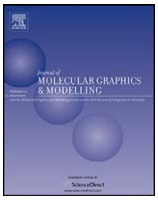
Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Robust modelling of solubility in supercritical carbon dioxide using Bayesian methods

Anna Tarasova [a], Frank Burden [a], Johann Gasteiger [b], David A. Winkler [a,*]

[a] CSIRO Molecular & Health Technologies, Private Bag 10, Clayton South MDC, Clayton, Victoria 3168, Australia
[b] Molecular Networks GmbH, IZMP, Henkestr. 91, D-91052 Erlangen, Germany

ARTICLE INFO

Article history:
Received 29 September 2009
Received in revised form 9 December 2009
Accepted 10 December 2009
Available online 16 December 2009

Keywords:
Bayesian methods
Structure–property relationships
Dyes
Solubility
Supercritical carbon dioxide

ABSTRACT

Two sparse Bayesian methods were used to derive predictive models of solubility of organic dyes and polycyclic aromatic compounds in supercritical carbon dioxide ($scCO_2$), over a wide range of temperatures (285.9–423.2 K) and pressures (60–1400 bar): a multiple linear regression employing an expectation maximization algorithm and a sparse prior (MLREM) method and a non-linear Bayesian Regularized Artificial Neural Network with a Laplacian Prior (BRANNLP). A randomly selected test set was used to estimate the predictive ability of the models. The MLREM method resulted in a model of similar predictivity to the less sparse MLR method, while the non-linear BRANNLP method created models of substantially better predictivity than either the MLREM or MLR based models. The BRANNLP method simultaneously generated context-relevant subsets of descriptors and a robust, non-linear quantitative structure–property relationship (QSPR) model for the compound solubility in $scCO_2$. The differences between linear and non-linear descriptor selection methods are discussed.

1. Introduction

Industrial processes use huge amounts of volatile organic solvents annually, posing a health and safety, environmental and potential climate hazard. Much of the estimated 20 million tons of volatile organic solvents per annum is released to the atmosphere. Supercritical carbon dioxide ($scCO_2$) is an environmentally benign solvent that shows much promise as a non-toxic, non-flammable, inexpensive, readily available, and easily removable solvent for a wide variety of processes such as chemical syntheses, compound extraction, and materials' processing. The $scCO_2$ is also an attractive solvent as it has a relatively low critical temperature and pressure (304 K and 73 bar, respectively) [1].

Our interest in $scCO_2$ lies in its potential use as a dye solvent for application in the area of organic electronics. A number of small, relatively homologous data sets, relating the molecular structure, temperature and pressure to the solubility of dyes in $scCO_2$, have been reported in the literature. An application of a mathematical model to interpret the available dye solubility data sets is an important method for a time and money saving prediction of solubility of existing and future dye structures. Quantitative structure–property relationship (QSPR) modelling is a mathematical relationship that has been used extensively in the pharmaceutical and environmental industries to model a very diverse range of biological and physicochemical properties of organic compounds, including solubility prediction of organic compounds in supercritical solvents [2–6]. There are a number of ways of generating QSPR models by linking mathematical representations of molecular structure to the materials' property of interest. The commonly used linear methods in QSPR studies of compound solubility in $scCO_2$ have been regression techniques such as multiple linear regression (MLR) [7] and partial least squares (PLS). For categorical rather than continuous data, methods such as decision trees and support vector machines are effective. For example, Fat'hi et al. [8] reported a simple equilibrium solvation model for four anthraquinone derivatives in $scCO_2$ that related solubility to temperature, pressure and $scCO_2$ density. This type of model was also used to model single anthraquinone derivatives [9,10], as well as a group of another seven anthraquinone analogues [11]. The equilibrium solvation approach is, however, limited by the need to measure quantities such as fluid density that may not be available for structurally complex solutes, and the applicability of the method to a single chemotype. In general, equation of state models can describe the solubility of compounds as a function of both temperature and pressure in the critical region of CO_2. However, these models require at least one empirical parameter derived from experimental solubility data. Although QSPR models require measured data to train them, unlike equation of state models they are capable of predicting the

[*] Corresponding author. Tel.: +61 3 9545 2477; fax: +61 3 9545 2447.
E-mail address: dave.winkler@csiro.au (D.A. Winkler).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.12.004

properties of new chemotypes not used in training, so long as the domain of applicability of the model is understood.

Since the relationship between molecular structure and properties is often complex, more sophisticated methods, such as back-propagation neural networks, have been adopted for QSPR modelling of solubilities of small series of dyes. Tabaraki et al. compared wavelet neural network (WNN), multiple linear regression (MLR), or artificial neural network (ANN) QSPR models of $scCO_2$ solubility of twenty-five anthraquinone [5] and twenty-one azo dyes [6] over a range of pressures and temperatures. Khayamian and Esteki [4] reported a similar WNN model for five polycyclic aromatic compounds. More recently, Hemmateenejad et al. [3] published a neural network QSPR model for $scCO_2$ solubility of twenty-nine anthraquinone, anthrone and xanthone derivatives.

Generally, the described neural networks suffer from a number of disadvantages such as over-training, over-fitting, and difficulty in finding the best neural network architecture. Intensive research into quantitative structure–activity relationships (QSAR) methods for the pharmaceutical industry has resulted in the discovery of powerful new methods in recent years that largely overcome these shortcomings. The Bayesian regularized neural networks have been shown to impose optimum parsimony on QSAR models, ensuring they have the best predictive power [12–16]. These new neural network methods have been shown to perform efficient feature selection and generate robust QSAR models without over-training or over-fitting, or risk of chance correlations. If a small number of descriptors are selected from a large pool of possibilities by forming different subsets repeatedly, a chance correlation can arise. Many methods for choosing context dependent features from large sets of descriptors have been published in the literature, but none are universally robust and free from error or bias. These include principal component analysis (PCA) for dimensional reduction of descriptors, clustering of descriptors by *k*-means and self-organized maps, and descriptor selection via genetic algorithms and forward selection/backward elimination

The sparse feature selection and modelling methods have proven invaluable for selecting the best set of molecular properties or descriptors for QSAR modelling [13,17]. For example, we have adapted the multiple linear regression with expectation maximization (MLREM) method to carry out relevant descriptor selection with very good results [17]. This method identified a sparse set of context-relevant descriptors from a larger pool without reducing the predictivity of the model. The Bayesian Regularized Artificial Neural Network with a Laplacian Prior (BRANNLP) method is another self-pruning network that is useful for modelling data sets where the QSAR is non-linear [13]. This is a classical feed forward method that employs a sparse Laplacian prior to carry out non-linear descriptor selection, as well as pruning of the weights (and complexity) of the neural network. This method provides optimum parsimony that balances *bias* (model too simple to reflect underlying relationship) with *variance* (overly complex model that also fits noise) ensuring highest predictivity. The method selects descriptors that have minimal correlation, can be chemically interpreted and suitable for explaining the response variable.

Within the scope of the authors' knowledge, linear and non-linear QSPR models for a large and chemically diverse data set of compound solubility in $scCO_2$ have not been reported in the literature. In this study we investigated whether robust predictive models could be derived using the modern linear and non-linear QSPR methods, MLREM and BRANNLP, respectively. The methods were employed to generate QSPR models of a solubility of chemically diverse dyes and polycyclic aromatic in $scCO_2$, at different temperatures (285.9–423.2 K) and pressures (60–1400 bar). The predictive properties of the models were evaluated using an external set of molecules not used in creating the models.

2. Experimental

2.1. Data set

We assembled a large data set of molecules soluble in $scCO_2$ with a broader chemical diversity than those used in earlier QSPR models [4–6]. Molecular structures of the compounds present in the data set are listed in Supplementary Information. The data set was relatively balanced, such that solubility data for one type of molecular structure did not dominate or bias the final data set which contained 67 molecular structures, with a total of 685 solubility data points (see Supplementary Information). The solubility measurements are reported as the log of mole fraction of solute in $scCO_2$ ($\log(S)$exp) which ranged from −8.233 (for Disperse Red 9 dye at 313.20 K and 100 bar) to −1.759 log units (for Disperse Yellow 7 dye at 393.15 K and 250 bar). The data set was randomly divided into a training set (80% of the data; 548 data points) and a test set (20% of the data; 137 data points). The measured solubilities in these studies had variable reported uncertainties of ±0.5 [18], ±2 [9], ±3 [8,19], ±4 [20,21], ±5 [10,11], ±10 [22] and ±15% [23]. The X variables (molecular descriptors) of the data set were auto-scaled (mean centered and divided by standard deviation) and the Y variable ($\log(S)$exp) scaled to lie between zero and one.

2.2. Molecular descriptors

Compound structures were generated using the Sybyl8.1 molecular modelling package (Tripos Associates, St Louis). A total of fifty-three topological, electronic, geometric and physicochemical descriptors were calculated, using Sybyl8.1 or CSIRO-Biomodeller to represent each compound (see Supplementary Information for a complete list of descriptors). The topological descriptors included the number of bonds and atom types and the molecular weight, and are contained within the atomistic (A) index [24]. The Burden index (B) encodes the connectivity of the molecules and nature of the valence electrons (eigenvalue descriptors) [15]. The binned charges (BC) index contained descriptors that describe the charge properties of the compounds (and indirectly the dipolar and hydrogen bonding properties) which were encoded by charge fingerprint descriptors [25]. The three described indices are complementary and have been shown previously to yield better models when used in combination rather than individually [26]. The geometric descriptors encode information that depends on the three-dimensional structure of the compound, such as surface area of the molecule, its molar refractivity and molecular volume. Example of a physicochemical descriptor is the hydrophobicity of a molecule. The compounds' both geometric and electronic properties are encoded in the PPSA1 and PNSA1 descriptors.

2.3. QSPR modelling

Using an in-house modelling package, CSIRO-Biomodeller, the MLR, MLREM and BRANNLP methods were used to derive QSPR models of the solubility data set. A commercial package, which can generate the descriptors used in this study, and a Bayesian regularized neural network method are available as part of the Know-It-All software package (Bio-RAD Corporation). The MLR method makes use of all the calculated descriptors in the model, whereas both the MLREM and BRANNLP methods pruned out the least informative descriptors by the sparse Laplacian prior feature selection method. The sparsity of the MLREM and BRANNLP methods was tuned progressively, using specific control parameters [13,17], until the quality of the derived models deteriorated

Table 1
Statistics of the best MLR, MLREM and BRANNLP-derived QSPR models of solubility of 64 types of compounds in $scCO_2$.

Method	N_{hidden}	$N_{d(i)}$	$N_{d(r)}$	Training set		Test set	
				SEE	r^2	SEP	q^2
MLR	–	53	53	0.56 ± 0.01	0.78 ± 0.01	0.59 ± 0.06	0.74 ± 0.04
MLREM	–	53	30	0.58 ± 0.01	0.76 ± 0.01	0.58 ± 0.06	0.73 ± 0.05
BRANNLP	2	53	38	0.40 ± 0.04	0.83 ± 0.05	0.45 ± 0.05	0.82 ± 0.05

N_{hidden}: number of nodes in the hidden layer; $N_{d(i)}$: number of initial descriptors; $N_{d(r)}$: number of relevant descriptors; SEE: standard error of estimation (training set); r^2: squared correlation coefficient of the training set; SEP: standard error of prediction (test set); q^2: squared correlation coefficient of the test set.

drastically, indicating that some of the most relevant descriptors have been removed.

The BRANNLP network used here consisted of input, hidden, and output layers. The number of nodes in the input layer is equal to the number of chosen descriptors and the output layer has one node that predicts compound solubility. The number of nodes in the hidden layer was varied over a limited range, being mindful that too few nodes may unduly restrict the complexity of the model. In addition, the Bayesian regularization automatically prevents the model from becoming overly complex if too many hidden layer nodes are specified. The neural networks were fully connected between the input and hidden layers and between the hidden and output layers. The number of effective weights is determined during the training of the network and tends to asymptote to a constant value as the number of hidden layer nodes is increased. The details of the Bayesian regularization applied to back-propagation neural networks can be found in previous publications [12,15]. Sigmoidal and linear transfer functions were used in the hidden and output layers, respectively. The training set was used to train the network using the BFGS method until a maximum was attained in the Bayesian evidence (maximum likelihood method). This avoids the need to use a validation set to determine when training should be stopped in order to prevent over-training of the network.

We used the standard error of estimation (SEE) for the training set and standard error of prediction (SEP) for the test set to assess the predictive quality of the derived models, reporting those models with the lowest SEP values. Other common statistical parameters, such as r^2, were listed but since these are dependent on the number of parameters in the model and size of the training set, they are not considered to be as robust estimators of models' quality.

3. Results and discussion

3.1. Performance of the models

The statistics of the MLR, MLREM and BRANNLP generated QSPR models of the data set are summarized in Table 1. The BRANNLP model gives the best overall training and test set statistics, with SEE and SEP values of 0.34 and 0.46, respectively, of the log(S)exp data. This translates to predicted $scCO_2$ solubilities (mol fractions) within a factor of two for the training set data and within a factor of three for the test set, across the diverse range of temperatures, pressures and solubilities. The MLR and MLREM derived models, on the other hand, had comparatively inferior statistics. The SEE value was 0.56 for both MLR and MLREM models and the derived SEP values were 0.58 and 0.60, respectively. This indicates that using these two models, the training and test solubility mol fractions could be predicted only within a factor of four.

The SEP values for MLR and MLREM models are very similar at ~0.6. From this it can be concluded that the overall performance of the MRLEM model is very similar to that of MLR model, despite the reduced number of descriptors (thirty descriptors retained by MLREM compared with the fifty-three descriptors used in the MLR method). The BRANNLP modelling of the data set reduced the descriptors to a total of thirty-eight (Table 1). The BRANNLP-derived QSPR model had improved predictivity of solubility compared to the linear models, with an SEP value of 0.42. A graphical representation of the predicted versus experimental solubility for the BRANNLP model is shown in Fig. 1. This result strongly suggests that the relationship between the compounds' solubility and the relevant descriptors, selected by the BRANNLP method, is reasonably non-linear. Superior non-linear models of $scCO_2$ solubility and selected descriptors were similarly obtained

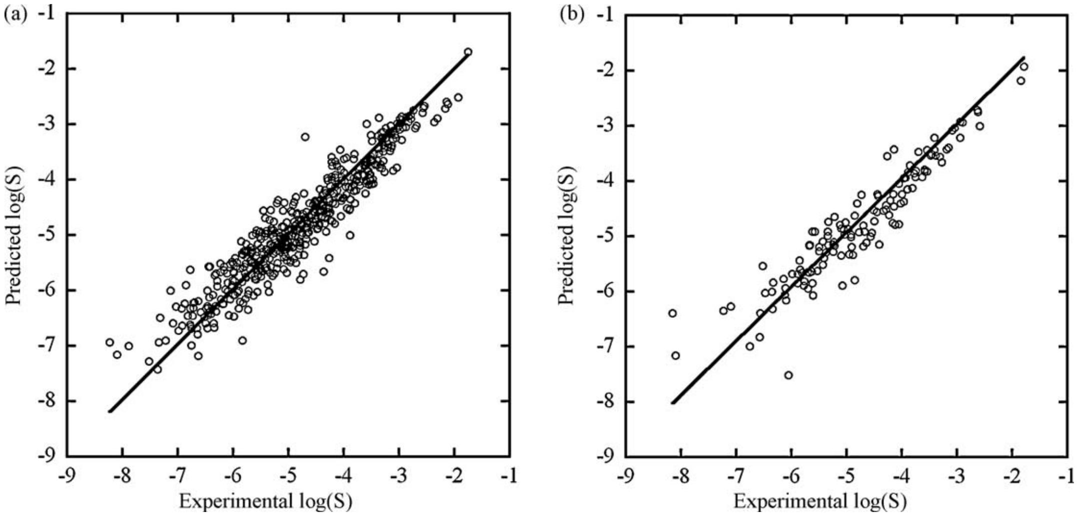


Fig. 1. A graphical representation of the performance of the BRANNLP model showing the BRANNLP predicted versus experimental solubility (log(S)) values for the (a) training and (b) test sets.

Table 2

The number of molecular descriptors ($N_{d(eff)}$), selected by the best sparse MLREM and BRANNLP methods, for QSPR modelling of compound solubility in $scCO_2$. Descriptors not common to both methods are italicized.

Method	$N_{d(eff)}$	Molecular descriptors[a]
MLREM	30	A6, *A8*, A10, A11, A15, A35, *G5*, E0, E1, E2, E4, E5, E6, E9, *BCGM1*, *BCGM2*, BCGM3, BCGM4, BCGM5, BCGM7, *BCGM8*, BCGM9, BCGM10, BCGM12, clogP, Pressure, Donor, Rotatable Bonds, PPSA1, CMR
BRANNLP	38	*A5*, A6, *A7*, *A10,* A11, *A14*, A15, A35, *A53*, *G6*, E0, E1, E2, E4, E5, E6, *E7*, *E8*, E9, BCGM3, BCGM4, BCGM5, *BCGM6*, BCGM7, BCGM9, BCGM10, BCGM12, clogP, *Temp*, *Total area*, *Ring Count*, Pressure, Donor, *PNSA1*, Rotatable Bonds, PPSA1, CMR, *Acceptor*

[a] The full names of the molecular descriptors are listed in Supplementary Information.

by Hemmateenejad et al., albeit for a less chemically diverse and smaller solubility data set of anthraquinone, anthrone and xanthone derivatives [9].

3.2. Significance of the descriptors used in the models

The descriptors selected via the linear and non-linear methods are listed in Table 2. An examination of these may be useful for providing some insight into the effect of the factors on the solubility of compounds in $scCO_2$.

The descriptors selected by the best sparse MLREM and BRANNLP models were the pressure, hydrophobicity (clogP), number of hydrogen-bond donor groups (donor), positively charged polar surface area (PPSA1), the number of rotatable bonds (a measure of molecular flexibility) and molar refractivity (CMR). The retention of the hydrophobicity descriptor is reasonable because of the hydrophobic nature of the supercritical carbon dioxide. For example, some anthraquinone derivatives have been shown to have increased solubility due to presence of hydrophobic substituents [11]. The molar refractivity (which describes the steric bulk of compounds), and the number of rotatable bonds, may also affect solubility by influencing the interactions between the solutes and carbon dioxide. It is likely that the hydrogen-bond donor descriptor was selected as the solutes may interact with one another or with carbon dioxide molecules. This formation of the inter- or intra-molecular hydrogen bonds was proposed by other studies, to influence the supercritical solubility [8,9,19,27–29].

The important atomic descriptors were identified by the best non-linear BRANNLP model to be the number of carbon atoms with two, three bonds and four bonds, number of nitrogen atoms with two and three bonds, number of oxygen atoms with one and two bonds, and number of chlorine atoms. These descriptors correlate positively (carbon and chlorine [28] atoms) or negatively (electronegative and partially charged heteroatoms [28]) with hydrophobicity, a property described earlier as important for $scCO_2$ solubility. Additional atomic descriptors selected by the MLREM method were aromatic carbon atoms that are also likely to correlate positively with the hydrophobicity descriptor.

3.3. Is non-linear feature selection important for non-linear models?

Overwhelmingly, QSAR and QSPR studies use linear feature selection methods, such as multiple linear regression (MLR) and principal components analysis, to select the most relevant descriptors, even when these are subsequently employed in non-linear models. However, linear selection of descriptors may not be the best method in situations where descriptors influence the outcome in a complex manner. In our study we know that both temperature and pressure affect $scCO_2$ density, a key factor in the solubility of compounds in supercritical fluids. We therefore expected all feature selection methods to choose temperature and pressure as relevant descriptors for modelling. However, the MLREM model with optimum sparsity did not retain the temperature as one of its relevant descriptors, whereas the best sparse BRANNLP model did. This is indicative of the complex and non-linear effects of the changing temperature and pressure values on the properties of the solute as well as the solvent. A number of $scCO_2$ solubility studies have found that the solubility-pressure isotherms of a group of solutes have been commonly shown to exhibit an intersection point for inversion of solubility with increase in temperature [8,10,11,19,21,28,30–33]. This behavior may be explained by occurrence of two opposing effects as temperature and pressure is increased. At a constant pressure, the increase in temperature results in a decrease of $scCO_2$ density and therefore solubility is generally decreased. The intersection point, and hence change in solubility, is caused by two opposing effects: (i) at low pressure values, the decrease of solvent density is dominant and the overall solubility decreases with increasing temperature, and (ii) at higher pressure values, solvent density is only slightly dependent on the temperature and the normal dependence of solubility on temperature dominates.

Therefore, the pruning of temperature as a relevant descriptor by the sparse linear MLREM method we attribute the difficulty in finding a relevant linear function to approximate a quite non-linear one like the relationship between temperature the solubility in $scCO_2$. The non-linear sparse Bayesian neural network method can accommodate this complex, non-linear behavior and temperature was therefore retained as an important parameter in this model. The inclusion of temperature as a relevant descriptor, and the fact that the BRANNLP model had higher predictive abilities of the data set, strongly supports complexity and non-linear relationship between solute solubility in $scCO_2$ and parameters such as temperature and pressure.

Other geometric and hybrid descriptors that were selected by either the MLREM or BRANNLP methods, but not both were the total molecular surface area (total area), ring count, partial negatively charged surface area (PNSA1) and hydrogen-bond acceptor groups. Caution must be used when comparing the descriptors that are not common to both descriptor selection methods as a number of the topological and atomistic descriptors also varied between the two methods. Some of these may correlate with the physicochemical descriptors selected or omitted by the linear or non-linear descriptor selection methods, complicating any simple comparison. The most interesting conclusion that can be drawn is that the non-linear feature selection method produced a different but largely overlapping list of descriptors to that generated by linear feature selection methods. Our work shows that it may be necessary to choose descriptors in a non-linear way for generation of a model that can predict complex and non-linear behaviors. The BRANNLP method provides a simple and robust method for achieving this.

3.4. Performance of the Bayesian modelling methods

Considering the structural diversity of the dyes used in this study, the performance of the BRANNLP-derived QSPR model was found to be very good when compared to other solubility models generated from much smaller and structurally homogeneous data sets. For instance, a QSPR model of twenty-five anthraquinone dyes, created using the wavelet neural network (WNN) method, had a root mean square error (RMSE) value of 0.34 for the validation set [5]. The BRANNLP method used here resulted in a robust model capable of predicting the solubility of structurally diverse dyes in a completely independent test set with an SEP of 0.46. Comparison of our work to another modelling study of the azo dyes [6], that also used the wavelet neural network (WNN) modelling, was not possible due to the incorrect experimental solubility values used in that study [19].

4. Conclusions

We have shown that two sparse Bayesian methods, the linear MLREM and the non-linear BRANNLP, enable the development of robust and predictive models for the solubility of dyes and polycyclic compounds in $scCO_2$. The QSPR models were derived from training data of substantially higher molecular diversity than previously published models. This work demonstrates that by using a Bayesian regularized non-linear neural network method with a Laplacian prior it is possible to devise a reasonably general model with a predictive and useful capability. The removal of temperature as a descriptor by the MLREM method, but not by the BRANNLP method, shows that temperature is weakly and non-linearly predictive of solubility of dyes and polycyclic compounds in $scCO_2$.

Acknowledgements

We gratefully acknowledge financial support from the CSIRO Advanced Materials Transformational Platform and the helpful discussions with Dr. Gavin Collis from CSIRO.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.12.004.

References

[1] W. Wu, W. Li, B. Han, Z. Zhang, T. Jiang, Z. Liu, A green and effective method to synthesize ionic liquids: supercritical CO_2 route, Green Chem. 7 (2005) 701.

[2] P. Battersby, J.R. Dean, W.R. Tomlinson, S.M. Hitchen, Predicting solubility in supercritical fluid extraction using a neural network, Analyst 119 (1994) 925.

[3] B. Hemmateenejad, M. Shamsipur, R. Miri, M. Elyasi, F. Foroghinia, H. Sharghi, Linear and nonlinear quantitative structure–property relationship models for solubility of some anthraquinone, anthrone and xanthone derivatives in supercritical carbon dioxide, Anal. Chim. Acta 610 (2008) 25.

[4] T. Khayamian, M. Esteki, Prediction of solubility for polycyclic aromatic hydrocarbons in supercritical carbon dioxide using wavelet neural networks in quantitative structure–property relationship, J. Supercrit. Fluids 32 (2004) 73.

[5] R. Tabaraki, T. Khayamian, A.A. Ensafi, Wavelet neural network modeling in QSPR for prediction of solubility of 25 anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide, J. Mol. Graphics Model. 25 (2006) 46.

[6] R. Tabaraki, T. Khayamian, A.A. Ensafi, Solubility prediction of 21 azo dyes in supercritical carbon dioxide using wavelet neural network, Dyes Pigments 73 (2007) 230.

[7] H.L. Engelhardt, P.C. Jurs, Prediction of supercritical carbon dioxide solubility of organic compounds from molecular structure, J. Chem. Inf. Comput. Sci. 37 (1997) 478.

[8] M.R. Fat'hi, Y. Yamini, H. Sharghi, M. Shamsipur, Solubilities of some recently synthesized 1,8-dihydroxy-9,10-anthraquinone derivatives in supercritical carbon dioxide, Talanta 48 (1999) 951.

[9] D. Tuma, B. Wagner, G.M. Schneider, Comparative solubility investigations of anthraquinone disperse dyes in near- and supercritical fluids, Fluid Phase Equilib. 182 (2001) 133.

[10] B. Wagner, C.B. Kautz, G.M. Schneider, Investigations on the solubility of anthraquinone dyes in supercritical carbon dioxide by a flow method, Fluid Phase Equilib. 158–160 (1999) 707.

[11] M. Shamsipur, A.R. Karami, Y. Yamini, H. Sharghi, Solubilities of some 1-hydroxy-9,10-anthraquinone derivatives in supercritical carbon dioxide, J. Supercrit. Fluids 32 (2004) 47.

[12] F.R. Burden, D.A. Winkler, Robust QSAR Models using Bayesian Regularized Neural Networks, J. Med. Chem. 42 (1999) 3183.

[13] F.R. Burden, D.A. Winkler, An Optimal Self-Pruning Neural Network and Nonlinear Descriptor Selection in QSAR, QSAR Comb. Sci. 9999 (2009), NA.

[14] D.A. Winkler, Neural, Networks as robust tools in drug lead discovery and development, Mol. Biotechnol. 27 (2004) 139.

[15] D.A. Winkler, F.R. Burden, Robust QSAR Models from Novel Descriptors and Bayesian Regularized Neural Networks, Mol. Simulat. 24 (2000) 243.

[16] F.R. Burden, D.A. Winkler, A Quantitative structure–activity relationships model for the acute toxicity of substituted benzenes to Tetrahymena pyriformis using Bayesian-Regularized Neural Networks, Chem. Res. Toxicol. 13 (2000) 436.

[17] F.R. Burden, D.A. Winkler, Optimal sparse descriptor selection for QSAR using Bayesian methods, QSAR Comb. Sci. 28 (2009) 645.

[18] T. Shinoda, K. Tamura, Solubilities of C.I. Disperse Red 1 and C.I. Disperse Red 13 in supercritical carbon dioxide, Fluid Phase Equilib. 213 (2003) 115.

[19] J. Fasihi, Y. Yamini, F. Nourmohammadian, N. Bahramifar, Investigations on the solubilities of some disperse azo dyes in supercritical carbon dioxide, Dyes Pigments 63 (2004) 161.

[20] A. Ferri, M. Banchero, L. Manna, S. Sicardi, An experimental technique for measuring high solubilities of dyes in supercritical carbon dioxide, J. Supercrit. Fluids 30 (2004) 41.

[21] K. Tamura, T. Shinoda, Binary and ternary solubilities of disperse dyes and their blend in supercritical carbon dioxide, Fluid Phase Equilib. 219 (2004) 25.

[22] H. Lin, C.C. Ho, M.J. Lee, Solubilities of disperse dyes of blue 79:1, red 82 and modified yellow 119 in supercritical carbon dioxide and nitrous oxide, J. Supercrit. Fluids 32 (2004) 105.

[23] H. Lin, C.Y. Liu, C.H. Cheng, Y.T. Chen, M.J. Lee, Solubilities of disperse dyes of blue 79, red 153, and yellow 119 in supercritical carbon dioxide, J. Supercrit. Fluids 21 (2001) 1.

[24] F.R. Burden, Using artificial neural networks to predict biological activity from simple molecular structural considerations, Quant. Struct.: Act. Relat. 15 (1996) 7.

[25] F.R. Burden, M.J. Polley, D.A. Winkler, Toward novel universal descriptors: charge fingerprints, J. Chem. Inf. Model. 49 (2009) 710.

[26] D.A. Winkler, F.R. Burden, A.J.R. Watkins, Atomistic topological indices applied to benzodiazepines using various regression methods, Quant. Struct.: Act. Relat. 17 (1998) 14.

[27] B. Guzel, A. Akgerman, Solubility of disperse and mordant dyes in supercritical CO_2, J. Chem. Eng. Data 44 (1999) 83.

[28] S.L. Draper, G.A. Montero, B. Smith, K. Beck, Solubility relationships for disperse dyes in supercritical carbon dioxide, Dyes Pigments 45 (2000) 177.

[29] H.-D. Sung, J.-J. Shim, Solubility of C.I. Disperse Red 60 and C.I. Disperse Blue 60 in supercritical carbon dioxide, J. Chem. Eng. Data 44 (1999) 985.

[30] U. Haarhaus, P. Swidersky, G.M. Schneider, High-pressure investigations on the solubility of dispersion dyestuffs in supercritical gases by VIS/NIR-spectroscopy. Part I. 1,4-Bis-(octadecylamino)-9,10-anthraquinone and disperse orange in CO_2 and N_2O up to 180 MPa, J. Supercrit. Fluids 8 (1995) 100.

[31] T. Shinoda, K. Tamura, Solubilities of C.I. Disperse Orange 25 and C.I. Disperse Blue 354 in supercritical carbon dioxide, J. Chem. Eng. Data 48 (2003) 869.

[32] P. Swidersky, D. Tuma, G.M. Schneider, High-pressure investigations on the solubility of anthraquinone dyestuffs in supercritical gases by VIS-spectroscopy. Part II. 1,4-Bis-(n-alkylamino)-9,10-anthraquinones and Disperse Red 11 in CO_2, N_2O, and CHF_3 up to 180 MPa, J. Supercrit. Fluids 9 (1996) 12.

[33] D. Tuma, G.M. Schneider, High-pressure solubility of disperse dyes in near- and supercritical fluids: measurements up to 100 MPa by a static method, J. Supercrit. Fluids 13 (1998) 37.