scientific reports

Check for updates

OPEN

Predicting drug solubility in supercritical carbon dioxide green solvent using machine learning models based on thermodynamic properties

Amir Hossein Sheikhshoaei[1] & Gholamhossein Sodeifian[2,3,4]✉

Reliable prediction of drug solubility in supercritical carbon dioxide ($scCO_2$) is crucial for the efficient design of pharmaceutical processes, including particle engineering and supercritical fluid-based extraction. Given that experimental determination of drug solubility in $scCO_2$ is costly and time-consuming, this study employs machine learning models to predict drug solubility in $scCO_2$, offering the advantage over thermodynamic models and empirical correlations of being able to predict the solubility of drugs beyond the model's training range. In this work, authors use CatBoost, XGBoost, LightGBM, and RF models to predict the solubility of a set of drugs (Sixty-eight) in $scCO_2$. Statistical errors and graphical analyses showed that the XGBoost model performed better than other models and had high reliability for predicting solubility. Among the evaluated models, XGBoost delivered the most accurate predictions, achieving a root mean square error (RMSE) of just 0.0605 and an R^2 value of 0.9984. Notably, 97.68% of the data points fell within the model's applicability domain, highlighting its strong predictive reliability. These outcomes underscore the capability of the XGBoost algorithm to serve as a robust and efficient approach for estimating solubility.

Keywords Machine learning, Solubility, Drugs, $ScCO_2$

Supercritical carbon dioxide ($scCO_2$) has emerged as a key player in green chemistry due to its unique properties, such as zero surface tension, low viscosity, high diffusivity, and tunable solubilization through adjustments in temperature, pressure, or cosolvent addition[1,2]. Its mild critical temperature (304.1 K) and pressure (7.4 MPa) make it an attractive and sustainable solvent across various industries, from dyeing and extraction to chromatography and cleaning[3–6]. In addition to being non-toxic and recyclable, $scCO_2$ enables efficient separation processes and the dissolution of a wide range of solutes, although its low polarity sometimes requires cosolvent enhancement[7,8].

In the pharmaceutical sector, $scCO_2$ has attracted attention as a green alternative to organic solvents, providing an effective medium for controlling drug solubility, facilitating particle formation, and enabling efficient supercritical fluid processing[9,10]. Applications include drug extraction, purification, crystal formation, and advanced drug delivery systems (DDSs) such as RESS, SAS, and PGSS methods. These technologies have the potential to reduce drug doses and administration frequency, enhance patient compliance, and support cleaner, safer production processes making $scCO_2$ a valuable tool for next-generation pharmaceuticals. Understanding the solubility of drugs in $scCO_2$ is essential because solubility directly affects the efficiency of supercritical processes, the stability and performance of DDSs, and the feasibility of using $scCO_2$ as a solvent, antisolvent, or solute medium[11–13]. Given that many current and pipeline drugs are poorly soluble (BCS class II and IV), enhancing their solubility in $scCO_2$ is critical for efficient particle formation, improved processability, controlled release profiles, and stable formulations, all of which are key priorities in pharmaceutical innovation[14,15].

[1]Petroleum and Petrochemical Engineering School, Hakim Sabzevari University, Sabzevar, Iran. [2]Department of Chemical Engineering, Faculty of Engineering, University of Kashan, Kashan 87317-53153, Iran. [3]Laboratory of Supercritical Fluids and Nanotechnology, University of Kashan, Kashan 87317-53153, Iran. [4]Modeling and Simulation Centre, Faculty of Engineering, University of Kashan, Kashan 87317-53153, Iran. ✉email: sodeifian@kashanu.ac.ir

While experimental determination of drug solubility in $scCO_2$ provides vital data for process design, it is often costly, time-consuming, and sometimes impractical under diverse conditions of temperature and pressure. To address these challenges, researchers have developed various simulation models, including correlation models, thermodynamic models, and equations of state (EoSs), which allow for more rapid, cost-effective, and flexible prediction of drug solubility[16–23]. Thermodynamic models, EoS approaches, and empirical correlations have long been used to predict drug 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. 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 enable significantly faster predictions compared to traditional experimental or simulation-based methods. While experimental solubility measurements in $scCO_2$ can take hours to days per condition, trained ML models can generate predictions in seconds to minutes for thousands of drug solvent condition combinations, depending on dataset size and model complexity. This rapid turnaround, combined with flexibility in handling diverse and heterogeneous datasets and the ability to include critical drug properties as input features, makes ML a powerful tool for efficient solubility estimation and process optimization.

Abdallah El Hadj et al. introduced a hybrid modeling strategy that integrates artificial neural networks (ANN) with particle swarm optimization (PSO) to estimate the solubility of solid drugs in $scCO_2$. Their ANN–PSO model demonstrated superior predictive capability compared to traditional density-based models and thermodynamic equations of state[24]. Similarly, Baghban et al. applied a least squares support vector machine (LSSVM) approach to forecast the logarithm of the solubility of 33 pharmaceutical compounds in $scCO_2$, utilizing key input variables such as temperature, pressure, CO_2 density, molecular weight, and melting point. Employing a radial basis function kernel, their LSSVM model achieved outstanding results with an average absolute relative deviation (AARD) of 5.61% and a coefficient of determination (R^2) of 0.9975, outperforming eight established empirical correlations[25]. Sodeifian et al. examined the solubility behavior of six drugs, including anti-HIV, anti-inflammatory, and anti-cancer agents, using four different modeling paradigms: cubic equations of state (SRK and modified-Pazuki), semi-empirical models (such as those proposed by Chrastil, Mendez-Santiago–Teja, Sparks et al., and Bian et al.), the regular solution theory with Flory–Huggins interaction parameters, and artificial neural networks. Their findings revealed that the ANN model exhibited the highest accuracy across all metrics (AARD, R^2, F-value), outperforming the other approaches in reproducing the experimental solubility values in arithmetic scale[26]. In another study, Euldji et al. developed a quantitative structure–property relationship (QSPR) model enhanced with artificial neural networks to estimate drug solubility in $scCO_2$. The study compiled a comprehensive dataset consisting of 3971 experimental data points from 148 drug-like compounds. Thirteen features comprising eleven molecular descriptors alongside temperature and pressure were used as inputs. The ANN model, structured as 13–10–1 and trained via Bayesian regularization (trainbr) with a log-sigmoid activation function, achieved strong predictive performance with AARD = 3.77%, RMSE = 0.5162, and a correlation coefficient r = 0.9761[27]. Furthermore, Euldji et al. also conducted a comparative assessment of seven meta-heuristic optimization algorithms for tuning the hyperparameters of a hybrid QSPR–Support Vector Regression (SVR) framework. Based on a dataset of 168 drug compounds and 4490 experimental data points, the study found that the hybrid HPSOGWO–SVR model delivered the most accurate solubility predictions, achieving an impressively low AARD of 0.706%, as validated through both statistical indices and graphical analysis[28]. Makarov et al. investigated the prediction of drug-like compound solubility in $scCO_2$ using machine learning (ML) approaches and compared them to a theoretical method based on classical density functional theory (cDFT). Two ML models based on the CatBoost algorithm were developed: one using alvaDesc descriptors and another using CDK descriptors plus drug melting points. The CatBoost-alvaDesc model showed strong predictive performance on 187 drugs, achieving an AARD of 1.8% and RMSE of 0.12 log units[29].

In this work, we predicted the solubility of 68 different drugs in $scCO_2$, using newly generated experimental data obtained by the authors and literature, and applied four advanced machine learning models: CatBoost, XGBoost, LightGBM, and Random Forest. Unlike previous studies that primarily relied on molecular descriptors or metaheuristic optimization techniques, our approach integrates critical drug-specific properties including critical temperature (Tc), critical pressure (Pc), acentric factor (ω), molecular weight (MW) and melting point (Tm) alongside commonly used state variables such as temperature (T), pressure (P), and density (ρ). This comprehensive set of input parameters allowed us to capture more nuanced relationships influencing solubility. The workflow involved systematic data preprocessing, hyperparameter tuning using mean square error (MSE) minimization, and performance evaluation through 10-fold cross-validation to ensure model robustness. Furthermore, we employed detailed statistical and graphical error analyses, complemented by outlier detection using William's plot, to rigorously define the applicability domain of the developed XGBoost model. Overall, this study not only advances predictive modeling for drug solubility in $scCO_2$ but also provides a practical tool for experimentalists. The developed model is predictive within the range of solubilities and conditions considered in this work, enabling more reliable design and optimization of supercritical fluid processes, and represents a clear improvement over earlier approaches.

Data collection

In this research, a total of 1726 experimental data points detailing the solubility of a set of drugs (Sixty-eight) in $scCO_2$ were compiled from previously published studies. Table 1 lists the names of the drugs used in this study, the number of data points for each, and the sources from which the data were collected Fig. 1 also shows

	Drug name	Number of data points	Reference
1	Chlorpromazine	45	[30]
2	Lamotrigine	36	[31]
3	Capecitabine	35	[32]
4	Aprepitant	32	[33]
5	Favipiravir	28	[34]
6	Ketoconazole	28	[35]
7	Ketotifen fumarate	28	[36]
8	Sertraline hydrochloride	28	[37]
9	Alendronate	28	[38]
10	Amantadine	28	[39]
11	Amiodarone hydrochloride	28	[40]
12	Sitagliptin phosphate	28	[41]
13	Fludrocortisone acetate	28	[42]
14	Gemifloxacin	28	[43]
15	lenalidomide	28	[44]
16	Metoprolol	28	[45]
17	Montelukast	28	[46]
18	Tramadol hydrochloride	28	[47]
19	Clozapine	27	[31]
20	Repaglinide	24	[48]
21	Oxcarbazepine	24	[49]
22	Imatinib mesylate	24	[50]
23	Loratadine	24	[51]
24	Methadone hydrochloride	24	[52]
25	Regorafenib monohydrate	24	[53]
26	Gefitinib hydrochloride	24	[17]
27	Ceftriaxone sodium	24	[54]
28	Niclosamide piperazine	24	[16]
29	5-fluorouracil	24	[55]
30	Gemcitabine	24	[56]
31	Oxycodone hydrochloride	24	[18]
32	Metformin	24	[57]
33	Esomeprazole	24	[58]
34	Dapagliflozin propanediol monohydrate	24	[19]
35	Sorafenib tosylate	24	[59]
36	Empagliflozin	24	[60]
37	Triamcinolone acetonide	24	[61]
38	Codeine phosphate	24	[62]
39	Buprenorphine hydrochloride	24	[63]
40	Nilotinib hydrochloride monohydrate	24	[64]
41	Fexofenadine hydrochloride	24	[65]
42	Hydroxychloroquine sulfate	24	[66]
43	Ibrutinib	24	[67]
44	Riluzole	24	[68]
45	Palbociclib	24	[69]
46	Rivaroxaban	24	[70]
47	Crizotinib	24	[71]
48	Prazosin hydrochloride	24	[72]
49	Pazopanib hydrochloride	24	[73]
50	Azathioprine	24	[74]
51	Metoclopramide hydrochloride	24	[52]
52	Teriflunomide	24	[75]
53	Pholcodine	24	[76]
54	Lansoprazole	24	[77]
55	Sodium Valproate	24	[78]
56	Triamterene (2,4,7-Triamino-6-phenylpteridine)	24	[79]
Continued			

	Drug name	Number of data points	Reference
57	Tamsulosin	24	[80]
58	Minoxidil	24	[81]
59	Amlodipine Besylate	24	[82]
60	Galantamine	24	[83]
61	Quetiapine hemifumarate	24	[84]
62	Sulfabenzamide	24	[85]
63	Clemastine fumarate	24	[86]
64	Dasatinib monohydrate	24	[87]
65	Sunitinib malate	24	[88]
66	Letrozole	20	[89]
67	Chlorothiazide	20	[90]
68	Dexamethasone	15	[91]

Table 1. Names of drugs used in this study, number of data points for each drug, and sources.

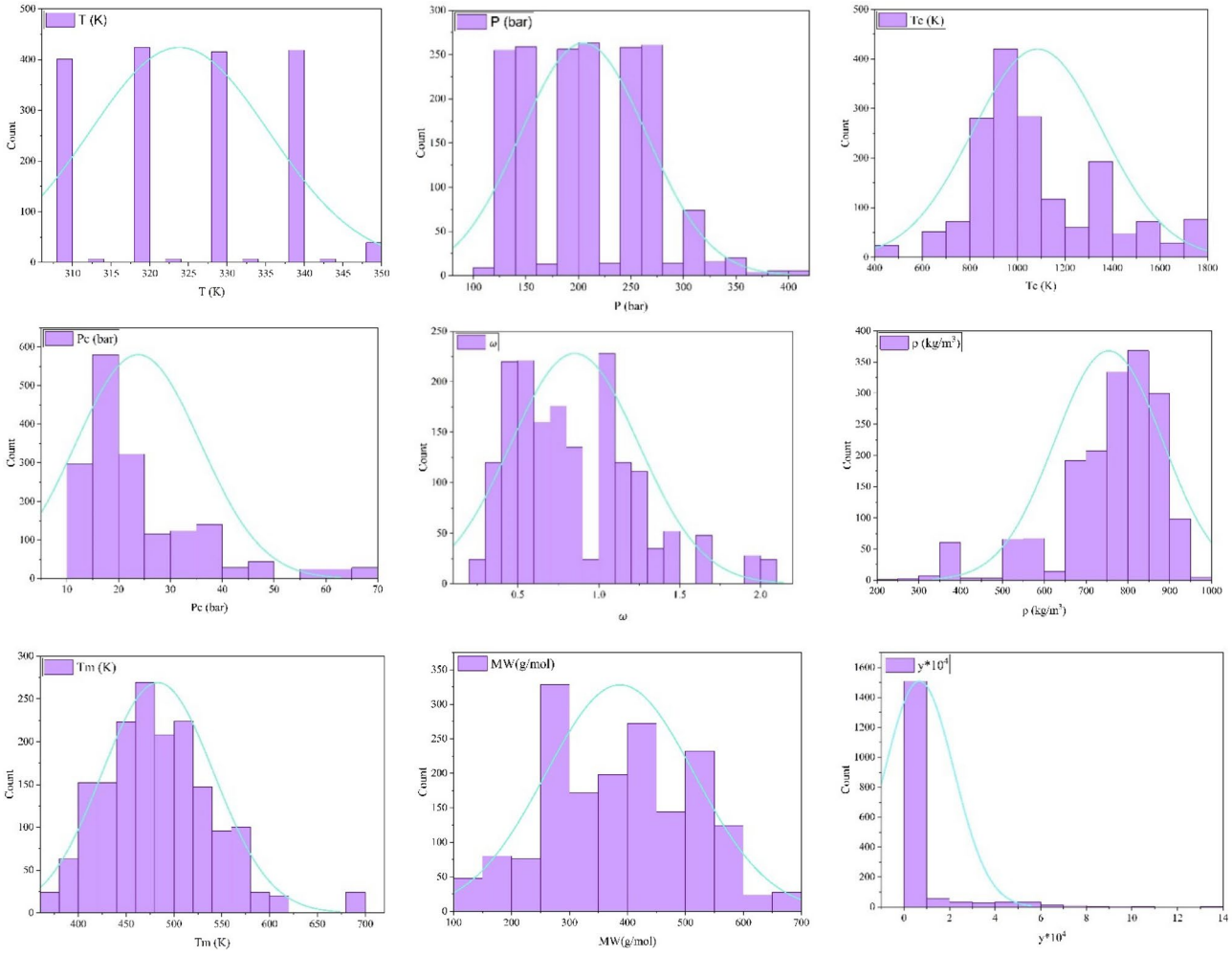

Fig. 1. Histogram plot demonstrating the distribution of the gathered database.

the distribution of the input and output features of the collected database. According to these figures, it can be seen that the amassed measurements cover comprehensive operational conditions. Table 2 provides a detailed statistical summary of the dataset, including parameters such as minimum, maximum, mean, median, skewness, and kurtosis.

Statistical parameters	Input								Output
	T (K)	*P* (bar)	Tc (K)	Pc (bar)	Tm (K)	ρ (kg/m^3)	ω	MW (g/mol)	$y \times 10^4$
Minimum	308.0	100.0	485.25	10.02	364.0	234.0	0.20	129.16	0.0007
Maximum	348.2	410.0	1789.25	66.82	698.65	976.43	2.09	681.77	13.016
Mean	323.81	204.83	1085.43	23.78	483.27	754.25	0.85	387.19	0.677
Median	328.0	210.0	1000.24	19.79	475.80	783.0	0.76	397.4	0.14
Skewness	0.06	0.33	0.70	1.65	0.68	−1.25	0.89	0.01	3.65
Kurtosis	−1.18	−0.44	0.03	2.64	1.27	1.36	0.64	−0.60	14.79

Table 2. Summary description of the performed database.

Statistical assessment of dataset

In this work, we used the input parameters T, P, Tc, Pc, ρ, ω, MW and Tm to predict the solubility of drugs in $scCO_2$.

Models development

Random forest (RF)

Random Forest (RF) is an influential ensemble-based machine learning technique developed by Leo Breiman in 2001[92]. It operates by constructing a large collection of decision trees during training and combining their outputs to improve predictive accuracy. For regression tasks, RF computes the average of predictions from all trees, while for classification, it selects the most frequent class label. Two central mechanisms underpin its effectiveness: bootstrap sampling where different subsets of the data are randomly drawn with replacement to train each tree, and randomized feature selection, in which only a random subset of features is considered at each split. These strategies help reduce model variance, enhance generalization, and mitigate overfitting, particularly when dealing with high-dimensional or complex datasets.

In regression settings, each tree yields a numeric prediction, and the RF aggregates these outputs by averaging. The trees are typically built using the CART (Classification and Regression Trees) methodology, with optimization often based on minimizing the mean squared error[93]. One of RF's advantages is that it functions effectively without the need for scaling or normalizing the input features, making it highly accessible and practical. Additionally, RF can estimate feature importance by analyzing the increase in prediction error when individual features are permuted, using out-of-bag samples for unbiased assessment. However, despite its strengths, RF can face limitations such as reduced performance with noisy datasets, sensitivity to class imbalance, and high computational costs when dealing with many large trees[94,95].

Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a high-performance ensemble learning algorithm that extends the gradient boosting technique with several enhancements aimed at increasing both accuracy and efficiency[96]. It constructs decision trees in sequence, where each new tree is trained to minimize the errors made by the previous ones. Unlike standard gradient boosting, XGBoost incorporates a second-order approximation of the loss function, utilizing both gradients and Hessians to improve the precision of model updates[97]. This second-order optimization allows the model to better capture complex patterns and nonlinear relationships in the data, making it especially effective for structured datasets. XGBoost stands out due to its scalability, ability to manage missing values natively, and high performance across diverse machine learning applications. The model employs a greedy search strategy to determine optimal splits in each tree and aggregates many shallow decision trees, specifically CARTs (Classification and Regression Trees), to form a strong predictive model. Because its hyperparameters (such as learning rate, regularization strength, and tree depth) interact with one another, careful tuning is critical to achieving reliable results without excessive computation. While XGBoost is renowned for its accuracy and robustness, its reliance on numerous decision trees may hinder interpretability, making the internal decision-making process less transparent than simpler models[97–99].

Categorical boosting (CatBoost)

CatBoost (Categorical Boosting) is a cutting-edge gradient boosting algorithm developed to natively handle categorical variables with high accuracy and minimal preprocessing[100]. Unlike traditional models that require techniques such as one-hot encoding to transform categorical data, CatBoost converts these features using target-based statistics while employing a special strategy called ordered boosting to prevent target leakage. This approach ensures that the model uses only past information when computing these statistics, which safeguards the training process against data leakage and helps produce more generalizable results. Built on the gradient boosting principle, CatBoost trains an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones. Its use of symmetric trees, combined with optimized depth control and learning rate settings, allows it to strike a balance between flexibility and regularization[101].

What makes CatBoost particularly advantageous is its ability to deliver high predictive power on datasets with mixed feature types, including high-cardinality categorical variables and sparse data. It is designed to work effectively with minimal data preprocessing and can accept raw data in various formats. Moreover, its architecture is engineered to mitigate overfitting through mechanisms like depth regulation and refined boosting techniques[102,103].

Light gradient boosting machine (LightGBM)

LightGBM (Lightweight Gradient Boosting Machine), introduced by Microsoft in 2017, is a highly efficient gradient boosting framework designed to improve training speed, reduce memory usage, and enhance prediction accuracy[104].Unlike traditional GBDT methods such as XGBoost, which rely on pre-sorted algorithms, LightGBM employs a histogram-based algorithm that bins continuous values into discrete intervals, significantly reducing computational complexity and memory requirements. A key innovation of LightGBM is its use of a leaf-wise tree growth strategy, where the algorithm splits the leaf with the highest potential to reduce error, as opposed to growing trees level by level. To control overfitting, LightGBM imposes a maximum depth constraint on trees. Additionally, LightGBM supports distributed training, enabling scalability for large datasets, and it accommodates various objective functions, including those for regression, classification, and ranking. Two core techniques further set LightGBM apart: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS prioritizes data points with large gradient values, which are more informative for learning, while randomly sampling from the remainder, thus reducing the volume of training data without sacrificing model accuracy. EFB addresses high-dimensional sparse datasets by combining mutually exclusive features those unlikely to be non-zero at the same time into a single bundled feature, thus reducing dimensionality and accelerating computation. LightGBM's innovations not only lead to faster training and lower memory overhead, but also maintain or even improve model accuracy compared to traditional boosting methods[104–107].

Predictive analytics

Fine-tuning the hyperparameters of each model is essential for achieving high predictive accuracy. Effective hyperparameter optimization enables each model to perform optimally on the given dataset. In this study, we utilize the Mean Squared Error (MSE) as the objective function to guide the hyperparameter tuning process. By minimizing the MSE, we determine the most suitable set of hyperparameters for each model.

Statistical error evaluation

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:

Mean Square Error (MSE)

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

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{i,\,pred} - y_{i,\,exp}\right| \tag{2}$$

Standard Deviation (SD)

$$SD = \sqrt{\frac{\sum_{i=1}^{n}\frac{\left(y_{i,\,exp} - y_{i,\,pred}\right)^2}{y_{i,\,exp}}}{n-1}} \tag{3}$$

Coefficient of Determination (R^2)

$$\boldsymbol{R}^2 = 1 - \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\left(\boldsymbol{y}_{\boldsymbol{i},\,\boldsymbol{exp}} - \boldsymbol{y}_{\boldsymbol{i},\,\boldsymbol{pred}}\right)^2}{\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\left(\boldsymbol{x}_{\boldsymbol{i},\boldsymbol{exp}} - \bar{\boldsymbol{x}}_{\boldsymbol{i},\,\boldsymbol{exp}}\right)^2} \tag{4}$$

Results and discussion

To evaluate the models' ability to predict drug solubility in $scCO_2$, both statistical indicators and graphical assessments were employed. The outcomes are discussed in the subsequent subsections.

Table 3 summarizes the performance of four machine learning models (CatBoost, RF, LightGBM, and XGBoost) in predicting drug solubility in supercritical CO_2, based on several statistical parameters. Across both training and test datasets, XGBoost and CatBoost consistently achieved the best results, with the lowest MSE and MAE, as well as the highest R^2. For example, XGBoost showed an almost perfect fit in the training set (MSE $\approx 1\times10^{-4}$, R^2 = 0.99999), and it maintained strong generalization ability on the test set (R^2 = 0.99013), outperforming the other models. CatBoost also delivered highly accurate predictions with test R^2 = 0.98386 and balanced performance between training and testing. In contrast, LightGBM showed relatively larger errors and wider variability, indicating lower robustness under test conditions.

The inclusion of 95% confidence intervals (CIs) and p-values provides further insight into the reliability and statistical significance of these results. Narrow CIs for XGBoost and CatBoost, particularly in MSE and MAE, confirm that these models produce stable predictions with minimal variability across different subsets of the data. On the other hand, LightGBM exhibited wider CIs, suggesting greater sensitivity to fluctuations in the dataset. The extremely small p-values (close to zero in all cases, often < 1e-300) demonstrate that the observed

Models		Statistical parameters				
		MSE (95% CI)	MAE (95% CI)	SD (95% CI)	R^2 (95% CI)	p-value
CatBoost	Train	0.00090 ± 0.00013	0.02000 ± 0.00117	1.13863 ± 0.39048	0.99963 ± 0.00009	0
	Test	0.02989 ± 0.01353	0.08524 ± 0.01714	1.33025 ± 0.26560	0.98386 ± 0.00743	1.8e-311
	Total	0.00671 ± 0.00270	0.03308 ± 0.00367	1.17913 ± 0.30039	0.99708 ± 0.00122	0
RF	Train	0.00926 ± 0.00328	0.03699 ± 0.00471	0.57803 ± 0.11110	0.99616 ± 0.00101	0
	Test	0.02960 ± 0.01308	0.07505 ± 0.01639	0.61579 ± 0.11106	0.98401 ± 0.00755	3.4e-311
	Total	0.01334 ± 0.00382	0.04462 ± 0.00481	0.58561 ± 0.09896	0.99420 ± 0.00149	0
LightGBM	Train	0.00321 ± 0.00074	0.03427 ± 0.00222	3.08629 ± 0.84939	0.99867 ± 0.00033	0
	Test	0.04761 ± 0.01991	0.12395 ± 0.01978	3.64764 ± 0.95239	0.97429 ± 0.01290	7.7e-276
	Total	0.01211 ± 0.00394	0.05225 ± 0.00443	3.20557 ± 0.73431	0.99473 ± 0.00188	0
XGBoost	Train	**$0.01 \times 10^{-4} \pm 0.000$**	**0.00078 ± 0.00004**	**0.05368 ± 0.01383**	**0.99999 ± 0.00000**	**0**
	Test	**0.01828 ± 0.00776**	**0.06019 ± 0.01257**	**0.63751 ± 0.18568**	**0.99013 ± 0.00349**	**0**
	Total	**0.00367 ± 0.00169**	**0.01269 ± 0.00270**	**0.28911 ± 0.07995**	**0.99841 ± 0.00072**	**0**

Table 3. Model accuracy evaluation using statistical indicators in the present study. Bold values indicate the lowest error and highest accuracy

Model		AARD
CatBoost	Train	0.36512
	Test	0.65682
	Total	0.42360
RF	Train	0.21400
	Test	0.34119
	Total	0.23950
LightGBM	Train	0.82083
	Test	1.57760
	Total	0.97254
XGBoost	Train	**0.01782**
	Test	**0.30635**
	Total	**0.07566**

Table 4. Assessing the accuracy of models using AARD. Bold values indicate the lowest error and highest accuracy

correlations between input variables and solubility are statistically significant and not due to random chance. Together, these findings show that XGBoost, followed closely by CatBoost, offers the most accurate, consistent, and statistically reliable predictions among the models evaluated.

Table 4 presents the average absolute relative deviation (AARD) for the four evaluated models, providing a measure of their predictive accuracy across the training, test, and total datasets. The XGBoost model exhibited the lowest AARD values (0.01782 for training, 0.30635 for test, and 0.07566 overall), indicating superior accuracy compared to CatBoost, RF, and LightGBM. CatBoost and RF also performed reasonably well, while LightGBM showed the highest deviations, particularly on the test set (AARD = 1.5776), reflecting less reliable predictions. It is important to note that the reported AARD values appear large due to the wide solubility range in the dataset (0.0007 to 13.016), and the deviations generally decrease as the solubility increases, highlighting improved predictive performance for compounds with higher solubility values.

Graphical error analysis

Graphical error analysis is a powerful tool for assessing model performance, especially when comparing the predictive accuracy of multiple models. In this study, several graphical methods were utilized to visualize and demonstrate the effectiveness of the developed models.

The cross plots provide a visual comparison between the predicted (Pred) and experimental (Exp) values, using the 45° diagonal line as a benchmark for perfect prediction. The predictive power of a model is reflected in how tightly its data points align with this reference line (45° diagonal line). As shown in Fig. 2, both CatBoost and XGBoost display a strong correspondence between predicted and measured solubility values across the training and testing sets. Only a few data points show noticeable deviation from the X = Y line. The dense concentration of points along the 45° line for these two models underscores their excellent performance in capturing the solubility patterns of the system, supporting the statistical findings reported in Table 3.

The error distribution plot provides a visual overview of the residual differences between predicted and experimental values, plotted against the corresponding experimental data points. In this type of plot, a tighter

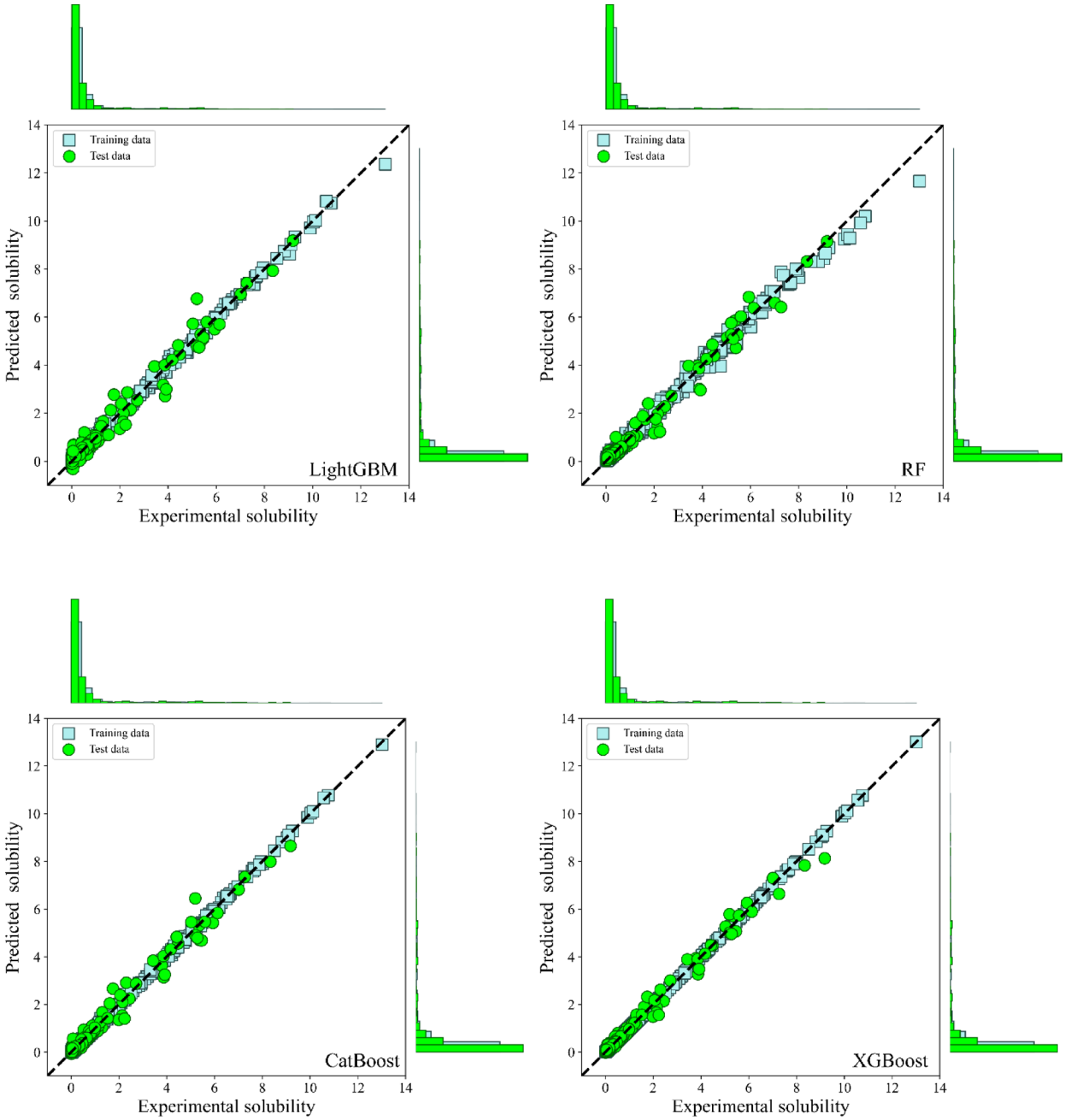

Fig. 2. Cross-plots used to assess model predictions of drug solubility in $scCO_2$.

clustering of points near the horizontal axis (Y = 0) indicates lower prediction errors and thus stronger model performance. The x-axis represents the experimental measurements, while the y-axis shows the residuals. As shown in Fig. 3, the XGBoost model exhibits the narrowest spread of error values across both the training and test datasets, highlighting its superior predictive accuracy compared to the other models.

Figure 4 depicts the cumulative frequency versus residual error for each evaluated model. This graphical representation shows the proportion of data points within defined error ranges, providing insight into the predictive reliability of each model. A steeper incline in the cumulative curve indicates that a larger proportion of predictions fall within a narrow error range, suggesting higher model precision. As shown, the XGBoost model outperforms the others, with nearly 90% of its predicted values exhibiting residual errors below 0.05, underscoring its high predictive consistency.

Figure 5 provides a comparative analysis of the prediction errors for the models assessed in this study. These errors reflect the discrepancies between the predicted and experimental solubility values. As illustrated, the XGBoost model demonstrates a narrower error range and superior accuracy in predicting solubility.

Group error plots are an effective method for evaluating the performance of models across a range of input features. In Fig. 6, these plots are presented for all models in relation to key input parameters: Tc, Pc, ρ, ω, MW,

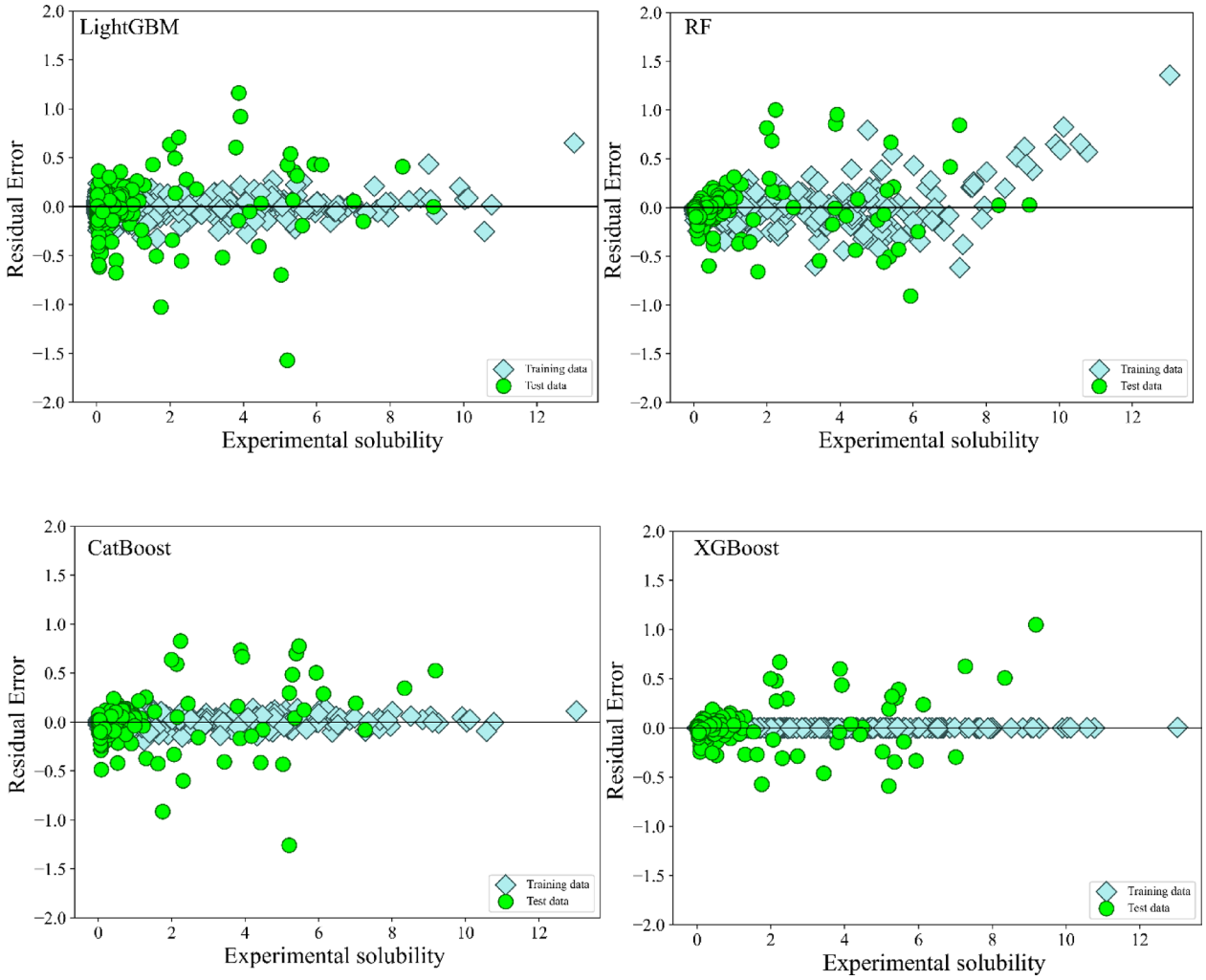

Fig. 3. Residual error distribution plots for the models predicting drug solubility in $scCO_2$.

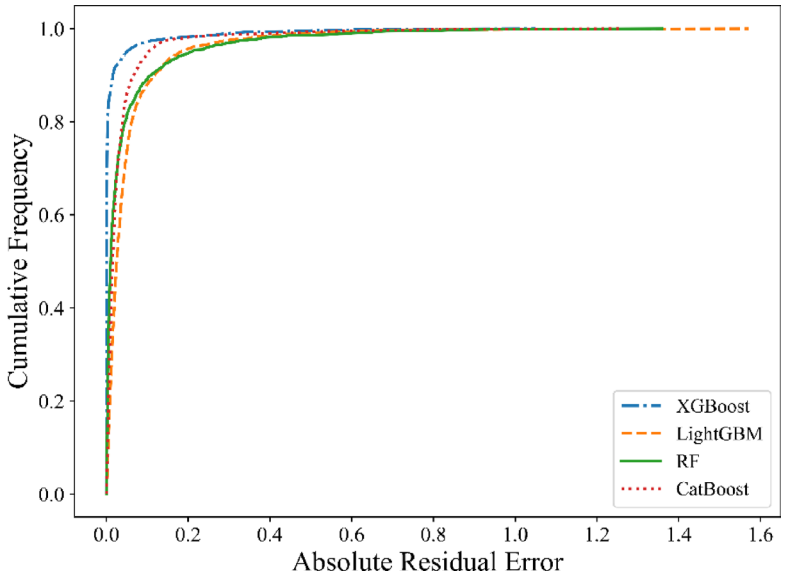

Fig. 4. Comparison of cumulative residual frequencies among the developed models.

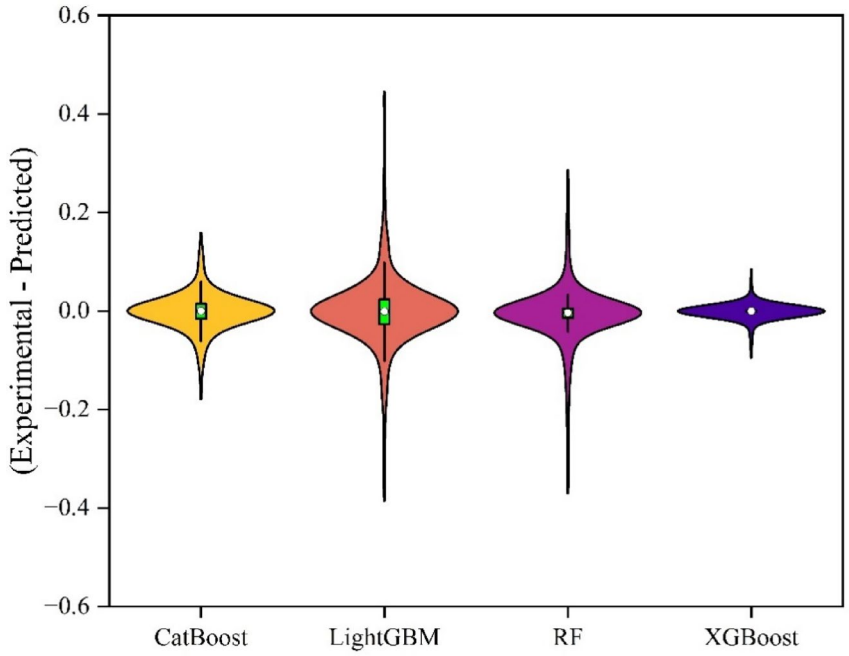

Fig. 5. Evaluation of model error behavior in solubility prediction tasks.

Tm, and the operational conditions of temperature and pressure. A visual comparison reveals that the XGBoost model consistently produces smaller prediction errors, demonstrating its superior accuracy compared to the other models.

Model trend analysis

Trend analysis provides a useful approach to explore how solubility responds to variations in input parameters. In this study, the XGBoost model, identified as the most accurate among the developed models, was employed to predict how solubility evolves with changes in density and temperature. Figure 7 illustrates the solubility behavior of hydroxychloroquine sulfate (HCQS) in $scCO_2$ as a function of temperature and $scCO_2$ density. As depicted, solubility rises with both increasing temperature and density trends that the XGBoost model accurately captured. Moreover, the close alignment between the experimental measurements and the model's predictions, as seen in the figure, further validates the strong predictive capability of the XGBoost model.

External validation and generalization assessment of drug solubility predictions in $scCO_2$

We collected gliclazide solubility data from Wang et al.[108] and performed an external validation using gliclazide as an independent drug, which was removed from the training dataset. The solubility was predicted at three temperatures (308, 318, and 328 K) across a pressure range of 100–180 bar. The results showed that the XGBoost model achieved the lowest MSE of 0.00022 and an MAE of 0.01282, demonstrating excellent accuracy in capturing the solubility behavior of a completely unseen drug. These findings confirm that the proposed model is highly generalizable, and its performance aligns with the objectives of one-drug-out cross-validation, validating the robustness and practical applicability of our approach.

Sensitivity analysis

Figure 8 displays SHAP summary plots that clarify how each input variable influences the XGBoost model's estimation of drug solubility in $scCO_2$. The plot on the right ranks features by their mean absolute SHAP values, reflecting their overall contribution to the model's predictions irrespective of whether the effect is positive or negative. A higher mean SHAP value signifies a greater influence on the model's output. The left-hand plot offers a pointwise breakdown of SHAP values, mapping how variations in individual feature values impact the predicted solubility. Feature values are color-coded, transitioning from green (low values) to purple (high values), allowing for intuitive visualization of value-dependent effects.

Among all features analyzed, Tm, P, andPc stand out with the most substantial influence on solubility predictions. The model identifies a strong positive relationship between pressure and solubility, aligning with fundamental thermodynamic laws such as Henry's law, which indicates that higher pressure generally increases gas solubility in liquids. Likewise, higher melting points are associated with greater solubility estimates, likely due to their role in modulating solid-state properties that affect dissolution behavior in supercritical media.

Other variables like Tc and the ω also exhibit non-negligible effects. The acentric factor, which captures molecular shape and polarity deviations from ideal behavior, plays a role in how well drug molecules interact with $scCO_2$. Conversely, MW and ρ appear to have a comparatively limited impact under the studied conditions, implying their influence on solubility is either indirect or less significant in this modeling context. Notably, T

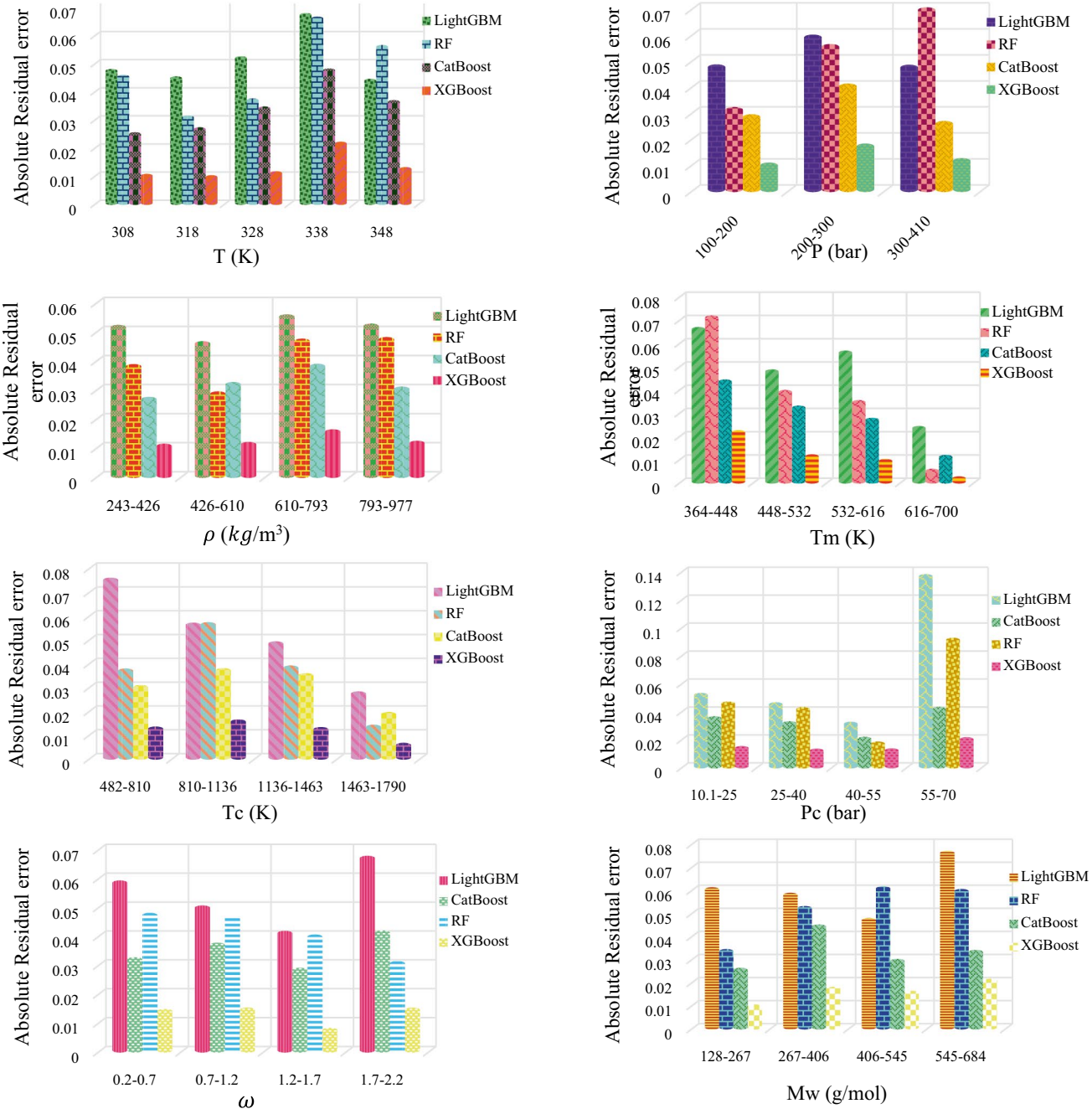

Fig. 6. Error distribution by input features for all proposed models.

contributes positively to solubility, indicating that as temperature rises, so does the predicted solubility. This trend is characteristic of $scCO_2$ systems in the most ranges, where higher temperatures enhance solute volatility and diffusivity, often outweighing density-related effects.

In summary, the SHAP results offer clear, model-agnostic explanations of the feature contributions, reinforcing the physical plausibility of the XGBoost model's internal logic. The dominant features identified by the model correspond well with established thermodynamic expectations, supporting its validity for solubility prediction in supercritical CO_2 systems.

Determining outliers and applicability domain of a technique

The 'Leverage Statistical Approach' is a widely adopted and efficient method for detecting potential anomalies data points that significantly differ from the rest of the dataset and for determining the validity range of established correlations. This technique generates a graph known as the "Williams Plot," which is constructed

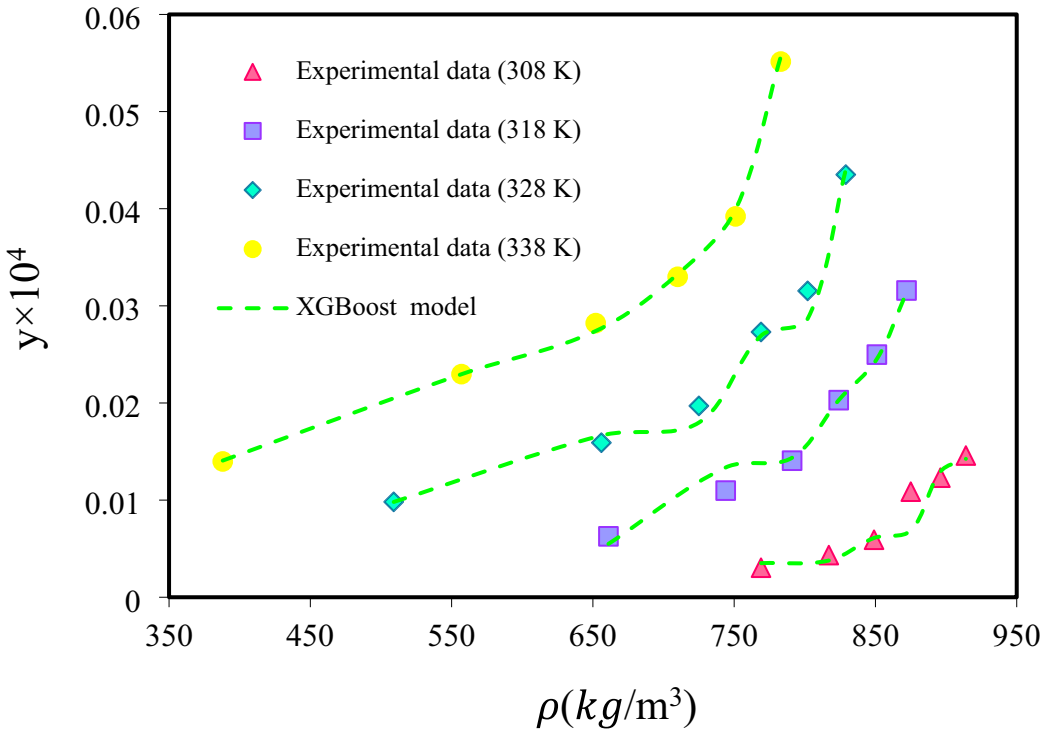

Fig. 7. HCQS solubility in $scCO_2$ versus density. Symbols are experimental data points. Solid lines are calculated from the XGBoost model.

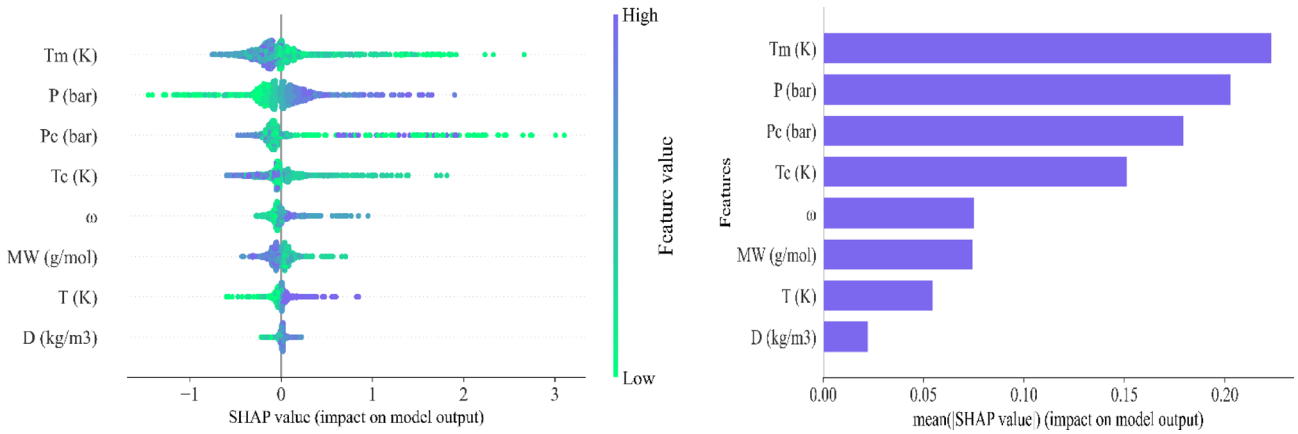

Fig. 8. Evaluation of the input parameters' impact on solubility.

by defining the Hat Matrix (H) and Standardized Residual (SR) (Fig. 9). The critical parameters required to construct this plot are calculated using the following formulas[109,110]:

• Hat matrix (H):

$$H = X\left(X^T X\right)^{-1} X^T \tag{5}$$

Here, X^T represents the transpose of the matrix X, which is a (y × d) matrix. In this case, y refers to the number of data points, and d refers to the number of input variables used by the model.

• Leverage limit (H^*):

$$H^* = \frac{3 \times (d+1)}{y} \tag{6}$$

• standardized residuals (SR):

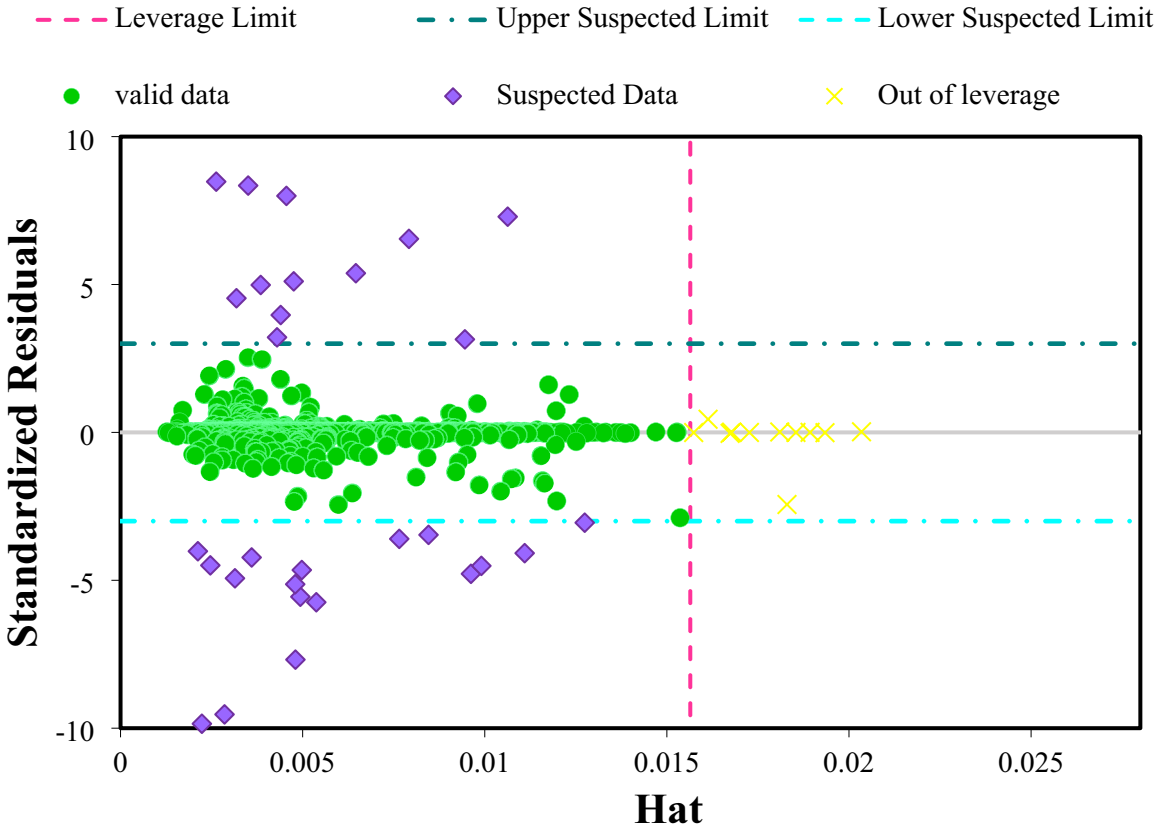

Fig. 9. William's plan for a leveraged review of the results of the XGBoost modelling.

$$R_i = \frac{z_i}{\sqrt{MSE\,(1 - H_{ii})}} \tag{7}$$

The variables z_i and H_{ii} represent the error and hat values associated with the *i-th* data point, respectively[111]. The region defined by $0 < H < H^*$ and $-3 < SR < 3$ indicates the valid domain where the model's predictions are statistically reliable (valid data region). As shown in Fig. 9, the majority of the data points (97.68%) fall within this range, confirming the strong predictive performance of the XGBoost model. However, points falling outside this domain, specifically in the areas where $SR > 3$ or $SR < -3$ and H is within the valid range, are classified as suspicious and are flagged as bad leverage points, accounting for 1.68% of the data. Additionally, points that fall within the range of $H^* < H$ and $-3 < SR < 3$ are categorized as good high leverage points and represent 0.63% of the data. Given that a significant portion of the data points are valid, it shows that the XGBoost model is very reliable for predicting drug solubility in $scCO_2$.As a complementary point in optimizing the parameters in the related sections, adjustable parameters in EoSs or semi-empirical models may be obtained by different methods including various algorithms like nonlinear regression methods [112-113].

Conclusions

In this study, we employed four machine learning models, CatBoost, XGBoost, LightGBM, and RF, to predict the solubility of a diverse set of drugs in $scCO_2$. Our dataset, compiled from 68 different drugs, included a total of 1,726 data points. To develop the predictive models, we used key input variables such as T, P, Tc, Pc, ρ, ω, MW, and Tm.

Based on statistical error metrics and graphical analyses, the XGBoost model consistently outperformed the other approaches, exhibiting the lowest prediction errors and highest accuracy in estimating solubility. Residual error analysis across the full range of input parameters further confirmed that XGBoost maintained superior performance regardless of temperature, pressure, or density ranges. Additionally, the model captured expected physical trends such as the increase in solubility with rising density at constant temperature and the enhancement of solubility with increasing temperature, reflecting its robustness and reliability.

