Quantitative Structure−Property Relationship Modeling of β-Cyclodextrin Complexation Free Energies

Alan R. Katritzky,*,§ Dan C. Fara,§,‡ Hongfang Yang,§ Mati Karelson,‡ Takahiro Suzuki,⊥ Vitaly P. Solov'ev,|| and Alexandre Varnek#

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611, Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu 51014, Estonia, Natural Science Laboratory, Toyo University, 11-10, Oka 2, Asaka-shi, Saitama 351-8510, Japan, Institute of Physiologically Active Compounds, Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russia, and Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France

Received August 30, 2003

CODESSA-PRO was used to model binding energies for 1:1 complexation systems between 218 organic guest molecules and β-cyclodextrin, using a seven-parameter equation with $R^2 = 0.796$ and $R_{cv}^2 = 0.779$. Fragment-based TRAIL calculations gave a better fit with $R^2 = 0.943$ and $R_{cv}^2 = 0.848$ for 195 data points in the database. The advantages and disadvantages of each approach are discussed, and it is concluded that a combination of the two approaches has much promise from a practical viewpoint.

INTRODUCTION

Cyclodextrins (CDs) are cyclic oligomers of α-D-glucose which result from the action of certain enzymes on starch. The family includes three well-known industrially produced members-α-CD (six glucose units), β-CD (seven units), and γ-CD (eight units)-as well as several other less well-known oligosaccharides. The α-, β-, and γ-CDs, commonly referred to as the native cyclodextrins, are crystalline, homogeneous, nonhygroscopic substances which form cylindrical or doughnut-shaped molecules with their hydroxyl groups on the outside of the molecule. Cyclodextrin molecules are shallow truncated cones rather than toruses. The primary hydroxyl rim of the cavity opening possesses a somewhat reduced diameter compared with the secondary hydroxyl rim.

The CD exterior, containing many hydroxyl groups, is fairly polar, whereas the interior of the cavity is nonpolar relative to the exterior and relative to water, which is the usual external environment.[1] In principle, in aqueous solution, the slightly apolar CD cavity is occupied by water molecules, which are energetically unfavored (polar−apolar interaction) and therefore can be readily substituted by appropriate "guest molecules" which are less polar than water. The dissolved CD is the host molecule, and the driving force of the complex formation is the substitution of the high-enthalpy water molecules by an appropriate guest molecule. One, two, or three CD molecules may contain one or more entrapped guest molecules.[1]

This host−guest property allows CDs to be used in numerous applications in industrial, pharmaceutical, agricultural, and other fields, including improving the solubility and stability of drugs and selectively binding materials that fit into the central cavity in affinity and chromatography purification methods.[2,3]

Many attempts have been made to rationalize the stability of noncovalent host−guest inclusion complexes in terms of the molecular shape and size of guests, intermolecular interactions, or solvation effects as driving forces for CD complexation. Previous studies have suggested five major interactions: (i) hydrophobic interactions, (ii) van der Waals interactions, (iii) hydrogen-bonding between polar groups of the "guests" and the hydroxyl groups of the "host", (iv) relaxation by release of high-energy water from the cyclodextrin cavity upon substrate inclusion, and (v) relief of the conformational strain in a cyclodextrin−water adduct.[1−3]

Computational chemistry has only recently been used as a tool for studying the stability of CD intrusion complexes. Its earlier neglect does not reflect a lack of interest from computational chemists, but rather a limitation of the tools available to carry out the calculations in a reasonable period of time, due to the fact that CDs and their derivatives are relatively large, flexible molecules that are often studied experimentally in aqueous environments. Consequently, application of quantum and molecular mechanics has been difficult. (i) It has required many assumptions and restrictions when methods restricted to the gas phase were applied to CDs in aqueous solution, where solvation and hydrophobic effects are very important. (ii) It has achieved good results mainly for small groups of closely related chemically compounds.[4] Recently, substantial increases in the power of computers and software has increased the possibility to apply successfully computational chemistry to CD complexation.

Applications of computational chemistry to the study of cyclodextrins have been well reviewed by Lipkowitz.[5] Group-contribution models, QSAR/QSPR methods (2D-QSAR, 3D-QSAR, CoMFA), molecular modeling computa-

* Corresponding author phone: (352) 392-0554; fax: (352) 392-9199; e-mail: katritzky@chem.ufl.edu.
§ University of Florida.
‡ University of Tartu.
⊥ Toyo University.
|| Russian Academy of Sciences.
Université Louis Pasteur.

10.1021/ci034190j CCC: $27.50
Published on Web 01/30/2004

tions (using quantum mechanics, molecular mechanics and dynamics, Monte Carlo simulations, etc.), statistical analysis tools, and artificial neural networks have all been applied to elucidate the most important factors influencing the host−guest interactions and to predict the thermodynamic stability of CDs inclusion complexes.[6−14] However, reliable and convenient methods for predicting the thermodynamic stability of CD complexes have not yet been published.

The aim of the present study is to build QSAR multiple regression models, which could correlate and predict the free energies of inclusion complexation between diverse guest molecules and CDs on the basis of the reported experimental binding data.[6] A large data set of 218 compounds containing great structural variability has been treated using two different QSPR approaches. The first of them (the Hansch-type approach[15]) uses as descriptors certain physicochemical parameters calculated either by quantum mechanical methods or by some empirical techniques. The second (the Free−Wilson-type approach[16]) uses counts of different molecular fragments as variables in a multiple regression analysis. Both techniques have their advantages and disadvantages.

METHODOLOGY

Descriptor Approach. CODESSA (comprehensive descriptors for structural and statistical analysis)-PRO[17] is a comprehensive program for developing quantitative structure−property relationships (QSPR), integrating all necessary mathematical and computational tools to (i) calculate a large variety of molecular descriptors on the basis of the 3D geometrical and/or quantum-chemical structural input of chemical compounds; (ii) develop (multi)linear and nonlinear QSPR models of the chemical, physical, or biological properties of individual compounds; (iii) perform cluster analyses of the experimental data and molecular descriptors; (iv) interpret the developed models; and (v) predict properties for compounds previously unknown or unavailable.

In the framework of the CODESSA-PRO program, diverse statistical structure−property correlation techniques can be used for the analysis of the given data set in combination with the calculated molecular descriptors. In particular, various algorithms based on stepwise statistical multilinear regression analysis are applicable for searching "the best" multiparameter correlation in the large spaces of the natural molecular descriptors. As only theoretically calculated descriptors are used in the resulting multiparameter correlation equations, the value of the property of interest can be predicted for any chemical structure.

The applicability of CODESSA-PRO methodology to various QSAR/QSPR problems has been again convincingly demonstrated in a series of our recent publications.[18−20]

Fragment Approach. Computation of fragmental counts does not require the knowledge of the 3D geometry and electronic structure of molecules; the regression coefficients related to the structural fragments are more easily interpretable than those for the topological indices. Molecular fragments are successfully used in diversity analysis of large databases[21,22] and in structure−property studies.[15,23−27] The recently developed PASS method,[24,28] used for estimation of a wide spectrum of biological activities, is also based on molecular fragments (augmented atoms). The disadvantage of QSPR methods based on fragments is related to the fact that they generally use more variables than those using "traditional" descriptors, thus leading to smaller values of Fischer criterion (less robust models). However, the experience of one of our groups shows[29−32] that, in most cases, fragment-based techniques lead to statistically stable and predictive models. Another problem of the fragment-based approach is related to molecules containing fragments of "rare" occurrence (i.e., found in a single molecule), which have to be excluded from the training or test sets, thus reducing the number of treated compounds.

The success of the fragmental 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 (SMF) method[29] incorporated 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 build QSPR models involving linear and nonlinear fitting equations. 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 in combination with three fitting equations. Using selected QSAR models, one can calculate average activities of the compounds from the test set, which smoothes the inaccuracies of particular individual models, thus improving the robustness of predictions.[31]

The TRAIL program[29−31] has been developed to calculate structure−property relationships on the basis of the SMF partitioning. At the first step of the calculations, TRAIL generates up to 147 structure−property models involving 49 types of fragments coupled with three fitting equations, and uses all of them at the training and test stages. Molecules containing fragments of "rare" occurrence (i.e., found in less than two molecules) are excluded from the training set. If some fragments are linearly dependent, they are treated as one extended fragment. The program then calculates statistical characteristics of models (correlation coefficient R, standard deviation s, Fischer's criterion F, cross-validation correlation coefficient R_{cv}, and cross-validation standard deviation s_{cv}) and performs statistical tests to select the best models. At the test stage, TRAIL calculates the predicted values using the fitted fragments' contributions obtained at the training stage.

Earlier, we showed that the SMF method represents an efficient tool to model the octanol/water partition coefficient, the stability constants of host−guest complexes in water and in nonaqueous solution, and the anti-HIV activity.[29−31]

The Present Work. We now report QSPR modeling studies of the free energy of complexation of β-cyclodextrin with 218 neutral guests in water, performed with the TRAIL program, which uses the fragmental (SMF) descriptors, and the CODESSA-PRO program, which applies up to 902 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors. All these descriptors are derived solely from molecular structure and do not require experimental data to be calculated.

We will show that a combination of these two different techniques (TRAIL and CODESSA-PRO) could lead to pertinent QSPR models, whose joint application may allow us to improve the robustness of predictions.

DATA SET AND CALCULATIONS

The thermodynamic stability of a CD inclusion complex is usually expressed using the binding constant/stability constant of the complex, *K*.

In the present work, a large data set of experimental ΔG free energies of complexation for β-CD in water at 25 °C (ΔG) was used, consisting of 218 organic compounds of different classes: aromatic hydrocarbons, alcohols, phenols, ethers, aldehydes, ketones, acids, esters, nitriles, anilines, heterocycles, nitro and sulfur compounds, steroids, and barbitals (see Table 1).[6] The number of carbon atoms in the molecules varies in the range from 1 to 26.

CODESSA-PRO Calculations. The structures of the compounds were drawn using the ISIS/Draw facility as implemented in the ISIS 2.4 packages and exported as MOL files. The molecular geometries were optimized using AM1 Hamiltonian calculations together with the eigenvector, following the geometry optimization procedure available in the quantum chemical program MOPAC 7.05, implemented in the CODESSA-PRO package.[17] 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., on the basis of 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. This result has the best representation of the property in the given descriptors pool.

In the development of QSPRs, a major decision is when to stop adding descriptors to the model during the stepwise regression procedure. An excessive number of descriptors leads to overcorrelated equations that are difficult to interpret in terms of interaction mechanisms. A simple procedure used to control the model expansion is the so-called "break point" in improvement of the statistical quality of the model. From analyses of the plot of the number of descriptors involved vs the squared correlation coefficient, and also the squared cross-validated correlation coefficient, using values corresponding to those models, it appears that the statistical improvement of the model is higher (steeper ascent of the relationship) until one point (the "break point") and after that the improvement is negligible (low ascent of the relationship). Consequently, the model corresponding to the break point is considered to be the best/optimum model.

TRAIL (SMF) Calculations. The SMF method[29] is based on the splitting of a molecular graph into fragments, and on the calculation of their contributions to a given property. Two different types of fragments are used: "sequences" (I) and "augmented atoms" (II). Three subtypes are defined for each type of fragments: those including atoms and bonds (AB), atoms only (A), or bonds only (B). Once a given molecular structure is split into constitutive fragments, the fragments are used as descriptors in linear or nonlinear QSPR models.

The SMF method recognizes nine different types of bonds: single, double, triple (in cycle or in chain), aromatic bonds, and two types of coordination bonds. Therefore, the *EdChemS* editor of 2D structures incorporated in TRAIL was used to modify the bond types in the aromatic fragments originally presented as Kekule structures. Then, using the *SDF Editor* also included in TRAIL, a SD file containing all 218 structures and experimental free energies of complexation was prepared and further used in structure−property modeling. All hydrogen atoms were omitted in TRAIL (SMF) calculations.

To study the influence of bonds presentation in similar molecular fragments in chains and cycles on the QSPR models, two types of calculations were performed: without any differentiation of those bonds (Type 1 calculations) and accounting for the difference between them (Type 2 calculations).

Internal Validation of QSPR Models. An important aspect of any QSPR study is the validation of the model. For the internal validation, the parent data set was divided into three subsets: the first, fourth, seventh, etc. entries go into the first subset (#1), the second, fifth, eighth, etc. go into the second subset (#2), and the third, sixth, ninth, etc. go into the third subset (#3). Then, three training sets—Set 1, Set 2, and Set 3—were prepared as a combination of two subsets, (#1 and #2), (#1 and #3), and (#2 and #3), respectively. Corresponding test sets relate to the remaining subsets (#3, #2, and #1, respectively).

For each training set, the correlation equation was derived with the same descriptors. The obtained equation was then used to predict ΔG values for the compounds from the corresponding test set. The efficiency of QSPR models to predict free energy of complexation was estimated using the cross-validation (*Leave One Out* method[33,34]) to determine the correlation both for the full set and for each training set. Correlation coefficients (R(pred.)) and standard deviations (s(pred.)) of linear correlations between experimental ΔG values and those predicted for the test sets were also calculated.

RESULTS

CODESSA-PRO Calculations. A QSPR model was developed for the full set of 218 compounds. The break point as additional descriptors are added is reached at the seventh descriptor, as indicated in Figure 1; improvement in the R^2 value drops below 0.02 after this point. Consequently, the seven-parameter equation, as the best model, is presented in Table 2 and plotted in Figure 2: $N = 218$, $n = 7$, $R^2 = 0.796$, $R_{cv}^2 = 0.779$, $F = 117.3$, and $s^2 = 5.66$.

The seven descriptors involved in the model can be classified as follows: (i) one as topological (average complementary information content of zeroth order, $^0\overline{CIC}$), (ii) four as charge-distribution-related (HACA-2/TMSA, HACA; H-acceptors FPSA, version 2, FPSA; maximum partial charge—Zefirov—for all atom types, Q_i; and WNSA-1-weighted PNSA, WNSA), (iii) one as geometrical (ZX shadow/ZX rectangle, $_SZX^R$), and (iv) one as semiempirical molecular orbital (LUMO energy, ϵ_{LUMO}).

Their direct interpretation is rather difficult, considering the complexity of the process involving conformational changes of the host, host−guest interactions, and solvation−desolvation effects. However, some indirect links between those descriptors and the physical phenomena involved in host−guest complexation might be suggested.

Table 1. Experimental and Calculated Values of the Complexation Free Energy ΔG of β-Cyclodextrin with 218 Organic Compounds in Water at 298 K[a]

		ΔG, kJ/mol						
			CODESSA-PRO		TRAIL (Type 1)		TRAIL (Type 2)	
no.	name of compound	exptl	calcd	exptl − calcd	calcd	exptl − calcd	calcd	exptl − calcd
1	carbon tetrachloride	−12.56	−12.43	−0.13	−12.56	0.0	−12.56	0.0
2	chloroform	−8.16	−11.09	2.93	−8.16	0.0	−8.16	0.0
3	methanol	2.8	4.91	−2.11	1.05	1.75	0.89	1.91
4	acetonitrile	1.54	3.79	−2.25	1.54	0.0	1.54	0.0
5	acetaldehyde	3.65	0.21	3.44	1.58	2.07	2.81	0.84
6	ethanol	0.17	−0.06	0.23	−1.63	1.80	−0.74	0.91
7	1,2-ethanediol	1.08	0.14	0.94	0.25	0.83	*b*	
8	acetone	−2.4	−4.67	2.27	−2.26	−0.14	−1.74	−0.66
9	1-propanol	−3.26	−4.96	1.70	−5.20	1.94	−4.27	1.01
10	2-propanol	−3.6	−6.16	2.56	−4.76	1.16	−3.76	0.16
11	1,3-propanediol	−3.82	−3.50	−0.32	−0.49	−3.33	−1.00	−2.82
12	tetrahydrofuran	−8.39	−6.86	−1.53	−9.45	1.06	−8.39	0.0
13	cyclobutanol	−6.74	−7.65	0.91	−11.63	4.89	−6.74	0.0
14	1-butanol	−6.99	−8.31	1.32	−7.68	0.69	−7.14	0.15
15	2-butanol	−6.79	−9.29	2.50	−7.06	0.27	−6.43	−0.36
16	2-methyl-1-propanol	−9.25	−9.63	0.38	−9.21	−0.04	−9.19	−0.06
17	2-methyl-2-propanol	−9.59	−9.84	0.25	−8.32	−1.27	−8.16	−1.43
18	1,4-butanediol	−3.65	−6.79	3.14	−2.96	−0.69	−3.87	0.22
19	diethylamine	−7.77	−8.81	1.04	−7.77	0.0	−7.77	0.0
20	cyclopentanol	−11.87	−9.74	−2.13	−16.08	4.21	−12.40	0.53
21	1-pentanol	−10.29	−10.19	−0.10	−10.15	−0.14	−10.01	−0.28
22	2-pentanol	−8.51	−11.03	2.52	−9.53	1.02	−9.30	0.79
23	3-pentanol	−7.71	−11.54	3.83	−9.36	1.65	−9.11	1.40
24	2-methyl-1-butanol	−11.87	−12.22	0.35	−10.41	−1.46	−11.20	−0.67
25	2-merthyl-2-butanol	−10.9	−12.86	1.96	−9.35	−1.55	−9.98	−0.92
26	3-methyl-1-butanol	−12.84	−11.87	−0.97	−10.59	−2.25	−11.40	−1.44
27	3-methyl-2-butanol	−10.96	−12.49	1.53	−9.80	−1.16	−10.49	−0.47
28	2,2-dimethyl-1-propanol	−15.48	−11.03	−4.45	−13.65	−1.83	−15.49	0.01
29	1,5-pentanediol	−6.97	−8.07	1.10	−5.43	−1.54	−6.74	−0.23
30	1,4-dibromobenzene	−16.98	−17.64	0.66	−16.00	−0.98	−16.04	−0.94
31	1,4-diiodobenzene	−18.12	−16.52	−1.60	−19.75	1.63	−19.81	1.69
32	3,5-dibromophenol	−14.61	−15.95	1.34	−16.50	1.89	−16.53	1.92
33	3,5-dichlorophenol	−11.82	−12.86	1.04	−13.31	1.49	−13.27	1.45
34	1-chloro-4-nitrobenzene	−12.27	−12.62	0.35	−13.64	1.37	−13.51	1.24
35	fluorobenzene	−11.18	−13.44	2.26	−10.46	−0.72	−10.48	−0.70
36	bromobenzene	−14.28	−16.47	2.19	−14.21	−0.07	−14.21	−0.07
37	iodobenzene	−16.71	−14.41	−2.30	−16.08	−0.63	−16.09	−0.62
38	3-fluorophenol	−9.7	−11.56	1.86	−10.96	1.26	−10.97	1.27
39	4-fluorophenol	−9.87	−11.20	1.33	−10.96	1.09	−10.97	1.10
40	3-chlorophenol	−13.01	−11.05	−1.96	−13.12	0.11	−13.07	0.06
41	4-chlorophenol	−14.92	−10.00	−4.92	−13.12	−1.80	−13.07	−1.85
42	3-bromophenol	−14.33	−14.35	0.02	−14.71	0.38	−14.69	0.36
43	4-bromophenol	−15.12	−13.39	−1.73	−14.71	−0.41	−14.69	−0.43
44	3-iodophenol	−16.73	−14.92	−1.81	−16.58	−0.15	−16.57	−0.16
45	4-iodophenol	−17	−14.84	−2.16	−16.58	−0.42	−16.57	−0.43
46	nitrobenzene	−11.64	−12.63	0.99	−13.45	1.81	−13.30	1.66
47	4-nitrophenol	−14.76	−11.57	−3.19	−13.95	−0.81	−13.79	−0.97
48	benzene	−12.72	−12.36	−0.36	−12.42	−0.30	−12.37	−0.35
49	phenol	−11.28	−13.05	1.77	−12.92	1.64	−12.86	1.58
50	hydroquinone	−11.72	−11.15	−0.57	−13.42	1.70	−13.34	1.62
51	4-nitroaniline	−14.18	−11.21	−2.97	−12.17	−2.01	−12.12	−2.06
52	aniline	−9.13	−10.91	1.78	−11.14	2.01	−11.19	2.06
53	sulfanilamide	−15.79	−10.82	−4.97	−15.79	0.0	−15.79	0.0
54	cyclohexanol	−15.27	−11.80	−3.47	−16.01	0.74	−14.42	−0.85
55	1-hexanol	−13.31	−12.32	−0.99	−12.63	−0.68	−12.88	−0.43
56	2-hexanol	−11.3	−14.07	2.77	−12.01	0.71	−12.17	0.87
57	2-methyl-2-pentanol	−11.36	−13.97	2.61	−11.82	0.46	−12.85	1.49
58	3-methyl-3-pentanol	−12.27	−14.72	2.45	−10.38	−1.89	−11.80	−0.47
59	4-methyl-2-pentanol	−11.64	−12.76	1.12	−12.44	0.80	−13.56	1.92
60	3,3-dimethyl-2-butanol	−15.7	−13.78	−1.92	−12.97	−2.73	−15.94	0.24
61	1,6-hexanediol	−9.65	−10.02	0.37	−7.91	−1.74	−9.61	−0.04
62	benzonitrile	−12.74	−9.01	−3.73	−10.52	−2.22	−10.55	−2.19
63	benzothiazole	−13.59	−13.60	0.01	*b*		*b*	
64	4-nitrobenzoic acid	−13.36	−13.91	0.55	−11.40	−1.96	−12.57	−0.79
65	benzaldehyde	−10.16	−13.15	2.99	−8.79	−1.37	−9.41	−0.75
66	benzoic acid	−12.13	−13.73	1.60	−10.37	−1.76	−11.64	−0.49
67	4-hydroxybenzaldehyde	−9.99	−11.81	1.82	−9.29	−0.70	−9.90	−0.09
68	4-hydroxybenzoic acid	−12.54	−12.20	−0.34	−10.87	−1.67	−12.13	−0.41
69	benzyl chloride	−13.95	−12.98	−0.97	*b*		*b*	

Table 1 (Continued)

		ΔG, kJ/mol						
			CODESSA-PRO		TRAIL (Type 1)		TRAIL (Type 2)	
no.	name of compound	exptl	calcd	exptl − calcd	calcd	exptl − calcd	calcd	exptl − calcd
70	toluene	−11.93	−12.22	0.29	−12.47	0.54	−12.71	0.78
71	benzyl alcohol	−9.76	−12.31	2.55	−13.01	3.25	−11.56	1.80
72	anisole	−13.24	−13.84	0.60	−12.85	−0.39	−12.67	−0.57
73	*m*-cresol	−11.3	−13.10	1.80	−12.97	1.67	−13.19	1.89
74	*p*-cresol	−13.68	−13.11	−0.57	−12.97	−0.71	−13.19	−0.49
75	4-methoxyphenol	−12.61	−13.19	0.58	−13.35	0.74	−13.16	0.55
76	3-methoxyphenol	−12.05	−13.23	1.18	−13.35	1.30	−13.16	1.11
77	4-hydroxybenzyl alcohol	−12.34	−12.19	−0.15	−13.51	1.17	−12.05	−0.29
78	hydrochlorothiazide	−10.04	−10.02	−0.02	*b*		*b*	
79	*N*-methylaniline	−12.07	−12.00	−0.07	−12.07	0.0	−12.07	0.0
80	1-butylimidazole	−12.5	−13.08	0.58	*b*		*b*	
81	1-heptanol	−16.27	−14.17	−2.10	−15.10	−1.17	−15.75	−0.52
82	phenylacetylene	−13.48	−9.80	−3.68	*b*		*b*	
83	thianaphthene	−18.44	−15.86	−2.58	−18.44	0.0	*b*	
84	4-fluorophenyl acetate	−12.05	−11.68	−0.37	−10.67	−1.38	−10.64	−1.41
85	3-fluorophenyl acetate	−10.9	−12.93	2.03	−10.67	−0.23	−10.64	−0.26
86	4-chlorophenyl acetate	−14.27	−11.37	−2.90	−12.82	−1.45	−12.74	−1.53
87	3-chlorophenyl acetate	−13.93	−11.18	−2.75	−12.82	−1.11	−12.74	−1.19
88	4-bromophenyl acetate	−15.3	−13.90	−1.40	−14.42	−0.88	−14.36	−0.94
89	3-bromophenyl acetate	−15.24	−14.52	−0.72	−14.42	−0.82	−14.36	−0.88
90	4-iodophenyl acetate	−17.13	−14.52	−2.61	−16.29	−0.84	−16.25	−0.88
91	3-iodophenyl acetate	−17.53	−15.93	−1.60	−16.29	−1.24	−16.25	−1.28
92	4-nitrophenyl acetate	−12.16	−13.69	1.53	−13.66	1.50	−13.46	1.30
93	acetophenone	−12.96	−12.58	−0.38	−12.22	−0.74	−12.72	−0.24
94	phenyl acetate	−11.99	−13.29	1.30	−12.63	0.64	−12.53	0.54
95	methyl benzoate	−14.28	−14.84	0.56	−15.13	0.85	−14.53	0.25
96	3-hydroxyacetophenone	−11.76	−11.94	0.18	−12.72	0.96	−13.20	1.44
97	4-hydroxyacetophenone	−12.44	−11.66	−0.78	−12.72	0.28	−13.20	0.76
98	acetoanilide	−12.54	−9.81	−2.73	−12.54	0.0	*b*	
99	*p*-xylene	−13.58	−12.87	−0.71	−12.52	−1.06	−13.04	−0.54
100	ethylbenzene	−14.77	−16.29	1.52	−15.81	1.04	−15.18	0.41
101	phenetole	−14.2	−12.82	−1.38	−13.43	−0.77	−13.65	−0.55
102	2-phenylethanol	−12.27	−14.11	1.84	−11.09	−1.18	−11.91	−0.36
103	3-ethylphenol	−14.84	−16.79	1.95	−16.31	1.47	−15.67	0.83
104	4-ethylphenol	−15.37	−16.48	1.11	−16.31	0.94	−15.67	0.30
105	4-ethoxyphenol	−13.3	−14.14	0.84	−13.93	0.63	−14.14	0.84
106	3-ethoxyphenol	−13.41	−13.44	0.03	−13.93	0.52	−14.14	0.73
107	3,5-dimethoxyphenol	−13.36	−13.75	0.39	−13.78	0.42	−13.46	0.10
108	*N*-ethylaniline	−13.33	−12.90	−0.43	−13.33	0.0	−13.33	0.0
109	*N*,*N*-dimethylaniline	−13.48	−12.70	−0.78	−13.48	0.0	−13.48	0.0
110	barbital	−10.19	−11.52	1.33	−11.01	0.82	−9.27	−0.92
111	cyclooctanol	−18.84	−16.31	−2.53	−17.15	−1.69	−19.15	0.31
112	1-octanol	−18.1	−15.76	−2.34	−17.57	−0.53	−18.62	0.52
113	2-octanol	−17.87	−16.60	−1.27	−16.95	−0.92	−17.91	0.04
114	quinoline	−12.1	−15.50	3.40	−12.10	0.0	−12.10	0.0
115	3-cyanophenyl acetate	−8.51	−11.39	2.88	−10.73	2.22	−10.70	2.19
116	4-hydroxycinnamic acid	−16.15	−14.55	−1.60	−16.67	0.52	−16.53	0.38
117	ethyl benzoate	−15.59	−16.57	0.98	−15.71	0.12	−15.51	−0.08
118	4′-hydroxypropiophenone	−15.01	−13.92	−1.09	−15.66	0.65	−13.26	−1.75
119	3′-hydroxypropiophenone	−14.9	−14.77	−0.13	−15.66	0.76	−13.26	−1.64
120	*p*-tolyl acetate	−14.21	−15.17	0.96	−12.68	−1.53	−12.86	−1.35
121	3-methylphenyl acetate	−12.61	−15.54	2.93	−12.68	0.07	−12.86	0.25
122	4-methoxyphenyl acetate	−13.99	−15.10	1.11	−13.06	−0.93	−12.83	−1.16
123	4-propylphenol	−20.26	−17.59	−2.67	−18.78	−1.48	−18.54	−1.72
124	3-propylphenol	−18.72	−18.01	−0.71	−18.78	0.06	−18.54	−0.18
125	4-isopropylphenol	−20.43	−19.09	−1.34	−20.08	−0.35	−19.53	−0.90
126	3-isopropylphenol	−19.64	−18.87	−0.77	−20.08	0.44	−19.53	−0.11
127	4-isopropoxyphenol	−16.33	−17.58	1.25	−14.95	−1.38	−16.50	0.17
128	2-norbornaneacetate	−20.5	−18.68	−1.82	−22.18	1.68	−20.49	−0.01
129	1-benzylimidazole	−14.92	−15.48	0.56	*b*		*b*	
130	*m*-methylcinnamic acid	−16.75	−17.39	0.64	−16.23	−0.52	−16.37	−0.38
131	4-ethylphenyl acetate	−16.15	−17.82	1.67	−16.01	−0.14	−15.34	−0.81
132	3-ethylphenyl acetate	−15.3	−17.33	2.03	−16.01	0.71	−15.34	0.04
133	4-ethoxyphenyl acetate	−14.5	−15.92	1.42	−13.64	−0.86	−13.81	−0.69
134	3-ethoxyphenyl acetate	−14.21	−17.04	2.83	−13.64	−0.57	−13.81	−0.40
135	allobarbital	−11.32	−11.26	−0.06	−11.32	0.0	*b*	
136	4-*n*-butylphenol	−22.66	−19.98	−2.68	−21.25	−1.41	−21.41	−1.25
137	3-*n*-butylphenol	−21.46	−20.07	−1.39	−21.25	−0.21	−21.41	−0.05
138	3-isobutylphenol	−24.03	−19.77	−4.26	−21.69	−2.34	−22.79	−1.24
139	4-*sec*-butylphenol	−23.86	−20.22	−3.64	−21.28	−2.58	−21.55	−2.31

Table 1 (Continued)

		ΔG, kJ/mol						
			CODESSA-PRO		TRAIL (Type 1)		TRAIL (Type 2)	
no.	name of compound	exptl	calcd	exptl − calcd	calcd	exptl − calcd	calcd	exptl − calcd
140	3-*sec*-butylphenol	−23.18	−20.06	−3.12	−21.28	−1.90	−21.55	−1.63
141	4-*tert*-butylphenol	−26.03	−19.73	−6.30	−24.28	−1.75	−24.78	−1.25
142	3-*tert*-butylphenol	−25.18	−20.37	−4.81	−24.28	−0.90	−24.78	−0.40
143	menadion	−12.95	−14.96	2.01	−12.95	0.0	*b*	
144	sulfapyridine	−15.43	−14.57	−0.86	−15.43	0.0	−15.43	0.0
145	sulfamonomethoxine	−14.16	−12.16	−2.00	−14.16	0.0	−14.16	0.0
146	sulfisoxazole	−13.25	−14.22	0.97	*b*		*b*	
147	4-*n*-propylphenyl acetate	−17.98	−18.67	0.69	−18.49	0.51	−18.21	0.23
148	3-*n*-propylphenyl acetate	−18.72	−18.81	0.09	−18.49	−0.23	−18.21	−0.51
149	4-isopropylphenyl acetate	−16.44	−19.52	3.08	−19.79	3.35	−19.20	2.76
150	3-isopropylphenyl acetate	−19.18	−19.95	0.77	−19.79	0.61	−19.20	0.02
151	4-*n*-amylphenol	−23.92	−21.28	−2.64	−23.73	−0.19	−24.28	0.36
152	4-*tert*-amylphenol	−26.83	−21.81	−5.02	−24.22	−2.61	−25.94	−0.89
153	carbutamide	−13.08	−14.81	1.73	−13.08	0.0	*b*	
154	pentobarbital	−17.22	−16.53	−0.69	−15.06	−2.16	−16.40	−0.82
155	amobarbital	−17.53	−17.53	0.00	−18.87	1.34	−19.27	1.74
156	thiopental	−18.71	−19.78	1.07	*b*		*b*	
157	dibenzofuran	−16.95	−16.46	−0.49	*b*		*b*	
158	dibenzothiophene	−19.87	−17.30	−2.57	−19.87	0.0	*b*	
159	phenazine	−13.76	−17.93	4.17	*b*		*b*	
160	thianthrene	−20.38	−17.27	−3.11	*b*		*b*	
161	carbazole	−13.93	−16.78	2.85	−13.93	0.0	*b*	
162	phenoxazine	−15.36	−16.75	1.39	*b*		*b*	
163	phenothiazine	−15.59	−16.30	0.71	*b*		*b*	
164	furosemide	−10.19	−10.94	0.75	*b*		*b*	
165	phenobarbital	−18.37	−11.07	−7.30	−17.77	−0.60	−17.54	−0.83
166	sulfisomidine	−12.02	−13.69	1.67	−12.39	0.37	−12.45	0.43
167	sulfamethomidine	−13.31	−12.89	−0.42	−12.75	−0.56	−12.66	−0.65
168	sulfadimethoxine	−12.89	−12.81	−0.08	−13.08	0.19	−13.11	0.22
169	4-*n*-butylphenyl acetate	−20.66	−20.87	0.21	−20.96	0.30	−21.08	0.42
170	3-*n*-butylphenyl acetate	−20.89	−20.17	−0.72	−20.96	0.07	−21.08	0.19
171	3-isobutylphenyl acetate	−21.86	−21.01	−0.86	−21.40	−0.46	−22.47	0.61
172	4-*tert*-butylphenyl acetate	−21.98	−21.00	−0.98	−23.99	2.01	−24.45	2.47
173	cyclobarbital	−15.47	−15.87	0.40	−16.07	0.60	−16.30	0.83
174	hexobarbital	−17.6	−16.99	−0.61	−17.00	−0.60	−16.77	−0.83
175	1-adamantaneacetate	−24.69	−23.51	−1.18	−24.94	0.25	−24.69	0.0
176	acridine	−13.3	−18.72	5.42	−13.30	0.0	−13.30	0.0
177	phenanthridine	−14.67	−16.39	1.72	−14.67	0.0	−14.67	0.0
178	xanthene	−15.47	−14.98	−0.49	−15.47	0.0	*b*	
179	*N*-phenylanthranilic acid	−16.53	−19.16	2.63	−16.53	0.0	*b*	
180	mephobarbital	−18.05	−12.93	−5.12	−18.65	0.60	−18.88	0.83
181	4-*n*-amylphenyl acetate	−21.69	−20.86	−0.83	−23.44	1.75	−23.95	2.26
182	flufenamic acid	−17.7	−15.63	−2.07	*b*		*b*	
183	meclofenamic acid	−15.27	−18.02	2.75	*b*		*b*	
184	nitrazepam	−11.26	−16.13	4.87	−11.26	0.0	−11.26	0.0
185	flurbiprofen	−21.07	−16.54	−4.53	−21.18	0.11	−20.69	−0.38
186	sulfaphenazole	−13.42	−15.60	2.18	*b*		*b*	
187	bendroflumethiazide	−10.83	−13.38	2.55	*b*		*b*	
188	mefenamic acid	−14.24	−19.93	5.69	−14.24	0.0	*b*	
189	acetohexamide	−16.77	−16.29	−0.48	−16.77	0.0	*b*	
190	fludiazepam	−13.31	−11.67	−1.64	−13.20	−0.11	−13.69	0.38
191	nimetazepam	−9.89	−14.36	4.47	−10.00	0.11	−9.51	−0.38
192	fenbufen	−15.02	−15.40	0.38	−14.91	−0.11	−15.40	0.38
193	ketoprofen	−16.31	−15.67	−0.64	−18.66	2.35	−18.01	1.70
194	medazepam	−13.72	−15.20	1.48	−13.72	0.0	−13.72	0.0
195	progabide	−14.48	−12.65	−1.83	−14.48	0.0	*b*	
196	griseofulvin	−8.39	−19.78	11.39	*b*		*b*	
197	tolnaftate	−21.9	−15.80	−6.10	*b*		*b*	
198	prostacyclin	−16.79	−20.13	3.34	*b*		*b*	
199	triamcinolone	−19.27	−17.94	−1.33	−18.44	−0.83	*b*	
200	cortisone	−19.1	−20.02	0.92	−19.77	0.67	−19.57	0.47
201	prednisolone	−20.3	−19.34	−0.96	−21.27	0.97	−20.58	0.28
202	hydrocortisone	−20.58	−22.22	1.64	−20.37	−0.21	−20.01	−0.57
203	corticosterone	−22	−22.52	0.52	−22.00	0.0	−22.00	0.0
204	dexamethasone	−20.84	−19.24	−1.60	−21.67	0.83	−20.84	0.0
205	betamethasone	−21.31	−20.11	−1.20	−21.31	0.0	*b*	
206	paramethasone	−19.44	−19.19	−0.25	−20.05	0.61	−19.44	0.0
207	cortisone-21-acetate	−20.65	−18.86	−1.79	−18.40	−2.25	−20.18	−0.47
208	prednisolone-21-acetate	−21.47	−19.35	−2.12	−19.89	−1.58	−21.19	−0.28
209	hydrocortisone-21-acetate	−20.05	−19.71	−0.34	−18.99	−1.06	−20.62	0.57

Table 1 (Continued)

		ΔG, kJ/mol						
			CODESSA-PRO		TRAIL (Type 1)		TRAIL (Type 2)	
no.	name of compound	exptl	calcd	exptl − calcd	calcd	exptl − calcd	calcd	exptl − calcd
210	fluocinolone acetonide	−19.85	−17.81	−2.04	−19.24	−0.61	−19.85	0.0
211	triamcinolone acetonide	−20.03	−20.69	0.66	−20.64	0.61	−20.03	0.0
212	spironolactone	−25.34	−24.66	−0.68	*b*		*b*	
213	dehydrocholic acid	−19.29	−19.90	0.61	−18.48	−0.81	−19.29	0.0
214	chenodeoxycholic acid	−24.9	−25.28	0.38	−26.12	1.22	−25.31	0.41
215	ursodeoxycholic acid	−25.71	−25.53	−0.18	−26.12	0.41	−25.31	−0.40
216	cholic acid	−19.97	−23.57	3.60	−20.26	0.29	−19.97	0.0
217	hydrocortisone-17-butyrate	−18.43	−21.03	2.60	−21.29	2.86	*b*	
218	cinnarizine	−20.77	−17.70	−3.07	*b*		*b*	

[a] Calculations for the full set were done using the seven-descriptor model (CODESSA) and the linear I(AB,2−4) model (TRAIL). Calculations with TRAIL were performed taking into account (Type 2) or without taking into account (Type 1) the difference between bonds in cycles and in chains. [b] Compound was excluded on the training stage because it contains unique molecular fragment(s).

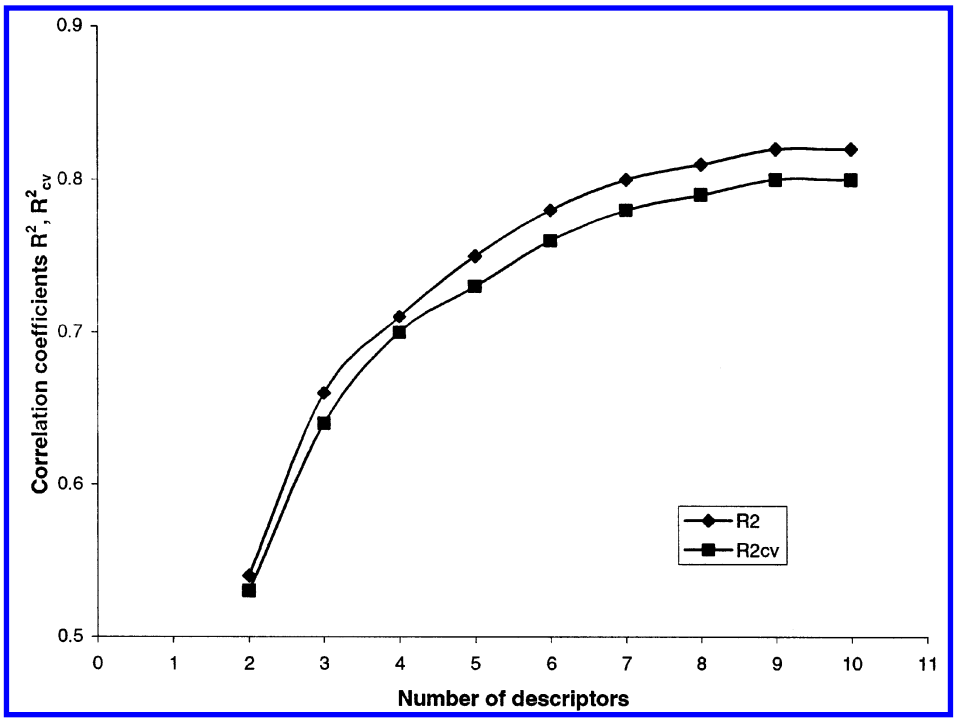

Figure 1. Correlation coefficients vs number of descriptors in the BMLR equation for the full set of compounds.

Table 2. The Best Seven-Parameter Correlation for the Full Set of 218 Organic Compounds ($F = 117.3$)[a]

no.	X	$\pm\Delta X$	t-test	R^2	R_{cv}^2	s^2	descriptor
0	$-3.062 \times 10^{+00}$	$2.383 \times 10^{+00}$	−1.285				intercept
1	$-8.902 \times 10^{+00}$	4.096×10^{-01}	−21.733	0.392	0.380	16.44	average complementary information content (order 0), ${}^0\overline{\mathrm{CIC}}$
2	$2.905 \times 10^{+00}$	1.953×10^{-01}	14.876	0.542	0.529	12.45	LUMO energy, ϵ_{LUMO}
3	1.081×10^{-01}	9.349×10^{-03}	11.567	0.657	0.643	9.37	WNSA-1-weighted PNSA (PNSA1*TMSA/1000), WNSA
4	$4.333 \times 10^{+02}$	4.410×10^{01}	9.823	0.693	0.680	8.41	HACA-2/TMSA (Mopac PC), HACA
5	$-5.009 \times 10^{+01}$	$7.567 \times 10^{+00}$	−6.619	0.747	0.732	6.97	maximum partial charge (Zefirov) for all atom types, Q_i
6	$-8.733 \times 10^{+00}$	1.634×10^{-02}	−5.346	0.774	0.759	6.26	H-acceptors FPSA (version 2), FPSA
7	1.401×10^{01}	$2.890 \times 10^{+00}$	4.848	0.796	0.779	5.66	ZX shadow/ZX rectangle, ${}_SZX^R$

[a] R, correlation coefficient; R_{cv}, cross-validated correlation coefficient; s^2, square of standard error; F, Fischer criterion value.

Thus, the average complementary information content of zeroth order (${}^0\overline{\mathrm{CIC}}$), based on information theory, may measure the loss of entropy of the guest species during the complexation.

The charge-distribution-related descriptors HACA-2/TMSA (HACA), H-acceptors FPSA, version 2 (FPSA), maximum partial charge—Zefirov—for all atom types (Q_i), and WNSA-1-weighted PNSA (WNSA) describe the electrostatic component of the host−guest interaction energy and account for formation of hydrogen bonds in the complexes of β-CD.

ZX shadow/*ZX* rectangle (${}_SZX^R$) is one of the molecular shadow indices defined as the projection of the molecule's van der Waals envelope on the *ZX* plane, where *Z* and *X* are the longest and shortest inertial axes of the molecule, respectively. The shadow areas are usually calculated by applying a two-dimensional square grid on the molecular projection and by the subsequent summation of the areas of

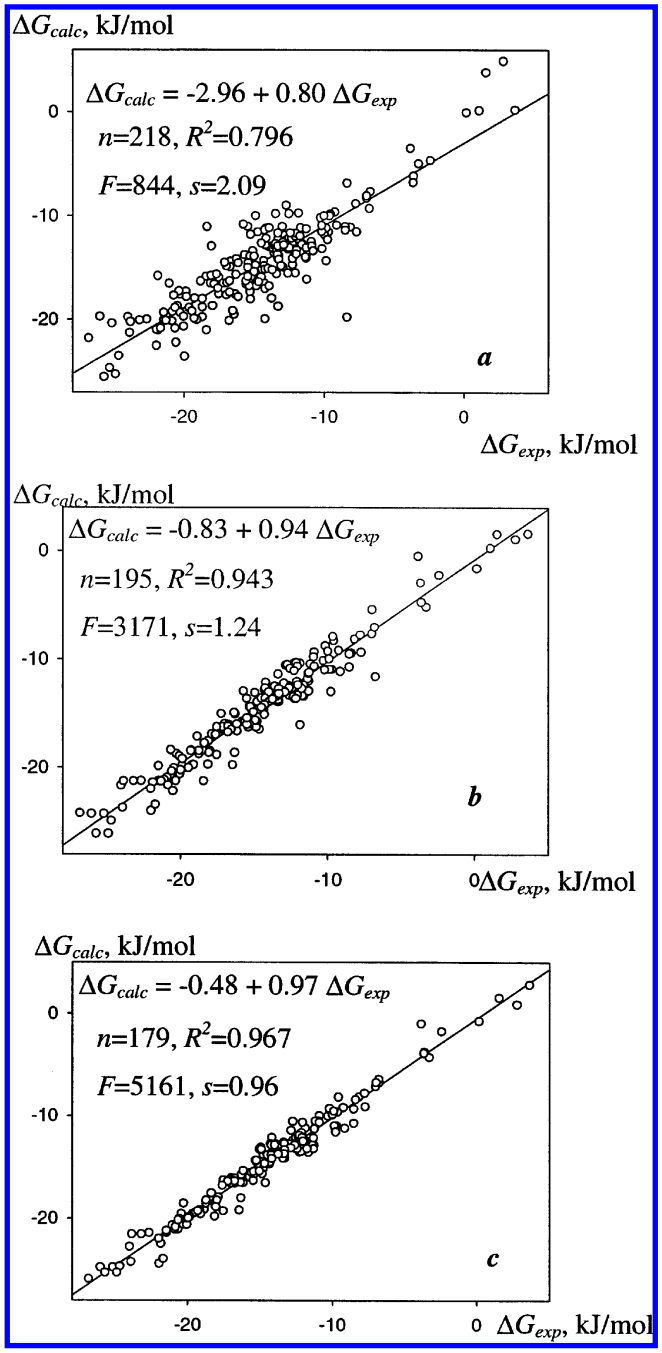

Figure 2. Calculated vs experimental complexation free energies of β-cyclodextrin with neutral guests in water. Calculations were performed for the full set of compounds with CODESSA-PRO using seven descriptors model (top) and with TRAIL using the linear I(AB,2−4) model, where the difference between bonds in cycles and in chains was not taken into account (middle) or where cyclic bonds and chain bonds were considered as different (bottom).

squares overlapped with the projection. The presence of this descriptor in our model underlines the importance of the molecular size for CDs complexation and accounts for the van der Waals component of the host−guest interaction energy.

The energy of the LUMO (expressed here by the descriptor LUMO energy, ϵ_{LUMO}) accounts for the electron affinity (EA) of a molecule and might also be related to polarization of one partner of the complexes in the electric field of another one. Although the polarization effects in the β-CD complexes are rather weak, they are not negligible.[34]

Table 3. Internal Validation of the Seven-Descriptor Model Calculated by Using the CODESSA-PRO Program for Three Training and Test Sets[a]

set to fit	N	R^2	s^2	set to predict	N	R^2(pred)	s^2(pred)
1 and 2	146	0.800	4.62	3	72	0.785	4.75
1 and 3	145	0.809	4.12	2	73	0.759	4.76
2 and 3	145	0.790	4.30	1	73	0.796	4.77
average		0.800	4.35			0.780	4.76

[a] N is the size of the set. R^2 and s^2 are the squared correlation coefficient and the squared standard deviation of calculations performed on the training sets. R^2(pred) and s^2(pred) are the squared correlation coefficient and the squared standard deviation of prediction calculations performed on the test sets.

Table 4. TRAIL Modeling for the Full Set of Compounds Using the Linear I(AB,2−4) Model[a]

type of calculation	n	k	R^2	F	s^2	R_{cv}^2	s_{cv}^2
Type 1[b]	195	80	0.943	23.9	2.72	0.848	9.42
Type 2[c]	179	90	0.967	29.2	1.88	0.877	13.49

[a] Molecules are represented without hydrogen atoms. Statistical parameters for the fitting on the training set: number of compounds (n), number of fitted coefficients (k), correlation coefficient (R), Fisher's criterion (F), standard deviation (s), cross-validation correlation coefficient (R_{cv}), and cross-validation standard deviation (s_{cv}). [b] The difference between bonds in cycles and in chains was not taken into account. [c] Bonds in cycles and chains are considered as different.

The cross-validated correlation of the model gives an R_{cv}^2 value 0.779. For the internal validation, the data set was divided into three subsets (the first, fourth, seventh, etc. entries go into the first subset, the second, fifth, eighth, etc. go into the second subset, and the third, sixth, ninth, etc. go into the third subset). For each of the three combinations, two of the subsets were combined into one and the correlation equation was derived with the same descriptors. The obtained equation was used to predict the remaining subset. The results show that predicted R^2 values are in good agreement with the original QSPR model, with average correlation coefficients of 0.800 and 0.780 for the fitted and predicted sets, respectively (Table 3).

TRAIL Calculations. Both Type 1 and Type 2 calculations performed on the parent set of 218 compounds led to about 40 models with $R^2 > 0.8$. However, only three models corresponded to R_{cv}^2 values larger than 0.5. By applying these three models for calculations on three training and test sets, we have found a single model which leads to reasonable statistical criteria for any set. This is a linear model I(AB,2−4) which involves the sequences of atoms and bonds containing from two to four atoms (Table 4).

In the Type 1 calculations on the full set, TRAIL has excluded 23 compounds having fragments of "rare" occurrence (less than 2), thus performing fitting on 195 molecules using 79 fragmental descriptors. This led to a linear correlation with $R^2 = 0.943$, $F = 23.9$, $s^2 = 2.72$, $R_{cv}^2 = 0.848$, and $s_{cv}^2 = 9.42$ (Table 4). Calculated values for the compounds from the full set are quite close to those obtained experimentally (Table 1; Figure 2, middle). The most significant deviation of calculated free energies from the experiment was found for cyclobutanol (4.89 kJ/mol) and for cyclopentanol (4.20 kJ/mol).

Table 5. Internal Validation of the Linear I(AB,2−4) Model Calculated by Using the TRAIL Program for Three Training and Test Sets[a]

	type of calculation	n	k	R^2	F	s^2	R_{cv}^2	s_{cv}^2	R^2(pred)	s^2(pred)
Set 1										
	Type 1	127	65	0.951	18.6	2.95	0.817	16.30	0.632	10.43
	Type 2	112	61	0.967	25.1	2.08	0.877	11.06	0.798	4.84
Set 2										
	Type 1	116	57	0.958	23.9	2.32	0.800	14.38	0.737	7.95
	Type 2	103	55	0.975	34.7	1.46	0.921	7.09	0.725	8.01
Set 3										
	Type 1	114	56	0.935	15.1	3.60	0.788	15.98	0.831	4.79
	Type 2	101	59	0.970	23.3	2.21	0.868	18.76	0.827	5.81
	average Type 1			0.948		2.96	0.801	15.55	0.733	8.80
	average Type 2			0.971		1.92	0.888	12.30	0.783	6.22

[a] See footnotes for Tables 3 and 4.

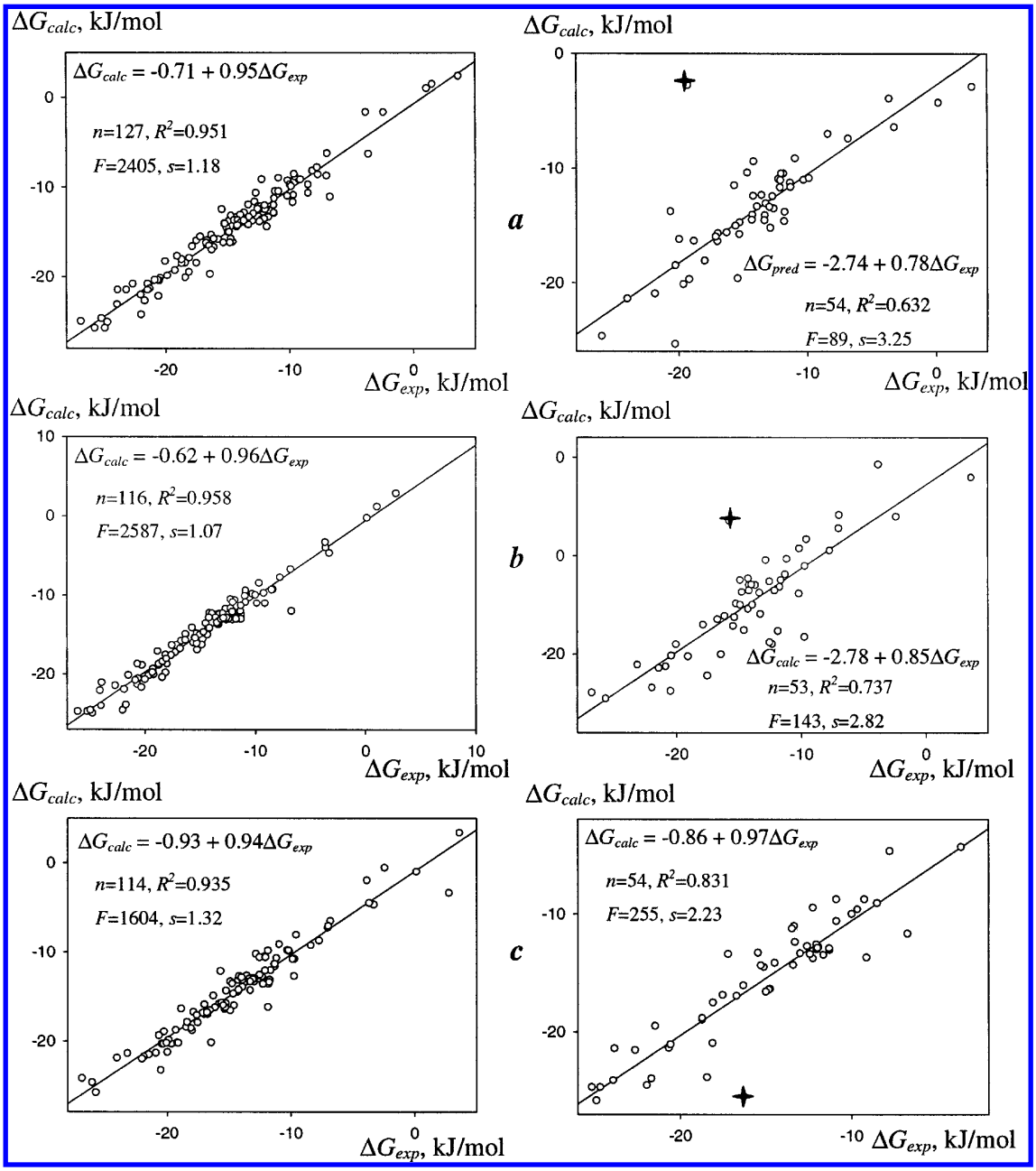

Figure 3. TRAIL modeling: linear correlation between experimental and calculated (left) and experimental and predicted (right) ΔG values for the three training/test sets using the linear I(AB,2−4) model. The difference between bonds in cycles and in chains was not taken into account (Type 1 calculations). The + symbol indicates the outliers in the test sets.

Differentiation of similar bonds in cycles and chains (Type 2 calculations) increases the number of variables to 90 and that of excluded molecules to 39 (Table 4). However, most statistical criteria of fitting and cross-validation become better

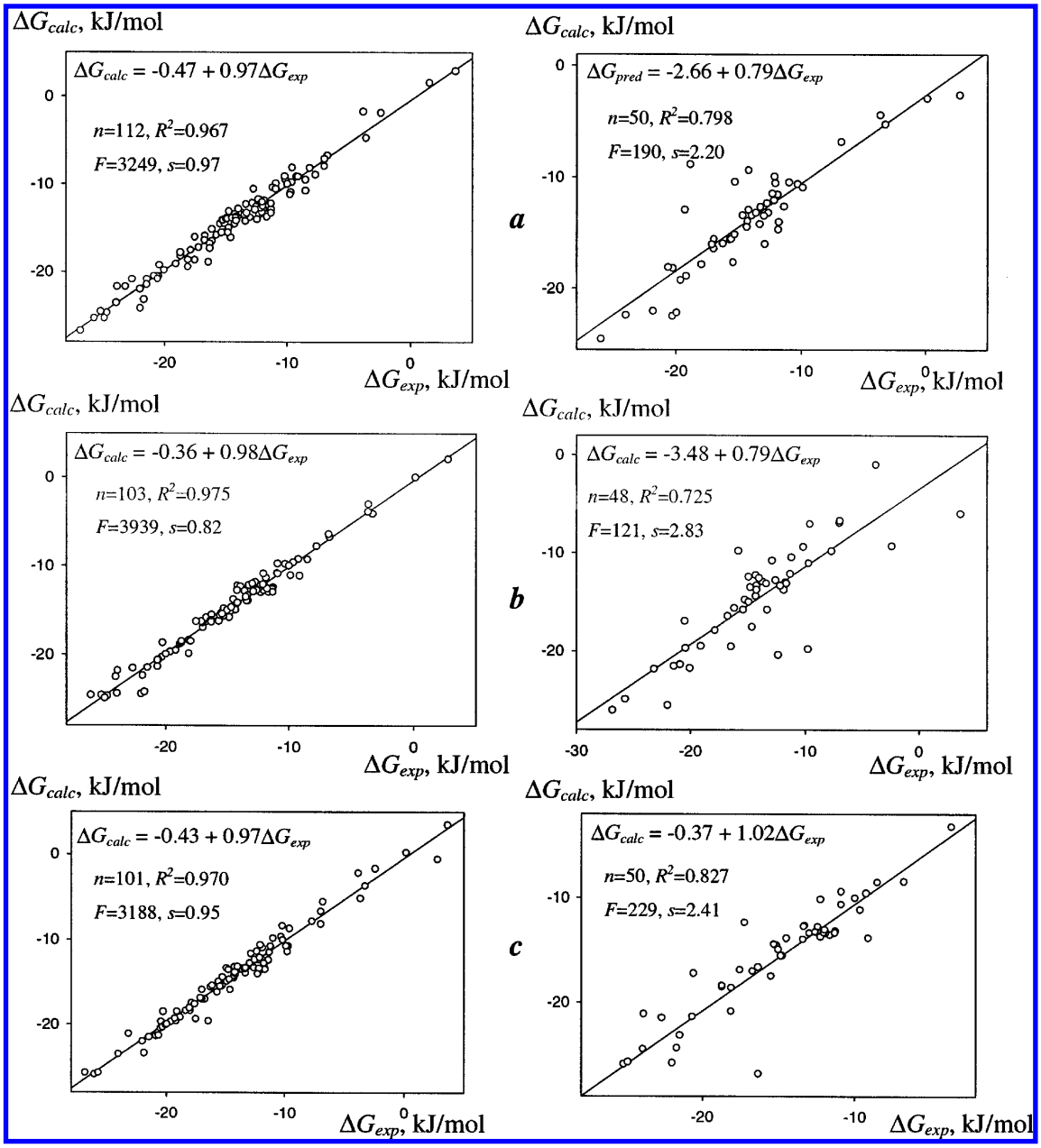

Figure 4. TRAIL modeling: linear correlation between experimental and calculated (left) and experimental and predicted (right) ΔG values for the three training/test sets using the linear I(AB,2−4) model. Bonds in cycles and bonds in chains are considered as different (Type 2 calculations).

than those for Type 1 calculations ($R^2 = 0.967$, $F = 29.2$, $s^2 = 1.88$, $R_{cv}^2 = 0.877$, and $s_{cv}^2 = 13.49$ (Table 4; Figure 2, bottom).

For a better comparison between these two different approaches, the same seven theoretical molecular descriptors involved in the CODESSA-PRO model were correlated with the ΔG of complexation of β-cyclodextrins for two reduced data sets: (i) 195 and (ii) 179 data points, respectively. The same molecules were eliminated from the full data set used as in the case of TRAIL approach (see Table 1). The statistical characteristics for these two models obtained with CODESSA-PRO are as follow: (i) $N = 195$, $n = 7$, $R^2 = 0.832$, $R_{cv}^2 = 0.817$, $F = 132.0$, and $s^2 = 4.90$, and (ii) $N = 179$, $n = 7$, $R^2 = 0.838$, $R_{cv}^2 = 0.822$, $F = 126.4$, and $s^2 = 4.78$. These show a slightly improvement of the correlations, which means that those seven parameters proposed to describe ΔG of complexation of β-cyclodextrins were well selected and they are not highly dependent on the size and/or composition of the data set.

Validation calculations with the I(AB,2−4) model were performed on three training/test subsets as in the CODESSA-PRO calculations. A certain number of compounds containing "rare" fragments were excluded at the training stage. TRAIL also excluded from the test sets molecules whose fragments were not presented in the compounds of corresponding training sets. The results given in Table 5 show that average statistical criteria obtained at the training ($R^2 = 0.948$, $s^2 = 2.96$ and $R^2 = 0.971$, $s^2 = 1.92$ for Type 1 and Type 2 calculations, respectively) and test (R^2(pred) = 0.733, s^2(pred) = 8.80 and R^2(pred) = 0.783, s^2(pred) = 6.22 for Type 1 and Type 2 calculations, respectively) stages are close to those obtained in CODESSA-PRO calculations.

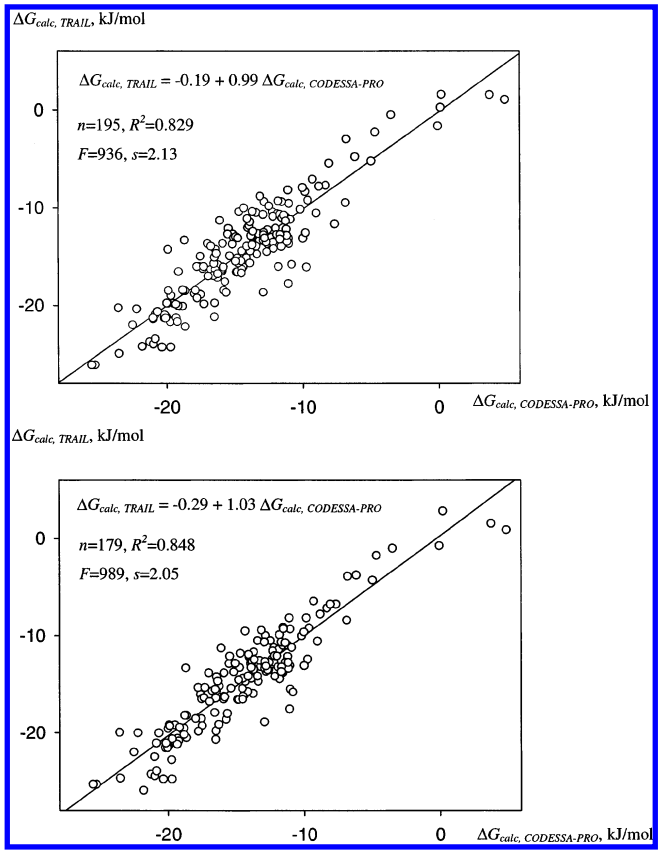

Figure 5. TRAIL vs CODESSA-PRO-calculated complexation free energy ΔG of β-cyclodextrin with neutral guest for the full set of compounds using the linear I(AB,2−4) model (TRAIL) and the seven-descriptors model (CODESSA-PRO). The two plots correspond to Type 1 (top) and Type 2 (bottom) calculations with TRAIL.

Table 6. Prediction of Free Energy of Complexation for β-CD with Organic Guests using the Seven-Descriptor Model (CODESSA-PRO) and the Linear I(AB,2−4) Model (TRAIL)

		ΔG, kJ/mol			
			predicted		
			CODESSA-PRO	TRAIL	
no.	guest	exptl[a]		Type 1	Type 2
1	3-aminobenzoic acid	−10.3	−11.7	−9.1	−10.5
2	3-hydroxybenzoic acid	−14.0	−13.0	−10.9	−12.1
3	3-nitrophenol	−13.9	−12.3	−14.0	−13.8

[a] Experimental data from Rekharsky et al.[35] The three molecules *do not* belong to the parent set of 218 compounds.

It should be noted that, in Type 1 calculations, each test set contains only one outlier (Figure 3), whose omission leads to substantial improvement of the statistical criteria of prediction calculations (R^2(pred) = 0.811, s^2(pred) = 5.28). No outliers were obtained with Type 2 calculations (Figure 4).

The free energies of complexation for the full set of compounds calculated with TRAIL correlate well with those obtained by CODESSA-PRO ($R^2 = 0.829$, $s^2 = 4.54$ and $R^2 = 0.848$, $s^2 = 4.20$ for Type 1 and Type 2 calculations, respectively; Figure 5). In principle, this may allow one to apply both models simultaneously to estimate the binding affinity of any other guest molecule similar to those used at the training stage.

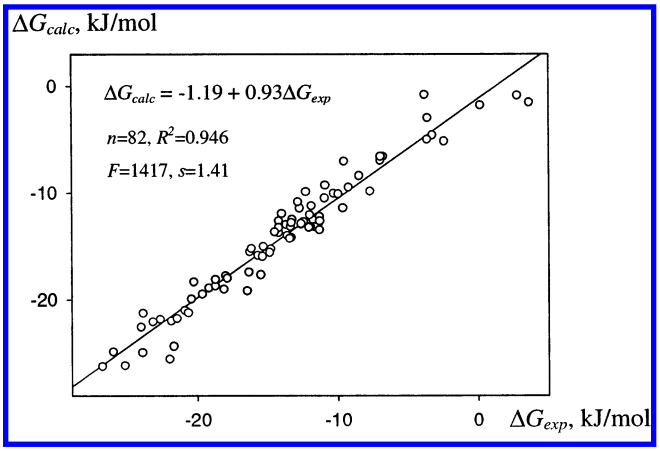

Figure 6. Linear correlation between experimental ΔG values and those predicted with the TRAIL program for the three test sets using the linear I(AB,2−4) (Type 1) model. Compounds whose fragments occur *less than in six structures* were excluded from the three training sets.

DISCUSSION

In this section we discuss general features of application of the TRAIL and CODESSA techniques to build statistically robust QSPR models.

The main limitation of all Free−Wilson-type methods, and in particular of the SMF method realized in TRAIL, is the requirement to have several examples of any particular molecular fragment in the training set. Molecules containing rarely occurring fragments have to be eliminated from the training set, thus reducing the number of treated compounds. On the other hand, fragmental descriptors are chemically easy to interpret. Indeed, calculations with TRAIL generate the list of the fragments' contributions to a given property. One can select the fragments whose contributions are considerable and give reasonable explanations based on chemical interactions between selected fragments (or any part of a molecule which is a combination of TRAIL fragments) and the biological or chemical target.

It is shown in this paper that, by using both whole molecule descriptors and the contribution of fragments, a given property for any molecule can be calculated easily.

To illustrate this, we used both CODESSA-PRO and TRAIL models to estimate ΔG of complexation of β-cyclodextrins for three organic molecules which were not in the list of 218 molecules. As one may see from Table 6, the calculated free energies for these compounds are not far from the experimental data.

It should be also noted that most of the "predictors" incorporated in commercial software (ChemOffice, ACD-Labs) are based on fragmental contributions. This is also a case with the well-known Leo and Hansch method of estimation of partition coefficients for octanol/water (log P).[15] Recently, one of our groups has modeled log P by TRAIL and obtained reasonable results for a set of test drug molecules.[29] In contrast to the Leo and Hansch method, where the fragments were selected manually, TRAIL does this job automatically, selecting the best fragmentation schemes.

Generally, the Free−Wilson-type method uses a larger number of variables compared to the Hansch-type method. With a Hansch-type method (CODESSA calculations), an

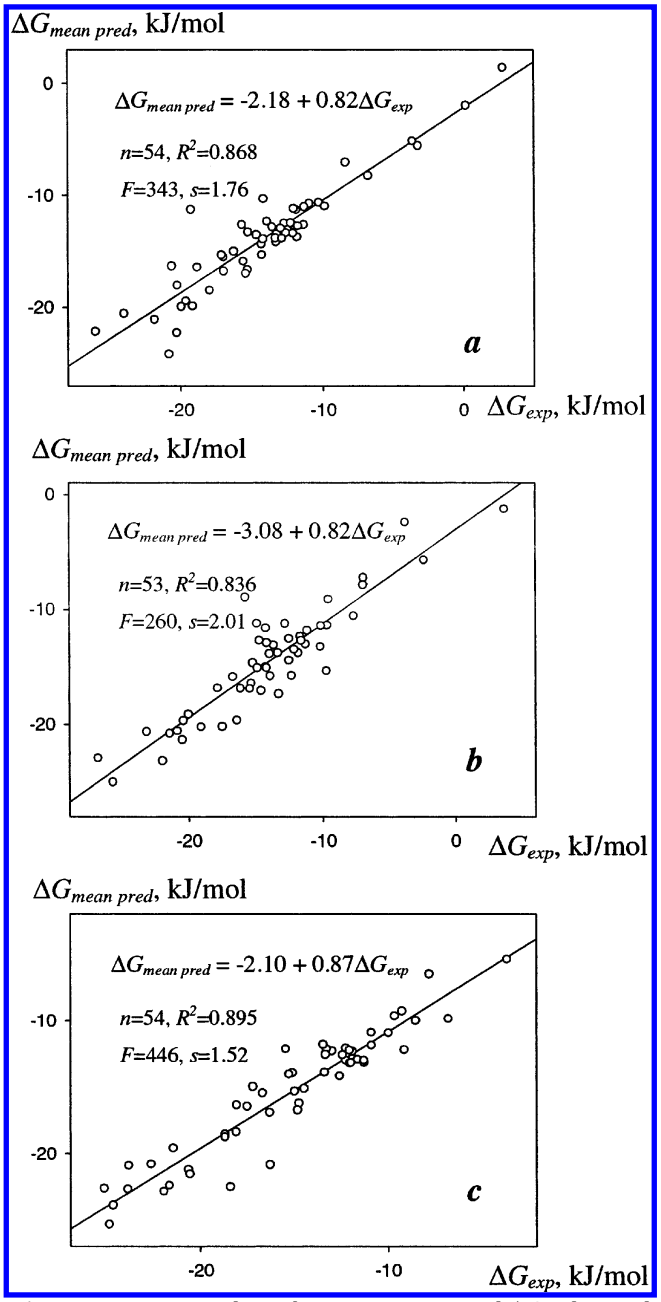

Figure 7. Linear correlation between experimental ΔG values and those predicted from the arithmetic mean by CODESSA-PRO/TRAIL (Type 1) for the three test sets. The average values of R^2 and s^2 over three tests sets are 0.866 and 3.15, respectively.

increase of the number of descriptors often leads to overfitted models when R^2(fit) becomes better but R^2(pred) decreases. Our experience shows that TRAIL produces statistically stable QSPR models despite the relatively large number of variables. Indeed, differentiation of bonds in cycles and in chains (see the Results section) led to an increase of the number of variables for the whole set, but the statistical criteria of fitting (R^2) and *Leave One Out* prediction (R_{cv}^2) become better (Table 4; Figure 2). Similar conclusions can be drawn when comparing statistical criteria of predictions for test sets 1−3 (Table 5; Figures 3 and 4).

Another possibility to improve the robustness of the TRAIL model is to change the threshold for acceptance of molecules having fragments of "rare occurrence". By default, a molecule is accepted by TRAIL if its constituent fragments occur in at least two molecules; otherwise it is excluded from the training set. Generally, an increase of the occurrence threshold leads to an improved quality of prediction, although the number of compounds excluded at the training stage also becomes larger. Thus, TRAIL modeling with occurrence threshold equal to 6 resulted in excellent correlation ($R^2 = 0.946$, $s^2 = 1.99$) between experimental and "predicted" values for three test sets (Figure 6).

The fragmental approach can be efficiently applied in "lead optimization" studies, particularly when one looks for "optimal" substituents, corresponding to molecules with desired properties. Combination of the TRAIL technique with the generator of virtual combinatorial libraries may allow one to generate relatively small focused libraries of compounds similar to the lead molecule.[32]

An important aspect of any QSPR study is the quality of predictions, which may vary from one particular method to another one. One could believe that averaging of values calculated with different approaches smoothes the inaccuracies of the predictions. Thus, the arithmetic mean of the ΔG values predicted by TRAIL and CODESSA-PRO for the three test sets correlates with the experimental results better than the values calculated by either of these programs alone. Indeed, the average values of R^2 and s^2 over three tests sets are 0.780 and 4.76 (CODESSA-PRO, Table 3), 0.733 and 8.80 (TRAIL, Table 5), and 0.866 and 3.15 (CODESSA-PRO/TRAIL, Figure 7).

Another question is related to the application of molecular fragments as external descriptors of CODESSA. Technically, CODESSA is able to incorporate any atom- or group-based descriptor, but a priori it is not clear how many there should be and what types of fragments are more appropriate. We believe that the fragments corresponding to the best TRAIL models can be efficiently applied as "external" descriptors for CODESSA-PRO. Thus, such a combination of TRAIL and CODESSA-PRO looks very promising, and we plan further work in this direction.

CONCLUSIONS

QSPR modeling of a large data set of experimental free energies of complexation for β-cyclodextrins with 218 organic compounds from different classes has been performed using two different techniques. One of them, realized in the program CODESSA-PRO, applies up to 902 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors. Another one, realized in the TRAIL program, uses several types of various fragmental descriptors, defined in the substructural molecular fragments method. The two approaches individually and in combination led to statistically stable and predictive QSPR models. Indeed, the best models selected at the training stage for the full set of compounds, and then applied for three training sets containing two-thirds of all molecules, led to correct predictions of free energies for the compounds of the corresponding test sets.

ACKNOWLEDGMENT

The authors thank Dr. Alexander Oliferenko for providing valuable comments that enhanced the quality of this study. Audray Le Ngoc is acknowledged for help with TRAIL calculations.

Supporting Information Available: Tables showing the modeling of free energy for the complexation of β-cyclodextrin with organic molecules in water at 298 K using the CODESSA-PRO program, and using the TRAIL program with or without taking into account the differences between bonds in cycles and in chains and with compounds whose fragments occur in less than six structures excluded from the training sets; a figure modeling the free energy for the complexation of β-cyclodextrin with organic molecules in water at 298 K with the CODESSA-PRO program. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Connors, K. A. The Stability of Cyclodextrin Complexes in Solution. *Chem. Rev.* **1997**, *97*, 1325−1357.
(2) Szejtli, J. Introduction and General Overview of Cyclodextrin Chemistry. *Chem. Rev.* **1998**, *98*, 1743−1753.
(3) Hedges, A. R. Industrial Applications of Cyclodextrins. *Chem. Rev.* **1998**, *98*, 2035−2044.
(4) Carpignano, R.; Marzona, M.; Cattaneo, E.; Quaranta, S. QSAR Study of Inclusion Complexes of Heterocyclic Compounds with β-Cyclodextrin. *Anal. Chim. Acta* **1997**, *348*, 489−493.
(5) Lipkowitz, K. B. Applications of Computational Chemistry to the Study of Cyclodextrins. *Chem. Rev.* **1998**, *98*, 1829−1873.
(6) Suzuki, T. A Nonlinear Group Contribution Method for Predicting the Free Energies of Inclusion Complexation of Organic Molecules with α- and β-Cyclodextrins. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1266−1273.
(7) Pérez, F.; Jaime, C.; Sánchez-Ruiz, X. MM2 Calculations on Cyclodextrins: Multimodel Inclusion Complexes. *J. Org. Chem.* **1995**, *60*, 3840−3845.
(8) Matsui, Y.; Nishioka, T.; Fujita, T. Quantitative Structure−Reactivity Analysis of the Inclusion Mechanism by Cyclodextrins. *Top. Curr. Chem.* **1985**, *128*, 61−89.
(9) Davis, D. M.; Savage, J. R. Correlation Analysis of the Host−Guest Interaction of α-Cyclodextrin and Substituted Benzenes. *J. Chem. Res. (S)* **1993**, 94−95.
(10) Park, J. H.; Nah, T. H. Binding Forces Contributing to the Complexation of Organic Molecules with β-Cyclodextrin in Aqueous Solution. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1359−1362.
(11) Klein, C. T.; Polheim, D.; Viernstein, H.; Wolschann, P. A Method for Predicting the Free Energies of Complexation between β-Cyclodextrin and Guest Molecules. *J. Inclusion Phenom. Macrocyclic Chem.* **2000**, *36*, 409−423.
(12) Liu, L.; Guo, Q.-X. Wavelet Neural Network and its Application to the Inclusion of β-Cyclodextrin with Benzene Derivatives. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 133−138.
(13) Suzuki, T.; Ishida, M.; Fabian, W. M. F. Classical QSAR and Comparative Molecular Field Analyses of the Host−Guest Interaction of Organic Molecules with Cyclodextrins. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 669−678.
(14) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(15) Hansch, C.; Leo, A.; Hoekman, D. H. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1995; p 580.
(16) Free, S. M., Jr.; Wilson, J. W. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(17) www.codessa-pro.com
(18) 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.
(19) Katritzky, A. R.; Maran, 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.
(20) Katritzky, A. R.; Oliferenko, 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.
(21) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. Ph.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.
(22) 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.
(23) Zefirov, N. S.; Palyulin, V. A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1112−1122.
(24) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432−2437.
(25) 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.
(26) Avidon, V. V. Criteria for a Comparison of Chemical Structures and Principles for the Construction of an Information Language for an Information-Logical System Covering Biologically Active Compounds. *Chim. Pharm. J. (Russ.)* **1974**, *8*, 22−25.
(27) Bawden, D. Computerized Chemical Structure-Handling Techniques in Structure−Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14−22.
(28) Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos, A. Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355.
(29) 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.
(30) Varnek, A.; Wipff, G.; Solov'ev, V. P. Towards an Information System on Solvent Extraction. *Solvent Extr. Ion Exch.* **2001**, *19*, 791−837.
(31) 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.
(32) Solov'ev, V. P.; Varnek, A. Anti-HIV Activity of HEPT, TIBO and Cyclic Urea Derivatives: Structure−Property Studies, Focused Combinatorial Library Generation and Hits Selection Using Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 1703−1719.
(33) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, New York, Chichester, 2000.
(34) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; J. Wiley & Sons: New York, 2000.
(35) Rekharsky, M. V.; Inoue, Y. Complexation Thermodynamics of Cyclodextrins. *Chem. Rev.* **1998**, *98*, 1875−1917.

CI034190J