

Full-length paper

## SVM approach for predicting LogP

Quan Liao, Jianhua Yao\* & Shengang Yuan

Department of Computer Chemistry and Chemoinformatics, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences

(\*Author for correspondence, E-mail: yaojh@mail.sioc.ac.cn, Tel.: +86-21-54925266, Fax: +86-21-54925264)

Received 8 October 2005; Accepted 27 October 2005

**Key words:** LogP prediction, multiple linear regression (MLR), partial least squares (PLS), support vector machines (SVM)

### Summary

The logarithm of the partition coefficient between n-octanol and water (logP) is an important parameter for drug discovery. Based upon the comparison of several prediction logP models, i.e. Support Vector Machines (SVM), Partial Least Squares (PLS) and Multiple Linear Regression (MLR), the authors reported SVM model is the best one in this paper.

**Abbreviations:** LogP, the logarithm of the partition coefficient between n-octanol and water; SVM, support vector machines; PLS, partial least squares; MLR, multiple linear regression

### Introduction

Lipophilicity is an important parameter related to drug properties, such as drug absorption, plasma protein binding, hydrophobic drug-receptor interactions and partly the pharmacokinetic behavior and toxicological properties [1, 2]. The logarithm of the partition coefficient between n-octanol and water (logP) is often used to represent molecular lipophilicity. Since the pioneer work of Hansch and Fujita [3], logP has become a crucial descriptor for the quantitative structure activity relationship (QSAR) studies.

Although the experimental determination of the logP value of a compound is relatively easy, the time and cost consumption cannot be ignored, especially when a large number of candidate molecules are screened and when they have not been synthesized yet. Therefore, there is an increasing need for reliable estimation of logP based only on the chemical structure.

During the past three decades, several works for prediction of logP had been published [1, 4–13], such as CLOGP [4], KLOGP [6], XLOGP [7], AUTOLOGP [8]. Like any other property prediction approach from chemical structures, logP prediction also relies on structure descriptors and computational algorithms. There are two classes of descriptors: structural fragments and numeric indices. Multiple Linear Regression (MLR), Partial Least Squares (PLS) and different types of Artificial Neural Networks (ANN) have been used as computational models for logP prediction.

The Support Vector Machines (SVM) [14–16] is more and more employed in chemistry [17–22]. The goal of this paper is to explain that SVM is the superior computational model for logP prediction among MLR, PLS and SVM models.

### Methods

In principle, construction of predictive models consists of several steps as follows: preparation of data set, acquirement of descriptors and analysis by statistic methods.

#### Data set

A high quality experimental data set is extremely important for developing a reliable predictive logP model. The data set used in this paper is compiled from the work of Hansch et al. [23]. In this compilation, the logP of compounds marked with star (\*) are considered more reliable than those not marked by the authors. We selected 8402 neutral organic compounds with the star. Their logP values are in the range of [–4.41, 8.42], and the mean and standard deviation are 1.80 and 1.68, respectively. These compounds are randomly divided into two data sets: training set (6722 compounds) and test set (1680 compounds).

#### Generation of substructure descriptors

The substructure descriptor is widely applied in QSAR/QSPR due to its advantages: easy to calculate and simple to interpret

Table 1. Summary of atom types and bond types

| No         | Code | Note           | No | Code | Note          |
|------------|------|----------------|----|------|---------------|
| Atom types |      |                |    |      |               |
| 1          | [C0] | Sp3 Carbon     | 9  | [H]  | Hydrogen      |
| 2          | [C1] | Sp2 Carbon     | 10 | [S]  | Sulfur        |
| 3          | [C2] | Sp Carbon      | 11 | [P]  | Phosphorus    |
| 4          | [N0] | Sp3 Nitrogen   | 12 | [F]  | Fluorine      |
| 5          | [N1] | Sp2 Nitrogen   | 13 | [Cl] | Chlorine      |
| 6          | [N2] | Other Nitrogen | 14 | [Br] | Bromine       |
| 7          | [O0] | Sp3 Oxygen     | 15 | [I]  | Iodine        |
| 8          | [O1] | Sp2 Oxygen     |    |      |               |
| Bond types |      |                |    |      |               |
| 1          | —    | Single bond    | 3  | \$   | Triple bond   |
| 2          | =    | Double bond    | 4  | #    | Aromatic bond |

models and be treated [24–27]. In this work, we used four kinds of substructure descriptors: Star, Path, Ring and Atom. The former three had been successfully employed in our previous work [28], and the last one (Atom) was newly added in this work. The procedure was slightly modified based on the previous work [28].

#### Pre-processing structural data

All hydrogen atoms were explicitly added to the chemical structure, hybridization degree of atoms, information about rings and the aromatic property were identified, and the atoms and bonds were assigned to different types. Atoms and bonds were classified in fifteen atom types and four bond types (see Table 1 for details)

#### Deriving substructures

All substructures in the four types were derived based on the following rules:

*Atom*: Every single atom was an Atom fragment.

*Star*: Each atom with connectivity more than two was considered as the center of a star. Starting from this atom, a substructure with one layer was generated as a Star fragment.

*Path*: Each atom was selected as the starting atom. Any paths with 1–4 bonds (2–5 atoms) was generated as a Path fragment.

*Ring*: Every ring was considered as a Ring fragment.

#### Encoding substructures

Every substructural fragment was encoded as a unique string by the simple coding rules as the following:

*Atom*: using the code listed in Table 1.

*Star*: [Center-atom](Bond<sub>1</sub>[Atom<sub>1</sub>])(Bond<sub>2</sub>[Atom<sub>2</sub>])... (Bond<sub>n</sub>[Atom<sub>n</sub>]). The subcodes (Bond<sub>i</sub>[Atom<sub>i</sub>]) were ordered by the alphabet.

*Path*: [Atom<sub>1</sub>]Bond(1,2)[Atom<sub>2</sub>]Bond(2,3)...Bond(n-1,n)[Atom<sub>n</sub>]. For one path, two codes (strings) were generated because of two end atoms. The alphabetically smaller one was selected as the descriptor.

*Ring*: [Atom<sub>1</sub>]Bond(1,2)[Atom<sub>2</sub>]Bond(2,3)...[Atom<sub>n</sub>]Bond(n,1). For a ring, 2n strings were generated because of n atoms in a ring. The alphabetically smallest one was selected as the descriptor.

where Atom<sub>i</sub> and Bond<sub>i</sub> were defined in Table 1.

The procedure of deriving descriptors is illustrated by an example in Figure 1.

#### Partial least squares (PLS)

PLS, also known as Projection to Latent Structures, is a powerful statistical method that can easily cope with larger numbers of correlated descriptors by projecting them into several orthogonal latent variables [29].

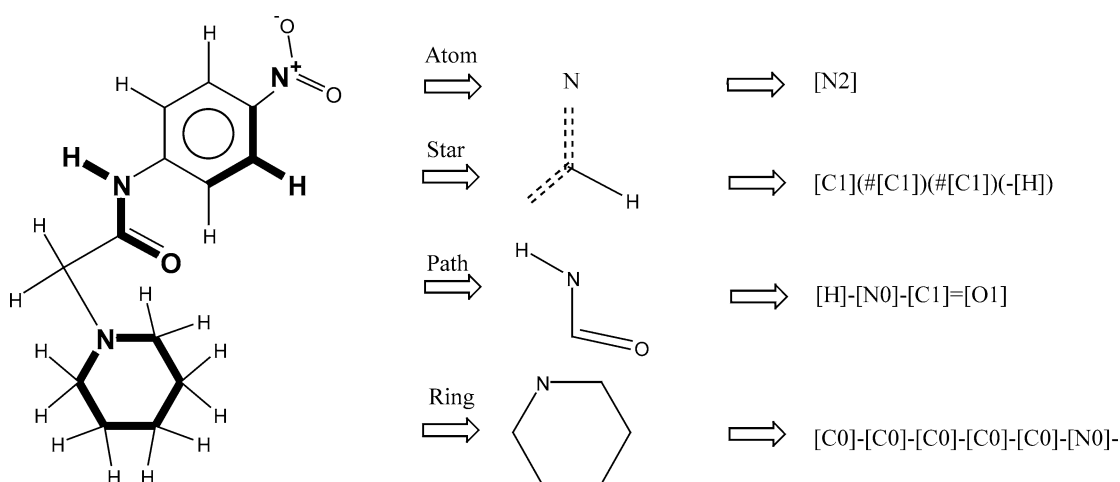


Figure 1. Some descriptors (in bold) of N-(p-Nitrophenyl)-3-N'-Piperidinoacetamide.

In order to determine how many latent variables should be selected, a 10-fold cross-validation was carried out. For 10-fold cross-validation, the training set were divided into 10 equal parts, and each of the different part was used once as a test set. The root-mean-square error (RMSE) between the experimental and tested logP values should be minimal when the optimal number of latent variables was used. The program called QSAR from Jay Ponder Lab [30] was modified in order to deal with larger numbers of data and construct the prediction model.

#### Multiple linear regression (MLR)

The statistical significance of a MLR model is highly depended on the number of descriptors, which should not exceed the first of five of the number of compounds [31]. Herein, the thousands of original substructure descriptors could not be directly used for MLR analysis. A subset of descriptors was necessary. The selection of descriptors was performed as follows:

1. The 15 types of atom descriptors were selected to construct an initial 15-descriptor model, and a residue variable (the difference between the experimental logP and calculated logP) was calculated.
2. Selecting the Nth significant substructure descriptor in rest descriptors. The Nth descriptor should have the maximal correlation coefficient ( $R^2$ ) with the residue variable of (N-1)-descriptor model.
3. A N-descriptor model was rebuilt, and the residue variable is recalculated.

The steps (2) and (3) were repeated. In each iteration, a descriptor was added to the model.

Similar to construction of PLS model, the 10-fold cross-validation was used to decide the optimal number of descriptors. Several functions in the book [32] were combined to a MLR program to construct the prediction model.

#### Support vector machines (SVM)

SVM is a very promising classification and regression method developed by Vapnik et al. [14–16]. Its main advantage is: adopting the structure risk minimization (SRM) principle minimizing an upper bound of the generalization error on the Vapnik-Chernoverkis dimension, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle minimizing the training error. Several works published [17–22] have proved its affectivity in classification and regression.

In general, the performance of SVM modeling is better than that of traditional machine learning approaches, including artificial neural networks. It has a special advantage: prevention of over-fitting.

In this work, the program LibSVM2.6 [33] was employed to construct the SVM model and run on a machine with a Pentium IV and 256M RAM. The prior regression models were obtained using the EPSILON-SVR method in LibSVM. The adjustable parameters are the capacity parameter C, the  $\varepsilon$  of  $\varepsilon$ -insensitive loss function, kernel function type and its corresponding parameters. The optimal value for  $\varepsilon$  depends on the noise in the training data. The  $\varepsilon$  was fixed to 0.4 in this study because the experimental error of logP is generally considered to be 0.4 log units. And the radial basis function (RBF) kernel was selected, with a kernel parameter  $\gamma$ . So, two parameters (C and  $\gamma$ ) should be optimized. Grid searches are performed to find the optimal value of parameters, which make the RMSE of 10-fold cross-validation minimal.

## Results and discussion

#### Substructure descriptor set

7779 unique substructures were derived from 6722 molecules in the training set. If a substructure occurred in N molecules, the corresponding descriptor would have N non-zero values and other 6722-N values were zero. The smaller the N value (occurrence in Table 2) was, the less informative the substructure descriptor was. Those substructures occurring in less than 6 molecules were not considered. Finally, 3730 frequent substructures were used to construct the descriptor set in this work.

Some substructure descriptors with different occurrences (N vaules) are listed in Table 2.










#### Results of PLS

All the 3730 descriptors were used in construction of PLS model. It consisted of the following steps. (1) Calculation of 10-fold cross-validation to find the optimal number of latent variables. The RMSE of cross-validation were used as the criteria. The relation of RMSE versus number of latent variables is shown in Figure 2. (2) Determining the optimal number of latent variables according to the relation, 18 is the optimal number in this work. (3) Construction of the PLS model expressed in (1):

$$\log P = b_0 + \sum_{i=1}^{3730} b_i x_i \quad (1)$$

where  $b_0$  is the regression constant ( $b_0 = 0.281$ ),  $b_i$  is the contribution of the  $i$ th substructure descriptor, and  $x_i$  is the occurrence of the  $i$ th substructure descriptor. For this model (PLS3730), the correlation coefficient ( $R^2$ ) and root-mean-square error (RMSE) of training are 0.956 and 0.352, the correlation coefficient ( $Q^2$ ) and root-mean-square error (RMSE) of cross-validation are 0.915 and 0.491.

Table 2. Examples of substructure descriptors

| No | Type | Substructure  | Code                          | Occurrence |
|----|------|---|-------------------------------|------------|
| 1  | Atom | H   | [H]                           | 6695       |
| 2  | Star |    | [C1](#[C1])(#[C1])(-[H])      | 4773       |
| 3  | Ring |    | [C1][C1][C1][C1][C1][C1][C1]  | 4336       |
| 4  | Path | C=O   | [C1]=[O1]                     | 3592       |
| 5  | Atom | P   | [P]                           | 147        |
| 6  | Path |    | [N0]-[C0]-[C0]-[C1]=[O1]      | 91         |
| 7  | Ring |    | [C0]-[C0]-[N0]-               | 54         |
| 8  | Star |    | [C1](-[H])(-[N0])(=[N1])      | 37         |
| 9  | Star |   | [P](-[O0])(-[O0])(-[S])(=[S]) | 17         |
| 10 | Path |  | [Br]-[C0]-[Br]                | 6          |
| 11 | Path |  | [C1]=[C1]-[I]                 | 3          |
| 12 | Ring |  | [C0]-[N0]-[C0]-[N0]-          | 1          |

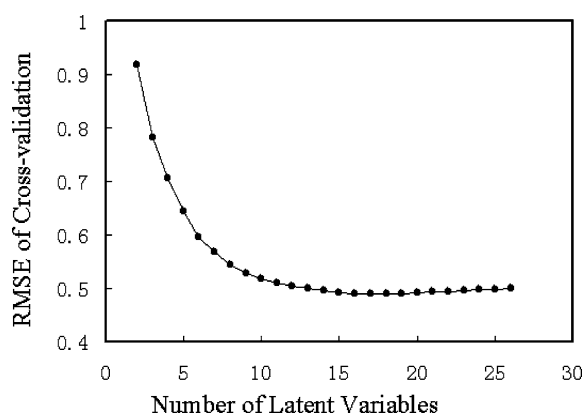


Figure 2. Cross-validation RMSE versus the number of latent variables of PLS.

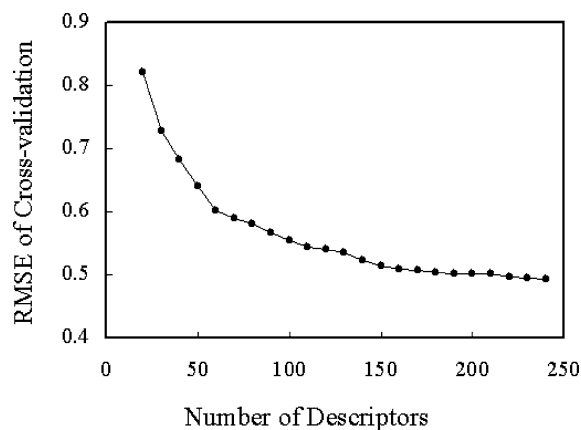


Figure 3. Cross-validation RMSE versus the number of descriptors of MLR.

### Results of MLR

According to the principle of the method mentioned above, a subset of descriptors was selected. The relationship

between RMSE of cross-validation and number of select descriptors is shown in Figure 3. We can propose that the set of 200 descriptors is proper. The final model contains 200 descriptors, including 15 atom types and 185 significant

Table 3. Examples of the first 10 selected correction factors in the MLR model

| ID | Type | Code                      | Example molecule <sup>a</sup> | Occurrence <sup>b</sup> | Coefficient |
|----|------|---------------------------|-------------------------------|-------------------------|-------------|
| 1  | Path | [H]-[N0]-[C0]-[C1]=[O1]   |                               | 377                     | -0.096      |
| 2  | Path | [H]-[C0]-[O0]-[H]         |                               | 744                     | -0.263      |
| 3  | Path | [N0]-[N1]                 |                               | 289                     | 0.925       |
| 4  | Path | [C0]-[N0]-[C1]#[N1]#[C1]  |                               | 94                      | 0.271       |
| 5  | Star | [C1](-[N0])(-[N0])(=[O1]) |                               | 408                     | 0.783       |
| 6  | Star | [C1]-[N0])(-[O0])(=[O1])  |                               | 301                     | 0.494       |
| 7  | Star | [C1](#[C1])(#[N1])(-[H])  |                               | 894                     | -0.256      |
| 8  | Star | [N2](#[C1])(#[C1])(=[O1]) |                               | 34                      | -2.501      |
| 9  | Path | [C1]#[C1]-[C1]-[O0]       |                               | 401                     | 0.187       |
| 10 | Path | [H]-[C0]-[S]=[O1]         |                               | 126                     | -0.069      |

<sup>a</sup>The selected substructure is drawn in bold, some of the hydrogen are not shown.<sup>b</sup>The occurrence number of molecules in the training set (total 6722 compounds).

substructure descriptors (correction factors). The model is expressed in (2):

$$\log P = b_0 + \sum_{i=1}^{15} b_i x_i + \sum_{j=16}^{200} b_j x_j \quad (2)$$

where  $b_0$  is the regression constant ( $b_0 = 0.069$ ),  $b_i$  is the contribution of the  $i$ th atom types,  $x_i$  is the occurrence of the  $i$ th atom types,  $b_j$  is the contribution of the  $j$ th correction factors, and  $x_j$  is the occurrence of the  $j$ th correction factors. The first selected 10 correction factors are listed in Table 3. For this model (MLR200),  $R^2$  and RMSE of training are 0.931 and 0.449; the  $Q^2$  and RMSE of cross-validation are 0.910 and 0.505.

The MLR200 model, as one novel type of atom-additive logP model, is quite different from those reported in references [7, 10], which often included many atom types and

only a few correction factors. The classification of atom types herein is very simple and does not consider the complex connectivity information around atoms. On the other hand, the interactions among atoms are represented by thousands of substructures, from which the correction factors are found automatically. Thus, the model depends less on empirical knowledge.

### Results of SVM

The 200 descriptors for MLR model and the 3730 descriptors for PLS model are all adopted in SVM model.

### Model based on 200 descriptors

Grid searches were made on the basis of 10-fold cross-validation. For the first search, the C should be varied from

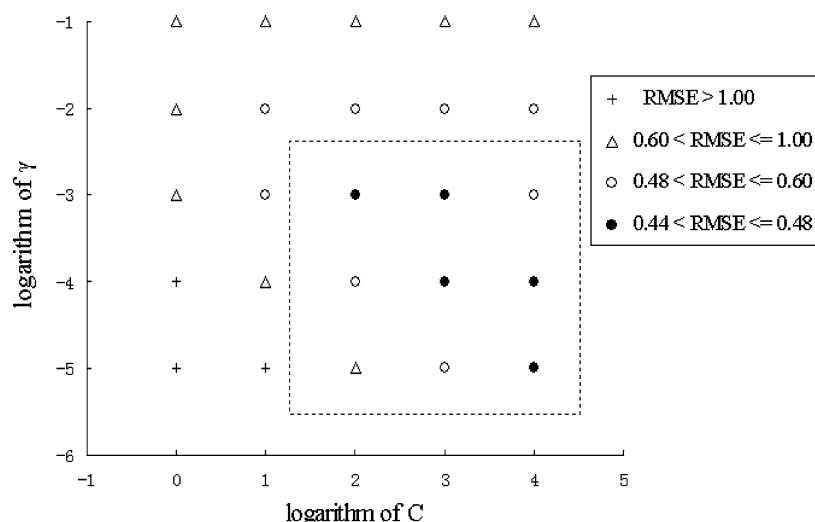


Figure 4. Cross-validation RMSE versus different C and  $\gamma$  value (for 200 descriptors).

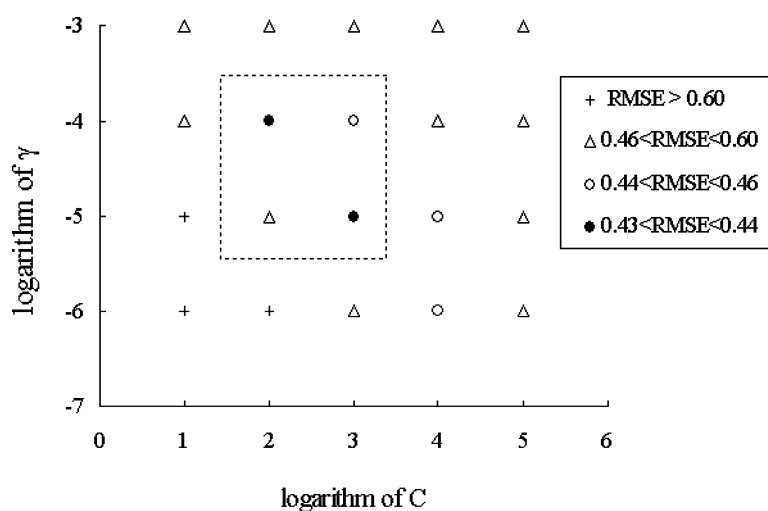


Figure 5. Cross-validation RMSE versus different C and  $\gamma$  value (for 3730 descriptors).

1 to  $10^4$  and the  $\gamma$  varied from  $10^{-5}$  to  $10^{-1}$ . As shown in Figure 4, an optimal subspace is explored for the next grid search. After several searches, the optimal C and  $\gamma$  were identified, which are 200 and 0.001 respectively. For this model (SVM200),  $R^2$  and RMSE of training are 0.979 and 0.245; the  $Q^2$  and RMSE of cross-validation are 0.929 and 0.446.

#### Model based on 3730 descriptors

Using the same process for the model from the 200 descriptors, another optimal parameters for C and  $\gamma$  are 100 and 0.0001 when 3730 descriptors were used (Figure 5). For this model (SVM3730),  $R^2$  and RMSE of training are 0.971 and 0.286; the  $Q^2$  and RMSE of cross-validation are 0.934 and 0.432.

The training results of the four models are listed in Table 4 and Figure 6.

#### Results of test

Based on different regression methods with different sets of descriptors, four computational models (PLS3730, MLR200, SVM200 and SVM3730) are derived from the training processes.  $R^2$  and  $Q^2$  of these models are calculated for statistical significance validation. The data outside the training data, 1680 diverse compounds were used to test these models. The test results are listed in Table 5. According to the criteria of Mannhold et al. [1], the estimation errors less than  $\pm 0.5$  are considered as acceptable; errors greater than  $\pm 0.5$  and less than  $\pm 1.0$  are considered as disputable; and errors exceeding  $\pm 1.0$  are considered as unacceptable. The test results were divided into 3 groups by the criterion. The percentages of these 3 groups for each model are listed in Table 5. The plots of experimental logP versus calculated logP are shown in Figure 7.

Table 4. Training results (6722 chemicals) of PLS, MLR and SVM models

| Model   | Method | Descriptor | Cross-validation |       | Training |       |
|---------|--------|------------|------------------|-------|----------|-------|
|         |        |            | $Q^2$            | RMSE  | $R^2$    | RMSE  |
| PLS3730 | PLS    | 3730       | 0.915            | 0.491 | 0.956    | 0.352 |
| MLR200  | MLR    | 200        | 0.910            | 0.505 | 0.931    | 0.449 |
| SVM200  | SVM    | 200        | 0.929            | 0.446 | 0.979    | 0.245 |
| SVM3730 | SVM    | 3730       | 0.934            | 0.432 | 0.971    | 0.286 |

Table 5. Test results (1680 chemicals) for PLS, MLR and SVM models

| Model   | $R^2$ | RMSE  | Acceptable (%) <sup>a</sup> | Disputable (%) <sup>b</sup> | Unacceptable (%) <sup>c</sup> |
|---------|-------|-------|-----------------------------|-----------------------------|-------------------------------|
| PLS3730 | 0.917 | 0.485 | 80.3                        | 14.7                        | 5.0                           |
| MLR200  | 0.911 | 0.502 | 76.7                        | 18.2                        | 5.1                           |
| SVM200  | 0.931 | 0.443 | 83.9                        | 12.1                        | 4.0                           |
| SVM3730 | 0.943 | 0.403 | 86.9                        | 10.4                        | 2.7                           |

<sup>a</sup>Estimated errors less than  $\pm 0.5$ .

