Generalized Fragment-Substructure Based Property Prediction Method

Matthew Clark*

Locus Pharmaceuticals Four Valley Square, 512 Township Line Road, Blue Bell, Pennsylvania 19422

Received August 17, 2004

The need for fast and accurate predictors of pharmaceutically important properties has been increasing due to pressure from high-throughput screening, in-silico screening, and the need to more rapidly identify potential pharmacokinetic issues before drugs advance to the more expensive clinical development stages. A novel method for making predictive models based on decomposing 2D structure into component structural fragments is used to model logP, water solubility, and melting point. The fragment orientation of the method facilitates understanding of how molecules might be altered to improve the desired properties. The 2D structure-based descriptor is computed by analysis of the target molecules with a substructure searching algorithm and a set of fragments selected for chemical and pharmaceutical relevance. These are combined with partial least squares to create predictive models. The correlation coefficients achieved are 0.86 for logP (SE = 0.68), 0.73 for logS (SE = 0.89), and 0.64 (SE = 48.9°) for melting point over diverse data sets of 11 447, 2427, and 5598 molecules, respectively. The models were verified via test sets of compounds not included in the training set.

INTRODUCTION

Interest in predicting ADME properties has been steadily increasing since the advent of high-throughout screening. Increasingly compounds proposed for synthesis or libraries must have the required physical properties to proceed to synthesis and optimization. Many models have been created for solubility, logP, and other properties separately. In this work we propose a model building method that can be applied uniformly to a wide range of properties based on the 2D structures of molecules.

Interest in solubility has been increasing dramatically recently, possibly driven by discussions of solubility problems in high throughput screening libraries. Many methods have been explored in the past five years for predicting solubilities and logP. These include 3D methods by Gasteiger,[1] one- and two-dimensional fingerprints,[2] neural networks,[3] Jur's ADAPT, multiple-regression,[4] QSPR,[5] fuzzy logic,[6] and topological descriptors.[7]

Methods for logP calculations have also abounded.[8,9] Much less has been done for melting point.[10,11] Melting point is significant in the pharmaceutical context as a surrogate for the cohesive energy of a solid that must be broken in order to dissolve a substance. This has been a harder issue to address than computation of the solvation energy of single molecules. Several models for melting point have been created; however, they have generally been for specific classes of compounds.[12,14]

The goal of using a fragment-based method for property prediction is to create a general method that can be applied to a large range of ADME properties as well as bioactivity to allow accurate estimation of properties.

Table 1 summarizes some recent approaches reported for prediction of physical properties. It is useful to categorize the descriptors into two classes, those using various computed property descriptors, and those using structure-based groups. The first class relies on a variety of popular descriptors to encode structural and electronic properties; bit strings based on topological paths, polar surface area, etc. Large numbers of these descriptors are used with correlation methods to extract models to predict the desired properties.

The second class of descriptor, and possibly the oldest, is the group and atom contribution. The original logP calculations and QSAR studies of Hansch typify early work using these methods.[42] These methods have the ability to give insight into the causality of the property and suggest modifications to the atoms and groups that will lead to improved properties. The relative disadvantage is that the training set must be large enough to provide statistically significant numbers of each fragment to develop a predictive model. This work describes a method to decompose molecules into a series of overlapping fragments which are then used with partial least squares regression to create a predictive model that relates the occurrence of each fragment to a property. Given a diverse training set the method can create a reliable model that also provides insight into avenues to optimize desired properties.

METHODS

Data Sets. Experimental data were derived from the PHYSPROP database, a large commercial database of over 25 000 compounds and associated measured and estimated physical properties in SD file format.[17] It contains a range of organic molecules, which include drug-like molecules as well as pesticides and industrial chemicals. The alternating single-double bond formalism was used instead of assigning an aromatic bond type. This data set was filtered to remove inorganic substances, those that represented mixtures of isomers, and those with formal charges noted in the SD file. Hydrogen atoms were added to fill all valences using the Cerius2 program.[18]

* Corresponding author e-mail: mclark@locuspharma.com.

10.1021/ci049744c CCC: $30.25
Published on Web 01/05/2005

Table 1. Summary of Recent Physical Property Models

ref	property predicted	data	no. of structures	method	statistics	result
Yan and Gasteiger[1]	solubility	Huuskonen[7]	1297	radial distribution	MLR	$r^2 = 0.79$
					neural net	$r^2 = 0.93$
Butina and Gola[13]	solubility	PHYSPROP[17]	3328	proprietary fingerprint descriptors	PLS	$r^2 = 0.71$
					cubist	$r^2 = 0.8$
Engkvist and Wrede[2]	solubility	PHYSPROP	5530	topological and physicochemical descriptors	neural net	$r^2 = 0.95$ test set $r^2 = 0.79$
Bruneau[3]	solubility	Huuskonen	1297	topological and physicochemical descriptors	neural net	$r^2 = 0.94$ SE 1.0
Hou[14]	solubility	Huuskonen	1297	atom contribution	MLR	$r^2 = 0.96$
Wildman and Crippen[41]	logP	Hansch[15]	9920	atom contribution	MLR	$r^2 = 0.918$
Xing[16]	logP	CRC	592	fingerprints, descriptors	PLS	$r^2 = 0.893$ SE = 0.653
Katritzky[5]	MP	CRC;	443	CODESSA descriptors	MLR	$r^2 = 0.84$
Yalkowsky[30]	MP	CRC	596	group contribution	MLR	$r^2 = 0.98$ SE 35 °C;
Yalkowsky[29]	MP	various	1600	group contribution	MLR	$r^2 = 0.99$ SE 37 °C

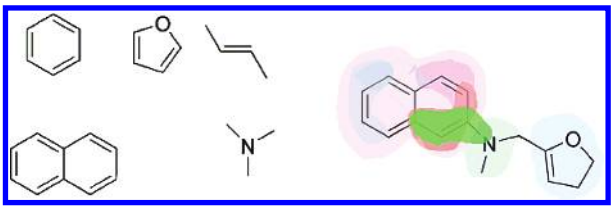

Figure 1. Example of fragments and overlapping instances in a molecule.

The data were filtered to remove compounds with data at the extrema of the measurements; compounds with logP values greater than 10 or less than −10 were removed from the data set. Compounds with solubility reported less than 1.0e-8 mg/L were removed, as were compounds with solubilities > 1e6 mg/L. In addition, only solubility values measured or estimated for the range between 20 and 30 °C were used in the study. Solubility data were converted from mg/L to molarity for the studies, and modeling was performed in log of the concentration. Only compounds with melting points above −50 °C and below 300 °C were used in the melting point model.

For validation 10% of each data set was selected randomly and removed from the training data to form test sets. Training models were made from these reduced data sets, and predictions were then made using the models created from the training sets.

Substructure Searching. The substructure search was applied iteratively to find all unique, nonoverlapping instances of each fragment in the molecule and the scores were summed. For example, the benzene fragment is found twice in biphenyl, and therefore the score for the benzene fragment would be 2.0 for biphenyl. Thus, the score marks not only the presence of the fragment but also the number of occurrences as well. This process was repeated for each fragment used in the analysis. Since the fragments are not "orthogonal", they form overlapping counts throughout the molecule. Figure 1 illustrates that the substructure search of the naphthyl moiety contains the benzene substructure as well as several butene substructure elements.

The score for each fragment was then accumulated for each compound to form a data matrix with columns for the target property and the total scores for each fragment and with each row corresponding to a molecule in the training set.

Correlations. The kernel partial least-squares method (PLS)[22] as implemented in the R package[23] was used to generate linear statistical models based on the fragment scores and the property to be predicted. The models were run with cross-validation to determine the optimal number of latent variables, and then a final model was produced using this optimal number.

The statistical models were built using 261 descriptors, one for each fragment, correlating to the dependent property. This resulted in a linear equation with a coefficient for each fragment that was used to predict properties as shown in eq 1.

$$\text{property} = [\sum_{\text{fragments}} \text{count}_{\text{frag}} * \text{coeff}_{\text{frag}}] + \text{offset}_{\text{property}} \quad (1)$$

Partial least squares has been shown to be very effective at separating noise from data in tests of chance correlation.[24] Indeed it was found that the real risk is in missing correlations rather than finding correlations in random data. When strong, cross-validated, correlations are found with PLS they are significant.

Fragments. Two classes of fragments were used. The first class of fragments is designed to find basic chemical functionalities, such as carbonyl group, primary, secondary, tertiary amines, and other fundamental organic moieties. The second group is composed of larger fragments selected for relevance to pharmaceutical properties—both physical and therapeutic.

RESULTS

Impact of the Number of Fragments. To investigate the impact of using various number of fragments, the fragments were classified into subsets and models for the log P were created with each subset. The first subset of 18 fragments consisted only of simple hydrocarbon motifs—primary, secondary, tertiary carbons, and various multiple bonds. The second set augmented this with motifs involving oxygen, primary, secondary alcohols, ketone, aldehyde, carboxylic acids, and oxygen-containing rings for a total of 37. The next set of fragments included the nitrogen analogues, bringing the total to 61. The next augmentation to 105 fragments includes more complex drug-like motifs, such as indole and derivatives as well as a variety of more complex

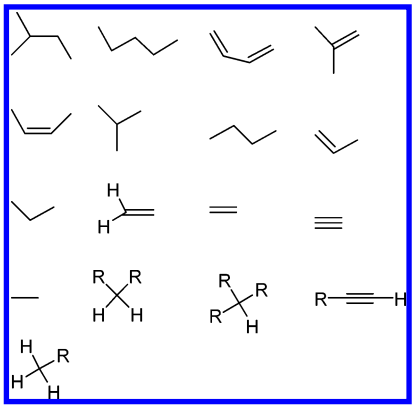

Figure 2. Examples of basic chemical functionalities (R = heavy atom).

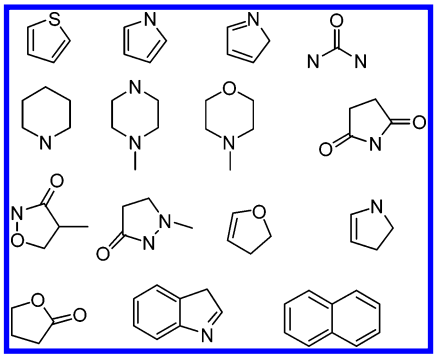

Figure 3. Examples of biorelevant functionalities.

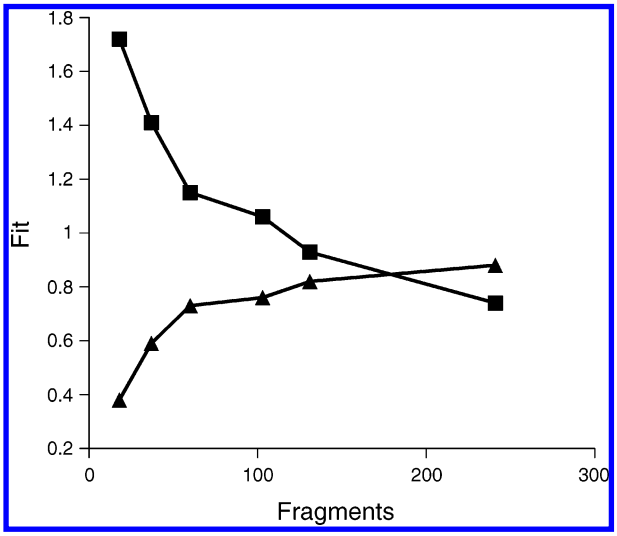

Figure 4. r^2 (△) and standard error (■) of prediction vs number of fragments used in log P model.

ring systems. The largest fragment set of 261 compounds includes more complex systems used for drug discovery. Figure 4 shows the increase in correlation and decrease in standard error associated with the increasing number of fragments in the solubility model.

LogP Model. LogP was identified as a critical parameter for ADME very early; the seminal work of Hansch in 1964 focused on predicting this value.[42] This work resulted in fragment-based methods typified by ClogP[25] as well as atom-type based methods typified by the AlogP method.[40]

The fragment based logP model was created using 11 447 compounds and 261 descriptor fragments. The r^2 produced was 0.876, and the standard error of prediction was 0.68 log

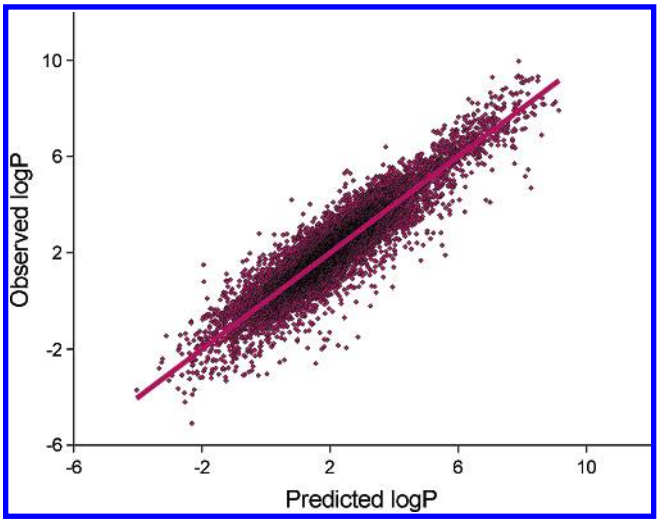

Figure 5. Predicted vs actual log P for training data.

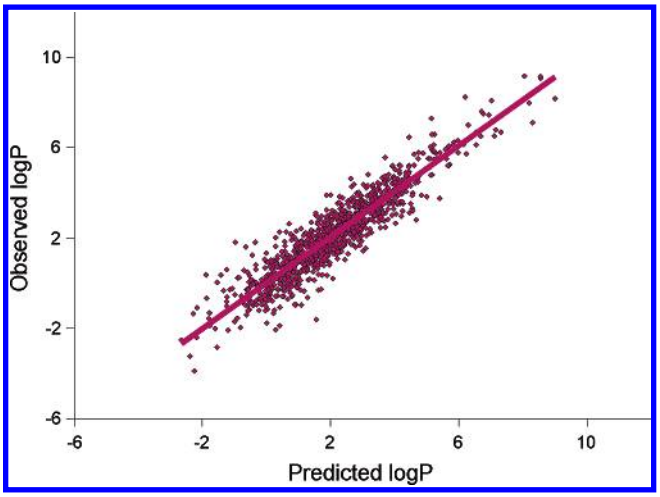

Figure 6. Predicted vs actual log P for test data.

units. The predicted vs actual data are shown in Figure 5. The results from the test set, composed of data not included in the training set, which provides a statistical fit equivalent to the training set, are shown in Figure 6; the statistics are summarized in Table 2.

Solubility Model. The solubility model was made using 2427 compounds and 257 fragments to predict the log of the solubility measured in molarity. The r^2 of this model was 0.73, and the standard error was 0.89. The predicted vs actual data are shown in Figure 7. The prediction on the test set is shown in Figure 8.

Hexatriacontane, with a PHYSPROP logS of −5.47 but predicted logS of −13.95, has the most underpredicted solubility. Hexatriacontane is a straight-chain hydrocarbon with 36 carbon atoms; the solubility has been reported to be 1.7e-3 mg/L,[27] as compared to the value of 1.802 mg/L provided by the PHYSPROP database. This error in the decimal point suggests the true logS should be −8.47, which is more consistent with the prediction.

The compounds used in the model and test set include a large number of compounds that are not considered drug-like. To test the predictive ability of the model a data set of more drug-like molecules used by Huuskonen and others was predicted.[7,14] The predictions, shown in Figure 9, provide statistics similar to the training and test data with a r^2 of 0.84 and a standard error of 0.82 log units.

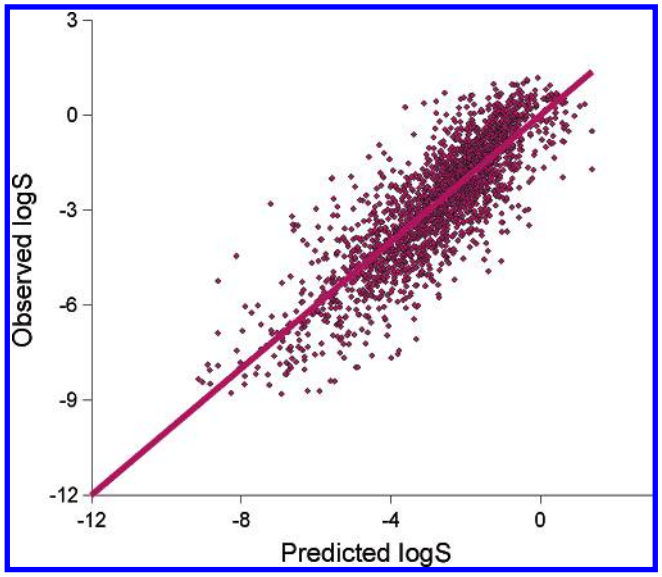

Figure 7. Predicted vs actual log solubility (molar) for the training set.

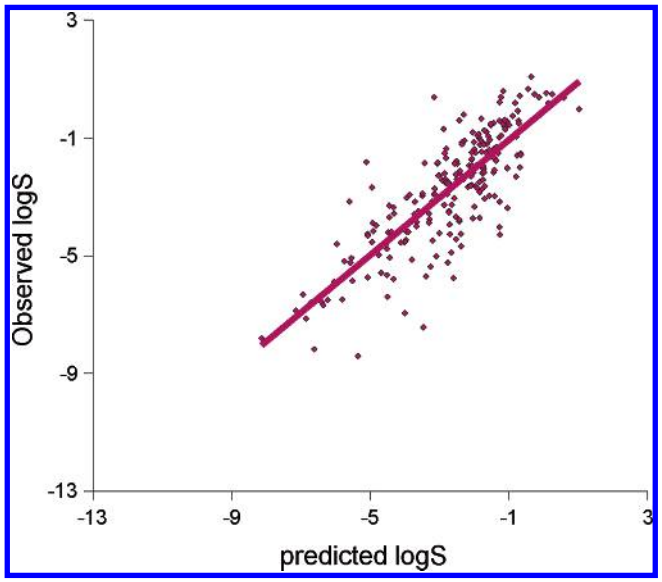

Figure 8. Predicted vs actual log solubility (molar) for the test set.

Table 2. Summary of Predictions for Test Sets

	n	data range	data SD	r^2	SE
logP	1149	13	1.8	0.86	0.67
log solubility	230	10	1.93	0.67	1.1
melting point	658	384 K	78 K	0.61	48.9 K

Melting Point Model. The energy required to break up the lattice energy of a solid is a significant contributor to the solubility of a compound. Thus the melting point is a surrogate measure for the cohesive force of the solid lattice free energy.[28] The melting point model was created from the 5598 compounds in the PHYSPROP database for which melting points were recorded above −50 °C and below 300 °C. The fragmentation and correlation process was carried out, and the model was used to predict the melting points of the compounds. The results are shown in Figure 10. The r^2 correlation coefficient for this model is 0.64, and the standard error of prediction is 49 °C. The statistics for the test set are shown in Figure 11 and summarized in Table 2. While the standard error is not impressive in comparison to the accuracy

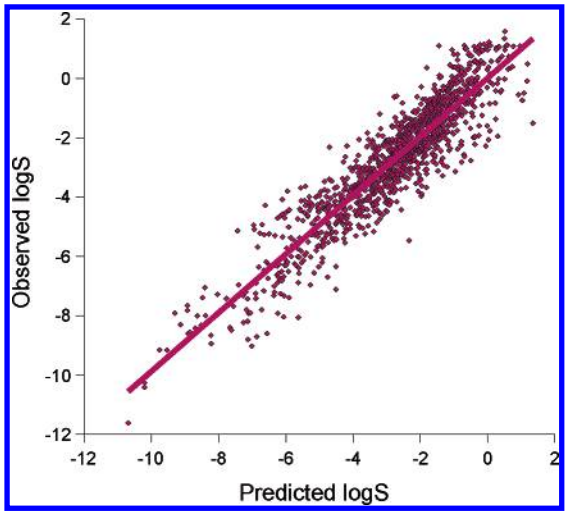

Figure 9. Predicted vs actual log solubility (molar) for the "Huuskonen" data set.

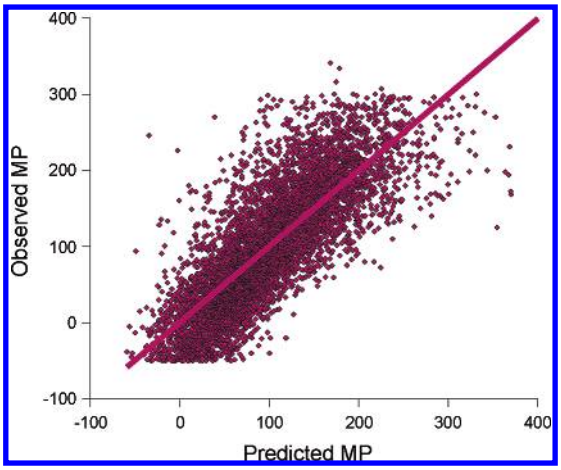

Figure 10. Predicted vs observed melting points (°C) for the training set.

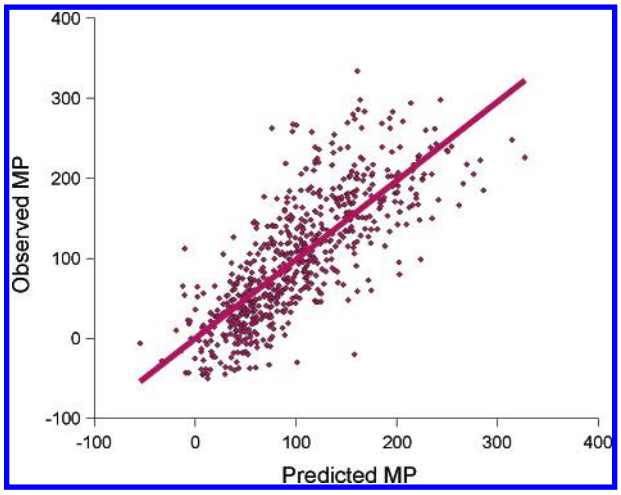

Figure 11. Predicted vs observed melting points (°C) for the test set.

possible in the measurement of melting points, it is comparable to models made with very specific classes of compounds. These models provided standard errors of 37°, for aromatic compounds,[29] and 35° for aliphatic non-hydrogen bonding compounds.[30]

Table 3 summarizes the statistical fits for the three properties. The count, *n*, of molecules used to make the

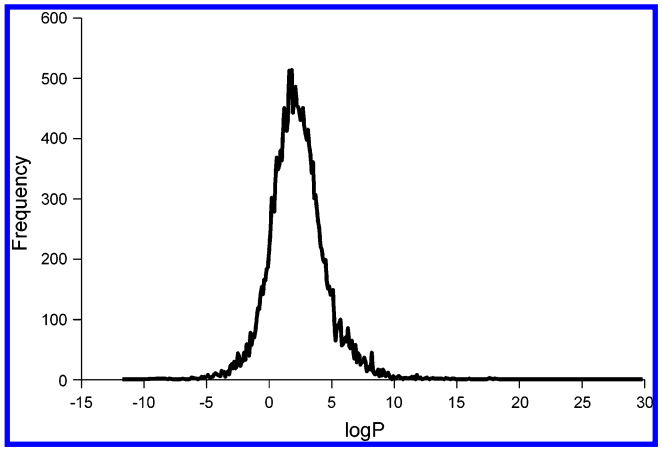

Figure 12. Distribution of logP data.

Table 3. Summary of Predicted Properties

	n	data range	data SD	r^2	SE
logP	11447	15	1.82	0.86	0.68
log solubility	2426	11	1.93	0.73	0.89
melting point	5598	471 K	80 K	0.64	48.9 K

model, the range, and standard deviation, SD, of the data are presented with the r^2 correlation and standard error, SE.

DISCUSSION

Data Set. To understand and evaluate the models, the training set must also be evaluated. If the standard deviation of the training set is small, assigning the mean of the property value to all compounds can result in a model with a low standard error of prediction. In other words, the narrower the distribution of the training set, the better a model can be made; however, its ability to predict outside of the distribution may be limited. Therefore the nature of the training set must be examined.

Analysis of the PHYSPROP data set is especially worthy since it has been used by several researchers to create the models outlined in Table 1. The data set contains a large diversity of organic and inorganic molecules. Even when the inorganic molecules are filtered out, the range of organic molecules is quite large and not limited to drug-like compounds. The data contain cases of typographical and other errors in the data. It is interesting that the model itself is a good detector of these errors as in the cases of hydroxylamine and hexatriacontane examples discussed above. In both cases it appears that an error was made on the exponent, and these were among the largest outliers. In the data preparation many errors in the structure files, and structure files that contained several isomers as a single unit, were detected by examination of the outliers; the data were checked against other sources before removing these outliers.

Several researchers have extracted the "drug-like" molecules to produce subsets that may be more similar to those that will be predicted. However, philosophically the attributes of drug-likeness are subjective and potentially biasing to the data. Table 3 summarizes the data range and standard deviation of the data sets used in this study.Figure 12 shows the distribution of logP in the data set. For the entire data set the standard deviation is 1.82 log units. Predicting all compounds to have a logP of 3 would result in a standard error of 1.82; therefore, this is a metric for the efficacy of

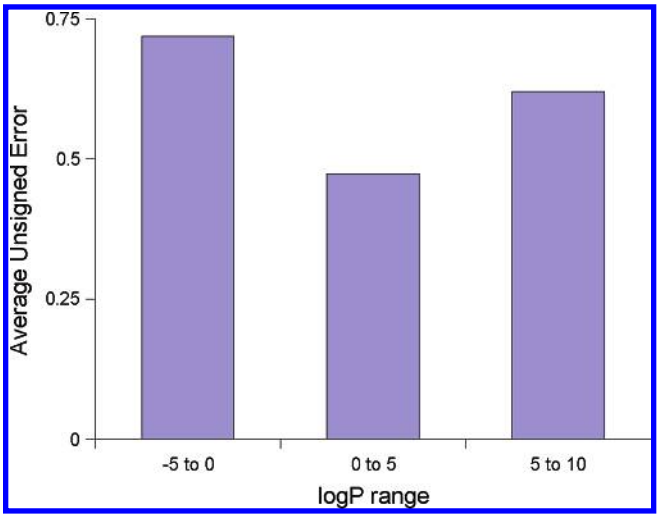

Figure 13. Average logP prediction error vs logP.

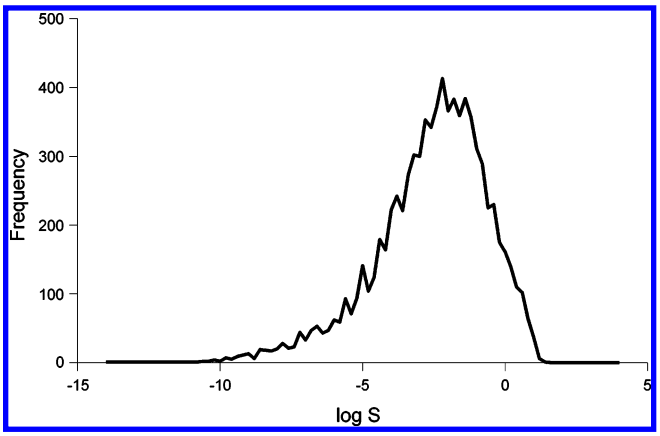

Figure 14. Distribution of logS (molar) data.

the model. The standard error of the model obtained in this study is smaller than the standard deviation of the data, 0.68 log units, suggesting that the model is useful. Moreover, as Figure 13 illustrates, the error of prediction increases only moderately at the extrema of the predictions. These errors may be due to both measurement errors and more limited statistical samples at the extrema.Figure 14 shows the distribution of the logS data. Here the standard deviation is about 1.9, and the standard error of prediction from the model is half of that, 0.89 log units. The data are skewed since there is an upper limit to the solubility. Among the most soluble compounds are trimethylamine and formaldehyde with solubilities at 1.18 log units, or about 15 molar solutions.

The use of the same data set by the numerous authors outlined in Table 1 provides an interesting opportunity to draw conclusions about the data itself. At some level the ability to make a correlation will be limited by the precision and accuracy of the measurements themselves. The correlations appear to approach an r^2 of 0.9, and a standard error of about 1 log unit. It is possible that the error in the measurements are near this level, creating a ceiling for all models that use this data set.The melting point data distribution is shown in Figure 15. The models were made on data selected in the −50 °C to 300 °C range. The standard deviation makes this the statistically most diverse data set. It is interesting that in all data sets studied here the standard error of prediction is approximately half of the standard deviation of the data. In this case the standard error is 56

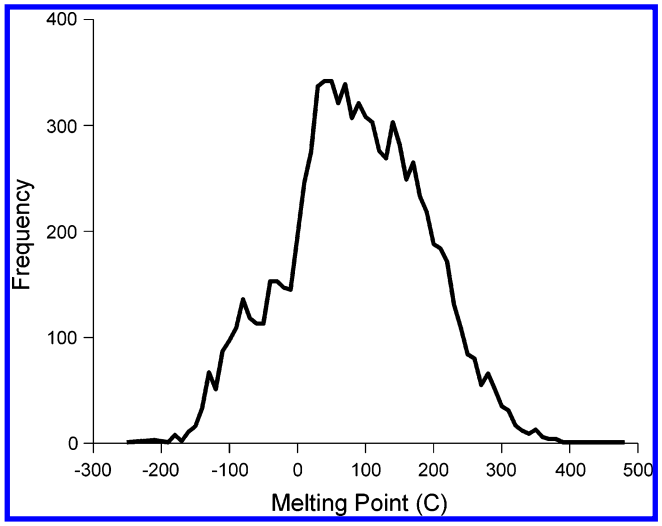

Figure 15. Distribution of melting point data.

°C. As in the other models, the error is nearly constant throughout the predicted range, as illustrated in Figure 10.

Of the three parameters of melting point, solubility, and logP, there are actually only two independent variables. When plotted against each other in 3D, the data points clearly align in a plane. This is in agreement with the "general solubility equation" developed by Yalkowski which relates the molar solubility, melting point, and octanol−water partition coefficient of organic solids:[31]

$$\log S_w^{solid} = 0.5 - 0.01(\text{MP } (^\circ\text{C}) - 25) - \log K_{ow}$$

The observed data in the PHYSPROP data set are in reasonable agreement with this relationship; the correlation coefficient of the PHYSPROP data for the relation in eq 3 is similar to those in the models of this work again suggesting that the variance in the reported data may be a limiting factor in creating predictive models with this resource.

The observation that a small number of intrinsic variables can explain a large number of observed properties was also made by Cramer in 1980.[32,33] In that work it was observed that 95% of the variance of four to 10 diverse physical properties of liquids could be explained with two latent variables, five variables provided the ability to make accurate predictions over a more diverse set. The interrelation of these properties can be used to both build models and to further explore the self-consistency of the data.

Fragment Based "Fingerprinting". The use of overlapping substructures is fundamentally different from methods that assign atom types or identify discrete substructures. The overlapping substructures provides more variation than counting only complete substructures.

Fingerprinting methods based on bit-strings, especially those found in commercial software systems, have been quite popular over the last five years.[34−36] However, in many cases these fingerprints were designed to speed the searching process rather than to describe molecules. Fragment based fingerprints as well as atom types provide an intuitive basis for similarity and analysis as well as allowing focus on chemical moieties of relevance to drug discovery. The contributions of the fragments can be analyzed in an intuitive way, and indeed properties can be optimized by fragment

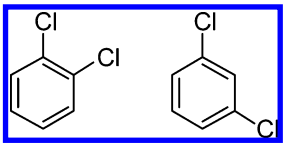

Figure 16. 1,2- and 1,3-dichlorobenzene.

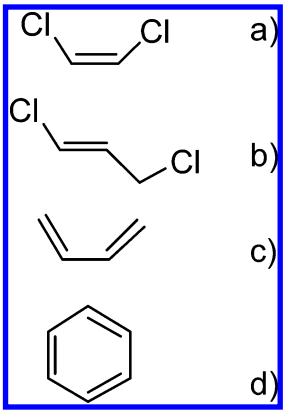

Figure 17. Example fragments.

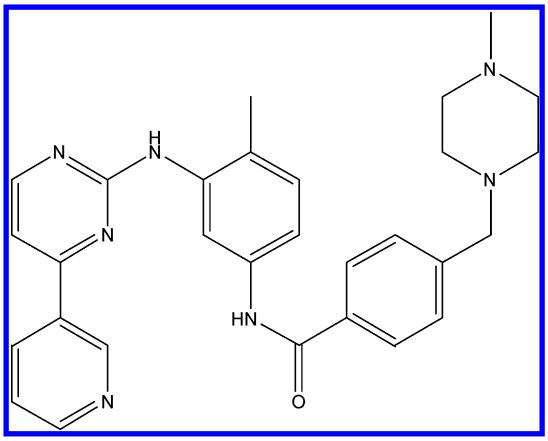

Figure 18. Imatinib (Gleevec).

changes. The overlapping fragment scores also take into account new connection patterns made when substitutions are made.

The use of fragments and atom types has had historical weaknesses. Gasteiger has pointed out that the use of fragments for prediction may not take into account the importance of how they are connected together.[37] However, the use of overlapping fragments allows distinction among molecules with the same components but substituted in different positions. For example, Figure 16 shows two isomers of dichlorobenzene. The use of overlapping fragments allows the descriptors for these two molecules to differ and allow, correctly, different property predictions. Figure 17 shows examples of fragments that would allow this.

While both molecules include one d) substructure, and 3 instances of substructure c) from Figure 17, 1,2-dichlorobenzene includes substructure a) but not b), and 1,3-dichlorobenzene includes b) but not the a) substructure. Judicious choice of fragments can allow fine distinction among a variety of isomers.

It is useful to examine a specific example of a fragment based prediction of an actual compound. Figure 18 shows the structure of imatinib, commercially known as Gleevec. Table 4 shows the most impactful fragments based on contribution to overall logS, which is the product of the count

Table 4. Fragment Based Prediction of the Solubility of Imatinib

fragment	count	coefficient	contribution	cumulative logS
C	29	−0.18	−5.23	−5.65
ethylene	9	−0.2	−1.76	−7.4
N	7	−0.21	−1.46	−8.86
dimethylamine	2	0.47	0.94	−7.92
trimethylamine	2	0.43	0.86	−7.06
pyridine	1	0.78	0.78	−6.28
2RNH	2	0.21	0.43	−5.79
RN═R	2	0.2	0.39	−5.36
R3N	2	0.18	0.36	−4.97
ethane	2	−0.18	0.35	−4.61
ethylamine	2	0.16	0.32	−4.97
aniline	1	0.28	0.28	−4.65
propene	2	0.25	0.25	−4.37
R═O	1	0.24	0.24	−4.12
1.4-disubstituted benzene	1	0.16	0.16	−3.96
1,3-butadiene	3	−0.05	−0.16	−4.12
2-aminopyrimidine	1	−0.12	−0.12	−4.24
R2CH2	2	0.05	0.09	−4.14
N-methylpiperazine	1	0.07	0.07	−4.07
benzene	2	−0.03	−0.06	−4.13
toluene	2	−0.03	−0.06	−4.2

[a] Includes offset of −0.42.

Table 5. Top 10 Fragments Impacting Solubility

fragment	coefficient	occurrences in training set
pyridine	0.78	103
naphthalene	−0.76	153
Br	−0.75	195
I	−0.62	53
S═O(═O)	0.56	154
acetone	0.51	80
dimethylamine	0.47	283
trimethylamine	0.43	107
ROH	0.42	1161
S	−0.42	518

Table 6. Top 10 Fragments Impacting logP

fragment	coefficient	occurrences in training set
decalin	−1.54	80
S═O(═O)	−1.53	886
1,3,5-triazine	1.28	124
beta lactone	−1.14	32
norbornane	0.96	44
R═O(═O)	0.92	1231
C(═O)O	0.88	3067
gem-dimethylcyclopentane	−0.83	222
pyridazine	−0.83	77
iodobenzene	−0.82	97

found in the molecule, and coefficient for the fragment. The total predicted solubility is −4.2 log units of the concentration in molarity. This compares well to literature descriptions that the logS is about −4.1 at pH 7.4.[38] Imatinib is often delivered as the mesylate to provide improved solubility over the neutral compound.

Among the observations is that the size of the molecule, as denoted by the number of carbon atoms, is the most impactful issue in the solubility model. Hydrophobic variations such as the ethylene substructure contribute to lowering the solubility while the inclusion of substructures of polar fragments such as amines and pyridine increase solubility. The count of nitrogen atoms alone appears to correlate with decreased solubility. The fragment *N*-methyl piperazine is predicted to contribute 0.07 log units of solubility to the molecule.

In general 2D structure-based descriptors have predominated for properties not highly dependent on conformation, such as solubility, while descriptors based on 3D structure have been widely used to predict obviously conformationally dependent properties such as protein−ligand binding. The reduction to two dimensions largely reduces the ability to rank properties that depend on the 3D shape. The main attractiveness of 2D structure-based descriptors is elimination of the need to determine relevant conformations for the property being predicted. The continuum of properties and their dependence on 3D conformation is illustrated in this study by the difference in fit between the logP, which is mildly affected by conformation, and the melting point which has a high dependence on the conformation and 3D aspects for packing into solid structures. The results reported are in agreement in that the ranking of models by r^2 is logP, solubility, and melting point. This is also roughly the ranking of the significance of the 3D structure to these properties and can be explained by examination of the relative influence of the 3D vs 2D structure for each property. In the logP model the variation of the lowest free energy conformations between octanol and water are lost in a 2D structure-based descriptor. For solubility, the difference between the condensed phase (solid or liquid) and aqueous solution conformation is more significant due to the larger change in the nature of these states than in logP. For melting point, the change in nature of the solid phase and liquid phase is even larger, since the cohesive force of the solid may favor a conformation far different than that found in the liquid and the overall entropy of the transition may depend on factors other than the 2D structure, and thus the prediction will be most dependent on 3D factors.

This does not intimate that 2D structure-based analysis is not useful or effective in ranking even protein−ligand interactions. A recent review of using 2D fingerprints to enrich libraries suggests strongly that while absolute prediction of activities may be difficult due to the diversity of interaction modes, 2D structure is related to bioactivity.[39] This same study suggests weaknesses of the fingerprinting methods in commercial software which may be reduced by the fragment-based method described here.

Tables 5−7 compare the fragments with the largest coefficients in the final predictive equations. The "R" in the formulas designates a heavy atom, otherwise any "free" valence may be a hydrogen or heavy atom in the match. The 2-amino pyridine and sulfonyl moieties appear significant in both solubility and logP. Otherwise there is no obvious commonality among the fragments with the largest coefficients.The fragments having the highest impact on melting point predictions are given in Table 7.

The improvement of fit with the increase of the number of fragments has rarely been discussed in the context of ADME property predictions except as the resulting descriptors in linear regression models. In this case the improvement in r^2 shown in Figure 4 follows a nearly logarithmic path that appears to asymptotically approach 1 as the number of descriptors approaches the number of molecules in the training set. The curve suggest that for this model about 100 descriptors, or fragments, are enough to produce most of the correlation over the 20 000 diverse molecules in the data set. This is consistent in general with the numbers of atom types used in molecular refractivity predictions[40] and the

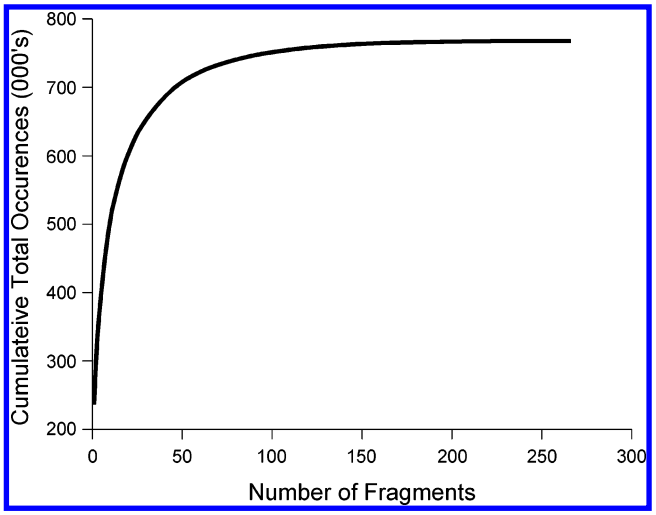

Figure 19. Cumulative total occurrences vs number of fragments.

Table 7. Top 10 Fragments Impacting Melting Point

fragment	coefficient	occurrences in training set
RNH_2	38.92	818
R_2NH	30.71	1206
ROH	29.38	2770
C	18.33	67796
S	17.9	959
C#N	−17.72	189
trimethylamine	17.14	378
neopentane	17.06	427
1,2-dichlorobenzene	16.38	147
I	15.42	161

AlogP calculation[41] which use approximately 100 types. Figure 19 shows a cumulative graph of the number of substructure matches for the logP subset of PHYSPROP vs the number of fragments.

The number of matches reaches 98% of the total in the first 100 fragments. The most common matched fragment, a carbon atom, is responsible for 18% of the total matches. In fingerprints designed for database searching the emphasis is generally on ways to identify unique attributes to filter out most of the molecules, and therefore common substructures such as carbon atoms or benzene rings are not considered information rich. In the current application, however, even the carbon atom is significant as the descriptor counts the *number* of carbon atoms in the molecule and relates that to the final properties (e.g. the number of carbon and other atoms makes molecular weight an implicit descriptor). For the logP model the carbon atom alone has a coefficient of 0.13, which confirms the intuition that in general having more carbon atoms increases logP and was counted 137 425 times in the data set. However the less common moieties are responsible for much of the differentiation among the molecules, as illustrated in Tables 5−7.

The melting and boiling point predictors of Yalkowsky used 40 and 24 fragments, although these models covered specific chemical classes.[20,29]

logP. logP has been studied extensively. High quality data sets have been curated to ensure that models were made on precise and accurate data. Indeed, the original work of Hansch provides the earliest examples of fragment-based correlation for property prediction.[42] Recent work comparing ClogP and new methods by Xing[16] showed a r^2 of 0.89 and a standard error of 0.66 log units. This is slightly better than the result of this work, but the model results of Xing were for a test set of 40 drug-like molecules and is therefore more limited in scope than the model presented here based on over 11 000 data points.

Solubility. The solubility model produced by this method is comparable to the works cited in Table 1, even though nearly the entire data set was used. The statistics of the model for predicting the solubilities of the compounds in the Huuskonen data set are equivalent to that of the larger data set. This gives confidence in the generality of the model for predicting solubility of both drug-like and nondrug-like compounds. In some respects it is surprising the model performs as well as it does because the data include the solubility of both solids and liquids. In the case of solids the solubility is also influenced by the lattice energy and entropy of the solid. The melting point model suggests that this contribution is modeled only poorly.

The correlation coefficient is somewhat less than that reported in recent atom-based models, where correlations of up to 0.96 were reported with a standard error of 0.59 log units.[14] However that study was on a selected subset of 1290 compounds.

Melting Point. The melting point model points out the limitations of this method in that melting point may not be well described by the 2D-structure. Although the melting point predictions are relatively poor, they are comparable to those reported previously using other methods. The errors reported in models limited to specific classes were approximately 36 °C; this study described here sacrifices some of the accuracy but provides a dramatically larger range of prediction in both diversity and temperature range. While the standard error of 49 °C is over 10 times the accuracy of measurement, the range of structures in this model is larger than previous studies.

CONCLUSIONS

The prediction of properties using the maximum common substructure score and partial least squares statistics is a powerful and general method for creating models to predict properties based on 2D structures. The method can be applied to a number of properties previously studied separately and may provide accurate predictive models for additional properties.

The predictive models created using our technique compare well to recently reported models. The precision of the data in the training set now widely used for solubility and logP models may be the limiting factor for the accuracy of these models as nearly all workers appear to be converging to the same level of accuracy when using the data set used in this study.

ACKNOWLEDGMENT

The author wishes to thank Tony Hopfinger of the University of Illinois at Chicago and Jeff Wiseman of Locus Pharmaceuticals for many insightful suggestions and comments while preparing the text.

Supporting Information Available: The fragment structures in SD format and the names of the molecules selected from PHYSPROP for the training and test sets. This

material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *4*, 429−434.

(2) Engkvist, O.; Wrede, P. High-Throughput, *In Silico* Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1247−1249.

(3) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*(6), 1605−1616.

(4) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1237−1247.

(5) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806−1814.

(6) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP Based on Quantitative Structure−Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177−1207.

(7) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(8) Zou, J.-W.; Zhao, W.-N.; Shang, Z.-C.; Huang, M.-L.; Guo, M.; Yu, Q.-S. A Quantitative Structure−Property Relationship Analysis of logP for Disubstituted Benzenes. *J. Phys. Chem. A* **2002**, *106*, 11550−11557.

(9) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. J. *Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.

(10) Wen, X.; Qiang, Y. Group Vector Space Method for Estimating Melting and Boiling Points of Organic Compounds. *Ind. Eng. Chem. Res.* **2002**, *41*, 5534−5537.

(11) Joback, K. G.; Reid, R. C. Estimation of Pure Component Properties From Group Contributions. *Chem. Eng. Commun.* **1987**, *57*, 233−243.

(12) Wei, J. Boiling Points and Melting Points of Chlorofluorocarbons. *Ind. Eng. Chem. Res*. **2000**, *39*, 3116−3119.

(13) Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 837−841.

(14) 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.

(15) BioByte Corp. Pomona, CA 1998.

(16) Xing, L.; Glen, R. C. Novel Methods for Prediction of logP, pK_a and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.

(17) PHYSPROP, Syracuse Research Corporation, Syracuse NY www.syrres.com

(18) Cerius2 Accellrys Inc. 9685 Scranton Road, San Diego CA 92121.

(19) Ullman, J. R. An algorithm for subgraph isomorphism. *J. ACM* **1976**, *23*, 31−42.

(20) Raymond, J. W.; Garidner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.

(21) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5−13.

(22) Stahle, L. Wold, S. *Prog. Med. Chem.* **1988**, *25*, 292−337.

(23) Venables, W. N.; Smith D. M. R Programming Environment for Data Analysis and Graphics 1.9.1 http://www.r-project.org/2004.

(24) Clark, M.; Cramer, R. D., III The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137−145.

(25) Leo, A. Calculating log Poct from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.

(26) *The Merck Index*, 11th ed.; Budavari, S., Ed.; Merck & Co.: Rahway, 1989; p 766, entry 4759.

(27) Baker, E. G. *Am. Chem. Soc. Div. Petrol. Chem., Preprints-Symposia* **1956**, *1*, No. 2, 5.

(28) Yalkowksy, S. H. Valvani, S. C.; Solubility and Partitioning I: Solubility of Nonelectrolytes in Water. *J. Pharm. Sci*. **1980**, *69*, 912−922.

(29) Simamora, P. Yalkowsky, S. H. Group contribution methods for predicting the Melting Points and Boiling Points of Aromatic Compounds. *Ing. Eng. Chem. Res.* **1994**, *33*, 1405−1409.

(30) Krzynaiak, J. F.; Myrdal, P. B.; Simamora, P. H. Yalkowsky, S. H. Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530−2535.

(31) Ran, Y.; Yalkowski, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Coput. Sci.* **2001**, *41*, 354−357.

(32) Cramer, R. D. BC(DEF) Parameters 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. *J. Am. Chem. Soc*. **1980**, *102*, 1837−1849.

(33) Cramer, R. D. BC(DEF) Parameters 2. An Empirical Based Structure-Based Scheme for the Prediction of Some Physical Properties. *J. Am. Chem. Soc*. **1980**, *102*, 1849−1859.

(34) Johnson, M.; Maggiora, G. *Concepts and Applications of Molecular Similarity*; 1990.

(35) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(36) Dean, P. M. *Molecular Similarity in Drug Design*; Chapman and Hall: 1994.

(37) Gastieger, J. Structure Representation for Toxicology Prediction ADMET1 Conference 2004, San Diego CA.

(38) *The Merck Index−An Encyclopedia of Chemicals, Drugs, and Biologicals*, 13th ed., O'Neil, M. I., Ed., Merck and Co., Inc.: Whitehouse Station, NJ, 2001.

(39) Martin, Y. C.; Kofron, J. J. Tarphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(40) Ghose, A. K.; Crippen G. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(41) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. J. *Chem. Inf. Comput. Sci*. **1999**, *39*, 868−873.

(42) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175.

CI049744C