AUTOLOGP™: A computer tool for simulating *n*-octanol-water partition coefficients

J. Devillers

CTIS, 3 Chemin de la Gravière, 69140 Rillieux La Pape, France

AUTOLOGP™ (Version 4.0) for Windows™ 3.1 or Windows™ 95 is a log *P* calculator using a neural network model designed from 7200 molecules described by means of the autocorrelation method. The characteristics of the software are underlined and its simulation performances are presented through different examples. Comparisons with other models are also performed.

Introduction

The logarithm of the partition coefficient between *n*-octanol and water (log *P*) is a leading physicochemical descriptor in many quantitative structure-activity relationship (QSAR) studies for modeling transports across biological membranes, biochemical and pharmacokinetic processes, and (eco)toxicity of organic compounds [1-4]. Although experimental log *P* data exist for several thousands of chemicals, this number is minuscule compared to the total number of compounds for which accurate log *P* values are desirable. Therefore, there has been continuing interest in deriving models for estimating partition coefficients from the chemical structure of organic molecules [5-10]. However, most of the available log *P* models have been developed from insufficiently large training sets and unrelevant algorithms to ensure acceptable application domains and reliability. In addition, these models generally involve the use of numerous correction factors which are often difficult to manipulate. To solve this problem, we recently developed a powerful neural network model [11-13] using autocorrelation descriptors [14-16] for encoding the molecules. It has been

Article available at http://analusis.edpsciences.org or http://dx.doi.org/10.1051/analusis:1999270023

embedded into a user-friendly software called AUTOLOGP™ (Version 4.0). The aim of the present paper is to present the main characteristics of this model and show the simulation performances of AUTOLOGP™ (Version 4.0).

Model design

Why a Neural Network?

Discovering valuable quantitative relationships between molecular descriptors and biological activities (QSAR) or physicochemical properties (QSPR) is always laborious and often error prone. Because many of today's QSAR and QSPR problems are complex, it is important to use robust statistical tools. A survey of the literature shows that linear methods are widely used for deriving QSAR and QSPR models [e.g., 1-3, 17-19]. However, their proper use requires different statistical assumptions such as: the data are normally distributed, the variables are independent, and so on. Artificial neural networks are not limited by these assumptions. Indeed, neural networks are nonlinear computing devices that attempt to use some "organizational" and functioning principles believed to be used in the human brain. They are composed of interconnected computing units called neurons. Each neuron performs a few simple operations and communicates the results to its neighbors. Typically the neurons are organized into layers with each neuron in one layer having a connection to each neuron in the next layer. Associated with each connection is a weight and each neuron has an activation function [20,21]. The most widely used neural network algorithm is the backpropagation neural network (BNN) (Fig. 1). The BNN algorithm adjusts weights by presenting example training pairs of input-target patterns. An input pattern is presented at the input layer and is propagated through all the neurons in the network to produce output(s) at the output layer. This output pattern is then compared with the "ideal" target pattern, and an error is propagated back through the network. The propagated error is used to adjust the weights of the connections. This training process is then repeated with a new training pair, and a new error is propagated backwards. This process is repeated many times until the error value is acceptable at which time the network has been trained [20,21]. It is important to note that for obtaining a BNN model presenting good generalization capabilities, it is required to tune different parameters, to follow rules (e.g., number of cycles) and validate the model on a testing test containing data not included in the training set. Nowadays, BNN appears as a strong predictive statistical engine for uncovering complex structure-activity and structure-property relationships even if it gives little insight into the underlying mechanisms that describe these relationships [22]. Thus, recently we demonstrated that a three-layer BNN was a better statistical tool than a regression equation for simulating lipophilicity [23]. This prompted us to use a BNN for developing the model which is now implemented in AUTOLOGP™ (Version 4.0).

Model Description

Autocorrelation method [14-16] was used for encoding the molecules of the training set, the testing set, and the validation set allowing to control the learning process of the BNN. For designing these descriptors, the structural formula of a molecule is represented as a graph with nodes and edges, and physicochemical properties are associated with each atom or functional group constituting the molecule. A calculation procedure yields a series of numbers (components of an autocorrelation vector) in relation with the different internodal distances in the molecular graph. A detailed description of this algorithm with examples of calculation can be found in a previous publication [16]. It is important to note that to optimize the descriptive power and the weak redundancy of the autocorrelation vectors, a new autocorrelation algorithm was designed. Its principal feature is that the first component of the autocorrelation vectors is obtained by simply summing the positive and negative contributions attributed to the atoms and functional groups constituting the studied molecule. In our study, molecules were described by means of four different autocorrelation vectors representing their lipophilicity (H), molecular refractivity (MR), and H-bonding acceptor (HBA) and H-bonding donor (HBD) abilities [11].

The optimal architecture and set of parameters for the BNN were determined by means of a trial and error procedure involving a huge number of experiments. We found that 35 input neurons (i.e., H_0 to H_{14}, MR_0 to MR_{14}, HBA_0 to

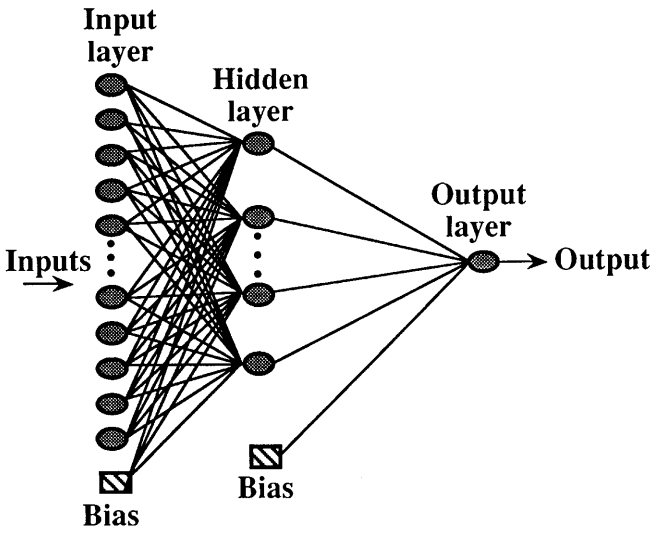


Figure 1. A backpropagation neural network.

HBA_3, and HBD_0) were required to correctly describe the molecules and model log *P*. The hidden layer consisted of 32 neurons. As regards the learning parameters, it was found that a learning rate (η) of 0.5 and a momentum term (α) of 0.9 always allowed us to obtain good BNN generalizations within c.a. 5500 cycles. As different models presenting a high predictive power were obtained, an ensemble of networks [24] constituted of interesting configurations was designed by testing several combinations. A composite network constituted of four configurations was selected as final model since it allowed us to obtain the best simulation results with the testing set. The final composite model applies to any uncharged organic chemical containing C, H, N, O, P, S, F, Cl, Br, and/or I. Its high modeling performances are depicted in figures 2 and 3 from the plots of the experimental versus calculated log *P* values for the compounds constituting the training and testing sets, respectively.

Simulation performances of AUTOLOGP™ (version 4.0)

Software description

AUTOLOGP™ (Version 4.0) is a user-friendly software for Windows™ 3.1 or Windows™ 95 implementing the model depicted in the previous section. It is obvious that it is not required to know the theory related to the development of the model to properly use the software. The software uses the classical SMILES notation [25] for input of the chemical structures and yields immediate estimation of log *P* values. AUTOLOGP™ (Version 4.0) proposes three calculation procedures. The first option is basically made for estimating the log *P* values of compounds one at a time by entering their SMILES strings (e.g., butane: CCCC; benzene: c1ccccc1). However, it presents other functionalities such as the possibility to import or export SMILES strings, the creation of a user's base and the recording of the results obtained during the user's session. The second option called "batch" allows the automatic computation of unlimited numbers of SMILES strings either entered with the program (User's files option) or contained in files generated by other software (Text files option). For running the batch computation, one has simply to select the type of files to treat and click on the "Start" button. Advancement in the calculations is visualized and can be stopped at any time. Results are displayed in a specific window and can be saved in a text file for exploitation by means of spreadsheet or text editors. The last option called "Simulation" allows to calculate the log *P* values of user-selected derivatives of a lead compound. To use this procedure, one has simply to enter the SMILES string of a molecule and indicate by a star (*****) the location of the substitution. The 150 available substituents are easily selected from a window. At the end of the calculations, the results are displayed in a specific window and can be saved in a text file.

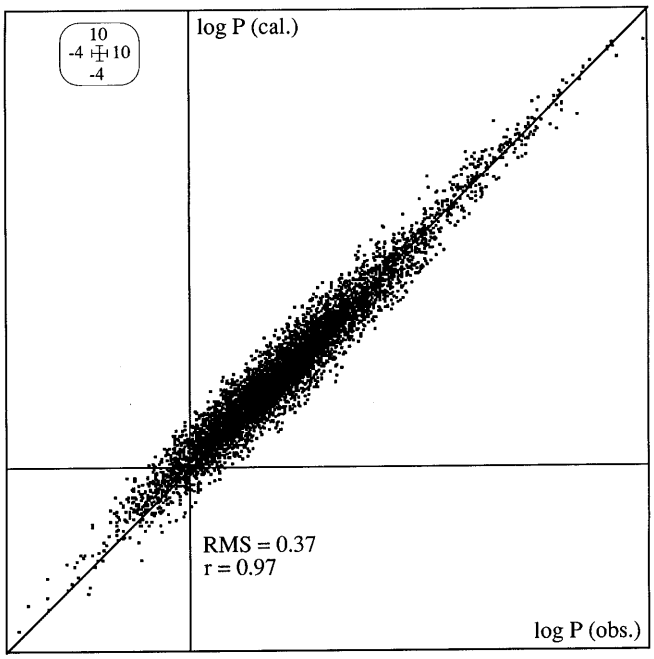


Figure 2. Calculated (cal.) log *P* values of the training set (7200 molecules) compared to experimental (obs.) data.

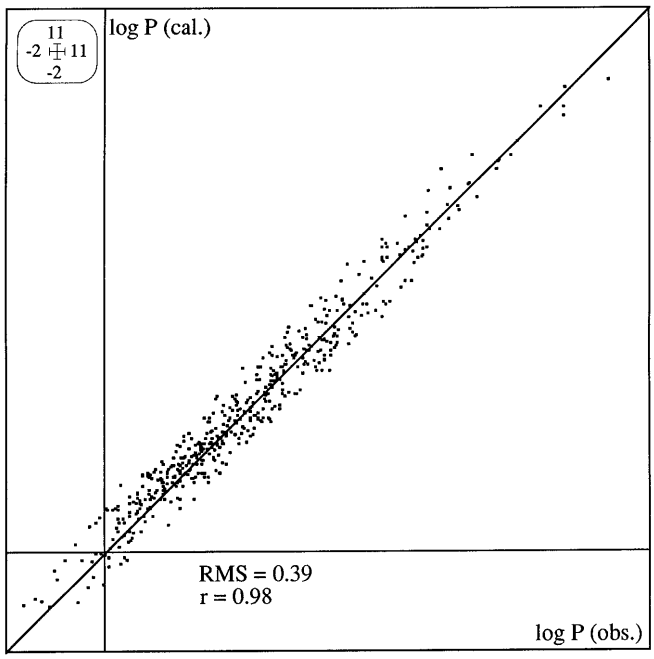


Figure 3. Calculated (cal.) log *P* values of the testing set (519 molecules) compared to experimental (obs.) data.

Figure 4 provides examples of simulations for 40 chemicals of pharmaceutical and environmental concerns. This figure confirms the good performances of the model

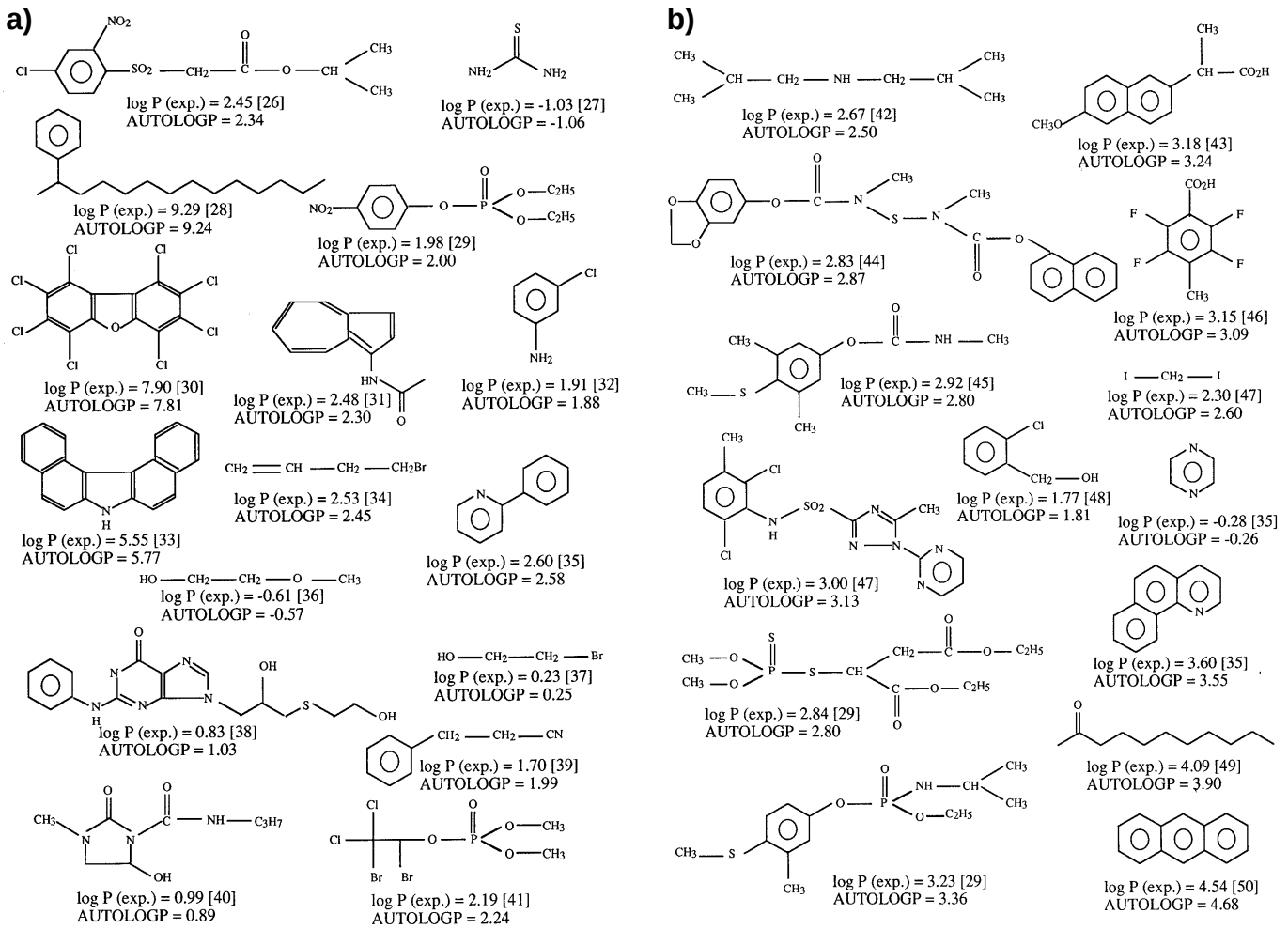


Figure 4. Comparison of experimental log *P* values with AUTOLOGP™ (Version 4.0) estimations for 40 structurally diverse compounds.

implemented in AUTOLOGP™ for highly diverse chemical structures.

Comparison with Other Models/Software

Many methods for estimating log *P* of organic molecules are available in the literature. The most common can be classified as "Atom/functional group contribution" methods in which the chemical structure of a molecule is divided into fragments "*f*" (i.e., atoms and functional groups) and their corresponding values are summed together to yield the log *P* estimate. These methods generally require the use of correction factors. The method of Rekker falls in this category [6]. Thus, the general Rekker's equation takes the following form:

$$\log P = \Sigma f + \Sigma k_n C_M$$

where "C_M" is the "magic constant" and "k_n" is the frequency of "C_M" occurring in a structure under consideration [6, p. 28]. Examples of calculation are given in tables I and II for diphenhydramine and Verapamil, respectively. It is obvious that on the basis of these two results we cannot claim that AUTOLOGP™ provides better simulations than the Rekker's methodology. However, it is important to stress that our model does not require correction factors while in the Rekker's method, they play a key role in the calculation procedure. This situation is particularly cumbersome because these correction factors are often difficult to manipulate and justify.

