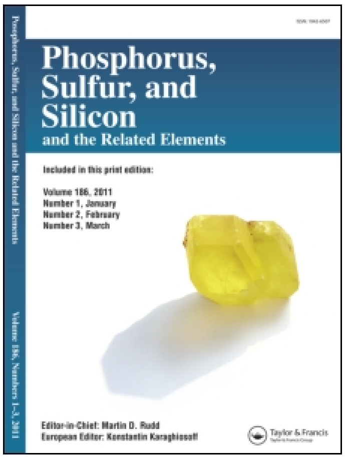

Phosphorus, Sulfur, and Silicon and the Related Elements
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713618290

A Molecular Approach for the Prediction of Sulfur Compound Solubility Parameters

Mehdi Mehrpooya[ab]; Farhad Gharagheizi[a]
[a] Department of Chemical Engineering, Faculty of Engineering, University of Tehran, Tehran, Iran [b] Center of Advanced Computing in Process Engineering, CACPEMP, Tehran, Iran

Online publication date: 28 December 2009

To cite this Article Mehrpooya, Mehdi and Gharagheizi, Farhad(2010) 'A Molecular Approach for the Prediction of Sulfur Compound Solubility Parameters', Phosphorus, Sulfur, and Silicon and the Related Elements, 185: 1, 204 — 210

To link to this Article: DOI: 10.1080/10426500902758394

URL: http://dx.doi.org/10.1080/10426500902758394

Phosphorus, Sulfur, and Silicon, 185:204–210, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 1042-6507 print / 1563-5325 online
DOI: 10.1080/10426500902758394

A MOLECULAR APPROACH FOR THE PREDICTION OF SULFUR COMPOUND SOLUBILITY PARAMETERS

Mehdi Mehrpooya[1,2] and Farhad Gharagheizi[1]
[1]Department of Chemical Engineering, Faculty of Engineering, University of Tehran, Tehran, Iran
[2]Center of Advanced Computing in Process Engineering, CACPEMP, Tehran, Iran

A quantitative structure–property relationship (QSPR) study was performed to construct a multivariate linear model and a three-layer feed-forward neural network model. This model relates the solubility parameters of 82 sulfur compounds to their structures. Molecular descriptors, which are extracted from the molecular structure of compounds, have been used as model parameters. The multivariate linear model was gained by a genetic algorithm–based multivariate linear regression; the results showed that the squared correlation coefficient (R^2) between predicted and experimental values was 0.964. Next, a three-layer feed-forward neural network model with optimized structure was employed; the results showed that the squared correlation coefficient (R^2) is 0.9874, and with this model we can predict the solubility parameter more accurately than the linear model.

Supplemental materials are available for this article. Go to the publisher's online edition of **Phosphorus, Sulfur, and Silicon and the Related Elements** ***to view the free supplemental file.***

Keywords Genetic algorithm; mercaptans; neural networks; QSPR; solubility parameter

INTRODUCTION

Sulfur compounds present in petroleum and natural gas cause many problems in chemical and petrochemical plants. These compounds are corrosive and they damage instruments and pipelines. Also, during reactions, these compounds cause catalyst poisoning. As a result, removing these compounds from petroleum and natural gas is needed. However, as these compounds are often environmental pollutants, it is necessary to decrease these compounds in petroleum and natural gas to international standards levels. Existence of such compounds in oil and gas plants feeds streams, and as a result, some units are designed for treating them. The situation of the treated streams (temperature, pressure, composition, etc.) is changed as they pass through these units, so applying optimization procedures is necessary for decreasing plant overall expenses.[1,2]

The major classes of sulfur compounds that exist in petroleum and natural gas usually include mercaptans, disulfides, sulfides, thiosulfides, benzothiophenes, and cyclic sulfur

Received 12 October 2008; accepted 14 January 2009.

Address correspondence to Mehdi Mehrpooya, Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran and current address: Center of Advanced Computing in Process Engineering, CACPEMP, P.O. Box 33316-19636, Tehran, Iran. E-mail: mehrpoya@ut.ac.ir; mehdiemp@gmail.com

analogs. There are many methods used to remove these compounds from petroleum and natural gas, such as iron oxide method, reactive adsorption method, merox method, sulfiran method, and solvent extraction method.[3–8]

In the solvent extraction method, selection of the proper solvent is of great importance. There are many parameters such as aniline point,[9,10] Hansen solubility parameter,[11–13] and solubility parameter[15–19] applied to select the proper solvent for a specific use.

One of the most widely used parameters is the solubility parameter. There are many methods to estimate solubility parameter from chemical structure of molecules. Of them, there is no accurate method to estimate solubility parameters of sulfur compounds needed in the solvent extraction method. In addition, the experimental values for these compounds are rare in the literature. As a result, the main aim of this article is to present an accurate method to predict solubility parameters of sulfur compounds.

One of the most widely used methods applied to relate physical and chemical properties to chemical structure of compounds is the well known quantitative structure–property relationship (QSPR) methodology. The major advantage of this method in comparison with other methods is to apply some chemical structure-based parameters called molecular descriptors. Molecular descriptors are parameters calculated only from the chemical structure of compounds. Then, using these molecular descriptors, the physical or chemical properties under consideration are correlated. The obtained correlations using this method have two major limitations:[20]

(1) The family of compounds used to derive the QSPR (the “training set”) should be chemically similar.
(2) Realistic predictions can only be made for compounds that are chemically related to some of those from which the QSPR model was derived, i.e., predictions should be of interpolations or short extrapolations.

In this article, some of most widely used techniques in QSPR such as genetic algorithm–based multivariate regression (GA-MLR) and feed-forward neural networks (FFNN) are used to develop an accurate molecular-based model to predict the solubility parameter of the sulfur compounds.

METHODOLOGY

Data Set

The data set used to develop the model in this study was provided from the DIPPR 801 database.[21] This database contains evaluated data for physical properties of pure compounds and has been recommended by the American Institute of Chemical Engineers (AIChE) for physical properties of pure compounds. After considering this database, 82 sulfur compounds were found, and their solubility parameters were extracted.

Determination of Molecular Descriptors

Molecular descriptors are basic molecular properties of a compound. These parameters are calculated based on the chemical structure of the component. There are many types of molecular descriptors.[22] Each type is related to the special types of interaction between chemical groups in a molecule. There are several software packages to calculate molecular descriptors from the chemical structure of molecules. One of the most widely used is Dragon software.[23] Dragon can calculate 1664 molecular descriptors for chemical structures. Since the values of many descriptors are related to the bonds length and bonds angles, the chemical structure of every molecule should be optimized before calculating its molecular descriptors. For this reason, chemical structures of all 82 sulfur compounds were drawn using the Hyperchem software[24] and optimized using the MM+ molecular mechanics force field.

After optimizing the chemical structures, the molecular descriptors were calculated using Dragon software. The inputs to this software are the optimized chemical structures of the molecules.

Definition of Problem

After calculation of the molecular descriptors, we can develop our models for prediction of the physical properties of sulfur compounds. In QSPR study, one of the main problems is to find a multivariate linear equation with minimum input parameters (molecular descriptors) to predict a desired property with the highest accuracy possible. This problem is classified in the field of subset variable selection. In other words, we should select the best subset variables from the set of 1664 molecular descriptors so that we can predict the desired property with the highest accuracy.

One of the best solutions of this problem is genetic algorithm–based multivariate linear regression (GA-MLR). In this method, a genetic algorithm is used to select the best subset variables with respect to an objective function. In this work, we use the GA-MLR algorithm proposed by Leardi et al.,[25] which has been successfully used by us in our previous work and the results were satisfactory.[10,19,26–37] This algorithm has been designed for feature selection by means of a genetic algorithm search that is suitable for this problem.

After obtaining the best multivariate linear model, we should study the predictive power of the model. There are many standard tests for checking the predictive power of the obtained model. A review of these tests is available.[20]

Of them, external validation technique and K-test were used. These two tests are commonly used to evaluate predictive power of the QSPR models.

In the external validation test, before starting calculations, the main data set is randomly divided into two sub-data sets: the first for training and the second for the test. The training set is used to develop the best multivariate linear model. Once the best model is obtained, the predictive power of this model is checked by the test set, which has not been used to develop the model. In this test, usually 20% of the main database is allocated to the test set, and 80% is allocated to the training set.[26]

Todeschini et al.[38] presented a quick rule for checking validity of the obtained model. This rule compares the multivariate correlation index K_x of X-block of the predictor variables with the multivariate correlation index K_{xy} obtained by augmented X-block matrix by adding the column of response variable. This rule says that if $K_{xy} > K_x$, then the model is predictive. This test is called “K-test.”

In many cases, it is possible that the obtained best multivariate linear model is not as accurate as we expected. As a result, the nonlinear behavior of this obtained multivariate linear model must be considered, as a complementary work.