SHAP analysis highlighted the Tm as the most influential factor among the input variables. Finally, the application of the Leverage approach for outlier detection showed that the majority of the data points fell within the defined applicability domain on the Williams plot, underscoring the reliability and generalizability of the XGBoost model.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 1 July 2025; Accepted: 10 October 2025
Published online: 17 November 2025

References

1. Bai, T., Kobayashi, K., Tamura, K., Jun, Y. & Zheng, L. Supercritical CO2 dyeing for nylon, acrylic, polyester, and casein buttons and their optimum dyeing conditions by design of experiments. *J. CO2 Utilization*. **33**, 253–261 (2019).
2. Lebedev, A., Katalevich, A. & Menshutina, N. Modeling and scale-up of supercritical fluid processes. Part I: supercritical drying. *J. Supercrit. Fluids*. **106**, 122–132 (2015).
3. Liu, Z. T. et al. Supercritical CO2 dyeing of Ramie fiber with disperse dye. *Ind. Eng. Chem. Res.* **45**, 8932–8938 (2006).
4. Qi, H., Gui, N., Yang, X., Tu, J. & Jiang, S. The application of supercritical CO2 in nuclear engineering: a review. *J. Comput. Multiph. Flows*. **10**, 149–158 (2018).
5. Ramsey, E., Qiubai, S., Zhang, Z., Zhang, C. & Wei, G. Mini-Review: green sustainable processes using supercritical fluid carbon dioxide. *J. Environ. Sci.* **21**, 720–726 (2009).
6. Aslanidou, D., Tsioptsias, C. & Panayiotou, C. A novel approach for textile cleaning based on supercritical CO2 and Pickering emulsions. *J. Supercrit. Fluids*. **76**, 83–93 (2013).
7. Patil, P. D., Dandamudi, K. P. R., Wang, J., Deng, Q. & Deng, S. Extraction of bio-oils from algae with supercritical carbon dioxide and co-solvents. *J. Supercrit. Fluids*. **135**, 60–68 (2018).
8. Zulkafli, Z. D., Wang, H., Miyashita, F., Utsumi, N. & Tamura, K. Cosolvent-modified supercritical carbon dioxide extraction of phenolic compounds from bamboo leaves (Sasa palmata). *J. Supercrit. Fluids*. **94**, 123–129 (2014).
9. Sodeifian, G., Usefi, M. M. B. & Solubility Extraction, and nanoparticles production in supercritical carbon dioxide: A mini-review. *ChemBioEng Reviews*. **10**, 133–166 (2023).
10. Liu, H., Liang, X., Peng, Y., Liu, G. & Cheng, H. Supercritical fluids: an innovative strategy for drug development. *Bioengineering* **11**, 788 (2024).
11. Cao, C. et al. WebTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* **50**, D1123–D1130 (2022).
12. Liu, K. et al. Triarylboron-doped Acenethiophenes as organic sonosensitizers for highly efficient sonodynamic therapy with low phototoxicity. *Adv. Mater.* **34**, 2206594 (2022).
13. Han, X. et al. Multifunctional TiO2/C nanosheets derived from 3D metal–organic frameworks for mild-temperature-photothermal-sonodynamic-chemodynamic therapy under photoacoustic image guidance. *J. Colloid Interface Sci.* **621**, 360–373. https://doi.org/10.1016/j.jcis.2022.04.077 (2022).
14. Kalepu, S. & Nekkanti, V. Insoluble drug delivery strategies: review of recent advances and business prospects. *Acta Pharm. Sinica B.* **5**, 442–453 (2015).
15. Zhang, L. et al. Homotypic targeting delivery of SiRNA with artificial cancer cells. *Adv. Healthc. Mater.* **9**, 1900772 (2020).
16. Sodeifian, G., Bagheri, H., Masihpour, F., Rajaei, N. & Arbab Nooshabadi, M. Niclosamide piperazine solubility in supercritical CO2 green solvent: A comprehensive experimental and modeling investigation. *J. CO2 Utilization*. **91**, 102995. https://doi.org/10.1016/j.jcou.2024.102995 (2025).
17. Sodeifian, G., Bagheri, H., Bargestan, M. & Ardestani, N. S. Determination of gefitinib hydrochloride anti-cancer drug solubility in supercritical CO2: evaluation of sPC-SAFT EoS and semi-empirical models. *J. Taiwan Inst. Chem. Eng.* **161**, 105569. https://doi.org/10.1016/j.jtice.2024.105569 (2024).
18. Sodeifian, G., Nateghi, H. & Razmimanesh, F. Mohebbi Najm Abad, J. Thermodynamic modeling and solubility assessment of oxycodone hydrochloride in supercritical CO2: Semi-empirical, EoSs models and machine learning algorithms. *Case Stud. Therm. Eng.* **55**, 104146. https://doi.org/10.1016/j.csite.2024.104146 (2024).
19. Sodeifian, G., Nateghi, H. & Razmimanesh, F. Measurement and modeling of Dapagliflozin propanediol monohydrate (an anti-diabetes medicine) solubility in supercritical CO2: evaluation of new model. *J. CO2 Utilization*. **80**, 102687. https://doi.org/10.1016/j.jcou.2024.102687 (2024).
20. Azim, M. M., Ushiki, I., Miyajima, A. & Takishima, S. Modeling the solubility of non-steroidal anti-inflammatory drugs (ibuprofen and ketoprofen) in supercritical CO2 using PC-SAFT. *J. Supercrit. Fluids*. **186**, 105626 (2022).
21. Euldji, I., Si-Moussa, C., Hamadache, M. & Benkortbi, O. QSPR modelling of the solubility of drug and drug-like compounds in supercritical carbon dioxide. *Mol. Inf.* **41**, 2200026 (2022).
22. Budkov, Y., Kolesnikov, A., Ivlev, D., Kalikin, N. & Kiselev, M. Possibility of pressure crossover prediction by classical DFT for sparingly dissolved compounds in scCO2. *J. Mol. Liq.* **276**, 801–805 (2019).
23. Noroozi, J. & Paluch, A. S. Microscopic structure and solubility predictions of multifunctional solids in supercritical carbon dioxide: A molecular simulation study. *J. Phys. Chem. B.* **121**, 1660–1674. https://doi.org/10.1021/acs.jpcb.6b12390 (2017).
24. el Abdallah, A., Laidi, M., Si-Moussa, C. & Hanini, S. Novel approach for estimating solubility of solid drugs in supercritical carbon dioxide and critical properties using direct and inverse artificial neural network (ANN). *Neural Comput. Appl.* **28**, 87–99. https://doi.org/10.1007/s00521-015-2038-1 (2017).
25. Baghban, A., Jalali, A., Mohammadi, A. H. & Habibzadeh, S. Efficient modeling of drug solubility in supercritical carbon dioxide. *J. Supercrit. Fluids*. **133**, 466–478. https://doi.org/10.1016/j.supflu.2017.10.032 (2018).
26. Sodeifian, G., Sajadian, S. A., Razmimanesh, F. & Ardestani, N. S. A comprehensive comparison among four different approaches for predicting the solubility of pharmaceutical solid compounds in supercritical carbon dioxide. *Korean J. Chem. Eng.* **35**, 2097–2116. https://doi.org/10.1007/s11814-018-0125-6 (2018).
27. Euldji, I., Si-Moussa, C., Hamadache, M. & Benkortbi, O. QSPR modelling of the solubility of drug and drug-like compounds in supercritical carbon dioxide. *Mol. Inf.* **41**, 2200026. https://doi.org/10.1002/minf.202200026 (2022).
28. Euldji, I., Belghait, A., Si-Moussa, C., Benkortbi, O. & Amrane, A. A new hybrid quantitative structure property relationships-support vector regression (QSPR-SVR) approach for predicting the solubility of drug compounds in supercritical carbon dioxide. *AIChE J.* **69**, e18115. https://doi.org/10.1002/aic.18115 (2023).
29. Makarov, D. M., Kalikin, N. N. & Budkov, Y. A. Prediction of Drug-like compounds solubility in supercritical carbon dioxide: A comparative study between classical density functional theory and machine learning approaches. *Ind. Eng. Chem. Res.* **63**, 1589–1603. https://doi.org/10.1021/acs.iecr.3c03208 (2024).
30. Alharby, T. N., Algahtani, M. M., Alanazi, J. & Alanazi, M. Advancing nanomedicine production via green thermal supercritical processing: laboratory measurement and thermodynamic modeling. *J. Mol. Liq.* **383**, 122042. https://doi.org/10.1016/j.molliq.2023.122042 (2023).
31. Hosseini, M. H., Alizadeh, N. & Khanchi, A. R. Solubility analysis of clozapine and lamotrigine in supercritical carbon dioxide using static system. *J. Supercrit. Fluids*. **52**, 30–35. https://doi.org/10.1016/j.supflu.2009.11.006 (2010).

32. Ardestani, N. S., Majd, N. Y. & Amani, M. Experimental measurement and thermodynamic modeling of capecitabine (an anticancer Drug) solubility in supercritical carbon dioxide in a ternary system: effect of different cosolvents. *J. Chem. Eng. Data*. **65**, 4762–4779. https://doi.org/10.1021/acs.jced.0c00183 (2020).
33. Sodeifian, G., Sajadian, S. A. & Ardestani, N. S. Determination of solubility of aprepitant (an antiemetic drug for chemotherapy) in supercritical carbon dioxide: empirical and thermodynamic models. *J. Supercrit. Fluids*. **128**, 102–111. https://doi.org/10.1016/j.supflu.2017.05.019 (2017).
34. Sajadian, S. A., Ardestani, N. S., Esfandiari, N., Askarizadeh, M. & Jouyban, A. Solubility of favipiravir (as an anti-COVID-19) in supercritical carbon dioxide: an experimental analysis and thermodynamic modeling. *J. Supercrit. Fluids*. **183**, 105539. https://doi.org/10.1016/j.supflu.2022.105539 (2022).
35. Sodeifian, G., Sajadian, S. A., Razmimanesh, F. & Hazaveie, S. M. Solubility of ketoconazole (antifungal drug) in SC-CO2 for binary and ternary systems: measurements and empirical correlations. *Sci. Rep.* **11**, 7546. https://doi.org/10.1038/s41598-021-87243-6 (2021).
36. Sodeifian, G., Saadati Ardestani, N., Sajadian, S. A. & Panah, H. S. Measurement, correlation and thermodynamic modeling of the solubility of ketotifen fumarate (KTF) in supercritical carbon dioxide: evaluation of PCP-SAFT equation of state. *Fluid. Phase. Equilibria*. **458**, 102–114. https://doi.org/10.1016/j.fluid.2017.11.016 (2018).
37. Sodeifian, G. & Sajadian, S. A. Experimental measurement of solubilities of Sertraline hydrochloride in supercriticalcarbon dioxide with/without menthol: data correlation. *J. Supercrit. Fluids*. **149**, 79–87. https://doi.org/10.1016/j.supflu.2019.03.020 (2019).
38. Abourehab, M. A. S. et al. Laboratory determination and thermodynamic analysis of alendronate solubility in supercritical carbon dioxide. *J. Mol. Liq.* **367**, 120242. https://doi.org/10.1016/j.molliq.2022.120242 (2022).
39. Arabgol, F., Amani, M., Ardestani, N. S. & Sajadian, S. A. Nanomedicine formulation using green supercritical processing: experimental solubility measurement and theoretical investigation. *Chem. Eng. Technol.* **47**, 318–326. https://doi.org/10.1002/ceat.202300398 (2024).
40. Sodeifian, G., Sajadian, S. A. & Razmimanesh, F. Solubility of an antiarrhythmic drug (amiodarone hydrochloride) in supercritical carbon dioxide: experimental and modeling. *Fluid. Phase. Equilibria*. **450**, 149–159. https://doi.org/10.1016/j.fluid.2017.07.015 (2017).
41. Ardestani, N. S., Sajadian, S. A., Esfandiari, N., Rojas, A. & Garlapati, C. Experimental and modeling of solubility of sitagliptin phosphate, in supercritical carbon dioxide: proposing a new association model. *Sci. Rep.* **13**, 17506. https://doi.org/10.1038/s41598-023-44787-z (2023).
42. Amani, M., Ardestani, N. S., Jouyban, A. & Sajadian, S. A. Solubility measurement of the fludrocortisone acetate in supercritical carbon dioxide: experimental and modeling assessments. *J. Supercrit. Fluids*. **190**, 105752. https://doi.org/10.1016/j.supflu.2022.105752 (2022).
43. Arabgol, F., Amani, M., Saadati Ardestani, N. & Sajadian, S. A. Experimental and thermodynamic investigation of Gemifloxacin solubility in supercritical CO2 for the production of nanoparticles. *J. Supercrit. Fluids*. **206**, 106165. https://doi.org/10.1016/j.supflu.2023.106165 (2024).
44. Sajadian, S. A., Amani, M., Saadati Ardestani, N. & Shirazian, S. Experimental analysis and thermodynamic modelling of Lenalidomide solubility in supercritical carbon dioxide. *Arab. J. Chem.* **15**, 103821. https://doi.org/10.1016/j.arabjc.2022.103821 (2022).
45. Alshahrani, S. M., Alsubaiyel, A. M., Abduljabbar, M. H. & Abourehab, M. A. S. Measurement of Metoprolol solubility in supercritical carbon dioxide; experimental and modeling study. *Case Stud. Therm. Eng.* **42**, 102764. https://doi.org/10.1016/j.csite.2023.102764 (2023).
46. Sajadian, S. A., Ardestani, N. S. & Jouyban, A. Solubility of Montelukast (as a potential treatment of COVID – 19) in supercritical carbon dioxide: experimental data and modelling. *J. Mol. Liq.* **349**, 118219. https://doi.org/10.1016/j.molliq.2021.118219 (2022).
47. Sodeifian, G., Bagheri, H., Razmimanesh, F. & Bargestan, M. Supercritical CO2 utilization for solubility measurement of Tramadol hydrochloride drug: assessment of cubic and non-cubic EoSs. *J. Supercrit. Fluids*. **206**, 106185. https://doi.org/10.1016/j.supflu.2024.106185 (2024).
48. Sodeifian, G., Hazaveie, S. M. & Sajadian, S. A. Saadati Ardestani, N. Determination of the solubility of the repaglinide drug in supercritical carbon dioxide: experimental data and thermodynamic modeling. *J. Chem. Eng. Data*. **64**, 5338–5348. https://doi.org/10.1021/acs.jced.9b00550 (2019).
49. Sodeifian, G., Hazaveie, S. M., Sajadian, S. A. & Razmimanesh, F. Experimental investigation and modeling of the solubility of oxcarbazepine (an anticonvulsant agent) in supercritical carbon dioxide. *Fluid. Phase. Equilibria*. **493**, 160–173. https://doi.org/10.1016/j.fluid.2019.04.013 (2019).
50. Sodeifian, G., Razmimanesh, F. & Sajadian, S. A. Solubility measurement of a chemotherapeutic agent (Imatinib mesylate) in supercritical carbon dioxide: assessment of new empirical model. *J. Supercrit. Fluids*. **146**, 89–99. https://doi.org/10.1016/j.supflu.2019.01.006 (2019).
51. Sodeifian, G., Razmimanesh, F., Sajadian, S. A. & Soltani Panah, H. Solubility measurement of an antihistamine drug (Loratadine) in supercritical carbon dioxide: assessment of qCPA and PCP-SAFT equations of state. *Fluid. Phase. Equilibria*. **472**, 147–159. https://doi.org/10.1016/j.fluid.2018.05.018 (2018).
52. Sodeifian, G., Hsieh, C. M., Derakhsheshpour, R., Chen, Y. M. & Razmimanesh, F. Measurement and modeling of Metoclopramide hydrochloride (anti-emetic drug) solubility in supercritical carbon dioxide. *Arab. J. Chem.* **15**, 103876. https://doi.org/10.1016/j.arabjc.2022.103876 (2022).
53. Sodeifian, G. et al. Determination of Regorafenib monohydrate (colorectal anticancer drug) solubility in supercritical CO2: Experimental and thermodynamic modeling. *Heliyon* **10**, https://doi.org/10.1016/j.heliyon.2024.e29049 (2024).
54. Sodeifian, G., Bagheri, H., Ashjari, M. & Noorian-Bidgoli, M. Solubility measurement of ceftriaxone sodium in SC-CO2 and thermodynamic modeling using PR-KM EoS and VdW mixing rules with semi-empirical models. *Case Stud. Therm. Eng.* **61**, 105074. https://doi.org/10.1016/j.csite.2024.105074 (2024).
55. Sodeifian, G., Alwi, R. S., Derakhsheshpour, R. & Ardestani, N. S. Determination of 5-fluorouracil anticancer drug solubility in supercritical CO2using semi-empirical and machine learning models. *Sci. Rep.* **15**, 4590. https://doi.org/10.1038/s41598-025-87383-z (2025).
56. Sodeifian, G., Markom, M., Ali, M., Mat Salleh, J., Derakhsheshpour, R. & M. Z. & Solubility of gemcitabine (an anticancer drug) in supercritical carbon dioxide green solvent: experimental measurement and thermodynamic modeling. *Sci. Rep.* **15**, 4451. https://doi.org/10.1038/s41598-025-88817-4 (2025).
57. Venkatesan, K. et al. Experimental - Theoretical approach for determination of Metformin solubility in supercritical carbon dioxide: thermodynamic modeling. *Case Stud. Therm. Eng.* **41**, 102649. https://doi.org/10.1016/j.csite.2022.102649 (2023).
58. Sodeifian, G., Derakhsheshpour, R. & Sajadian, S. A. Experimental study and thermodynamic modeling of Esomeprazole (proton-pump inhibitor drug for stomach acid reduction) solubility in supercritical carbon dioxide. *J. Supercrit. Fluids*. **154**, 104606. https://doi.org/10.1016/j.supflu.2019.104606 (2019).
59. Sodeifian, G., Razmimanesh, F., Sajadian, S. A. & Hazaveie, S. M. Experimental data and thermodynamic modeling of solubility of Sorafenib tosylate, as an anti-cancer drug, in supercritical carbon dioxide: evaluation of Wong-Sandler mixing rule. *J. Chem. Thermodyn.* **142**, 105998. https://doi.org/10.1016/j.jct.2019.105998 (2020).
60. Sodeifian, G., Garlapati, C., Razmimanesh, F. & Nateghi, H. Experimental solubility and thermodynamic modeling of empagliflozin in supercritical carbon dioxide. *Sci. Rep.* **12**, 9008. https://doi.org/10.1038/s41598-022-12769-2 (2022).

61. Sodeifian, G., Alwi, R. S., Nooshabadi, A., Razmimanesh, M., Roshanghias, A. & F. & Solubility measurement of triamcinolone acetonide (steroid medication) in supercritical CO2: experimental and thermodynamic modeling. *J. Supercrit. Fluids*. **204**, 106119. https://doi.org/10.1016/j.supflu.2023.106119 (2024).
62. Sodeifian, G., Garlapati, C., Arbab Nooshabadi, M., Razmimanesh, F. & Roshanghias, A. Studies on solubility measurement of Codeine phosphate (pain reliever drug) in supercritical carbon dioxide and modeling. *Sci. Rep.* **13**, 21020. https://doi.org/10.1038/s41598-023-48234-x (2023).
63. Sodeifian, G., Arbab Nooshabadi, M., Razmimanesh, F. & Tabibzadeh, A. Solubility of buprenorphine hydrochloride in supercritical carbon dioxide: study on experimental measuring and thermodynamic modeling. *Arab. J. Chem.* **16**, 105196. https://doi.org/10.1016/j.arabjc.2023.105196 (2023).
64. Nateghi, H., Sodeifian, G. & Razmimanesh, F. Mohebbi Najm Abad, J. A machine learning approach for thermodynamic modeling of the statically measured solubility of nilotinib hydrochloride monohydrate (anti-cancer drug) in supercritical CO2. *Sci. Rep.* **13**, 12906. https://doi.org/10.1038/s41598-023-40231-4 (2023).
65. Sodeifian, G., Bagheri, H., Arbab Nooshabadi, M., Razmimanesh, F. & Roshanghias, A. Experimental solubility of Fexofenadine hydrochloride (antihistamine) drug in SC-CO2: evaluation of cubic equations of state. *J. Supercrit. Fluids*. **200**, 106000. https://doi.org/10.1016/j.supflu.2023.106000 (2023).
66. