JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Calculation of Aqueous Solubility of Crystalline Un-Ionized Organic Chemicals and Drugs Based on Structural Similarity and Physicochemical Descriptors

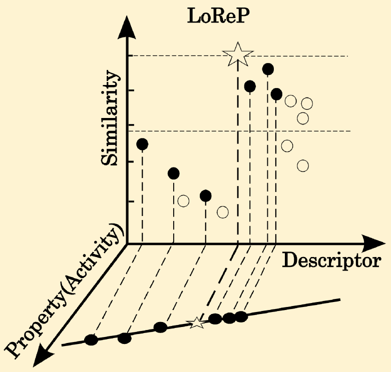
Oleg A. Raevsky,[*,†,§] Veniamin Yu. Grigor'ev,[†] Daniel E. Polianczyk,[†] Olga E. Raevskaja,[†] and John C. Dearden[‡]

[†]Institute of Physiologically Active Compounds, Russian Academy of Science, Chernogolovka, Russia

[‡]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool L3 3AF, United Kingdom

Ⓢ *Supporting Information*

ABSTRACT: Solubilities of crystalline organic compounds calculated according to AMP (arithmetic mean property) and LoReP (local one-parameter regression) models based on structural and physicochemical similarities are presented. We used data on water solubility of 2615 compounds in un-ionized form measured at 25 ± 5 °C. The calculation results were compared with the equation based on the experimental data for lipophilicity and melting point. According to statistical criteria, the model based on structural and physicochemical similarities showed a better fit with the experimental data. An additional advantage of this model is that it uses only theoretical descriptors, and this provides means for calculating water solubility for both existing and not yet synthesized compounds.



INTRODUCTION

The solubility of chemicals and drugs in the water phase has a strong influence on the extent of their absorption and transport in a body. Hence, solubility is considered to be a very important parameter in current ADMET (absorption, distribution, metabolism, excretion, toxicity) research. Water solubility is also a key determinant of the environmental impact of agrochemicals and pollutants. So, there is no question that the derivation of predictive models of solubility is extremely useful. There have been many publications presenting such quantitative structure−property relationship (QSPR) models based on fragment contribution schemes and different physicochemical and quantum chemical descriptors. These have been described and discussed in a number of reviews.[1−6]

Despite the abundance of publications on QSPR prediction of water solubility, some problems remain, especially for crystalline organic compounds.

For such chemicals, the approach describing water solubility as a function of partition coefficient and melting point is very popular.[7] However, although this method yields reasonable results, it is based on experimental values of two physicochemical properties and therefore has an essential shortcoming, which is that to calculate the solubility of a compound one needs to synthesize it and measure its partition coefficient and melting point. Hence, it is preferable to calculate solubility and other ADMET (absorption, distribution, metabolism, excretion, and toxicity) parameters based not on experimental properties but on theoretical descriptors. It may be noted that partition coefficients can be estimated well with structural fragments[7] and various descriptors[6,8] and can be used in water solubility calculations instead of experimental values. The calculation of melting point, however, is still far from satisfactory, and replacement of its experimental value with a calculated descriptor value worsens the statistical criteria of the corresponding equation. For example, replacement of measured melting point by topographical polar surface area[9] resulted in a poorer model.

Previously, we have carried out QSPR modeling of Henry's law constant of organic compounds,[10] water solubility of liquid organic compounds,[11] and crystalline organic compounds[12] using HYBOT[13] descriptors in linear regression equations. We did not obtain satisfactory results using multiple regression with physicochemical descriptors.

The present work is devoted to the creation of stable predictive models of aqueous solubility by means of AMP and LoReP QSPR models that are based on a combination of concepts of structural similarity, "read-across", and local lazy regression procedures.[14−16]

MATERIALS AND METHODS

As stated above, there are many QSPR publications concerning water solubility. Usually data on solubility of organic compounds in pure water are used mainly at a temperature of 25 °C.

However, chemicals containing acidic and basic groups can, depending on their concentration in water, be in the ionized or

Received: November 25, 2013

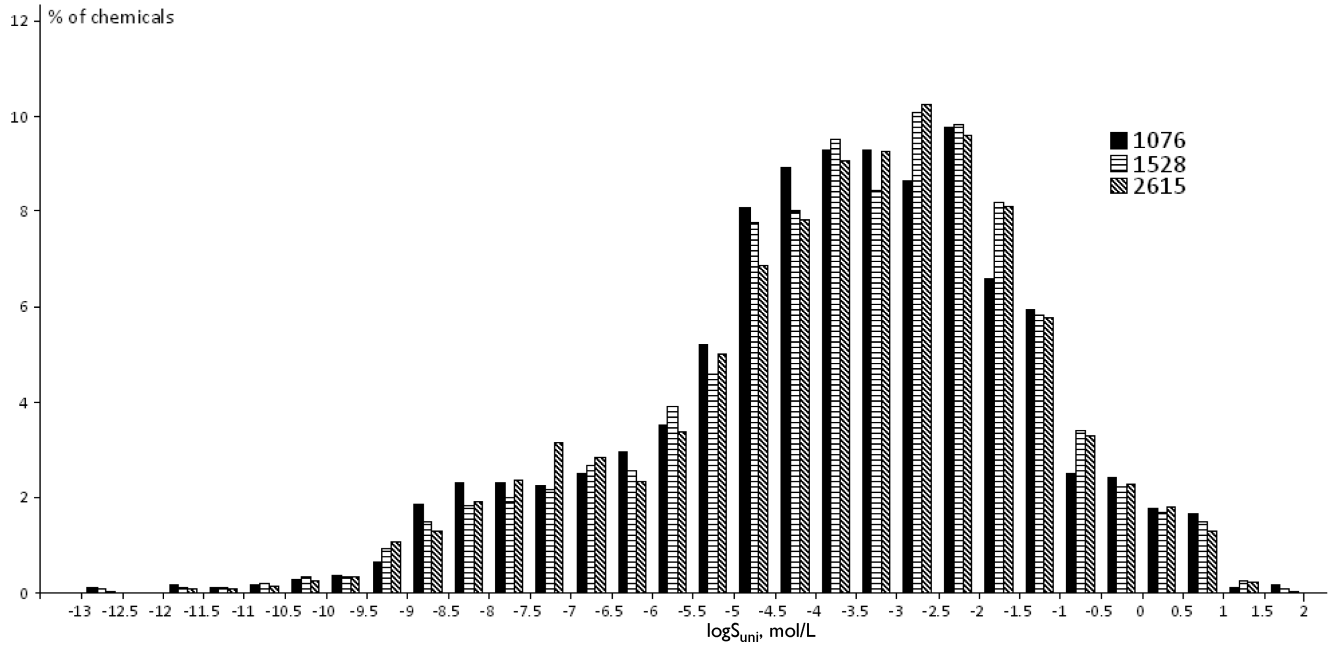


Figure 1. Solubility distribution in the three data sets.

