SOLUBILITY OF ORGANIC COMPOUNDS IN SUPERCRITICAL CARBON DIOXIDE: A QSPR STUDY

S. Verma and R. N. Behera✉

Department of Chemistry, Birla Institute of Technology & Science, Pilani- K. K. Birla Goa
Campus, Zuarinagar - 403726, Goa, India.

✉Corresponding Author: rbehera@goa.bits-pilani.ac.in

ABSTRACT

The quantitative-structure-property-relation (QSPR) model has been used to predict the solubility of a wide range of compounds (with a broader chemical diversity) in supercritical carbon dioxide. The dataset contains 26 molecules (with 324 data points) covering food, textile, pharmaceuticals and petrochemical industries. The concept of Artificial Neural Network with open source software and in-house programs have been used to create the model avoiding the use of commercial and expensive software. The model is found to be robust and performs better compared to many earlier models developed.

Keywords: Supercritical Fluid, SC-CO₂, Structure-property, Neural Network, Machine Learning.

RASAYAN J. Chem., Special Issue, 2021

INTRODUCTION

Supercritical carbon dioxide (SC-CO₂), and in general supercritical solvents, are gaining increasing importance as tunable solvents and reaction media in various industries, such as pharmaceutical, chemical and food industries.¹⁻⁵ SC-CO₂ possess many exciting properties, including good solvation and transport properties. Its low critical temperature (critical parameters 31.1°C and 7.38 MPa) prevents damage of thermally reactive solutes compared to conventional extraction. SC-CO₂ finds its use in extracting compounds (e.g., extracting caffeine from the coffee beans, extracting petrochemical waste etc.), in Pharmaceuticals, as a solvent for organic dyes, etc. Knowledge of the solubility of compounds in SC-CO₂ at various thermodynamic conditions will help in minimizing experimental and operational costs. Thus, there is a requirement for reliable models that can correlate the solubility of organic compounds in supercritical carbon dioxide.⁶

Solubility prediction of compounds in SC-CO₂ started mostly with empirical and semi-empirical models, or using theoretical equations of state (EOS). The EOS-based models, e.g. the cubic equations of state, needs complicated methods and/or specific data (critical properties, sublimation pressure etc.), which may not be available for many compounds.⁷ Most of the semi-empirical models used for predicting solubility perform reasonably well but within an error of 12 to 40 percent.⁸⁻¹³ Moreover, some parameters of these equations themselves must be obtained either from experiments or to be predicted due to lack of data.¹¹ For example, in the Chrastil equation,

$$\log(w_{2,sat}) = \log(w_2^o) + (k - 1)\log\left(\frac{\rho_1}{\rho_1^o}\right) - \frac{\Delta H}{2.303R}\left(\frac{1}{T} - \frac{1}{T^o}\right)$$

one need to have experimental/predicted data for the association number (k), the density (ρ), the total heat (ΔH) to get solubility (w).^{8,11-13}

In recent years, quantitative structure-property relationship (QSPR) models have been used extensively to predict various biological and physicochemical properties.¹⁴⁻¹⁸ In the QSPR model, the desired property is obtained from the chemical structures only, and is not dependent on obtaining input value experimentally. QSPR are the supervised machine learning models where studies are performed on the basis of the correlation that exists between the experimental values (i.e. Property) and the descriptors ($x_1, x_2, \dots, x_n$) which are obtained from the molecular structure of respective compounds,¹⁴

$$\text{Property} = f(x_1, x_2, \dots, x_n) + \text{constant},$$

Where the property is a function (f) of the descriptors $x_1, x_2, \dots, x_n$ used. Depending upon whether the structure-property relationship (f) is linear or non-linear, the QSPR may use linear models (e.g. least-squares etc.) or non-linear models (e.g. polynomial, bilinear, neural network method, etc). The artificial neural networks (ANN) are versatile machine learning methods of modeling almost any nonlinear QSPR relationship.¹⁹ The advantages of ANN include, (i) detection of complex nonlinear relationships between dependent and independent variables, (ii) identification of all possible interactions between input variables, etc. ANN processes information via nodes, to find the complex relationships between inputs and outputs.²⁰ Each connection is associated with a weight value by which signals between the two nodes are controlled. The input signals are combined linearly and a sigmoid function is applied (to hidden layers) to squash the result.

Most of the literature on the QSPR model using ANN (for estimating solubility of the compound in SC-CO₂) predict solubility for one or similar type of compounds. Famini et al. and Engelhardt et al. modeled solubility of aromatic compounds (22 and 58 numbers, respectively) at a particular pressure and temperature.²¹⁻²² Tabaraki et al. compared solubility of 25 anthraquinone and 21 azo dyes in a range of temperatures and pressures using a wavelet neural network and confirmed that a non-linear ANN model performed better than a linear model.²³⁻²⁴ Later, they developed a QSPR for predicting solubility in supercritical CO₂, ethane and trifluoromethane in various conditions (pressure range 5.1–36.2 MPa, temperature range 308–343 K).²⁵ Khayamian and Esteki studied five polycyclic aromatic compounds using a similar model.²⁶ Using a neural network QSPR model for SC-CO₂ solubility, Hemmateenejad et al., studied 29 compounds of anthraquinone, anthrone and xanthone derivatives.²⁷ Tarasova et al. used Bayesian methods to derive models for predicting solubility of organic dyes and polycyclic aromatic compounds in SC-CO₂.²⁸ They have used 67 compounds with 685 data points (in temperature range 285.9–423.2 K and pressure range 6–140 MPa). Later, Valenzuela et al. developed QSPR models (using 2436 data points in the range 20–150°C, 6–49 MPa) to predict the solubility of polyaromatic hydrocarbons' solubility anthroquinones and their derivatives in SC-CO₂.²⁹ The descriptors were obtained from molecular dynamics simulations. They have used 30 compounds for training and testing and predicted the solubility for three of them accurately. Thus, QSPR modeling can be used to predict the solubility of new compounds, at least in some cases, using sufficient literature data.²⁹ However, the QSPR modeling predictions are limited with very similar types of compounds, e.g. organic dyes/polycyclic aromatic compounds or polyaromatic hydrocarbons, anthroquinones and their derivatives etc.^{28,29} There is a need to develop a robust model which can predict solubility for not-so-similar compounds. The present work is an attempt on this line. Here we have used ANN to create a QSPR model for predicting solubility for a range of compounds with a broader chemical diversity covering major applications of SC-CO₂ than those used in earlier QSPR models. The selected compounds found application in different fields like caffeine extraction, textile industries, pharmaceuticals and extraction of petroleum wastes. Another aim of the present study is to use the open-source software (with minor in-house programs) to develop a QSPR model.

EXPERIMENTAL

Dataset Preparation

The first step in QSPR modeling is preparing a dataset containing a set of chemical structures and experimental solubility data in SC-CO₂. We have selected compounds from its application in different fields like caffeine extraction, textile industries, pharmaceuticals and extraction of petroleum wastes. Our dataset contains 26 molecules with a total of 324 data points. The solubility values reported as the logarithm of the mole fraction of solute in SC-CO₂ ($\log(S)$) for various Temperatures and Pressure (80–433 K, 9–49 MPa) are obtained from the literature.²⁸⁻³¹ Optimized 3D structures of the selected compounds were obtained from PubChem (Open Chemistry database) site.³² List of the studied compounds are given in Table-1.

Table-1: List of Compounds Studied in the Current Work

S.No.	Name	S.No.	Name
1	Benzoic acid	14	1,3,7-Trimethyl-2,6-dioxopurine
2	1,2-dihydroxanthraquinone	15	1,4-dihydroxy-9,10-anthraquinone
3	Biphenyl	16	N-(4-Hydroxyphenyl)acetamide
4	Naphthalene	17	1-amino-2-methylanthraquinone

5	2-nitroso-1-naphthol	18	1-hydroxy-9,10-anthraquinone
6	β-Carotene	19	1-hydroxy-2-methyl-9,10-anthraquinone
7	Phenanthrene	20	1,4-bis(methylamino)anthraquinone
8	2,2'-Methylenebiphenyl	21	1,4-bis(butylamino)anthraquinone
9	1,2,3,4-Dibenzonaphthalene	22	4-(N, N-dimethylamino)4'-nitroazobenzene
10	4-amino-4'-nitroazobenzene	23	3-[N-ethyl-4-(4-nitrophenylazo)phenylamino] propionitrile
11	2-Methyl-1,4-naphthoquinone	24	2,2'-(4-((4-aminophenyl)diazanyl)phenylazane diyl)diethanol
12	1-(methylamino)anthraquinone	25	2-(3-hydroxy-2-quinolyl)-1,3-indanedione
13	1,3-Phthalandione	26	4-(naphthalen-1-yl diazenyl)benzene-1,3-diamine