<sup>b</sup>Errors greater than  $\pm 0.5$  and less than  $\pm 1.0$ .

<sup>c</sup>Errors greater than  $\pm 1.0$ .

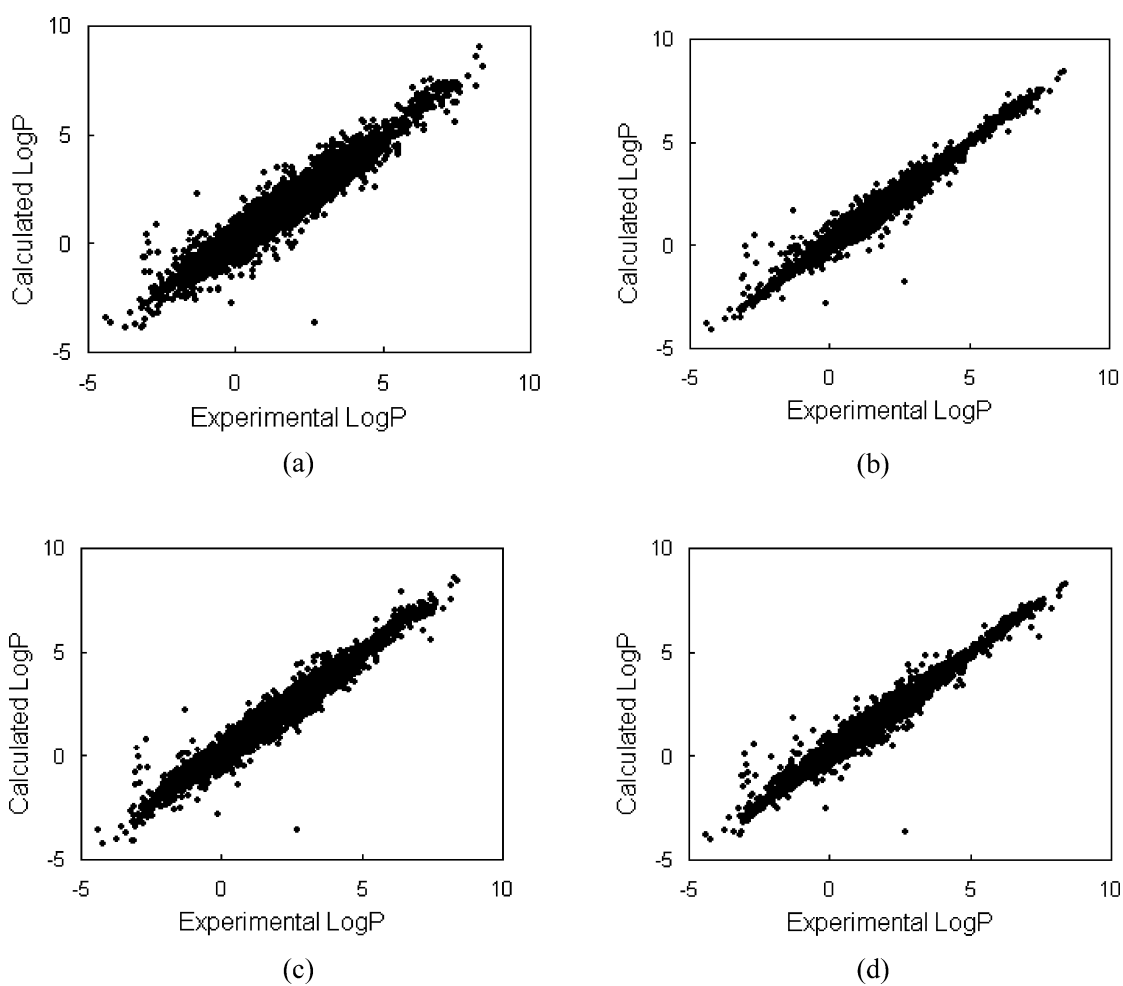


Figure 6. Plots of the calculated LogP values versus experimental LogP values (training). (a) MLR200 model, (b) SVM200 model, (c) PLS3730 model and (d) SVM3730 model.

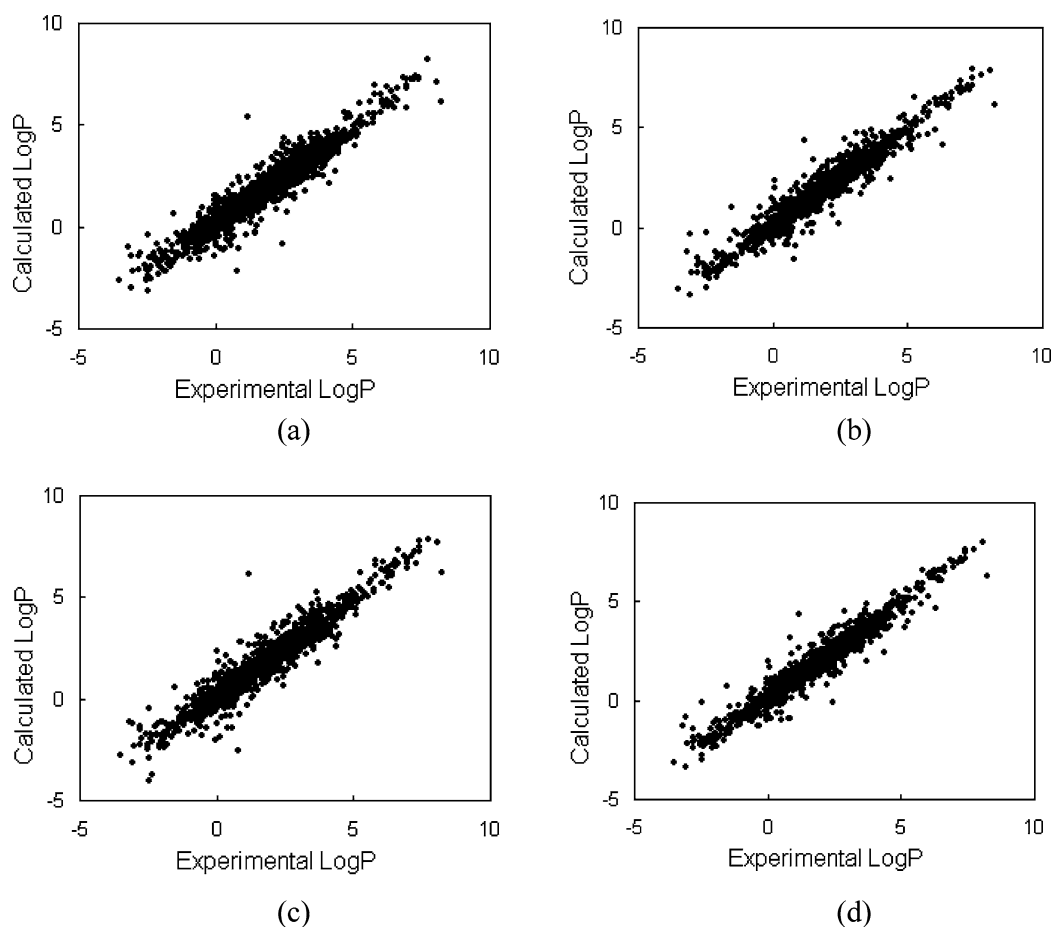


Figure 7. Plots of the calculated LogP values versus experimental LogP values (test). (a): MLR200 model, (b): SVM200 model, (c): PLS3730 model, (d): SVM3730 model.

#### Comparison of the four models

From the values of  $Q^2$  of the cross-validation of the four models (in Table 4) and  $R^2$  of the test of the four models (in Table 5), we can conclude that the performance of the four models is in this sequence: SVM3730>SVM200>PLS3730>MLR200. SVM3730 is the best model, its  $R^2$  is 0.943 and RMSE is 0.403. Only 2.7% of the 1680 test compounds are badly predicted by this model.

The relationship between logP and chemical structure should be nonlinear in nature. The predictability of SVM models is better than that of MLR and PLS model. Our study clearly shows that SVM is a better model to make prediction of logP from both high or low-dimensional data space.

#### Conclusion

In this paper, we have implemented four models, MLR200, PLS3730, SVM200 and SVM3730, to predict logP from substructure descriptors. These four models were compared. From these computational experiments, we observed:

1. All models can produce reasonable results comparing with the work reported in references [1, 4–13]. The substructure descriptors developed in our group are representative for predicting logP.
2. Models generated from SVM are better than those generated from PLS and MLR approaches. According to the test results, SVM which is used with substructure descriptors developed in our group is a better option to predict logP.

#### Acknowledgment

This work was supported in part by the National Basic Research Program (also called 973 Program) of China, through Grants 2003CB114400; by National Natural Science Foundation of China through Grants 20473112; by Chinese Academy of Sciences, through Grants KGCX2-SW-213-05 and KGCX2-SW-213-01.

#### References

1. Mannhold, R. and Dross, K., *Calculation procedures for molecular lipophilicity: a comparative study*, Quant. Struct.-Act. Relat., 15 (1996) 403–409.



2. Testa, B., Crivori, P., Reist, M. and Carrupt, P.A., *The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples*, Perspect. Drug Disc. Design, 19 (2000) 179–211.
3. Hansch, C. and Fujita, T., *Correlation of biochemical activity of phenoxycetic acids with Hammett substituent constants and partition coefficients*, Nature, 194 (1962) 178–180.
4. Leo, A., *Calculating logP<sub>oct</sub> from structures*, Chem. Rev., 93 (1993) 1281–1306.
5. Suzuki, T. and Kudo, Y., *Automatic logP estimation based on combined additive modeling methods*, J. Comput.-Aided Mol. Des., 4 (1990) 155–198.
6. Klopman, G., Li, J.Y., Wang, S. and Dimayuga, M., *Computer automated logP calculations based on an extended group contribution approach*, J. Chem. Inf. Comput. Sci., 34 (1994) 752–781.
7. Wang, R., Fu, Y. and Lai, L., *A new atom-additive method for calculating partition coefficients*, J. Chem. Inf. Comput. Sci., 37 (1997) 615–621.
8. Devillers, J., Domine, D. and Guillon, C., *Autocorrelation modeling of lipophilicity with a back-propagation neural network*, Eur. J. Med. Chem., 33 (1998) 659–664.
9. Mannhold, R. and Petrauskas, A., *Substructure versus whole molecule approaches for calculating logP*, QSAR Comb. Sci., 22 (2003) 466–475.
10. Sun, H., *A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption*, J. Chem. Inf. Comput. Sci., 44 (2004) 748–757.
11. Chuman, H., Mori, A., Tanaka, H., Yamagami, C. and Fujita, T., *Analyses of the partition coefficient, logP, using ab initio MO parameter and accessible surface area of solute molecules*, J. Pharm. Sci., 93 (2004) 2681–2697.
12. In, Y., Chai, H.H. and No, K.T., *A partition coefficient calculation method with the SFED model*, J. Chem. Inf. Model., 45 (2005) 254–263.
13. Schnackenberg, L.K. and Beger, R.D., *Whole-molecule calculation of logP based on molar volume, hydrogen bonds, and simulated <sup>13</sup>C NMR spectra*, J. Chem. Inf. Model., 45 (2005) 360–365.
14. Vapnik, V.N. (Ed.) *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
15. Cristianini, N. and Shawe-Taylor, J. (Eds.) *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.
16. Burges, C.J.C., *A tutorial on Support Vector Machine for pattern recognition*, Data Min. Knowl. Disc., 2 (1998) 121–167.
17. Burbidge, R., Trotter, M., Buxton, B. and Holden, S., *Drug design by machine learning: Support Vector Machines for pharmaceutical data analysis*, Comput. Chem., 26 (2001) 5–14.
18. Song, M., Breneman, C.M., Bi, J., Sukumar, N., Bennett, K.P., Cramer, S. and Tugcu, N., *Prediction of protein retention times in anion-exchange chromatography systems using Support Vector Regression*, J. Chem. Inf. Comput. Sci., 42 (2002) 1347–1357.
19. Kramer, S., Frank, E. and Helma, C., *Fragment generation and Support Vector Machines for inducing SARs*, SAR QSAR Environ. Res., 13 (2002) 509–523.
20. Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P. and Pletnev, I.V., *Drug discovery using Support Vector Machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions*, J. Chem. Inf. Comput. Sci., 43 (2003) 2048–2056.
21. Yao, X.J., Panaye, A., Doucet, J.P., Zhang, R.S., Chen, H.F., Liu, M.C., Hu, Z.D. and Fan, B.T., *Comparative study of QSAR/QSPR correlations using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression*, J. Chem. Inf. Comput. Sci., 44 (2004) 1257–1266.
22. Luan, F., Zhang, R.S., Zhao, C.Y., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., *Classification of the carcinogenicity of N-nitroso compounds based on Support Vector Machines and Linear Discriminant Analysis*, Chem. Res. Toxicol., 18 (2005) 198–203.
23. Hansch, C., Leo, A. and Hoekman, D. (Eds.) *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, Vol 2, American Chemical Society, Washington, DC, 1995.
24. Zefirov, N.S. and Palyulin, V.A., *Fragmental approach in QSPR*, J. Chem. Inf. Comput. Sci., 42 (2002) 1112–1122.
25. Hurst T. and Heritage T., *HQSAR – A highly predictive QSAR technique based on molecular holograms*, 213th ACS Natl. Meeting, San Francisco, CA, (1997), CINF 019.
26. Merlot, C., Domine, D., Cleve, C. and Church, D.J., *Chemical substructures in drug discovery*, Drug. Discovery Today, 8 (2003) 594–602.
27. Clark, M., *Generalized fragment-substructure based property prediction method*, J. Chem. Inf. Model, 45 (2005) 30–38.
28. Liao, Q., Yao, J.H., Li, F., Yuan, S.G., Doucet, J.P., Panaye, A. and Fan, B.T., *CISOC-PSCT: A predictive system for carcinogenic toxicity*, SAR QSAR Environ. Res., 15 (2004) 217–235.
29. Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (Ed.) *Multi- and Megavariate Data Analysis Principles and Applications*, Umetrics Academy: Kinnelon, NJ, 2001.
30. <ftp://dasher.wustl.edu/pub/qsar/>.
31. Topliss, J.G. and Edwards, R.P., *Chance factors in studies of quantitative structure-activity relationships*, J. Med. Chem., 22 (1979) 1238–1244.
32. Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (Eds.) *Numerical Recipes in C: the Art of Scientific Computing*, 2nd Ed., Cambridge University Press, Cambridge, 1995, 676–681.
33. Chang, C.C. and Lin, C.J., *LIBSVM – A library for Support Vector Machines*, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>.