Bioorganic & Medicinal Chemistry 16 (2008) 7944–7955

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

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

New QSPR study for the prediction of aqueous solubility of drug-like compounds

Pablo R. Duchowicz [a,*], Alan Talevi [a,b], Luis E. Bruno-Blanch [b], Eduardo A. Castro [a]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, La Plata 1900, Argentina
[b] Medicinal Chemistry, Biological Sciences Department, Faculty of Exact Sciences, La Plata National University (UNLP), 47 y 115, La Plata 1900, Argentina

ARTICLE INFO

Article history:
Received 8 July 2008
Revised 22 July 2008
Accepted 23 July 2008
Available online 29 July 2008

Keywords:
QSPR theory
Aqueous solubility
DRAGON molecular descriptors
Replacement method variable subset selection
Group contribution method

ABSTRACT

Solubility has become one of the key physicochemical screens at early stages of the drug development process. Solubility prediction through Quantitative Structure–Property Relationships (QSPR) modeling is a growing area of modern pharmaceutical research, being compatible with both High Throughput Screening technologies and limited compound availability characteristic of early stages of drug development. We resort to the QSPR theory for analyzing the aqueous solubility exhibited by 145 diverse drug-like organic compounds (0.781 being the average Tanimoto distances between all possible pairs of compounds in the training set). An accurate and generally applicable model is derived, consisting on a linear regression equation that involves three DRAGON molecular descriptors selected from more than a thousand available. Alternatively, we apply the linear QSPR to other 21 commonly employed validation compounds, leading to solubility estimations that compare fairly well with the performance achieved by previously reported Group Contribution Methods.

1. Introduction

1.1. Importance of solubility in the early stages of a drug development program

In the past, traditional drug development scheme focused on biological activity and potency of drugs. As a consequence, many drug development programs usually failed, because of toxicological and pharmaco-kinetical issues, at late stages of the development process, when large investments had already been made. Thus, modern drug development paradigm (usually referred as 'fail early, fail cheap') includes determination and/or estimation of physicochemical properties related to bioavailability at the very first stages of drug development, when a lead compound is being sought.[1] For many reasons, solubility stands out among such properties (along with pK_a, lipophilicity and stability) as one of the key physicochemical screens in early compound screening, which explains why solubility determination and estimation have been subjects of several recent publications and reviews in Medicinal Chemistry and Pharmaceutical specialized journals.[2–7] Among these reasons we may list:

1. To elicit their pharmacological activity, orally administered drugs should exhibit certain solubility in physiological intestinal fluids to be present in the dissolved state at the site of absorption. Aqueous solubility is a major indicator for the solubility in the intestinal fluids and its contribution to bioavailability issues. Note that 56 out of 100 product launches between 1995 and 2002 belong either to class II or class IV of the Biopharmaceutical Classification System, which means their oral bioavailability may be improved by enhancing their solubility.[2]
2. Determination of the true concentration of the free drug is critical in the in vitro assays; wrong conclusions regarding efficacy or toxicity may be drawn if unexpected low solubility or precipitation of the drug occurs.[2,6–8] Achievement of solution state is usually also needed for adequate in vivo testing.
3. Low solubility of compounds contributes to extent timelines, since material engineering of the drug or formulation efforts should be used to produce dosage forms that consistently deliver the desired dose of the drug in the site of absorption.[2–7]
4. Compounds with high solubility are more easily metabolized and eliminated from the organism, thus leading to lower probability of adverse effects and bioaccumulation.[9]

1.2. Solubility measurement and prediction

Solubility measurements determine either the thermodynamic or the kinetic solubility of the compounds. Thermodynamic solu-

* Corresponding author. Tel.: +54 221 425 7430/7291; fax: +54 221 425 4642.
E-mail addresses: pduchowicz@gmail.com, pabloducho@gmail.com (P.R. Duchowicz).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.07.067

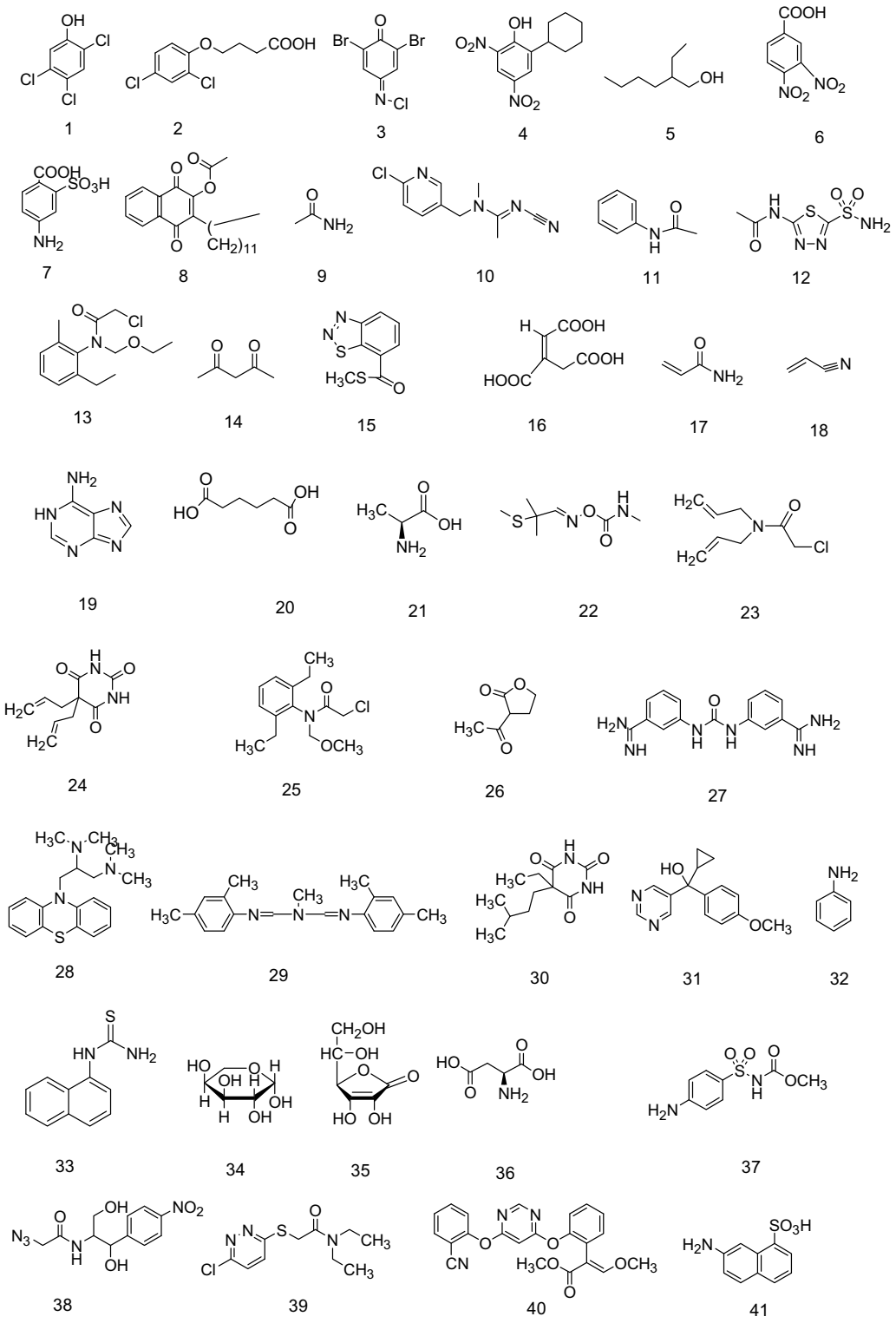

Figure 1. Molecular structures for the training set compounds ($N = 97$).

bility measurements are performed by dispensing a purified crystalline solid compound in a liquid, allowing an incubation time (typically, 24–48 h) to ensure equilibrium.[6,7] The necessary time to measure thermodynamic solubility is not compatible with modern High Throughput Screening (HTS) technologies: standard equilibrium solubility measures are restricted to about 25–50 compounds a week if handled by one specialist. Moreover, they demand 3–10 mg of purified compound, at an early stage when usually only a few milligrams of product are available which should also be used to measure other important absorption, distribution, metabolization, and elimination (ADME) parameters and biological activity.[6] True HTS solubility assays are only available in a few specialized companies; and they involve complex task such as automatically handling powders with different characteristics, with the consequent cross-contamination potential, power loss during movement of dosing heads and difficulties in equipment cleaning.

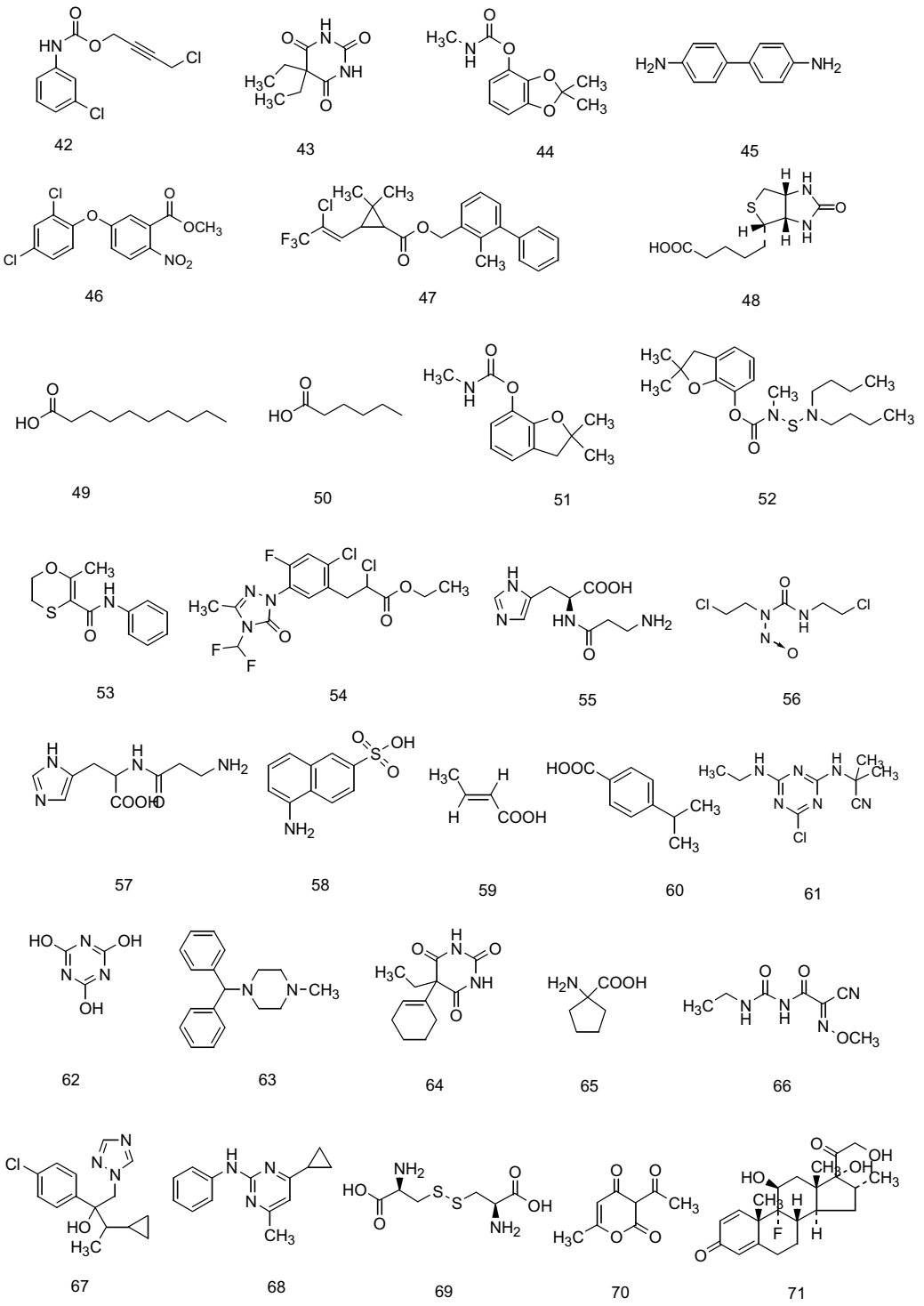

Figure 1. (*continued*)

Kinetic solubility measurement starts from a pre-dissolved sample of the compound, usually in dimethyl-sulfoxide (DMSO). Small volumes of the stock solution are added incrementally to the aqueous solution of interest until the solubility limit is reached, the resulting precipitation being detected optically. Although faster than thermodynamic solubility measurement, the DMSO might well operate as a co-solvent, dramatically enhancing the solubility of lipophilic compounds.[6] Because of these reasons and also because the sample is in amorphous state, kinetic solubility tends to overestimate thermodynamic solubility.

With this background, the use of QSPR methodologies to predict aqueous solubility appears as an interesting, increasingly popular alternative to solubility measurement: they are compatible with both HTS technologies and limited compound availability typical of early stages of development, since none of the samples of compound is needed for the estimation of solubility and relatively few computational time is needed for the predictions. Balakin et al. have proposed the following classification of in silico approaches for the assessment of aqueous solubility[3]:

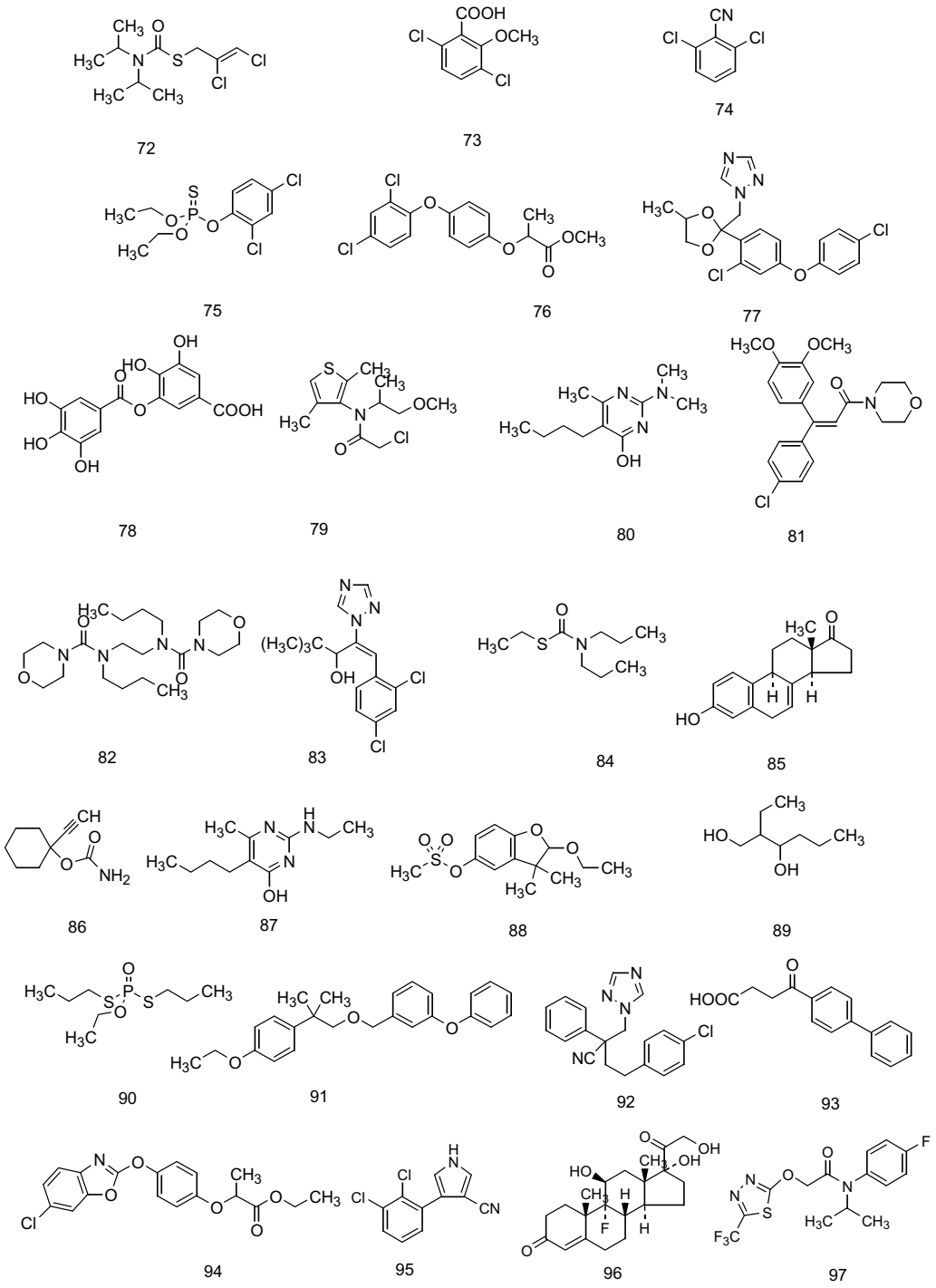

Figure 1. (*continued*)

Table 1
Different linear methods applied on the same 21-test set compounds

Lead author	Method	Type of descriptors	Number of parameters	rms	*N*/*d*	Ref.
Klopman	GCM	2D substructures	34	1.213	0.62	18
Yan	MLR	3D descriptors	40	1.286	0.53	50
Hou	GCM	atomic	78	0.664	0.27	51
Huuskonen	MLR	topologicals	30	0.810	0.70	52
Duchowicz	MLR	Dragon	3	1.202	7.00	This study

Table 2
Experimental and predicted values for $\log_{10}$Sol (mg ml^{-1})

No.	Chemical name	Exp.	Pred. Eq. (3)
Training set			
1	2,4,5-Trichlorophenol	0.079	−0.943
2	2,4-DB	−1.337	−1.29
3	2,6-Dibromoquinone-4-chlorimide	−1.230	−1.194
4	2-Cyclohexyl-4,6-dinitrophenol	−1.823	−1.275
5	2-Ethyl-1-hexanol	−0.056	0.321
6	3,4-Dinitrobenzoic acid	0.826	-0.503
7	4-Amino-2-sulfobenzoic acid	0.477	0.134
8	Acequinocyl	−4.173	−4.506
9	Acetamide	3.352	2.929
10	Acetamiprid	0.623	−0.6
11	Acetanilide	0.806	0.363
12	Acetazolamide	−0.009	1.765
13	Acetochlor	−0.652	−0.812
14	Acetylacetone	2.221	1.978
15	Acibenzolar-*S*-methyl	−2.113	−0.212
16	Aconitic acid	2.698	1.021
17	Acrylamide	2.806	2.46
18	Acrylonitrile	1.872	2.396
19	Adenine	0.013	1.706
20	Adipic acid	1.414	0.951
21	Alanine	2.214	2.441
22	Aldicarb	0.780	−0.115
23	Allidochlor	1.294	−0.022
24	Allobarbital	0.258	0.468
25	Alochlor	−0.620	−0.208
26	Alpha-acetylbutyrolactone	2.301	1.458
27	Amicarbalide	0.700	−0.282
28	Aminopromazine	−3.239	−1.933
29	Amitraz	−3.000	−2.534
30	Amobarbital	−0.220	0.497
31	Ancymidol	−0.187	−0.719
32	Aniline	1.556	0.979
33	ANTU	−0.222	−0.663
34	Arabinose	2.698	2.168
35	Ascorbic acid	2.522	2.005
36	Aspartic acid	0.912	2.095
37	Asulam	0.699	−0.106
38	Azidamfenicol	1.301	0.258
39	Azintamide	0.699	−0.762
40	Azoxystrobin	−2.000	−3.393
41	Badische acid	−0.225	−0.817
42	Barban	−1.958	−2.248
43	Barbital	0.873	0.857
44	Bendiocarb	−0.585	−0.596
45	Benzidine	−0.495	−1.026
46	Bifenox	−3.397	−3.207
47	Bifentrhin	−4.000	−3.5
48	Biotin	−0.658	−0.217
49	Capric acid	−1.209	−0.511
50	Caproic acid	1.012	1.128
51	Carbofuran	−0.495	−0.836
52	Carbosulfan	−3.522	−2.296
53	Carboxin	−0.701	−0.274
54	Carfentrazone-ethyl	−1.657	−2.226
55	Carisoprodol	−0.523	1.088
56	Carmustine	0.602	0.597
57	Carnosine	1.914	0.791
58	1,6-Cleve's acid	0.000	−0.577
59	Crotonic Acid	1.934	1.788
60	Cumic Acid	−0.821	−0.404
61	Cyanazine	−0.767	−0.417
62	Cyanuric Acid	0.301	1.614
63	Cyclizine	0.000	−1.525
64	Cyclobarbital	0.204	0.241
65	Cycloleucine	1.698	1.183
66	Cymoxanil	−0.051	1.391
67	Cyproconazole	−0.854	−1.399
68	Cyprodinil	−1.886	−1.58
69	Cystine	−0.951	0.781
70	Dehydroacetic Acid	−0.161	0.997
71	Dexamethasone	−1.051	−0.785
72	Diallate	−1.853	−1.154
73	Dicamba	−0.080	−1.003
74	Dichlobenil	−1.673	−0.705
75	Dichlofenthion	−3.610	−2.456
76	Diclofop-methyl	−3.096	−3.08
77	Difenoconazole	−1.823	−3.469
78	Digallic Acid	−0.301	−0.87
79	Dimethenamid	0.079	−0.951
80	Dimethirimol	0.079	0.019
81	Dimethomorph	−1.728	−1.966
82	Dimorpholamine	2.698	1.118
83	Diniconazole	−2.397	−1.725
84	EPTC	−0.426	−0.065
85	Equilin	−2.850	−2.385
86	Ethinamate	0.398	−0.064
87	Ethirimol	−0.699	−0.26
88	Ethofumesate	−1.301	−1.044
89	Ethohexadiol	1.623	0.772
90	Ethoprop	−0.125	−0.377
91	Etofenprox	−6.000	−3.456
92	Fenbuconazole	−3.699	−2.248
93	Fenbufen	−2.656	−2.072
94	Fenoxaprop-ethyl	−3.046	−3.228
95	Fenpiclonil	−2.318	−1.35
96	Fludrocortisone	−0.854	−1.204
97	Flufenacet	−1.252	−1.213
Test set val			
98	Flufenamic acid	−2.041	−1.585
99	Flumioxazin	−2.747	−2.17
100	Fluspirilene	−2.000	−4.587
101	Fluthiacet-methyl	−3.070	−2.002
102	Folic acid	−2.795	−2.532
103	Fumaric acid	0.845	2.145
104	Furametpyr	−0.648	−0.611
105	Furazolidone	−1.397	0.07
106	Ganciclovir	0.633	0.997
107	Gluconolactone	2.770	1.776
108	Glutamic acid	0.933	1.717
109	Glycine	2.396	2.883
110	Glyphosate	1.079	1.729
111	Guaifenesin	1.698	0.936
112	Haloperidol	−1.853	−2.916
113	Heptabarbital	−0.602	−0.651
114	Hexazinone	1.519	−0.268
115	Histidine	1.658	1.705
116	Hydrocortisone	−0.495	−1.325
117	Hydroflumethiazide	−0.523	−0.956
118	Hydroquinone	1.857	1.128
119	Hydroxyphenamate	1.397	0.05
120	Hydroxyproline	2.557	1.917
121	Hymexazol	1.929	2.116
122	Idoxuridine	0.301	0.688
123	Imazapyr	1.053	−0.193
124	Imazaquin	−1.045	−1.316
125	Imazethapyr	0.146	−0.625
126	Iridomyrmecin	0.301	−0.272
127	Isoflurophate	1.187	0.614
128	Isoleucine	1.536	1.384
129	Isoniazid	2.146	1.64
130	Isophorone	1.079	0.542
131	Ketanserin	−2.000	−2.602
132	Khellin	0.017	−0.417
133	Lenacil	−2.221	−0.913
134	Linuron	−1.124	−1.373
135	Methomyl	1.763	0.648
136	PABA	0.769	0.586
137	*p*-Fluorobenzoic acid	0.079	0.349
138	Phthalazine	1.698	0.776
139	Phthalic Acid	0.846	0.228
140	Phthalimide	−0.444	0.303
141	*p*-Hydroxybenzoic Acid	0.699	0.645
142	Picloram	−0.367	−0.035
143	Picric Acid	1.103	−0.426
144	Pirimicarb	0.431	−0.424
145	Thionazin	0.057	0.222
Test set 21			
146	2,2′,4,5,5′-PCB	−5.376	−3.932
147	Benzocaine	−0.102	0.104
148	Theophylline	0.886	1.316

Table 2 (*continued*)

No.	Chemical name	Exp.	Pred. Eq. (3)
149	Antipyrine	2.665	0
150	Atrazine	−1.216	−0.509
151	Phenobarbital	0.026	−0.208
152	Diuron	−1.392	−0.713
153	Nitrofurantoin	−1.003	0.511
154	Phenytoin	−1.588	−0.581
155	Testosterone	−1.610	−1.955
156	Lindane	−2.136	−2.381
157	Parathion	−1.826	−1.453
158	Phenolphthalein	−0.397	−2.581
159	Malathion	−0.841	−0.523
160	Chlorpyrifos	−3.125	−1.833
161	Prostaglandin E2	0.077	−2.344
162	DDT	−5.530	−4.244
163	Chlordane	−4.247	−4.302
164	Diazepam	−1.301	−2.282
165	Aspirin	0.663	−0.074
166	Diazinon	−1.397	−0.535

1. Solubility methods based on other experimental measurements, such as the melting point and the experimental log*P* value. Although they present good accuracy, the greatest drawback of these methodologies is the requirement of experimentally measure one or more physicochemical properties, which in some cases might be difficult or impossible to determine (e.g., compounds with very low or very high log*P* values and compounds with very high melting points that decompose before melting).
2. Methods exploring 3D structure, which suppose either low speed of calculation (when ab initio approaches are employed) or previous optimization of molecular structures.
3. Methods using low dimensional descriptors (1D–2D). These include the group contribution methods (GCM) and QSAR approaches relying on topological descriptors. They are not computationally demanding, neither they require optimization of the molecular structure. GCM are easy to apply, relying solely on the sum of contributions of each molecular structure fragment to the aqueous solubility.[10–12] The basic assumption of this approach is the transferability concept for a group; if this hypothesis does not hold, then GCM can be corrected with experimental data when available to achieve better predictions. The methods proposed by Nirmalakhandan et al.,[13] Suzuki et al.,[14] Kuhne et al.,[15] Lee et al.,[16] and Klopman et al.[17,18] belong to this category. Among all these approaches, only Klopman et al.'s approach is a pure and general group contribution model without using additional experimental parameters. Although GCM have a simple and practical implementation, some common drawbacks of this methodology are the following: (a) they require a large data set to obtain a contribution of each functional group; (b) in its basic form (without corrections) it cannot model isomeric structures; (c) they may contain a 'missing fragment' problem, which means that if a compound contains a missing fragment which cannot be defined by the group contribution model, its aqueous solubility cannot be precisely predicted; (d) measured data are not always available to extend these methods to strange compounds such as molecules containing fused aromatic rings or to organo-metallic compounds.

Table 1 summarizes different linear estimation methods for solubility prediction in terms of type and number of structural descriptors used to derive the model, and the root mean square error (rms) against a common test set of 21 'classic compounds' found in many solubility prediction papers.[18] It has been pointed out that solubility modeling efforts have suffered from some basic concerns, among them: training sets that are not drug-like, lack of structural diversity, unknown experimental error, incorrect tautomers or structures, neglect of ionization and crystal packing effects, over-sampling of compounds with low molecular weight and range in solubility data that is not pharmaceutically relevant.[4,19] In this paper, we present a QSPR that answers to some of these issues, since it was developed from a structural diverse training set composed by drug-like compounds with more than half the data set presenting solubility values below 1 mg ml^{-1}. Note that low solubility compounds are actually the ones one would like to obtain more accurate predictions,[4,19] since they have higher probability of presenting difficulties in pre-clinic and clinic assays and formulation stages. We were also careful not to over-represent compounds with low molecular weight. In the present analysis, we decided to use Multivariable Linear Regression (MLR)-based methods instead of the GCM approach for analyzing the aqueous solubilities of 166 organic compounds. A great number of theoretical molecular descriptors are simultaneously explored by including definitions of all classes. For this task, we employ the linear variable subset selection approach Replacement Method (RM),[20–23] and we draw conclusions by contrasting our results with other previously reported linear models of the literature.

2. Methods

2.1. Data set

The experimental aqueous solubilities (Sol) measured at 298 K and expressed in mg ml^{-1} for 145 structurally diverse drug-like organic compounds were extracted from Merck Index 13th.[24] Solubility data were checked at ChemID Plus (National Library of Medicine, National Institute of Health).[25] No differences in solubility data were found between Merck Index and ChemID records except for crotonic acid (ΔSol = 0.053 log units), cyanazine (ΔSol = 0.003), dexamethasone (ΔSol = 0.050) and PABA (ΔSol = 0.016). In those cases ChemID data were considered. None solubility record at 25° was found in ChemID Plus for 4-amino-2-sulfobenzoic acid, acequinocyl, aconitic acid, amicarbalide, aminopromazine, ascorbic acid, axocystrobin, ethirimol and furametpyr. For modeling purposes, these data are converted into logarithm units ($\log_{10}$Sol) and are presented in Table 2; all the molecular structures are drawn in Figures 1 and 2. The molecular set was split into a 97-compound training set (train) and a 48-compounds test set (val), selecting the members of each set in such a way to share similar structural characteristics of the compounds. Additionally, we also used an external molecular set (test set 21) that was not involved during the model design, composed of 21 well-known compounds found in many solubility prediction papers,[4,18] in order to further examining the model's validation. In a recent work, we have already used this 145 data set (plus aspirin, diazepam, and diazinon, which in the current study are part of the classic 21-compound test) for modeling of aqueous solubility through the RM.[26] In that opportunity, however, we conditioned the model to include at least one out of twelve descriptors inspired in Lipinski rules.[27] In the present study, we imposed no restriction regarding the descriptors included in the models.

Note that most of the drugs that compose the training and test sets accomplish several drug-likeness criteria. It can be noticed that more than 99% of the data set observes the Lipinski-rule criteria for estimating drug oral bioavailability,[28] while more than 93% accomplish Veber et al. rule.[29] More than 99% of the data set also accomplishes more general rules for evaluating drug-likeness extracted from several recent publications[30–32]: $100 \leqslant$ molecular weight $\leqslant 800$ g mol^{-1}; log$P \leqslant 7$; number of H bond acceptors $\leqslant 10$; number of H bond donors $\leqslant 5$; rotatable bonds $\leqslant 15$; halogen

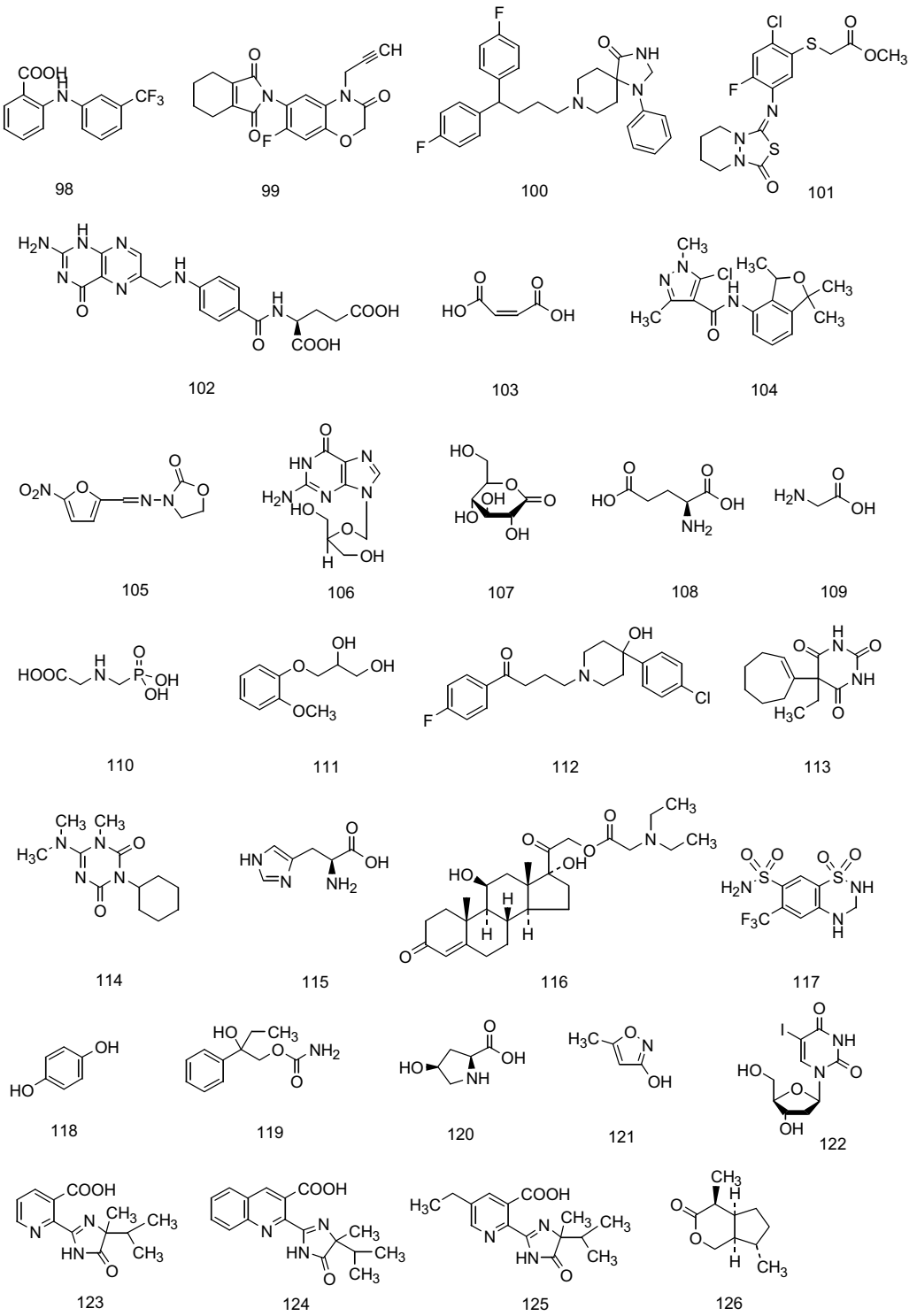

Figure 2. Molecular structures for the test set compounds ($N = 48$).

atoms $\leqslant 7$; alkyl chains $\leqslant (CH_2)_6CH_3$; no perfluorinated chains: $CF_2CF_2CF_3$; no big size ring with more than seven members; no presence of other atoms than C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca or Li and; presence of at least one N or O atom. Moreover, note that low molecular weight compounds are not over-represented in this molecular set. The structural diversity of the training set was assessed through calculation of the average Tanimoto intermolecular distances (based on atom pairs) for all the possible pairs of structures that could be derived from the training set. For this purpose, we used de PowerMV software provided by the National Institute of Statistical Sciences.[33] According to the results, the average Tanimoto intermolecular distance for the training set is 0.781 with a SD of 0.412, which confirms the high structural diversity of the training set. Figure 3 includes a histogram representing the distribution of the 166 aqueous solubilities under study, which suggests that the experimental sample is normally distributed over more than four logarithmic units and can thus be employed in regression analysis.

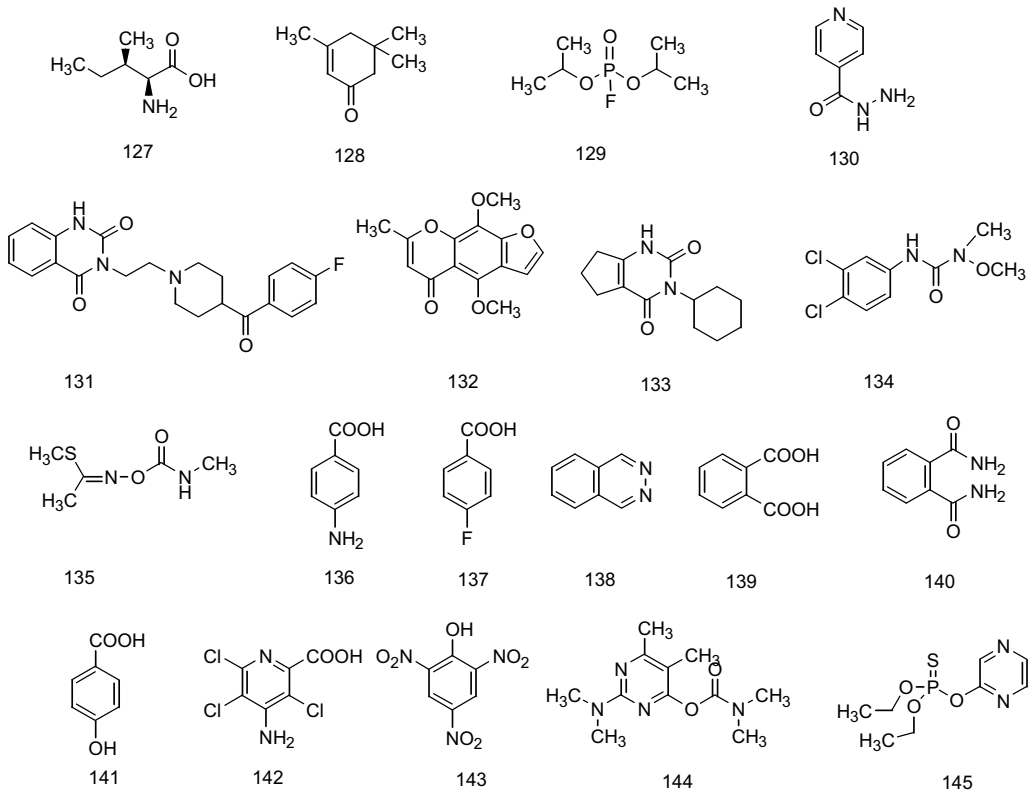

Figure 2. (*continued*)

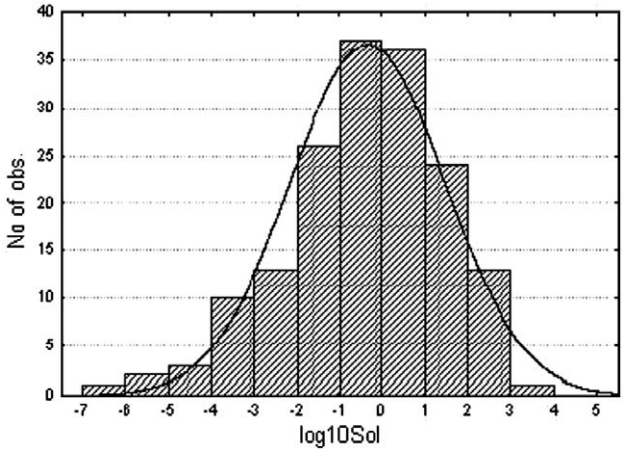

Figure 3. Normal distribution of the experimental $\log_{10}$Sol values under analysis $N = 166$.

2.2. Molecular descriptors

The structures of the compounds were firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in the Hyperchem 6.03 package,[34] and the resulting geometries were further refined by means of the Semi-Empirical Molecular Orbitals Method PM3 (Parametric Method-3) using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal Å^{-1}.

We computed 1497 molecular descriptors using the software Dragon 5.0,[35] including descriptors of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centred Fragments, Empirical and Properties.[36] Furthermore, four molecular descriptors were derived taking into consideration Lipinski's rule, based on combinations of the detour index *dd* from the Chemical Graph Theory (calculated as the ratio between the half sum of the elements of the Detour Matrix (DD) and molecular features related to solubility such as the number of H donors (D), the number of H acceptors (A), and the number of hetero-atoms (H) present in the molecular structure).[26,27,37] We also considered the square and cubic roots of these last descriptors. Finally, five quantum-chemical descriptors not provided by the program Dragon were added to the pool: molecular dipole moments, total energies, homo–lumo energies, and homo–lumo gap (Δhomo–lumo) calculated at the PM3 level. The total pool of explored descriptors consisted on $D = 1514$ variables.

2.3. Model search

The computer system Matlab 5.0 was used in all our calculations.[38] Our purpose was to search the optimal subset of d descriptors from the total number of D descriptors which to accomplish the following criterion: $d \ll D$ and d with minimum standard deviation S:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} \mathrm{res}_i^2 \qquad (1)$$

where N is the number of molecules in the training set, and res_i the residual for molecule i (difference between the experimental and predicted property **p**). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where **d** is a point in a space of $D!/[d!(D-d)!]$

Table 3
Linear QSPR models established for the training set of aqueous solubilities ($N = 97$)

d[a]	Descriptors involved	R[b]	S[c]	FIT[d]	R_{loo}[e]	S_{loo}[f]	R_{val}[g]	S_{val}[h]
1	DP03	0.722	1.257	1.053	0.708	1.283	0.794	1.047
2	DP03, MLOGP	0.831	1.016	2.071	0.817	1.054	0.798	0.983
3	**X1sol, RDF060u, MLOGP**	**0.871**	**0.903**	**2.747**	**0.849**	**0.971**	**0.848**	**0.899**
4	X1sol, RDF060u, RDF020e, MLOGP	0.889	0.844	3.078	0.870	0.911	0.838	0.986
5	Sp, nR09, H3D, Mor04u, MLOGP	0.895	0.829	2.991	0.878	0.890	0.891	0.758

The best relationship found appears in bold.
[a] d: number of descriptors in the linear regression.
[b] R: correlation coefficient of the model.
[c] S: standard deviation of the model.
[d] FIT: Kubinyi function.
[e] R_{loo}: R of Leave-One-Out.
[f] S_{loo}: S of Leave-One-Out.
[g] R_{val}: R of validation test set.
[h] S_{val}: S of validation test set.

Table 4
Symbols for molecular descriptors involved in different models

Molecular descriptor	Dim[a]	Type	Description
DP03	3D	Randic molecular profiles	Molecular profile No. 3
MLOGP	1D	Properties	Moriguchi octanol–water partition coefficient
X1sol	2D	Topological	Solvation connectivity index chi-1
RDF060u	3D	Radial Distribution Function	Radial distribution function – 6.0/unweighted
RDF020e	3D	Radial Distribution Function	Radial distribution function – 2.0/weighted by atomic Sanderson electronegativities
Sp	0D	Constitutional	Sum of atomic polarizabilities (scaled on carbon atom)
nR09	0D	Constitutional	Number of nine-membered rings
H3D	3D	Geometrical	3D-Harary index
Mor04u	3D	3D-MoRSE	3D-MoRSE-signal 04/unweighted

[a] Dim, dimensionality of the descriptor.

ones. Usually, a full search (FS) of optimal variables is unfeasible because it requires $D!/[d!(D-d)!]$ linear regressions. Some time ago we proposed the Replacement Method (RM) that produces linear QSPR-QSAR models that are quite close the FS ones with much less computational work.[20–23] This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$. The RM gives models with better statistical parameters than the Forward Stepwise Regression procedure and similar ones to the more elaborated Genetic Algorithms.[39,40]

The Kubinyi function (FIT)[41] is a statistical parameter that closely relates to the Fisher ratio (F), but avoids the main disadvantage of the latter that is too sensitive to changes in small d values and poorly sensitive to changes in large d values. The $FIT(\mathbf{d})$ criterion has a low sensitivity to changes in small d values and a substantially increasing sensitivity for large d values. The greater the FIT value the better the linear equation. It is given by the following equation, where $R(\mathbf{d})$ is the correlation coefficient for a model with d descriptors.

$$FIT = \frac{R^2(N-d-1)}{(N+d^2)(1-R^2)} \tag{2}$$

2.4. Model internal validation

The theoretical 'internal validation' practiced over each developed linear model is based on the Leave-More-Out Cross-Validation procedure (l-n%-o),[42] with n% representing the percentage of molecules removed from the training set. The number of cases for random data removal analyzed in every l-n%-o is of 5,000,000. The percentage n% depends simultaneously upon the number of compounds (N), as one cannot remove many molecules from the training set if a small sample is analyzed as the normality condition of the fitted data has to be obeyed, and upon their structural diversity, since if the molecules are structurally very different, more compounds would have to be removed from the set for checking the predictive performance of the model. We choose the value of $n\% = 10\%$ (10 compounds) in Cross-Validation in order to properly validate the QSAR equations.

In addition, we applied the y-randomization technique[43] with the purpose of demonstrating that the model established does not result from happenstance but involves a real structure–property relationship. This method consists on scrambling the experimental property of each compound in such a way that it does not correspond to the respective compound. After analyzing 5,000,000 cases of y-randomization for each developed QSPR, the smallest S value obtained using this procedure turned out to be a poorer value when compared to the one found when considering the true calibration.

2.5. Orthogonalization procedure

We employ the orthogonalization procedure introduced several years ago by Randic[44,45] as a way of improving the statistical interpretation of the model built by interrelated indices. From our point of view, the co-linearity of the molecular descriptors should be as low as possible, because the interrelatedness among the different descriptors can lead to highly unstable regression coefficients, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficients of the model. The crucial step of the orthogonalization process is the choice of an appropriate order of orthogonalization, which in present analysis is the order that maximizes the correlation between each orthogonal descriptor and the observed aqueous solubilities. From now on, an orthogonalized descriptor will be represented with notation $\mathbf{\Omega}$.

3. Results and discussion

The application of the RM method on the training set of 97 heterogeneous drugs leads to the best 1–5 variables linear regression models listed in Table 3, while the specific details for all the molecular descriptors reported in this article are provided in Table 4. A close inspection of Table 3 reveals that the best linear QSPR equation found for modeling the aqueous solubility of the organic compounds includes the following satisfactory three molecular descriptors relationship:

$$\log_{10}Sol = -0.435(\pm0.03)\cdot\Omega(\text{X1sol}) - 0.503(\pm0.06)\cdot\Omega(\text{MLOGP}) + 0.0767(\pm0.01)\cdot\Omega(\text{RDF060u}) + 2.970(\pm0.3) \quad (3)$$

$N_{train} = 97$, $N_{train}/d = 32.333$, $R = 0.871$, $S = 0.903$, $FIT = 2.747$
$R_{loo} = 0.849$, $S_{loo} = 0.971$, $R_{l-10\%-o} = 0.809$, $S_{l-10\%-o} = 1.090$, $p < 10^{-4}$
$N_{val} = 48$, $R_{val} = 0.848$, $S_{val} = 0.899$

Here, the absolute errors of the regression coefficients are provided in parentheses, p is the significance of the model, and loo sub-index stands for the Leave-One-Out Cross-Validation technique.[42] The QSPR derived does not incorporate redundant structural information, as it involves orthogonal descriptors. Furthermore, by means of a proper standardization[39] of such orthogonal variables it is feasible to assign a greater importance to those molecular descriptors that exhibit larger absolute standardized coefficients (st.coeff.). The order of appearance of each descriptor within the QSPR of Eq. 3 corresponds to its order of importance in the established relationship, and each variable includes the following standardized coefficients: Ω(X1sol): 0.71, Ω(MLOGP): 0.42, and Ω(RDF060u): 0.28.

Table 2 also includes the predicted residuals as obtained via Eq. (3) for the training and test sets, while the plot of predicted versus experimental aqueous solubilities shown in Figure 4a suggests that the 97 training and 48 test set val compounds follow a straight line. The behavior of the plotted residuals in terms of the predictions in Figure 4b leads to a normal distribution. This figure includes two calibration outliers with a residual exceeding the value $2S = 1.806$: compounds **15** (Acibenzolar-*S*-Methyl, 1.902) and **91**(Etofenprox, −2.545), while none of the training compounds exceed the value $3S = 2.709$; the presence of these outliers may be attributed exclusively to be a pure consequence of the limited number of structural descriptors participating in Eq. 3, since this model have a high ratio of number of observations to number of parameters ($N/d = 32.333$). The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion or exclusion of compounds, as measured by the loo parameters $R_{loo} = 0.849$ and $S_{loo} = 0.971$, and by the more severe test of higher percentage of compounds exclusion $R_{l-10\%-o} = 0.809$ and $S_{l-10\%-o} = 1.090$. These results are in the range of a validated model: $R_{l-n\%-o}$ must be greater than the value of 0.50, according to the specialized literature.[46] Furthermore, the predictive capability of the so-established equation is demonstrated by its performance in the test set val, leading to $R_{val} = 0.848$ and $S_{val} = 0.899$. Finally, after analyzing 5,000,000 cases for *y*-randomization, the smallest S value obtained using this procedure was 1.650, a poorer value when compared to the one found considering the true calibration (S 0.903). In this way, the robustness of the model could be assessed, showing that the calibration was not a fortuitous correlation and therefore results in a structure–activity relationship.

The three structural descriptors mentioned in Eq. 3 quantify different aspects of the molecular geometry and can be classified as follows: (i) a topological 2D-descriptor: X1sol, the solvation connectivity index chi-1, (ii) a Property 1D-descriptor: MLOGP, the Moriguchi octanol–water partition coefficient; and (iii) a Radial Distribution Function 3D-descriptor: RDF060u, the radial distribution function −6.0/unweighted. As can be appreciated, different definitions of descriptors are needed to correctly represent the structures for the drug-like heterogeneous compounds. Figure 5 includes the histograms of the 166 organic compounds for each of the three descriptors appearing in the optimal QSPR equation found.

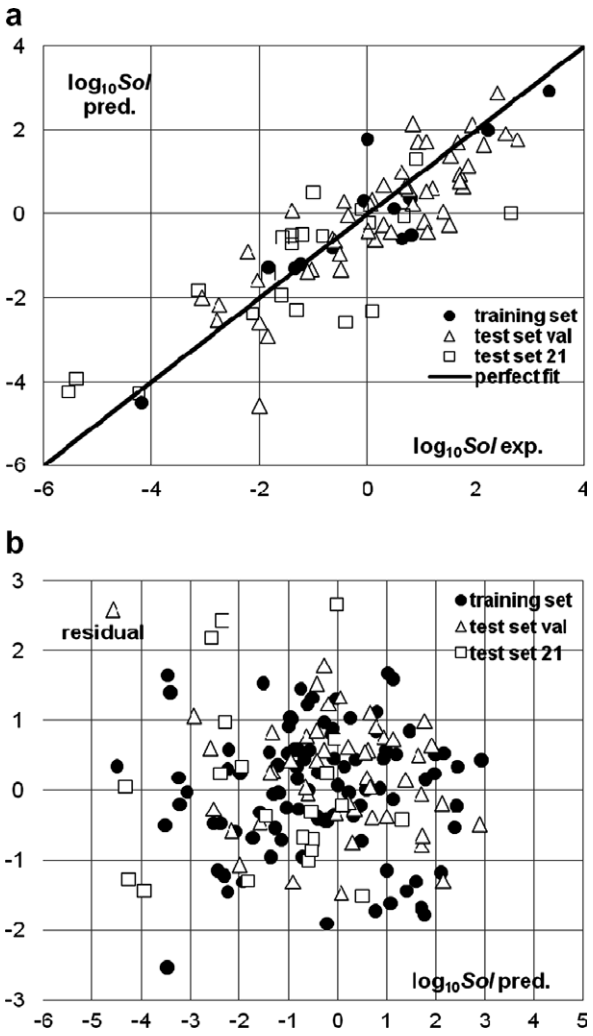

Figure 4. (a) Predicted (Eq. 3) versus experimental $\log_{10}$Sol for the training and test sets. (b) Dispersion plot of the residuals for the training and test sets according to Eq. 3.

The most important structural factor of the model, the bidimensional descriptor X1sol, was proposed by Zefirov and Palyulin[47] in 1991 in order to treat the enthalpies of non-specific solvation. For instance, the solvation enthalpy of propane ($CH_3CH_2CH_3$) and di-methyl-mercury (CH_3HgCH_3) differs enormously, but both of these molecules are represented by the same hydrogen depleted graph, and, hence, have the identical topological indices which do not take into account atom types. The solvation index was created exactly to differentiate such cases, having the following general formula when calculated for hydrogen- and fluorine-depleted molecular graphs:

$$\text{Xmsol} = (1/2^{m+1})\sum \frac{Z_i Z_j \ldots Z_k}{(\delta_i\delta_j\ldots\delta_k)^{1/2}} \quad (4)$$

where m is the order of index; summation is over all sub-graphs of order m; $\delta_i\delta_j\ldots\delta_k$ are connectivities of vertexes of sub-graph; and

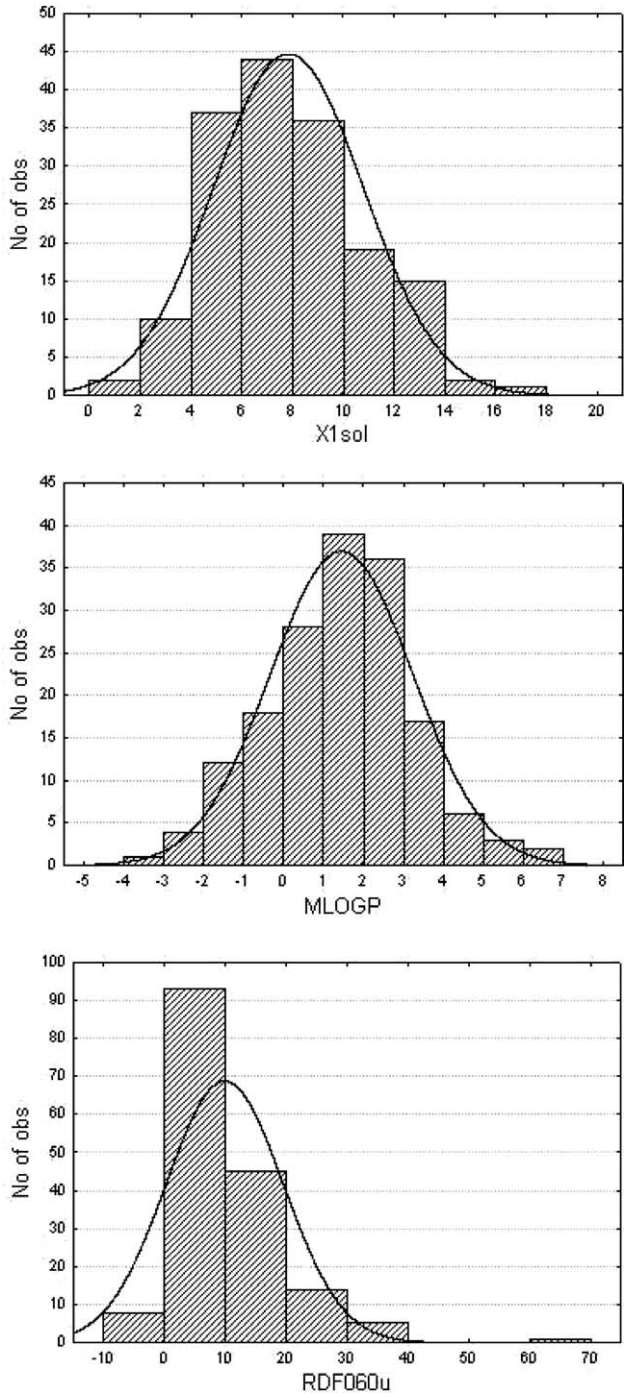

Figure 5. Histograms for the molecular descriptors appearing in the QSPR solubility model ($N = 166$).

$Z_iZ_j \ldots Z_k$ are coefficients characterizing the atom size, which coincide to the number of the period in the Periodic Table. The term $1/2^{m+1}$ just normalizes values of Xmsol to provide their coincidence with the connectivity index Xm for the elements of the second row. The second important descriptor involved in Eq. 3 corresponds to the Moriguchi octanol–water partition coefficient,[48] revealing that a compound's hydrofobicity plays a crucial role in explaining the aqueous solubility data. Finally, the contribution of a 3D-Radial Distribution Function[49] helps to improve the predictive power of the QSPR. Such a kind of molecular descriptor defined for an ensemble of atoms may be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, incorporating different types of atomic properties in order to differentiate the nature and contribution of atoms to the property being modeled. For the case of RDF060u, the sphere radius is of 6.0 Å and no atomic property is employed, thus characterizing the molecular size.

It is feasible to discuss the numerical effect of the optimal subset of structural descriptors selected in Eq. 3 on the aqueous solubility predictions. Since the orthogonal descriptor Ω(X1sol) is numerically positive for all the structures under study, its contribution to $\log_{10}$Sol results in a negative quantity, according to the regression coefficient (−0.435). This causes that chemical compounds displaying greater values of Ω(X1sol) would tend to exhibit lower predicted values of aqueous solubilities. For the case of the orthogonal variable Ω(MLOGP), drugs manifesting higher positive values of this descriptor would tend to manifest their preference to the octanol lipophilic phase rather than to the water phase, and according to the sign of the regression coefficient in Eq. 3 (−0.503) would lead to a lower prediction of the aqueous solubilities. Finally, the tri-dimensional descriptor Ω(RDF060u) would tend to lead to higher predictions of $\log_{10}$Sol whenever it presents higher numerical values.

Applying now the designed QSPR model of Eq. 3 to the classical test set 21, whose data are considered 'unknown' and that do not participate during the model development (as is the case of test set val), leads to a square root mean quadratic residual (rms) of 1.202. The statistical quality achieved on this test set is comparable to that obtained by the previously reported models for aqueous solubilities in Table 1, and the main advantage here is that only three molecular descriptors are employed to model the physical property and thus leads to a favorable ratio $N/d = 7$. This equation results in a superior predictive quality than that obtained by the GCM of Klopman (rms = 1.213) involving 34 parameters,[18] and also outperforms the MLR of Yan (rms = 1.286) using 40 parameters.[50]

4. Conclusions

The chemical information encoded by three theoretical molecular descriptors of the one-, two-, and three-types participating in a linear QSPR model enabled to explain the variation of the experimental aqueous solubilities in a satisfactory extent, and allowed a proper characterization of structurally heterogeneous drug-like organic compounds from both the training and test sets. The QSPR designed involved molecular descriptors that have a quite direct interpretation, and this relationship proved to have general applicability. The statistical parameters of the proposed model compare fairly well with others published previously based on Group Contribution methods. Furthermore, among the different linear regression based-algorithms, the Replacement Method continues demonstrating to be an efficient technique for the search of a reduced set of numerical variables from a huge number of them. This has application for the analysis of any physicochemical, biological, or pharmacological property of interest.

Acknowledgments

P.R.D. and A.T. thank to the National Council of Scientific and Technological Research (CONICET) for supporting this work. L.B.B. thanks ANPCyT and Universidad Nacional de La Plata for supporting this work.

References and notes

1. Schuster, D.; Laggner, C.; Langer, T. *Curr. Pharm. Des.* **2005**, *11*, 3545.
2. Stegemann, S.; Leveiller, F.; Franchi, D.; de Jong, H.; Lindén, H. *Eur. J. Pharm. Sci.* **2007**, *31*, 249.
3. Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. *Curr. Med. Chem.* **2006**, *13*, 226.
4. Delaney, J. S. *Drug Discov. Today* **2005**, *10*, 289.
5. Goodwin, J. J. *Drug Discov. Today Technol.* **2006**, *3*, 67.

6. Alsenz, J.; Kansy, M. *Adv. Drug Deliv. Rev.* **2007**, *59*, 546.
7. Bhattachar, S. N.; Deschenes, L.; Wesley, J. A. *Drug Discov. Today* **2006**, *11*, 1012.
8. Di, L.; Kerns, E. H. *Drug Discov. Today* **2006**, *11*, 446.
9. Smith, C. J.; Hansch, C. *Food Chem. Toxicol.* **2000**, *38*, 637.
10. Artist, http://www.ddbst.de/new/Win_DDBSP/frame_Artist.htm.
11. ChemEng Software Design, http://www.cesd.com/chempage.htm.
12. Predict, http://www.mwsoftware.com/dragon/desc.html.
13. Nirmalakhandan, N. N. P.; Speece, R. E. *Environ. Sci. Technol.* **1989**, *23*, 708.
14. Suzuki, T. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149.
15. Kuhne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schuurmann, G. *Chemosphere* **1995**, *30*, 2061.
16. Lee, Y.; Myrdal, P. B.; Yalkowsky, S. H. *Chemosphere* **1996**, *33*, 2129.
17. Klopman, G.; Zhu, H. *J. Chem. Inf. Model.* **2001**, *41*, 439.
18. Klopman, G.; Wang, S.; Balthasar, D. M. *J. Chem. Inf. Model.* **1992**, *32*, 474.
19. Johnson, S. R.; Zheng, W. *AAPS J.* **2006**, *8*, E27.
20. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. *Chem. Phys. Lett.* **2005**, *412*, 376.
21. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179.
22. Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. *Bioorg. Med. Chem.* **2006**, *14*, 5876.
23. Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. *Chemometr. Intell. Lab.* **2006**, *81*, 180.
24. The Merck Index An Encyclopedia of Chemicals, Drugs, and Biologicals; Merck & Co.: NJ, 2001.
25. Division of Specialized Information Services, National Institute of Health. ChemID Plus. http://chem.sis.nlm.nih.gov/chemidplus/.
26. Duchowicz, P. R.; Talevi, A.; Bellera, C.; Bruno-Blanch, L. E.; Castro, E. A. *Bioorg. Med. Chem.* **2007**, *15*, 3711.
27. Talevi, A.; Castro, E. A.; Bruno-Blanch, L. E. *J. Arg. Chem. Soc.* **2006**, *44*, 129.
28. Lipinski, C. A.; Lombardo, F.; Dominy, D. W.; Feeney, P. J. *Adv. Drug Deliver. Rev.* **2001**, *46*, 3.
29. Veber, D. F.; Johnson, S. R.; Cheng, H.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615.
30. Charifson, P. S.; Walters, W. P. *J. Comput. Aided Mol. Des.* **2002**, *16*, 311.
31. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. *Mol. Divers.* **2006**, *10*, 339.
32. Walters, W. P.; Murcko, M. A. *Adv. Drug Deliv. Rev.* **2002**, *54*, 255.
33. Liu, K.; Feng, J.; Young, S. S. *J. Chem. Inf. Model.* **2005**, *45*, 515. PowerMV v.0.61. http://www.niss.org/PowerMV.
34. Hyperchem 6.03 (Hypercube) http://www.hyper.com.
35. Dragon 5.0, Evaluation Version, http://www.disat.unimib.it/chm.
36. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
37. Harary, F. *Graph Theory*; Addison-Wesley, 1969.
38. Matlab 7.0, The MathWorks Inc.
39. Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley& Sons: New York, 1981.
40. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
41. Kubinyi, H. *Quant.-Struct.-Act. Relat.* **1994**, *13*, 393.
42. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Model.* **2003**, *43*, 579.
43. Wold, S.; Eriksson, L. *Chemometrics Methods in Molecular Design*; VCH: Weinheim, 1995.
44. Randic, M. *J. Chem. Inf. Model.* **1991**, *31*, 311.
45. Randic, M. *New J. Chem.* **1991**, *15*, 517.
46. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Model.* **2002**, *20*, 269.
47. Antipin, I. S.; Arslanov, N. A.; Palyulin, V. A.; Konovalov, A. I.; Zefirov, N. S. *Dokl. Akad. Nauk. SSSR* **1991**, *316*, 925 (*Chem. Abstr. 115*, 91390).
48. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsuchita, Y. *Chem. Pharm. Bull.* **1992**, *40*, 127.
49. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Model.* **2002**, *42*, 693.
50. Yan, A.; Gasteiger, J. *J. Chem. Inf. Model.* **2003**, *43*, 429.
51. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. *J. Chem. Inf. Model.* **2004**, *44*, 266.
52. Huuskonen, J. *J. Chem. Inf. Model.* **2000**, *40*, 773.