

Original article

Computational aqueous solubility prediction for drug-like compounds in congeneric series

Lei Du-Cuny^a, Jörg Huwyler^b, Michael Wiese^c, Manfred Kansy^{a,*}^a F. Hoffmann-La Roche Ltd., Pharmaceuticals Division, Grenzacherstrasse, CH-4070 Basel, Switzerland^b University of Applied Sciences Northwestern Switzerland, Institute for Pharma Technology, Gründenstrasse 40, CH-4132 Muttenz, Switzerland^c Rheinische Friedrich-Wilhelms-Universität, Pharmazeutische Chemie, An der Immenburg 4, D-53121 Bonn, Germany

Received 29 January 2007; received in revised form 10 April 2007; accepted 12 April 2007

Available online 6 May 2007

Abstract

It was the aim of the present work to develop a quantitative structure–property relationship (QSPR) model for predicting the aqueous solubility of drug-like compounds in congeneric series. Lipophilicity combined with structural fragment information, fragmental based correction factors and congeneric series indices were used as descriptors for a principal component analysis (PCA) followed by multivariate partial least squares regression statistics (PLS). The derived PLS regression model for the prediction of solubility parameters was based on an in-house data set of 2473 drug-like compounds. The generated PLS model had a coefficient of determination (R^2) = 0.844 and a root-mean-square (rms) error of 0.51 log units. It predicted the solubility of the test data set with a high degree of accuracy (R^2 = 0.81). In addition, the PLS model was successful in predicting the solubility of new congeneric test sets when solubility values of corresponding scaffolds were accessible.

© 2007 Elsevier Masson SAS. All rights reserved.

Keywords: Quantitative structure–property relationship (QSPR); Aqueous solubility; Drug design; Principal component analysis (PCA); PLS regression

1. Introduction

Aqueous solubility of a chemical is defined as the amount of solute dissolved in a saturated aqueous solution under equilibrium conditions. Solubility is not only a property of interest to many areas of academic research, but also a key parameter when it comes to drug design and formulation development in the pharmaceutical industry [1]. Solubility may have a major impact on intestinal absorption of drugs. For example, improved gastrointestinal solubility of the lipophilic and poorly water-soluble immunosuppressive agent cyclosporine A in presence of surface-active ingredients leads to a significant increase in oral bioavailability at high doses [2].

With the advent of high-throughput techniques, it is possible to measure a large amount of solubility data accurately [3]. Nevertheless, data generation only is not sufficient. Computational

solubility prediction tools may help to understand the relationship between chemical structure and solubility and to guide efforts in medicinal chemistry. Computational models for the prediction of aqueous solubility from electrotopology, molecular surface areas, lipophilicity, and hydrophilic measures have been devised, and several of these show impressive statistics [4–9]. However, the most tools commercially available or published by academia, are usually calibrated with organic compounds from the AQUASOL database [10]. In comparison with the majority of compounds stored in AQUASOL, drug-like compounds usually have higher molecular weight, including a larger fraction of aromatic atoms and are characterized by a lower average solubility. Therefore, commercially available prediction tools for solubility based on calibration sets from AQUASOL are more suited to deal with small organic molecules but often might fail to predict the solubility of drug-like compounds with sufficient accuracy.

In this work, a general solubility model of higher accuracy for drug-like compounds in congeneric series is described. Lipophilicity was thereby used as descriptor in combination

* Corresponding author. Tel.: +41 616885874; fax: +41 616887408.

E-mail address: manfred.kansy@roche.com (M. Kansy).

with information from structural fragments. The derived model tried to overcome difficulties of commercially available prediction tools for solubility by focusing on structurally related series of drug-like compounds. Using experimental solubility data for new chemical scaffolds, the present model could be adopted to related series of compounds not covered by the original data sets. In addition, rules were derived from the present prediction model, which may be used by chemists or interested scientists as a rough guideline on the contribution of structural fragments on solubility.

2. Materials and methods

2.1. Solubility data set and solubility measurements

Thermodynamic equilibrium solubilities of 2473 compounds in 81 congeneric series were obtained from an in-house data set. Solubility data for this Roche proprietary database were generated by either a traditional saturation shake-flask method in 50 mM phosphate buffer at pH 6.5 and room temperature or a potentiometric pH titration method [11]. Among these 2473 compounds, 983 were uncharged, 166 had measured pK_a values and for 1324 compounds, the pK_a values were assigned according to structural similarity comparisons. Additionally, the 2473 compounds were classified using clustering algorithms [12] and singletons were eliminated. Partition coefficients were calculated using ClogP v4.71 (Daylight Chemical Information Systems, Irvine, CA). It is important to note that the data set used for the development of the present prediction tools was selected based on the following criteria to ensure a high quality level: First, pH value and temperature of the saturated solution used for equilibrium solubility measurements was registered. Second, the solubility of compounds available as salts were considered only, when no pH shift was observed for saturated solutions. Third, aqueous solubility used for prediction was expressed as $\log 1/S_0$, where S_0 is the molarity of the unionized molecular species. Thus, for all molecules ionization was considered and calculated according to the Henderson–Hasselbalch equation using information on pK_a values.

2.2. Data analysis

Multivariate data analysis was performed using the program SIMCA [13]. Variable preprocessing was performed. Thus, all the descriptors were mean-centered and scaled to unit variance (UV). Descriptors with a higher skewness than 1.5 were log-transformed. Principal component analysis (PCA) was performed to get an overview on the data sets. The information contained in original variables was summarized by calculation of new latent variables. The compounds, which could not be well explained with the latent variables were classified as outliers in PCA. Outliers conforming to the overall correlation structure, but occupying extreme characteristics were strong and were identified using the 95% tolerance interval signified as ellipse in the PCA loading plot [13]. Outliers found by inspecting residuals for each observation were moderate and were identified by the “distance to the model in X space”

(DModX) plot [13]. Furthermore, PCA loading plots were used to detect reason for the outliers in PCA and were sometimes helpful in explanation of results. Multivariate partial least square regression statistics (PLS) was performed to predict solubility. The goodness of fit of a PLS model was given by a regression coefficient R^2 . The goodness of prediction was evaluated by a cross-validated R^2 , designated as Q^2 . The Q^2 value was the main criterion for assessing the quality of a model. In general, a model with a Q^2 value of 0.3 or higher is statistically meaningful, while a Q^2 value greater than 0.5 is regarded as a good model and 0.8 or above is excellent. Variable influence on projection (VIP) estimated the influence of every original variable on the matrix Y . Variables with larger VIPs were the most relevant for explaining Y , and those with VIPs less than 0.8 were of lesser importance.

Once a model was chosen, it was validated by a permutation test using scrambled Y values to ensure that the model was not obtained by chance. The result of the response permutation test was summarized in the validation plot in SIMCA. The R^2 - and Q^2 -intercepts in the validated plot are interpretable as measures of the significance of the model's predictive power. A model with R^2Y -intercept below 0.3–0.4 and the Q^2 -intercept below 0.05 can be assumed not to be overfitted.

In case of large data sets ($N > 100$, as in the present study), the data set was divided into a training data set and a test data set. The PLS model was built by only using the training data set. The obtained model was tested with an independent test data set. The predictive power of the model was further tested using additional data from the literature.

3. Results and discussion

3.1. Experimental solubility data

Solubility data of 2473 compounds in 81 congeneric series were used to develop an improved model for the prediction of intrinsic solubility of drug-like organic molecules. Standardized experimental solubility measurements were carried out using either a traditional saturation shake-flask method at pH 6.5 or a potentiometric pH titration method, which offers an increased throughput due to partial automation. Both methods provide information on the thermodynamic equilibrium solubility of a given molecule. Previous validation studies demonstrated the reliability and robustness of the two methods as well as the comparability of the obtained experimental data [14]. The solubility data used for the present study represent a subset of data from a Roche proprietary database and were selected based on stringent quality criteria including the availability of information on pH value, temperature of the saturated solution used for the saturation shake-flask measurements and the availability of information on pK_a values.

3.2. Data analysis

The solubility data set was analyzed by principal component analysis (PCA) followed by multivariate partial least squares regression statistics (PLS). The multiple linear regression is taken

here as a rational method for solubility prediction, because in comparison to neural networks, multiple linear regression provides further insights in the nature of the major properties or features governing solubility. The presentation of the multiple linear regression is graphically oriented and many diagnostics and parameters are available for model interpretation and validation. From our data analysis, no three-dimensional descriptors were identified as important for the solubility prediction. However, lipophilicity, structural fragment information, fragmental based correction factors and congeneric series indices emerged as the driving molecular parameters and were therefore used as descriptors for the derived PLS model. In the final model, 170 structural fragments plus four fragmental based correction factors were used (see Table 1).

With respect to fragment based descriptors, the structural fragmentation scheme of ClogP was found to be the easiest way to obtain molecular fragments. In ClogP, the molecules are dissected according to the rule of “Isolating Carbon”. An “Isolating Carbon” atom (IC) is a carbon which is not double- or triple-bonded to a hetero atom. Isolating carbons can, however, be multiply bonded to one another, such as those in $\text{CH}_3\text{CH}=\text{CH}_2$. An IC is an atomic fragment that, for calculation purposes at least, is always hydrophobic. Any hydrogen atom attached to an isolating carbon (ICH) is also a hydrophobic atomic fragment. All atoms or groups of covalently bonded atoms that remain after removal of ICs and ICHs are polar fragments. Thus a polar fragment contains no ICs but each has one or more bonds to ICs. These bonds are used to label the environments of a polar fragment, and are usually designated as A for aliphatic, Z for benzyl, V for vinyl, Y for styryl and a for aromatic. Smarts strings (Table 2) are defined for these five binding environments and they provide a complete description of each fragment listed in the Table 1.

ClogP fragments are defined in a way that each heavy atom in the molecule belongs to one fragment only. Thus, the presence of fragments can be easily checked since the total number of heavy atoms in a molecule has to be equal to the sum of the number of the heavy atoms in the fragments. Provided that all fragments of a given molecule are defined in a database, simple test methods can be devised to identify missing fragments. Therefore ClogP fragments and not KowWin LogP [5] fragments were used as descriptors for the development of the present solubility prediction tool. In addition to the 170 structural fragments (Table 1), four fragments were used as correction factors to improve the predictive power of the new solubility tool. It should be noted that while 170 structural fragments were enough for the solubility prediction of 2473 compounds in 81 congeneric series, this number of fragments is still far from being enough to cover the whole ‘medicinal chemical space’. For the prediction of extended data sets and to further improve the accuracy of the present model, it might therefore be necessary to add further structural fragments.

3.3. The PLS regression model

In the present study, the data set was randomized and divided into a training data set and a test data set. The PLS

analysis yielded a model with three principal components. The PLS analysis were performed using the program SIMCA. It is important to note that this approach to PLS analysis is different from other methods in that the percentage of cumulative variation and the number of latent vectors is not included in the output. The way in which the SIMCA PLS model fits into a framework among other latent variable models has been discussed by Burnham et al. [15,16]. And, the way in which conventional PLS relates to canonical PLS and continuum power regression, was reported by De Jong et al. [17].

A PLS model (Eq. (1)) with $R^2 = 0.844$, $Q^2 = 0.79$ and $\text{rmse} = 0.510$ was obtained for 1515 compounds in the training data set. The quality of the model was tested with 958 compounds in the test data set and $R^2 = 0.813$ was obtained. Using this model, the solubility of most compounds was predicted within an error of one log unit (Fig. 1). The standard error of the predicted solubilities is 0.42 log units, as evident from the plot of the solubility residues of all the compounds (Fig. 2).

For comparison, the analysis of the same data set of drug-like compounds using the commercial program WsKow [18] shows no correlation ($R^2 = 0.1014$). However, it is important to note that the WsKow model is calibrated with organic compounds from the AQUASOL database [10] and is therefore designed to deal with small organic molecules. It is therefore not surprising, that the latter program fails to predict the solubility of our drug-like compounds, which includes a larger fraction of aromatic atoms and have often a higher molecular weight. This example illustrates the fact that the predictive power of computational models depends very much on the selection of representative training data sets.

$$\log \frac{1}{S_0} = 0.131493 \times \text{C log P} + \sum_{i=1}^{n=174} b_i \times \text{frag}_i + \sum_{i=1}^{n=81} c_i \times f_{\text{series},i} + 3.7551 \quad (1)$$

where S_0 is the intrinsic solubility of a given compound with unit in mol/L, ClogP is the calculated partition coefficient, b_i is a cohesive energy contribution descriptor of a fragment (frag_i) to $\log 1/S_0$, c_i is a cohesive energetical contribution descriptor of a molecular scaffold (f_{series}) to $\log 1/S_0$. Note that frag_i and f_{series} adopt the values 0 or 1, depending on the absence or presence of a given fragment or scaffold in the target molecule.

Eq. (1) uses ClogP to describe the liquid–liquid interaction in the solvation process and the fitted coefficients b_i to study the cohesive energy caused by each fragment in the solid state. Thus, the solubility value of a fragment is calculated as the sum of the fragmental contribution to lipophilicity and to crystal packing. The solubility values of fragments as well as the b_i coefficients are listed in Table 1 and can be used for the solubility prediction of drug-like test compound. The congeneric series indices c_i of a molecular scaffold (f_{series}) describe the contribution of specific molecular scaffolds within congeneric series. They can be determined experimentally by testing the solubility of selected representatives of new series as discussed below.

Table 1
List of fragment associated parameters, which define the present PLS regression model


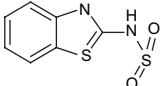
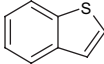
Nr.	Name	CE	Smart	b_i	Coeff. (ClogP)	FC to log $1/S_0$
Frag1	Tertiary amine	AZZ	AN(Z)Z	−0.505804	−2.2	−0.78789
Frag2	Tertiary amine	AAa	A[N&X3](A)a	0.246447	−1.12	0.102841
Frag3	Tertiary amine	AAA	AN(A)A	0.112752	−2.37	−0.19113
Frag4	Tertiary amine	AAZ	AN(A)Z	0.03326	−1.98	−0.22062
Frag5	Secondary amine	AA	A[NH]A	0.283721	−1.77	0.056772
Frag6	Secondary amine	Aa	A[NH]a	−0.0704758	−1.03	−0.20254
Frag7	Secondary amine	aa	a[NH]a	0.0130401	−0.09	0.0015
Frag8	Secondary amine	AZ	A[NH]Z	0.320647	−1.69	0.103955
Frag9	Secondary amine	Za	[NH](Z)a	−0.0950652	−1.15	−0.24252
Frag10	Secondary amine	ZZ	[NH](Z)Z	−0.0924969	−2.1	−0.36176
Frag11	Primary amine	A	A[NH2]	0.88698	−1.54	0.689521
Frag12	Primary amine	Z	[NH2]Z	0.800753	−1.35	0.627656
Frag13	Primary amine	a	a[NH2]	−0.293951	−1	−0.42217
Frag14	Acid hydrazide-NH	aa	a[NH][NH]C(a)=O	−0.21722	−2.3	−0.51213
Frag15	Aromatic amide	aa	a[nH]c(a)=O	−0.148826	−2	−0.40527
Frag16	Acid imide	Aza	AN(C(Z)=O)C(a)=O	−1.05145	−1.72	−1.27199
Frag17	Amide	AAA	AN(A)C(A)=O	−0.568526	−3.19	−0.97755
Frag18	Amide	AAa	AN(A)C(a)=O	−0.328139	−2.82	−0.68972
Frag19	Amide	AaA	AN(a)C(A)=O	−0.495233	−1.4	−0.67474
Frag20	Amide	aaa	aN(a)C(a)=O	−0.0949459	−0.33	−0.13726
Frag21	Amide	Aaa	AN(a)C(a)=O	0.0412647	−2.09	−0.22672
Frag22	Amide	AaZ	AN(a)C(Z)=O	−0.459037	−2.12	−0.73086
Frag23	Amide	AAZ	AN(A)C(Z)=O	−0.860587	−2.99	−1.24396
Frag24	Amide	AZA	AN(Z)C(A)=O	−0.105184	−2.99	−0.48856
Frag25	Amide	AZa	AN(Z)C(a)=O	−0.3433	−2.2	−0.62538
Frag26	Amide	AZZ	AN(Z)C(Z)=O	−0.533844	−2.87	−0.90184
Frag27	Formylamine	AA	AN(A)[CH]=O	−0.103085	−2.67	−0.44543
Frag28	NH-amide	AA	A[NH]C(A)=O	−0.260609	−2.71	−0.60809
Frag29	NH-amide	Aa	A[NH]C(a)=O	−0.148322	−1.81	−0.3804
Frag30	NH-amide	aA	a[NH]C(A)=O	−0.0877811	−1.51	−0.28139
Frag31	NH-amide	aa	a[NH]C(a)=O	0.0820121	−1.06	−0.0539
Frag32	NH-amide	AV	A[NH]C(V)=O	−0.416968	−2.26	−0.70675
Frag33	NH-amide	AZ	A[NH]C(Z)=O	−0.414681	−2.51	−0.73651
Frag34	NH-amide	aZ	a[NH]C(Z)=O	−0.124235	−1.54	−0.32169
Frag35	NH-amide	aV	a[NH]C(V)=O	−0.530449	−1.3	−0.69714
Frag36	NH-amide	ZA	[NH](Z)C(A)=O	−0.58768	−2.25	−0.87618
Frag37	NH-amide	Za	[NH](Z)C(a)=O	−0.0663947	−1.41	−0.24718
Frag38	Formamine-NH	A	a[NH][CH]=O	0.0694213	−0.75	−0.02674
Frag39	Urea (tetrasub)	AAAA	AN(A)C(=O)N(A)A	−0.598173	−3.01	−0.98412
Frag40	1,1,3-Urea	Aaa	A[NH]C(=O)N(a)a	−0.085475	−2.16	−0.36243
Frag41	1,1,3-Urea	aAA	a[NH]C(=O)N(A)A	−0.282532	−2.77	−0.6377
Frag42	1,1,3-Urea	aAZ	a[NH]C(=O)N(A)Z	−0.492871	−2.09	−0.76085
Frag43	<i>N,N'</i> -Urea	Aa	A[NH]C(=O)[NH]a	−0.137129	−1.57	−0.33843
Frag44	<i>N,N'</i> -Urea	Za	[NH](Z)C(=O)[NH]a	0.298589	−1.37	0.122928
Frag45	NH-urea	A	a[NH]C([NH2])=O	−0.374615	−1.07	−0.51181
Frag46	NH-carbamate	aA	a[NH]C(=O)OA	0.288844	−1.06	0.152931
Frag47	NH-carbamate	aZ	a[NH]C(=O)OZ	1.01701	−1.06	0.881097
Frag48	NH ₂ -amide	A	aC([NH2])=O	0.0208799	−1.26	−0.14068
Frag49	NH ₂ -amide	A	AC([NH2])=O	−0.411489	−1.99	−0.66665
Frag50	NH ₂ -amide	Z	C(Z)([NH2])=O	−0.131347	−1.99	−0.3865
Frag51	Thioamide-NH	aA	a[NH]C(A)=S	−0.420511	−0.96	−0.5436
Frag52	Thioamide-NH ₂	A	AC([NH2])=S	0.0493423	−1.13	−0.09555
Frag53	Ester	AA	AOC(A)=O	−0.275125	−1.45	−0.46104
Frag54	Ester	Aa	AOC(a)=O	−0.203746	−0.56	−0.27555
Frag55	Ester	AY	AOC(Y)=O	−0.086602	−0.96	−0.20969
Frag56	Ester	AZ	AOC(Z)=O	−0.124924	−1.38	−0.30187
Frag57	Ester	Za	O(Z)C(a)=O	−1.40291	−0.3	−1.44138
Frag58	Carboxy (ZW-)	A	AC([OH])=O	0.509152	−1.07	0.371957
Frag59	Carboxy (ZW-)	a	aC([OH])=O	1.08236	−0.03	1.078513
Frag60	Carboxy	Z	C(Z)([OH])=O	−0.295626	−1.03	−0.42769
Frag61	Carbonyl	Aa	AC(a)=O	0.104273	−1.09	−0.03549
Frag62	Carbonyl	aa	aC(a)=O	−0.0367111	−0.53	−0.10467
Frag63	Carbonyl	AA	AC(A)=O	−0.61826	−1.84	−0.85418
Frag64	Aldehyde	a	a[CH]=O	0.00267908	−0.42	−0.05117

Table 1 (continued)

Nr.	Name	CE	Smart	b_i	Coeff. (ClogP)	FC to log 1/S ₀
Frag65	Ether	AA	AOA	−0.12718	−1.82	−0.36054
Frag66	Ether	Aa	AOa	−0.00697339	−0.61	−0.08519
Frag67	Ether	aa	aOa	0.167299	0.53	0.235256
Frag68	Ether	AY	AOY	−0.0610615	−1.3	−0.22775
Frag69	Ether	AZ	AOZ	0.242133	−1.28	0.078011
Frag70	Ether	aZ	aOZ	0.153457	−0.41	0.100887
Frag71	Alcohol or hydroxy	A	A[OH]	−0.372527	−1.64	−0.58281
Frag72	Alcohol or hydroxy	a	a[OH]	−0.331856	−0.44	−0.38827
Frag73	Alcohol or hydroxy	Z	[OH]Z	0.00490034	−1.34	−0.16691
Frag74	Sulfide	AA	A[S&X2]A	0.415272	−0.7	0.325518
Frag75	Sulfide	Aa	A[S&X2]a	−0.093221	0.03	−0.08937
Frag76	Sulfide	aa	a[S&X2]a	0.14817	0.77	0.246899
Frag77	Sulfide	AZ	A[S&X2]Z	−0.613268	−0.35	−0.65815
Frag78	Sulfide	VV	V[S&X2]V	−0.416968	0.18	−0.39389
Frag79	Sulfide	Za	[S&X2](Z)a	0.591607	0.03	0.595454
Frag80	Azo	A	AN=[N+]=[N−]	−0.479841	0.62	−0.40034
Frag81	Nitro	a	a[N+](=O)[O−]	0.00594268	−0.03	0.002096
Frag82	Nitrile	a	aC#N	0.255116	−0.34	0.211521
Frag83	Nitrile	A	AC#N	−0.0841117	−1.27	−0.24695
Frag84	Nitrile	Z	C(Z)#N	0.742307	−0.88	0.629473
Frag85	Fluoride	A	AF	−0.0522532	−0.38	−0.10098
Frag86	Fluoride	a	aF	0.05092	0.37	0.098361
Frag87	Fluoride	Z	FZ	0.019207	−0.18	−0.00387
Frag88	Chloride	a	aCl	0.0639783	0.94	0.184505
Frag89	Chloride	Z	ClZ	0.304997	0.26	0.338334
Frag90	Bromide	a	aBr	0.364067	1.09	0.503827
Frag91	Iodide	a	aI	0.0205602	1.35	0.193657
Frag92	Sulfoxide	AA	A[S&X3](A)=O	−0.5968	−3.01	−0.98274
Frag93	Sulfonyl	AA	AS(A)(=O)=O	0.0359061	−3.01	−0.35004
Frag94	Sulfonyl	Aa	AS(a)(=O)=O	0.633814	−2.17	0.355577
Frag95	Sulfonamide	AAa	AN(A)S(a)(=O)=O	−0.179356	−2.09	−0.44734
Frag96	Sulfonamide	Aaa	AN(a)S(a)(=O)=O	0.122288	−1.67	−0.09184
Frag97	Sulfonamide	AAA	AN(A)S(A)(=O)=O	−0.388641	−1.37	−0.5643
Frag98	Sulfonamide	AAZ	AN(A)S(Z)(=O)=O	−0.295066	−2.76	−0.64895
Frag99	Sulfonamide	AZa	AN(Z)S(a)(=O)=O	−0.362315	−1.89	−0.60465
Frag100	NH-sulfonamide	Aa	A[NH]S(a)(=O)=O	−0.150474	−1.75	−0.37486
Frag101	NH-sulfonamide	aA	a[NH]S(A)(=O)=O	−2.31194	−1.72	−2.53248
Frag102	NH-sulfonamide	aa	a[NH]S(a)(=O)=O	−0.213922	−1.13	−0.35881
Frag103	NH-sulfonamide	aZ	a[NH]S(Z)(=O)=O	−0.228931	−1.6	−0.43408
Frag104	NH-sulfonamide	AA	A[NH]S(A)(=O)=O	−1.21692	−2.5	−1.53747
Frag105	NH-sulfonamide	AZ	A[NH]S(Z)(=O)=O	−0.181414	−2.42	−0.49171
Frag106	NH-sulfonamide	Za	[NH](Z)S(a)(=O)=O	−0.542045	−1.55	−0.74079
Frag107	NH ₂ -sulfonamide	a	aS([NH2])(=O)=O	0.035859	−1.61	−0.17058
Frag108	Sulfamide, tetrasubst.	AAAA	AN(A)S(=O)(=O)N(A)A	−0.287803	−4.05	−0.80709
Frag109	Sulfondiamide, trisubst.	AAA	A[NH]S(=O)(=O)N(A)A	−0.761375	−3.4	−1.19732
Frag110	Sulfondiamide, trisubst.	aAA	a[NH]S(=O)(=O)N(A)A	0.428803	−2.043	0.16685
Frag111	Sulfondiamide, trisubst.	ZAA	[NH](Z)S(=O)(=O)N(A)A	−0.0114224	−1.545	−0.20952
Frag112	Thiadiazole dioxide	AA	A[NH]S(=O)(=O)[NH]A	−0.942546	−1.775	−1.17014
Frag113	N-Carboxysulfonamide	aa	aC(=O)[NH]S(a)(=O)=O	−1.69519	−0.97	−1.81956
Frag114	Sulfonylurea, N(disubst-amino)	AAa	AN(A)[NH]C(=O)[NH]S(a)(=O)=O	0.0325044	−4.34	−0.52397
Frag115	1-Sulfonyl-3-urea	Aa	A[NH]C(=O)[NH]S(a)(=O)=O	−0.519879	−2.26	−0.80966
Frag116	FragA	AA	A[NH]S(=O)(=O)[NH]C(=O)OA	−1.24472	−1.745	−1.46846
Frag117	FragB	AAa	AN(A)C=NS(a)(=O)=O	−0.318364	−1.745	−0.54211
Frag118	FragC	Aaaa	An(a)c(=O)n(a)S(a)(=O)(=O)	−1.04836	−2.728	−1.39814
Frag119	FragD	Zaaa	n(Z)(a)c(=O)n(a)S(a)(=O)(=O)	0.774524	−2.728	0.42474
Frag120	FragE	aaa	[nH](a)c(=O)n(a)S(a)(=O)(=O)	−0.154406	−2.424	−0.46521
Frag121	Thiophosphorothioate	AAA	AOP(=S)(OA)SA	−0.051248	0.1	−0.03843
Frag122	FragF	A	AOP([OH])([OH])=O	1.15179	−2.174	0.87304
Frag123	FragG	AAa	A[N+](A)(a)[O−]	−2.09328	−1.349	−2.26625
Frag124	Cyanoguanidyl	aAA	a[NH]C(=NC#N)N(A)A	−1.56706	−1.104	−1.70861
Frag125	Oxanilic ester	aA	a[NH]C(=O)C(=O)OA	0.249812	−1.72	0.029274
Frag126	Amidine	a	aC([NH2])=[NH]	0.46317	−1.27	0.300331
Frag127	FragH	aa	aC([NH2])=NC(a)=O	0.537292	−1.137	0.391506
Frag128	FragI	a	aC([NH2])=NO	−0.0988768	−0.891	−0.21312

(continued on next page)

Table 1 (continued)

Nr.	Name	CE	Smart	b_i	Coeff. (ClogP)	FC to log 1/ S_0
Frag129	Dicarbonylhydrazine (sym)	aa	aC(=O)[NH][NH]C(a)=O	−0.392074	−1.49	−0.58312
Frag130	Acid hydrazide-NH ₂	A	AC(=O)[NH][NH2]	0.212569	−2.5	−0.10798
Frag131	<i>N,N</i> -Carboxamide, alpha-keto	AZA	AN(Z)C(=O)C(A)=O	0.046828	−3.105	−0.3513
Frag132	Formocarbamide	Aa	AN[CH]=O)C(a)=O	0.18389	−1.43	0.000535
Frag133	Acid imide	Aaa	AN(C(a)=O)C(a)=O	0.00251297	−1.05	−0.13212
Frag134	Tertiary imine	Aaa	AN=C(a)a	−0.28748	−1.65	−0.49904
Frag135	<i>N,N</i> -Carbamate	AAA	AOC(=O)N(A)A	−0.00739098	−1.95	−0.25742
Frag136	<i>N</i> -Carboxyguanidyl	Aa	AOC(=O)N=C(a)[NH2]	0.357515	−1.5	0.165185
Frag137	Carbonate	AA	AOC(=O)OA	0.0425553	−1.93	−0.20491
Frag138	Iminoxy	Aa	AON=Ca	0.0416457	−0.6	−0.03529
Frag139	FragJ	aa	a[n+](a)[O−]	−0.740466	−1.745	−0.96421
Frag140	1-Pyrrole	Aaa	An(a)a	−0.0874169	−1.09	−0.22718
Frag141	1-Pyrrole	aaa	an(a)a	−0.167806	−0.56	−0.23961
Frag142	1-Pyrrole	Zaa	n(Z)(a)a	−0.181305	−0.89	−0.29542
Frag143	Ring amide, <i>N</i> -subst.	aaa	an(a)c(a)=O	−0.754877	−2.35	−1.05619
Frag144	Ring amide, <i>N</i> -subst.	aZa	an(Z)c(a)=O	−0.525923	−2.39	−0.83237
Frag145	Arom.1-(3 <i>H</i>)diaz-2,4-dioxo	Zaa	n(Z)(a)c(=O)[nH]c(a)=O	−1.09721	−2.79	−1.45494
Frag146	Disubst. pyrimidin-dione	Zaaa	n(Z)(a)c(=O)n(a)c(a)=O	−0.0664098	−1.91	−0.31131
Frag147	FragK	AAaa	AN(A)n(a)c(a)=O	−0.0620699	−3.297	−0.48481
Frag148	1-Amino-2-pyridone	aa	an([NH2])c(a)=O	−0.842206	−1.6	−1.04736
Frag149	Tetrazolyl	Ya	n1(Y)annn1	0.174292	−1.77	−0.05266
Frag150	2,3,4-Trisubst. urazole	ZZa	n1(Z)n(Z)c(=O)n(a)c1=O	−0.708604	−2.207	−0.99159
Frag151	2-Tetrazolyl	Aa	An1nann1	−0.651707	−1.65	−0.86327
Frag152	2-Tetrazolyl	Za	n1(Z)nann1	−0.0943726	−1.65	−0.30594
Frag153	2-Pyrimidinone	aZa	a[n&X2]c(=O)n(Z)a	0.643812	−3.12	0.243766
Frag154	Triazole	aaa	annn(a)a	0.293345	−1.25	0.13307
Frag155	Isoxazolyl	aa	a[n&X2]oa	−0.237556	−0.95	−0.35937
Frag156	Isotiazole #1	aa	a[n&X2]sa	−0.342363	−0.2	−0.36801
Frag157	1,3,4-Triazinone	a	O=c1[nH]an[nH]1	1.36702	−1.01	1.237518
Frag158	Aromatic diazo (type 2)	aa	a[n&X2][n&X2]a	−0.483229	−2.16	−0.76018
Frag159	Diazole- <i>N</i> -subst.	aaa	a[n&X2]n(a)a	−0.0462078	−1.1	−0.18725
Frag160	Diazole- <i>N</i> -subst.	aYa	a[n&X2]n(Y)a	−0.0872313	−1	−0.21545
Frag161	Diazole- <i>N</i> -subst.	aAa	a[n&X2]n(A)a	−0.289211	−1.69	−0.5059
Frag162	Diazole- <i>N</i> -subst.	aZa	a[n&X2]n(Z)a	−0.363835	−1.69	−0.58053
Frag163	Aromatic NH	aa	a[nH]a	−0.0163266	−0.68	−0.10352
Frag164	Aromatic oxygen	aa	a[o&X2]a	0.108018	−0.11	0.093914
Frag165	Thiophenyl	aa	a[s&X2]a	−0.0664319	0.36	−0.02027
Frag166	Aromatic-nitrogen-type2	aa	a[n&X2]a	0.031963	−1.14	−0.11421
Frag167	Aliphatic carbon		[C;!\$(=,#[!#6])]	0.0163697	0.195	0.041373
Frag168	Aromatic carbon		[c;!\$(=,#[!#6])]	0.0435754	0.13	0.060244
Frag169	NH-amide	ZZ	Z[NH]C(Z)=O			
Frag170	Tertiary Imine	aAa	AN=C(A)a			
CorrFrag1	Aliphatic ring			0.0503823		
CorrFrag2	Trifluoromethyl		C(F)(F)F	0.0649662		
CorrFrag3		A	aS(=O)(=O)[NH]c1sc2ccccc2n1	1.24005		
CorrFrag4			s1ccc2ccccc12	0.528395		

One hundred and seventy fragments and four correction factors were identified as descriptors of the molecular structures of 2473 drug-like compounds in 81 congeneric series. CE: connection environment; FC: fragment contribution.

Table 2

Smarts notations for the five connection environments of the “Isolating Carbon”

Type	Symbol	Smarts
Alkyl	A	[C; !\$(=, #[!#6]); !\$(C(-*)a; !\$(=C)]
Benzyl	Z	[C; !\$(C=); \$(C(-*)a)]
Vinyl	V	[C; !\$(=C); !\$(=Ca); !\$(=C)a]
Styryl	Y	[C; !\$(=Ca); !\$(=C)a]
Aromatic	a	[c; !\$(=, #[!#6])]

Smarts notations were used to reproduce the five connection environments defined in ClogP in order to enable flexible and efficient fragment search.

Structure based solubility rules were derived from the contribution of fragments to solubility. Molecular properties important for the solubility enhancement can be identified by inspection of the structure based solubility rules. A small section from the structure based solubility rules is taken here as example to visualize the influence of small structural changes on intrinsic solubility as shown in Fig. 3. The following conclusions were drawn from this structure based analysis. Molecules containing aliphatic fragments are more soluble than aromatic ones. Dipole moment enhances solubility. Compounds containing polar fragments are more soluble than non-polar ones. Compounds containing strong basic and acid fragments have lower intrinsic solubility than neutral ones. Solubility decreases with increasing molecular weight.

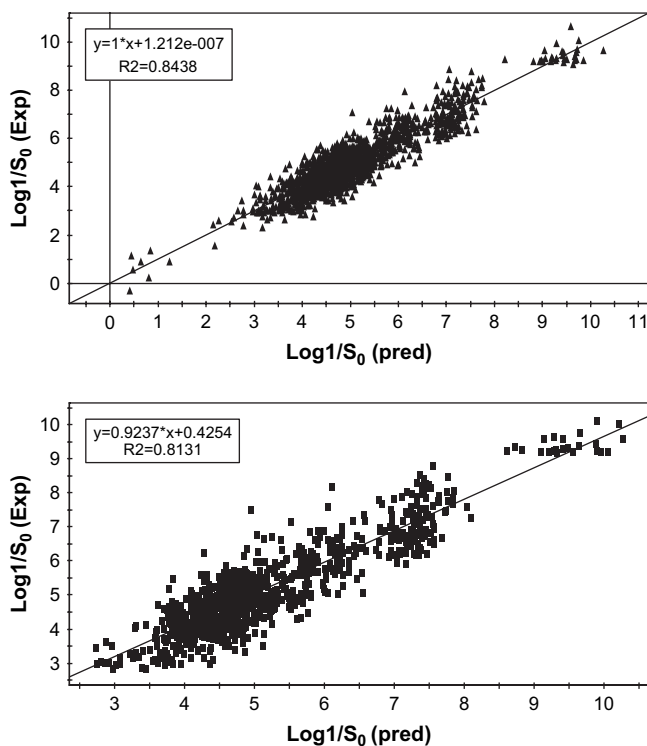


Fig. 1. Correlation between measured solubility (Exp) and predicted solubilities for drug-like compounds using computational solubility models. Upper panel: predictivity of the presented PLS regression solubility model (Pred) for 1515 compounds from a training data set (Exp). Lower panel: predictivity of the model for 958 compounds from a test data set.

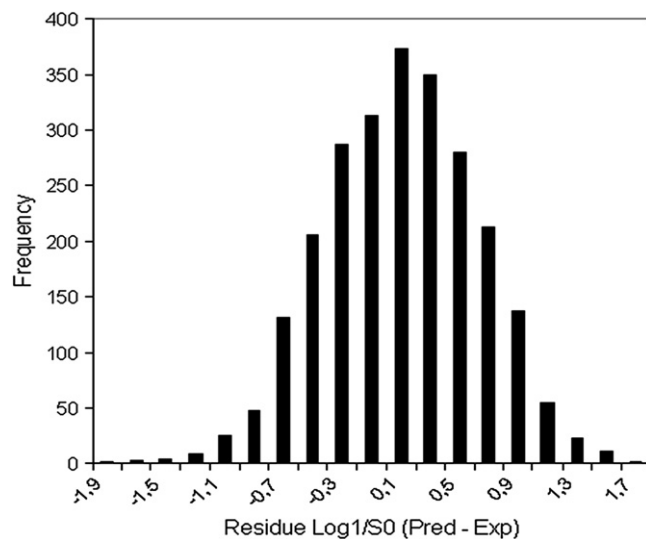


Fig. 2. Plot of the solubility residues of the analyzed drug-like compounds.

3.4. Performance with the AQUASOL data set

The question arises if the present PLS model developed for the solubility prediction of drug-like molecules could be used as well for the solubility prediction of organic compounds from the AQUASOL database. Four hundred and sixty seven organic compounds (for which the fragments were present in the newly developed fragmental database) were therefore selected from AQUASOL database and used as test data set. As expected, the solubility prediction using the newly developed prediction tool was poor with a $R^2 = 0.37$. The same data set was analyzed as well with the commercial program WsKow [18], which achieved a superior prediction ($R^2 = 0.83$). These results can be explained by different molecular properties of the compounds of the respective data sets (Fig. 4). Drug-like compounds contain mostly aromatic atoms and have a higher molecular weight. In addition, a more compact crystal packing is often observed for drug-like compounds due to

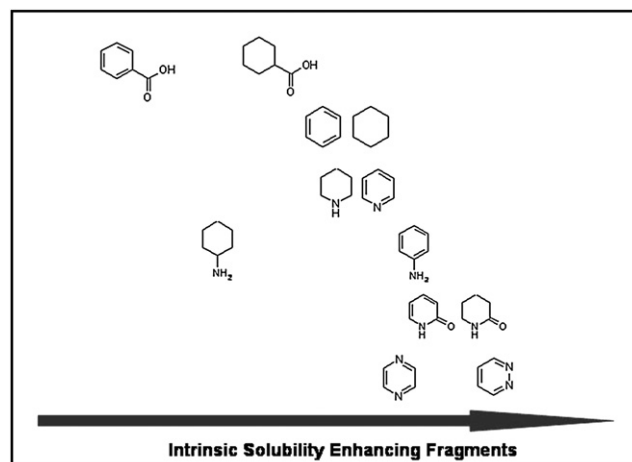


Fig. 3. Examples for intrinsic solubility enhancing fragments.

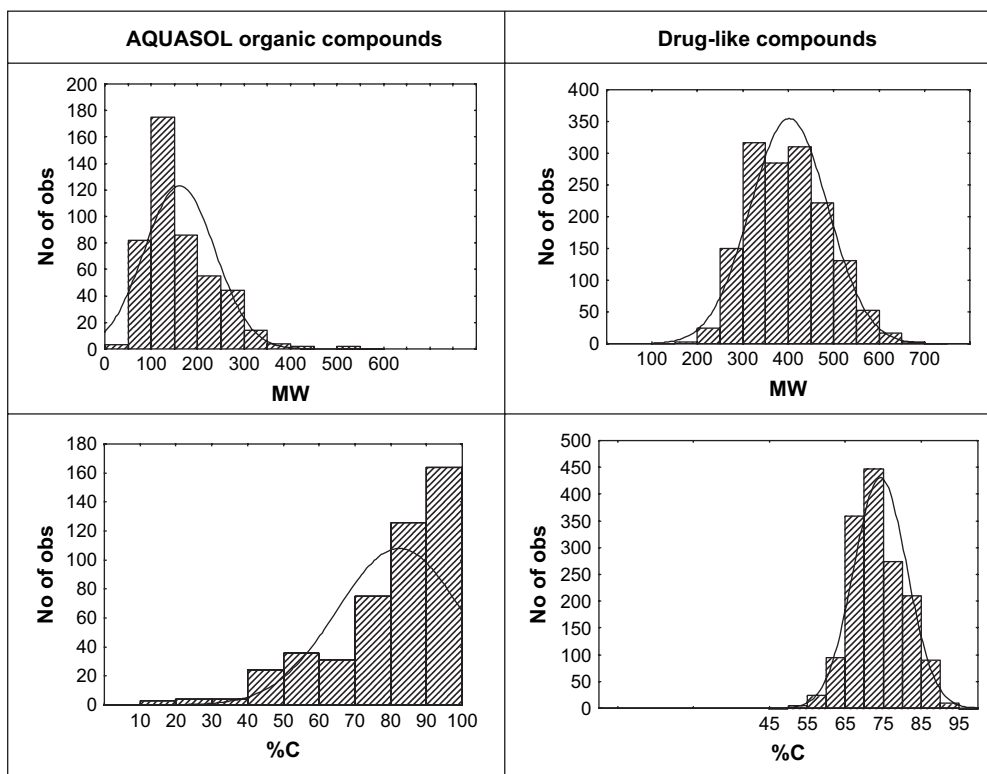


Fig. 4. Analysis of distribution of molecular weight (MW) and the percentage of aliphatic carbon atoms in the molecule (%C) of 467 organic molecules from the AQUASOL database (left panel) and 1515 drug-like compounds used in the present study (right panel).

intermolecular hydrogen bonding. In contrast to the drug-like compounds, 35% of organic compounds from the AQUASOL data set contain only aliphatic carbons. Such molecules are held together in the crystal state through van der Waals interactions. Therefore, the contribution of an aliphatic carbon atom to the solubility is different for simple organic compounds in comparison to drug-like molecules and cannot be predicted by the present PLS model.

To further explore the differences between the two data sets, PCA analysis was performed for the AQUASOL data set. Five components were calculated for PCA and its first three components described 78.1% of the x space. The dominating descriptors for this separation were molecular weight and the percentage of aliphatic carbon atoms in the molecule (%C) (Fig. 4). These descriptors were clearly different from the descriptors identified for drug-like molecules.

3.5. Impact of crystal structure and solid state

The question arises if solid state information such as melting points and crystal structures could be used to further improve the present solubility prediction model. The PLS model was therefore applied to a solubility data set based on two congeneric series of commercial compounds with crystal structures registered in the Cambridge Structure Database (CSD) [19]. The analyzed compounds were from sulphonamide and benzodiazepine series and are presented as ‘case studies’ in the Sections 3.5.1 and 3.5.2.

3.5.1. Sulfonamides

Prediction of lipophilicity of sulphonamides revealed two outliers, sulfadiazine (compound 6, Table 3) and sulfamethazine (compound 2, Table 3), with the largest deviation between their experimental and predicted solubilities. Sulfadiazine and sulfamethazine have similar 2D structures, differing by two methyl groups only. However, by losing two methyl substituents, the molecular moiety containing the pyrimidine ring in case of sulfadiazine is flatter than that of sulfamethazine. The pyrimidine rings of sulfadiazine can therefore be stacked on top of each other and thereby building six intermolecular hydrogen bonds, which lead to a higher density in the crystal packing and a higher energetic cost for crystal lattice degradation (Fig. 5). Thus, the higher melting point of sulfadiazine is a result of a very dense crystal packing leading to lower solubility (despite its lower lipophilicity as compared to the better soluble sulfamethazine). We conclude that the significant prediction error for sulfadiazine and sulfamethazine was a result of insufficient consideration of solid state properties.

3.5.2. Benzodiazepines

The molecular properties of benzodiazepines are listed in Table 4. The aqueous solubility of benzodiazepines was well predicted using the newly developed tool. It is interesting to note, however, that the higher the molecular weight, the higher the solubility of benzodiazepines. This finding does not agree with the empirical rule that increase in molecular weight often results in reduced solubility [5]. This abnormal solubility

Table 3

Chemical structure, lipophilicity, experimental (Exp) and predicted (Pred) solubilities of sulphonamides

Nr.	Name	Structure	ClogP	log <i>P</i> (Exp)	log 1/ <i>S</i> ₀ (Exp)	log 1/ <i>S</i> ₀ (Pred)
1	Sulfisomidine		1.097	−0.3	2.78	3.00
2	Sulfamethazine		1.097	0.89	2.30	3.15
3	Sulfisoxazole		0.222	1.15	3.45	2.84
4	Sulfamethoxazole		0.563	1.75	2.87	2.90
5	Sulfadoxine		1.231	1.06	3.59	3.16
6	Sulfadiazine		0.1	−0.13	3.67	2.98
7	Sulfamethoxypyridazine		0.41	0.4	2.55	2.51
8	Sulfamerazine		0.599	0.13	3.02	3.03
9	Sulfameter		0.648	0.46	3.63	3.07
10	2-Methyl-4-methoxy-6-sulfanilamidopyrimidine		1.547	0.61	3.05	3.10

Experimental lipophilicity is used here for the predictions.

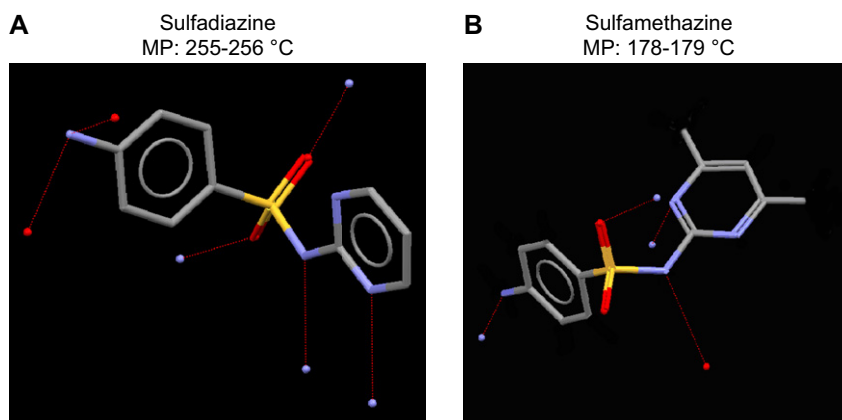


Fig. 5. Crystal structure and intermolecular hydrogen bonds of sulfadiazine and sulfamethazine. (A) 3D crystal structure of sulfadiazine. (B) 3D crystal structure of sulfamethazine.

phenomenon observed for the benzodiazepines can be explained by comparing the compounds' crystal structures and melting points in Table 4: The amido hydrogen atom in oxazepam (compound 7, Table 4) and nordiazepam (compound 2) is a good hydrogen donor, which is responsible for strong hydrogen and dipolar interaction within the crystal lattice. Thus, the N-alkylation of the amide group in temazepam (compound 4) and diazepam (compound 5) leads to lower melting points and higher solubility despite the concomitant increase in molecular weight. Therefore we conclude that molecular weight is not always a good predictor for solubility especially if small changes in congeneric series are evaluated.

3.6. Application of the PLS regression model to new congeneric series

The derived fragment related coefficients from the newly developed solubility tool can be applied to predict the solubility of any external congeneric series. Two methods were evaluated to derive solubilities for compounds with similar structures.

First, the experimental solubility of a compound in the congeneric series is taken as a starting point. The scaffold solubility value of this compound is calculated by subtracting the solubility values of substituents from the compound's experimental solubility value. The required solubility prediction value of any other compound will be a result of the calculated value of the scaffold and the solubility values of substituents derived from the generated solubility model.

The procedure of this method can be demonstrated as follows using solubility values of pyridopyrimidine trifluoromethyl ketones measured by Edwards et al. [20] (Table 5). The methyl derivative of pyridopyrimidine trifluoromethyl ketones is taken here as the starting point for the calculation. The solubility value of this scaffold ($\log 1/S_0$) is equal to 3.989, which is derived from the subtraction of the solubility value of the aromatic bonded methyl group ($\log 1/S_0 = 0.041$) from the experimental value of the methyl derivative ($\log 1/S_0 = 4.03$). The solubility of the methoxybenzyl derivative ($\log 1/S_0 = 4.346$) is calculated as the sum of its

scaffold values including the two aliphatic carbons ($\log 1/S_0 = 0.041$), six aromatic carbons ($\log 1/S_0 = 0.060$) and one ether fragment with aliphatic and aromatic binding environments ($\log 1/S_0 = -0.085$). Hence the predicted value of the methoxybenzyl derivative is 4.346.

Second, the scaffold solubility value is first calculated for each compound in the congeneric series, by subtracting the solubility values of substituents from the compound's experimental solubility value. Finally, the mean value of the scaffold is assigned to all compounds. The predicted solubility value of any other compound will then be the mean value of the scaffold in addition to its substituents' fragmental solubility values.

It is important to emphasize that both strategies outlined above are combination methods, i.e., they use both measured experimental data as well as computational predictions. These two strategies were applied to several published solubility data sets for congeneric series [20–24]. In all studied cases, the newly developed solubility prediction tool showed a high predictive power, which was superior to the one of commercially available tools [25].

4. Conclusions

In conclusion, a general solubility model of high accuracy was obtained for drug-like compounds in congeneric series when lipophilicity was used in combination with structural fragment descriptors and congeneric series indices. Rules were derived from the prediction models of solubility which can be used by medicinal chemists or interested scientists as a rough guideline on the contribution of structural fragments on solubility. However, there is still room for improvement. First, incorporating information reflecting solid state, e.g., melting point, crystal structure and density, would result in better solubility predictions. Second, the model flexibly allows extending the initial data set by addition of further structural fragments in order to enhance the predictive power. Third, solubility predictions could be extended from aqueous to other solvent systems, because different solvent and solvent mixtures are frequently used in the pharmaceutical industry for formulation and crystallization.

Table 4

Chemical structure, molecular weight, melting point, lipophilicity, experimental (Exp) and predicted (Pred) solubilities of benzodiazepines

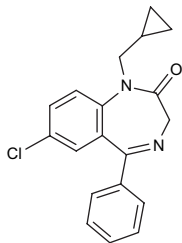
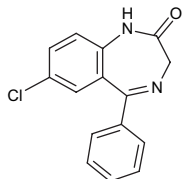
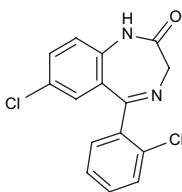
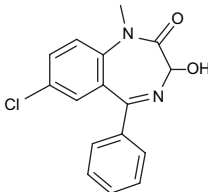
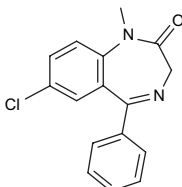
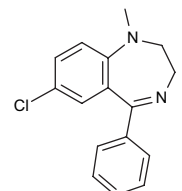
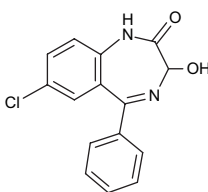
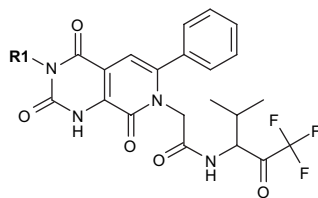
Nr.	Compound	Structure	MW	Melting point (°C)	ClogP	log 1/ <i>S</i> ₀ (Exp)	log 1/ <i>S</i> ₀ (Pred)
1	Prazepam		324.81		4.143	4.67	4.00
2	Nordiazepam		270.72	216	3.021	4.23	4.15
3	7-Chloro-5-(<i>o</i> -chlorophenyl)-1,3-dihydro-2 <i>H</i> -1,4-benzodiazepin-2-one		305.16		3.084	3.97	4.22
4	Temazepam		300.74	119	2.549	3.51	3.32
5	Diazepam		284.75	132	3.17	3.83	3.77
6	Medazepam		270.78		3.71	4.41	4.60
7	Oxazepam		286.72	197	2.305	4.12	3.68

Table 5

The solubility of pyridopyrimidine trifluoromethyl ketones measured in 0.01 M sodium phosphate buffer at pH = 7.4



Nr.	Substituent R1	pK _a	MW	S (mg/mL)	log 1/S ₀	log 1/S ₀ (pred)
1 ^a	CH ₃		478	0.044	4.03	4.03
2			584	0.23	3.40	4.35
3			521	0.13	3.60	3.65
4			535	0.1	3.73	3.34
5			538	0.42	3.11	4.27
6			536	0.008	4.83	3.80
7		7.04	577	0.30	3.44	3.74

^a Compound 1 was taken as starting point for the prediction with the newly developed solubility tool.

References

- [1] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, *Adv. Drug Deliv. Rev.* 46 (2001) 3–26.
- [2] R.C. Bravo González, J. Huwyler, I. Walter, R. Mountfield, B. Bittner, *Int. J. Pharm.* 245 (2002) 143–151.
- [3] A. Avdeef, B. Testa, *Cell. Mol. Life Sci.* 59 (2002) 1681–1689.
- [4] M.H. Abraham, J. Le, *J. Pharm. Sci.* 88 (1999) 868–880.
- [5] W.M. Meylan, P.H. Howard, *Perspect. Drug Discovery Des.* 19 (2000) 67–84.
- [6] W.L. Jorgensen, E.M. Duffy, *Bioorg. Med. Chem. Lett.* 10 (2000) 1155–1158.
- [7] S.H. Yalkowsky, *Aqueous Solubility. Methods of Estimation for Organic Compounds*, Marcel Dekker, New York, 1992.
- [8] I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva, A.E.P. Villa, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1488–1493.
- [9] A. Klamt, F. Eckert, M. Hornig, M.E. Beck, T. Burger, *J. Comput. Chem.* 23 (2002) 275–281.
- [10] Y. Ran, Y. He, G. Yang, J.L. Johnson, S.H. Yalkowsky, *Chemosphere* 48 (2002) 487–509.
- [11] A. Avdeef, C.M. Berger, *Eur. J. Pharm. Sci.* 14 (2001) 281–291.
- [12] N.E. Shemetulskis, J.B.J. Dunbar, B.W. Dunbar, D.W. Moreland, C. Humblet, *J. Comput. Aided Mol. Des.* 9 (1995) 407–416.
- [13] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, *Multi- and Megivariate Data Analysis: Principles and Applications*, Umetrics Academy, Umea, 1999.
- [14] A. Avdeef, C.M. Berger, C. Brownell, *Pharm. Res.* 17 (2000) 85–89.
- [15] A.J. Burnham, J.F. MacGregor, R. Viveros, *J. Chemom.* 13 (1999) 49–65.
- [16] A.J. Burnham, R. Viveros, J.F. MacGregor, *J. Chemom.* 10 (1996) 31–45.
- [17] S. De Jong, B.M. Wise, N.L. Ricker, *J. Chemom.* 15 (2001) 85–100.
- [18] W.M. Meylan, P.H. Howard, R.S. Boethling, *Environ. Toxicol. Chem.* 15 (1996) 100–106.
- [19] F.H. Allen, J.E. Davies, J.J. Galloy, O. Johnson, O. Kennard, C.F. Macrae, E.M. Mitchell, G.F. Mitchell, J.M. Smith, D.G. Watson, *J. Chem. Inf. Comput. Sci.* 31 (1991) 187–204.
- [20] P.D. Edwards, D.W. Andisik, A.M. Strimpler, B. Gomes, P. Tuthill, *J. Med. Chem.* 39 (1996) 1112–1124.
- [21] C. Goosen, J. Laing Timothy, J. du Plessis, C. Goosen Theunis, L. Flynn Gordon, *Pharm. Res.* 19 (2002) 13–19.
- [22] V. Bavetsias, L.A. Skelton, F. Yafai, F. Mitchell, S.C. Wilson, B. Allan, A.L. Jackman, *J. Med. Chem.* 45 (2002) 3692–3702.
- [23] T. Rosen, D.T. Chu, I.M. Lico, P.B. Fernandes, K. Marsh, L. Shen, V.G. Cepa, A.G. Pernet, *J. Med. Chem.* 31 (1988) 1598–1611.
- [24] P.R. Bernstein, D. Andisik, P.K. Bradley, C.B. Bryant, C. Ceccarelli, J.R. Damewood Jr., R. Earley, P.D. Edwards, S. Feeney, B.C. Gomes, *J. Med. Chem.* 37 (1994) 3313–3326.
- [25] W.M. Meylan, P.H. Howard, *J. Pharm. Sci.* 84 (1995) 83–92.