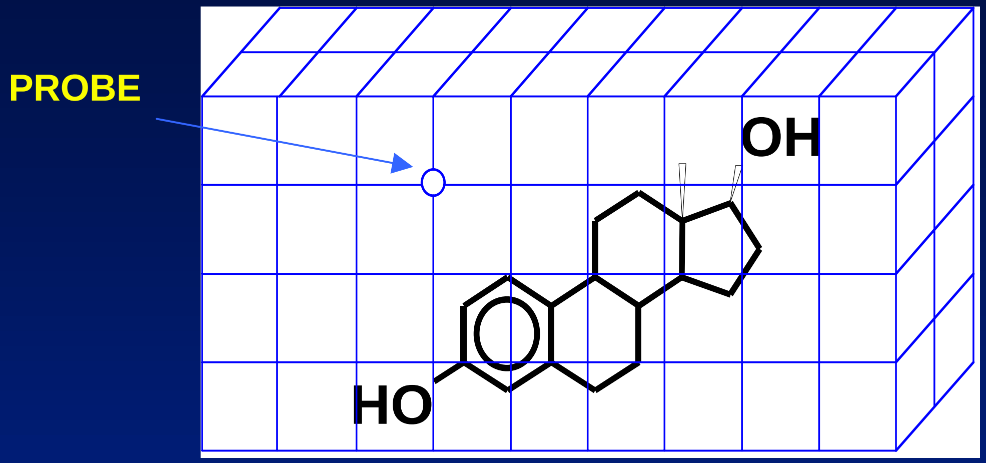
Volume Learning Artificial Neural Network

Igor V. Tetko[1,2] & Vasyl V. Kovalishyn[2]

IBPC, Ukrainian Academy of Sciences,
Murmanskaya 1, Kiev-660, Ukraine & Institute
for Bioinformatics, MIPS, GSF , Ingolstaedter
Landstrasse 1, D-85764 Neuherberg (Munich),
Germany

itetko@vcclab.org

CoMFA



Number	BioActivity	S001	S002… S9999	E001… E9999
M1	9.87			
M2	6.15			
Mn	0.03			

CoMFA converts 3D lattice into table. The first column contains biological activities. The remaining columns are parameters corresponding to the energies of ‘PROBE’ interactions with the molecule at the lattice of points around it. Usually there are 30-100 cases and 1,000-20,000 parameters.

