Model Selection Based on Structural Similarity−Method Description and Application to Water Solubility Prediction

Ralph Kühne, Ralf-Uwe Ebert, and Gerrit Schüürmann*

Department of Ecological Chemistry, UFZ Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany

Received September 5, 2005

A method is introduced that allows one to select, for a given property and compound, among several prediction methods the presumably best-performing scheme based on prediction errors evaluated for structurally similar compounds. The latter are selected through analysis of atom-centered fragments (ACFs) in accord with a *k* nearest neighbor procedure in the two-dimensional structural space. The approach is illustrated with seven estimation methods for the water solubility of organic compounds and a reference set of 1876 compounds with validated experimental values. The discussion includes a comparison with the similarity-based error correction as an alternative approach to improve the performance of prediction methods and an extension that enables an ad hoc specification of the application domain.

INTRODUCTION

Water solubility (S_w) is one of the key properties for the environmental fate of organic compounds. Furthermore, it acts as a threshold for the aquatic toxicity, imposing an upper limit on the amount of xenobiotics dissolved in aqueous solution under thermodynamic conditions. There are many experimental S_w data available with uncertainties typically up to 0.5 logarithmic units. Nonetheless, prediction methods are important both for filling data gaps and screening virtual compound libraries as well as for quality checks.

While there are many S_w prediction models available[1−4] and even quite a larger number of new methods published only in 2004 and 2005,[5−11] there is usually little known about the associated application domains and the performance when applying the schemes to new compounds. Obviously, models may differ in their prediction capabilities due to different descriptors, computational techniques, and complexity. To select an appropriate model, the intended accuracy will be considered as well as available computational opportunities for the method-specific descriptors and their physicochemical meaning. Moreover, intrinsic factors such as known performance differences for different compound classes may serve as selection criteria. Note, however, that an unambiguous assignment to certain compound classes is usually not possible for multifunctional chemicals and substances containing several heteroatoms.

Another aspect of model selection is the application domain. Each model is developed by means of a more or less limited training set and usually should not be applied to compounds outside the respective chemical domain. Extrapolations may be possible to some extent depending on the model. An idea of the chemical domain may be obtained by published training sets, but this is not the case in general. And even with a known training set, the coverage of the application domain is not straightforward due to the complexity of chemicals as outlined above, and because there is no unique way to define the chemical domain associated with a set of compounds.

If for a given target property several prediction methods are available, the overall best performance for a given set of compounds would be achieved by identifying and using the method with the smallest prediction error for each compound. While the compound-specific performance of a given method is usually not known in advance, it can be estimated from the prediction errors of the method when applied to sufficiently similar compounds. The latter requires an operational definition of molecular similarity and a somewhat larger set of reference compounds with experimental data for the property of interest. The performance of the new approach is illustrated taking water solubility and seven respective prediction methods as example.

MATERIALS AND METHODS

Data Set. The data set consists of thoroughly validated experimental values for the water solubility S_w at 25 °C of 1876 pure organic chemicals taken from our in-house database.[12] It covers molar masses from 16 to 666 and a broad range of compound classes − 203 hydrocarbons, 249 halogen hydrocarbons, 595 compounds containing O and partly halogens, 644 compounds with at least one N atom and possibly containing O and/or halogens, and 186 compounds with at least one of S or P in addition.

Particular attention has been devoted to include complex compounds. This is reflected by the number of 1030 chemicals in the training set with more than one functional group. In addition, there are 643 substances with a single functional group and 203 hydrocarbons. 1047 compounds contain at least one aromatic ring. With regard to the physical condition, there are 1140 solids, 694 liquids, and 42 gases. The water solubility data range covers almost 15 orders of magnitude from −12.8 to 1.92 in log units of mol/L. Since the focus of this study is the principle technique for the model

* Corresponding author phone: +49-341-235-2309; fax: +49-341-235-2401; e-mail: gerrit.schuurmann@ufz.de.

10.1021/ci0503762 CCC: $33.50 © 2006 American Chemical Society
Published on Web 02/04/2006

selection but not the water solubility model itself, the data set is not included here but will be part of a more detailed comparative analysis of the prediction performances of a larger variety of S_w estimation methods.

Individual Models. Seven models have been selected that are based on 2D descriptors and do not require the melting point in case of solids. All methods as well as the model selection procedure are implemented in our software system ChemProp[12] and can be run automatically.

Meylan et al.[13,14] − this method calculates log S_w from the logarithmic octanol/water partition coefficient (log K_{ow}), the molar mass, and some correction factors from the 2D structure. The associated training set consists of 1450 organic compounds − solids and liquids, including chemicals totally miscible with water − in a total range from −12 to +2 in logarithmic units, and the external validation set of 817 compounds. Because the training set is not published, the application domain cannot be determined in detail. Formally, a valid result for a compound is achieved as long as the associated log K_{ow} is known or can be calculated. The version applied here is part of the online EPI-Suite with two exceptions: The preferred model employing the melting point is not considered, and the log K_{ow} is computed by other methods, because the log K_{ow} model of Meylan et al.[14,15] is not published in full detail. In our implementation, the log K_{ow} model of Marrero et al.[16] is applied. When it fails for formal reasons, the model of Viswanadhan[17] has been applied.

Klopman et al.[18] − this method uses fragments and correction factors with a nonlinear smoothing of extreme results ("stereographic projection"). The original training set covered 1168 liquid and solid compounds, with explicit exclusion of gases, salts, and mixtures. Neither the training set nor the data range are published. Formally, compounds with structural features not covered by the method fragments are outside the application domain. Examples for the latter include acetylene, formaldehyde, nitro compounds, nitrates, and chemicals with NN═O. The mathematical technique sets the lower limit of the possible results to −10.75 and avoids extreme results at the upper end to a large extent.

Marrero et al.[16] − this model is a linear combination of first-, second-, and third-order fragments and the molar mass. The training set includes 2087 organic compounds from C_3 to C_{70} but again is not published. An external set of 238 data was applied for the model validation. Again, the application domain is given only formally in terms of the fragment coverage of the compounds.

Hou et al.[19] − in this case, a quite complex fragment method is combined with two general correction factors and the molar mass. The training set of 1290 compounds is entirely published in terms of chemical structures and data. In principle, all major compound classes are represented. The log S_w [mol/L] range is from −12 to +2. A test set of 21 chemicals was used for validation.

Huuskonen[20] − here, a linear combination of electrotopological indices was fitted to 674 chemicals (349 liquids and 325 solids) in a data range from −12 to +3 log units of S_w [mol/L]. For validation, a test set of 71 compounds was applied.

Tetko et al.[21] − this is another model based on electrotopological states. The training set consists of 879 compounds, and the test set of 412 compounds, with the data being presented in another paper.

The training sets of the models of Huuskonen and of Tetko et al. are not published explicitly but refer to other publications and thus could be reconstructed in principle. Formally, these models yield valid results for any organic compound (with atom types for which electrotopological indices are defined) except methane. In the Tetko model, however, there is an additional formal limitation to a maximum of 50 atoms except hydrogen in one molecule.

Abraham et al.[22] − the model applies the LSER (linear solvation energy relationship) equation. The published training set covers only 594 compounds, ranging from −9 to +2 log units of S_w [mol/L]. In our approach to apply calculated descriptors only, the five descriptors required for each compound are estimated from structure by available fragment methods.[23,24] The model yields formally valid results whenever the descriptors can be calculated. Note, however, for a more elaborated applicability check the compounds of interest would need to be compared to the training set compounds of the original LSER model.

Model Selection. To select an appropriate model for a compound of interest, a k nearest neighbors approach is applied based on a quantitative similarity measure defined in the 2D structural space of the compounds. In our case, k is equal to 5, and molecular similarity is specified in terms of atom-centered fragments (ACFs) as outlined in the next section. The approach requires an initial run of all methods with a reference set of compounds with validated (typically experimental) data (in our case: experimental S_w at 25 °C for 1876 organic compounds as described above) and a storage of all respective model errors. For a compound of interest, the five most similar reference compounds are identified, and the method with the associated smallest average absolute error for this compound-specific subset is selected for predicting the property under investigation.

A necessary condition for the model selection is that at least one reference compound that meets the similarity threshold belongs to the (at least formally defined) application domain of the prediction method. If no sufficiently similar compound is available, a predefined default method (typically the method with the overall best performance as identified in the initial run) is selected. Alternatively, the similarity-based compound selection could also be used to introduce an ad hoc definition of the applicability domain that goes beyond the formal model constraints, which will be discussed below.

Similarity Measure. Molecular similarity is defined through atom-centered fragments (ACFs).[25] More precisely, first-order ACFs are used that are built from the hydrogen-suppressed connection table of a molecule by center atoms and their first neighbors. Each non-hydrogen atom is the center atom of one ACF, which results in as many ACFs per molecule as non-hydrogen atoms. Extending the − to our knowledge − first mentioning of ACFs,[26] these may vary with respect to the degree of accuracy in specifying the atom types of the center atom as well as of the neighbor atoms, taking into account arbitrary atom properties (e.g. halogen, chlorine, ring atom) and bond properties (e.g. aromatic). For the model selection procedure, a hierarchical ACF definition has been applied, which comprises five differently weighted levels and is summarized in Table 1. Taking level 1 as an

Table 1. Hierarchical ACF Definition of Molecular Similarity for Eq 1

level	relative weight	center atom type	center atom properties	bond types	neighbor atom type	neighbor atom properties
1	1.00	except H	number of H, aromaticity	single, double, triple, aromatic	except H	aromaticity
2	0.50	except H	aromaticity	single, double, triple, aromatic	except H	aromaticity
3	0.25	only halogen, not distinguished		single, double, triple, aromatic	except H	aromaticity
4	0.10	except H, halogen not distinguished	aromaticity			
5	0.05	except H, halogen not distinguished				

example,center atoms are characterized in terms of the atom type, the number of attached hydrogens and the aromaticity status, bonds between the center atom and first non-hydrogen neighbors in the usual manner (single, etc.), and first non-hydrogen neighbor atoms in terms of atom types and aromaticity status.

The hierarchy setup was primarily led by mechanistic considerations on partitioning processes and may be subject to future refinements. Model runs revealed the superiority of a multilevel approach as compared to a single-level definition of ACF similarity in the context of the presently introduced model selection procedure. The single-level approach may yield an insufficient number of similar compounds when applying reasonably stringent similarity thresholds. Decreasing the similarity threshold in such cases may extend the range of compounds to be selected in a quite unsystematic way, while inclusion of additional selection levels with less strict similarity definitions has turned out to be more powerful in providing additional compounds still based on structural considerations. Similarity level 2 essentially is level 1 without considering attached hydrogens, level 3 explicitly addresses similarity in halogen patterns, and level 4 and 5 account for simple atom type similarity (center atom only) with and without aromaticity. Down from level 3, halogens are treated generically without distinguishing between different halogen atom types. The weights were selected to reflect the respective considerations, e.g., the most differentiating level 1 effectively contributes to 52% of the calculated similarity, and the weakest levels 4 and 5 together account for 7%.

To compare two molecules through their ACFs, the Dice similarity was slightly modified to reflect the hierarchical weights:

$$\text{similarity}_{A,B} = \sum_i w_i \cdot \frac{2 \cdot \text{ACF}_i^{A,B}}{\text{ACF}_i^{A} + \text{ACF}_i^{B}} \qquad (1)$$

Here, i is the respective hierarchical level with weight w_i (as defined in Table 1), ACF_i^{A} and ACF_i^{B} denote the numbers of ACFs defined according to level i in the molecules A and B, and $\text{ACF}_i^{A,B}$ is the number of ACFs defined according to level i that are common to both molecules. In the present investigation, a similarity of 0.7 was used as minimum threshold. Accordingly, for each chemical only compounds with a ACF-based similarity of at least 0.7 were proper candidates for the model performance check, resulting in less than five hits in some cases. The similarity threshold of 0.7 has proven as empirical optimum by several trial-and-error runs.

Algorithm. In summary, the following steps are required to perform estimations by the multimodel technique:

1. Training set preparation

1.1. Set up a database of structures and corresponding validated experimental data.

1.2. Provide implementations of the respective individual models.

1.3. Run all models for all compounds.

1.4. From the results, calculate all absolute errors and put them into the database.

1.5. For all compounds, analyze and store all atom-centered fragments.

2. Prediction runs

2.1. Analyze the atom-centered fragments of the prediction compound.

2.2. Select the *n* (here: 5) most similar compounds from the database. Do not consider compounds with a similarity below a certain threshold (here: 0.7).

2.3. For all individual models, calculate the signed averaged error (bias) for these *n* database compounds.

2.4. Select the model with the lowest bias.

2.5. Apply the model to the prediction compound.

3. Application domain

3.1. If application domain information for individual models is available, apply. Do not consider models with application domain violations.

3.2. If task 2.2 cannot identify *n* compounds, use less than *n*. Keep in mind, the estimation result may be less reliable.

3.3. If task 2.2 cannot identify at least one compound, a default individual model hierarchy may be applied depending on a general analysis of the results from 1.4. However, in this case there is a severe risk of being outside the application range even for the individual models (cf. discussion below).

Indeed the steps of 1. and 2. require several software modules. For the individual methods (1.2 and 1.3), it depends on the models themselves. A number of chemical database systems are available for the training set treatment (1.1 and 1.4). The calculation steps (1.4 and 2.2−2.5) can easily be implemented in standard spreadsheet software. The most difficult task is the ACF treatment (1.5 and 2.1). Some molecular modeling packages offer respective opportunities. In our case, the entire procedure (1.1−2.5) has been performed by means of the in-house software system ChemProp.[12]

RESULTS AND DISCUSSION

Performance of Individual Methods. The statistical performance of the seven models when applied to the reference set of 1876 compounds with experimental S_w data is summarized in the upper part of Table 2. Interestingly, only the Hou method is formally applicable to all compounds, while there are 93 compounds with missing fragments for

Table 2. Statistic for the Individual Models and the Multimethod Model Approach[a]

model	*n*	q^2	*se*	bias	max ne	max pe
Meylan[13−17]	1866	0.83	0.86	+0.07	−4.00	+3.94
Hou[19]	1876	0.82	0.87	−0.33	−6.12	+3.38
Tetko[21]	1874	0.80	0.93	−0.07	−5.20	+4.48
Marrero[16]	1783	0.79	0.96	+0.13	−3.85	+4.63
Klopman[18]	1851	0.68	1.19	−0.16	−4.50	+5.03
Huuskonen[20]	1875	0.58	1.36	+0.12	−9.96	+14.9
Abraham[22−24]	1874	0.34	1.69	+0.12	−23.1	+6.09
multimethod model	1876	0.88	0.71	−0.01	−3.85	+4.15
worst-selection model	1876	−0.17	2.26	+0.05	−23.1	+14.9
theoretical optimum	1876	0.98	0.31	<0.01	−2.41	+1.79

[a] Abbreviations: *n* − number of formally valid results, q^2 − predictive squared regression coefficient, *se* − standard error, bias − systematic error, max ne − maximum negative error, max pe − maximum positive error. Data are in logarithmic units of mol/L.

the Marrero model. The overall best statistics are achieved with the Meylan method with a predictive squared correlation coefficient q^2 (for an explanation in contrast to the conventional squared correlation coefficient r^2, see ref 27) of 0.83 and an associated standard error of 0.86 log units of S_w [mol/L]. Note that this is the only method that makes use of the (predicted) log K_{ow} of the compounds, while the other 6 models are confined to fragments and other substructural features of the 2D structure of the molecules (which holds also true for the Abraham method, because here all LSER parameters are estimated from 2D structural features of the compounds).

The prediction performance of the Hou method is pretty close to the Meylan method except for a significantly larger bias (−0.33 vs 0.07). The latter suggests that with this method, there is room for improvement through recalibration of the fragmentation scheme. Interestingly, the LSER approach is inferior to all other methods, probably because of a limited performance of the fragment scheme to predict the LSER parameters from molecular structure.[23,24] Both the Huuskonen and the Abraham method yield individual prediction errors above 10 log units, which reflects pitfalls of the formal definition of the application domain as given by the respective fragmentation schemes. However, also with the other 5 methods there are individual prediction errors above 3 log units and thus factors of ca. 10 above the experimental accuracy. These findings indicate that with current fragment schemes to predict water solubility, the formal applicability of the method is alone not sufficient to guarantee a reasonable prediction quality.

