Journal of CO_2 Utilization 26 (2018) 262–270

Contents lists available at ScienceDirect

Journal of CO_2 Utilization

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

A new chemical structure-based model to estimate solid compound solubility in supercritical CO_2

Alireza Baghban[a,**], Jafar Sasanipour[b], Zhien Zhang[c,d,*]

[a] Chemical Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Mahshahr Campus, Mahshahr, Iran
[b] Gas Engineering Department, Petroleum University of Technology, Ahwaz, Iran
[c] School of Chemistry and Chemical Engineering, Chongqing University of Technology, Chongqing 400054, China
[d] Key Laboratory of Low-grade Energy Utilization Technologies and Systems, Ministry of Education of China, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Keywords:
Drug
Solubility
Chemical structure
Supercritical CO_2
Least square support vector machine

ABSTRACT

Utilization of new approaches in the determination of drug solubility in supercritical fluids can reduce the computation time and represent reliable results. This also leads to more applications of the supercritical technology in the field of drug manufacturing. A least-square support vector machine (LSSVM) approach is employed in this study in order to predict 33 different drug solubility in supercritical CO_2. The solubility of the drugs is estimated as a function of temperature, pressure, supercritical CO_2 density, and 20 different chemical substructures. LSSVM results are then compared to those obtained from 8 previously reported semi-empirical correlations. Satisfying predictions are performed by the proposed LSSVM with an average absolute relative deviation of 4.92% and determination coefficient of 0.998 for the testing dataset. Therefore, the proposed LSSVM can be applied as a reliable predictive tool to estimate the drugs' solubility, if drugs' chemical structures are given.

1. Introduction

Utilization of supercritical fluids as alternatives for organic solvents are considered due to adjustable properties of the supercritical fluids with pressure and temperature [1]. These fluids are applied in different fields such as separation, processes, reactions, purification, and particle sizing of pharmaceuticals [2]. Increasing energy demand in recent decades resulted in further consumption of fossil fuels and increasing carbon dioxide (CO_2) emissions from energy producing sources (such as power plants). Carbon dioxide capture and storage (CCS) is concerned with separation, transportation, and long-term isolation of the emitted CO_2 using various technologies (i.e. post-combustion capture, integrated gasification combined cycle, and oxyfuel). Captured CO_2 can be theoretically stored in oceans, and geological sub-surfaces regardless of the captured amount and also can be used for mineral carbonation or industrial uses. However, in reality, coupling of the captured CO_2 from large scale emitters and geological storage is most likely to be employed as commercial implementation of the CCS. Capture technologies imply different strategies to increase the molar concentration of the flue gas from combustion so that the compression to supercritical state of the captured CO_2 would be viable for geological storage. In geological CO_2 storage, the captured CO_2 will be stored in a storage with desired rates of injectivity, storage capacity, and containment security. Injection of the supercritical CO_2 to oil reservoirs is known as a common enhanced oil recovery (EOR) method [3]. The supercritical CO_2 has also attracted many attentions in pharmaceutical applications due to its non-toxic, environmentally safe, inflammable, and economical characteristics. Furthermore, possession of accurate predictions for solubility of the solid solute in a supercritical fluid plays an important role in developing any supercritical fluid technology. Investigation of pharmaceutical compounds' solubility under different temperature and pressure conditions requires determination of the thermophysical properties through reliable predictive models [4–6]. Reliable experimental data on drugs' solubility in the supercritical region is required to prevent disturbances to the equilibrium [5]. Therefore, the necessity of a reliable predictive tool is clarified in order to estimate the solubility of the drugs in supercritical CO_2 used for designation of the corresponding processes.

Different methods, based on equations of state or density are applied in order to develop correlations based on experimental data [2,7–13]. Semi-empirical correlations are often applied due to their ease of application and nonrequirement of employing directly undetectable physiochemical properties (such as critical properties and sublimation pressure) which require application of different methods of

* Corresponding author at: School of Chemistry and Chemical Engineering, Chongqing University of Technology, Chongqing 400054, China.
** Corresponding author.
E-mail addresses: Alireza_baghban@alumni.ut.ac.ir (A. Baghban), zhienzhang@hotmail.com, zhienzhang@cqut.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.jcou.2018.05.009
Received 5 April 2018; Received in revised form 3 May 2018; Accepted 8 May 2018

Table 1
Ranges of experimental data on drug solubility in supercritical CO_2.

System no.	Solute	Temperature (K)	Pressure (bar)	No. of data points	References
1	Rosuvastatin	308–348	121.6–354.6	45	[20]
2	Simvastatin	308–348	121.6–354.6	45	[20]
3	Atorvastatin	308–348	121.6–354.6	45	[20]
4	Lovastatin	308–348	121.6–354.6	45	[20]
5	Fluvastatin	308–348	121.6–354.6	45	[20]
6	Anastrazole	308–348	122–355	45	[21]
7	Letrozole	308–348	122–355	45	[21]
8	Exemestane	308–348	122–355	45	[21]
9	Atropine	308–348	122–355	45	[22]
10	Diazepam	308–348	122–355	45	[22]
11	Codeine	308–348	122-355	45	[22]
12	Carbamazepine	308–348	243–355	38	[22]
13	Benzocaine	308–348	122–355	40	[23]
14	Naproxen	308–348	122–355	40	[23]
15	Metronidazole benzoate	308–348	122–355	40	[23]
16	Methylparaben	308–348	122–355	40	[24]
17	Bisacodyl	308–348	122–355	39	[24]
18	Methimazole	308–348	122–355	39	[24]
19	Budesonide	338–358	213–385	21	[25]
20	Lamotrigine	318–348	121.6–354.6	36	[26]
21	Clozapine	318–348	121.6–354.6	27	[26]
22	Zopiclone	313–333	100–250	21	[27]
23	Nimodipine	313–333	100–250	21	[27]
24	alpha-Tocopherol	313–353	199–349	24	[28]
25	δ-Tocopherol	313–353	199–349	24	[28]
26	Retinol	313–353	200–350	20	[28]
27	beta-Carotene	313–353	200–320	23	[28]
28	Vitamin D3	313–353	200–350	23	[28]
29	Vitamin D2	313–353	200–280	19	[28]
30	Vitamin K1	313–353	200–350	24	[28]
31	Cefixime trihydrate	308–328	183–335	18	[29]
32	DA	308–348	122–355	45	[30]
33	CP	308–348	122–355	44	[30]

calculation [6].

Artificial intelligence algorithms have found many applications in different scientific fields such as oil and gas [14,15]. The solubilities of different gases such as CO_2, H_2S, and NH_3 in amine solutions and ionic liquids are investigated by Baghban et al. [15–19].

This study investigates the applicability of the least square support vector machine approach in predicting 33 different drugs' solubility in supercritical CO_2. Model development is based on experimental data reported in previous papers [20–30]. Hyperparameters of the proposed model are determined using particle swarm optimization (PSO) method coupled with the LSSVM. This proposed model estimates the target variable (drug's solubility) as a function of five independent variables (i.e. temperature, pressure, supercritical CO_2 density, and 20 different substructures of drug). Comparing the results from the proposed model with 8 semi-empirical correlations [2,7–13] reveals the better performance of the LSSVM. To the best of our knowledge, no records are available on the application of LSSVM to predict the drugs' solubility in supercritical CO_2, so far.

2. Theory

2.1. Least square support vector machine (LSSVM)

Support vector machine (SVM) is an intelligent approach based on the concepts of statistical learning theory (SLT) and structural risk minimization (SRM) [31,32] which has found many applications in different regression, classification, and pattern recognition problems. This approach was firstly proposed by Vapnik [33]. In SVM, the non-linear input area is transformed into a high-dimensional properties area and a hyperplane is found using a non-linear mapping. Furthermore, the solution to the SVM problem is available through solving a quadratic programming. This will lead to a time-consuming computation due to difficulties in solving a set of non-linear equations. In order to overcome the deficiencies of the SVM approach, Suykens and Vandewalle [34] represented the least square version of the support vector machine. The LSSVM benefits from SVM characteristics while diminishes the SVM deficiencies by substituting the SVM's non-linear constraints with the linear ones resulting in simple computation methods.

For a given training dataset of {xk, yk}, k = 1, 2, …, N, where xk $\in R^n$ denotes the kth input data, yk is the corresponding output value and N refers to the number of data points in the training dataset. The LSSVM utilizes the non-linear function to map the training data set from input space to the high dimensional space and uses the following expression to estimate the non-linear relationship of input and output variables:

$$y = \omega^T \varphi(x) + b \; with \; \omega \in R^{n_h}, b \in R, \varphi(.) \in R^n \rightarrow R^{n_h} \quad (1)$$

where ω represents the weight factor and b is the bias term. n and nh are dimensions of the data space and the unknown feature space [35,36]. LSSVM minimizes the following cost function (2) subjected to the corresponding constraints (3):

$$J(\omega, e) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 \quad (2)$$

$$y_k = \omega^T\phi(x_k) + b + e_k \; k = 1,2, \ldots,N \quad (3)$$

where γ represents the regularization parameter responsible for balancing the complexity of the model and training error, and ek is regression error. The Lagrangian form of the LSSVM is given by:

$$L(\omega, b, e, \alpha) = J(\omega, e) - \sum_{k=1}^{N} \alpha_k \{\omega^T\phi(x_k) + b + e_k - y_k\} \quad (4)$$

Lagrangian multipliers are denoted by α_k. The solution to Eq. (4) is available through equating the equation's derivatives to zero: