Modeling solubility in supercritical carbon dioxide using quantitative structure–property relationships

Loreto M. Valenzuela *, Andrea G. Reveco-Chilla, José M. del Valle

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

A R T I C L E   I N F O

A B S T R A C T

Prediction of solubility in supercritical CO_2 (SC-CO_2) as a function of system pressure and temperature aids selection of process condition for extraction processes. Previously, other groups have developed semi-empirical models to predict solubility of different compounds in SC-CO_2. We built quantitative structure–property relationships (QSPRs) to predict the parameters of the solubility equation of Chrastil using a small set of descriptors obtained from their equilibrated 3D structure after molecular dynamic simulations in implicit CO_2. Using 30 compounds for training and testing the model, leave-one-out strategy was used to select the descriptors, and leave-10%-out was used for crossvalidation. These models, that give very good correlations ($R^2 > 0.91$ for test set), were tested against five new compounds for validation giving accurate predictions of the solubility as function of pressure and temperature, for three of them. In an alternative approach, accurate predictions of solubility as function of pressure and temperature were obtained for the five compounds if the experimental value of the solubility at a reference temperature was known and only the effect of pressure and temperature was obtained through the model.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Supercritical CO_2 (SC-CO_2) is currently used in extraction processes at industrial scale of high-value compounds in vegetable substrates, due to its convenient liquid-like solvent properties and gas-like transport properties. The solubility of the compound on SC-CO_2 at operational pressure (P) and temperature (T) affects extraction.

Several equations have been 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. The equation that fits the data the best depends on the studied compound [1,2].

In this study, we used the equation of Chrastil [3] because it is popular, easy to use, 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 ($w_{2,sat}$, mg kg^{-1} solute/CO_2 or ppm) as a function of three independent factors, namely the solubility at a reference condition ($w_2°$, ppm), a correction by SC-CO_2 density (ρ_1, kg m^{-3}), and a correction by absolute temperature (T, K):

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

where $w_2°$ is the solubility at a reference pressure $P°$ (MPa) and a reference temperature $T°$; $\rho_1°$ is the reference density of SC-CO_2 at $T°$ and $P°$; $(k-1)$, where k is an association number or number of solvent molecules combining with a single solute molecule to form a solvato complex [3]; ΔH (kJ mol^{-1}) 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 (0.008314 kJ mol^{-1} K^{-1}).

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 material's 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 sets of compounds in SC-CO_2 as reviewed by Le et al. [5]. Most of these works predict solubility for individual

* Corresponding author. Tel.: +56 2 23544927; fax: +56 2 23545803.
  E-mail address: lvalenzr@ing.puc.cl (L.M. Valenzuela).

Table 1
Summary 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
Famini and Wilson [6]	Aromatic compounds	22	22	4
Engelhardt and Jurs [7]	Organic compounds	58	58	7
Khayamian and Esteki [8]	Polycyclic aromatic hydrocarbons	5	89	6
Tabaraki et al. [9,10]	Anthraquinone dyes	25	760	8
Hemmateenejad et al. [11]	Anthraquinone, anthrone, and xanthone derivatives	29	1190	20
Tarasova et al. [12]	Organic dyes and polycyclic aromatic compounds	67	685	>30
Li et al. [13]	n-alkanes	7	7	1

experimental values as single and independent inputs [6–12] (Table 1). Famini and Wilson [6] and Engelhardt and Jurs [7] modeled solubility at one single pressure and temperature, so it is not possible to extrapolate to other operational conditions. Most authors [8–12] used a large ratio of data points to compounds because they used single experimental points as independent entries to train, test, and validate their models. Furthermore, they used P and T as descriptors to build their models. Among them, Tarasova et al. [12] used the most comprehensive set to date, which includes 67 compounds and 685 experimental data points, at different temperatures and pressures. Li et al. [13] correlated a parameter of the Carnahan–Starling–van der Waals equation with the carbon number of seven n-alkanes. 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, for 30 compounds and 2436 experimental data points, corresponding to 167 solubility isotherms.

Our objective is to develop a semi-empirical model to predict the solubility of a group of compounds in SC-CO_2, under different P and T 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 (w_2°, $(k-1)$, and ΔH) (Eq. (1)).

