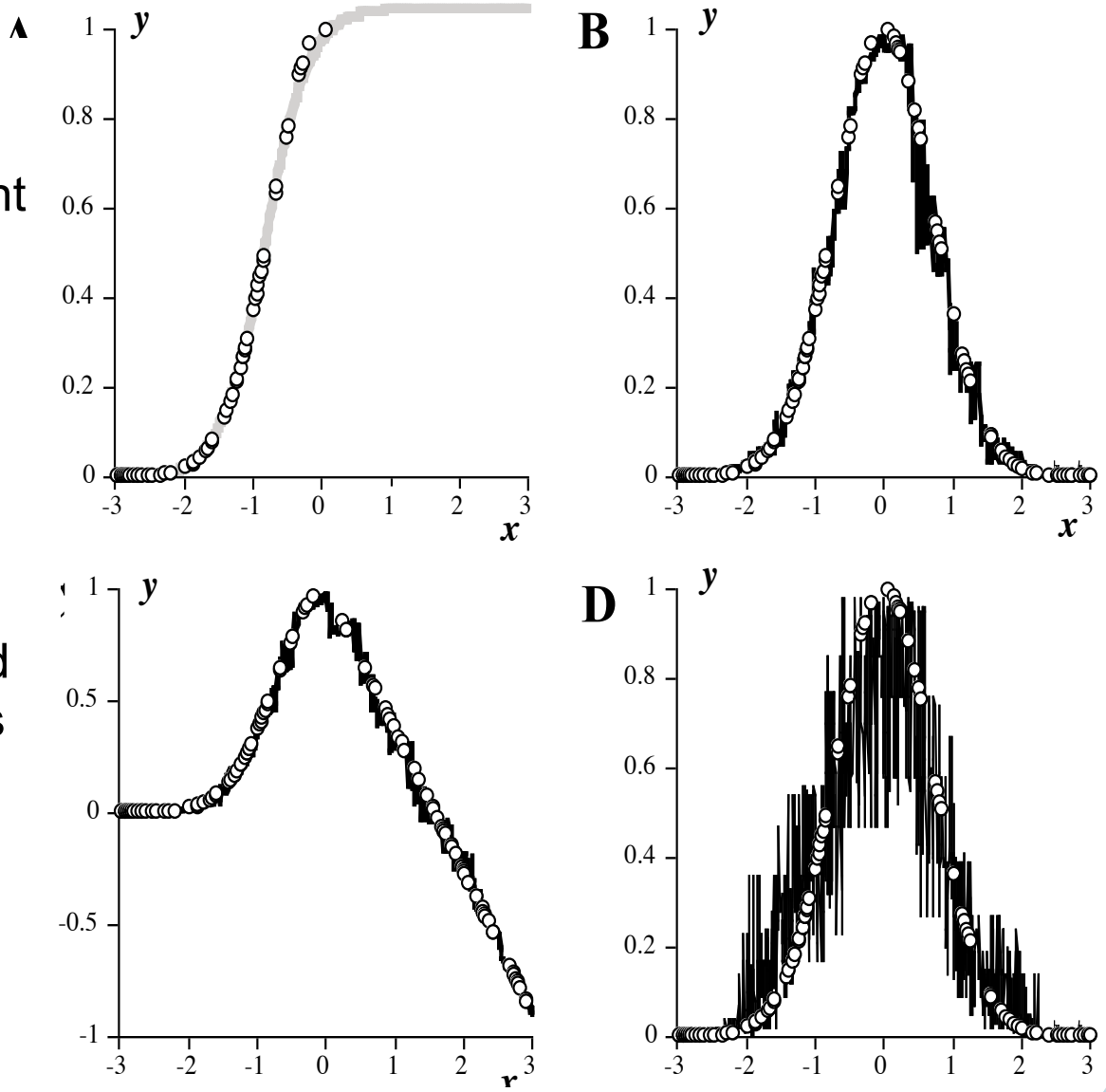
1 - Schwenk and Bengio, *Neural Computation*, **12**, 2000, 1867-1887.
2 - Tetko, *Neural Processing Letters*, 2002, **16**, 187-199.

gsf

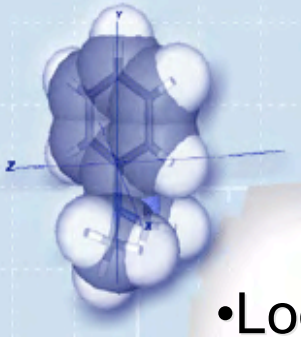
Gauss function extrapolation (correction of extrapolation bias with “fresh data”)

LIBRARY mode: new data are used for kNN correction (enlargement of the ASNN memory) without rebuilding the neural network model!

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



Notice: $x=x_1+x_2$



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)

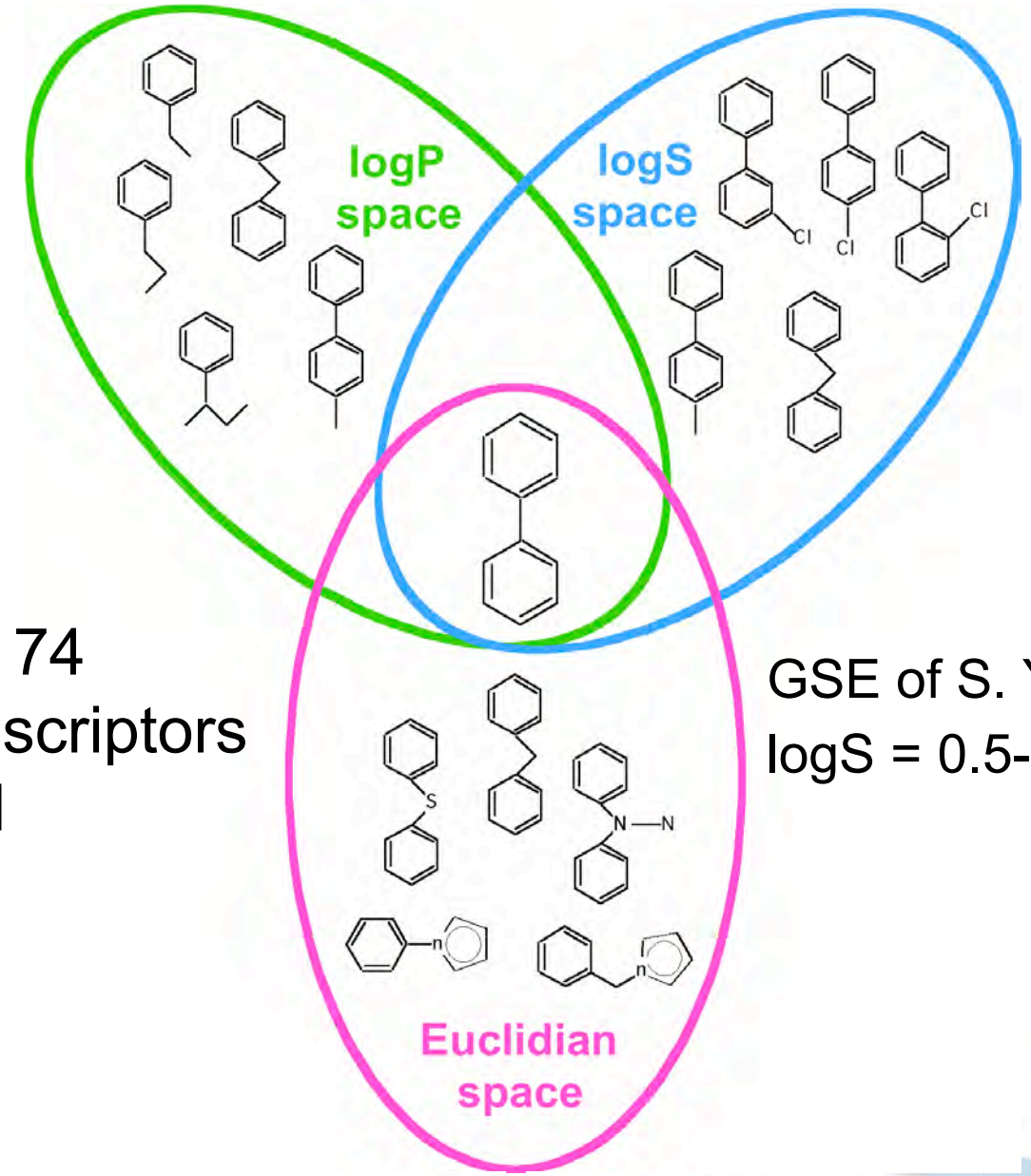
•LogS: 33 input E-state indices, 1291 molecules in the database, 64 neural networks in the ensemble. Calculated results RMSE=0.49, MAE=0.35, n=18 outliers (>1.5 log units)

- 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

Nearest neighbors in different spaces

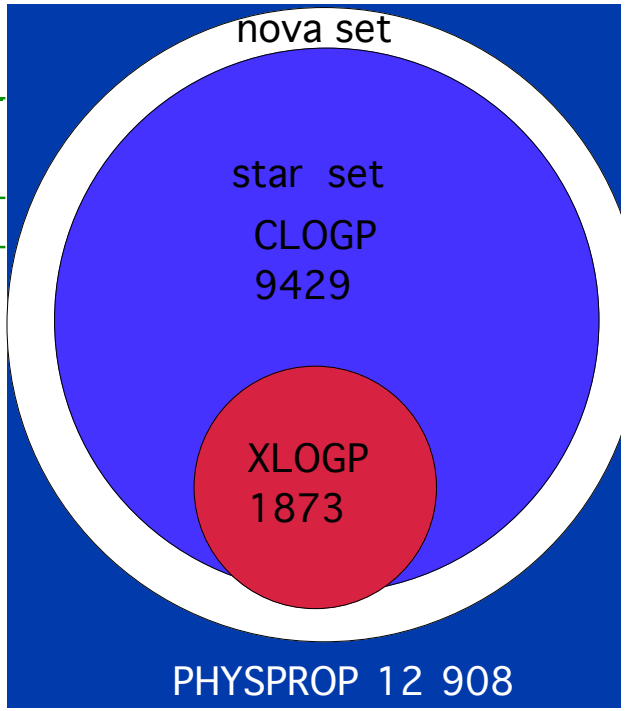


The same 74
E-state descriptors
were used

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

Accuracy of predictors developed using whole set (ALOGPS) and “star” set only

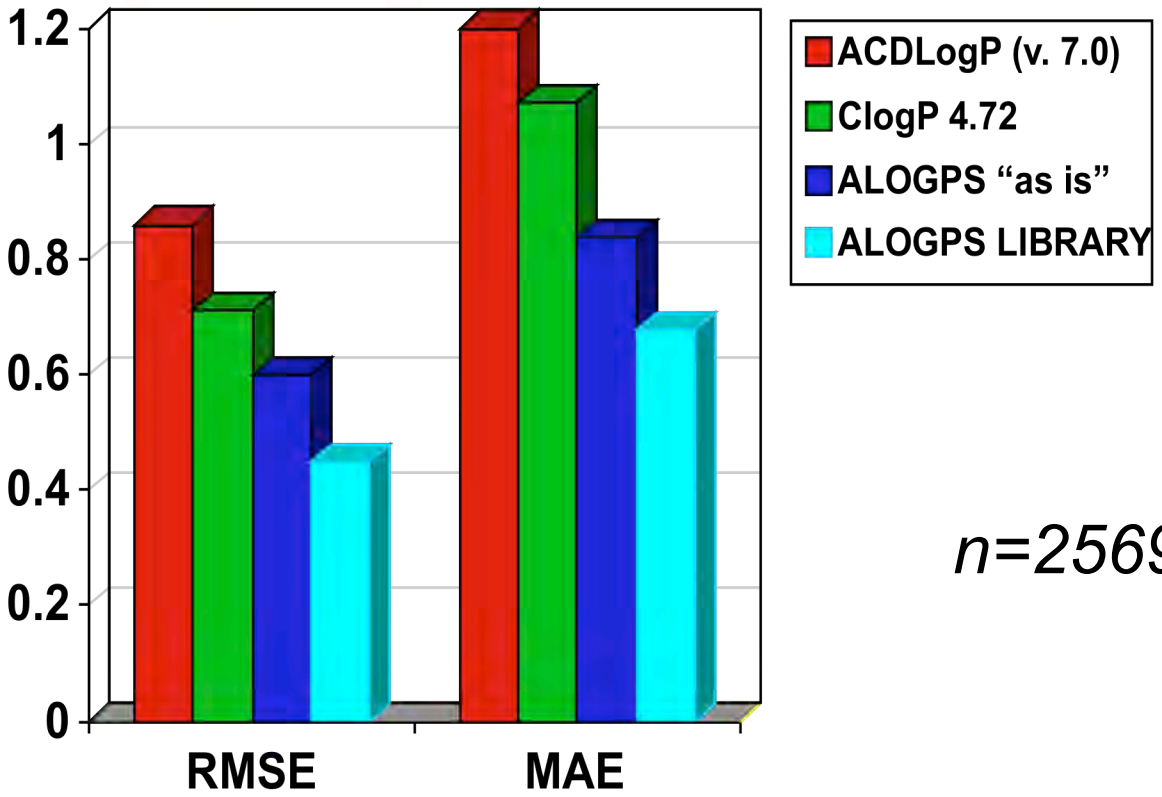
network	training set, *N*	ANN, "nova set" prediction			ASNN, "nova set" used as LIBRARY, LOO		
		RMSE	MAE	outliers	RMSE	MAE	outliers
ALOGPS	**12908**	**0.49**	**0.38**	**68**	**0.43**	**0.32**	**50**
ANN trained on XLOGP set	1853	0.65	0.52	647	0.47	0.36	141
ANN trained on “star” set	9429	0.59	0.47	480	0.46	0.35	98



Notice in “LIBRARY” new data (nova set) were used for the *kNN* corrections without rebuilding neural network models

Tetko & Tanchuk, *JCICS*, **2002**, *42*, 1136-1145.

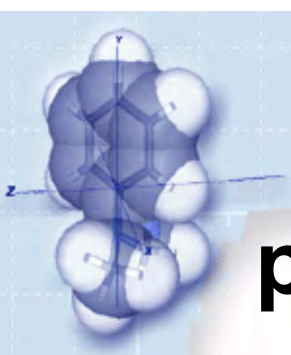
Prediction of AstraZeneca logP set



n=2569

ACDlogP (v. 7.0): *MAE = 0.86, RMSE=1.20*
CLOGP (v. 4.71): *MAE = 0.71, RMSE=1.07*
ALOGPS BLIND: *MAE = 0.60, RMSE=0.84*
ALOGPS LIBRARY: *MAE = 0.45, RMSE=0.68*

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



Can we predict some other similar properties with ALOGPS, e.g. logD?

- logP is defined for neutral compounds
- logD is defined for charged compounds

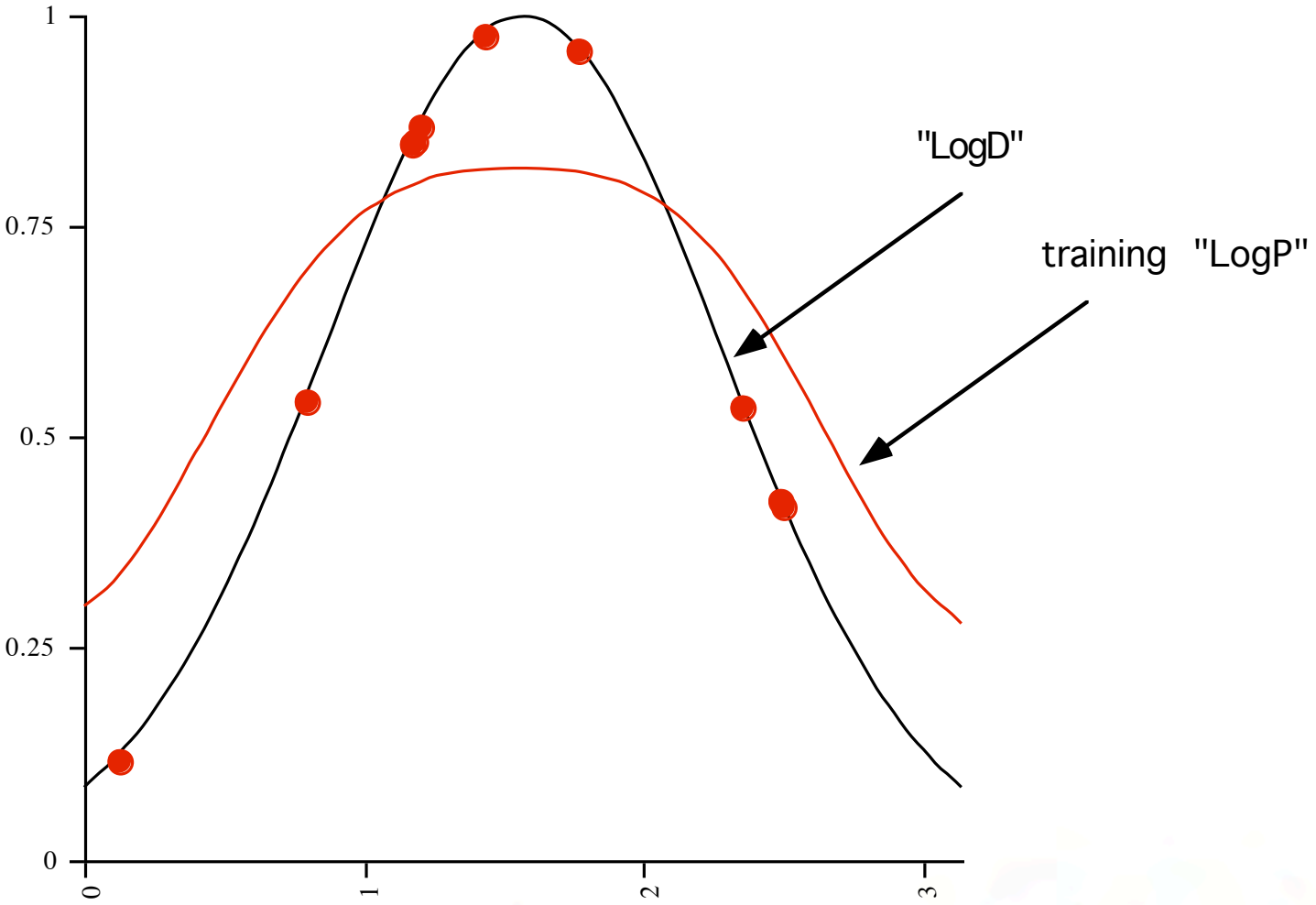
$logD(pH) = logP - log(1+10^{(pH-pKa)\delta i})$,
where δ_i = {1,-1} for acids and bases, respectively

logD is more difficult to predict

a) since it can accumulate errors due to both logP and pKa predictions
b) there is no good compilation of logD data measured for a diverse collection of compounds (however, a lot of data is available commercially)

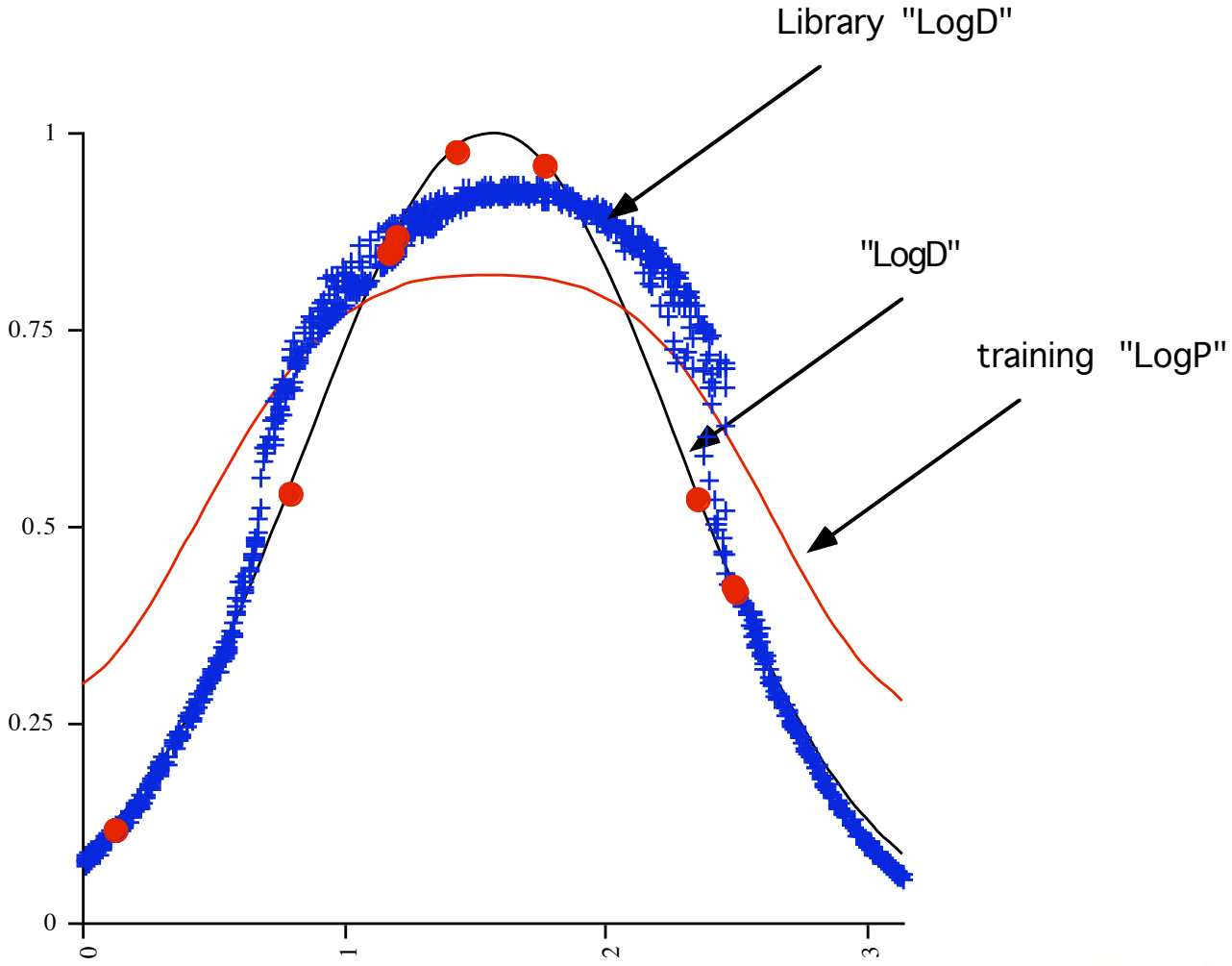
gsf

ASNN: prediction of data that are inconsistent with the training set

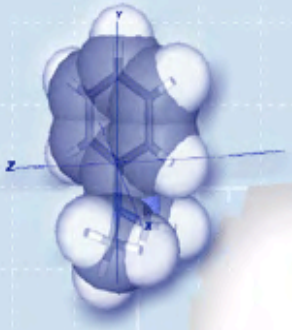


"logP" set -- 100 points; "logD" set -- 8 points, **x**=x_1+x_2

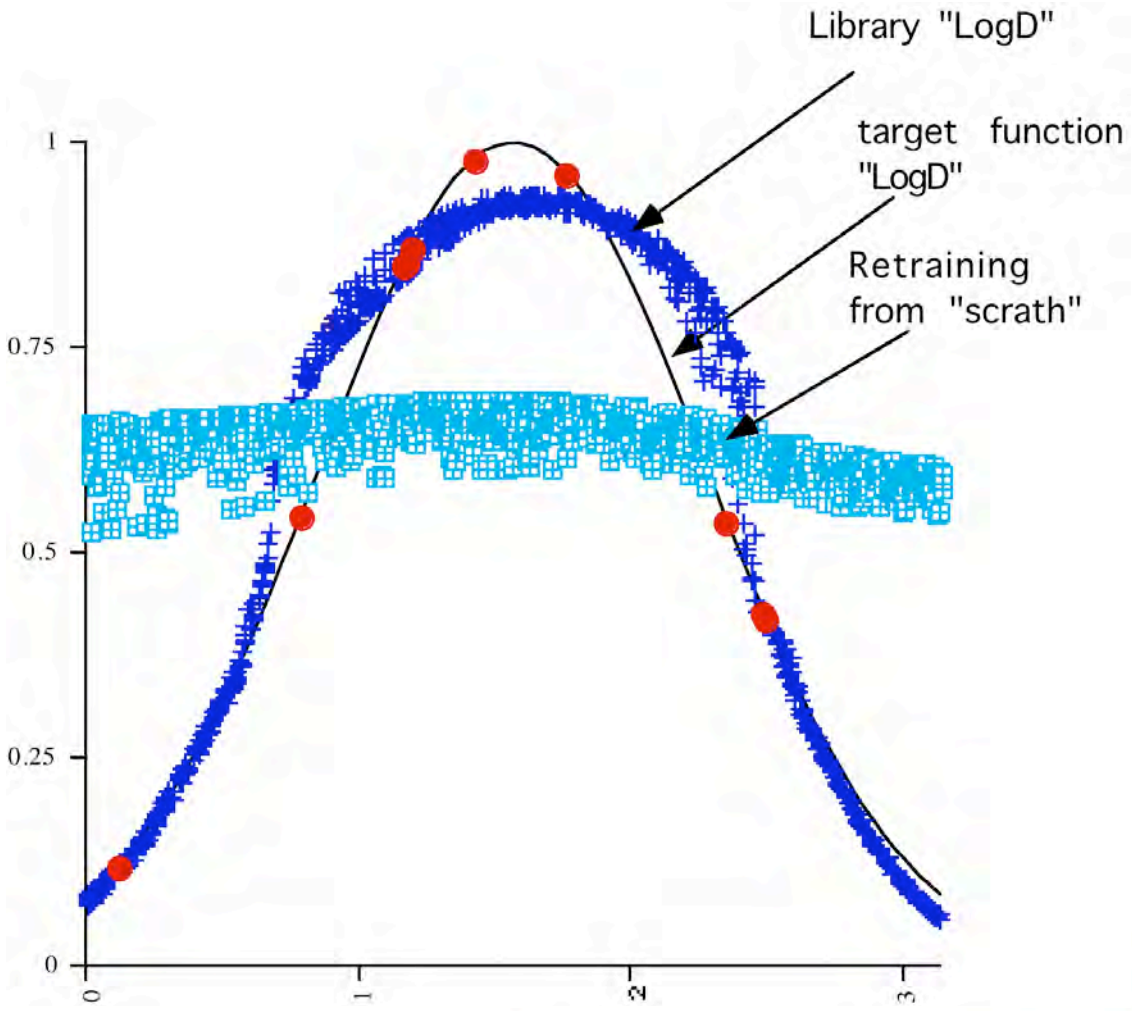
LIBRARY mode (no retraining)



N.B.! Training using both “old” and “new” data will not work!!!



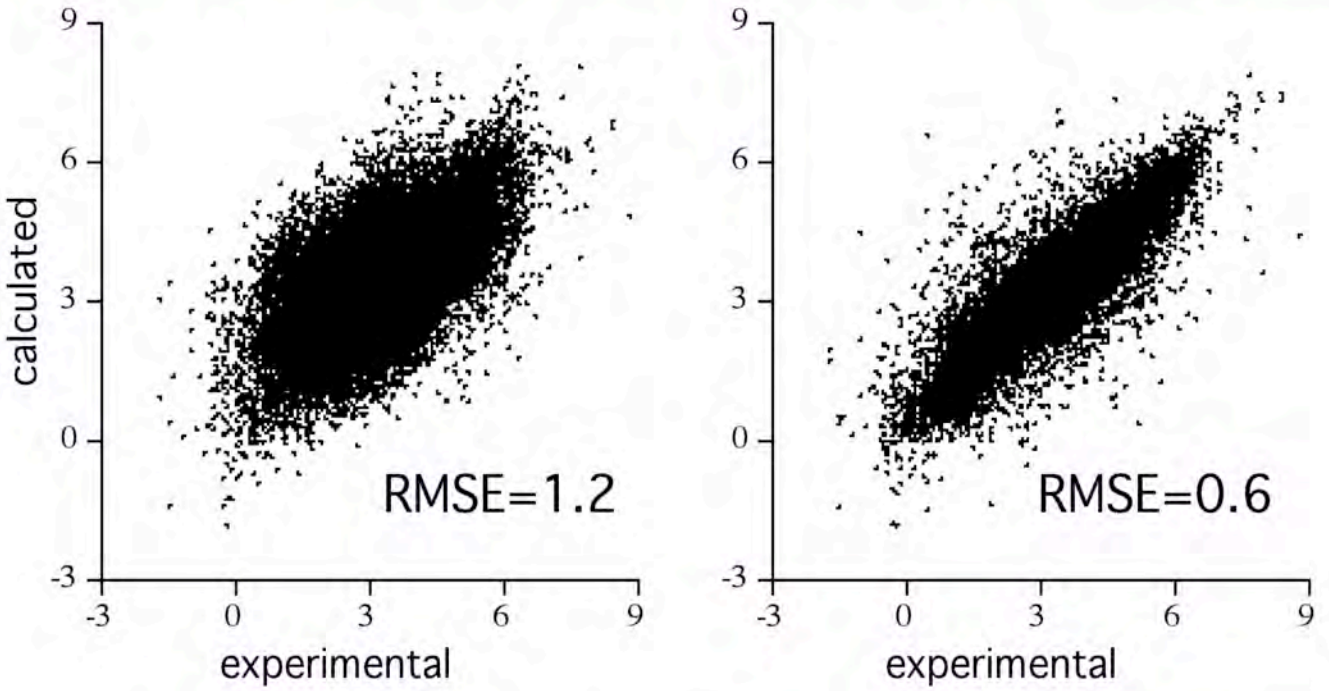
Library mode vs training using “logD” data



gsf

"Self-learning" Pfizer logD data using logP model (LIBRARY)

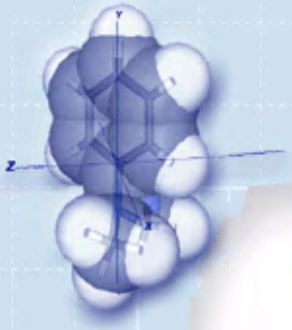
ALOGPS prediction for ElogD set of 17,861 compounds



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.



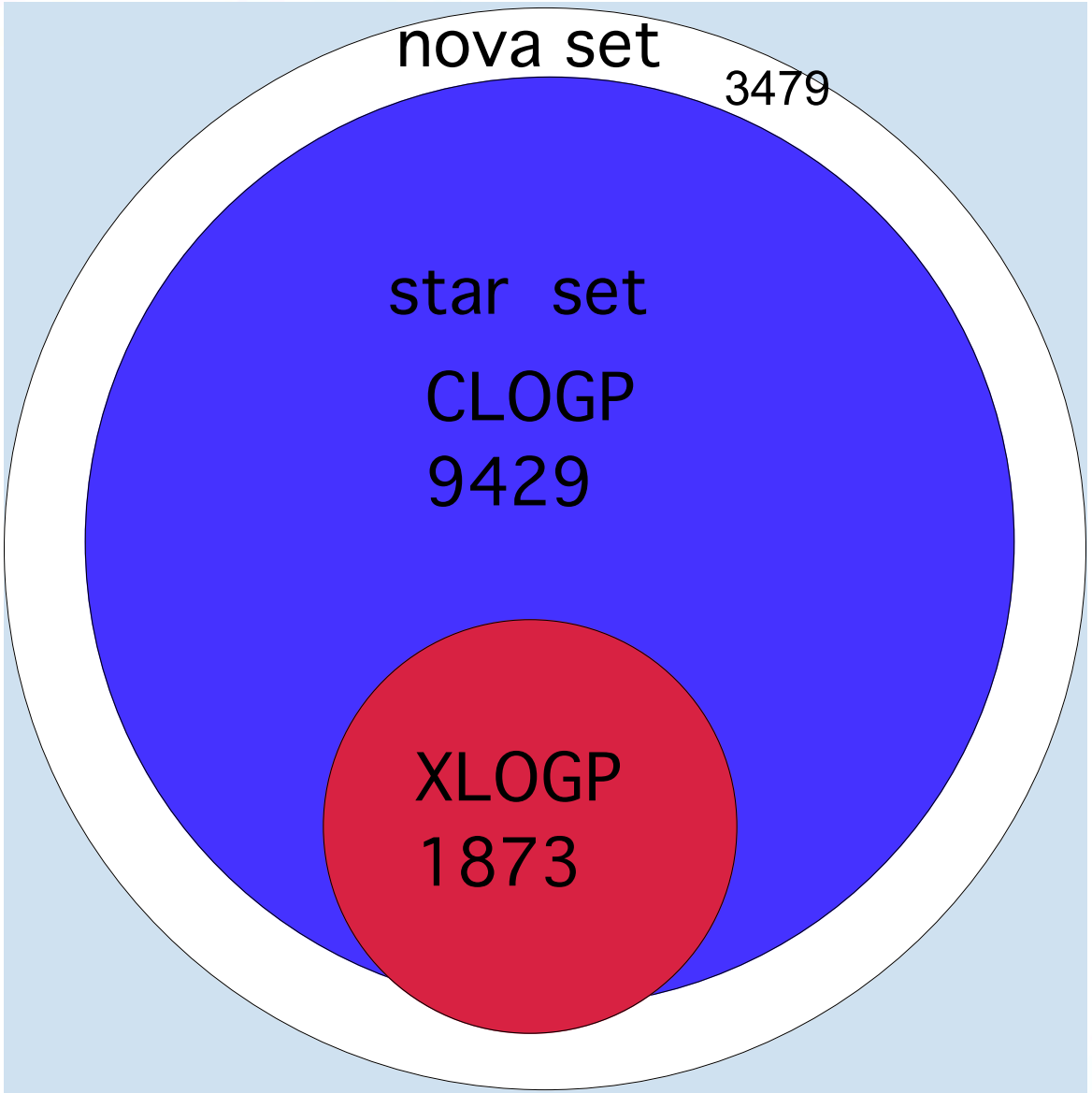
Layout of presentation

- Problems with chemoinformatics models developed by Academia
- Introduction to the Associative Neural Network
- Properties of the ASNN: bias estimation and correction
- ✓ Properties of the ASNN: applicability domain of models
- Properties of the ASNN: secure sharing of data

gsf

PHYSPROP data set

Total: 12908



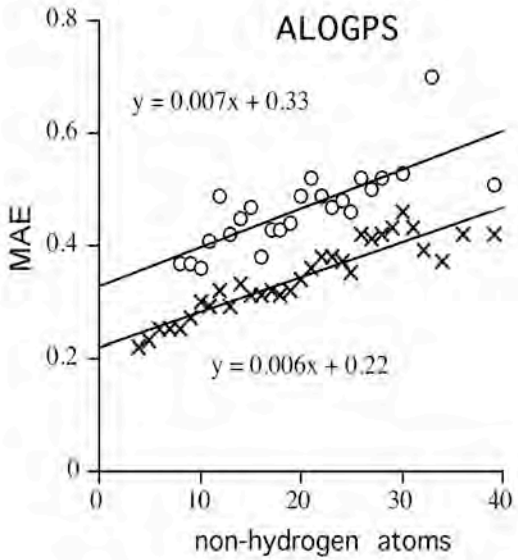
training
“nova” -->
prediction
star set

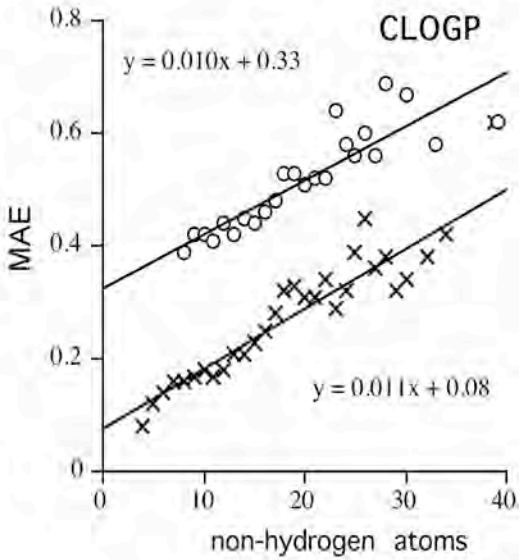
Mean Average Error (MAE) as function of the number of non-hydrogen atoms

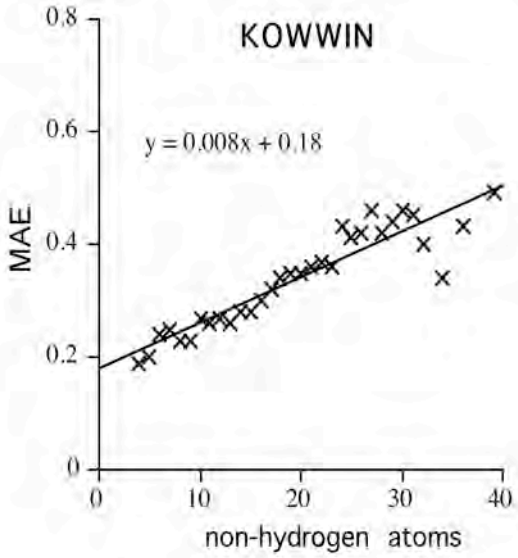
Methods trained using “star” set provided a low prediction ability for the “nova” set because of the different chemical diversity in both sets

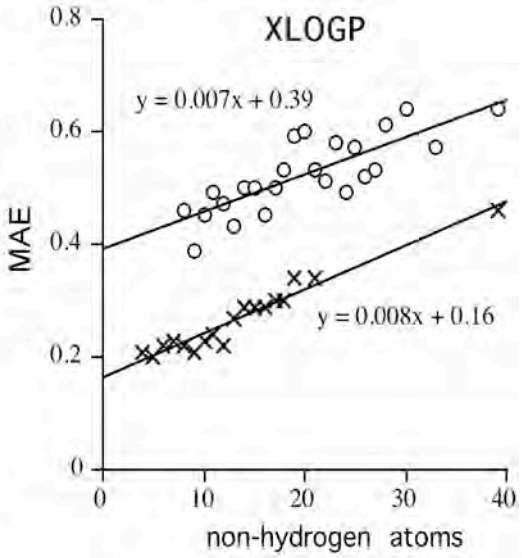
x - training set performance

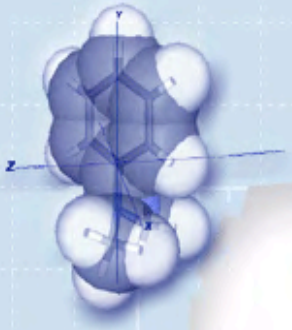
O - test set performance



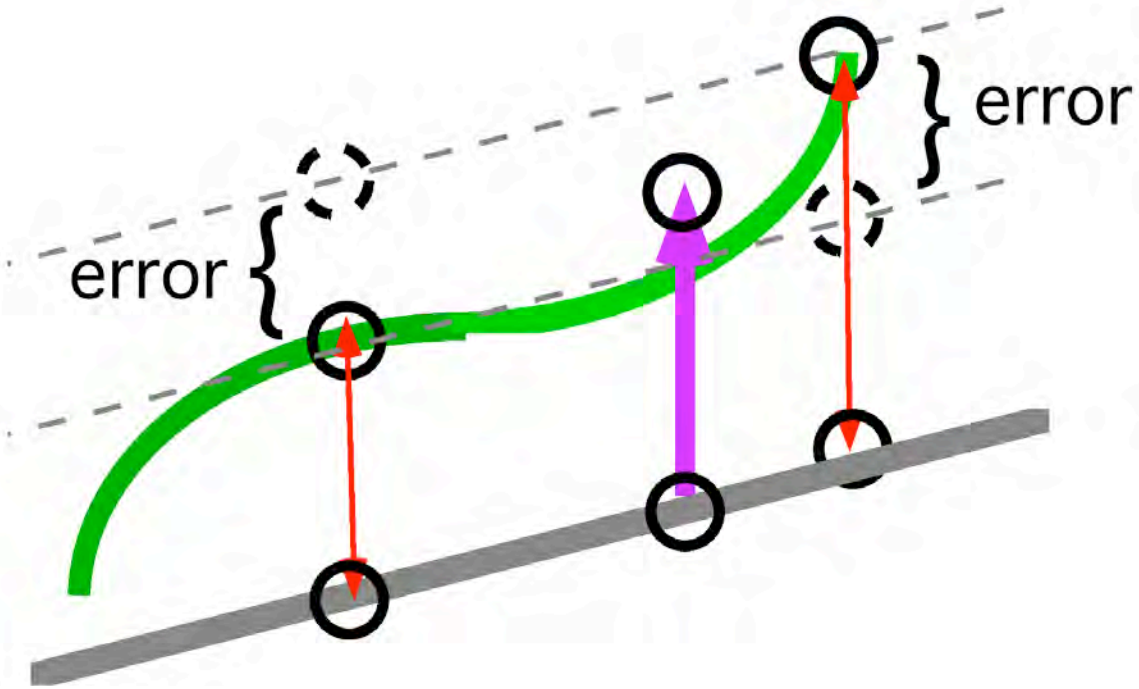








Estimation of the model accuracy by the nearest neighbors

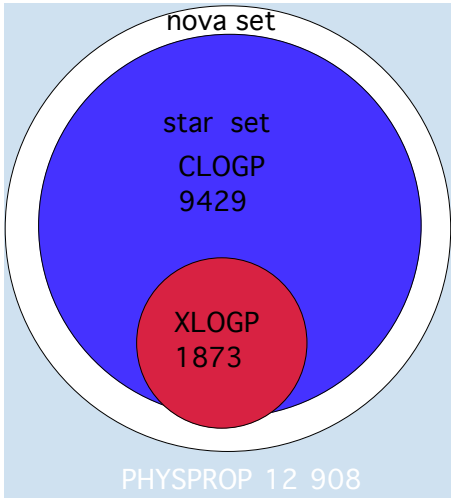


Calibration: For each query molecule we find the most similar molecule (max correlation in the property-based space: “property-based similarity”) in the training set. We plot prediction accuracy of the query molecules as function of their property based similarities (calibration curve).

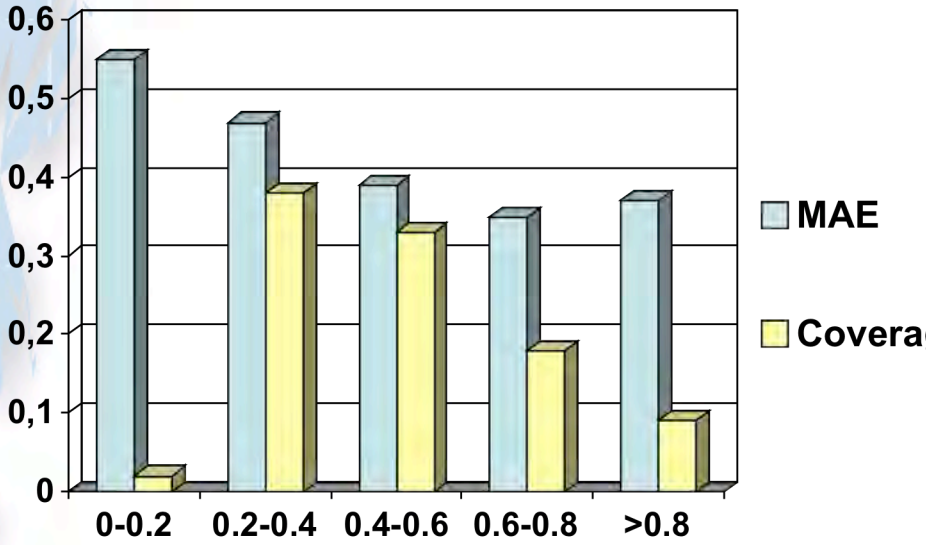
Usage: For each test molecule we again find the most similar molecule and estimate its prediction accuracy using calibrated value for the actual value of “property-based similarity” .

gsf

Prediction performance as a function of the "property-based" similarity



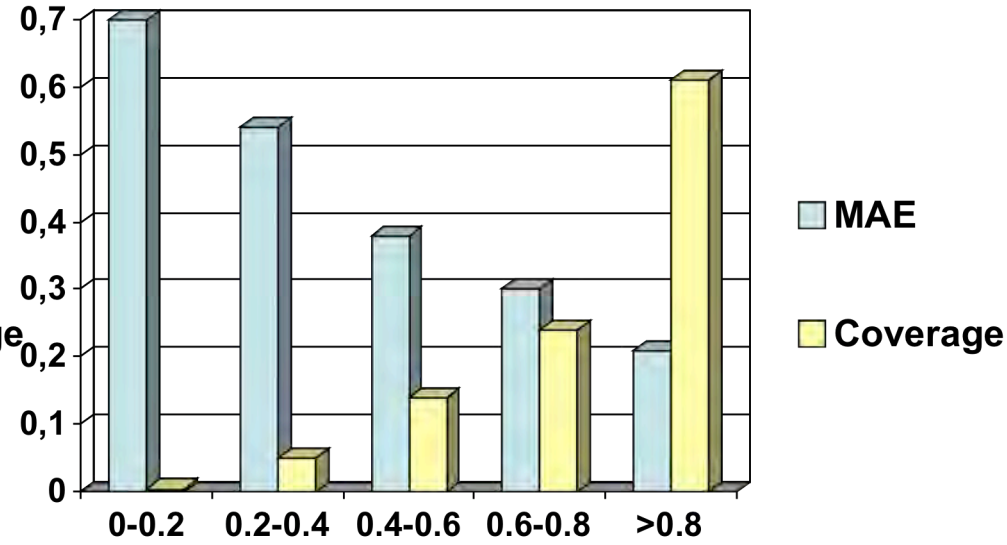
Blind prediction



"property-based similarity":
max correlation coefficient of a test molecule to **"star"** set

MAE=0.43

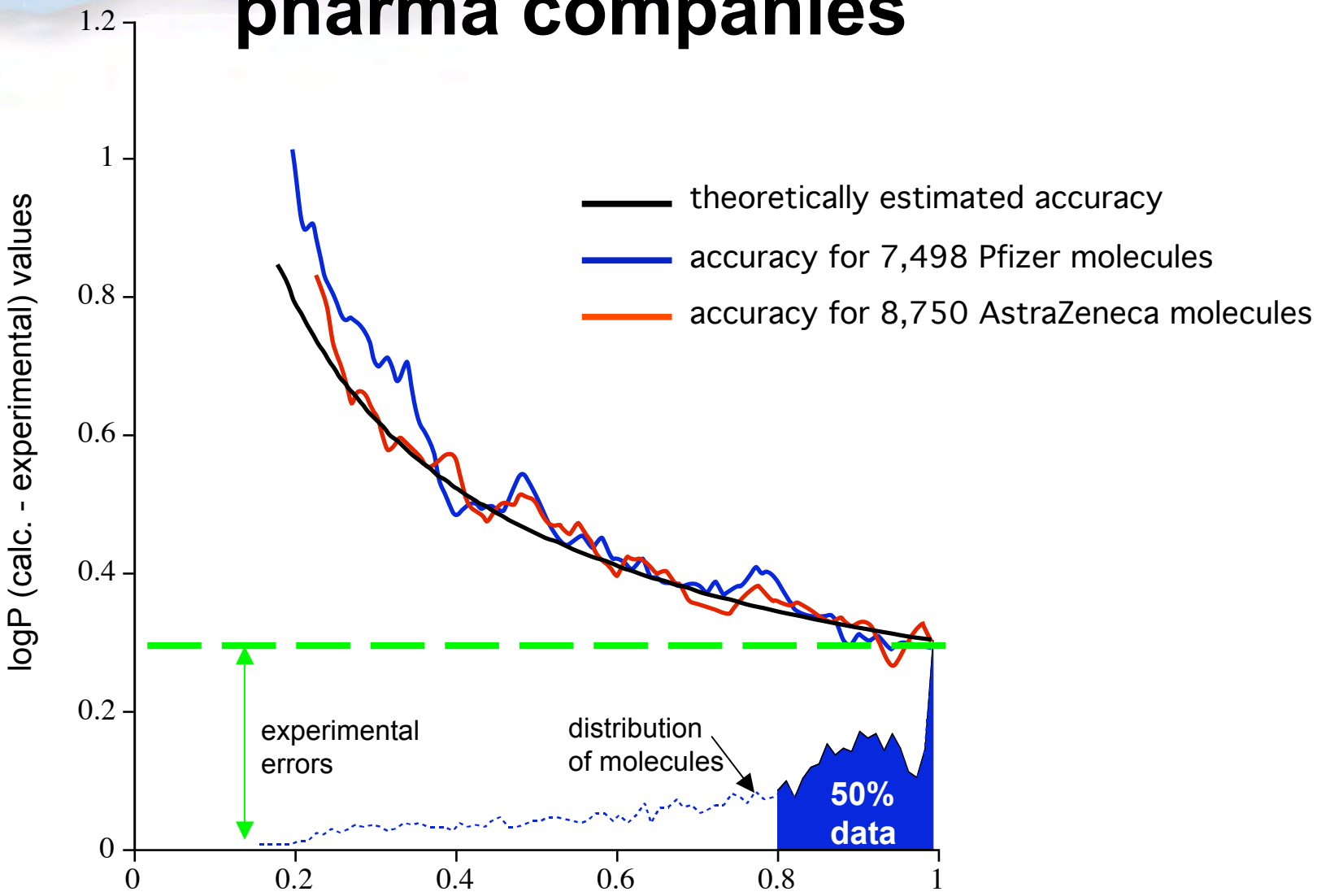
LIBRARY mode



"property-based similarity":
max correlation coefficient of a test molecule to **"star" + "nova"** sets

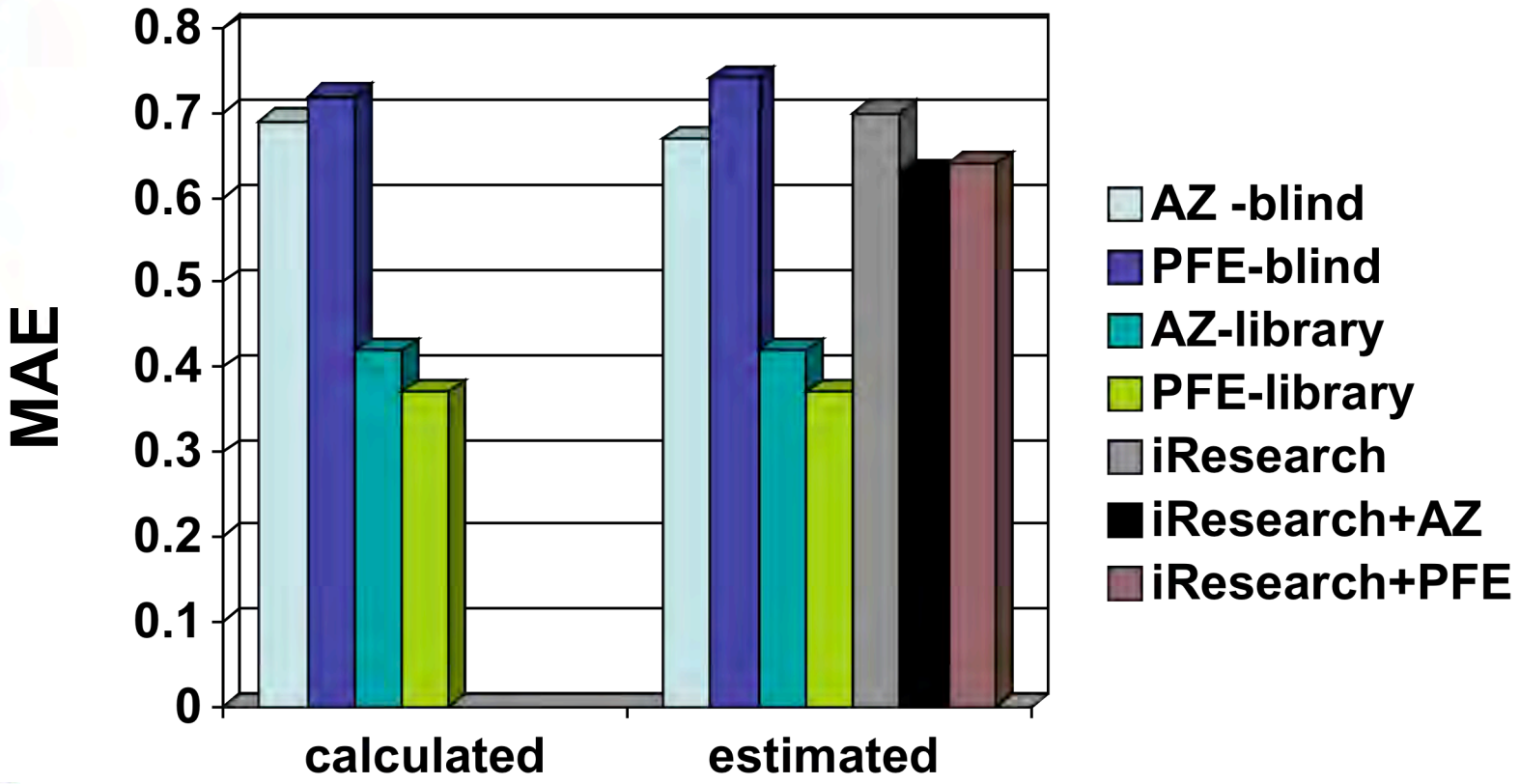
MAE=0.28 (0.26)

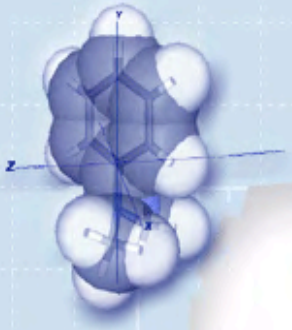
Lipophilicity (logP) prediction of pharma companies



- We can reliably estimate which compounds can/can't be reliably predicted.
- ✓ ASNN can save costs for measurements of up to 50% of molecules.

Estimated and calculated error for AstraZeneca (AZ), Pfizer (PFE) and iResearch Library sets (>1.5*10^7 molecules)

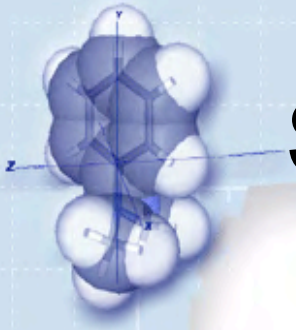




Layout of presentation

- Problems with chemoinformatics models developed by Academia
- Introduction to the Associative Neural Network
- Properties of the ASNN: bias estimation and correction
- Properties of the ASNN: applicability domain of models
- ✓ Properties of the ASNN: secure sharing of data

gsf

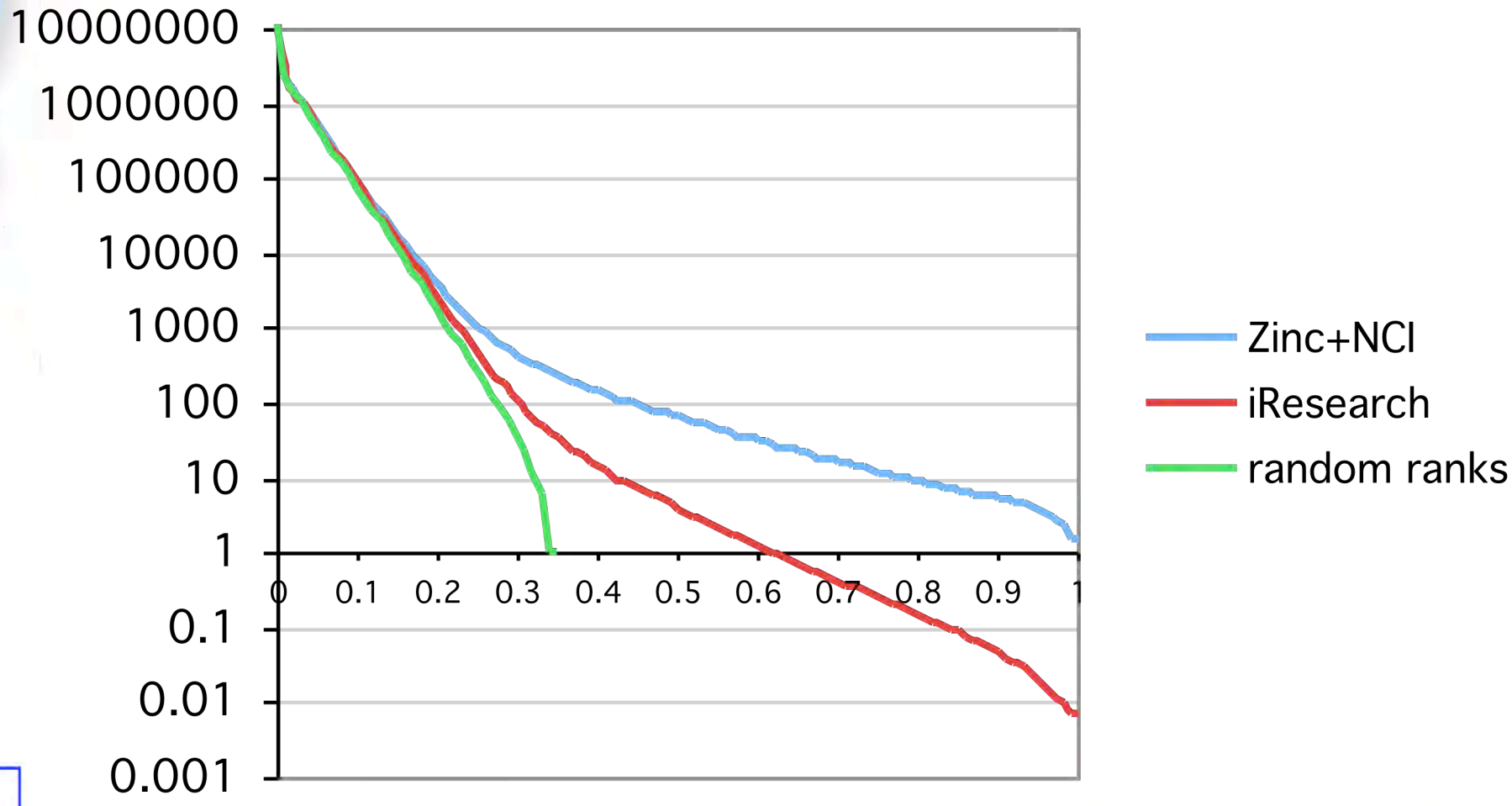


Secure sharing of information but not molecules

- Symposium organized by T. Oprea at 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 ... but only until it is “hacked”
- Sharing molecular descriptors of a target molecule is a difficult business
- But …. let us share reliably predicted molecules!
- These are the molecules with significant high ***R*** in property space to the target molecule