One of the best methods for considering the nonlinear behavior is application of neural networks. There are many types of neural networks, but of all of them, feed-forward neural networks (FFNN)[39] are widely used in QSPR studies. Therefore, we use this type of neural network for this study.

Table I The coefficients of obtained best multivariate linear model for prediction of solubility parameter that was obtained by GA-MLR and their statistical parameters

Standard regression coefficient	Confidence intervals (0.95)	Errors of regression coefficient	Regression coefficient	Variable	ID
0	0	1.546635	27.47175696	intercept	0
1.94808	0.649416791	0.3612313	4.161082828	Ms	1
1.584471	–0.086482722	2.58*E*–02	–4.86*E*–02	SRW05	2
1.969403	–0.499535345	2.75417	–24.13847356	X0A	3
1.917207	0.15403703	0.8778788	2.437226835	X1Av	4
1.192101	–0.14717237	0.5612934	–2.394492918	Mor27v	5
1.732672	0.319850668	2.036319	12.9893081	HATS1m	6
1.314489	0.375384677	0.367235	3.623874864	nROH	7
1.155668	–0.121502797	0.2501464	–0.908778318	H-048	8

GA-MLR Calculations and Results

Input parameters of our program contain the molecular descriptors and the number of desired molecular descriptors. In the first step, we started with one-molecular descriptor models. Based on the training set, the best multivariate one-molecular descriptor linear model was obtained. In next step, we examined two-molecular descriptor multivariate linear models. After obtaining the best two-molecular descriptors model by the GA-MLR technique, we compared the best one-molecular descriptor and best two-molecular descriptor multivariate linear models. In this comparison, we considered the effect of increase in number of molecular descriptors on the increasing accuracy of these two models. If the increase in the number of molecular descriptors showed a sensible increase in the accuracy of the model, we continued this procedure and increased the number of molecular descriptors to three, and then we should find the best three molecular descriptor multivariate linear model by the GA-MLR technique. If the increase in the number of molecular descriptors did not show any sensible effect on the accuracy of the model, then the best model has been obtained.

This sequence of events was repeated in order to find the best three-, four-, five-, etc. molecular descriptor multivariate linear models.

The best multivariate linear model obtained for the solubility parameter of sulfur compounds is presented in Table I. In the table, the coefficients and statistical parameters of the obtained best multivariate linear model are presented. The physical meanings of the

Table II The molecular descriptors that were entered in multivariate linear model and their physical meanings

Physical meaning	Molecular descriptors
Mean electrotopological state (Constitutional descriptors)	Ms
Self-returning walk count of order 05 (Walk and path counts)	SRW05
Average connectivity index chi-0 (Connectivity indices)	X0A
Average valence connectivity index chi-1 (Connectivity indices)	X1Av
3D-MoRSE—signal 27/weighted by atomic van der Waals volumes (3D-MoRSE descriptors)	Mor27v
Leverage-weighted autocorrelation of lag 1/weighted by atomic masses	HATS1m
Number of hydroxyl groups (Functional group counts)	nROH
H attached to C(sp3)/C(sp2)/C(sp)	H-048

Table III The statistical parameters of the best obtained tultivariate linear model and overall data set and over est set

Value	Statistical parameters
0.9523	Squared correlation coefficient for Training set (R^2)
0.9361	Squared correlation coefficient for Test set ($R_{test}{}^2$)
0.6681	Standard deviation error in calculation (SDEC)
0.7727	Standard deviation error in calculation (SDEP)
0.5168	Root mean square error (RMS)

entered molecular descriptors in the multivariate linear equation are presented in Table II. The statistical parameters of fitting of these three equations are presented in Table III. In addition, the statistical parameters are defined in Appendix (available online in the Supplemental Materials). The predicted values of the solubility parameters of sulfur compounds, in comparison with the DIPPR 801 data, are shown in Figure 1.

FFNN Calculations and Results

The three-layer feed-forward neural networks with the sigmoidal (hyperbolic tangent) transfer function have been the standard techniques used in QSPR modeling.[39] In order to consider the nonlinear behavior of different entered molecular descriptors into the best multivariate linear models, which were obtained in the previous section, three-layer feed-forward neural networks (FFNN) were used. This part of the calculations was performed using Neural Network Toolbox, available in the MATLAB software (Mathworks Inc. Software).

The main dataset was divided into two datasets: the first functioned as the training network and the second for testing it. Neural networks are good at fitting functions, and there is proof that a simple neural network can fit any data set very well. As a result, for checking the prediction power of the neural network, use of the test set is needed. The test

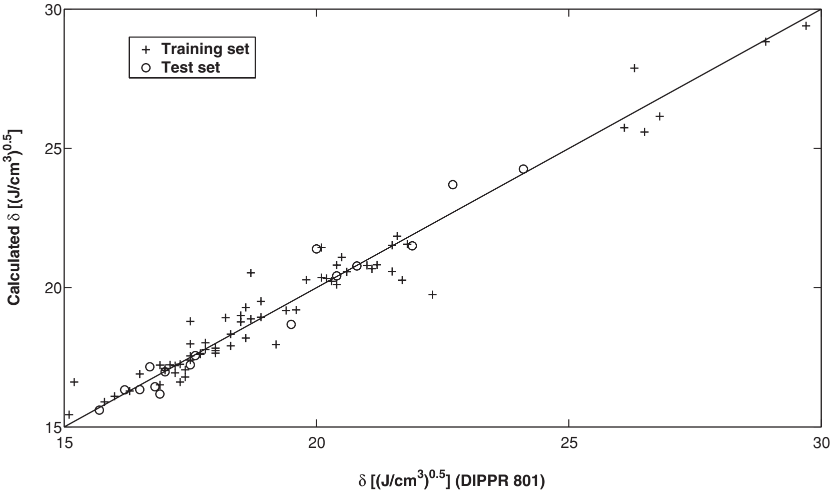

Figure 1 Predicted solubility parameters by multivariate linear model vs. the DIPPR 801 data.

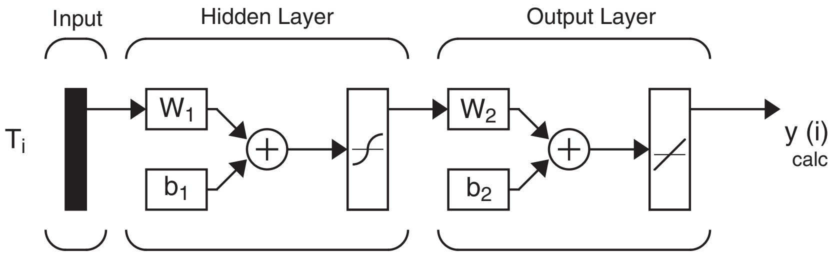

Figure 2 The schematic structure of the three-layer feed-forward neural network that was used in this study.

set is only used for checking the produced neural network and is not used to train it. In this part of study, 20% of the main database is allocated to the test set and 80% is allocated to the training set. These training set and test set differ from those used in previous parts of the study. The reason for this difference is to assure more checking independency of the obtained model to an especially divided training set and test set.

The schematic of the used three-layer FFNN in this work is shown in Figure 2. This type of FFNN has been used in our previous work, and the detailed explanations about FFNN used in this study have been presented are available.[10,28,31,33,35,36,40]

Usually, all inputs and outputs of FFNN are normalized between –1 and +1, to decrease the accounting error. This work was performed by means of the minimum and maximum values of every descriptor and also every output value.

The values of W_1, W_2, b_1, and b_2 are obtained by minimization of an objective function, which is commonly the mean squared error between the outputs of the neural network and the target values. This minimization is usually performed using the Levenberg–Marquart algorithm. This algorithm is rapid and accurate in the process of training neural networks.[39] An extensive table of solubility parameters for sulfides (Table VI) can be found in the Supplemental Materials (available online).

CONCLUSION

In the present study, an accurate model based on feed-forward neural networks has been presented to predict solubility parameters of sulfur compounds. The parameter of the obtained models is molecular descriptors, which are calculated based on chemical structure of any molecule.

In the first step, the GA-MLR technique was used to select the most statistically effective molecular descriptors on the solubility parameter of sulfur compounds from a pool of 1664 molecular descriptors. The predictive power of the model was checked using two methods (external validation, and K-test). Using these selected molecular descriptors, a FFNN was generated for prediction of the solubility parameter of sulfur compounds.

This model can be used to predict the solubility parameter of any regular sulfur compound according to the limitations presented in the Introduction.

REFERENCES

1. M. Mehrpooya, A. Jarrahian, and M. R. Pishvaie, *Int. J. Energy Res.*, **30,** 1336 (2006).
2. M. Mehrpooya, F. Gharagheizi, and A. Vatani, *Chem. Eng. Technol.*, **29,** 1469 (2006).

3. R. F. Zaykina, Y. A. Zaykin, T. B. Mamonova, and N. K. Nadirov, *Radiat. Phys. Chem.*, **63,** 621 (2002).
4. K. Mohammadbeigi and M. Tajerian, *Petrol. Coal*, **46,** 17 (2004).
5. A. M. Farhat, A. Al-Malki, B. El-Ali, G. N. Martinie, and M. Siddiqui, *Fuel*, **85,** 1354 (2006).
6. E. Ito and J. A. R. van Veen, *Catal. Today*, **116,** 446 (2006).
7. S. A. Bedell, L. L. Pirtle, J. M. Griffin, *Hydrocarb. Process.*, **86,** 49 (2007).
8. UOP company website, http://www.uop.com
9. G. Wypych, Ed., *Handbook of Solvents* (ChemTech Publishing, Toronto, 2001).
10. F. Gharagheizi, B. Tirandazi, and R. Barzin, *Ind. Eng. Chem. Res.*, **28,** 1678 (2009).
11. C. M. Hansen, *Hansen Solubility Parameters: A User's Handbook* (CRC Press, Boca Raton, FL, 2000).
12. F. Gharagheizi, *J. Appl. Polym. Sci.*, **103,** 31 (2007).
13. F. Gharagheizi and M. T. Angaji, *J. Macromol. Sci.*, **B45,** 285 (2006).
14. F. Gharagheizi, M. Sattari, and M. T. Angaji, *Polym. Bull.*, **57,** 377 (2007).
15. J. H. Hildebrand and R. L. Scott, *Solubility of Non-Electrolytes*, 3rd ed. (Reinhold, New York, 1964).
16. J. H. Hildebrand and R. L. Scott, *Regular Solutions* (Prentice-Hall, Englewood Cliffs, NJ, 1962).
17. J. H. Hildebrand and R. L. Scott, *Regular and Related Solutions* (Van Nostrand-Reinhold, Princeton, NJ, 1970).
18. A. F. M. Barton, *Handbook of Solubility Parameters and Other Cohesion Parameters* (CRC Press, Boca Raton, FL, 1983).
19. F. Gharagheizi, *QSAR Comb. Sci.*, **27,** 165 (2007).
20. A. R. Katritzky and D. C. Fara, *Energy Fuels*, **19,** 922 (2005).
21. Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties Relationships (DIPPR) (American Institute of Chemical Engineers (AIChE), 2006).
22. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, Germany, 2000).
23. Talete srl., Dragon for Widows (Software for Molecular Descriptor Calculation). Version 5.4 (2006), http://www.talete.mi.it/.
24. Hyperchem Release 7.5 for Windows., Molecular Modeling System, Hypercube, Inc. (2002).
25. R. Leardi, R. Boggia, and M. Terrile, *J. Chemometr.*, **6,** 267 (1992).
26. F. Gharagheizi, *Comput. Mat. Sci.*, **40,** 159 (2007).
27. F. Gharagheizi and M. Mehrpooya, *Energy Convers. Manage.*, **48,** 2453 (2007).
28. F. Gharagheizi, *e-Polymers*, Article Number 114, 2007.
29. F. Gharagheizi, *Thermochim. Acta*, **469,** 8 (2008).
30. F. Gharagheizi, *Chemometr. Intell. Lab. Sys.*, **91,** 177 (2008).
31. F. Gharagheizi and A. Fazeli, *QSAR Comb. Sci.*, **27,** 758 (2008).
32. F. Gharagheizi and R. F. Alamdari, *QSAR Comb Sci.*, **27,** 679 (2008).
33. F. Gharagheizi and R. F. Alamdari, *Fuller. Nanotub. Car. N.*, **16,** 40 (2008).
34. A. Vatani, M. Mehrpooya, and F. Gharagheizi, *Int. J. Mol. Sci.*, **8,** 407 (2007).
35. M. Sattari and F. Gharagheizi, *Chemosphere*, **72,** 1298 (2008).
36. F. Gharagheizi and M. Mehrpooya, *Mol. Divers.*, **12,** 143 (2008).
37. F. Gharagheizi, *Energy & Fuels*, **22,** 3037 (2008)
38. R. Todeschini, V. Consonni, and A. Maiocchi, *Chemometr. Intell. Lab. Syst.*, **46,** 13 (1999).
39. J. Taskinen and J. Yliruusi, *Adv. Drug. Deliver. Rev.*, **55,** 1163 (2003).
40. F. Gharagheizi, R. F. Alamdari, and M. T. Angaji, *Energy & Fuels*, **22,** 1628 (2008).