2. Materials and methods

2.1. Materials

Thirty compounds were used in this study as shown in Table 2, classified as polyaromatic hydrocarbons (PAHs) and derivatives, and anthroquinones and derivatives. The model was validated for two PAHs, two anthrones, and one xanthone. All of them have 50 atoms or less, due to restrictions on the computational methods [14].

Solubility data was obtained from published literature (Table 2) and re-computed as a weight fraction in a solute-free basis (w_2, mg kg^{-1} solute/CO_2). Typically, solubility values are reported as molar fractions (y_2) of solute (component 2) in a saturated solution of CO_2 (component 1) as a function of system temperature and pressure using Eq. (2):

$$w_2 = \frac{y_2}{1-y_2}\frac{M_2}{44.01} \qquad (2)$$

where M_2 is the molar weight of the solute and $M_1 = 44.01$ is the molar weight of CO_2. We estimated system density under the assumption that the density of pure CO_2 was not affected by the dissolved solute [3] which is possibly valid in low-solubility systems such as the ones analyzed in this work. This density (ρ_1, kg m^{-3}) was estimated as a function of system pressure and temperature using NIST database [55]. High-solubility data ($0.001 w_2 \rho_1 > 150$ g m^{-3} solute/CO_2) was discarded because the assumption that the density of the saturated solution equals that of pure CO_2 under equivalent system temperature and pressure fails when the solubility is above 100–200 g m^{-3} [3]. The model of Chrastil predicts that

solubility increases monotonously with system pressure for a solubility isotherm. Inconsistent low-solubility data were identified in a few data sets at low pressure when solubility decreased as pressure increased at constant temperature. In these cases all values below the pressure where a minimal was identified, were also disregarded. Most data sets had no discarded data (Table 2, Supplementary content), but in the case of biphenyl, percent of discarded values reached a top value of 25.6% for measurements of McHugh and Paulaitis [15]. Discarded data are signaled in plots of measured versus predicted values in the figures in the Supplementary content, and percent discarded data is reported in Table S3.

2.2. Fitting of Chrastil's equation

Solubility data from different sources were correlated using the model of Chrastil, Eq. (1). As the reference conditions we used $P^\circ = 14$ MPa and $T^\circ = 308$ K (35 °C), conditions for which the density of pure CO_2 is $\rho_1^\circ = 802.5$ kg m^{-3}. This is the same condition for the models of Famini and Wilson [6], and of Engelhardt and Jurs [7], and it is a condition that is included in most studies in Table 2. The mean error in log(w_2) of each solubility isotherm was computed. Best-fit parameters were selected by minimizing the average mean error of all isotherms of all sources. For each literature source two indicators of the goodness-of-fit were computed: (1) the average mean error of all solubility isotherms tested was required to be below a solute-dependent threshold; and, (2) an average bias (the average of individual bias values) to identify sources that systematically over-predicted (average bias = +1.00) or under-predicted (average bias = −1.00) the solubility of the solute as compared to the other sources. The bias of a single observation is +1.00 if the solubility predicted by the best-fit model for the system conditions is higher than experimentally measured at those conditions, −1.00 if the solubility predicted by the best-fit model is lower than experimentally measured, and 0.00 if the two values coincide [4]. Data sources that failed either criterion were discarded and model parameters were best-fitted again to the valid data sets still remaining. Because the best-fit model depends on the data sets that are considered in the regression (set of valid data sets), the two criteria were applied to the remaining (valid) and discarded (invalid) data sets in each step. Thus, following every step, a reshuffling of the data sets could be required for data sets used in model data fitting that failed one of the indicators of goodness-of-fit, and discarded for data sets that fulfilled both. This procedure was repeated as many times as it was required to properly classify the data sets (data of single author(s)) as valid (considered in data correlations) or invalid (disregarded from data correlation). The mean error of log(w_2) of a valid data set was below a solute-dependent threshold, and its average bias was between −1.00 and +1.00 (Table S1 in Supplementary content). In the case on an invalid data set, on the other hand, the mean error of log(w_2) was above the solute-dependent threshold, or the average bias was either −1.00 or +1.00, or both (Table S3 in Supplementary content).

One possible explanation for a data set being invalid is an experimental method with a systematic error. Because of that in many