I&EC research
Industrial & Engineering Chemistry Research

pubs.acs.org/IECR

Article

Using COSMO-RS to Predict Hansen Solubility Parameters

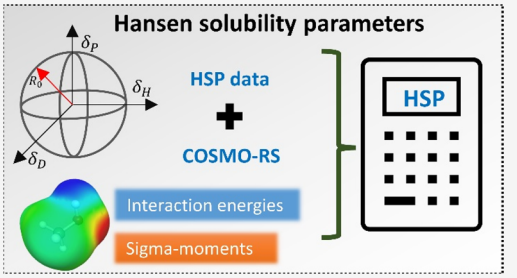
José Pedro Wojeicchowski, Ana M. Ferreira, Tifany Okura, Marlus Pinheiro Rolemberg, Marcos R. Mafra,* and João A. P. Coutinho*

Cite This: https://doi.org/10.1021/acs.iecr.2c01592

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Hansen solubility parameters (HSP) provide essential information on the nature of solvents, being a useful tool in their selection for product and process design. In this work, linear models were developed to estimate HSP based on the use of COSMO-RS (conductor-like screening model for realistic solvents) descriptors. Hansen solubility parameters for 195 compounds were obtained from the literature and classified into two categories: training set (133 compounds) and testing set (62 compounds). Then a factorial regression was carried out to predict dispersion, hydrogen-bonding interaction, and polar contribution of HSP (δ_D, δ_H, and δ_P, respectively) based on σ-moments and energy COSMO-RS descriptors. The description of the dispersion contribution was the least successful due to the low variability of the experimental data, despite the database containing compounds of a widely different chemical nature. The models obtained proved to be more than simple mathematical equations since a physical meaning is achieved when COSMO-RS descriptors are used. The models developed were then applied to the test database. An excellent performance was observed, with δ_H showing the highest R^2 (0.90), and the MAE obtained were 0.98, 1.74, and 1.44 MPa$^{1/2}$ for δ_D, δ_H, and δ_P, respectively, which are lower than those found in previous works. Finally, HSP data for caffeine, nicotine, paracetamol, and D-camphor were estimated. The results were close to the literature data and those predictive by the HSPiP software. Moreover, it was shown that δ_P and δ_H drive the molecules' polarity, impacting their log K_{ow} (a quantitative measure of polarity). This work shows that COSMO-RS descriptors are a tool to predict HSP through linear models, opening new doors in the screening, design, and selection of solvents to be used in chemical processes.



INTRODUCTION

Solvent selection is one of the most critical steps in chemical product and process design.[1] Besides water (the greenest and safest solvent)[2] and petroleum-based solvents, several alternatives have been proposed in the past few years, such as biobased solvents,[3,4] supercritical CO_2,[2,5] liquefied gases,[6] ionic liquids,[7] and deep eutectic solvents (DES).[2,7] Indeed, the recent increase of the range of solvents available is a major progress allowing the development of more sustainable processes; however, the choice of the most suitable solvents has been often carried out based on experimental screening. Thus, tools to support solvent selection are required to identify the best solvents rather than time-consuming and material-intensive trial-and-error approaches. There is still a gap in rapid and effective methods to predict the ability of a solvent (or a mixture of solvents), in particular, the most recent generation of alternative solvents to be used to solubilize a given solute.

The empirical "like dissolves like" rule can be the first approach to a solvent selection.[8] However, it is necessary to go further and define the "like" to have a quick and efficient method for choosing suitable solvents. In line with this idea, Hildebrand and Scott[9] were the pioneers of introducing the solubility parameter (δ) as a way to determine the affinity between solute and solvent. The solubility parameter is defined as the square root of the cohesive energy density

$$\delta = (E/V)^{1/2} \quad (1)$$

where V is the molar volume of the pure solvent and E is its energy of vaporization. However, δ was limited to molecular interactions based on van der Waals forces. In order to overcome this limitation, Hansen[8,10] proposed an extension of the unidimensional solubility parameter concept to a three-dimensional approach. More specifically, Hansen improved this concept through the development of a theory able to describe the total energy of vaporization of a liquid based on three individual energies: E_D, (atomic) dispersion forces; E_P (molecular) permanent dipole-permanent dipole force; and, E_H (molecular) hydrogen bonding. Dividing the individual contribution by the molar volume gives the respective Hansen cohesion energy parameters, eq 2:

$$\frac{E_T}{V} = \frac{E_D}{V} + \frac{E_P}{V} + \frac{E_H}{V} \quad (2)$$

Special Issue: Thermophysical Properties for Chemical Industry

Received: May 5, 2022
Revised: June 28, 2022
Accepted: July 1, 2022

Considering that δ was defined by eq 1, the square of the total Hansen solubility parameter can be defined as the sum of the squares of the Hansen individual components, eq 3:

$$\delta_T^2 = \delta_D^2 + \delta_P^2 + \delta_H^2 \quad (3)$$

The three-dimensional Hansen solubility scale provides information on the relative strengths of solvents, allowing the identification of the most suitable solvents to dissolve a specific solute where the distance between a solute i and a solvent j, "distance" R_a, depends on their respective solubility parameter components (eq 4):[8,11]

$$R_a = \sqrt{4(\delta_{D_i} - \delta_{D_j})^2 + (\delta_{P_i} - \delta_{P_j})^2 + (\delta_{H_i} - \delta_{H_j})^2} \quad (4)$$

Equation 4 was developed from experimental data where the constant "4" correctly represents the solubility data as a sphere encompassing the good solvents.[11] Thus, this solubility sphere is defined as the zone of solvent−solute solubilization, and the radius of the sphere is named as "interaction radius" R_0. Combining R_a and R_0, the relative energy difference (RED) is obtained and could be expressed as

$$\mathrm{RED} = R_a/R_0 \quad (5)$$

The RED values reveal the ability of a solvent in the solubilization of a solute. In other words, RED $\leq$ 1 indicates high affinity and RED > 1 low affinity between a given solute−solvent. However, it should be kept in mind that R_0 is based on experimental data resulting from the observation of the interaction between studied solutes and known solvents and thus can only be used when solubility experiments can be performed.

HSP has been demonstrated to be a key tool in the selection of the most promising solvents, being thus used in a wide range of solubility applications, such as pharmaceutical and chemical industries,[12,13] materials,[14] agricultural and food science,[8,15] and dentistry.[16] Due to the broad applicability of HSP, both experimental and theoretical methods for estimating these parameters have been proposed. HSP calculated through an equation of state[17] derived from statistical thermodynamics, or using a group contribution method,[18] are probably the most popular. Nevertheless, although the definition of the solubility parameters is simple, their experimental determination is not always easy, especially for nonvolatile compounds. Moreover, the experimental determination of the HSP requires pure materials and is generally expensive. Along this line, software like HSPiP make easier the selection of an extraction solvent and the HSP predictions from hundreds of compounds (https://www.hansen-solubility.com/HSPiP/). However, some limitations in the HSP simulation can be found, especially for solvents with large molecules or complex multicomponent systems, since the group contribution methods require the knowledge of all chemical group contributions, which is difficult, especially for alternative solvents (e.g., ionic liquids) or even mixtures involving strong molecular association (e.g., DES).[19] Dallos and co-workers[19] in 2011 proposed a quantitative structure−property relationship (QSPR) multivariate model based on artificial neural network (ANN) to predict the three-dimensional HSP using COSMO-RS (conductor-like screening model for realistic solvents) σ-moment descriptors. In a similar line, Panayiotou et al.[20] and Niederquell et al.[21] also demonstrated that COSMO-RS combined with other approaches (such as QSPR) for the calculation of the HSP allows for broadening the application and achieving better calculation accuracy. These works show relevant and promising results in the prediction of the three-dimensional HSP, mainly when COSMO-RS is applied, but are usually based on nonlinear methods, such as artificial neural networks, that are difficult to interpret and can only be used in one direction, i.e., used to predict HSP but cannot be readily used to design a solvent with a given HSP. Therefore, the development of a simple and effective predictive model for HSP, that can be used in both directions, remains a challenge.

The COSMO-RS theory is an effective way to link the molecular and the thermodynamic levels.[22,23] This quantum chemistry-based thermodynamic model is a variant of the dielectric continuum solvation modes, which considers the molecular surface's screening charge densities (σ) to calculate molecular interactions.[19] It has been shown that COSMO-RS descriptors tend to be excellent regressors relating some important material characteristics to molecular properties.[19,24] COSMO-RS theory has been applied to deduce the Abraham solute parameters[25] and the solvatochromic response of Reichardt's dye.[26] The outputs of this model, like misfit and van der Waals interaction energies, moment hydrogen bond donor, and moment hydrogen bond acceptor, can be correlated to other solubility scales, such as Kamlet−Taft.[27−29] Kundu et al.[30] developed predictive models of acidity (α) and basicity (β) Kamlet−Taf parameters of DES. Moreover, Wojeicchowski et al.[31] used COSMO-RS regressors to predict α, β, and dipolarity/polarizability (π^*) for organic and DES. Thus, COSMO-RS descriptors seem to be excellent regressors, allowing the development of simple linear predictive correlation models to estimate molecular properties.

Considering the need for simple, rapid, and efficient predictive methodologies to determine HSP, a predictive methodology using COSMO-RS descriptors is developed in this work to estimate HSP, based on linear models with physical meaning. For that, this work was divided into three steps. First, linear correlations are developed using only a part of the compounds taken from the literature (133 from 195 compounds). Then the methodology is tested on the remaining 62 compounds to demonstrate its applicability. Finally, the prediction models of the solubility parameters developed in this work and the typical software HSPiP will be compared in the estimation of HSP of caffeine, nicotine, paracetamol, and D-camphor.

METHODS

Training and Testing Sets of Molecules with HSP Data. A training/testing set of 195 compounds (Table S1, Supporting Information) was selected from the literature. It presents compounds with a wide range of chemistries with δ_D values ranging from 7.3 to 24.7 MPa$^{1/2}$, δ_H from 0 to 21.4 MPa$^{1/2}$, and δ_P from 0 to 42 MPa$^{1/2}$. It is important to work with this kind of diversity/quantity of data to develop modeling equations.

COSMO-RS. As previously introduced, this thermodynamic model provides several descriptors and outputs regarding the interaction ability of the molecules, with valuable physical meaning, to be explored in the prediction of molecular properties.[19,24] Molecules are considered as a group of surface segments, and each segment has a particular screening charge. To estimate thermodynamic quantities, COSMO-RS relies on pairwise interactions between each segment pair. However, first, the individual molecules' geometry and charge density must be optimized. In this work, each molecule was optimized

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

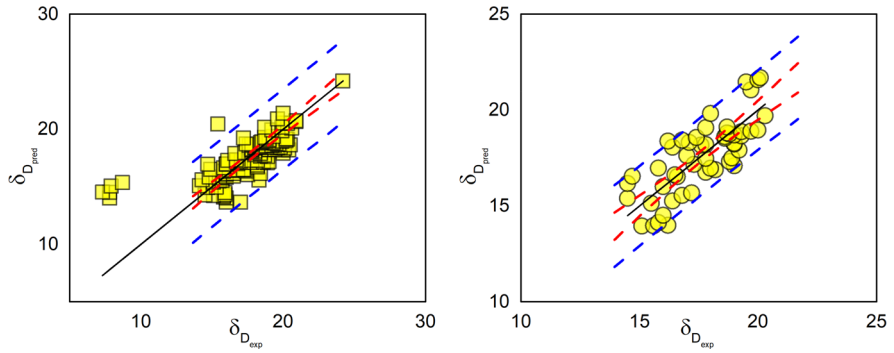


Figure 1. Experimental ($\delta_{D_{exp}}$) vs predicted ($\delta_{D_{pred}}$) values of dispersion HSP parameter, using correlations developed in this work, for the training (left) and testing (right). Dashed lines are the 95% prediction (blue) and confidence (red) intervals.

using the COSMO-BP-TZVP template of the TmoleX software package[32] (interface of TURBOMOLE), which includes a def-TZVP basis set, DFT with the B-P83 functional level of theory, and the COSMO solvation model (infinite permittivity). All COSMO-RS calculations were performed using the software COSMOtherm[33] with the BP_TZVP_C30_19.ctd parametrization, using the files ".cosmo" generated after the geometry optimization. Some of the σ-moments provided by COSMO-RS were selected as regressors to be applied in the modeling step, and the complete list is available in the Supporting Information (Table S2).

