

# Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility

Diansong Zhou, Yun Alelyunas, and Ruifeng Liu\*

Department of Development DMPK & Bioanalysis and Department of Chemistry AstraZeneca,  
1800 Concord Pike, Wilmington, Delaware 19850

Received January 18, 2008

QSPR studies, using scores of SciTegic's extended connectivity fingerprint as raw descriptors, were extended to the prediction of melting points and aqueous solubility of organic compounds. Robust partial least-squares models were developed that perform as well as the best published QSPR models for structurally diverse organic compounds. Satisfactory performance of the QSPR models indicates that the scores of extended connectivity fingerprint are high performance molecular descriptors for QSAR/QSPR studies. Performance of the fingerprint-based descriptors is further validated by the satisfactory prediction of aqueous solubility of nearly 1300 organic compounds (squared correlation coefficient of 0.83 and RMSE of 0.85 log unit) with Yalkowsky's general solubility equation using both calculated melting points and calculated octanol–water partition coefficients. It demonstrates for the first time that it is feasible to predict aqueous solubility of structurally diverse organic compounds with the general solubility equation using both the calculated melting points and the partition coefficients.

## INTRODUCTION

With the aim of bringing new drugs to market faster and at a lower cost, in silico prediction of molecular properties is playing increasingly more important roles in modern drug discovery.<sup>1,2</sup> Compounds designed for drug discovery programs are routinely profiled by in silico models before they are synthesized. Obviously the quality of prediction models is of crucial importance. Among other things, the quality of a prediction model depends on the molecular descriptors that quantify structural features relevant to the properties under investigation. Many molecular descriptors were designed to capture different aspects of structural information.<sup>3</sup> Among them, molecular fingerprints give perhaps the most detailed description of molecular structures, as evidenced by the fact that they are routinely used and very efficient in molecular similarity analysis.<sup>4</sup> However, molecular fingerprints are rarely used as descriptors in quantitative structure–activity/property (QSAR/QSPR) studies, because conventional fingerprints encode the presence of structural features but not the frequency that a structural feature appears in a molecule, and high performance fingerprint features are usually overlapping and correlated.

Recently we explored the feasibility of using the scores of SciTegic's extended connectivity fingerprint (ECFP) as descriptors in the QSPR study of lipophilicity.<sup>5</sup> The results are encouraging as a very robust logP prediction model was derived based on the partial least-squares (PLS) technique.<sup>6</sup> In this paper, we report the results of our study using the same approach for the prediction of melting points and aqueous solubility of organic compounds. The objective is to better understand the performance of the fingerprint scores as QSPR descriptors and to gain insight into intrinsic factors affecting QSPR prediction of the important physical properties.

Aqueous solubility is one of the most extensively studied properties in the QSAR community.<sup>7,8</sup> From a large number of publications on this topic, it is clear that even though some models may perform reasonably for certain classes of compounds, reliability of existing models still do not meet drug discovery needs.<sup>9</sup> Although melting point is also an important physical property and it is intrinsically related to solubility,<sup>10</sup> there have been only a few QSPR studies on it.<sup>11–17</sup> Several recent publications on a large number of structurally diverse organic compounds indicate that currently the best QSPR models can predict the melting points of organic compounds with a standard error of about 40 °C.<sup>13,16,17</sup>

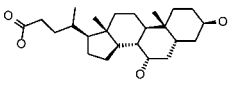
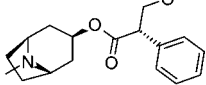
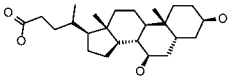
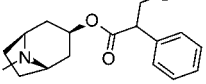
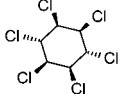
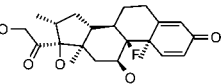
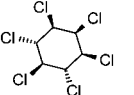
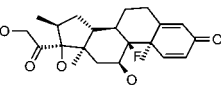
## COMPUTATIONAL DETAILS

**1. Statistical Method and Descriptor Selection.** Scores of SciTegic's ECFP<sub>2</sub> fingerprint,<sup>18</sup> called ECFC<sub>2</sub> fingerprint, are used as raw descriptors for both melting point and aqueous solubility in the current study. The ECFP<sub>2</sub> fingerprint encodes each atom and its molecular environment within a circle with a diameter of two chemical bonds.<sup>5</sup> As in our previous study, to balance performance and computational cost, the ECFC<sub>2</sub> fingerprints were folded, using the logical OR function, into a fixed length of 1024 bits.

Kernel-based PLS<sup>6</sup> as implemented in the R packages<sup>19</sup> was used in this study to overcome difficulties associated with a large number of descriptors and descriptor correlation. PLS can be considered as an extension of principal component analysis (PCA). It combines PCA and least-squares regression to extract information relevant to the property under investigation and reduce dimensionality. This is achieved by combining raw descriptors via a linear combination to produce orthogonal latent descriptors. Statistical analysis is applied to identify the most relevant latent descriptors to build the final model. In the present study,

\* Corresponding author e-mail: Ruifeng.Liu@astrazeneca.com.

**Table 1.** CAS Numbers, Molecular Structures, and logS of Stereoisomer Pairs in Huuskonen Data Set

CAS#	Structure	logS	CAS#	Structure	logS
128-13-2		-4.29	101-31-5		-1.91
474-25-9		-3.64	51-55-8		-2.12
319-86-8		-4.51	378-44-9		-3.77
58-89-9		-4.6	50-02-2		-3.64

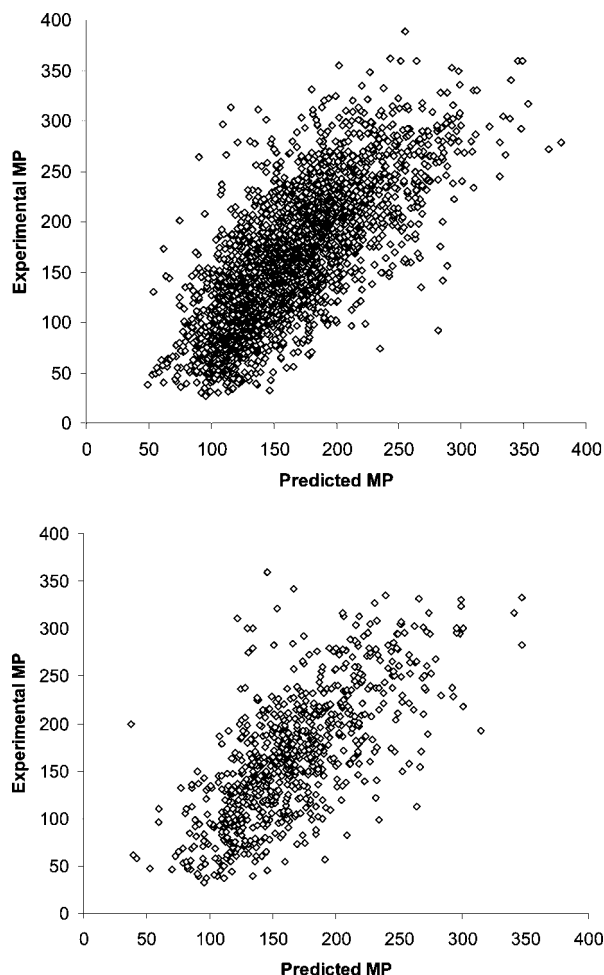
leave-10%-out cross-validation was performed. It was done by splitting the training set randomly into 10 equal-sized data sets, and nine of them were used to build a model. The model was used to predict the values of the tenth set left out in model building. The deviations between the predicted and the experimental values were calculated for the left-out set. This process was repeated until each and every set was left out once. The root mean square error (RMSE) of prediction and the estimate of predictive ability,<sup>20</sup>  $q^2$ , were calculated. They are used as criteria for finding the appropriate number of latent variables to build the final PLS model.

**2. Melting Point Data Set.** To build a robust QSPR model, one needs a quality training set of significant size and sufficient molecular structural diversity. Most early QSPR studies of melting points used small training sets up to a few hundred compounds with limited structural diversity. Only very recently studies were published on training set size of a few thousand compounds. Among them, Clark<sup>13</sup> used over 5000 compounds retrieved from the PHYSPROP database.<sup>21</sup> Bender et al. used over 4000 compounds retrieved from the Molecular Diversity Preservation International (MDPI) database.<sup>22</sup> In the current study, we used the same MDPI database, as it is free to any one via Web download and therefore easy for interested parties to compare the performance of different approaches. Not all compounds in the database with melting points were selected. Similar to Bender et al.,<sup>16,17</sup> we excluded inorganic compounds and organometallic compounds as well as compounds with a reported melting point range over 5 °C. We also excluded compounds denoted to sublime or decompose, compounds cocrystallized with organic solvents, and organic salts. The salts were removed because the number of them is small and not sufficient for a reliable model to be derived. This may be different from Bender et al., who retained organic salts but “washed” them by removing the counterions and standardized the organic ions (setting formal charge and protonate/deprotonate accordingly assuming pH = 7) to generate the parent molecules.<sup>16</sup> By doing so, they are effectively assuming that melting points of organic salts are the same as melting points of the corresponding parent organic molecules.

We further excluded stereoisomers with melting points more than 5 °C apart. This is because the descriptors we use are all 2D descriptors that do not encode stereochemistry information. For stereoisomers with melting points less than 5 °C apart, the structure with the higher melting point was excluded. This is different from the approach of Bender et al.<sup>16</sup> who retained stereoisomers and calculated some 3D descriptors based on a single conformer optimized by MMFF94 force field. As it is hard to judge that the single conformer generated by MMFF94 geometry optimization is reasonably close to the conformer of the compound in the solid state, we felt the approach of Bender et al. may bring in as much noise as useful 3D information. This is corroborated by the results of Bender et al. who observed that their 3D descriptors contain less relevant information than their 2D descriptors.<sup>16</sup>

After “cleaning” the experimental data set as described above, there are a total of 3804 structurally unique organic compounds with melting points between 27 and 389 °C and a molecular weight range from 84 to 1031. These compounds were randomly separated into a training set of 3000 compounds and a test set of 804 compounds.

**3. Aqueous Solubility Data Set.** The solubility data set used in this study is that of Huuskonen<sup>23</sup> who collected solubility data for ~1300 organic compounds from several sources. This data set was used in many QSPR studies.<sup>23–32</sup> Yan made some corrections and additions to the data set and made the data available at <http://www.vcclab.org>.<sup>27</sup> Data downloaded from this Web site at the end of 2006 have 1310 entries of SMILES strings with corresponding aqueous solubility values. Examination of the data indicates that there are four pairs of stereoisomers. Molecular structures of the stereoisomers, CAS registration numbers, and corresponding logS values are shown in Table 1. As mentioned above, since the descriptors we use are all 2D descriptors, they do not encode 3D information and therefore cannot distinguish the stereoisomers. We therefore excluded these eight compounds from the data set. Thus a total of 1302 structurally unique compounds are in the solubility data set of the current study.



**Figure 1.** Experimental and predicted MP of the training set (upper) and the test set (lower) compounds.

## RESULTS AND DISCUSSION

**1. Melting Point.** All calculations were performed using SciTegic's Pipeline Pilot.<sup>33</sup> The first step is to calculate raw molecular descriptors—the ECFC<sub>2</sub> fingerprints. For the training set of 3000 compounds, the calculation indicates that 142 of the 1024 fingerprint bits are off for all the molecules (i.e., the structural features these bits represent do not appear in any of the compounds). These 142 bits were removed because their values are zeros for all the molecules (zero variance descriptors). Thus the total number of raw descriptors used in this study is 882. Based on RMSE of prediction and estimate of predictive ability calculated from leave-10%-out cross-validation of the training set, the top nine latent descriptors were selected to build the final model. Melting points of the training set and the test set compounds were predicted by the nine-descriptor PLS model, and the results are compared with the experimental values in Figure 1.

Errors of prediction and correlation coefficients between the predicted and measured melting points are compared with those of Clark and Bender et al. in Table 2. Since different compounds are in the training and the test sets in different studies, data in Table 2 can only give an approximate performance measure of different approaches. However since a large number of structurally diverse compounds (on the order of thousands) were used in the studies, the data in Table 2 suggest that our PLS model, even though using only nine latent descriptors, is at least as good as the published

models. The performance of the PLS model indicates that the raw descriptors we use, the ECFC<sub>2</sub> fingerprint, captured relevant molecular structure information effectively and perform very well compared to the descriptors used in other studies for melting point prediction.

Melting point is a fundamental physical property. Its experimental measurement for organic compounds is relatively easy compared to measurement of many other physical properties. For most pure organic compounds, the uncertainty of melting point measurement is within a couple degrees. Table 2 shows that the uncertainty of QSPR predictions is on the order of 30–40 °C, much wider than the uncertainty of experimental measurements. We agree with Karthikeyan et al. that much of the model errors should be attributed to the fact that the single molecule-based descriptors used in the studies cannot properly account for long-range interactions such as intra- and intermolecular hydrogen bonding and crystal packing. These long-range interactions are crucial factors affecting melting points. To illustrate this point, molecular structures of a few compounds and their corresponding melting points retrieved from the MDPI database are shown in Figure 2.

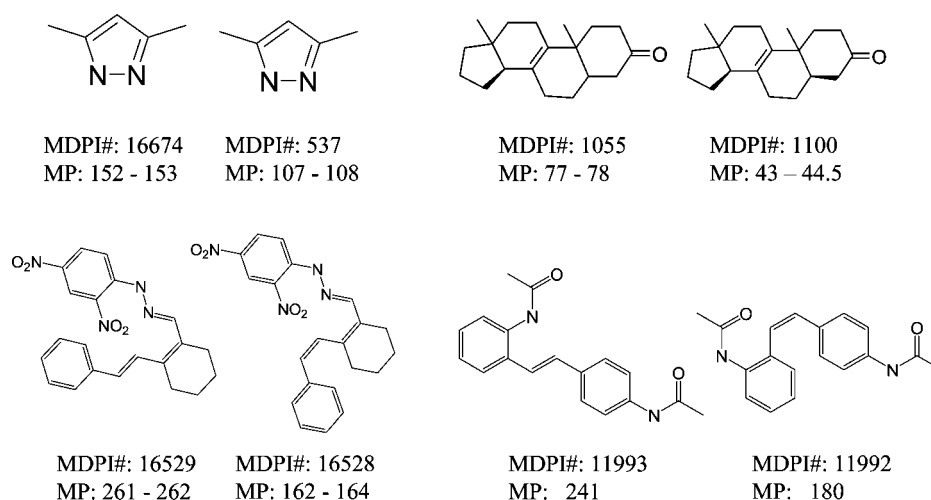
The first example is provided by the melting points of two crystalline forms of 3,5-dimethylpyrazole. This is a very simple molecule, yet the difference in melting points of the two crystal forms is about 50 °C. Without specifically modeling the effect of crystal packing, none of the QSPR approaches is able to predict the melting points of the different crystalline states. The second example is a pair of stereoisomers, MDPI# 1055 and MDPI# 1100, differing in chirality at a single carbon atom. The melting point difference for the two compounds is over 30 °C. As none of the 2D descriptors used in this study is able to distinguish the two stereoisomers, it is understandable that the PLS model is unable to deal with this situation. The last two pairs of compounds in Figure 2 are cis- and trans-isomer pairs involving benzene rings. In the MDPI database, there are over 15 pairs of cis–trans isomer pairs with aromatic ring substituents. In most of the cases the melting points of the trans isomers are higher than those of the cis isomers. This is easy to understand as the trans isomers are expected to be more planar because of spatial crowdedness of the cis form. As a result we can expect tighter  $\pi$ - $\pi$  stacking in the trans form and therefore higher crystal lattice energy. These examples highlight some of the issues theoretical approaches have to deal with in order to be able to predict melting points more accurately. Without taking care of these issues, data in Table 2 probably reflect what can be achieved in a QSPR approach for melting point prediction.

**2. Aqueous Solubility.** The experimental data set used in this study has only 1302 structurally diverse organic compounds. The number of raw ECFC<sub>2</sub> descriptors is up to 1024. Thus the experimental data set appears quite small. We attempted to randomly separate the data set into a 1000-compound training set and a 302-compound test set. However no matter how we do it randomly, there are always some compounds with unique structural features which appear in the test sets only. Because of missing ECFC<sub>2</sub> descriptors in the training set, the model derived from the training set should not be expected to perform well for these compounds unless the missing structural features are irrelevant to solubility. This is a well-known deficiency of the QSPR

**Table 2.** Comparison of Model Performance for Melting Point Prediction of Structurally Diverse Organic Compounds

	method	training set size	training set $R^2$	training set RMSE <sup>a</sup>	training set MAE <sup>b</sup>	test set size	test set $R^2$	test set RMSE	test set MAE
Karthikeyan <sup>c</sup>	ANN	2089	0.66	48.0	37.6	1042	0.66	49.3	38.2
Nigsch <sup>d</sup>	k-NN <sup>e</sup>	4119	0.49	46.2					
Clark <sup>f</sup>	PLS	5598	0.64	48.9		658	0.61	48.9	
current work	PLS	3000	0.54	43.4	33.7	804	0.47	48.3	37

<sup>a</sup> Root mean square error. <sup>b</sup> Mean absolute error. <sup>c</sup> Reference 16. <sup>d</sup> Reference 17. <sup>e</sup> K-nearest neighbor. <sup>f</sup> Reference 13.

**Figure 2.** Structures and melting points of some representative compounds with significantly different melting points in different crystalline states or different stereoisomers.

approach: the predictability of a model deteriorates for compounds away from the chemistry space of the training set. In order to build a model with broad applicability, we need to make sure the training set covers as much chemistry space as possible. It was achieved in two steps. In the first step, we clustered the 1302 compounds by an exclusive sphere method<sup>34</sup> using Daylight fingerprint and a Tanimoto similarity threshold of 0.70. Compounds with a similarity coefficient 0.70 or higher are grouped into a cluster. It generated 188 clusters. The cluster size ranges from 2 to 54 compounds. A total of 902 compounds are in the clusters. The rest of the compounds, 400 of them, are singletons (structurally dissimilar to other compounds). In the second step, all singletons are selected into the training set. In addition, we randomly selected one compound from each cluster into the training set. This resulted in a training set of 588 compounds. To better represent the chemistry space occupied by the clusters and arrive at a training set of 1000 compounds, we repeated random selection of compounds from each cluster a few more times. In the end, the training set includes all singletons and all compounds in clusters with less than 5 members. The compounds in the test set are leftovers from large clusters.

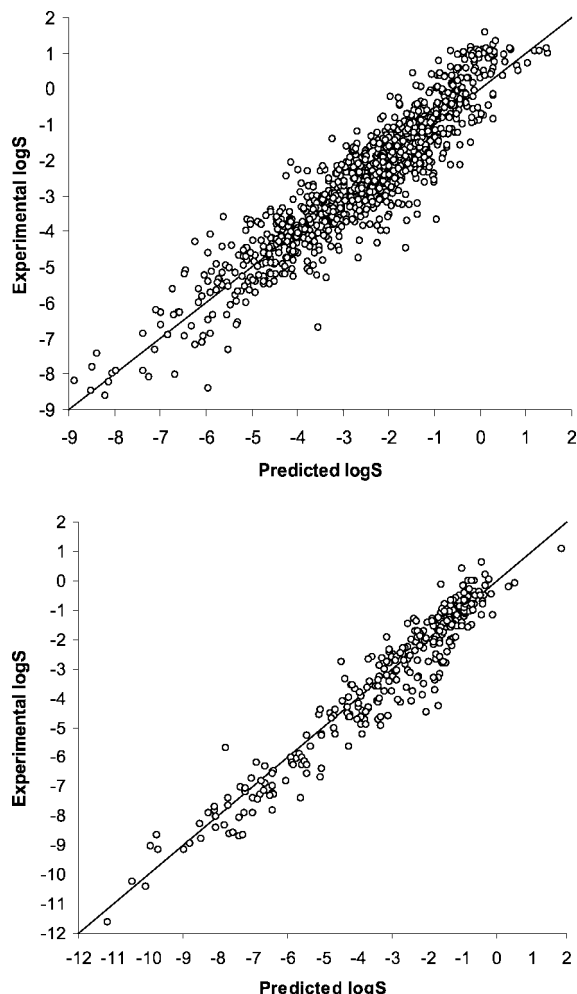
Leave-10%-out cross-validation under PLS framework identified again top nine latent descriptors as the most relevant for aqueous solubility prediction. The final PLS model for aqueous solubility was derived based on the nine latent descriptors and the 1000-compound training set. Calculated solubilities for the training and test set compounds are compared with their experimental values in Figure 3. It shows that even though only nine latent descriptors are used, the PLS model is very robust. The squared correlation coefficient between the calculated and experimental logS is

0.85 for the training set. The correlation coefficient for the test set appears better because the chemistry space of the test set compounds is well represented by the training set. Comparison of statistical parameters of some published QSPR models built on more or less the same data set is given in Table 3.

Two types of models are shown in Table 3: those based on artificial neural networks (ANN) and those based on least-squares regression (MLR and PLS). Based on squared correlation coefficients and standard errors of prediction, it seems ANN models outperform MLR and PLS based models. However it should be noted that ANN models employ a much higher number of regression parameters than MLR or PLS. Four out of the five ANN models employed over 100 hidden parameters. A large number of hidden parameters and a limited number of observations can lead to overfitting, which is a well-known challenge for artificial neural networks. Reliability of the experimental logS data used in this study is not known, but a conservative estimate is no better than a half-log unit. Katritzky et al. analyzed the experimental solubility values of 411 compounds (not the data set used in the current study) and concluded that the average standard deviation is 0.6 log unit.<sup>35</sup> We suspect that overtraining of the neural net might have played a role for some neural net models with a standard error of less than 0.6 log unit.

Table 3 shows that even though only nine latent descriptors were employed in the current PLS approach, the error of prediction of the PLS model is similar to those of the ANN and MLR models. The satisfactory performance of the PLS model indicates that ECFC\_2 fingerprints can serve as high performance molecular descriptors for QSPR study of aqueous solubility.





**Figure 3.** Comparison of the calculated and experimental solubility of the training set (upper) and the test set (lower).

**3. Relationship between Lipophilicity, Melting Point, and Aqueous Solubility.** Starting from the van't Hoff equation for the ideal solubility of a solid and assuming Walden's rule ( $\Delta S_m = 56.5 \text{ J/Kmol}$ ) applies, Yalkowsky derived a general solubility equation (GSE) for nonelectrolyte organic compounds relating lipophilicity, melting point, and aqueous solubility as given below:<sup>10</sup>

$$\text{LogS} = 0.5 - \text{logP} - 0.01(\text{MP} - 25)$$

In this equation, logP is the logarithm of n-octanol/water partition coefficient, MP is the melting point in °C, and logS is the logarithm of molar solubility. Using experimental melting points, predicted logP by the ClogP program, or experimentally measured logP values available, Yalkowsky et al. were able to predict solubilities of 380 nonelectrolyte organic compounds included in the Huuskonen data set with a root-mean-square error of 0.76 log unit.<sup>36</sup> This is significant because many reasonably good logP prediction models are available, and the melting point of an organic compound is easily measurable provided a sample is available. Therefore the aqueous solubility of an organic compound can be reasonably predicted without using information and regression parameters from aqueous solubilities themselves. On the other hand, in modern drug discovery we usually need to profile many molecular properties before synthesis is carried out. Thus it will be more useful, at least for drug design purposes, to be able to use predicted melting points in GSE.

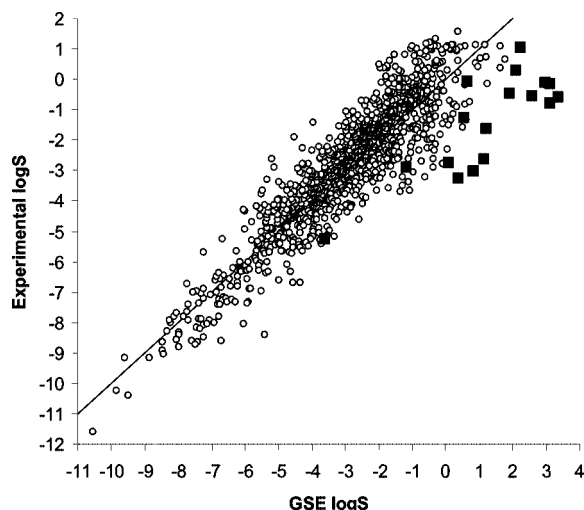
To investigate the feasibility of using predicted melting points and logP in GSE, we calculated melting points of the 1302 compounds in the solubility data set using the melting point model developed in this study and calculated the logP values for these compounds using the logP model we developed earlier.<sup>5</sup> The calculated MP and logP values were then used to estimate aqueous solubility using the GSE. LogS values for the 1302 compounds calculated by GSE are compared with the experimental values in Figure 4. Overall, the GSE calculated logS are in reasonable agreement with experimental values. The squared correlation coefficient is 0.81, and the root-mean-square error of prediction is 0.90 log unit. It should be noted that other than errors in calculated melting points and logP, the fact that some compounds in the data set are strong electrolytes also introduces significant errors, because the GSE was derived for nonelectrolytes. Indeed we found that solubilities of all amino acids (strong electrolytes) in the data set are overestimated by GSE. There are 17 amino acids in the data set. They are represented by the solid squares in Figure 4. When the 17 amino acids are excluded, the squared correlation coefficient for the 1285 compounds becomes 0.83, and the root-mean-square error is 0.85 log unit. Compared to the data in Table 3, it seems that GSE with predicted melting points and logP is only slightly worse than the best QSPR solubility models. The good performance not only further validates GSE for organic compounds but also validates performance of the logP and melting point prediction models and indicates that the ECFC fingerprints are high performance QSPR descriptors.

In the literature, there are numerous reports of QSPR studies utilizing powerful regression techniques to reproduce training set experimental values as close as possible, resulting in models that seem to be highly accurate but may be unreliable for nontraining set compounds. The validity of GSE helps highlight the reliability limit of QSPR models for solubility prediction. Figure 2 shows that there are plenty of cases crystal packing can make a difference in melting points of as much as over 100 °C. GSE shows that the impact on predicted solubility by every 10 °C error in melting point is 0.1 log unit. As the crystalline states of many experimental solubility data are unknown, uncertainty in experimental solubility data may well be higher than reported. Thus limits to the accuracy of conventional QSPR solubility prediction are both theoretical and experimental. From the theoretical point of view, one should not reasonably expect conventional QSPR solubility models to be highly accurate for structurally diverse solid compounds without taking into account the effects of crystal packing. Experimentally, crystalline states of solubility data are under-reported; this is especially true of data from pharmaceutical companies. Due to increasingly smaller scales of organic synthesis in the discovery setting of pharmaceutical companies, many new compounds synthesized are not recrystallized. Solubility is measured from whatever polymorph of the solid sample is on hand without consideration whether it is the only or most stable crystalline state. In reality, many organic compounds are known to exhibit polymorphism.<sup>37</sup> Considering the existence of polymorphism in organic compounds and the fact that melting depends strongly on intermolecular forces, the level of correlation between single molecule descriptors and melting points as shown in Figure 1 and previous publications<sup>13,16,17</sup> is perhaps the best one can reasonably expect. Any further

**Table 3.** Comparison of Solubility Models Derived from the Same Experimental Data Set

reference	method	no. of descriptors	no. of model parameters	comps in training set	training set $R^2$	training set RMSE	comps in test set	test set $R^2$	test set RMSE
Huuskonen <sup>a</sup>	ANN	30	372	884	0.94	0.47	413	0.92	0.60
Yan <sup>b</sup>	ANN	40	328	797	0.86	0.50	496	0.85	0.59
Wegner <sup>c</sup>	ANN	9	150	1016	0.94	0.52	253	0.93	0.54
Liu <sup>d</sup>	ANN	7	16	1033	0.86	0.70	258	0.86	0.71
Tetko <sup>e</sup>	ANN	33	136	879	0.94	0.47	412	0.91	0.60
Huuskonen <sup>a</sup>	MLR	30	30	884	0.89	0.67	413	0.88	0.71
Hou <sup>f</sup>	MLR	78	78	878	0.92	0.59	412	0.90	0.63
Yan <sup>b</sup>	MLR	40	40	797	0.79	0.93	496	0.82	0.79
Tetko <sup>e</sup>	MLR	33	33	879	0.86	0.75	412	0.85	0.81
current work	PLS	9	9	1000	0.85	0.71	302	0.93	0.71

<sup>a</sup> Reference 23. <sup>b</sup> Reference 27. <sup>c</sup> Reference 26. <sup>d</sup> Reference 24. <sup>e</sup> Reference 25. <sup>f</sup> Reference 29.

**Figure 4.** Comparison of GSE calculated and experimental logS of the solubility data set (solid squares are amino acids).

improvement will very likely require elaborate modeling of crystal forms. This helps to understand what can be reasonably achieved by a conventional QSPR study of aqueous solubility using single molecule descriptors.

### SUMMARY

Robust QSPR models for melting point and aqueous solubility of structurally diverse organic compounds were derived using partial least-squares technique and the ECFC fingerprint as raw descriptors. Satisfactory performance of these QSPR models demonstrates that the ECFC fingerprints are high performance molecular descriptors for QSPR studies. The good performance of the fingerprint-based descriptors can be attributed to the ability of the molecular fingerprints to give accurate description of molecular structures. Conventional QSPR descriptors are also based on molecular structures, but they usually reflect only certain aspects of molecular structure and are therefore inferior to fingerprint based descriptors.

The soundness of the QSPR model for melting point prediction developed from this study and the octanol–water partition coefficient prediction model developed in a previous study is validated by satisfactory prediction of aqueous solubility of a large number of organic compounds using Yalkowsky's general solubility equation. The relationship between aqueous solubility and melting point as revealed by the general solubility equation dictates that for more

accurate prediction of aqueous solubility of structurally diverse solid-state organic compounds, effects of crystal packing have to be taken into account by the prediction models. Conventional QSPR approaches using single molecule based descriptors are unlikely to be able to properly account for crystal packing effects and therefore cannot be expected to give a highly accurate prediction of melting points and aqueous solubility of structurally diverse compounds.

### ACKNOWLEDGMENT

We are grateful to members of SciTegic Technical Support team for their timely assistance in resolving many issues encountered in the course of this study.

### REFERENCES AND NOTES

- (1) Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 381–401.
- (2) Banik, G. M. In silico ADME-Tox prediction: The more, the merrier. *Curr. Drug Discovery* **2004**, *4*, 31–34.
- (3) Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363–372.
- (4) Herta, J.; Willet, P.; Wilton, D. J.; Acklinb, P.; Azzaouib, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- (5) Liu, R.; Zhou, D. Using Molecular Fingerprint as Descriptors in QSPR Study of Lipophilicity. *J. Chem. Inf. Model.* **2008**, *48*, 542–549.
- (6) Wehrens, R. and Mevik, B. PLS: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). R package version 2.0-1. <http://mevik.net/work/software/pls.html> (accessed Dec. 17, 2007).
- (7) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241.
- (8) Johnson, S. R.; Zheng, W. Recent progress in the computational prediction of aqueous solubility and absorption. *AAPS J.* **2006**, *8*, E27–E40.
- (9) Bergström, C. A. S. In silico Predictions of Drug Solubility and Permeability: Two Rate-limiting Barriers to Oral Drug Absorption. *Basic Clin. Pharmacol. Toxicol.* **2005**, *96*, 156–161.
- (10) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90* (2), 234–252.
- (11) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. Perspective on the Relationship between Melting Points and Chemical Structure. *Cryst. Growth Des.* **2001**, *1* (4), 261.
- (12) Jain, A.; Yang, G.; Yalkowsky, S. H. Estimation of Melting Points of Organic Compounds. *Ind. Eng. Chem. Res.* **2004**, *43*, 7618–7621.
- (13) Clark, M. Generalized fragment-structure based property prediction method. *J. Chem. Inf. Model.* **2005**, *45*, 30–38.
- (14) Modarresi, H.; Dearden, J. C.; Modarressi, H. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 930–936.

- (15) Godavorthy, S. S., Jr.; Gasem, A. M. An improved structure - property model for predicting melting point temperatures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5117–5126.
- (16) Karthikeyan, M.; Glen, R. C.; Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *J. Chem. Inf. Model.* **2005**, *45*, 581–590.
- (17) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422.
- (18) Pipeline Pilot Basic Chemistry Component Collection, SciTegic Inc., 9655 Chesapeake Drive, Suite 401, San Diego, CA 92123.
- (19) R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL <http://www.R-project.org> (accessed Dec 17, 2007).
- (20) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (21) The Physical Properties Database (PHYSPROP), Syracuse Research Corporation. <http://www.syrres.com/esc/physprop.htm>.
- (22) Molecular Diversity Preservation International (MDPI), Basel, Switzerland. <http://www.mdpi.org> (accessed December 1, 2006).
- (23) 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.
- (24) Liu, R.; So, S. S. Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
- (25) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.
- (26) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.
- (27) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
- (28) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
- (29) 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.
- (30) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds by topological descriptors. *QSAR Comb. Sci.* **2003**, *22*, 821–829.
- (31) Sun, H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748–757.
- (32) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.
- (33) SciTegic, 10188 Telesis Court, Suite 100, San Diego, CA 92121. <http://www.scitegic.com> (accessed January 6, 2008). Accelrys offers Pipeline Pilot at no cost to students and research groups in academic institutions. Details can be found at <http://www.accelrys.com/products/scitegic/pp-student/> (accessed February 12, 2008).
- (34) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. *Rev. Comput. Chem.* **2004**, *18*, 1–40.
- (35) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720–725.
- (36) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208–1217.
- (37) Grant, D. J. W. Theory and origin of polymorphism. In *Polymorphism in Pharmaceutical Sciences, Drugs and the Pharmaceutical Sciences*; Brittain, H., Ed.; Marcel Dekker: New York; 1999; Vol. 95, pp 1–33.

CI800024C