Journal of the Serbian Chemical Society
JSCS@tmf.bg.ac.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. 75 (4) 483–495 (2010)
JSCS–3981

UDC 54–145.2+615.2:54.02:001.891.53
Original scientific paper

The importance of the accuracy of the experimental data for the prediction of solubility

SLAVICA ERIĆ[1]*#, MARKO KALINIĆ[1], ALEKSANDAR POPOVIĆ[1], HALID MAKIĆ[2], ELVISA CIVIĆ[2] and MEJRA BEKTAŠEVIĆ[2]

[1]*Faculty of Pharmacy, University of Belgrade, Vojvode Stepe 450, 11000 Belgrade, Serbia and* [2]*Biotechnical Faculty, University of Bihać, Kulina Bana 2, 77000 Bihać, Bosnia and Herzegovina*

(Received 9 August, revised 7 October 2009)

Abstract: Aqueous solubility is an important factor influencing several aspects of the pharmacokinetic profile of a drug. Numerous publications present different methodologies for the development of reliable computational models for the prediction of solubility from structure. The quality of such models can be significantly affected by the accuracy of the employed experimental solubility data. In this work, the importance of the accuracy of the experimental solubility data used for model training was investigated. Three data sets were used as training sets – data set 1, containing solubility data collected from various literature sources using a few criteria (n = 319), data set 2, created by substituting 28 values from data set 1 with uniformly determined experimental data from one laboratory (n = 319), and data set 3, created by including 56 additional components, for which the solubility was also determined under uniform conditions in the same laboratory, in the data set 2 (n = 375). The selection of the most significant descriptors was performed by the heuristic method, using one--parameter and multi-parameter analysis. The correlations between the most significant descriptors and solubility were established using multi-linear regression analysis (MLR) for all three investigated data sets. Notable differences were observed between the equations corresponding to different data sets, suggesting that models updated with new experimental data need to be additionally optimized. It was successfully shown that the inclusion of uniform experimental data consistently leads to an improvement in the correlation coefficients. These findings contribute to an emerging consensus that improving the reliability of solubility prediction requires the inclusion of many diverse compounds for which solubility was measured under standardized conditions in the data set.

Keywords: aqueous solubility prediction; experimental data; model training; heuristic method.

* Corresponding author. E-mail: seric@pharmacy.bg.ac.yu
# Serbian Chemical Society member.
doi: 10.2998/JSC090809022E

INTRODUCTION

The solubility of drugs and drug-like compounds has been the subject of extensive studies aimed at finding a way to predict solubility from molecular structure. The aqueous solubility of a drug is an important factor that influences its absorption by, distribution in and elimination from the body.[1] Since poor pharmacokinetics is one of the major causes for late stage drug development failures,[2] it is clear that properties such as solubility need to be considered very early in the drug discovery process. Therefore, a reliable tool for the prediction of solubility from structure alone would be of great importance to help in the elimination of unsuitable candidates and reduction of overall development attrition rates.

A considerable number of *in silico* models for the prediction of solubility have been proposed over the past decade.[3–22] These utilize an ever-growing variety of approaches that differ in the way structure is represented, in the nature of the descriptors or properties that are derived from molecular structure and in the regression methods used. The sheer volume of publications on novel methods for the prediction of solubility seems to indicate that none of the existing models is fully satisfactory.[23] Consistent with this, very few of the proposed models for prediction of solubility have found practical implications in the drug discovery process, probably due to low prediction reliability. Whilst most of the published models perform satisfactorily with the test sets used for their validation, their performance with more diverse data plummet considerably.[24]

While significant progress has been made in developing new modelling techniques, there is, nevertheless, an emerging consensus that moving forward will require focusing on altogether different issues affecting the performance of existing models. Most of the recent reviews on solubility prediction indicate that solubility modelling efforts have suffered from some basic faults, such as training sets that are not drug-like, unknown or high experimental error, lack of structural diversity, incorrect tautomers or structures, neglect of ionization, no consideration of salt and/or common ion issues, avoidance of crystal packing effects and range of solubility data that is not pharmaceutically relevant.[25]

One of the issues is that the design of a good quality training set is something often overlooked.[26] It is worth noting that any model is only as good as the data used for its generation. A training set codifies the relationship between the relevant property and chemical structure; therefore, the applicability domain as well as model reliability will depend heavily upon the choice of the training set. The most limiting factor in the choice of a proper training set is the accuracy of the experimental solubility data. Katritzky *et al*. demonstrated that the average standard deviation for solubility measurements originating from different sources is as high as 0.6log *S* units.[27] Occasionally, the solubility values reported for one compound may differ by 2–3 log units; this large difference may originate from different experimental protocols. There are differences in sample concentration,

co-solvent presence, co-solvent concentration, incubation times, thermodynamic methods, kinetic methods, *etc*.[28] In addition, inclusion of values that were not distinctly reported, such as those for intrinsic solubility, and other unintentional errors all contribute significantly to the overall experimental error that can plague a data set.[29,30] In a recent work by Taskinen and Norinder, in which over 30 models from the literature were reviewed, it was concluded that improving the accuracy and applicability will "require more consideration of the consistency of the experimental solubility data and the training set composition".[24] Other authors also stressed that further development will require large, diverse sets of high-quality, uniformly determined experimental data.[30,31]

