Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure−Property Relationship Studies of Metal Complexation with Ionophores

Igor V. Tetko[†]

Institute of Bioorganic & Petrochemistry, Kiev, Ukraine

Vitaly P. Solov'ev

Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a,
119991 Moscow, Russia

Alexey V. Antonov

Institute for Bioinformatics, Neuherberg D-85764, Germany

Xiaojun Yao, Jean Pierre Doucet, and Botao Fan

Université Paris 7-Denis Diderot, ITODYS-CNRS UMR 7086, 1, rue Guy de la Brosse, Paris 75005, France

Frank Hoonakker, Denis Fourches, Piere Jost, Nicolas Lachiche, and Alexandre Varnek*

Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur,
4, rue B. Pascal, Strasbourg 67000, France

Received September 24, 2005

A benchmark of several popular methods, Associative Neural Networks (ANN), Support Vector Machines (SVM), k Nearest Neighbors (kNN), Maximal Margin Linear Programming (MMLP), Radial Basis Function Neural Network (RBFNN), and Multiple Linear Regression (MLR), is reported for quantitative−structure property relationships (QSPR) of stability constants logK_1 for the 1:1 (M:L) and logβ_2 for 1:2 complexes of metal cations Ag^+ and Eu^{3+} with diverse sets of organic molecules in water at 298 K and ionic strength 0.1 M. The methods were tested on three types of descriptors: molecular descriptors including E-state values, counts of atoms determined for E-state atom types, and substructural molecular fragments (SMF). Comparison of the models was performed using a 5-fold external cross-validation procedure. Robust statistical tests (bootstrap and Kolmogorov-Smirnov statistics) were employed to evaluate the significance of calculated models. The Wilcoxon signed-rank test was used to compare the performance of methods. Individual structure−complexation property models obtained with nonlinear methods demonstrated a significantly better performance than the models built using multilinear regression analysis (MLRA). However, the averaging of several MLRA models based on SMF descriptors provided as good of a prediction as the most efficient nonlinear techniques. Support Vector Machines and Associative Neural Networks contributed in the largest number of significant models. Models based on fragments (SMF descriptors and E-state counts) had higher prediction ability than those based on E-state indices. The use of SMF descriptors and E-state counts provided similar results, whereas E-state indices lead to less significant models. The current study illustrates the difficulties of quantitative comparison of different methods: conclusions based only on one data set without appropriate statistical tests could be wrong.

INTRODUCTION

An important branch of supramolecular chemistry is the chemistry of ionophore−molecules possessing high affinity toward metal cations in solutions. Their ability to bind cations selectively is widely used in practice for the separation and concentration of metals (solvent extraction) and in analytical devices (ion-selective electrodes, CHEMFETs, etc.).[1]

Experimental measurements of stability constants of ionophore−metal complexes and related free energies of complexation reactions represent rather difficult and costly tasks. That is why a theoretical quantitative estimation of complexes stabilities might become an important complement of experimental studies thus providing researchers a way to reduce the number of experiments and to indicate the strategy of "optimization" of known metal binders.

The thermodynamic complexation properties depend on many parameters: the nature of the metal, structure of ionophore, solvent, conterion(s), temperature, and background compounds. In experiments, even small inaccuracies in measuring species concentration or temperature may lead to errors in complexation constants up to several log units.[2,3]

One can mention different theoretical approaches to assess free energies of complexation. Quantum Mechanics calculations in the gas phase could be hardly recommended for these

* Corresponding author e-mail: varnek@chimie.u-strasbg.fr.

[†] Current address: Institute for Bioinformatics, Neuherberg D-85764, Germany. http://www.vcclab.org.

10.1021/ci0504216 CCC: $33.50 © 2006 American Chemical Society
Published on Web 01/17/2006

purposes because of the crucial role of solvent effects on metal−ionophore interactions. Molecular Dynamics or Monte Carlo simulations in solution coupled with the free energy perturbation technique[4] looks like a promising way to estimate complexes stabilities. However, this approach is far from becoming a reliable predictive tool because of too simplified an energy representation within the empirical force fields approach, complexity of modeled systems, and large simulation times.

There exist many empirical approaches based on correlations of experimentally available stability constants with some chemical and physical parameters: cationic radii, the number of ionophore's coordination centers, chemical "hardness" and "softness", electrostatic potential distribution, etc.[5−10] Although these correlations attempted to account for the mechanism of metal−ionophore interactions, they were built on rather small data sets involving restricted families of ionophores. Therefore their utilization as predictive tools is very limited. Another type of empirical correlations concerns those linking stability constants of the same series of ionophores (metals) with two different metals (ionophores). These relationships can be used for predictions only if an experimental value of a stability constant is available for one metal (ionophore) involved.

Here, to assess stabilities of metal−ionophore complexes we use Quantitative Structure−Property Relationships (QSPR) which is the most realistic way to develop robust structure−stability constant models suitable to be used for the theoretical design of new efficient metal binders. Two main aspects of any QSPR study are related to the selection of the type of descriptors (topological, physicochemical, molecular fragments, etc.) and the mathematical method used to develop the models (linear or nonlinear techniques). A selection of pertinent descriptors and a method is crucial to develop robust models with superior performance.

The question arises what type of descriptors and methods could be recommended for QSPR studies of metal−ionophore complexes? To answer this question, we have carried out systematic studies on four structurally diverse data sets containing stability constants for the M:L = 1:1 complexes ($\log K_1$) and for 1:2 complexes ($\log\beta_2$) of cations Ag^+ and Eu^{3+} with ionophores belonging to different chemical classes (macrocyclic, heterocyclic, and acyclic agents bearing acidic, basic, or neutral functions).[11] Silver(I) and europium(III) bearing different electric charges and displaying big difference in coordination numbers[12] and coordination polyhedrons[13] were chosen to model stabilities of the 1:1 and 1:2 complexes of metal cations of essentially dissimilar nature.

Two type of 2D descriptors were compared: substructural molecular fragments (SMF)[14−16] and E-state indices.[17] Earlier, the SMF descriptors have been successfully used for the assessment of complexation constants[14−16,18,19] and the design of new extractans of UO_2^{2+}.[18,20] Although, E-state indices were never used for structure−property modeling of metal complexation, they were widely used in QSAR studies of aqueous solubility, lipophilicity, and some biological activities.[21] Multiple Linear Regression Analysis, Support Vector Machine, Radial Basis Function, k Nearest Neighbors, Maximal Margin linear programming, and Artificial Neural Networks were tested to obtain QSPR models.

Table 1. Composition of Analyzed Data Sets

			SMF descriptors		
no.	data set	size	type	no. of descriptors	no. of E-state indices
1	$\log K_1(Ag^+)$	161	I(AB,2−4)	216	59
2	$\log\beta_2(Ag^+)$	112	II(Hy)	79	48
3	$\log K_1(Eu^{3+})$	241	I(AB,3−4)	215	57
4	$\log\beta_2(Eu^{3+})$	81	I(AB,4−4)	88	44

Another goal of this article was to propose some statistical framework to estimate the significance of calculated models and to compare results obtained with different methods. Indeed, one can meet in literature the conclusions about "better performance" or "superiority" of some methods, which are not supported by any statistical tests. We describe and consistently apply several statistical tests (bootstrap, Kolmogorov-Smirnov statistics, and Wilcoxon signed-rank test[22]) to compare performances of linear and nonlinear methods using exactly the same cross-validation approach, training and test sets, and descriptors.

DATA SETS

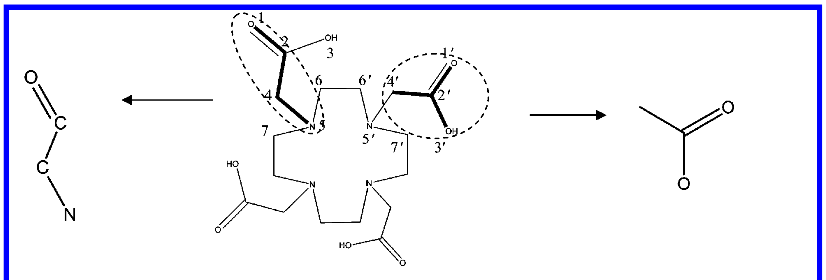
Experimental stability constants $\log K_1$ for the 1:1 (M:L) and $\log\beta_2$ for 1:2 complexes of the metal cations with organic molecules in water were critically selected at standard temperature 298 K and ionic strength $I = 0.1$ M from IUPAC Stability Constants Database (http://www.acadsoft.co.uk, SC DB, licensed version 5.33, April 2004, Academic Software). Additional data were collected by adjusting the $\log K_1$ and $\log\beta_2$ values to the standard conditions. Devies equation[2,11] for the mean ionic activity coefficient was applied to adjust stability constants to a specified ionic strength ($I = 0.1$ M), whereas the van't Hoff equation[23] was used to adjust stability constants to a specified temperature ($T = 298$ K) assuming the enthalpy is independent of temperature.

Four structurally diverse data sets including 161 (Ag^+) and 241 (Eu^{3+}) $\log K_1$ and 112 (Ag^+) and 81 (Eu^{3+}) $\log\beta_2$ values for the complexation of Ag^+ and Eu^{3+} with organic molecules in water were prepared (Table 1). Selected ionophores belong to different classes of organic compounds presented in SC DB: macrocyclic, heterocyclic, and acyclic agents bearing acidic, basic, or neutral functions.

Data Normalization. Prior to analysis with all methods with an exception of Multiple Linear Regression analysis (MLRA), all input variables were normalized to the (0,1) interval range. The experimental values were normalized on the (0.2,0.8) interval for the Associative Neural Network (ASNN). Before running global optimization we nonsystematically tried several other normalizations (normalization on (−1,1) interval, normalization of both input descriptors and target values) and did not observe any significant differences in calculated performances of the methods.

MOLECULE DESCRIPTORS

Fragmental Descriptors. The Substructural Molecular Fragment (SMF) method[14,18] which is part of the ISIDA software[24] was used to generate molecular fragments (Figure 1). Two different classes of fragments were used: "sequences" (**I**) and "augmented atoms" (**II**). Three subtypes **AB**, **A**, and **B** were defined for each class. For fragments I, they represented sequences of atoms and bonds (**AB**), of



SMF

SEQUENCES **I**		**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 (-) (-) (=)

E-state

INDIVIDUAL

atom no	index name	index 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

CUMULATED

index no	index name	index value	count
1	SdO	44.36	4
2	SdO(acid)	44.36	4
3	SdssC	-4.08	4
4	SsOH	36.32	4
5	SsOH(acid)	36.32	4
6	SssCH2	1.52	12
7	SsssN	6.56	4
8	SsssN(al)	6.56	4

Figure 1. Used descriptor systems. *Substructural molecular fragments*: shortest path sequences (**I**) and augmented atoms (**II**) including atoms and bonds (**AB**), only atoms (**A**), or only bonds (**B**). From top to bottom: the sequences (**I**) correspond to the **I**(**AB**, 2−4), **I**(**A**, 2−4), and **I**(**B**, 2−4) types involving paths between each pair of atoms. The **II**(**Hy**) augmented atoms correspond to the **II**(**A**) type, where hybridization of the atom is taken into account. *E-state indices and counts*. E-state indices were calculated according to ref 17. Since the molecule is symmetrical, all symmetric atoms have the same individual E-state values, e.g. 1 and 1′, 2 and 2′, etc. The cumulated atom-type E-state values and counts are sums over all the atoms of the considered type. The extended atom types[25,26] are shown in parentheses.

atoms only (**A**), or of bonds (**B**) only. The number of atoms in these sequenced was varied from 2 to 6, and only the shortest paths for each pair of atoms were used. Atomic hybridization was also taken into account for augmented atoms of the **A**-type. Totally 49 types of fragments were generated, and the fragments counts were used as descriptors. For each data set we selected one type of fragment as described in the Methods section. We will refer to this type of descriptors as fragmental or SMF descriptors.

Atom Type E-state Indices. The electrotopological state (E-state) indices introduced by Hall and Kier[17,21] combine together both electronic and topological characteristics of the analyzed molecules (Figure 1). For each atom type in a molecule the E-state indices are summed and are used in a group contribution manner. In this study we used an extended set of atom-type E-state indices which was developed to better cover functional groups and the neighborhood of nitrogen and oxygen atoms.[25,26] Similar to our previous studies[26,27] we also included the molecular weight (MW) and the number of non-hydrogen atoms (NA) as two additional parameters for E-state indices and their counts.

Atom Type E-state Counts. Some researchers recently argued[28−30] that a use of atom counts corresponding to atom types determined for E-state indices may generate models with similar prediction ability to the models developed using the E-state indices. Therefore in addition to the E-state indices we also included counts of atoms corresponding to E-state indices as additional descriptors. To distinguish both systems we will refer to E-state indices as “E-state values” and to atom counts corresponding to them as “E-state counts”. The atom-type E-state indices and their counts were calculated using the http://www.vcclab.org/lab/pclient program.[31] The E-state counts actually correspond to the II(**B**) type of the SMF descriptors but have a larger number of atom types.

METHODS

Model Validation and Optimization of Method Parameters. Double cross-validation which included internal (optimization of model parameters) and external cross-validations (testing the models performance) was used (Figure 2). All compounds in each initial data set were

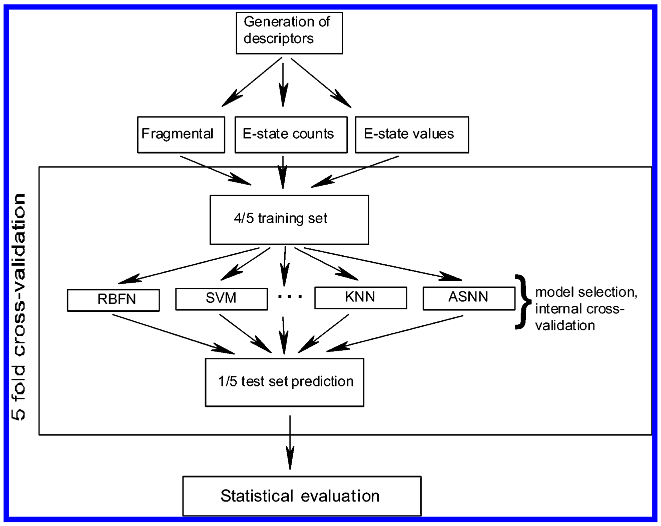


Figure 2. Dataflow for benchmarking analyzed methods.

randomly shuffled to avoid possible artificial ordering due to data preparation. The received sets were split five times on subsets containing 4/5 of all molecules (to be further referred to as "cross-validation training sets", CVTS) and their complements containing 1/5 of all molecules ("cross-validation test sets"). The same cross-validation training and test sets splits were used for all methods. The CVTS were employed to optimize internal parameters of each method, and the selected parameters were applied to predict corresponding test sets. Thus, each method performed a "blind prediction" of cross-validation test sets.

Multiple Linear Regression Analysis and Three ISIDA Models. MLRA based on the singular value decomposition (SVD) method was used as implemented in the ISIDA/QSPR program.[24] The program exploits SVD as a computational engine which allows one to build nonlinear models by incorporating second-order terms and cross-terms products in the regressions.[14,15,18,19,32−34] At the training stage, the ISIDA/QSPR program builds up to 196 structure−property models involving linear (including constant term or not) and nonlinear fitting equations and 49 types of fragment descriptors (45 sequences' types and 4 augmented atoms' types). If some of the variables are linearly dependent or if a given fragment occurs in a relatively small number of molecules, the standard deviation for the fragment contributions can be large enough to lead to the corresponding t-test being smaller than the tabulated value (t_0). The following procedure was applied to improve the robustness of the models. First, the program selects the variable with the smallest $t < t_0$, then it performs a new fitting excluding that variable. This procedure is repeated until $t \geq t_0$ for selected variables or if the number of variables reaches the user's defined value. First, 49 fragments types without the generation of nonlinear terms were considered for each data set. The types of fragments which provided the best internal leave-one-out cross-validation parameters for linear models ("ISIDA-single" models) (cross-validation coefficient Q^2, R_H-factor of Hamilton,[35] standard deviation s) were selected (Table 1), and the corresponding fragments counts were used as descriptors to build models by all other methods.

Usually not one but several models provide high internal cross-validation parameters. Thus, another model, "ISIDA-5", included five (data sets 1−3, Table 1) or four (data set 4) best models, which were characterized with the best internal cross-validation parameters. The third model, "ISIDA-average", was calculated as an average of all 196 linear and nonlinear models after excluding tail predictions using the t-test.[35] The drawback of the fragment descriptors is related to the property's prediction for compounds containing some fragments, which are not present in the training set—these compounds are not predicted. To make possible the direct comparison of the "ISIDA-single" model, we stepped out from our usual procedure and always predicted all molecules in the test set.

Support Vectors Machine. Support Vectors Machine (SVM) is a nonlinear classification and regression method developed by Vapnik.[36] We used the open source LibSVM package.[37] SVMs were originally developed for classification problems. Later on they were also extended to solve nonlinear regression problems by the introduction of ϵ-insensitive loss function. In support vector regression, the input variables are first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. The kernel functions often used in SVM include linear, polynomial, radial basis function (RBF), and sigmoid function. The generalization performance of SVM depends on the selection of several internal parameters of the algorithm (C and ϵ), the type of kernel, and the parameters of the kernel. It was shown that a RBF kernel under a proper selection of width of radial basis function can be equivalent to other types of kernels.[38] For the purposes of the current study we restricted our analysis to the RBF kernel and selected the parameters of the algorithm using internal 5-fold cross-validation on corresponding CVTS. The grid search considered three parameters $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ $\epsilon = 0.0001, 0.001, ..., 10$ and width of the RBF kernel $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$ as recommended in the LibSVM manual. Thus for each data set we performed $5*11*6*9 = 3300$ SVM program runs.

k-Nearest Neighbor Method (kNN). The kNN method predicts activity of the target compounds as an average value of activities of its k-nearest neighbors in the space of input descriptors. The Euclidian distance was used. The number of neighbors, k, was optimized using corresponding CVTS and then applied to predict the cross-validation test sets. The kNN method was programmed in house.

Associative Neural Network (ASNN). The associative neural network (ASNN) represents a combination of an ensemble of feed-forward neural networks and the kNN. This method uses the correlation between ensemble responses (each molecule is represented in space of neural network models as a vector of model predictions) as a measure of distance amid the analyzed cases for the nearest neighbor technique. Thus ASNN performs kNN in space of ensemble residuals. This provides an improved prediction by the bias correction of the neural network ensemble.[39,40] The neural networks ensemble of 100 networks with one hidden layer was used. The optimization was performed to select an optimal number of neurons in the hidden layer (neurons = 1, 2, 3, 5, 7) and the type of distance in the ASNN correction $D = 0, 1, 2$. The second parameter corresponded to use of the ensemble average without kNN analysis ($D = 0$), simple ($D = 1$), and weighted correction ($D = 2$).[41] The results for $D = 0$ were calculated simultaneously with $D = 1$ or $D = 2$. Thus, for each data set we performed $5*6*2 = 60$ ASNN analyses corresponding to the training of 6000 neural networks. The calculations were performed using a batch-system at http://www.vcclab.org/lab/asnn and were run in parallel on 14 computers.

Radial Basis Function Network. The Radial Basis Function Neural Network (RBFNN) available in the WEKA program[42] was used. The RBFNN and SVM with the RBF kernel have the same basic architecture. However, the differences appear due to the training approaches. The RBFNN uses k-means clustering to select basis functions and uses symmetric multivariate Gaussians to fit the data from each cluster. The linear regression is further used on the top of these basis functions. The adjustable parameters included the number of clusters, $B = 2, 3, 5, 7, 10, 20, 30, 50, 100, 200$, and the ridge parameter for linear regression, $R = 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100$. Thus for each data set we performed $5*9 \times 10 = 450$ WEKA runs.

Maximal Margin Linear Programming Method (MMLP). The Maximal Margin Linear Programming method (MMLP) was initially developed for the classification of tumor samples in data sets having tens of thousands of genes (input variables) and only a few tens of measurements.[43] This procedure detects groups of genes and constructs sparse models (features) that strongly correlate with particular tumor types. The method detects a solution with maximum margin, and the models detected by this approach are expected to have high prediction ability. Later on we developed an extension of this approach to perform an optimization of regression tasks.[44] To find a solution the method requires a set of variables, which significantly exceed the number of samples. To increase the dimensionality of the input space we additionally used nonlinear terms, x^{-3}, x^{-2}, x^{-1}, x^2, x^3. There is only one parameter, J, which controls the upper limit of model accuracy, the loss during the training process. The following values $J = 0.01, 0.04, 0.07, 0.10, 0.13, 0.16$ were tested using an internal cross-validation procedure. Thus for each data set we performed $5*6 = 30$ runs of the algorithm.

Statistical Parameters. ***Bootstrap Test.*** The comparison of approaches was done using Mean Absolute Error (MAE). A traditional approach to compare results of two methods is to use the bootstrap test.[45] The general idea of bootstrapping is to use the single data set and to design a sort of Monte Carlo experiment in which the data are used to generate an approximation of analyzed statistical parameters. A bootstrap sample $\mathbf{x}^*$ of a data set $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a replica of $\mathbf{x}$ generated by sampling with replacement n elements from the original data. The advantage of this test is that it can be applied without assumptions about the underlying distribution of analyzed values. For each set of predicted or experimental values we generated 10 000 bootstrap replicas. The MAE values were calculated for each replica and were used to estimate confidence values of the original set (i.e., for $p = 0.05$ we selected MAE values corresponding to the low and top 5% of MAE values calculated for 10 000 bootstrap replicas). The bootstrap test can be applied to estimate confidence values for other statistical parameters, e.g. Root Mean Squared Error (RMSE) or R^2.

The Wilcoxon signed-rank test[22] was used to evaluate the performance of methods. This test compared paired MAE values calculated for the same data sets. The statistical significance was based on the number of times one method calculated higher or lower MAE values compared to another one. All pairs of MAE values were used, independently if the compared models were significant or not significant according to the bootstrap test. Thus the results of the Wilcoxon signed-rank test were independent from those of the bootstrap set

Regression Error Characteristic (REC) Curve. Recently, the Regression Error Characteristic Curve (REC) was introduced as a generalization of Receiver Operator Curve (ROC) used in classification.[46] These curves plot the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis. The REC can be plotted using an absolute deviation $|y - f(x)|$ or squared residual $(y - f(x))^2$. It allows a quick visual estimation of the relative merit of many regression curves by examining their relative positions. The Area-Over-the-Curve (AOC) provides a biased estimation of the model performance, and it is similar to MAE and RMSE for absolute and squared curves, respectively. The authors of the method proposed to estimate a significant difference between RECs using the Kolmogorov-Smirnov (KS) two-sample test.[22] We found that this test is more sensitive than the bootstrap test, but it requires some caution. This test should be only applied if one curve dominates the second one for the most period of time, apart from the beginning and the end where all curves converge. If REC curves are crossing one another, each curve outperforms another one in a particular region of REC and the test become invalid. Thus, we decided to use KS only in cases when the bootstrap test failed to find significant values and a visual inspection of data indicated that one curve was always on top of another one.

Arithmetic Average Model (AAM). To state that a given model is significant according to the bootstrap or the KS test, one should have some model as a base or null hypothesis, i.e., “no model”, for this comparison. A straightforward approach consists of using the average complexation constant of all molecules as the predicted value for all compounds. Indeed, such a model provides $R^2 = 0$ with the target activity and perfectly fits our intuition of a model without any prediction power. We refer to it as an “arithmetic average model” (AAM) and used it as the null hypothesis of “no model”. A model was considered as meaningful (significant) if it calculated a significantly smaller ($p < 0.05$) mean absolute error compared to the MAE of the “arithmetic average model” according to the bootstrap test. If the bootstrap test failed, it was also considered as a significant one if its REC curve was always at the top of the AAM curve, and both curves were significantly different ($p < 0.05$) according to the KS test.

RESULTS

The first analysis was performed to compare different systems of indices for prediction of the activities of molecules. Each method was applied to selected types of SMF fragments indicated in Table 1, E-state values, and counts using exactly the same training and test sets. In the following, all models were tested against the AAM model. Only models found significantly different from the AAM model were discussed.

Comparison of Descriptors Types. *SVM* models for $\log K_1(\mathrm{Ag}^+)$ and $\log K_1(\mathrm{Eu}^{3+})$ with an exception of the SMF

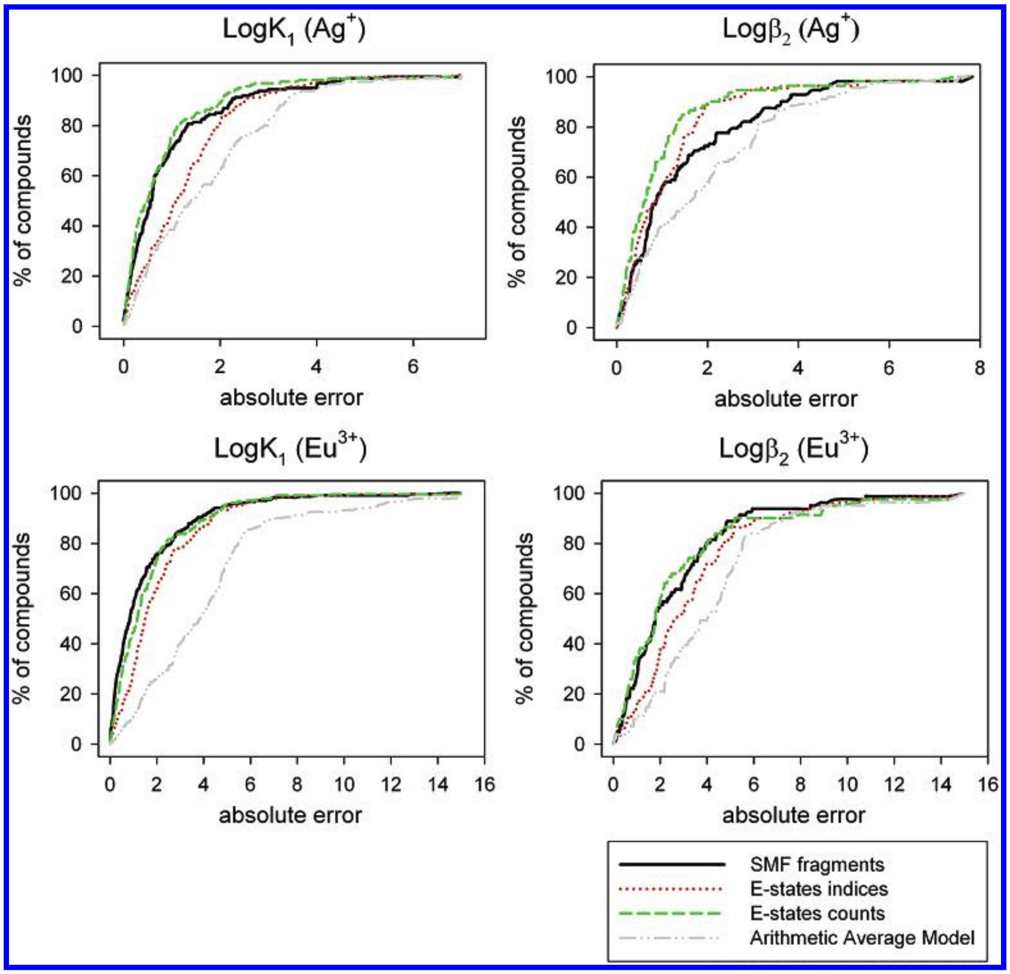


Figure 3. REC curves for prediction of data sets using SVM. Green lines correspond to models calculated using E-state counts; red lines correspond to models calculated using E-state values; black lines correspond to models calculated using SMF fragments; and gray lines correspond to arithmetic average models (AAM). The area between AAM and the corresponding calculated curve reflects the quality of the models (the large area corresponds to better models), and it is maximal for the logK_1(Eu^{3+}) model.

fragment-based model for logβ_2(Ag$^+$) were significant according to the bootstrap test ($p < 0.01$, Figure 3). The model built with fragmental descriptors was significant according to the KS test ($p < 0.05$). Since it did not cross the AAM model, it was considered as a significant one. Two models, based on SMF fragments and E-state counts, were significant according to the bootstrap test for logβ_2(Eu^{3+}). To some surprise SVM demonstrated a significantly better performance according to the bootstrap test ($p < 0.05$) for E-state counts and SMF fragments compared to the E-state values for logK_1(Ag$^+$) and logK_1(Eu^{3+}) data sets. The E-state count model for logβ_2(Ag$^+$) was also significantly better compared to the model based on SMF fragments according to the same test ($p < 0.05$). Thus, SVM demonstrated a strong tendency to the E-state count as the descriptor system.

ASNN showed similar results using all three systems of indices (Figure A1 in the Supporting Information). All models, except for logβ_2(Eu^{3+}) were significantly different ($p < 0.01$) from the AAM according to the bootstrap test. KS indicated that all models for logβ_2(Eu^{3+}) were significant ($p < 0.001$). The visual inspection, however, indicated that the models REC curves crossed the AAM curve, and we concluded that this test was not applicable. No significant statistical differences between SMF fragments and the E-state descriptors were detected.

kNN provided significant models ($p < 0.01$) for logK_1(Ag$^+$), logβ_2(Ag$^+$), and logK_1(Eu^{3+}) data sets. The E-state count model was also significant for logβ_2(Eu^{3+}) set ($p < 0.05$). The KS test indicated significant differences between models calculated using different descriptors, but all these results were discarded since respective REC curves crossed each other (Figure A2, Supporting Information). The number of nearest neighbors for the kNN method ranged from 1 to 6, and their average number was 2.3.

RBFNN. Except for the model calculated using SMF fragments, which was not significant for logβ_2(Ag$^+$), all models were significant for the first three data sets according to the bootstrap test ($p < 0.01$). The E-state values model was also significant for the logβ_2(Eu^{3+}) data set ($p < 0.05$) according to the same test (Figure A3, Supporting Information).

*MMLP***.** The significant models according to the bootstrap test were calculated for three data sets, and no significant models were found for logβ_2(Eu^{3+}). The MMLP models calculated using E-state counts were significantly better compared to the E-state values for logK_1(Ag$^+$) and logβ_2(Ag$^+$) data sets ($p < 0.05$) according to the bootstrap test. For the logK_1(Ag$^+$) set the E-state count-based models were also significantly better ($p < 0.05$) compared to the model calculated using SMF fragmental descriptors. Thus, similar to the SVM method MMLP demonstrated a strong tendency to the E-state count as the descriptor system (Figure A4, Supporting Information).

ISIDA-average and ISIDA-5 models (Figures A5 and A6, Supporting Information) calculated for first three data sets were significant ($p < 0.01$) according to the bootstrap test. No significant models were calculated for $\log\beta_2(Eu^{3+})$. Notice that both models were calculated using only SMF fragments, as indicated in the Method section. ISIDA-single models (Figure A7, Supporting Information) were calculated with all three descriptor systems. All models were significant for the $\log K_1(Eu^{3+})$ data set ($p < 0.01$). Apart from this, only the model based on SMF fragments for $\log K_1(Ag^+)$ and the model based on E-state counts for $\log\beta_2(Ag^+)$ were significant ($p < 0.05$). Thus, the pure linear regression approach produced the smallest number of significant models compared to nonlinear methods.

Analysis of Top-Ranked Models. A comparison of indices and methods performances using several systems of descriptors in general is difficult. Definitely, each method can provide the best results using one or another set of descriptors. Moreover, the “best” descriptors can be different for different data sets. To simplify this comparison we ordered results calculated with different methods and descriptors in the descending order of their MAE. Only models significantly different from the AAM, i.e., “significant models”, were included in the lists (Table 2). We also tested the top ranked model for each data set against all other significant models for the same set.

The comparison of top-ranked models showed a difference in their number for analyzed data sets. Most of the models calculated using all of the methods for all of the descriptors systems were significant for the first three data sets. Contrary to that most of the models were not significant for the $\log\beta_2(Eu^{3+})$ set. Moreover, one can easily notice that models for $\log K_1$ have in general lower MAE compared to $\log\beta_2$, because of the lower quality of the experimental data in the later models, as it is explained in the Discussion.

The analysis of models in Table 2 indicated complexities with comparison methods using just one data set. A combination of model + descriptors, which provided top and very significant results for one data set, was at the bottom of the list for another data set. For example, a combination MMLP+E-state counts provided the two best models for $\log K_1(Ag^+)$ and $\log\beta_2(Ag^+)$ sets. The model built using the same combination of method/descriptor was at the bottom of the significant model list for the $\log K_1(Eu^{3+})$ model and nonsignificant for $\log\beta_2(Eu^{3+})$. Another example, a combination of SVM+SMF descriptors and SVM+counts both provided top and consistent models for $\log K_1(Eu^{3+})$ and $\log\beta_2(Eu^{3+})$ but had a very different performance for $\log\beta_2(Ag^+)$. In fact, the SVM+SMF and SVM+counts models for $\log\beta_2(Ag^+)$ are significantly different at $p < 0.05$ according to the bootstrap test. Thus, it is rather dangerous to extrapolate results obtained with a particular method and the type of descriptors on any other case. Such generalizations should be always based on at least several sets and, moreover, be accompanied with a clear description of the protocol used to optimize the model parameters and statistical tests used to compare the models.

Of course, both the bootstrap and the KS test represent a very convenient way to quantitatively support or reject

Table 2. Statistical Parameters of Significant Models

model no	method	descriptors	MAE	RMSE	R^2
		(A) $\log K_1(Ag^+)$			
1	MMLP	E-state counts	0.79	1.33	0.64
2	SVM	E-state counts	0.8	1.3	0.65
3	ISIDA-5	SMF	0.85	1.4	0.62
4	ASNN	E-state values	0.86	1.39	0.6
5	ASNN	E-state counts	0.86	1.44	0.57
6	kNN	SMF	0.92	1.61	0.48
7	ASNN	SMF	0.93	1.45	0.57
8	SVM	SMF	0.93	1.47	0.56
9	ISIDA-average	SMF	0.98	1.62	0.47
10	WEKA	E-state counts	1.04	1.61	0.47
11	WEKA	E-state values	1.04	1.63	0.46
12	WEKA	SMF	1.05	1.69	0.43
13	kNN	E-state counts	1.1	1.76	0.42
14	KNN[a]	E-state values	1.13	1.77	0.4
15	MMLP[a]	E-state values	1.15	2.01	0.37
16	MMLP[a]	SMF	1.16	1.94	0.34
17	ISIDA-single[a]	SMF	1.2	2.39	0.44
18	SVM[a]	E-state values	1.29	1.69	0.56
19	AAM		1.71	2.24	0
		(B) $\log\beta_2(Ag^+)$			
1	MMLP	E-state counts	0.84	1.17	0.79
2	kNN	E-state values	0.96	1.74	0.56
3	SVM	E-state counts	0.97	1.59	0.62
4	kNN	E-state counts	0.98	1.82	0.53
5	ISIDA-average	SMF	1.04	1.62	0.6
6	ASNN	E-state values	1.07	1.85	0.48
7	MMLP	SMF	1.08	1.61	0.62
8	ASNN	E-state counts	1.08	1.82	0.5
9	WEKA	E-state counts	1.08	1.9	0.47
10	ASNN	SMF	1.1	1.54	0.64
11	ISIDA-5	SMF	1.11	1.92	0.5
12	WEKA	E-state values	1.12	1.96	0.43
13	SVM[a]	E-state values	1.18	1.74	0.54
14	kNN	SMF	1.19	1.82	0.5
15	MMLP[a]	E-state values	1.32	2.48	0.42
16	ISIDA-single[a]	SMF	1.4	1.94	0.44
17	SVM[b]	SMF	1.5	2.14	0.31
18	AAM		1.95	2.58	0
		(C) $\log K_1(Eu^{3+})$			
1	SVM	SMF	1.51	2.46	0.77
2	ASNN	SMF	1.55	2.39	0.77
3	SVM	E-state counts	1.66	2.46	0.77
4	ASNN	E-state counts	1.66	2.55	0.75
5	ISIDA-5	SMF	1.7	2.63	0.73
6	ASNN	E-state values	1.79	2.73	0.72
7	WEKA	SMF	1.83	2.91	0.68
8	KNN	E-state counts	1.86	2.79	0.7
9	KNN	SMF	1.95	2.98	0.67
10	SVD[a]	E-state counts	1.95	2.99	0.67
11	MMLP	SMF	1.96	3.22	0.62
12	KNN[a]	E-state values	1.97	2.98	0.67
13	WEKA[a]	E-state counts	1.98	3.08	0.64
14	SVM[a]	E-state values	2.06	2.83	0.72
15	ISIDA-average[a]	SMF	2.06	3.4	0.57
16	WEKA[a]	E-state values	2.07	3.09	0.64
17	ISIDA-single[a]	SMF	2.12	3.27	0.65
18	MMLP[a]	E-state counts	2.23	3.89	0.5
19	SVD[a]	E-state values	2.26	3.62	0.57
20	MMLP[a]	E-state values	2.51	4.59	0.44
21	AAM		4.14	5.19	0
		(D) $\log\beta_2(Eu^{3+})$			
1	SVM	SMF	2.72	3.99	0.48
2	SVM	E-state counts	2.86	4.7	0.36
3	WEKA	E-state values	2.94	4.56	0.33
4	kNN	E-state counts	2.97	4.31	0.43
5	AAM		4.34	5.56	0

[a] The model is significantly different ($p < 0.05$) compared to the top-ranked model of the corresponding set using the bootstrap test (10 000 replicas). [b] The model is significantly different ($p < 0.05$) compared to an arithmetic average model (AAM) according to the Kolmogorov-Smirnov test, but it is not significantly different compared to AAM according to the bootstrap test.[22]

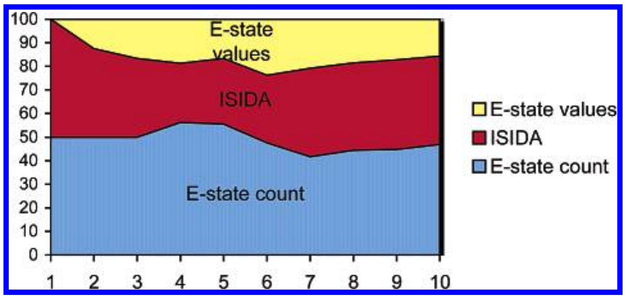


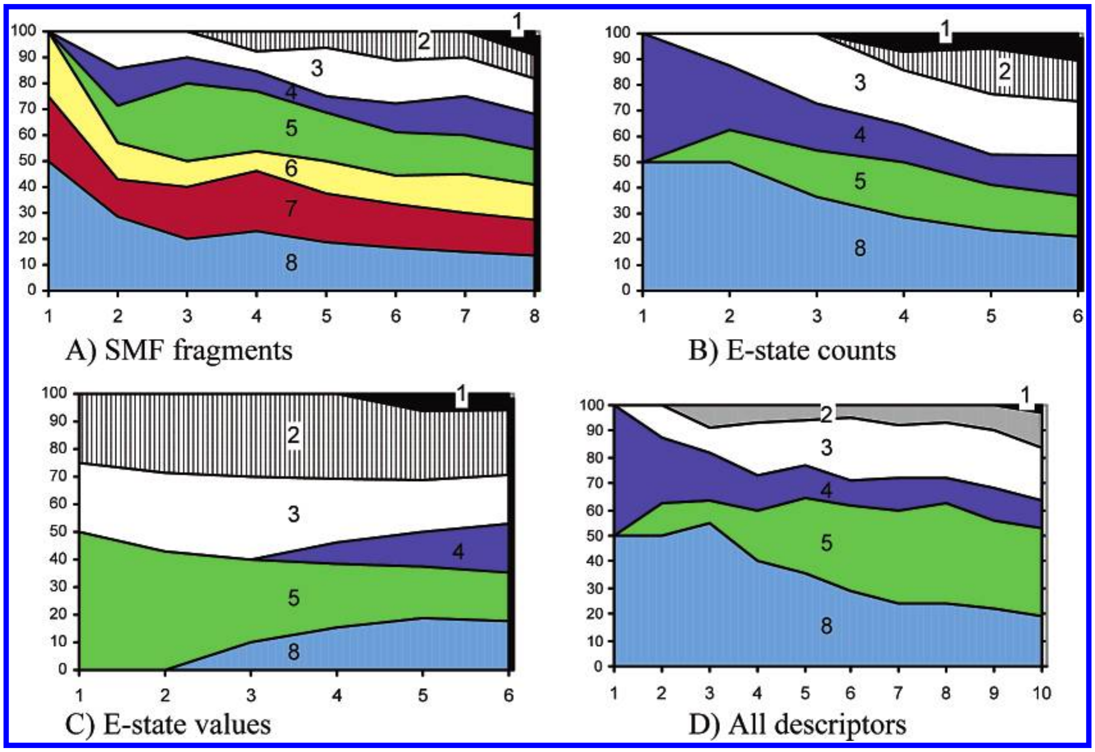
Figure 4. Percentage of models (*y* axis) calculated using the corresponding descriptor system as a function of the number of best models (*x* axis) per each data type. For each data set we selected the *n*-best models (Table 2) and the calculated percents of models generated using each descriptor system. ISIDA-5 and ISIDA-average models generated using SMF fragments were not included in the analysis.

subjective opinions about "apparent performances" of one or another approach.

Cumulative Plots of Descriptors Contributions to Top-Ranked Models. To provide some general estimations of all results, we selected *n* top-ranked models per data set characterized by the smallest MAE (see Table 2) and the analyzed contribution of different descriptors and methods types to these models. Figure 4 show which types of descriptors were more frequently used in the top-ranked models. The plot clearly demonstrates that E-state counts and SMF dominated over the models calculated using E-state values. The performance of all models calculated using the E-state count and the E-state values were significantly different ($p < 0.01$) according to the Wilcoxon signed-rank test thus providing a quantitative confirmation to this observation. The other comparison, i.e., E-state counts or values vs SMF fragments, did not indicate any significant difference. Thus, both SMF fragments and E-state counts, which almost completely correspond to one type of SMF fragments (see Methods), provided similar results.

Cumulative Plots of Methods Contributions to Top-Ranked Models. Figure 5 summarizes the methods, which contributed to the top-ranked models for each type of descriptor. The ISIDA-5 and ISIDA-average models are shown only for a plot corresponding to the SMF fragment types. Both these models provided a considerable number of significant models in the plot. The graphs indicated an apparent preference of some methods. The SVM and MMLP methods have a large number of significant models for SMF fragments and E-state counts, while the RBFNN method produced most of its significant models for E-state indices. ASNN and kNN methods did not demonstrate a clear preference and calculated significant models for all descriptors sets. Of course, one should notice that both ISIDA-5 and ISIDA-averaged were each underrepresented in these plots, since they contributed models with only SMF fragments.

The ISIDA-single models constituted a very small number of significant models for all plots. The Wilcoxon signed-rank test showed that this type of model was significantly different compared to any other models with $p < 0.05$. Thus, any nonlinear approach produced significantly better models compared to a multiple linear regression analysis. One should also notice that SMF fragments used in the comparison study were selected using MLRA, i.e., in some sense these descriptors were optimal for this method. Nevertheless, even this did not improve the prediction ability of the MLRA.



A) SMF fragments

B) E-state counts

C) E-state values

D) All descriptors

Figure 5. 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 the *n*-best models (Table 2) and counted the percents of models contributed using each method. A, B, and C correspond to models calculated using a single descriptor type. 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). SVM provided >50% of all $n = 3$ top-ranked models and SVM + ASNN accounted for >60% of all $n = 5$ top-ranked models.

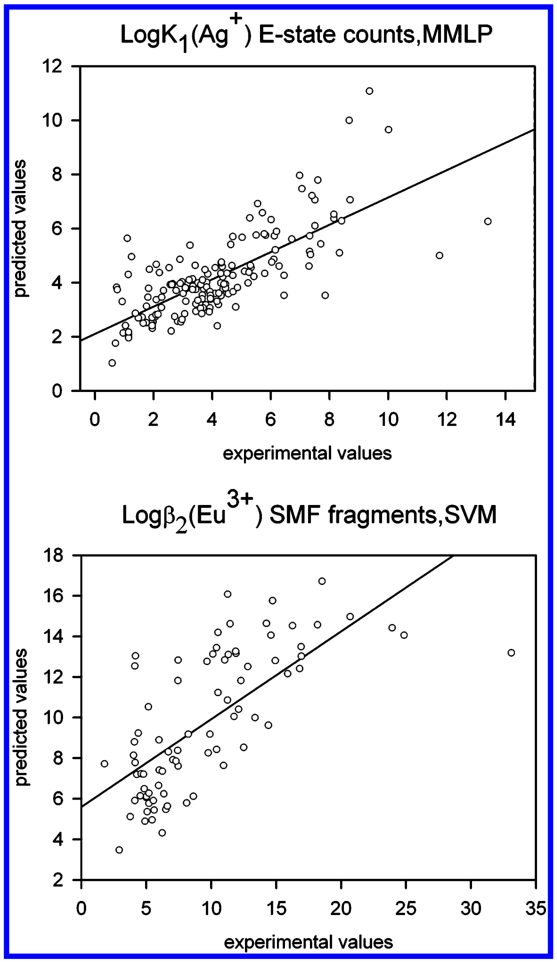


Figure 6. Experimental versus predicted values for models of logK_1(Ag$^+$) and logβ_2(Eu^{3+}). Despite the apparent difference in the quality of both models, the outlying molecules in each model can be easily observed.

On the other hand, the same linear approach but applied in ISIDA-5 or ISIDA-average models provided similar models compared to the best ones obtained with nonlinear approaches. The apparent controversy of these results can be explained by "refusing to predict" molecules with fragments underrepresented in the training set. The policy to predict such "difficult molecules" using another type of fragment dramatically improved the prediction ability of the ISIDA-5 and ISIDA-average models. This interesting strategy can be tested with nonlinear methods, and it may improve their prediction abilities.

ASNN and SVM contributed to more than 60% of the five best models for all analyzed data sets. This result clearly indicates a higher prediction ability and robustness of both these approaches compared to other analyzed methods.

Filtering of Outlying Molecules. One of the factors limiting the quality of models could be a low accuracy of experimental data. As it was mentioned already, the experimental values could contain considerable outliers appearing due to data preparation, experimental conditions, differences in protocols, and simply mistakes. Figure 6 shows examples of some outliers calculated using two top-ranked methods.

Table 3. Outlying Molecules in the Analyzed Data Sets

		predicted value averaged over all models	
molecule no.	exptl value	before outliers filtering	after outliers filtering
		logK_1(Ag$^+$)	
17	11.75	3.97	5.7
60	13.40	6.81	6.27
		logβ_2(Ag$^+$)	
21	14.10	7.06	7.91
96	14.10	7.04	7.61
		logK_1(Eu^{3+})	
10	8.90[a]	14.97	14.32
53	18.87	11.96	12.29
103	26.21	17.16	17.35
177	12.69	6.23	6.26
203	13.75	3.62	4.48
223	22.85	6.96	6.23
		logβ_2(Eu^{3+})	
1	1.78[a]	9.28	8.78
13	24.85	5.33	10.17
14	4.12[a]	12.46	12.79
19	4.17[a]	14.87	15.44
23	14.40	7.13	8.62
41	16.80	10.34	9.1
56	16.92	10.73	8.64
65	33.11	11.09	11
74	23.92	19.1	16.69
77	12.81	8.19	9.63

[a] The reported experimental value of the compound is lower than the predicted value.

We identified molecules as outliers if for at least half of all models they had a difference in experimental and predicted values more than 3 standard errors or if this difference was larger than 6 log units. The outlying molecules, their experimental and average predicted values with all methods, are indicated in Table 3. One can easily notice that most of these molecules have large experimental values (logK_1 > 7 and logβ_2 > 12). Indeed, to measure large constants is very difficult work, as it is discussed in the Discussion, and such molecules tend to have the largest experimental errors.

After filtering outlying molecules we repeated all analyses with "outlier-free" data sets to quantify the effect of outliers on data modeling accuracy. Since different sets are characterized with different MAE errors, we calculated a relative rather than absolute change in the prediction performance of all algorithms, defined as

$$\Delta\mathrm{MAE}(\%) = \frac{100\%}{N}\sum_{i=i,\ldots,N}\frac{\mathrm{MAE}_i' - \mathrm{MAE}_i}{0.5*(\mathrm{MAE}_i' + \mathrm{MAE}_i)}$$

where N is the number of models, and MAE_i and MAE_i' correspond to the error of model i calculated using all the data and the outlier-free set, respectively.

The filtering of outliers did not change the predicted values for compounds that were excluded (Table 3) but improved statistical results for models as shown in Table 4. In general filtering of outliers decreased MAE of the models both for the whole and the outlier-free set for about 5% compared to prediction of the same set using all molecules (Table 4). These decreases were significant according to the bootstrap test at $p < 0.01$.

DISCUSSION

The benchmarking of different methods is a difficult task. Indeed, such comparison requires the same validation

Table 4. ΔMAE(%) of Models Following Outlier Filtering Procedure

descriptors	full data sets	data sets with excluded outliers
SMF fragments	−3.36[a]	−5.84
E-state values	−1.76	−3.91
E-state counts	−4.61	−4.33
average value	−3.39	−4.69

[a] The negative values indicate that MAE decreased following the outlier filtering procedure.

protocol for all methods. It may involve optimization of tons of parameters, including some nonapparent ones such as data normalization. Since it is impossible to try all possible combinations of parameters for any method, particularly if there are more than 3 parameters, any such comparison provides only a suboptimal result. A use of incorrect parameters of methods or nonstrict following the protocol (i.e., if the parameters for some method are selected by its performance on external rather than on internal cross-validation sets) may completely invalidate conclusions of any benchmarking study. In this work we tried our best to compare the different approaches applying the model selection used by the expert in the field and strictly following the guidelines provided by the authors of the original approaches. Moreover, the contributors of this study are the authors of several approaches (ISIDA, ASNN, MMLP) or are experts in them (SVM and RBFNN). The MLRA approach used in this study is part of the TRAIL[14,15,18,19,32,33] and now the ISIDA program.[24]

The use of a rigorous validation procedure allowed us to conclude that nonlinear methods provide significantly better results compared to the MLRA using the same set of descriptors. It should be noticed, however, that we sometimes indeed observed a considerable discrepancy in cross-validation results calculated in "internal" and "external" cross-validation procedures. Certainly, sometimes the internal validation errors were 2−3 times lower than the external one. This result simply demonstrates that one should not use the internal cross-validation as a measure of the prediction ability of the (nonlinear) method. For example, after 3300 grid search optimizations to select the best parameters for SVM using the whole data set, one should not use the received "best cross-validation" MAE as a real measure of the prediction performance of the SVM! While the model selected in this way is possibly good, the received MAE estimation is completely wrong. It is biased by the model optimization procedure. One can get a much better estimation of the actual performance of the analyzed method by running external *n* cross-validation, i.e., by selecting $1/n$ part of the data as a test set, optimizing the method on the remaining $(n - 1)/n$ data set using the same grid search, predicting the $1/n$ test set, and repeating this procedure *n* times. Such an estimation is not overfitted due to the model parameter selection procedure. In fact, the same overfitting problem appearing due to optimization of model parameters was already analyzed by us with respect to the artificial neural network a decade ago.[47] The only difference that a role of neural network overfitting is performed by the parameter selection procedure itself.

The used methods have different computational power requirements. The longest time was required to run ASNN. The calculation of ASNN results required one weekend on a cluster with 14 computers (Athlon/Pentium 2−3 GHz). On the contrary, similar calculations with all other methods, except MMLP, were completed in less than 12 h on a single Athlon64 3000 notebook. The longest time was required for the WEKA RBFNN function optimization and the shortest for both kNN and MLRA. The MMLP method required several hours to perform its analysis on a single Athlon 3 GHz computer. Taking into account method performance/computational speed, the SVM algorithm overperforms the other approaches.

The quality of experimental data is crucial for a good performance of methods. As it was mentioned in the Results section, most of the outliers had large reported experimental values. In fact, studies of equilibrium corresponding to large complexation constants are very difficult.[2,3] This requires extreme accuracy in the preparation of initial solutions. Sometimes, a tiny error (less than 1%) in initial concentrations of the reagents may provoke a systematical error in the constant value of more than one unit in log scale.[48] It should be also noted that only a few experimental methods (e.g., potentiometry) could be applied to measure relatively large constants ($\log K_1 > 7$). The methods such as calorimetry and NMR and IR spectroscopy may be applied only in the framework of some special techniques, e.g., competition complexation experiments. The measurements of overall constant $\log\beta_2$ provides further complications since it requires even more accurate assumptions about the chemical model of equilibrium in solution concerning a certain number of species and their stoichiometry. Therefore, it is not surprising that experimental data for $\log\beta_2$ have lower quality, and as a result models are less accurate and may contain a larger number of outliers compared to $\log K_1$ data, as it is exemplified at Figure 6.

CONCLUSIONS

A benchmark study of several popular methods, Associative Neural Networks, Support Vector Machines, k Nearest Neighbors, Maximal Margin linear programming, Radial Basis Function, and Multiple Linear Regression, has been performed for QSPR studies of stability constants $\log K_1$ for the 1:1 (M:L) and $\log\beta_2$ for 1:2 complexes of metal cations Ag^+ and Eu^{3+} with diverse sets of organic molecules in water at 298 K and ionic strength 0.1 M. Three types of descriptors, E-state indices and counts and SMF descriptors, were used. Comparison of the models was performed using a 5-fold external cross-validation procedure.

Individual structure−complexation property models obtained with nonlinear methods calculated a significantly better performance than the models built using multilinear regression analysis ($p < 0.05$). However, the averaging of several MLRA models based on SMF descriptors provided as good of a prediction as the most efficient nonlinear techniques. No overfitting of nonlinear methods and no significant differences in performances of nonlinear approaches were detected. The filtering of outlying molecules increased statistical parameters of models ($p < 0.05$). Support Vector Machines and Associative Neural Networks contributed the largest number of significant models. Considering a relatively short computation time among nonlinear techniques, SVM and kNN could be proposed as a good choice for the modeling of complexation constants of metals.

Models based on fragments (SMF descriptors and E-state counts) had a higher prediction ability than those based on E-state indices. The use of SMF descriptors and E-state counts provided similar results, whereas E-state indices lead to less significant models. Thus, one may recommend fragment descriptors for QSPR studies of metal complexation.

Glossary and Related References. ALOGPS, program to predict lipophilicity and aqueous solubility of compounds;[27,41] AAM, Arithmetic Average Model; ASNN, Associative Neural Networks;[39,41] CVTS, Cross-Validation Training Set; E-state count, count of atoms corresponding to atom types in E-state indices; E-state values, descriptors developed by Kier and Hall which incorporate topological and electronic structure of molecules;[17,21] ISIDA, software program for QSPR/QSAR studies;[24] ISIDA-5, model calculated as average of 5 best models selected by internal cross-validation using the ISIDA program; ISIDA-average, model calculated as average of all models calculated in the ISIDA program; ISIDA-single, MLRA SVD model calculated using one descriptor set; kNN, k Nearest Neighbors method; KS, Kolmogorov Smirnov statistics;[22] LibSVM, implementation in C++/Java of SVM approach;[37] MAE, Mean Absolute Error; MLRA, Multiple Linear Regression Analysis; MMLP, Maximal Margin Linear Programming;[43,44] QSAR, Quantitative Structure−Activity Relationships; QSPR, Quantitative Structure−Property Relationships; R^2, square of correlation coefficient between predicted and experimental values; RBFNN, Radial Basis Function Neural Network; REC, Regression Error Characteristic Curve;[46] RMSE, Root Mean Squared Error; SC DB, Stability Constants Database, http://www.acadsoft.co.uk; SMF, Substructural Molecular Fragments;[14] SVD, Singular Value Decomposition; SVM, Support Vectors Machine;[36] TRAIL, software program for QSPR studies, which preceded ISIDA;[14,49] VCCLAB, Virtual Computational Chemistry Laboratory site,[31,50] http://www.vcclab.org; WEKA, software package to perform machine learning data mining.[42]

ACKNOWLEDGMENT

The authors thanks Dr. N. Bukar for help with the preparation of data sets used in this article and Université Louis Pasteur for the Invited Professor position to I.V.T. F.H. thanks “NOVALYST DISCOVERY” company for the Ph.D. fellowship. Professor L. Pettit and Dr. G. Pettit are acknowledged for the help of subsets selection from SC-DB. Part of this work has been performed in the framework of French−Russian collaborative project GDRE “SupraChem”.

Supporting Information Available: The data sets with calculated descriptors of molecules and REC curves for all methods. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) *Comprehensive Coordination Chemistry II: From Biology to Nanotechnology*; Elsevier: Amsterdam, 2003; Vol. 1−10.

(2) Beck, M.; Nagypal, I. *Chemistry of Complex Equilibria*; Akademiai Kiado: Budapest, 1989.

(3) Hartley, F. R.; Burgess, C.; Alcock, R. M. *Solution Equilibria*; Ellis Horwood: New York, 1980.

(4) Leach, A. R. *Molecular Modelling. Principles and Applications*; Longman: Singapore, 1996.

(5) Dimmock, P. W.; Warwick, P.; Robbins, R. A. Approaches to predicting stability constants. *Analyst* **1995**, *120* (8), 2159−70.

(6) Hancock, R. D.; Martell, A. E. Ligand Desing for Selective Complexation of Metal Ions in Aqueous Solution. *Chem. Rev.* **1989**, *89* (8), 1875−1914.

(7) Hancock, R. D. Approaches to Predicting Stability Constants. A Critical Review. *Analyst* **1997**, *122*, 51R−58R.

(8) Varnek, A. A.; Glebov, A. S.; Kuznetsov, A. N. Charge Density Distribution, Electrostatic Potential and Complex Formation Ability of Some Neutral Agents. *Portugal Phys.* **1988**, 59−61.

(9) Varnek, A. A.; Kuznetsov, A. N.; Petrukhin, O. M. Electrostatic Potential Distribution and Extraction Ability of Some Organophosphorus Compounds. *Zh. Strukt. Khim. (Rus.)* **1989**, *30*, 44−48.

(10) Varnek, A. A.; Kuznetsov, A. N.; Petrukhin, O. M. Calculation of the Indexes of Extractability of Some Neutral Organo-Phosphorus Compounds Within the Framework of Electron Density Functional Method. *Koord. Khim. (Russ.)* **1991**, *17*, 1038−1043.

(11) IUPAC Stability Constants Database, Version 5.33; http://www.acadsoft.co.uk/(accessed June 2004), 2004.

(12) Solov’ev, V. P.; Stuklova, M. S.; Koltunova, E. V.; Kochanova, N. N. Coordination numbers of central atoms in coordination compounds. *Russ. J. Coord. Chem.* **2003**, *29* (9), 660−668.

(13) *The Cambridge Structural Database*; http://www.ccdc.cam.ac.uk/products/csd/:2005.

(14) Solov’ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 847−858.

(15) Varnek, A.; Wipff, G.; Solov’ev, V. P.; Solotnov, A. F. Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 812−829.

(16) Solov’ev, V. P.; Varnek, A. A. Structure−property modeling of metal binders using molecular fragments. *Russ. Chem. Bull.* **2004**, *53* (7), 1434−1445.

(17) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039−1045.

(18) Varnek, A.; Wipff, G.; Solov’ev, V. P. Towards an information system on solvent extraction. *Solvent Extr. Ion Exch.* **2001**, *19* (5), 791−837.

(19) Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov’ev, V. P.; Varnek, A. Quantitative Structure−Property Relationship Modeling of *beta*-Cyclodextrin Complexation Free Energies. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 529−541.

(20) Varnek, A.; Fourches, D.; Solov’ev, V. P.; Baulin, V. E.; Turanov, A. N.; Karandashev, V. K.; Fara, D.; Katritzky, A. R. “In Silico” Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1365−1382.

(21) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; London, 1999.

(22) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++. The Art of Scientific Computing*, 2nd ed.; Cambridge, 2002; p 1002.

(23) Kiss, T.; Sovago, I.; Gergely, A. Critical Survey of Stability Constants of Complexes of Glycine. *Pure Appl. Chem.* **1991**, *63* (4), 597−638.

(24) Varnek, A. 2005, http://infochim.u-strasbg.fr/recherche/isida/.

(25) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indexes. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 947−955.

(26) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1488−1493.

(27) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1407−1421.

(28) Taskinen, J.; Yliruusi, J. Prediction of physicochemical properties based on neural network modelling. *Adv. Drug. Deliv. Rev.* **2003**, *55* (9), 1163−1183.

(29) Butina, D. Performance of Kier-Hall E-state Descriptors in Quantitative Structure Activity Relationship (QSAR) Studies of Multifunctional Molecules. *Molecules* **2004**, *9*, 1004−1009.

(30) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13* (2), 223−241.

(31) Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **2005**, *10* (22), 1497−1500.

(32) Solov'ev, V. P.; Varnek, A. Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure−Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1703−1719.

(33) Varnek, A.; Solov'ev, V. P. "In Silico" Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Comb. Chem. High Throughput Screening* **2005**, *8* (5), 403−416.

(34) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* 2005.

(35) Muller, P. H.; Neumann, P.; Storm, R. *Tafeln der mathematischen Statistik*; VEB Fachbuchverlag: Leipzip, 1979.

(36) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(37) Chang, C. C.; Lee, C. J. LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.: 2001.

(38) Keerthi, S. S.; Lin, C. J. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.* **2003**, *15* (7), 1667−89.

(39) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 717−728.

(40) Tetko, I. V. Associative neural network. *Neural Process. Lett.* **2002**, *16*, (2), 187−199.

(41) 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), 1136−1145.

(42) Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **2004**, *20* (15), 2479−81.

(43) Antonov, A. V.; Tetko, I. V.; Mader, M. T.; Budczies, J.; Mewes, H. W. Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics* **2004**, *20* (5), 644−652.

(44) Antonov, A. V. Manuscript in preparation.

(45) Good, P. I. *Permutation, Parametric and Bootstrap Tests of Hypotheses. A Practical Guide to Resampling Methods for Testing Hypotheses*, 3rd ed.; Springer-Verlag: New York, 2005; p 315.

(46) Bi, J.; Bennett, K. P. *Regression Error Characteristic Curves, Proceedings of the 20th International Conference on Machine Learning (ICML 2003), Washington, DC, 2003*; Washington, DC, 2003.

(47) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (5), 826−833.

(48) Solov'ev, V. P.; Vnuk, E. A.; Strakhova, N. N.; Raevsky, O. A. *Thermodynamics of Complexation of the Macrocyclic Polyethers with Salts of Alkali and Alkaline-Earth Metals (Russ.)*; VINITI: Moscow, 1991.

(49) Raevskii, O. A.; Sapegin, A. M.; Chistyakov, V. V.; Solov'ev, V. P.; Zefirov, N. S. The Forming of the Models of the Relationships Structure-Complexing Ability. *Koord. Khim.* **1990**, *16* (9), 1175−1184.

(50) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D. J.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry Laboratory − Design and Description. *J. Comput.-Aided Mol. Des.* **2005**, *19* (6), 453−463.

CI0504216