Generation of Molecular Descriptors

We have used the E-Dragon software (free version) to generate the molecular descriptors from the optimized structure of selected compounds.^{33,34} This software provides a total of around 1600+ descriptors from 20 broader categories. We have carefully selected a variety of descriptors (e.g. constitutional block, functional group count, molecular flexibility, H-Bond Donor atoms, Aromatic ratio, etc.) that may impact solubility. A total of 42 descriptors were used in this study to represent each compound for solubility prediction. They are given below:

Numbers of atoms (nAT), non-H atoms (nSK), bonds (nBT), non-H bonds (nBO), multiple bonds (nBM), rotatable bonds (RBN), double bonds (nDB), aromatic bonds (nAB), Hydrogen atoms (nH); Carbon atoms (nC), Nitrogen atoms (nN), Oxygen atoms (nO), rings (cyclomatic number) (nCIC), circuits25 (nCIR), 5-membered rings (nR05), 6-membered rings (nR06), benzene-like rings (nBnz), donor atoms for H-bonds (nHDon), acceptor atoms for H-bonds (nHAcc), intramolecular H-bonds with N, O, F (nHBonds), aromatic C (nCar), aromatic carboxylic acids (nArCOOH), aromatic esters (nArCOOR), aromatic nitro groups (nArNO2).

Sums of atomic van der Waals volumes scaled on carbon atom (Sv), atomic Sanderson electronegativities scaled on carbon atom (Se), atomic polarizabilities scaled on carbon atom (Sp), conventional (H-depleted) bond orders (SCBO).

Pressure (P), Temperature (Temp), Molecular weight (MW), Aromatic ratio (ARR), Mean square (Balaban) distance index (MSD), Radial Distribution Function - 095 / weighted by mass (RDF095m), CH₃R / CH₄ (C-001), CH₂R₂ (C-002), phenol / enol / carboxyl OH (O-057), O=O (O-058), R--N--R / R--N--X (N-075), Hydrophilic factor (Hy), topological polar surface area using N, O polar contributions (TPSA(NO)), Ghose-Crippen molar refractivity (AMR).

We have written a java program that runs through all the selected compounds and extracts the value of the selected descriptor from the output data of Dragon software and appends it with Temperature, Pressure and Solubility data created already.³⁵

The Artificial Neural Network Model

The ANN model was created using Neuroph Studio (Open source software for Neural Network implementation in Java).^{36,37} The neural network created consists an input layer having 42 nodes (for the selected 42 descriptors). It contains two hidden layers having 14 and 8 nodes, respectively, and an output layer containing of 1 node for solubility. We have used a supervised learning training phase, i.e. it involved a teacher who teaches ANN by improving its predictions by a set of experimentally available data. This was mainly done using the backpropagation algorithm, which learns by example. This algorithm adjusted the network's weights so that the desired output can be produced for a particular input after finishing the training.

For training the model, we first need to normalize all the input parameters (the descriptors) so that all the parameter lies in a given range. The input parameters are normalized using the formula,

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

27

Where $x = (x_1, x_2 \dots x_n)$ and y_i is the i th normalized data. A java program was written for normalizing our dataset, which goes through each input descriptor and normalizes it according to the above equation.³⁵ The output solubility values, viz., $\log(S)$, which range from -1.759 to -8.699, was scaled appropriately (divided by -10) to lie in between 0 to 1 because of the requirement of sigmoid function used in ANN. Once the dataset is normalized and scaled, we split it into training, validation and test set. We randomly selected 65% of total data for training, 12% for validation, and the remaining 23% for testing. We trained our model with a learning rate value of 0.2. The training was done to a maximum allowed error of 0.00009. Reducing the error further led to over-fitting, as was observed with the increase in mean-square-error (MSE) of the validation set.³⁵ The total network error generated while training is shown in Fig.-1.

RESULTS AND DISCUSSION

We carried out a training run series with different input values for the training parameters (as discussed in the previous section). The trained model was selected, which gave the least MSE on the validation dataset. First, we have analyzed the performance of the model on the training data. The total root means square error (RMSE) and the square of correlation coefficient (R^2) value for the training dataset obtained are 0.124 and 0.992, respectively. We have compared the predicted solubility of our model for the training dataset with the experimental solubility in Fig.-2(a). We can see that the added trend line equation is very close to $y = x$ (slope value = 0.9946 and intercept = -0.014), which means that the predicted values are almost equal to the experimental values. Also, for our training set, the R^2 value is equal to 0.992, which indicates the excellent performance of the model on training data.

