Comparison of applicability domains of QSAR models: Application to the modelling of the environmental toxicity against *T. pyriformis*

Tetko, I.V., Virtual Computational Chemistry Laboratory, http://www.vcclab.org, Neuherberg/D(GSF), Tropsha, A., Chapel Hill/USA (UNC), Zhu, H., Chapel Hill/USA, Papa, E., Varese/I (UI), Gramatica, P., Varese/I, Öberg, T., Kalmar/S (UK), Sushko, I., Neuherberg/D (GSF), Pandey, A.K. Neuherberg/D (GSF), Fourches, D., Strasbourg/FR (ULP), Varnek, A., Strasbourg/FR

http://vcclab.org

Goal: to compare different methods to define applicability domain of models (distances to the models)

Data

- Training set 644 molecules
- Validation set 1: 339 compounds
- Validation set 2: 110 compounds

Activity

Logarithm of 50% growth inhibitory concentration (pIGC50) of *T. pyriformis*

Analyzed Models

nn	group	modeling techniques	descriptors	abbreviation	distance to models	
					descriptor space	property-based space
1	UNC	ensemble of 192 kNN models	MolconnZ	kNN-MZ	EUCLID	STD
2	UNC	ensemble of 542 kNN models	Dragon	kNN-DR	EUCLID	STD
3	GSF	ensemble of 100 neural networks	E-state indices	ASNN-ESTATE		CORREL, STD
4	ULP	kNN	ISIDA Fragments	kNN-FR	EUCLID, TANIMOTO	
5	ULP	MLR	ISIDA Fragments	MLR-FR	EUCLID, TANIMOTO	
6	UI	OLS	Dragon	OLS-DR	LEVERAGE	
7	UK	PLS	Dragon	PLS-DR	LEVERAGE	PLSEU
8	UNC	SVM	MolconnZ	SVM-MZ		
9	UNC	SVM	Dragon	SVM-DR		
10	ULP	SVM	ISIDA Fragments	SVM-FR		
11	ULP	MLR	Molecular properties (CODESSA-Pro)	MLR-COD		
12	Average of all models		-	CONS		STD

Distances to Models (DM)

$$EUCLIDIAN = \min \| \mathbf{x}^i, \mathbf{x} \|$$

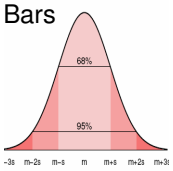
$$LEVERAGE = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$$

$$CORREL = \max_j R^2(\mathbf{Y}_{calc}, \mathbf{Y}^j_{calc})$$

$$STD = \frac{1}{N-1}\sum\left(Y_{calc} - \bar{Y}_{calc}\right)^2$$

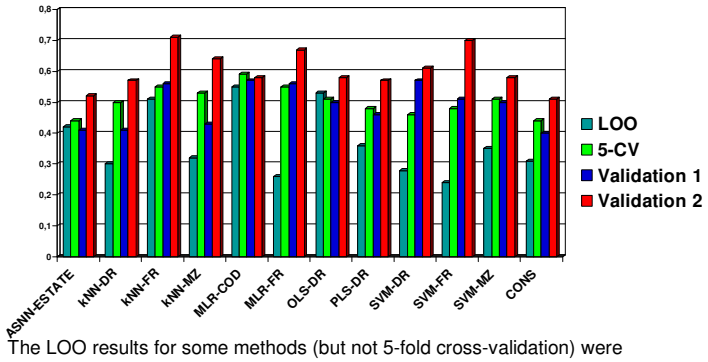
$$Tanimoto(a,b) = \frac{\sum x_{a,i}x_{b,i}}{\sum x_{a,i}x_{a,i} + \sum x_{b,i}x_{b,i} - \sum x_{a,i}x_{b,i}}$$

Error Bars



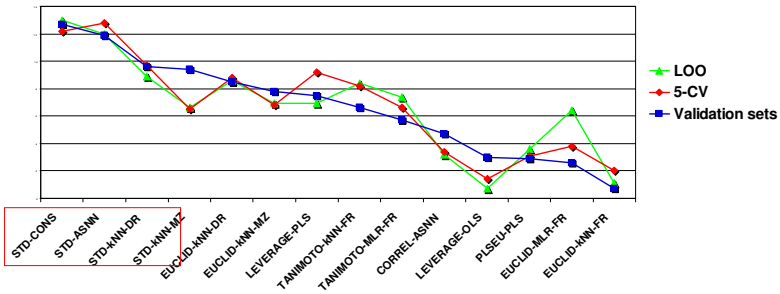
Fraction plots: number of molecules within confidence intervals vs theoretical number. The more close are both curves - the better is prediction of the errors.

RMSE of the Models



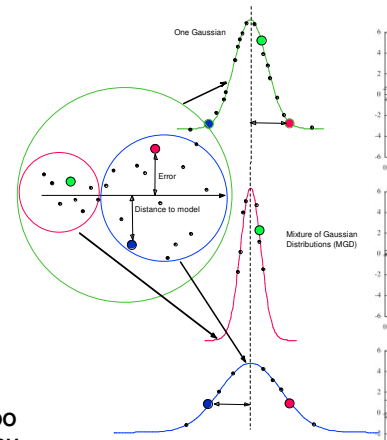
The LOO results for some methods (but not 5-fold cross-validation) were significantly different ($p<0.05$ according to the bootstrap test) compared to results for the validation set 1 (overfitting).

Ranking of the DMs

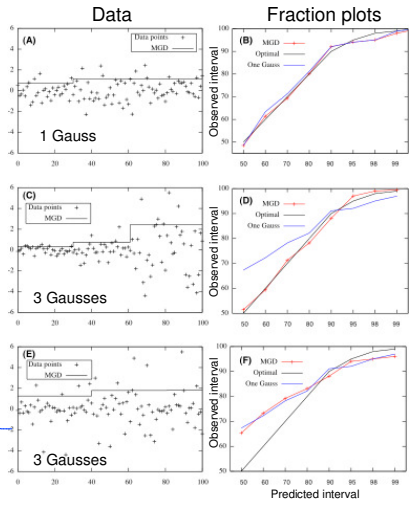


The higher ranks of DM correspond to better discrimination of molecules with large and small errors.

Mixture of Gaussian Distributions: maximizes *Likelihood score:* $\Pi N(0,\sigma^2(e_i))$

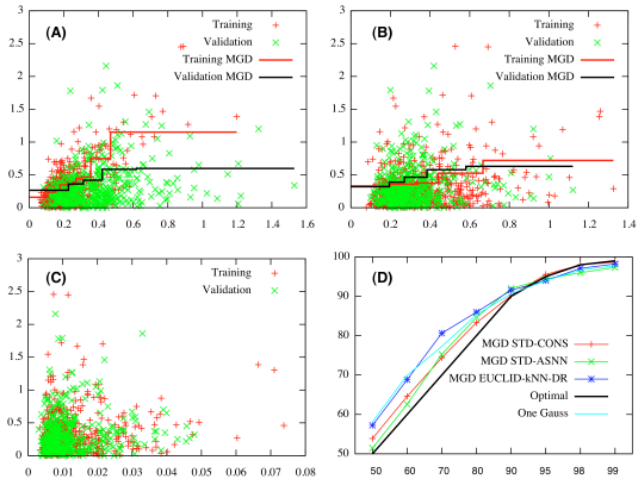


Significance of likelihood score is tested by comparison with score for a Gaussian distribution.



A significant MGD can be detected only for the second dataset where the DM (position in the dataset) allows to discriminate different Gaussians distributions

The accuracy of the ASNN model as a function of DM



A) DM with the best MGD (STD-CONS) provide very good discrimination of molecules B) DM with less significant MGD C) DM is not correlated with the accuracy of prediction D) Corresponding fraction plots

Conclusions

MGD and likelihood score can be used to compare DMs
The STD-based DMs provided the best discrimination of molecules with small and large errors
The cross-validation after the variable selection may bias the the estimation of the prediction accuracy of a model

The UNC co-authors wish to thank Dr. Alexander Golbraikh for his help with programming and interpretation of results. The ULP co-authors thank Prof. Alan R. Katritzky for providing them with the CODESSA-Pro program. This study was supported in part in part by the NIH RoadMap grant GM076059 and by EPA STAR grant RD832720 (UNC) and partially supported by the Go-Bio BMBF grant 0313883 (AZ-31P4556) (GSF).