Factorial Regression. A factorial regression allows the analysis of the high-order interactive effects of multiple regressors. Considering the complexity of HSP modeling, especially using linear models,[19] the use of a factorial regression may be a good option, since the interaction between regressors improves the models' performance. Thus, δ_D, δ_H, and δ_P were the dependent variables, while M_i^X were the regressors provided by COSMO-RS, where M_i^X is the ith σ-moment (9 in all). According to Klamt et al.,[34] some of the σ-moments have a simple physical meaning. M_0^X is the total surface area of the molecule X. The first-order σ-moment is the total COSMO polarization charge on the surface. Note that these variables were well correlated to essential properties, such as log K_{ow} (octanol–water partition coefficient) working as QSPR or QSAR (quantitative structure–activity relationship) descriptors.[34] The building of models was based on the stepwise forward method, which starts with an empty model and the variables are added one by one, according to the statistical criteria of significance ($p < 0.05$), using Statistica 14 (TIBCO Software, Inc., 2020). Note that the data set was split into a training set (133 compounds) used to develop the models and a and testing set (62 compounds) for validation. The models were statistically evaluated by the squared correlation coefficient (R^2) of the experimental versus fitted or predicted values, mean absolute error (MAE), and mean relative error (MRE), according to eqs 6 and 7, respectively:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\delta_{i,j,\mathrm{predicted}} - \delta_{i,j,\mathrm{experimental}}| \quad (6)$$

$$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\delta_{i,j,\mathrm{predicted}} - \delta_{i,j,\mathrm{experimental}}|}{\delta_{i,j,\mathrm{experimental}}} \quad (7)$$

where $i = 1, ..., N$, where N is the number of cases, and j = D, P, or H, which means one of the Hansen parameters under analysis.

RESULTS AND DISCUSSION

Hansen solubility parameters of 195 compounds were obtained from the literature[19,35−40] (Table S1, Supporting Information). From this data set, 133 compounds were used as a training set to develop the models, and the remaining 62 became the testing set. A COSMO-RS descriptors database (Table S2, Supporting Information) was created based on these molecules, with the M_i^X parameters, besides E_{vdW}, E_{MF}, E_{HB}, and E_{int}, i.e., van der Waals, misfit, hydrogen bonding, and interaction energies. Some of these energy descriptors were previously used in the development of predictive correlation models, such as in the estimation of the Kamlet–Taft parameters for organic solvents, ionic liquids, and deep eutectic solvents.[31,41,42] COSMO-RS treats the molecules as a group of surface segments with a particular screening charge. Since this thermodynamic model relies on pairwise interactions, the parameters can quantify the specific interaction abilities of the molecules. Some of the M_i^X parameters have a clear physical meaning, for example M_2^X is highly correlated with the total COSMO polarization energy; i.e., it is a measure of the overall ability of the solute to interact electrostatically with a polarizable continuum.[43] M_3^X, in turn, indicates a kind of skewness of the σ-profile of compound X, while M_{Hba3}^X and M_{Hbd3}^X represent the acceptor and donor function, respectively.[19,43] E_{int} is the interaction enthalpy of the compound, which is separated into different contributions arising from electrostatic misfit (E_{MF}), hydrogen bonding (E_{HB}), and van der Waals interactions (E_{vdW}). Note that E_{int} may not be the sum of the mentioned contributions, especially when using conformers. The COSMO-RS descriptors chosen to correlate the Hansen parameters in this work were: surface area (area), M_2^X, M_3^X, M_S^X, M_{Hba3}^X, M_{Hbd3}^X, E_{int}, E_{vdW}, E_{MF}, and E_{HB}.

Modeling of δ_D. Considering the difficulties of developing linear HSP models, as related by Járvas and co-workers,[19] in this work a factorial regression was used. Despite being a linear regression, it considers the interaction between the factors, which could improve the models performance. Note that from the entire data set the dispersion parameter shows a specific characteristic with a mean of measured δ_D equal to 17.47 ± 2.56 $MPa^{1/2}$, which indicates a low coefficient of variation of the experimental data (ratio of the standard deviation to the mean). Thus, since there is a wide dispersion of COSMO-RS descriptor values but a low variability of δ_D, its modeling is difficult. After some preliminary tests, the area, M_2^X, M_S^X, M_{Hba3}^X, M_{Hbd3}^X, E_{int}, E_{vdW}, E_{MF}, and E_{HB} were selected as independent parameters on the δ_D factorial regression. Using these nine initial parameters on the factorial regression resulted in 511

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 1. Adjusted Coefficients for δ_D, δ_P, and δ_H Models Based on Training Dataset and *t*-Values of the Estimated Parameters

coefficients of the parameters	δ_D	δ_D *t*-value	δ_P	δ_P *t*-value	δ_H	δ_H *t*-value
x_1	-1.220×10^{00}	−4.35	-8.376×10^{-02}	−9.47	-2.116×10^{-02}	−3.76
x_2	2.903×10^{00}	8.14	1.624×10^{-01}	16.88	5.608×10^{-02}	4.49
x_3	1.446×10^{00}	4.18	1.264×10^{00}	6.90	-3.457×10^{00}	−12.89
x_4	8.993×10^{-01}	2.86	8.259×10^{-02}	6.86	8.302×10^{-03}	5.66
x_5	2.081×10^{-01}	4.98	1.280×10^{-01}	9.28	1.397×10^{-02}	9.97
x_6	9.456×10^{-02}	2.34	1.300×10^{01}	10.09	4.561×10^{-06}	3.77
x_7	3.199×10^{-03}	3.57			4.655×10^{00}	5.06
x_8	1.135×10^{-05}	3.10				
x_9	-2.564×10^{-04}	−2.44				
x_{10}	1.312×10^{01}	11.27				
MAE ($MPa^{1/2}$)	**1.09**		**1.78**		**2.01**	
MRE	**0.02**		**0.44**		**0.21**	
R^2	**0.51**		**0.77**		**0.84**	

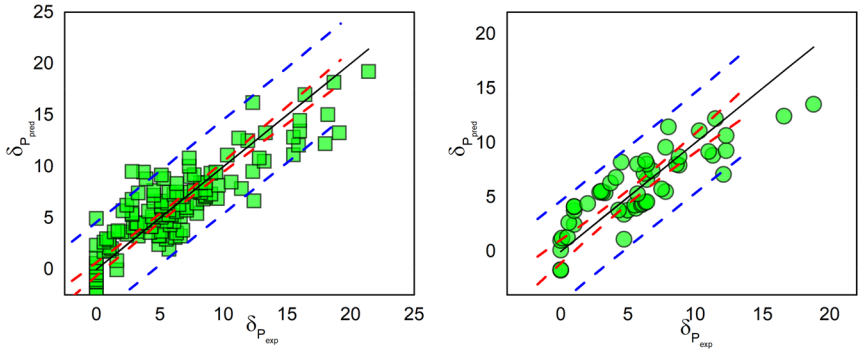


Figure 2. Experimental ($\delta_{P_{exp}}$) vs predicted ($\delta_{P_{pred}}$) values of dispersion HSP parameter, using correlations developed in this work, for the training (left) and testing (right). Dashed lines are the 95% prediction (blue) and confidence (red) intervals.

evaluated effects, including all levels of linear combinatorial effect. Just nine effects should be considered according to the Pareto chart (Figure S1) (Supporting Information), since they are statistically significant factors ($p < 0.05$) for δ_D modeling, besides the intercept effect.

Figure 1 shows the regression for both training (left) and testing (right). Predicted δ_D obtained for the training set indicates that most of the data were inside the prediction interval. The outliers were not removed at this time, resulting in a determination coefficient (R^2) equal to 0.51, as shown in Table 1, which also reveals the fitted coefficients of eq 8. This equation was applied to the testing database, resulting in Figure 1 (right).

$$\delta_D = x_1 E_{int} + x_2 E_{MF} + x_3 E_{HB} + x_4 E_{vdW} + x_5 M_{Hba3} M_{Hbd3} + x_6 M_{Hbd3} E_{vdW} + x_7 E_{vdW} M_5^X + x_8 \text{area} M_2^X E_{vdW} + x_9 M_{Hba3} E_{vdW} M_5^X + x_{10} \quad (8)$$

Moreover, this validation was also done without considering the outliers, and R^2 was equal to 0.68, with a MAE value of 0.98 $MPa^{1/2}$ and MRE value of 0.06 (Table S3, Supporting Information). As discussed before, due to the low variability of dispersion data, the fair performance of this model was expected. Anyway, a straight Pearson correlation (r) between experimental and estimated data was shown, around 0.82, which is considered high.

Modeling of δ_P. The Hansen polar solubility parameter data set showed higher variability in comparison to δ_D. The mean value of δ_P was 6.15 ± 4.67 $MPa^{1/2}$, indicating a coefficient of variation equal to 76%, with values ranging from 0 to 21.4 $MPa^{1/2}$. This could be positive in this modeling approach because COSMO-RS descriptors can be properly correlated to δ_P. For example, when δ_P is relatively high, M_2^X increases. Initially, area, M_2^X, M_{Hba3}^X, M_{Hbd3}^X, E_{int}, E_{vdW}, E_{MF} and, E_{HB} were selected as independent parameters on the δ_P factorial regression. As expected, the eq 9 obtained to estimate δ_P was simpler in comparison to δ_D model. Of the 256 calculated effects, only five were statistically significant, besides the intercept, as confirmed by the Pareto chart (Figure S2, Supporting Information). An R^2 equal to 0.77 was achieved for the training set, as presented in Table 1 and depicted in Figure 2 (left).

$$\delta_P = x_1 \text{area} + x_2 M_2^X + x_3 E_{int} + x_4 E_{int} M_{Hbd3} + x_5 E_{int} E_{vdW} + x_6 \quad (9)$$

It was clear how spread the data are in Figure 2, indicating the high variability previously mentioned. Figure 2 (right) shows the results of testing data with a Pearson correlation, between experimental and predicted, equal to 0.88, after removing the outliers, with an MAE value of 1.74 $MPa^{1/2}$ and MRE equal to 0.93 (Table S3, Supporting Information). Other works also obtained better R^2 values for δ_P than δ_D, followed by lower δ_D MAE errors.[19,44,45] Note that the classification of testing and training points were the same for all Hansen solubility parameters, and it was made randomly through Statistica 14 software. Since COSMO-RS is a thermodynamic model based on quantum chemistry, its descriptors allow the construction of physically meaningful models. From Figure S2 and Table 1 (δ_P *t*-value) it is possible to note that M_2^X is the

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

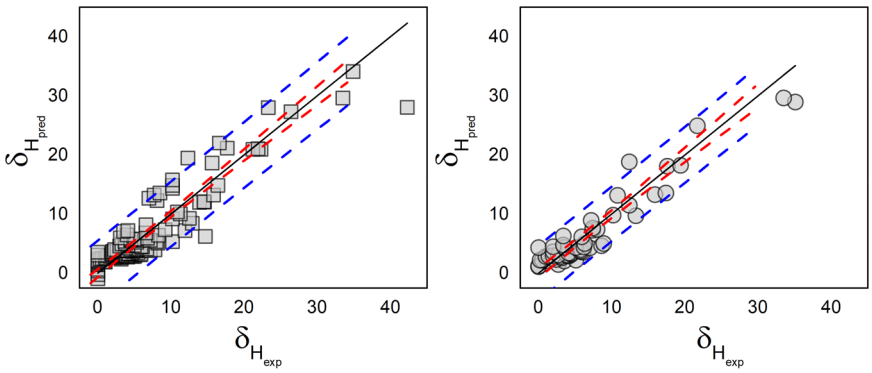


Figure 3. Experimental ($\delta_{H_{exp}}$) vs predicted ($\delta_{H_{pred}}$) values of dispersion HSP parameter, using correlations developed in this work, for the training (left) and testing (right). Dashed lines are the 95% prediction (blue) and confidence (red) intervals.

most important factor to explain δ_P, although the coefficient associated with it was not the highest. Thus, since δ_P is related to polar interaction, this high correlation to M_2^X makes sense, that is, the ability of "X" to interact electrostatically with a polarizable media. In other words, the higher the M_2^X the higher is δ_P. Besides that, M_2^X was not an effect when modeling δ_D, which is a Hansen parameter associated with nonpolar interaction. It means that, even though initially M_2^X was available to be a part of the predictive model for δ_D, its absence of physical meaning to δ_D was captured by the statistical analysis, which indicated M_2^X, and its individual contribution and combinatorial contribution to the other effects, were, in fact, not significant ($p > 0.05$).

Modeling of δ_H. As observed in the δ_P, Hansen hydrogen bonding data present high variability, ranging from 0 to 42.3 $MPa^{1/2}$. Despite the high extreme value, its mean was 7.42 ± 7.25 $MPa^{1/2}$, which represents the highest coefficient of variation, 102%. Thus, based on our achievements on the other Hansen parameters, we expected a high correlation between COSMO-RS descriptors and δ_H, due to the variability of both dependent and independent variables. Nevertheless, this is not a general rule. In this study, according to Table S4 (Supporting Information), δ_H exhibited the most statistically significant correlation (positive and negative) to the selected COSMO-RS descriptors. On the other hand, the Pearson correlation between δ_D and the descriptors were not as good as those from δ_H and δ_P. This supports the lower R^2 values when predicting the dispersion contribution on HSP. Moreover, Table S4 revealed an important insight, i.e., some individual factors, such as M_{Hbac3} and M_{Hbd3} are not significant to δ_D. However, their combination M_{Hbac3} and M_{Hbd3} were statistically relevant on factorial regression, which had a positive effect on the prediction of δ_D (Table 1 and Figure S1 in the Supporting Information). It means that, even though an individual independent variable is not correlated to a variable, such as δ_D, its combination with other variables can become relevant.

After some initial tests, area, M_2^X, M_3^X, M_5^X, M_{Hba3}^X, M_{Hbd3}^X, and E_{HB} were selected as independent parameters on the δ_H factorial regression, resulting in 126 evaluated effects. The Pareto chart (Figure S3, Supporting Information) and Table 1 (δ_H t-value) showed that on the model prediction of δ_H only six parameters were statistically significant, besides the intercept (eq 10). Moreover, the coefficient of E_{HB} was the most significant effect and its value is negative (−3.46). Note that the negative signal of this coefficient does not mean a negative impact on the dependent variable because E_{HB} is also negative. The obtained R^2 value for the training data set (Figure 3, left) was 0.84. For the testing set, R^2 was greater, 0.90, after removal of the outliers, with an MRE equal to 0.62 and an MAE value of 1.44 $MPa^{1/2}$. These results indicated a good performance of the proposed model for the correlation between COSMO-RS descriptors and Hansen hydrogen bonding solubility parameter.

$$\delta_H = x_1\text{area} + x_2 M_2^X + x_3 E_{HB} + x_4 M_2^X M_{Hba3} + x_5 \text{area} E_{HB} + x_6 M_3^X M_5^X M_{Hba3} E_{HB} + x_7 \quad (10)$$

After predicting δ_D, δ_P, and δ_H the square of the total solubility parameter was estimated through eq 3, and the results are depicted in Figure 4. The majority of the predicted

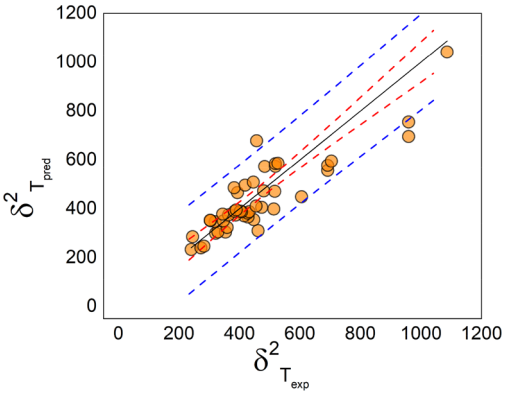


Figure 4. Experimental ($\delta^2_{T_{exp}}$) vs predicted ($\delta^2_{T_{pred}}$) values of dispersion HSP parameter, using correlations developed in this work, for the testing. Dashed lines are the 95% prediction (blue) and confidence (red) intervals.

δ_T^2 were between the prediction lines, indicating a good description ($R^2 = 0.71$) of the developed HSP models. The R^2 has this value due to the performance on the estimation of δ_D (δ_D^2) as discussed previously. In summary, a simple way to correlate HSP with COSMO-RS descriptors was demonstrated by using linear models based on factorial regression. Moreover, most of the estimated values were close to or within the experimental error associated with the determination of solubility parameters, as shown in Table S3 (Supporting Information).

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

A simple comparison of the estimation errors of the results obtained in this work and other HSP prediction methods was carried out. Járvas et al.[19] obtained 1.37, 1.85, and 2.58 $MPa^{1/2}$ as errors for δ_D, δ_P, and δ_H, respectively. Using a similar (about 62% of common data), but bigger, data set, this work resulted in lower deviations, 0.98, 1.74, and 1.44 $MPa^{1/2}$ for δ_D, δ_P, and δ_H, respectively (Table S3). Indeed, the models developed in this work achieved better performance than similar works found in the literature,[17,44,45] which further highlight our approach.

Estimation of HSP for Different Solutes. According to the "like-dissolve-like" rule, different compounds with similar HSP values mean should have a certain degree of affinity. This principle could be verified through eqs 4 and 5. Considering a minimal difference between i and j, R_a tends to a minimum value, and thus, RED results in $\ll 1$, which means high affinity. Based on this, a three-dimensional plot was made in order to compare the solubility power of the 195 studied compounds (circles) with four solutes (stars), namely: caffeine, nicotine, paracetamol, and D-camphor; see Figure 5. The model solutes

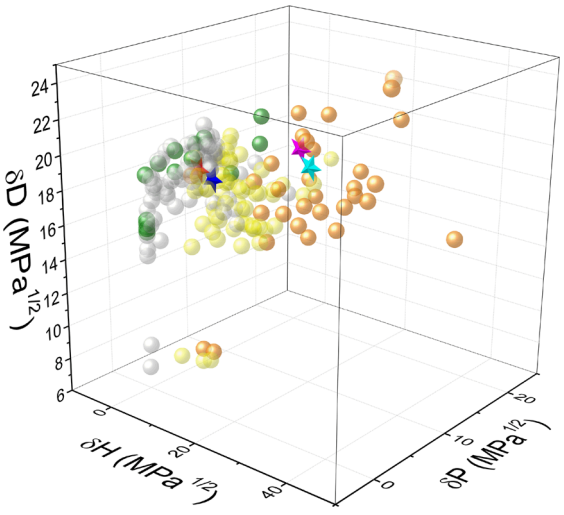


Figure 5. Experimental HSP data (circles) and predicted (stars) according to the following classification: log K_{ow} < 0 (orange circle); 0 < log K_{ow} < 2 (yellow circle); 2 < log K_{ow} < 4.5 (gray circle), log K_{ow} > 4.5 (green circle), caffeine (pink star), nicotine (dark blue star), paracetamol (light blue star), and D-camphor (red star).

were selected due to their importance to the chemical and pharmaceutical industries.[46−53] For the estimation of their HSP, the models developed in this work were used. Regarding the compounds, they were grouped taking into account their polarity level according to log K_{ow} values (a quantitative measure of solvent polarity); i.e., different colors mean different polarity levels. log K_{ow} values are usually between −3 (very hydrophilic) and +10 (extremely hydrophobic).[54] Furthermore, this property is quite important for predicting the distribution in the environment (water, soil, biota, etc.), e.g., chemicals with high log K_{ow} (>4.5) raise more concerns because they may have the potential to bioconcentrate in living organisms, which may induce toxicity.[55] Along this line, Veeger and co-workers[56] developed a general rule, where the efficiency of biocatalysis in organic solvents is low for compounds with log K_{ow} < 2, moderate for compounds with 2 < log K_{ow} < 4.5, and high for compounds log K_{ow} > 4.5. Thus, in Figure 5, the HSP of 195 compounds (data) were grouped according to their classification of the log K_{ow} discussed above.

First, the polar solvents (orange circles, with log K_{ow} < 0) are concentrated in a specific range of HSP values, where caffeine and paracetamol are inside them. These results are in agreement with the log K_{ow} of these solutes, with caffeine (−0.07) being more hydrophilic than paracetamol (0.58). On the other hand, nicotine and camphor are far from the other two solutes and close to a more lipophilic zone (yellow and light gray circles). Thus, it means that to solubilize d-camphor a hydrophobic solvent should be applied, including some of the more apolar ones (log K_{ow} > 4.5). This simple analysis allows an initial guess about solute−solvent affinity. Moreover, the results of log K_{ow} are aligned to HSP data, indicating that δ_P and δ_H drive the molecules' polarity. The prediction obtained by this work was close to the results from the HSPiP database and also predicted by the "DIY- HSPiP" tool, Table S5.

CONCLUSION

Previous works reported the difficulties of modeling HSP through linear regression. However, in this work, linear models were obtained using a factorial regression approach based on the predictive power of COSMO-RS. This quantum-based thermodynamic model provided σ-moments and energy descriptors, resulting in correlation models to predict HSP, with satisfactory adjustments and high physical meaning. The determination coefficient (R^2) for the training models were equal to 0.51, 0.77, and 0.84 for δ_D, δ_H, and δ_P, respectively. The obtained mean absolute errors (MAE) were similar to those found in previous works. A very good performance was observed in the testing set after removal of the outliers. The MAE were 0.98, 1.74, and 1.44 $MPa^{1/2}$ for δ_D, δ_H, and δ_P, respectively. Finally, HSP data for important solutes for chemical and pharmaceutical industries caffeine were estimated. The results were close to the literature data and those predictive by the HSPiP software. Moreover, it was shown that δ_P and δ_H drive the molecules polarity, impacting their log K_{ow}. This work demonstrated the potential to combine the predictive power of COSMO-RS and HSP, providing simple linear correlation models, which can be applied in the screening, design, and selection of solvents to be used in chemical processes.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.2c01592.

Database of 195 Hansen Solubility Parameters (HSP) from the literature; data for COSMO-RS descriptors for the created database; Pareto chart of t-values for the coefficients of the δ_D model; experimental and predicted HSP values for the testing set after removing outliers; Pareto chart of t-values for the coefficients of the δ_P model; Pareto chart of t-values for the coefficients of the δ_H model; correlations between dependent and independent variables; HSP of different solutes predicted in this work in comparison to results estimate by the "Do it yourself (DIY) tool" and data from HSPiP database (PDF)

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Authors

Marcos R. Mafra − *Department of Chemical Engineering, Federal University of Paraná (UFPR), Curitiba, PR 81531-990, Brazil*; Phone: +55 3361-3586; Email: marcos.mafra@ufpr.br

João A. P. Coutinho − *CICECO − Aveiro Institute of Materials, Department of Chemistry, University of Aveiro (UA), Aveiro 3810-193, Portugal*; orcid.org/0000-0002-3841-743X; Phone: +351 234 370 200; Email: jcoutinho@ua.pt

Authors

José Pedro Wojeicchowski − *Department of Chemical Engineering, Federal University of Paraná (UFPR), Curitiba, PR 81531-990, Brazil; CICECO − Aveiro Institute of Materials, Department of Chemistry, University of Aveiro (UA), Aveiro 3810-193, Portugal*; orcid.org/0000-0002-6790-8510

Ana M. Ferreira − *CICECO − Aveiro Institute of Materials, Department of Chemistry, University of Aveiro (UA), Aveiro 3810-193, Portugal*; orcid.org/0000-0003-3057-5019

Tifany Okura − *Department of Chemical Engineering, Federal University of Paraná (UFPR), Curitiba, PR 81531-990, Brazil*

Marlus Pinheiro Rolemberg − *Science and Technology Institute, Federal University of Alfenas (UNIFAL), Poços de Caldas, MG 37715-400, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.2c01592

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partly developed within the scope of the project CICECO-Aveiro Institute of Materials, UIDB/50011/2020, UIDP/50011/2020 & LA/P/0006/2020, financed by national funds through the FCT/MEC (PIDDAC). J. P. Wojeicchowski is grateful for the scholarship (88881.361904/2019-01) provided by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil) and by Banco Santander S.A (Brazil). M. R. Mafra is grateful to the Brazilian National Council for Scientific and Technological Development (CNPq Grant 308517/2018-0).

REFERENCES

(1) Herrero, M.; Ibáñez, E. Green Processes and Sustainability: An Overview on the Extraction of High Added-Value Products from Seaweeds and Microalgae. *J. Supercrit. Fluids* **2015**, *96*, 211−216.

(2) Claux, O.; Santerre, C.; Abert-Vian, M.; Touboul, D.; Vallet, N.; Chemat, F. Alternative and Sustainable Solvents for Green Analytical Chemistry. *Curr. Opin. Green Sustain. Chem.* **2021**, *31*, 100510.

(3) Farooque, S.; Rose, P. M.; Benohoud, M.; Blackburn, R. S.; Rayner, C. M. Enhancing the Potential Exploitation of Food Waste: Extraction, Purification, and Characterization of Renewable Specialty Chemicals from Blackcurrants (Ribes Nigrum L.). *J. Agric. Food Chem.* **2018**, *66* (46), 12265−12273.

(4) Ravi, H. K.; Breil, C.; Vian, M. A.; Chemat, F.; Venskutonis, P. R. Biorefining of Bilberry (Vaccinium Myrtillus L.) Pomace Using Microwave Hydrodiffusion and Gravity, Ultrasound-Assisted, and Bead-Milling Extraction. *ACS Sustain. Chem. Eng.* **2018**, *6* (3), 4185−4193.

(5) Ferrentino, G.; Giampiccolo, S.; Morozova, K.; Haman, N.; Spilimbergo, S.; Scampicchio, M. Supercritical Fluid Extraction of Oils from Apple Seeds: Process Optimization, Chemical Characterization and Comparison with a Conventional Solvent Extraction. *Innov. Food Sci. Emerg. Technol.* **2020**, *64*, 102428.

(6) Kanda, H.; Hoshino, R.; Murakami, K.; Wahyudiono; Zheng, Q.; Goto, M. Lipid Extraction from Microalgae Covered with Biomineralized Cell Walls Using Liquefied Dimethyl Ether. *Fuel* **2020**, *262*, 116590.

(7) Vanda, H.; Dai, Y.; Wilson, E. G.; Verpoorte, R.; Choi, Y. H. Green Solvents from Ionic Liquids and Deep Eutectic Solvents to Natural Deep Eutectic Solvents. *Comptes Rendus Chimie*; Elsevier Masson SAS June 1, 2018; pp 628−638. DOI: 10.1016/j.crci.2018.04.002.

(8) Sánchez-Camargo, A. del P.; Bueno, M.; Ballesteros-Vivas, D.; Parada-Alfonso, F.; Cifuentes, A.; Ibañez, E. Hansen Solubility Parameters for Selection of Green Extraction Solvents. *Comprehensive Foodomics*; Elsevier, 2021; pp 710−724. DOI: 10.1016/b978-0-08-100596-5.22814-x.

(9) Hildebrand, J. H.; Scott, R. L. *The Solubility of Nonelectrolytes*, 3rd ed.; Reinhold: New York, 1950.

(10) Hansen, C. *The Three Dimensional Solubility Parameter and Solvent Diffusion Coefficient, Their Importance in Surface Coating Formulation*; Techinical University of Denmark, 1967.

(11) Hansen, C. M. *Hansen Solubility Parameters: A User's Handbook*, 2nd ed.; CRC Press, 2007. DOI: 10.1201/9781420006834.

(12) Petříková, E.; Patera, J.; Gorlová, O. Influence of Active Pharmaceutical Ingredient Structures on Hansen Solubility Parameters. *Eur. J. Pharm. Sci.* **2021**, *167*, 106016.

(13) Yu, S.; Sharma, R.; Morose, G.; Nagarajan, R. Identifying Sustainable Alternatives to Dimethyl Formamide for Coating Applications Using Hansen Solubility Parameters. *J. Clean. Prod.* **2021**, *322*, 129011.

(14) Gao, M.; Zhang, Z.; Zhang, W.; Cao, Q.; Tang, Z.; Zhao, W. Understanding the Top-down Fragmentation of 2D Material in Miscible Liquid Environment Based on Hansen Solubility Parameters Theory. *FlatChem.* **2022**, *32*, 100346.

(15) Terrell, E. Estimation of Hansen Solubility Parameters with Regularized Regression for Biomass Conversion Products: An Application of Adaptable Group Contribution. *Chem. Eng. Sci.* **2022**, *248*, 117184.

(16) Toledano, M.; Vallecillo-Rivas, M.; Aguilera, F. S.; Osorio, M. T.; Osorio, E.; Osorio, R. Polymeric Zinc-Doped Nanoparticles for High Performance in Restorative Dentistry. *J. Dent.* **2021**, *107*, 103616.

(17) Stefanis, E.; Tsivintzelis, I.; Panayiotou, C. The Partial Solubility Parameters: An Equation-of-State Approach. *Fluid Phase Equilib.* **2006**, *240* (2), 144−154.

(18) Stefanis, E.; Panayiotou, C. A New Expanded Solubility Parameter Approach. *Int. J. Pharm.* **2012**, *426* (1−2), 29−43.

(19) Járvás, G.; Quellet, C.; Dallos, A. Estimation of Hansen Solubility Parameters Using Multivariate Nonlinear QSPR Modeling with COSMO Screening Charge Density Moments. *Fluid Phase Equilib.* **2011**, *309* (1), 8−14.

(20) Panayiotou, C. Equation-of-State Models and Quantum Mechanics Calculations. *Ind. & Eng. Chem. Res.* **2003**, *42* (7), 1495−1507.

(21) Niederquell, A.; Wyttenbach, N.; Kuentz, M. New Prediction Methods for Solubility Parameters Based on Molecular Sigma Profiles Using Pharmaceutical Materials. *Int. J. Pharm.* **2018**, *546* (1−2), 137−144.

(22) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc. Perkin Trans. 2* **1993**, No. 5, 799−805.

(23) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102* (26), 5074−5085.

(24) Marsh, K. N. COSMO-RS from Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design. By A. Klamt. Elsevier: Amsterdam, The Netherlands, 2005. 246 Pp. $US 165. ISBN 0−444−51994−7. *J. Chem. Eng. Data* **2006**, *51* (4), 1480−1480.
(25) Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Harris, J. M.; Taft, R. W. Linear Solvation Energy Relationships. Part 37. An Analysis of Contributions of Dipolarity−Polarisability, Nucleophilic Assistance, Electrophilic Assistance, and Cavity Terms to Solvent Effects on t-Butyl Halide Solvolysis Rates. *J. Chem. Soc. Perkin Trans. 2* **1987**, No. 7, 913−920.
(26) Palomar, J.; Torrecilla, J. S.; Lemus, J.; Ferro, V. R.; Rodríguez, F. A COSMO-RS Based Guide to Analyze/Quantify the Polarity of Ionic Liquids and Their Mixtures with Organic Cosolvents. *Phys. Chem. Chem. Phys.* **2010**, *12* (8), 1991−2000.
(27) Taft, R. W.; Kamlet, M. J. The Solvatochromic Comparison Method. 2. The α-Scale of Solvent Hydrogen-Bond Donor (HBD) Acidities. *J. Am. Chem. Soc.* **1976**, *98* (10), 2886−2894.
(28) Kamlet, M. J.; Abboud, J. L.; Taft, R. W. The Solvatochromic Comparison Method. 6. The Π* Scale of Solvent Polarities1. *J. Am. Chem. Soc.* **1977**, *99* (18), 6027−6038.
(29) Kamlet, M. J.; Taft, R. W. The Solvatochromic Comparison Method. I. The β-Scale Of Solvent Hydrogen-Bond Acceptor (HBA) Basicities. *J. Am. Chem. Soc.* **1976**, *98* (2), 377−383.
(30) Kundu, D.; Rao, P. S.; Banerjee, T.; Rao, S.; Banerjee, T. First Principle Prediction of Kamlet-Taft Solvatochromic Parameters of Deep Eutectic Solvent Using COSMO-RS Model. *Ind. Eng. Chem. Res.* **2020**, *59* (24), 11329−11339.
(31) Wojeicchowski, J. P.; Abranches, D. O.; Ferreira, A. M.; Mafra, M. R.; Coutinho, J. A. P. Using COSMO-RS to Predict Solvatochromic Parameters for Deep Eutectic Solvents. *ACS Sustain. Chem. Eng.* **2021**, *9* (30), 10240−10249.
(32) Steffen, C.; Thomas, K.; Huniar, U.; Hellweg, A.; Rubner, O.; Schroer, A. TmoleX-A Graphical User Interface for TURBOMOLE. *J. Comput. Chem.* **2010**, *31* (16), 2968−2970.
(33) Eckert, F.; Klamt, A. Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach. *AIChE J.* **2002**, *48* (2), 369−385.
(34) Klamt, A. The σ-Moment Approach. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Leverkusen, 2005; pp 137−147. DOI: 10.1016/B978-044451994-8/50009-2.
(35) Hansen, C. M.; Smith, A. L. Using Hansen Solubility Parameters to Correlate Solubility of C60 Fullerene in Organic Solvents and in Polymers. *Carbon N. Y.* **2004**, *42* (8−9), 1591−1597.
(36) Novo, L. P.; Curvelo, A. A. S. Hansen Solubility Parameters: A Tool for Solvent Selection for Organosolv Delignification. *Ind. Eng. Chem. Res.* **2019**, *58* (31), 14520−14527.
(37) Batista, M. *Determinação Dos Parâmetros de Solubilidade de Hansen de Ésteres Graxos Etílicos*; Universidade Estadual de Campinas, 2010.
(38) Azevedo, G. R. *Avaliação Dos Parâmetros de Solubilidade de Hansen Do Paracetamol e de Óleos Essenciais*; Universidade Federal do Maranhão, 2013.
(39) Lambourne, R. Solvents, Thinners, and Diluents. In *Paint and Surface Coatings*; Woodhead Publishing, 1999; pp 166−184. DOI: 10.1533/9781855737006.166.
(40) Villa, R. D.; De Oliveira, A. P.; Pupo Nogueira, R. F. Avaliação Dos Parâmetros de Solubilidade de Hildebrand/Hansen Na Seleção de Solventes Para a Extração de Pesticidas Organoclorados Do Solo. *Quim. Nova* **2011**, *34* (9), 1501−1506.
(41) Cláudio, A. F. M.; Swift, L.; Hallett, J. P.; Welton, T.; Coutinho, J. A. P.; Freire, M. G. Extended Scale for the Hydrogen-Bond Basicity of Ionic Liquids. *Phys. Chem. Chem. Phys.* **2014**, *16* (14), 6593−6601.
(42) Kurnia, K. A.; Lima, F.; Cláudio, A. F. M.; Coutinho, J. A. P.; Freire, M. G. Hydrogen-Bond Acidity of Ionic Liquids: An Extended Scale. *Phys. Chem. Chem. Phys.* **2015**, *17* (29), 18980−18990.
(43) Klamt, A. *COSMO-RS from Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, 1st ed.; Elsevier: Amsterdam, 2005. DOI: 10.1016/b978-044451994-8/50000-6.
(44) Belmares, M.; Blanco, M.; Goddard, W. A., III; Ross, R. B.; Caldwell, G.; Chou, S.-H.; Pham, J.; Olofson, P. M.; Thomas, C. Hildebrand and Hansen Solubility Parameters from Molecular Dynamics with Applications to Electronic Nose Polymer Sensors. *J. Comput. Chem.* **2004**, *25* (15), 1814−1826.
(45) Stefanis, E.; Panayiotou, C. Prediction of Hansen Solubility Parameters with a New Group-Contribution Method. *Int. J. Thermophys.* **2008**, *29* (2), 568−585.
(46) Cai, C.; Li, F.; Liu, L.; Tan, Z. Deep Eutectic Solvents Used as the Green Media for the Efficient Extraction of Caffeine from Chinese Dark Tea. *Sep. Purif. Technol.* **2019**, *227*, 115723.
(47) Rigueto, C. V. T.; Nazari, M. T.; De Souza, C. F.; Cadore, J. S.; Brião, V. B.; Piccin, J. S. Alternative Techniques for Caffeine Removal from Wastewater: An Overview of Opportunities and Challenges. *J. Water Process Eng.* **2020**, *35*, 101231.
(48) Archie, S. R.; Sharma, S.; Burks, E.; Abbruscato, T. Biological Determinants Impact the Neurovascular Toxicity of Nicotine and Tobacco Smoke: A Pharmacokinetic and Pharmacodynamics Perspective. *Neurotoxicology* **2022**, *89*, 140−160.
(49) Wagner, F. F.; Comins, D. L. Recent Advances in the Synthesis of Nicotine and Its Derivatives. *Tetrahedron* **2007**, *63* (34), 8065−8082.
(50) Werner, N. S.; Duschek, S.; Schandry, R. D-Camphor-Crataegus Berry Extract Combination Increases Blood Pressure and Cognitive Functioning in the Elderly − A Randomized, Placebo Controlled Double Blind Study. *Phytomedicine* **2009**, *16* (12), 1077−1082.
(51) Wongchompoo, W.; Buntem, R. Microencapsulation of Camphor Using Trimethylsilylcellulose. *Carbohydr. Polym. Technol. Appl.* **2022**, *3*, 100194.
(52) Boumya, W.; Taoufik, N.; Achak, M.; Barka, N. Chemically Modified Carbon-Based Electrodes for the Determination of Paracetamol in Drugs and Biological Samples. *J. Pharm. Anal.* **2021**, *11* (2), 138−154.
(53) Chen, X.; Li, Y.; Li, X.; Li, R.; Ye, B. Transition Metal Copper Composite Ionic Liquid Self-Built Ratiometric Sensor for the Detection of Paracetamol. *Anal. Chim. Acta* **2021**, 338992.
(54) Cumming, H.; Rücker, C. Octanol−Water Partition Coefficient Measurement by a Simple 1H NMR Method. *ACS Omega* **2017**, *2* (9), 6244−6249.
(55) European Chemical Agency. Chapter R.7c: Endpoint Specific Guidance Draft Version 3.0. In *Guidance on Information Requirements and Chemical Safety Assessment*; Helsinki, 2017; p 274.
(56) Laane, C.; Boeren, S.; Vos, K.; Veeger, C. Rules for Optimization of Biocatalysis in Organic Solvents. *Biotechnol. Bioeng.* **1987**, *30* (1), 81−87.

https://doi.org/10.1021/acs.iecr.2c01592
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX