Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge

Tiejun Cheng, Yuan Zhao, Xun Li, Fu Lin, Yong Xu, Xinglong Zhang, Yan Li, and Renxiao Wang*

State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, P. R. China

Luhua Lai

State Key Laboratory of Structural Chemistry of Stable and Unstable Species, College of Chemistry, Peking University, Beijing, P. R. China

Received July 18, 2007

We have developed a new method, i.e., XLOGP3, for log*P* computation. XLOGP3 predicts the log*P* value of a query compound by using the known log*P* value of a reference compound as a starting point. The difference in the log*P* values of the query compound and the reference compound is then estimated by an additive model. The additive model implemented in XLOGP3 uses a total of 87 atom/group types and two correction factors as descriptors. It is calibrated on a training set of 8199 organic compounds with reliable log*P* data through a multivariate linear regression analysis. For a given query compound, the compound showing the highest structural similarity in the training set will be selected as the reference compound. Structural similarity is quantified based on topological torsion descriptors. XLOGP3 has been tested along with its predecessor, i.e., XLOGP2, as well as several popular log*P* methods on two independent test sets: one contains 406 small-molecule drugs approved by the FDA and the other contains 219 oligopeptides. On both test sets, XLOGP3 produces more accurate predictions than most of the other methods with average unsigned errors of 0.24−0.51 units. Compared to conventional additive methods, XLOGP3 does not rely on an extensive classification of fragments and correction factors in order to improve accuracy. It is also able to utilize the ever-increasing experimentally measured log*P* data more effectively.

INTRODUCTION

Since the pioneering work of Hansch et al.,[1−3] the logarithm of the partition coefficient between *n*-octanol and water (log*P*) has been widely used in quantitative structure−activity relationship (QSAR) studies as a parameter for characterizing lipophilicity. These QSAR studies cover a wide range of themes, including ligand−protein interaction, transportation process, cellular uptake, and so on.[4,5] In recent years, the role of log*P* has been "rediscovered" in modeling ADMET properties. Some surveys have revealed that approximately half of the drug candidates which eventually fail to reach market are due to unsatisfactory pharmacokinetic properties or toxicity.[6] Therefore, it will help to reduce the overall cost of drug discovery if such compounds can be identified as early as possible. Many approaches have been developed for evaluating ADMET properties or "druglikeness" based merely on the chemical structures of organic compounds,[7−9] in which log*P* often acts as a key descriptor. For example, among the descriptors used in Lipinski's "rule of five",[10] only log*P* cannot be deduced directly from chemical structure. Reliable computation of log*P* is thus much desired especially by the high-throughput studies in this field.

Since the 1970s, various methods for log*P* computation have been proposed, which have been reviewed from time to time.[11−13] Those methods can be roughly divided into two major categories. (i) Property-based methods: They compute log*P* as a function of molecular physicochemical properties, such as molecular surfaces, volumes, dipoles, partial charges, HOMO/LUMO energies, and others. Various topological and electrostatic indices may also be used as descriptors. While early methods normally employ linear equations to combine these descriptors, more recent methods tend to employ sophisticated statistical models, such as the associative neural network in ALOGPS.[14,15] The major shortcoming of such methods is perhaps their theoretical basis. Although it is intuitively acceptable to assume correlation between log*P* and molecular physicochemical properties, the rationale of mixing some molecular properties in a certain combination to compute log*P* is often unclear. Besides, many of these methods are validated on relatively small sets of organic compounds. It is unclear if they are applicable in a larger chemical space. (ii) Additive methods: Since the physicochemical properties of a molecule are in principle determined by its chemical structure, these methods use basic structural building blocks directly as descriptors. They compute the log*P* value of a given molecule by summing up the contributions from all building blocks of its structure. In addition, so-called correction factors are introduced when the results produced by the above approach deviate significantly from

* Corresponding author phone: 86-21-54925128; e-mail: wangrx@mail.sioc.ac.cn. Corresponding author address: Shanghai Institute of Organic Chemistry, 354 Fenglin Road, Shanghai 200032, P. R. China.

10.1021/ci700257y CCC: $37.00
Published on Web 11/07/2007

experimental values. Popular additive methods include CLOGP,[11,16] ALOGP,[17−19] ACD/LogP,[20,21] KOWWIN,[22,23] and KLOGP.[24,25] Additive methods are normally calibrated on large data sets. They also produce more accurate results than property-based methods according to some comparative evaluations.[26,27] At present, additive methods are dominant in practice.

We developed an additive method, called XLOGP,[28] about a decade ago. The current release of this method is XLOGP2,[29] which uses 90 basic atom types and 10 corrections as descriptors. Along the way of developing and applying additive methods, we have learned that such methods have their own problems. We believe that their fundamental problem lies in the assumption of additivity. Additivity in log*P* is certainly observed on a wide range of organic compounds, which in fact accounts for the success of additive methods. However, additivity may fail even in very simple cases. In order to illustrate this issue, we plot in Figure 1 the correlation between the log*P* values of a series of *n*-carboxylic acids and the lengths of their hydrocarbon chains. One can see that although a unit contribution of the methylene group can be derived from region *A* and *B*, respectively, it is not quite possible to derive a unit contribution of the methylene group across the entire series of compounds. One would expect that the additivity assumption is more likely to fail on more complicated structures. Indeed, additive methods are observed to produce larger errors on them.[26,30]

The second problem with conventional additive methods is that it is unclear how they can be further improved. An additive method relies on an extensive classification of molecular fragments, assuming that a properly defined fragment will have a unit contribution to log*P*. One possible approach for improving additive methods is to pursue an even more extensive classification scheme. As a matter of fact, some current additive methods already use a large number of fragments and correction factors, ranging from several hundred (such as CLOGP) to several thousand (such as ACD/LogP). An even more extensive classification of fragments, for example, by considering remote atoms in addition to neighboring atoms, will certainly increase the total number of possible fragments to the next level. This will cause technical problems since calibration of an additive method requires a sufficient number of log*P* data. Available log*P* data are probably not going to increase fast enough to support this approach. In addition, given the problem in additivity assumption mentioned above, we doubt that it will be helpful even if this approach is practical.

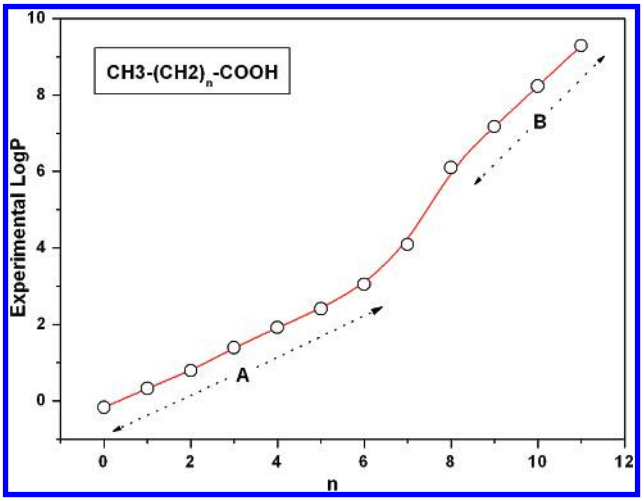

Figure 1. Experimentally measured log*P* values of *n*-carboxylic acids. Additivity of methylene groups is observed in region *A* and *B* but not globally.

In this article, we will describe a new method for log*P* computation by combining a knowledge-based approach with a conventional additive model. Our basic assumption is that compounds with similar chemical structures have similar properties, a strategy that has been successfully applied in many areas.[31] Thus, the log*P* value of a given compound can be computed more reliably from the known log*P* value of an appropriate reference compound. Our method addresses the problems in conventional additive methods and produces more accurate results than them. Detailed descriptions of our method will be given in the following sessions.

METHODS

Data Set Preparation. The training set used in our study is based on Hansch's compilation.[32] It provides experimentally measured log*P* values of over 16 700 organic compounds, about half of which are identified by Hansch et al. as reliable (indicated by * or √ in their compilation). Only those compounds with reliable log*P* values are considered in our study. The chemical structure of each compound is sketched according to its name and formula given in Hansch's compilation. If a compound has possible stereoisomers (*Z*/*E* or *R*/*S*), it is carefully sketched in the correct isomer if such information is specified; otherwise it is sketched in an arbitrary isomer. A number of compounds are discarded during this process because they are either inorganic compounds or contain undesired elements, such as metal atoms. Three-dimensional structural models of the remaining 8418 compounds are constructed with the Sybyl software.[33] Each compound is constructed in its most extended conformation and then optimized with the MMFF94 force field. Note that we do not attempt to obtain the lowest-energy conformation of each compound through extensive conformational sampling since our method actually does not rely on three-dimensional structures in computation. All final models are saved in Tripos Mol2 format for the convenience of automatic processing.

Two independent test sets are also compiled. Since a method for computing log*P* would better be tested on a variety of organic compounds of pharmaceutical interests, small-molecule organic drugs approved by the Food and Drug Administration (FDA) of the United States are considered in our study. We download a list of such compounds from DrugBank[34] and then search for their experimentally measured log*P* values in Hansch's compilation or the online PHYSPROP database at http://www.syrres.com/esc/physdemo.htm. Finally, experimentally measured log*P* values of a total of 406 compounds are collected, and they are assembled as the first test set. The other test set used in our study is designed to consist of a specific category of organic compounds. For this purpose, we retrieve all of the oligopeptides out of the 8418 compounds selected from Hansch's compilation, 219 in total, as the second test set.

The remaining 8418 − 219 = 8199 compounds are used as the training set for calibrating the atom-additive model in our method. It also serves as the knowledge set for finding

Table 1. Some Properties of the Data Sets Used in This Study

property[a]	training set $N = 8199$	test set 1 $N = 406$	test set 2 $N = 219$
molecular weight	222.7 (±89.2)	306.1 (±119.7)	365.0 (±86.4)
no. of heavy atoms	15.0 (±5.9)	21.2 (±8.5)	25.9 (±6.3)
no. of oxygen and nitrogen atoms	3.8 (±2.3)	5.1 (±2.8)	7.9 (±1.8)
log*P*	1.84 (±1.63)	1.85 (±1.99)	−0.86 (±1.18)

[a] Numbers outside brackets are mean values; numbers inside brackets are standard deviations.

reference compounds through similarity search (see the descriptions below). Some additional information on the training set and test sets used in our study is summarized in Table 1.

Overall Strategy. The key idea of our method, called XLOGP3, is to compute the log*P* of a given compound from the known log*P* of a structural analog, i.e., the reference compound. A conventional additive model computes log*P* as

$$\log P = \sum_{i=1}^{M} a_i A_i + \sum_{j=1}^{N} c_j C_j \qquad (1)$$

where a_i and A_i are the contribution and the occurrence of the *i*th atom/group type in the given compound, respectively, while c_j and C_j are the contribution and the occurrence of the *j*th correction factor in the given compound, respectively. Log*P* of the reference compound can be computed with the same additive model as

$$\log P^0 = \sum_{i=1}^{M} a_i A_i^0 + \sum_{j=1}^{N} c_j C_j^0 \qquad (2)$$

By subtracting eq 2 from eq 1, one gets

$$\log P = \log P^0 + \sum_{i=1}^{M} a_i (A_i - A_i^0) + \sum_{j=1}^{N} c_j (C_j - C_j^0) \qquad (3)$$

Equation 3 indicates how XLOGP3 computes the log*P* value of a given compound by using a reference compound as a starting point. A real example of this process is given in Figure 2.

The compound showing the highest structural similarity with the given compound in a knowledge set will be selected as the reference compound. In order to avoid using an irrelevant compound as reference, a similarity threshold of 50% is currently applied in XLOGP3. If a qualified reference compound cannot be found in the knowledge set, the given compound will be computed with the pure additive model (eq 1) for instead.

Additive Model. In our study, we prefer an atom-additive model to a group-additive model for its simplicity. An atom-additive model is certainly easier to implement since a molecule can be dissected into atoms without any ambiguity. In addition, an atom-additive model, if using a properly designed atom typing scheme, will not have the "missing fragment" problem that a group-additive model may have.

A total of 83 basic atom types are implemented in XLOGP3 to classify carbon, nitrogen, oxygen, sulfur, phosphorus, and halogen atoms. The classification of a given atom is made by considering (i) its element type, (ii) its hybridization state, (iii) its accessibility to solvent, characterized by the number of attached hydrogen atoms on this atom, (iv) the nature its direct neighboring atoms, (v) whether it is connected to a conjugated system with π electrons, and (vi) whether it is in a ring. Compared to what is in XLOGP2,[29] this atom typing scheme has been redesigned and optimized considerably. In addition to atom types, four terminal groups are also included in our classification scheme. Details of this classification scheme are summarized in the Supporting Information.

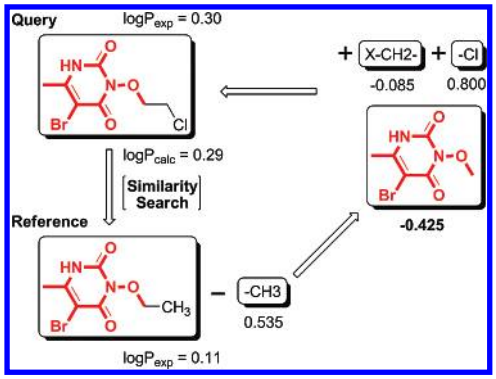

Figure 2. The basic computational procedure of XLOGP3.

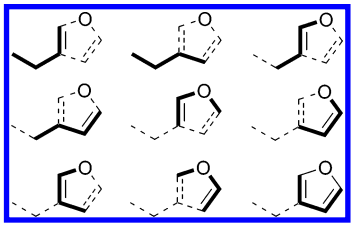

Figure 3. Topological torsion descriptors in 3-ethylfuran.

XLOGP3 also uses two correction factors, both of which have clear physical meanings and are found to have significant impacts on certain classes of compounds. The first correction factor accounts for internal hydrogen bonds, which makes a molecule less hydrophilic than what is indicated by its chemical structure. The second correction factor is used on organic compounds with amino acid moieties. Under a neutral pH condition, such a compound exists primarily in zwitterionic form instead of neutral form and thus exhibits a much lower apparent partition coefficient. Details of these two correction factors are also given in the Supporting Information.

With our new classification scheme of atom/group types and correction factors, the total number of descriptors has been reduced from 100 in XLOGP2 (90 atom types plus 10 correction factors) to 89 in XLOGP3 (87 atom/group types plus 2 correction factors). For the sake of convenience, the additive model described above will be referred to as XLOGP3-AA throughout the rest of this article.

Similarity Search Algorithm. In our method, the reference compound is required to be a structural analog to the given compound. The similarity between any two organic molecules is quantified by the number of common fragments found on their chemical structures. We adopt the so-called topological torsion descriptor (TTD) proposed by Nilakantan et al.[35] for this purpose. The topological torsion descriptor was originally defined as a set of four consecutively bonded non-hydrogen atoms (Figure 3), and each atom was characterized by its type, the number of non-hydrogen

branches attached to it, and its number of π electron pairs. In our study, we have extended the concept of topological torsion descriptor by characterizing each component atom with its XLOGP3 atom type so that more structural details are encoded implicitly. All possible topological torsion descriptors in a given structure are generated and recorded. The similarity score Sim_{AB} between two molecules *A* and *B* is then calculated as a Tanimoto coefficient[36]

$$Sim_{AB} = \frac{\sum_i w_i TTD_i^{AB}}{\sum_i w_i TTD_i^A + \sum_j w_j TTD_j^B - \sum_k w_k TTD_k^{AB}} \quad (4)$$

where TTD_i^A is the *i*th topological torsion descriptor in *A*; TTD_j^B is the *j*th topological torsion descriptor in B; and TTD_k^{AB} is the *k*th topological torsion descriptor common in *A* and *B*. Sim_{AB} ranges from 0 to 1. A score of 1 indicates the highest structural similarity between *A* and *B*.

In our method, the contribution of each topological torsion descriptor is also regulated by an adjustable weight factor *w*. By using weight factors, different topological torsion descriptors may have different contributions to the final similarity score, which is a popular option for computing structural similarity.[36] Our past experience suggests that large errors in computed log*P* values are often associated with chemical moieties containing heteroatoms. This is understandable since heteroatoms are generally more electronegative and polar than carbon atoms and thus play a more important role in the interactions with solvent molecules. Thus, it will be helpful for finding the right reference compound in log*P* computation if similarity score is biased somewhat toward heteroatom-containing moieties. In our algorithm, the weight factor of each topological torsion descriptor is computed as $(H+4)/N$, where *H* is the total number of heteroatoms in its contents, and *N* is the total occurrence of this descriptor in the given molecule. The weight factor is divided by *N* so that repeating topological torsion descriptors will not have exaggerated contributions.

Other log*P* Methods under Evaluation. Two methods which were originally developed by us are evaluated in this study, including XLOGP2 (the previous release of XLOGP)[29] and PLOGP (a method particularly developed for computing log*P*/log*D* values of peptides).[37] Other methods also evaluated in this study include CLOGP,[11,16] HINTLOGP,[38] TOPKAT,[39] AlogP98,[30] ALOGPS,[14,15] and KOWWIN.[22,23] Among them, CLOGP and HINTLOGP are implemented in the Sybyl software (version 7.2); AlogP98 is implemented in the Discovery Studio software (version 1.7); TOPKAT is from the standalone TOPKAT package (version 6.2); ALOGPS (version 2.1) is acquired directly from its authors; while KOWWIN is available online for testing at http://www.syrres.com/esc/est_kowdemo.htm. One should be aware that different implementations of the same method may produce somewhat different results. Therefore, the results reported in this study are valid strictly on the implementations indicated above.

Program Description. The XLOGP3 program is written in the C++ language and has been tested on Unix/Linux and Windows platforms. Besides log*P*, it also computes some other basic properties required by druglikeness rules, such as molecular weight and the number of hydrogen bond donors and acceptors. The XLOGP3 program is available online for testing at http://www.sioc-ccbg.ac.cn/software/xlogp3/. The standalone release of XLOGP3 is available by contacting the authors.

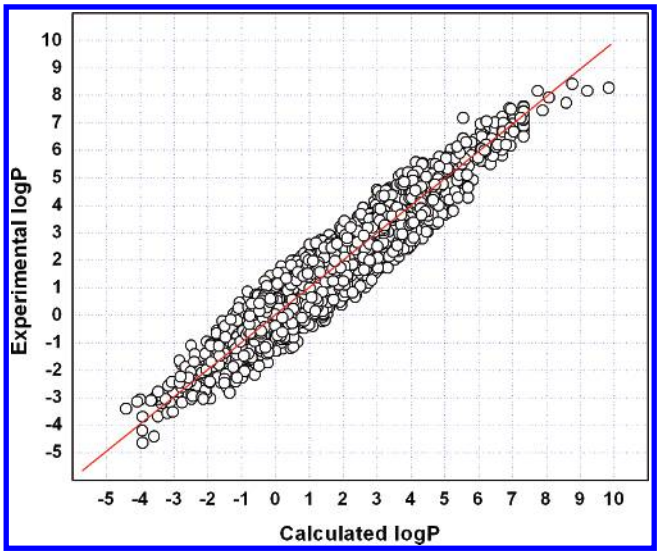

Figure 4. Experimentally measured log*P* values versus computed values by XLOGP3-AA on the training set ($N = 8199$, $R^2 = 0.913$, SD = 0.48, $\log P_{exp} = 0.993 \times \log P_{calc} + 0.024$).

RESULTS

The Additive Model. A total of 89 descriptors are used in XLOGP3-AA, including 87 atom/group types and two correction factors. The contributions of all descriptors are obtained through a multivariate linear regression analysis on the 8199 compounds in the training set (see the Supporting Information). The regression analysis produces a correlation coefficient between experimental and fitted values (R^2) of 0.913, a standard deviation between experimental and fitted values (SD) of 0.48, and a Fisher value (*F*) of 970 (Figure 4).

In order to test the predictive power of this regression model, an eight-fold cross-validation test is conducted. A total of 1000 compounds are randomly selected from the entire data set as a test set; the regression model is trained on the remaining 7199 compounds and is subsequently applied to the test set. The above trial is repeated for 1000 times to obtain robust statistical results. The average outcomes out of 1000 individual trials are as follows: a correlation coefficient between experimental and predicted values (R^2) of 0.904, a root-mean-squared error (RMSE) of 0.50, and a mean unsigned error (MUE) of 0.39. The results from this cross-validation test are very close to the ones from the regression analysis, indicating that XLOGP3-AA is not an overfitted model.

It is appropriate to make a comparison of XLOGP3-AA with XLOGP2 since the latter is also a pure additive model. A multivariate linear regression analysis of XLOGP2 on the training set used in this study produced the following: $R^2 = 0.885$, SD = 0.56, and $F = 631$. It is clear that XLOGP3-AA is a better regression model than XLOGP2. Apparently, the refined atom typing scheme in XLOGP3-AA should account for its improved performance over XLOGP2. Considering that XLOGP3-AA uses fewer descriptors than

Table 2. Results of Some log*P* Methods on the Training Set (N = 8199)

method[a]	R^2	RMSE[b]	MUE[c]
XLOGP3	0.937	0.41	0.31
CLOGP	0.931	0.45	0.30
XLOGP3-AA[d]	0.904	0.50	0.39
ALOGPS	0.894	0.54	0.34
AlogP98	0.854	0.62	0.46
XLOGP2	0.831	0.68	0.48
HINTLOGP	0.419	1.88	1.22

[a] Methods are ranked by the correlation coefficients produced by them. [b] Root-mean-squared error between experimental and predicted values. [c] Mean unsigned error between experimental and predicted values. [d] Eight-fold cross-validation results.

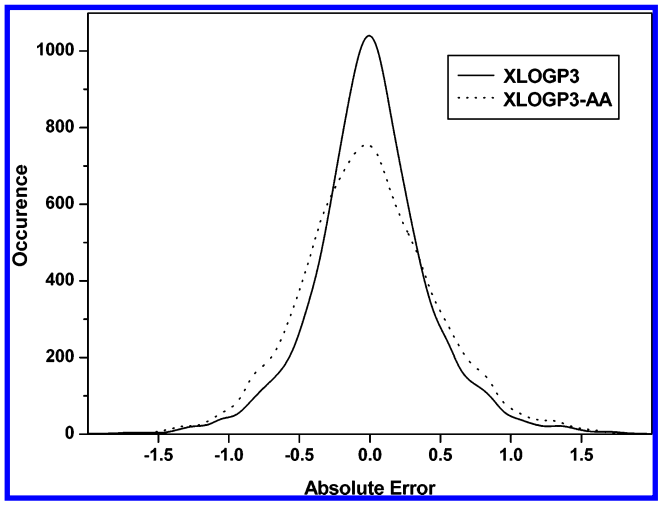

Figure 5. Error distributions of XLOGP3 and XLOGP3-AA on the training set.

XLOGP2 (89 versus 100), its improved performance on a large diverse set of organic compounds is inspiring.

The Full Model. The full model of XLOGP3 (eq 3) is also applied to the entire training set. For any given compound, its reference compound is selected among the N-1 remaining compounds in the training set. Since a similarity threshold of 50% is applied in XLOGP3, only 6256 compounds among all 8199 compounds in the training set, accounting for 76.3% of the entire population, find qualified reference compounds. Other compounds are still computed with XLOGP3-AA. Even so, XLOGP3 still demonstrates improved accuracy as compared to XLOGP3-AA: the correlation coefficient (R^2) between experimental and predicted values is increased from 0.911 to 0.937, while the mean unsigned error (MUE) in prediction is reduced from 0.38 to 0.31 (Table 2 and Figure 5).

Several popular methods for log*P* computation, including CLOGP, HINTLOGP, AlogP98, and ALOGPS, are also applied to the entire training set. Their statistical results are summarized in Table 2. The average error produced by XLOGP3 is arguably the smallest among all methods under test. It is encouraging to observe that in this test XLOGP3 is at least comparable to CLOGP, which is perhaps the most popular log*P* algorithm. Considering that our training set is in fact a subset of the data set used by CLOGP, the advantage of XLOGP3 may be more obvious than what is indicated in Table 2.

Performance on FDA-Approved Drugs. In order to evaluate its predictive power, we have applied XLOGP3 to

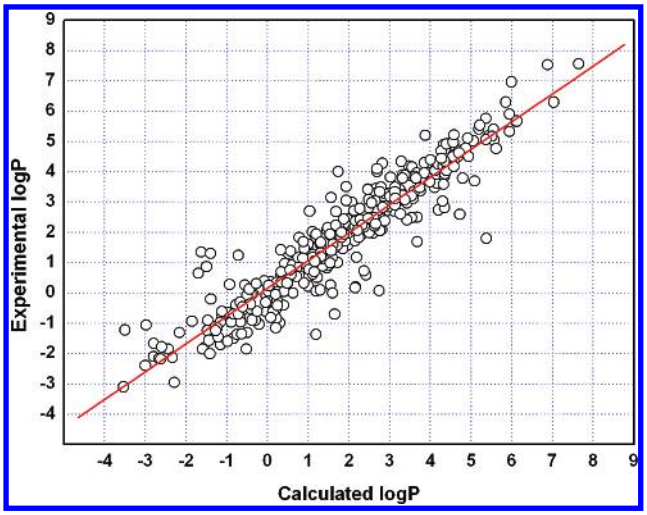

Figure 6. Experimentally measured log*P* values versus predicted values by XLOGP3 for 406 small-molecule drugs approved by the FDA (R^2 = 0.872, RMSE = 0.72, MUE = 0.51).

Table 3. Results of Some log*P* Methods on 406 FDA Approved Small-Molecule Drugs

method[a]	R^2	RMSE[b]	MUE[c]
ALOGPS	0.908	0.60	0.42
XLOGP3	0.872	0.72	0.51
XLOGP3-AA	0.847	0.80	0.57
CLOGP	0.838	0.88	0.51
TOPKAT	0.815	0.88	0.56
AlogP98	0.802	0.90	0.64
XLOGP2	0.777	0.95	0.68
KOWWIN	0.771	1.10	0.63
HINTLOGP	0.491	1.93	1.30

[a] Methods are ranked by the correlation coefficients produced by them. [b] Root-mean-squared error between experimental and predicted values. [c] Mean unsigned error between experimental and predicted values.

the 406 FDA approved small-molecule drugs in the first test set. This test set is challenging because its structural complexity is greater than that of the training set (see Table 1). The log*P* values of these compounds also span over nearly 12 units. The scatter plot of experimentally measured log*P* values versus predicted values by XLOGP3 is given in Figure 6. XLOGP3 makes reasonable predictions for most compounds in this test set except for about a dozen of significant outliers.

Statistical results produced by other log*P* methods in this test set are summarized in Table 3. One can see that the performance of XLOGP3 is considerably better than its predecessor XLOGP2. It also outperforms some other popular methods, such as CLOGP and TOPKAT. XLOGP3 is only second to ALOGPS with a marginally larger average error. It should be mentioned that calibration of XLOGP3 is completely independent of this test set, while ALOGPS uses the PHYSPROP database[40] as training set, which in fact contains all of the compounds in this test set. It is unclear how well ALOGPS will perform if these compounds are removed from its training set.

Performance on Oligopeptides. Our second test set consists of 219 oligopeptides. This test set is used to validate the predictive power of our method on a series of compounds sharing the same structural scaffold. The statistical results

Table 4. Results of Some log*P* Methods on 219 Oligopeptides

method[a]	R^2	RMSE[b]	MUE[c]
XLOGP3[d]	0.932	0.32	0.24
PLOGP	0.832	0.46	0.28
XLOGP3-AA	0.824	0.72	0.59
ALOGPS	0.765	0.73	0.54
XLOGP3[e]	0.758	0.71	0.55
TOPKAT	0.706	0.74	0.43
CLOGP	0.536	0.97	0.75
AlogP98	0.086	2.07	1.62
XLOGP2	0.078	2.24	1.78
KOWWIN	0.075	2.18	1.70
HINTLOGP	0.007	2.92	2.34

[a] Methods are ranked by the correlation coefficients produced by them. [b] Root-mean-squared error between experimental and predicted values. [c] Mean unsigned error between experimental and predicted values. [d] When peptides are included in the knowledge set. [e] When peptides are not included in the knowledge set.

produced by XLOGP3 as well as other log*P* methods are summarized in Table 4.

The correlation between experimentally measured log*P* values versus predicted values by XLOGP3 on this test set produces $R^2 = 0.758$ and MUE = 0.55. An interesting observation is that this performance is even worse than that of XLOGP3-AA ($R^2 = 0.824$ and MUE = 0.59). It should be mentioned that there are no peptides in our training set since we remove all of them intentionally. This result indicates that if an irrelevant reference compound is used in log*P* computation, it will not necessarily lead to an improved prediction as compared to a pure additive model. We conduct another test by incorporating the oligopeptides in this test set into the training set so that XLOGP3 is able to use an appropriate reference compound in computation. After this treatment, XLOGP3 indeed demonstrates much improved accuracy with $R^2 = 0.932$ and MUE = 0.24 (Figure 7 and Table 4). This level of accuracy is very promising and clearly better than all of the other log*P* methods under our test. In particular, XLOGP3 also outperforms PLOGP,[37] a method especially developed for computing log*P* or log*D* values of oligopeptides. The power of XLOGP3 in handling congeneric compounds based on known knowledge is clearly demonstrated in this test.

DISCUSSION

On the Advantages over Conventional Additive Methods. The new strategy for log*P* computation implemented in XLOGP3 has some obvious advantages over conventional additive methods. The first advantage is that it relies less on the additivity assumption. XLOGP3 computes the log*P* value of a given compound by using the known log*P* value of a reference compound as a starting point. Nonadditive features, which are presumably the primary origin of errors in log*P* computation, will cancel out at least on the common substructures between the query compound and the reference compound (Figure 2), which will result in a reduced error. One would expect that the assumption of additivity tends to fail on complex molecules. In Figure 8, we plot the average errors of XLOGP3, XLOGP3-AA, and CLOGP on our training set as a function of molecular size. One can see that additive methods, i.e., XLOGP3-AA and CLOGP, generally produce larger errors on larger molecules. In contrast, the accuracy of XLOGP3 is not very sensitive to molecular size, which justifies our statement above.

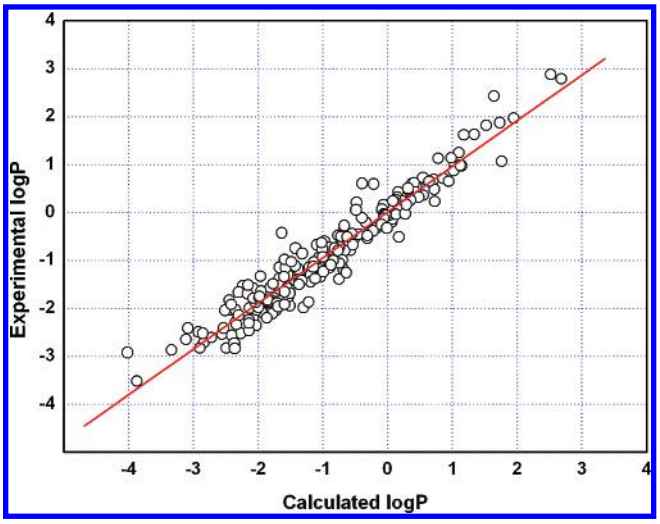

Figure 7. Experimentally measured log*P* values versus predicted values by XLOGP3 for 219 oligopeptides ($R^2 = 0.932$, RMSE = 0.32, MUE = 0.24).

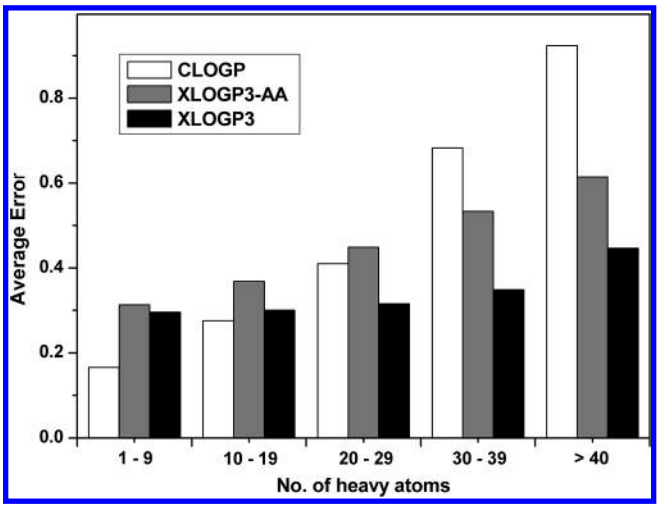

Figure 8. Average errors as a function of molecular size observed on the training set.

The improved performance of our strategy can also be explained conceptually as the following. According to XLOGP3

$$\log P^X_{\text{calc}}(\text{XLOGP3}) = \log P^Y_{\text{exp}} + \log P^X_{\text{calc}}(AA) - \log P^Y_{\text{calc}}(AA) \quad (5)$$

where *X* and *Y* denote the query compound and the reference compound, respectively, and *AA* represents a pure additive model, e.g., XLOGP3-AA. The above equation can be rewritten as

$$\log P^X_{\text{calc}}(\text{XLOGP3}) = \log P^Y_{\text{exp}} + [\log P^X_{\text{exp}} + \Delta E^X_{\text{calc}-\text{exp}}(AA)] - [\log P^Y_{\text{exp}} + \Delta E^Y_{\text{calc}-\text{exp}}(AA)] =$$

$$\log P^X_{\text{exp}} + \Delta E^X_{\text{calc}-\text{exp}}(AA) - \Delta E^Y_{\text{calc}-\text{exp}}(AA) \quad (6)$$

Thus, the absolute error of XLOGP3 on *X* is

$$\log P^X_{\text{calc}}(\text{XLOGP3}) - \log P^X_{\text{exp}} = \Delta E^X_{\text{calc}-\text{exp}}(AA) - \Delta E^Y_{\text{calc}-\text{exp}}(AA) \quad (7)$$

In comparison, the absolute error of a conventional additive model on X is

$$\log P^{X}_{\mathrm{calc}}(AA) - \log P^{X}_{\mathrm{exp}} = \Delta E^{X}_{\mathrm{calc-exp}}(AA) \qquad (8)$$

Thus, if the absolute errors produced by an additive model on X and Y have the same sign and are also at the same range, the absolute error of XLOGP3 (eq 7) will be smaller than the one produced by an additive model (eq 8). It is reasonable to expect that the above assumption is more likely to be valid if the reference compound resembles the query compound more closely. Currently, a similarity threshold of 50% is applied in XLOGP3. This similarity threshold certainly can be raised once a more comprehensive knowledge set is available. We expect that this will help XLOGP3 make even more accurate predictions.

The second advantage of our strategy is that it does not rely on an extensive classification of fragments and correction factors in order to improve accuracy. The additive model in XLOGP3 uses only 89 descriptors, a very modest number as compared to hundreds to thousands of descriptors used in some other additive methods. In particular, it seems to be critical for conventional additive methods to use corrections factors to achieve an acceptable level of accuracy. Some correction factors have relatively clear physical meanings, while others simply serve as "patches" when computed values deviate significantly from experimental values. In fact, a correction factor is always associated with a special pattern in molecular structure. In our method, if a query compound has such a pattern, the reference compound selected by similarity search, ideally, should have the same pattern. If so, the contribution of this pattern to log$P$ will cancel out between the query compound and the reference compound and thus does not need to be considered explicitly using a correction factor. This explains why our method does not need many correction factors. Furthermore, in a conventional additive method, a special structural pattern will draw attention and be handled with a correction factor only if it is observed on a good number of compounds. Otherwise, its impact on log$P$ will not be statistically significant in regression analysis and thus will be neglected. In contrast, the impact of a special structural pattern on the query compound may be taken into account by our method even if only one compound in the knowledge set contains this particular pattern. From this point of view, using the known log$P$ value of a structural analog as reference is in principle more effective than using correction factors.

The third advantage of our strategy is its ability of utilizing the ever-increasing log$P$ data more effectively. XLOGP3 can be considered as an interpolation method. It is well-known that the predicted values by an interpolation method will be more accurate if a greater number of known points exist in the solution space. The knowledge set used by the current release of XLOGP3 includes all of the compounds in our training set and two test sets, a total of nearly 9000 organic compounds with known log$P$ values. It is built as an external module so that it can be expanded conveniently once more log$P$ data are available. If this knowledge set is expanded, XLOGP3 will certainly have a better chance to find an appropriate reference compound for a query compound, and it in turn will lead to an improved accuracy. Experimentally measured log$P$ data are certainly increasing constantly. XLOGP3 will benefit from this expansion almost in a proportional manner. In contrast, conventional additive methods use experimental log$P$ data primarily for deducing the contribution of each descriptor through regression analyses. It is a common sense that once the size of the training set exceeds what is enough, a regression model will tend to converge, and its quality will not continue to improve automatically.

Using an external knowledge set brings XLOGP3 another appealing feature. Some researchers may maintain their in-house collections of log$P$ data. For understandable reasons, such data are not always available to public. We provide auxiliary tools in the XLOGP3 package so that the users may process their own data sets into the format accepted by XLOGP3 and then supply them as the knowledge set used by XLOGP3. The importance of this feature should not be underestimated since in-house log$P$ data are perhaps many times more than publicly available data. In contrast, conventional additive methods are normally provided to users as is and lack the ability of utilizing users' in-house data.

On Other Approaches with Similar Ideas. Development of XLOGP3 is inspired by some pioneering studies. For example, the *k*-nearest neighbors algorithm (*k*-NN) is a popular method for classifying objects based on the closest training examples in a feature space, which has long been applied to QSAR studies.[41,42] In XLOGP3, the so-called reference compound is in fact the nearest neighbor of a query compound in the structural space. A standard *k*-NN approach, however, predicts the features of a new object by weighing the corresponding features of its closest training examples, while XLOGP3 quantifies the difference between a query compound and its reference compound with a mathematical model.

In the field of log$P$ computation, some studies have successfully applied similar ideas as ours. Two of them should be mentioned in particular. The first one is the "Experimental Value Adjusted (EVA)" algorithm proposed by Meylan and Howard with their KOWWIN method.[23] The basic idea is to use an analog compound with known log$P$ as reference, while the difference between the query compound and its reference compound is computed by an additive model. They reported that EVA produced improved results over the pure additive KOWWIN method. Nevertheless, they did not describe in their publication how to choose appropriate analog compounds in order to apply EVA. They mentioned that "*our current computer program ... requires the user to manually select and enter the similar compound and known logP*". The EVA algorithm is not implemented in the online demo version of KOWWIN evaluated in our study. It is not clear to us if the EVA algorithm has been automated in any other releases of KOWWIN. Recently, Sedykh and Klopman published a new method for log$P$ computation.[25] Their method also adopts the same idea as the EVA algorithm. A major new development is that they have provided an automatic approach for selecting an analog to the query compound based on structural similarity. They also reported that this new method led to improved performance over conventional additive methods.

Our method is conceptually similar to Klopman's method although it is conceived independently. Both methods follow the idea originally proposed by Meylan and Howard; both

methods compute the similarity between two organic compounds by counting the common substructures on them; and both methods search among a large number of compounds with known log*P* values for reference compounds. Nevertheless, our method is technically different from Klopman's method virtually at every aspect. Compared to Klopman's method, the most remarkable feature of XLOGP3 is its simplicity. First, Klopman's method uses 102 fragments and 36 correction factors in its additive model; while XLOGP3 only uses 87 fragments and two correction factors. As we have discussed in this article, using a reference compound in log*P* computation should make most correction factors unnecessary. It seems that Klopman's method has not fully taken advantage of this feature. Second, although both Klopman's method and XLOGP3 decompose chemical structures into substructures, the size of substructure itself is an adjustable parameter in Klopman's method; while in XLOGP3, each substructure is formed uniformly by four non-hydrogen atoms, i.e., a topological torsion descriptor. Third, the training set used in Klopman's study is carefully selected so that every compound has at least one structural analog under certain similarity thresholds. Selection of this training set seems to be critical for achieving the reported accuracy in Klopman's study.[25] In our study, we emphasize on XLOGP3's ability of utilizing external data sets. The knowledge set used by XLOGP3 is simply an assembly of diverse organic compounds with known log*P* data. In practice, external data sets supplied by users of XLOGP3 do not need any special editing either.

A comparative evaluation of Klopman's method on our test sets has not been made in our study since it is not available to us. According to Klopman's report, their method produces standard deviations of 0.50−0.97 units on some independent test sets, which are smaller than the ones produced by CLOGP up to 0.20 units. Based on the results obtained in our own tests (Tables 2−4), we believe that the overall accuracy of XLOGP3 is at least comparable to that of Klopman's method. As for empirical methods, a simpler method is certainly more appealing if it is able to produce comparable results as a more sophisticated method. A simpler method is also easier for other researchers to reproduce.

Finally, we would like to point out that the strategy implemented in XLOGP3 can also be applied to the computation of other physicochemical properties of organic compounds. For example, Ralph Kühne et al. recently reported a similar approach to the computation of water solubility.[43] In fact, this strategy is in principle applicable wherever a large quantity of experimental data has been accumulated.

CONCLUSION

We have developed a new method for log*P* computation, i.e., XLOGP3. XLOGP3 computes the log*P* value of a query compound by using the known log*P* value of a reference compound as a starting point. The difference between the query compound and the reference compound is then estimated by an additive method. This strategy has several obvious advantages over conventional additive methods. First, it relies less on the assumption of additivity. Second, it in principle does not need a more and more extensive classification of fragments and correction factors in order to improve accuracy. Third, it is able to utilize the ever-increasing log*P* data effectively. Our tests demonstrate that XLOGP3 produces more accurate results than its predecessor as well as some other methods. We believe that XLOGP3 and similar approaches like KOWWIN-EVA and Klopman's method collectively represent an inspiring direction for log*P* computation.

ACKNOWLEDGMENT

The authors are grateful for the financial support from the Chinese National Natural Science Foundation (Grant No. 20502031), the Chinese Ministry of Science and Technology (the 863 high-tech project, Grant No. 2006AA02Z337), and the Science and Technology Commission of Shanghai Municipality (the Pu-Jiang Talents program, Grant No. 06PJ14115). Technical aid provided by Chunni Lu and Weiqi Zhang at the Shanghai Institute of Organic Chemistry is also appreciated.

Supporting Information Available: Detailed descriptions of the atom/group typing scheme and correction factors implemented in XLOGP3. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(2) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.
(3) Leo, A. J.; Hansch, C.; Elkins, D. Partition Coefficients and their Uses. *Chem. Rev.* **1971**, *71*, 525−616.
(4) Hansch, C.; Leo, A. *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology;* American Chemical Society: Washington, DC, 1995.
(5) Leo, A. J.; Hansch, C. Role of Hydrophobic Effects in Mechanistic QSAR. *Perspect. Drug Discovery Des.* **1999**, *17*, 1−25.
(6) van de Waterbeemd, H.; Grifford, E. ADMET In Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(7) Clark, D. E.; Grootenhuis, P. D. Progress in Computational Methods for the Prediction of ADMET Properties. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 382−390.
(8) Beresford, A. P.; Segall, M.; Tarbit, M. H. In Silico Prediction of ADME Properties: Are We Making Progress? *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 36−42.
(9) Davis, A. M.; Riley, R. J. Predictive ADMET Studies, the Challenges and the Opportunities. *Curr. Opin. Chem. Biol.* **2004**, *8*, 378−386.
(10) 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 Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(11) Leo, A. J. Calculating $\log P_{oct}$ from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.
(12) Carrupt, P. A.; Testa, B.; Gaillard, P. Computational Approaches to Lipophilicity: Methods and Applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley: New York, 1997; Vol. 11, pp 241−315.
(13) Eros, D.; Kovesdi, I.; Orfi, L.; Takacs-Novak, K.; Acsady, G.; Keri, G. Reliability of logP Predictions Based on Calculated Molecular Descriptors: A Critical Review. *Curr. Med. Chem.* **2002**, *9*, 1819−1829.
(14) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(15) Tetko, I. V.; Bruneau, P. Application of ALOGPS to Predict 1-Octanol/Water Distribution Coefficients, logP, and logD, of AstraZeneca In-House Database. *J. Pharm. Sci.* **2004**, *93*, 3103−3110.
(16) Chou, J. T.; Jurs, P. C. Computer-Assisted Computation of Partition Coefficients from Molecular Structures Using Fragment Constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172−178.
(17) Ghose, A. K.; Crippen, G. M. Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(18) Viswanadhan, V. N.; Ghose, A. K.; Reyankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(19) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(20) Petrauskas, A. A.; Kolovanov, E. A. ACD/LogP Method Description. *Perspect. Drug Discovery Des.* **2000**, *19*, 99−116.
(21) Walker, M. J. Training ACD/LogP with Experimental Data. *QSAR Comb. Sci.* **2004**, *23*, 515−520.
(22) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(23) Meylan, W. M.; Howard, P. H. Estimating logP with Atom/Fragments and Water Solubility with logP. *Perspect. Drug Discovery Des.* **2000**, *19*, 67−84.
(24) Zhu, H.; Sedykh, A.; Chakravarti, S. K.; Klopman, G. A New Group Contribution Approach to the Calculation of LogP. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 3−9.
(25) Sedykn, A. Y.; Klopman, G. A Structural Analogue Approach to the Prediction of the Octanol-Water Partition Coefficient. *J. Chem. Inf. Model.* **2006**, *46*, 1598−1603.
(26) 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.
(27) Mannhold, R.; Petrauskas, A. Substructure versus Whole-molecule Approaches for Calculating LogP. *QSAR Comb. Sci.* **2003**, *22*, 466−475.
(28) Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−621.
(29) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-additive Method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47−66.
(30) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(31) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(32) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR, Hydrophobic Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995; Vol. 2, pp 3−193.
(33) *The SYBYL software (version 7.2)*; Tripos Inc.: St. Louis, MO 63144, U.S.A.
(34) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.
(35) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(36) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Search. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983−996.
(37) Tao, P.; Wang, R.; Lai, L. Calculating Partition Coefficients of Peptides by the Addition Method. *J. Mol. Model*. **1999**, *5*, 189−195.
(38) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−552.
(39) Gombar, V. K.; Enslein, K. Assessment of n-octanol/water Partition Coefficient: When Is the Assessment Reliable? *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127−1134.
(40) *Physical/Chemical Property Database (PHYSPROP)*; SRC Environmental Science Center: Syracuse, NY, 1994.
(41) Itskowitz, P.; Tropsha, A. k Nearest Neighbors QSAR Modeling as a Variational Problem: Theory and Applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.
(42) Ajmani, S.; Jadhav, K.; Kulkarni, S. A. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* **2006**, *46*, 24−31.
(43) Kühne, R.; Ebert, R. U.; Schüürmann, G. Model Selection Based on Structural Similarity-Method Description and Application to Water Solubility Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 636−641.

CI700257Y