PLS analysis of CoMFA data

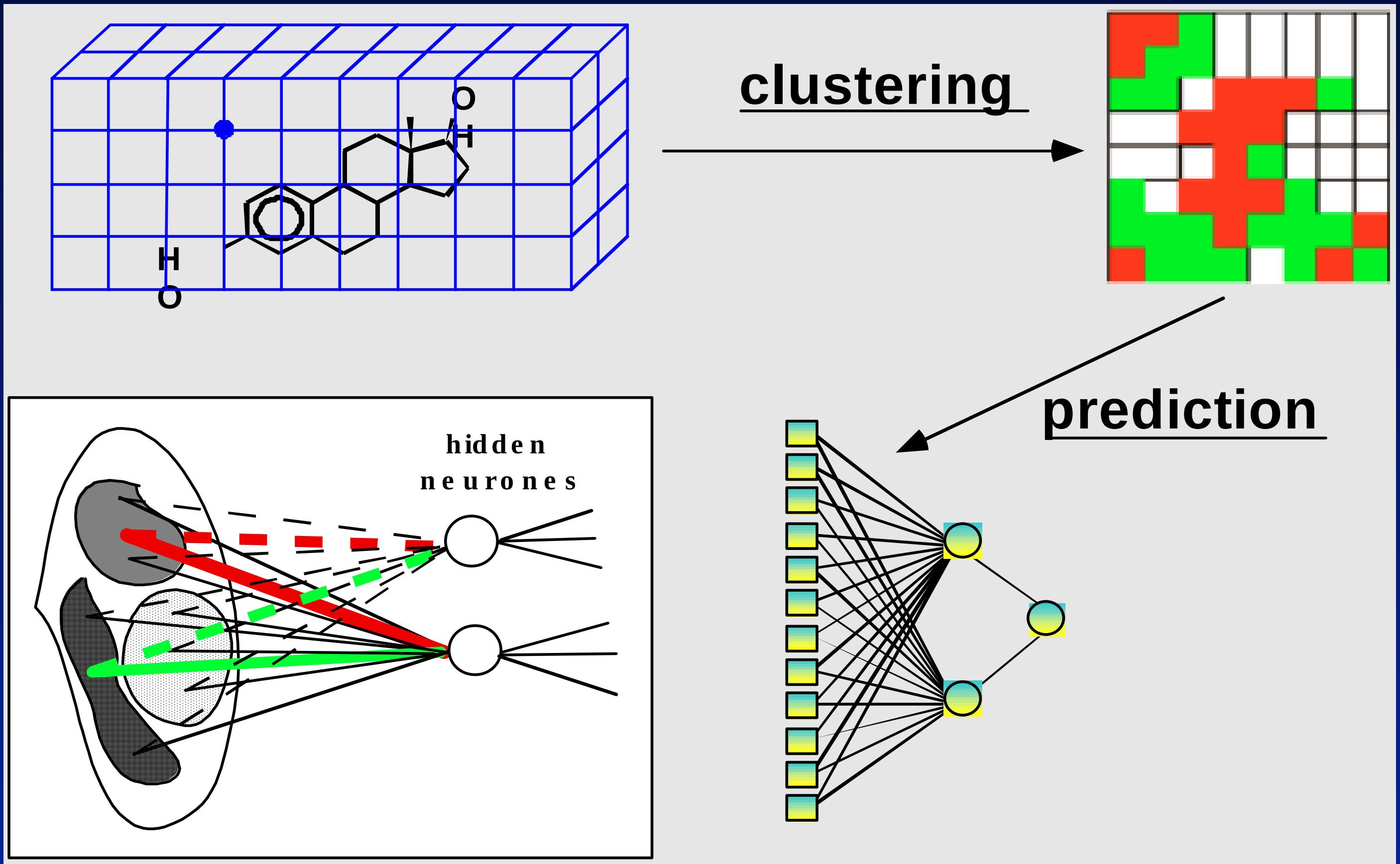
- X is matrix of input variables (kxm), Y is response vector (k)
- $Y=XB+E$ -- standard regression does not work, since the number of samples $k<<m$ parameters. Here E is noise and B are the regression coefficients.

- $Y=TQ+E$ -- regression model, Q is matrix of regression coefficients (loadings) and T is matrix of so-called latent variables (kxh) such as $h<k$
- $T=XW$, W is weight matrix that is computed to maximise the covariance between response vector Y and latent variables T

⇒PLS reduces dimension of input space (X=>T) and performs MLRA analysis. A similar idea can be also used for neural networks!

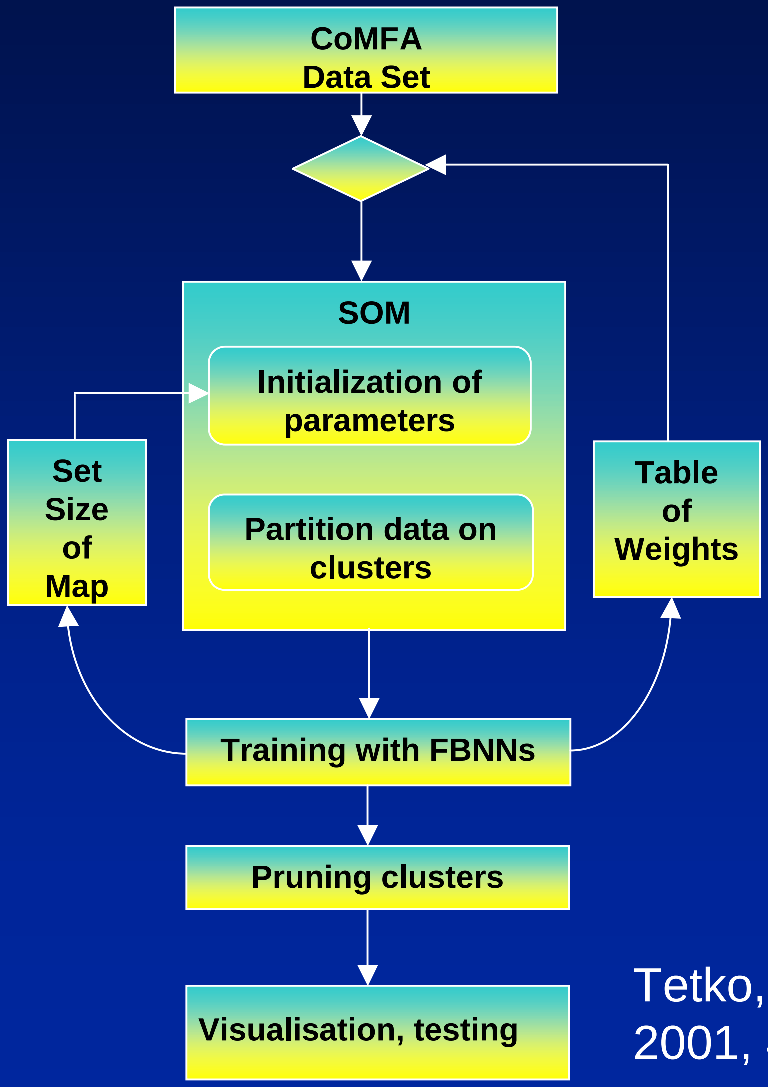
PLS references: look for articles of S. Wold + PLS in PubMed!