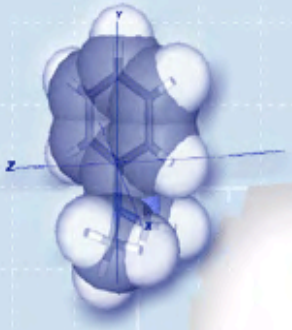
gsf

Number of molecules as a function of the property-based similarity

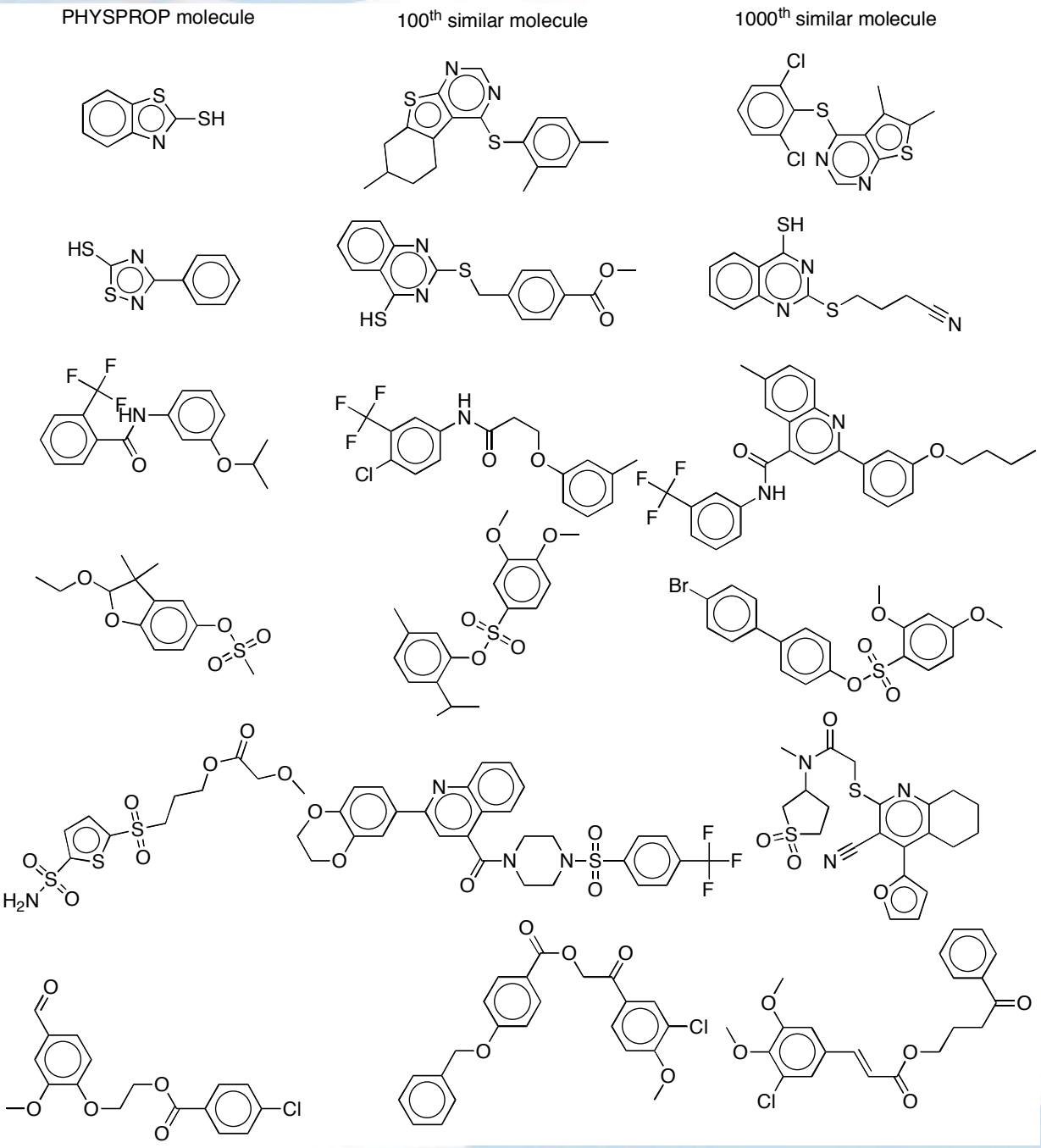


The average number of molecules in respective databases with higher or equal correlation coefficient to a molecule in the PHYSPROP database

Real and surrogate molecules



PHYSPROP molecule	100th similar molecule	1000th similar molecule

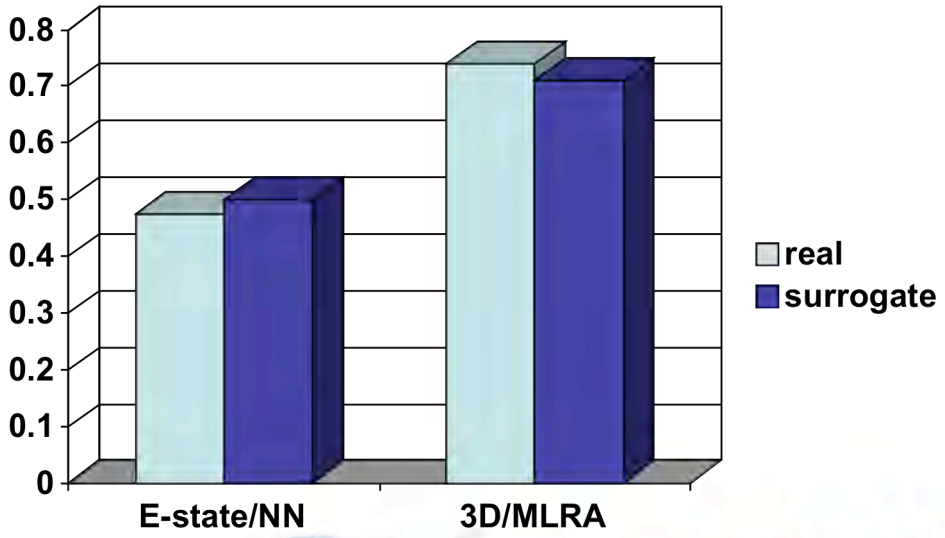


Tetko, Abagyan,Oprea
J. Comp. Aid. Mol. Des.
2005,19, 749.

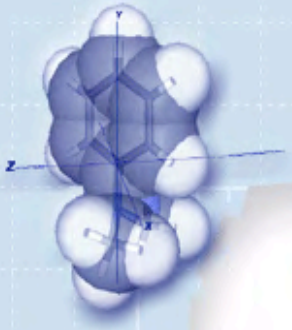
Models for logP prediction developed with real & surrogate data

- Take a “real” molecule from PHYSPROP logP dataset
- Find for it 100th (1000th) 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” to develop models
- Predict all 12908 PHYSPROP molecules using both models

Dataset sizes
Real = surrogate =
1949 molecules



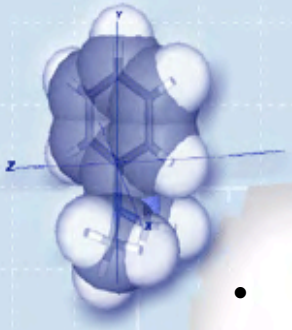
N.B.! This is a property-specific data sharing!!!



Conclusions

- Use of ASNN approach allows to
  - Estimate bias of the model
  - Correct bias of the model
  - Improve accuracy of predictions
  - Instantly update models with new data without retraining global (neural network) model
    - Drug discovery
    - Control systems (movement)
  - Allow to estimate applicability domain and accuracy of prediction of models
    - novelty detection
    - detection of outlying data points
    - detection of non-stationary measurements
    - experimental design
    - secure data sharing

gsf

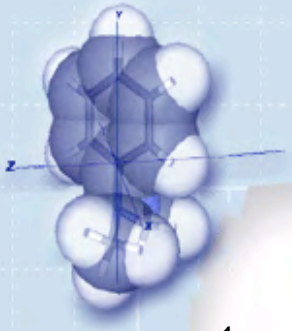


Neuro-physiological roots

- Spatio-temporal coding of information processing
- A lot of correlations in the brain, fast reaction (100s ms for visual stimulus)

- A signal with maximum correlation to the stimulus from the long-term memory will be reinforced
- Some part of the brain provides modulation (increase of excitability) of regions that will be used for search of the proto-types and thus switching the “property-based similarity” depending on the context of the query or/and desired action
- **Explains presence of two levels of behavior**
  - Long term skills (fast reaction, parallel processing of information -- **procedural memory**)
    - Genetically programmed skills
    - Skills developed on the level of vegetative neural system
  - Short term skills (sequential processing of information -- **declarative memory**)
    - Skills developed by learning from few examples, cognition, observation of similar situation
    - Used to correct the long term skills
- Following training declarative memory changes to the procedural memory

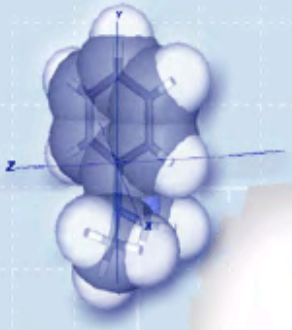
gsf



Literature

1. Tetko, I. V.; Villa, A. E. P. In Unsupervised and Supervised Learning: Cooperation toward a Common Goal, ICANN'95, International Conference on Artificial Neural Networks NEURONIMES'95, Paris, France, 1995; F., F.-S., Ed. EC2 & Cie: Paris, France, 1995; pp 105-110.
2. Tetko, I. V.; Villa, A. E. P., Efficient partition of learning data sets for neural network training. Neural Networks 1997, 10, (8), 1361-1374.
3. Tetko, I. V., Associative Neural Network, CogPrints Archive, cog00001441. 2001.
4. Tetko, I. V., Associative neural network. Neural Processing Letters 2002, 16, (2), 187-199.
5. Tetko, I. V., Neural network studies. 4. Introduction to associative neural networks. J Chem Inf Comput Sci 2002, 42, (3), 717-28.
6. Tetko, I. V.; Tanchuk, V. Y., Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J Chem Inf Comput Sci 2002, 42, (5), 1136-45.
7. Tetko, I. V.; Poda, G. I., Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. J Med Chem 2004, 47, (23), 5601-4.
8. Tetko, I. V.; Bruneau, P., Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. J Pharm Sci 2004, 93, (12), 3103-10.
9. 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, (15-16), 700-707.

gsf



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 (USA) for providing compounds from the iResearch Library (ChemNavigator).

Thank you for your attention!

gsf