Associative Neural Network

Igor V. Tetko

Institute for Bioinformatics,

GSF - Forschungszentrum fuer Umwelt und Gesundheit, GmbH, Ingolstaedter Landstrasse 1, D-85764 Neuherberg (Munich), Germany

itetko@vcclab.org

Supervised regression methods

- Memoryless: **multiple linear regression analysis, neural networks, polynomial neural networks, usually these are** global models
- Memory-based: **k-nearest neighbours (KNN), Parzen-window regression, memory-based reasoning, usually these are** local models

Associative Neural Network (ASNN)

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

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 o 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

$$[\mathbf{x}_i]\bullet[\mathbf{ANNE}]_{10}=\begin{bmatrix}0.5\\0.3\\0.4\\0.5\\0.6\\0.7\\0.5\\0.4\\0.5\\0.6\end{bmatrix},\begin{bmatrix}0.9\\0.8\\0.7\\0.9\\0.8\\0.9\\0.8\\0.7\\0.5\\0.4\end{bmatrix},\begin{bmatrix}0.2\\0.1\\0.3\\0.1\\0.1\\0.3\\0.2\\0.3\\0.2\\0.1\end{bmatrix}$$

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

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

Test case:

$$[\mathbf{x}_t]\bullet[\mathbf{ANNE}]_{10}=\begin{bmatrix}0.7\\0.4\\0.4\\0.6\\0.7\\0.8\\0.9\\0.7\\0.4\\0.6\end{bmatrix}$$

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

ASNN result:

$r(x_1,x_t)=0.55$

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

$r(x_3,x_t)=0.16$

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

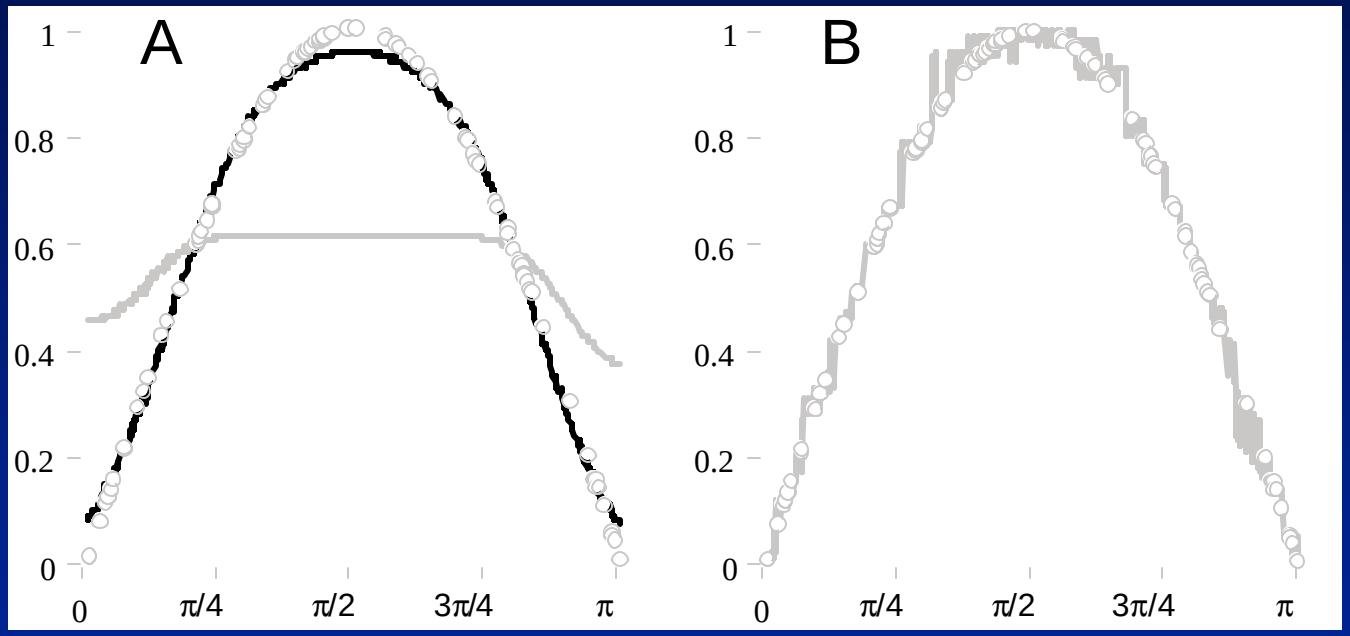
$$=>\bar{z}_t'=0.62+\frac{1}{2}\left((0.4-0.5)+(0.6-0.74)\right)$$

