"In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library, and Experimental Tests

A. Varnek* and D. Fourches

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

V. P. Solov'ev and V. E. Baulin

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

A. N. Turanov

Institute of Solid State Physics, Russian Academy of Sciences,
142432 Chernogolovka, Moscow Region, Russia

V. K. Karandashev

Institute of Microelectronics Technology and High Purity Materials,
142432 Chernogolovka, Moscow Region, Russia

D. Fara and A. R. Katritzky

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida,
Gainesville, Florida 32611

Received January 7, 2004

This paper is devoted to computer-aided design of new extractants of the uranyl cation involving three main steps: (*i*) a QSPR study, (*ii*) generation and screening of a virtual combinatorial library, and (*iii*) synthesis of several predicted compounds and their experimental extraction studies. First, we performed a QSPR modeling of the distribution coefficient (log*D*) of uranyl extracted by phosphoryl-containing podands from water to 1,2-dichloroethane. Two different approaches were used: one based on classical structural and physicochemical descriptors (implemented in the CODESSA PRO program) and another one based on fragment descriptors (implemented in the TRAIL program). Three statistically significant models obtained with TRAIL involve as descriptors either sequences of atoms and bonds or atoms with their close environment (augmented atoms). The best models of CODESSA PRO include its own molecular descriptors as well as fragment descriptors obtained with TRAIL. At the second step, a virtual combinatorial library of 2024 podands has been generated with the CombiLib program, followed by the assessment of log*D* values using developed QSPR models. At the third step, eight of these hypothetical compounds were synthesized and tested experimentally. Comparison with experiment shows that developed QSPR models successfully predict log*D* values for 7 of 8 compounds from that "blind test" set.

1. INTRODUCTION

Solvent extraction is a widely used technique for selective separation and concentration of metals in biphasic water/organic solvent systems. It involves a cation−ligand complexation in one of the liquid phases or at the liquid/liquid interface, accompanied by transfer of the complexes into bulk organic phase. Development of new extraction systems with desirable properties generally proceeds in empirical manner because of complexity of studied processes. Indeed, thermodynamic parameters of extraction depend on many variables (the nature of metal(s), conterion(s), ligand(s), pH, organic solvent, and background compounds), and, therefore, their theoretical modeling represents a very difficult task.

In fact, in silico design of new extraction systems with desired characteristics could be possibly based on an informational system involving (*i*) a comprehensive database, (*ii*) an expert system which models quantitative structure−property relationships (QSPR), and (*iii*) a generator of combinatorial libraries. Figure 1 illustrates links between these modules: experimental data collected in the database are treated by the expert system which establishes relationships between structure of compounds and their extraction properties. Then, structure−property models are applied to screen a virtual combinatorial library leading to potential

* Corresponding author phone: 33-390-241549; fax: 33-390-241545; e-mail: varnek@chimie.u-strasbg.fr.

10.1021/ci049976b CCC: $27.50 © 2004 American Chemical Society
Published on Web 07/26/2004

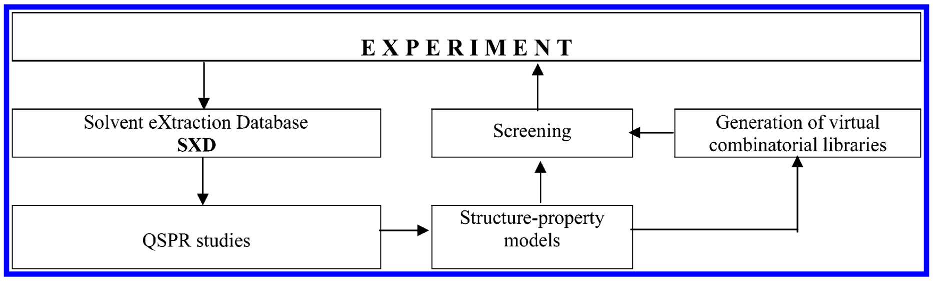
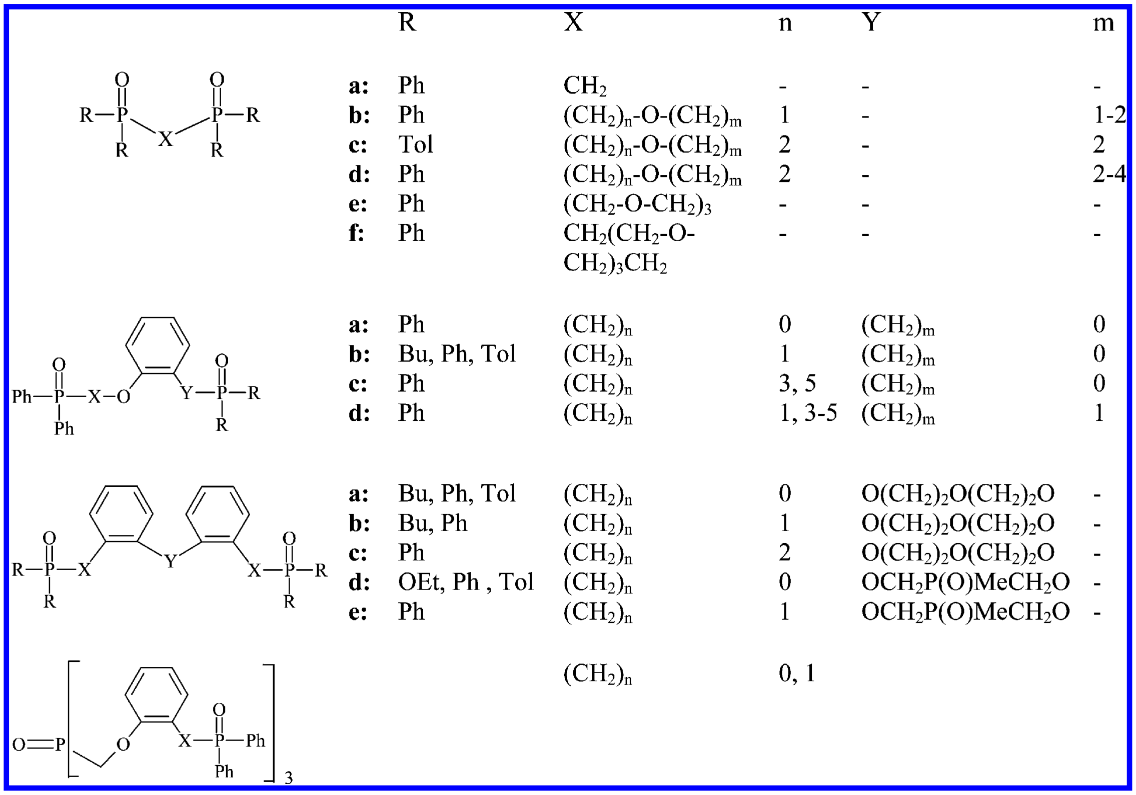

EXPERIMENT
Solvent eXtraction Database
SXD
QSPR studies
Structure-property models
Screening
Generation of virtual combinatorial libraries

Figure 1. The strategy applied for “in silico” design of new ligands for solvent extraction.



	R	X	n	Y	m
a:	Ph	CH_2	-	-	-
b:	Ph	$(CH_2)_n$-O-$(CH_2)_m$	1	-	1-2
c:	Tol	$(CH_2)_n$-O-$(CH_2)_m$	2	-	2
d:	Ph	$(CH_2)_n$-O-$(CH_2)_m$	2	-	2-4
e:	Ph	$(CH_2$-O-$CH_2)_3$	-	-	-
f:	Ph	$CH_2(CH_2$-O-$CH_2)_3CH_2$	-	-	-
a:	Ph	$(CH_2)_n$	0	$(CH_2)_m$	0
b:	Bu, Ph, Tol	$(CH_2)_n$	1	$(CH_2)_m$	0
c:	Ph	$(CH_2)_n$	3, 5	$(CH_2)_m$	0
d:	Ph	$(CH_2)_n$	1, 3-5	$(CH_2)_m$	1
a:	Bu, Ph, Tol	$(CH_2)_n$	0	$O(CH_2)_2O(CH_2)_2O$	-
b:	Bu, Ph	$(CH_2)_n$	1	$O(CH_2)_2O(CH_2)_2O$	-
c:	Ph	$(CH_2)_n$	2	$O(CH_2)_2O(CH_2)_2O$	-
d:	OEt, Ph , Tol	$(CH_2)_n$	0	$OCH_2P(O)MeCH_2O$	-
e:	Ph	$(CH_2)_n$	1	$OCH_2P(O)MeCH_2O$	-
		$(CH_2)_n$	0, 1		

Figure 2. Phosphoryl-containing podands studied in this work.

actives which then can be experimentally tested, in turn providing the database with new information.

An expert system is an important element of the informational system, since the reliability of predictions of new compounds depends on robustness of structure−property models. To our knowledge, in the literature there are very few publications concerning QSPR studies of extraction systems. Thus, equilibrium constants of actinide elements extracted by some neutral phosphoryl-containing ionophores were correlated with descriptors such as group electronegativities or Taft parameters of molecular fragments,[1] atomic charges,[1,2] energies of 1s-orbitals of oxygen atoms,[1] electrostatic potential distribution,[2,3] chemical softness, and donor−acceptor interaction energies.[4] A regression model based on charge parameters (a sum of the point charges on atoms of amidic function) and calculated association energy for the metal−ligand complexes has been used by Rabbe et al.[5] for the modeling of log*D* of U(VI) extracted by monoamides. Voelkel and Szymanowski[6] applied the distance-based topological indexes and connectivity indexes to build QSPR models of the partition data of Cu extracted by various homologues of hydroxyoximes from acidic sulfate solutions. Yoshizuka et al.[7] have found a relationship linking the complexation strain energy difference between the lanthanide cations and La(III) with their relative extractability. The *Substructural Molecular Fragments* (SMF) method has been applied to assess extraction constants for the complexes of uranyl cation with phosphoryl-containing ligands[8] and distribution coefficients for Hg, In, and Pt extracted by phosphoryl-containing monopodands, for uranyl extracted by monoamides.[9]

The goal of this work is computer-aided design of new phosphoryl-containing podands which efficiently extract the uranyl cation from water to the organic solvent.

Phosphoryl-containing podands (Figure 2) are acyclic molecules with polyether spacer(s) linking two (in monopodands), three (in dipodands and tripodands), or four (in tripodands) terminal phosphine oxide groups.[10] Variation of the length of the spacers, of molecular topology, and of the substituents at the phosphorus atoms may lead to desirable complexation[11−13] or extraction[14−17] properties of podands for selected metal cations. In particular, these molecules extract the uranyl cation forming 1:1 complexes[15] whose structures were not studied so far. Analysis of the data from the Solvent eXtraction Database[9] shows that the set of distribution coefficients for uranyl extraction for 32 podands[15] (Figure 2) used in our study is one of the largest available in the literature data sets where extraction properties of the compounds were studied strictly at the same conditions. Some preliminary results on QSPR modeling of extraction properties of these podands were reported in ref 9.

This study involves three main steps: (*i*) QSPR modeling of the logarithm of the distribution coefficient (log*D*) of UO_2^{2+} which quantitatively measures extraction efficiency of ligands, (*ii*) generation, enumeration, and screening of a virtual combinatorial library, and (*iii*) synthesis of theoretically predicted compounds and their experimental extraction studies.

Two different QSPR approaches were used in this work: the Hansch-type approach[18] which uses as descriptors some physicochemical parameters calculated either by quantum mechanical methods or by some empirical techniques and the Free-Wilson-type approach[19] which uses molecular fragments as variables in a multiple regression analysis. The first type of calculations was performed with the CODESSA-PRO program which applies up to 902 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors. The calculations with fragment descriptors (atom/bond sequences or augmented atoms) were performed using Substructural Molecular Fragments method implemented into the TRAIL program.[8,9] These descriptors are derived solely from molecular structure and do not require experimental data or expensive theoretical calculations to be obtained.

Here we show that the joint application of these two different techniques (TRAIL and CODESSA-PRO) is a promising way to improve the robustness of predictions. Moreover, a mixed set of fragment (from TRAIL) and "classical" (from CODESSA PRO) descriptors leads to QSPR models statistically more significant than those obtained by TRAIL or CODESSA PRO alone.

Derived by TRAIL and CODESSA PRO structure-property models were applied to 8 compounds selected from generated in this study virtual combinatorial library. Then, those compounds were synthesized, and the log*D* values for them were measured using the same protocol as for initial data set of 32 podands. Comparison of theoretical and experimental log*D* values for new podands shows that our QSPR models reasonably predict log*D* values for 7 compounds from 8.

2. METHOD

2.1. Data Preparation. Distribution coefficients (log*D*) of uranyl cation extracted from 2 M HNO_3 aqueous solution in 1,2-dichloroethane by podands (0.01 M) at 291 ± 2 K were taken from ref 15. Structure-property models were built both using a data set containing 32 compounds, and two different training/test sets were selected from that parent set. To prepare test sets, we followed recommendations of Oprea et al.:[20] (*i*) experimental methods for determination of activities in the training and test sets should be similar; (*ii*) the activity values should span several orders of magnitude but should not exceed activity values in the training set by more than 10%; and (*iii*) the balance between active and inactive compounds should be respected for uniform sampling of the data. Here, each test set contained 3 compounds, i.e. about 10% of the compounds from the corresponding parent set. The molecules in a particular test set differ completely from other sets (Table 1). Other data from the initial parent set were included in the corresponding training sets (Table 1).

2.2. Computations. *2.2.1. Calculations with TRAIL.* The TRAIL program is based on the *Substructural Molecular Fragments* (SMF) method[8,9] which involves the splitting of a molecular graph into fragments (subgraphs) followed by calculations of their contributions to a given property $\boldsymbol{Y}$ (here $\boldsymbol{Y} = \log D$). Two different classes of fragments are used: "sequences" (**I**) and "augmented atoms" (**II**). Three subtypes **AB**, **A**, and **B** are defined for each class. For the fragments **I**, they represent sequences of atoms and bonds (**AB**), of atoms only (**A**), or of bonds only (**B**). The number of atoms in these sequences varies from 2 to 6, and only shortest paths from one atom to the other are used. An "augmented atom" represents a selected atom with its 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. Totally, TRAIL generates 49 types of fragments including 45 different atoms/bonds, atoms only or bonds only sequences, and 4 types of augmented atoms.

Once a molecular graph is split into constitutive fragments, any quantitative physical or chemical property $\boldsymbol{Y}$ is calculated from the fragments contributions using linear (1) or nonlinear (2) and (3) fitting equations

$$\boldsymbol{Y} = a_o + \sum_i a_i N_i \tag{1}$$

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

$$\boldsymbol{Y} = a_o + \sum_i a_i \mathrm{N}_i + \sum_{i,k} b_{ik} N_i N_k \tag{3}$$

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

At the training stage, TRAIL builds up to 147 structure-property models involving 3 linear and nonlinear fitting equations and 49 types of fragment descriptors. Molecules containing fragments of "rare" occurrence (i.e. found in less than 2 molecules) are excluded from the training set. If some fragments are linearly dependent, they are treated as one extended fragment. Using the singular value decomposition method,[21] TRAIL fits the a_i and b_i terms in eqs 1-3, calculates corresponding statistical characteristics (correlation coefficient (*R*), standard deviation (*s*), Fischer's criterion (*F*), cross-validation correlation coefficient (*Q*), Kubyni's criterion (*FIT*), R_H-factor of Hamilton and matrix of pair correlations (covariation matrix) for the terms a_i and b_i), and performs statistical tests[22] to select the best models. On the validation stage, TRAIL calculates the "predicted" values using the fitted fragments contributions for the "best" models.

A significant advantage of the SMF method is the possibility to select during the training stage several best fit models (instead of a single QSPR model) related to different fragmentation schemes combined with three fitting equations. Using selected QSPR models, one can calculate average activities of the compounds from the test set, which smooth inaccuracies of particular individual models, thus improving the robustness of predictions.[23]

Table 1. Modeling of Distribution Coefficients (log*D*) of U(VI) Extracted by Phosphoryl Containing Podands in 1,2-Dichloroethane: Experimental and Calculated log*D* Values for the Full Data Set[a]

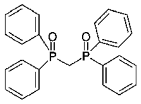
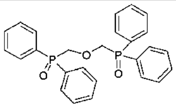
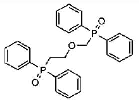
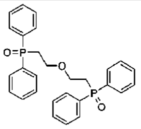
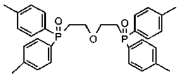
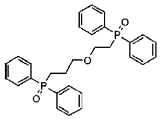
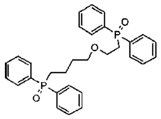
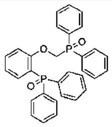
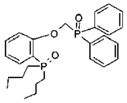
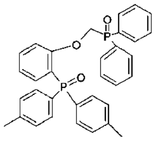
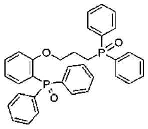
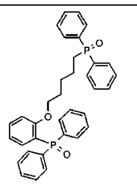
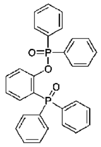
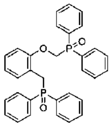
no.	Compound	Exp.	TRAIL[b]		CODESSA PRO[b]	
			Calc	Exp - Calc	Calc	Exp - Calc
1		2.58	–[c]	–[c]	2.59	-0.01
2		0.10	–[c]	–[c]	0.20	-0.10
3		0.51	0.81	-0.30	0.58	-0.07
4[e]		0.75	0.67	0.08	0.59	0.16
5		1.05	1.21	-0.16	0.86	0.19
6		0.85	0.61	0.24	0.52	0.33
7		0.45	0.55	-0.10	0.61	-0.16
8		0.84	0.53	0.31	0.84	0.00
9		0.95	1.14	-0.19	0.89	0.06
10[d]		1.20	0.80	0.40	0.99	0.21
11[e]		0.33	0.33	0.00	0.33	0.00
12		0.42	0.21	0.21	0.28	0.14
13[e]		0.47	0.53	-0.06	0.41	0.06
14		0.67	0.64	0.03	0.23	0.44

Table 1 (Continued)

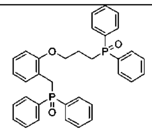
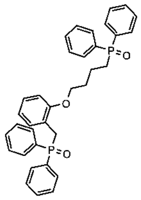
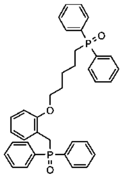
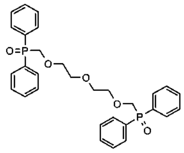
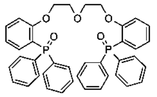
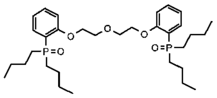
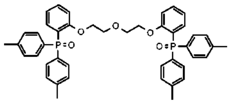
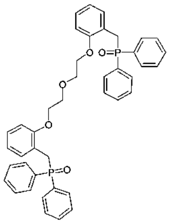
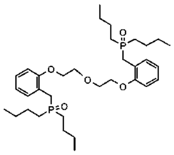
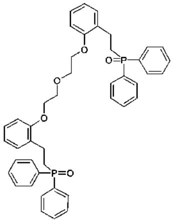
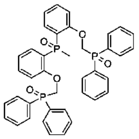
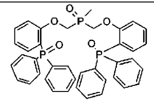
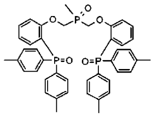
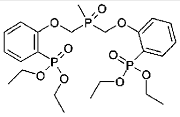
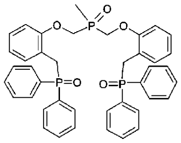
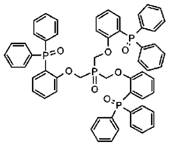
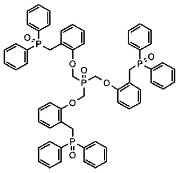
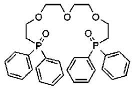
no.	Compound	Exp.	TRAIL [b]		CODESSA PRO [b]	
			Calc	Exp - Calc	Calc	Exp - Calc
15		0.02	0.43	-0.41	0.32	-0.30
16		0.26	0.37	-0.11	0.14	0.12
17		0.26	0.31	-0.05	0.19	0.07
18		0.31	0.36	-0.05	0.26	0.05
19		-0.87	-0.47	-0.40	-0.43	-0.44
20		0.62	0.74	-0.12	0.89	-0.27
21		0.30	0.07	0.23	0.35	-0.05
22 [d]		-0.20	-0.27	0.07	0.03	-0.23
23		1.20	0.94	0.26	1.20	0.00
24		0.65	- [c]	- [c]	0.36	0.29
25		1.98	1.48	0.50	1.90	0.08

Table 1 (Continued)

no.	Compound	Exp.	TRAIL[b]		CODESSA PRO[b]	
			Calc	Exp - Calc	Calc	Exp - Calc
26		1.42	1.48	-0.06	1.51	-0.09
27		1.67	2.02	-0.35	1.86	-0.19
28		0.16	-[c]	-[c]	0.29	-0.13
29[d]		1.72	1.68	0.04	1.53	0.19
30		1.16	1.27	-0.11	1.16	0.00
31		1.62	1.58	0.04	1.68	-0.06
32		0.09	0.08	0.01	0.37	-0.28

[a] Experimental data were taken from ref 15. [b] Average over three (TRAIL) or two (CODESSA PRO) best models fitted on the full set. [c] Compound was excluded by TRAIL at the training stage because it contains unique molecular fragment(s). [d] Compounds belonging to the test set 1. [e] Compounds belonging to the test set 2.

Structures and experimental data for podands have been retrieved from the Solvent eXtraction Database (SXD)[9] and exported into SD file used as an input for TRAIL.

2.2.2. Calculations with CODESSA PRO. CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) PRO is a comprehensive program for developing quantitative structure/property relationships. The RC3 version of the program used in this work integrates all necessary mathematical and computational tools to (*i*) calculate a large variety of molecular descriptors from the 3D structure of chemical compounds; (*ii*) develop (multi)linear QSPR models of the chemical, physical or biological properties of individual compounds; and (*iii*) predict properties for compounds previously unknown or unavailable. The applicability of CODESSA PRO methodology to various QSAR/QSPR problems has been convincingly demonstrated in refs 24−26.

The 2D structures of the podands were imported as MOL files. The molecular geometries were optimized using AM1 Hamiltonian calculations together with eigenvector following geometry optimization procedure available in the quantum chemical program MOPAC 7.05 implemented in the CODESSA PRO package. The gradient norm 0.01 kcal/Å was applied to the geometry optimization.

The best multilinear regression (BMLR) procedure was used to find the best correlation models from selected noncollinear descriptors. The BMLR selects the best two-parameter regression equation, the best three-parameter regression equation etc., based on the highest R^2 value in the stepwise regression procedure. During the BMLR procedure the descriptor scales are normalized and centered automatically, and the final result is given in natural scales.

The optimal number of descriptors in QSPR model was determined from the plot of statistical criteria of models (R^2 and Q^2) vs the number of descriptors involved. As a rule, addition of one more descriptor improves the statistical criteria, but starting with one point ("break point") that improvement is negligible. Consequently, the model corresponding to the break point is considered as the best/optimum model.

2.2.2.1. Preselection of Pertinent Descriptors of CODESSA PRO. For a given molecule, its 3D structure generated by CODESSA PRO from a MOL file corresponds to one of its low energy conformers. For a flexible molecule, possessing many low energy conformers, this may provoke an ambiguity with the calculation of some descriptors varying with geometry and, consequently, with the developed QSPR models (if no reference structure is identified or no alignment or energy requirements are specified).

To select descriptors either conformationally invariant or weakly depending on molecular conformations,[27] a series of test calculations was performed on 30 low energy conformers of a typical podand (molecule 32, Table 1). The 3D structures of those conformers were generated by means of Monte Carlo conformational search option of the MacroModel 5.5 program[28] using MMFF force field, followed by their energy optimization with semiempirical AM1 method incorporated in the SPARTAN 5.1.3 program.[29] Resulting structures were imported in CODESSA PRO (RC3 version) in order to calculate molecular descriptors for each conformers. Quantitatively, the relative variation for *i*th descriptor as a function of conformation was estimated as $\delta d_i = \Delta d_i / \langle d_i \rangle$, where $\langle d_i \rangle$ was the average value of the given descriptor for all 30 conformations and Δd_i was a standard deviation. So, 262 descriptors for which $\delta d_i < 0.1$ were finally selected for QSPR modeling.

2.2.3. Development of Predictive QSPR Models. The key question of any QSPR study is related to reliability of developed structure−property models. Here, the following strategy has been applied in order to improve the robustness of our models for log*D*.

1. Recently, we have shown[23,30] that simultaneous utilization of several models (instead of a single one) at the validation stage improves the predictability of property calculations. Indeed, any particular model involving a relatively large number of variables may lead to excellent statistical parameters for the training set but poor correlation with experimental data at the validation stage. An average data set resulted from several “best” QSPR models, smoothes inaccuracies of data calculated with individual models. Therefore, we select several QSPR models from calculations both with TRAIL and CODESSA PRO.

2. At the training stage, TRAIL prepares the list of models sorted according to R^2*(fitting)*. Then, several best fit models are selected using user’s defined threshold (typically, R^2*(fitting)* > 0.8).

3. The BMLR option of CODESSA PRO produces only one model for a given number (*m*) of descriptors involved. Keeping the same *m*, different models can be obtained in calculations on the full set and selected training sets.

4. The reliability of the best models selected at the training stage with TRAIL or CODESSA PRO are examined according to the rules suggested by Golbraikh and Tropsha:[31] (*i*) a Leave One Out cross-validation correlation coefficient $Q^2 > 0.5$; (*ii*) the correlation coefficient for the linear regression $\mathbf{Y_{calc}}$ vs $\mathbf{Y_{exp}}$, $R^2 > 0.6$; (*iii*) the correlation coefficients R_0^2 or $R'_0{}^2$ should be close to R^2, i.e., $[(R^2 - R_0^2)/R^2] < 0.1$ or $[(R^2 - R'_0{}^2)/R^2] < 0.1$; and (*iv*) the slopes *k* and *k*′ should be between 0.85 and 1.15. Here, R_0^2 and $R'_0{}^2$ are, respectively, the correlation coefficients for linear regressions $\mathbf{Y_{calc}} = k\mathbf{Y_{exp}}$ (R_0^2) or $\mathbf{Y_{exp}} = k'\mathbf{Y_{calc}}$ ($R'_0{}^2$) passing through the origin, whereas *k* and *k*′ are slopes of those regressions.

5. To check the statistical stability of a given model, a set of descriptors selected at the fitting stage for a given data set, was applied for the fitting on other data sets (e.g., the set of *k* descriptors selected by CODESSA PRO or TRAIL upon fitting the model on the full data set was further applied for the fitting on the training sets 1 and 2). The model is stable if (*i*) Q^2 and R^2 are large enough for any fitting and (*ii*) a variation δa_i of the coefficients a_i (b_i) is comparable to the standard deviations Δa_i obtained in different fitting procedures.

Table 2. TRAIL Calculations: Statistical Criteria of Selected Models[a]

fragment set	*n*	*k*	R^2	*s*	*F*	R_H, %	Q^2	s_{PRESS}
				Full Set				
II(**B**)	28	8	0.850	0.29	16.2	25.6	0.712	0.40
II(**A**)	27	12	0.876	0.30	9.6	23.3	0.702	0.47
I(**AB**, 2−3)	27	12	0.876	0.30	9.6	23.3	0.702	0.47
				Training Set 1				
II(**B**)	25	8	0.847	0.28	13.5	25.6	0.642	0.44
II(**A**)	24	12	0.869	0.31	7.2	23.8	0.590	0.55
I(**AB**, 2−3)	24	12	0.869	0.31	7.2	23.8	0.590	0.55
				Training Set 2				
II(**B**)	25	8	0.847	0.31	13.5	26.0	0.692	0.44
II(**A**)	25	12	0.876	0.32	8.4	23.4	0.668	0.53
I(**AB**, 2−3)	25	12	0.876	0.32	8.4	23.4	0.668	0.53

[a] Experimental distribution ratio log*D* of U(VI) extracted by phosphoryl-containing podands in 1,2-dichloroethane are taken from ref 15. Statistical parameters calculated for the training set: the number (*n*) of points (compounds), the number of fitted coefficients (*k*) in eq 1, correlation coefficient (*R*), standard deviation (*s*), Fisher’s criterion (*F*), factor of Hamilton (R_H), cross-validation correlation coefficient (*Q*), and cross-validation standard deviation (s_{PRESS}).

6. The models retained at the previous step are applied to “predict” the property of the compounds from test sets. The most predictive models selected at that stage can be further applied for screening of virtual libraries, either individually or all together in order to calculate average values of the modeled property.

2.2.4. Generation of Virtual Combinatorial Libraries. The *CombiLib* program has been prepared to generate virtual combinatorial libraries based on Markush structures.[32] It uses its own *EdChemS* editor of 2D structures allowing user to prepare the molecular core as MOL file, to select the attachment points and to prepare collections of substituents as SD file. Attachment of substituents to the molecular core can be performed by either connecting two atoms belonging to the two fragments (“atom−atom attachment”), by overlapping two bonds of these fragments (“bond−bond attachment”), or by inclusion of an atom of the core into a cyclic substituent (“spiro attachment”).

3. RESULTS

3.1. TRAIL Calculations. At the preliminary stage, TRAIL excluded from the data set 4 molecules containing fragments of “rare” occurrence (i.e. found in less than 2 molecules). Calculations on the full set containing 28 molecules and on the training sets 1 and 2 containing 25 molecules resulted in three statistically significant linear models based on the **II**(**B**) and **II**(**A**) fragments which represent augmented atoms involving respectively only bonds or only atoms of their environment, and the **I**(**AB**, 2−3) fragments which represent a sequence of atoms and bonds from 2 to 3 atoms. Fitting the models on the compounds of the full set led to statistical criteria (R^2*(fit)* = 0.850−0.876, *s* = 0.29−0.30, Q^2 = 0.702−0.712, s_{PRESS} = 0.40−0.47) slightly better than those obtained for training sets 1 and 2 (R^2*(fit)* = 0.847−0.876, *s* = 0.28−0.32, Q^2 = 0.590−0.692, s_{PRESS} = 0.44−0.55) (Table 2). For all those three models,

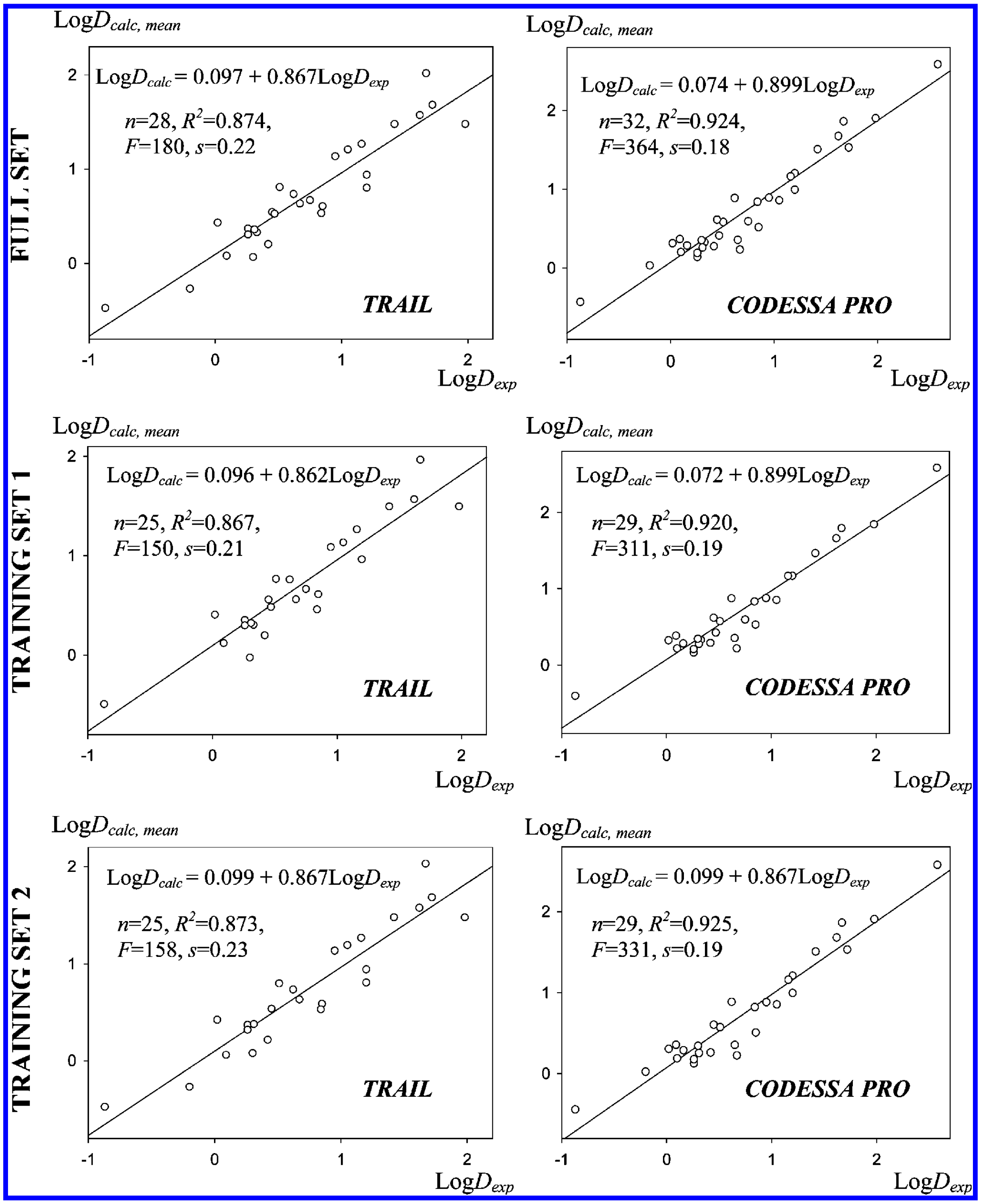


Figure 3. Correlation between experimental and calculated log*D* values for the full set (*top*), training set 1 (*middle*), and training set 2 (*bottom*). The log*D* values were calculated as an average of values obtained by TRAIL with **II(B)**/eq 1, **II(A)**/eq 1, and **I(AB**, 2−3)/eq 1 models (*left*) or obtained by CODESSA PRO using **FS** and **TS1** models (*right*).