In the same way, most of the log *P* simulators are not able to distinguish isomers while our model encodes this crucial information from the autocorrelation method used for describing the molecules. This limitation is problematic since it is well known that the position of an atom or functional group in a molecule can significantly affect its log *P* value. This is clearly illustrated in table III where experimental log *P* values of biphenyl and 15 polychlorinated biphenyls are reported with simulation results obtained from six different software [58]. Table III also underlines another

c)

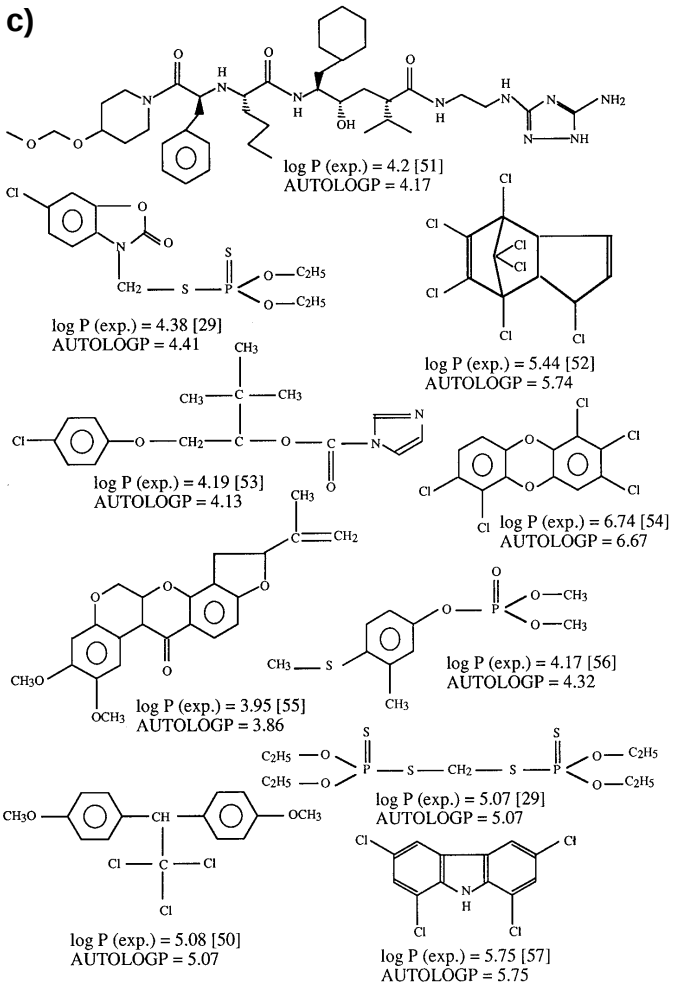
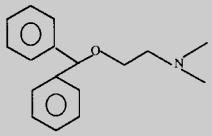


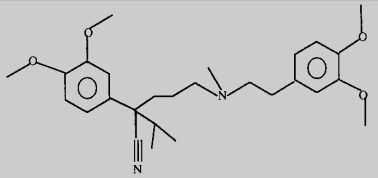
Table I. Calculated and observed log *P* values of diphenhydramine.



Fragment	*Calculation*	*procedure**
N(al.)	1 × (–2.074)	–2.074
O(al.)	1 × (–1.545)	–1.545
C	17 × (0.110)	1.870
H	21 × (0.204)	4.284
CM	4 × (0.219)	0.876
Total (Rekker)		3.41
AUTOLOGP™		3.05
Experimental		3.27 [6]

*Based on the "gross-formula approach" [6, p. 85 and 86].

Table II. Calculated and observed log *P* values of Verapamil.



Fragment	*Calculation*	*procedure**
O(ar.)	4 × (–0.450)	–1.800
CN(al.)	1 × (–1.031)	–1.031
N(al.)	1 × (-2.074)	–2.074
C	26 × (0.110)	2.860
H	38 × (0.204)	7.752
CM	2 × (0.219)	0.438
Total (Rekker)		6.15
AUTOLOGP™		4.25
Experimental		3.79 [6]

*Based on the "gross-formula approach" [6, p. 85 and 86].

shortcoming of the majority of the log *P* models. Indeed, based on a simple "additivity" principle or linear relationships, they cannot correctly mimic the effects of an increase of the degree of substitution of a molecule. More specifically, overestimations are often noted with molecules presenting a high degree of substitution.

Our model has been designed from a large training set of 7200 molecules encoded with information-rich autocorrelation descriptors. A backpropagation neural network has been used for discovering the best QSPR between the structure of the molecules and their lipophilicity. Under these conditions, it is not surprising to show that this model provides reliable calculation of log *P* values for uncharged organic compounds.

References

1. Karcher, W.; Devillers, J. Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology, Kluwer Academic Publishers, Dordrecht, 1990.
2. Devillers, J. Comparative QSA*R*, Taylor & Francis, Washington, DC, 1998.

Table III. Observed* and calculated log *P* values of biphenyl and 15 polychlorinated biphenyls.**

Substitution	*Exp.*	*M1*	*M2*	*M3*	*M4*	*M5*	*AUTOLOGP™*
Biphenyl	3.98	4.03	3.76	3.98	3.97	4.09	3.81
4-Cl	4.61	4.74	4.40	4.55	4.71	4.02	4.49
2,2'-Cl	4.73	4.96	5.05	4.93	5.45	4.68	4.99
4,4'-Cl	5.58	5.46	5.05	5.12	5.45	4.68	5.16
2,4,5-Cl	5.81	5.80	5.69	5.40	6.19	5.34	5.66
2,2',5-Cl	5.60	5.67	5.69	5.43	6.19	5.34	5.57
2,2',5,5'-Cl	6.09	6.38	6.34	5.92	6.93	5.99	6.06
2,3,4,5-Cl	6.41	6.27	6.34	5.73	6.93	5.99	6.19
2,2',4,5,5'-Cl	6.44	7.00	6.98	6.38	7.67	6.65	6.60
2,3,4,5,6-Cl	6.52	6.62	6.98	6.04	7.67	6.65	6.49
2,2',4,4',5,5'-Cl	6.80	7.57	7.62	6.82	8.41	7.30	7.14
3,3',4,4',5,5'-Cl	7.55	7.83	7.62	6.90	8.41	7.30	7.35
2,2',3,3',4,4',6-Cl	6.99	8.25	8.26	7.16	9.15	7.96	7.51
2,2',3,3',5,5',6,6'-Cl	7.15	8.25	8.91	7.41	9.89	8.61	7.54
2,2',3,3',4,5,5',6,6'-Cl	8.16	8.73	9.56	7.76	9.89	9.27	7.91
Deca-Cl	8.26	9.20	10.2	8.10	11.4	9.92	8.33

*Recommended values from Sangster [58; Tab. 6.10, p. 148].
**Calculated values [58; Tab. 6.10, p. 148] with M1 (A. Leo and C. Hansch, ClogP for Windows 1.0.0), M2 (WM Meylan and PH Howard, KOWWIN 1.53 [10]), M3 (ACD/Log*P* 1.0. This model uses 532 group contributions, 21 carbon atom type contributions, and 2206 intramolecular correction factors [58, p. 116]), M4 (Rekker's method, Prolog*P* 5.1), M5 (Klog*P* contained in ToxAlert 1.2 (see [9] for more information)).

3. Hansch, C.; Leo, A. Exploring QSAR. Fundamentals and applications in chemistry and biology, American Chemical Society, Washington, DC, 1995.
4. Pliska, V.; Testa, B.; van de Waterbeemd, H. Lipophilicity in drug action and toxicology, VCH, Weinheim, 1996.
5. Nys, G. G.; Rekker, R. F. *Eur. J. Med. Chem. - Chim. Ther.* **1974**, *9*, 361-375.
6. Rekker, R. F.; Mannhold, R. Calculation of drug lipophilicity. The hydrophobic fragmental constant approach, VCH, Weinheim, 1992.
7. Leo, A. *Chem. Rev*. **1993**, *93*, 1281-1306.
8. Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem. -Chim. Ther.* **1984**, *19*, 71-78.
9. Klopman, G.; Li, J. Y.; Wang, S.; Dimayuga, M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752-781.
10. Meylan, W. M.; Howard, P. H. *J. Pharm. Sci.* **1995**, *84*, 83-92.
11. Devillers, J.; Domine, D.; Guillon, C.; Bintein, S.; Karcher, W. *SAR QSAR Environ. Res*. **1997**, *7*, 151-172.
12. Devillers, J.; Domine, D. *SAR QSAR Environ. Res.* **1997**, *7*, 195-232.
13. Domine, D.; Devillers, J. *Sci. Comput. Autom.* 1998, March, 55-63.
14. Moreau, G.; Broto, P. *Nouv. J. Chim*. **1980**, *4*, 359-360.
15. Moreau, G.; Broto, P. *Nouv. J. Chim*. **1980**, *4*, 757-764.
16. Broto, P.; Devillers, J. In: Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology; Karcher, W.; Devillers, J. Eds., Kluwer Academic Publishers, Dordrecht, 1990; pp 105-127.
17. Kubinyi, H. QSAR: Hansch analysis and related approaches, VCH, Weinheim, 1993.
18. Livingstone, D. Data analysis for chemists, Oxford University Press, Oxford, 1995.
19. Ford, M. G.; Greenwood, R.; Brooks, G. T.; Franke, R. Bioactive compound design: Possibilities for industrial use, SCI Bios Scientific, UK, 1996.
20. Goonatilake, S. In: Intelligent systems for finance and business; Goonatilake, S.; Treleaven, P., Eds., John Wiley & Sons, Chichester, 1995; pp 1-28.
21. Devillers, J. Neural networks in QSAR and drug design, Academic Press, London, 1996.
22. Devillers, J. In: Neural networks in QSAR and drug design; Devillers, J. Ed., Academic Press, London, 1996; pp 1-46.
23. Domine, D.; Devillers, J.; Karcher, W. In: Neural networks in QSAR and drug design; Devillers, J. Ed.; Academic Press, London, 1996; pp 47-63.
24. Hansen, L. K.; Salamon, P. *IEEE Trans. Pattern Anal. Machine Intell*. **1990**, *12*, 993-1001.
25. Weininger, D.; Weininger, J. L. In: Comprehensive medicinal chemistry; Ramsden, C. A. Ed., Pergamon Press, Oxford, 1990; Vol. 4, pp 59-82.
26. Hong, H.; Han, S.; Wang, X.; Wang, L.; Zhang, Z.; Zou, G. *Environ. Sci. Technol.* **1995**, *29*, 3044-3048.
27. Govers, H.; Ruepert, C.; Stevens, T.; van Leeuwen, C. J. *Chemosphere* **1986**, *15*, 383-393.
28. Sherblom, P. M.; Gschwend, P. M.; Eganhouse, R. P. *J. Chem. Eng. Data* **1992**, *37*, 394-399.
29. Bowman, B. T.; Sans, W. W. *J. Environ. Sci. Health* **1983**, *B18*, 667-683.
30. Doucette, W. J.; Andren, A. W. *Chemosphere* **1988**, *17*, 345-359.
31. Lichtenwalner, M. R.; Speaker, T. J. *J. Pharm. Sci.* **1980**, *69*, 337-339.
32. de Bruijn, J.; Busser F.; Seinen, W.; Hermens, J. *Environ. Toxicol. Chem.* **1989**, *8*, 499-512.
33. Bond, J. A.; Baker, S. M.; Bechtold, W. E. *Toxicology* **1985**, *36*, 285-295.
34. Tewari, Y. B.; Miller, M. M.; Wasik, S. P.; Martire, D. E. *J. Chem. Eng. Data* **1982**, *27*, 451-454.
35. de Voogt, P.; Wegener, J. W. M.; Klamer, J. C.; van Zijl, G. A.; Govers, H. *Biomed. Environ. Sci.* **1988**, *1*, 194-209.
36. Tanii, H.; Saito, S.; Hashimoto, K. *Arch. Toxicol.* **1992**, *66*, 368-371.
37. Dillingham, E. O.; Mast, R. W.; Bass, G. E.; Autian, J. *J. Pharm. Sci.* **1973**, *62*, 22- 30.
38. Xu, H.; Maga, G.; Focher, F.; Smith, E. R.; Spadari, S.; Gambino, J.; Wright, G. E. *J. Med. Chem.* **1995**, *38*, 49-57.
39. Tanii, H.; Hashimoto, K. *Arch. Toxicol.* **1984**, *55*, 47-54.
40. Yanase, D.; Chiba, M.; Andoh, A.; Gotoh, T. *Pestic. Sci.* **1995**, *43*, 279-285.
41. Saito, H.; Koyasu, J.; Yoshida, K.; Shigeoka, T.; Koike, S. *Chemosphere* **1993**, *26*, 1015-1028.
42. Gagnaire, F.; Azim, S.; Simon, P.; Cossec, B.; Bonnet, P.; de Ceaurriz, J. *J. Appl. Toxicol.* **1993**, *13*, 129-135.
43. Unger, S. H.; Cook, J. R.; Hollenberg, J. S. *J. Pharm. Sci.* **1978**, *67*, 1364-1367.

44. Wallace, G. C.; Zerba, E. N. *Pestic. Sci.* **1989**, *26*, 215-225.
45. Briggs, G. G. *J. Agric. Food Chem.* **1981**, *29*, 1050-1059.
46. Tench, A. J.; Williams, R. H.; Bromilow, R. H.; Chamberlain, K. *Pestic. Sci.* **1993**, *37*, 31-37.
47. Percival, A. *Pestic. Sci.* **1991**, *31*, 569-580.
48. Miyake, K.; Kitaura, F.; Mizuno, N.; Terada, H. *Chem. Pharm. Bull.* **1987**, *35*, 377-388.
49. Tanii, H.; Tsuji, H.; Hashimoto, K. *Toxicol. Lett.* **1986**, *30*, 13-17.
50. Karickhoff, S. W.; Brown, D. S.; Scott, T. A. *Water Res.* **1979**, *13*, 241-248.
51. Boyd, S. A.; Fung, A. K. L.; Baker, W. L.; Mantei, R. A.; Stein, H. H.; Cohen, J.; Barlow, J. L.; Klinghofer, V.; Wessale, J. L.; Verburg, K. M.; Polakowski, J. S.; Adler, A. L.; Calzadilla, S. V.; Kovar, P.; Yao, Z.; Hutchins, C. W.; Denissen, J. F.; Grabowski, B. A.; Cepa, S.; Hoffman, D. J.; Garren, K. W.; Kleinert, H. D. *J. Med. Chem.* **1994**, *37*, 2991-3007.
52. Veith, G. D.; DeFoe, D. L.; Bergstedt, B.V. *J. Fish Res. Board Can.* **1979**, *36*, 1040-1048.
53. Imai, T.; Uchida, T.; Yamaguchi, K.; Takao, H.; Goto, T. *Pestic. Sci.* **1994**, *40*, 9-16.
54. Sijm, D. T. H. M.; Wever, H.; de Vries, P. J.; Opperhuizen, A. *Chemosphere* **1989**, *19*, 263-266.
55. Hermens, J.; Leeuwangh, P. *Ecotoxicol. Environ. Safety* **1982**, *6*, 302-310.
56. de Bruijn, J.; Hermens, J. *Environ. Toxicol. Chem.* **1991**, *10*, 791-804.
57. Burkhard, L. P.; Kuehl, D. W. *Chemosphere* **1986**, *15*, 163-167.
58. Sangster, J. Octanol-water partition coefficients: Fundamentals and physical chemistry, John Wiley & Sons, Chichester, 1997.