PAPER

Check for updates

Integrating experimental data and machine learning models for solubility prediction of yellow 23 in supercritical carbon dioxide

Seyed Ali Sajadian, [ID] *[a] Amir Hossein Sheikhshoaei, [ID] [b] Nadia Esfandiari [ID] [c] and Adel Noubigh [ID] [d]

This study reports, for the first time, the solubility investigation of yellow 23 in supercritical carbon dioxide over a pressure range of 12–30 MPa and a temperature range of 313–343 K. Yellow 23's experimental molar solubilities in supercritical carbon dioxide were found to be between 6.67×10^{-5} to 20.55×10^{-5} (313 K), 4.32×10^{-5} to 23.58×10^{-5} (323 K), 3.41×10^{-5} to 27.37×10^{-5} (333 K) and $2.29.3 \times 10^{-5}$ to $3.840.4 \times 10^{-5}$ (343 K). Four semiempirical correlations (MST, Chrastil, Bartle *et al.*, and K-J) were used to calculate the solubility of yellow 23 in supercritical carbon dioxide. The machine learning models (Multilayer Perceptron, Gaussian process regression, Random Forest) models were considered for modeling in this research. The K-J model proved to be the most suitable for fitting the experimental data, exhibiting the lowest mean absolute relative deviation of 6.39%. All three machine learning models have impressive act on approximation of yellow 23 solubility. However, model MLP with the highest R^2 (99.7) and lowest MSE (0.001) was selected as the best among the three models.

1 Introduction

Tartrazine (yellow 23) is a synthetic lemon-yellow azo dye commonly used as a coloring agent in a variety of industries such as food, pharmaceutical, cosmetic, textile and paper.[1–5] During dyeing processes with this substance, about 50% of the azo dyes used are wasted and sent to wastewater. However, its common use and water solubility present serious environmental challenges, as it is resistant to biodegradation and can be toxic to aquatic organisms.[6] Conventional dyeing and wastewater treatment methods involving yellow 23 commonly result in the making of large capacity of contaminated water and chemical waste. It is possible to extract organic dyes from wastewater using supercritical carbon dioxide (scCO$_2$). In recent years, supercritical fluid dyeing has received attention. In this method, scCO$_2$ is used instead of water. Using scCO$_2$ instead of water will reduce water consumption, eliminate the need for drying after dyeing, and allow for convenient and economical recycling of carbon dioxide, without producing toxic wastewater, and without environmental pollution. To perform this

process, data on the solubility of yellow 23 in scCO$_2$ is required. Sometimes it is necessary to produce colored nanoparticles for better solubility in other solvents. If the way of creating nanoparticles is to use supercritical fluid, the solubility data of the dye in CO$_2$ should be studied. Because the choice of the method of producing nanoparticles is affected by the solubility data.[7–11]

Supercritical carbon dioxide has gained substantial attention as a promising solvent for extraction, purification, and dyeing applications, primarily due to its unique combination of liquid-like density providing enhanced solvation capability and gas-like diffusivity and low viscosity, which facilitate mass transfer. Additionally, its tunable solvent strength through adjustments in pressure and temperature, along with its environmentally benign characteristics such as non-toxicity and ease of separation from products, further support its application as a sustainable processing medium. This methodology can be applied in many medical applications.[12–18] Despite these established advantages frequently highlighted in the literature, a critical knowledge gap persists concerning the fundamental mechanisms governing solute dissolution in scCO$_2$, particularly for polar or structurally complex organic dyes. The solubility behavior of dyes in scCO$_2$ cannot be attributed solely to variations in solvent density or operating pressure. Instead, it is governed by a combination of molecular-level factors, including dye polarity, hydrogen-bonding capacity, molecular size, conformational rigidity, and the inherent compatibility between the dye and the solvent environment. Prior studies on disperse dyes have demonstrated that solubility generally increases with CO$_2$ density typically achieved through elevated

[a] Department of Chemical Engineering, Faculty of Engineering, University of Kashan, Kashan, Iran. E-mail: seyedali.sajadian@gmail.com

[b] Petroleum and Petrochemical Engineering School, Hakim Sabzevari University, Sabzevar, Iran

[c] Department of Chemical Engineering, Marv.C., Islamic Azad University, Marvdasht, Iran

[d] Center for Scientific Research and Entrepreneurship, Northern Border University, 73213, Arar, Saudi Arabia

pressures whereas the influence of temperature is more nuanced. At higher temperatures, reduced solvent density tends to diminish solvation capability, while simultaneously increased vapor pressure of the solute enhances its tendency to dissolve. The interplay of these competing effects often results in the well-documented crossover phenomenon observed in solubility isotherms for many dye–CO_2 systems.[19] Moreover, for polar or ionic dyes, affinity for CO_2 is often very low unless a co-solvent or modifier is employed, since CO_2 is a weakly polar, Lewis-acidic solvent and has limited capacity to stabilize strong dipoles or ionic species.

In this context, generating reliable solubility data for specific dye molecules in scCO$_2$ is not merely supplementary, but rather a critical requirement for the rational design and optimization of industrial processes, including extraction, dyeing, encapsulation, and purification. Although several studies have reported solubility values for selected dye families, existing datasets remain fragmented and inconsistent, with significant variability across experimental conditions, methodologies, and modeling approaches. This issue is particularly pronounced for polar and ionic dyes, where the inherently low affinity toward non-polar scCO$_2$ leads to limited solubility data and, in some cases, contradictory findings. Therefore, beyond expanding the thermodynamic database, a clear and justified rationale for the selection of the target dye is essential one that explains its industrial relevance, physicochemical challenges, and the potential scientific or technological impact of obtaining accurate solubility measurements in scCO$_2$ systems.[20]

Much research has been conducted on collecting laboratory information on the solubility of materials in scCO$_2$.[21–28] While experimental specification of material solubility in scCO$_2$ provides vital data for process design, it is often pricey, complex, and sometimes impractical under diverse conditions of temperature and pressure. To address these challenges, researchers have developed various simulation models, including semi-empirical density-based models (Mendez-Santiago & Teja (MST), Chrastil, Bartle *et al.*, Kumar & Johnston (K-J), and Alwi-Garlapati), thermodynamic models, and

equations of state (EoSs), which allow for more rapid, cost-effective, and flexible prediction of material solubility.[29–36]

Thermodynamic models, equation of state approaches, and empirical correlations have long been used to predict material solubility in scCO$_2$, but they come with notable limitations. These models often rely on simplifying assumptions and idealizations that can compromise accuracy, especially when applied to complex or structurally diverse compounds. Empirical correlations, while simpler to apply, are typically system-specific and struggle to generalize across different datasets.[37–41] Moreover, many of these traditional models require detailed knowledge of system parameters and involve computationally intensive, iterative calculations, making them less practical for large-scale applications. In contrast, machine learning models can directly learn complex, nonlinear relationships from data without relying on predefined physical equations. This allows them to achieve higher predictive accuracy and better generalization across a wide range of drug-solvent systems. Machine learning approaches also offer faster predictions, greater flexibility in handling large and heterogeneous datasets, and the ability to incorporate critical drug properties as input features, further enhancing their predictive capabilities.[42–45]

In this investigation, the solubility of yellow 23 in scCO$_2$ has been calculated and modeled at several temperatures and pressures. The present study investigated four density-based models (MST, Chrastil, Bartle *et al.*, and K-J) and three machine learning models (Multilayer Perceptron, Gaussian process regression, and Random Forest).

2 Theory and methodology

2.1. Materials

The chemical reagents used in this work included yellow 23 (CAS 604-75-1, Merck). High-purity carbon dioxide (CO_2, 99.99%, CAS 124-38-9) was supplied by Aboughadareh Co. (Shiraz, Iran). The specifications of all materials are summarized in Table 1.

Table 1 The specification of yellow 23

| Compound	Formula	Structure	M_W	CAS number	λ_{max} (nm)
Yellow 23	C$_{16}$H$_9$N$_4$Na$_3$O$_9$S$_2$		534.357	1934-21-0	427
Carbon dioxide	CO_2		44.01	124-38-9	
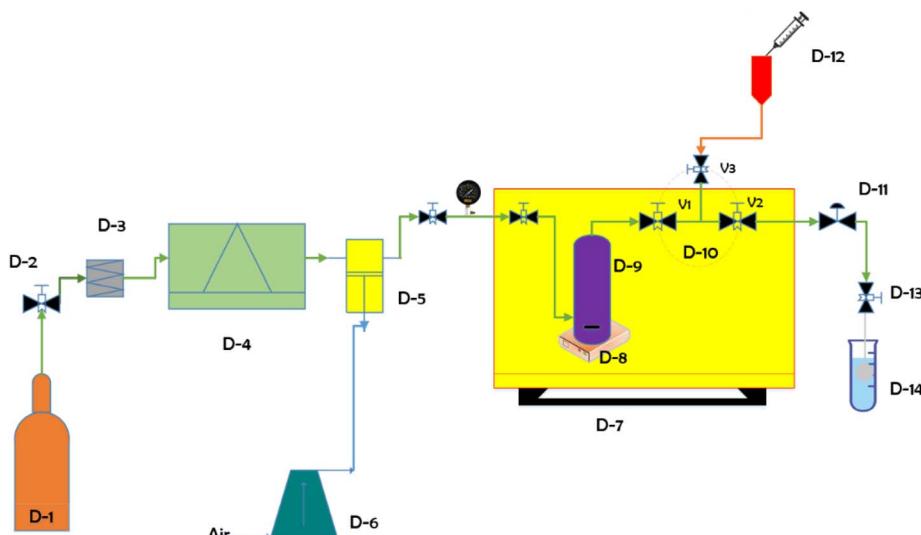
Fig. 1 Test tools diagrammatic.

2.2. Solubility measurement

The experimental setup used to determine the solubility of yellow 23 in supercritical carbon dioxide (as illustrated in Fig. 1) includes a UV-Vis spectrophotometer, a carbon dioxide supply tank, an air-driven compressor (Finac, China), and a high-pressure CO_2 pump (specifically, an air-driven liquid pump, model M64 from Shine East). Additional components comprise a refrigeration unit, a magnetic stirrer operating at 250 rpm, a filter, various flow control valves (including needle, back-pressure, and metering valves), an equilibrium high-pressure cell, and an oven. By passing through a molecular filter, the carbon dioxide that was introduced from the tank was free of any possible contaminants. Before entering the high-pressure pump, the carbon dioxide flow was liquefied by cooling into a frig operating at −10 °C. The binary (yellow 23 and scCO$_2$) systems were homogenized using a magnetic stirrer.

1000 mg of yellow 23 was placed to the cell (300 mL) for each experiment. The solution in scCO$_2$ was uniformed using a stirrer. The temperature was controlled using an oven. The system was subsequently compressed to the operating pressure. Prior to being delivered to the cell, the carbon dioxide was compressed to the appropriate pressure. According to the prior test, the fixed time was 240 minutes. A 3-valve system moves scCO$_2$ from the balance cell to a 300 μL pattern loop once equilibrium has been reached. A micrometer valve is used to depressurize the sample loop in order to stop liquid ethanol from ejecting. Ethanol is injected into the sampling loop using a syringe. 5 mL of solution are contained in the collection container. A spectrophotometer (PerkinElmer) equipped with quartz chamber and a 3 cm path length is used to measure the solubility of yellow 23 under different operating conditions. Using calibration curves, the concentration of yellow 23 in the solution is determined, and the highest wavelength UV absorption analysis is used to measure it in the collecting vial.

The equilibrium solubility of yellow 23 in supercritical carbon dioxide is determined by the following equations, which are explained in more detail in previous articles:[21,23,24,39,41,46,47]

$$y = \frac{n_{solute}}{n_{solute} + n_{CO_2}} \tag{1}$$

Here:

$$n_{solute} = \frac{C_s(g\ L^{-1}) \times V_s\ (L)}{M_s(g\ mol^{-1})} \tag{2}$$

$$n_{CO_2} = \frac{V_1(L) \times \rho(g\ L^{-1})}{M_{CO_2}(g\ mol^{-1})} \tag{3}$$

V_s (L) and V_1 (L) show the volumes of the collection vial and sample loop, M_s (g mol^{-1}) is the solute's molar mass and M_{CO_2} (g mol^{-1}) is the molar mass of carbon dioxide, both in units of g mol^{-1}, respectively, and n_{solute} and n_{CO_2} indicate the number of moles of the yellow 23 and carbon dioxide in the sample loop. C_s (g L^{-1}) indicates the yellow 23 concentration in the collection vial, as determined by the calibration curve. When eqn (2) and (3) are substituted in eqn (1), the result is:

$$y = \frac{C_s(g\ L^{-1}) \times V_s\ (L) \times M_{CO_2}(g\ mol^{-1})}{C_s(g\ L^{-1}) \times V_s\ (L) \times M_{CO_2}(g\ mol^{-1}) + V_1(L) \times \rho(g\ L^{-1}) \times M_s(g\ mol^{-1})} \tag{4}$$

Eqn (5) displays the yellow 23's equilibrium solubility S $(\mathrm{g\,L^{-1}})$ in scCO$_2$.

$$S(\mathrm{g\ L^{-1}}) = \frac{C_{\mathrm{s}}(\mathrm{g\ L^{-1}}) \times V_{\mathrm{s}}\,(\mathrm{L})}{V_{\mathrm{l}}(\mathrm{L})} \qquad (5)$$

2.3. Empirical (density-based) models

In this academic work, empirical models that depend on density to find a correlation between the solubility data of yellow 23 in super critical carbon dioxide gained through experimentation. In addition to SCF density, temperature, and pressure, these models also include constants and movable parameters, which stand in for independent variables. These models have the benefit of not requiring the estimation of the solid's properties. In particular, Chrastil, MST, Bartle *et al.*, and K-J were the density models chosen for this investigation. These models were selected due to their proven ability to forestall material solubility in scCO$_2$. The empirical models sequential were found by applying regression analysis to the experimental data. MATLAB's simulation annealing was utilized to obtain the adjustable parameters. Table 2 list the empirical models used to correlate the yellow 23.

Two trustworthy statistical measures, AARD% and correlation coefficient (R_{adj}), were used in this investigation to evaluate how well the thermodynamic models used to correlate the solubility of yellow 23 in scCO$_2$. In eqn (6), the molar solubility of yellow 23 in scCO$_2$ is represented by the symbol y^{exp}. The theoretical solubility derived from the suggested thermodynamic models are represented by the symbol y^{cal}.

$$\mathrm{AARD}\,\% = \frac{100}{N_t - Z} \sum_{i=1}^{N_t} \frac{|y^{\mathrm{cal}} - y^{\mathrm{exp}}|}{y^{\mathrm{exp}}} \qquad (6)$$

R_{adj} was calculated using eqn (7). In eqn (7), Q is the number of independent changeable, and N is the number of report points for each set.

$$R_{\mathrm{adj}} = \sqrt{\left| R^2 - \left(Q(1 - R^2)/(N - Q - 1) \right) \right|} \qquad (7)$$

2.4. Machine learning models

2.4.1. Multilayer perceptron (MLP). The Multilayer Perceptron (MLP) is a feedforward artificial neural network used for regression and classification tasks.[52] It consists of an input layer, one or more hidden layers, and an output layer, with fully connected neurons that apply weighted sums and nonlinear activation functions to capture complex patterns.[53] The model is trained using forward propagation, error computation (*e.g.*, MSE), and backpropagation to iteratively adjust weights and biases, enabling it to approximate highly complex functions.[54,55]

2.4.2. Gaussian process regression (GPR). Gaussian Process Regression (GPR) is a non-parametric, Bayesian regression technique that models the relationship between inputs and outputs as a distribution over functions.[56] It uses a mean function to estimate central trends and a covariance (kernel) function to capture similarities between input points, allowing it to model complex, non-linear patterns.[57]

GPR provides not only predictions but also uncertainty estimates, adapts flexibly to data without assuming a fixed form, and is effective for small or noisy datasets. Kernel parameters are optimized during training, and although computationally intensive for large datasets, GPR is interpretable, robust, and less prone to overfitting.[58,59]

2.4.3. Random forest (RF). Random Forest (RF) is a robust supervised learning method based on ensemble learning, designed to improve prediction accuracy and reduce overfitting.[60] It builds multiple decision trees (CART) using bootstrapped data samples and random subsets of features, with final predictions obtained by averaging (regression) or majority voting (classification).[61,62] This randomness enhances diversity, lowers variance, and improves generalization. RF effectively handles high-dimensional, heterogeneous, and multicollinear data, capturing complex non-linear relationships. Key hyperparameters, like the number of trees and tree depth, can be tuned for optimal performance, and its independent tree construction allows parallel processing, making it scalable for large datasets.[63,64]

2.4.4. Predictive analytics. In this study, 80% of the dataset was randomly selected for training, while the remaining 20% was used for testing, and the same split was applied across all models to ensure a fair and consistent comparison. Prior to model development, all input variables were normalized using the StandardScaler, which transforms each feature to have zero mean and unit variance. This normalization step was essential given the small dataset size, as it prevents features with larger numerical ranges from dominating the learning process and improves model stability during training. To further reduce the risk of overfitting, especially given the limited data volume, relatively simple model structures with a small number of hyperparameters were chosen. Hyperparameter tuning was performed using GridSearchCV combined with 5-fold cross-

Table 2 The empirical models used in this work

Models	Formula
Chrastil[48]	$\ln\,S = a_0 + a_1\,\ln(\rho_1) + \dfrac{a_2}{T}$
Méndez - Santiago & Teja (MST)[49]	$T\ln(y_2 P) = a_0 + a_1\rho_1 + a_2 T$
Bartle *et al.*[50]	$\ln\left(\dfrac{y_2 P}{P_{\mathrm{ref}}}\right) = a_0 + a_1(\rho_1 - \rho_{\mathrm{ref}}) + \dfrac{a_2}{T}$
Kumar-Johnston (K-J)[51]	$\ln\,y_2 = a_0 + a_1\rho_1 + \dfrac{a_2}{T}$

Table 3 Hyperparameters of developed models

Models	Parameters	Best value
GPR	Alpha	0.1
	Kernel	$1^2 \times \mathrm{RBF}(1)$
RF	N_estimators	150
	Max_depth	None
	Min_samples_split	2
MLP	Activation	tan$h$
	Alpha	0.005
	Hidden_layer_sizes	(4)
	Learning_rate_init	0.001
	Solver	lbfgs

validation, ensuring that every data point contributed to both training and validation across the folds. This approach prevents the model from being tailored to a single train–test split and provides a more reliable estimate of generalization performance. The Mean Squared Error (MSE) was used as the objective function during hyperparameter search to identify the optimal configuration for each model. This systematic procedure consistent data splitting, feature normalization, controlled model complexity, and rigorous cross-validation helped prevent overfitting and ensured robust and unbiased model comparison. The final optimized hyperparameters and corresponding performance metrics are summarized in Table 3.

2.4.5. Statistical error analysis. In this study, we used GPR, RF, and MLP models to predict the solubility of yellow 23 in $scCO_2$. The models' accuracy was assessed by comparing the predicted drug solubility in $scCO_2$ (y_{pred}) with the corresponding experimental values (y_{exp}). To comprehensively evaluate model performance, several statistical error analyses were conducted, as detailed in the following sections. The output variable (mole fraction) was scaled as $y \times 10^6$; thus, MAE and SD are expressed in units of (mole fraction $\times 10^6$), while MSE is reported in (mole fraction)$^2 \times 10^{12}$. R^2 is dimensionless.

Mean Square Error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_{i,\text{pred}} - y_{i,\text{exp}} \right)^2 \qquad (8)$$

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_{i,\text{pred}} - y_{i,\text{exp}} \right| \qquad (9)$$

Standard Deviation (SD)

$$\text{SD} = \sqrt{\frac{\sum\limits_{i=1}^{n} \dfrac{\left(y_{i,\text{exp}} - y_{i,\text{pred}} \right)^2}{y_{i,\text{exp}}}}{n-1}} \qquad (10)$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n} \left(y_{i,\text{exp}} - y_{i,\text{pred}} \right)^2}{\sum\limits_{i=1}^{n} \left(y_{i,\text{exp}} - \overline{y}_{i,\text{exp}} \right)^2} \qquad (11)$$

3 Results and discussion

3.1. Solubility of yellow 23 in $scCO_2$

Specific details about yellow 23's solubility in $scCO_2$ are given in Table 4. To enhance the reliability of the experimental data, each experiment was replicated three times, and the mean values were reported. The solubility data for yellow 23 are

Table 4 The experimental data of yellow 23 in $scCO_2$ at various conditions of temperature and pressure[a]

Temperature (K)	Pressure (MPa)	Density (kg m^{-3})	$y \times 10^6$ (mole fraction)	Standard deviation of the mean, SD ($\bar{y}$) $\times$ (10^6)	Expanded uncertainty $\times 10^7$	$S \times 10$ (solubility (g per l))
313	12	719	0.667	0.123	0.281	0.0623
	15	781	0.848	0.192	0.421	0.0842
	18	820	0.974	0.263	0.562	0.1005
	21	850	1.189	0.375	0.789	0.1264
	24	873	1.563	0.563	1.170	0.1701
	27	893	1.721	0.697	1.438	0.1910
	30	910	2.055	0.925	1.897	0.2320
323	12	587	0.432	0.080	0.182	0.0347
	15	701	0.826	0.186	0.408	0.0747
	18	758	1.198	0.323	0.692	0.1151
	21	797	1.544	0.486	1.023	0.1546
	24	826	1.934	0.696	1.447	0.1998
	27	851	2.148	0.870	1.794	0.2276
	30	871	2.358	1.061	2.176	0.2551
333	12	435	0.341	0.064	0.145	0.0211
	15	606	0.746	0.170	0.372	0.0595
	18	688	1.328	0.359	0.767	0.1169
	21	740	1.686	0.538	1.130	0.1575
	24	777	2.116	0.762	1.584	0.2062
	27	806	2.429	0.984	2.029	0.2445
	30	830	2.737	1.232	2.526	0.2828
343	12	346.1	0.229	0.047	0.105	0.0108
	15	507.5	0.574	0.139	0.302	0.0389
	18	613.5	1.426	0.388	0.829	0.1130
	21	678.9	1.978	0.623	1.310	0.1706
	24	724.9	2.334	0.840	1.747	0.2130
	27	760.3	2.632	1.066	2.199	0.2504
	30	788.9	3.084	1.388	2.846	0.3032

[a] Standard uncertainty u are $(T) = 0.1$ K; $u(p) = 0.1$ MPa.

Table 5 The correlation results of the yellow 23 scCO$_2$ system presented by the semi empirical models

Model	a_0	a_1	a_2	AARD%	R_{adj}	AICc
Chrastil	4.95	−4396.3	−23.66	9.81	0.986	−652
Bartle *et al.*	12.61	0.00884	−6953.01	10.71	0.961	−631
MST	−8824.2	2.90	12.44	9.88	0.983	−649
K-J	−3.80	0.2648	−4645.9	6.39	0.988	−694

presented in Table 5. In addition, to validate the solubility measurements, we measured the solubility of capecitabine and naphthalene at different temperatures and pressures. Then, we compared these measurements with existing data. The experimental data for capecitabine and naphthalene in supercritical carbon dioxide were reported in our previous article.[65] The density of carbon dioxide, mole fraction, and solubility of yellow 23 in scCO$_2$ were measured in triplicate at pressures between 12 and 30 MPa and temperatures between 313 and 343 K. The outcome of scCO$_2$ density and pressure on yellow 23 solubility is depicted in Fig. 2. Solubility of yellow 23 in scCO$_2$, as shown in Table 4 and Fig. 2, showed a rising tendency as pressure increased. This is explained by the solvent density increasing in tandem with the pressure increase.

The crossover pressure observed at approximately 18 MPa (Fig. 2) represents a transition in the dominant thermodynamic mechanism controlling the solubility of yellow 23 in scCO$_2$. Below this pressure, the solubility decreases with increasing temperature, indicating an enthalpy-dominated regime. In this region, the reduction in CO$_2$ density with temperature results in a weaker solvation environment and higher Gibbs free energy of dissolution. Consequently, the dissolution process becomes energetically unfavorable, leading to a negative temperature solubility relationship. Similar behavior has been reported for other polar or structurally rigid dye molecules in scCO$_2$ systems. Above the crossover pressure, a different behavior emerges: the solubility increases with temperature, demonstrating an entropy-driven regime. At higher pressures, the solvation power of CO$_2$ remains sufficiently high despite decreases in density, and the dominant driving force becomes the increased vapor (or sublimation) pressure of the dye with temperature. As temperature rises, the solid–fluid phase equilibrium shifts toward dissolution, lowering the energetic barrier for phase transition and enhancing molecular mobility and mixing entropy. This competing interplay between solvent density-controlled enthalpic effects and solute vapor-pressure-driven entropic effects is widely recognized as the governing mechanism behind crossover behavior in supercritical dissolution systems.

Several studies support this mechanistic interpretation. Kalikin *et al.* demonstrated that crossover points are not fixed but shift depending on temperature and pressure, highlighting the dynamic balance between enthalpic and entropic contributions in supercritical systems.[66] De Melo *et al.* analyzed the effect of solute properties such as molar volume, sublimation pressure, and crystal lattice energy on crossover pressure, confirming that these factors determine the balance of enthalpic penalties and entropic gains.[67] Sodeifian *et al.* further demonstrated similar trends in pharmaceutical compounds, linking solubility changes to Gibbs free energy, enthalpy, and entropy contributions.[68]

The exact position of the crossover point depends on thermophysical properties of the solute including molar volume, sublimation pressure, crystal lattice strength, and enthalpy of sublimation as well as solvent compressibility near the operating temperature. These factors collectively determine the balance between enthalpic penalties associated with cavity formation in scCO$_2$ and the entropic gains from disorder and molecular dispersion upon dissolution. For yellow 23, the measured mole fraction ranged between 2.29×10^{-6} and 3.084×10^{-5}, with the lowest solubility observed at 12 MPa and 343 K, consistent with a low-pressure, enthalpically dominated regime. The observed crossover behavior confirms that despite its polarity and relatively low affinity for scCO$_2$, yellow 23 exhibits dissolution characteristics aligned with established thermodynamic models for solid solutes in supercritical media.
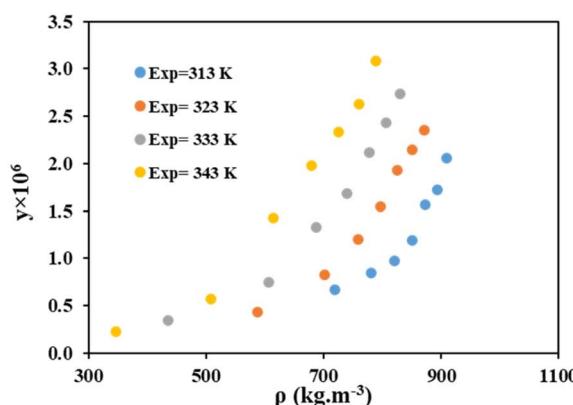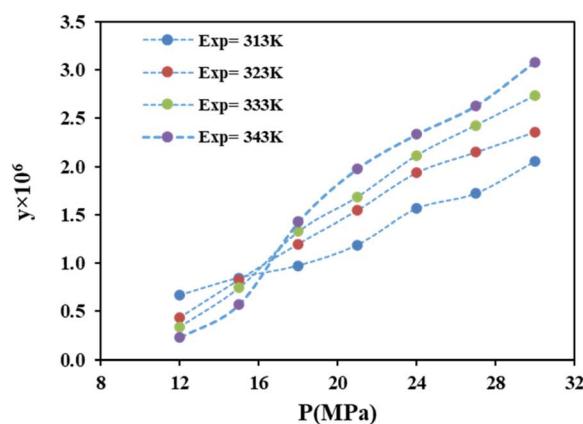
Fig. 2 The impression of (a) pressure and (b) density of the scCO$_2$ on yellow 23 solubility at various temperatures.

3.2. Empirical models

This work employed four semi-empirical correlations to simulate the solubility of yellow 23 in scCO$_2$ inside a binary system. The models have three sets of parameters. The equations employed in this investigation are summarized in Table 5, and the corresponding AARD%, corrected correlation coefficient (R_{adj}), and parameters (a_0, a_1, a_2) are given in Table 5. Chrastil, Bartle et al., MST, and K-J had average absolute relative deviations of 9.8126%, 10.7169%, 9.8865%, and 6.3902%, respectively. The R_{adj} were 98.65%, 96.08%, 98.34%, and 98.81% for Chrastil, Bartle et al., MST, and K-J, respectively.

The parameters of the four empirical correlations in Table 5 provide meaningful insight into the mechanisms that govern the interaction between yellow 23 and scCO$_2$. In the Chrastil model, the positive value of parameter k reflects the effective association number between dye molecules and CO$_2$. Its moderate magnitude indicates weak, yet non-negligible, solute–solvent clustering, which is consistent with yellow 23's limited polarity compatibility. The negative intercept parameter corresponds to an endothermic dissolution process overall at low

pressures, which aligns with the observed crossover behavior. In the MST model, the negative slope of the vs. density relationship indicates that increasing CO$_2$ density enhances solubility, which is a thermodynamically expected effect linked to stronger solvent–solute interactions. The Bartle parameters have small magnitudes, which reflect the empirical model's limited sensitivity to molecular interactions for highly polar solutes in nonpolar supercritical carbon dioxide. For the K-J equation, the positive, density-dependent parameter confirms that solubility increases with solvent density. The negative constant term suggests a significant energetic barrier to dissolution at low density. Overall, the sign and magnitude of the extracted parameters indicate that yellow 23 experiences weak solvation in scCO$_2$, with solubility primarily dominated by density-driven packing effects rather than strong specific interactions.

Fig. 3 shows the outcomes of the models based on the aforementioned semi-empirical correlations. The MST, Chrastil, Bartle et al., and K-J models were also used to analyze the experimental data's self-consistency tests. The experimental findings of $T \times (\ln(y \times P) - a_2)$ were plotted against
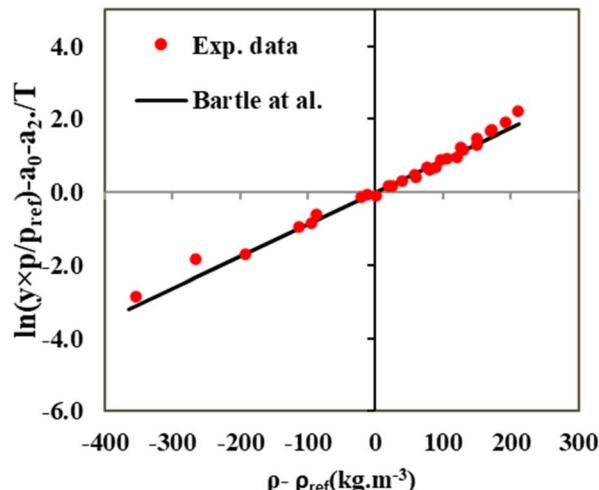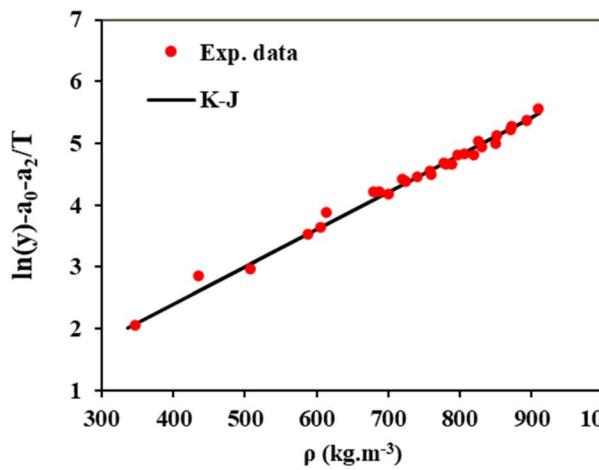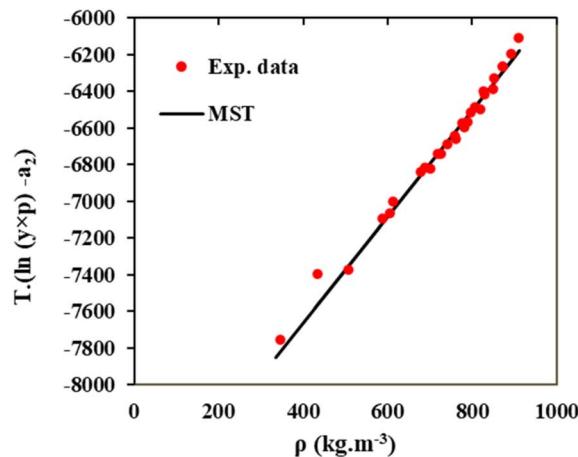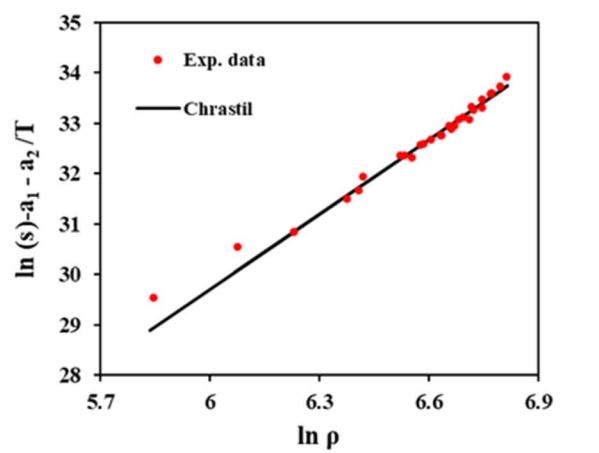

Fig. 3 Self-consistency of the test information of yellow 23 with different empirical models, (a) Chrastil, (b) MST, (c) K-J, and (d) Bartle et al.

the density of $scCO_2$ for the self-consistency experiment from the MST model, and Fig. 3 should show a linear curve. Fig. 3 shows the results of the self-consistency test employing all

models. The measured solid solubility inputs are consistent across all test situation, as demonstrated by the linear behavior in Fig. 3.

Based on this agreement between the experimental results and model predictions, the fitted correlations may be considered reliable for interpolation within the studied pressure and temperature domain. However, the present results do not demonstrate the capability of these models to accurately perform extrapolation beyond the experimental boundaries. Reliable extrapolation requires external validation or the extension of the experimental dataset over a broader thermodynamic window.

This interpretation is consistent with previous studies, where density-based solubility models showed strong interpolation accuracy but limited predictive reliability when applied outside the calibrated experimental range. Similar observations have been reported for disperse dyes, pharmaceuticals, and organic solutes in supercritical CO_2 systems, highlighting that the empirical nature of such correlations constrains their

Table 6 Analytical review of the model outputs using statistical indicators

Models		Statistical parameters			
		MSE	MAE	SD	R^2
GPR	Train	0.006	0.060	0.108	0.988
	Test	0.006	0.050	**0.133**	0.989
	Total	0.006	0.058	0.111	0.989
RF	Train	0.007	0.071	0.078	0.987
	Test	0.016	0.102	0.443	0.971
	Total	0.009	0.078	0.202	0.984
MLP	Train	**0.001**	0.033	**0.036**	**0.996**
	Test	**0.001**	0.028	0.135	**0.997**
	Total	**0.001**	0.032	**0.066**	**0.997**
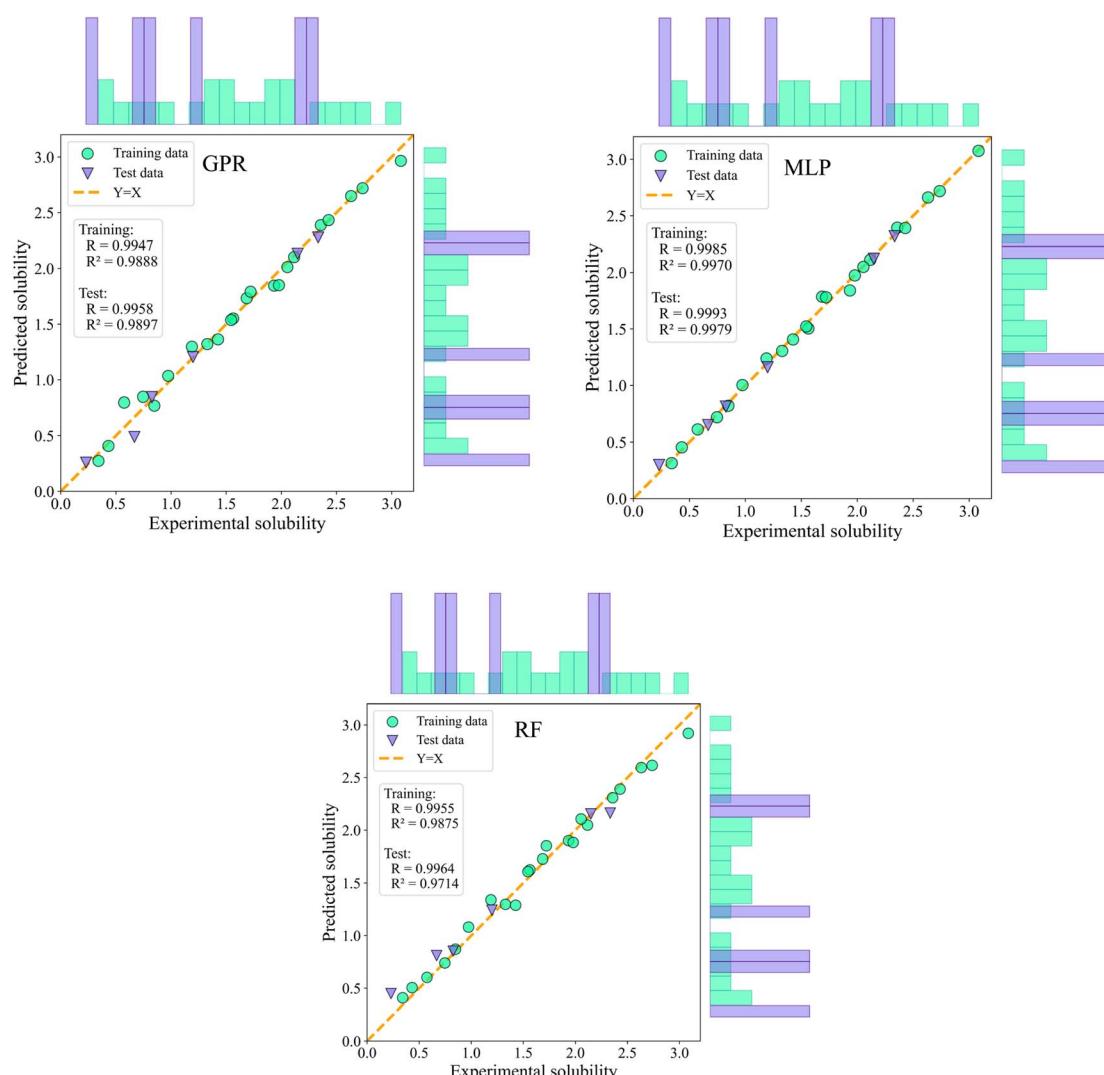
Fig. 4 Predicted solubility values *versus* experimental solubility values for machine learning models.

extrapolation performance unless thermodynamic constraints or independent datasets are incorporated for verification.[66–68]

3.3. Machine learning models

Table 6 presents a detailed statistical summary of the GPR, RF, and MLP models across the training, testing, and overall datasets. Among these models, the MLP exhibits the highest performance in all datasets, achieving the lowest MSE and MAE values and the highest R^2 score (0.997). The minimal difference between the training, testing, and total errors indicates that the MLP model does not suffer from overfitting and is capable of generalizing beyond the training data. This strong generalization ability, combined with its high accuracy on unseen data, underscores the robustness and effectiveness of the MLP model for predicting solubility behavior, even outside the original training range. The superior performance of the MLP can be attributed to its deep learning architecture, which comprises multiple layers of interconnected neurons capable of learning complex and highly nonlinear relationships within the data. This enables the MLP to effectively capture intricate interactions among input variables such as temperature, pressure, and density factors that are critical for accurate solubility prediction. The GPR model also demonstrates strong performance, with consistent results across the training and testing sets, reflecting its robustness and stability, particularly in handling small datasets. In contrast, the RF model, while performing well on

the training set ($R^2 = 0.987$), shows a noticeable decline in performance on the test set ($R^2 = 0.971$) along with the highest standard deviation (0.443), suggesting a tendency toward overfitting and reduced reliability when applied to unseen data.

3.3.1. Graphical error analysis. Alongside statistical analysis, graphical error assessment serves as a powerful tool for evaluating model performance. It is particularly beneficial for comparing multiple models. This study employed various graphical techniques to highlight the effectiveness of the developed model.

Cross-plots offer a visual comparison between predicted (Pred) and experimental (Exp) values, with the 45-degree diagonal line representing perfect prediction. The predictive accuracy of a model is reflected by how closely the data points align with this reference line the closer the points lie to the line, the more accurate the model's predictions. Fig. 4 presents the cross-plots for the RF, GPR, and MLP models. As shown, the MLP model exhibits a powerful adjustment between the correlated and experimental solubility data in both the training and test sets. This highlights the high correctness and generalization capability of the MLP model in forecasting the solubility of yellow 23.

The error distribution plot shows the residuals between predicted and experimental values, allowing assessment of model accuracy across different solubility values and identification of potential biases or systematic deviations. Data points clustered around the $Y = 0$ line indicate low prediction errors
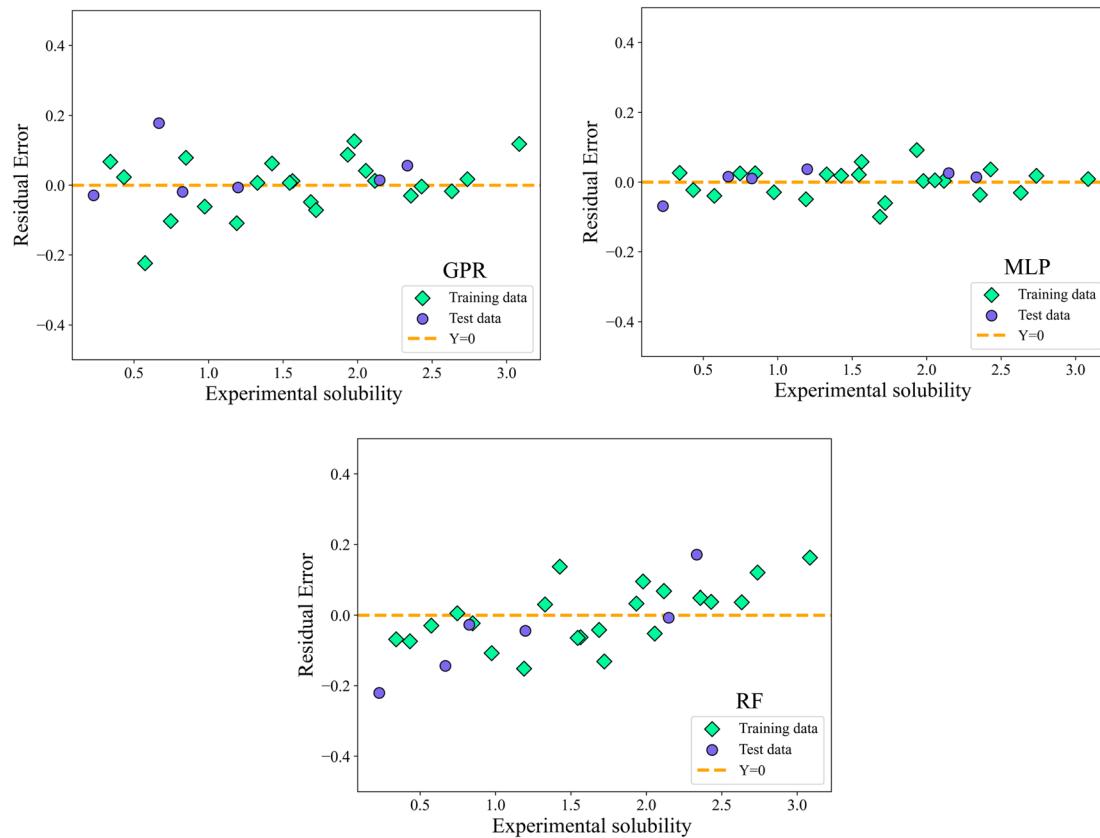

**Fig. 5** Residual error distribution plots for machine learning models.
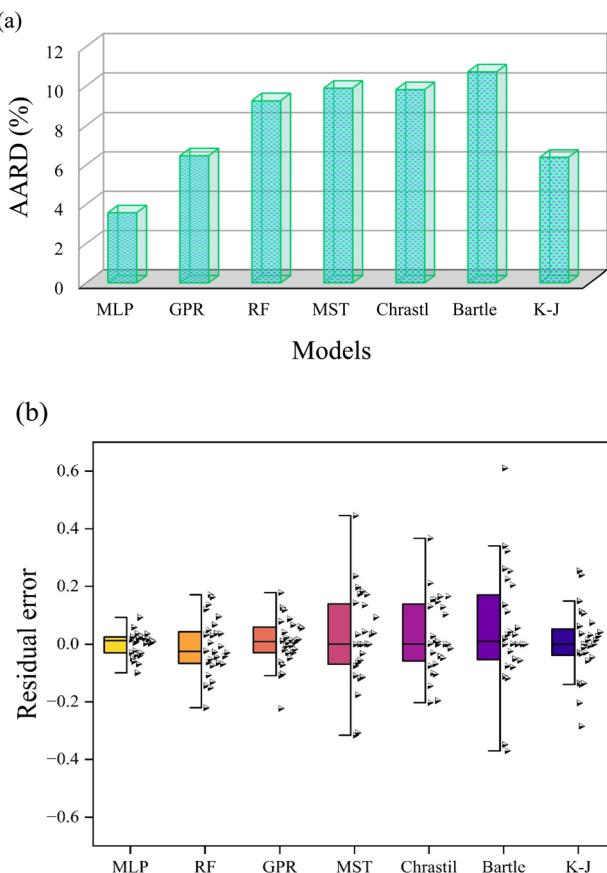
Fig. 6 Comparison of model performance based on (a) (AARD%) and (b) residual error distributions for all models evaluated in this study.

and higher model accuracy. Fig. 5 presents the error distribution curves for the machine learning models predicting the solubility of yellow 23 in $scCO_2$. The MLP model exhibits the highest concentration of points near $Y = 0$, confirming its superior accuracy and reliability.

Fig. 6a illustrates the Average Absolute Relative Deviation (AARD) values corresponding to all the predictive models assessed in this investigation. Notably, the MLP model exhibits the lowest AARD % value of 3.572, underscoring its superior capability in capturing the underlying data patterns with minimal relative error. Fig. 6b presents the residual error distributions, where the MLP model again outperforms its counterparts, demonstrating the most compact error range, confined between −0.099 and 0.092. This narrow dispersion of residuals highlights the model's precision and stability in capturing the underlying patterns of the solubility data. Collectively, these graphical results corroborate the statistical outcomes detailed in Table 6, affirming the MLP model's robustness, generalization capacity, and exceptional reliability in accurately predicting the solubility of yellow 23.

3.3.2. Model trend analysis. Trend analysis serves as an effective approach to evaluate the responsiveness of solubility to variations in key input parameters. Fig. 7a illustrates the solubility profile of yellow 23 in $scCO_2$ *versus* temperature and density. The results clearly demonstrate that solubility

increases with both rising temperature and density a trend accurately captured by all models. However, the MLP model exhibits superior consistency and predictive precision across the entire temperature and density spectrum, whereas the other models show limited predictive capacity, often performing well only within specific regions of the parameter space.

Similarly, Fig. 7b depicts the solubility behavior of yellow 23 in $scCO_2$ *versus* temperature and pressure. As depicted, solubility increases with rising temperature and pressure an expected thermodynamic trend that is effectively captured by all the models. However, the MLP model demonstrates notably higher predictive accuracy across the entire temperature and pressure domain. In contrast, the performance of the other models appears constrained, with reliable predictions limited to narrower, specific regions of the temperature pressure space. This observation highlights the superior generalization capability of the MLP model in modeling complex, nonlinear solubility responses under varying thermodynamic conditions.

3.3.3. Sensitivity analysis. Fig. 8 displays the SHAP (SHapley Additive exPlanations) summary plots that illustrate the contribution of each input variable to the MLP model's predictions of yellow 23 solubility in $scCO_2$. Fig. 8a ranks the features based on their mean absolute SHAP values, reflecting their average overall influence on the model's output regardless of whether the influence is positive or negative. A higher mean SHAP value corresponds to a stronger effect on the model's prediction. Fig. 8b further provides a detailed distribution of SHAP values for all data points, with feature magnitudes color-coded from green (low values) to turquoise (high values), enabling an intuitive interpretation of how changes in temperature, pressure, and density affect individual solubility predictions. Among the three input parameters, density exhibits the highest impact on the model's predictions, as indicated by its dominant SHAP value in both plots. This finding is fully consistent with density-based thermodynamic correlations, where solubility in supercritical fluids is strongly governed by the solvent's density. Higher density corresponds to closer molecular packing and stronger intermolecular interactions, thereby increasing the solvating power of $scCO_2$. Consequently, the SHAP dominance of density is not only statistically supported by the model but also aligned with established physical principles. Additionally, the SHAP distributions show that temperature, pressure, and density exert overall positive influences on solubility, which agrees with the expected behavior in supercritical systems: increases in temperature or pressure often lead to changes in fluid density, which in turn enhance solubility. Furthermore, the SHAP findings reinforce the physical expectation that solubility varies more sensitively with density than with temperature or pressure alone, because density acts as an integrated state variable capturing the combined effects of both parameters on fluid structure. The consistency between the SHAP ranking and classical density-based solubility correlations provides strong confidence that the model is learning meaningful, physically interpretable relationships rather than relying on spurious statistical patterns.[69]
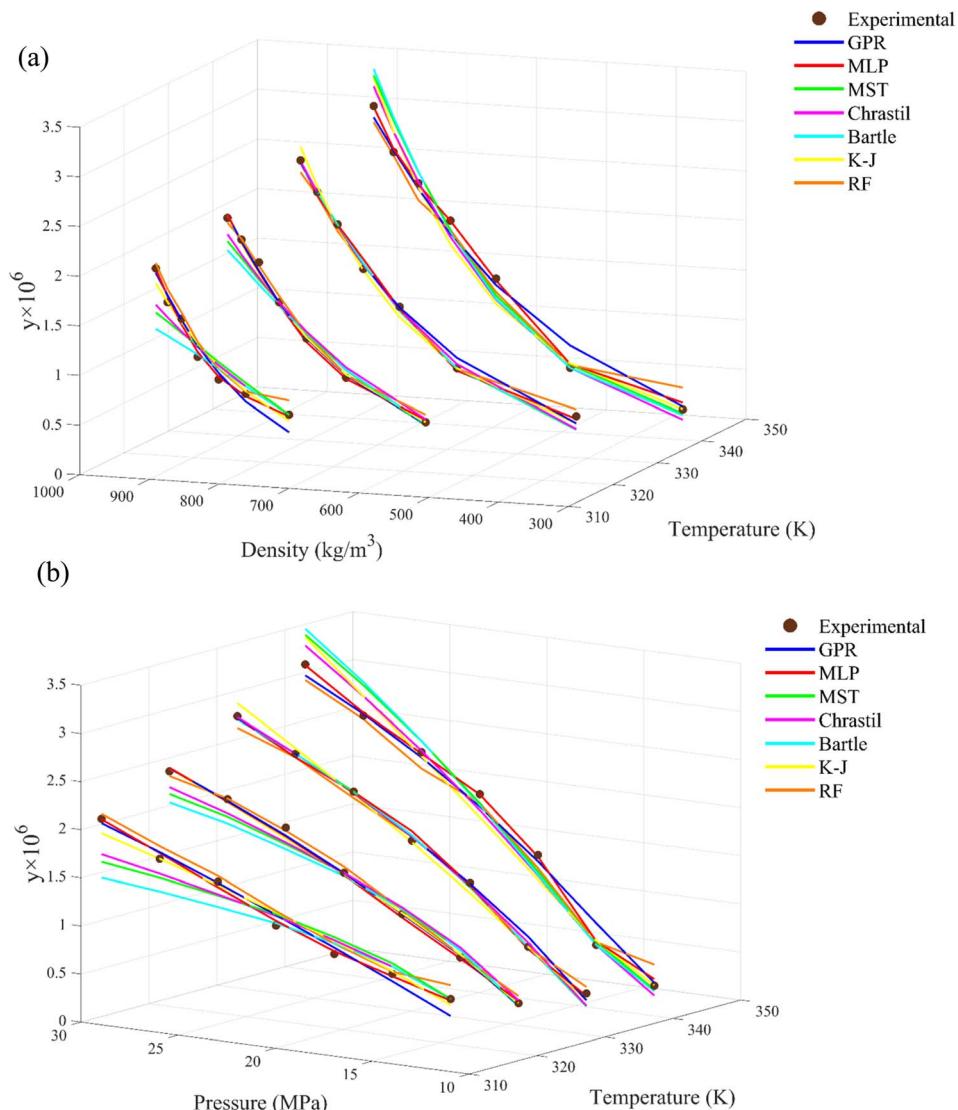
Fig. 7 Comparison between experimental and predicted solubility values of yellow 23 in scCO$_2$: (a) *versus* temperature and density. (b) *Versus* temperature and pressure.
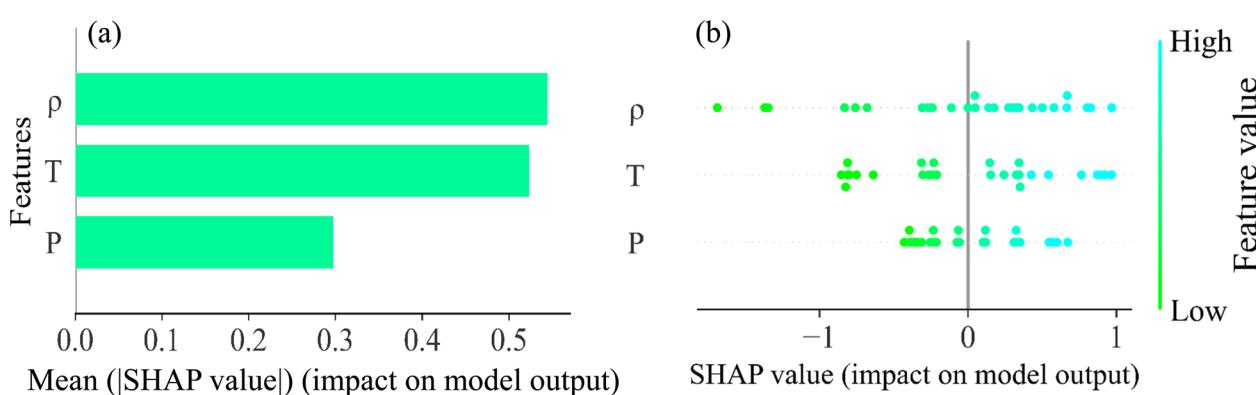

Fig. 8 SHAP summary plots illustrating feature contributions to MLP model predictions of yellow 23 solubility in scCO$_2$.

4 Conclusion

The solubility of yellow 23 in supercritical CO_2 was recorded at several temperatures and pressures. The experimental data were fitted with four density-based semi-empirical models (Chrastil, MST, Bartle et al., and K-J). Among them, the K-J model showed the best agreement with the experimental results (AARD = 6.39%). Simultaneously, three machine-learning strategies (MLP, GPR, RF) were implemented for solubility prediction, and the MLP model was found to be the most accurate (MSE = 0.001, R^2 = 0.997) with good generalization over the entire dataset. These findings can be used to improve the efficiency of supercritical-fluid processes with dyes, such as dye extraction from wastewater, the production of solvent-free coloration systems, and the fabrication of dye-loaded nanoparticles where solubility determines particle formation pathways. Accurate solubility prediction enables identifying the optimal operating windows that maximize CO^2 solvent strength while minimizing experimental efforts. The sensitivity analysis revealed that density and temperature have almost the same impact on the MLP model's predictions as their SHAP values differ only slightly. This suggests that the model does not rely on one variable to achieve the best performance but rather captures the combined nonlinear effects of variables. The study is limited to a single dye under a restricted thermodynamic range. The dataset of dyes can be diversified and the conditions can be extended to improve model generalizability further.

Author contributions

S. A. S.: investigation, modeling, conceptualization, project administration, and review and editing. A. H. S.: writing original draft, investigation, modeling, software, formal analysis. N. E.: investigation, validation, methodology, and review and editing. A. N.: methodology, validation, and investigation, review and editing.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

Acknowledgements

References

1 L. Amaral De Faria Silva, M. Ferreira Alves, D. Florêncio Filho, J. Aparecida Takahashi, L. Soares Santos and S. Almeida De Carvalho, *Food Chem.*, 2022, **389**, 132967.

2 P. Barciela, A. Perez-Vazquez and M. A. Prieto, *Food Chem. Toxicol.*, 2023, **178**, 113935.

3 S. Albazi, E. T. Al-Samarrai and L. H. Alwan, *Green Anal. Chem.*, 2024, **11**, 100150.

4 M. Gonçalves, S. Brito, C. Song, Y. Han, B.-H. Bin and B. M. Weon, *Mater. Today Bio*, 2025, **31**, 101618.

5 M. Chehouri, E. Pedron, B. Genard, K. Doiron, S. Fortin, W. Bélanger, J.-S. Deschênes and R. Tremblay, *The Microbe*, 2025, **8**, 100448.

6 E. Forgacs, T. Cserháti and G. Oros, *Environ. Int.*, 2004, **30**, 953–971.

7 N. Esfandiari, *J. Supercrit. Fluids*, 2015, **100**, 129–141.

8 N. Esfandiari and S. A. Sajadian, *Arab. J. Chem.*, 2022, **15**, 104164.

9 M. Askarizadeh, N. Esfandiari, B. Honarvar, S. A. Sajadian and A. Azdarpour, *ChemBioEng Rev.*, 2023, **10**, 1006–1049.

10 S. A. Sajadian, N. Esfandiari, M. Najafi and M. Rahmanzadeh Derisi, *Chem. Thermodyn. Therm. Anal.*, 2022, **8**, 100094.

11 S. A. Sajadian, N. Esfandiari and L. Padrela, *J. CO2 Util.*, 2024, **84**, 102832.

12 N. Esfandiari, A. Rojas, A. Babhadiashar, M. J. Galotto, N. Saadati Ardestani and S. A. Sajadian, *Process*, 2023, **11**, 11.

13 M. Fathi, G. Sodeifian and S. A. Sajadian, *J. Supercrit. Fluids*, 2022, **188**, 105674.

14 I. Álvarez, C. Gutiérrez, J. F. Rodriguez, A. Lucas and M. García, *J CO2 Util.*, 2020, 41.

15 A. Ameri, G. Sodeifian and S. A. Sajadian, *J. Supercrit. Fluids*, 2020, **164**, 104892.

16 M. Champeau, J.-M. Thomassin, T. Tassaing and C. Jérôme, *J. Controlled Release*, 2015, **209**, 248–259.

17 G. Tkalec, M. Pantić, Z. Novak and Ž. Knez, *J. Mater. Sci.*, 2015, **50**, 1–12.

18 M. Kuddushi, X. Deng, J. Nayak, S. Zhu, B. B. Xu and X. Zhang, *ACS Appl. Bio Mater.*, 2023, **6**, 3810–3822.

19 Y. Yamini, M. Moradi, M. Hojjati, F. Nourmohammadian and A. Saleh, *J. Chem. Eng. Data*, 2010, **55**, 3896–3900.

20 T. Abou Elmaaty and E. Abd El-Aziz, *Text. Res. J.*, 2018, **88**, 1184–1212.

21 M. Askarizadeh, N. Esfandiari, B. Honarvar, S. A. Sajadian and A. Azdarpour, *Fluid Phase Equilib.*, 2025, **590**, 114284.

22 S. A. Sajadian, N. Esfandiari, N. Saadati Ardestani, M. Amani and L. A. Estévez, *Chem. Eng. Technol.*, 2024, **47**, 811–821.

23 S. A. Sajadian, H. Bagheri, P. Gurikov, A. Rojas, N. Esfandiari and A. Jouyban, *J. Chem. Eng. Data*, 2024, **69**, 1718–1730.

24 A. Rojas, S. A. Sajadian, F. Razmimanesh, G. Aguila, N. Esfandiari and A. Jouyban, *Fluid Phase Equilib.*, 2024, **585**, 114165.

25 M. Askarizadeh, N. Esfandiari, B. Honarvar, S. Ali Sajadian and A. Azdarpour, *Arab. J. Chem.*, 2024, **17**, 105707.

26 N. Esfandiari, N. Saadati Ardestani, R. S. Alwi, A. Rojas, C. Garlapati and S. A. Sajadian, *Sci. Rep.*, 2023, **13**, 17089.

27 C.-L. Cui, W. Shi and J.-J. Long, *J. Taiwan Inst. Chem. Eng.*, 2018, **91**, 213–223.

28 R. Wang, T. Feng, H. Sun, L. Li, K. Yu and J. Yin, *J. Supercrit. Fluids*, 2025, **222**, 106601.

29 M. Bazaei, B. Honarvar, N. Esfandiari, S. A. Sajadian and Z. A. Aboosadi, *Fluid Phase Equilib.*, 2023, **573**, 113877.

30 N. Esfandiari and S. Ali Sajadian, *J. Mol. Liq.*, 2022, **360**, 119467.

31 N. Esfandiari and S. A. Sajadian, *Fluid Phase Equilib.*, 2022, **556**, 113408.

32 A. Raghavan, R. S. Alwi, G. D. M and C. Garlapati, *Chem. Thermodyn. Therm. Anal.*, 2025, **18**, 100177.

33 Y. Hiraga and I. Ushiki, *Fluid Phase Equilib.*, 2025, **595**, 114417.

34 E. Ansari, B. Honarvar, S. A. Sajadian, Z. A. Aboosadi and M. Azizi, *Sci. Rep.*, 2023, **13**, 13402.

35 G. Sodeifian, S. A. Sajadian and F. Razmimanesh, *Fluid Phase Equilib.*, 2017, **450**.

36 G. Sodeifian, S. A. Sajadian and N. S. Ardestani, *J. Supercrit. Fluids*, 2017, **128**, 102.

37 G. Sodeifian, H. Nateghi and F. Razmimanesh, *J. CO$_2$ Util.*, 2024, **80**, 102687.

38 G. Sodeifian, F. Razmimanesh and S. A. Sajadian, *J. Supercrit. Fluids*, 2019, **146**, 89.

39 S. A. Sajadian, N. Esfandiari, A. Rojas, S. Hemmati, A. Jouyban, G. Aguila and C. Garlapati, *Sci. Rep.*, 2025, **15**, 3870.

40 M. Askarizadeh, N. Esfandiari, B. Honarvar, S. Ali Sajadian and A. Azdarpour, *Arab. J. Chem.*, 2024, **17**, 105707.

41 S. A. Sajadian, N. Esfandiari, S. Ardestani, N. Amani and L. A. Estévez, *Chem. Eng. Technol.*, 2024, **47**, 811.

42 F. M. A. Altalbawy, N. N. A. Jafar, D. Sur, A. Yadav, S. Ganesan, A. Shankhyan, M. Ravi Kumar, G. C. Sharma, I. Shernazarov, S. Q. Badraldin, U. A.-R. Hussein, K. Muzammil and H. M. Asl, *J. CO$_2$ Util.*, 2025, **92**, 103021.

43 J.-E. Li, S.-C. Chien and C.-M. Hsieh, *J. Mol. Liq.*, 2024, **395**, 123884.

44 E. F. Oghenemaro, T. Alghazali, S. Uthirapathy, H. Doshi, I. Ahmad, K. N. Raja Praveen, A. Gupta, D. Verma, A. Kumar and A. Hjazi, *J. Mol. Liq.*, 2025, **434**, 128000.

45 A. A. Almehizia, A. M. Naglah, H. M. Alkahtani, U. Hani and M. Ghazwani, *J. Mol. Liq.*, 2023, **392**, 123466.

46 S. A. Sajadian, N. Esfandiari, N. Saadati Ardestani, M. Amani and L. A. Estévez, *Chem. Eng. Technol.*, 2024, **47**, 811–821.

47 S. A. Sajadian, A. Noubigh, M. Askarizadeh and L. A. Estévez, *Chem. Eng. Technol.*, 2025, **15**, 34648.

48 J. Chrastil, *J. Phys. Chem.*, 1982, **13**, 3016–3021.

49 J. Méndez-Santiago and A. S. Teja, *Fluid Phase Equilib.*, 1999, **158–160**, 501–510.

50 K. D. Bartle, A. A. Clifford, S. A. Jafar and G. F. Shilstone, *J. Phys. Chem. Ref. Data*, 1991, **20**, 713–756.

51 S. K. Kumar and K. P. Johnston, *J. Supercrit. Fluids*, 1988, **1**, 15–22.

52 S. K. Pal and S. Mitra, *IEEE Trans. Neural Network.*, 1992, **3**, 683–697.

53 B. Karlik and A. V. Olgac, *Int. J. Artif. Intell. Expet. Syst.*, 2011, **1**, 111–122.

54 H. Ramchoun, Y. Ghanou, M. Ettaouil and M. A. Janati Idrissi, 2016.

55 A. H. Sheikhshoaei and R. Zabihi, *Sci. Rep.*, 2025, **15**, 35273.

56 C. E. Rasmussen, in *Summer School on Machine Learning*, Springer, 2003, pp. 63–71.

57 R. Grbić, D. Kurtagić and D. Slišković, *Expert Syst. Appl.*, 2013, **40**, 7407–7414.

58 J. Quinonero-Candela and C. E. Rasmussen, *J. Mach. Learn. Res.*, 2005, **6**, 1939–1959.

59 A. H. Sheikhshoaei and A. Sanati, *Sci. Rep.*, 2025, **15**, 22672.

60 M. Z. Asadzadeh, H.-P. Gänser and M. Mücke, *Appl. Eng. Sci.*, 2021, **6**, 100049.

61 H. R. Pourghasemi and N. Kerle, *Environ. Earth Sci.*, 2016, **75**, 185.

62 L. Breiman, *Classification and Regression Trees*, Routledge, 2017.

63 A. H. Sheikhshoaei, A. Khoshsima and D. Zabihzadeh, *Chem. Thermodyn. Therm. Anal.*, 2025, **17**, 100154.

64 A. H. Sheikhshoaei, A. Sanati and A. Khoshsima, *Sci. Rep.*, 2025, **15**, 26445.

65 N. Esfandiari and S. A. Sajadian, *Fluid Phase Equilib.*, 2022, **556**, 113408.

66 N. N. Kalikin, R. D. Oparin, A. L. Kolesnikov, Y. A. Budkov and M. G. Kiselev, *J. Mol. Liq.*, 2021, **334**, 115997.

67 S. A. B. Vieira de Melo, G. M. N. Costa, A. C. C. Viana and F. L. P. Pessoa, *J. Supercrit. Fluids*, 2009, **49**, 1–8.

68 G. Sodeifian, C. Garlapati, F. Razmimanesh and H. Nateghi, *Sci. Rep.*, 2022, **12**, 7758.

69 A. H. Sheikhshoaei and A. Sanati, *Energy Fuels*, 2025, **39**, 17506–17521.

© 2025 The Author(s). Published by the Royal Society of Chemistry