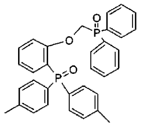
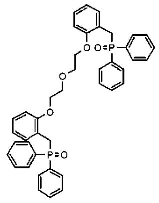
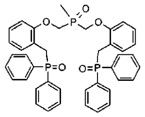
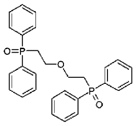
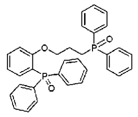
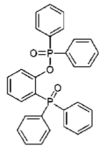
we have calculated the correlation coefficients for the linear regression logD_{calc} vs logD_{exp} (R^2) and for those passing through the origin log$D_{\mathrm{calc}} = k$ logD_{exp} (R_0^2) and log$D_{\mathrm{exp}} = k'$ logD_{calc} (R'^2_0) as well as slopes of those correlations (k and k', respectively). It has been found that the $[(R^2 - R_0^2)/R^2]$ and $[(R^2 - R'^2_0)/R^2]$ values do not exceed 0.015 and corresponding k and k' values are of 0.94 and 1.003, respectively. This shows that the three models correspond to the robustness criteria of Golbraikh and Tropsha.[31]

An average values of log*D* calculated over three best fit models correlate well with the experimental data for the compounds of the full set as well as those of training sets 1 and 2 ($R^2 = 0.867-0.874$, $s = 0.21-0.23$, Figure 3).

Calculations on the test sets 1 and 2 show a good convergence of selected TRAIL models: the standard deviation s_{av} for average log*D* values varies from 0.01 to 0.17 (Table 3), which is smaller than standard deviations of fitting. Both log*D* values calculated with particular models individually and those averaged over all three models well correlate with experimental data ($R^2 = 0.843-0.978$, $s = 0.04-0.44$, Table 3).

3.2. CODESSA PRO Calculations. The calculations on the full set were performed using 3D structures of podands as they were generated from the MOL files. No statistically significant models were obtained when 262 “own” descriptors of CODESSA PRO were used. Therefore, 16 fragment descriptors corresponding to the **I**(**AB**, 2−3)/eq 1 model obtained with TRAIL were added as external descriptors. Calculations with that new set of 278 descriptors resulted in several reasonable models involving from 2 to 10 descriptors.

Table 3. Experimental and Predicted by TRAIL log*D* Values for Compounds of the Two Test Sets[a]

		Log*D*				
			predicted			
no.	compound	exp.	**II**(**B**)	**II**(**A**)	**I**(**AB**, 2-3)	*mean* [b]
			Test Set 1			
1		1.20	0.69	0.70	0.70	0.70 (0.01)
2		-0.20	-0.42	-0.22	-0.22	-0.29 (0.12)
3		1.72	1.51	1.80	1.80	1.70 (0.17)
	R^2 [c]		0.971	0.907	0.907	0.931
	s [c]		0.23	0.44	0.44	0.37
			Test Set 2			
1		0.75	0.76	0.58	0.58	0.64 (0.10)
2		0.33	0.35	0.30	0.30	0.32 (0.03)
3		0.47	0.54	[d]	[d]	0.54
	R^2 [c]		0.978			0.843
	s [c]		0.04			0.09

[a] Experimental data are taken from ref 15. Calculations are performed using linear eq 1. [b] Average log*D* value calculated from the three best fit models; the standard deviation s_{av} is given in parentheses. [c] Statistical characteristics (*R* and *s*) for the correlation between experimental and predicted log*D* values. [d] Fragment contributions for this compound were not available at the training stage.

Variation of R^2 and Q^2 as a function of the number of descriptors allowed us to fix the "break point" corresponding to the model involving 6 descriptors (Figure 4). This number well corresponds to the "rule of thumb" in QSAR/QSPR studies which states that the number of compounds per descriptor should not be smaller than 5.[33]

Having fixed the optimal number of descriptors as 6, the calculations were performed on the compounds of training sets 1 and 2. This led to two models (**TS1** and **TS2** for the training sets 1 and 2, respectively) different compared to that (**FS**) obtained in calculations on the full set (Table 4). Among six descriptors involved in the **FS** model, two descriptors (*Relative Positive Charge* and *Maximal antibonding contribution of one MO*) could be directly related to intermolecular interactions,[34] two descriptors describe chemical bonds in podands (*Maximal valency for atom C* and *Maximal σ−σ bond order*), and two fragment descriptors (*the number of C−P−C* and *the number of P−C$_{ar}$−C$_{ar}$ fragments*). In the **TS1** model, two descriptors are common with the **FS** model (*Maximal valency for atom C* and *the number of C−P−C fragments*), and the others (*Average bond order for atom P*, *Maximal bond order for atom H*, *Minimal atomic state energy for atom O* and *Minimal valency for atom H*) characterize electronic structure of studied molecules. Five of six descriptors involved in the **TS2** model are common either with the **FS** or **TS1** model; the only different from the above models descriptor is *Minimal (>0.1) bond order for atom O* (Table 4).

To check statistical stability of the **FS**, **TS1**, and **TS2** models, a set of descriptors selected at the training stage for a given data set was applied for the fitting on another data set. Here, the **FS** descriptors were used for the fitting on the

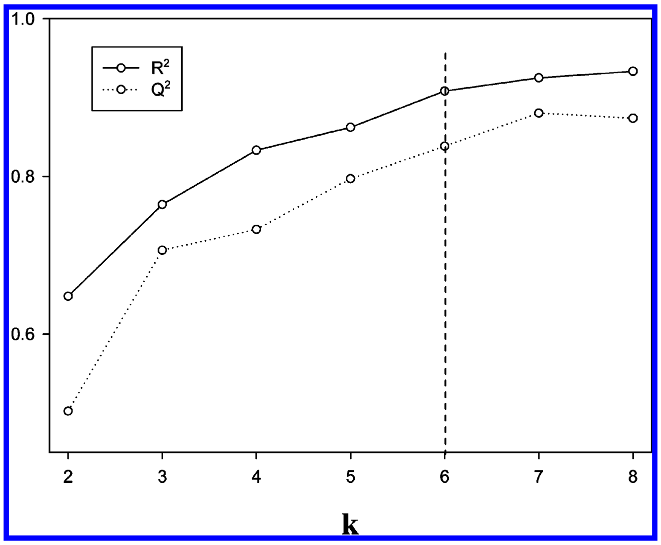


Figure 4. CODESSA-PRO calculations: statistical parameters of fitting (R^2 and Q^2) as a function of the number of descriptors involved in the QSPR models.

Table 4. CODESSA-PRO Calculations: Descriptors (d_i), Fitted Coefficients (a_i), and Their Standard Deviations (Δa_i) for Multilinear Regression log$D = a_0 + \Sigma_{i=1}^{6} a_i d_i$ Corresponding to **FS**, **TS1**, and **TS2** Models

i	d_i	a_i	Δa_i
	Model **FS**		
0		−34.73	14.48
1	number of fragment C−P−C	0.52	0.05
2	max valency for atom C	1.17	0.13
3	max $\sigma-\sigma$ bond order	60.99	11.01
4	relative positive charge	10.16	1.67
5	max antibonding contribution of one MO	17.45	4.43
6	number of fragment $P-C_{ar}-C_{ar}$	−0.08	0.02
	Model **TS1**		
0		−203.65	40.10
1	number of fragment C−P−C	0.24	0.03
2	max valency for atom C	0.74	0.20
3	av bond order for atom P	12.90	3.06
4	max bond order for atom H	52.64	8.63
5	min atomic state energy for atom O	0.4655	0.1111
6	min valency for atom H	−7.76	2.30
	Model **TS2**		
0		−51.98	11.59
1	number of fragment C−P−C	0.19	0.02
2	max valency for atom C	1.18	0.13
3	max bond order for atom H	55.41	6.83
4	av bond order for atom P	21.71	2.97
5	min (>0.1) bond order for atom O	5.76	0.96
6	max antibonding contribution of one MO	14.73	4.34

Table 5. CODESSA-PRO Calculations: Statistical Criteria of Selected 6 Descriptors' Models[a]

no.	QSPR models	*n*	R^2	*s*	*F*	Q^2
			Full Set			
1	***FS***[b]	*32*	*0.908*	*0.24*	*41.3*	*0.838*
2	**TS1**[c]	32	0.890	0.26	33.7	0.785
3	**TS2**[c]	32	0.783	0.36	15.0	0.279
			Training Set 1			
4	**FS**[c]	29	0.896	0.25	31.6	0.796
5	***TS1***[b]	*29*	*0.903*	*0.24*	*34.0*	*0.747*
6	**TS2**[c]	29	0.812	0.34	15.9	0.360
			Training Set 2			
7	**FS**[c]	29	0.914	0.24	38.9	0.865
8	**TS1**[c]	29	0.889	0.27	29.5	0.731
9	***TS2***[b]	*29*	*0.931*	*0.22*	*49.8*	*0.849*

[a] See footnotes for Table 1. [b] At the first step, the **FS**, **TS1**, and **TS2** models were obtained by fitting, respectively, on the full set and training sets 1 and 2. Those models involve six descriptors selected by CODESSA PRO among 278 descriptors available. [c] At the second step, the fitting on a given data set was performed using only six descriptors obtained at the first step for other data sets. For example, the descriptors involved in the **FS** model were applied for fitting on the training sets 1 and 2.

training sets 1 and 2, whereas the **TS1 (TS2)** descriptors were applied to fit models for the compounds of the full set and for the training set 2 (training set 1). Thus, three different sets of descriptors were applied for each data set. Statistical criteria of multilinear correlations (Table 5) show that the **TS2** descriptors do not provide the user with statistically significant models when applied for the full set or training set 1 ($Q^2 = 0.279$ and 0.360, respectively). On the other hand, correlations involving the **FS** and **TS1** descriptors have reasonable statistical criteria ($R^2 = 0.889-0.914$, $s = 0.24-0.27$, $Q^2 = 0.731-0.865$). Statistically, both models are stable with respect to variation of the number or type of compounds in the data set used for the fitting since only small variations of the coefficients at the variables in multilinear equations was observed while fitted on the full set or on the training sets 1 and 2. Average values of log*D* calculated over **FS** and **TS1** models correlate well with the experimental data for the compounds of a full set as well as those of the training sets 1 and 2 ($R^2 = 0.920-0.925$, $s = 0.18-0.19$, Figure 3).

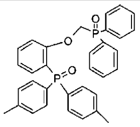
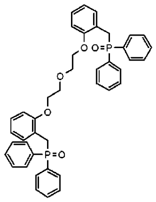
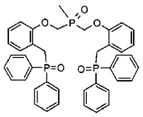
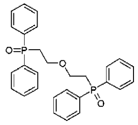
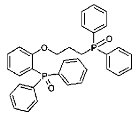
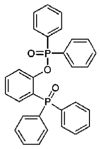
For all selected models, $[(R^2 - R_0^2)/R^2]$ and $[(R^2 - R'_0{}^2)/R^2]$ values are less than 0.007 and corresponding k and k' values are of about 0.95 and 1.01, respectively, just showing that they meet requirements of robustness suggested in ref 31.

Being applied for the estimation of log*D* values of the compounds from the test sets 1 and 2, the models involved **FS** and **TS1** descriptors provide a good correlation between experimental and calculated log*D* ($R^2 = 0.901-0.998$, $s = 0.01-0.27$, Table 6). The values obtained by averaging of log*D* calculated with those models also well correlate with the experiment ($R^2 = 0.981-0.998$, $s = 0.01-0.14$, Table 6).

3.3. "In Silico" Design of New Podands: Generation of the Combinatorial Library, Hits Selection, and Experimental Tests. A challenge for any QSPR study is related to prediction ability of structure−property models when they are applied for compounds which were not included in the parent data set. Selection of the candidate molecules for such a "blind test" was performed in four steps: (*i*) choice of the molecular core and preparation of its "optimal" substituents, (*ii*) generation of combinatorial library, (*iii*) evaluation of activity of virtual compounds, and (*iv*) hits selection.

The combinatorial library has been generated for the Markush structure $R_1R_2R_3P{=}O$ (Figure 5) for which a set of "optimal" substituents R_1, R_2, and R_3 was prepared using fragment contributions a_i calculated by TRAIL for the three best fit models on the training stage. To build R_1, R_2, and R_3, we have selected molecular fragments (Figure 5) most of which correspond to large possible positive a_i. Selection of the "optimal" substituents is illustrated in Figure 6. Here, when replacing the H atom with the Me group in the aromatic ring of a monopodand (Figure 6), for the linear **II**(**A**)/eq 1

Table 6. Experimental and Predicted by CODESSA-PRO log*D* Values for Compounds of the Two Test Sets

		Log*D*			
			predicted		
no.	compound	exp.	**TS1**	**FS**	*mean*[b]
			Test Set 1		
1		1.20	0.74	1.11	0.93 (0.26)
2		-0.20	0.19	-0.11	0.04 (0.21)
3		1.72	1.42	1.59	1.51 (0.12)
	R^2 [a]		*0.901*	*0.998*	*0.981*
	s [a]		*0.27*	*0.02*	*0.14*
			Test Set 2		
1		0.75	0.67	0.49	0.58 (0.13)
2		0.33	0.44	0.18	0.31 (0.18)
3		0.47	0.55	0.27	0.41 (0.20)
	R^2 [a]		*0.973*	*0.998*	*0.998*
	s [a]		*0.03*	*0.01*	*0.01*

[a] Statistical characteristics (*R* and *s*) for the correlation between experimental and predicted log*D* values. [b] Average log*D* value calculated from the two best fit models; the standard deviation s_{av} is given in parentheses.

model, TRAIL determines four new fragments: two C(C, C, C) and two C(C) augmented atoms whose contributions are 2*0.036 and 2*0.028, respectively. This replacement leads also to the disappearance of two C(C, C) augmented atoms whose contribution 2*(−0.063) is negative (Table 7). Thus, the overall effect is 2*0.036 + 2*0.028 − 2*(−0.063) = 0.25 log*D* units, which corresponds to the increase of the extraction ability (Figure 6). In such a way, most of the candidates for R_1, R_2, and R_3 were suggested (Figure 5), then a small virtual combinatorial library of 2024 podands has been generated. Then, activity of each compound in the library was evaluated by TRAIL using three best fit models (**II**(**A**), **II**(**B**), and **I**(**AB**, 2−3)) followed by calculations of the average data set for these models.

Among the variety of potential candidates, we selected eight molecules which span the range of log*D* variation for experimentally studied molecules (log*D* = −0.05−3.05) and estimated their extraction ability using the best models of CODESSA PRO. Then, those podands were prepared and experimentally studied as uranyl extractants using the same protocol as in previous studies of the compounds from the parent set.[15] Thus, one can speak about a *real blind test*, since prediction calculations were performed before the experimental studies were done.

Comparison of experimental log*D* values with those predicted by TRAIL (Table 8) and CODESSA PRO (Table 9) is given in Figure 7. One can see that both QSPR approaches reasonably estimate log*D* for 7 from 8 compounds from the “blind test” set. Theoretical calculations have overestimated log*D* only for the $R_3P{=}O$, R = $CH_2-O-C_6H_4-P(O)Bu_2$, podand: experimental value (1.58) is about twice smaller than those predicted by TRAIL (3.10, Table 8) and by CODESSA PRO (2.85, Table 9). We attribute that result to poor population (3 of 32) of the

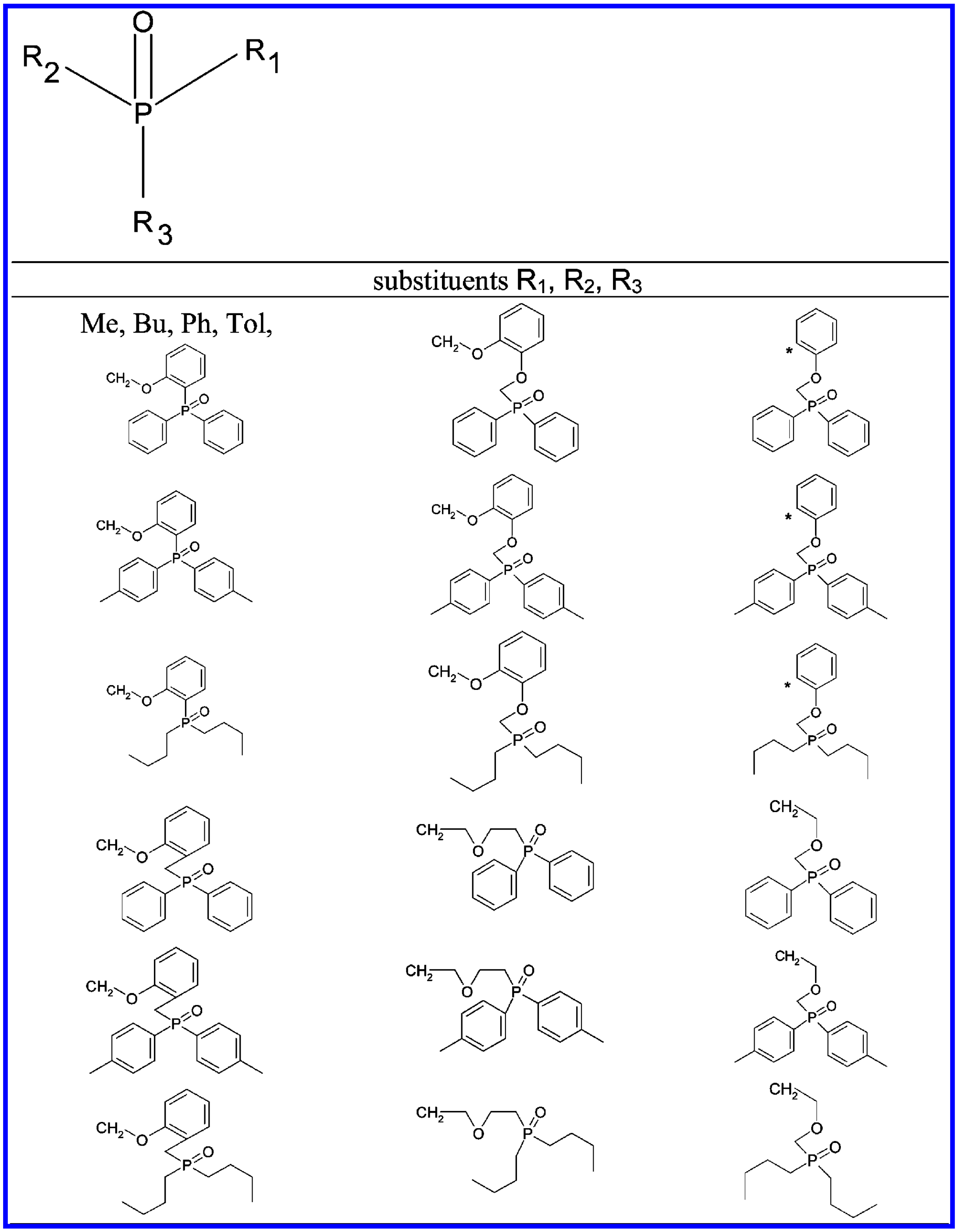


Figure 5. Molecular core and the set of selected substituents used to generate the focused combinatorial library containing 2024 virtual mono-, di-, and tripodands. Connecting atoms are indicated as * or CH_2.

molecules having butyl substituents at the phosphorus atom in the data set used.

Generally, one can notice a good convergence between the log*D* values calculated for the blind test compounds by TRAIL and CODESSA PRO: on average, they differ by about 0.3 log*D* units (Figure 7, Tables 8 and 9), which correspond to standard deviations of the models obtained at the fitting stage.

It should be noted, that we have also selected several hits whose hypothetical activities are larger than the largest activity ($\log D_{exp} = 2.58$, Table 1) observed experimentally. Actually, one of us (V.B.) is working on the synthesis of those suggested molecules.

4. DISCUSSION

4.1. Preselection of Descriptors for QSPR Modeling of Metal Complexation Involving Flexible Ligands. CODESSA PRO is a powerful structure−property tool operating with up to 902 descriptors, most of which vary as a function of molecular geometry. However, not all of those descriptors can be used for the modeling of flexible metal binders. Indeed, a 3D structure generated from a MOL file corresponds to one of the low energy conformers of a given molecule. If no reference structure is used or no alignment nor energy requirements are specified, any conformer (of many possible conformers) can be generated from the MOL file. This leads to an ambiguity with the 3D descriptors which

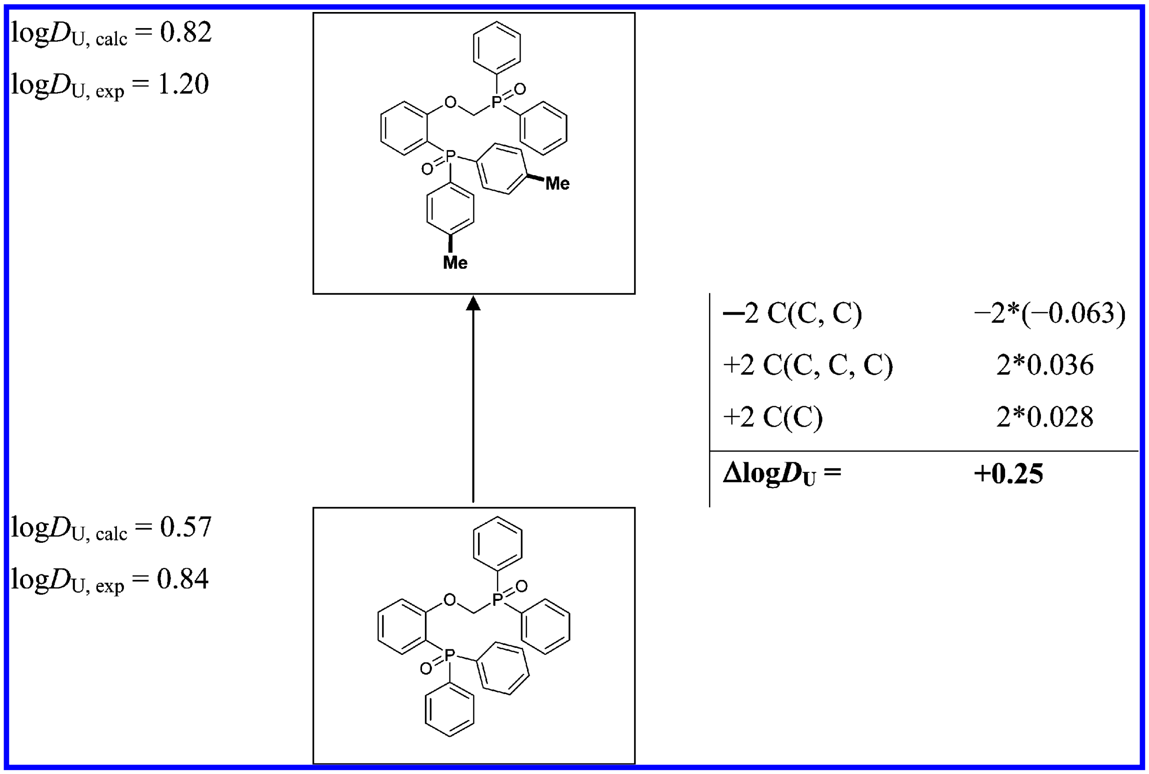

$\log D_{U,\,calc} = 0.82$
$\log D_{U,\,exp} = 1.20$
O
P
O
P
O
Me
Me
−2 C(C, C) −2*(−0.063)
+2 C(C, C, C) 2*0.036
+2 C(C) 2*0.028
$\Delta\log D_U$ = +0.25
$\log D_{U,\,calc} = 0.57$
$\log D_{U,\,exp} = 0.84$
O
P
O
P
O

Figure 6. Variation of log*D* due to an attachment of Me groups into aromatic fragments. Calculations are performed by TRAIL using fragment contributions for the **II**(**A**)/eq 1 model (see Table 7).

Table 7. Fragment Contributions (a_i) to Distribution Ratio log*D* for the Linear **II**(**A**)/Eq 1 Model

no.	fragment[c]	contribution a_i
1	P(C, C, C, O); O(P)[a]	0.371
2	C(C, P)	0.240
3	C(C, O)	−0.161
4	O(C, C)	0.015
5	C(O, P)	0.256
6	C(C, C, P)	0.158
7	C(C, C)	−0.063
8	C(C, C, C)	0.036
9	C(C)	0.028
10	C(C, C, O)	−0.065
11	C(P)	0.461
12	a_o [b]	0.351

[a] Linearly dependent fragments in training set form one group as an extended fragment. [b] The a_o term is fragment independent. [c] For a given central atom, the list of its neighboring atoms is given in parentheses.

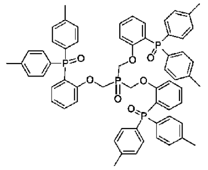
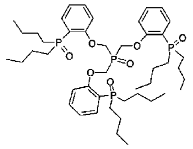
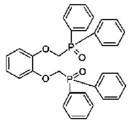
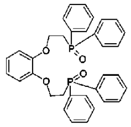
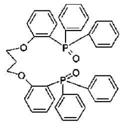
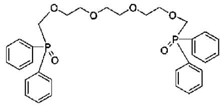
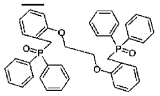
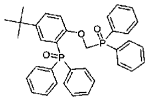
more or less vary from one conformer to another. To select the descriptors which weakly vary with conformation, an empirical threshold corresponding to 10% of variation of the average descriptor's value has been introduced. Accordingly, in section 2.2.2.1, all 555 descriptors were calculated for 30 low energy conformers of the typical podand 32 (Table 1) from which 262 descriptors have been selected.

Nevertheless, a preselection based on the conformational analysis of a single molecule may not be sufficient to exclude conformationally dependent descriptors. Thus, we have performed a new series of CODESSA PRO calculations, using the 3D structures in which some torsional angles in the linkers between P═O groups were modified compared to the initial geometry. Unlike reported in section 3.2, the calculations with 262 CODESSA's descriptors resulted in a statistically significant model (R^2 = 0.893, R_{cv}^2 = 0.807, s = 0.25) involving six descriptors: *Average Bonding Information content (order 0)*, *Average valency for atom C*, *Minimal valency for atom O*, *Maximal bond order for atom H*, *Maximal atomic state energy for atom C*, and *Shadow plane YZ*. Since, that model has not been previously found, the values of the descriptors involved were carefully compared for the initial and modified 3D structures. It has been found that one descriptor, *Shadow plane YZ*, significantly (more than 10%) varies as a function of conformation for the molecules 19, 20, 23, and 24 (Table 1) but not for molecule 32. Since the descriptors' preselection procedure was performed on the conformers of the molecule 32, the *Shadow plane YZ* descriptor has not been excluded. When 16 fragment descriptors were added, the calculations initiated with 278 descriptors resulted in a reasonable six descriptors model (R^2 = 0.916, R_{cv}^2 = 0.859, s = 0.23), also involving *Shadow plane YZ*. Thus, both models issued from the calculations with modified 3D structures included a conformationally dependent descriptor, and, therefore, they were not retained.

It should be noted that models based on such conformationally dependent descriptors are generally less predictive than those involving conformationally invariant descriptors. Indeed, when the new models involved *Shadow plane YZ* were applied for eight compounds from the "blind test" set, the statistical criteria for the correlation $(\log D)_{calc}$ vs $(\log D)_{exp}$ were less good (R^2 = 0.790 for the initial set of 278 descriptors) than those obtained with the models containing no conformationally dependent descriptors (R^2 = 0.877, Table 9).

The above results show that the preselection of pertinent descriptors is an important step to improve the quality of predictions, especially for flexible metal binders. A reliable protocol of such preselection should be further developed.

Table 8. Experimental and Predicted by TRAIL log*D* Values for 8 Compounds from the "Blind Test" Set[a]

no.	compound	LogD_U				
		exp.	average predicted[b]			
			Full Set	Training Set 1	Training Set 2	Overall mean[c]
1		1.81 ± 0.05	2.08 ± 0.17	1.97 ± 0.18	2.10 ± 0.15	2.05 (0.16)
2		1.58 ± 0.05	3.08 ± 0.11	3.15 ± 0.18	3.08 ± 0.10	3.10 (0.12)
3		-0.27 ± 0.02	0.50 ± 0.21	0.38 ± 0.15	0.50 ± 0.20	0.46 (0.17)
4		0.39 ± 0.02	0.22 ± 0.08	0.18 ± 0.07	0.19 ± 0.04	0.19 (0.06)
5		-0.09 ± 0.01	-0.24 ± 0.03	-0.28 ± 0.03	-0.24 ± 0.03	-0.25 (0.03)
6		0.28 ± 0.02	0.06 ± 0.01	0.05 ± 0.02	0.09 ± 0.01	0.07 (0.02)
7		-0.19 ± 0.01	0.03 ± 0.09	-0.02 ± 0.12	0.02 ± 0.08	0.01 (0.09)
8		1.03 ± 0.05	0.37	0.61	0.51	0.50 (0.12)
	R^2[d]		0.688	0.728	0.714	0.712
	s[d]		0.71	0.67	0.68	0.68
	R^2[e]		0.673	0.781	0.715	0.726
	s[e]		0.48	0.38	0.45	0.44

[a] Experimental studies followed by prediction calculations. [b] Average over **II**(**B**)/eq 1, **II**(**A**)/eq 1, and **I**(**AB**, 2−3)/eq 1 linear models for a given data set. [c] Average over **II**(**B**)/eq 1, **II**(**A**)/eq 1, and **I**(**AB**, 2−3)/eq 1 linear models and over all data sets. [d,e] Statistical criteria for the correlation between experimental and predicted log*D* values were calculated for all "blind test" compounds[d] or excluding the outliers (compound 2)[e].

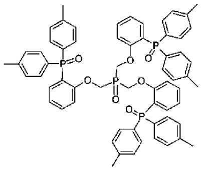
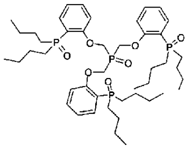
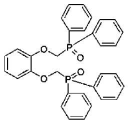
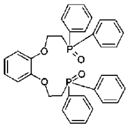
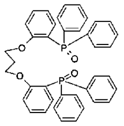
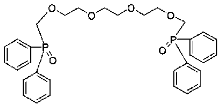
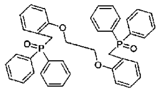
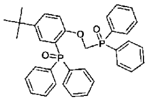
4.2. Fragments vs Other Descriptors. In this study, we used two different QSPR approaches: the Hansch-type approach[18] which uses as descriptors some physicochemical parameters calculated either by quantum mechanical methods or by some empirical techniques and the Free-Wilson-type approach[19] which uses molecular fragments as descriptors. Both techniques have their advantages and disadvantages.

Computation of fragment descriptors does not require the knowledge of the geometry and electronic structure of molecules; structural fragments are more easily interpretable than topological indices. Molecular fragments are successfully used in diversity analysis of large databases[35,36] and in structure−property studies.[18,37−40] The disadvantage of QSPR methods based on fragments is related to the fact that they generally use more variables than those using the CODESSA-PRO descriptors, thus leading to smaller values of Fischer criterion. However, our experience shows[8,9,23,30] that in most cases, fragments based techniques lead to statistically stable and predictive models.

The success of the fragment approach in QSPR studies depends on the diversity of structural fragments as well as on the flexibility of atom/bond classification. Here we used the *Substructural Molecular Fragments* method[8] implemented in the TRAIL program, which represents a very flexible structure−property tool since it uses many different types of fragments (atom/bond sequences and augmented atoms) in order to builds QSPR models involving linear and nonlinear fitting equations.

Using molecular fragments together with "classical" descriptors is an interesting way to improve the robustness of a QSPR model. Thus, including the molecular fragments generated by TRAIL (**I**(**A**, 2−3) sequences) to the list of

Table 9. Experimental and Predicted by CODESSA-PRO log*D* Values for the 8 Compounds of the “Blind Test” Set[a]

		LogD_U				
			average predicted[d]			
no.	compound	exp.	Full Set	Training Set 1	Training Set 2	*Overall mean*
1		1.81±0.05	2.01 ± 0.35	1.97 ± 0.32	2.01 ± 0.35	2.00±0.26
2		1.58±0.05	2.86 ± 0.06	2.80 ± 0.05	2.88 ± 0.04	2.85±0.05
3		-0.27±0.02	-0.06 ± 0.14	-0.03 ± 0.13	-0.07 ± 0.11	-0.05±0.10
4		0.39±0.02	0.58 ± 0.12	0.56 ± 0.08	0.58 ± 0.14	0.57±0.09
5		-0.09±0.01	0.04 ± 0.18	0.04 ± 0.15	0.03 ± 0.21	0.04±0.14
6		0.28±0.02	0.26 ± 0.27	0.27 ± 0.23	0.25 ± 0.25	0.26±0.19
7		-0.19±0.01	0.02 ± 0.27	0.04 ± 0.24	0.01 ± 0.25	0.02±0.20
8		1.03±0.05	0.99 ± 0.02	0.94 ± 0.05	0.99 ± 0.01	0.97±0.04
	R^2 [b]		0.877	0.871	0.875	0.874
	s [b]		0.40	0.40	0.41	0.41
	R^2 [c]		0.979	0.975	0.979	0.977
	s [c]		0.12	0.13	0.12	0.12

[a] Experimental studies followed by prediction calculations. [b] Statistical characteristics (R^2 and s) for the correlation between experimental and predicted log*D* values for all compounds or [c]excluding the outliers. [d] Average over the **FS** and **TS1** models.

descriptors of CODESSA PRO allowed us to build more statistically significant models than those obtained by TRAIL. Thus, the statistical criteria of linear regression logD_{calc} vs logD_{exp} calculated for the compounds from the parent data set and the training sets 1 and 2 with CODESSA PRO ($R^2 = 0.920 - 0.925$, $s = 0.18 - 0.19$, Figure 3) are better than those obtained with TRAIL ($R^2 = 0.867 - 0.874$, $s = 0.21-0.23$, Figure 3). Analogously, correlation coefficients (R^2) for the test sets 1 and 2 obtained with TRAIL are 0.843−0.931 (Table 3), and those calculated with CODESSA PRO are 0.981−0.998 (Table 6). At last, for the podands from the “blind test” set, the R^2 values are 0.712 (Table 8) and 0.874 (Table 9) for TRAIL and CODESSA PRO, respectively.

4.3. Iterative Improvement of QSPR Models. As we have earlier mentioned, the problems with an overestimated log*D* value for the compound 2 from the blind test set (Figure 7) is related to insufficient size of the data set used for the training of QSPR models. A natural way to improve reliability of QSPRs is enlargement of the number of compounds in the training set due to addition of newly synthesized and tested molecules, as it is shown in Figure 1. To improve developed QSPR models, the eight new podands were added to the initial parent set of 32 compounds thus forming a new data set of 40 molecules. To develop the QSPR models we used the protocol described in Method section. The best model of CODESSA PRO (**FS40**) built for the new data set involves 6 descriptors *(the number of the C−P−C fragments, maximal valency for atom C, minimal valency for atom H, maximal valency for atom H, average complementary information content (order 2) and maximal partial charge (Zefirov) for atoms C)* from which only two and three descriptors are common, respectively, with the **FS** and **TS1** models developed for the set of 32

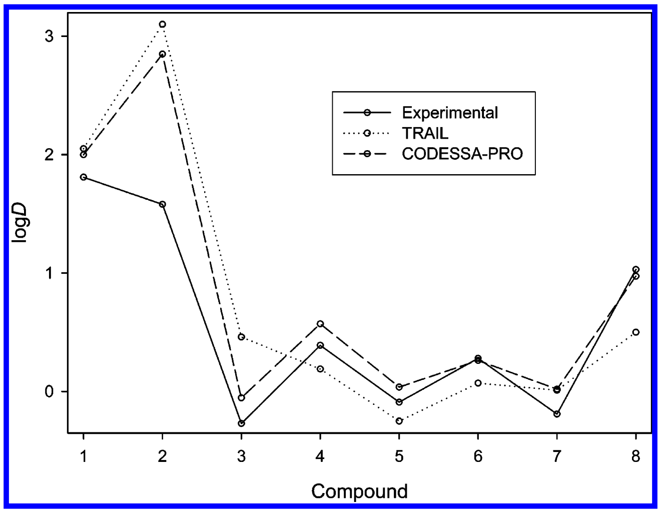


Figure 7. Experimental and predicted log*D* values for the 8 compounds from the "blind test" set. The log*D* values were calculated as an average of values obtained by TRAIL with **II(B)**/eq 1, **II(A)**/eq 1, and **I(AB**, 2−3)/eq 1 models and of those obtained by CODESSA PRO using **FS** and **TS1** models.

compounds. Statistical parameters of the **FS40** model (R^2 = 0.896, s = 0.25, Q^2 = 0.845, F = 47.4) are better than those for the models obtained for the set of 40 compounds with **FS** (R^2 = 0.875, s = 0.28, Q^2 = 0.739, F = 38.4) and **TS1** (R^2 = 0.850, s = 0.30, Q^2 = 0.742, F = 31.2) descriptors. Comparison of correlation coefficients for linear regressions between experimental and calculated log*D* values (Figure 8) shows that the **FS40** model describes experimental data better than the **FS** and **TS1** models (R^2 = 0.896, 0.854, and 0.837 for **FS40**, **FS**, and **TS1** models, respectively). Thus, the "second iteration" according to the scheme in Figure 1 really improves a robustness of the model.

As for the initial set of 32 podands, calculations with TRAIL on the enlarged data set resulted in linear **II**(**B**)/eq 1, **II**(**A**)/eq 1, and **I**(**AB**, 2−3)/eq 1 models corresponding to reasonable statistical criteria ($R^2(fit)$ = 0.798−0.828, s = 0.29−0.35, Q^2 = 0.530−0.557, s_{PRESS} = 0.51−0.58).

5. CONCLUSION

A QSPR modeling of the distribution coefficient (log*D*) of uranyl extracted by phosphoryl-containing podands was performed using two different approaches: one based on classical physicochemical descriptors (implemented in the CODESSA PRO program) and another one based on fragment descriptors (implemented in the TRAIL program). Taking into account conformational flexibility of podands, only conformationally invariant or weakly conformationally dependent descriptors were used in CODESSA PRO calculations. Several robust models were obtained with CODESSA PRO which involved its "own" descriptors together with fragment descriptors generated by TRAIL. Using TRAIL alone, three statistically significant models involving as descriptors either sequences of atoms and bonds or atoms with their close environment (augmented atoms) were developed. Obtained QSPR models were applied for the estimation of log*D* values for a virtual combinatorial library of 2024 podands generated with the CombiLib program. Eight of these hypothetical compounds which span the range of log*D* variation for experimentally studied molecules were then synthesized and tested experimentally. Comparison of calculated and new experimental results shows that developed QSPR models have successfully predicted log*D* values for 7 of 8 compounds from that "blind test" set.

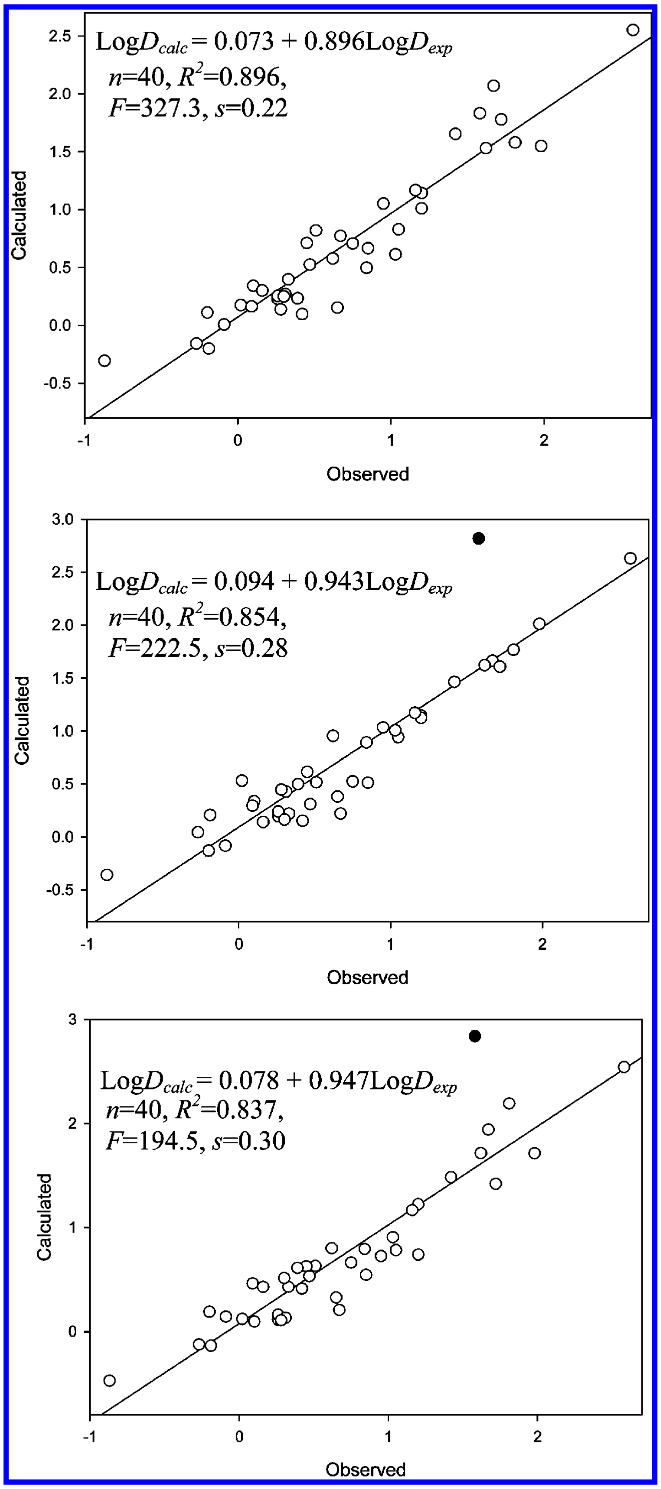


Figure 8. Correlation between experimental and calculated log*D* values obtained for an enlarged data set of 40 podands using **FS40** (*top*), **FS** (*middle*), and **TS1** (*bottom*) models. The outliers are indicated in black.

6. EXPERIMENTAL SECTION

6.1. Synthesis. Below we report the synthesis of new podands 1, 2, and 8 (Table 9). Their structures were established by analytical data and ^{1}H and ^{31}P spectra. NMR spectra were recorded on a Bruker CXP-200 spectrometer in $CDCl_3$ with tetramethylsilane as internal and H_3PO_4 as external references. Melting points (uncorrected) were measured on a Boetius PNMK-05 instrument.

Other compounds were prepared as reported earlier by Baulin et al.[41] (podands 3 and 6), by Evreinov et al.[42] (podand 4), by Tsvetkov et al.[43] (podand 5), and by Evreinov et al.[44] (podand 7).

Tris{[o-di(p-tolyl)phosphinyl]phenoxy}phosphine oxide (compound 1, Table 9). To a suspension of 3.223 g (0.010 mmol) of *o*-diphenylphosphinylphenol[45] and 3.260 g (0.010 mmol) of anhydrous Cs_2CO_3 in 50 mL of anhydrous dioxane was added 0.65 g (0.0033 mol) of tris(chlormethyl)phoshine oxide. The mixture was heated to 95−101 °C for 24 h. After cooling to room temperature 150 mL of H_2O was added, and then 6 N HCl was added until pH 1 was reached. The mixture was extracted by $CHCl_3$ (3 × 50 mL), and the solvent was removed in vacuo. The residue was purified by column chromatography [$SiO_2/CHCl_3$, $CHCl_3$−EtOH(20:1)] to yield 1.80 g (52%) (comp. 1, Table 9) as colorless glass. $C_{63}H_{60}O_7P_4$ mol. mass 1053.07, calcd C 71.86; H 5.74; P 11.77; found C 71.79; H 5.56; P 11.67. ^{1}H NMR (200 MHz, $CDCl_3$) $\delta = 1.20$ (s, 18H, Ar−CH_3), 4.60 (d, 6H $^2J_{HP}$ 6.5 Hz, OCH_2P), 6.78−8.00 (m, 36H Ar−H). ^{31}P NMR (200 MHz, $CDCl_3$) $\delta = 28.02$, 36.98.

Tris[o-(di-butylphosphinyl)phenoxy]phosphine oxide (compound 2, Table 9) was obtained analogous to the compound 1 from 2.543 g (0.010 mmol) of dibutylphosphinylphenol,[45] 3.260 g (0.010 mmol) of anhydrous Cs_2CO_3, and 0.65 g (0.0033 mol) of tris(chlormethyl)phoshine oxide in 50 mL of anhydrous dioxane. Yield 1.76 g (63%); mp 205−207 °C (benzene). $C_{45}H_{72}O_7P_4$ mol. mass 848.97, calcd C 63.67; H 5.55; P 14.59; found C 63.50; H 5.54; P 14.33. ^{1}H NMR (200 MHz, $CDCl_3$) $\delta = 0.7$−1.88 (m, 54H, C_4H_9), 4.67 (d, 6H $^2J_{HP}$ 6.5 Hz, OCH_2P), 6.88−8.10 (m, 12H Ar−H). ^{31}P NMR (200 MHz, $CDCl_3$) $\delta = 36.98$, 41.32.

Diphenylphosphinoyl-[(2-diphenylphoshinoyl-4-tert-butyl)-phenoxy}methan (compound 8, Table 9) was obtained analogously to compound 1 (Table 9) from 3.500 g (0.010 mmol) of 2-diphenylphosphinoyl-4-*tert*-butylphenol,[45] 3.260 g (0.010 mmol) of anhydrous Cs_2CO_3, and 3.26 g (0.010 mol) of diphenylphosphinoylmethyl *p*-toluenesulfonate in 50 mL of anhydrous dioxane. Yield 3.55 g (63%); colorless glass. $C_{35}H_{34}O_3P_2$ mol mass 564.61; calcd C 74.46; H 6.07; P 10.97; found C 74.24; H 6.00; P 10.75. ^{1}H NMR (200 MHz, $CDCl_3$) $\delta = 1.28$ (s, 9H, *t*-C_4H_9), 4.60 (d, 6H $^2J_{HP}$ 6.5 Hz, OCH_2P), 7.35−7.70 (m, 23H Ar−H). ^{31}P NMR (200 MHz, $CDCl_3$) $\delta = 28.98$, 31.32.

6.2. Extraction Experiments. Distribution coefficients log*D* of uranyl cation extracted by new 8 podands (Table 9) have been performed using the same protocol as used for the parent set of 32 compounds in ref 15. Solutions of uranyl cation were prepared from one-element standards (*Merck*, Germany; *Inorganic Ventures, Inc.*, U.S.A.); solutions of 2 M HNO_3 were prepared from analytical grade HNO_3; and 1,2-dichloroethane of chemical grade was used in the experiments.

The extractions were carried out in glass ampules at 291 ± 2 K. The volume of organic (1,2-dichloroethane) and aqueous phases were both 2 mL. The initial concentrations of the uranyl cation and podands were 10^{-5} M and 0.01 M, respectively. The solutions were mixed at 60 rpm for 1 h to achieve an equilibrium. One-half milliliter of the aqueous phase was taken for further analysis. The concentrations of elements in the initial and equilibrium aqueous solutions were determined by inductively coupled plasma mass-spectrometry (ICP-MS).

The reliability of new extraction experiment was checked by measuring distribution coefficients for the podands 30 and 31 from initial data set (Table 1). The new values (log*D* 1.23 ± 0.05 and 1.66 ± 0.05, respectively) correspond well to those reported previously[15] (log*D* 1.16 and 1.62, respectively).

ACKNOWLEDGMENT

V.P.S. gratefully acknowledges A.V. for the opportunity to have spent a term in 2002−2003 as an Invited Professor at the ULP, Strasbourg. GDR PARIS is acknowledged for support.

REFERENCES AND NOTES

(1) Rozen, A. M.; Krupnov, B. V. Dependence of the Extraction Ability of Organic Compounds on Their Structure. *Uspekhi Khim. (Rus.)* **1996**, *65*, 1052−1079.

(2) 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.

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

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

(5) Rabbe, C.; Sella, C.; Madic, C.; Godard, A. Molecular Modeling Study of Uranyl Nitrate Extraction with Monoamides. II. Molecular Mechanics and Lipophilicity Calculations. Structure−Activity Relationships. *Solvent Extr. Ion Exch.* **1999**, *17*, 87−112.

(6) Voelkel, A.; Szymanowski, J. Structure−Activity Relationships for Hydroxyoxime Metal Extractants. *J. Chem. Technol. Biotechnol.* **1993**, *56*, 279−288.

(7) Yoshizuka, K.; Inoue, K.; Ohto, K.; Gloe, K.; Stephan, H.; Rambusch, T.; Comba, P. QSPR of Extractability Trends of the Lanthanoid Series Using Novel Molecular Mechanics Calculations. *Solvent Extraction for the 21st Century, Proceedings of ISEC '99, Barcelona, Spain, July 11−16, 1999*, 2001; pp 687−692.

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

(9) Varnek, A.; Wipff, G.; Solov'ev, V. P. Towards an Information System on Solvent Extraction. *Solvent Extr. Ion Exch.* **2001**, *19*, 791−837.

(10) Kron, T. E.; Tsvetkov, E. N. Neutral, Acyclic Analogues of Crown Ethers and Cryptands and Their Complex-Forming Properties. *Uspekhi Khim. (Rus.)* **1990**, *59*, 483−508.

(11) 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. Obshchei Khim. (Rus.)* **1990**, *60*, 1506−1511.

(12) 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. Obshchei Khim. (Rus.)* **1995**, *65*, 223−231.

(13) Solov'ev, V. P.; Baulin, V. E.; Strakhova, N. N.; Kazachenko, V. P.; Belsky, V. K.; Varnek, A. A.; Volkova, T. A.; Wipff, G. Complexation of Phosphoryl-Containing Mono-, Bi- and Tri-Podands with Alkali Cations in Acetonitrile. Structure of the Complexes and Binding Selectivity. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1489−1498.

(14) Turanov, A. N.; Karandashev, V. K.; Baulin, V. E. Extraction Properties of Neutral Phosphoryl Containing Podands in Hydrochloride Media. *Zh. Neorg. Khim. (Rus.)* **1998**, *43*, 1734−1749.

(15) Turanov, A. N.; Karandashev, V. K.; Baulin, V. E. Extraction of Uran and Thorium with Neitral Phosphoryl Containing Podands from Nitric Acid Solutions. *Radiokhimiya (Rus.)* **1998**, *40*, 36−43.

(16) Turanov, A. N.; Karandashev, V. K.; Baulin, V. E. Extraction of Metal Species from HNO3 Solutions by Phosphoryl-Containing Podands. *Solvent Extr. Ion Exch.* **1999**, *17*, 525−552.

(17) Turanov, A. N.; Karandashev, V. K.; Evseeva, N. K.; Baulin, V. E.; Ushakova, A. P. Structure Influence of Phosphoryl-Containing Podands on Europium and Americium Extraction from Nitric Acid Solutions. *Radiokhimiya (Rus.)* **1999**, *41*, 219−224.

(18) Hansch, C.; Leo, A. *Exloring QSAR. Fundamentals and Applications in Chemistry and Biology*; ACS Prof. Ref. Book: Washington, 1995; p 557.

(19) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(20) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-Dimensional Quantitative Structure−Activity Relationship of Human Immunodeficiency Virus (I) Protease Inhibitors. 2. Predictive Power Using Limited Exploration of Alternate Binding Modes. *J. Med. Chem.* **1994**, *37*, 2206−2215.

(21) Forsythe, G. E.; Malcolm, M. A.; Moler, C. B. *Computer Methods for Mathematical Computations*; Prentice Hall, Inc.: Englewood Cliffs, NY, 1977; p 280.

(22) Kendall, M. G.; Stuart, A. *The Advanced Theory of Statistics*; Griffin: London, 1966.

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

(24) Katritzky, A. R.; M., U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(25) Katritzky, A. R.; O. A.; Lomaka, A.; Karelson, M. Six-Membered Cyclic Ureas as HIV-1 Protease Inhibitors: A QSAR Study Based on CODESSA PRO Approach. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3453−3457.

(26) Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3−10.

(27) Charton, M. The Upsilon Steric Parameter X: Definition and Determination. *Steric Effects in Drug Design*; Springer-Verlag: Berlin, 1983; pp 57−91.

(28) *MacroModel 5.5*; Department of Chemistry, Columbia University, New York 10027, 1996.

(29) *Spartan*; version 5.1.3; Wavefunction, Inc.: Irvine, CA, 1999.

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

(31) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(32) Barnard, J. M.; Downs, J. M.; von-Scholley-Pfab, A.; Brown, R. D. Use of Markush Structure Analysis Techniques for Descriptor Generation and Clustering of Large Combinatorial Libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 452−463.

(33) Tute, M. S. *History and Objectives of Quantitative Drug Design*; Pergamon Press: Oxford, New York, Beijing, Frankfurt, 1990; pp 1−31.

(34) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; J. Wiley & Sons: New York, 2000; p 430.

(35) Klopman, G.; Tu, M. Diversity analysis of 14 156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure−activity relational expert system MCASE. *J. Med. Chem.* **1999**, *42*, 992−998.

(36) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.

(37) Zefirov, N. S.; Palyulin, V. A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112−1122.

(38) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.

(39) Avidon, V. V. The Criteria of Chemical Structures Similarity and the Principles for Design of Description Language for Chemical Information Processing of Biologically Active Compounds. *Chim. Pharm. J. (Rus.)* **1974**, *8*, 22−25.

(40) Bawden, D. Computerized Chemical Structure-Handling Techniques in Structure−Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14−22.

(41) Baulin, V. E.; Evreinov, V. I.; Vostroknutova, Z. N.; Bondarenko, N. A.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands. 9. Synthesis of Bis(diphenylphosphinyl)ethyl Esters of Oligoethylene Glycols and Their Complex-Forming Properties with Alkali Metal Cations in Weakly Polar Solvents. *Izvestiya Akademi Nauk, Seriya Khimicheskaya (Rus.)* **1992**, 1161−1167.

(42) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Tsvetkov, E. N. Phosphorus-Containing Podands. 10. An Improved Method for Synthesis of Oligo(ethylene glycol)bis[2-(diphenylphosphinoyl)ethyl] Ethers and Their Complex-Forming Properties with Respect to Alkali Metal Cations in Anhydrous Acetonitrile. *Izvestiya Akademii Nauk, Seriya Khimicheskaya (Rus.)* **1993**, 518−522.

(43) Tsvetkov, E. N.; Evreinov, V. I.; Baulin, V. E.; Ragulin, V. V.; Bondarenko, N. A.; Vostroknutova, Z. N.; Safronova, Z. V. Phosphorus-Containing Podands. XIII. Ligands with Lithium−Sodium Selectivity − Oligoethylene Glycol Bis[o-(2-diphenylphosphinylethyl)phenyl]-ethers and Their Analogues. *Zh. Obshchei Khim. (Rus.)* **1995**, *65*, 1421−1431.

(44) Evreinov, V. I.; Baulin, V. E.; Vostroknutova, Z. N.; Bondarenko, N. A.; Syundyukova, V. K.; Tsvetkov, E. N. Phosphorus-Containing Podands. 4. Influence of the Length of the Polyether Chain of Bis-[(o-diphenylphosphinylmethyl)phenyl] ethers of Oligo(ethylene glycols) on Their Complexation and Selectivity Properties in Relation to Alkali-Metal Cations. *Izvestiya Akademii Nauk SSSR, Seriya Khimicheskaya (Rus.)* **1989**, 1990−1997.

(45) Tsvetkov, E. N.; Syundyukova, V. K.; Baulin, V. E. Neutral Mono- and Dipodands with Terminal Phosphinylphenyl Groups. *Zh. Obshchei Khim. (Rus.)* **1987**, *57*, 2456−2461.

CI049976B