$$=0.62-0.12=0.5$$

Interpolation of $y=sin(x=x_1+x_2)$

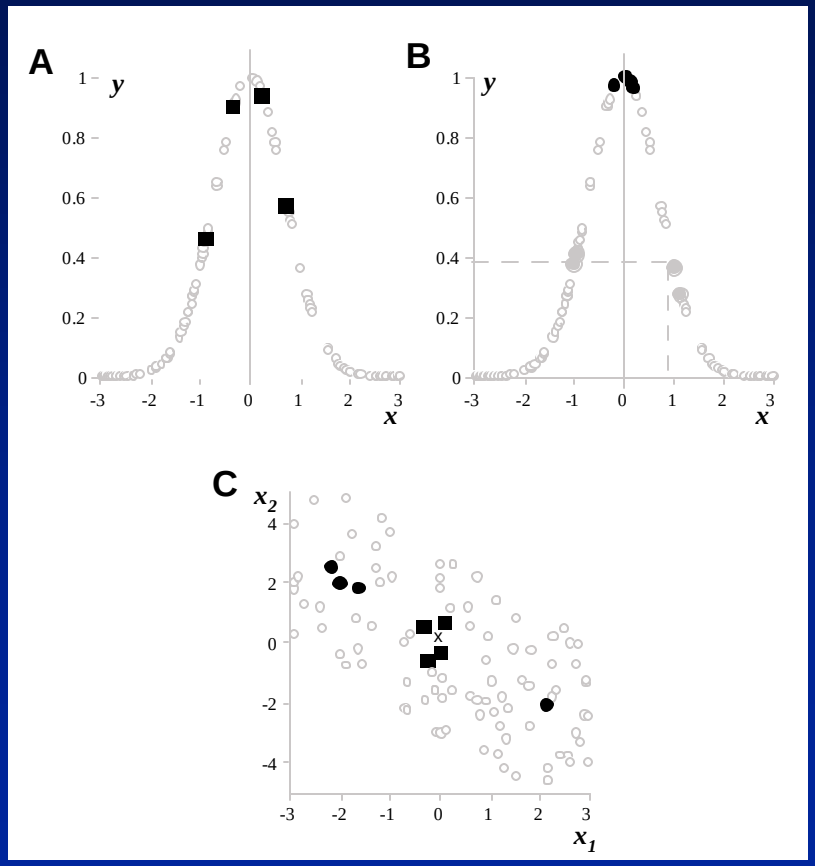
Simple ensemble average

ASNN (one hidden neuron)



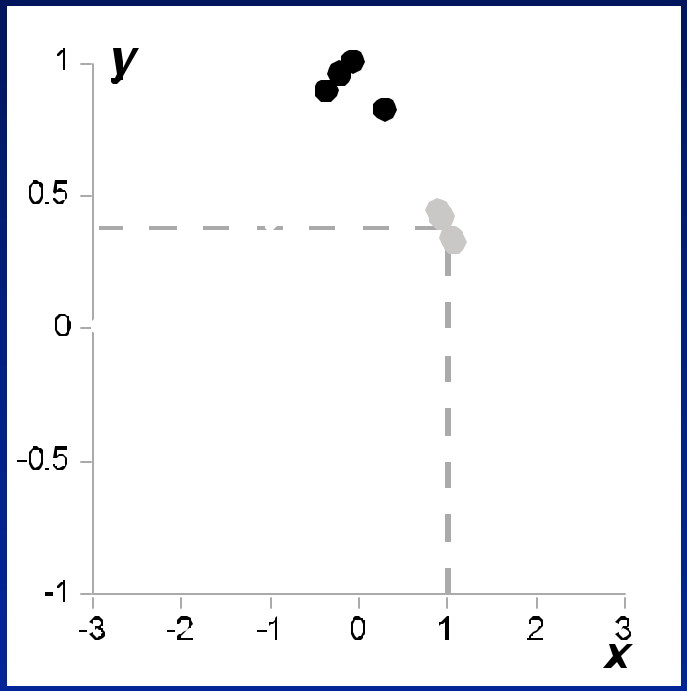
Gray (black) line corresponds to neural networks with one (two) hidden neurons. The bias problem (underfitting) is more prominent for one-hidden neuron networks. ASNN dramatically decrease bias of the network prediction.

