SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity

Oleg A. Raevsky,* Sergey V. Trepalin, Helen P. Trepalina, Vadim A. Gerasimenko, and Olga E. Raevskaja

Department of Computer-Aided Molecular Design, Institute of Physiologically Active Compounds, Russian Academy of Science, 142432, Chernogolovka, Moscow Region, Russia, and MOLPRO Project Group, 142432, Chernogolovka, Moscow Region, Russia

Received September 10, 2001

A new approach for predicting the lipophilicity (log P), solubility (log Sw), and oral absorption of drugs in humans (FA) is described. It is based on structural and physicochemical similarity and is realized in the software program SLIPPER-2001. Calculated and experimental values of log P, log Sw, and FA for 42 drugs were used to demonstrate the predictive power of the program. Reliable results were obtained for simple compounds, for complex chemicals, and for drugs. Thus, the principle of "similar compounds display similar properties" together with estimating incremental changes in properties by using differences in physicochemical parameters results in "structure - property " predictive models even in the absence of a precise understanding of the mechanisms involved.

INTRODUCTION

Creating and improving approaches to calculate compound properties such as lipophilicity, aqueous solubility, and permeability through biological membranes is important in the design of effective drugs. Central to this problem is ADMET (absorption, distribution, metabolism, elimination, and toxicity).

Lipophilicity is a fundamental property related to the ability of a drug to distribute itself throughout an organism. This property is usually defined as the partition coefficient (P) of the compound distributed between octanol and water phases and is commonly expressed as log P, its logarithmic form. There are many approaches to predict lipophilicity, and nearly as many corresponding computer programs to calculate partition coefficients.[1] Some of them are based on structural fragment methods, some on atom-based approaches, some on conformationally dependent approaches, and some on combined fragmental and atom-based approachs.[2] However, despite these many programs, lipophilicity is still difficult to calculate for complex organic chemicals and drugs.

In 1968, Hansch et al.[3] published a good correlation between the solubilities of liquid organic compounds in water and their corresponding lipophilicities (P_{oct}). For the solubilities of crystalline compounds there is the additional problem related to the effect of crystal lattice energy. To estimate the solubilities of such compounds there are several different approaches. Some of them use the entropy of fusion and melting point;[4−8] the others employ group contribution approaches[9−11] or neural networks.[12−15] The application of "mobil order theory" for estimating aqueous solubility is described in detail in the review.[16] Quantum-chemical descriptors have been used.[17] Monte Carlo simulation has been carried out also.[18] Abraham and Joelle presented an amended solvation energy relationship for the aqueous solubility of liquid and solids.[19]

Modern predictive models for intestinal drug absorption are also mainly based on physicochemical descriptors. For example, the "rule of five", which for potential drug molecules takes into account molecular weight, lipophilicity, and the number of hydrogen bond donor and acceptor groups.[20] The simplest model considers only the influence of molecular weight on absorption.[21] Recently, a more elaborate nonlinear six-descriptor neural network model was proposed.[22] Quantum-chemical descriptors have also been used[23] as well as a novel numerical molecular representation called the "molecular hashkey".[24] A nonaqueous partitioning system has been used to predict the oral absorption of peptides.[25] Drug liposome partitioning served as a tool to predict human passive intestinal absorption.[26] High-throughput permeability results have been correlated with gastrointestinal absorption in humans.[27,28] In recent years the dynamic polar molecular surface area (PSA_d) has been used as a predictor of drug absorption.[29−34] PSA is computed as the area of the van der Waals surface of oxygen and nitrogen atoms and of the hydrogen atoms attached to heteroatoms. Thus, PSA_d can be described as a number that is conformationally dependent on hydrogen bonding.

Despite the obvious progress in creating quantitative structure−activity relationship models for calculating lipophilicity, solubility, and drug absorption in humans, there remain significant problems, particularly for complex organic compounds and drugs that contain multiple functional groups. In our view, creating totally reliable quantitative predictive models based on general structure fragments or physicochemical descriptors is not likely because of widely varying structural features in different compound classes. One way to overcome this problem is to introduce indicator variables to take into account specific structural differences between

* Corresponding author phone/fax: (007) 095-785-70-24; e-mail: raevsky@ipac.ac.ru.

10.1021/ci010097o CCC: $22.00
Published on Web 03/12/2002

subclasses of related compounds, e.g. assign to an indicator variable the value of one (1) to members of a subset that contain, say, a hydroxy group at a certain structural site and a value of zero (0) to the others. Another way to create a general predictive model is to separate the compounds into distinct subsets and determine the QSAR within each group.[35,36]

The development of similarity concepts affords a new opportunity to create still further models to predict the physical properties of compounds. These concepts have been used to discover similar compounds in large libraries of compounds containing many different chemical classes, the aim being mainly to find compounds with similar biological activity. Willett et al. wrote a detailed review on various approaches to similarity searching.[37] By applying the principle of "similar compounds display similar properties", there is also the possibility to use similarity principles to predict physicochemical properties as well.[38−40]

The present paper describes this new approach as it is realized in the software program SLIPPER-2001. It predicts the lipophilicities, the solubilities, and the human oral absorptions of complex chemicals and drugs on the basis of physicochemical descriptors (including H-bond factors) and on structural and physicochemical similarity.

GENERAL DESCRIPTION AND CALCULATION METHOD

General Description of the Approach. QSARs are based on linear or nonlinear functional dependence (f) between a property (Pro) and descriptors characterizing the structure (S) of a compound in a training set:

$$\mathrm{Pro} = \mathrm{f(S)} \tag{1}$$

Such relationships work enough well in the case of relatively similar compounds where interpolation presents no problem. In sets where the compounds have significantly different structures, and in cases that require extrapolation, the predictive power of such models is decreased substantially.

Sometimes the similarity concept is used as an alternative to QSAR. For example, the observed property of the nearest neighbor (NN) is used as an estimate for a property of the compound-of-interest,[41,42] or the mean of the observed property for the K nearest neighbors can be used.

The approach presented in this work combines traditional QSAR with a structural similarity analysis to improve predictions. Formally, this combination can be represented as

$$\mathrm{Pro} = \mathrm{Pro_{nn}} + \Delta\mathrm{Pro} \tag{2}$$

where $\mathrm{Pro_{nn}}$ is the property value of the nearest neighbor, and ΔPro is the increment of the property related to the difference in descriptor values (Δ*S*) for the compound-of-interest and its neighbor. A calculation of ΔPro is carried out on the basis of the eq 3

$$\Delta\mathrm{Pro} = \mathrm{f}(\Delta S) \tag{3}$$

where the coefficient values of the independent descriptors are taken to be the same as those in eq 1. The required training sets must have simple structures so that there is a higher probability to correctly estimate the principal functional dependence between the property considered and the independent parameters. In using eq 2 to predict the properties of compounds containing several functional groups, all possible structural features must be taken into consideration.

From a statistical point of view, it is desirable to consider and to apply only a few nearest neighbors in this approach. In this case, the following equation uses a simple averaging scheme to weight the contributions of the nearest neighbors

$$\mathrm{Pro} = [\sum_{i=1}^{n}(\mathrm{Pro}\ i + \Delta\mathrm{Pro})]/n \tag{4}$$

where *n* is the number of nearest neighbors.

As to similarity, it is desirable to define the factors that influence the selection of nearest neighbors and thus the final results of property prediction. First let us consider molecular fragmentation procedures, then similarity types, similarity indexes, thresholds of similarity, and finally the optimum number of neighbors.

Molecular Fragmentation Procedure. In our work we use spherical fragments with different radii of the sphere around a chosen atom. The radius of the sphere is the topological distance between the central atom and the atom maximally remote from it; each bond length is assumed to be equal to 1. To divide the molecule into spherical fragments, the center on each atom in the molecule is chosen in turn. Then, fragments with spherical radii equal to 1, 2, etc. are taken into account. Further details of this procedure have been published.[43,44] It is obvious that the number and types of fragments depend on the sphere radius chosen. Table 1 contains similarity information about a drug (pholcodine) from the lipophilicity data set used in this work. This training set contains 10 937 chemicals and drugs. Where the sphere radius is equal to 1, this training set has 489 fragments; there are 140 170 fragments when the sphere radius is equal to 5. The types of fragments also depend on the sphere radius chosen. Where the sphere radius is equal to 1, only small fragments are obtained. With increasing spherical radii, ever-larger fragments are formed. So for the proper application of this approach, it is necessary to select a sphere radius that allows an optimal ratio between the number and the type of fragment. For example, we selected a sphere radius equal to 2 for the above-mentioned training set. In this case, the number of fragments (10 364) is approximately equal to the number of compounds in the training set (10 937). The fragments contain detailed information about the structures of the compounds. This choice also accords with the conclusion about preferences of this spherical radius for optimal fragmentation of organic compounds.[44]

Similarity Types and Indexes. In this work, Tanimoto (Tc), Euclidian (Ec), and Cosine (Cc) similarity indexes were used as quantitative measures of similarity

$$T_c = \mathrm{N(A\&B)/[N(A) + N(B) - N(A\&B)]} \tag{5}$$

$$E_c = \mathrm{1\text{-}SQRT\{[N(A) + N(B)\text{-}2*N(A\&B)]/2N(tot)\}} \tag{6}$$

$$C_c = \mathrm{N(A\&B)/SQRT[N(A)*N(B)]} \tag{7}$$

where N(A) is the number of fragments in molecule A, N(B)

Table 1. Similarity Indexes Values for the Nearest 10 Relative Structures of Pholcodine in the logP Data Base Containing 10 937 Compounds

		sphere radius													
		1 -- 489[a]			2 -- 10 364[a]			3 -- 47 707[a]			4 -- 100 840[a]			5 -- 140 170[a]		
NN	name	T_c	E_c	C_c	T_c	E_c	C_c	T_c	E_c	C_c	T_c	E_c	C_c	T_c	E_c	C_c
1	ethylmorphine	0.944	0.955	0.972	0.844	0.974	0.916	0.778	0.982	0.876	0.745	0.984	0.856	0.714	0.984	0.835
2	codeine	0.889	0.936	0.941	0.800	0.971	0.890	0.524	0.971	0.690	0.347	0.972	0.517	0.277	0.972	0.437
3	morphine	0.941	0.955	0.970	0.796	0.971	0.888	0.692	0.979	0.824	0.608	0.980	0.761	0.540	0.980	0.707
4	nalorphine	0.833	0.922	0.910	0.653	0.960	0.791	0.537	0.972	0.700	0.420	0.974	0.594	0.342	0.974	0.512
5	diacetyl morphine	0.700	0.889	0.824	0.640	0.958	0.781	0.554	0.972	0.714	0.466	0.975	0.637	0.399	0.975	0.571
6	drocode	0.833	0.922	0.910	0.633	0.958	0.776	0.494	0.970	0.664	0.361	0.972	0.533	0.288	0.972	0.449
7	myrophine	0.750	0.899	0.858	0.618	0.956	0.765	0.510	0.969	0.678	0.426	0.972	0.600	0.357	0.971	0.530
8	norcodeine	0.700	0.889	0.824	0.600	0.956	0.751	0.494	0.970	0.664	0.395	0.973	0.569	0.318	0.973	0.485
9	normorphine	0.737	0.899	0.849	0.592	0.956	0.746	0.458	0.969	0.632	0.336	0.972	0.507	0.273	0.972	0.434
10	nicomorphine	0.667	0.880	0.800	0.582	0.953	0.736	0.489	0.968	0.658	0.407	0.972	0.580	0.345	0.972	0.514

[a] Fragment number in the data base.

is the number of fragments in molecule B, N(A&B) is the number of common fragments in molecules A and B, and N(tot) is the total number of fragments in the database. The MOLDIVS program package was used to calculate similarity.[45,46]

Similarity index values for the 10 nearest neighbors of pholcodine are presented in Table 1. For all evaluated spherical radii and for all evaluated indexes, the first nearest neighbor for pholcodine is the same, ethylmorphine. And in the case of spherical radius equal to 2 the order of all nearest neighbors is the same for all evaluated indexes. For our work, we used a spherical radius equal to 2 and the Cosine coefficient as a quantitative measure of similarity.

Threshold of Similarity. There are not any strict rules to determine thresholds between similar and dissimilar structures. Our experience suggests that similar compounds have Tanimoto indexes of at least 0.54. The corresponding value for the Cosine index is 0.70. These thresholds conform to the presence of seven common fragments in a pair of molecules with 10 fragments in each of its partners or of 70 common fragments in a pair of complex compounds with 100 fragments in each of its partners. The ranges of values in the Euclidian index are not useful in this study because they are too narrow and because they are too close to 1.000 to make reliable distinctions. For this reason this index was not considered further in the current work.

Optimum Number of Neighbors. From the statistical point of view, it is desirable to use the maximum number of neighbors with similarity index values above the desired thresholds. However, in the case of large databases the number of such compounds may be too many. In contrast, there may be only a few similar molecules in small training sets. Our experience shows that using only the first nearest neighbor is not enough for accurate predictions. On the other hand, it is possible to obtain satisfactory results using only three nearest neighbors. We will return to this problem in further detail in the Results and Discussion section. Here it is only necessary to indicate that we automatically limit the calculation procedure in our program SLIPPER-2001 except of special studies to five nearest neighbors in using the lipophilicity database (training set has 10 937 compounds), by three nearest neighbors for solubility prediction (training set has 1502 compounds), and by two nearest neighbors for absorption prediction (training set has only 257 drugs). In cases where the compound-of-interest has no related structures ($C_c < 0.70$) in the training set for the above-mentioned properties, we predicted its value directly from eq 1. There is a higher probability of getting poor predictions where dissimilar partners are used.

Of course, the similarity concept is very complex and includes many other factors that can influence the results. For example, we estimated that in cases where the similarity indexes values are close to each other, a better prediction is realized where the neighbors have descriptor values closer to compound-of-interest. There are at least two reasons for this effect. The smaller the differences between the descriptor values for compound-of-interest and its neighbors, the smaller the errors in the estimates. At the same time, the close physicochemical descriptor values between those of compound-of-interest and its neighbors reflect their "physicochemical similarity".

The Scheme of SLIPPER-2001 Program Package. SLIPPER-2001 is designed to run under the Windows-95, -98, -2000 and Windows NT operating systems using the chemical database shell CheD.[43] This shell affords an opportunity to take advantage of its unique system to manipulate databases and to apply our methods for predicting lipophilicity, solubility, and human drug absorption on the basis of compound structural similarity and physicochemical properties.

To calculate log P, log Sw, and FA in SLIPPER-2001 the following scheme is used:

- For any structure, the search for nearest neighbors is carried out on the basis of Cosine similarity indexes by means of a similarity search algorithm[44] either in the internal databases of the SLIPPER-2001 (the database on log P contains about 10 937 entries; on log Sw, about 1502 entries; and on FA, 257 entries) or in the user's own databases. The number of nearest neighbors can be chosen to be anywhere from 1 to 10.
- Lipophilicity and solubility pH-dependent profiles are calculated for ionizable compounds from their calculated pK_a values.
- The results can be printed or saved as an Excel file (in case of a single structure) or saved as an SDF file (in the case of many compounds).

Figure 1 presents the algorithmic scheme for SLIPPER-2001. As an example, the protocol used to calculate the properties of imipramine is presented in Table 2. The computed pH-dependent profiles of logP, logS, and FA for this drug and the results for predicting its properties from a MOL file are presented in Figure 2.

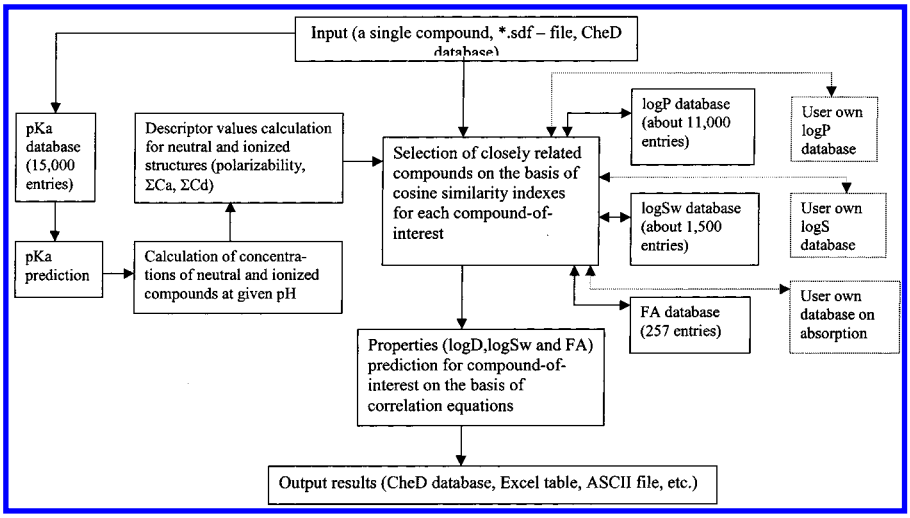

Figure 1. Block-scheme of SLIPPER-2001 program package.

Table 2. Protocol of Properties Calculations for Imipramine by Means of Slipper-2001

	neighbor	name	C_c	logP (exp)	logS (exp)	FA (exp)	logP (eq 14, 1 nn)[a]	logS (eq 16, 1 nn)[a]	FA (eq 19, 1 nn)[a]
logP	first	chlorimipramine	0.8898	5.19			4.735		
	second	depramine	0.8721	4.26			4.470		
	third	ketipramine	0.8611	4.24			5.033		
logS	first	desimpramine	0.8315		−3.66			−4.866	
	second	promazine	0.7182		−4.30			−4.762	
	third	amitriptyline	0.7064		−4.46			−3.846	
FA	first	desimpramine	0.9366			0.99			0.997
	second	diphenhydramine	0.9012			0.85			0.890
	third	carbamazepine	0.8928			0.70			0.873
							4.746 (eq 14, 3 nn)[b]	−4.491 (eq 16, 3 nn)[b]	0.920 (eq 19, 3 nn)[b]
							4.800[c]	−4.190[c]	1.000[c]

[a] Predicted properties values for imipramine. [b] Mean calculated values for imipramine. [c] Experimental values for imipramine.

RESULTS AND DISCUSSION

1. Lipophilicity. Fifteen years ago it was proposed[47,48] that the partition coefficient (P) encodes two major structural contributions that, mathematically stated, results in a volume-related term (describing bulk effects) and a term that reflects intermolecular interactions such as hydrogen bonding. At the time of this proposal, the assumptions could not be verified because there was no way to quantify hydrogen-bonding strength. Since then, we systematically investigated the thermodynamics of hydrogen bonding.[49] This allowed us to create HYBOT (Hydrogen Bond Thermodynamics), an original program to calculate quantitative hydrogen bond descriptors.[50−52] HYBOT in turn allowed us to develop a concept to quantitatively describe solubility, lipophilicity, and permeability of compounds based on volume-related, steric terms and hydrogen bond strength.

In 1995, an equation was published[53] that took into account molecular volume (MV) and hydrogen bonding to describe the partition coefficients for 38 *neutral* carbonyl and hydroxyl compounds in the octanol/water system. For the hydrogen bonding part, the authors used ΣC_a and C_d (HYBOT descriptors) where the former term is the sum of H-bond acceptor factors for the molecule and the latter, the donor factor for the molecule (only one or no donor per molecule was used in this study). On the basis of this equation it was possible to calculate the contribution of each term to log P. Work with other classes of compounds showed that when MV and ΣC_a were used to correlate with log P, there were nonzero values for the intercepts in the equations. For this reason, different volume-related terms were evaluated along with hydrogen bond descriptors in training sets containing different types of compounds (nitriles, amines, carbonyl compounds, ethers, esters, alcohols, phenols, compounds with phenyl, nitro and halogen groups, acids). Five volume-related terms (molecular weight (MW), surface area (SA), molecular volume (MV), molecular refractivity (MR), and polarizability (α)) were used together with $C_{a\,max}$, ΣC_a, ΣC_a/MW, $C_{d\,max}$, ΣC_d, and ΣC_d/MW. Only one volume-related term at-a-time could be used among the aforementioned because each such term was highly correlated with the others.[54] The best results were obtained by using polarizability (α) and hydrogen bond acceptor strength (ΣC_a). Equation 8 is the result obtained for a large set of *simple* neutral compounds:

$$\log P = 0.267\alpha - 1.00\Sigma Ca \qquad (8)$$

$$n = 2850,\ r = 0.970,\ s = 0.23$$

This equation is a mathematical model for the distributions of compounds in the system octanol/water. In the hypothetical cases where both polarizability and hydrogen bond factors are equal to zero and where $0.267\alpha = 1.00\ \Sigma C_a$, log P = 0.00 which implies that the solute is equally distributed in both solvent phases. Because polarizability can have only positive values, it can only make a positive contribution to logP and hence lead to larger concentrations of the solute in the octanol phase. Similarly, hydrogen bond acceptor factors

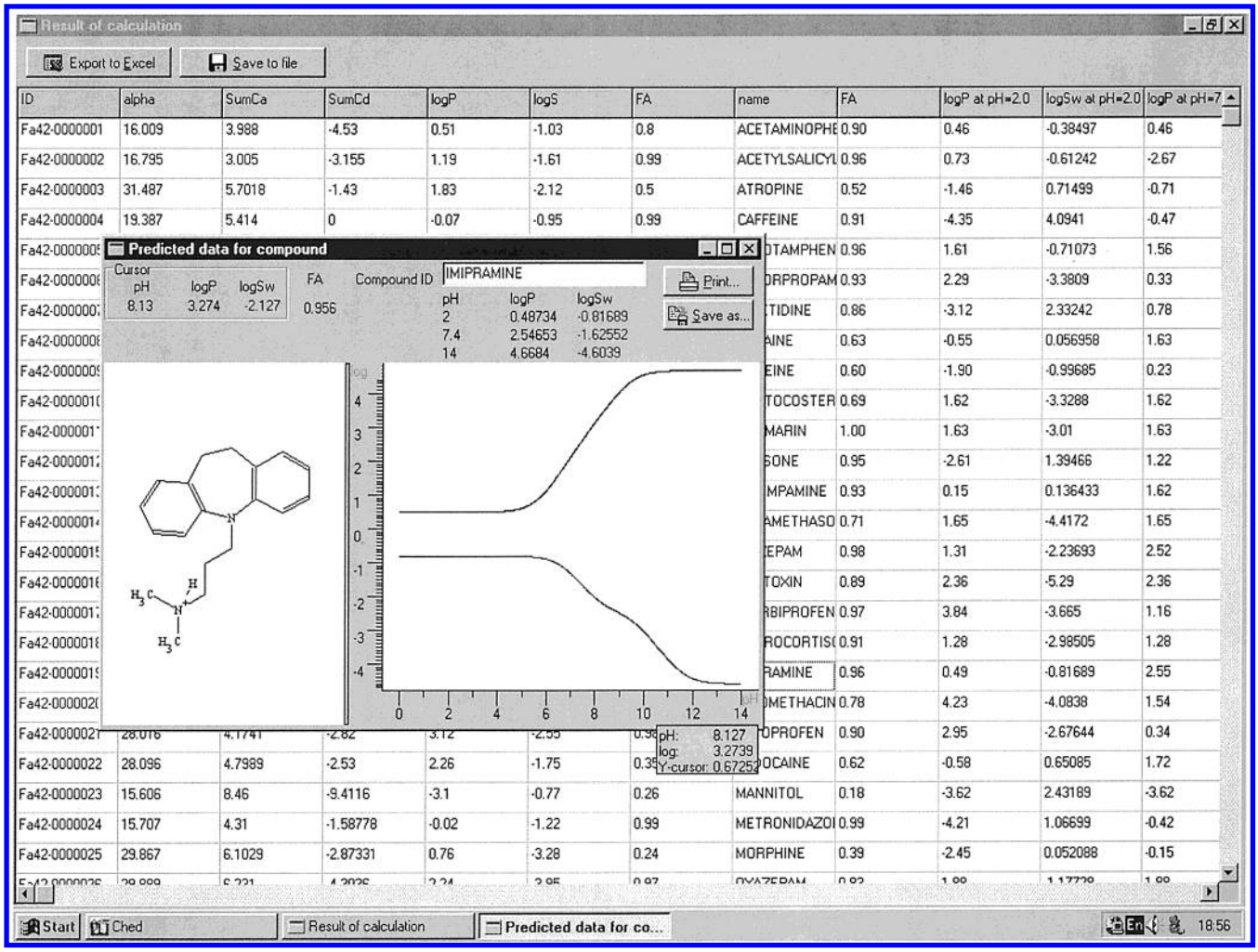

Figure 2. Calculation from SDF file performed by SLIPPER-2001. For any compound the user may get full pH dependent lipophilicity and solubility profiles.

can only be positive, but because its term in eq 8 is preceded by a negative sign, H-bond acceptors can only increase the solute's distribution in the water phase. We suggested analogous relationships also for ionized structures.[54]

In solution, ionizable compounds exist in equilibria with neutral and ionized species. Their contributions are taken into account by calculating an effective (or apparent) partition coefficient (D)[55]

$$\log D = \log \sum_{i=1}^{n} fiPi \qquad (9)$$

where *fi* is the fraction of a component *i*.

In 1998, on the basis of these equations, we created the program SLIPPER-98 that calculates lipophilicity[56,57] including the logD-pH profile.

To check the results of our lipophilicity calculations we carried out log P predictions for the 138 compounds used by Mannhold and Dross.[1] The results comparing the experimental (log P_{exp}) to the calculated (log P_{calc}) values for the entire database are given by the following regression equation, forced through the origin:[54]

$$\log P_{exp} = 0.950(\pm 0.009)\ \log P_{calc} \qquad (10)$$

$$n = 138,\ r = 0.970,\ msd = 0.152,\ s = 0.369,\ F = 2148$$

For the subset of 90 simple organic compounds:

$$\log P_{exp} = 0.960(\pm 0.009)\ \log P_{calc} \qquad (11)$$

$$n = 90,\ r = 0.980,\ msd = 0.093,\ s = 0.289,\ F = 2106$$

and for the subset of 48 drugs:

$$\log P_{exp} = 0.935(\pm 0.019)\ \log P_{calc} \qquad (12)$$

$$n = 48,\ r = 0.955,\ msd = 0.262,\ s = 0.480,\ F = 478$$

Despite the good statistical criteria, detailed analyses of these results show that the problem of predicting lipophilicity is especially sensitive for organic compounds and drugs containing several functional groups. It is apparent that this is associated with additional intramolecular interactions that were not taken into account by the simple additivity schemes employed.

To improve the calculation procedure for lipophilicity, a new approach based on structural similarity and physicochemical parameters was recently suggested.[58] In this approach, the partition coefficient value of a nearest neighbor in the training set is used as a first approximation of the lipophilicity of the compound of interest. Further, the differences in the polarizability values and in the sums of

H-bond acceptor factors between this compound and its nearest neighbor are taken into consideration:

$$\log P_i = \log P_{nn} + 0.267(\alpha_i - \alpha_{nn}) - 1.00(\sum Ca_i - \sum Ca_{nn}) \quad (13)$$

From a statistical point of view, it is desirable to use several nearest neighbors for property prediction. In analyzing the prediction results using eq 13 for different neighbors, we could not discern any dependence between the order of neighbors on a similarity scale (first, second, and so on) and the accuracy of prediction. Therefore, a simple averaging scheme to weight the nearest neighbors' values was used to predict in the case several neighbors:

$$\log P = \sum_{j=1}^{N} [\log Pj + 0.267(\alpha - \alpha j) - 1.00(\sum Ca - \sum Caj)\ ]/N \quad (14)$$

where N is the of number nearest neighbors employed.

On the basis of eq 14, log P values for the previously mentioned 138 compounds[1] were calculated by using a database containing experimental lipophilicity values for about 7000 compounds.[58] This significantly improved the statistical criteria for the correlation between experimental and calculated values. This method gave results that were better than any found by Mannhold and Dross.[1] Compounds that gave acceptable values (Δ = logPexp − logPcalc < |0.5|) were 37; those with marginal values ($|0.5| \leq \Delta \leq |1.0|$) were 10; and those with unacceptable values ($\Delta > |1.0|$) were 1 (one). When eq 8 was used the corresponding numbers were 31, 13, and 4. These results demonstrate the advantages of a new approach as compared to previous methods.

As a result we created a new version of SLIPPER (now called SLIPPER-2001 and presented in this publication) to predict lipophilicity, solubility in water, and absorption in humans.

Because we significantly increased the training set for lipophilicity to 10 937 compounds, an additional test set of compounds was studied to verify the lipophilicity calculation procedure. Here, we used 42 complex drugs with several functional groups for which there were experimental values of log P, log Sw, and FA in the three training sets of SLIPPER-2001.

Table 3 summarizes the prediction results for those drugs. Calculations were carried out for 1−10 nearest neighbors. However, careful examination of the results showed that three closely related structures were enough to get good lipophilicity predictions. The statistical results for 1−3 nearest neighbors are presented in Table 4. For lipophilicity, the number of acceptable compounds was 35 (83.33%), marginal compounds were 7 (16.67%), and there were no unacceptable compounds. Thus, it can be concluded that SLIPPER-2001 predicts lipophilicity from eq 14 very well. The correlation coefficient between experimental and calculated logP values on the basis of eq 14 for the test set of 42 drugs is 0.965 already in the case of the use of three nearest neighbors. This is a little less than the correlation coefficient found between experimental and calculated log P for all 10 937 compounds of the training set (0.972). The graphical comparison of experimental and calculated logP values for all 10 937 chemicals and drugs in the case of the use of eqs 8 and 14 (three nearest neighbors) is presented on Figure 3.

In particular, we also found that where the compound-of-interest did not have any neighbors in the training set with Cosine coefficient values above 0.70, the use of eqs 13 and 14 did not give good predictive results. In these cases, it is better to use the direct correlation of logP from physicochemical parameters, eq 8. In contrast, where the compound-of interest has closely related compounds in the training set, the use of eqs 13 and 14 gives especially good results; under these same circumstances the direct correlation of logP with physicochemical parameters cannot ensure satisfactory predictions. Structural features of the compound-of-interest that are not taken into account by direct correlation can explain this. In the framework of the present approach, using experimental logP values of closely related compounds allows us to consider these structural features in a hidden form.

We also analyzed the influence of the number of neighbors on predictive power. Table 5 contains the results of predicting logP for two compounds, salmefanol and dagapamil, using up to 10 NN with eqs 13 and 14, and three dissimilar compounds. In the both cases eq 8 gave unsatisfactory results. Using the average logP values for the five nearest neighbors in eq 14 allows us to predict logP correctly within the probable experimental error of measurement. The use of dissimilar compounds (the bottom three compounds of Table 5) gave unsatisfactory results.

2. Aqueous Solubility. The advantage of using HYBOT descriptors for predicting water solubility has been described in several articles. McFarland et al. obtained a satisfactory correlation for 22 crystalline drugs.[59] Using 45 neutral polar compounds, Raevsky et al.[51] showed that besides polarizability, hydrogen bond acceptor factors, and hydrogen bond donor factors also influenced water solubility. Later,[60] the training set was increased to 142 compounds. Recently, the same descriptors were used to compute the solubility of 493 simple liquid organic compounds of diverse structures:[61]

$$\log S = 0.42 - 0.275\alpha + 0.96\sum C_a - 0.27\sum C_d \quad (15)$$

$$n = 493,\ r = 0.966,\ s = 0.35,\ R_{cv} = 0.965$$

A comparison of eqs 8 and 15, for lipophilicity and for solubility, respectively, shows that the magnitudes of the coefficients for the polarizability terms are approximately equal. The same is true for the hydrogen bond acceptor term. However, those coefficients in eqs 8 and 15 have opposite signs. Thus, as a first approximation, the solubility of a liquid compounds can be considered as the reciprocal of the corresponding lipophilicity. An essential difference between these equations is the additional contribution to solubility made by the sum of hydrogen bond donor factors.

The use of eq 15 to predict the solubility of crystalline compounds was not successful. The insertion of the cross term $\sum C_a \cdot \sum C_d$[62] did not improve matters. There are at least two reasons for these failures: (i) intramolecular interactions are not adequately described by these schemes and (ii) the energy of crystal lattice in the dissolution process is not taken into account.

We tried to take the influence of these factors into account by using the solubilities of closely related compounds and

Table 3. Calculations Results of Properties Prediction for 42 Drugs

					$logP_{calc}$			logP	$logS_{calc}$			logS	FA_{calc}			FA
NN	name	α	$\sum C_a$	$\sum C_d$	1 nn	2 nn	3 nn	exp	1 nn	2 nn	3 nn	exp	1 nn	2 nn	3 nn	exp
1	acetaminophen	16	4	−4.5	0.730	0.435	0.329	0.51	−0.818	−0.885	−0.864	−1.03	0.921	0.932	0.832	0.8
2	acetylsalicylic acid	16.8	3	−3.2	0.087	0.522	1.003	1.19	−1.201	−1.265	−1.278	−1.61	0.986	0.95	0.963	0.99
3	atropine	31.5	5.7	−1.4	1.619	1.496	1.551	1.83	−1.257	−1.9	−2.335	−1.93	0.947	0.557	0.532	0.5
4	caffeine	19.4	5.4	0	−0.257	−0.103	0.041	−0.07	−0.418	−0.375	−0.321	−0.95	0.988	0.877	0.917	0.99
5	chloramphenicol	28.2	6.2	−5.5	2.136	1.845	1.859	1.14	−0.709	−1.262	−1.829	−1.9	0.976	0.978	0.957	0.9
6	chlorpropamide	26.6	5.3	−3.7	2.351	2.3	2.124	2.27	−2.911	−2.828	−2.841	−3	0.92	0.946	0.939	0.9
7	cimetidine	27.4	5.1	−5.8	0.716	1.123	0.985	0.4	−0.75	−1.064	−2.007	−1.55	0.99	0.872	0.897	0.85
8	cocaine	31.3	4.7	0	2.796	2.594	2.703	2.3	−1.804	−1.843	−2.93	−2.278	0.444	0.657	0.735	0.57
9	codeine	31.7	6	−1.5	1.317	1.084	1.053	1.14	−1.555	−1.621	−1.81	−1.57	0.462	0.597	0.44	0.5
10	corticosterone	37.6	6.8	−3.3	1.484	1.459	2.047	1.94	−1.657	−3.353	−3.262	−3.24	0.993	0.996	0.99	0.99
11	coumarin	15.4	2.2	0	2.111	1.75	1.864	1.39	−2.528	−2.403	−2.426	−1.8	0.999	0.999	0.99	0.99
12	dapsone	26.9	5.8	−6.5	1.213	1.344	1.393	0.97	−3.571	−3.528	−3.302	−2.8	0.986	0.966	0.953	0.93
13	desimpramine	33.2	4	−2	3.967	3.972	4.078	4.9	−2.984	−3.245	−3.674	−3.66	0.996	0.896	0.808	0.99
14	dexamethasone	39.8	8.1	−5.8	0.883	1.567	1.724	1.83	−3.465	−3.577	−3.288	−3.77	0.982	0.938	0.872	0.98
15	diazepam	31.1	5.5	0	2.618	2.377	2.42	2.8	−3.614	−3.90	−3.911	−3.76	0.98	0.989	0.992	0.97
16	digitoxin	77.1	16.5	−7.7	2.644	2.668	2.176	1.74	−5.67	−5.045	−4.763	−5.29	0.962	0.887	0.883	0.9
17	flurbiprofen	25.9	2.9	−2.8	3.577	3.682	3.754	4.16	−3.209	−3.454	−3.441	−4.5	0.994	0.988	0.971	0.92
18	hydrocortisone	38.2	8.6	−5.1	1.073	1.117	1.366	1.61	−2.746	−2.993	−2.659	−2.97	0.981	0.858	0.761	0.89
19	imipramine	35	3.9	0	4.735	4.603	4.737	4.8	−4.762	−4.304	−4.567	−4.19	0.997	0.935	0.92	0.999
20	indomethacin	37.3	5.5	−2.5	4.759	4.211	4.389	4.27	−5.33	−4.138	−4.915	−4.48	0.351	0.662	0.733	0.98
21	ketoprofen	28	4.2	−2.8	2.98	2.868	2.785	3.12	−3.039	−2.897	−3.212	−2.55	0.945	0.864	0.902	0.98
22	lidocaine	28.1	4.8	−2.5	1.636	2.384	2.336	2.26	−1.24	−0.974	−1.626	−1.75	0.516	0.436	0.427	0.35
23	mannitol	15.6	8.5	−9.4	−3.594	−3.88	−4.045	−3.1	−1.062	−2.005	−0.676	−0.03	0.02	0.026	0.222	0.26
24	metronidazole	15.7	4.3	−1.6	−1.169	−0.529	−0.486	−0.02	−2.009	−0.928	−0.411	−1.22	0.981	0.991	0.977	0.99
25	morphine	29.9	6.1	−2.9	0.583	0.439	0.668	0.76	−4.209	−3.314	−2.793	−3.28	0.269	0.173	0.282	0.24
26	oxazepam	29.9	6.2	−4.3	1.939	1.699	1.881	2.24	−3.695	−3.691	−3.792	−3.95	0.925	0.76	0.818	0.97
27	phenazone	21.8	3.6	0	0.21	0.719	1.307	0.38	2.681	1.16	0.028	0.43	0.999	0.995	0.961	0.97
28	phenobarbital	24.1	5	−3.4	2.309	1.804	1.649	1.47	−2.413	−2.155	−1.965	−2.33	0.817	0.792	0.857	0.99
29	phenylbutazone	35.4	5.9	0	3.558	3.025	2.795	3.16	−1.98	−2.503	−3.355	−3.8	0.998	0.929	0.839	0.9
30	phenitoin	28.1	5.8	−3.7	2.286	2.312	2.265	2.47	−3.228	−4.077	−3.991	−3.99	0.898	0.894	0.813	0.9
31	prednisolone	38	8.1	−5.1	0.173	0.932	1.311	1.62	−2.425	−2.915	−3.105	−3.18	0.813	0.87	0.805	0.988
32	progesterone	36.3	4.2	0	3.756	3.684	3.725	3.87	−4.722	−4.647	−4.673	−4.42	0.989	0.774	0.842	0.91
33	quinidine	37.7	6.7	−1.5	3.39	2.664	2.838	3.44	−3.15	−2.791	−3.263	−3.12	0.735	0.781	0.837	0.8
34	salicylic acid	13	1.4	−4.8	1.818	1.923	2.069	2.26	−0.141	−1.166	−1.599	−1.89	0.993	0.965	0.967	0.99
35	spironolactone	44.2	5.7	0	4.323	2.982	2.862	2.78	−5.298	−5.256	−5.263	−4.3	0.728	0.825	0.57	0.25
36	sulfadiazine	25.5	7.2	−5.3	−0.174	−0.146	0.239	−0.09	−2.491	−2.82	−2.577	−3.4	0.921	0.909	0.906	0.98
37	sulfamethoxazole	24.7	5.6	−5.3	−0.339	0.85	0.642	0.89	−3.365	−2.711	−2.792	−2.64	0.973	0.984	0.981	0.98
38	sulfisoxazole	26.6	6.1	−5.3	2.648	1.148	1.022	1.01	−3.429	−3.067	−2.752	−3.02	0.97	0.982	0.975	0.96
39	testosterone	33.1	3.8	−1.4	3.416	3.173	3.111	3.32	−3.477	−3.487	−4.104	−4.08	0.844	0.902	0.912	0.98
40	theophylline	17.6	5.2	−2.1	0.42	0.313	0.304	−0.02	−2.759	−1.991	−1.775	−1.36	0.967	0.590	0.725	0.96
41	tolbutamide	28.4	5.1	−3.7	2.057	2.21	2.454	2.34	−3.639	−3.511	−3.510	−3.55	0.912	0.944	0.94	0.93
42	triamcinolone acetonide	43.3	9	−3.9	1.782	1.759	1.909	2.53	−2.483	−3.361	−3.647	−4.31	0.142	0.418	0.315	0.23

by calculating the solubility increment of the compound-of-interest in comparison to the solubilities of the nearest neighbors. Calculation of the increments was based on differences in the polarizabilities, the hydrogen bond acceptor and hydrogen bond donor factors, and the by use of coefficient values for those descriptors from eq 15

$$\log Sw = \sum_{j=1}^{N}[\log Sj + 0.275(\alpha - \alpha j) + 0.96(\sum Ca - \sum Caj) - 0.27(\sum Cd - \sum Cdj)]/N \quad (16)$$

The results obtained for the solubilities crystalline compounds and drugs by means of this equation show the important role that hydrogen donor factors ($\sum C_d$) value play in predicting solubilities. The use of neighbors with small differences in $\sum C_d$ with respect to the compound-of-interest ensures good predictions. So it is not unreasonable to assume that hydrogen bond donors make important contributions to the crystal lattice energy.

Table 6 presents solubility values for 1,2,3,4,6,7,8-heptachlorodibenzo-p-dioxane calculated by eqs 15 and 16. The large difference between the experimental value and that calculated by eq 15 shows that the linear scheme is inadequate. It is likely that the main contribution to the solubility of this compound is the influence of crystal lattice energy, a factor that can change this property a 1000-fold or more. As shown by the other information in Table 6, the nearest neighbor approach takes this effect into account in a hidden form by using the properties of closely related structures.

Table 4. Statistic Criteria Linear Correlation of Experimental and Calculated Property Values $\log Y_{exp} = k \log Y_{calcc}$

property	no. of nearest neighbors	linear coeff k (±)	no. of compds n	correlation coeff r	SD s
logP	1	0.943 (±0.043)	42	0.895	0.68
	2	1.033 (±0.030)	42	0.957	0.44
	3	1.014 (±0.026)	42	0.965	0.40
logS	1	0.972(±0.048)	42	0.706	0.93
	2	1.017(±0.032)	42	0.887	0.61
	3	0.994(±0.024)	42	0.931	0.48
FA	1	0.975(±0.029)	42	0.745	0.16
	2	1.013(±0.026)	42	0.796	0.15
	3	1.035(±0.019)	42	0.908	0.10

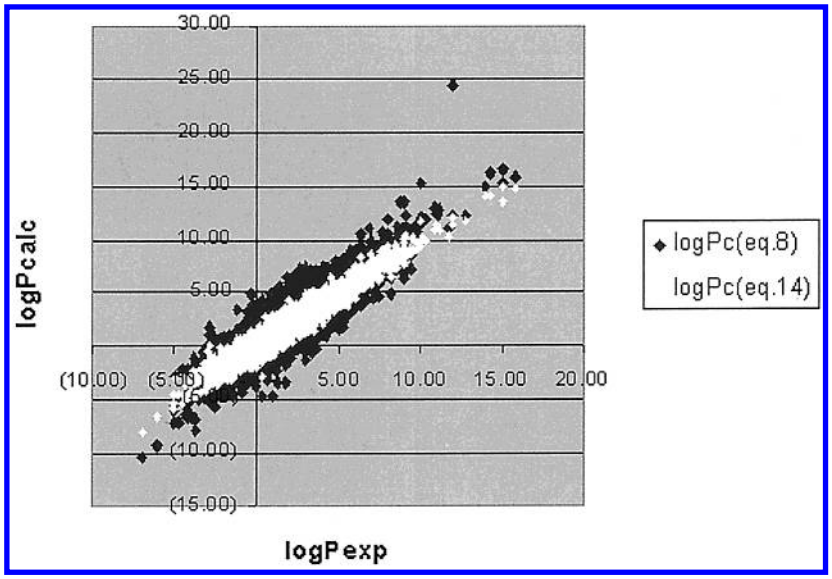

Figure 3. The graphical comparison of experimental and calculated logP values for 10 937 chemicals and drugs. Dark points correspond to values calculated by eq 8 and light points correspond to eq 14.

Table 5. logP Prediction Results for Two Compounds on the Basis of Eqs 8, 13, and 14

NN nn	compd name	C_c	α	∑Ca	exp	logP eq 8	salmefanol calc eq 13	salmefanol calc eq 14	compd name	C_c	α	∑Ca	exp	logP eq 8	dagapamil calc eq 13	dagapamil calc eq 14
	salmefanol	1.0000	36.11	7.55	**1.05**	2.09			dagapamil	1.0000	66.61	7.56	**8.00**	10.22		
1	naminterol	0.8807	37.32	7.39	0.49		0.01	0.01	anipamil	0.9718	62.39	6.15	9.00		8.72	8.72
2	formoterol	0.8111	37.11	8.04	1.11		1.33	0.67	devapamil	0.8733	47.98	7.04	4.00		8.45	8.59
3	etanterol	0.8056	35.48	7.55	−0.01		0.16	0.50	gallopamil	0.8575	52.50	8.47	3.01		7.77	8.31
4	fenoterol	0.7874	32.08	6.56	0.83		0.92	0.61	ronipamil	0.8498	58.17	5.07	9.50		9.26	8.55
5	metiprenaline	0.7737	23.98	5.14	0.57		1.40	0.76	verapamil	0.8418	50.09	7.94	3.79		8.82	8.60
6	butopamine	0.7698	33.64	6.10	2.03		1.24	0.84	emopamil	0.7191	41.66	5.09	4.63		8.72	8.62
7	protokylol	0.7660	33.41	7.05	1.36		1.58	0.95	symetine	0.6667	56.37	6.53	8.02		9.72	8.78
8	sulfinalol	0.7642	40.65	8.58	1.71		1.53	1.02	tetrandrine	0.6586	67.13	10.20	8.40		10.90	9.05
9	albuterol	0.7579	26.18	6.36	0.11		1.57	1.08	trepipam	0.6537	34.06	4.76	3.18		9.07	9.05
10	fenalcomine	0.7366	37.04	5.12	3.62		0.94	1.07	famipamil	0.6499	45.42	8.66	1.16		8.81	9.02
	$1,4(OCH_3)_2$-C_6H_4	0.5000	14.65	1.89	2.03		2.10		AF-2259	0.5000	35.48	3.20	5.13		9.92	
	recainame	0.4000	31.41	4.64	1.13		−0.53		cinoxate	0.4000	26.26	4.32	2.72		9.09	
	quazoline	0.3000	24.17	4.54	1.98		2.16		amineptine	0.3000	39.45	4.68	4.74		9.62	

Table 6. Physicochemical Descriptors and Experimental and Calculated Values logS Values for 1,2,3,4,6,7,8-Heptachlorodibenzo-p-dioxine on the Basis of Direct Calculation by Eq 15 and by Structural and Physicochemical Similarity for Closely Related Structures: 1,2,3,4,7,8-Hexachlorodibenzo-p-dioxine (First Nearest Neighbor), 1,2,3,4,7,8-Hexachlorodibenzo-p-dioxine (Second Nearest Neighbor), and 1,2,3,7-Tetrachlorodibenzo-p-dioxine (Third Nearest Neighbor)

name	α	ΣC_a	ΣC_d	$\log S_{w\ exp}$	$\log S_{w\ calc}$ for 1,2,3,4,6,7,8-heptachlorodibenzo-p-dioxine
1,2,3,4,6,7,8-heptachlorodibenzo-p-dioxine	33.4	2.3	0.0	−11.250	−6.557 [eq 15]
1,2,3,4,7,8- hexachlorodibenzo-p-dioxine	31.4	2.3	0.0	−10.950	−11.474 [eq 16], 1 nearest neighbor)
1,2,3,7-tetrachlorodibenzo-p-dioxine	27.6	2.4	0.0	−8.890	−11.041 [(eq 16], 2 nearest neighbors)
TCDD	27.6	2.3	0.0	−10.220	−11.302 [eq 16], 3 nearest neighbors)

Table 3 contains the results of solubility calculations based on eq 16 for 42 drugs using from 1−3 nearest neighbors. Despite the fact that the database on solubility was limited about 1500 compounds, the results are quite respectable. Table 4 gives the results of the statistical analysis correlating the experiment values with those calculated. The number of compounds with acceptable computed values was 29 (69.05%); marginal, 12 (28.57%); and unacceptable, 1 (2.38%).

It is realistic to expect that the predictivity of the nearest neighbor model, epitomized by eq 16, will be improved as the relevant databases of experimental values grow and physicochemical similarity models develop further.

3. Intestinal Drug Absorption in Human. We recently published a stable nonlinear model that, on the basis of physicochemical parameters, quantitatively estimated drug absorption in humans for passively transported compounds.[63] Using only a descriptor that characterizes the total ability of a compound to form hydrogen bonds (ΣC_{ad}) led to good results:

$$FA = 1/(1 + 10^{-[5.30(\pm 1.86) - 0.33(\pm 0.12)\Sigma C_{ad}]}) \quad (17)$$

$$n = 31,\ R = 0.947,\ s = 0.12,\ R_{cv} = 0.923$$

This demonstrates the great importance of hydrogen bonding in the absorption and transport of chemicals and drugs in humans. When the composite descriptor $\sum C_{ad}$ was

separated into its parts an even better result was obtained:

$$FA = 1/(1 + 10^{-[5.47(\pm1.50) - 0.39(\pm0.10)\Sigma C_a + 0.29(\pm0.10)\Sigma C_d]}) \quad (18)$$

$$n = 31,\ R = 0.978,\ s = 0.09,\ R_{cv} = 0.957$$

Later we attempted[64] to use eqs 17 and 18 directly to predict the human intestinal absorption of 100 drugs not used in the original study. This was not successful. Of the 100, the absorptions of 23 drugs were incorrectly predicted.

Such disappointing results could be related to the presence of drugs with different mechanisms of transport in the test set. The most commonly recognized method of drug permeation is passive diffusion. Other methods of transcellular transport include carrier-mediated and vesicular transport mechanisms. Paracellular transport, i.e., the aqueous extracellular route, and active efflux and active influx systems also play important roles.

To improve predictions of human oral drug absorption, we decided to use the similarity approach described above. In this case, we get an equation of the following form:

$$FA = \sum_{i=1}^{N} [1/(1 + 10^{\log(1-FAi)/FAi + 0.39(\Sigma Ca - \Sigma Cai) - 0.29(\Sigma Cd - \Sigma Cdi)})]/N \quad (19)$$

Table 3 summarizes the results of calculations using from 1−3 nearest neighbors for human oral absorption of 42 drugs. Table 4 gives the results of the statistical analysis in comparing the experimental against the calculated values for each case. The number of compounds with acceptable values for human oral absorption ($\Delta = FA_{exp} - FA_{calc} \leq 0.1$) is 33 (78.6%); marginal ($0.1 < \Delta \leq 0.2$), 7 (16.7%); and unacceptable ($\Delta > 0.2$), 2 (4.7%).

The material presented above demonstrates advantages of a prediction procedure based on a combination of traditional QSAR and Similarity. For any property or activity, the corresponding equations can be different. For example, both in the form and contents, eq 19 differs from eq 14 (for lipophilicity) and from eq 16 (for solubility). In the latter cases, there is a linear dependence of properties on the descriptors; in the case of absorption, the dependence is sigmoid.

The first terms of eqs 14, 16, and 19, those related to the experimental properties of the nearest neighbors, play important roles in that they take into account much of what is not understood about how the various structural features contribute to the property of interest. We sometimes refer to this as dealing with such unknowns in a "hidden form". In the case of lipophilicity, this has allowed us to deal with intramolecular interactions that are not well taken into account by the usual additivity schemes. In the case of the solubility of solid chemicals and drugs we used nearest neighbors to account the influence of crystal lattice energy, which even now cannot be estimated quantitatively by a direct way. In the case of complex phenomena, such as absorption, the mechanism of action is often unknown. Here, the principle "similar compounds display similar properties" allows one to bypass this problem by assuming that closely related compounds have similar mechanisms of absorption. Thus, the methodology considered in this paper allows one to create reliable predictive models of "structure-property" even in the absence of a precise understanding of the mechanisms involved.

CONCLUSION

In this paper, we presented a novel approach to reliably predict the properties of complex chemicals and drugs on the basis of the experimental values of closely related chemical structures and physicochemical descriptors. The predictive power such models will be defined by the information content of the databases employed. Thus, careful analysis and data processing play significant roles. In this approach, supplemental data should improve the accuracy of the predictions. Obviously, additional developments in the similarity concept could also lead to further improvements. These include estimating not only similarity on the basis of two-dimensional indices but also enhancements in the descriptors concerning a molecular structure, e.g. indices of both chemical structure and physicochemical parameters.

ACKNOWLEDGMENT

The authors thank Dr. J. W. McFarland, Dr. K.-J. Schaper, and both reviewers for valuable comments that helped us to complete this work for publication.

REFERENCES AND NOTES

(1) Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: a Comparative Study. *Quant. Struct.-Act. Relat.* **1996**, *15*, 403−409.
(2) Waterbeemd, H.; Mannhold, R. Programs and Methods for Calculation of logP-values. *Quant. Struct.-Act. Relat.* **1996**, *15*, 410−412.
(3) Hansch, C. F.; Quinlan. J. E.; Lawrence. G. L. The linear free energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.
(4) Irmann, F. Eine einfasche korrelation zwischen wasserloslichkeit und strukture von kohlenwasserstoffen und halogenkohlenwasserstoffen. *Chem. Ing. Tech*. **1965**, *37*, 789−798.
(5) Yalkovsky, S. H.; Valvani, S. C. Solubility and partitioning 1. Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **1980**, *69*, 912−922.
(6) Yalkovsky, S. H.; Valvani, S. C.; Rosemann, T. J. Solubility and partitioning VI: octanol solubility and octanol−water partition coefficients. *J. Pharm. Sci*. **1983**, *72*, 866−870.
(7) Yalkowsky, S. H. Solubility and Solubilization. In *Aqueous Media*; Oxford University: Oxford, 1999.
(8) Nouwen, J.; Hansen, B. Correlation analysis between watersolubility, octanol−water partition coefficient and melting point based on clustering. *Quantum Struct.-Act. Relat* **1996**, *15*, 17−30.
(9) Wakita, K.; Yoshimoto, M.; Mijamoto, S.; Watanabe, H. A. A method for calculation of aqueous solubility of organic compounds using new fragment solubility constants. *Chem. Pharm. Bull*. **1986**, *34*, 4663−4681.
(10) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 474−482.
(11) Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurman, G. Group contribution methods to estimate water solubility of organic compounds. *Chemosphere* **1995**, *20*, 2061−2077.
(12) Bodor, N.; Harget, A.; Huang, M.-J. Neural network studies.1. Estimation of the aqueous solubility of organic compounds. *J. Am. Chem. Soc*. **1991**, *113*, 9480−9483.
(13) Nelson, T. M.; Jurs, P. S. Prediction of aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 601−609.
(14) Chow, H.; Chen, H.; Ng, T.; Myrdal, P.; Yalkowsky, S. H. Using back-propagation networks for the estimation of aqueous activity coefficients of aromatic organic compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 723−728.
(15) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for Diverse set of heteroatom-containing organic compounds using a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 100−107.
(16) Ruelle, P. Undestanding the volume−solubility dependence: the mobile order and disorder view. *J. Phys. Org. Chem*. **1999**, *12*, 769−786.

(17) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR sdudies on vapor pressure, aqueous solubility and the prediction of water−air partition coefficients. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 720−725.

(18) Jorgensen, W. L.; Duffy, E. M. Prediction of drugs solubility from Monte Carlo Simulations. *Bioorg., Med. Chem. Lett*. **2000**, *10*, 1155−1158.

(19) Abraham, M. H.; Joelle, L. The correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci*. **1999**, *88*, 868−880.

(20) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug Del. Rev*. **1997**, *23*, 3−25.

(21) Fecik, R. A.; Frank, K. E.; Gentry, E. J.; Menon, S. R.; Mitscher, L. A.; Telikepalli, H. The search for orally active medications through combinatorial chemistry. In *Combinatorial Chemistry*; John Wiley & Sons: New York, 1998; pp 149−185.

(22) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muscal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(23) Norinder, U.; Osterberg, T.; Artursson, P. Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parametrization and PLS statistics. *Eur. J. Pharm. Sci*. **1999**, *8*, 49−56.

(24) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. Molecular hashkeys: A novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem*. **1999**, *42*, 1739−1748.

(25) Paterson, D. A.; Conradi, R. A.; Hilgers, A. R.; Vidmar, T. J.; Burton, P. S. A nonaqueous partitioning system for predicting the oral absorption potential of peptides. *Quant. Struct.-Act. Relat.* **1994**, *13*, 4−10.

(26) Balon, K.; Riebesehl, B. U.; Müller, B. W. Drug liposome partitioning as a tool for prediction of human passive intestinal absorption. *Pharm. Res.* **1999**, *16*, 882−888.

(27) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes. *J. Med. Chem*. **1998**, *41*, 1007−1010.

(28) Wohnsland, F.; Faller, B. High-throughput permeability pH profile and high-throughput alkane/water log *P* with artificial membranes. *J. Med. Chem*. **2001**, *44*, in press.

(29) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.

(30) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in human. *Pharm. Res.* **1997**, *14*, 568−571.

(31) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem*. **1998**, *41*, 5382−5392.

(32) Stenberg, P.; Luthman, K.; Artursson, P. Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties. *Pharm. Res*. **1998**, *16*, 205−212.

(33) Krarup, L. H.; Christensen, I. T.; Hovgaad, L.; Frøkjaer, S. Predicting drug absorption from molecular dynamics simulations. *Pharm. Res*. **1998**, *7*, 972−978.

(34) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci*. **1999**, *88*, 807−814.

(35) Wold, S.; Dunn, W. J.; Hellberg, S. Pattern recognition as atool for drug design. In *Drug design: Fact or Fantasy*; Jolles, G., Woldridge, K., Eds.; Academic Press: London, 1984; pp 95−117.

(36) Raevsky, O. A.; Sapegin, A.; Zefirov, N. S. Discrete-regression model. *Quant. Struct.-Relat.* **1995**, *15*, 403−409.

(37) Willett, P.; Barnard J. M.; Downs, G. M Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(38) *Concepts and applications of molecular similarity*; Johnson, M., Maggiora, G. M., Eds.; Wiley: New York, 1990.

(39) *Molecular similarity in drug design*; Dean, P. M., Ed.; Chapman and Hall: Glasgow, 1994.

(40) Downs, G. M. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* **1997**, *7*, 1−66.

(41) Basak, S. C.; Grunwald, G. D. Tolerance space and molecular similarity. *SAR QSAR Environ. Res*. **1995**, *3*, 265−277.

(42) Filimonov, D.; Poroikov, V. V.; Borodina, Y.; Glorizova, T. Chemical similarity assessment through multilevel neighborhoods of atoms. Definition and comparizon with other descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 983−996.

(43) 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.

(44) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. Ph.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* In press.

(45) Gerasimenko, V. A.; Trepalin, S. V.; Raevsky, O. A. MOLDIVS − a new program for molecular similarity and diversity calculations. In: *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jörgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 423−424.

(46) Raevsky, O. A.; Gerasimenko, V. A.; Trepalin, S. V. 1999. Program Package MOLDIVS (MOLecular DIVersity & Similarity), Patent No. 990093 (26.02.99) of Russian State Patent and Trade Mark Department, commercially available version: http://www.ibmh.msk.su/molpro.

(47) Waterbeemd, H.; Testa, B. In *Advances in Drug Research*; Testa, B., Ed.; Academic Press: London, 1987; Vol. 16, pp 87−225.

(48) El Tayar, N.; Testa, B. In *Trends in QSAR and Molecular Modelling 92*; Wermuth. C., Ed.; ESCOM: Leide, 1993; pp 101−108.

(49) Raevsky, O. A. Molecular structure descriptors in the computer-aided design of biologically active compounds. *Russ. Chem. Rev.* **1999**, *68*, 505−524

(50) Raevsky, O. A. Quantification of noncovalent interactions on the basis of the thermodynamic hydrogen bond parameters. *J. Phys. Org. Chem*. **1997**, *10*, 405−413.

(51) Raevsky, O. A. Hydrogen bond strength estimation by means of HYBOT. In *Computer-Assisted Lead Finding and Optimization*; Waterbeemd, H. van de, Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997; pp 367−378.

(52) Raevsky, O. A.; Grigor'ev, V. Ju.; Trepalin S. V. Program Package HYBOT (HYdrogen BOnd Thermodynamics), Patent No. 990090 (26.02.99) of Russian State Patent and Trade Mark Department, commercially available version: raevsky@ipac.ac.ru and reckon.dat@ibm.net (for U.S.A. and Canada) and J. C. Dearden@livjm.ac.uk (UK), 1999.

(53) Raevsky, O. A.; Schaper, K.-J.; Seydel, J. K. H-bond contribution to octanol−water partition coefficients of polar compounds. *Quant. Struct.-Act. Relat.* **1995**, *14*, 433−436.

(54) Raevsky, O. A.; Trepalina, E. P.; Trepalin, S. V. SLIPPER − New Program for Lipophilicity, Solubility & Liposome Permeability Prediction. *Chem.-Farm. Z.(Rus)* **2000**, *34*, 34−37.

(55) Avdeef, A. In *Lipophilicity in Drug Action and Toxicology*; Pliska, V., Testa, B., Waterbeemd, H., Eds.; VCH: Weinheim, 1996; pp 109−139.

(56) Raevsky, O. A.; Trepalin S. V.; Trepalina E. P. Program Package SLIPPER (Solubility, LIPophilicity & PERmeability), Patent No. 990089 (26.02.99) of Russian State Patent and Trade Mark Department, commercially available version: http://www.ibmh.msk.su/molpro, 1999.

(57) Raevsky, O. A.; Trepalina, E. P.; Trepalin, S. V. SLIPPER − a new program for water solubility, lipophilicity and permeability prediction. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, Boston, Dordrecht, London, Moscow, 1998; pp 489−490.

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

(59) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimation the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1355−1359.

(60) Raevsky, O. A.; Schaper, K.-J.; Waterbemd, H. van de; McFarland, J. Hydrogen bond contributions to properties and activities of chemicals and drugs. In *Molecular Modelling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 221−228.

(61) Schaper, K. J.; Kunz, B.; Trepalina, E.; Raevsky, O. A. (private communication).

(62) Abraham, M. H.; Joelle, L. The correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci*. **1999**, *88*, 868−880.

(63) Raevsky, O.; Fetisov, V.; Trepalina, E.; McFarland, J.; Shaper, K. Quantitative estimation of drug absorption in humans for passively transported compounds on the basis of their physicochemical parameters. *Quant. Struct.-Act. Relat.* **2000**, *19*, 366−374.

(64) Raevsky, O.; Schaper, K.; Artursson, P.;McFarland, J. On a General Predictive Model for the Intestinal Absorption of Drugs in Humans based on the Hydrogen Bond Descriptors and Structural Similarity. *Quant. Struct.-Act. Relat.* In press.

CI010097O