Fig.-1: Plot of the Total Network Error generated during the Training Dataset

While training, we keep checking the performance on a validation dataset not over to fit. When the RMSE reached the minimum for validation data, we stopped training.³⁵ The RMSE on the validation set was 0.276. We then tested the performance of the model on the test dataset. The RMSE and R^2 values obtained for the test dataset were 0.308 and 0.9547, respectively. The predicted solubility of our model for the test dataset with the experimental solubility is shown in Fig.-2(b). In this case, also, we observe that the fitted trend line equation lies very close to the ideal $y = x$, with slope value = 0.9828 and intercept = 0.0121. The R^2 value was found to be 0.9547. These suggest that our model performs exceptionally well on the test dataset also.

A comparison of the performance of our model with some known models is shown in Table-2.^{28,29} In Comparing the R^2 values with the model created by Tarasova et al., our model performs better on both training and test data.²⁸ Their R^2 value on the training data set is 0.83, whereas ours is 0.9927. On test data their R^2 value is 0.82 while ours is 0.9547. Also, their standard error of prediction, which is the RMSE value for test data is 0.45, whereas for us, it is 0.308. For the model developed by Valenzuela et al., the R^2 values are 0.9918 and 0.9507 (compared to ours 0.9927 & 0.9547) for the training dataset and test dataset, respectively.²⁹ The corresponding standard errors for the training dataset and test dataset are 0.03 & 0.49

(compared to ours 0.124 & 0.308), respectively. So both the models perform similarly as far as R^2 is concerned, but our model gave a lower standard error on the test dataset. Although the number of data points considered here is towards the lower side, but considering the diversity of the compound used, the performance of our model is quite significant.

Fig.-2: Comparison of the predicted Solubility with the Experimental Solubility of Compounds for (a) Training Dataset and (b) Test Dataset

Table-2: Comparison of the Performance for Predictions of Solubility in SC-CO₂ for Different Models

Training Dataset		Test Dataset		Model
R^2	RMSE	R^2	RMSE	
0.9927	0.124	0.9547	0.308	Current work
0.9918	0.03	0.9507	0.49	Valenzuela et al. ²⁹
0.83	0.40	0.82	0.45	Tarasova et al. ²⁸

CONCLUSION

We have developed an artificial neural network-based QSPR model for predicting the solubility of a range of organic compounds in SC-CO₂. Compared to previous models, the dataset we created has sufficient versatility covering major applications of SC-CO₂. Moreover, the model has been created using open source software and java programs written by us, in contrast to the similar work done using commercial and expensive software. This model is found to be quite robust and performs well in predicting solubility. We hope that our model will help predict solubility and similarly modeling other physical properties of interest.

REFERENCES

- Y. Marcus, Supercritical Carbon Dioxide, Nova Science Publishers (2019).
- E. Kiran, S. Levelt, M.H. Johanna (Eds.), Supercritical Fluids: Fundamentals for Application, Springer (1994).
- M. Mukhopadhyay, Natural Extracts Using Supercritical Carbon Dioxide, CRC Press (2000).
- M. F. Kemmere, T. Meyer (Edn), Supercritical Carbon Dioxide in Polymer Reaction Engineering, Wiley (2006).
- K. Brun, P. Friedman and R. Dennis (Edn), Fundamentals and Applications of Supercritical Carbon Dioxide (SCO₂) Based Power Cycles, Woodhead Publishing, Elsevier (2017).
- M. Li, J. Zhang, Y. Zou, F. Wang, B. Chen, L. Guan, Y. Wu, *Materials Today Communication*, **25**, 101277 (2020), <https://doi.org/10.1016/j.mtcomm.2020.101277>
- Z. Tang, J. Jin, Z. Zhang and H. Liu, *Industrial & Engineering Chemistry Research*, **51**, 5515 (2012), <https://doi.org/10.1021/ie2016224>
- J. Chrastil, *Journal of Physical Chemistry*, **86**, 3016 (1982), <https://doi.org/10.1021/j100212a041>
- H.D. Sung, J.J. Shim, *Journal of Chemical & Engineering Data*, **44**, 985 (1999), <https://doi.org/10.1021/je990018t>
- A. Jouyban, H.K. Chan, N.R. Foster, *Journal of Supercritical Fluids*, **24**, 19(2002), [https://doi.org/S0896-8446\(02\)00015-3](https://doi.org/S0896-8446(02)00015-3)
- J. M. del Valle, A.G. Reveco-Chilla, L. M. Valenzuela, J. C. de la Fuente, *Journal of Supercritical Fluids*, **157**, 104688 (2020), <https://doi.org/10.1016/j.supflu.2019.104688>

12. X. Zha, S. Han, W. Wang, Z. Jiao, *Journal of Chemical Thermodynamics*, **131**, 104(2019), <https://doi.org/10.1016/j.jct.2018.10.032>
13. A. G. Reveco-Chilla, L. M. Valenzuela, J. M. del Valle, E. J. Maginn, *Fluid Phase Equilibria*, **433**, 112 (2017), <https://doi.org/10.1016/j.fluid.2016.10.017>
14. S.P. Gupta, *QSAR and Molecular Modeling*, Springer (2018).
15. S. P. Gupta, *Chemical Reviews*, **112**, 3171(2012), <https://doi.org/10.1021/cr200097p>
16. S. P. Gupta, *Chemical Reviews*, **115**, 6427(2015), <https://doi.org/10.1021/cr500483r>
17. B. Geetha, G. Swarnalatha and G.V. Subba Reddy, *Rasayan Journal of Chemistry*, **12**, 1063(2019), <https://doi.org/10.31788/RJC.2019.1235165a>
18. A.F. Zahrt, S.V. Athavale, and S.E. Denmark, *Chemical Reviews*, **120**, 1620(2020), <https://doi.org/10.1021/acs.chemrev.9b00425>
19. D. J. Livingstone, *Artificial Neural Networks: Methods and Applications*, Springer-Verlag New York (2008).
20. B.G. Sumpter, C. Getino, and D.W. Noid, *Annual Review of Physical Chemistry*, **45**, 439 (1994), <https://doi.org/10.1146/annurev.pc.45.100194.002255>
21. G.R. Famini, L.Y. Wilson, *Journal of Physical Organic Chemistry*, **6**, 539(1993), <https://doi.org/10.1002/poc.610061002>
22. H.L. Engelhardt, P.C. Jurs, *Journal of Chemical Information and Computer Sciences*, **37**, 478 (1997), <https://doi.org/10.1021/ci960085g>
23. R. Tabaraki, T. Khayamian, A.A. Ensafi, *Journal of Molecular Graphics and Modelling*, **25**, 46 (2006), <https://doi.org/10.1016/j.jmglm.2005.10.012>
24. R. Tabaraki, T. Khayamian, A.A. Ensafi, *Dyes and Pigments*, **73**, 230(2007), <https://doi.org/10.1016/j.dyepig.2005.12.003>
25. R. Tabaraki, A. Toulabi, *Fluid Phase Equilibria*, **383**, 108 (2014), <https://doi.org/10.1016/j.fluid.2014.10.008>
26. T. Khayamian, M. Esteki, *Journal of Supercritical Fluids*, **32**, 73(2004), <https://doi.org/10.1016/j.supflu.2004.02.003>
27. B. Hemmateenejad, M. Shamsipur, R. Miri, M. Elyasi, F. Foroghnia, H. Sharghi, *Analytica Chimica Acta*, **610**, 25(2008), <https://doi.org/10.1016/j.aca.2008.01.011>
28. A. Tarasova, F. Burden, J. Gasteiger, D.A. Winkler, *Journal of Molecular Graphics and Modelling*, **28**, 593(2010), <https://doi.org/10.1016/j.jmglm.2009.12.004>
29. L.M. Valenzuela, A.G. Reveco-Chilla, J.M. Del Valle, *Journal of Supercritical Fluids*, **94**, 113(2014), <https://doi.org/10.1016/j.supflu.2014.06.022>
30. S.-K. Liao and P.-S. Chang, *American Journal of Analytical Chemistry*, **3**, 923(2012), <https://doi.org/10.4236/ajac.2012.312A122>
31. R.B. Gupta and J-J. Shim, *Solubility in Supercritical Carbon Dioxide*, CRC Press (2007).
32. <https://pubchem.ncbi.nlm.nih.gov/> , National Center for Biotechnology Information. PubChem Compound Database.
33. I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, *Journal of Computer-Aided Molecular Design*, **19**, 453 (2005), <https://doi.org/10.1007/s10822-005-8694-y>
34. <http://www.vcclab.org>, Virtual Computational Chemistry Laboratory (VCCLAB) 2005.
35. S. Verma, M. Sc. Thesis, Department of Chemistry, BITS Pilani – K. K. Birla Goa Campus, Goa, India (2017).
36. https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm
37. <https://sourceforge.net/projects/neuroph/files/neuroph-2.9> Z. Sevarac, Neuroph Studio (Version 2.9). [RJC-6613/2021]