Convenient QSAR model for predicting the complexation of structurally diverse compounds with β-cyclodextrins

ARTICLE *in* BIOORGANIC & MEDICINAL CHEMISTRY · DECEMBER 2008

Impact Factor: 2.79 · DOI: 10.1016/j.bmc.2008.11.040 · Source: PubMed

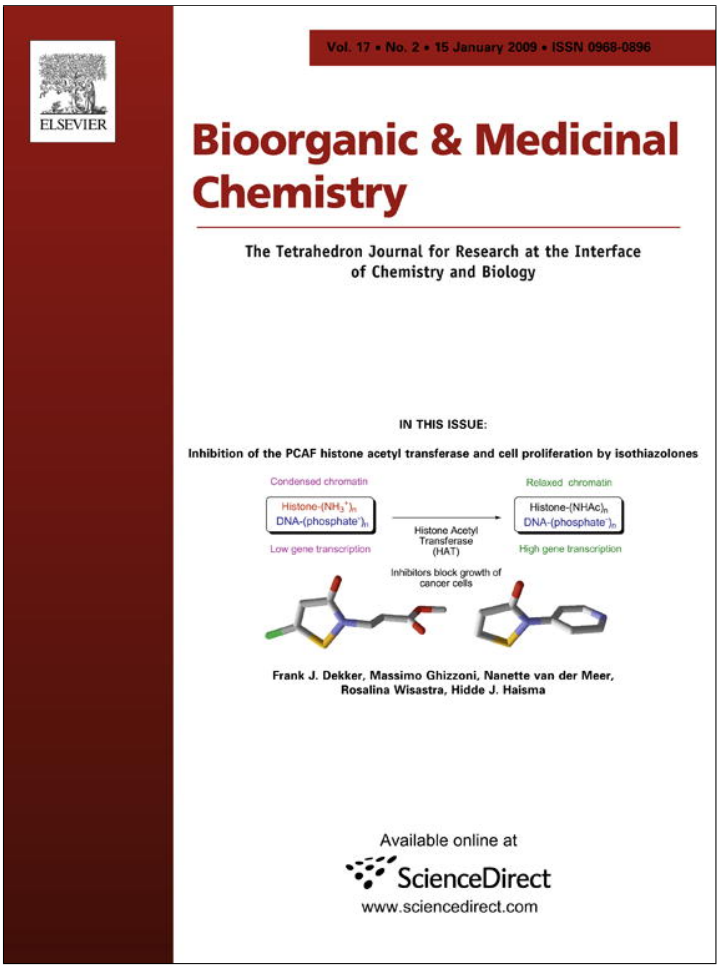
5 AUTHORS, INCLUDING:

Alfonso Pérez-Garrido
Universidad Católica San Antonio de Murcia
16 PUBLICATIONS **128** CITATIONS
SEE PROFILE

Aliuska Morales Helguera
50 PUBLICATIONS **1,041** CITATIONS
SEE PROFILE

Adela Abellán
Universidad Católica San Antonio de Murcia
9 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Natália D. S. Cordeiro
University of Porto
245 PUBLICATIONS **2,847** CITATIONS
SEE PROFILE



Bioorganic & Medicinal Chemistry 17 (2009) 896–904

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Convenient QSAR model for predicting the complexation of structurally diverse compounds with β-cyclodextrins

Alfonso Pérez-Garrido [a,b,*], Aliuska Morales Helguera [c,d,e], Adela Abellán Guillén [b]
M. Natália D. S. Cordeiro [e], Amalio Garrido Escudero [a]

[a] Enviromental Engineering and Toxicology Dpt., Catholic University of San Antonio, Murcia, C.P., Guadalupe 30107, Spain
[b] Department of Food and Nutrition Technology, Catholic University of San Antonio, Murcia, C.P., Guadalupe 30107, Spain
[c] Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba
[d] Molecular Simulation and Drug Design Group, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba
[e] REQUIMTE, Chemistry Department, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 29 July 2008
Revised 4 November 2008
Accepted 12 November 2008
Available online 24 November 2008

Keywords:
QSAR
Topological descriptors
β-Cyclodextrins
Complex stability constant

ABSTRACT

This paper reports a QSAR study for predicting the complexation of a large and heterogeneous variety of substances (233 organic compounds) with β-cyclodextrins (β-CDs). Several different theoretical molecular descriptors, calculated solely from the molecular structure of the compounds under investigation, and an efficient variable selection procedure, like the Genetic Algorithm, led to models with satisfactory global accuracy and predictivity. But the best-final QSAR model is based on Topological descriptors meanwhile offering a reasonable interpretation. This QSAR model was able to explain ca. 84% of the variance in the experimental activity, and displayed very good internal cross-validation statistics and predictivity on external data. It shows that the driving forces for CD complexation are mainly hydrophobic and steric (van der Waals) interactions. Thus, the results of our study provide a valuable tool for future screening and priority testing of β-CDs guest molecules.

1. Introduction

Cyclodextrins (CDs) are cyclic oligomers composed of either six (α-cyclodextrin), seven (β-cyclodextrin), eight (γ-cyclodextrin), or more α-D-glucopyranose units linked in a toroidal structure by α-(1-4)glycosidic bonds (Fig. 1). Their overall molecular shape is normally portrayed in terms of a truncated cone with primary and secondary hydroxyl groups crowning the narrower rim and wider rim, respectively[1] (Fig. 1). CDs are among the most frequently used host molecules in a wide range of applications in industrial, pharmaceutical, agricultural, and other fields, including improving the solubility and stability of drugs and biopharmaceuticals, and selectively binding materials that fit into the central hole in affinity purification and chromatography methods.[2,3] Experimental determination of the complex binding constant is often difficult and time consuming because of the low solubility of the guest molecules in aqueous solution. For instance, 10 days were required for gathering data related to the equilibrium system of digitoxin and β-cyclodextrin.[4] The employment of QSAR/QSPR methodology allows cost savings by reducing the laboratory resources needed, and the time required to create and investigate new compounds with better complexing profile. For this reason, QSAR/QSPR is a useful alternative tool in the research of novel compounds.

Computer-based methods have already been applied as tools for predicting CD binding constants and for studying the driving forces involved in the encapsulation phenomena. These applications have been excellently reviewed by Lipkowitz in the late nineties.[5] Molecular modeling using quantum mechanics calculations, Monte Carlo or molecular dynamics simulations, etc., group-contribution models; quantitative-structure–activity/property relationship (QSAR/QSPR) techniques based on 2D, 3D molecular descriptors and on comparative molecular field analysis; statistical analysis tools; and artificial neural networks have whole been used to predict the thermodynamic stability of CDs inclusion complexes and to elucidate the most important factors influencing the host–guest interactions.[6–15]

The general picture that emerges from the joint analysis of the large body of available experimental and theoretical work reveals that there are five major interactions between CD-hosts and guest molecules, namely (i) hydrophobic interactions, (ii) van der Waals interactions, (iii) hydrogen-bonding between polar groups of the guest 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. CD complex formation usually results from different combinations of these forces.

* Corresponding author. Tel.: +34 968 278 755.
E-mail address: Aperez@pdi.ucam.edu (A. Pérez-Garrido).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.11.040

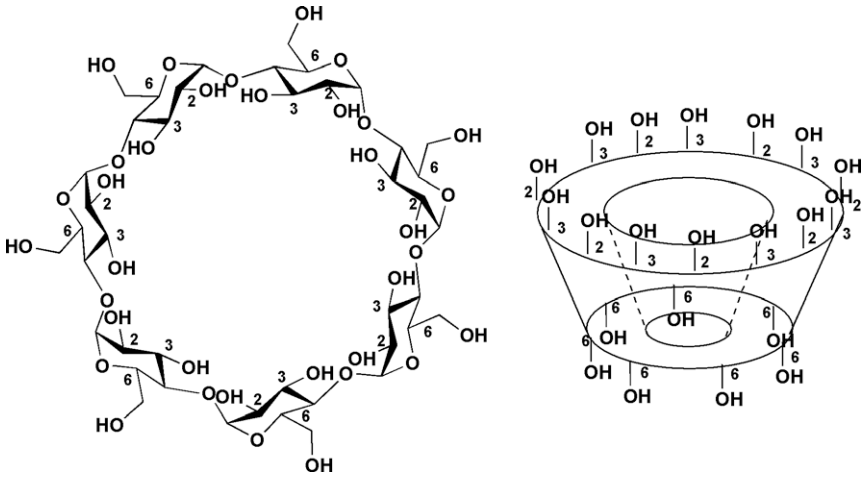


Fig. 1. Chemical structure of β-cyclodextrin.

The aim of the present study is to build a multiple linear regression QSAR model, able to correlate and predict the complex stability constant between diverse guest molecules and β-CDs, since these are the most commonly used. Special emphasis will be given to elucidate the driving forces leading to the complexation of the set of molecules under study. For this purpose, we resorted to the free-software package DRAGON, available at the internet site: www.vcclab.org/lab/pclient/.[16] DRAGON contains more than 1600 molecular descriptors divided into several families: 0D (constitutional descriptors), 1D (e.g., functional group counts), 2D (e.g., topological descriptors and connectivity indices), and 3D (e.g., GETAWAY, WHIM, RDF and 3D-MoRSE descriptors). These descriptors have proved to be particularly useful in QSAR/QSPR modeling studies, and to provide satisfactory correlation between the modeled target and molecular parameters.[17–32]

2. Materials and methods

2.1. Data set

The overall data set of 233 substances comprised a large number of classes of organic compounds: aromatic hydrocarbons, alcohols, phenols, ethers, aldehydes, ketones, acids, esters, nitriles, anilines, halogenated compounds, heterocycles, nitro, sulfur and steroids and barbitals compounds. This set of guest molecules was extracted from the work of Suzuki,[6] and the experimental endpoint to be predicted is the β-CD complex stability constant (K) in water at 298 K taken from references therein. Two of such guest molecules are stereoisomers chemicals 214 and 215, which could not be distinguished by the present 2D descriptors but had nevertheless different K values. Thus, one of the isomers was discarded (chemical 215), being only the other one (chemical 214) considered in our study with an averaged value of K. Moreover, all K values were log-transformed ($\log K$) for being of practical use in the following QSAR modeling. Table 1 displays a complete list of the chemicals along with the reported experimental data.

2.2. Model selection and validation

The structures of the compounds were first drawn with the aid of ISIS/Draw software ver. 2.5.[33] Molecular structures were then fully optimized with the molecular mechanics method (MM2)[34] followed by the PM3 semi-empirical Hamiltonian[35,36] implemented in MOPAC ver. 6.0.[37] Subsequently, different families of descriptors were calculated using the web-DRAGON.[16] Input variables with constants or closed to constants values were immediately eliminated. To validate the models, k-means cluster analysis was used to split the original dataset of chemicals into training and test sets. Mathematical models were obtained afterwards by means of multiple linear regression analysis along with a variable subset selection procedure.

2.3. *k*-Means cluster analysis

Developing rational approaches for the selection of training and test set compounds is an active area of research. These approaches range, for instance, from straightforward random selection[38] to various clustering techniques.[39] The main goal of k-means cluster analysis (k-MCA) is to partition the original series of compounds into several statistically representative classes of chemicals, among which one might then select the training and test set compounds. Here, we have decided that the training set should contain 80% (186/233) of the original data and the test set the remaining 20%, to guarantee that any kind of substance as determined by the clusters derived from k-MCA was represented in each set.

Starting from all descriptors of all 0–3D family types, those that produce the greatest separation of clusters meanwhile ensuring a statistically acceptable data partition were selected. In so doing, we took into account the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible). The k-MCA split the compounds into four clusters comprising 76, 53, 69 and 35 members with standard deviations of 0.08, 0.14, 0.2 and 0.2, respectively. Selection of the training and test sets was carried out by taking compounds belonging to each cluster, proportionally to the size of the cluster. We also made an inspection of the standard deviation between and within clusters, the respective Fisher ratio and p level of significance (ought to be lower than 0.05)[40,41] (Table 2).

2.4. Variable selection

Presently, there is a vast amount and wide range of molecular descriptors with which one can model the activity of interest. This makes the search for gathering the most suitable subset quite complicated and time consuming because of the many possible combinations, especially if one tries to define an accurate, robust, and (above all) interpretable model. For this reason, we applied the genetic algorithm (GA)[42] procedure for selecting the variables, as implemented in Mobydigs software ver. 1.0.[43] The particular GA simulation applied here resorted to the generation of 100 regression models, ordered according to their increased internal predictive performance (verified by cross-validation). Fist of all, models

Table 1
Names, observed and predicted activity[a], and leverage values for the compounds used in this study

No	Name	Log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.	No	Name	log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.
1	Carbon tetrachloride	2.20	2.25	Test	0.058	[10]	37	Iodobenzene	2.93	2.44	Training	—	[12]
2	Chloroform	1.43	1.39	Training	—	[10]	38	3-Fluorophenol	1.70	1.60	Training	—	[8]
3	Methanol	−0.49	−0.74	Training	—	[8]	39	4-Fluorophenol	1.73	1.66	Training	—	[8]
4	Acetonitrile	−0.27	−0.38	Training	—	[10]	40	3-Chlorophenol	2.28	2.41	Training	—	[8]
5	Acetaldehyde	-0.64	-0.38	Training	—	[10]	41	4-Chlorophenol	2.61	2.47	Training	—	[12]
6	Ethanol	−0.03	−0.08	Test	0.114	[8]	42	3-Bromophenol	2.51	2.52	Test	0.013	[8]
7	1,2-Ethanediol	-0.19	0.17	Training	—	[8]	43	4-Bromophenol	2.65	2.58	Test	0.015	[12]
8	Acetone	0.42	0.40	Training	—	[10]	44	3-Iodophenol	2.93	2.65	Training	—	[8]
9	1-Propanol	0.57	0.54	Test	0.062	[8]	45	4-Iodophenol	2.98	2.71	Training	—	[12]
10	2-Propanol	0.63	0.77	Training	—	[10]	46	Nitrobenzene	2.04	1.75	Training	—	[10]
11	1,3-Propanediol	0.67	0.70	Training	—	[8]	47	4-Nitrophenol	2.39	1.80	Training	—	[12]
12	Tetrahydrofuran	1.47	0.90	Test	0.034	[10]	48	Benzene	2.23	1.40	Test	0.025	[64]
13	Cyclobutanol	1.18	1.25	Training	—	[8]	49	Phenol	1.98	1.71	Training	—	[12]
14	1-Butanol	1.22	1.07	Training	—	[64]	50	Hydroquinone	2.05	1.93	Test	0.015	[12]
15	2-Butanol	1.19	1.34	Training	—	[8]	51	4-Nitroaniline	2.48	2.14	Test	0.020	[12]
16	2-Methyl-1-propanol	1.62	1.35	Training	—	[8]	52	Aniline	1.60	1.92	Training	—	[10]
17	2-Methyl-2-propanol	1.68	1.33	Training	—	[8]	53	Sulfanilamide	2.76	2.72	Training	—	[11]
18	1,4-Butanediol	0.64	1.12	Training	—	[8]	54	Cyclohexanol	2.67	2.55	Training	—	[64]
19	Diethylamine	1.36	1.22	Training	—	[10]	55	1-Hexanol	2.33	1.79	Training	—	[64]
20	Cyclopentanol	2.08	1.87	Training	—	[8]	56	2-Hexanol	1.98	2.28	Training	—	[8]
21	1-Pentanol	1.80	1.49	Training	—	[64]	57	2-Methyl-2-pentanol	1.99	2.55	Training	—	[8]
22	2-Pentanol	1.49	1.87	Training	—	[8]	58	3-Methyl-3-pentanol	2.15	2.27	Training	—	[8]
23	3-Pentanol	1.35	1.70	Training	—	[8]	59	4-Methyl-2-pentanol	2.04	2.40	Training	—	[8]
24	2-Methyl-1-butanol	2.08	1.71	Training	—	[8]	60	3,3-Dimethyl-2-butanol	2.75	2.28	Training	—	[8]
25	2-Methyl-2-butanol	1.91	1.93	Training	—	[8]	61	1,6-Hexanediol	1.69	1.65	Training	—	[8]
26	3-Methyl-1-butanol	2.25	1.88	Test	0.021	[8]	62	Benzonitrile	2.23	1.84	Training	—	[12]
27	3-Methyl-2-butanol	1.92	1.82	Training	—	[8]	63	Benzothiazole	2.38	2.59	Training	—	[65]
28	2,2-Dimethyl-1-propanol	2.71	1.94	Test	0.074	[64]	64	4-Nitrobenzoic acid	2.34	1.71	Training	—	[12]
29	1,5-Pentanediol	1.22	1.43	Test	0.120	[8]	65	Benzaldehyde	1.78	1.89	Training	—	[10]
30	1,4-Dibromobenzene	2.97	3.14	Training	—	[12]	66	Benzoic acid	2.12	2.05	Training	—	[64]
31	1,4-Diiodobenzene	3.17	3.38	Training	—	[12]	67	4-Hydroxybenzaldehyde	1.75	2.04	Training	—	[8]
32	3,5-Dibromophenol	2.56	3.20	Training	—	[8]	68	4-Hydroxybenzoic acid	2.20	2.09	Training	—	[12]
33	3,5-Dichlorophenol	2.07	2.99	Test	0.020	[8]	69	Benzyl chloride	2.45	2.70	Training	—	[12]
34	1-Chloro-4-nitrobenzene	2.15	2.39	Training	—	[12]	70	Toluene	2.09	2.03	Training	—	[10]
35	Fluorobenzene	1.96	1.37	Test	0.020	[12]	71	benzyl alcohol	1.71	2.25	Training	—	[10]
36	Bromobenzene	2.50	2.31	Training	—	[12]	72	Anisole	2.32	2.11	Training	—	[12]
73	*m*-Cresol	1.98	2.24	Training	—	[8]	109	*N*,*N*-Dimethylaniline	2.36	2.80	Training	—	[12]
74	*p*-Cresol	2.40	2.30	Training	—	[12]	110	Barbital	1.78	2.39	Training	—	[11]
75	4-Methoxyphenol	2.21	2.27	Training	—	[12]	111	cyclooctanol	3.30	3.31	Training	—	[8]
76	3-Methoxyphenol	2.11	2.19	Training	—	[8]	112	1-Octanol	3.17	2.13	Test	**0.212**[c]	[8]
77	4-Hydroxybenzyl alcohol	2.16	2.39	Training	—	[12]	113	2-Octanol	3.13	2.74	Training	—	[8]
78	Hydrochlorothiazide	1.76	1.94	Training	—	[11]	114	Quinoline	2.12	2.47	Training	—	[65]
79	*N*-Methylaniline	2.12	2.38	Training	—	[12]	115	3-Cyanophenyl acetate	1.49	2.24	Training	—	[8]
80	1-Butylimidazole	2.19	2.96	Training	—	[66]	116	4-Hydroxycinnamic acid	2.83	2.61	Training	—	[11]
81	1-Heptanol	2.85	2.00	Test	**0.164**[3]	[8]	117	Ethyl benzoate	2.73	2.53	Training	—	[12]
82	Phenylacetylene	2.36	2.07	Training	—	[12]	118	4′-Hydroxypropiophenone	2.63	2.70	Training	—	[8]
83	Thianaphthene	3.23	2.83	Training	—	[65]	119	3′-Hydroxypropiophenone	2.61	2.61	Training	—	[8]
84	4-Fluorophenyl acetate	2.11	2.24	Test	0.031	[8]	120	*p*-Tolyl acetate	2.49	2.78	Training	—	[8]
85	3-Fluorophenyl acetate	1.91	2.13	Training	—	[8]	121	3-Methylphenyl acetate	2.21	2.69	Training	—	[8]
86	4-Chlorophenyl acetate	2.50	2.93	Training	—	[8]	122	4-Methoxyphenyl acetate	2.45	2.45	Training	—	[8]
87	3-Chlorophenyl acetate	2.44	2.84	Training	—	[8]	123	4-Propylphenol	3.55	3.14	Training	—	[8]
88	4-Bromophenyl acetate	2.68	3.02	Training	—	[8]	124	3-Propylphenol	3.28	3.05	Training	—	[8]
89	3-Bromophenyl acetate	2.67	2.94	Test	0.018	[8]	125	4-Isopropylphenol	3.58	3.18	Training	—	[8]
90	4-Iodophenyl acetate	3.00	3.15	Training	—	[8]	126	3-Isopropylphenol	3.44	3.08	Training	—	[8]
91	3-Iodophenyl acetate	3.07	3.06	Training	—	[8]	127	4-Isopropoxyphenol	2.86	3.08	Training	—	[8]
92	4-Nitrophenyl acetate	2.13	1.91	Training	—	[8]	128	2-Norbornaneacetate	3.59	3.42	Test	0.040	[64]
93	Acetophenone	2.27	2.34	Training	—	[12]	129	1-Benzylimidazole	2.61	3.12	Training	—	[66]
94	Phenyl acetate	2.10	2.39	Training	—	[8]	130	*m*-Methylcinnamic acid	2.93	2.95	Training	—	[11]
95	Methyl benzoate	2.50	2.24	Training	—	[12]	131	4-Ethylphenyl acetate	2.83	2.97	Test	0.014	[8]
96	3-Hydroxyacetophenone	2.06	2.35	Training	—	[8]	132	3-Ethylphenyl acetate	2.68	2.82	Training	—	[8]
97	4-Hydroxyacetophenone	2.18	2.44	Training	—	[12]	133	4-Ethoxyphenyl acetate	2.54	2.63	Training	—	[8]
98	Acetoanilide	2.20	2.65	Test	0.011	[12]	134	3-Ethoxyphenyl acetate	2.49	2.47	Test	0.019	[8]
99	*p*-Xylene	2.38	2.61	Training	—	[12]	135	Allobarbital	1.98	2.28	Training	—	[11]
100	Ethylbenzene	2.59	2.55	Training	—	[12]	136	4-*n*-Butylphenol	3.97	3.44	Test	0.027	[8]
101	Phenetole	2.49	2.57	Training	—	[12]	137	3-*n*-Butylphenol	3.76	3.35	Training	—	[8]
102	2-Phenylethanol	2.15	2.72	Training	—	[8]	138	3-Isobutylphenol	4.21	3.45	Training	—	[8]
103	3-Ethylphenol	2.60	2.66	Training	—	[8]	139	4-*sec*-Butylphenol	4.18	3.41	Training	—	[8]
104	4-Ethylphenol	2.69	2.75	Training	—	[12]	140	3-*sec*-Butylphenol	4.06	3.31	Training	—	[8]
105	4-Ethoxyphenol	2.33	2.66	Test	0.013	[8]	141	4-*tert*-Butylphenol	4.56	3.69	Test	0.032	[8]
106	3-Ethoxyphenol	2.35	2.58	Test	0.010	[8]	142	3-*tert*-Butylphenol	4.41	3.58	Training	—	[8]
107	3,5-Dimethoxyphenol	2.34	2.25	Training	—	[8]	143	Menadion	2.27	2.42	Test	0.023	[11]
108	*N*-Ethylaniline	2.34	2.83	Test	0.016	[12]	144	Sulfapyridine	2.70	2.68	Training	—	[11]
145	Sulfamonomethoxine	2.48	1.87	Training	—	[11]	181	4-*n*-Amylphenyl acetate	3.80	3.35	Training	—	[8]
146	Sulfisoxazole	2.32	2.58	Training	—	[11]	182	Flufenamic acid	3.10	2.75	Training	—	[64]
147	4-*n*-Propylphenyl acetate	3.15	3.13	Training	—	[8]	183	Meclofenamic acid	2.67	3.38	Training	—	[64]
148	3-*n*-Propylphenyl acetate	3.28	2.96	Training	—	[8]	184	Nitrazepam	1.97	1.97	Training	—	[11]

Table 1 (*continued*)

No	Name	Log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.	No	Name	log K_{obs}	Log K_{pred}	Partition	Leverage	Ref.
149	4-Isopropylphenyl acetate	2.88	3.26	Training	—	[8]	185	Flurbiprofen	3.69	3.02	Training	—	[11]
150	3-Isopropylphenyl acetate	3.36	3.09	Training	—	[8]	186	Sulfaphenazole	2.35	2.17	Training	—	[11]
151	4-*n*-Amylphenol	4.19	3.65	Test	0.039	[8]	187	Bendroflumethiazide	1.90	2.40	Training	—	[11]
152	4-*tert*-Amylphenol	4.70	3.84	Training	—	[8]	188	Mefenamic acid	2.49	2.40	Training	—	[11]
153	Carbutamide	2.29	2.82	Training	—	[11]	189	Acetohexamide	2.94	3.18	Test	0.047	[11]
154	Pentobarbital	3.01	2.79	Test	0.042	[11]	190	Fludiazepam	2.33	2.45	Training	—	[11]
155	Amobarbital	3.07	3.01	Training	—	[64]	191	Nimetazepam	1.73	1.99	Training	—	[11]
156	Thiopental	3.28	3.40	Training	—	[11]	192	Fenbufen	2.63	3.19	Training	—	[11]
157	Dibenzofuran	2.97	2.77	Training	—	[65]	193	Ketoprofen	2.85	2.77	Training	—	[11]
158	Dibenzothiophene	3.48	3.39	Training	—	[65]	194	Medazepam	2.40	3.09	Training	—	[11]
159	Phenazine	2.41	2.69	Training	—	[65]	195	Progabide	2.53	2.98	Test	0.080	[11]
160	Thianthrene	3.57	3.82	Test	0.039	[65]	196	Griseofulvin	1.47	1.56	Training	—	[11]
161	Carbazole	2.44	3.01	Training	—	[65]	197	Tolnaftate	3.83	3.38	Training	—	[11]
162	Phenoxazine	2.69	2.85	Test	0.021	[65]	198	Prostacyclin	2.94	3.70	Training	—	[11]
163	Phenothiazine	2.73	3.20	Training	—	[65]	199	Triamcinolone	3.37	3.32	Test	0.087	[67]
164	furosemide	1.78	2.47	Test	0.071	[11]	200	cortisone	3.35	3.49	Training	—	[11]
165	Phenobarbital	3.22	2.50	Test	0.033	[64]	201	Prednisolone	3.56	3.65	Test	0.065	[67]
166	Sulfisomidine	2.10	2.32	Test	0.108	[11]	202	Hydrocortisone	3.60	3.77	Training	—	[11]
167	Sulfamethomidine	2.33	1.94	Test	0.106	[11]	203	Corticosterone	3.85	3.89	Test	0.073	[67]
168	Sulfadimethoxine	2.26	1.50	Training	—	[11]	204	Dexamethasone	3.65	3.63	Training	—	[11]
169	4-*n*-Butylphenyl acetate	3.62	3.26	Training	—	[8]	205	Betamethasone	3.73	3.82	Training	—	[67]
170	3-*n*-Butylphenyl acetate	3.66	3.08	Training	—	[8]	206	Paramethasone	3.40	3.59	Training	—	[67]
171	3-Isobutylphenyl acetate	3.83	3.24	Training	—	[8]	207	Cortisone-21-acetate	3.62	3.45	Training	—	[67]
172	4-*tert*-Butylphenyl acetate	3.85	3.72	Training	—	[8]	208	Prednisolone-21-acetate	3.76	3.63	Training	—	[67]
173	Cyclobarbital	2.71	2.90	Training	—	[11]	209	Hydrocortisone-21-acetate	3.51	3.69	Training	—	[67]
174	Hexobarbital	3.08	3.02	Training	—	[11]	210	Fluocinolone acetonide	3.48	2.97	Training	—	[67]
175	1-Adamantaneacetate	4.32	4.04	Training	—	[64]	211	Triamcinolone acetonide	3.51	3.39	Training	—	[67]
176	Acridine	2.33	2.91	Training	—	[65]	212	Spironolactone	4.44	3.79	Training	—	[67]
177	Phenanthridine	2.57	2.82	Training	—	[65]	213	Dehydrocholic acid	3.38	3.39	Training	—	[67]
178	Xanthene	2.71	2.99	Training	—	[65]	214	Chenodeoxycholic acid	4.36[b]	4.74	Training	—	[67]
179	*N*-Phenylanthranilic acid	2.89	2.85	Training	—	[64]	215	Ursodeoxycholic acid	4.51[b]		Training	—	[67]
180	Mephobarbital	3.16	2.53	Training	—	[11]	216	Cholic acid	3.50	4.38	Test	0.121	[67]
217	Hydrocortisone-17-butyrate	3.23	3.25	Training	—	[67]	226	L-a-*O*-benzylglycerol	2.11	3.22	Test	0.019	[66]
218	Cinnarizine	3.64	3.71	Training	—	[11]	227	Sulfamerazine	1.97	2.37	Training	—	[11]
219	Cycloheptanol	3.23	2.94	Training	—	[8]	228	Butyl 4-hydroxybenzoate	3.39	2.86	Training	—	[11]
220	2-Methoxyethanol	0.22	0.58	Test	0.068	[8]	229	Butyl 4-aminobenzoate	3.19	3.14	Training	—	[11]
221	3-Hydroxycinnamic acid	2.56	2.54	Training	—	[11]	230	Benzidine	3.35	3.54	Test	0.021	[11]
222	Ethyl 4-hydroxybenzoate	3.01	2.49	Training	—	[11]	231	Triflumizole	2.66	2.60	Training	—	[11]
223	Ethyl 4-aminobenzoate	2.69	2.81	Test	0.012	[11]	232	Diazepam	2.33	2.75	Training	—	[11]
224	4-Methylcinnamic acid	2.65	3.04	Training	—	[11]	233	Prostaglandine E2	3.09	2.91	Training	—	[11]
225	Sulfadiazine	2.52	2.25	Test	0.077	[11]							

[a] β-CD complex stability constant (K), then log-transformed (log$K$).
[b] Chemicals 214 and 215 were replaced by only one compound (chemical 214) with an averaged log$K$ value (=4.44).
[c] Chemicals 81 and 112 have leverage values above the threshold (0.14) and, for that reason, its predictions were not taken into account when calculating Q^2_{EXT}.

Table 2
Standard deviation between and within clusters, degrees of freedom (df), Fisher ratio (F) and level of significance (p) of the variables in the k-means cluster analysis

	Between SS	df	Within SS	df	F	p
VEZ1	208.9675	3	23.03249	229	692.5517	$<10^{-5}$
VEm1	208.9593	3	23.04073	229	692.2767	$<10^{-5}$
VEv1	209.6369	3	22.36308	229	715.5670	$<10^{-5}$
VEe1	209.1965	3	22.80353	229	700.2717	$<10^{-5}$
VEp1	209.6248	3	22.37521	229	715.1377	$<10^{-5}$
Xu	211.0286	3	20.97139	229	768.1187	$<10^{-5}$

with one to two variables were developed by the variable subset selection procedure in order to explore all low combinations. The number of descriptors was subsequently increased one by one, and new models formed. The GA was stopped when further increments in the size of the model did not increase internal predictivity in any significant degree. Furthermore, the following conditions were used on our GA simulation: the maximum number of variables in a model was 10, the number of best retained models for each size was 5, the trade off between crossovers and mutation parameter (T) was from 0.3 to 0.7, and selection bias (B%) was from 30 to 90.

Table 3
Best models derived using from 2 to 10 variables for each family of descriptors

	Variables	N^a	F	s	Q^2_{CV-LOO}	In domain[b] (%)	Q^2_{boot}
Topologic	ZM1, S1K, ZM1V, SMTIV, LPRS PHI, J, Xu, T(N..S), T(O..O)	10	93.72	0.362	0.821	95.74	0.814
GETAWAY	HGM, H3m, H0v, HATSp, R6v R4v+, R7e, R4p, R6p, R8p+	10	69.34	0.414	0.776	95.74	0.763
Eigenvalues-based	AEige, SEige, VRA1, VRv2, VRm2 SEigm, VRA2, VEA1, SEigv, VRp1	10	74.07	0.398	0.769	97.87	0.760
Conectivity	χ_0, χ_1, χ_{1A}, χ_{2A}, χ_{0v} χ_{MOD}, χ_3, χ_{3A}, χ_{3sol}, RDCHI	10	68.63	0.416	0.771	93.62	0.731
Burden eigenvalues	BEHm1, BELm4, BELm5, BELm7, BEHv1 BEHv8, BELv8, BEHe1, BELe7, BELe8	10	66.80	0.420	0.765	97.87	0.754
Molecular propieties	Ui, Hy, AMR, MLOGP2, GVWAI-80 Inflam-50, Hypert-50, Infect-80, Infect-50, BLTF96	10	55.48	0.452	0.727	95.74	0.712
3DMoRSE	Mor01u, Mor02m, Mor03m, Mor04m, Mor01v Mor07v, Mor31e, Mor01p, Mor04p, Mor05p	10	56.77	0.448	0.726	95.74	0.705
WHIM	L2m, L1p, L3s, E1s, Tp Au, Ae, As, Du, De	10	48.40	0.476	0.689	93.62	0.667
RDF	RDF010u, RDF015u, RDF020u, RDF085u, RDF020m RDF040m, RDF015v, RDF030e, RDF050p, RDF060p	10	40.18	0.508	0.654	93.62	0.632
Randić molecular profile	DP01, DP02, DP08, DP17, SP01 SP04, SP05, SP07, SP17, SHP2	10	26.16	0.584	0.557	95.74	0.532
Galvez topological charge	JGI1, GGI6, GGI1, JGI6, GGI4 JGI7, GGI2, GGI5, JGI5, JGI4	10	17.65	0.644	0.441	97.87	0.415

[a] Number of variables.
[b] Percentage of chemicals from the training set within the applicability domain.

2.5. Model validation

Goodness of fit of the models was assessed by examining the determination coefficient (R^2), the standard deviation (s), Fisher ratio (F) and the ratio between the number of cases and the number of adjustable parameters in the model (known as the ρ statistics; notice that ρ should be $\geqslant 4$).[44] Other important statistics, namely the Kubinyi function (FIT)[45,46] and Akaike's information criteria (AIC)[47,48] were taken into account, as they give enough criteria for comparing models with different parameters, numbers of variables and numbers of chemicals. As to the robustness and predictivity of the models, these were evaluated by means of cross-validation, basically leave-one-out (CV-LOO) and bootstrapping testing techniques calculated with the training set, by looking to the outcome statistics of both techniques (i.e., Q^2_{CV-LOO} and Q^2_{boot}) as well as to the Q^2_{EXT} values obtained with the test set substances that fall within the applicability domain of the model. Further, the stability under heavy perturbations in the training set was checked by examining the outcome statistics of a response randomization procedure (Y scrambling) for the training and test sets ($a(R^2)$ and $a(Q^2)$ values). All these calculations were carried out with software Mobydigs ver. 1.0.[43]

To sum up, good quality of the models is indicated by high F, FIT and ρ values, by low s and AIC values, as well as by values closed to one for R^2, Q^2_{CV-LOO}, Q^2_{boot} and Q^2_{EXT} (save for $a(R^2)$ and $a(Q^2)$ values, which check random correlations).

Table 4
Symbols and description of the topological descriptors involved in the QSAR model (Eq. 1)

Symbols	Descriptor definition
ZM1	First Zagreb index M1
ZM1V	First Zagreb index by valence vertex degrees
SMTIV	Schultz MTI by valence vertex degrees
Xu	Xu index
J	Balaban distance connectivity index
S1K	1-Path Kier alpha-modified shape index
T(N..S)	Sum of topological distances between N..S
T(O..O)	Sum of topological distances between O..O
PHI	Kier flexibility index
LPRS	Log of product of row sums (PRS)
VEZ1	Eigenvector coefficient sum from Z weighted distance matrix (Barysz matrix)
VEm1	Eigenvector coefficient sum from mass weighted distance matrix
VEv1	Eigenvector coefficient sum from van der Waals weighted distance matrix
VEe1	Eigenvector coefficient sum from electronegativity weighted distance matrix
VEp1	Eigenvector coefficient sum from polarizability weighted distance matrix

2.6. Model orthogonalization

The main drawback of collinearity from the point of view of QSAR/QSPR modeling is that it increases the standard errors associated with the individual regression coefficients, thereby decreasing their value for purposes of interpretability. To overcome this problem, we employed here the Randić method of orthogonalization.[49–53] The first step for orthogonalizing the molecular descriptors is to select the appropriate order of orthogonalization, which, in this case, is the order of significance of the variables in the model. The first variable (v_1) is taken as the first orthogonal descriptor and the second one is orthogonalized with respect to it by taking the residual of its correlation with v_1. The process is repeated until all variables are completely orthogonalized, after which they are further standardized. Orthogonal standardized variables are then used to obtain a new model.

2.7. Applicability domain of the model

Given that the real utility of a QSAR model relies on its ability to accurately predict the modeled activity for new chemicals, careful assessment of the model's true predictive power is a must. This includes the model validation but also the definition of the applicability domain of the model in the space of molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models[54,55] but the most common one encompasses determining the leverage values for each compound.[56] A Williams plot, that is, the plot of standardized residuals versus leverage values (h), can then be used for an

Table 5
Correlation matrix for intercorrelations among the ten variables of the QSAR model (Eq. 1)

	Xu	*ZM1V*	*LPRS*	*SMTIV*	*S1K*	*ZM1*	*T(N..S)*	**PHI**	*J*	*T(O..O)*
Xu	1.00	–	–	–	–	–	–	–	–	–
ZM1V	**0.95**	1.00	–	–	–	–	–	–	–	–
LPRS	**0.99**	**0.94**	1.00	–	–	–	–	–	–	–
SMTIV	**0.92**	**0.90**	**0.97**	1.00	–	–	–	–	–	–
S1K	**0.97**	**0.92**	**0.98**	**0.95**	1.00	–	–	–	–	–
ZM1	**0.98**	**0.94**	**0.99**	**0.95**	**0.96**	1.00	–	–	–	–
T(N..S)	0.28	0.36	0.26	0.26	0.27	0.26	1.00	–	–	–
PHI	0.68	0.57	0.68	0.67	**0.77**	0.58	0.15	1.00	–	–
J	−0.54	−0.53	−0.56	−0.54	−0.44	−0.57	−0.17	−0.18	1.00	–
T(O..O)	**0.73**	**0.72**	**0.79**	**0.86**	**0.80**	**0.79**	0.04	0.56	−0.37	1.00

Table 6
Step-by-step analysis of the forward stepwise orthogonalization process

Step	$^{1}\Omega Xu$	$^{2}\Omega ZM1V$	$^{4}\Omega ZM1$	$^{3}\Omega LPRS$	$^{9}\Omega J$	$^{8}\Omega PHI$	$^{7}\Omega T(N..S)$	$^{10}\Omega T(O..O)$	$^{6}\Omega SMTIV$	$^{5}\Omega S1K$	Intercept	$R^2(Adj.)$	$\Delta R^2(Adj.)$	*p*-Level
1	0.0763	–	–	–	–	–	–	–	–	–	1.595	0.303	0.303	$<10^{-5}$
2	0.0763	−0.0109	–	–	–	–	–	–	–	–	1.595	0.502	0.199	$<10^{-5}$
3	0.0763	−0.0109	0.0703	–	–	–	–	–	–	–	1.595	0.650	0.148	$<10^{-5}$
4	0.0763	−0.0109	0.0703	−0.0389	–	–	–	–	–	–	1.595	0.724	0.074	$<10^{-5}$
5	0.0763	−0.0109	0.0703	−0.0389	−1.0263	–	–	–	–	–	1.595	0.756	0.032	$<10^{-5}$
6	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	–	–	–	–	1.595	0.781	0.025	$<10^{-5}$
7	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	–	–	–	1.595	0.803	0.022	$<10^{-5}$
8	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	–	–	1.595	0.825	0.022	$<10^{-5}$
9	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	0.0001	–	1.595	0.833	0.008	0.003450
10	0.0763	−0.0109	0.0703	−0.0389	−1.0263	−0.2921	−0.0492	−0.0132	0.0001	0.0465	1.595	0.834	0.001	0.123876

immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within ± *x* standard deviations and a leverage threshold h^* (h^* is generally fixed at $3\kappa/n$, where n is the number of training compounds and κ the number of model parameters, whereas $x = 2$ or 3), lying outside this area (vertical lines) the outliers and (horizontal lines) influential chemicals. For future predictions, only predicted complex stability constant data for chemicals belonging to the chemical domain of the training set should be proposed and used.[57] So, calculations of Q^2_{EXT} were performed only for those substances that had a leverage value below the threshold h^*.

3. Results and discussion

3.1. QSAR models

Several QSAR models for predicting β-cyclodextrins complex stability constants were developed, using the same training set and routine for variable selection. This was accomplished by finding regression models for the *k*-MCA chosen training set (185 compounds) based on GA selection (between 2 and 10 variables), in conjunction with the following eleven sets of molecular descriptors: topological, GETAWAY, eigenvalue-based indices, connectivity indices, Burden eigenvalues, molecular properties, 3D-MoRSE, WHIM, RDF, Randić molecular profiles and Galvez topological charges indices. The best QSAR-models are given in Table 3.

There are substantial differences in the explanation of the experimental variance given by the topological model, when compared with the rest of the models. Thus, while the topological model is able to explain more than 84% of experimental variance, the other models, at best, can only explain 79.9% of such variance. The predictive ability-expressed as Q^2_{CV-LOO} and Q^2_{EXT}- of the topological model is also higher than the other descriptors' models, even for those based on 3D descriptors (3D-MoRSE, WHIM, RDF, GETAWAY and Randić) that showed lower scores.

Topological descriptors, unlike three-dimensional descriptors, do not consider information on conformational aspects, such as bond lengths, bond angles and torsion angles, but do encode important information on adjacency, branching and relative distance among different functionalities in a numerical form. Thus, these molecular descriptors determine a wide range of physicochemical properties of molecules. In addition, they can be derived from molecular structures using low computational resources, making them remarkably attractive in molecular modeling.

Successful correlations between β-CDs-complex stability constants and topological indices have also been found in the literature[58,11] but with lesser number of ligands. Our resulting best-fit topological-QSAR model (a 10-variable equation) is given below together with the statistical parameters of the regression. As can be seen, this model is reasonable in both statistical significance and goodness of fit or prediction.

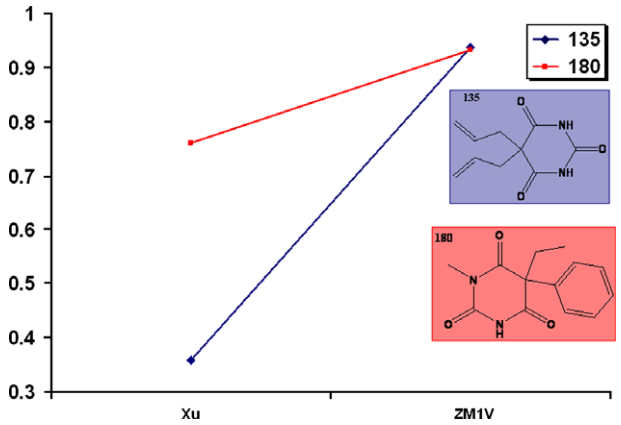


Fig. 2. Contributions from each of the variables to the final value of log*K* for allobarbital (chemical 135) and mephobarbital (chemical 180).

$$\log K = 1.08(\pm 0.063)Xu - 1.51\cdot 10^{-2}(\pm 9.23\cdot 10^{-4})ZM1V - 0.38(\pm 2.61\cdot 10^{-2})LPRS + 7.81\cdot 10^{-4}(\pm 6.68\cdot 10^{-5})SMTIV + 0.93(\pm 8.16\cdot 10^{-2})S1K + 6.65\cdot 10^{-2}(\pm 8.08\cdot 10^{-3})ZM1 - 8.70\cdot 10^{-2}(\pm 1.12\cdot 10^{-2})T(N..S) - 0.41(\pm 5.85\cdot 10^{-2})PHI - 1.20(\pm 0.17)J - 1.32\cdot 10^{-2}(\pm 2.67\cdot 10^{-3})T(O..O) - 0.77(\pm 0.28) \quad (1)$$

$$N = 185;\quad R^2 = 0.843;\quad Q^2_{(CV-LOO)} = 0.821;\quad s = 0.361;$$
$$F = 93.72;\quad AIC = 0.147;\quad FIT = 3.371\quad Q^2_{boot} = 0.813;$$
$$a(R^2) = 0.02;\quad a(Q^2) = -0.113;\quad Q^2_{EXT} = 0.756$$

The meaning of each of the topological descriptor variable involved in the cluster analysis and thereby used in the model above is shown in Table 4.

An aspect deserving special attention is the degree of collinearity among the variables of the model, which can be readily diagnosed by analyzing the cross-correlation matrix. As seen in Table 5, the pairs of descriptors (*Xu; ZM1V*), (*Xu; LPRS*), (*Xu; SMTIV*), (*Xu; S1K*), (*Xu; ZM1*), (*T(O..O); ZM1V*), (*T(O..O); LPRS*), (*T(O..O); SMTIV*), (*T(O..O); S1K*), (*T(O..O); ZM1*), (*T(O..O); Xu*), (*ZM1V*; *LPRS*), (*ZM1V*; *SMTIV*), (*ZM1V*; *S1K*), (*ZM1V*; *ZM1*), (*LPRS*; *SMTIV*), (*LPRS*; *S1K*), (*LPRS*; *ZM1*), (*SMTIV*; *S1K*), (*SMTIV*; *ZM1*), (*S1K*; *ZM1*), and (*S1K*; *PHI*) are correlated with each other. Rather than deleting any of these descriptors, it is of interest to examine the performance of orthogonal complements in modeling β-CD complexation.

Following the Randić technique, we have determined orthogonal complements for all variables of the non-orthogonalized model (see Table 6). As a result, variable $^5\Omega S1K$ was found to be statistically non-significant ($p = 0.124$; Table 6), may be because the information contained in this variable is common to the information contained in the other descriptor variables. Furthermore, the significance of adding $^5\Omega S1K$ to the model remains unclear as seen from the modest improvements in $R^2(Adj.)$ (adjusted determination coefficient) on going from step 9 to 10 (see in Table 6, $R^2(Adj.)$ from step 9 to 10). Thus, by removing it, we obtained the following QSAR model, which is given below after to be standardized.

$$\log K = 0.488(\pm 0.027)^1\Omega Xu - 0.392(\pm 0.026)^2\Omega ZM1V - 0.239(\pm 0.027)^3\Omega LPRS + 0.076(\pm 0.026)^4\Omega SMTIV + 0.337(\pm 0.026)^6\Omega ZM1 - 0.134(\pm 0.027)^7\Omega T(N..S) - 0.161(\pm 0.030)^8\Omega PHI - 0.160(\pm 0.027)^9\Omega J - 0.127(\pm 0.026)^{10}\Omega T(O..O) + 2.535(\pm 0.027) \quad (2)$$

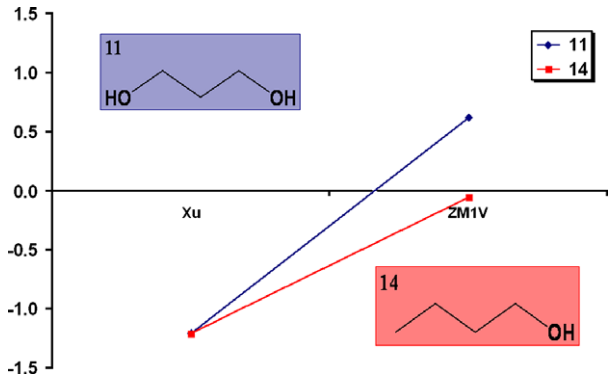


Fig. 3. Contributions from each of the variables to the final value of log*K* for 1,3-propanediol (chemical 11) and 1-butanol (chemical14).

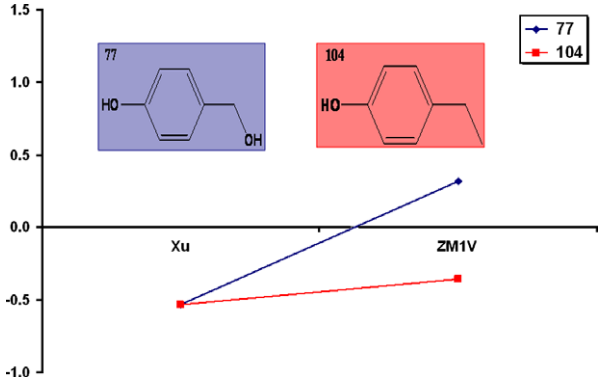


Fig. 4. Contributions from each of the variables to the final value of log*K* for 4-hydroxybenzyl alcohol (chemical 77) and 4-ethylphenol (chemical 104).

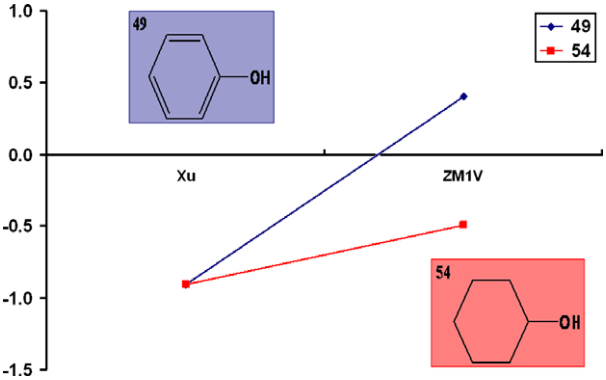


Fig. 5. Contributions from each of the variables to the final value of log*K* for phenol (chemical 49) and cyclohexanol (chemical 54).

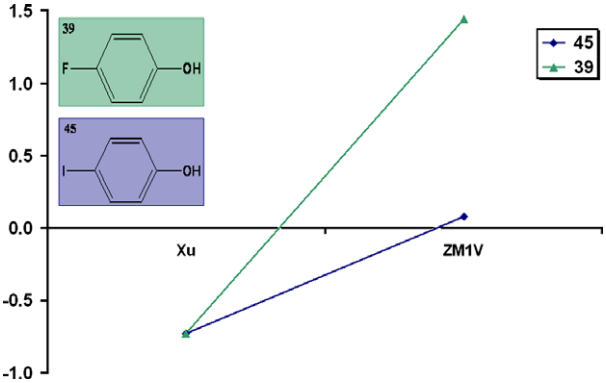


Fig. 6. Contributions from each of the variables to the final value of log*K* for 4-fluorophenol (chemical 39) and 4-iodophenol (chemical 45).

$$N = 185;\quad R^2 = 0.841;\quad Q^2_{(CV-LOO)} = 0.821;\quad s = 0.363;$$
$$F = 103.05;\quad AIC = 0.147;\quad FIT = 3.487$$
$$Q^2_{boot} = 0.812;\quad a(R^2) = 0.014;\quad a(Q^2) = -0.105;\quad Q^2_{EXT} = 0.764$$

As can be seen in Table 6, removal of $^5\Omega S1K$ had little effect on the overall fitness of the model as the statistics are as robust as before, and further, by comparing Eq. 1 with Eq. 2, one can see that there are no changes in either the sign of the regression coefficients. Nevertheless, the relative contributions of the variables in the orthogonal-descriptor model are quite different to those related to the non-orthogonalized model. Therefore, for purposes of QSAR interpretability, we shall use the orthogonal-descriptor model defined in Eq. 2.

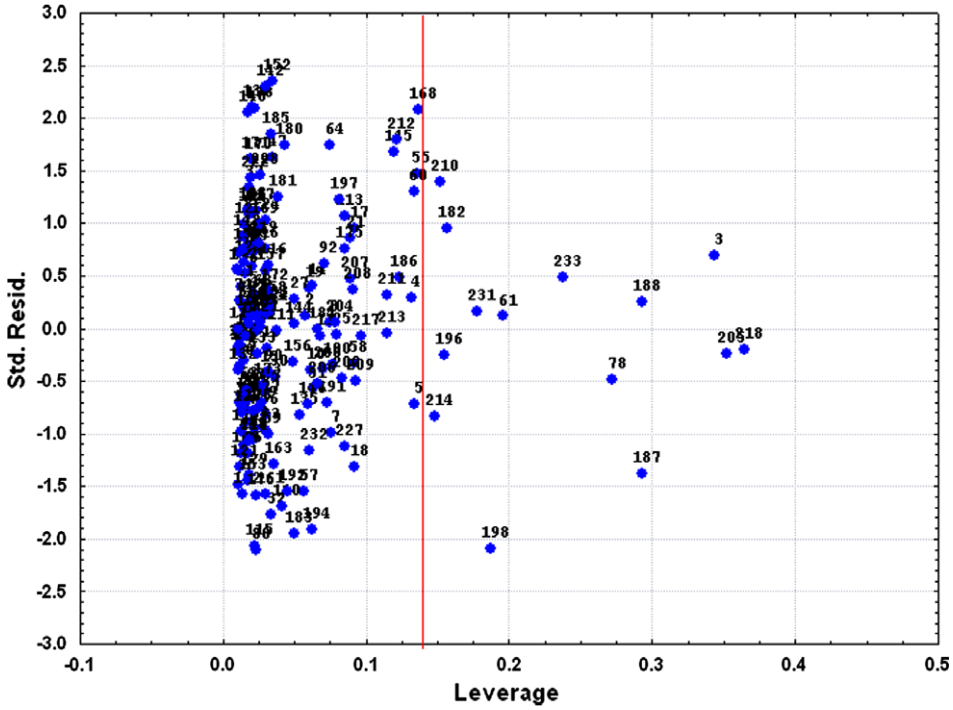


Fig. 7. Applicability domain of the model of Eq. 2.

The meaning of topological indices compared with other parameters is not explicit.[59] They essentially condense a large amount of physicochemical information into a single number. For this reason, physical interpretation of these topological indices is complicated. Despite this potential drawback; however, these descriptors have been extensively used in Medicinal Chemistry.[60,24] According to Eq. 2, variables $^{1}\Omega Xu$, $^{4}\Omega SMTIV$ and $^{6}\Omega ZM1$ contribute positively to complexation while $^{2}\Omega ZM1V$, $^{3}\Omega LPRS$, $^{7}\Omega T(N..S)$, $^{9}\Omega J$, $^{8}\Omega PHI$ and $^{10}\Omega T(O..O)$, they do it negatively. These variables encode molecular factors such as size, branching, cyclization, presence of heteroatoms, flexibility and molecular volume. Due to the complexity, interpretation of our model will rely merely on the influence of the most significant variables. Here, it should be remarked that the major contribution to the stability of the inclusion complexes, therefore to encapsulation, comes from descriptor variables $^{1}\Omega Xu$ and $^{2}\Omega ZM1V$, followed by the remaining variables (Table 6).

The *Xu* descriptor is related to structural factors such as molecular dimension and shape. This parameter does correlate with the van der Waals volume, so we might say that complexation with β-CDs for this group of substances is primarily induced by van der Waals forces.[61] The size of the guest is another important factor determining the stability of the CD complex. A good matching between the solute size and the CD cavity is essential for efficient intermolecular contacts which lead to an increased stability of the final inclusion complex.[10] In Figure 2, we can see a clear example of this type of interaction, where an increase of the molecular size caused by the presence of a benzene ring stabilizes the complex due to stronger van der Waals interactions.

Regarding *ZM1V*, this descriptor is determined by the values of the valence vertex degrees. Therefore, its values will augment by an increase of the branching and by the presence of heteroatoms (since the greater the atomic number, the greater the influence). One can observe across the *ZM1V* variation that an increase on the number of hydroxyl groups (more hydrophilic) affects negatively the complexation (Figs. 3 and 4). This may be due to the fact that such an increase hinders access of the molecule in the hydrophobic cavity of β-CD.[62]

Another important feature that can be highlighted by variable *ZM1V* concerns the presence of benzene ring. This moiety has a more reduced ability to form complexes with β-CD than its hydrogenated counterpart. Along with this, the value of variable *ZM1V* (which contributes negatively) in the model for cyclohexanol is smaller than it is for phenol, reflecting a decline in the valence vertex degree (due to the rise in hydrogenation), which in turn improves the inclusion in the β-CD (Fig. 5).

For cases like the ones mentioned above, we can say that variable *ZM1V* is inversely proportional to hydrophobicity.

Moving now to halogenated derivatives of the phenyl group (Fig. 6), one can see that complexation is enhanced by a more voluminous substituent, which in turn goes along with the *ZM1V* contributions because of the increment on the atomic number. This nicely confirms what has been observed by Liu *et al.*,[63] that is to say, where an increased volume and polarizability of the guest substituent increased the stability of the complex due to stronger van der Waals interactions.

To sum up, we can conclude that for this set of molecules steric (van der Waals) and hydrophobic interactions are of prime importance in the inclusion processes on β-CD.

3.2. Applicability domain

It would be very interesting to have a predictive model for the vast majority of chemicals, particularly for those who have not been tested and, therefore, with unknown log*K* values. Since this is usually not possible, one should define the applicability domain of the QSAR model, that is, the range within which it bears a new compound. For that purpose, we built a Williams plot using the leverage values calculated for each compound. As seen in Figure 7, most of the compounds of the test set are within the applicability domain covered by ±3 times the standard residual (σ) and the leverage threshold h^{*} (=0.14), save for compounds 3, 61, 78, 182,

187, 188, 196, 198, 205, 210, 214, 218, 231 and 233. Even so, the latter should not be considered outliers but influential chemicals.[54]

Nevertheless, all evaluations pertaining to the external set were performed by taking into account the applicability domain of our QSAR model. So, if a chemical belonging to the test set had a leverage value greater than h^*, we consider that this means that the prediction is the result of substantial extrapolation and therefore may not be reliable.[55]

4. Conclusions

The forces affecting the phenomenon of complexation of chemicals with CDs are numerous and active in combination. In this study, we have examined the ability of a large and diverse set of substances to provide statistically sound and predictive QSAR models of β-CD complexation. We have thoroughly evaluated regression models in conjunction with a variety of structure representations, codifying a number of topological, physicochemical and three-dimensional aspects. For the present training set, topological descriptors provided the best model, as judged by extensive cross-validation and external-prediction. This topological-QSAR model was found to be superior to models derived using other 2D descriptors Burden eigenvalues, Galvez topological charge indices, connectivity indices, and eigenvalue-based indices- or even 3D descriptors Randić molecular profiles, RDF, 3DMoRSE, GETAWAY, and WHIM- or molecular properties.

Moreover, the driving forces for CD complexation ascertained by the model are hydrophobic and steric (van der Waals) interactions mainly. Finally, this is a simple model that might be used in the prediction of β-CD complex stability constants of compounds inside the applicability domain. It may thus constitute an alternative and particular useful tool for screening large libraries of compounds.

Acknowledgments

A.M.H. acknowledges the *Portuguese Fundaão para a Ciência e a Tecnologia* (FCT - Lisboa) (SFRH/BD/22692/2005) for financial support.

References and notes

1. Saenger, W.; Jacob, J.; Gessler, K.; Steiner, T.; Daniel, S.; Sanbe, H.; Koizumi, K.; Smith, S. M.; Tanaka, T. *Chem. Rev.* **1998**, *98*, 1787–1802.
2. Szejtli, J. *Chem. Rev.* **1998**, *98*, 1743–1753.
3. Hedges, A. R. *Chem. Rev.* **1998**, *98*, 2035–2044.
4. Yoshida, A.; Yamamoto, M.; Hirayama, F.; Uekawa, K. *Chem. Pharm. Bull.* **1988**, *36*, 4075–4080.
5. Lipkowitz, K. B. *Chem. Rev.* **1998**, *98*, 1829–1873.
6. Suzuki, T. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1266–1273.
7. Pérez, F.; Jaime, C.; Sánchez-Ruiz, X. *J. Org. Chem.* **1995**, *60*, 3840–3845.
8. Matsui, Y.; Nishioka, T.; Fujita, T. *Top. Curr. Chem.* **1985**, *128*, 61–89.
9. Davis, D. M.; Savage, J. R. *J. Chem. Res. (S)* **1993**, 94–95.
10. Park, J. H.; Nah, T. H. *J. Chem. Soc.* **1994**, *Perkin Trans. 2*, 1359–1362.
11. Klein, C. T.; Polheim, D.; Viernstein, H.; Wolschann, P. *J. Inclusion Phenom. Macrocyclic Chem.* **2000**, *36*, 409–423.
12. Liu, L.; Guo, Q.-X. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 133–138.
13. Suzuki, T.; Ishida, M.; Fabian, W. M. F. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 669–678.
14. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
15. Katritzky, A. R.; Fara, D. C.; Yang, H. F.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529–541.
16. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aid. Mol. Des.* **2005**, *19*, 453–463.
17. Saíz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2007**, *15*, 3565–3571.
18. Saíz-Urra, L.; González, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2006**, *14*, 7347–7358.
19. González, M. P.; Terán, C.; Teijeira, M.; Helguera, A. M. *Bull. Math. Bio.* **2007**, *69*, 347–359.
20. Saíz-Urra, L.; González, M. P.; Fall, Y.; Gómez, G. *Eur. J. Med. Chem.* **2007**, *42*, 64–70.
21. González, M. P.; Suárez, P. L.; Fall, Y.; Gómez, G. *Bioorg. Med. Chem.* **2005**, *15*, 5165–5169.
22. Helguera, A. M.; Perez, M. A. C.; González, M. P. *J. Mol. Model.* **2006**, *12*, 769–780.
23. Helguera, A. M.; Cordeiro, M. N. D. S.; Perez, M. A. C.; Combes, R. D.; González, M. P. *Bioorg. Med. Chem.* **2008**, *16*, 3395–3407.
24. Helguera, A. M.; Rodríguez-Borges, J. E.; García-Mera, X.; Fernández, F.; Cordeiro, M. N. D. S. *J. Med. Chem.* **2007**, *50*, 1537–1545.
25. Gupta, M. K.; Prabhakar, Y. S. *J. Chem. Inf. Model.* **2006**, *46*, 93–102.
26. Gupta, S.; Singh, M.; Madan, A. K. *J. Chem Inf. Comput. Sci.* **1999**, *39*, 272–277.
27. Pirrung, M. C.; Tumey, L. N.; McClerren, A. L.; Raetz, C. R. H. *J. Am. Chem. Soc.* **2003**, *125*, 1575–1586.
28. McElroy, N. R.; Jurs, P. C. *J. Med. Chem.* **2003**, *46*, 1066–1080.
29. Hayatshahia, S. H. S.; Abdolmalekia, P.; Ghiasib, M.; Safarian, S. *FEBS Lett.* **2007**, *581*, 506–514.
30. Kline, T. et al *J. Med. Chem.* **2002**, *45*, 3112–3129.
31. Sakowski, J.; Böhm, M.; Sattler, I.; Dahse, H. M.; Schlitzer, M. *J. Med. Chem.* **2001**, *44*, 2886–2899.
32. Kleinman, E. F.; Campbell, E.; Giordano, L. A.; Cohan, V. L.; Jenkinson, T. H.; Cheng, J. B.; Shirley, J. T.; Pettipher, E. R.; Salter, E. D.; Hibbs, T. A.; DiCapua, F. M.; Bordner, J. *J. Med. Chem.* **1998**, *41*, 266–270.
33. ISIS/Draw, Symyx MDL, San-Leandro, California, USA.
34. Allinger, N. L.; Zhou, X. F.; Bergsma, J. *J. Mol. Struct. (Theochem.)* **1994**, *118*, 69–83.
35. Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 209–220.
36. Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 221–264.
37. Frank, J. *Seiler Research Laboratory*; US Air Force Academy: Colorado, Springs Co., 1993.
38. Yasri, A.; Hartsough, D. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227.
39. Gore, P. A. J. Handbook of applied multivariate statistics and mathematical modeling. In *Chapter Cluster analysis*; Tinsley, H. E. A., Brown, S. D., Eds.; Academic Press: USA, 2000; pp 298–318.
40. McFarland, J. W.; Gans, D. J. Chemometric methods in molecular design. In *Chapter Cluster Significance Analysis*; van Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 295–307.
41. Johnson, R. A.; Wichern, D. W. *Applied MultiVariate Statistical Analysis*; Prentice-Hall: New York, 1988.
42. Goldberg, D. *Genetic Algorithms in Search*; Addison-Wesley: USA, 1989.
43. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. Mobydigs Computer Software, 1.0 ed.; 2004.
44. Garcia-Domenech, R.; Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445–449.
45. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 285–294.
46. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 393–401.
47. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*; Akademiai Kiado, Budapest, 1973.
48. Akaike, H. *IEEE Trans. Automat. Contr.* **1974**, *AC-19*, 716–723.
49. Lucic, B.; Nikolic, S.; Trinajstic, N.; Juric, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532–538.
50. Klein, D.; Randi, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quantum Chem.* **1997**, *63*, 215–222.
51. Randic, M. *N.J. Chem.* **1991**, *15*, 517–525.
52. Randic, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
53. Randic, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45–59.
54. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environmental Health Perspect.* **2003**, *111*, 1361–1375.
55. Netzeva, T. I. et al *ATLA* **2005**, *33*, 155–173.
56. Gramatica, P. *QSAR Comb. Sci.* **2007**, *00*, 1–9.
57. Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.
58. Estrada, E.; Perdomo-López, I.; Torres-Labandeira, J. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1561–1568.
59. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers: Australia, 1999.
60. Devillers, J. Topological indices and related descriptors in QSAR and QSPR. In *Chapter no-free-lunch molecular descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Publishers: Australia, 1999; pp 1–17.
61. Ren, B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139–143.
62. Rekharsky, M. V.; Inoue, Y. *Chem. Rev.* **1998**, *98*, 1875–1917.
63. Liu, L.; Guo, Q.-X. *J. Inclusion Phenom. Macrocyclic Chem.* **2002**, *42*, 1–14.
64. Inoue, Y.; Hakushi, T.; Liu, Y.; Tong, L.-H.; Shen, B.-J.; Jin, D.-S. *J. Am. Chem. Soc.* **1993**, *115*, 475–481.
65. Carpignano, R.; Marzona, M.; Cattaneo, E.; Quaranta, S. *Anal. Chim. Acta* **1997**, *348*, 489–493.
66. Rekharsky, M. V.; Goldberg, R. N.; Schwarz, F. P.; Tewari, Y. B.; Ross, P. D.; Yamashoji, Y.; Inoue, Y. *J. Am. Chem. Soc.* **1995**, *117*, 8830–8840.
67. Wallimann, P.; Marti, T.; Fürer, A.; Diederich, F. *Chem. ReV.* **1997**, *97*, 1567–1608.