MODELLING OF CAROTENOIDS SOLUBILITY IN SUPERCRITICAL CARBON DIOXIDE USING QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIPS

Loreto Valenzuela[1,*]; José Manuel del Valle[1]; Juan de la Fuente[2]

[1] Chemical and Bioprocess Engineering Department
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile

[2] Departamento de Procesos Químicos, Biotecnológicos y Ambientales
Universidad Técnica Federico Santa María, Chile

Email: lvalenzr@ing.puc.cl

Abstract. Prediction of solubility in SuperCritical CO_2 (SC-CO_2) as a function of system pressure (*P*) and temperature (*T*) aids selection of process condition for extraction processes. Several equations have being used to correlate solubility as a function of *P* and *T*, but best-fitting procedures typically demand a large set of experiments. Previously, other groups have developed semi-empirical models to predict solubility of different compounds in SC-CO_2. Our objective is to develop a semi-empirical model to predict the solubility of carotenoids in SC-CO_2 under different pressure and temperature conditions, using a small set of descriptors obtained from their equilibrated 3D structure. Experimental solubility of selected carotenoids in SC-CO_2 at different pressure and temperature was used to build the model, using their solubility parameters according to Chrastil equation. Descriptors were calculated from the solute structures after molecular dynamic simulations in implicit CO_2, from which carotenoids were separated in clusters. Descriptors were ranked, and Quantitative Structure-Property Relationships (QSPRs) were built using a small set of descriptors and a subset of the experimental values for solubility, using both linear regression and Artificial Neural Networks. Further experimental data is required to validate the model and to be able to predict outside of the training set, without the need to run a very large set of experiments.

Keywords: Carotenoids, QSPR, Semiempirical Modeling, Solubility, Supercritical CO_2.

1. Introduction

SuperCritical CO_2 (SC-CO_2) is currently used in extraction processes of high-value compounds in vegetable substrates, due to its convenient liquid-like solvent properties and gas-like transport properties. One of the parameters that control the extraction is the compound solubility on SC-CO_2 at operational pressure (*P*) and temperature (*T*).

Several equations have being used to correlate solubility at different operational conditions. They are best-fitted to experimental measurements of phase equilibrium of the compound of interest and SC-CO_2 at different *P* and *T* values. Depending on the model, best fitting parameters do not necessarily mean something specific about solvation phenomena. Different equations have shown to be the most accurate, depending on the compound of study [1, 2].

In this study, we used the equation of Chrastil [3] because it is popular, and it allows a separation of the contributions of system conditions to the solubility. Indeed, upon re-parametrization [4], Chrastil's equation predicts the solubility of a solute in SC-CO_2 (c_{sat}) as a function of three independent factors, namely the solubility at a reference condition, a correction by absolute temperature (*T*), and a correction by SC-CO_2 density (ρ):

$$\log(c_{sat}) = \log(c^o) + k' \log\left(\frac{\rho}{\rho^o}\right) - \frac{\Delta H}{2.303R}\left(\frac{1}{T} - \frac{1}{T^o}\right) \quad (1)$$

where c^o is the solubility at a reference temperature T^o and reference pressure P^o; ρ^o is the reference density of SC-CO_2 at T^o and P^o; $k' = k-1$, where k is an association number or number of solvent molecules combining with a single solute molecule to form a solvato complex [3]; ΔH is the total heat (heat of vaporization plus heat of dissolution in SC-CO_2) required to synthesize the solvato complex [3]; and R is the universal gas constant.

Even for Chrastil's equation, parameters must be fitted from experimental data, which can lead to long and expensive experiments. As an alternative, Quantitative Structure-Property Relationships (QSPRs) are mathematical functions that allow predicting materials properties reducing the need of extensive experiments after being trained with reliable and comparable experimental data.

Previously, QSPR models have been developed to predict solubility of different set of compounds in SC-CO_2 using all experimental values as single and independent inputs for each model [5-10] (Table 1). Engelhardt and Jurs [5] modeled solubility at one single pressure and temperature, so it is not possible to extrapolate to other operational conditions. The other authors used single experimental points as independent entries to train, test, and validate their models. In this work, we explore building QSPR models of solubility in SC-CO_2 using the parameters of Chrastil's equation, instead of single solubility values as the entries.

Table 1. Sumary of QSPR models for organic compound solubility in SC-CO_2.

Reference	Type of compounds	Number of compounds	Number of data points	Number of descriptors
Engelhardt and Jurs [5]	Organic compounds	58	58	7
Khayamian and Esteki [6]	Polycyclic aromatic hydrocarbons	5	89	6
Tabaraki et al. [7, 8]	Anthraquinone dyes	25	760	8
Hemmateenejad et al. [9]	Anthraquinone, anthrone, and xanthone derivatives	29	1190	20
Tarasova et al. [10]	Organic dyes and polycyclic aromatic compounds	67	685	>30

All of these models use large data sets derived to relatively smaller number of compounds. Also, all of these models used large number of descriptors, which in general include P and T. Our objective is to develop a semi-empirical model to predict the solubility of a group of compounds in SC-CO_2, under different pressure and temperature conditions, using a small set of descriptors obtained from their equilibrated 3D structure. Instead of modeling for all data points separately, we built models for the parameters of the Chrastil equation (k' and c^o) (Equation 1).

2. Materials and Methods

Materials. We modeled the solubility of β-carotene in SC-CO_2 using the data from Sakaki [11], Johanssen & Brunner [12], Subra et al. [13], Mendes et al. [14], Sovova et al. [15], Kraska et al. [16], and Araus et al. [17], who reported values that were consistent. Data of Stahl et al. [18] and Skerget et al. [19] were not included because they consistently over-predicted the solubility of β-carotene as compared to the selected authors. On the other hand, data of Cygnarowics et al. [20], Jay et al [21], and Hansen et al. [22] were not included because their experimental values differed significantly from predictions of the adopted model. We observed that in the case of β-carotene solubility values measured by Stahl et al. [18] differed by a fixed factor from values predicted by our model which we imputed to a systematic error (bias) in their experimental method. We assumed that this bias affected also solubilities measured by the same authors for apocarotenal, apocarotenate ethyl ester, cantaxanthin, and zeaxanthin, and made corrections prior to their correlation using Chrastil's equation. Stalh et al. [18] measured solubilities at a single temperature, and because of this we could not study the effect of temperature on solubility (term $\Delta H/R$ in Equation 1). Table 2 summarizes the experimental values extracted from the literature, and used in this study.

Table 2. Chrastil parameters from (Equation 1) for carotenoids used in this study

Compound	k'	$\Delta H/R$ (K)[a]	c^o (μg/kg)	Reference
Apocarotenal	9.369	-	69.212	[18][b]
Apocarotenate ethyl ester	9.076	-	98.125	[18][b]
Astaxanthin	2.715	5260	1.650	[23]
β-carotene	5.617	4177	2.326	[11, 12, 14, 16, 17, 24]
Cantaxanthin	6.632	-	0.270	[18][b]
Capsanthin	4.915	2483	12.734	[25]
Lutein	4.241	2306	17.200	[26]
Lycopene	5.617	3829	4.325	[23]
Zeaxanthin	5.693	-	0.130	[18][b]

[a] No available data for all compounds.
[b] data was adjusted for being comparable with each other.

Computational methods. Descriptors were obtained from 3D structures of the compounds, using implicit CO_2 as solvent. Only four descriptors were selected for each parameter of Chrastil's equation (Equation 1). A nonlinear model, using ANNs was used to correlate these descriptors and Chrastil parameters for the 9 compounds of the study. Further steps for this research include external validation and prediction of solubility for other compounds, which could be done having a larger training set. A summary of these methods is shown in Figure 1.

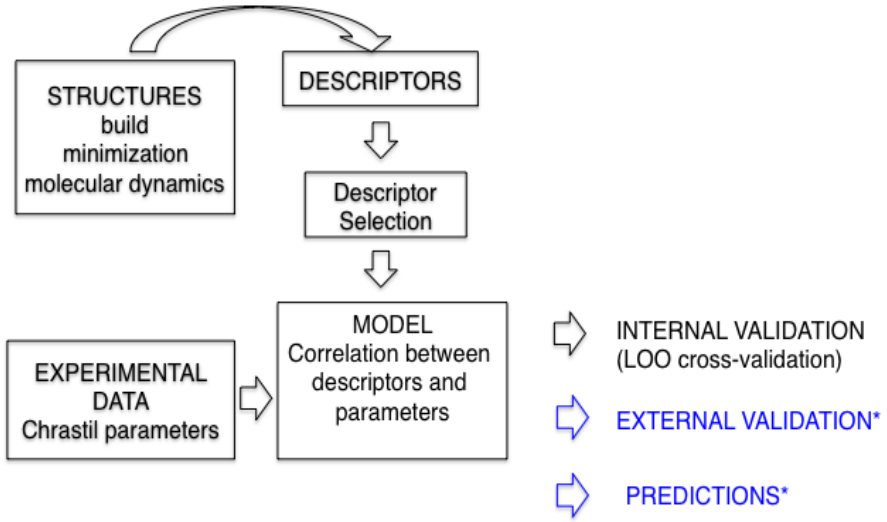

Figure 1. Scheme of the methodology for the semiempirical modeling of this research.
*External validation and predictions are not included in this work.

Two-dimensional descriptors for the set of compounds were obtained using the basic molecular structure derived from the chemical formulae. ChemBio 3D (CambridgeSoft, Cambridge, MA) was used to represent the molecular structure of each compound. Energy was minimized for 10 fs by Molecular Dynamics Simulations, with a time step of 2 fs, and using the Molecular Modeling Force Field (MMFF94) [27]. The Brooks-Beeman algorithm [28, 29] for integrating the equations of motion was used to compute new positions and velocities of each atom at each step. Minimum RMS gradient of 0.1 kcal/mol was used to determine model convergence. Minimization was performed using implicit solvent, with a dielectric constant of 1.57 (SC-CO_2 at the reference conditions of T^o=313 K, P^o=30 MPa, and ρ^o=909.9 kg/m^3) [30].

Dragon (Milano Chemometrics and QSAR Research Group) [31] was used to obtain the molecular descriptors from each structure. For each compound, 3224 descriptors were calculated, which include 0D (constitutional descriptors), 1D (e.g., functional group counts, charge descriptors, molecular properties), 2D (topological descriptors, walk and path counts, connectivity indices), and 3D (e.g., Randic molecular profiles, geometrical, RDF, WHIM, GETAWY and 3D-MoRSE descriptors) descriptors.

The data-mining package WEKA (Waikato Environment for Knowledge Analysis) [32] was used in this study. From the normalized descriptors obtained by Dragon, descriptors that did not vary among the

compounds of interest, had a variance over 99%, are duplicated from other descriptors, were eliminated. For each parameters, the less correlated descriptors were eliminated by Principal Component Analysis (PCA).

For each model, expectation-maximization (EM) [32] cluster analysis was employed to categorize the solubility parameter of interest. The most significant descriptors were selected using a J48 Decision Tree [33] to select descriptors that correctly partition each solubility parameter according to the EM cluster analysis, and linear regression to select those descriptors with higher linear or inverse correlation with the parameter of interest.

A multilayer perceptron (MLP) was used to build ANN models for each parameter with the selected descriptors. No hidden layers (nodes) were used. Output nodes were unthresholded linear units [32]. Backpropagation by gradient descent was used as MLP learning method. All input variables were scaled to the unit interval while the learning rate and the momentum applied for updating the weights were 0.3 and 0.2, respectively. Training time was set on 1,000 epochs, which showed to be enough for model convergence. Due to the reduced number of compounds to train the model, leave one out (LOO) crossvalidation was performed, in all possible combinations. Model accuracy was evaluated by the correlation coefficient (R^2).

3. Results and discussion

Using EM analysis, values of Chrastil parameters, k' and c^o, were grouped in two and three clusters, respectively (Table 3). For apocarotenal and apocarotenal ethyl ester, both parameters were in the "high" level, implying that the effect of density and the reference were the highest of the set of compounds on study. For capsanthin and lutein, k' was in the "low" level, while c^o was in the "medium" one. The remaining 5 compounds present both parameters in the "low" level.

Table 3. Clusters of Chrastil parameters for selected carotenoids.

Compound	k'	Clusters for k'	c^o (μg/kg)	Clusters for $c^o \cdot 10^3$
Apocarotenal	9.369	high	69.212	high
Apocarotenate ethyl ester	9.076	high	98.125	high
Astaxanthin	2.715	low	1.650	low
β-carotene	5.617	low	2.326	low
Cantaxanthin	6.632	low	0.270	low
Capsanthin	4.915	low	12.734	medium
Lutein	4.241	low	17.200	medium
Lycopene	5.617	low	4.325	low
Zeaxanthin	5.693	low	0.130	low

In the selection of the most correlated descriptors for each parameter, 37 descriptors were selected for k', and 348 for c^o. From those descriptors, 4 were selected to build the QSPR models of these parameters, from the highest direct and inverse correlated from linear regression analysis, and decision tree analysis (Figures 2 and 3).

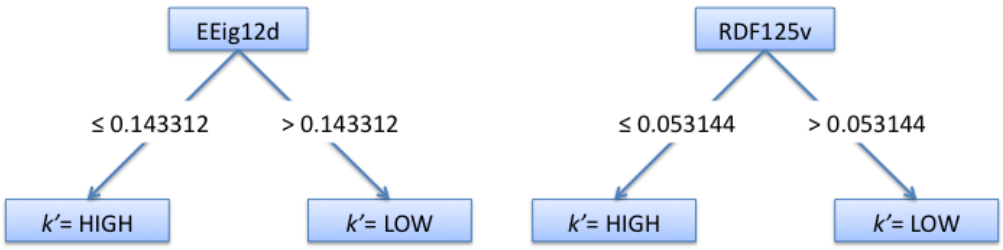

Figure 2. Decision trees for k'. Both trees fully classify the levels of the nine compounds for this parameter.

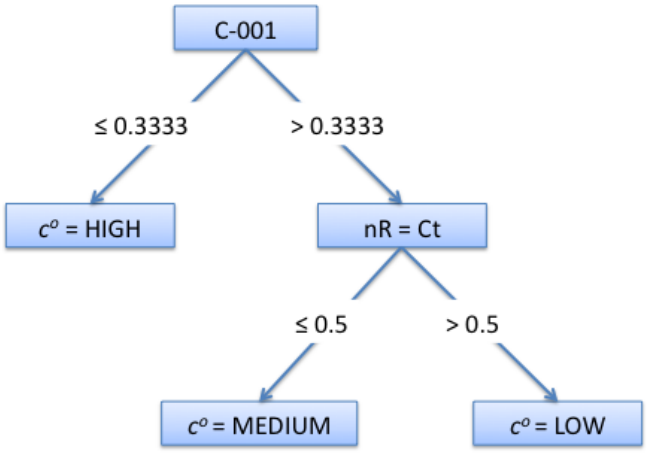

Figure 3. Decision tree for c^o. It fully classifies the levels of the nine compounds for this parameter.

As shown in Table 4, selected descriptors vary from 1D to 3D. One-dimension descriptors are C-001 (atom centered descriptor on CH_3R/CH_4) and nR=Ct (number of aliphatic tertiary C(sp2)), and they account for the difference of chemical structure of the studied compounds. Two-dimension descriptors include edge adjacency indices and frequency fingerprints, which describe topological features of molecules. Three-dimension descriptors include radial distribution functions (RDF), R-GETAWAY, and 3D-MoRSE descriptors. RDF descriptors are built from the radial distribution function of an ensemble of atoms, which provides information about interatomic distances within the molecule and other information such as bond distances, ring types, planar and non-planar systems, and atom types [34]. R-GETAWAY descriptors account for the local aspects of the molecule such as branching, cyclicity, and conformational changes [35]. 3D-MoRSE descriptors provide structural information of the molecules in the space [36], and it has been suggested that this information is related with the free volume of molecules [37, 38].

Table 4. Descriptors and correlation coefficient of models for each parameter.

Parameter	Descriptors	Type of descriptor	Correlation coefficient (R^2) with LOO crossvalidation
k'	ESpm06d	edge adjacency index (2D)	0.9350
	RDF125v	RDF descriptor (3D)	
	R7m	R-GETAWAY descriptor (3D)	
	EEig12d	edge adjacency index (2D)	
c^o (μg/kg)	C-001	atom-centered fragment (1D)	0.7323
	nR=Ct	functional group count (1D)	
	F02[O-O]	frequency fingerprint (2D)	
	Mor13e	3D-MoRSE descriptor (3D)	

With the selected four descriptors and the values derived from the experimental measurements for the nine compounds, an ANN model was built for each parameter of the Equation 1. Using LOO cross-validation, the correlation coefficients for the models of k' and $c^o \cdot 10^3$ were 0.935 and 0.732, respectively (Table 4 and Figure 4). Predictions of the effect of density on the solubility k' were very accurate, implying that there is a strong correlation between this parameter and the molecular structure of the compounds, in terms of topology and geometry. The solubility of reference, c^o, was less accurately predicted using similar descriptors, showing that this property is correlated with molecular structure, but there are other factors that may be affecting it at the same time.

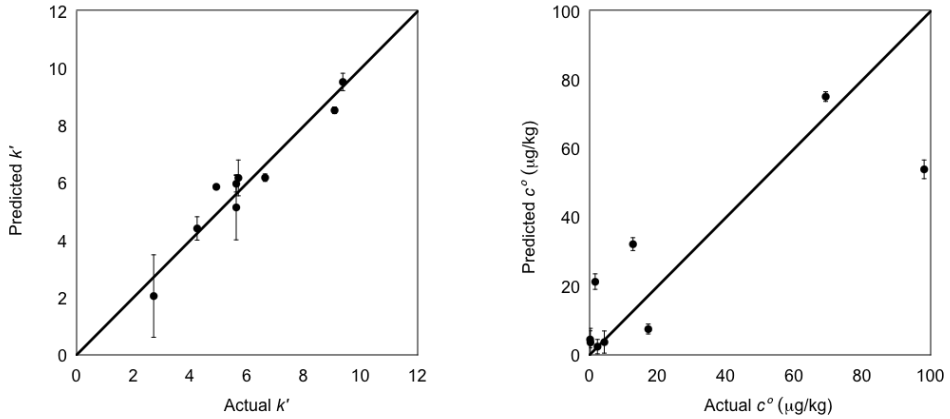

Figure 4. Prediction versus experimental values for k' and c^o, respectively. Black line represents x=y. Values are presented as mean value ± SD of predictions.

These two parameters are useful to partially understand the differences in solubility among a set of compounds. However, reliable and comparable experimental data is needed in order to built a model for the third parameter of the Chrastil equation, ΔH, which accounts for the effect of temperature on the solubility.

4. Conclusions

Models for two parameters of Chrastil's equation for solubility on SC-CO_2 were obtained for nine carotenoids using experimental data from the literature and four molecular descriptors calculated from relaxed 3D structures. Although having a small set of data, accurate predictions of the effect of density on the solubility were obtained. Improvements on these models will need more reliable experimental data, which has to be comparable among different authors.

Acknowledgements

This work was funded by Chilean agency Fondecyt (Regular Project 111-1008). The help of Andrea Reveco in obtaining molecular descriptors was very much appreciated.

References

[1] Tang, Z.; Jin, J.-s.; Zhang, Z.-t.; Liu, H.-t., New experimental data and modeling of the solubility of compounds in supercritical carbon dioxide. Industrial & Engineering Chemistry Research 51 (2012) 5515-5526.
[2] Chang, F.; Jin, J.; Zhang, N.; Wang, G.; Yang, H.-J., The effect of the end group, molecular weight and size on the solubility of compounds in supercritical carbon dioxide. Fluid Phase Equilibria 317 (2012) 36-42.
[3] Chrastil, J., Solubility of solids and liquids in supercritical gases. The Journal of Physical Chemistry 86 (1982) 3016-3021.
[4] del Valle, J. M.; de la Fuente, J. C.; Uquiche, E., A refined equation for predicting the solubility of vegetable oils in high-pressure CO2. The Journal of Supercritical Fluids 67 (2012) 60-70.
[5] Engelhardt, H. L.; Jurs, P. C., Prediction of supercritical carbon dioxide solubility of organic compounds from molecular structure. Journal of Chemical Information and Computer Sciences 37 (1997) 478-484.
[6] Khayamian, T.; Esteki, M., Prediction of solubility for polycyclic aromatic hydrocarbons in supercritical carbon dioxide using wavelet neural networks in quantitative structure property relationship. Journal of Supercritical Fluids 32 (2004) 73-78.
[7] Tabaraki, R.; Khayamian, T.; Ensafi, A. A., Solubility prediction of 21 azo dyes in supercritical carbon dioxide using wavelet neural network. Dyes and Pigments 73 (2007) 230-238.
[8] Tabaraki, R.; Khayamian, T.; Ensafi, A. A., Wavelet neural network modeling in QSPR for prediction of solubility of 25 anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide. Journal of Molecular Graphics & Modelling 25 (2006) 46-54.

