# Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments

V. P. Solov'ev

Institute of Physiologically Active Compounds, Russian Academy of Sciences, 142432 Chernogolovka, Moscow region, Russia

A. Varnek and G. Wipff*

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

Received October 28, 1999

A substructural molecular fragment (SMF) method has been developed to model the relationships between the structure of organic molecules and their thermodynamic parameters of complexation or extraction. The method is based on the splitting of a molecule into fragments, and on calculations of their contributions to a given property. It uses two types of fragments: atom/bond sequences and "augmented atoms" (atoms with their nearest neighbors). The SMF approach is tested on physical properties of $C_2-C_9$ alkanes (boiling point, molar volume, molar refraction, heat of vaporization, surface tension, melting point, critical temperature, and critical pressures) and on octanol/water partition coefficients. Then, it is applied to the assessment of (i) complexation stability constants of alkali cations with crown ethers and phosphoryl-containing podands, and of $\beta$-cyclodextrins with mono- and 1,4-disubstituted benzenes, and (ii) solvent extraction constants for the complexes of uranyl cation by phosphoryl-containing ligands.

## INTRODUCTION

Binding of metal cations by specific organic molecules (ionophores) in homogeneous solutions (complexation) or in biphasic liquid−liquid systems (solvent extraction) is involved in important processes widely used in industry and in research laboratories. Many efforts have been devoted to design "the best" ionophores that selectively bind a given metal, proceeding in an essentially empirical manner. A question that can be asked today is whether it is possible to use current computer tools to design an ionophore with particular desired characteristics. We feel that this is possible if a suitable information platform involving a collection of available experimental data on thermodynamics of complexation or extraction, coupled with the programs that establish quantitative relationships between structure of ionophores and their properties, were to exist. Such an information platform should contain three main elements: (i) an ensemble of databases, (ii) quantitative structure−property relationships (QSPR), and (iii) molecular modeling methods.

Unfortunately, such an ensemble of computational tools is still far from being established, although some of its elements are already well-developed. Thus, several databases on complexation of metals[1−4] as well as handbooks and reviews on complexation in solution[5−12] and on solvent extraction[13−18] are available. Modern methods of molecular modeling (molecular dynamics, free energy perturbation techniques, quantum mechanics[19]) allow one to study molecular recognition patterns and specific features of host−guest interactions using explicit representations of all molecules, including solvent(s). However, current computer power limits the size of simulated systems to up to 12000−15000 atoms, which corresponds to 2000−3000 solvent molecules and several ionophore molecules, metal cations, and counterions. Sometimes, this is not sufficient to model real systems containing complex molecular aggregates (micelles, vesicles, etc.). Molecular dynamics simulations still consume too much computer time to design efficiently new complexation and extraction systems from scratch without preliminary knowledge of experimental data. Continuing advances in computer technology render such ab initio approaches a viable prospect of the future.

Phenomenological QSPR may be alternatively used to accelerate the search for ionophores with desired properties and may serve as a "bridge" between experimental information stored in databases and molecular modeling methods. These methods have reached tremendous success in pharmacology and in many areas of chemistry, but have little been used in the fields of complexation and extraction. Most related publications concern the development of relationships (correlations) between constants of complexation and of extraction and some physical or chemical properties.[20−22] Thus, Shi and McCullough[23] performed a multiple linear regression analysis using various experimental and calculated parameters (radius and electronegativity for metal cations, dielectric constant for solvents, total energy and its components for ligands, and some others) as descriptors. Their studied systems included up to 314 metal cation−solvent−ligand combinations leading to standard deviations ranging from 0.36 to 1.42 log $\beta$ units. To assess the stability of complexes of some crown ethers with alkali cations, Schneider et al.[24] used H-bonding electron-donor factors[25,26] of binding sites of macrocycles as descriptors. Experimental data on extraction of actinide elements by phosphoryl-containing ionophores were correlated with indices such as group

10.1021/ci9901340 CCC: $19.00 
Published on Web 03/21/2000

electronegativities or Hammett−Taft parameters of molecular fragments,[22] atomic charges,[22,27] electrostatic potential distribution,[27,28] and donor−acceptor interaction energies.[29]

Another type of modeling based on neural networks method has been used by Gakh et al.[30] to assess complexation stability constants of alkali cations with crown ethers in methanol. At the first step of calculations, the variable parameters were fitted for compounds from the learning set; then stability constants were "predicted" within experimental error for the validation set. As input, only the number of aromatic moieties, the ring size, and the number of ether oxygens were used.[30]

Although in all cited studies reasonable correlations between a target property and a number of descriptors were obtained, it is not clear whether they can be used as a predictive tool for estimation of thermodynamics properties of complexation for various classes of ionophores.

The goal of this study is to develop a substructural molecular fragments (SMF) method and the related software tools, to model relationships between the structure of ionophores and their complexation and extraction properties. This method is based on the representation of a molecule by its fragments and on the calculation of their contributions to a given property. It uses two types of fragments: (i) the sequences of atoms and/or bonds (atom and/or bond paths up to specified maximal length) and (ii) "augmented atoms" represented by a selected atom with its environment. In fact, it represents an extension of empirical methods used to calculate physical or chemical properties of molecules using atomic or bond increments.[31−35]

We describe the SMF method used to model a given property and the related computer program "TRAIL". Then, we report test calculations on well-studied systems including an assessment of eight physical properties of alkanes and of octanol/water partition coefficients for organic molecules. Then, the modeling of thermodynamic parameters of some complexation and extraction systems is reported. For these purpose, different types of systems were selected: (i) complexation of an alkali cation ($Na^+$) with macrocyclic host molecules (crown ethers) in polar protic solvent (MeOH), (ii) complexation of an alkali cation ($K^+$) with noncyclic polydentate ionophores (podands) in aprotic solvent (mixture of tetrahydrofurane with chloroform), (iii) complexation of neutral molecular guests (1,4-disubstituted benzenes) by a large macrocyclic host ($\beta$-cyclodextrin) in water, and (iv) extraction of the uranyl cation by phosphoryl-containing ligands.

## 2. METHODS

The *substructural molecular fragments* (SMF) method is based on the splitting of a molecule into fragments, and on the calculation of their contributions to a given property ***X***. Two different types of fragments are considered (Figure 1): "sequences" (**I**) and "augmented atoms" (**II**). For each type of fragment one can define three subtypes **AB**, **A**, and **B** (Figure 1). For the fragments **I**, they represent sequences of atom and bond types (**AB**), of atom types only (**A**), or of bond types only (**B**). These fragments correspond to sequential set of atoms linked by chemical bonds where either atom types (C, N, O, ...) or bond types (single, double, ...), or both, of them are considered explicitly. In the following, we specify the number of atoms of a given sequence. For instance, **I**(**AB**, 2-6) refers to all sequences containing from 2 to 6 atoms connected by bonds of specified type. For **I**(**A**, 2-6), the definition is similar, but bond types are omitted, whereas for **I**(**B**, 2-6) only the types of bonds are considered. Only shortest paths from one atom to the other are used, as shown in Figure 1. An "augmented atom" represents a selected atom with its nearest environment including either neighboring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). Atomic hybridization (**Hy**) can be taken into account for augmented atoms of the **A**-type.

Once a given compound is split into constitutive fragments, any corresponding quantitative physical or chemical property ***X*** is calculated from the fragments contributions using linear (1) or nonlinear (2) and (3) fitting equations.

$$\boldsymbol{X} = a_0 + \sum_i a_i N_i \tag{1}$$

$$\boldsymbol{X} = a_0 + \sum_i a_i N_i + \sum_i b_i (2N_i^2 - 1) \tag{2}$$

$$\boldsymbol{X} = a_0 + \sum_i a_i N_i + \sum_{i,k} b_{ik} N_i N_k \tag{3}$$

where $a_i$ and $b_i$ ($b_{ik}$) are fragment contributions and $N_i$ is the number of fragments of $i$ type. The $a_0$ term is fragment-independent; it is fitted by default, but optionally it can be omitted.

**2.1. Software Development.** The TRAIL program has been developed to calculate structure−property relationships based on the SMF partitioning. The program reads input data in the structure−data file (SDF)[36] format containing structural and properties data. It also accepts SYBYL MOL2[37] format, which contains structural information; in that case a file containing properties data should be input separately. The graphical interface of TRAIL allows one to attribute data to the learning or to the validation sets and to set up the parameters of calculations (type of fragments, minimal and maximal number of atoms/bonds in the sequences, type of environment for augmented atoms, type of equation (eqs 1−3), etc.).

At the first step of calculations, TRAIL generates substructural molecular fragments from the molecules for the learning set. For the sequences, the program uses the Floyd algorithm[38] to calculate the shortest paths between each pair of atoms (Figure 1). The number of atoms (bonds) in these sequences varies from two to six. Augmented atoms are generated directly from the connectivity table. Then TRAIL creates a list of independent fragments, considering, for example, the C−C−N and N−C−C sequences as one fragment. Molecules containing fragments of "rare" occurrence (i.e., found in less than two molecules) are excluded from the learning set. If some fragments are linearly dependent, they form a single group defined as an extended fragment. In most cases, hydrogen atoms are not taken into account, but they can be optionally included into calculations.

At the second step, TRAIL fits the $a_i$ and $b_i$ terms in eqs 1−3 using the singular value decomposition method,[39] calculates corresponding statistical characteristics (correlation coefficient ($R$), standard deviation ($s$), Fischer's criterion ($F$), $R_H$ factor of Hamilton, and matrix of pair correlations (covariation matrix) for the terms $a_i$ and $b_i$) and performs

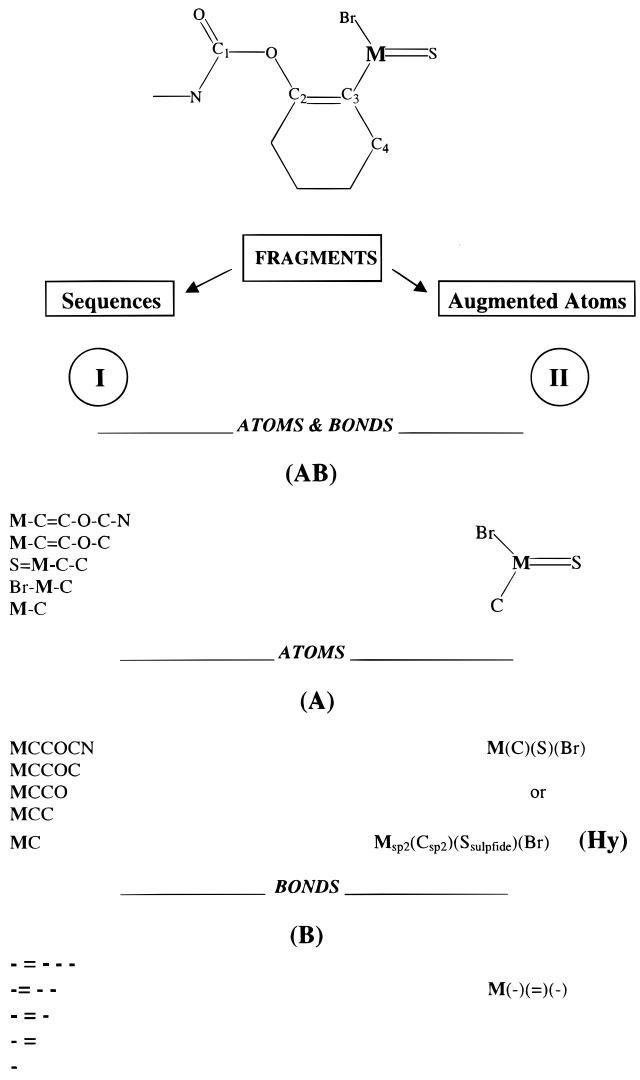


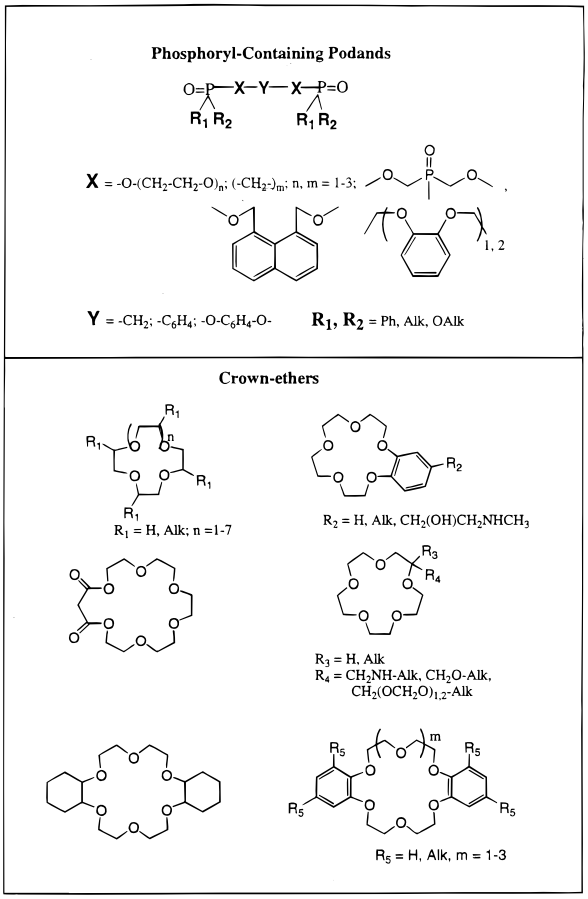

**Figure 1.** Different types of substructural molecular fragments: shortest path sequences (**I**) and augmented atoms (**II**) including atoms and bonds (**AB**), only atoms (**A**) or only bonds (**B**). The shortest pathway from N to N is **M**-$C_3$═$C_2$-O-$C_1$-N. From top to bottom: the sequences (**I**) correspond to the **I**(**AB**, 2-6), **I**(**A**, 2-6) and **I**(**B**, 2-6) types.

statistical tests[40] to select the best models. Finally, using the fitted fragments contributions, the program calculates the "predicted" values ***X*** for the compounds from the validation set.

TRAIL can be used in automatic or in manual modes. In automatic mode, the program uses all possible types of fragments coupled with one of three fitting equations (1)−(3), which corresponds to about 150 models of calculations. If the total number of fitted terms $a_i$ and $b_i$ for to a given computational scheme is more than the number of molecules $n$, the program skips this model and moves to another one. At the end of the calculations, TRAIL rejects models for which Fischer's criterion is not large enough. Selection of the most reasonable model is based on comparison of the statistical parameters $R$, $F$, and $s$ according to standard procedure given in ref 40. In manual mode, all parameters (type of fragments, minimal and maximal number of atoms/bonds in the sequences, type of environment for augmented atoms, type of equation (eqs 1−3), etc.) are user-defined.

In the output files, TRAIL presents a list of substructural fragments, their occurrence in the whole set and in each

**Chart 1.** Typical Podands and Crown Ethers Calculated by the SMF Method




molecule, fitted values $a_i$ and $b_i$, statistical characteristics, a matrix of pair correlations, a list of experimental and calculated properties ***X***, and the plots of calculated ($\boldsymbol{X}_{calc}$) vs experimental ($\boldsymbol{X}_{exp}$) values as well as of residuals[41] ($\boldsymbol{X}_{calc} - \boldsymbol{X}_{exp}$) vs $\boldsymbol{X}_{exp}$.

The program has been written using the DELPHI programming platform[42,43] for the WINDOWS 95/98/NT environment.

**2.2. Preparation of Input Data. 2.2.1. Alkanes.** First, a text file containing an array of eight properties (boiling and melting points, molar volume, molar refraction, heat of vaporization, surface tension, critical temperature, and critical pressure) for 74 normal and branched alkanes from ref 44 has been prepared. The 2D structures of all compounds were made with the ISIS/Draw program[45] in MOL format.[36] Finally, a file converter has been developed to merge the above two files into an input file for TRAIL in SDF format.[36]

**2.2.2. Octanol/Water Partition Coefficient.** An input file for TRAIL in SDF format has been prepared by converting the text file with experimental log $P$ values for 1831 organic molecules and the SYBYL MOL2[36] files with their 3D structures. The initial files were delivered to us by Dr. Wang and Professor Lai.[31]

**2.2.3. Complexation Data.** Complexation data (stability constants) for podands,[46−56] crown ethers,[7] (Chart 1) and $\beta$-cyclodextrin[57] were critically selected from the THECOMAC database,[3] which can export data into SDF format.

An input SDF file containing equilibrium constants for *extraction* of the uranyl cation by 12 organophosphorous

**Table 1.** Physical Properties[a] of Alkanes: Statistical Criteria[b] of Models (Eq 1) with Different Types of Fragments[c]

| no. | property | type of fragment[c] | *n* | *R* | *F* | *s* |
|---|---|---|---|---|---|---|
| 1 | bp | **I**(**AB**, 2-5) | 67 | 0.995 | 1553 | 4.9 |
| | | **II**(**AB**) | 67 | 0.990 | 808 | 6.8 |
| 2 | MV | **I**(**AB**, 2-5) | 62 | 0.9992 | 9224 | 0.69 |
| | | **II**(**AB**) | 62 | 0.988 | 580 | 2.7 |
| 3 | MR | **I**(**AB,** 2-5) | 62 | 0.99996 | 156705 | 0.052 |
| | | **II**(**AB**) | 62 | 0.9993 | 10172 | 0.20 |
| 4 | $H_V$ | **I**(**AB**, 2-5) | 62 | 0.998 | 3243 | 0.37 |
| | | **II**(**AB**) | 62 | 0.993 | 1069 | 0.64 |
| 5 | $T_C$ | **I**(**AB**, 2-5) | 67 | 0.988 | 622 | 9.5 |
| | | **II**(**AB**) | 67 | 0.979 | 358 | 12 |
| 6 | $P_C$ | **I**(**AB**, 2-5) | 67 | 0.981 | 389 | 0.90 |
| | | **II**(**AB**) | 67 | 0.937 | 112 | 1.6 |
| 7 | ST | **I**(**AB**, 2-5) | 61 | 0.982 | 369 | 0.39 |
| | | **II**(**AB**) | 61 | 0.896 | 57 | 0.90 |
| 8 | mp | **I**(**AB**, 2-5) | 49 | 0.600 | 6 | 30 |
| | | **II**(**AB**) | 49 | 0.626 | 7 | 29 |

[a] Boiling point (bp), molar volume (MV), molar refraction (MR), heat of vaporization ($H_V$), surface tension (ST), critical temperature ($T_C$), critical pressure ($P_C$), and melting point (mp). Experimental data are taken from ref *44*. [b] Number of compounds in the learning set (*n*), correlation coefficient (*R*), Fisher's criterion (*F*), and standard deviation (*s*). [c] See Figure 1 for fragments definition. The total number of fragment types is four in all calculations.

compounds[13] was prepared by merging the text file containing an array of data and the MOL file[36] containing 2D structures, using a developed file converter.

## 3. RESULTS

**3.1. Test Calculations. 3.1.1. Physical Properties of Alkanes.** Alkanes represent an especially attractive class of compounds as a starting point for the application of modeling techniques based, for instance, on Wiener distance indices,[58] information indices,[59] connectivity indices,[60] ad hoc descriptors,[61] weighted path numbers,[62] and some others.[59] A few years ago, Needham et al.[44] used connectivity indices[60] and ad hoc descriptors[61] to assess eight physical properties for 74 normal and branched alkanes: boiling and melting points, molar volume, molar refraction, heat of vaporization, surface tension, critical temperature, and critical pressure. We used an array of experimental data from Needham's article[44] to test the reliability of the SMF method to model structure−property relationships, taking 7 compounds for the validation set and treating the other 67 as the learning set.

The first series of fitting calculations was performed on the **I**(**AB**, 2-5) and **II**(**AB**) fragments using the linear fitting equation in eq 1. Results given in Table 1show that augmented atoms are less satisfactory descriptors than atom/bond sequences because they lead to smaller *R* and *F* values and to larger values of *s* for any physical property.

The second series of calculations (Table 2) was done only for the **I**(**AB**, 2-5) and **I**(**A**, 2-6) atom/bond sequences[63] using linear (eq 1) or nonlinear (eq 3) fitting equations. Since in these calculations we used longer sequences than for those given in Table 1, the number of variables was also larger. Resulted statistical characteristics are, at least, of the same level of accuracy as those obtained by Needham et al.[44] for alkanes using connectivity indices and ad hoc descriptors. For the melting point our results are even better. Thus, *R*, *F*, and *s* values obtained for the nonlinear model with the **I**(**AB**, 2-5) fragments (0.924, 14, and 16, respectively, Table 2) are more satisfactory than those obtained with connectivity indices (0.755, 13, and 23.8) or with ad hoc descriptors (0.606, 10, and 287.3).[44]

**Table 2.** Modeling of Physical Properties[a] of $C_2-C_9$ Alkanes Using **I**(**AB**, 2-6) Atom/Bond Sequences

| no. | property | *n* | *R* | *F* | *s* |
|---|---|---|---|---|---|
| 1 | bp | 67 | 0.998 | 3769 | 2.82 |
| 2 | MV | 62 | 0.9994 | 8639 | 0.64 |
| 3 | MR | 62 | 0.99996 | 131762 | 0.050 |
| 4 | $H_V$ | 62 | 0.998 | 2795 | 0.35 |
| 5 | $T_C$ | 67 | 0.994 | 1040 | 6.6 |
| 6 | $P_C$ | 67 | 0.983 | 360 | 0.83 |
| 7 | ST | 61 | 0.987 | 420 | 0.33 |
| 8 | mp | 49 | 0.627 | 6 | 29 |
| 8[b] | mp | 49 | 0.924 | 14 | 16 |

[a] See footnotes for Table 1. If it is not specified, eq 1 has been used. The total number of fragments is five in all calculations. [b] Calculations with the **I**(**AB**, 2-5) fragments and eq 3.

Plots of calculated ($\boldsymbol{X}_{calc}$) vs experimental ($\boldsymbol{X}_{exp}$) values and statistical characteristics of their correlations (Figure 2) illustrate the quality of these calculations: for all properties (except the melting point) *R* varies from 0.983 to 0.99996, and *F* varies from 1918 to 705 868. For the melting point, both statistical characteristics are less satisfactory but still acceptable (0.924 and 274, respectively).

Comparison of statistical characteristics for given property $\boldsymbol{X}$ in Tables 1 and 2 shows that increasing the number of terms in eq 1 does not always lead to better results. Thus, calculations with five fragments (Table 2) instead of four fragments (Table 1) leads to larger correlation coefficients and Fisher's criterion for boiling and melting points, critical pressure and temperature, and surface tension; however, calculation for other properties become worse.

The performance of the SMF method is demonstrated in Table 3 for the compounds from the validation set where experimental values are compared with those calculated using **I**(**A**, 2-6) fragments. This comparison shows that the SMF method represents a reliable tool of structure−property modeling for alkanes.

**3.1.2. Octanol/Water Partition Coefficients.** Octanol/water partition coefficient (log *P*) is an important parameter widely used in QSPR studies in chemistry and biology.[64,65] It can be a particularly useful descriptor of hydrophobicity of complexant and extractant molecules.[66] Most of empirical approaches of log *P* calculations are based on the contributions of molecular fragments (augmented atoms, functional groups).[31,33,65−68] To our knowledge, no atom/bond sequences of more than two atoms were so far used for this purpose.

For the learning set we used experimental log *P* values of 1831 organic molecules containing various functional groups, as did Wang et al.[31] Two input sets in the SDF format were prepared for TRAIL: one with hydrogen atoms omitted and another one with all atoms represented explicitly. Depending on the model, molecules containing "rare" fragments were excluded from the learning set. Calculations using eqs 1−3 and different types of fragments have been performed in the automatic mode; statistical parameters of some representative models are given in Table 4.

Among different computational schemes applied, only a few of them reasonably model log *P* (Table 4). The best results were obtained for the **I**(**AB**, 2-4)[63] and **II**(**Hy**)

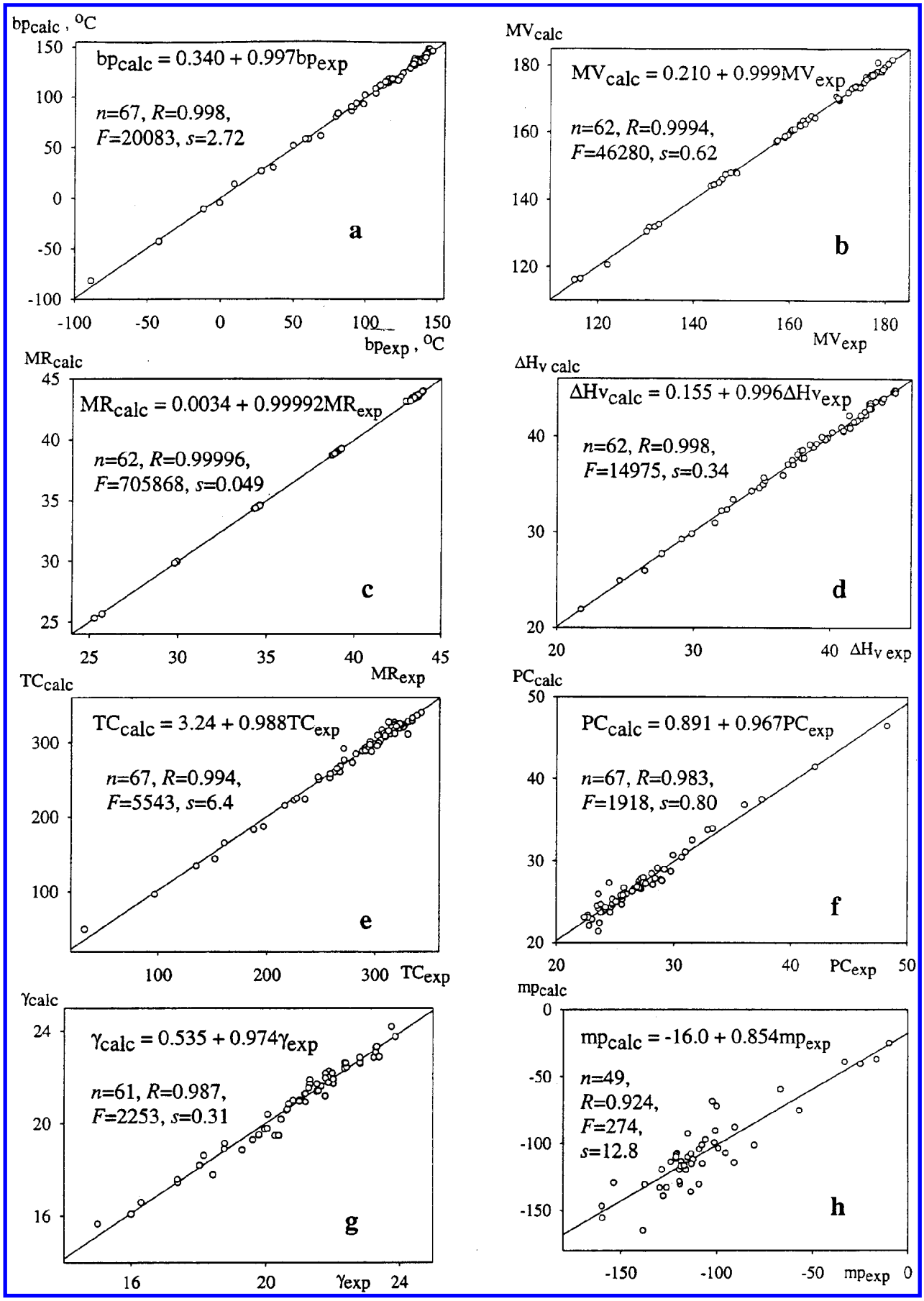



**Figure 2.** $C_2-C_9$ alkanes: correlation between calculated and experimental physical properties. (a) boiling points (°C), (b) molar volumes, (c) molar refractions, (d) heats of vaporization, (e) critical temperatures, (f) critical pressures, (g) surface tensions, and (h) melting points.

**Table 3.** Physical Properties of Alkanes: Experimental and Predicted Values for the Validation Set[a,b]

| | | bp | | MV | | MR | | $H_V$ | | $T_C$ | | $P_C$ | | ST | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no. | compound | exp. | pred. | exp. | pred. | exp. | pred. | exp. | pred. | exp. | pred. | exp. | pred. | exp. | pred. |
| 1 | 3-Me-pentane | 63.28 | 62.24 | 129.72 | 130.14 | 29.8016 | 29.8089 | 30.27 | 30.23 | 231.2 | 228.1 | 30.83 | 31.17 | 18.12 | 18.02 |
| 2 | 3,3-$Me_2$-pentane | 86.06 | 87.12 | 144.53 | 144.77 | 34.3323 | 34.3636 | 33.02 | 33.24 | 263.0 | 261.2 | 30.0 | 28.80 | 19.59 | 19.32 |
| 3 | 2,5-$Me_2$-hexane | 109.10 | 106.58 | 164.70 | 164.58 | 39.2596 | 39.2748 | 37.86 | 38.17 | 279.0 | 273.9 | 25.0 | 25.26 | 19.73 | 19.63 |
| 4 | *n*-nonane | 150.80 | 155.28 | 178.71 | 179.01 | 43.8423 | 43.8358 | 46.44 | 45.90 | 322.0 | 333.1 | 22.74 | 21.02 | 22.92 | 22.89 |
| 5 | 2,6-$Me_2$-heptane | 135.21 | 141.72 | 180.91 | 179.96 | 43.9258 | 43.8949 | 42.82 | 43.31 | 306.0 | 317.2 | 23.7 | 21.70 | 20.83 | 21.63 |
| 6 | 2,2,4-$Me_3$-hexane | 126.54 | 124.97 | 179.22 | 179.18 | 43.7638 | 43.8105 | 40.57 | 40.58 | 301.0 | 300.7 | 23.39 | 23.98 | 20.51 | 20.63 |
| 7 | 2,4-$Me_2$,3-Et-pentane | 136.73 | 139.62 | 173.80 | 172.94 | 43.4037 | 43.3521 | 42.93 | 42.68 | 324.2 | 328.4 | 25.46 | 26.03 | 22.8 | 23.05 |

[a] See footnotes of Table 1. [b] Calculations with the **I(AB**, 2-6) fragments (Figure 1) using eq 1.

**Table 4.** Assessment of Octanol/Water Partition Coefficients Using Different Models[a,b]

| no. | fragment | fitting equation[d] | $N$[c] | $n$ | $R$ | $F$ | $s$ |
|---|---|---|---|---|---|---|---|
| 1 | **I(AB**, 2-2) | 1 | 30 | 1828 | 0.893 | 244 | 0.68 |
| | *e* | 1 | 26 | 1828 | 0.816 | 144 | 0.88 |
| | | 2 | 61 | 1828 | 0.908 | 138 | 0.65 |
| | | 3 | 496 | 1813 | 0.955 | 28 | 0.53 |
| 2 | **I(AB**, 2-3) | 1 | 128 | 1813 | 0.942 | 104 | 0.53 |
| | *e* | 1 | 102 | 1814 | 0.929 | 107 | 0.58 |
| | | 2 | 275 | 1813 | 0.958 | 62 | 0.47 |
| 3 | **I(A**, 2-3) | 1 | 71 | 1823 | 0.930 | 159 | 0.57 |
| | *e* | 1 | 51 | 1823 | 0.884 | 127 | 0.72 |
| | | 2 | 151 | 1823 | 0.944 | 92 | 0.52 |
| 4 | **I(AB**, 3-3) | 1 | 104 | 1813 | 0.941 | 129 | 0.52 |
| | *e* | 1 | 81 | 1814 | 0.868 | 66 | 0.77 |
| | | 2 | 215 | 1813 | 0.954 | 76 | 0.48 |
| **5** | **I(A, 2-4)** | **1** | **156** | **1804** | **0.962** | **133** | **0.43** |
| | *e* | 1 | 103 | 1809 | 0.910 | 81 | 0.65 |
| | | 2 | 347 | 1804 | 0.972 | 71 | 0.40 |
| **6** | **I(AB, 2-4)** | **1** | **354** | **1760** | **0.976** | **81** | **0.36** |
| | *e* | 1 | 264 | 1766 | 0.966 | 80 | 0.42 |
| | | 2 | 803 | 1760 | 0.85 | 38 | 0.36 |
| 7 | **11(A)** | 1 | 111 | 1800 | 0.942 | 122 | 0.52 |
| | *e* | 1 | 82 | 1807 | 0.919 | 116 | 0.61 |
| | | 2 | 231 | 1800 | 0.952 | 65 | 0.50 |
| 8 | **II(AB)** | 1 | 141 | 1790 | 0.948 | 105 | 0.50 |
| | *e* | 1 | 138 | 1791 | 0.948 | 108 | 0.50 |
| | | 2 | 291 | 1790 | 0.957 | 57 | 0.48 |
| 9 | **II(B)** | 1 | 31 | 1830 | 0.904 | 269 | 0.65 |
| | *e* | 1 | 42 | 1830 | 0.913 | 220 | 0.62 |
| | | 2 | 61 | 1830 | 0.914 | 149 | 0.62 |
| **10** | **II(Hy)** | **1** | **293** | **1723** | **0.976** | **97** | **0.36** |
| | *e* | 1 | 285 | 1725 | 0.976 | 101 | 0.36 |
| | | 2 | 625 | 1723 | 0.981 | 44 | 0.37 |

[a] If not specified, all molecules are represented with hydrogen atoms; the most reasonable models are given in bold. [b] See Table 1 and Figure 1 for the definitions. [c] The number of fitted parameters. [d] Equations 1, 2, or 3 (see text). [e] Hydrogens are excluded.

fragments with hydrogens included. The standard deviation in these models ($s = 0.36$) is similar to the one reported in ref 31. Reasonable results are also obtained with the **I(A**, 2-4) fragments represented "with hydrogens" ($R = 0.962$, $s = 0.43$) where the number of variables is much smaller than for the best models **I(AB**, 2-4) and **II(Hy)**.

Some interesting conclusions could be drawn from the statistical characteristics of different models given in Table 4. Calculations made with all-atom models lead in most cases to better results than those with hydrogens omitted. Thus, omitting hydrogens for the calculations with **I(AB**, 2-4) fragments results in decreasing of $R$ from 0.976 to 0.966; this can be explained by the reduction of the number of variables from 353 to 263. Analogously, for the **I(A**, 2-4) fragment sets, $R$ drops from 0.962 to 0.910 and $s$ increases from 0.43 to 0.65, respectively, when hydrogens are removed. On the contrary, for the models that use the **II(Hy)** fragments, $R$, $F$, and $s$ values are similar for both sets of molecules.

Nonlinear models which use eqs 2 and 3 have a much larger number of variables than linear ones. Although their $R$ and $s$ characteristics become slightly better compared to those of linear models, the related Fischer's criteria $F$ drops down (Table 4). This shows that eq 1 is more appropriate than eqs 2 and 3 to model log $P$.

For the validation test of log $P$, we used 19 drug molecules from ref 31 for which calculations were performed with the **I(AB**, 2-4), **I(A**, 2-4), and **II(Hy)** fragments (Table 5). Formally, the best fit is found for the **II(Hy)** fragments, which is not surprising since "traditional" models[31,33,65−68] for the log $P$ assessment use additive schemes based on the augmented atoms. However, calculations with the **II(Hy)** fragments fail for six molecules that contain "rare" fragments. For the sequences, the **I(AB**, 2-4) fragment type looks most reasonable: although it cannot be applied for one of molecule (cimetidine), its statistical characteristics for the validation set ($R = 0.933$, $s = 0.62$) are much better those for the **I(A**, 2-4) fragments ($R = 0.866$, $s = 0.80$). Comparison with earlier developed models (Table 5) shows that the **I(AB**, 2-4) fragments yields better results than Suzuki−Kudo's[33] and Rekker's[66] methods. It behaves similarly to Wang's[31] and Moriguchi's[68] approaches, but it is not as good as Hansch−Leo's[65,67] method.

**3.2. Modeling of Complexation and Extraction Properties. 3.2.1. Complexation of Phosphoryl-Containing Podands with the Potassium Cation.** Phosphoryl-containing podands are acyclic ligands containing two terminal phosphine oxide groups connected by a spacer (Chart 1). They possess a remarkable binding affinity for alkali cations, which varies as a function of the nature of constitutive fragments. A set of 84 stability constants (log $\beta$) for the 1:1 complexes of podands and with $K^+$ in the mixed solvent THF:$CHCl_3$ (4:1 volume) at 298 K[46−56] has been selected from the THECOMAC database.[3,7] The learning set excluding rare fragments, contained 70 or 71 podands depending on the decomposition scheme, while the validation set contained 5 podands. The TRAIL program skipped the models that generate more variables ($N$) than the number of compounds ($n$) at the learning stage, or those having small Fischer's criteria. Statistical characteristics of the 18 "best" models ($R > 0.9$) are given in Table 6. Molecular structure and calculated and experimental stability constants for each molecule from the learning set are available in the Supporting Information (Table SM-7). On the basis of statistical criteria,[40] comparison of these models shows that calculations

**Table 5.** Octanol/Water Partition Coefficient: Experimental and Predicted Values for 19 Compounds of the Validation Set[a]

| | compound | experiment[b] | **I**(**AB**, 2-4)[c] | **I**(**A**, 2-4)[e] | **II**(**Hy**)[e] | Lai[b] | Moriguchi[b] | Rekker[b] | Hansch−Leo[b] | Suzuki−Kudo[b] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | atropine | 1.83 | 2.01 | 1.69 | 2.17 | 2.29 | 2.21 | 1.88 | 1.32 | 0.03 |
| 2 | chloramphenicol | 1.14 | 0.67 | 1.62 | *d* | 1.46 | 1.23 | 0.32 | 0.69 | -0.75 |
| 3 | chlocothiazide | −0.24 | −0.82 | −0.96 | −0.65 | −0.58 | −0.36 | −0.68 | −1.24 | −0.44 |
| 4 | chlorpromazine | 5.19 | 5.26 | 4.68 | 5.00 | 4.91 | 3.77 | 5.10 | 5.20 | 3.89 |
| 5 | cimetidine | 0.4 | *d* | 2.17 | *d* | 0.20 | 0.82 | 0.63 | 0.21 | 3.33 |
| 6 | diazepam | 2.99 | 2.80 | 2.97 | 2.69 | 2.98 | 3.36 | 3.18 | 3.32 | 1.23 |
| 7 | diltiazem | 2.70 | 2.62 | 2.10 | *e* | 3.14 | 2.67 | 4.53 | 3.55 | 1.96 |
| 8 | diphenhydramine | 3.27 | 3.33 | 3.46 | 3.04 | 3.74 | 3.26 | 3.41 | 2.93 | 3.35 |
| 9 | flufenamic acid | 5.25 | 3.93 | 4.05 | 4.33 | 4.45 | 3.86 | 5.81 | 5.58 | 5.16 |
| 10 | haloperidol | 4.30 | 4.06 | 4.42 | 4.48 | 4.35 | 4.01 | 3.57 | 3.52 | 3.43 |
| 11 | imipramine | 4.80 | 5.16 | 4.06 | 4.66 | 4.26 | 3.88 | 4.43 | 4.41 | 3.38 |
| 12 | lidocaine | 2.26 | 1.12 | 2.58 | 2.10 | 2.47 | 2.52 | 2.30 | 1.36 | 0.91 |
| 13 | phenobarbital | 1.47 | 1.74 | 1.19 | *d* | 1.77 | 0.78 | 1.23 | 1.37 | 1.29 |
| 14 | phenytoin | 2.47 | 0.98 | 1.67 | *d* | 2.23 | 1.80 | 2.76 | 2.09 | 2.01 |
| 15 | procainamide | 0.88 | 1.19 | 1.32 | 0.84 | 1.27 | 1.72 | 1.11 | 1.11 | 0.65 |
| 16 | propranolol | 2.98 | 2.24 | 2.63 | 2.80 | 2.98 | 2.53 | 3.46 | 2.75 | 2.15 |
| 17 | tetracaine | 3.73 | 3.59 | 3.36 | 3.96 | 2.73 | 2.64 | 3.55 | 3.65 | 2.90 |
| 18 | trimethoprim | 0.91 | 1.02 | 1.71 | 1.17 | 0.72 | 1.26 | −0.07 | 0.66 | 0.57 |
| 19 | verapamil | 3.79 | 4.46 | 5.76 | *d* | 5.29 | 3.23 | 6.15 | 3.53 | 6.49 |
| | *R* | | 0.933 | 0.866 | 0.982 | 0.942 | 0.933 | 0.917 | 0.970 | 0.737 |
| | *s* | | 0.62 | 0.80 | 0.34 | 0.56 | 0.53 | 0.77 | 0.42 | 1.25 |

[a] Calculations are performed using linear model (eq (1). Statistical characteristics *(R* and *s)* for each set are given below the log *P* values. [b] Experimental and calculated by different methods log *P* values are taken from ref 31. [c] Hydrogen atoms are not included. [d] Learning set does not have enough fragments. [e] Molecules are represented with explicit hydrogens.

**Table 6.** Assessment of Stability Constants (log $\beta$) for the Complexes of Phosphoryl Containing Podands with $K^+$ in THF:$CHCl_3$ (4:1 vol.) at 298 K Using Different Types of Fragments[a,b,e]

| no. | fragment type | fitting equation | $N^c$ | *R* | *F* | *s* |
|---|---|---|---|---|---|---|
| 1 | **I**(**AB**, 2-4) | 1 | 38 | 0.986 | 30.2 | 0.32 |
| 2 | **I**(**AB**, 2-3) | 1 | 17 | 0.956 | 35.5 | 0.44 |
| 3 | **II**(**AB**)[d] | 1 | 19 | 0.960 | 33.1 | 0.42 |
| 4 | **I**(**A**, 2-3) | 2 | 27 | 0.958 | 18.9 | 0.47 |
| 5 | **I**(**AB**, 2-2) | 3 | 28 | 0.957 | 17.51 | 0.48 |
| 6 | **I**(**AB**, 2-3) | 2 | 43 | 0.971 | 11.1 | 0.49 |
| 7 | **II**(**Hy**)[d] | 2 | 33 | 0.966 | 16.1 | 0.46 |
| 8 | **II**(**A**)[d] | 2 | 33 | 0.966 | 16.1 | 0.46 |
| 9 | **II**(**AB**)[d] | 2 | 39 | 0.973 | 14.4 | 0.45 |
| 10 | **II**(**Hy**)[d] | 1 | 16 | 0.944 | 29.7 | 0.48 |
| 11 | **II**(**A**)[d] | 1 | 16 | 0.944 | 29.7 | 0.48 |
| 12 | **I**(**A**, 2-2) | 1 | 5 | 0.908 | 77.7 | 0.56 |
| 13 | **I**(**A**, 2-2) | 2 | 9 | 0.927 | 47.1 | 0.52 |
| 14 | **I**(**A**, 2-2) | 3 | 15 | 0.933 | 26.9 | 0.52 |
| 15 | **I**(**A**, 2-3) | 1 | 13 | 0.940 | 36.7 | 0.49 |
| 16 | **I**(**AB**, 2-2) | 1 | 7 | 0.913 | 53.2 | 0.55 |
| 17 | **I**(**AB**, 2-2) | 2 | 13 | 0.924 | 28.2 | 0.54 |
| 18 | **II**(**B**) | 2 | 23 | 0.943 | 17.6 | 0.52 |

[a] See footnotes of Table 4. [b] Only models with $R > 0.9$ are included. [c] The number of fitted coefficients in eqs 1, 2, or 3. [d] Calculations performed for 70 compounds in the learning set. [e] Unless specified, calculations performed for 71 compounds in the learning set.

with **I**(**AB**, 2-4) atom−bond sequences[63] using eq 1 lead to the best statistical characteristics *R* and *s* (0.986 and 0.32, respectively). Models 2−18 can be classified in three main groups presented in Table 6. The first group (models 2−5) corresponds to models resulting in $R > 0.95$ and a reasonably small number of variables compared to the total number of the molecules in the learning set. The largest Fischer's criterion ($F = 35.5$) is obtained with the **I**(**AB**, 2-3) atom−bond sequences including 17 variables. This is close to calculations with the **II(AB**) augmented atoms, including 19 variables ($F = 31.1$). Both decomposition schemes give a reasonable standard deviation $s = 0.42$ (log $\beta$ units), which is similar to typical experimental errors.[46−56]

The second group of models (6−9) mostly involves the calculations within nonlinear fitting equations, eqs 2 and 3, which include larger number of variables compared to the models from the first group. According to statistical criteria,[40] nevertheless, the quality of the calculations for the models from the first and second groups is similar. The third group (models 10−18) involves calculations of relatively low accuracy ($0.90 < R < 0.95$) using the **I (A**, 2-2), **I** (**A**, 2-3), and **I** (**AB**, 2-2) fragments.

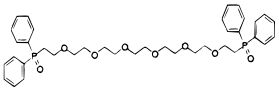
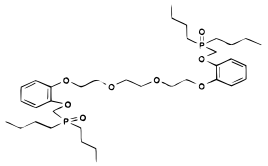
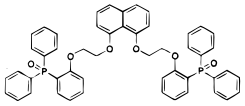
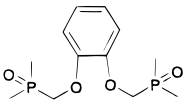
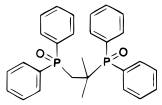
Nonlinear models do not improve the quality of correlations; on the contrary, their Fischer's criterion decreases because of increasing the number of variables (Table 6).

The **I**(**AB**, 2-3) and **I**(**AB**, 2-4) fragments have been used to validate the stability constants of complexation of $K^+$ with five "new" podands that differ both by their terminal groups and by the connecting spacer (Table 7). Log $\beta$ values calculated with both models are rather close to the experimental ones. Standard deviations for these correlations (0.34 and 0.28 for **I**(**AB**, 2-3) and **I**(**AB**, 2-4), respectively) are within experimental error (0.3−0.4).[46−56]

**3.2.2. Complexation of Crown Ethers with the Sodium Cation.** The learning set contains stability constants (log $\beta$) of sodium complexes of 56 crown ethers including unsubstituted macrocycles ($(-CH_2-CH_2-O-)_n$, $n$ = 4-12), 19-crown-6, 20-crown-6, and their benzo, cyclohexyl, and lariate derivatives (Chart 1) taken from reviews of Izatt et al.[6] and Soloviev et al.[7] This set (see Table SM-8 of the Supporting Information) represents a much larger variety of molecules than the set of crown ethers studied by Gakh et al.[30] using a neural network method.

The first attempt to model log $\beta$ values failed because of large deviations between calculated and experimental values for 15- and 18-member rings. We assumed that this discrepancy might be attributed to the "macrocyclic effect",[69,70] when the stabilities of metal complexes are systematically

**Table 7.** Complexation of Phosphoryl-Containing Podands with $K^+$ in THF:$CHCl_3$ (4:1 volume) at 298 K[a]

| No. | Compound | log$\beta$ | | |
|---|---|---|---|---|
| | | Exp. | Predicted | |
| | | | **I**(**AB**, 2-3) | **I**(**AB**, 2-4) |
| 1 |  | 5.3 | 5.40 | 5.54 |
| 2 |  | 4.7 | 4.55 | 4.66 |
| 3 |  | 3.6 | 4.16 | 4.38 |
| 4 |  | 2.54 | 3.22 | 3.03 |
| 5 |  | 1.5 | 1.71 | 2.23 |
| | *R* | | 0.978 | 0.984 |
| | *s* | | 0.34 | 0.28 |

[a] Experimental and "predicted" stability constants (log $\beta$) for the molecules from the validation set.

larger for macrocyclic compounds than that for their open-chain analogues. To take this specificity into account, we introduced a "cyclicity" descriptor *($N_{cycl}$)* as an additional term in eq 1:

$$X = a_0 + \sum_i a_i N_i + a_{cycl} N_{cycl} \qquad (4)$$

In order to test this model, we performed calculations on the subset containing nine unsubstituted crown ethers $(-CH_2-CH_2-O-)_m$, $m = 4-12$.[71] Since these molecules contain the same repetitive unit $(-CH_2-CH_2-O-)$, any fragment decomposition scheme (**I** or **II)** involves linearly dependent fragments, which, therefore, form one extended fragment. Thus, for the given set of unsubstituted crown ethers, eq 4 has only three fitted parameters: $a_0$, $a_l$, and $a_{cycl}$. These calculations resulted in reasonable statistical characteristics ($R = 0.995$, $s = 0.10$, and $F = 294$) for $N_{cycl}$ equal to 2 (for 15-crown-5 and its derivatives), 3 (for 18-crown-6 and its derivatives), and 0 for other compounds. Then, these values of $N_{cycl}$ were used in calculations of 56 molecules from the learning set[63,72] using linear fitting equation eq 4 and nonlinear fitting equations eqs 5 and 6:

$$X = a_0 + \sum_i a_i N_i + \sum_i b_i (2N_i^2 - 1) + a_{cycl} N_{cycl} \qquad (5)$$

$$X = a_0 + \sum_i a_i N_i + \sum_{i,k} b_{ik} N_i N_k + a_{cycl} N_{cycl} \qquad (6)$$

As for podands, the **I**(**AB**, 2-4) atom/bond sequences represent the best linear model ($R = 0.958$, $s = 0.28$, eq 4) for complexes of crown ethers with $Na^+$ (Table 8). Other linear and nonlinear models lead to smaller values of $R$ (Table 8).

The best linear models (**I**(**AB**, 2-4), **I(AB**, 2-3), or **II**(**B**) decomposition schemes used with eq 4) well predict the stability constants for five new crown ethers from the validation set, which differ by the size of the cycle and by substituents (Table 9).

**3.2.3. Complexation of $\beta$-Cyclodextrin with Mono- and 1,4-Disubstituted Benzenes.** Although this work mostly concerns the complexation of metal cations, some efforts have been made to assess stabilities of the complexes of the large macrocyclic $\beta$-cyclodextrin molecule with nonspherical neutral molecular guests (mono- and 1,4-disubstituted benzenes, see Table SM-9 of the Supporting Information). The initial learning set included complexation data for 40 substituted benzenes taken from ref 57. Then, for a given model, this number decreased to 29 owing to exclusion of the molecules containing rare fragments.

Calculations performed with the fitting eq 1 lead to four reasonable models **I**(**AB**, 2-4),[63,72] **I(AB**, 2-5), **I(AB**, 2-3), and **II(Hy)** with similar statistical characteristics (Table 10). Among the different decomposition schemes, **I**(**AB**, 2-3) and **II**(**Hy**) best "predict" the stability constants of the complexes with five molecules of the validation set (Table 11). This conclusion, nevertheless, can be changed if learning and validation sets would contain larger number of molecules.

**3.2.4. Extraction of $UO_2^{2+}$.** Thermodynamics parameters of solvent extraction processes vary as a function of a composition of both liquid (aqueous and nonaqueous) phases including diluents, ligands, metal cations, counterions, and background compounds. To compare extraction constants of a given metal by the series of ionophores, one therefore has to keep all other components of studied extraction systems the same. Unlike complexation processes, there are relatively few extraction data sets available that fit the above requirements. In our calculations, we used a set of 12 effective extraction constants (log $K_{ex}$) for $UO_2^{2+}$ by diphosphine oxide ligands from water to chloroform (Table 12).[13] Three values including rare fragments were discarded from the learning set. In this series, $K_{ex}$ varies by 8 orders of magnitude, which is much larger than for the complexation stability constants studied above.

Since the total number of molecules was not large enough, we were not able to select any molecule for the validation set. The only calculation concerns an assessment of different models for the whole set. Results presented in Table 12 show that the **I**(**AB**, 2-2) and **I**(**AB**, 2-4) sequences used with eq 1[63] give suitable statistical characteristics and reasonably reproduce experimental data.

## 4. DISCUSSION

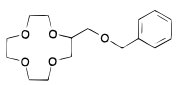
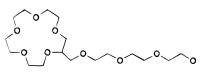
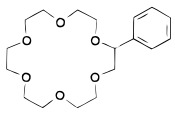
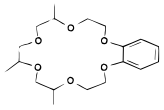
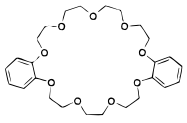
In this section we discuss some specific features of the SMF method compared to other QSPR approaches used earlier for modeling of partition coefficients, complexation, and extraction processes. We have shown that the SMF

**Table 8.** Assessment of Stability Constants (Log $\beta$) for 56 Complexes of Crown Ethers with $Na^+$ in MeOH at 298 K Using Different Types of Fragments[a,b]

| no. | fragment type | fitting equation | $N^c$ | $R$ | $F$ | $s$ |
|---|---|---|---|---|---|---|
| 1 | **I(AB**, 2-4) | 4 | 22 | 0.958 | 17.9 | 0.28 |
| 2 | **II(B)** | 4 | 13 | 0.910 | 17.3 | 0.36 |
| 3 | **I(AB**, 2-3) | 4 | 13 | 0.901 | 15.5 | 0.37 |
| 4 | **II(B)** | 5 | 26 | 0.957 | 13.2 | 0.30 |
| 5 | **I(A**, 2-3) | 5 | 20 | 0.912 | 9.4 | 0.39 |
| 6 | **I(AB**, 2-3) | 5 | 32 | 0.934 | 5.3 | 0.41 |
| 7 | **I(B**, 2-3) | 6 | 37 | 0.916 | 2.7 | 0.52 |

[a] See footnotes of Table 4. [b] Only models with $R > 0.9$ are included. [c] The number of fitted coefficients in eqs 4, 5, or 6.

**Table 9.** Complexation of Crown Ethers with $Na^+$ in MeOH at 298 K[a]

| No. | Compound | Log$\beta$ Exp. | Predicted I(**AB**, 2-4) | Predicted I(**AB**, 2-3) | Predicted II(**B**) |
|---|---|---|---|---|---|
| 1 |  | 1.35 | 1.61 | 1.81 | 1.89 |
| 2 |  | 3.04 | 3.12 | 3.30 | 3.28 |
| 3 |  | 4.17 | 4.12 | 3.93 | 3.89 |
| 4 |  | 3.53 | 3.92 | 3.92 | 3.63 |
| 5 |  | 2.41 | 2.51 | 2.49 | 2.62 |
| | $R$ | | 0.988 | 0.972 | 0.992 |
| | $s$ | | 0.18 | 0.25 | 0.12 |

[a] Experimental and "predicted" stability constants (log $\beta$) for the molecules from the validation set.

**Table 10.** Assessment of Stability Constants (Log $\beta$) for Complexes of $\beta$-Cyclodextrin with 1,4-Disubstituted Benzenes in Water Using Different Types of Fragments[a,b]

| no. | fragment type | model[c] | $n$ | $N^d$ | $R$ | $F$ | $s$ |
|---|---|---|---|---|---|---|---|
| 1 | **I(AB**, 2-4) | 1 | 29 | 24 | 0.909 | 1.0 | 0.71 |
| 2 | **I(AB**, 2-5) | 1 | 27 | 22 | 0.905 | 1.1 | 0.71 |
| 3 | **II(Hy)** | 1 | 29 | 22 | 0.898 | 1.4 | 0.66 |
| 4 | **I(AB**, 2-3) | 1 | 29 | 19 | 0.895 | 2.2 | 0.54 |

[a] See footnotes for Table 1. [b] Only models with $R \geq 0.9$ are included. [c] Equation 1. [d] The number of fitted coefficients in eq 1.

method implemented in the TRAIL program represents a very flexible structure−property modeling tool. Unlike traditional additive schemes that use a limited number of augmented atoms, bonds, or functional groups, TRAIL

**Table 11.** Complexation of $\beta$-Cyclodextrin with Mono- and 1,4-Disubstituted Benzenes (X−$C_6H_4$−Y) in Water[a]

| | X | Y | log $\beta$ exp. | log $\beta$ predicted I(**AB**, 2-4) | II(**AB**, 2-5) | I(**AB**,2-3)I | II(**Hy**) |
|---|---|---|---|---|---|---|---|
| 1 | $NO_2$ | H | 5.63 | 5.30 | 5.30 | 5.21 | 5.45 |
| 2 | I | H | 6.74 | 6.16 | 6.16 | 6.19 | 6.33 |
| 3 | OH | OH | 4.73 | 5.35 | 5.36 | 5.10 | 5.10 |
| 4 | $CH_3CO$ | OH | 5.02 | 5.50 | 5.51 | 5.38 | 5.31 |
| 5 | $CH_3$ | OH | 5.52 | 5.46 | 5.46 | 5.34 | 5.32 |
| | | $R$ | | 0.838 | 0.830 | 0.894 | 0.960 |
| | | $s$ | | 0.22 | 0.22 | 0.22 | 0.16 |

[a] Experimental and "predicted" stability constants (log $P$) for the molecules from the validation set.

**Table 12.** Extraction of $UO_2^{2+}$ by Phosphoryl-Containing Ligands $T_2P(O)CH_2P(O)T_2$ from Water to Chloroform: Assessment of Effective Extraction Constants (log $K_{ex}$) Using Different Types of Fragments[a]

| | T | exp. | **I(AB**, 2-2) | **I(AB**, 2-4) |
|---|---|---|---|---|
| 1 | 4-$CH_3OC_6H_4$ | 9.90 | 8.80 | 8.95 |
| 2 | 3-$CH_3OC_6H_4$ | 8.00 | 8.80 | 8.95 |
| 3 | 4-$CH_3C_6H_4$ | 9.04 | 8.16 | |
| 4 | $C_6H_5$ | 7.83 | 8.71 | 7.83 |
| 5 | 4-$ClC_6H_4$ | 5.08 | 3.67 | 3.67 |
| 6 | 3-$ClC_6H_4$ | 2.26 | 3.67 | 3.67 |
| 7 | 4-$BrC_6H_4$ | 3.70 | 2.87 | 2.87 |
| 8 | 3-$BrC_6H_4$ | 2.04 | 2.87 | 2.87 |
| 9 | *b* | 6.84 | 7.13 | |
| | $n$ | 9 | 9 | 7 |
| | $N^c$ | | 6 | 5 |
| | $R$ | | 0.933 | 0.935 |
| | $F$ | | 4.0 | 3.5 |
| | $s$ | | 1.7 | 1.9 |

[a] See footnotes for Table 1. [b] Ligand $T_2P(O)CH_2P(O)T'_2$ with two different terminal groups $T' = 4\text{-}CH_3OC_6H_4$ and $T'' = 4\text{-}CF_3C_6H_4$. [c] The number of fitted coefficients in eq 1.

generates *several types of fragmentation* of a molecular structure and calculates a given property using each of them together with linear or nonlinear fitting equations, choosing afterward the best model.

For the first time, our method uses atom and/or bond sequences. Although the atom/bond sequences of variable length are widely used for chemical structure search in databases,[73−75] to our knowledge, they were never used for assessment of molecular properties. Unlike molecular fragments used in the CAS database,[73] the SMF method generates the shortest path between each pair of atoms (Figure 1).

We show that long enough sequences work better for complexation and extraction than augmented atoms. For all considered cases of complexation as well as for extraction, the **I**(**AB**, 2-4) fragments lead to the best correlation between calculated and experimental values (Tables 6−11, Figure 4). For the modeling of log $P$ values, this type of fragment is as good as augmented atoms (Tables 4, 5, Figure 3).

The SMF calculations require relatively little computational resources even for large arrays of data. Thus, fitting of log $P$ performed with a linear model involving the **I**(**AB**, 2-4) fragments on the set of 1831 molecules using 156 variables took only 5 min on an IBM PII 350 MHz. This represents a clear advantage compared to the neural networks methods, which are generally very CPU time consuming.[30,76]

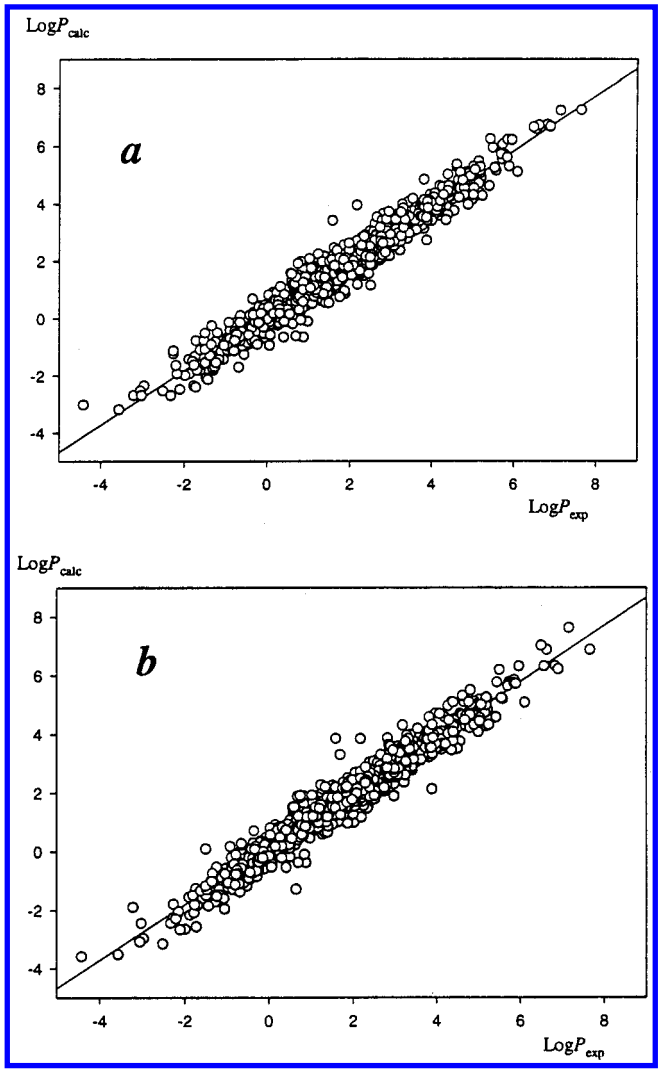



**Figure 3.** Octanol/water partition coefficients (log *P*). Calculated vs experimental values for learning set using (a) **I**(**AB**, 2-4) fragments (log $P_{calc} = 0.08 + 0.95$ 1og $P_{exp}$, $n = 1760$, $R = 0.976$, $F = 35\,949$, $s = 0.32$) and (b) **II**(**Hy**) fragments (log $P_{calc} = 0.09 + 0.95$ 1og $P_{exp}$, $n = 1723$, $R = 0.976$, $F = 34\,204$, $s = 0.32$).

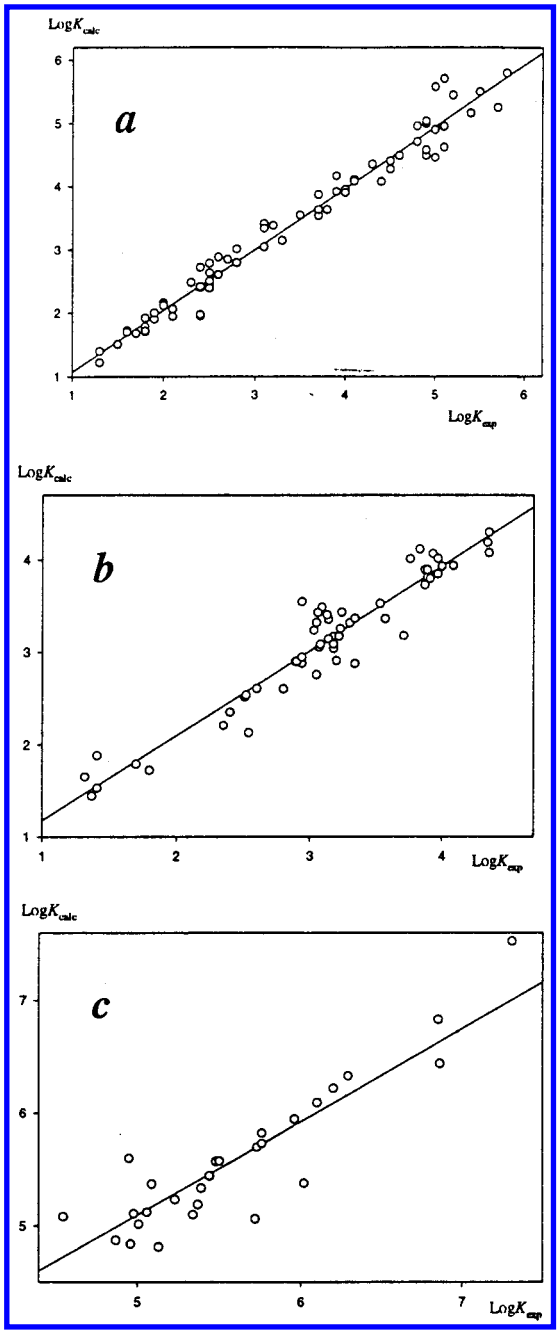



**Figure 4.** Assessment of stability constants (log $\beta$) of 1:1 host−guest complexes in solution at 298 K using the **I**(**AB**, 2-4) model. Calculated vs experimental values for complexation (a) of podands with $K^+$ in THF:$CHCl_3$ (4:1 volume) (log $\beta_{calc} = 0.10 + 0.97$ 1og $\beta_{exp}$, $n = 71$, $R = 0.986$, $F = 2339$, $s = 0.22$), (b) of crown ethers with $Na^+$ in MeOH (log $\beta_{calc} = 0.26 + 0.92$ 1og $\beta_{exp}$, $n = 56$, $R = 0.958$, $F = 597$, $s = 0.21$), and (c) of $\beta$-cyclodextrin with 1,4-disubstituted benzenes in water (log $\beta_{calc} = 0.96 + 0.83$ 1og $\beta_{exp}$, $n = 29$, $R = 0.909$, $F = 129$, $s = 0.28$).

From the results given in Figures 2-4, one may observe for some compounds a substantial deviation of calculated properties from experimental ones (significantly larger than *s*). Such a deviation may not be necessarily related to the defaults of the applied model but to some specific features of a given molecule compared to other molecules in the learning set. In such a case, additional descriptors could be developed in order to improve the structure−property modeling. Thus, in this work, introduction of a cyclicity descriptor allowed us to take the macrocyclic effect[69,70] into account and to reasonably assess the stabilities of $Na^+$ complexes of crown ethers.

Generally speaking the development of descriptors incorporating 3D structural information is expected to further improve the structure−property relationships, as recently shown by log *P* calculations of Wang et al.,[31] who considered specific intramolecular interactions. We believe that development of new types of descriptors including 3D characteristics of molecular structures combined with atom/bond sequences could extend the SMF approach.

In fact, the current version of the TRAIL program allows one to model host−guest interactions in solution only for a given series of hosts (or guests), when all other components of the studied systems (partners of interactions, solvent, background compounds, etc.) are the same. A further extension of the SMF method might treat systems including different host and guest species, solvents, or other components.

Finally we would like to emphasize that our method and developed software tools may already serve as an expert system to assess the complexation and extraction data stored in databases and, thus, may be used to design new com-

pounds with desired complexation or extraction properties.

## CONCLUSIONS

A substructural molecular fragment (SMF) method has been developed to model physical and chemical properties of organic molecules. It is based on the splitting of a molecule into fragments, and on the calculations of their contributions to a given property. It uses two different types of fragments: atom/bond sequences or "augmented atoms" (atoms with their nearest neighbors). The structure−properties relationships can be fitted using additive and additive−multiplicative models. The TRAIL program has been developed to perform such calculations, first, on the learning set to select an appropriate model, and then on the validation set to "predict" properties of the compounds.

The method was successfully tested on classical systems such as (i) boiling point, molar volume, molar refraction, heat of vaporization, surface tension, melting point, critical temperature, and critical pressures of $C_2-C_9$ alkanes and (ii) octanol/water partition coefficients of organic molecules.

For the first time an approach based on molecular fragments is applied in the field of complexation and extraction properties of organic molecules. We modeled (i) the complexation of alkali cations with crown ethers and phosphoryl-containing podands, and of $\beta$-cyclodextrins with mono- and 1,4-disubstituted benzenes, and (ii) extraction of the uranyl cation by phosphoryl-containing ligands. For the molecules from the validation sets, these properties were reproduced within the experimental accuracy.

As the SMF method uses limited computational resources, assessment of complexation or extraction constants can be easily performed with a large variety of ligands in complex conditions.

## ACKNOWLEDGMENT

The authors thank N. Strakhova, V. Kazachenko, and L. Solov'eva for the help with preparation of the THECOMAC database and Dr. P. Jost for the help with the statistical treatment of the calculated results and for fruitful discussions. Professor Lai and Dr. Wang are acknowledged for delivering for us experimental and structural data on partition coefficients of octanol/water. V.P.S. gratefully acknowledges W.G. for the opportunity to have spent a term in 1999 as an Invited Professor at the Laboratory MSM, ULP, Strasbourg.

**Supporting Information Available:** Tables of physical properties of C2−C9 alkanes, fragment contributions to octanol/water partition coefficients, to log $\beta$, and to log $K$, and experimental and calculated stability constants for various complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) The NIST Critically Selected Stability Constants of Metal Complexes Database. Version 5.0; http://www.nist.gov/srd/nist46.htm.
(2) IUPAC Stability Constants Database. http://www.acadsoft.co.uk/.
(3) The THECOMAC (thermodynamics on complexation of macrocycles) database, developed by Dr. V. P. Soloviev, contains about 10 000 records on thermodynamics of complexation of metal cations with macrocyclic ligands and their analogues in solutions.
(4) Wagman, D. D. Data Bases: Past, Present and Future. *Pure Appl. Chem.* **1992**, *64*, 37−48.
(5) Cooper, S. R. *Crown Compounds: Towards Future Applications;* VCH: New York, Weinheim, Cambridge, 1992.
(6) Izatt, R. M.; Pawlak, K.; Bradshaw, J. S.; Bruening, R. L. Thermodynamic and Kinetic Data for Macrocycle Interaction with Cations and Anions. *Chem. Rev.* **1991**, *91*, 1721−2085.
(7) Solov'ev, V. P.; Vnuk, E. A.; Strakhova, N. N.; Raevsky, O. A. *Thermodynamics of Complexation of the Macrocyclic Polyethers with Salts of Alkali and Alkaline-Earth Metals* (in Russian); VINITI: Moscow, 1991.
(8) Gokel, G. W. *Crown Ethers and Cryptands;* The Royal Society of Chemistry: Cambridge, 1991.
(9) *Cation Binding by Macrocycles*; Inoue, Y., Gokel, G. W., Eds.; Marcel Dekker: New York, 1990.
(10) Lindoy, L. F. *The Chemistry of Macrocyclic Ligand Complexes;* Cambridge University Press: New York, 1989.
(11) *Comprehensive Supramolecular Chemistry, Vol. 1; Molecular Recognition: Receptors for Cationic Guests;* Atwood, J. L., Devies, J. E. D., MacNicol, D. D., Vögtle, F., Lehn, J.-M., Eds.; Pergamon Press: New York, 1998.
(12) Izatt, R. M.; Bradshaw, J. S.; Nielsen, S. A.; Lamb, J. D.; Christensen, J. J.; Sen, D. Thermodynamic and Kinetic Data for Cation−Macrocycle Interaction. *Chem. Rev.* **1985**, *85*, 271−339.
(13) *Handbook on Solvent Extraction* (in Russian); Rozen, A. M., Ed.; ATOMIZDAT: Moscow, 1976; Vols. 1−3.
(14) Martinov, B. V. *Extraction by Organic Acids and their Salts* (in Russian); ENERGOATOMIZDAT: Moscow, 1989.
(15) Mejov, E. A. *Handbook on Extraction by Amines and Quaternary Ammonium Bases* (in Russian); ENERGOATOMIZDAT: Moscow, 1987.
(16) Nikolotova, Z. I. *Handbook on Extraction of Lanthanides and Actinides by Neutral Extractants* (in Russian); ENERGOATOMIZDAT: Moscow, 1987.
(17) Mejov, E. A. *Handbook on Extraction by Amines and Quaternary Ammonium Bases* (in Russian); ENERGOATOMIZDAT: Moscow, 1999.
(18) Nikolotova, Z. I. *Extraction of Lanthanides and Actinides by Neutral Extractants* (in Russian); ENERGOATOMIZDAT: Moscow, 1999.
(19) Leach, A. R. *Molecular Modeling. Principles and Applications;* Longman: Singapore, 1996.
(20) Beck, M.; Nagypal, I. *Chemistry of Complex Equilibria;* Akademiai Kiado: Budapest, 1989.
(21) Christensen, J. J.; Izatt, R. M. *Handbook of Metal Ligand Heats*; Marcel Dekker: New York, 1983.
(22) Rozen, A. M.; Krupnov, B. V. Dependence of the Extraction Ability of Organic Compounds on their Structure. *Russ. Chem. Rev.* (in Russian) **1996**, *65*, 1052−1079.
(23) Shi, Z. G.; McCullough, E. A., Jr. A Computer Simulation−Statistical Procedures for Predicting Complexation Equilibrium Constants. *J. Inclusion Phenom.* **1994**, *18*, 9−26.
(24) Schneider, H. J.; Rudiger, V.; Raevsky, O. A. The Incremental Description of Host−Guest Complexes: Free Energy Increments Derived from Hydrogen Bonds Applied to Crown Ethers and Cryptands. *Org. Chem.* **1993**, *58*, 3648−3653.
(25) Raevsky, O. A.; Grigor'ev, V. Y.; Solov'ev, V. P. The Estimation of Electron-Donor and -Acceptor Functions of Ionic Groups in Biologically Active Compounds on the Base of Thermodynamic Data. *Khim. Farma. Zh.* (in Russian) **1984**, 578−582.
(26) Raevsky, O. A.; Grigor'ev, V. Y.; Solov'ev, V. P. Modeling Structure−Activity Relationship. II. The Estimation Electron-Donor and -Acceptor Functions of Active Centres in the Molecules of Physiologically Active Compounds. *Khim. Farma. Zh.* (in Russian) **1989**, 1294−1300.
(27) Varnek, A. A.; Glebov, A. S.; Kuznetsov, A. N. Charge Density Distribution, Electrostatic Potential and Complex Formation Ability of Some Neutral Agents. *Portugal Phys.* **1988**, 59−61.
(28) Varnek, A. A.; Kuznetsov, A. N.; Petrukhin, O. M. Electrostatic Potential Distribution and Extraction Ability of Some Organophosphorus Compounds. *J Struct. Chem.* (in Russian) **1989**, *30*, 44−48.
(29) Varnek, A. A.; Kuznetsov, A. N.; Petrukhin, O. M. Calculation of the Indices of Extractability of Some Neutral Organophosphorus Compounds within the Framework of Electron Density Functional Method. *Coord. Chem.* (in Russian) **1991**, *17*, 1038−1043.
(30) 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. Inclusion Phenom.* **1997**, *27*, 201−213.
(31) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
(32) Chou, J. T.; Jurs, P. C. Computed assisted computation of partition coefficient from molecular structures using fragment constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172−178.
(33) Suzuki, T.; Kudo, Y. Automatic Log P Estimation Based on Combined Additive Modeling Methods. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 155−198.

(34) Estrada, E. Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 3. Molecules Containing Cycles. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23−27.
(35) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids;* McGraw-Hill Book Co.: New York, 1977.
(36) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(37) *Sybyl 6.3 Molecular Modeling Software*; Tripos Associates, Inc., St. Louis, MO.
(38) Swamy, M. N. S.; Thulasiraman, K. *Graphs, Networks, and Algorithms*; John Wiley & Sons: New York, 1981.
(39) Forsythe, G. E.; Malcolm, M. A.; Moler, C. B. *Computer Methods for Mathematical Computations*; Prentice Hall, Inc.: Englewood Cliffs: NJ, 1977.
(40) Kendall, M. G.; Stewart, A. *The Advanced Theory of Statistics*; Griffui: London, 1966; Vols. 1−3.
(41) Seber, G. A. *Linear Hypothesis: A General Theory*; Hafner: New York, 1966.
(42) Thurrott, P.; Brent, G.; Bagdazian, R.; Tendon, S. *DELPHI 3. Superbible*; Waite Group Press, A Division of Sams Publishing: Corte Madera, CA, 1997.
(43) Swan, T. *Foundations of DELPHI Development for WINDOWS 95;* IDG Books Worldwide, Inc: Key West, FL, 1995.
(44) Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194.
(45) *ISIS Draw Software*; Molecular Design Ltd.: San Leandro, CA.
(46) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Safronova, Z. V.; Bondarenko, N. A.; Tsvetkov, E. N. Phosphorus-Containing Podands. 12. Effect of Alkyl and Phenyl Substituents Near Phosphorus Atoms on the Complexating Ability of Neutral Monopodands—Anomalous Alkyl Effect. *Zh. Obsch. Khim.* (in Russian) **1995**, *65*, 223−231.
(47) Bovin, A. N.; Evreinov, V. I.; Safronova, Z. V.; Tsvetkov, E. N. Phosphorus-Containing Podands. 11. Synthesis of Bis(Ortho-Diphenylphosphinyl)Benzyl Ethers of Oligoethylene Glycols and Their Complexing Properties with Respect to Alkali-Metal Cations. *Russ. Chem. Bull.* (in Russian) **1993**, *42*, 912−916.
(48) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Tsvetkov, E. N. Phosphorus-Containing Podands. 10. An Improved Method for Synthesizing Oligo(Ethylene Glycol) Bis(2-(Diphenylphosphinoyl)-Ethyl) Ethers and Their Complex-Forming Properties with Respect to Alkali-Metal-Cations in Anhydrous Acetonitrile. *Russ. Chem. Bull.* (in Russian) **1993**, *42*, 472−477.
(49) Baulin, V. E.; Evreinov, V. I.; Vostroknutova, Z. N.; Bondarenko, N. A.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands. 9. Synthesis of Oligoethylene Glycol Bis(Diphenylphosphinylethyl) Esters and Their Complexing Properties with Respect to Alkali-Metal Cations in a Low-Polarity Solvent. *Bull. Acad. Sci. USSR, Chem. Sci.* (in Russian) **1992**, *41*, 914−918.
(50) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Safronova, Z. V.; Krashakova, I. B.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands. 7. Complexing Properties of Ortho-Diphenylphosphinyl-Substituted Diphenyl Ethers of Oligoethylene Glycols with Respect to Alkali-Metal Cations. *Bull. Acad. Sci. USSR, Chem. Sci.* (in Russian) **1991**, *40*, 1759−1766.
(51) Evreinov, V. I.; Vostroknutova, Z. N.; Bovin, A. N.; Degtyarev, A. N.; Tsvetkov, E. N. Phosphorus-Containing Podands—Structure of Terminal Groups and Complexing Ability. *Zh. Obsch. Khim.* (in Russian) **1990**, *60*, 1506−1511.
(52) Bovin, A. N.; Evreinov, V. I.; Vostroknutova, Z. N.; Tsvetkov, E. N. Effect of a Catechol Segment in a Polyether Chain on the Complexing Ability of Some Phosphonate and Quinoline Monopodands. *Bull. Acad. Sci. USSR, Chem. Sci.* (in Russian) **1989**, *38*, 2398−2401.
(53) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Bondarenko, N. A.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands. 4. Effect of Polyether Chain-Length of Oligoethylene Glycol Bis(Ortho-Diphenylphosphinylmethyl)Phenyl Ethers on Their Complex-Forming and Selective Properties with Respect to Alkali-Metal Cations. *Bull. Acad. Sci. USSR, Chem. Sci.* (in Russian) **1989**, *38*, 1828−1834.
(54) Evreinov, V. I.; Degtyarev, A. N.; Bovin, A. N.; Tsvetkov, E. N.; Vostroknutova, Z. N. Phosphorus-Containing Podands. 3. Effect of the Length of the Polyether Chain of Bis(Orthophenyl(Diethoxyphosphinylmethoxy)Phenyl) Ethers of Oligoethylene Glycols on Their Complexing Ability Towards Alkali-Metal Action. *Bull. Acad. Sci. USSR, Chem. Sci.* (in Russian) **1989**, *38*, 50−53.
(55) Evreinov, V. I.; Tsvetkov, E. N.; Syundyukova, V. K.; Vostroknutova, Z. N.; Baulin, V. E.; Bondarenko, N. A. Phosphorus-Containing Podands—Effect of the Structure of Phosphoryl-Containing Fragments of Monoethyleneglycol Diesters on Their Complexability Towards Cations of Alkali-Metals. *Zh. Obsch. Khim.* (in Russian) **1989**, *59*, 67−72.
(56) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands—Effect of the Length of Polyester Chain of Bis(Ortho-(Diethoxyphosphoryl)-Phenyl)Ethers of Oligoethyleneglycol on Their Complexability towards Cations of Alkali-Metals. *Zh. Obsch. Khim.* (in Russian) **1989**, *59*, 73−77.
(57) Liu, L.; Guo, Q. X. Wavelet Neural Network and Its Application to the Inclusion of $\beta$-Cyclodextrin with Benzene Derivatives. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 133−138.
(58) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 2636.
(59) Bonchev, D. *Information Theoretical Indices for Characterisation of Chemical Structure;* Wiley-Interscience: New York, 1983.
(60) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(61) Seybold, P. G.; May, M. A.; Bagal, U. A. Molecular Structure−Property Relationships. *J Chem. Educ.* **1987**, *64*, 575-581.
(62) Randic, M.; Basak, S. C. Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf: Comput. Sci.* **1999**, *39*, 261−266.
(63) Fragment contributions are given in the Supporting Information.
(64) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Bilology;* Wiley: New York, 1979.
(65) Leo, A. L. Calculating Log *P* from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(66) Rekker, R. F. *The Hydrophobic Fragment Constant;* Elsevier: New York, 1977.
(67) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525−616.
(68) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, L.; Matsushita, Y. Simple Method of Calculating Octanol/Water Partition Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(69) Cabbines, D. K.; Margerum, D. W. Macrocyclic Effect on the Stability of Copper(II) Tetramine Complexes. *J. Am. Chem. Soc.* **1969**, *91*, 6540−6541.
(70) Tsivadze, A. Y.; Varnek, A. A.; Khutorsky, V. E. *Coordination Compounds of Metals with Crown-Ligands* (in Russian); Nauka: Moscow, 1991.
(71) Inoue, Y.; Liu, Y.; Tong, L. H.; Ouchi, M.; Hakushi, T. Complexation Thermodynamics of Crown Ethers. Part 3. 12-Crown-4 to 36-Crown-12: from Rigid to Flexible Ligand. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 1947−1950.
(72) Calculated and experimental stability constants for each molecule from the learning set are available in the Supporting Information.
(73) Fisanick, W.; Lipkus, A. H.; Rusinko III, A. Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf.: Comput. Sci.* **1994**, *34*, 130−140.
(74) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.
(75) Domokos, L. Beilstein Ring Search System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 663−667.
(76) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. A Neural Device for Searching Direct Correlations between Structures and Properties of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715−721.

CI9901340