Volume Learning Algorithm



The Volume Learning Algorithm (VLA) uses SOM to cluster input data. The average cluster values are used to train supervised neural network.

Data Analysis in VLA



- 1) Kohonen’s Self-Organizing Map is used to find clusters in the input data.
- 2) The centres of clusters are used as inputs for FBNN neural nets and the optimal clustering of input space is detected.
- 3) The optimised centres of clusters are used to develop the final model and to predict new test patterns

Tetko, Kovalishyn, Livingstone, J. Med. Chem, 2001, 44, 2411-2420; Kovalishyn et al., ANNIE’2002

Some formulas

k – number of samples $(\mathbf{x}_{\mathrm{i}})$, m - dimension of each sample

initial SOM training using CoMFA params :

$$\begin{cases} m - \text{number of samples of dimension } k \\ \qquad\qquad X_j = (x_1^j, \ldots, x_k^j) \end{cases}$$

ANN training M = 100, each network has IxHx1 architecture

Weights of ANN with minimal early-stopping error were saved

SOM training using ANN weights :

$$\begin{cases} m - \text{number of samples of dimension H * M} \\ \qquad X_j = (w_{j1}^1, \ldots, w_{jH}^1, \ldots, w_{j1}^M, \ldots, w_{jH}^M) \end{cases}$$

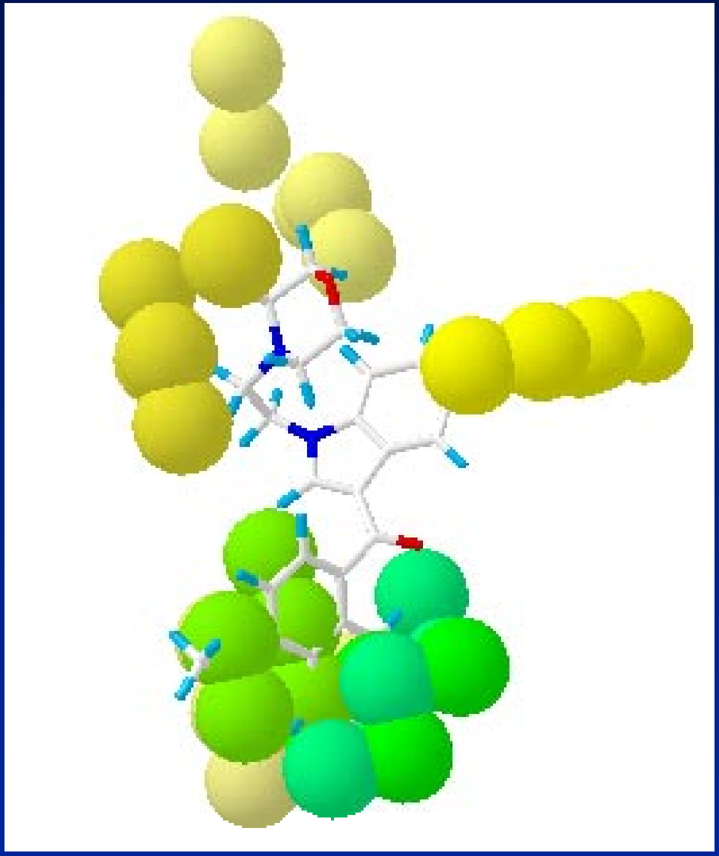
Cross-validated q^2 values calculated

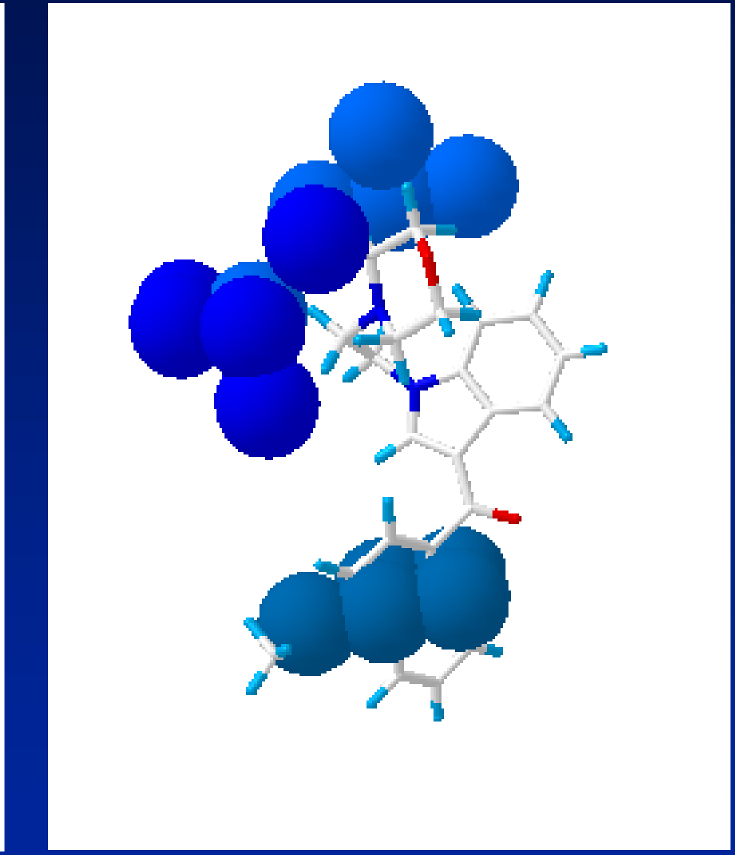
for cannabimimetic amino-alkyl indoles

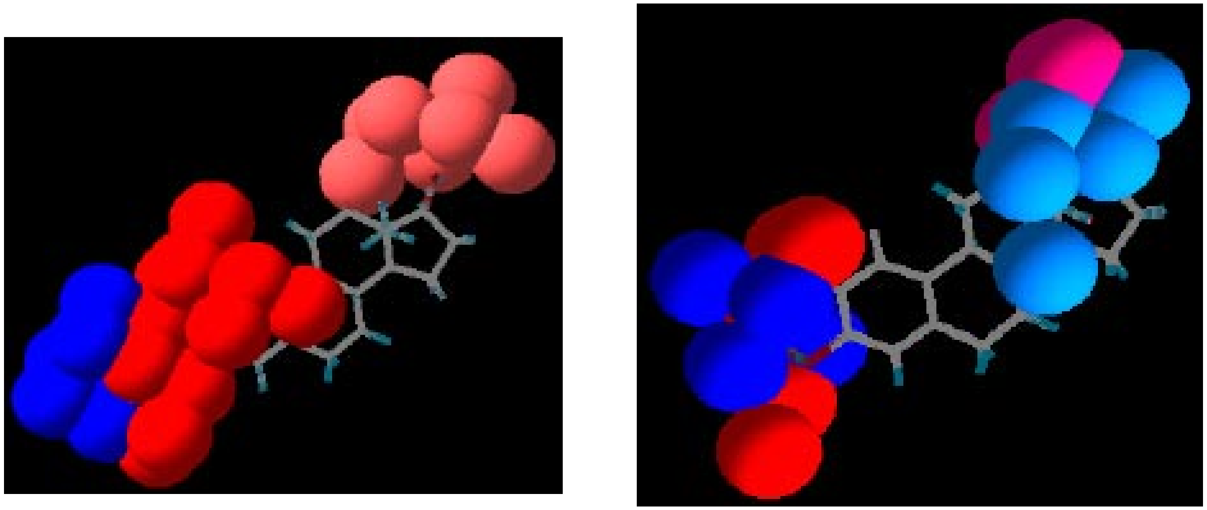
Molecular	Initial CoMFA params		ANNs weights	
params	clusters	q^2	clusters	q^2
Steric	28	0.47	14	0.78
Electrost.	16	0.28	8	0.43
S.+E.	83	0.39	16	0.75

The non-significant clusters are eliminated using pruning alorithms described in Tetko et al., JCICS, 1996, 36(4), 794-803.

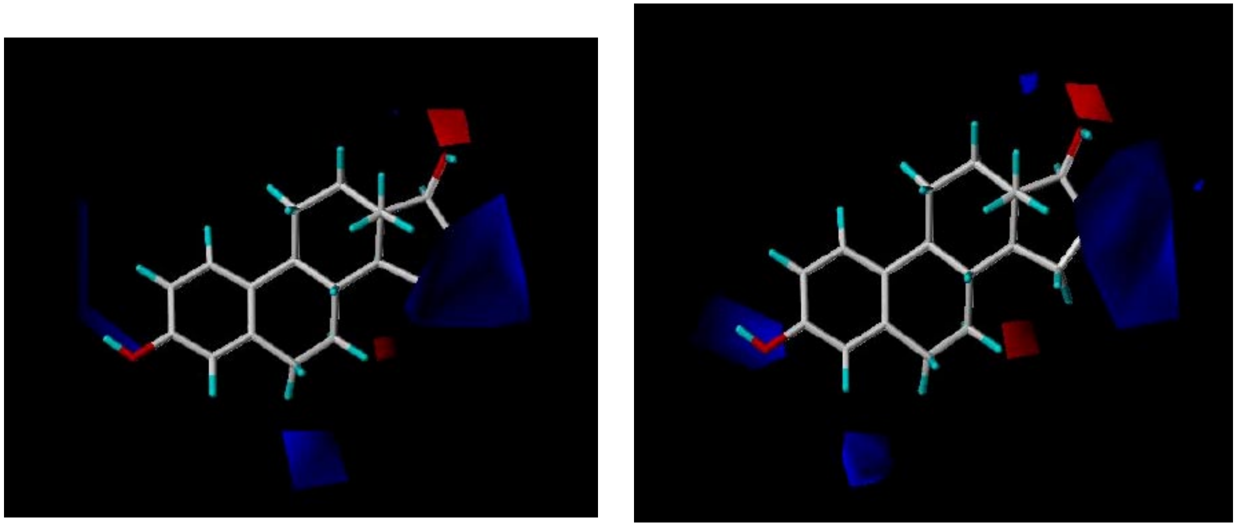
Steric and electrostatic contour plots for cannabimimetic amino-alkyl indoles, agonists of cannabinoid receptor







VLA clusters calculated for electrostatic fields of Estrogen α and β receptors.



Similar plots calculated for the same receptors using CoMFA/PLS method.

Cross-validated q^2 coefficients calculated for QSAR examples

Field	[a]VLA				[b]PLS	
	All clusters		Pruning results		Latent variables	Cross validated dataset
	Number of clusters	Cross validated dataset	Number of selected clusters	Cross validated dataset		
1. Aminoalkyl indoles						
Steric	14	0.78±0.01	10	0.78±0.01	5	0.53
Electr.	8	0.43±0.02	4	0.49±0.02	4	0.31
S.+E.	16	0.75±0.02	10	0.76±0.03	6	0.56
2. Estrogen Receptor (ER) α Subtype						
Ster.	7	0.80±0.02	4	0.80±0.02	–	–
Electr.	8	0.57±0.02	4	0.61±0.02	–	–
S.+E	15	0.79±0.02	7	0.81±0.02	4	0.52
3. Estrogen Receptor (ER) β Subtype						
Ster.	8	0.75±0.02	7	0.76±0.02	–	–
Electr.	6	0.64±0.02	5	0.63±0.02	–	–
S.+E	14	0.72±0.02	7	0.77±0.02	4	0.54

VLA references

- Tetko, I.V.; Kovalishyn, V. V.; Livingstone, D.J. Volume Learning Algorithm Artificial Neural Networks for 3D QSAR studies, *J. Med. Chem*., 2001, 44, 2411-2420. -- **description of the algorithm**

- Tetko, I.V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* 1996, *36(4)*, 794-803. -- **description of the variable selection methods (pruning algorithms) used in the VLA**

These articles + posters are available at

http://vcclab.org/lab/pdf

Acknowledgement

The main part of this presentation was done during my work at the University of Lausanne (Switzerland). This study was also supported with the
Virtual Computational Chemistry Laboratory
INTAS-INFO 00-0363 and INTAS-UA 95-0060 projects.

I also thank Prof. F. Masulli for an opportunity to participate to this school.

Thank you for your attention!