

# Stability constants of complexes of $Zn^{2+}$ , $Cd^{2+}$ , and $Hg^{2+}$ with organic ligands: QSPR consensus modeling and design of new metal binders

Vitaly Solov'ev · Igor Sukhno · Vladimir Buzko ·  
Aleksy Polushin · Gilles Marcou · Aslan Tsivadze ·  
Alexandre Varnek

Received: 31 December 2010 / Accepted: 26 April 2011 / Published online: 21 May 2011  
© Springer Science+Business Media B.V. 2011

**Abstract** QSPR modeling of the stability constant  $\log K$  of the complexes of  $Zn^{2+}$ ,  $Cd^{2+}$  and  $Hg^{2+}$  with various 556 ( $Zn^{2+}$ ), 347 ( $Cd^{2+}$ ) and 76 ( $Hg^{2+}$ ) organic ligands in water for the  $M^{2+} + L = (M^{2+})L$  equilibrium at 298 K and an ionic strength 0.1 M was performed. Two machine-learning methods were used: Multiple Linear Regression Analysis (MLR) and Partial Robust M-regression Algorithm (PRM). The PRM method was realized for consensus modeling using substructural molecular fragments (SMF) as descriptors. Using different types of SMF, ensembles of individual predictive MLR and PRM models were prepared to build consensus models (CM). The root mean squared error of test set predictions of fivefold cross-validations is 1.8 and 1.9 ( $Zn^{2+}$ ), 1.9 and 2.2 ( $Cd^{2+}$ ), 2.7 and 2.8 ( $Hg^{2+}$ ) for the MLR and PRM approaches correspondingly. Experimental  $\log K$  values vary in the range of 0.8–21.9 ( $Zn^{2+}$ ), 0.9–23.3 ( $Cd^{2+}$ ) and 1.6–29.7 ( $Hg^{2+}$ ). Extra validation of the models has been performed on a set of ligands recently reported in the literature. The QSPR models are

sampled for the design of new binders of the  $Zn^{2+}$ ,  $Cd^{2+}$   $Hg^{2+}$  cations.

**Keywords** QSPR modeling of stability constants · Design of metal binders · Complexes of  $Zn^{2+}$ ,  $Cd^{2+}$ , and  $Hg^{2+}$  with organic ligands in water

## Introduction

The interest in the coordination chemistry of zinc, cadmium and mercury is related not only to the widespread industrial uses of their compounds, but also to their toxicity and health effects [1]. Zinc is one of the most abundant divalent metals in living organisms. It is an essential cofactor of many metabolic enzymes and transcription factors. Some biological metal binding sites are vulnerable to attacks by nonbiogenic “alien” cations such as Cd and Hg [2]. Cadmium compounds are regarded as carcinogenic to humans [1]. The removal of toxic heavy metal contaminants is an environmental issue of great importance. Therefore, great attention is addressed to the design and synthesis of ligands able to bind specific metals [3]. For a construction of ligands with specific behavior (e.g., changing of luminescence at the complexation [4–7]), it is important to know ligand-binding properties to metal cations. The design of new metal binders represents an area of current interest in supramolecular chemistry [6].

Chemoinformatics approaches open opportunities for computer-aided design of new metal binders with desire stability of their complexes and metal selectivity [8, 9]. The number of publications on QSPR modeling of stability constants of metal–ligand complexes in solutions is restricted by two tens of papers [8]. The works on QSPR modeling of stability constants of cation–ligand complexes

**Electronic supplementary material** The online version of this article (doi:10.1007/s10847-011-9978-6) contains supplementary material, which is available to authorized users.

V. Solov'ev (✉) · A. Tsivadze  
Institute of Physical Chemistry and Electrochemistry, Russian  
Academy of Sciences, Leninskiy prospect, 31a, Moscow,  
Russian Federation 119991  
e-mail: solovev-vp@mail.ru

I. Sukhno · V. Buzko · A. Polushin  
Kuban State University, 149 Stavropolskaya st., Krasnodar,  
Russian Federation

G. Marcou · A. Varnek  
Laboratoire d'Informatique, UMR 7177 CNRS, Université de  
Strasbourg, 4, rue B.Pascal, 67000 Strasbourg, France

differ in a variety of classes and the number of analyzed ligands, as well as a discrepancy between approaches to test predictive models. The modeling of restricted classes of ligands was carried out for the 1:1 (M:L) complexation of  $\text{Na}^+$  [10–14],  $\text{K}^+$  [11–15],  $\text{Cs}^+$  [11, 13, 14],  $\text{Ca}^{2+}$  [10],  $\text{Zn}^{2+}$  [10],  $\text{Gd}^{3+}$  [16],  $\text{Cu}^{2+}$  and  $\text{Ni}^{2+}$  [17] with crown ethers [10–15], aza-crowns [16], phosphoryl-containing podans [12, 14], cryptands [10] and spherands [10], cyclic and acyclic aminocarboxylates [16], fructose-amino acids [17] in water [16, 17], methanol [11–15], different pure and mixed solvents [10],  $\text{CDCl}_3$  [10],  $\text{THF}:\text{CHCl}_3$  (4:1 vol.) [12, 14]. More wide variety of the ligands presents the studies of the 1:1 complexation of  $\text{Ca}^{2+}$  [18–20],  $\text{Mg}^{2+}$ ,  $\text{Mn}^{2+}$ ,  $\text{Fe}^{2+}$ ,  $\text{Co}^{2+}$  [19, 21, 22],  $\text{Ni}^{2+}$ ,  $\text{Cu}^{2+}$ ,  $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$  [19–22],  $\text{Al}^{3+}$ ,  $\text{Pb}^{2+}$  [20], lanthanide cations [23] with amino acids, adenosine and its phosphate derivatives, heterocyclic and aromatic rings [19, 21, 22], aminocarboxylates and aza-crown ethers [23] and the ligands containing carboxylate, phenol, amine, ether, and alcohol functional groups [20] in water. Predictive QSPR models have been developed for the stability constants of the 1:1 [9, 24, 25] and 1:2 [25] complexes of  $\text{Sr}^{2+}$  [24],  $\text{Ag}^+$  [25],  $\text{Eu}^{3+}$  [25] and 13 lanthanide cations [9] with quite diverse organic molecules [9, 24, 25] in water at 298 K and an ionic strength 0.1 M. The acyclic and macrocyclic, acidic, basic and neutral organic molecules were studied. As a rule, those organic molecules bear several electron-donor groups. Some of carboxy, carbonyl, hydroxy, phosphono, phosphinyl, amino, amido, sulfo, ether, mercapto, thioether, nitro, imidazolyl or pyridyl groups are included in different combinations [9, 24, 25]. The number of the ligands per metal varies from 130 ( $\text{Sr}^{2+}$  [24]) to 308 ( $\text{Gd}^{3+}$  [9]) and exceeds the data sets of aforementioned separate classes of the ligands. Predictive power of models is estimated by calculations for test set(s) independent of training data [8]. Often one [10–12, 14, 16, 19, 20] or 3–4 [13, 24] test sets are utilized. In order to avoid any uncertainties related to a selection of a particular test set, a more severe  $n$ -fold cross-validation technique is recommended [8] and used [9, 25]. For restricted classes of the ligands, standard deviation and root mean squared error in the stability constant  $\log K$  values for validated data sets are similar to experimental errors and vary from 0.2 to 0.3 [11–14] to 0.6–0.7 [13, 14, 19]. For diverse organic ligands, they vary from 0.7 to 1.40 [10, 20, 24, 25] to 2.3–2.4 [9, 25]. A combination of machine-learning methods and consensus modeling [8] could be used to increase the reliability of predictions for diverse ligands. From practical point of view, it is important to have tools for an application of developed QSPR models for the ligand design and stability constant estimations. As a rule using available experimental data on the stability constant, QSPR modeling enable to predict the stability constant values of the

complexes of new ligands. In contrast to quantum chemistry and force-field simulations, chemoinformatics approaches are much less pay attention to complex geometry and the nature of the donor atoms and chemical groups that interact with the metal cation. However, certain descriptors of structure–property models such as topological substructure fragments may shed light on binding sites of the ligands.

In this paper, we report the QSPR consensus modeling of the stability constant  $\log K$  of the 1:1 (M:L) complexes of metal cations  $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$  and  $\text{Hg}^{2+}$  with diverse sets of organic molecules in aqueous solution at 298 K and an ionic strength 0.1 M via Multiple Linear Regression (MLR) approach of the ISIDA program package and newly realized Partial Robust M-regression (PRM) method. Using different types of substructural molecular fragments of molecular graphs of the ligands as descriptors, we build hundreds of individual predictive MLR and PRM models to prepare consensus models (CM). Predictive ability of the CM models is analyzed using the fivefold external cross-validation procedure and an extra test set of ligands recently reported in the literature. The QSPR models are sampled for the design of new metal binders with desired complexation properties using the 2D sketcher EdChemS interactively and the predictor COMET via Internet.

## Methods

### Descriptors

Substructural molecular fragments (SMF) of the ISIDA program [12, 26] as subgraphs of molecular graph were used as descriptors which are independent variables in the QSPR models. A fragment occurrence is a descriptor value. The descriptors were derived solely from 2D chemical structures. Molecules were represented with implicit hydrogen atoms. Two subclasses of the SMF descriptors were utilized: shortest topological paths with explicit presentation of atoms and bonds (*i*), and terminal groups as shortest path sequences defined by length and explicit indication of beginning atom and bond and ending bond and atom (*ii*). For searching the shortest paths, the Floyd algorithm [27] was used. We distinguished single, double, triple and aromatic bonds. Moreover, single, double and triple bonds were considered different in acyclic and cyclic non-aromatic parts of molecules. For each subclass of the sequences, the minimal ( $n_{\min} \geq 2$ ) and maximal ( $n_{\max} \leq 15$ ) numbers of constituent atoms are defined. The values of  $n_{\min}$  and  $n_{\max}$  varied from 2 to 15 for MLR, and from 2 to 4 ( $n_{\min}$ ) and from 6 to 15 ( $n_{\max}$ ) for PRM. The notations IAB ( $n_{\min} - n_{\max}$ ) and IAB( $n_{\min} - n_{\max}$ )*t* represent types of two subclasses *i* and *ii* of the SMF

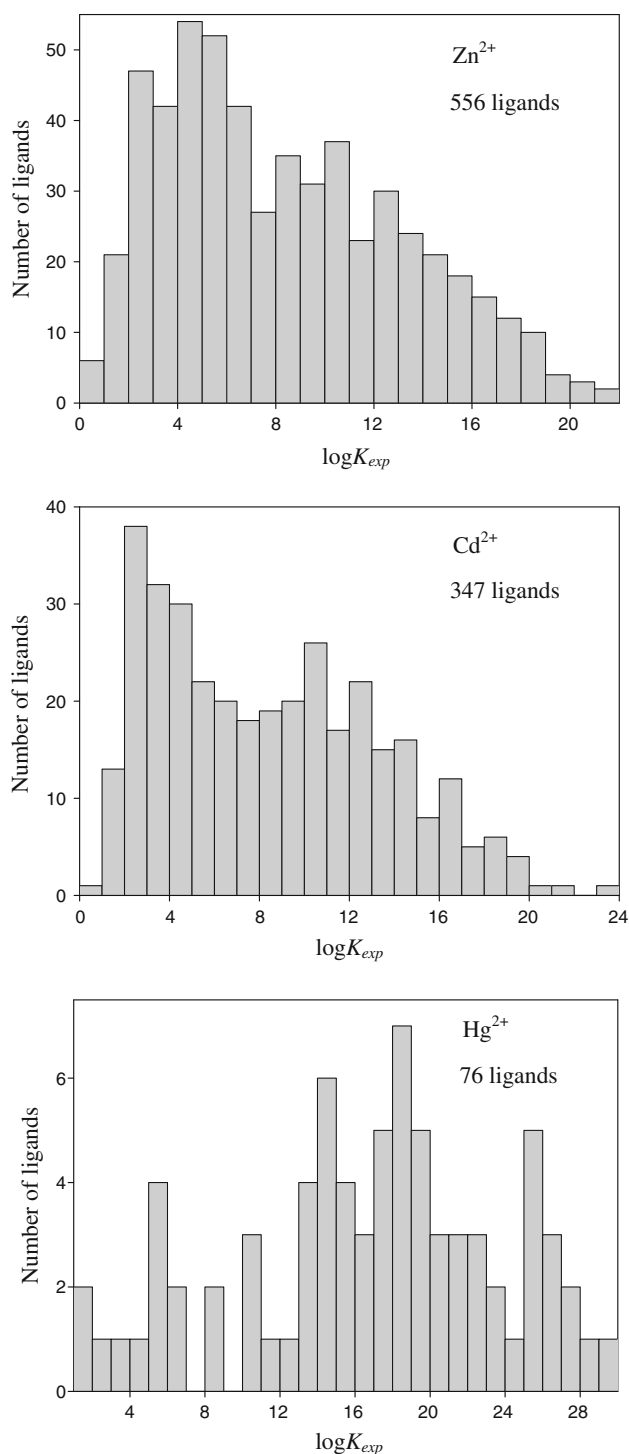
descriptors, which include all intermediate shortest paths with  $n$  atoms, for which  $n_{\min} \leq n \leq n_{\max}$ . Varying the values of  $n_{\min}$  and  $n_{\max}$ , 210 types of the sequences of two subclasses were prepared for the modeling. One type of the SMF descriptors is used to build one (PRM) and two (MLR) individual models. SMF of both subclasses were used in the modeling by MLR, and shortest path sequences  $i$  were merely applied in the PRM modeling.

#### Data sets

Experimental stability constant values ( $\log K$ ) for the 1:1 (M:L) complexes of  $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$  and  $\text{Hg}^{2+}$  cations with divers organic ligands in water were critically selected from IUPAC Stability Constants Database (SC DB) [28] (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 corrected 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 [24, 26, 29] into Structure Data Files (SDF) readable by the MLR and PRM programs of the ISIDA package [30, 31]. If several values of the stability constant  $\log K$  were available for a particular ligand, for selections we followed the recommendations of IUPAC [32]; in some cases the most recent data or the data consistent with respect to different experimental methods were chosen. At the pretreatment stage of the modeling, some specific ligands were excluded if for the SMF type with  $n_{\min} = 2$  and  $n_{\max} = 6$ , they bring fragments occurred less than in 3 ligands for given metal cation. Finally, 556 ( $\text{Zn}^{2+}$ ), 347 ( $\text{Cd}^{2+}$ ) and 76 ( $\text{Hg}^{2+}$ ) 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 0.8–21.9 ( $\text{Zn}^{2+}$ ), 0.9–23.3 ( $\text{Cd}^{2+}$ ) and 1.6–29.7 ( $\text{Hg}^{2+}$ ).

The names of the ligands and the stability constant values are presented as supporting information in Tables SM1–SM3. Large majority of the organic ligands can be classified on acyclic and macrocyclic, acidic, basic and neutral compounds. As a rule, the organic ligand bears several electron-donor groups. Some of the carboxy, carbonyl, hydroxy, phosphono, phosphinyl, amino, amido, sulfo, ether, mercapto, thioether, nitro, imidazolyl or pyridyl groups can be included in different combinations. The sets of the ligands include amino and hydroxy derivatives of carboxylic acids; different aminoacids 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



**Fig. 1** Distribution of experimental values of the stability constant ( $\log K$ ) for the 1:1 (M:L) complexes of organic ligands with  $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$  and  $\text{Hg}^{2+}$  in water at temperature 298 K and an ionic strength 0.1 M

dipodands of ternary amines; amino derivatives of phenols; crown-ethers, thia-, and aza-crown-ethers with neutral and acidic lariat groups, cryptands, etc. (see supporting information: Tables SM1–SM3).

## Machine learning methods

To estimate an ability of a model to predict reliably modeling property, the fivefold cross validation was used [25, 33]. In this procedure, an entire dataset is divided in 5 non-overlapping pairs of training and test sets. Each training set covers 4/5th of the dataset while the related test set covers the remaining 1/5th. Predictions are prepared for all molecules of the initial dataset, since each of them belongs to one of the test sets. As criterions of robustness of models, squared coefficient of determination ( $R_0^2$ ), root mean squared error (RMSE) and mean absolute error (MAE) for training ( $Y = Y_{\text{calc}}$ ) and test ( $Y = Y_{\text{pred}}$ ) sets are used

$$R_0^2 = 1 - \frac{\sum(Y_{\text{exp}} - Y)^2}{\sum(Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2},$$

$$\text{RMSE} = \left( \frac{\sum(Y_{\text{exp}} - Y)^2}{n} \right)^{1/2} \quad \text{and}$$

$$\text{MAE} = \frac{\sum|Y_{\text{exp}} - Y|}{n},$$

where  $Y_{\text{calc}}$ ,  $Y_{\text{pred}}$  and  $Y_{\text{exp}}$  are calculated, predicted and experimental values of the stability constant correspondingly and  $Y = \log K$ .

For the predictions of the properties, we used consensus models (CM). CM combines the predictions issued from many individual models originated from different types of the SMF descriptors. CM allows one to smooth inaccuracies of individual models and ensures more reliable predictions [14, 24, 31, 33]. Thus for each compound from the test set, the program computes the property as an arithmetic mean of values obtained with a collection of selected on training stage individual models excluding those leading to outlying values according to Tompson's rule and a method of ranked series [34], and taking into account an applicability domain (AD) of each model. We used the collections of best individual models for which leave one out (LOO) cross-validation correlation coefficient  $Q^2 > Q_{\text{lim}}^2$  (MLR) or determination coefficient  $R_0^2 > R_{0, \text{lim}}^2$  (PRM) for the models. Here  $Q_{\text{lim}}^2$  and  $R_{0, \text{lim}}^2$  are user defined thresholds.

When applying an individual model for CM, the program checks its AD [9, 29] which measures a similarity between a test compound and the compounds from the training set. If the test compound is identified as being outside AD, the prediction by given model is not included for a preparation of CM. We took into account two AD approaches conjointly: bounding box considering as AD a multi-dimension descriptor space confined by minimal and maximal values of counts of SMF descriptors involved in individual model, and fragment control rejecting a prediction for the test compound containing unknown SMF fragment, which is non-existent in the initial SMF pool for given model preparation. Prediction calculations were made both with AD by the PRM and MLR methods and without AD by the PRM method also.

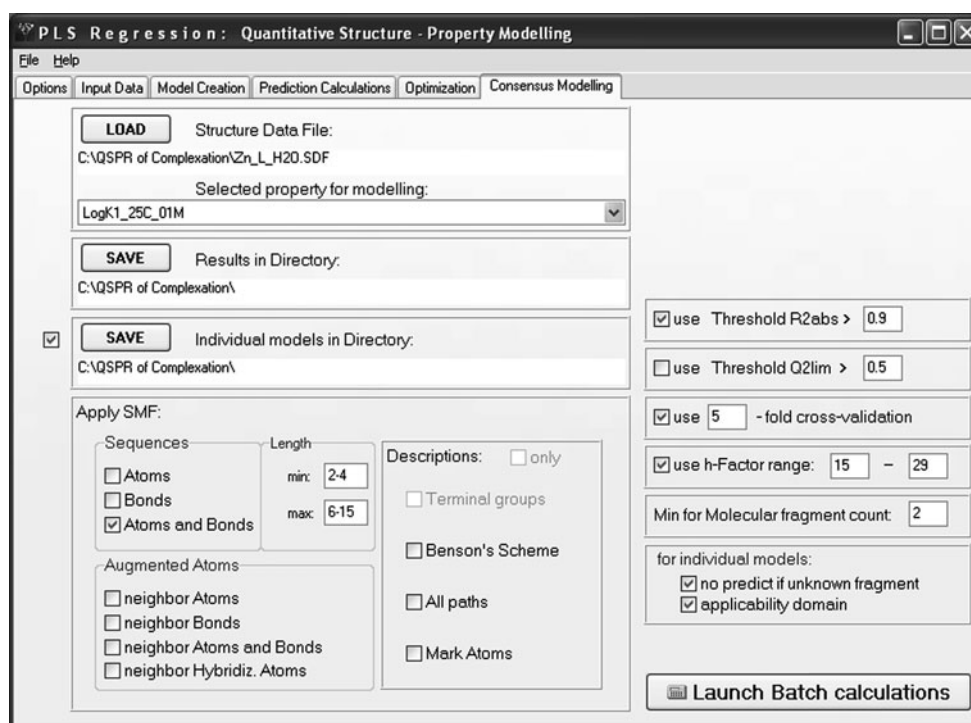
The ISIDA (In Silico design and Data Analysis) program package [29–31] (<http://infochim.u-strasbg.fr/recherche/Download/Download.php>) has been used for structure–property modeling, which was performed using two machine learning methods: Multiple Linear Regression Analysis (MLR) [12] and Partial Robust M-regression Algorithm (PRM).

### Multiple Linear Regression Analysis (MLR)

The SMF descriptors are independent variables used to build multi-linear correlation equations  $Y = a_0 + \sum a_i X_i$ , where  $Y$  is modeling property,  $X_i$  is the count of the  $i$ th SMF,  $a_i$  is its contribution, and  $a_0$  is the descriptor independent term. For each SMF type, two types of the equations were prepared: including the  $a_0$  term or without it. Using the training set, the coefficients  $a_0$  and  $a_i$  have been fitted by the Singular Value Decomposition method [35]. The robust models were selected at the training stage according to LOO cross-validation correlation coefficient  $Q^2 > Q_{\text{lim}}^2$ , where  $Q_{\text{lim}}^2$  is a user defined threshold. In this work, we used  $Q_{\text{lim}}^2 = 0.5$  for the studied metal cations.

For the MLR method, forward and backward stepwise techniques [14, 33, 36] have been utilized for selections of pertinent variables  $X$  from initial pools of the SMF descriptors. In the beginning, original Variable Selection Suite (VSS) program eliminates variables  $X_i$  which have small correlation coefficient with the property ( $|R_{Y,i}| < R_{Y,i}^0$ ) or those highly correlated with other variables  $X_j$  ( $|R_{i,j}| > R_{i,j}^0$ ), which were already selected for the model [33, 36]. In this work, the boundary values  $R_{Y,i}^0 = 0.001$  and  $R_{i,j}^0 = 0.99$  were used. Concatenated fragments always occurring in the same combination in each compound of the training set are interpreted as one extended fragment. Infrequent fragments (i.e., found in less than  $m$  molecules, here  $m < 2$ ) were excluded. Then, forward stepwise iterative procedure on each step selects two variables  $X_i$  and  $X_j$  giving maximal correlation coefficient ( $R_{Y,ij} = (R_{Y,i}^2 + R_{Y,j}^2 - 2R_{Y,i}R_{Y,j}R_{ij})/(1 - R_{2ij}^2)$ ) with a residual of the property  $Y^{(p)}$ . At each step  $p$ , fitted residual is  $Y^{(p)} = Y^{(p-1)} - Y_{\text{calc}}$ , where  $Y^{(0)} = Y_{\text{exp}}$  for the first step ( $p = 1$ ) and  $Y_{\text{calc}} = c_0 + c_i X_i + c_j X_j$  is calculated function for corresponding  $Y^{(p-1)}$  by the two-variable model with selected variables  $X_i$  and  $X_j$ . This loop is repeated until the number  $k$  of variables reaches a user-defined value [33, 36]; in this work,  $k = 0.6 N$ , where  $N$  is the number of data points (here ligands) in the training set. The final backward stepwise variable selection is based on using of the Student's  $t$ -criterion [14]. The program eliminates the variables with low  $t_i = a_i/\Delta a_i$  values, where  $\Delta a_i$  is standard deviation for the coefficient  $a_i$  at the  $i$ th variable in the MLR model. First, the program selects the variable with the minimal  $t_{i,\text{min}} < t_0$ , then it builds a new model

**Fig. 2** Graphical interface for the partial robust M-regression program showing consensus modeling



excluding that variable. This procedure is repeated until  $t \geq t_0$  for all remaining variables. Here the tabulated value  $t_0 = 1.96$  of Student's criterion was employed.

#### Partial Robust M-regression (PRM)

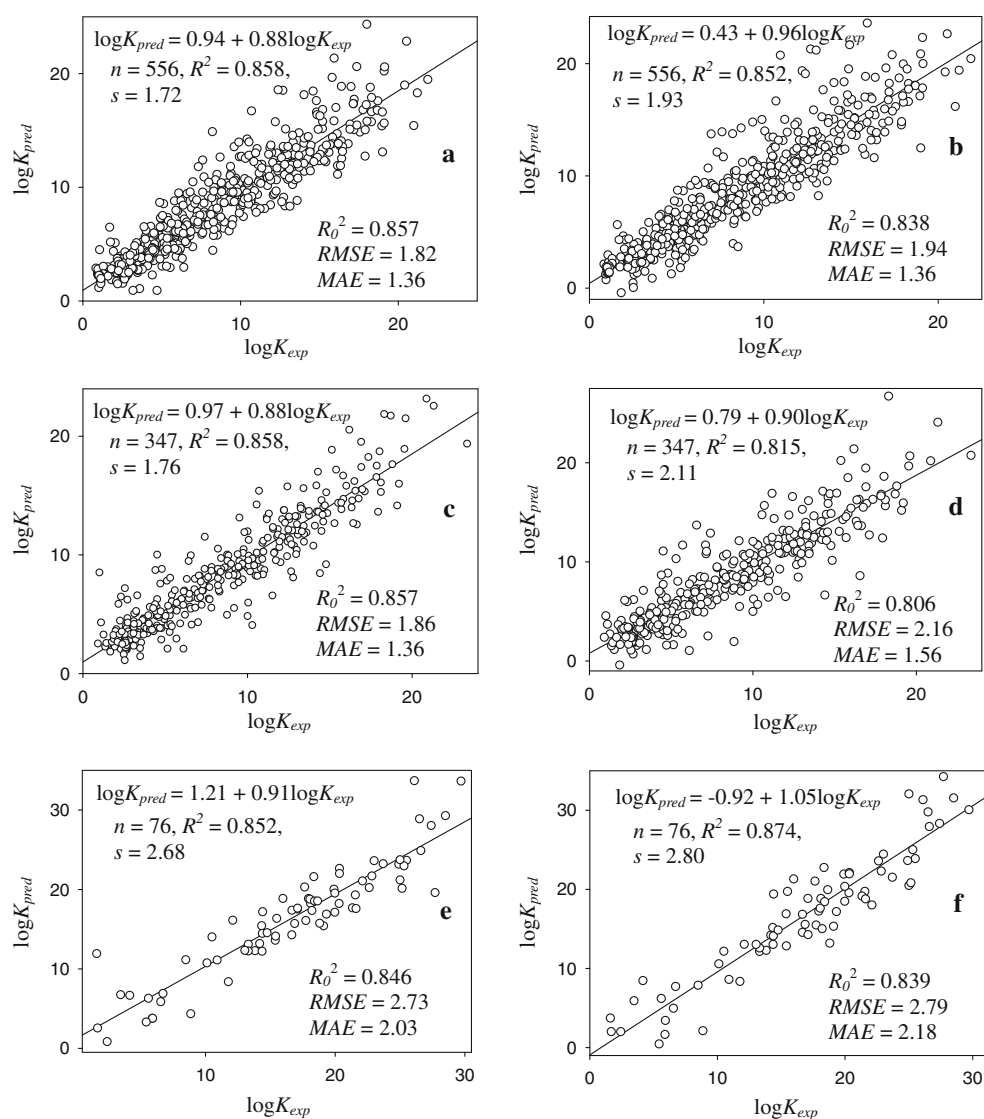
The implemented in the ISIDA software PRM method was realized for consensus modeling to combine predictions issued from many individual models using SMF as descriptors. PRM is a modification of the widely used Partial Least Squares (PLS) statistical tool. PRM outperforms some another methods for robust PLS regression in terms of statistical precision, computational speed and robustness with respect to outliers [37]. It is important for the SMF descriptor application in the QSPR modeling since the method empowers to solve a problem of descriptor multicollinearity and to keep descriptor pool completeness in a model when the number of variables exceeds substantially the number of observations. The underlying idea is that PLS summarizes the often high-dimensional independent variables (the SMF descriptors in our case) into a smaller set of uncorrelated, so-called latent variables ( $h$ ), which have a maximal covariance to the modeling property. PRM gives a protection against both vertical outliers and leverage points [37]. Vertical outliers are outliers in the error terms. The leverage points are observations in the descriptor space far away from the big majority of the data. In our case, the number of descriptors  $p$  (independent variables) is essentially large compared to

the number of the observations  $n$  (data points). According to [37], computations are sped up by carrying out preliminary singular value decomposition (SVD) on the data matrix  $p \times n$  then applying the PRM iteration scheme on the reduced data matrix having size  $n \times n$ .

The program realized partial robust M-regression algorithm has been designed using Object Pascal and DELPHI programming platform for WINDOWS (Fig. 2). For molecular structures and property input, we use the SDF format. User can select a collection of the SMF descriptor types and the range of the numbers of the latent variables  $h$  to build hundreds of individual models for a preparation of the consensus model. One can select applicability domain method(s). We carried out the calculations with involving of the AD methods and without their use. The robust models can be selected according to thresholds of LOO cross-validation correlation coefficient  $Q^2 > Q_{lim}^2$  or/and determination coefficient  $R_0^2 > R_{0, lim}^2$ . Here  $Q_{lim}^2$  and  $R_{0, lim}^2$  are user defined thresholds. In this work, we used  $R_0^2 > R_{0, lim}^2 = 0.9$ .

#### Results and Discussion

QSPR consensus modeling of the stability constant  $\log K$  was performed for the complexation  $M^{2+} + L = (M^{2+})L$  of  $Zn^{2+}$ ,  $Cd^{2+}$  and  $Hg^{2+}$  with 556 ( $Zn^{2+}$ ), 347 ( $Cd^{2+}$ ) and 76 ( $Hg^{2+}$ ) structurally diverse organic ligands in water at 298 K and an ionic strength 0.1 M. The MLR



**Fig. 3** Predicted versus experimental values of the stability constant ( $\log K$ ) for the 1:1 (M:L) complexation of organic ligands with  $\text{Zn}^{2+}$  (a, b),  $\text{Cd}^{2+}$  (c, d) and  $\text{Hg}^{2+}$  (e, f) in water at temperature 298 K and ionic strength 0.1 M. Results were obtained using the ISIDA/MLR (a,

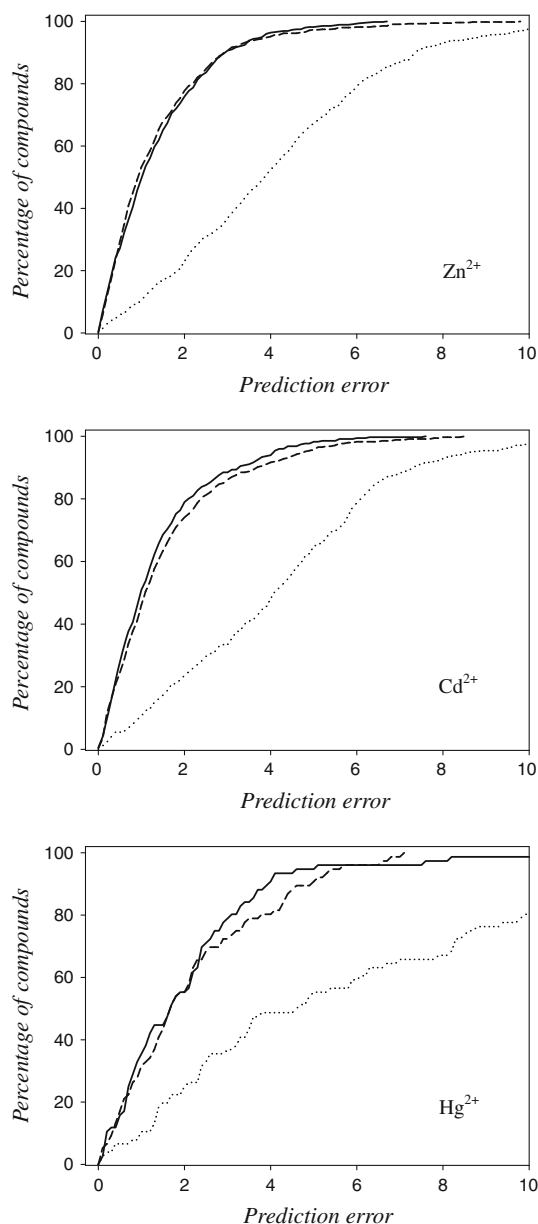
c, e) and ISIDA/PRM (b, d, f) consensus models. The predicted data represent a combination of all five external test sets of the fivefold cross-validation procedure

and PRM machine learning methods were used. Altogether for each metal cation, 2,100 (MLR) and 2,250 (PRM) individual models have been built. The number of latent variables in the PRM models was varied from 15 to 29. For the preparation of the consensus models, collections of robust individual models were selected, for which LOO cross-validation correlation coefficient  $Q^2 > 0.5$  (MLR) or determination coefficient  $R_0^2 > 0.9$  (PRM). The collections contain 210–240 ( $\text{Zn}^{2+}$ ), 260–310 ( $\text{Cd}^{2+}$ ) and 240–260 ( $\text{Hg}^{2+}$ ) MLR models and 300–340 ( $\text{Zn}^{2+}$ ), 320–380 ( $\text{Cd}^{2+}$ ), 400–450 ( $\text{Hg}^{2+}$ ) PRM models for the training sets of the fivefold cross-validations.

The determination coefficient  $R_0^2$ , root mean squared error RMSE and mean absolute error MAE obtained

between experimental and predicted stability constant values  $\log K$  have been considered as criteria of the robustness of the MLR and PRM consensus models. The predicted data represent the combinations of the external test sets of the fivefold cross-validation procedure. RMSE of the test set predictions is 1.8 and 1.9 ( $\text{Zn}^{2+}$ ), 1.9 and 2.2 ( $\text{Cd}^{2+}$ ), 2.7 and 2.8 ( $\text{Hg}^{2+}$ ), MAE is 1.4 and 1.4 ( $\text{Zn}^{2+}$ ), 1.4 and 1.6 ( $\text{Cd}^{2+}$ ), 2.0 and 2.2 ( $\text{Hg}^{2+}$ )  $\log K$  units for the MLR and PRM approaches correspondingly (Fig. 3). The squared determination coefficient  $R_0^2$  of the predictions varies from 0.81 to 0.86 for both methods and three metal cations. The consensus models demonstrate reasonable predictive ability for  $\text{Zn}^{2+}$  and  $\text{Cd}^{2+}$ . The modeling on the relative small data set for  $\text{Hg}^{2+}$  led to larger RMSE and

MAE values compared to the  $\text{Zn}^{2+}$  and  $\text{Cd}^{2+}$  data sets. The ISIDA/MLR technique demonstrates slightly better performance over PRM. The prediction calculations (Fig. 3) were carried out using the applicability domain (AD) approaches for the MLR method and without AD for PRM. Using of the AD option for the PRM validation calculations demonstrates smallest RMSE of the test set predictions: 1.7 ( $\text{Zn}^{2+}$ ), 1.9 ( $\text{Cd}^{2+}$ ) and 2.3 ( $\text{Hg}^{2+}$ ), however in this case, predicted log  $K$  values are rejected by the AD methods for 23% ( $\text{Zn}^{2+}$ ), 16% ( $\text{Cd}^{2+}$ ) and 26% ( $\text{Hg}^{2+}$ ) of



**Fig. 4** Percentage of compounds versus absolute prediction error  $|\log K_{\text{exp}} - \log K_{\text{pred}}|$ ; continuous line corresponds to MLR CM; broken curve corresponds to PRM CM; dotted line corresponds to “no model”: arithmetic mean of experimental constant log  $K_{\text{exp}}$  of all ligands is as the predicted value for any ligand

the studied ligands. The Regression Error Curves demonstrate (Fig. 4) that for  $\text{Zn}^{2+}$  and  $\text{Cd}^{2+}$  absolute prediction error is below 1.0 for 50% of the ligands. This error  $|\log K_{\text{exp}} - \log K_{\text{pred}}| \approx 1$  is appropriate to the discrepancy in experimental log  $K$  values measured by different methods [28]. For diverse organic ligands and severe fivefold cross-validation technique, RMSE of the predictions is similar ( $\text{Hg}^{2+}$ ) or lower ( $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$ ) than corresponding RMSE for the complexation of lanthanide cations [9, 25].

**Table 1** The SMF types and statistical parameters of best MLR models according to five training sets of the fivefold cross validation procedure

No.	SMF type <sup>a</sup>	$n$	$m$	$s$	$Q^2$
$\text{Zn}^{2+}$					
1	IAB(2–10)	444–445	97–121	0.9–1.1	0.901–0.932
2	IAB(3–10)	444–445	97–119	0.9–1.1	0.912–0.930
3	IAB(2–15)t	444–445	74–98	1.0–1.3	0.901–0.939
$\text{Cd}^{2+}$					
4	IAB(3–15)	277–278	81–97	0.6–0.9	0.896–0.953
5	IAB(3–13)	277–278	67–97	0.7–1.1	0.899–0.956
6	IAB(2–15)	277–278	81–97	0.6–0.9	0.901–0.952
7	IAB(2–12)	277–278	56–93	0.7–1.3	0.900–0.940
$\text{Hg}^{2+}$					
8	IAB(3–10)	61	21–27	0.9–1.1	0.966–0.969
9	IAB(4–13)t	59–60	21–25	0.9–1.1	0.950–0.961

Statistical parameters of the MLR models: the number data point in training set ( $n$ ), the number of SMF variables ( $m$ ), standard deviation ( $s$ ), squared LOO cross-validation correlation coefficient ( $Q^2$ )

<sup>a</sup> SMF type: see the notation in “Methods” section: descriptors

**Table 2** The SMF types and statistical parameters of best PRM models according to five training sets of the fivefold cross validation procedure

no.	SMF type	$n$	$M$	$h$	$s$	$R_0^2$
$\text{Zn}^{2+}$						
1	IAB(2–15)	444–445	3,121–3,499	26–29	0.7–1.0	0.961–0.976
2	IAB(3–14)	445	2,966–3,259	26–29	0.8–0.9	0.965–0.973
3	IAB(3–15)	444–445	3,101–3,478	26–29	0.8–0.9	0.965–0.973
$\text{Cd}^{2+}$						
4	IAB(2–11)	277–278	1,215–1,312	27–29	0.7–0.9	0.969–0.978
5	IAB(3–14)	277–278	1,352–1,510	26–29	0.7–0.9	0.968–0.977
6	IAB(3–11)	277–278	1,200–1,297	26–29	0.7–1.0	0.964–0.977
$\text{Hg}^{2+}$						
7	IAB(2–9)	60–61	291–314	24–29	0.6–1.0	0.976–0.992
8	IAB(2–8)	60–61	249–275	20–29	0.9–1.2	0.972–0.991
9	IAB(2–15)	61	355–414	24–29	0.7–1.1	0.974–0.990

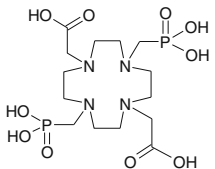
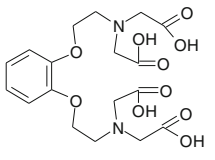
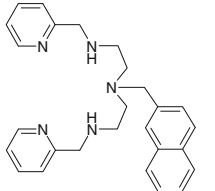
See the footnotes in Table 1;  $h$  is the number of latent variables,  $R_0^2$  is squared coefficient of determination

**Table 3** Experimental and predicted stability constant values  $\log K$  for extra test set

No.	Ligand	Cation	$\log K$		
			exp.	pred. MLR <sup>a</sup>	pred. PRM <sup>b</sup>
1		Zn <sup>2+</sup>	5.10 <sup>c</sup>	6.4 (1.5)	6.06 (0.10)
		Cd <sup>2+</sup>	8.10 <sup>c</sup>	6.80 (0.74)	6.84 (0.54)
		Hg <sup>2+</sup>	10.1 <sup>c</sup>	12.04 (0.35)	10.87 (0.57)
2		Cd <sup>2+</sup>	7.90 <sup>d</sup>	4.2 (1.3)	4.96 (0.46)
		Hg <sup>2+</sup>	8.65 <sup>d</sup>	8.6 (2.0)	8.57 (0.56)
3		Cd <sup>2+</sup>	5.8 <sup>c</sup>	3.3 (1.1)	6.23 (0.88)
4		Zn <sup>2+</sup>	8.95 <sup>c</sup>	4.78 (0.93)	7.33 (0.42)
		Cd <sup>2+</sup>	8.08 <sup>c</sup>	5.2 (1.4)	7.22 (0.22)
		Hg <sup>2+</sup>	9.62 <sup>c</sup>	10.2 (2.3)	10.10 (0.86)
5		Zn <sup>2+</sup>	8.14 <sup>c</sup>	5.30 (0.95)	7.74 (0.41)
		Cd <sup>2+</sup>	8.20 <sup>c</sup>	4.77 (0.82)	7.89 (0.48)
		Hg <sup>2+</sup>	10.88 <sup>c</sup>	10.2 (2.3)	8.65 (0.90)
6		Zn <sup>2+</sup>	7.5 <sup>c</sup>	9.9 (1.3)	12.84 (0.19)
		Cd <sup>2+</sup>	9.0 <sup>c</sup>	8.1 (1.2)	11.6 (1.1)
		Hg <sup>2+</sup>	11.8 <sup>c</sup>	15.3 (3.5)	15.5 (1.8)
7		Zn <sup>2+</sup>	5.4 <sup>c</sup>	4.18 (0.82)	7.49 (0.17)
		Hg <sup>2+</sup>	7.8 <sup>c</sup>	10.2 (2.2)	11.73 (0.41)
8		Zn <sup>2+</sup>	16.15 <sup>e</sup>	13.9 (1.7)	16.06 (0.70)
		Cd <sup>2+</sup>	17.20 <sup>e</sup>	9.75 (0.51)	16.93 (0.72)
9		Zn <sup>2+</sup>	17.9 <sup>e</sup>	13.12 (0.92)	16.93 (0.99)
		Cd <sup>2+</sup>	18.83 <sup>e</sup>	10.7 (2.2)	16.71 (0.90)
		Hg <sup>2+</sup>	30.28 <sup>e</sup>	24.6	28.8 (1.0)
10		Zn <sup>2+</sup>	7.13 <sup>f</sup>	5.77 (0.98)	6.74 (0.29)
		Cd <sup>2+</sup>	9.12 <sup>f</sup>	8.12 (0.38)	6.76 (0.70)
		Hg <sup>2+</sup>	10.68 <sup>f</sup>	12.9	8.3 (1.3)
11		Zn <sup>2+</sup>	8.5 <sup>f</sup>	7.08 (0.80)	10.24 (1.8)
		Cd <sup>2+</sup>	10.3 <sup>f</sup>	7.1 (1.3)	8.79 (0.42)
		Hg <sup>2+</sup>	13.1 <sup>f</sup>	15.5	10.6 (2.0)
12		Zn <sup>2+</sup>	9.58 <sup>d</sup>	10.02 (0.42)	11.77 (0.14)
		Cd <sup>2+</sup>	9.91 <sup>d</sup>	10.27 (0.15)	10.33 (0.51)



**Table 3** continued

No.	Ligand	Cation	log <i>K</i>		
			exp.	pred. MLR <sup>a</sup>	pred. PRM <sup>b</sup>
13		Zn <sup>2+</sup>	22.5 <sup>g</sup>	22.2 (1.5)	23.42 (0.72)
14		Zn <sup>2+</sup>	13.27 <sup>h</sup>	18.0 (2.6)	18.96 (0.65)
15		Zn <sup>2+</sup>	13.01 <sup>i</sup>	12.6 (1.1)	12.6 (1.0)
	$R_0^2$			0.629	0.823
	RMSE			3.2	2.2

Experimental data are given at 298 K and ionic strength 0.1 M excepting ligand 13, for which ionic strength is 0.15 M

<sup>a, b</sup> Predicted stability constant values log  $K_{\text{pred}}$  are computed using the consensus models of the MLR and PRM methods and standard deviations are given in parentheses

<sup>c</sup> Reference [5]

<sup>d</sup> Reference [38]

<sup>e</sup> Reference [39]

<sup>f</sup> Reference [6]

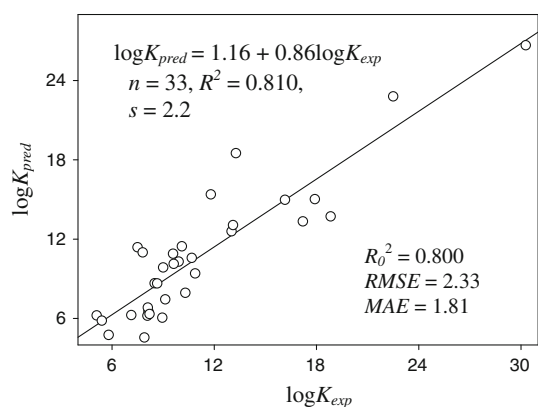
<sup>g</sup> Reference [40]

<sup>h</sup> Reference [41]

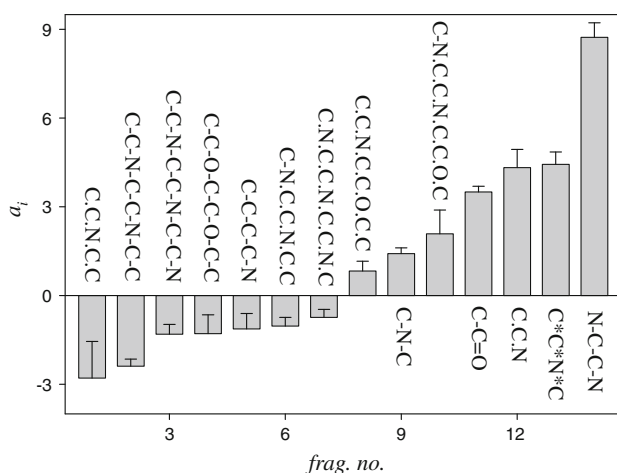
<sup>i</sup> Reference [42]

Among the SMF types applied, few of them enabled to build the models with best statistical parameters for most, if not for all training sets of the fivefold cross validation (Tables 1 and 2). Such results were obtained for the IAB(2–10) (Zn<sup>2+</sup>) and IAB(3–10) (Zn<sup>2+</sup>, Hg<sup>2+</sup>), IAB(3–15) and IAB(3–13) (Cd<sup>2+</sup>) fragment types using the MLR method. Squared LOO cross-validation correlation coefficient of these models is reasonable:  $Q^2 = 0.90$ – $0.97$  (Table 1). Notable PRM results are obtained with the IAB(2–15) (Zn<sup>2+</sup>), IAB(3–14) (Zn<sup>2+</sup>, Cd<sup>2+</sup>), IAB(2–11) (Cd<sup>2+</sup>) and IAB(2–9) (Hg<sup>2+</sup>) fragments (Table 2). The models are characterized by remarkable values of squared coefficient of determination  $R_0^2 = 0.96$ – $0.99$ , and the number of variables 2,966–3,499 (Zn<sup>2+</sup>), 1,200–1,510 (Cd<sup>2+</sup>) and 249–414 (Hg<sup>2+</sup>) is ten times more than in the best MLR models. The standard deviation of these models ( $s = 0.6$ – $1.2$ ) is very similar to the one of MLRs (Tables 1 and 2).

The MLR and PRM consensus models were applied to the prediction of the stability constant values log  $K$  and their standard deviations for an extra test set of the complexes of Zn<sup>2+</sup>, Cd<sup>2+</sup> and Hg<sup>2+</sup> ions with 15 recent synthesized organic ligands. The predictions were performed for the comparison with 33 experimental log  $K$  values [5, 6, 38–42] which were not presented in the sets used for the modeling. Mainly, the ligands are 12- and 15-membered macrocycles with the N, O and S donor atoms (Table 3). The predicted log  $K_{\text{pred}}$  values demonstrate accordance with the experimental log  $K_{\text{exp}}$  data:  $R_0^2 = 0.823$  and RMSE = 2.2 for PRM and  $R_0^2 = 0.629$  and RMSE = 3.2 for MLR correspondingly (Table 3). For the extra test set, the PRM technique demonstrates better performance over MLR due to relatively poor MLR assessments for Cd<sup>2+</sup> and 2,5,8-triaza-[9]-10,23-phenanthroline and 5-aminoethyl-2,5,8-triaza-[9]-10,23-phenanthroline ligands (no. 8 and 9 in Table 3). Figure 5 illustrates



**Fig. 5** Ensemble modeling by means the MLR and PRM consensus models: arithmetic mean predicted versus experimental values of the stability constant  $\log K$  for extra test set in Table 3



**Fig. 6** QSPR modeling of the stability constant  $\log K$  of the  $\text{Hg}^{2+}$  complexes with diverse organic ligands in water at 298 K and an ionic strength 0.1 M: molecular fragment contributions ( $a_i$ ) in the MLR model  $\log K = \sum a_i N_i$ , where  $N_i$  is an occurrence of the  $i$ th fragment. The SMF type is IAB (3–10); bond notation: ‘-’ single in chain, ‘.’ single in non-aromatic cycle, ‘=’ double in chain, ‘\*’ aromatic

ensemble modeling: the stability constant values were calculated as arithmetic mean of the predictions by means MLR and PRM CM. The predictions are characterized by reasonable values of  $R_0^2 = 0.800$  and  $\text{RMSE} = 2.3$ .

The individual MLR and PRM models can be interpreted, taking into account values and sign of fragment contributions. For instance, the stability constant  $\log K$  of the mercury complexes can be calculated using 14 fragment descriptors in one MLR model (Fig. 6). Some of fragments bring high positive (N–C–C–N in chain,  $\text{C}_{\text{ar}}\text{--}\text{C}_{\text{ar}}\text{--}\text{N}_{\text{ar}}\text{--}\text{C}_{\text{ar}}$  in aromatic cycle) or negative (C–C–N–C–C in non-aromatic cycle) contributions into  $\log K$ , whereas others (C–C–N–C–C–O–C–C in non-aromatic cycle) are less important (Fig. 6). The SMF contributions can be

recalculated to contributions of usual chemical groups [24, 31, 43] or individual atoms which are easily interpretable. Thus, relative contributions in  $\log K$  for heteroatoms in macrocyclic moieties for ligands **6** and **12** (Table 3) decrease in the order  $\text{N} > \text{O} > \text{S}$ . This is in agreement with experimentally observed trends in the stability constants for the  $(\text{Zn}^{2+})\text{L}$  complexes when the oxygen atom is replaced with the sulfur atom (see the ligands **2** and **3**, Table 3) or with the nitrogen atom (see the ligands **1** and **12**, Table 3). For the molecule **6**, the contribution of the N atom in the side chain is found smaller than that for the N atom of the macrocycle which allows us to suggest the larger role of the macrocyclic moiety in the complex formation.

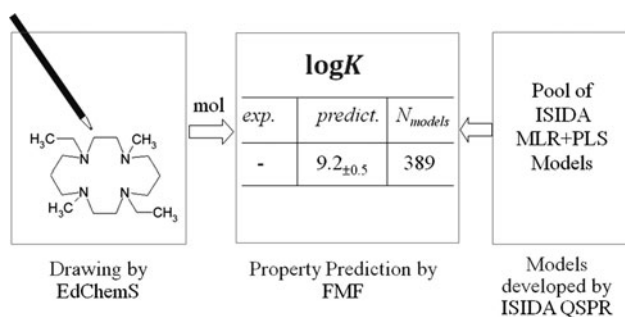
Various numbers of different fragments characterize the individual MLR models. The analysis of the fragment contributions of the models in a fragment library shows that these portions are about constant or vary in enough narrow limits. For example, shortest topological paths S–C–C–C=O ( $\text{Zn}^{2+}$ ), S–C–C–N ( $\text{Cd}^{2+}$ ) and O–C–C–C–O ( $\text{Hg}^{2+}$ ) in chain parts of the ligands contribute 4.6, 6.0 and 5.1  $\log K$  units into the stability constant according to 32, 49 and 27 models correspondingly (Table 4 and supplementary illustrations there). The fragments and their contributions are convenient tools for the rationale design of the ligands with desirable thermodynamic stability of

**Table 4** Selected molecular fragments and their mean contributions into  $\log K$  according to sets of individual MLR models

No.	SMF <sup>a</sup>	$\langle a_i \rangle$	$N_{\text{model}}$	$N_{\text{mol}}$
<b><math>\text{Zn}^{2+} + \text{L}</math>, 556 Ligands</b>				
1	S–C–C–C=O	4.59 (0.59)	32	14
2	O=[5]-S	3.08 (0.35)	58	14
3	N.[5].N	2.7 (1.0)	41	66
<b><math>\text{Cd}^{2+} + \text{L}</math>, 347 Ligands</b>				
4	S–C–C–O	5.7 (1.4)	63	10
5	S–C–C–N	5.96 (0.63)	49	14
6	N-[4]-S	2.41 (0.74)	57	14
<b><math>\text{Hg}^{2+} + \text{L}</math>, 76 Ligands</b>				
7	O–C–C–C–C–O	5.1 (2.4)	27	10
8	N-[7]-N	4.9 (1.7)	32	4
9	N-[4]-N	5.3 (1.8)	30	34

$\langle a_i \rangle$  is fragment contribution (arithmetic mean) and its standard deviation (in parentheses) according to  $N_{\text{model}}$  models and  $N_{\text{mol}}$  ligands

<sup>a</sup> substructural molecular fragments (SMF): S–C–C–C=O, S–C–C–O, S–C–C–N and O–C–C–C–C–O are shortest topological paths with explicit presentation of atoms and bonds and O=[5]-S, N.[5].N, N-[4]-S, N-[7]-N and N-[4]-N are terminal groups as shortest path sequences defined by length (in square brackets) and explicit indication of beginning atom and bond and ending bond and atom; bonds: ‘-’ and ‘=’ are single and double in chain, ‘.’ is single in non-aromatic cycle



**Fig. 7** Interactive design of chemical structure with desirable property value using the 2D sketcher EdChemS and the “Forecast by Molecular Fragments” (FMF) module (<http://infochim.u-strasbg.fr/recherche/Download/Download.php>)

their complexes. For this aim, the “Forecast by Molecular Fragments” (FMF) program was developed. FMF interacts with the 2D sketcher EdChemS [24, 26, 44] (<http://infochim.u-strasbg.fr/recherche/Download/Download.php>) (Fig. 7). If molecular structure is edited on the screen by EdChemS, FMF predicts the property interactively using loaded CM or selected individual model(s). For structure modifications, one can use own ideas and the library of fragments and their contributions. To demonstrate this option, 15 virtual ligands were designed for which the stability constant  $\log K$  on  $Zn^{2+}$  varies from 1 to 15 with the step about 1 (Table 5). MLR and PRM CMs and the AD methods were applied.

Recently, the COMET (Complexation of METals) predictor was developed [9] to apply the QSPR models for the predictions of the stability constants of the complexes of metal cations with organic ligands in solutions by means of Internet (<http://infochim.u-strasbg.fr/cgi-bin/predictor.cgi>). The COMET predictor contains the models for the complexation of alkaline-earth and lanthanide cations in water [9]. At present, the best models for the complexation of  $Zn^{2+}$ ,  $Cd^{2+}$  and  $Hg^{2+}$  obtained by the MLR and PRM methods are included in the COMET predictor. The models can be applied either individually or altogether to form the consensus model as an arithmetic mean over all individual models taking into account the applicability domain methods. It enables to design theoretically new organic ligands, thus providing experimentalists with structures of new potential metal binders.

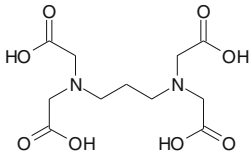
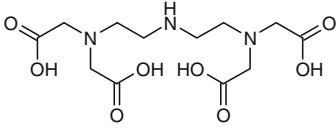
## Conclusions

A comparative study of Multiple Linear Regression Analysis of the ISIDA program package and newly realized Partial Robust M-regression has been performed for QSPR consensus modeling of the stability constant  $\log K$  of the 1:1 (M:L) complexes of metal cations  $Zn^{2+}$ ,  $Cd^{2+}$  and

**Table 5** Designed virtual ligands for which the stability constant  $\log K$  value on  $Zn^{2+}$  varies from 1 to 15 with the step about 1

No.	Ligand	$\log K_{\text{pred}}$	$s$	$N_m$
1		1.2	0.3	337
2		2.1	0.5	362
3		2.9	0.5	355
4		4.3	0.4	305
5		5.1	0.5	391
6		6.0	0.3	314
7		7.0	0.4	374
8		8.0	0.6	399
9		9.2	0.5	389
10		9.9	0.6	388
11		11.3	0.4	305
12		12.2	0.8	444
13		13.1	0.5	347

**Table 5** continued

No.	Ligand	log $K_{\text{pred}}$	$s$	$N_m$
14		14.0	0.4	320
15		15.4	0.5	331

Predicted stability constant log  $K_{\text{pred}}$ , standard deviation  $s$  and the number of individual models  $N_m$  of MLR and PRM CMs using the AD methods

$\text{Hg}^{2+}$  with the diverse sets of organic molecules in water at 298 K and an ionic strength 0.1 M. Two subclasses of the SMF descriptors were utilized: shortest topological paths with explicit presentation of atoms and bonds, and terminal groups as shortest path sequences defined by length and explicit indication of beginning atom and bond and ending bond and atom. The variation of the minimal and maximal numbers of constituent atoms in the sequences gives a multitude of the SMF types to build collections of numerous robust individual models for the preparation of the consensus models. Predictive performance of the consensus models was assessed using the fivefold external cross-validation procedure. Consensus MLR and PRM modeling with using the SMF descriptors represents a reliable tool for the prediction of the stability constants of the complexes of metal ions with organic ligands in water. The QSPR models are sampled for the design of new ligands with desired complexation properties. For this aim, the 2D sketcher EdChemS and the predictor FMF interactively or the predictor COMET via Internet can be applied. To demonstrate these options, 15 virtual ligands were designed for which the stability constant log  $K$  on  $\text{Zn}^{2+}$  varies from 1 to 15 with the step about 1.

**Acknowledgments** We thank GDR PARIS, the ARCUS project, CNRS France and the Russian Foundation for Basic Research (project no. 09-03-93106) for the support.

## References

1. Ferreiros-Martinez, R., Esteban-Gomez, D., Platas-Iglesias, C., de Blas, A., Rodríguez-Blas, T.: Zn(II), Cd(II) and Pb(II) complexation with pyridinecarboxylate containing ligands. *Dalton Trans.*, pp. 5754–5765 (2008)
2. Dudev, T., Lim, C.: Principles governing Mg, Ca, and Zn binding and selectivity in proteins. *Chem. Rev.* **103**(3), 773–787 (2003)
3. Ambrosi, G., Formica, M., Fusi, V., Giorgi, L., Guerri, A., Lucarini, S., Micheloni, M., Paoli, P., Rossi, P., Zappia, G.: Coordination behavior toward Copper(II) and Zinc(II) ions of three ligands joining 3-Hydroxy-2-pyridinone and polyaza fragments. *Inorg. Chem.* **44**(9), 3249–3260 (2005)
4. Clares, M.P., Aguilar, J., Aucejo, R., Lodeiro, C., Albelda, M.T., Pina, F., Lima, J.C., Parola, A.J., Pina, J., Seixas De Melo, J., Soriano, C., García-España, E.: Synthesis and H<sup>+</sup>, Cu<sup>2+</sup>, and Zn<sup>2+</sup> coordination behavior of a bis(fluorophoric) bibrachial lariat aza-crown. *Inorg. Chem.* **43**(19), 6114–6122 (2004)
5. Aragoni, M.C., Arca, M., Bencini, A., Blake, A.J., Caltagirone, C., Decortes, A., Demartin, F., Devillanova, F.A., Faggi, E., Dolci, L.S., Garau, A., Isaia, F., Lippolis, V., Prodi, L., Wilson, C., Valtancoli, B., Zaccheroni, N.: Coordination chemistry of *N*-aminopropyl pendant arm derivatives of mixed N/S-, and N/S/O-donor macrocycles, and construction of selective fluorimetric chemosensors for heavy metal ions. *Dalton Trans.*, no. 18, pp. 2994–3004 (2005)
6. Blake, A.J., Bencini, A., Caltagirone, C., De Filippo, G., Dolci, L.S., Garau, A., Isaia, F., Lippolis, V., Mariani, P., Prodi, L., Montalti, M., Zaccheroni, N., Wilson, C.: A new pyridine-based 12-membered macrocycle functionalised with different fluorescent subunits; coordination chemistry towards CuII, ZnII, CdII, HgII, and PbII. *Dalton Trans.*, no. 17, pp. 2771–2779 (2004)
7. Tamanini, E., Flavin, K., Motevalli, M., Piperno, S., Gheber, L.A., Todd, M.H., Watkinson, M.: Cyclam-Based “Clickates”: homogeneous and heterogeneous fluorescent sensors for Zn(II). *Inorg. Chem.* **49**(8), 3789–3800 (2010)
8. Varnek, A., Solov'ev, V.: Quantitative structure–property relationships in solvent extraction and complexation of metals. In: Sengupta, A.K., Moyer, B.A. (eds.) *Book: Ion exchange and solvent extraction, a series of advances*, vol. 19. pp. 319–358. CRC Press Taylor and Francis Group, Boca Raton (2009)
9. Varnek, A., Fourches, D., Kireeva, N., Klimchuk, O., Marcou, G., Tsvadze, A., Solov'ev, V.: Computer-aided design of new metal binders. *Radiochim. Acta* **96**, 505–511 (2008)
10. Shi, Z.G., McCullough, E.A.: A computer-simulation—statistical procedure for predicting complexation equilibrium-constants. *J. Incl. Phenom. Mol. Recognit. Chem.* **18**(1), 9–26 (1994)
11. Gakh, A.A., Sumpter, B.G., Noid, D.W., Sachleben, R.A., Moyer, B.A.: Prediction of complexation properties of crown ethers using computational neural networks. *J. Incl. Phenom. Mol. Recognit. Chem.* **27**(3), 201–213 (1997)
12. 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**(3), 847–858 (2000)
13. 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**(4), 812–829 (2002)
14. Solov'ev, V.P., Varnek, A.A.: Structure-property modeling of metal binders using molecular fragments. *Russ. Chem. Bull.* **53**(7), 1434–1445 (2004)
15. Ghasemi, J., Saaidpour, S.: QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression. *J. Incl. Phenom. Macrocycl. Chem.* **60**(3–4), 339–351 (2008)
16. Qi, Y.-H., Zhang, Q.-Y., Xu, L.: Correlation analysis of the structures and stability constants of gadolinium(III) complexes. *J. Chem. Inf. Comput. Sci.* **42**(6), 1471–1475 (2002)
17. Miličević, A., Raos, N.: Prediction of stability of copper(II) and nickel(II) complexes with fructose-amino acids from the molecular graph models developed on amino acid chelates. *Croat. Chem. Acta* **80**(3–4), 557–563 (2007)

18. 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.* **16**(9), 1175–1184 (1990)
19. 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**(9), 611–617 (2004)
20. Cabaniss, S.E.: Quantitative structure–property relationships for predicting metal binding by organic ligands. *Environ. Sci. Technol.* **42**(14), 5210–5216 (2008)
21. 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**(8), 574–578 (2001)
22. 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**(12), 877–880 (2002)
23. Svetlitski, R., Lomaka, A., Karelson, M.: QSPR modelling of lanthanide-organic complex stability constants. *Sep. Sci. Technol.* **41**(1), 197–216 (2006)
24. Solov'ev, V.P., Kireeva, N.V., Tsvadze, A.Y., Varnek, A.A.: Structure–property modelling of complex formation of strontium with organic ligands in water. *J. Struct. Chem.* **47**(2), 298–311 (2006)
25. 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**(2), 808–819 (2006)
26. 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. Design* **19**(9–10), 693–703 (2005)
27. Swamy, M.N.S., Thulasiraman, K.: *Graphs, networks, and algorithms*. Wiley, New York (1981)
28. Pettit, G., Pettit, L.: IUPAC stability constants database. <http://www.acadsoft.co.uk/> (2010)
29. 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 Design* **4**(3), 191–198 (2008)
30. Varnek, A.: ISIDA (In silico design and data analysis) program. <http://infochim.u-strasbg.fr/recherche/isida/index.php> (2010)
31. Varnek, A., Solov'ev, V.P.: “In Silico” design of potential anti-HIV actives using fragment descriptors. *Comb. Chem. High Throughput Screen* **8**(5), 403–416 (2005)
32. 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**(1), 71–102 (2003)
33. 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**(3), 1111–1122 (2007)
34. Muller, P.H., Neumann, P., Storm, R.: *Tafeln der mathematischen Statistik*. VEB Fachbuchverlag, Leipzig (1979)
35. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. *Numer. Math.* **14**, 403–420 (1970)
36. 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**(3), 927–939 (2007)
37. Sermeels, S., Croux, C., Filzmoser, P., Van Espen, P.J.: Partial robust M-regression. *Chemom. Intell. Lab. Syst.* **79**, 55–64 (2005)
38. Caltagirone, C., Bencini, A., Demartin, F., Devillanova, F.A., Garau, A., Isaia, F., Lippolis, V., Mariani, P., Papke, U., Tei, L., Verani, G.: Redox chemosensors: coordination chemistry towards CuII, ZnII, CdII, HgII, and PbII of 1-aza-4,10-dithia-7-oxacyclododecane ([12]aneNS2O) and its N-ferrocenylmethyl derivative. *Dalton Trans.*, issue no. 5, pp. 901–909 (2003)
39. Bazzicalupi, C., Bencini, A., Berni, E., Bianchi, A., Borsari, L., Giorgi, C., Valtancoli, B., Lodeiro, C., Lima, J.C., Parola, A.J., Pina, F.: Protonation and coordination properties towards Zn(II), Cd(II) and Hg(II) of a phenanthroline-containing macrocycle with an ethylamino pendant arm. *Dalton Trans.*, issue no. 4, pp. 591–597 (2004)
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**(9), 3851–3862 (2008)
41. Baranyai, Z., Bombieri, G., Meneghetti, F., Tei, L., Botta, M.: A solution thermodynamic study of the Cu(II) and Zn(II) complexes of EBTA: X-ray crystal structure of the dimeric complex [Cu2(EBTA)(H2O)3]2. *Inorg. Chim. Acta* **362**, 2259–2264 (2009)
42. 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**(14), 6173–6183 (2008)
43. 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.* **44**(4), 1365–1382 (2004)
44. 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.* **44**(2), 529–541 (2004)