Despite the importance of this issue, there are very few published works dealing primarily with proposing carefully designed training sets. In one aspect this is understandable – consistent solubility data are not widely available and determining them for a "QSPR-significant" number of compounds would be a time--consuming, laborious and expensive endeavour. On the other hand, addressing the issue of a more selective collection of data from the literature is feasible. One such example is the data set proposed by Rytting *et al*., who set out clear criteria for the inclusion of experimental data.[32]

Increasing interest in the importance of high-quality experimental data for modelling purposes was perhaps best exemplified in a recent paper by Llinas *et al*., in which researchers where challenged to develop a model based on 100 reliable solubility measurements and to use it to predict the solubility of 32 additional molecules provided.[33] As the authors remarked, the findings of the challenge provided an overall perspective as to the current ability to estimate aqueous solubility.[34]

Based on the importance of this issue, the aim of this study was to investigate whether the implementation of solubility data obtained under standardized experimental conditions can make a significant contribution to the process of establishing new or optimizing existing QSPR models for the prediction of solubility.

EXPERIMENTAL

Data sets

The set of 322 structurally diverse "drug-like" molecules proposed by Rytting *et al*.[32] served as the basis for the first set used in this study (data set 1). The solubility data originated from various literature sources and were collected following several criteria: (*i*) the given compound is a drug or drug-like molecule, solid at room temperature; (*ii*) the reported value is that of the intrinsic solubility at around 25 °C; (*iii*) for solubility measurement, the equilibrium must have been achieved over time, excess solid must be present at the end of testing and acceptable analytical methods must have been used for quantification. Due to geometrical optimisation issues, three molecules were excluded from the Rytting set; hence, data set 1 consisted of 319 compounds. For the investigation of the implementation of experimental solubility data obtained under the uniform experimental conditions, the Sirius data set,[35]

consisting of 84 diverse drug molecules for which the solubility was determined using the Sirius CheqSol technique,[36] was used. This potentiometric method for the measurement of intrinsic solubility is very accurate and allows for rapid equilibration of the experiment and collection of the precipitate during the experiment for characterization.[37]

Data set 2 was created by substituting 28 solubility values from data set 1 with those from the Sirius data set. Data set 3 was created by adding the remaining 56 molecules from the Sirius data set to data set 2 (Table I). The Table with structures of all components included in the Sirius data set can be obtained on request.[35]

Table I. Data sets composition

Data set	Solubility values of compounds
1	319 from ref. 32
2	291 from ref. 32 and 28 from ref. 35
3	291 from ref. 32 and 84 from ref. 35

Structure optimisation and descriptor calculation

All structures were constructed using Spartan software.[38] Geometry optimisation was performed by the AM1 semi-empirical method implemented in the Spartan software. Calculation of descriptors was performed using the Codessa programme (comprehensive descriptors for structural and statistical analysis).[39] A total of 728 descriptors were calculated and divided into five groups: constitutional, topological, geometrical, electrostatic and quantum-chemical.

Correlation analysis

The heuristic method implemented in the Codessa software was used for the selection of the most significant descriptors for the prediction of solubility. The most significant descriptors were selected in each group of descriptors using the heuristic method. Heuristic 5-parameter analysis was also used for selection of the five most significant descriptors among all descriptors calculated, which were then used for establishing a QSPR (quantitative structure––property relationship) equation for each of the sets using the multiple linear regression (MLR) method. Prior to this, descriptor intercorrelation analysis was performed, so that no two descriptors appearing in the final equations have an intercorrelation coefficient larger than 0.5.

RESULTS AND DISCUSSION

The potential limit of the accuracy of experimental data on the predictability of solubility models should be addressed before turning to the purely computational methods, therefore the Sirius data set was used to investigate the extent to which the implementation of solubility data measured under the same standardized method may influence the quality of the potential model training. The creation of data sets 1, 2 and 3 are described in the Experimental section.

The heuristic method was applied for the selection of the most significant descriptors for the prediction of solubility to all three data sets. The most significant descriptors for all three sets from each of the groups of descriptors obtained by the heuristic method are presented in Table II. The most highly correlated descriptor in all data sets was the partition coefficient (log *P*). Most of the selected electrostatic and quantum-chemical descriptors are derived from mole-

cular surface areas (H-donor/acceptor surface areas) and relative electrostatic charges. This would indicate that the descriptors selected using the heuristic method proved to be related to the solvation mechanism of the molecules.

Table II. The most significant descriptors among the five groups of descriptors, selected using the heuristic method

Descriptor[a]	Data set 1		Data set 2		Data set 3	
	R^2	F	R^2	F	R^2	F
Constitutional descriptors						
log P	0.5013	318.67	0.5337	362.77	0.5479	451.95
N_{BR}	0.3848	198.28	0.4116	221.73	0.4390	291.88
R_C	0.3587	177.33	0.3562	175.41	0.3289	182.81
R_{BR}	0.3133	144.65	0.3184	148.09	0.2978	158.16
N_C	0.3013	136.71	0.3294	155.70	0.3639	213.34
N_R	0.2651	114.34	0.2675	115.74	0.2765	142.53
N_{AB}	0.1885	73.63	0.2062	82.36	0.2686	136.97
R_N	0.1526	57.08	0.1570	59.05	–	–
GI_{all}	0.1351	49.53	0.1479	55.04	0.2051	96.23
MW	0.1283	46.65	0.1409	51.97	0.1962	91.05
Topological descriptors						
^{0}ABIC	0.2833	125.31	0.2943	132.19	0.2526	126.06
^{0}ACIC	0.2478	104.41	0.2661	114.94	0.2640	133.81
^{1}ACIC	0.2196	89.22	0.2339	96.76	–	–
^{1}ABIC	0.2072	82.83	0.2147	86.68	–	–
^{2}ACIC	0.1949	76.73	0.2089	83.72	–	–
^{3}Randic	0.1926	75.62	0.2038	81.12	0.2446	120.76
^{2}CIC	0.1895	74.14	0.2224	90.68	0.2254	108.54
^{1}K&H	0.1884	73.59	0.2061	82.31	0.2576	129.40
^{1}Randic	0.1829	70.96	0.2023	80.39	0.2580	129.72
^{2}Randic	–	–	0.1938	76.19	0.2426	119.49
^{0}K&H	–	–	–	–	0.2327	113.13
^{2}K&H	–	–	–	–	0.2199	105.16
Geometrical descriptors						
MSA	0.2312	95.31	0.2565	109.36	0.3125	169.57
XY Shadow	0.2259	92.50	0.2495	105.38	0.2978	158.17
I_C	0.2176	88.19	0.2289	94.11	0.2550	127.69
I_B	0.1778	68.55	0.1868	72.80	0.2142	101.66
MV	0.1613	60.99	0.1829	70.95	0.2378	116.35
I_A	0.1407	51.92	0.1500	55.93	0.1719	81.36
ZX Shadow		32.46	0.1104	39.34	0.1685	75.56
Electrostatic descriptors						
^{e}FPSA-3	0.2603	111.56	0.2676	115.80	0.2759	142.10
^{ed}HDSA-1/TMSA	0.2590	110.81	0.2678	115.96	0.2649	134.42
^{ed}HDSA-2/TMSA	0.2555	108.82	0.2621	112.62	0.2554	127.93
^{e}HASA-2/TMSA	0.2282	93.74	0.2401	100.17	0.2454	121.31
^{e}HACA-2/TMSA	0.2189	88.82	0.2279	93.56	0.2333	113.48
^{ed}HDSA-2/SQRT	0.2001	79.32	0.2037	81.09	–	–

TABLE II. Continued

Descriptor[a]	Data set 1		Data set 2		Data set 3	
	R^2	F	R^2	F	R^2	F
Electrostatic descriptors						
^{ed}HDCA-1/TMSA	0.1984	78.48	0.2072	82.86	0.2091	98.60
^{ed}HDCA-2/TMSA	0.1979	79.20	0.2040	81.22	0.2040	95.57
^{e}HASA-2/SQRT	0.1935	76.06	0.2035	81.01	–	–
^{e}HASA-1/TMSA	0.1905	74.59	0.2026	80.54	0.2164	103.01
^{ew}WNSA-1	–	–	–	–	0.2497	124.17
^{e}TMSA	–	–	–	–	0.2437	120.18
Quantum-chemical descriptors						
ALFA-pol	0.2657	114.70	0.2873	127.79	0.3458	197.12
^{qd}HDSA-1/TMSA	0.2579	110.18	0.2673	115.67	0.2642	133.95
^{qd}HDSA-2/TMSA	0.2574	109.85	0.2645	113.99	0.2579	129.64
^{q}HASA-2/TMSA	0.2282	93.74	0.2401	100.17	0.2454	121.31
^{q}RNCG	0.2281	93.69	0.2393	99.72	0.2433	119.23
PMI_C	0.2176	88.19	0.2289	94.11	0.2550	127.31
^{qd}HDSA-2/SQRT	0.2005	79.52	0.2047	81.60	–	–
^{q}HASA-2/SQRT	0.1935	76.06	0.2035	81.01	–	–
^{q}HASA-1/TMSA	0.1905	74.59	0.2026	80.54	–	–
PMI_C/#	0.1892	73.99	–	–	–	–
^{q}TMSA	–	–	–	–	0.2437	120.16
$Etot_{2\text{-c ex}}$	–	–	–	–	0.2385	116.80
^{qw}WNSA-1	–	–	–	–	0.2274	109.80

[a]Symbols of descriptors used; **constitutional**: N_{BR} – number of benzene rings, N_C – number of C atoms, R_{BR} – relative number of benzene rings, R_C – relative number of C atoms, R_N – relative number of N atoms, N_R – number of rings, N_{AB} – number of aromatic bonds, GI_{all} – gravitational index (all bonds), MW – molecular weight; **topological**: ^{0}ACIC – average complementary information content (order 0), ^{1}ACIC – average complementary information content (order 1), ^{2}ACIC – average complementary information content (order 2), ^{0}ABIC – average bonding information content (order 0), ^{1}ABIC – average bonding information content (order 1), ^{2}CIC – complementary information content (order 2), ^{1}Randic – Randic index (order 1), ^{2}Randic – Randic index (order 2), ^{3}Randic – Randic index (order 3), ^{0}K&H – Kier & Hall index (order 0), ^{1}K&H – Kier & Hall index (order 1), ^{2}K&H – Kier & Hall index (order 2); **geometrical**: XY Shadow – area of the shadow of the molecule projected on a plane defined by X and Y axes, ZX Shadow - area of the shadow of the molecule projected on a plane defined by Z and X axes, MSA – molecular surface area, MV – molecular volume, I_A – moment of inertia A, I_B – moment of inertia B, I_C – moment of inertia C; **electrostatic** (Zefirov's PC): ^{e}FPSA-3 – FPSA-3 fractional PPSA (PPSA-3/TMSA), ^{ed}HDSA-1/TMSA – HA dependent HDSA-1/TMSA, ^{ed}HDSA-2/TMSA – HA dependent HDSA-2/TMSA, ^{ed}HDSA-2/SQRT – HA dependent HDSA-2/SQRT(TMSA), ^{ed}HDCA-1/TMSA – HA dependent HDCA-1/TMSA, ^{ed}HDCA-2/TMSA – HA dependent HDCA-2/TMSA, ^{e}HASA-1/TMSA – HASA-1/TMSA, ^{e}HASA-2/TMSA – HASA-2/TMSA, ^{e}HASA-2/SQRT – HASA-2/SQRT(TMSA); ^{e}HACA-2/TMSA – HACA-2/TMSA, ^{e}TMSA – TMSA total molecular surface area, ^{ew}WNSA-1 – WNSA-1 weighted PNSA (PNSA1*TMSA/1000); **quantum-chemical**: ALFA-pol – ALFA polarizability (DIP), ^{qd}HDSA-1/TMSA – HA dependent HDSA-1/TMSA [Semi-MO PC], ^{qd}HDSA-2/TMSA – HA dependent HDSA-2/TMSA (semi-MO PC), ^{qd}HDSA-2/SQRT – HA dependent HDSA-2/SQRT(TMSA) (semi-MO PC), ^{q}HASA-1/TMSA – HASA-1/TMSA (semi-MO PC), ^{q}HASA-2/TMSA – HASA-2/TMSA (semi-MO PC), ^{q}HASA-2/SQRT – HASA-2/SQRT(TMSA) (semi-MO PC), PMI_C – principal moment of inertia C, PMI_C/# – principal moment of inertia C/# of atoms, ^{q}RNCG – RNCG relative negative charge (QMNEG/QTMINUS) (semi-MO PC), $Etot_{2\text{-c ex}}$ – total molecular 2-center exchange energy, ^{q}TMSA – TMSA total molecular surface area (semi-MO PC).

Comparing data sets 1 and 2, which differ only in the solubility values of 28 compounds, the order of the most significant descriptors selected by the heuristic method in the one-parameter correlation among each of the five groups of descriptors remained relatively unchanged. However, there was an increase in the correlation coefficient in all the descriptors involved. Comparing data sets 1 and 3, the order of the descriptors was slightly changed. Although the correlations of the individual descriptors decreased in a small number of instances, there was an overall improvement of the correlation in all descriptor groups. This would suggest that if no additional components are included in the set, with just the experimental data being altered, individual descriptors could retain their respective capacities for describing the correlation between solubility and structure. This does not hold true, however, when establishing a multi-parameter correlation, as is later demonstrated.

Inclusion of additional components in data set 3 increased the diversity of the set. This can result in a change of relative importance of different molecular properties governing solubility (within a set), which in turn affects the significance of selected molecular descriptors. In the present case, this was reflected in the reordering of closely related descriptors among each group of descriptors.

The present observations showed that changing the experimental data in multi-parameter correlations tended to have even a greater impact. A five-parameter heuristic correlation analysis and subsequent MLR yielded the following QSPR equations for data sets 1, 2 and 3, respectively:

$$\log S = -0.63647\log P - 0.01454\,{}^{q}HBCA + 24.65157 I_C + \\ + 0.010761\,{}^{2}AIC - 19.7268 R_{NR},\ n = 319,\ R^2 = 0.6616 \quad (1)$$

$$\log S = -0.42352\log P + 0.559774\,Etot_{2\text{-}c\ ex}/\# - 0.00756\,{}^{q}DPSA\text{-}3 + \\ + 9.912633\,{}^{q}RNCG + 0.197134\ \min(\#HA, \#HD),\ n = 319,\ R^2 = 0.6689 \quad (2)$$

$$\log S = -0.60863\log P - 0.04033 N_{AB} - 14.2527 R_{NR} - \\ 0.01145\,{}^{q}DPSA\text{-}3 + 2.020366\,{}^{e}RNCG,\ n = 375,\ R^2 = 0.7045 \quad (3)$$

where: log *P* – partition coefficient;
qHBCA – HBCA H-bonding charged surface area (semi-MO PC);
I_C – moment of inertia C;
${}_2AIC$ – average information content (order 2);
R_{NR} – relative number of rings;
Etot2-c ex/# – total molecular 2-center exchange energy/# of atoms;
qDPSA-3 – DPSA-3 difference in CPSAs (PPSA3-PNSA3) (semi-MO PC);
qRNCG – RNCG relative negative charge (QMNEG/QTMINUS) (semi-MO PC);
min(#HA, #HD)- minimum number of H-acceptors/donors;
N_{AB} – number of aromatic bonds;
R_{NR} – relative number of rings;
qDPSA-3– DPSA-3 difference in CPSAs (PPSA3-PNSA3) (semi-MO PC);

eRNCG - RNCG Relative negative charge (QMNEG/QTMINUS).

The plots of the experimental *vs*. the predicted solubility values using these three equations for the three training sets are shown in Figs. 1–3.

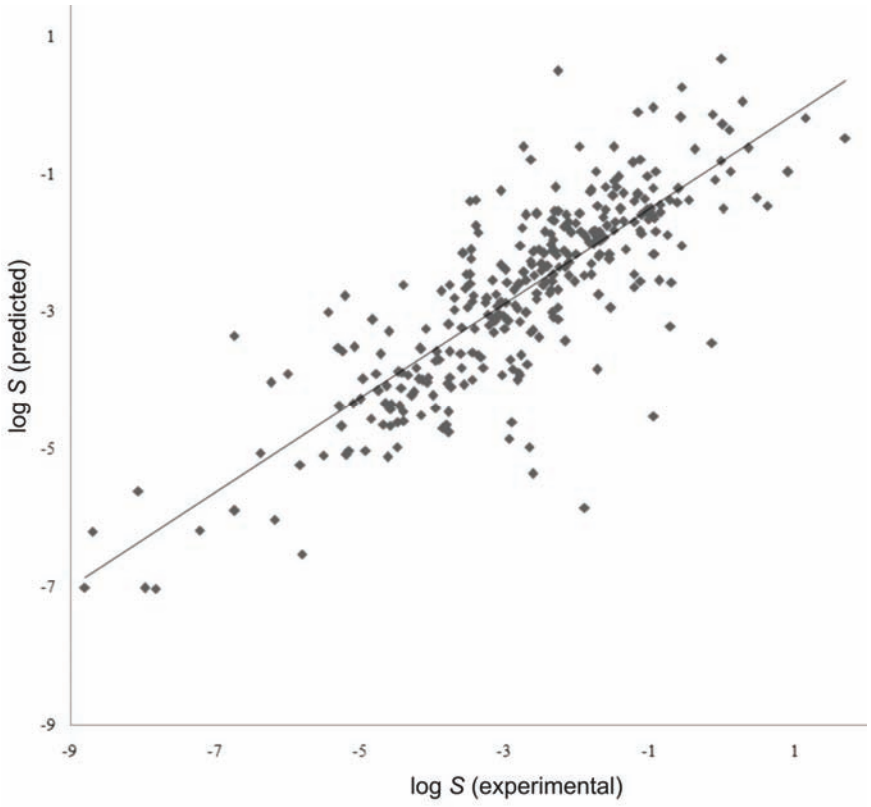

Fig. 1. The correlation between the experimental and predicted log *S* values for data set 1 (n = 319, R^2 = 0.6616).

Equation (1), derived from data set 1 (n = 319), has the smallest correlation coefficient of the three (R^2 = 0.6616, *RMSE* = 0.9641). In comparison, Eq. (2) shows a somewhat improved performance, with a slightly better correlation coefficient and a remarkable reduction in the root mean squared error (n = 319, R^2 = = 0.6689, *RMSE* = 0.8623). Equation (3), derived from the supplemented set (n = = 375), has the best correlation coefficient and a slightly reduced *RMSE*, compared to Eq. (1) (R^2 = 0.7045, *RMSE* = 0.9382). Introduction of uniform experimental data consistently leads to an increase in the correlation coefficient. This can be attributed to both the correction of outliers and the improvement of overall data consistency.

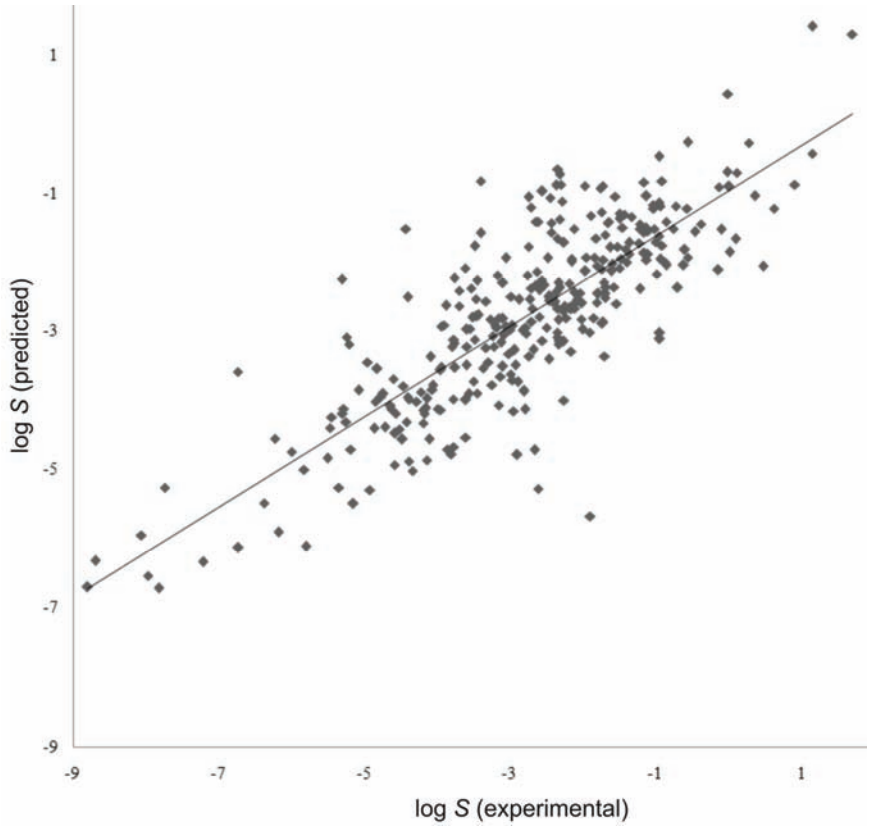

Fig. 2. The correlation between the experimental and predicted log *S* values for data set 2 (n = 319, R^2 = 0.6689).

Moving from single to multi-parameter correlations, the difference in selection of the most significant descriptors using the three data sets became more evident. The equation corresponding to data set 1 is composed of 2 constitutional, 1 geometrical, 1 topological and 1 quantum-chemical descriptor. On the other hand, Eq. (2) was established using 1 constitutional and 4 quantum-chemical descriptors, which together account for several aspects of the solvation process, especially polar interactions and the possibility of H-bond formation. Thus, while the R^2 values for Eqs. (1) and (2) are similar, the interpretability of these equations is significantly affected by the changes in the solubility values of the data set. On average, these values for the 28 substitutions made between data set 1 and 2 (Table III) differ by 0.57log *S*. This is largely consistent with observations made by Katritzky *et al.*[27] Differences in excess of 1.5log *S* are also present in some instances, *e.g.*, phenylbutazone, propranolol, sulfamerazine and notably terfenadine, for which the values differ by 3log *S*. Such large-scale differences can clearly affect the selection of the most significant descriptors. Data Set 3 is struc-

turally more diverse than the previous two, thus the corresponding Eq. (3) also features a different combination of descriptors. It is composed of 3 constitutional and 2 quantum-chemical descriptors. These descriptors encompass molecular properties that relate to hydrophobicity as well as those that facilitate solvation. In summary, both the structural diversity of the training set and the standardized experimental solubility data included in the training significantly influence not just the statistical performance but also the interpretability of a prospective model.

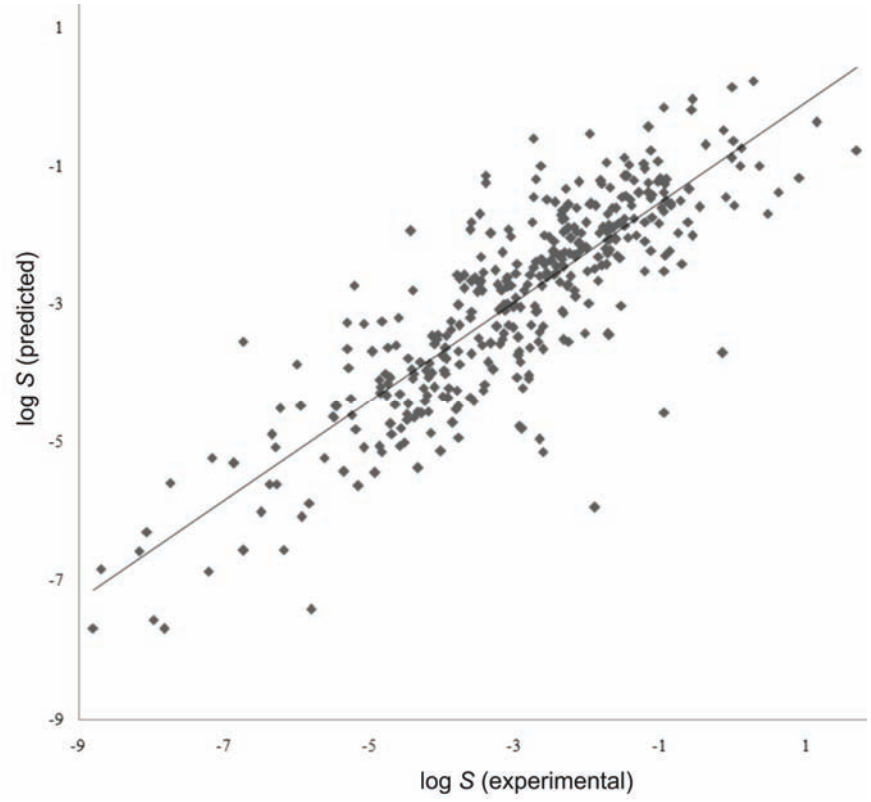

Fig. 3. The correlation between the experimental and predicted log *S* values for data set 3 (n = 375, R^2 = 0.7045).

TABLE III. Rytting solubility data used in data set 1, substituted with Sirius solubility data used in data Set 2 (No. of compounds: 28)

Compound	log *S*	
	Rytting (data set 1)	Sirius (data set 2)
Amitriptyline	–4.4560	–4.3900
Benzocaine	–2.6160	–2.2300
Benzoic acid	–1.5550	–1.6100
Chlorzoxazone	–2.8310	–2.6100

TABLE III. Continued

Compound	log *S*	
	Rytting (data set 1)	Sirius (data set 2)
Diclofenac	–5.0970	–5.4500
Flufenamic acid	–4.6230	–5.3500
Flurbiprofen	–3.7400	–4.1100
Folic acid	–5.4410	–5.3100
Haloperidol	–4.4290	–5.4700
Hydrochlorothiazide	–2.6890	–2.6800
Ibuprofen	–3.4200	–3.6100
Lidocaine	–1.7680	–1.8500
Metoclopramide	–3.1760	–3.5900
Nadolol	–1.0080	–1.5700
Naproxen	–4.1550	–4.1400
Nitrofurantoin	–3.4770	–3.3300
Norfloxacin	–3.0570	–2.7500
Paracetamol	–1.0740	–1.0000
Phenobarbital	–2.3660	–2.2800
Phenylbutazone	–2.6440	–4.3900
Prochlorperazine	–4.3980	–4.8700
Promethazine	–4.2600	–4.1900
Propranolol	–0.7140	–3.5000
Quinine	–2.7900	–2.8100
Sulfamerazine	–1.2180	–3.1000
Sulfathiazole	–2.8050	–2.7000
Sulindac	–5.0000	–4.5200
Terfenadine	–4.6740	–7.7400

CONCLUSIONS

Solubility is a difficult property to predict, and one reason for this is the absence of a high-quality data set of reliable and reproducible solubility measurements. Hopefully, by measuring many compounds under standardized conditions, current predictive models can be improved. In this work, an attempt was made to demonstrate the importance of implementing such data to improve the confidence of the training of the models.

It was successfully shown that the usage of uniform experimental data can significantly improve the correlation in the training set. The results also showed that updating existing data sets with such data leads to changes in the selection of the most significant descriptor, which would require the given model to be additionally optimized. Continuously updated models would be a valuable tool for preliminary solubility screening and could be developed alongside solubility measurement devices as added value software.

Acknowledgments. This work was supported by the Ministry of Science and Technological Development of the Republic of Serbia under Project No 142071 and the Federal Ministry of Education of the Federation of Bosnia and Herzegovina under Project No. 03-39-159-13.

ИЗВОД

ВАЖНОСТ ПРЕЦИЗНОСТИ ЕКСПЕРИМЕНТАЛНИХ ПОДАТАКА ЗА ПРОЦЕНУ РАСТВОРЉИВОСТИ

СЛАВИЦА ЕРИЋ[1], МАРКО КАЛИНИЋ[1], АЛЕКСАНДАР ПОПОВИЋ[1], HALID MAKIĆ[2], ELVISA CIVIĆ[2] и MERJA BEKTAŠEVIĆ[2]

[1]*Фармацеутски факултет, Универзитет у Београду, Војводе Степе 450, 11000 Београд и* [2]*Biotechnical Faculty, University of Bihać, Kulina Bana 2, 77000 Bihać, Bosnia and Herzegovina*

Растворљивост лека у води је значајан фактор који утиче на више аспеката његовог фармакокинетичког профила. Бројне публикације презентују различите методологије за развој поузданих компјутерских модела за предвиђање растворљивости на основу структуре једињења. Квалитет модела за предвиђање растворљивости битно зависи од тачности експерименталних вредности за растворљивост које су коришћене за тренирање модела. У овом раду проучаван је значај примене експерименталних података добијених под стандардизованим, униформним условима за тренирање модела за предвиђање растворљивости. Коришћена су три сета података – испитивани сет 1 који је добијен одабиром експерименталних вредности за растворљивост под одређеним критеријумима из различитих литературних извора (n = 319), затим испитивани сет 2 који је добијен заменом 28 вредности за растворљивост из испитиваног сета 1 вредностима за растворљивост добијеним стандардизованом експерименталном методом у једној лабораторији (n = 319) и испитивани сет 3 који је добијен додатком још 56 компонената у испитивани сет 2, за које су вредности растворљивисти такође одређене под стандардизованим условима у истој лабораторији (n = 375). Затим је примењена хеуристичка метода за селекцију најзначајнијих дескриптора, коришћењем једнопараметарских и вишепараметарских анализа. Постављене су корелације између најзначајнијих дескриптора и растворљивости коришћењем мултилинеарне регресионе анализе за сва три испитивана сета података. Уочена је значајна разлика између једначина које су добијене коришћењем различитих сетова података, што указује на то да је након увођења нових експерименталних података неопходно додатно оптимизовати постојеће моделе. Показано је да коришћење униформних експерименталних података условљава побољшање коефицијената корелације. Ови резултати говоре у прилог све заступљенијем ставу да је за побољшање поузданости предвиђања растворљивости потребно користити сетове података великог броја различитих једињења чија је растворљивост мерена под стандардизованим условима.

(Примљено 9. августа, ревидирано 7. октобра 2009)

REFERENCES

1. S. Stegemann, F. Leveiller, D. Franchi, H. de Jong, H. Lindén, *Eur. J. Pharm. Sci.* **31** (2007) 249
2. H. van de Waterbeemd, E. Gifford, *Nat. Rev. Drug Discovery* **2** (2003) 192
3. J. Huuskonen, *J. Chem. Inf. Comput. Sci.* **40** (2000) 773
4. W. L. Jorgensen, E. M. Duffy, *Bioorg. Med. Chem. Lett.* **10** (2000) 1155
5. I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, A. E. P. Villa, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1488
6. P. Bruneau, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1605
7. C. A. S. Bergström, U. Norinder, K. Luthman, P. Artursson, *Pharm. Res.* **19** (2002) 182
8. O. Engkvist, P. Wrede, *J. Chem. Inf. Comput. Sci.* **42** (2002) 1247
9. J. K. Wegner, A. Zell, *J. Chem. Inf. Comput. Sci.* **43** (2003) 1077
10. A. Cheng, K. M. Merz, *J. Med. Chem.* **46** (2003) 3572

11. J. S. Delaney, *J. Chem. Inf. Model.* **44** (2004) 1000
12. T. J. Hou, K. Xia, W. Zhang, X. J. Xu, *J. Chem. Inf. Model.* **44** (2004) 266
13. C. A. S. Bergström, C. M. Wassvik, U. Norinder, K. Luthman, P. Artursson, *J. Chem. Inf. Model.* **44** (2004) 1477
14. A. Yan, J. Gasteiger, M. Krug, S. Anzali, *J. Comput.-Aided Mol. Des.* **18** (2004) 75
15. C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *J. Chem. Inf. Model.* **45** (2005) 170
16. A. Schwaighofer, T. Schroeter, S. Mika, J. Laub, A. ter Laak, D. Sülzle, U. Ganzer, N. Heinrich, K. R. Müller, *J. Chem. Inf. Model.* **47** (2007) 407
17. J. Wang, G. Krudy, T. Hou, W. Zhang, G. Holland, X. Xu, *J. Chem. Inf. Model.* **47** (2007) 1395
18. D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, *J. Chem. Inf. Model.* **47** (2007) 150
19. J. Huuskonen, D. J. Livingstone, D. T. Manallack, *SAR QSAR Environ. Res.* **19** (2008) 191
20. L. Du-Cuny, J. Huwyler, M. Wiese, M. Kansy, *Eur. J. Med. Chem.* **43** (2008) 501
21. P. R. Duchowicz, A. Talevi, L. E. Bruno-Blanch, E. A. Castro, *Bioorg. Med. Chem.* **16** (2008) 7944
22. J. Wang, T. Hou, X. Xu, *J. Chem. Inf. Model.* **49** (2009) 571
23. B. Faller, P. Ertl, *Adv. Drug Delivery Rev.* **59** (2007) 533
24. J. Taskinen, U. Norinder, *Comprehensive Medicinal Chemistry II*, Vol. 5, Elsevier, Amsterdam, 2007, p. 627
25. S. R. Johnson, W. Zheng, *AAPS J.* **8** (2006) E27
26. J. Taskinen, J. Yliruusi, *Adv. Drug Delivery Rev.* **55** (2003) 1163
27. A. R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson, *J. Chem. Inf. Comput. Sci.* **38** (1998) 720
28. D. Edwards, in *Proceeding of Phys. Chem. Forum 3*, Forest Row, UK, 2007, p. 10
29. K. V. Balakin, N. P. Savchuk, I. V. Tetko, *Curr. Med. Chem.* **13** (2006) 223
30. W. L. Jorgensen, E. M. Duffy, *Adv. Drug Delivery Rev.* **54** (2002) 355
31. C. A. S. Bergström, *Basic Clin. Pharmacol. Toxicol.* **96** (2005) 156
32. E. Rytting, K. A. Lentz, X. Q. Chen, F. Qian, S. Venkatesh, *AAPS J.* **7** (2005) E78
33. A. Llinàs, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model* **48** (2008) 1289
34. A. J. Hopfinger, E. X. Esposito, A. Llinàs, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.* **49** (2009) 1
35. K. Box, J. Mole, T. Gravestock, J. Comer, R. Allen, in *Proceedings of AAPS Annual Meeting*, San Diego, CA, 2007, T3104
36. M. Stuart, K. Box, *Anal. Chem.* **77** (2005) 983
37. G. Völgyi, K. Box, E. Baka, M. Stuart, K. Takács-Novák, J. Comer, *J. Pharm. Sci.* **95** (2006) 1298
38. *Spartan '02 for Linux, Wavefunction, Inc.* Irvine, CA, 2002
39. A. R. Katritzky, W. S. Lobanov, M. Karelson, *Codessa Reference Manual (version 2.0)*, Gainesville, FL, 1994, p. 2.