[9] Hemmateenejad, B.; Shamsipur, M.; Miri, R.; Elyasi, M.; Foroghinia, F.; Sharghi, H., Linear and nonlinear quantitative structure-property relationship models for solubility of some anthraquinone, anthrone and xanthone derivatives in supercritical carbon dioxide. Analytica Chimica Acta 610 (2008) 25-34.
[10] Tarasova, A.; Burden, F.; Gasteiger, J.; Winkler, D. A., Robust modelling of solubility in supercritical carbon dioxide using Bayesian methods. Journal of Molecular Graphics & Modelling 28 (2010) 593-597.
[11] Sakaki, K., Solubility of Beta-Carotene in Dense Carbon-Dioxide and Nitrous-Oxide from 308 to 323-K and from 9.6 to 30 Mpa. Journal of Chemical and Engineering Data 37 (1992) 249-251.
[12] Johannsen, M.; Brunner, G., Solubilities of the fat-soluble vitamins A, D, E, and K in supercritical carbon dioxide. Journal of Chemical and Engineering Data 42 (1997) 106-111.
[13] Subra, P.; Castellani, S.; Ksibi, H.; Garrabos, Y., Contribution to the determination of the solubility of beta-carotene in supercritical carbon dioxide and nitrous oxide: Experimental data and modeling. Fluid Phase Equilibria 131 (1997) 269-286.
[14] Mendes, R. L.; Nobre, B. P.; Coelho, J. P.; Palavra, A. F., Solubility of beta-carotene in supercritical carbon dioxide and ethane. Journal of Supercritical Fluids 16 (1999) 99-106.
[15] Sovova, H.; Stateva, R. P.; Galushko, A. A., Solubility of beta-carotene in supercritical CO2 and the effect of entrainers. Journal of Supercritical Fluids 21 (2001) 195-203.
[16] Kraska, T.; Leonhard, K. O.; Tuma, D.; Schneider, G. M., Correlation of the solubility of low-volatile organic compounds in near- and supercritical fluids. Part 1: applications to adamantane and beta-carotene. Journal of Supercritical Fluids 23 (2002) 209-224.
[17] Araus, K. A.; Canales, R. I.; del Valle, J. M.; de la Fuente, J. C., Solubility of beta-carotene in ethanol- and triolein-modified CO(2). Journal of Chemical Thermodynamics 43 (2011) 1991-2001.
[18] Stahl, E.; Quirin, K. W.; Gerard, D., Dense gases for extraction and refining. M.R.F. Ashworth: New York, NY, 1988; p Medium: X; Size: Pages: 249.
[19] Skerget, M.; Knez, Z.; Habulin, M., Solubility of Beta-Carotene and Oleic-Acid in Dense Co2 and Data Correlation by a Density Based Model. Fluid Phase Equilibria 109 (1995) 131-138.
[20] Cygnarowicz, M. L.; Maxwell, R. J.; Seider, W. D., Equilibrium Solubilities of Beta-Carotene in Supercritical Carbon-Dioxide. Fluid Phase Equilibria 59 (1990) 57-71.
[21] Jay, A. J.; Steytler, D. C.; Knights, M., Spectrophotometric studies of food colors in near-critical carbon dioxide. The Journal of Supercritical Fluids 4 (1991) 131-141.
[22] Hansen, B. N.; Harvey, A. H.; Coelho, J. A. P.; Palavra, A. M. F.; Bruno, T. J., Solubility of capsaicin and ss-carotene in supercritical carbon dioxide and in halocarbons. Journal of Chemical and Engineering Data 46 (2001) 1054-1058.
[23] de la Fuente, J. C.; Oyarzun, B.; Quezada, N.; del Valle, J. M., Solubility of carotenoid pigments (lycopene and astaxanthin) in supercritical carbon dioxide. Fluid Phase Equilibria 247 (2006) 90-95.
[24] Sovova, H.; Stateva, R. P.; Galushko, A. A., Essential oils from seeds: solubility of limonene in supercritical CO2 and how it is affected by fatty oil. Journal of Supercritical Fluids 20 (2001) 113-129.
[25] Araus, K. A.; del Valle, J. M.; Robert, P. S.; de la Fuente, J. C., Effect of triolein addition on the solubility of capsanthin in supercritical carbon dioxide. The Journal of Chemical Thermodynamics 51 (2012) 190-194.
[26] del Valle, J. M.; Araus, K. A.; de la Fuente, J. C.; Robert, P. S., Effect of ethanol addition on the solubility of lutein in supercritical carbon dioxide. Journal of Chemical Thermodynamics (submitted).
[27] Halgren, T. A., Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry 17 (1996) 490-519.
[28] Levitt, M.; Hirshberg, M.; Sharon, R.; Daggett, V., Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Computer Physics Communications 91 (1995) 215-231.
[29] Beck, D. A. C.; Daggett, V., Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods 34 (2004) 112-120.
[30] Hourri, A.; St-Arnaud, J. M.; Bose, T. K., Solubility of solids in supercritical fluids from the measurements of the dielectric constant: Application to CO(2)-naphthalene. Review of Scientific Instruments 69 (1998) 2732-2737.
[31] Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Dragon Web version*, v.3.0; Milano, Italy, 2003.
[32] Witten, I. H.; Frank, E., Data Mining: Practical machine learning tools and techniques with JAVA implementations. 1st ed.; Academic Press: San Diego, 2000; p 371.
[33] Quinlan, J. R., C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.: 1993; p 302.
[34] Hemmer, M. C.; Steinhauer, V.; Gasteiger, J., Deriving the 3D structure of organic molecules from their infrared spectra. Vibrational Spectroscopy 19 (1999) 151-164.
[35] Consonni, V.; Todeschini, R.; Pavan, M., Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. Journal of Chemical Information and Computer Sciences 42 (2002) 682-692.
[36] Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V., Finding the 3D structure of a molecule in its IR spectrum. Fresenius Journal of Analytical Chemistry 359 (1997) 50-55.
[37] Liu, W., Prediction of glass transition temperatures of aromatic heterocyclic polyimides using an ANN model. Polymer Engineering & Science 50 (2010) 1547-1557.

[38] Mattioni, B. E.; Jurs, P. C., Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. Journal of Chemical Information and Computer Sciences 42 (2002) 232-240.