

Binary Classification of Aqueous Solubility Using Support Vector Machines with Reduction and Recombination Feature Selection

Tiejun Cheng, Qingliang Li, Yanli Wang,* and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, United States

S Supporting Information

ABSTRACT: Aqueous solubility is recognized as a critical parameter in both the early- and late-stage drug discovery. Therefore, *in silico* modeling of solubility has attracted extensive interests in recent years. Most previous studies have been limited in using relatively small data sets with limited diversity, which in turn limits the predictability of derived models. In this work, we present a support vector machines model for the binary classification of solubility by taking advantage of the largest known public data set that contains over 46 000 compounds with experimental solubility. Our model was optimized in combination with a reduction and recombination feature selection strategy. The best model demonstrated robust performance in both cross-validation and prediction of two independent test sets, indicating it could be a practical tool to select soluble compounds for screening, purchasing, and synthesizing. Moreover, our work may be used for comparative evaluation of solubility classification studies ascribe to the use of completely public resources.

INTRODUCTION

Aqueous solubility is one of the most fundamental physicochemical properties of drug candidates.¹ Highly active compounds can be totally silent due to the lack of desirable solubility, which is directly relevant to absorption and eventual bioavailability.^{2,3} Thus, eliminating compounds with unfavorable solubility as early as possible at the screening stage will reduce costs and save time for drug discovery. However, solubility measurement can be laborious, especially when dealing with a large library of compounds. Therefore, considerable efforts have been devoted to developing computational tools for fast and accurate estimation of solubility.^{3,4}

Recent modeling studies of solubility (Supporting Information, Table S1) have employed methods, such as artificial neural networks (ANN), multilinear regression (MLR), support vector machines (SVM), partial least-squares (PLS), random forest (RF), *k*-nearest neighbor (KNN), and recursive partitioning (RP).^{5–18} Though less prevalent, there are also solubility classification studies in which a class label (e.g., soluble or insoluble) is assigned to a given compound.^{19–23}

A common feature in the above studies is that they are based on relatively small data sets. For example, the largest data set ever used consists of less than 6000 compounds (Supporting Information, Table S1). Though good results can still be achieved, data diversity is limited by using a small data set. As a result, the real predictive power of derived model for an independent test set is also weakened. We notice that the data sets used in most previous studies are derived primarily or at least partially from two commercial databases (AQUASOL and PHYSPROP) or from in-house collections, which often makes it difficult to conduct comparative evaluation using the same data sets. On the other hand, public data sets are becoming increasingly popular, as they are readily available to all researchers. Therefore, results obtained on public data sets from different studies can be possibly compared on the same ground.

Unlike previous studies, we took advantage of a high-quality data set containing over 46 000 compounds with known solubility, which is believed to be so far the largest public one. In this study, we considered the binary classification of solubility by using the SVM, an established machine learning method that has succeeded in many areas, such as pattern recognition and pharmacokinetic property prediction.^{24–29} Our SVM model was optimized in conjunction with a reduction and recombination feature selection strategy.³⁰ In particular, we constructed a hybrid fingerprint from three existing structural and/or physicochemical fingerprints. Our best model employing this fingerprint produced promising results not only in cross-validation but also in the prediction of two independent test sets.

METHODS

Data Set. The Burnham Center for Chemical Genomics (BCCG) has launched a screening campaign for aqueous solubility against the NIH Molecular Libraries Small Molecule Repository (MLSMR), which contains more than 350 000 compounds. The resultant bioassay (PubChem AID: 1996) was deposited publicly in the PubChem BioAssay database.³¹ As of June 18, 2010, this bioassay stored experimental solubility data for 47 567 compounds. The solubility data can be downloaded from the PubChem FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/>). All compounds were measured using a standard protocol under the same conditions.³² We consider that data set compiled from a single source, e.g., those used in this work, is more advantageous for statistical studies than those compiled from various sources (Supporting Information, Table S1).

The 47 567 compounds were processed as follows: First, compounds with multiple components, such as mixtures and

Received: September 13, 2010

Published: January 07, 2011

Table 1. Data Sets Used in This Study

data set	type	total compounds	soluble compounds	insoluble compounds	soluble/insoluble ratio
I	training set	41 501	28 921	12 580	2.30: 1
II	internal test set	4510	3177	1333	2.38: 1
III	external test set	32	25	7	3.57: 1

salts, were discarded. Second, compounds with conflicting or redundant information were minimized. For instance, if two compounds could be characterized with the same fingerprint and their solubility class labels (soluble or insoluble) were inconsistent, then both compounds were discarded to avoid conflict; if their solubility class labels were identical, then only one compound was retained to avoid redundancy. In total, 41 501 compounds were compiled and used as the training set for SVM model construction (Table 1, data set I). The solubility of each compound is expressed in $\mu\text{g}/\text{mL}$ unit. As we considered the binary classification of solubility in this study, compounds with solubility $\geq 10 \mu\text{g}/\text{mL}$ were regarded as soluble, while those $< 10 \mu\text{g}/\text{mL}$ were regarded as insoluble. This criterion is in accordance with that specified by the original BCCG depositors, although there are considerable debates in the literature on defining the boundary of a soluble/insoluble class.³³

While this manuscript was in preparation, another 4795 compounds with experimental solubility data were added to the PubChem BioAssay database under the same bioassay (PubChem AID: 1996, updated on July 15, 2010). They were processed as above and served as an internal test set (4510 compounds in total) to assess the performance of our SVM model (Table 1, data set II). In addition, 32 drug-like compounds with reliably measured intrinsic solubility from a recent solubility prediction challenge³⁴ were used as an external test set (Table 1, data set III) to provide a comparative evaluation of our model with those previous methods. The same criterion as above was applied to classify soluble and insoluble compounds.

Fingerprints and Feature Selection. Molecular fingerprints are widely applied in substructure/similarity searching,³⁵ compound clustering,³⁶ and classification.²² In this study, considering both their popularity and public availability, we adopted the MDL MACCS key³⁷ and the PubChem fingerprint.³⁸ The MACCS key is a binary vector of 166 structural and/or physicochemical features (MACCS166), while the PubChem fingerprint represents the presence/absence of 881 substructures (PC881). We also considered one additional fingerprint consisting of six physicochemical properties (ADD6), which were previously found to be relevant to solubility modeling.^{22,39} With respect to these physicochemical properties, data sets I and II are rather diverse (Figure 1). Regardless of the minimal and maximal values, both data sets have similar distributions with respect to most of these properties. This is probably because the MLSMR is a compound library designed for screening purposes. Data set III demonstrates a better diversity in terms of these properties. The PC881 and ADD6 were downloaded from the PubChem Compound database. The MACCS166 keys were generated by using the Open Label.⁴⁰

Feature selection has been broadly applied to select a subset of features from a given fingerprint.^{41–44} In this study, we adopted a simple strategy based on F-score, which measures the discrimination of two sets of numbers.⁴⁵ Given a binary classification task

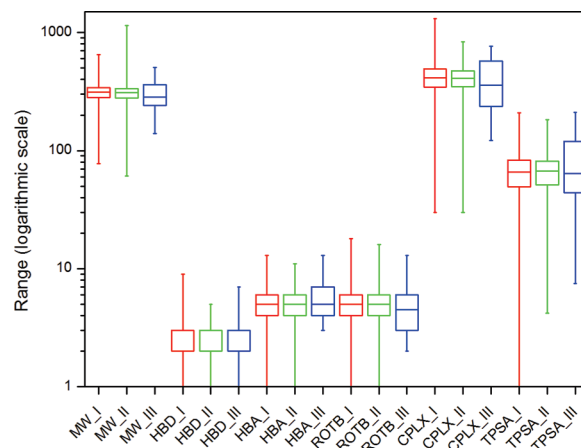


Figure 1. Six additional physicochemical properties (ADD6) used in this study. The box plot shows the minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum of each property. MW: molecular weight; HBD: number of hydrogen-bond donors; HBA: number of hydrogen-bond acceptors; ROTB: number of rotatable bonds; CPLX: molecular complexity; and TPSA: topological polar surface area. The properties of training set (data set I) are suffixed with I, while those of two test sets (data set II and III) are suffixed with II and III, respectively. Note that the statistics for all properties have been increased by one to fit in the logarithmic coordination, because the minimal values of some properties (e.g., HBD) are zeros, which would become infinity in the logarithmic scale.

and a data set, in which the compound is characterized by an m -feature fingerprint, the F-score of the i^{th} feature is defined as

$$F_i = \frac{(\bar{x}_i^{[+]} - \bar{x}_i)^2 + (\bar{x}_i^{[-]} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{[+]} - \bar{x}_i^{[+]})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{[-]} - \bar{x}_i^{[-]})^2} \quad (1)$$

where n_+ and n_- are the numbers of soluble and insoluble samples within a data set; \bar{x}_i , $\bar{x}_i^{[+]}$ and $\bar{x}_i^{[-]}$ are the average of the i^{th} feature of all, soluble, and insoluble samples, respectively; and $x_{k,i}^{[+]}$ and $x_{k,i}^{[-]}$ are the i^{th} feature of the k^{th} soluble and insoluble samples, respectively.

In principle, the larger an F-score is, the more likely a feature is more discriminative. In this study, for each of the three fingerprints MACCS166, PC881, and ADD6, the F-score of each feature was calculated from the distribution of soluble and insoluble samples in data set I. Features were ranked in a descending order of their F-scores. Our aim is to select the most discriminative features so that computational efficiency can be improved, though information may be lost to some extent. Considering both sides, we chose the F-score of 0.001 as a threshold to select only the top-ranked features from each parent fingerprint. We adopted this F-score-based feature selection because it is very straightforward to implement and generally

quite effective as well.⁴⁵ Besides, F-score can be calculated in advance and thus is independent of the chosen classifier.

SVM Modeling and Evaluation. All SVM calculations in this work were conducted by using the LIBSVM.⁴⁶ The 10-fold cross-validation was applied to evaluate model performance. Briefly, data set I was randomly split into 10 folds in a stratified way so that the ratio of soluble/insoluble samples in each fold was kept identical. In each round, one fold was chosen as a test subset, while the remaining nine folds were combined into a training subset. An SVM model was then built using this training subset, which in turn was used to predict the test subset. The above procedures were repeated for each of the 10 folds. The results from 10 rounds were averaged to give a final assessment of model performance. In addition, two independent test sets (data set II and III) were also used to provide additional evaluations. The following metrics were calculated

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (5)$$

where TP, FP, TN, and FN denote the predicted true positive, false positive, true negative, and false negative, respectively. In addition, G-mean that tries to maximize the accuracy on the both sides of two classes was also calculated

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (6)$$

RESULTS

SVM Modeling with Default Parameters. The linear kernel and radical basis function (RBF) kernel are two common kernel functions in the LIBSVM. To get an overview of their general performance, we first investigated a few simple SVM models with default parameters within the LIBSVM. The 10-fold cross-validation results given by these models are listed in Table 2. While others have found SVM models with RBF kernel outperform those with linear kernel,^{47,48} we observed that they were similar in performance. For example, when MACCS166 was employed, both kernels reported comparable G-means (69.7 vs 70.9%). The linear kernel gave marginally better results than RBF kernel when PC881 was used (76.6 vs 74.7%). We consider that the default parameters in the LIBSVM might not be suitable for RBF kernel in this case. Actually, several studies have shown that selecting optimal parameters is critical for RBF kernel.^{49,50} Some researchers also indicate that linear kernel is a special case of RBF kernel for some parameters.⁵¹ Therefore, RBF kernel is more commonly used and was also adopted by us.

When there are multiple fingerprints available, it is important to choose an appropriate one. It is clear from Table 2 that the SVM models employing PC881 demonstrate significantly superior results than those employing MACCS166. For example, PC881 outperformed MACCS166 by nearly 4% (74.7 vs 70.9%) when RBF kernel was applied. This is imaginable since the PC881

Table 2. The 10-Fold Cross-Validation Using G-Mean as a Metric for SVM Models with Default Parameters

fingerprint ^a	SVM (%) ^b	
	linear kernel ^c	RBF kernel
MACCS166	69.7 (53.3)	70.9 (51.0)
PC881	76.6 (70.6)	74.7 (63.2)
MACCS166 + ADD6		72.8
PC881 + ADD6		74.8
PC881 + MACCS166		75.6
PC881 + MACCS166 + ADD6		75.7
PC307 + MACCS90 + ADD5		75.6

^a MACCS166: the MDL MACCS 166 keys; PC881: the PubChem fingerprint; ADD6: the six additional physicochemical properties described in Figure 1; and PC307, MACCS90, and ADD5 are the truncated versions of their parent fingerprints whose component features have F-scores above 0.001. The trailing digit indicates the length of the corresponding fingerprint. ^b The number inside the parentheses is generated by the SVM model in which data imbalance has not been considered. ^c Relevant metrics for SVM models with linear kernel have not been calculated for the last five fingerprints.

fingerprint is more than five times (881/166) as long as the MACCS166 key. The much longer PC881 is believed to be more information-rich, making it more discriminative than MACCS166 in our binary classification.

Data imbalance is known to have a great impact on most classifiers, including SVM.^{47,48,52} As shown in Table 1, data set I shows partial data imbalance with a ratio of soluble to insoluble samples of 2.30. To address this issue, biased weights were assigned respectively to soluble and insoluble classes during model construction. The weights were determined from the proportion of soluble to insoluble samples in data set I, by imposing a larger penalty on the classification error for the minor class (0.435 and 1.000 for soluble and insoluble classes, respectively). As seen in Table 2, there was a significant decrease in performance (10–20%) if data imbalance was not taken into account. In the following analysis, data imbalance was always considered.

Optimizing SVM Models with Feature Selection. The reduction and recombination feature selection strategy successfully enhanced compound recall and structural diversity for hits discovery.³⁰ This inspired us to mix the three fingerprints of MACCS166, PC881, and ADD6. The underlying assumption is that different fingerprints can encode different aspects of information for the problem of interest, so they may complement each other to yield better performance.

We first investigated the fingerprint combination strategy (without reduction). Table 2 shows the four different combinations of MACCS166, PC881, and ADD6. As one can see, the SVM models employing combined fingerprint consistently outperformed those employing individual fingerprint. For instance, the SVM model employing MACCS166 + ADD6 outperformed the one employing MACCS166 by about 2% (72.8 vs 70.9%). This supports previous findings that the six additional physicochemical properties comprised in the ADD6 fingerprint are relevant to solubility.^{22,39} An interesting observation is that model performance increased as combined fingerprint became longer. This is in line with our previous observation that the longer PC881 performed better than MACCS166. On the other hand, the performance of SVM models tended to converge as the

Table 3. Prediction of Independent Test Sets

data set	model	soluble compounds ^a				insoluble compounds ^b							
		TP	FN	sensitivity (%)	FNR (%)	TN	FP	specificity (%)	FPR (%)	precision (%)	recall (%)	accuracy (%)	G-mean (%)
II (N = 4510)	SVM ^c	2622	555	82.5	17.5	1115	218	83.6	16.4	92.3	82.5	82.9	83.1
	SVM ^d	2705	472	85.1	14.9	1084	249	81.3	18.7	91.6	85.1	84.0	83.2
III (N = 32)	SVM ^c	22	3	88.0	12.0	2	5	28.6	71.4	81.5	88.0	75.0	50.1
	SVM ^d	22	3	88.0	12.0	3	4	42.9	57.1	84.6	88.0	78.1	61.4
	SVM ^{c,e}	19	3	86.4	13.6	2	4	33.3	66.7	82.6	86.4	75.0	53.6
	ASM-ATC-LOGP ^f	24	1	96.0	4.0	3	4	42.9	57.1	85.7	96.0	84.4	64.1
	MLR ^g	21	4	84.0	16.0	5	2	71.4	28.6	91.3	84.0	81.2	77.5
	ANN ^g	24	1	96.0	4.0	4	3	57.1	42.9	88.9	96.0	87.5	74.1
	category ^g	24	1	96.0	4.0	2	5	28.6	71.4	82.8	96.0	81.2	52.4
	ChemSilico ^g	24	1	96.0	4.0	1	6	14.3	85.7	80.0	96.0	78.1	37.0
	optibrium ^g	24	1	96.0	4.0	3	4	42.9	57.1	85.7	96.0	84.4	64.1
	pharma algorithms ^g	24	1	96.0	4.0	1	6	14.3	85.7	80.0	96.0	78.1	37.0
	Simulations Plus ^g	22	3	88.0	12.0	3	4	42.9	57.1	84.6	88.0	78.1	61.4
	original consensus ^g	23	2	92.0	8.0	2	5	28.6	71.4	82.1	92.0	78.1	51.3
SPARC ^g	15	10	60.0	40.0	6	1	85.7	14.3	93.7	60.0	65.6	71.7	

^a TP: true positive; FN: false negative; and FNR: false negative rate = FN/(FN + TP). ^b TN: true negative; FP: false positive; and FPR: false positive rate = FP/(FP + TN). ^c Model is based on the selected feature set, i.e., PC307 + MACCS90 + ADD5. ^d Model is based on the complete feature set, i.e., PC881 + MACCS166 + ADD6. ^e Results are based on a clean version of data set III by removing the four common samples in data set I and III. ^f Data are cited from ref 16. ^g Data are cited from the Supporting Information of ref 55.

length of combined fingerprint increased. For example, Table 2 shows that the gained performance was merely 1% by extending PC881 to PC881 + MACCS166 + ADD6 (74.7 vs 75.7%). Therefore, elongating a fingerprint by incorporating more features may not necessarily improve a model effectively. Moreover, issues, such as feature intercorrelation and feature redundancy, may arise when integrating different fingerprints.

The best result of 10-fold cross-validation was given by the SVM model employing PC881 + MACCS166 + ADD6 (75.7%, Table 2). However, using such a long fingerprint (1053 features) would be computationally expensive, especially in the grid search for the optimal parameters of RBF kernel. Therefore, we utilized the reduction and recombination strategy to make a shorter fingerprint from existing ones. We believed this strategy could alleviate, if not fully solve, the issue of feature redundancy. Only the top-ranked features with F-score above 0.001 from each of the PC881, MACCS166, and ADD6 were retained, which resulted in three truncated fingerprints: PC307, MACCS90, and ADD5. They were then recombined together to yield a new fingerprint: PC307 + MACCS90 + ADD5. Compared to its full-length parent PC881 + MACCS166 + ADD6, there is only negligible information loss (75.6 vs 75.7%, Table 2), and this new fingerprint is much shorter (402 features). The above results provided us with confidence to use the reduction and recombination strategy for feature selection. Further optimization of SVM model was based on this new PC307 + MACCS90 + ADD5.

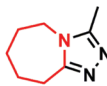
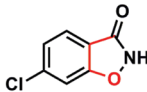
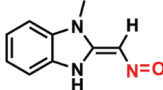
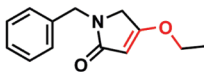
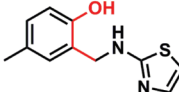
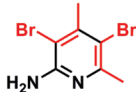
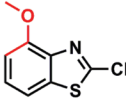
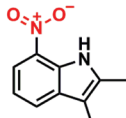
The two parameters of C and γ in RBF kernel are critical to a SVM model.^{49,50} To seek the optimal pair of C and γ , grid search in the parameter space was conducted along with five-fold cross-validation, which turned out to be the most inefficient step during model construction. In this work, it took about two hours to accomplish a typical five-fold cross-validation task on a 16 CPU × 2.60 GHz Linux cluster (using only one CPU) with a maximal

memory of 224 MB used. We started from a coarse grid ($C \in [0, 12]$ and $\gamma \in [-12, 0]$, both in log 2 units) with a grid spacing of 1.0. A subregion ($C \in [1, 3]$ and $\gamma \in [-7, -5]$) showing relatively better performance was identified. Further grid search was restricted in this subregion with a finer grid spacing of 0.25 to identify an even better subregion. This procedure was repeated until the optimal parameters ($C = 2.43$ and $\gamma = -6.34$) were determined. Built on these two parameters, our final SVM model achieved a G-mean of 80.3% by 10-fold cross-validation.

Predicting an Internal Test Set (Data Set II). More often than not, a model fails to predict an independent test set, although it can perform extremely well during training. A common mistake in the applications of feature selection, as pointed by Smialowski et al.,⁵³ is that some researchers first use the whole data set for feature selection, then split it into training and test sets, with the former to build a classifier and the latter to evaluate model performance. We strongly agree that more rigorous evaluation should be provided, since in such procedure the trained classifier has already taken advantage of the information leaked from test set.

In this study, data set II, which was excluded entirely from feature selection and model construction, was used as an internal test set to evaluate the performance of our final SVM model. We preprocessed this test set using the 402 features of PC307 + MACCS90 + ADD5 as applied to data set I. The prediction results by our model are listed in Table 3. The G-mean was 83.1%, which is close to that of the 10-fold cross-validation (80.3%), indicating the robustness of our model. As for soluble compounds, our model successfully recognized 2622 out of the 3177 soluble compounds, giving a sensitivity of 82.5%. This result may not be surprising since our data sets are imbalanced toward soluble compounds (Table 1), and thus classifiers tend to label samples as major class.⁵⁴ Nevertheless, when focusing on the classification of insoluble compounds, our model also gave a low false positive rate (16.4%). This can be ascribed to the application

Table 4. Top 10 Features That Contribute Most to Classification

No.	Feature	Weight	Source	Description or SMARTS	Example structure in data set I ^a	PubChem CID	Solubility (μg/mL)
1	1,053	2.915	ADD6	Topological polar surface area		Not applicable	
2	979	1.891	MACCS166	[!C;!c;!#1]1~*~*~*~*~*~*1		4364917	= 11.9
3	542	1.871	PC881	O-C:C-C		2781852	> 25.4
4	944	1.653	MACCS166	[N,n]=[O,o]		6019741	> 26.3
5	477	1.370	PC881	C-O-C=C		563187	> 32.6
6	1,048	-4.920	ADD6	Molecular weight		Not applicable	
7	591	-2.745	PC881	C-C:C-O-[#1]		891102	= 7.4
8	510	-1.686	PC881	Br-C:C-C		223259	= 6.2
9	566	-1.579	PC881	C:C-O-C		2049862	= 7.4
10	951	-1.564	MACCS166	[!C;!c;!#1]~[N,n]~[O,o]		38742	= 1.2

^a Example fragment of respective SMARTS is depicted with red.

of biased weights to soluble/insoluble classes during model training. As a result, the hyper-plane of SVM classifier was pushed toward minor class (insoluble samples), giving a promising specificity (83.6%). In addition, using G-mean as a quality control in cross-validation, the performance of our SVM model was maximized for both soluble and insoluble compounds. The overall classification accuracy is 82.9%, which is comparable to those reported in previous studies (Supporting Information, Table S1). This level of performance is satisfactory, considering the large-size test set used here.

In the above analysis, the optimal parameters of *C* and γ were applied to the SVM model employing a selected subset of features (PC307 + MACCS90 + ADD5). What if the same parameters were applied to the SVM model employing the full-length PC881 + MACCS166 + ADD6? One can see from Table 3 that slightly better results were obtained in terms of accuracy and G-mean. Therefore, information loss occurred after feature selection, but it was rather marginal. For example, the reported G-mean for the SVM models with and without feature selection are 83.1 and 83.2%, respectively. It is thus interesting to observe that the optimal parameters derived from the SVM

model with feature selection are also applicable to that without feature selection, although they may not be truly optimal for the latter. This might also indicate that SVM models are more sensitive to the chosen parameters than the employing features of a fingerprint, which may be responsible for the universal success of SVM applications.

Predicting an External Test Set (Data Set III). This data set consists of 32 pharmaceutical chemicals from a recent solubility prediction challenge³⁴ and was used to provide an external evaluation of our SVM model. A number of previous studies have reported their predictions for the same test set,^{16,55} making it possible to compare our model with theirs on the same ground. The comparative results are also listed in Table 3. Our SVM model employing PC307 + MACCS90 + ADD5 gave a moderate accuracy of 75.0%, while slightly better results were obtained when PC881 + MACCS166 + ADD6 was applied. It should be noted that four compounds (Supporting Information, Table S2) in this test set were also contained in data set I (i.e., training set), making the prediction not completely independent. Comparable or slightly better results were obtained when these four common compounds were removed from data

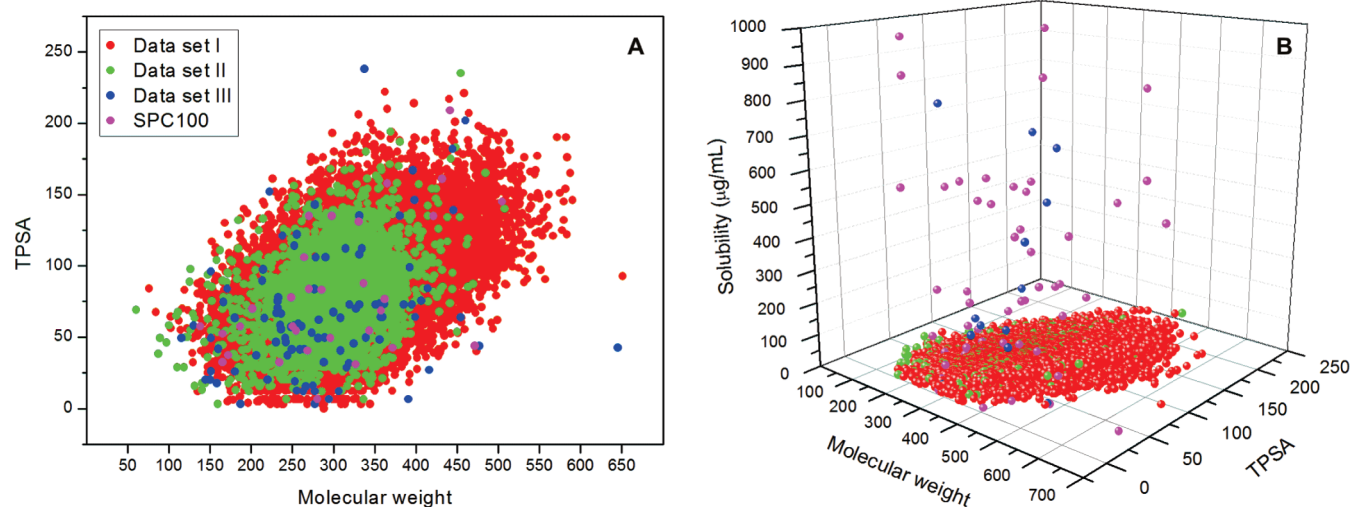


Figure 2. Diversity analysis of data sets I–III and the 100 compounds from the training set of the solubility prediction challenge (SPC100).³⁴ (A) Chemical space defined by molecular weight and TPSA. Note that one data point (1139.8, 133) from data set II is not included in this figure. (B) Distribution of solubility in a chemical space defined by molecular weight and TPSA. Both figures use the same color scheme.

III. It is notable that one compound (PubChem CID: 3108) was incorrectly classified though it was included in data set I. Further investigation indicates that this compound was reported as soluble in data set I, while insoluble in data set III. Therefore, the inconsistency in the experimental determination of solubility for this compound finally led to the misclassification by our model. This indicates again the importance of data quality, especially when compiling from multiple sources. In comparison, our model achieved comparable performance to some previous methods (e.g., ChemSilico and SPARC). Relevant discussion is given below.

DISCUSSION

Features and Physical Meanings of the Fingerprints. In this work, we have employed the reduction and recombination feature selection strategy to select the most discriminative features. It thus would be very helpful to interpret the predictability as well as the physical meanings of our SVM model from the perspective of these features. The description, F-score and weight of all the 1053 features from PC881, MACCS166, and ADD6 were provided (Supporting Information, excel file). The weight (i.e., relative contribution to classification) of each feature was derived from a linear SVM model by using the svm-weight.⁵⁶ In particular, the top 10 features that contributed most to classification are listed in Table 4. A greater positive weight indicates a larger contribution of this feature to the classification of soluble samples and vice versa. As one can see, these top 10 features came from PC881, MACCS166, or ADD6, implying that all three fingerprints indeed played a key role in our model. It can be observed in Table 4 that the most significant feature for the classification of soluble samples is the 1053rd feature (topological polar surface area). This is anticipated because the larger polar surface area a compound has, the more likely it is soluble in water. Similarly, compounds containing the 944th feature (nitroso group) also tend to be soluble, which is in accordance with the previous findings that this functional group makes a negative contribution to hydrophobicity.^{57,58} Likewise,

the 1048th feature (molecular weight) contributes most to insolubility classification. This is true for many chemicals. For example, the solubility of alcohol in water decreases as the molecular size increases. However, the relationship between molecular weight and solubility is not always that straightforward. Other features, such as the 510th feature, can also be interpretable for insolubility classification since it basically encodes hydrophobic substructures. Nevertheless, this does not mean that compounds containing such negatively contributing features suggested in this work are necessarily insoluble or vice versa. Solubility or insolubility should always consider a molecule as a whole.

Diversity and Chemical Space of Data Sets. Data diversity should always be addressed when building a computational model. That is the reason why we emphasized the use of large data sets in this work. We plotted in Figure 2A the chemical space of data sets I–III, which is defined by molecular weight and topological polar surface area. These two coordinates were chosen because they were found in the above analysis to be relevant to solubility classification. As one can see, both data sets II and III (test sets) share a similar chemical space of data set I (training set), which may account for the reasonably good prediction of our SVM model on both test sets. However, data sets that are within a similar low-dimension chemical space may not necessarily distribute similarly in a higher dimension chemical space. As shown in Figure 2B, the experimental solubility of data set III is more sparsely scattered than that of data set II, implying that the former is a more challenging test set for our SVM model as well as for other methods.^{34,59} This is in accordance with the relatively lower performance of our SVM model for data set III. In contrast, some other methods, such as MLR and ANN (Table 3), were calibrated by using the 100 compounds from the training set of the solubility prediction challenge,³⁴ whose chemical space (Figure 2B) is more similar to that of data set III. This might contribute to their relatively better performance than ours for data set III. Another possible reason is that the choice of 10 µg/mL as a binary cutoff for solubility

classification may not be suitable for data set III, as the solubility of compounds therein was measured using a completely different experiment. Data set I covers a very small portion of the chemical space of the MLSMR and an even smaller portion of the chemical space of the PubChem BioAssay database (Supporting Information, Figure S1). Thus, the predictability of our model for a data set that is beyond the training chemical space of our model should not be anticipated without caution, which is true for any supervised machine learning methods.

CONCLUSIONS

In this study, we have presented a binary classification model of aqueous solubility using the SVM. A reduction and recombination feature selection strategy was applied to design a new fingerprint by selecting and recombining the most discriminative features from three existing fingerprints. Based on this new fingerprint (PC307 + MACCS90 + ADD5), an SVM model was constructed and optimized using a large and diverse training set (data set I, $N = 41\,501$). For an internal test set (data set II, $N = 4510$), our model correctly classified both soluble and insoluble samples with an overall accuracy of 82.9%. For an external drug-like test set (data set III, $N = 32$), the performance of our SVM model was found to be comparable to that of some other methods, such as MLR and ANN. Therefore, our model may be used as a practical tool for fast and accurate classification of solubility for untested compounds, which may facilitate compound selection and library design at the early stage of drug discovery. Our study may also provide insights into building predictive models based on very large data sets. In addition, using completely public resources (data sets, software, and methods) in this work will facilitate others to reproduce or compare with our results. The performance of our SVM classification model may be further improved when more experimental solubility data become available.

ASSOCIATED CONTENT

S Supporting Information. Recent modeling studies of aqueous solubility. Four common compounds in data set I and III. Chemical space for the compounds in data sets I–III as well as for those in the MLSMR and the PubChem BioAssay database. Definition, F-score, and weight of all the 1053 features from PC881, MACCS166, and ADD6. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Y.W.: e-mail ywang@ncbi.nlm.nih.gov. S.H.B.: e-mail bryant@ncbi.nlm.nih.gov.

ACKNOWLEDGMENT

We thank the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) for funding support. We also thank the NIH Fellows Editorial Board (FEB) for manuscript revision.

REFERENCES

(1) 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.* **2001**, *46*, 3–26.

(2) van de Waterbeemd, H.; Gifford, E. ADMET in silico modeling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.

(3) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1*, 31–52.

(4) Johnson, S. R.; Zheng, W. Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption. *AAPS J.* **2006**, *8*, E27–40.

(5) 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.

(6) 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.

(7) 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.* **2003**, *44*, 266–275.

(8) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.

(9) Xia, X.; Maliski, E.; Cheetham, J.; Poppe, L. Solubility Prediction by Recursive Partitioning. *Pharm. Res.* **2003**, *20*, 1634–1640.

(10) Votano, Joseph R.; Parham, M.; Hall, Lowell H.; Kier, Lemont B.; Hall, L. M. Prediction of Aqueous Solubility Based on Large Datasets Using Several QSPR Models Utilizing Topological Structure Representation. *Chem. Biodiversity* **2004**, *1*, 1829–1841.

(11) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170–176.

(12) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30–38.

(13) Jain, N.; Yang, G.; Machatha, S. G.; Yalkowsky, S. H. Estimation of the aqueous solubility of weak electrolytes. *Int. J. Pharm.* **2006**, *319*, 169–171.

(14) 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.* **2006**, *47*, 150–158.

(15) Zhou, D.; Alelyunas, Y.; Liu, R. Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *48*, 981–987.

(16) Wang, J.; Hou, T.; Xu, X. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571–581.

(17) Carpenter, G. A.; Grossberg, S.; Reynolds, J. H. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks* **1991**, *4*, 565–588.

(18) Carpenter, G. A.; Grossberg, S.; Markuzon, N.; Reynolds, J. H.; Rosen, D. B. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. Neural Networks* **1992**, *3*, 698–713.

(19) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.

(20) Fredsted, B.; Brockhoff, Per B.; Vind, C.; Padkjær, Søren B.; Refsgaard, Hanne H. In Silico Classification of Solubility using Binary k -Nearest Neighbor and Physicochemical Descriptors. *QSAR Comb. Sci.* **2007**, *26*, 452–459.

(21) Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. Straightforward Recursive Partitioning Model for Discarding Insoluble Compounds in the Drug Discovery Process. *J. Med. Chem.* **2008**, *51*, 2891–2897.

(22) Zhang, H.; Xiang, M.-L.; Ma, C.-Y.; Huang, Q.; Li, W.; Xie, Y.; Wei, Y.-Q.; Yang, S.-Y. Three-class classification models of logS and logP derived by using GA-CG-SVM approach. *Mol. Diversity* **2009**, *13*, 261–268.

(23) Kramer, C.; Beck, B.; Clark, T. Insolubility Classification with Accurate Prediction Probabilities Using a MetaClassifier. *J. Chem. Inf. Model.* **2010**, *50*, 404–414.

- (24) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.
- (25) Chen, H.-F. In Silico LogP Prediction for a Large Data Set with Support Vector Machines, Radial Basis Neural Networks and Multiple Linear Regression. *Chem. Biol. Drug Des.* **2009**, *74*, 142–147.
- (26) Liao, Q.; Yao, J.; Yuan, S. SVM approach for predicting LogP. *Mol. Diversity* **2006**, *10*, 301–309.
- (27) Trotter, Matthew W. B.; Holden, Sean B. Support Vector Machines for ADME Property Classification. *QSAR Comb. Sci.* **2003**, *22*, 533–548.
- (28) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*; ACM: Pittsburgh, PA, 1992; pp 144–152.
- (29) Vapnik, V. N. *The nature of statistical learning theory*; Springer-Verlag, Inc.: New York, 1995; p 188.
- (30) Nisius, B.; Bajorath, J. Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of Hits. *Chem. Biol. Drug Des.* **2010**, *75*, 152–160.
- (31) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010**, *38*, D255–266.
- (32) Bhattachar, S. N.; Wesley, J. A.; Seadeek, C. Evaluation of the chemiluminescent nitrogen detector for solubility determinations to support drug discovery. *J. Pharm. Biomed. Anal.* **2006**, *41*, 152–157.
- (33) Yu, L. X.; Amidon, G. L.; Polli, J. E.; Zhao, H.; Mehta, M. U.; Conner, D. P.; Shah, V. P.; Lesko, L. J.; Chen, M.-L.; Lee, V. H. L.; Hussain, A. S. Biopharmaceutics Classification System: The Scientific Basis for Biowaver Extensions. *Pharm. Res.* **2002**, *19*, 921–925.
- (34) Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.
- (35) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (36) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL “Keys” as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
- (37) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- (38) *PubChem fingerprint*, version 1.3; National Center for Biotechnology Information (NCBI): Bethesda, MD; ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. Accessed July 13, 2010.
- (39) Tulp, I.; Dobchev, D. A.; Katritzky, A. R.; Acree, W.; Maran, U. A General Treatment of Solubility 4. Description and Analysis of a PCA Model for Ostwald Solubility Coefficients. *J. Chem. Inf. Model.* **2010**, *50*, 1275–1283.
- (40) *Open Babel*, version 2.2.3; Department of Chemistry, University of Arizona: Tucson, AZ; <http://openbabel.org>. Accessed July 13, 2010.
- (41) Yang, Y.; Pedersen, J. O. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: Cambridge, MA, 1997; pp 412–420.
- (42) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
- (43) Ding, Y.; Wilkins, D. Improving the Performance of SVM-RFE to Select Genes in Microarray Data. *BMC Bioinf.* **2006**, *7*, S12.
- (44) Wood, I. A.; Visscher, P. M.; Mengersen, K. L. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics* **2007**, *23*, 1363–1370.
- (45) Chen, Y.-W.; Lin, C.-J. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction*; Springer: Berlin/Heidelberg, 2006; Vol. 207/2006, pp 315–324.
- (46) *LIBSVM*, version 2.9.1; <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. Accessed on July 13, 2010.
- (47) Tang, Y.; Zhang, Y.-Q. Granular support vector machines with data cleaning for fast and accurate biomedical binary classification. In *Proceedings from 2005 IEEE International Conference on Granular Computing*, Beijing, China, July 25–27, 2005; IEEE: New York, 2005; pp 262–265.
- (48) Li, Q.; Wang, Y.; Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **2009**, *25*, 3310–3316.
- (49) Fröhlich, H.; Chapelle, O.; Schölkopf, B. Feature Selection for Support Vector Machines by Means of Genetic Algorithms. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, Sacramento, CA, November 3–5, 2005; IEEE Computer Society: New York, 2003; p 142.
- (50) Huang, C.-L.; Wang, C.-J. A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst. Appl.* **2006**, *31*, 231–240.
- (51) Keerthi, S. S.; Lin, C.-J. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.* **2003**, *15*, 1667–1689.
- (52) Tang, Y.; Zhang, Y. Q.; Chawla, N. V.; Krasser, S. SVMs Modeling for Highly Imbalanced Classification. *IEEE Trans. Syst. Man Cybern. B: Cybern.* **2009**, *39*, 281–288.
- (53) Smialowski, P.; Frishman, D.; Kramer, S. Pitfalls of supervised feature selection. *Bioinformatics* **2010**, *26*, 440–443.
- (54) Guha, R.; Schürer, S. Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 367–384.
- (55) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49*, 2572–2587.
- (56) *svm-weight*; <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/calw/>. Accessed on July 13, 2010.
- (57) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140–2148.
- (58) Meylan, W.; Howard, P. Estimating log P with atom/fragments and water solubility with log P. *Perspect. Drug Discovery Des.* **2000**, *19*, 67–84.
- (59) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *49*, 1–5.