Benchmarking of linear and non-linear approaches for QSPR studies of metal complexation with ionophores

Igor V. Tetko,1,2* Vitaly P. Solov'ev,3 Alexey V. Antonov,1 Xiaojun Yao,4 Jean Pierre Doucet,4 Botao Fan,4 Frank Hoonakker,5 Denis Fourches,5 Piere Jost,5 Nicolas Lachiche,5 and Alexandre Varnek,5

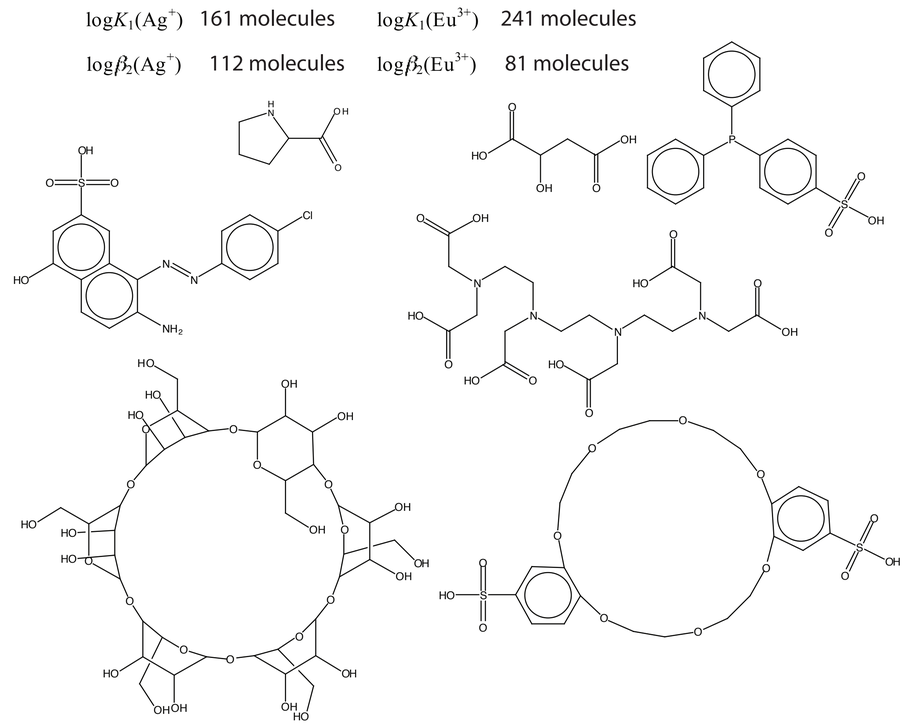
1- GSF - National Centre for Environment and Health, Institute for Bioinformatics(MIPS), 85764 Neuherberg, Germany
2- Institute of Bioorganic & Petrochemistry, National Ukrainian Academy of Sciences, 02094, Kyiv, Ukraine, http://www.vcclab.org
3- Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a, 119991 Moscow, Russia
4- Université Paris 7-Denis Diderot, ITODYS-CNRS UMR 7086, 1, rue Guy de la Brosse, Paris 75005, France
5- Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France

Objectives

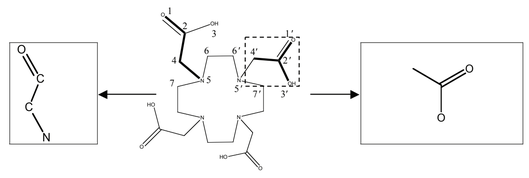
Can we predict complexation constants using QSPR?
What are the best descriptors?
What are the best methods?
Do non-linear methods add some value?
Can we compare results of different methods in an objective way?

Data Sets

$\log K_1(Ag^+)$ 161 molecules $\log K_1(Eu^{3+})$ 241 molecules
$\log \beta_2(Ag^+)$ 112 molecules $\log \beta_2(Eu^{3+})$ 81 molecules



Descriptors



SEQUENCES	AUGMENTED ATOMS II
ATOMS and BONDS (AB)	
O=C-C-N; C-C-N; C-N; O=C-C; C=O; C-C	C (-C) (-O) (=O)
ATOMS (A)	
O C C N; C C N; C N; O C C; C O; C C (**Hy**)	C (C) (O) (O) or $C_{sp2}(C_{sp3})(O_{sp3})(O_{sp2})$
BONDS (B)	
= - -; - -; - -; = -; = -; -	C (-) (-) (=)

(C) E-state indices

atom no	index name	value	count
1	dO	11.09	1
1	dO(acid)	11.09	1
2	dssC	-1.02	1
3	sOH	9.08	1
3	sOH(acid)	9.08	1
4	ssCH2	-0.229	1
5	sssN	1.64	1
5	sssN(al)	1.64	1
6	ssCH2	0.305	1
7	ssCH2	0.305	1

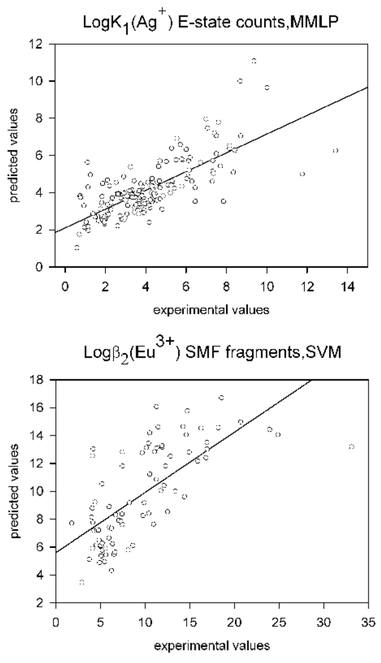
(D) Atom-type E-state indices and counts

index no	index name	values	counts
1	SdO	44.35	4
2	SdO(acid)	44.35	4
3	SdssC	-4.08	4
4	SsOH	36.30	4
5	SsOH(acid)	36.30	4
6	SssCH2	1.52	12
7	SsssN	6.57	4
8	SssssN(al)	6.57	4

Analyzed approaches

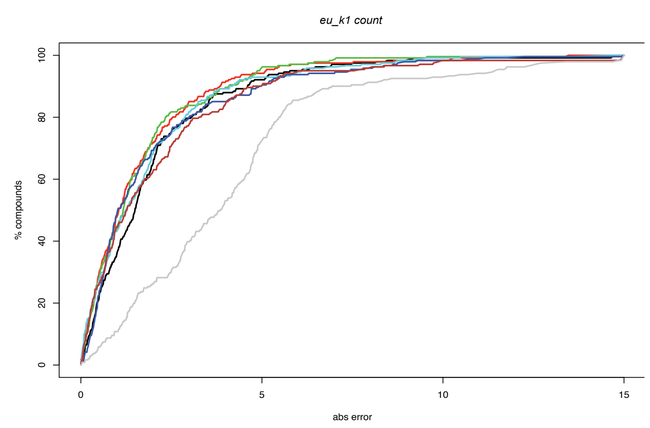
Singular Value Decomposition (MLRA/SVD) http://infochim.u-strasbg.fr/recherche/isida/
Associative Neural Network (ASNN) http://www.vcclab.org/lab/asnn
Radial Basis Function Network (RBFN) http://www.cs.waikato.ac.nz/~ml/weka
Maximal Margin Linear Programming Method (MMLP) http://mips.gsf/proj/mdcs
k-Nearest Neighbor Method (kNN)
Support Vectors Machine http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Traditional plot



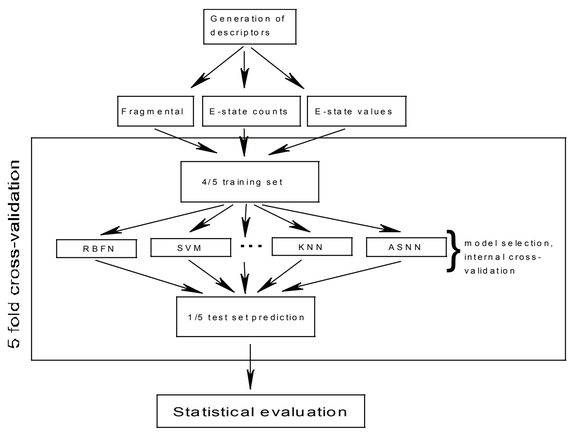
Experimental versus predicted values for models of $\log K_1(Ag^+)$ and $\log \beta_2(Eu^{3+})$. Despite apparent difference in quality of both models, the outlying molecules in each model can be easily observed.

Regression Error Curve

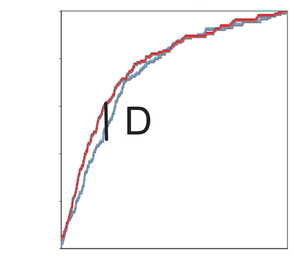


REC allows to compare results from several methods on one plot

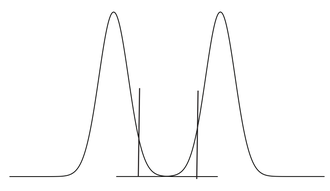
Data Analysis: double 5-fold crossvalidation



Testing of Statistical Significance



Kolmogorov-Smirnov (KS)



Bootstrap significance test

Statistical assesment of results

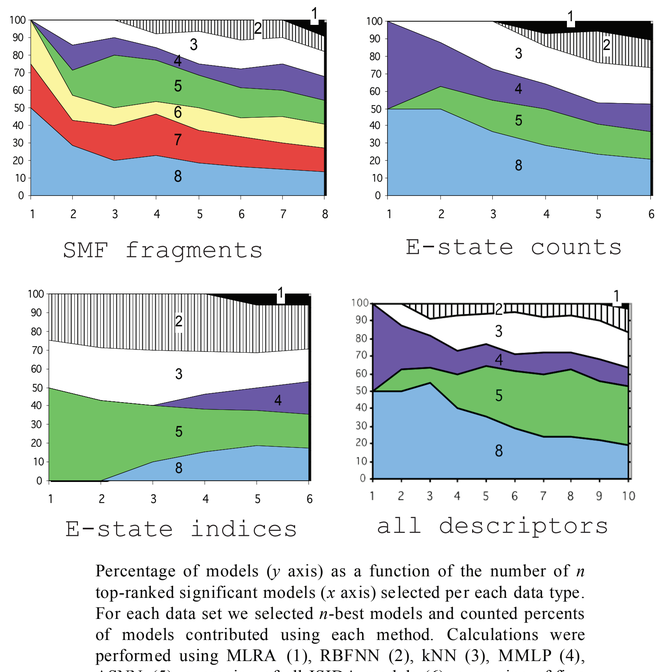
```
METHOD   REC     RMSE    MAE
SVD      0.133   3.4     2.06 -- black
ASNN     0.11    2.55    1.65 -- red
SVM      0.11    2.46    1.65 -- green
KNN      0.124   2.79    1.85 -- cyan
RBFN     0.132   3.07    1.98 -- blue
MMLP     0.142   3.89    2.22 -- brown
AVERAGE  0.274   5.19    4.13 -- gray

BOOSTRAP: asnn > mmlp average p<0.001
BOOSTRAP: svm > mmlp average p<0.001
BOOSTRAP: knn > average p<0.001
BOOSTRAP: weka > average p<0.001
BOOSTRAP: svd > average p<0.001
BOOSTRAP: mmlp > average p<0.001

KS: svd != asnn 0.0081
    svd != svm 0.0147
    svd != weka 0.0258
    svd != average  p<0.0001
KS: asnn != average  p<0.0001
KS: svm != mmlp 0.0258
    svm != average  p<0.0001
KS: knn != average  p<0.0001
KS: weka != average  p<0.0001
KS: mmlp != average  p<0.0001
```

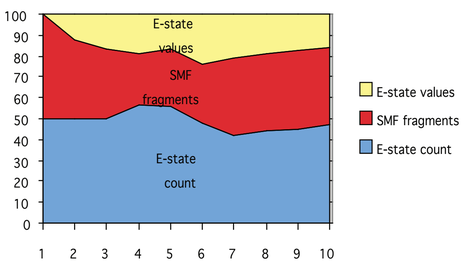
Statistical analysis provides an objective comparison of different methods

Comparison of Methods



Percentage of models (*y* axis) as a function of the number of *n* top-ranked significant models (*x* axis) selected per each data type. For each data set we selected *n*-best models and counted percents of models contributed using each method. Calculations were performed using MLRA (1), RBFNN (2), kNN (3), MMLP (4), ASNN (5), averaging of all ISIDA models (6), averaging of five first ranked ISIDA models (7) and SVM (8).

Comparison of Descriptors



Percentage of best models (y axis) calculated using corresponding descriptor system and all methods.

Conclusions

Models based on fragments (SMF, E-state counts) > E-state indices
Non-linear approaches > multiple linear regression (MLRA) ($p<0.05$)
But ensemble of several MLRA ≈ non-linear approaches
No-significant differences in performance of non-linear models
SVM and ASNN provided largest number of "best" models
kNN was the fastest method

Acknowledgement IVT was supported with Invited Professor position from Université Louis Pasteur.
The part of this work has been performed in the framework of French-Russian collaborative project GDRE “SupraChem”.