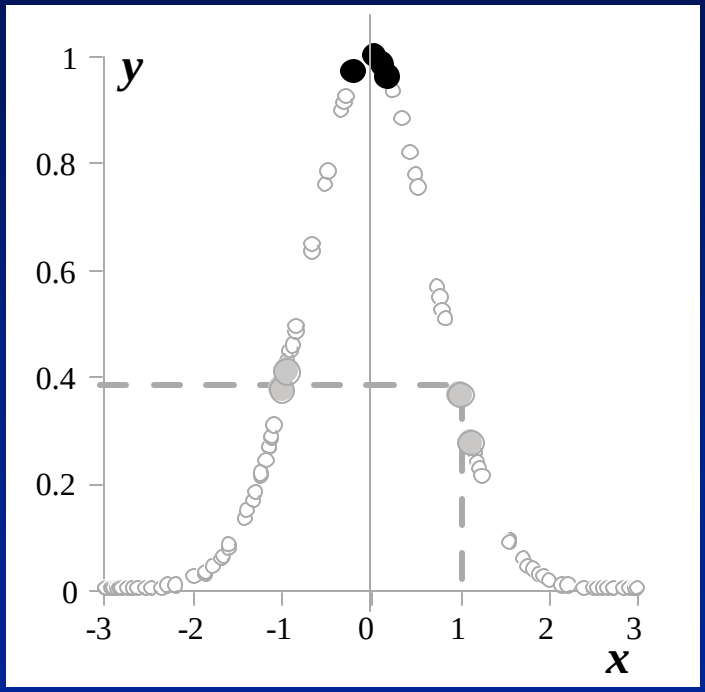
Similarities in input/output space



$Y=Gauss(x_1+x_2)$

Similarities of symmetric & non-symmetric functions



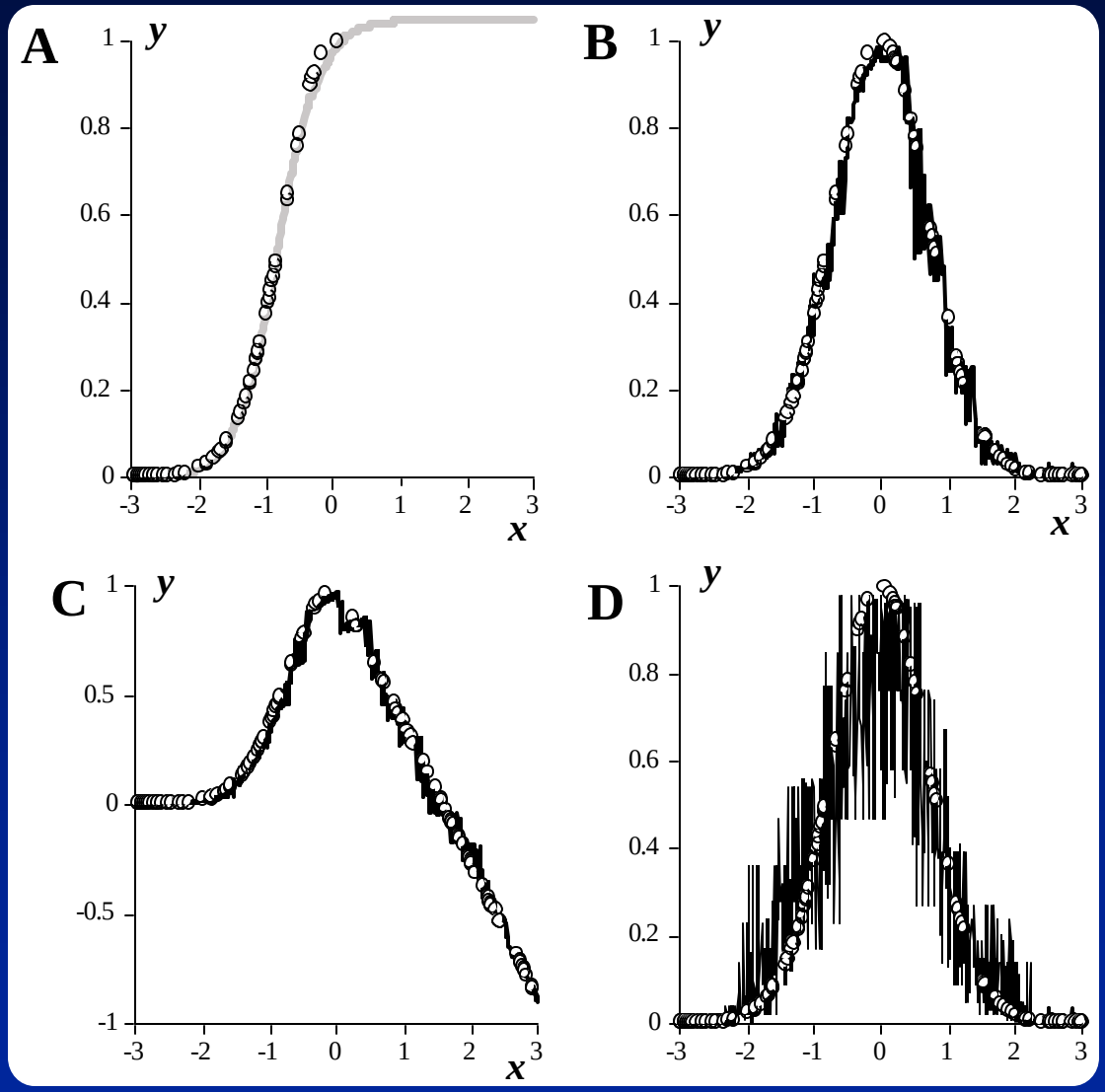


Nearest neighbors of case $(x_1,x_2)=(0,0)$ are shown as black circles.
Nearest neighbors of case $(x_1,x_2)=(1,0)$ are shown as gray circles.

Gauss function interpolation with fresh data

Features:

fast, no weights retraining; correction is not limited by the range of values in the training set.



N.B! KNN in the output space works better, since it takes into account invariance $x=x_1+x_2$!

ALOGPS - program to predict lipophilicity (logP) and aqueous solubility (logS) of chemicals

LogP: 75 input variables corresponding to electronic and topological properties of atoms (E-state indices), 12908 molecules in the database, 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)

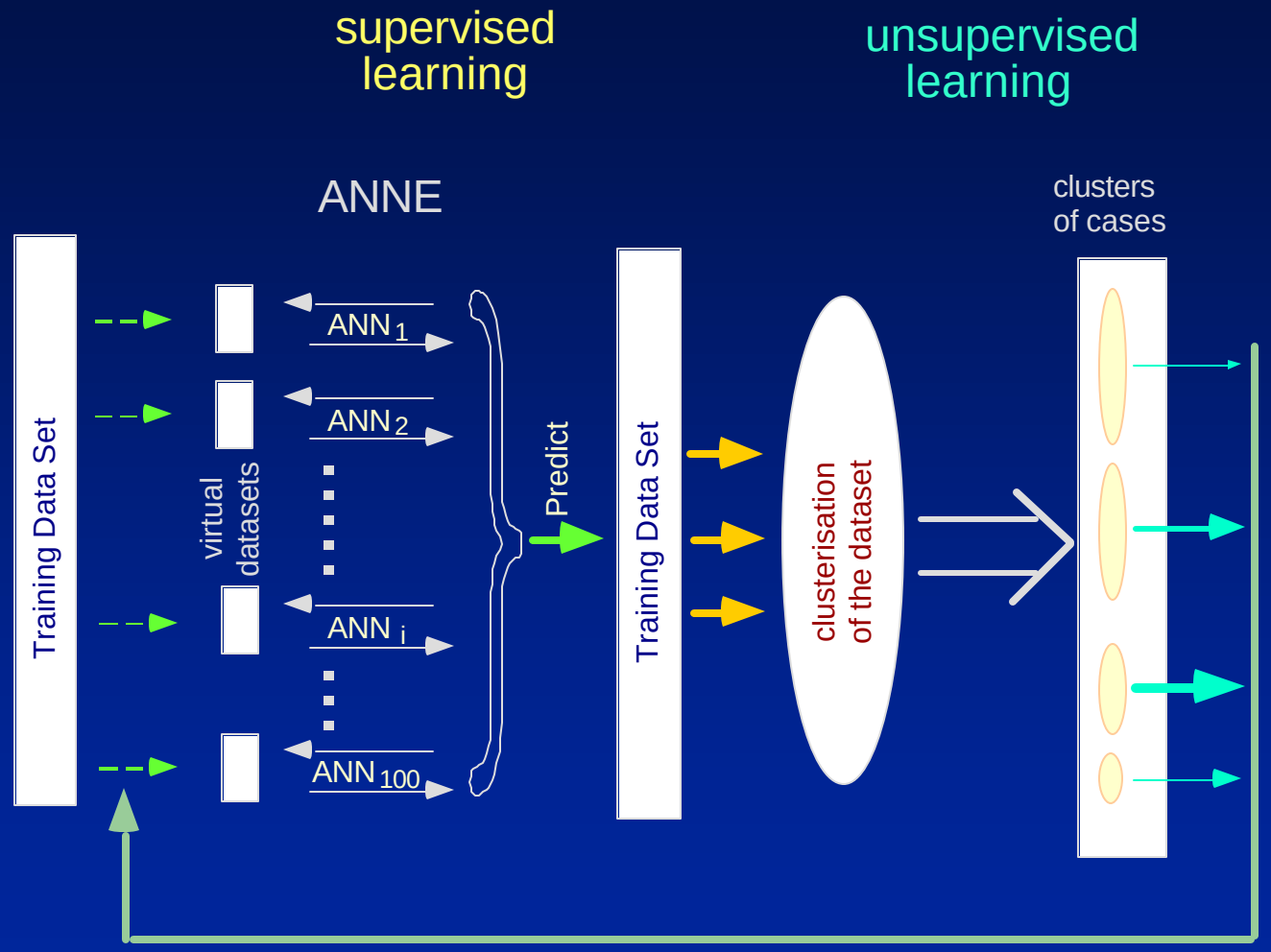
Tetko et al, JCICS, 2001, 41, 1488-1493 & 1407-1421

Percentage of molecules within indicated error range for lipophilicity prediction

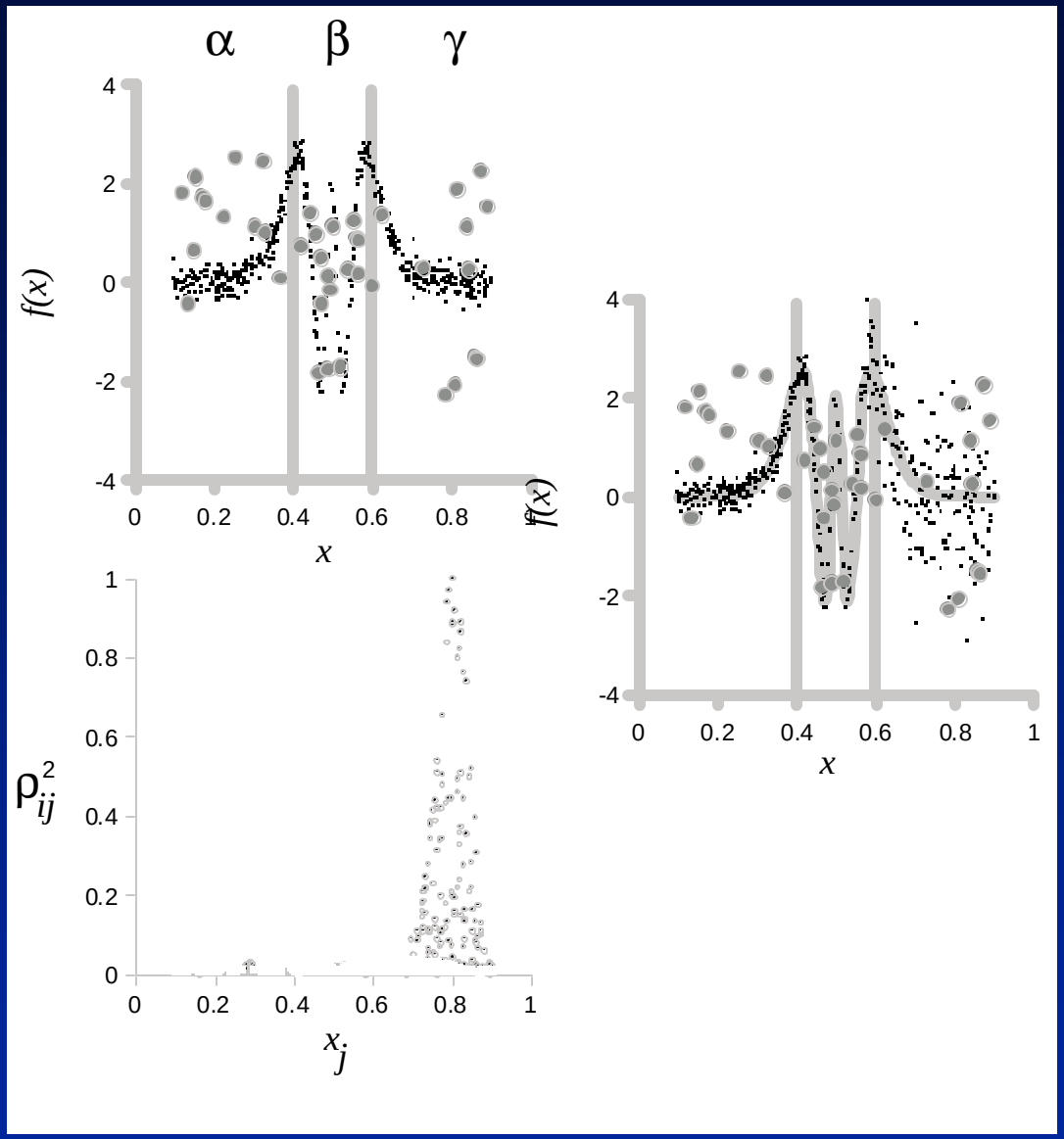
$\lvert logP_{pred} - logP_{exp} \rvert$	LOO for the training set	BASF, 6100 "as is"	BASF, 6100 LOO[1]
0--0.3	63%	30%	60%
0--0.5	81	49	80
0--1.0	96	82	96
0--2.0	99	98	99

[1]Tetko, 2002, JCICS, 42, 717-728.

What are the Roots of ASNN?
Efficient Partition Algorithm!



Tetko & Villa, ICANN'95, and Neural Networks, 1997



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

ASNN & logP

- **More theoretical articles:**
- Tetko, I.V. Neural Network Studies. 4. Introduction to Associative Neural Networks, *J. Chem. Inf. Comput. Sci.*, 2002, 42, 717-728.
- Tetko, I.V. Associative Neural Network, *Neural Processing Letters* 2002, 16, 187-199.
- Tetko, I.V.; Villa, A. E. P. Efficient Partition of Learning Datasets for Neural Network Training, *Neural Networks* 1997, 10, 1361-1374.
- **More applied one:**
- Tetko, I.V.; Tanchuk, V. Yu. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 program, *J. Chem. Inf. Comput. Sci.,* 2002, in press.

- **These articles + posters are available at http://vcclab.org/lab/pdf**
- **ASNN is available at http://vcclab.org/lab**

Acknowledgement

Part of this presentation was done during my work in the University of Lausanne (Switzerland), Institute for Bioinformatics, MIPS (Germany) and also thanks to the Virtual Computational Chemistry Laboratory INTAS-INFO 00-0363 project.

I thank Prof. Hugo Kubinyi for testing ALOGPS program at BASF, R. Borisyuk, I. Litvinyuk for their remarks and Prof. F. Masulli for an opportunity to participate to the school.

And I am very grateful to M.J. Castro Bleda, W. Diaz Villanueva and J.L. Dominguez Rubio who agreed to switch my poster (no. 15) with their poster (no. 2).

Thank you for your attention!