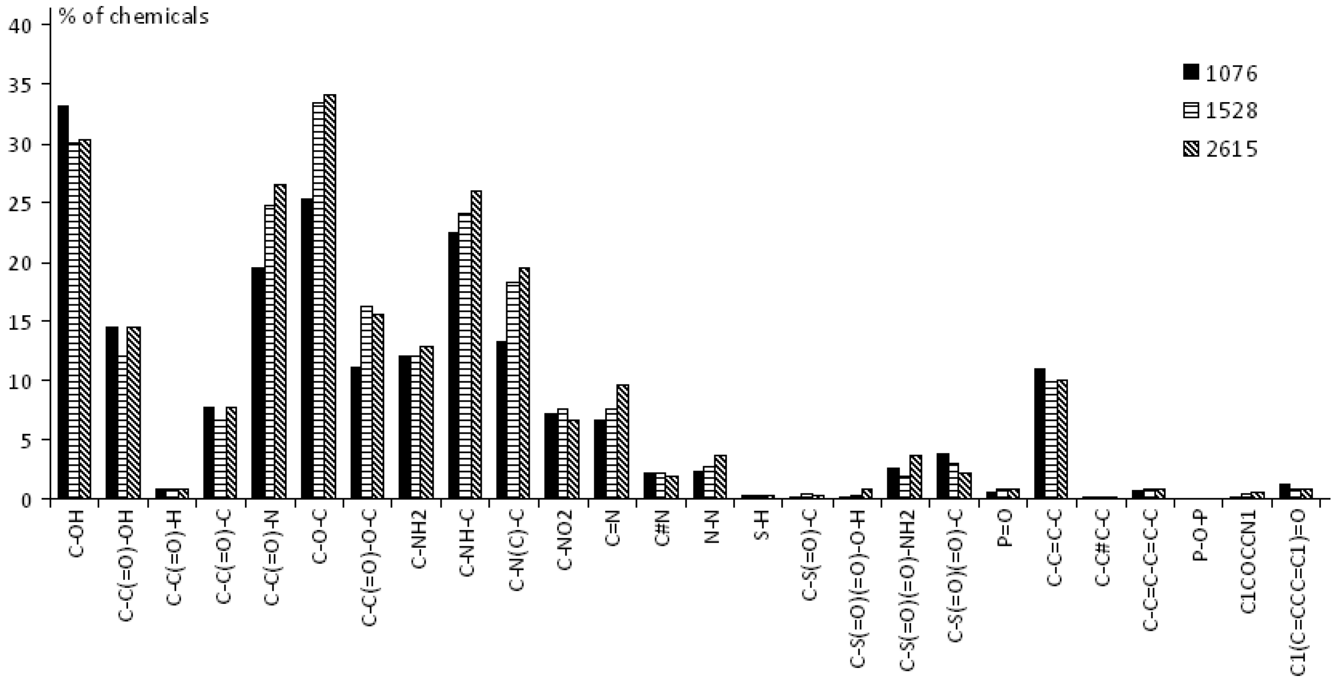


Figure 2. Distribution of compounds according to their chemical functional groups.

neutral state or in equilibrium between both. To ignore this in the development of a QSPR breaks one of the main principles of the "structure property relationship", namely, the adequate description of structure.

Therefore, in this work, we have used data on water solubility of crystalline chemicals (with units of mol/L) in the un-ionized form, logS$_{uni}$, taken from an extensive database created by Schaper and Raevskaja.[12] On the basis of these data, three data sets were constructed (Figure 1): (1) 1076 chemicals with solubility measured at 25 °C and for which there are in the specified database experimental values of melting points and partition coefficients. This data set was created for comparison of QSPR models based on experimental physicochemical parameters and theoretical descriptors. (2) 1528 chemicals with solubility measured in the range 25 ± 5 °C and for which also there are in the specified database experimental values of melting points and partition coefficients. This data set was created for an assessment of the use of solubility data in the range 25 ± 5 °C and comparison of this model and the model constructed with data obtained at a fixed temperature of 25 °C. (3) 2615 chemicals with solubility measured in the range 25 ± 5 °C. All data found and processed by us on solubility of un-ionized organic compounds are included in this set.

The Supporting Information contains complete information about the 2615 compounds studied, namely, Smile, Name, MW (molecular weight), T_{mp} (melting point, °C), log P (experimental

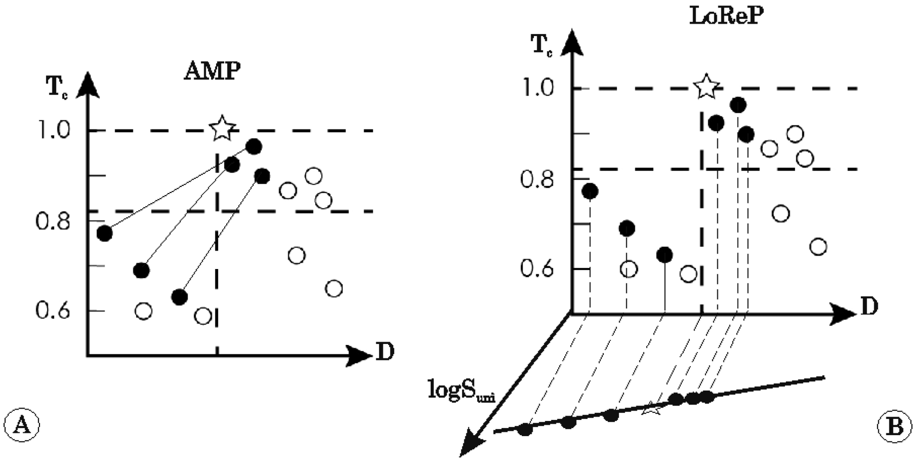


Figure 3. Comparison of AMP and LoReP models. (A) AMP uses three pairs of neighbors denoted by black circles and chemical of interest by an asterisk. Arithmetic means of property values of those chemicals is assumed as calculated chemical of interest property value. (B) LoReP uses the same neighbors for creation of the one parameter regression equation, which is used to predict chemical of interest property values.

value of partition coefficient), pK_1, pK_2, pK_3, pH (determined by the buffer), pH$_{calc}$ (calculated according to ref 11), f_{ui} (fraction un-ionized), log(1/f_{ui}), α (molecular polarizability), A log P (by Ghose−Crippen), and solubility values (in g/L, Mol/L, log S_{exp}, and log S_{uni}).

Enantiomers have identical structural descriptor values. Therefore, they were excluded from the three data sets to avoid overoptimistic results of calculations and to avoid the problem of more than one species contributing to the solubility.

For each data set, the range of solubilities (as log S_{uni}, S in mol/L) is −12.79 to 1.70. For 95% of all compounds, this interval is slightly less (−9.00 to 1.00). The distribution of solubilities takes a bell-like form for all three data sets, peaking in the range of −3.00 to −1.50. The solubility range (14.5 logarithmic units) greatly exceeds the experimental error (from 0.5 to 0.6 log unit[5]). This observation, together with the fact that all compounds are in the un-ionized state, conforms to QSPR modeling requirements.

Another important requirement of data for QSPR is a good range of structural variation of compounds (Figure 2).

It is shown from Figure 2 that there are five prevailing chemical groups: C−OH, C−O−C, C(= O)−N, C−NH−C, and C−N(C)−C. The percentages of each functional group are approximately the same in all three data sets.

The assessment of structural similarity of molecules was made on the basis of fragments of structural formulas. Spherical fragments with various spherical radii around the chosen atom were used. The radius of the sphere is the topological distance between the chosen radius and the atom most removed from it; each bond length is taken to be equal to one unit. To divide molecules into spherical fragments, each atom is considered as the sphere center. Then the number of fragments with spherical radii equal to 1, 2, etc. are counted. The subsequent details of the process are described in ref 17. Structural similarity was estimated using the Tanimoto similarity index (T_c=N_{AB}/(N_A + N_B − N_{AB}), where N_{AB} is the number of common fragments, and N_A and N_B are the number of fragments for A and B) obtained by means of the MOLDIVS (MOLecular DIVersity and Similarity) program.[18]

A set of physicochemical descriptors from HYBOT[13] and DRAGON[19] software packages was used for QSPR modeling. These descriptors characterize steric and electrostatic interactions and H-bond factors and also Moriguchi (MlogP) and Ghose−Crippen (AlogP) partition coefficients. Values of each of the specified descriptors cover wide ranges and thus are adequate for QSPR.

Each descriptor was used in QSPR separately, obviating the need to examine pairwise correlations.

The following QSPR methods were used: (1) The equation published in refs 20,21 with the fixed values of coefficients at independent variables. (2) A linear regression model based on experimental values of melting point and partition coefficient. (3) A linear regression model with any combination of the calculated descriptors of the HYBOT and DRAGON programs. (4) k-nearest neighbors (k-NN) based on structural similarity. (5) AMP (Arithmetical mean property).[22] In this case, first, for each chemical of interest, all the other chemicals were ordered in accordance with their structural similarity. This procedure was carried out by the MOLDIVS program using Tanimoto index values. Then a set of descriptors were calculated for all chemicals. For each chemical, from a cluster of structural neighbors one or few pairs of chemicals were selected, such that each pair contained one chemical with a higher and another chemical with a lower descriptor value compared with descriptor value for the chemical of interest. The arithmetic mean value of the property of those pairs was used as the calculated property value of the chemical of interest (Figure 3). The number of pairs is selected empirically and depends on the size and quality of the training set (a large number of pairs is possible only in the case of very large data sets). It should also be remembered that increasing the number of neighbor pairs approximates to the standard k-NN procedure. The main difference between AMP and k-NN relates to the selection type of **nearest** structural neighbors. (6) LoReP (local one-parameter regression) with application to the same structural neighbors as in model 5.

■ RESULTS AND DISCUSSION

A large data set in which melting point and partition coefficient were previously used to model aqueous solubility led to the following QSPR[20]

$$\log S = 0.3814 - 1.0223 \log P - 0.00961(T_{mp} - 25)$$
$$n = 1026, R^2 = 0.96, \text{sd} = 0.4957 \quad (1)$$

Table 1. Results for AMP Model (T_c ≥0.00) for 32 Descriptors, Calculated by the HYBOT and DRAGON Programs[a]: log S_{uni}(exp) = a_0 ± err$_0$ + (a_1 ± err$_1$)log S_{uni}(calc)

descriptor[b]	N	R^2	sd	F	Q^2	sd_cv	a_0 ± err$_0$	a_1 ± err$_1$
A log P	1070	0.826	0.95	5070.3	0.826	0.95	0.10 ± 0.06	1.04 ± 0.01
M log P	1070	0.803	1.01	4342.5	0.803	1.01	0.04 ± 0.07	1.03 ± 0.02
α	1070	0.801	1.02	4296.1	0.801	1.02	0.15 ± 0.07	1.05 ± 0.02
MW	1070	0.797	1.03	4202.4	0.796	1.03	0.16 ± 0.07	1.04 ± 0.04
$\sum E_d/\alpha$	678	0.556	1.09	847.2	0.556	1.09	−0.02 ± 0.11	1.02 ± 0.04
$\sum C_d/\alpha$	678	0.539	1.11	789.6	0.538	1.11	−0.06 ± 0.11	1.02 ± 0.04
$\sum C_{ad}/\alpha$	1069	0.732	1.19	2910.0	0.731	1.19	−0.08 ± 0.08	1.02 ± 0.02
$\sum E_{ad}/\alpha$	1070	0.732	1.19	2920.4	0.731	1.19	−0.06 ± 0.08	1.02 ± 0.02
$\sum C_a/\alpha$	1069	0.730	1.19	2883.6	0.729	1.19	−0.18 ± 0.08	1.01 ± 0.02
$\sum Q^-/\alpha$	1066	0.719	1.20	2725.7	0.718	1.20	−0.12 ± 0.08	1.01 ± 0.02
ΣQ^-	1070	0.725	1.20	2821.7	0.724	1.21	0.09 ± 0.08	1.04 ± 0.02
max(E_a)×max(E_d)	678	0.462	1.21	580.3	0.459	1.21	−0.13 ± 0.12	1.02 ± 0.04
$\sum E_a/\alpha$	1070	0.720	1.22	2744.1	0.719	1.22	−0.14 ± 0.08	1.01 ± 0.02
PSA$_{cd}$	678	0.453	1.22	558.9	0.450	1.22	−0.04 ± 0.13	1.05 ± 0.04
$\sum C_d$	678	0.452	1.22	558.0	0.449	1.22	−0.03 ± 0.13	1.05 ± 0.04
PSA$_{ed}$	678	0.450	1.23	554.1	0.447	1.23	−0.04 ± 0.13	1.04 ± 0.04
$\sum E_d$	678	0.449	1.23	551.5	0.446	1.23	−0.05 ± 0.13	1.04 ± 0.04
max(E_d)	678	0.446	1.23	544.7	0.443	1.23	−0.04 ± 0.13	1.08 ± 0.05
max(C_a)×max(C_d)	678	0.430	1.24	510.0	0.426	1.24	−0.22 ± 0.13	0.98 ± 0.04
$\sum E_a$	1070	0.707	1.25	2581.6	0.706	1.25	−0.02 ± 0.08	1.01 ± 0.02
max(C_d)	678	0.410	1.27	469.4	0.406	1.27	−0.16 ± 0.13	1.03 ± 0.05
$\sum E_{ad}$	1070	0.699	1.27	2482.2	0.698	1.27	−0.09 ± 0.08	1.01 ± 0.02
max(C_a)	1070	0.693	1.27	2410.3	0.692	1.27	−0.10 ± 0.08	1.03 ± 0.02
PSA$_c$	908	0.548	1.29	1098.8	0.546	1.30	−0.09 ± 0.10	1.03 ± 0.03
$\sum C_a$	1070	0.680	1.30	2271.8	0.679	1.31	−0.08 ± 0.09	1.00 ± 0.02
PSA$_{ca}$	908	0.538	1.31	1056.6	0.536	1.31	−0.02 ± 0.11	1.04 ± 0.03
$\sum C_{ad}$	1070	0.670	1.33	2165.1	0.668	1.33	−0.21 ± 0.09	0.98 ± 0.02
PSA$_e$	908	0.519	1.33	976.9	0.516	1.34	0.14 ± 0.12	1.10 ± 0.04
PSA$_{ea}$	908	0.511	1.34	946.1	0.508	1.35	0.19 ± 0.12	1.11 ± 0.04
max(E_a)	1068	0.645	1.37	1937.8	0.644	1.37	0.55 ± 0.11	1.27 ± 0.03
TPSA(Tot)	909	0.490	1.37	870.2	0.487	1.38	0.33 ± 0.13	1.12 ± 0.04
TPSA(NO)	908	0.484	1.38	848.6	0.481	1.39	0.31 ± 0.13	1.12 ± 0.04

[a] N is the number of compounds, R is the correlation coefficient, sd is the standard deviation, Q is the cross-validated leave-one-out correlation coefficient, sd_cv is the cross-validated leave-one-out standard deviation, and F is the Fisher criterion. [b] Descriptors: molecular polarizability (α), molecular weight (MW), Moriguchi octanol−water partition coefficient (M log P), Ghose−Crippen octanol−water partition coefficient (A log P), sum of all negative atomic charges in a molecule ($\sum Q^-$), maximal enthalpy H-bond acceptor atom factor in a molecule (max(E_a)), maximal enthalpy H-bond donor atom factor in a molecule (max(E_d)), maximal free energy H-bond acceptor atom factor in a molecule (max(C_a)), maximal free energy H-bond donor atom factor in a molecule (max(C_d)), sum of enthalpy H-bond acceptor atom factor in a molecule ($\sum E_a$), sum of enthalpy H-bond donor atom factor in a molecule ($\sum E_d$), sum of free energy H-bond acceptor atom factor in a molecule ($\sum C_a$), sum of enthalpy H-bond acceptor and donor atom factor in a molecule ($\sum E_{ad}$), sum of free energy H-bond acceptor and donor atom factors in a molecule ($\sum C_{ad}$), topological polar surface area using N, O, S, and P polar contributions TPSA(Tot), topological polar surface area using N and O polar contributions TPSA(NO), van der Waals donor and acceptor surface area, proportional to enthalpy H-bond factor (PSA$_e$= PSA$_{ea}$ + PSA$_{ed}$), and van der Waals donor and acceptor surface area, proportional to free energy H-bond factor (PSA$_c$= PSA$_{ca}$ + PSA$_{cd}$).

Table 2. Six Neighbors with T_c ≥ 0.3 to Chemical of Interest

descriptor	N	R^2	sd	a_0 ± err$_0$	a_1 ± err$_1$
A log P	670	0.877	0.80	0.19 ± 0.07	1.03 ± 0.02
M log P	663	0.871	0.82	0.13 ± 0.07	1.02 ± 0.02
α	656	0.867	0.83	0.14 ± 0.07	1.03 ± 0.02
MW	653	0.864	0.84	0.11 ± 0.07	1.02 ± 0.02

Table 3. Six Neighbors with T_c ≥ 0.5 to Chemical of Interest

descriptor	N	R^2	sd	a_0 ± err$_0$	a_1 ± err$_1$
MW	253	0.940	0.61	0.08 ± 0.09	1.01 ± 0.02
A log P	288	0.939	0.63	0.26 ± 0.08	1.04 ± 0.02
M log P	271	0.938	0.63	0.17 ± 0.09	1.03 ± 0.02
α	271	0.934	0.63	0.09 ± 0.09	1.01 ± 0.02

In the present work, eq 1 was used for the calculation of solubility of 1076 chemicals. This yielded the following correlation

$$\log S_{\text{uni}}(\text{exp}) = -0.43(\pm 0.06) + 0.94(\pm 0.01)\log S_{\text{uni}}(\text{calc})$$

$$n = 1076,\ R^2 = 0.820,\ \text{sd} = 0.98,\ Q^2 = 0.820,\ F = 4905.3 \quad (2)$$

Clearly, the statistical criteria of eq 2 are worse than those of eq 1. This is possibly because eq 1 was based on a training set that contained not only solid but also about 500 liquid chemicals.

Direct regression of solubility values of our 1076 compound data set with experimental values of melting point and partition coefficient yielded the following QSPR

Table 4. Consensus AMP Models for 1076 Compounds with Different Similarities

$T_c \geq$	descriptors	N	R^2	sd	$a_0 \pm \mathrm{err}_0$	$a_1 \pm \mathrm{err}_1$
0	α, $A\log P$	1076	0.836	0.93	0.25 ± 0.06	1.08 ± 0.02
	MW, $A\log P$	1075	0.835	0.93	0.27 ± 0.06	1.07 ± 0.01
	MW, $M\log P$	1075	0.824	0.96	0.23 ± 0.06	1.07 ± 0.02
	α, $M\log P$	1075	0.823	0.97	0.21 ± 0.06	1.07 ± 0.02
0.3	α, $A\log P$	736	0.871	0.81	0.22 ± 0.07	1.04 ± 0.02
	α, $M\log P$	724	0.867	0.82	0.19 ± 0.06	1.04 ± 0.02
	α, MW	690	0.866	0.83	0.14 ± 0.07	1.03 ± 0.02
	MW, $A\log P$	739	0.866	0.83	0.21 ± 0.07	1.04 ± 0.01
0.5	α, MW	289	0.936	0.62	0.11 ± 0.08	1.02 ± 0.02
	MW, $A\log P$	319	0.936	0.64	0.22 ± 0.08	1.03 ± 0.02
	α, $A\log P$	324	0.932	0.65	0.21 ± 0.08	1.03 ± 0.02
	α, $M\log P$	316	0.931	0.66	0.20 ± 0.08	1.03 ± 0.02

Table 5. Consensus LoReP Models for 1076 Compounds with Different Similarities

$T_c \geq$	descriptors	N	R^2	sd	$a_0 \pm \mathrm{err}_0$	$a_1 \pm \mathrm{err}_1$
0	α, $A\log P$	1076	0.874	0.82	0.12 ± 0.05	1.05 ± 0.01
	MW, $A\log P$	1076	0.871	0.83	0.15 ± 0.05	1.04 ± 0.01
	α, $M\log P$	1075	0.858	0.87	0.11 ± 0.06	1.05 ± 0.01
	MW, $M\log P$	1075	0.858	0.87	0.14 ± 0.06	1.05 ± 0.01
	$A\log P$	1073	0.852	0.88	−0.05 ± 0.05	1.00 ± 0.01
	α	1070	0.827	0.95	−0.01 ± 0.06	1.01 ± 0.01
	$M\log P$	1075	0.821	0.97	−0.11 ± 0.06	0.99 ± 0.01
	MW	1076	0.799	1.03	−0.06 ± 0.07	0.98 ± 0.01
0.3	$A\log P$	670	0.902	0.72	0.11 ± 0.06	1.02 ± 0.01
	α	656	0.890	0.75	0.02 ± 0.06	1.01 ± 0.01
	$M\log P$	663	0.887	0.77	0.04 ± 0.06	1.00 ± 0.01
	α, $A\log P$	736	0.882	0.78	−0.02 ± 0.06	1.00 ± 0.01
	MW	653	0.880	0.79	0.01 ± 0.07	1.01 ± 0.02
	MW, $A\log P$	739	0.879	0.79	0.00 ± 0.06	1.00 ± 0.01
	α, MW	690	0.875	0.80	0.02 ± 0.06	1.00 ± 0.01
	α, $M\log P$	724	0.872	0.81	−0.03 ± 0.06	1.00 ± 0.01
0.5	α	271	0.955	0.52	0.00 ± 0.07	1.01 ± 0.01
	MW	253	0.953	0.55	−0.04 ± 0.08	1.01 ± 0.01
	$A\log P$	288	0.951	0.57	0.11 ± 0.07	1.02 ± 0.01
	$M\log P$	271	0.946	0.59	0.02 ± 0.08	1.00 ± 0.02
	α, $A\log P$	324	0.944	0.59	0.04 ± 0.07	1.01 ± 0.01
	MW, $A\log P$	319	0.942	0.61	0.04 ± 0.07	1.01 ± 0.01
	MW, $M\log P$	307	0.939	0.63	−0.07 ± 0.08	0.99 ± 0.01
	α, $M\log P$	316	0.930	0.67	−0.08 ± 0.08	0.98 ± 0.02

$$\log S_{\mathrm{uni}} = -0.03(\pm 0.08) - 0.96(\pm 0.01)\log P - 0.0093(\pm 0.0004)(T_{\mathrm{mp}} - 25)$$

$$n = 1076,\ R^2 = 0.821,\ \mathrm{sd} = 0.98,\ Q^2 = 0.820,\ F = 2451.7 \quad (3)$$

This equation has almost zero intercept, while the coefficients of $\log P$ and T_{mp} are close to the corresponding coefficients in eq 1.

We believe that eq 3 is better for our purposes than eq 1 as it is based on data on solubility only of crystalline solids un-ionized in solution. Its statistical criteria can be considered as satisfactory although the standard deviation (sd = 0.98) is considerably higher than the experimental error.[5]

Of course, eq 3, like eq 1, has an essential shortcoming—the use of experimental values of chemical parameters. Therefore, the next step in our work was the use of theoretically calculated descriptors, in particular molecular polarizability (α) and the sum of H-bond acceptor ($\sum C_a$) and donor ($\sum C_d$) factors, which well describe the solubility of liquid chemicals.[15] An equation with rather poor statistical parameters was obtained ($R^2 = 0.704$, sd = 1.25). The use of calculated values of the partition coefficient ($A\log P$[23]) and molecular polarizability gave a slightly better result ($R^2 = 0.760$, sd = 1.13).

Next we used the k-NN method, with four, six, and ten neighbors (eqs 4−6)

$$\log S_{\mathrm{uni}}(\exp) = -0.29(0.08) + 0.95(0.02)\log S_{\mathrm{uni}}(\mathrm{calc\_4nn})$$

$$n = 1076,\ R^2 = 0.715,\ \mathrm{sd} = 1.23,\ F = 2694.9 \quad (4)$$

$$\log S_{\mathrm{uni}}(\exp) = -0.20(0.08) + 0.97(0.02)\log S_{\mathrm{uni}}(\mathrm{calc\_6nn})$$

$$n = 1076,\ R^2 = 0.696,\ \mathrm{sd} = 1.27,\ F = 2460.2 \quad (5)$$

$$\log S_{\mathrm{uni}}(\exp) = -0.16(0.09) + 0.99(0.02)\log S_{\mathrm{uni}}(\mathrm{calc\_10nn})$$

$$n = 1076,\ R^2 = 0.651,\ \mathrm{sd} = 1.36,\ F = 2003.4 \quad (6)$$

Table 6. Single and Consensus AMP and LoReP Models for 1528 Compounds with Different Similarities

method	$T_c \geq$	descriptors	N	R^2	sd	$a_0 \pm \text{err}_0$	$a_1 \pm \text{err}_1$
AMP	0	α, $A \log P$	1528	0.821	0.96	0.15 ± 0.05	1.06 ± 0.01
		$A \log P$	1522	0.814	0.97	0.08 ± 0.05	1.03 ± 0.01
		α	1522	0.803	1.00	0.10 ± 0.05	1.04 ± 0.01
	0.3	$A \log P$	1006	0.865	0.83	0.15 ± 0.05	1.03 ± 0.01
		α, $A \log P$	1099	0.862	0.84	0.17 ± 0.05	1.04 ± 0.01
		α	982	0.856	0.86	0.08 ± 0.06	1.02 ± 0.01
	0.5	$A \log P$	434	0.933	0.68	0.21 ± 0.07	1.03 ± 0.01
		α	407	0.931	0.69	0.17 ± 0.07	1.03 ± 0.01
		α, $A \log P$	489	0.928	0.69	0.23 ± 0.06	1.03 ± 0.01
LoReP	0	α, $A \log P$	1528	0.861	0.85	0.06 ± 0.04	1.03 ± 0.01
		$A \log P$	1522	0.842	0.90	−0.04 ± 0.05	1.00 ± 0.01
		α	1522	0.831	0.93	−0.05 ± 0.05	1.00 ± 0.01
	0.3	$A \log P$	1006	0.889	0.76	0.07 ± 0.05	1.02 ± 0.01
		α	982	0.878	0.79	−0.02 ± 0.05	1.00 ± 0.01
		α, $A \log P$	1099	0.875	0.80	−0.02 ± 0.05	1.00 ± 0.01
	0.5	α	407	0.950	0.58	0.06 ± 0.06	1.02 ± 0.01
		$A \log P$	434	0.947	0.60	0.09 ± 0.06	1.02 ± 0.01
		α, $A \log P$	489	0.944	0.61	0.04 ± 0.05	1.01 ± 0.01

Table 7. Single and Consensus AMP and LoReP Models for 2615 Compounds with Different Similarity Levels

method	$T_c \geq$	descriptors	N	R^2	sd	$a_0 \pm \text{err}_0$	$a_1 \pm \text{err}_1$
AMP	0	$A \log P$	2609	0.823	0.96	0.09 ± 0.04	1.03 ± 0.01
		α	2609	0.805	1.00	0.05 ± 0.04	1.03 ± 0.01
		α, $A \log P$	2613	0.825	0.95	0.11 ± 0.04	1.04 ± 0.01
	0.3	α	1886	0.859	0.85	0.09 ± 0.04	1.02 ± 0.01
		$A \log P$	1914	0.849	0.89	0.10 ± 0.04	1.02 ± 0.01
		α, $A \log P$	2069	0.854	0.87	0.13 ± 0.04	1.03 ± 0.01
	0.5	$A \log P$	907	0.922	0.71	0.19 ± 0.05	1.03 ± 0.01
		α	862	0.921	0.72	0.11 ± 0.05	1.02 ± 0.01
		α, $A \log P$	1020	0.917	0.72	0.21 ± 0.05	1.03 ± 0.01
LoReP	0	$A \log P$	2609	0.842	0.90	−0.03 ± 0.04	1.00 ± 0.01
		α	2609	0.827	0.94	−0.07 ± 0.04	1.00 ± 0.01
		α, $A \log P$	2613	0.855	0.86	0.02 ± 0.03	1.02 ± 0.01
	0.3	α	1886	0.878	0.79	0.00 ± 0.04	1.00 ± 0.01
		$A \log P$	1914	0.871	0.82	0.02 ± 0.04	1.01 ± 0.01
		α, $A \log P$	2069	0.867	0.83	−0.05 ± 0.04	1.00 ± 0.01
	0.5	α	862	0.935	0.65	0.04 ± 0.05	1.01 ± 0.01
		$A \log P$	907	0.935	0.65	0.10 ± 0.04	1.02 ± 0.01
		α, $A \log P$	1020	0.929	0.67	0.03 ± 0.04	1.01 ± 0.01

where $\log S_{\text{uni}}$(calc_4nn), $\log S_{\text{uni}}$(calc_6nn), and $\log S_{\text{uni}}$(calc_10nn) are solubility values calculated by k-NN using four, six, and ten structural neighbors respectively.

Clearly the k-NN method led to poor statistical parameters even in comparison with eq 3.

We then used our own AMP method. Initially, we used separately 32 descriptors calculated by the HYBOT and DRAGON programs with a full range of the similarity index from 0 to 1 ($T_c \geq 0.00$). For each chemical, three pairs of neighbors were considered, with higher and lower descriptor values than those of the chemical of interest. This procedure allowed us to obtain results for almost all chemicals (Table 1).

Statistical parameters of the AMP method with a single descriptor are wide ranging: R^2 from 0.484 to 0.822; sd from 0.96 to 1.38.

The best results were obtained with AlogP, α, MW and MlogP. Thus use of AlogP within the AMP model led to better results, not only within direct regression, but also in comparison with eq 3.

We next considered classification of subsets, based on a level value of the Tanimoto index. This means that the higher the similarity, the closer is the property value of the compound of interest with those of its neighbors. In a given data set there are compounds with different similarities to their neighbors, and so the accuracy of solubility calculations will differ, and will depend on the structural and physicochemical similarity of neighbors. Results for compounds with neighbors with various structural similarity levels are presented.

It can be seen that increasing the similarity increases the predictive ability of the model.

A consensus model is based on the use of the arithmetic mean of the values calculated for all compounds in each single model.[24] Descriptor correlation is unimportant here because the arithmetic mean is the result of calculations. For 32 descriptors, there are 496 pairwise combinations, the best of which are listed in Table 4.

The consensus model gave improved statistic parameters. Consideration of consensus models based on three descriptors

Table 8. Single and Consensus AMP and LoReP Models for Training and Test Sets with Different Similarities

$T_c \geq$	descriptors	set	N	R^2	sd	F	Q^2	sd_cv	$a_0 \pm \text{err}_0$	$a_1 \pm \text{err}_1$
0	α, $A \log P$	training	2092	0.845	0.90	11371.0	0.844	0.90	0.03 ± 0.04	1.02 ± 0.01
		test	521	0.858	0.85	3137.8	0.856	0.85	−0.05 ± 0.07	1.00 ± 0.02
	$A \log P$	training	2087	0.829	0.94	10073.0	0.828	0.94	−0.03 ± 0.04	1.00 ± 0.01
		test	517	0.853	0.86	2989.0	0.852	0.86	0.00 ± 0.08	1.01 ± 0.02
	α	training	2087	0.815	0.98	9171.5	0.814	0.98	−0.06 ± 0.04	1.00 ± 0.01
		test	517	0.831	0.92	2530.5	0.830	0.93	−0.11 ± 0.08	0.98 ± 0.02
0.3	α	training	1430	0.875	0.80	9983.5	0.875	0.81	0.00 ± 0.04	1.00 ± 0.01
		test	358	0.881	0.79	2634.7	0.880	0.79	−0.11 ± 0.08	0.99 ± 0.02
	$A \log P$	training	1451	0.865	0.84	9321.8	0.865	0.84	0.03 ± 0.05	1.01 ± 0.01
		test	362	0.880	0.79	2643.6	0.879	0.79	−0.06 ± 0.08	1.00 ± 0.02
	α, $A \log P$	training	1579	0.862	0.84	9885.1	0.862	0.84	−0.05 ± 0.04	1.00 ± 0.01
		test	395	0.870	0.81	2638.8	0.869	0.82	−0.18 ± 0.08	0.97 ± 0.02
0.5	α	training	605	0.938	0.65	9179.6	0.938	0.65	0.06 ± 0.06	1.01 ± 0.01
		test	150	0.950	0.57	2791.9	0.948	0.58	−0.25 ± 0.10	0.98 ± 0.02
	$A \log P$	training	627	0.932	0.68	8596.9	0.932	0.68	0.10 ± 0.06	1.02 ± 0.01
		test	157	0.949	0.57	2872.7	0.948	0.58	−0.06 ± 0.10	1.00 ± 0.02
	α, $A \log P$	training	727	0.920	0.72	8327.2	0.919	0.73	0.03 ± 0.05	1.01 ± 0.01
		test	180	0.943	0.59	2947.7	0.942	0.60	−0.10 ± 0.09	0.99 ± 0.02

Table 9. Single and Consensus AMP and LoReP Models for External Test Sets with Different Similarities

$T_c \geq$	descriptors	set	N	R^2	sd	F	Q^2	sd_cv	$a_0 \pm \text{err}_0$	$a_1 \pm \text{err}_1$
0	α, $A \log P$	training	2613	0.853	0.87	15095.0	0.852	0.87	0.01 ± 0.03	1.02 ± 0.01
		external	43	0.702	0.89	96.4	0.668	0.94	−0.29 ± 0.37	0.97 ± 0.10
0.3	α, $A \log P$	training	2069	0.869	0.82	13658.0	0.868	0.82	−0.04 ± 0.04	1.00 ± 0.01
		external	27	0.833	0.56	124.8	0.813	0.59	−0.38 ± 0.29	0.91 ± 0.08
0.5	α, $A \log P$	training	1020	0.930	0.66	13584.0	0.930	0.66	0.07 ± 0.04	1.02 ± 0.01
		external	9	0.619	0.58	11.4	0.405	0.73	−0.48 ± 0.57	0.88 ± 0.26

(4960 models) did not show improved results compared to the model built on a single descriptor.

Our next step was to consider LoReP based on data for the same six neighbors, which form three pairs in the AMP (Table 5).

LoReP and AMP based on molecular polarizability (α) and Ghose−Crippen partition coefficient ($A \log P$) have better predictive ability than with the use of other descriptors.

Thus, it is shown that the use of theoretical descriptors can build a more sustainable model compared with those based on experimental parameters.

For the second set of 1528 compounds, the solubilities of which were measured at 25 ± 5 °C, we also built a regression equation based on the experimental data of melting point and partition coefficient and compared the results of our models.

$$\log S_{\text{uni}} = -0.08(\pm 0.06) - 0.96(\pm 0.01)\log P - 0.0094(\pm 0.0004)(T_{\text{mp}} - 25)$$

$$n = 1528,\ R^2 = 0.814,\ \text{sd} = 0.98,\ Q^2 = 0.814,\ F = 3340.2 \qquad (7)$$

The use of solubility data for the extended temperature range led to a substantially coincident QSPR model to that for the first data set.

The results of calculations for a single descriptor model and a consensus model are listed in Table 6.

The results in Table 6 show QSPRs almost identical with the results obtained at fixed temperature.

The QSPR results of the solubility prediction of 2615 compounds are presented in the Table 7.

These results show that the AMP and LoReP models work well for calculating the solubility of un-ionized crystalline organic compounds. If compounds have structural neighbors at $T_c \geq 0.5$, LoReP allows calculation of $\log S_{\text{uni}}$ with an accuracy of ±0.65, which is close to the experimental error.[4]

The stability of the models obtained was examined with the use of internal and external test subsets.

It is very important to note that the solubility values of the chemicals of interest are never included in the calculations made by the LoReP model. Thus, any chemical in the studied data set or other external database may to be considered as an independent test chemical in the assessment of the LoReP model. Nevertheless, we selected training and test sets from the studied data set by using MOLDIVS. In this case, all chemicals were put in order of their dissimilarity by applying the following algorithm: (a) selection of a chemical that is most dissimilar from all others, (b) selection of a second chemical that is most dissimilar from the first, (c) selection of a third chemical that is most dissimilar from the first two, and so on. Then each fifth chemical was selected in the test set, and the remaining chemicals were included in the training set. The test set chemicals were predicted by looking for neighbors only among compounds in the training set. The results are given in the Table 8.

To test the method, we also used an external test set[25] with data on $\log S_{\text{uni}}$ of crystalline compounds that were not included in our database (43 chemicals) (Table 9).

The distribution of internal training and test sets, and also external test set chemicals, is given in Figure 4. It is clear that almost all compounds of the internal and external test sets are located in the training set space. Hence, our constructed models are correct regarding applicability domain.

The results of this calculation are listed in the Table 8 and show that the models obtained are stable and allow for

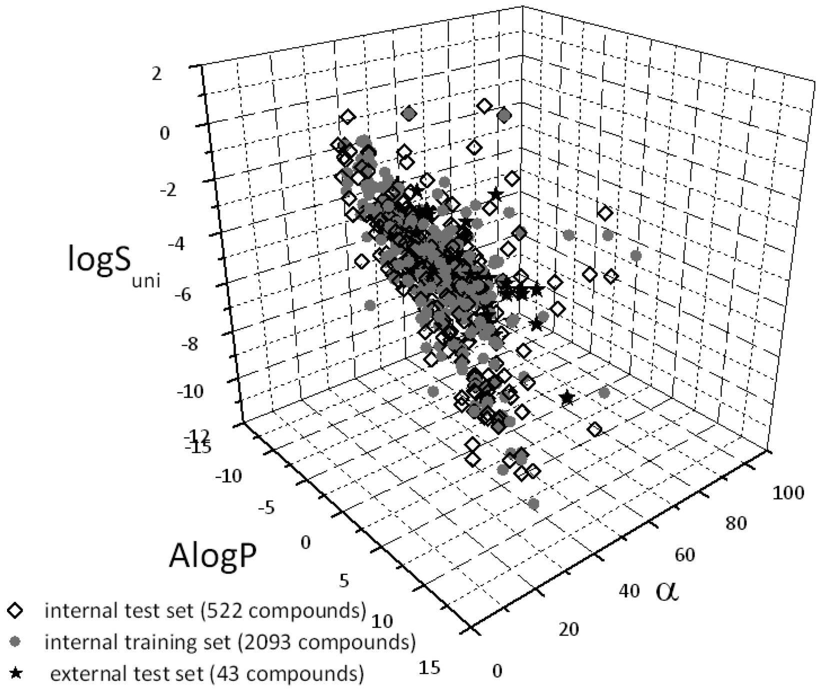


Figure 4. Un-ionized solubility distribution of internal training (gray circles) and test sets (white rhombs), external test set (stars) in molecular polarizability (α), and partition coefficient ($A \log P$).

calculation of solubility of organic chemicals especially compounds having "good" structural neighbors.

In conclusion, we note that our best models are based on molecular polarizability (α) and partition coefficient ($A \log P$) as independent variables. The first descriptor relates to steric interactions. Partition coefficient may be regarded as a composite descriptor relating to steric and H-bonding interactions.[26] Thus, these two types of interactions can be regarded as fundamental to the solubility of gaseous,[10] liquid,[11] and crystalline[12] compounds.

CONCLUSIONS

The aqueous solubility of crystalline chemicals has been calculated using a method combining concepts of structure similarity, "read-across", and multiple regression procedures as an alternative to a popular method using experimental values of two physicochemical properties and therefore having an essential shortcoming, namely, that to calculate the solubility of a compound one would need to synthesize it and measure its partition coefficient and melting point.

During our work, we tested models for a few subsets of chemicals having neighbors with different levels of similarity. Calculations on all chemicals (2093 chemicals in the training set and 522 chemicals in the test set) were performed without limitation of similarity level for neighbors ($T_c \geq 0.0$). For such subsets, standard errors were in the range from 0.85 to 0.90. The use of neighbor similarity on the level $T_c \geq 0.3$ reduced the numbers of chemicals involved in calculations but increased precision (sd from 0.79 to 0.80). At the level $T_c \geq 0.5$, the remaining chemicals were calculated with sd from 0.57 to 0.65 (that is, close to the error of experimental solubility determination). First, those data give a clear demonstration of the correctness of the principle that related chemicals have related properties. Second, those calculations allow the possibility of estimation of error of calculation for any new chemical on the base of the similarity level of its neighbors among the chemicals of the training set.

ASSOCIATED CONTENT

Supporting Information

Excel file with complete information about the 2615 compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: raevsky@ipac.ac.ru.

Present Address

§O. A. Raevsky: 142432, Russia, Chernogolovka, Severniy proezd 1.

Author Contributions

Prof. Dr. O. A. Raevsky is an author of considered approach and a head of all researches in this manuscript. Dr. V. Yu. Grigor'ev is an author of the homemade program AMP and LoReP. The last one was created especially for this research. Young researcher D. E. Polianczyk made all calculations in this work. Researcher O. E. Raevskaya collected solubility data and prepared training and test sets. Prof. J. C. Dearden participated in discussions of all aspects of this research and in preparation of the manuscript.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Lyman, W. J. Solubility in Water. In *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; American Chemical Society: Washington, DC, 1990; pp 2.1−2.52.

(2) Mackay, D. Solubility in Water. In *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*; Boethling, R. S., Mackay, D., Eds.; CRC Press: Boca Raton, FL, 2000; pp 125−139.

(3) Livingstone, D. J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003**, *3*, 1171−1192.

(4) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1*, 31−52.

(5) Dearden, J. C.; Rotureau, P.; Fayet, G. QSPR prediction of physicochemical properties for REACH. *SAR QSAR Environ. Res.* **2013**, *24*, 279−318.

(6) Grime, K. H.; Barton, P.; McGinnity, D. F. Application of in silico, in vitro and preclinical pharmacokinetic data for the effective and efficient prediction of human pharmacokinetics. *Mol. Pharmaceutics* **2013**, *10*, 1191−1206.

(7) Mannhold, R.; Ostermann, C. Prediction of Log *P* with Substructure-based Methods. In *Molecular Drug Properties: Measurement and Prediction*; Mannhold, R., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; pp 357−380.

(8) Tetko, I. V.; Poda, G. I. Prediction of Log *P* with Property-based Methods. In *Molecular Drug Properties: Measurement and Prediction*; Mannhold, R., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; pp 381−406.

(9) Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B. Revisiting the general solubility equation: In silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. *J. Chem. Inf. Model.* **2012**, *52*, 420−428.

(10) Raevsky, O. A.; Raevskaja, O. E.; Schaper, K.-J. Analysis of water solubility data on the basis of HYBOT descriptors. Part 1. Partitioning of volatile chemicals in the water−gas phase system. *QSAR Comb. Sci.* **2003**, *22*, 926−942.

(11) Schaper, K.-J.; Kunz, B.; Raevsky, O. A. Analysis of water solubility data on the basis of HYBOT descriptors. Part 2. Solubility of liquid chemicals and drugs. *QSAR Comb. Sci.* **2003**, *22*, 943−958.

(12) Raevsky, O. A.; Raevskaja, O. E.; Schaper, K.-J. Analysis of water solubility data on the basis of HYBOT descriptors. Part 3. Solubility of crystal neutral chemicals and drugs. *QSAR Comb. Sci.* **2004**, *23*, 327−343.

(13) Raevsky, O. A.; Grigor'ev, V. Y.; Trepalin, S. V. HYBOT program, registration by Russian State Patent Agency No. 990090 of 26.02.99.

(14) Raevsky, O. A. Molecular the partition coefficient calculations of chemically heterogeneous chemicals and drugs on the basis of structural similarity and physicochemical parameters. *SAR QSAR Environ. Res.* **2001**, *12*, 367−381.

(15) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* **2006**, *46*, 1984−1995.

(16) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local lazy regression: Making use of the neighbourhood to improve QSAR prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1836−1847.

(17) Trepalin, S. V.; Yarkov, A. V. CheD: Chemical database compilation tool, internet server, and client for SQL servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100−107.

(18) Gerasimenko, V. A.; Trepalin, S. V.; Raevsky, O. A. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, F. S., Eds; Kluwer Academic/Plenum Publishers: New York, 2000; pp 423−424.

(19) *DRAGON*, version 5.5; Talete srl: Milano, Italy, 2011.

(20) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.

(21) Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H. Estimation of aqueous solubility by the general solubility equation (GSE) the easy way. *QSAR Comb. Sci.* **2003**, *22*, 258−262.

(22) Raevsky, O. A.; Grigor'ev, V.Ju.; Modina, E. A.; Worth, A. P. Prediction of acute toxicity to mice by the arithmetic mean toxicity (AMT) modelling approach. *SAR QSAR Environ. Res.* **2010**, *21*, 265−275.

(23) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(24) Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ. Res.* **2004**, *15*, 19−32.

(25) Hansen, N. T.; Kouskoumvekaki, I.; Jørgensen, F. S.; Brunak, S.; Jónsdóttir, S. O. Prediction of pH-dependent aqueous solubility of druglike molecules. *J. Chem. Inf. Model.* **2012**, *46*, 2601−2609.

(26) Raevsky, O. A.; Schaper, K.-J.; van de Waterbeemd, H.; McFarland, J. W. Hydrogen Bond Contributions to Properties and Activities of Chemicals and Drugs. In *Molecular Modelling and Prediction of Bioactivity*; Gundertofe, K., Jorgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 221−227.