For the model selection procedure, the overall best performing Meylan method is selected as the default method for those cases where no sufficiently similar compound with a formal applicability of at least one method is available, and the Hou method is taken as the substitute default method in cases where the Meylan method cannot be applied due to missing fragments for the log K_{ow} models used.

Multimethod Model Performance. As outlined above, the model selection is based on an evaluation of the performances of the individual methods for subsets of the most similar reference compounds. Here, molecular similarity is defined through the Dice coefficient applied on ACFs (eq 1). For the application of this approach to the data set of 1876 compounds, however, too optimistic results would be achieved, because each individual compound would also belong to its associated subset of most similar compounds (every compound has maximum similarity to itself). As a consequence, a leave-one-out procedure was applied, making sure that for each compound of the data set, the model selection is based strictly on the performances of up to 7 methods with the (up to) 5 most similar compounds except the compound itself (and except those cases where the default method mode applies).

Table 3. Multimethod Approach: Numbers of Individual Model Selections as Compared to Their Actual Performances

model	no. of selections	no. of truly best results	q^2 from Table 1
Meylan[13−17]	424	325	0.83
Hou[19]	211	241	0.82
Tetko[21]	205	232	0.80
Marrero[16]	244	262	0.79
Klopman[18]	274	304	0.68
Huuskonen[20]	316	287	0.58
Abraham[22−24]	202	225	0.34

Inspection of Table 2 shows that the respective multimethod model is significantly superior to all individual methods except for the maximum positive prediction error (4.15 log units). As compared to the overall best individual method, q^2 has increased by 0.05 units, and the standard error has decreased by 0.15 log units. However, there is still room for improvement, as can be seen from the statistics of the theoretical optimum (last row in Table 2). The latter refers to the ideal situation where for each compound, the truly best model has been selected (which is possible in our case because we know the prediction error for each compound and method). Interestingly, the standard error of the theoretical optimum is indeed in the range of the experimental error, while the maximum prediction error is still greater by a factor of 5.

The substantial variation in individual model performances is reflected by the performance of the worst-model selection (which is again possible due to the known true prediction errors). The associated q^2 of −0.17 indicates that this worst-method model is even inferior to using the average log S_w for all predictions (cf. ref 27).

Model Selection Results. Details of the model selection are summarized in Table 3. The overall best Meylan model yields the truly best results for 325 of the 1876 compounds and was selected by our ACF-based procedure in 424 cases. Even the overall worst Abraham method is superior to the other 6 methods in 225 cases and was selected 202 times. The latter indicates that for a given compound, the local method performance is relevant. Our presently introduced model selection procedure addresses this feature, because it is designed to identify, for each compound, the locally best performing method with the help of ACF-based molecular similarity.

A more detailed analysis of the model selection performance is given in Table 4. For 472 of the 1876 compounds (25%), the truly best model (judged from the individual prediction errors) was in fact identified as the best local model for the subsets of (up to) 5 most similar reference compounds. While this may look disappointing at first sight, it should be noted that in many cases, the differences between the prediction errors of several methods are relatively small. When lumping together the cases where the selected model is the truly best, second best, or third best model, Table 4

Table 4. Multimethod Model: Numbers of Individual Model Selections as Compared to Their True Compound-Specific Ranking

true rank	no. of selections	true rank	no. of selections
1	472	5	192
2	389	6	134
3	303	7	126
4	260		

shows that this is achieved for 1164 compounds (62%). By contrast, only for 452 compounds (24%) one of the actually three worst methods was selected using the multimethod approach.

While the model selection procedure is clearly not perfect, its leads to a significant increase in the overall performance of the property prediction, as is demonstrated in Table 2. At the same time, there are several opportunities for improving the multimethod model while maintaining the principle approach. First, an increased set of reference compounds will of course increase the chances to identify more similar reference compounds and thus provide a better basis for the comparative evaluation of the individual methods with respect to their expected performance for the compounds under investigation. Second, the ACF-based definition of molecular similarity (eq 1) contains opportunities for refinement with regard to both the ACF definition itself and the hierarchical structure, both of which is beyond the scope of the present study.

Comparison with Similarity-Based Error Correction. A related approach to improve the performance of existing prediction methods is the similarity-based error correction.[26] Here, for a given (single) method identification of the k (in our case: $k = 5$) most similar compounds is followed by the calculation of the method-specific bias for this subset, and the prediction for the compound of interest is corrected by this bias. In this way, local systematic errors can be corrected, while nonsystematic errors remain. As an example, the Meylan et al. model estimates −5.78 for fluphenazin. The five most similar compounds, trifluoperazin, perphenazin, trifluopromazin, prochlorperazine, and thiopropazate, yield errors from −0.35 to −2.25, with an average error (bias) of −1.56 for the five chemicals. Correction of the fluphenazin result by −1.56 results in −4.22. With an experimental value of −4.15, the corresponding estimation error of the corrected result is −0.07 in contrast to −1.63 without correction.

When confining the ACF-based similarity to level 1 of Table 1, the local error correction based on method-specific biases of (up to) the 5 most similar compounds results in significantly improved prediction statistics for most of the 7 methods under investigation. For example, for the Klopman method q^2 increases from 0.675 to 0.818, and for the Hou method from 0.824 to 0.893, and the associated *se* values decrease from 1.191 to 0.891 (Klopman) and from 0.875 to 0.681 (Hou), respectively. The ad hoc improvement of the Hou method is particularly striking and leads to a performance that is even slightly superior to the multimethod model. This result suggests that for the methods and compounds under investigation, the prediction errors are mainly caused by local biases, the latter of which can be eliminated ad hoc through similarity-based error corrections.

At the same time, it is notable that the newly introduced model selection procedure achieves an essentially similar performance without any actual correction but only through a judicious choice of the method to be used for each compound of interest. Note that for this procedure, the absolute average error achieved for the subset of (up to) the 5 most similar compounds is decisive, while the similarity-based error correction is driven by the local bias (sum of the signed prediction errors). There are also opportunities for combining both strategies, which is beyond the scope of the present investigation.

Application Domain. For a suite of prediction methods, the model selection procedure can also be used for a local similarity-based specification of the associated application domain. The idea is as follows: If for a given compound sufficiently similar reference compounds can be found, the associated reference data allow one to perform a local performance check. If, however, no sufficiently similar reference compound can be found, it is likely that there is little experience with the method performance for chemicals structurally similar to the compound of interest. In this case, the compound would be considered as being outside the chemical domain of the multimethod model. Knowledge of the application domains of the individual methods would provide additional method-specific constraints. As noted earlier, however, most of the individual application domains are currently not available.

When applying the ACF-based hierarchical similarity scheme (Table 1) with a threshold of 0.7 (eq 1), the multimethod model identifies 140 compounds, where no sufficiently similar reference compounds can be found. For this subset, the default mode (application of Meylan or Hou if Meylan cannot be applied) yields $q^2 = 0.663$ and $se = 1.135$, which is substantially inferior to the theoretical optimum (best-model selection, cf. Table 2) with $q^2 = 0.945$ and $se = 0.460$. For the complementary set of 1736 (= 1876−140) compounds all of which have at least one sufficiently similar reference compound, the default mode (Meylan/Hou) and the theoretical optimum yield q^2 values of 0.841 and 0.979 and *se* values of 0.831 and 0.301, respectively. Here, the corresponding results of the model selection procedure are $q^2 = 0.897$ and $se = 0.671$. These findings show that, on the average, the prediction performance is indeed significantly inferior for chemicals without sufficiently similar reference compounds than for chemicals where the reference set contains at least one sufficiently similar (but structurally still different) compound. As a consequence, ACF-based similarity appears useful as structure-related domain check, which may be used in addition to fragmentation-based (formal) definitions of method-specific application domains.

CONCLUSIONS

The model selection procedure can be understood as a computerized decision-support system that allows one to select, for a given compound and suite of prediction methods, the presumably best performing calculation scheme. Evaluation of molecular similarity based on atom-centered fragments (ACFs) in combination with a larger reference set leads to prediction results significantly superior to the ones of the individual methods, without any change of calculation schemes or prediction results. In contrast to the similarity-based error correction, the model selection procedure ad-

dresses both local biases and nonsystematic errors. Besides improving the prediction performance, the multimethod model enables a run-time check of its applicability domain. Opportunities for improvement include refinements of the ACF concept of molecular similarity, consideration of higher-order ACFs, and optimization of procedure-specific parameters such as similarity threshold, the number of most similar compounds to be evaluated, and alternative hierarchical structures of the ACF-based similarity.

ACKNOWLEDGMENT

This study was partly funded by the German Federal Environmental Agency (Umweltbundesamt) Dessau, Germany (FKZ 2046762/05), and by the European Commission (Integrated Project NOMIRACLE, Contract No. 003956-GOCE).

REFERENCES AND NOTES

(1) Huuskonen, J. Estimation of Aqueous Solubility in Drug Design. *Comb. Chem. High Throughput Screening* **2001**, *4*, 311−316.
(2) Tolls, J. Sorption of Veterinary Pharmaceuticals in Soils: a Review. *Environ. Sci. Technol.* **2001**, *35*, 3397−3406.
(3) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.
(4) Taskinen, J. Prediction of Aqueous Solubility in Drug Design. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 102−107.
(5) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170−176.
(6) Clarc, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.
(7) Raevsky, O.; Andreeva, E.; Raevskaya, O.; Skvortsov, V.; Schaper, K. QSPR Analysis of the Partitioning of Vaporous Chemicals in a Water-Gas-Phase System and the Water Solubility of Liquid and Solid Chemicals on the Basis of Fragment and Physicochemical Similarity and HYBOT Descriptors. *SAR QSAR Environ. Res.* **2005**, *16*, 191−202.
(8) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.
(9) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly From Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(10) Estrada, E.; Delgado, E. J.; Alderete, J. B.; Jaña, G. A. Quantum-Connectivity Descriptors in Modeling Solubility of Environmentally Important Organic Compounds. *J. Comput. Chem.* **2004**, *14*, 1787−1796.
(11) Hilal, S. H.; Karickhoff, S. W.; Carreira, S. W. Prediction of the Solubility, Activity Coefficient and Liquid/Liquid Partition Coefficient of Organic Compounds. *QSAR Comb. Sci.* **2004**, *23*, 709−720.
(12) Schüürmann, G.; Kühne, R.; Kleint, F.; Ebert, R.-U.; Rothenbacher, C.; Herth, P. A Software System for Automatic Chemical Property Estimation From Molecular Structure. In *Quantitative Structure-Activity Relationships in Environmental Sciences - VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 93−114.
(13) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility From Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.
(14) Meylan, W. M.; Howard, P. H. Estimating Log P With Atom/Fragments and Water Solubility With Log P. *Perspect. Drug. Discovery Des.* **2000**, *19*, 67−84.
(15) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(16) Marrero, J.; Gani, R. Group-Contribution-Based Estimation of Octanol/Water Partition Coefficient and Aqueous Solubility. *Ind. Eng. Chem. Res.* **2002**, *41*, 6623−6633.
(17) Viswanadhan, V. N.; Ghose, A. K.; Wendoloski, J. J. Estimating Aqueous Solvation and Lipophilicity of Small Organic Molecules: A Comparative Overview of Atom/Group Contribution Methods. *Perspect. Drug. Discovery Des.* **2000**, *19*, 85−98.
(18) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445. Errata **2001**, *41*, 1096−1097.
(19) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.
(20) Huuskonen, J. Estimation of Water Solubility From Atom-Type Electrotopological State Indices. *Environ. Toxicol. Chem.* **2001**, *20*, 491−497.
(21) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(22) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(23) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.
(24) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid Chromatography. *Chromatographia* **1987**, *23*, 243−246.
(25) Kühne, R.; Kleint, F.; Ebert, R.-U.; Schüürmann, G. Calculation of Compound Properties Using Experimental Data From Sufficiently Similar Chemicals. In *Software Development in Chemistry 10*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt, 1996; pp 125−134.
(26) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centred Fragments. *J. Chem. Soc.* **1971**, 1971, 3702−3706.
(27) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Prediction of the Temperature Dependency of Henry's Law Constant From Chemical Structure. *Environ. Sci. Technol.* **2005**, *39*, 6705−6711.

CI0503762