J Incl Phenom Macrocycl Chem (2013) 76:159–171
DOI 10.1007/s10847-012-0185-x

ORIGINAL ARTICLE

QSPR ensemble modelling of alkaline-earth metal complexation

V. P. Solov'ev · N. Kireeva · A. Yu. Tsivadze · A. Varnek

Received: 17 August 2011 / Accepted: 25 May 2012 / Published online: 13 June 2012

Abstract QSPR ensemble modeling of the stability constant log *K* of the complexes of Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} with diverse 273 (Mg^{2+}), 284 (Ca^{2+}), 147 (Sr^{2+}) and 198 (Ba^{2+}) organic ligands in water for the $M^{2+} + L = (M^{2+})L$ equilibrium at 298 K and an ionic strength 0.1 M has been performed. For each compound, predicted log *K* was calculated as an arithmetic average over the outputs of individual multiple linear regression models based on fragment descriptors. The root mean squared errors in fivefold cross-validation are 0.75 (Mg^{2+}), 0.77 (Ca^{2+}), 0.72 (Sr^{2+}) and 0.87 (Ba^{2+}). Additional external validation of the models has been performed on the complexes of 11 ligands recently reported in the literature. Several methodological developments related to (i) descriptors selection for an individual model and (ii) discarding redundant models have been proposed. Developed models have been integrated in the COmplexation of METals (COMET) predictor available as WEB application.

Keywords QSPR modeling and prediction of stability constants · Design of metal binders · Selectivity · Complexes of Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} with organic ligands in water

V. P. Solov'ev · N. Kireeva · A. Yu. Tsivadze
Institute of Physical Chemistry and Electrochemistry,
Russian Academy of Sciences, Leninskiy Prospect, 31a,
119991 Moscow, Russian Federation

N. Kireeva · A. Varnek (✉)
Laboratoire d'Infochimie, UMR 7177 CNRS, Université
de Strasbourg, 4, rue B. Pascal, 67000 Strasbourg, France
e-mail: varnek@chimie.u-strasbg.fr

Introduction

Thermodynamic stability of complexes of alkaline-earth metal ions (M^{2+}) with organic ligands is important for selective separation of Ba^{2+} (Ca^{2+}) ion from Sr^{2+} (Mg^{2+}) [1], for recovery radioactive ^{90}Sr [1], and for quantitative assessments of interactions of M^{2+} with bio ligands in living organisms [2, 3]. Theoretical assessment of stabilities of the metal/ligand complexes [4] provides researchers with a way to reduce the number of experiments, to find the strategy of "optimization" of known ligands and to design new selective metal binders [4, 5].

Up to now, a few QSPR linear modeling of alkaline-earth metal ions complexation have been reported. Shi et al. [6] reported the models built on 314 stability constants of Ca^{2+}, Na^{+} and Zn^{2+} complexes with crown ethers, cryptands and spherands in different pure and mixed solvents using as descriptors some force field energy components, internal strain energy of ligands, surface tension and dipole moments of solvent, charge and ionic radii of metal cations. Different types of organic ligands were involved in the modeling of calcium [7–9] and magnesium [8, 10, 11] complexes in water: amino acids, adenosine derivatives, heterocyclic compounds [8, 10, 11], as well as the ligands containing carboxylate, phenol, amine, ether, and alcohol functional groups [9]. Raevsky et al. [7] used molecular fragments, topological indices and some physicochemical descriptors for QSPR modeling of complexation of 56 ligands with Ca^{2+} in water. Toropov et al. [8, 10, 11] developed the models based on topological indices for the data sets containing 110 [11] and 150 [8] compounds. Fragment descriptors [9] were used for QSPR modeling of the complexation of Ca^{2+} with 42 organic ligands. Ensemble models for the stability constants of the 1:1 complexes of Sr^{2+} with 130 organic ligands have been developed by Solov'ev et al. [12] using substructural molecular fragments (SMF) as descriptors. The models described above were either not validated at all or their validation have been performed only on one selected test set. The applicability domain of the models have never been used. These, certainly, weaken the practical application of the reported models for computer-aided design of new metal binders.

In this paper, we report the QSPR ensemble modeling of the stability constant log *K* of the 1:1 (M:L) complexes of metal cations Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} with organic ligands in aqueous solution using multiple linear regression (MLR) approach and SMF descriptors. The diversity and the size of the involved datasets significantly exceed those used in references [6, 8, 9, 12]. Compared to previous studies, three main achievements have been reached:

(i) External cross-validation procedure has been used to assess the predictive performance of the models.
(ii) The applicability domain has been assessed for each individual model.
(iii) The models are available for the end users via WEB interface.

Ensemble modeling implies generation of several QSPR models, selection of the most pertinent ones, followed by their simultaneous application to a given test compound. In this case, the predicted value is estimated as an arithmetic average of those calculated by selected individual models (IM). The performance of this "consensus" model (CM) depends on that of each individual model, on one hand, and on the composition of the subset of selected IM, on the other hand. The first issue concerns the selection of the most relevant molecular descriptors from the initial descriptors pool whereas the second one investigates the redundancy (collinearity) of different QSPR models. Here, we describe some methodological developments related to these two key aspects of QSPR ensemble modeling.

Method

Data preparation

Experimental stability constant values (log *K*) for the 1:1 (M:L) complexes of Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} cations with organic ligands in water were critically selected from IUPAC stability constants database (SC DB) [13] (version 5.33, Academic Software) at standard temperature 298 K and an ionic strength $I = 0.1$ M. Some of the log *K* values were adjusted to specified temperature and ionic strength using the procedures included in SC DB.

2D structures of the ligands, names of the metal ions as well as corresponding experimental log *K* values were converted by the EdiSDF data manager [12, 14, 15] into structure data files (SDF) used as an input in the MLR module of the in silico design and data analysis (ISIDA) package [16, 17]. If several values of the stability constant log *K* were available for a given ligand, we followed the recommendations of IUPAC [18] to select the most appropriate one. In some cases the most recent data or the data consistent with respect to different experimental methods were chosen. 273 (Mg^{2+}), 284 (Ca^{2+}), 147 (Sr^{2+}) and 198 (Ba^{2+}) organic ligands were involved in the QSPR modeling. Distributions of the experimental values log *K* in the data sets are given in Fig. 1. For the studied complexes, the values log *K* vary in the range of 0.1–11.2 (Mg^{2+}), 0.1–14.1 (Ca^{2+}), 1.1–12.4 (Sr^{2+}) and 0.2–11.8 (Ba^{2+}).

The names of the ligands and the stability constant values are given as supporting information in Tables SM1–SM4. As a rule, the organic ligand bears several electron-donor groups. The sets of the ligands include amino and hydroxy derivatives of carboxylic acids; different amino-acids and their oligomers, alkylated derivates of phosphoric acid; alkyl- and aminophosphonic acids; acyclic polydentate ligands with the terminal carboxy groups separated by various cyclic or acyclic spacers; derivatives of diphosphonic acids; ternary amines with phosphono and carboxy groups; mono- and dipodands of ternary amines; amino derivatives of phenols; crown-ethers, thia-, and aza-crown-ethers with neutral and acidic lariat groups, cryptands, etc. (see Tables SM1–SM4).

Descriptors

SMF of the ISIDA package [14, 19] were used as descriptors in the QSPR models. Each fragment represents a subgraph of a molecular graph, whereas its occurrence is a descriptor value. Molecules were represented with implicit hydrogen atoms. Two subclasses of the SMF descriptors were used: (i) shortest topological paths with explicit representation of all atoms and bonds and (ii) shortest topological paths with explicit representation of only terminal atoms and bonds. The Floyd algorithm [20] was used for finding the shortest paths in the molecular graphs. Single, double, triple and aromatic bonds were recognized, where bonds in cycles and in chains were considered differently. For every subclass of the sequences, the minimal ($n_{min} \geq 2$) and maximal ($n_{max} \leq 15$) numbers of atoms are defined. Thus, IAB(n_{min}–n_{max}) and IAB(n_{min}–n_{max})t represent two subclasses of the SMF descriptors that include all intermediate shortest paths with *n* atoms, for which $n_{min} \leq n \leq n_{max}$. Varying the values of n_{min} and n_{max}, 210 types of the sequences of two subclasses were generated. Concatenated fragments always occurring in the same combination in each compound of the training set were considered as one extended fragment.

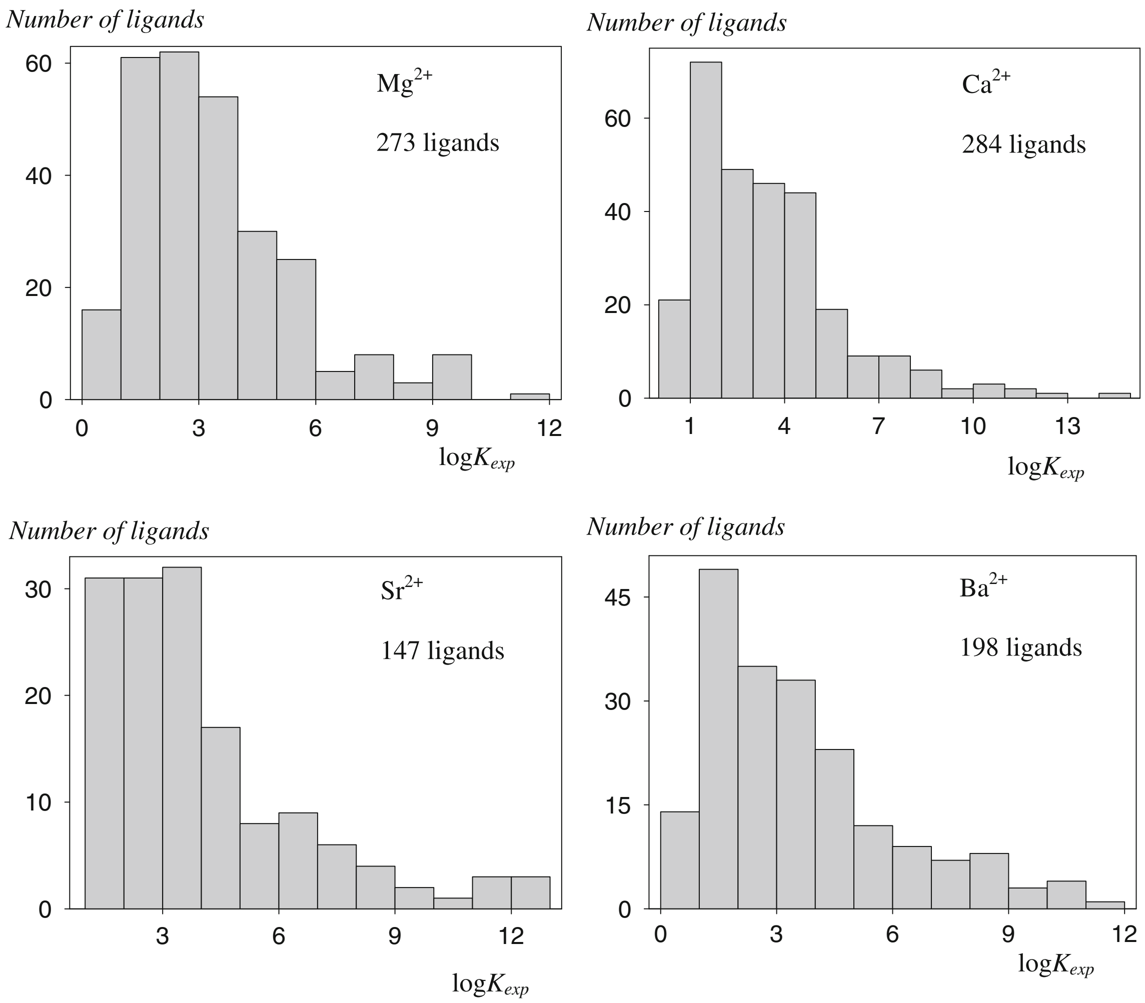


Fig. 1 Distribution of experimental values of the stability constant log *K* for the 1:1 (M:L) complexes of organic ligands with Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} in water at 298 K and ionic strength 0.1 M

Obtaining, selection and validation of the models

The ISIDA/MLR program [15–17] has been used for the modeling. MLR establishes a linear relationship between two or more independent variables (molecular descriptors) and a response variable (property) $Y = a_0 + \Sigma a_i X_i$ to observed data, where every descriptor value X_i (in our case, occurrence of fragment descriptor) is associated with a property value, a_i is descriptor contribution, and a_0 is the independent term. Here, the singular value decomposition method [21] is used to search the adjustable coefficients a_i of input variables so as to minimize the squared difference between the values calculated by the models and actual observed values of the property in the training set.

The ISIDA/MLR program generates many MLR models, each of them corresponds to different initial subset of the SMF descriptors. The leave-one-out (LOO) cross–validation correlation coefficient *Q,* corresponds to the stability of models and is accepted as a criterion of model selection: only the models for which $Q^2 > Q^2_{lim}$, where Q^2_{lim} is a user defined threshold, are selected. In this work, $Q^2_{lim} = 0.5$ was used.

The traditional technique for model validation implies the split of the initial data set into training and test sets. The training set is used in model development whereas the test set is used only for the model validation. In this work, the fivefold external cross validation (5-CV) was used to evaluate the predictive performance of models [5, 22]. In this procedure, the entire dataset is split into five non-overlapping pairs of training and test sets. Each training set covers 4/5th of the initial dataset while the related test set covers the remaining 1/5th. The models developed on the *i*-th training set have to be applied to the corresponding test set, thus, all the molecules of the initial data set are predicted.

Coefficient of determination R^2 and root mean squared error *RMSE* were used to evaluate the model ability to reproduce quantitatively the experimental data:

$$R^2 = 1 - \Sigma(Y_{exp} - Y_{pred})^2/\Sigma(Y_{exp} - <Y>_{exp})^2,$$

$$RMSE = (\Sigma(Y_{exp} - Y_{pred})^2/n)^{1/2} \quad \text{and} \quad MAE = \Sigma \mid Y_{exp} - Y_{pred} \mid /n,$$

where Y_{pred} and Y_{exp} are predicted and experimental values (here, $Y = \log K$).

Ensemble modeling implies generation of several QSPR models, selection among them the most pertinent ones, followed by their joint application to a given test compound. Thus, for each compound from the test set, the program computes the property as an arithmetic average of values obtained with an ensemble of selected at the training stage IM. IM leading to outlying values according to the ranked series method [24] are excluded. The application of such "consensus" model (CM) compensates inaccuracies of the individual model predictions [12, 17, 22, 23, 25, 26]. In order to make a decision whether or not a QSPR model can be applied to a given compound, the concept of applicability domain (AD) of models is used. Generally, the AD of the model is associated with the area of chemical space occupied by the training set. Applying an individual model, the program checks its AD measuring the similarity between the test compound and the compounds from the training set. If given compound is identified as being outside AD, the predicted value for this compound by given model is excluded from CM. In this work, two AD approaches were applied: (i) bounding box which considers AD as a multidimensional descriptor space confined by minimal and maximal values of occurrences of SMF descriptors involved in individual model [15], and (ii) fragment control [15] which discards the query compounds containing fragments different from those in the training set.

Variable selection algorithms in ISIDA/MLR

Combined forward and backward stepwise techniques have been used to select the most pertinent variables from initial pool of the SMF descriptors. Previous versions of the ISIDA/MLR program used either entire initial descriptors pool [19, 23, 27] or a part of descriptors selected by backward stepwise variable selection (BVS) algorithm [23]. BVS eliminates the variables with low $t_i = a_i/\Delta a_i$ values, where Δa_i is a standard deviation for the coefficient a_i at the i-th variable in the model. First, the program selects the variable with the minimal $t_{min} < t_0$, then it builds a new model excluding this variable. This procedure is repeated until $t \geq t_0$ for all selected variables. Here t_0 is the tabulated value of Student's criterion. By default, t_0 equals 1.96. For non-collinear descriptors, the used SVD method [21] allows one to calculate the values Δa_i, if the initial number of descriptors (M) does not exceed the size (N) of the training set.

In order to apply the BVS technique [5, 14, 22, 23, 28–31] to large initial set of descriptors ($M >> N$), a new forward stepwise variable selection (FVS) algorithm has been developed to pre-select the user-defined number of the most relevant variables ($M_p < N$). The FVS employs the known equations for the correlation coefficients between the response variable Y and one- two- and three variables [32] in combination with the FSMLR algorithm [33]. Accordingly, three sub-algorithms (FVS-1, FVS-2 and FVS-3) have been developed. At step p, the FVS procedure defines a new response variable $Y^{(p)} = Y^{(p-1)} - Y_{calc}$, where $Y_{calc} = c_0 + c_iX_i$ (FVS-1), $Y_{calc} = c_0 + c_iX_i + c_jX_j$ (FVS-2), $Y_{calc} = c_0 + c_iX_i + c_jX_j + c_kX_k$ (FVS-2), $p = 1, 2, 3,\ldots$ and $Y^{(0)} = Y_{exp}$. Thus, one (X_i), two (X_i, X_j) or three variables (X_i, X_j and X_k) are selected to maximize the correlation coefficients ($R_{y,i}$, $R_{y,ij}$ or $R_{y,ijk}$ correspondingly) between the variable(s) and $Y^{(p)}$. This is repeated until the number of selected variables M_p reaches a user-defined value. It should be noted that the sub-algorithm FVS-1 corresponds to fast stepwise multiple linear regression [33]. Optionally, variables X_m with small correlation coefficient with $Y^{(p)}$ ($|R_{y,m}| < R^0_{y,m}$), those highly correlated with other variables X_i ($|R_{i,m}| > R^0_{i,m}$) or "rare" fragments (*i.e.*, found in less than q molecules, here $q < 2$) and can be eliminated.

The efficiency of the FVS procedure was compared with different implementations of Genetic Algorithm. They were applied to the Selwood data set [34] comprising 31 compounds and to the QSAR modeling of different types of anti-HIV activity for three families of compounds [31]. The results show similar predictive performance of computationally expensive GA-based approaches and FVS calculations.

The optimization of the FVS algorithm parameters was carried out where the predictive performance of CMs was evaluated as a function of the FVS sub-algorithm, the number of preselected descriptors M_p and a collinearity of descriptors using 5-CV procedure. The number of preselected descriptors $M_p = kN$ was systematically varied in the range of k from 0.1 to 0.9, where N is the training set size. The QSPR modeling was performed using the FVS-1, FVS-2 and FVS-3 sub-algorithms on the data sets of experimental stability constants log K for the (Ca^{2+})L and (Sr^{2+})L complexes. For the sub-algorithms, optimal number of preselected descriptors M_p can be recommended to kN, where $k = 0.6$ and N is the number of data points in the training set (Figs. 2). In the range of k from 0.3 to 0.7, the sub-algorithm FVS-2 is preferable because of the reasonable performance of predictions as compared with FVS-1 (Fig. 2) and time of calculations as compared with FVS-3. FVS-2 is in several times faster than FVS-3.

Collinearity of descriptors The presence of highly correlated variables can cause the instability of IM. However,

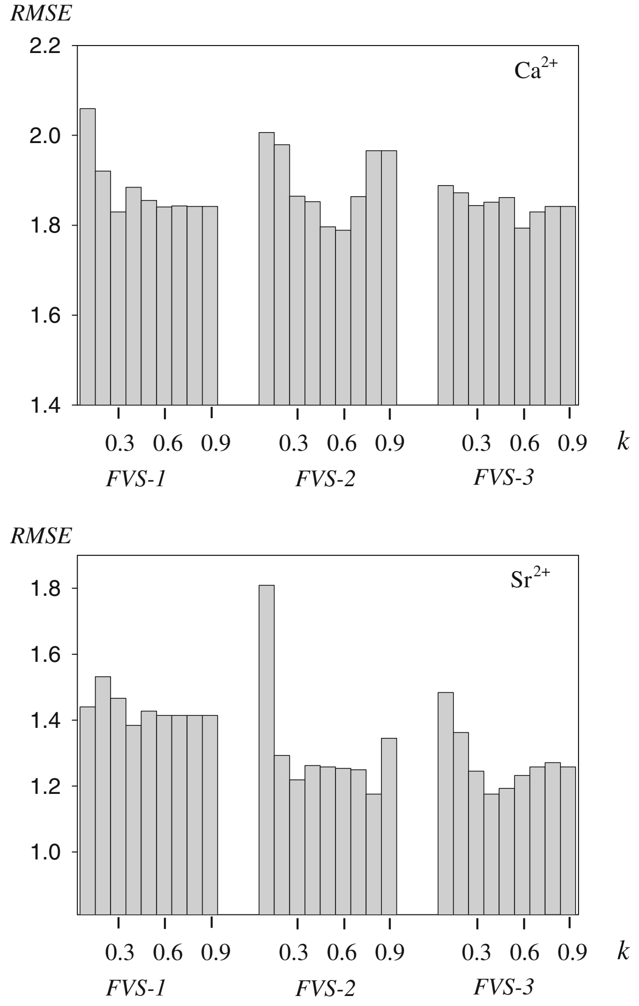


Fig. 2 Root mean squared error *RMSE* of predicted stability constant log K_{pred} (5-CV procedure) for the 1:1 (M:L) complexation of organic ligands with Ca^{2+} and Sr^{2+} as a function of the type of the FVS sub-algorithm and the number of preselected descriptors $N_d = kN$, where N is the training set size and k is the coefficient varied in the range from 0.1 to 0.9

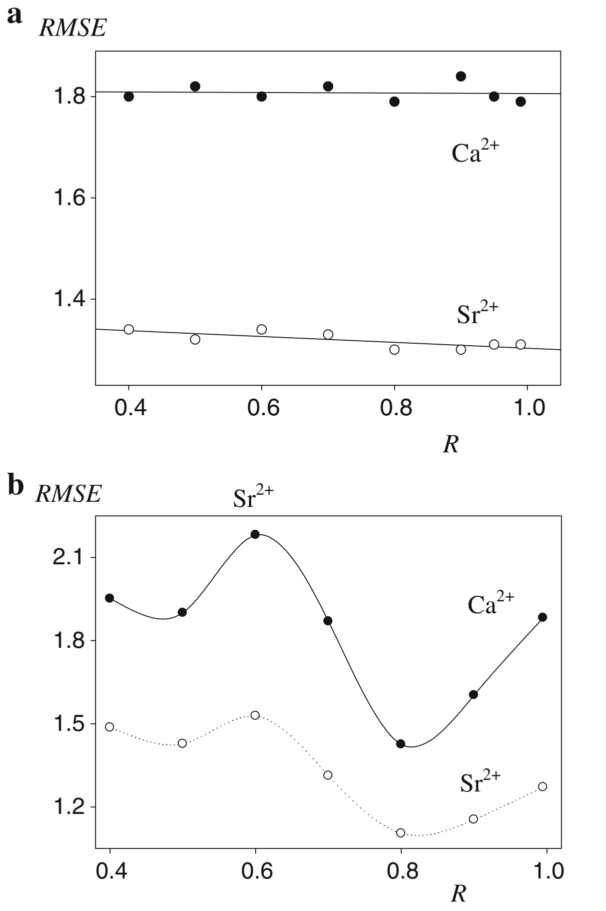


Fig. 3 Root mean squared error *RMSE* of predicted stability constant log K_{pred} (5-CV procedure) for the 1:1 (M:L) complexes of organic ligands with Ca^{2+} and Sr^{2+} as a function of: **a** descriptor collinearity and **b** the collinearity of the IM. Here, R is the Pearson' correlation coefficient

this problem was never studied for the consensus model involving ensemble of IMs. Here, we measured the consensus model's performance (*RMSE* in 5-CV) as a function of the Pearson' correlation coefficient R_{ij} between two descriptor vectors which determines descriptors' collinearity. Thus, variables X_j correlated with already preselected ones X_i were eliminated if $|R_{ij}| > R^0$. The threshold value R^0 was systematically varied in the range from 0.40 to 0.99. Figure 3a shows that *RMSE* practically doesn't vary with R_{ij} just showing that the descriptor collinearity has very little impact on the predictive ability of CMs. One may assume that negative influence of multicollinearity on the predictive ability of each individual model is smoothen in CM involving several IMs.

Collinearity of IM

The ISIDA/MLR program generates an ensemble of IM, some of which can be redundant. The question arises how does collinearity of IM impact on predictive ability of CM? The Pearson' correlation coefficient between residuals vectors of two models R_{rm} has been used to measure the IMs collinearity. For a given IM, the residual vector contains N components (N is the training set size) in which its *i*-th component is the difference between experimental and predicted property values for *i*-th molecule. An individual model *r* is excluded from CM if it correlates ($|R_{rm}| > R^0$) with another already selected model *m*. The threshold R^0 was systematically varied in the range from 0.40 to 0.99. One can see (Fig. 3) that discarding highly correlated IM reduces *RMSE* thus leading to the improvement of the CM performance. Further removal of correlated models decreases the accuracy of predictions. Thus, $R^0 \approx 0.8$ looks

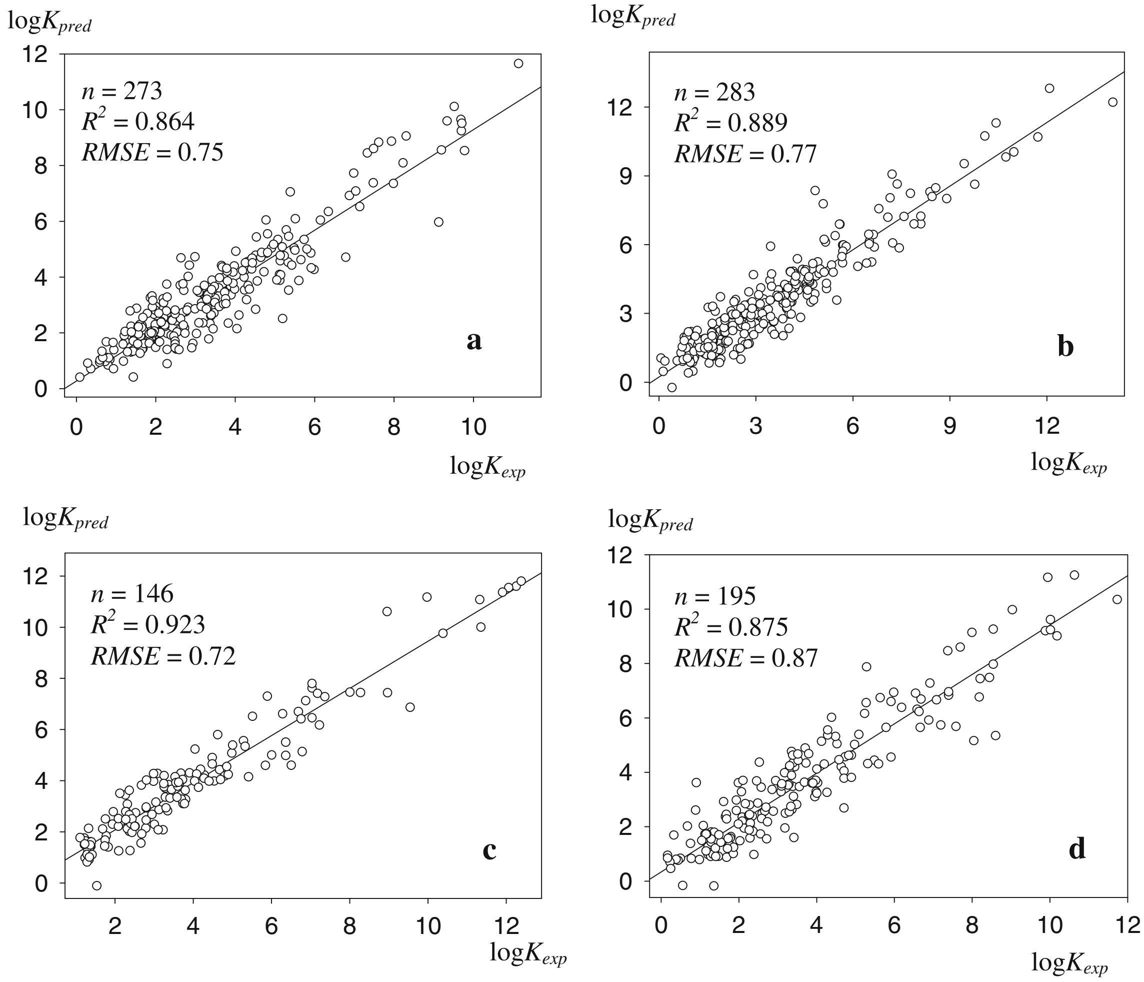


Fig. 4 Predicted *vs*. experimental values (5-CV procedure) of the stability constant (log *K*) for the 1:1 (M:L) complexation of organic ligands with **a** Mg^{2+}, **b** Ca^{2+}, **c** Sr^{2+} and **d** Ba^{2+}. Determination coefficient (R^2) and root mean squared error (*RMSE*) correspond to $y = x$

as an optimal degree of the models' collinearity, which is used in the further calculations.

Results and discussion

For every metal, 4200 IM were initially built with the ISIDA/MLR program. After discarding meaningless ($Q^2 < 0.5$) or highly correlated models, the remaining models were used for consensus calculations. Their number varies from one-fold in 5-CV to another one: 571–628 (Mg^{2+}), 236–278 (Ca^{2+}), 655–712 (Sr^{2+}) and 659–665 (Ba^{2+}) IM. Obtained consensus models demonstrate a reasonable predictive ability in log *K* predictions on the ensemble of test sets in 5-CV: *RMSE* of predicted log *K* values varies from 0.72 to 0.89 log *K* units and squared determination coefficient R^2 varies from 0.864 to 0.923 as a function of metal (Fig. 4). For 80 % of studied compounds, observed absolute prediction error is less than one log *K* unit; this is comparable with the

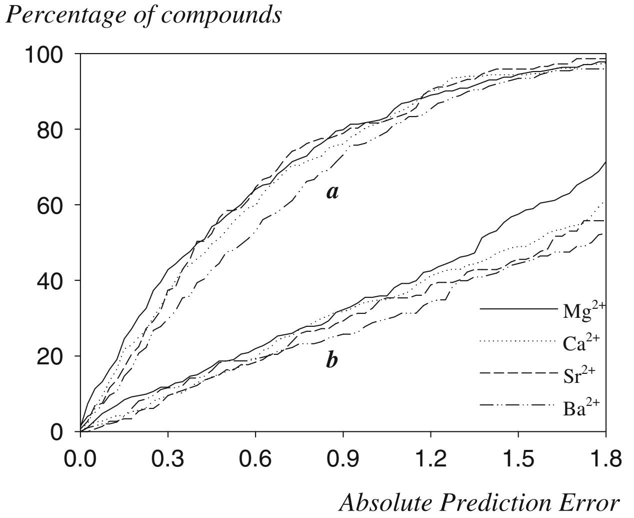


Fig. 5 The Regression Error Curves displaying percentage of compounds *vs* absolute prediction error |log K_{exp}–log K_{pred}| for the complexes of the Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} cations: corresponding ***a*** to predictions in 5-CV procedure, and ***b*** to "no model" calculations

Table 1 Experimental and predicted stability constant values log *K* of the 1:1 (M:L) complexation for the external test set

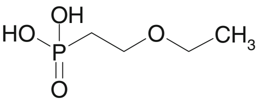
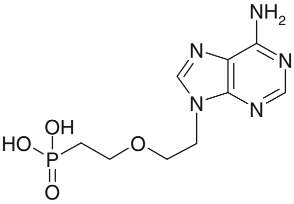
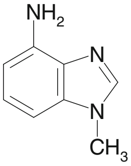
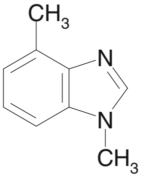
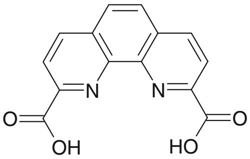
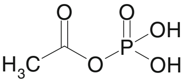
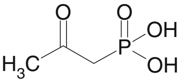
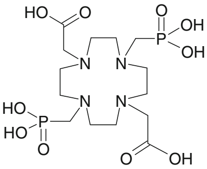
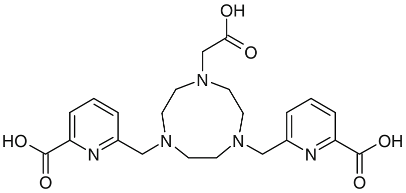
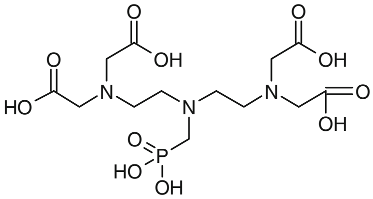
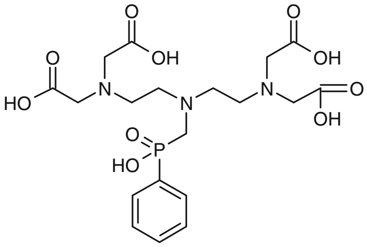
No.	Ligand	Cation	log *K*		
			Exp.	Pred.[a]	N_m[b]
1		Mg^{2+}	1.73[c]	1.37 (0.25)	140
		Ca^{2+}	1.51[c]	1.31 (0.20)	77
		Sr^{2+}	1.26[c]	1.11 (0.51)	214
		Ba^{2+}	1.24[c]	1.02 (0.45)	184
2		Mg^{2+}	1.74[c]	1.44 (0.33)	86
		Ca^{2+}	1.52[c]	1.08 (0.40)	48
		Sr^{2+}	1.27[c]	1.30 (0.60)	136
		Ba^{2+}	1.20[c]	1.09 (0.31)	85
3		Mg^{2+}	−0.02[d]	0.93 (0.52)	127
		Ca^{2+}	−0.07[d]	1.09 (0.70)	98
		Sr^{2+}	−0.11[d]	1.11 (0.77)	126
		Ba^{2+}	−0.20[d]	0.77 (0.65)	154
4		Mg^{2+}	−0.04[d]	0.85 (0.44)	108
		Ca^{2+}	−0.20[d]	0.42 (0.29)	44
		Sr^{2+}	−0.28[d]	0.94 (0.79)	140
		Ba^{2+}	−0.11[d]	0.51 (0.40)	111
5		Mg^{2+}	3.5[e]	2.6 (1.3)	285
		Ca^{2+}	7.3[e]	3.68 (0.62)	39
		Sr^{2+}	5.6[e]	2.80 (0.87)	116
		Ba^{2+}	5.4[e]	2.28 (0.52)	217
6		Mg^{2+}	1.51[f]	2.17 (0.21)	119
		Ca^{2+}	1.55[f]	2.14 (0.19)	48
		Sr^{2+}	1.47[f]	1.82 (0.15)	89
		Ba^{2+}	1.53[f]	1.68 (0.20)	97
7		Mg^{2+}	1.73[f]	2.03 (0.03)	38
		Ca^{2+}	1.55[f]	1.88 (0.18)	54
		Sr^{2+}	1.31[f]	1.45 (0.62)	159
		Ba^{2+}	1.23[f]	1.62 (0.53)	165
8		Ca^{2+}	15.1[g]	12.8 (1.1)	64
9		Ca^{2+}	8.18[h]	9.3 (1.6)	46

Table 1 continued

No.	Ligand	Cation	log *K*		
			Exp.	Pred.[a]	N_m^b
10		Ca^{2+}	10.7[i]	9.03 (0.35)	62
11		Ca^{2+}	9.38[i]	9.72 (0.93)	75
*R*²				*0.880*	
RMSE				*1.2*	

Experimental data are given at 298 K and ionic strength 0.1 M excepting: for the ligands 3 and 4 an ionic strength is 0.5 M, for the ligand 8 an ionic strength is 1.0 M

[a] Predicted stability constant values log K_{pred} are computed using the MLR consensus model, standard deviations are given in parentheses

[b] For the given ligand, the number of IM in CM using AD

[c] Ref. [36]

[d] Ref. [37]

[e] Ref. [38]

[f] Ref. [39]

[g] Ref. [40]

[h] Ref. [41]

[i] Ref. [42, 44]

variation in experimental log *K* values measured by different methods for the same metal–ligand complex [35] (Fig. 5). Several arsenic containing compounds (one for the Ca^{2+} and Sr^{2+} sets and three for the Ba^{2+} set) were found out of the applicability domain of all IM and, therefore, they were excluded from the test sets.

In spite of severe validation procedure, our models outperform those reported in previous studies [7–11]. Thus, failing to obtain reasonable models on the training data of 49 Ca^{2+} ligands, Raevski et al. [7] split the training data (49 compounds) into several small clusters on which some “local” models were developed. The validation of these models was performed on the test set containing only seven compounds which lead to poor correlation between the predicted and experimental data ($R^2 = 0.4$). In a number of studies [8, 10, 11], the training and test sets contained many common compounds, which could hardly be considered as valuable validation procedure. In a study by Cabaniss [9], the standard deviation of prediction (0.93 in log *K* units) was calculated for the small test set containing 15 compounds only.

The predictive ability of the developed models was assessed on the external test set of 32 complexes of Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} with 11 new organic ligands (Table 1) taken from references [36–42] which were not included in the initial modeling set. They contain the derivatives of phosphonic acid, benzimidazole and 1,10-phenanthroline, the aza-macrocycles with the lateral functional groups as well as the acyclic multidentate ligands with the carboxylic and phosphonic groups (Table 1). The statistical parameters ($R^2 = 0.880$ and $RMSE = 1.2$, Table 1 and Fig. 6) demonstrate a reasonable agreement between the experimental and predicted log *K* values; they are similar to those obtained in 5-CV on the modeling set.

The obtained results allow one to assess the ligand selectivity of metal M^a with respect to M^b measured by the logarithm of a ratio of their stability constants ($\log(K_M a_L/K_M b_L)$) [43]. For several selective ligands for Mg^{2+}, Ca^{2+}

Table 2 Predicted values of selectivity for several considered ligands

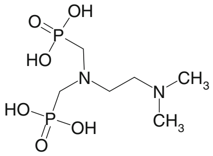
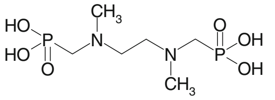
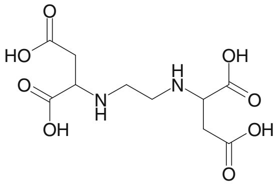
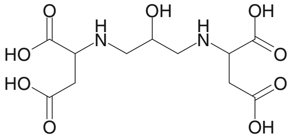
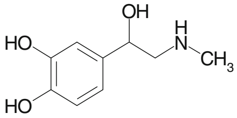
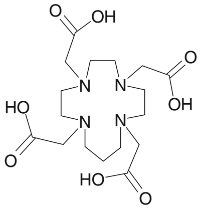
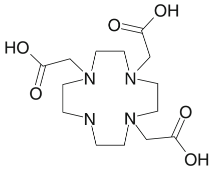
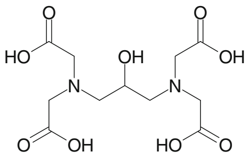
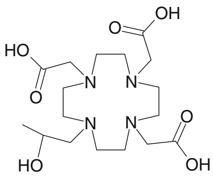
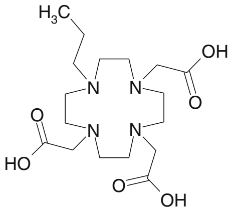
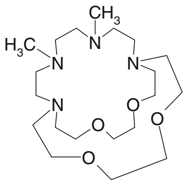
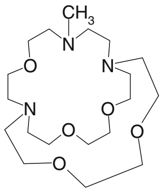
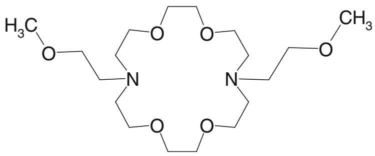
No.	Ligand	Selectivity, $\log(K_M a_L/K_M b_L)$			
		M^a	M^b	*Exp.*[a]	*Pred.*
1		Mg^{2+}	Ca^{2+}	0.5	1.3
		Mg^{2+}	Sr^{2+}	1.5	1.2
		Mg^{2+}	Ba^{2+}	2.0	1.7
2		Mg^{2+}	Ca^{2+}	1.4	1.2
		Mg^{2+}	Sr^{2+}	2.4	1.8
		Mg^{2+}	Ba^{2+}	2.6	1.9
3		Mg^{2+}	Ca^{2+}	1.2	1.1
		Mg^{2+}	Sr^{2+}	3.0	2.2
		Mg^{2+}	Ba^{2+}	3.7	2.9
4		Mg^{2+}	Ca^{2+}	0.7	0.9
		Mg^{2+}	Sr^{2+}	0.9	1.3
		Mg^{2+}	Ba^{2+}	1.7	1.9
5		Mg^{2+}	Ca^{2+}	0.6	2.0
		Mg^{2+}	Sr^{2+}	1.3	0.8
		Mg^{2+}	Ba^{2+}	2.0	1.6
6		Ca^{2+}	Mg^{2+}	8.2	2.5
		Ca^{2+}	Sr^{2+}	6.4	2.8
		Ca^{2+}	Ba^{2+}	7.8	4.2
7		Ca^{2+}	Mg^{2+}	3.6	3.3
		Ca^{2+}	Sr^{2+}	4.4	2.4
		Ca^{2+}	Ba^{2+}	6.0	4.0
8		Ca^{2+}	Mg^{2+}	1.8	2.2
		Ca^{2+}	Sr^{2+}	1.5	2.8
		Ca^{2+}	Ba^{2+}	1.9	2.9
9		Ca^{2+}	Mg^{2+}	4.4	3.8
		Ca^{2+}	Sr^{2+}	1.8	1.6
		Ca^{2+}	Ba^{2+}	4.0	4.3

Table 2 continued

No.	Ligand	Selectivity, log($K_M a_L/K_M b_L$)			
		M^a	M^b	*Exp.*[a]	*Pred.*
10		Ca^{2+}	Mg^{2+}	1.3	3.6
		Ca^{2+}	Sr^{2+}	1.9	1.6
		Ca^{2+}	Ba^{2+}	2.9	4.5
11		Ba^{2+}	Mg^{2+}	4.1	4.4
		Ba^{2+}	Ca^{2+}	2.4	2.4
		Ba^{2+}	Sr^{2+}	0.6	1.7
12		Ba^{2+}	Mg^{2+}	7.1	4.4
		Ba^{2+}	Ca^{2+}	4.4	3.0
		Ba^{2+}	Sr^{2+}	1.6	1.6
13		Ba^{2+}	Mg^{2+}	1.7	2.4
		Ba^{2+}	Ca^{2+}	1.3	0.8
		Ba^{2+}	Sr^{2+}	0.1	0.8

[a] The experimental selectivity is calculated using experimental values of the stability constants [13] at 298 K and ionic strength 0.1 M

and Ba^{2+} a qualitative agreement between the predicted and experimental selectivity values has been observed (Table 2 and Fig. 7). The selectivity values for 2-[1,7,10-tris(carboxymethyl)-1,4,7,10-tetraazacyclotridecan-4-yl]acetic acid were systematically underestimated which results in the relatively small R^2 ($R^2 = 0.41$) and high RMSE ($RMSE = 1.26$) value.

The obtained models for log K have been incorporated into COmplexation of METals (COMET) software [25] (Fig. 8) which is freely available at http://infochim.ustrasbg.fr/cgi-bin/predictor.cgi. It doesn't need any specific installation and can be used through any WEB browser. Any new compound can be submitted as an SDF (or MOL) file or prepared online. For each metal, COMET applies an ensemble of MLR models. A combination of bounding box and fragment control applicability domains is applied to each individual model in order to decide whether this can be included in consensus calculations. Predicted log K can be exported as an EXCEL file which containing the information for each individual model. The current version of COMET predicts stability constants of 1:1 complexes in water of alkaline-earth (Ca^{2+}, Sr^{2+}, Ba^{2+}, Mg^{2+}), lanthanides (Ce^{3+}, Pr^{3+}, Nd^{3+}, Sm^{3+}, Eu^{3+}, Gd^{3+}, Tb^{3+}, Dy^{3+}, Ho^{3+}, Er^{3+} , Tm^{3+}, Yb^{3+}, Lu^{3+}) and transition metals (Ag^{+}).

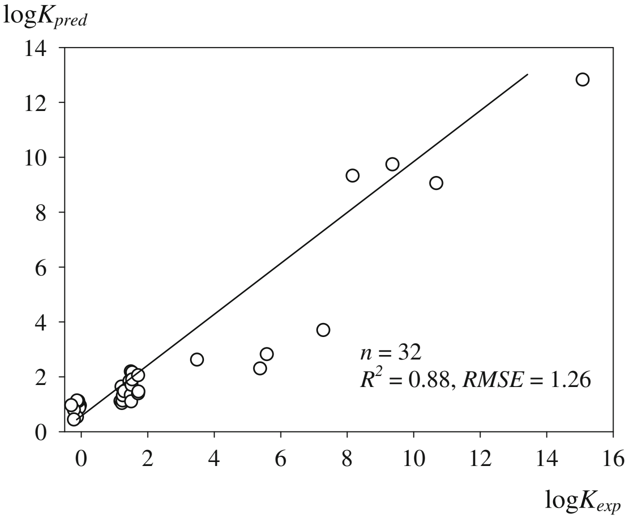


Fig. 6 Predicted versus experimental stability constant values (log K) for the 1:1 (M:L) complexation for the external test set

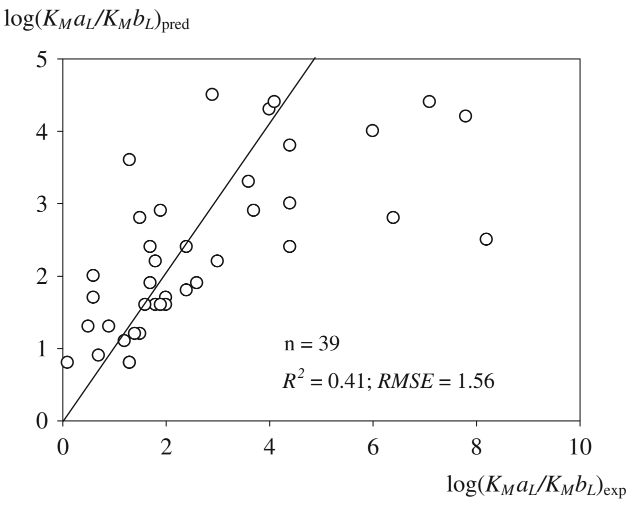
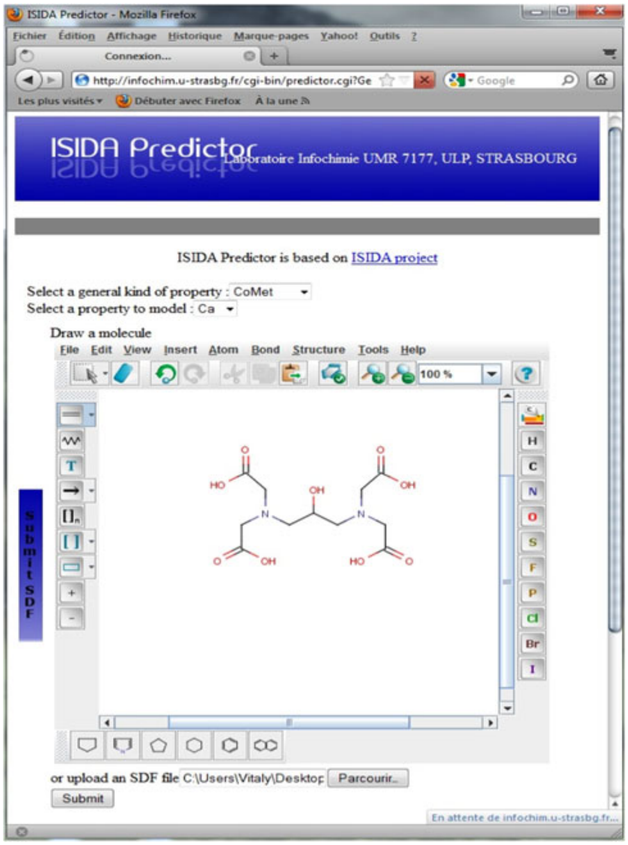


Fig. 7 Predicted versus experimental values of selectivity ($\log(K_M a_L/K_M b_L)$) of the 1:1 (M:L) complexation for the external test set


ISIDA Predictor - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Yahoo! Outils ?
Connexion...
http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi?Ge
Google
Les plus visités Débuter avec Firefox À la une
ISIDA Predictor
Laboratoire Infochimie UMR 7177, ULP, STRASBOURG
File Edit View Insert Atom Bond Structure Tools Help
100 %
HO
O
OH
O
OH
N
N
O
OH
HO
O
H
C
N
O
S
F
P
Cl
Br
I
Submit SDF
Parcourir..
Submit
En attente de infochim.u-strasbg.fr...

ISIDA Predictor is based on ISIDA project

Select a general kind of property : CoMet

Select a property to model : Ca

Draw a molecule

or upload an SDF file C:\Users\Vitaly\Desktop

Fig. 8 Screenshot of the WEB interface of COMET software

Conclusions

QSPR ensemble modeling of the stability constant log *K* of the complexes of Mg^{2+}, Ca^{2+}, Sr^{2+} and Ba^{2+} with diverse 273 (Mg^{2+}), 284 (Ca^{2+}), 147 (Sr^{2+}) and 198 (Ba^{2+}) organic ligands in water for the $M^{2+} + L = (M^{2+})L$ equilibrium at 298 K and an ionic strength 0.1 M has been performed. The IM were obtained using MLR methods and SMF descriptors. For each compound, predicted log *K* is calculated as an arithmetic average over values calculated by ensemble of IM. Several new algorithms of variables selection and models redundancy control have been applied.

Developed models outperform those reported in previous studies. Thus, root mean squared errors for 5-CV procedure are 0.75 (Mg^{2+}), 0.77 (Ca^{2+}), 0.72 (Sr^{2+}) and 0.87 (Ba^{2+}) which is comparable with the variation of experimental log *K* values measured for the given metal–ligand complex by different methods. Additional external validation of the models was performed on the data recently reported in the literature. Developed models have been integrated in the COMET predictor available for the end users via the internet (http://infochim.ustrasbg.fr/cgi-bin/predictor.cgi).

Acknowledgments We thank GDRE PARIS, the ARCUS project, CNRS France, the French Embassy in Russia and the Russian Foundation for Basic Research (project no. 09-03-93106) for the support.

Supporting Information Available Tables SM1–SM4 contain the names of the organic ligands (L) and the experimental stability constant values (log *K*) for the equilibrium $M^{2+} + L = (M^{2+})L$ (M = Mg, Ca, Sr, Ba) in water at 298 K and an ionic strength 0.1 M. Table SM5 contains the statistical parameters of the best IM.

References

1. Tretyakov, Y.D., Martynenko, L.I., Grigoryev, A.N., Tsivadze, A.Y.: Inorganic Chemistry. Chemistry of Elements. Book 1 (Rus.). Himia, Moscow. http://www.ozon.ru/context/detail/id/3768786/ (2001)
2. Bhattacharya, P.K.: Metal Ions in Biochemistry, p. 228. Alpha Scince International, Harrow (2005)
3. Sigel, A., Sigel, H. (eds.): Metal Ions in Biological Systems: Metal Ions and Their Complexes in Medication, vol. 41, p. 530. Marcel Dekker, Basel (2004)
4. Varnek, A., Solov'ev, V.: Quantitative structure-property relationships in solvent extraction and complexation of metals. In: Sengupta, A.K., Moyer, B.A. (eds.) Ion Exchange and Solvent Extraction, A Series of Advances, vol. 19, pp. 319–358. CRC Press, Taylor and Francis Group, Boca Raton (2009)
5. Tetko, I.V., Solov'ev, V.P., Antonov, A.V., Yao, X.J., Fan, B.T., Hoonakker, F., Fourches, D., Lachiche, N., Varnek, A.: Benchmarking of linear and non-linear approaches for quantitative structure-property relationship studies of metal complexation with organic ligands. J. Chem. Inf. Model. **46**, 808–819 (2006)
6. Shi, Z.G., McCullough, E.A.: A computer-simulation-statistical procedure for predicting complexation equilibrium-constants. J. Incl. Phenom. Mol. Recogn. **18**, 9–26 (1994)
7. Raevskii, O.A., Sapegin, A.M., Chistyakov, V.V., Solov'ev, V.P., Zefirov, N.S.: Development of a model for the relation between structure and complex forming ability. Koord. Khim. (Russ.) **16**, 1175–1184 (1990)

8. Toropov, A.A., Toropova, A.P., Nesterova, A.I., Nabiev, O.M.: QSPR modeling of complex stability by correlation weighing of the topological and chemical invariants of molecular graphs. Russ. J. Coord. Chem. **30**, 611–617 (2004)
9. Cabaniss, S.E.: Quantitative structure-property relationships for predicting metal binding by organic ligands. Environ. Sci. Technol. **42**, 5210–5216 (2008)
10. Toropov, A.A., Toropova, A.P.: QSPR modeling of stability of complexes of adenosine phosphate derivatives with metals absent from the complexes of the teaching access. Russ. J. Coord. Chem. **27**, 574–578 (2001)
11. Toropov, A.A., Toropova, A.P.: QSPR modeling of complex stability by optimization of correlation weights of the hydrogen bond index and the local graph invariants. Russ. J. Coord. Chem. **28**, 877–880 (2002)
12. Solov'ev, V.P., Kireeva, N.V., Tsivadze, A.Y., Varnek, A.A.: Structure-property modelling of complex formation of strontium with organic ligands in water. J. Struct. Chem. **47**, 298–311 (2006)
13. Pettit, G., Pettit, L.: IUPAC stability constants database. http://www.acadsoft.co.uk/ (1993–2005). Accessed 7 June 2012
14. Varnek, A., Fourches, D., Hoonakker, F., Solov'ev, V.P.: Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J. Comp. Aided Mol. Des. **19**, 693–703 (2005)
15. Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I.V., Marcou, G.: ISIDA—platform for virtual screening based on fragment and pharmacophoric descriptors. Curr. Comput. Aided Drug Des. **4**, 191–198 (2008)
16. Varnek, A. ISIDA (in silico design and data analysis) program. http://infochim.u-strasbg.fr/recherche/isida/index.php (2005–2012). Accessed 7 June 2012
17. Varnek, A., Solov'ev, V.P.: "In silico" design of potential anti-HIV actives using fragment descriptors. Comb. Chem. High Throughput Screen. **8**, 403–416 (2005)
18. Arnaud-Neu, F., Delgado, R., Chaves, S.: Critical evaluation of stability constants and thermodynamic functions of metal complexes of crown ethers (IUPAC technical report). Pure Appl. Chem. **75**, 71–102 (2003)
19. Solov'ev, V.P., Varnek, A.A., Wipff, G.: Modeling of ion complexation and extraction using substructural molecular fragments. J. Chem. Inf. Comput. Sci. **40**, 847–858 (2000)
20. Swamy, M.N.S., Thulasiraman, K.: Graphs, Networks, and Algorithms. Wiley, New York (1981)
21. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numer. Math. **14**, 403–420 (1970)
22. Varnek, A., Kireeva, N., Tetko, I.V., Baskin, I.I., Solov'ev, V.P.: Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? J. Chem. Inf. Model **47**, 1111–1122 (2007)
23. Solov'ev, V.P., Varnek, A.A.: Structure-property modeling of metal binders using molecular fragments. Russ. Chem. Bull. **53**, 1434–1445 (2004)
24. Muller, P.H., Neumann, P., Storm, R.: Tafeln der Mathematischen Statistik, p. 280. VEB Fachbuchverlag, Leipzip (1979)
25. Varnek, A., Fourches, D., Kireeva, N., Klimchuk, O., Marcou, G., Tsivadze, A., Solov'ev, V.: Computer-aided design of new metal binders. Radiochimica Acta **96**, 505–511 (2008)
26. Solov'ev, V., Oprisiu, I., Marcou, G., Varnek, A.: Quantitative structure-property relationship (QSPR) modeling of normal boiling point temperature and composition of binary Azeotropes. Ind. Eng. Chem. Res. **50**, 14162–14167 (2011)
27. Varnek, A.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. **42**, 812–829 (2002)
28. Katritzky, A.R., Kuanar, M., Fara, D.C., Karelson, M., Acree Jr, W.E., Solov'ev, V.P., Varnek, A.: QSAR modeling of blood:air and tissue: air partition coefficients using theoretical descriptors. Bioorg. Med. Chem. **13**, 6450–6463 (2005)
29. Katritzky, A.R., Dobchev, D.A., Fara, D.C., Hur, E., Tamm, K., Kurunczi, L., Karelson, M., Varnek, A., Solov'ev, V.P.: Skin permeation rate as a function of chemical structure. J. Med. Chem. **49**, 3305–3314 (2006)
30. Katritzky, A.R., Kuanar, M., Slavov, S., Dobchev, D.A., Fara, D.C., Karelson, M., Acree Jr, W.E., Solov'ev, V.P., Varnek, A.: Correlation of blood—brain penetration using structural descriptors. Bioorg. Med. Chem. **14**, 4888–4917 (2006)
31. Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C., Varnek, A.: Stochastic versus stepwise strategies for quantitative structure-activity relationship generation-how much effort may the mining for successful QSAR models take? J. Chem. Inf. Model. **47**, 927–939 (2007)
32. Kubinyi, H.: Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. Quant. Struct. Act. Relat. **13**, 393–401 (1994)
33. Zhokhova, N.I., Baskin, I.I., Palyulin, V.A., Zefirov, A.N., Zefirov, N.S.: Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies. Doklady Chem. **417**, 282–284 (2007)
34. Selwood, D.L., Livingstone, D.J., Comley, J.C., O'Dowd, A.B., Hudson, A.T., Jackson, P., Jandu, K.S., Rose, V.S., Stables, J.N.: Structure-activity relationships of ant filarial antimycin analogs: a multivariate pattern recognition study. J. Med. Chem. **33**, 136–142 (1990)
35. Christensen, J.J., Izatt, R.M.: Handbook of Metal Ligand Heats and Related Thermodynamic Quantities. Marcel Dekker Inc., New York (1983)
36. Fernandez-Botello, A., Griesser, R., Holy, A., Moreno, V., Sigel, H.: Acid-base and metal-ion-binding properties of 9-[2-(2-phosphonoethoxy)ethyl]adenine (PEEA), a relative of the antiviral nucleotide analogue 9-[2-(phosphonomethoxy)ethyl]adenine (PMEA). An exercise on the quantification of isomeric complex equilibria in solution. Inorg. Chem. **44**, 5104–5117 (2005)
37. Kapinos, L.E., Holy, A., Günter, J., Sigel, H.: Metal ion-binding properties of 1-methyl-4-aminobenzimidazole (=9-Methyl-1,3-dideazaadenine) and 1,4-dimethylbenzimidazole (=6,9-dimethyl-1,3-dideazapurine). Quantification of the steric effect of the 6-amino group on metal ion binding at the N7 site of the adenine residue. Inorg. Chem. **40**, 2500–2508 (2001)
38. Melton, D.L., VanDerveer, D.G., Hancock, R.D.: Complexes of greatly enhanced thermodynamic stability and metal ion size-based selectivity, formed by the highly preorganized non-macrocyclic ligand 1,10-phenanthroline-2,9-dicarboxylic acid. A thermodynamic and crystallographic study. Inorg. Chem. **45**, 9306–9314 (2006)
39. Sigel, H., DaCosta, C.P., Song, B., Carloni, P., Gregan, F.: Stability and structure of metal ion complexes formed in solution with acetyl phosphate and acetonylphosphonate: quantification of isomeric equilibria. J. Am. Chem. Soc. **121**, 6248–6257 (1999)
40. Kálmán, F.K., Baranyai, Z., Tóth, I., Bányai, I., Király, R., Brücher, E., Aime, S., Sun, X., Sherry, A.D., Kovács, Z.: Synthesis, potentiometric, kinetic, and NMR studies of 1,4,7,10-tetraazacyclododecane-1,7-bis(acetic acid)-4,10-bis(methylenephosphonic acid) (DO2A2P) and its complexes with Ca(II), Cu(II), Zn(II) and lanthanide(III) Ions. Inorg. Chem. **47**, 3851–3862 (2008)
41. Nonat, A., Gateau, C., Fries, P.H., Mazzanti, M.: Lanthanide complexes of a picolinate ligand derived from 1,4,7-triazacyclononane with potential application in magnetic resonance imaging and time-resolved luminescence imaging. Chem. Eur. J. **12**, 7133–7150 (2006)

42. Kotek, J., Kálmán, F.K., Hermann, P., Brücher, E., Binnemans, K., Lukeš, I.: Study of thermodynamic and kinetic stability of transition metal and lanthanide complexes of DTPA analogues with a phosphorus acid pendant arm. Eur. J. Inorg. Chem. **2006**, 1976–1986 (2006)
43. Solov'ev, V.P., Baulin, V.E., Strakhova, N.N., Kazachenko, V.P., Belsky, V.K., Varnek, A.A., Volkova, T.A., Wipff, G.: Complexation of phosphoryl-containing mono-, bi- and tri-podands with alkali cations in acetonitrile. Structure of the complexes and binding selectivity. J. Chem. Soc. Perkin Trans. **2**, 1489–1498 (1998)
44. Rodriguez, L., Lima, J.C., Parola, A.J., Pina, F., Meitz, R., Aucejo, R., Garcia-Espana, E., Llinares, J.M., Soriano, C., Alarcon, J.: Anion detection by fluorescent Zn(II) complexes of functionalized polyamine ligands. Inorg. Chem. **47**, 6173–6183 (2008)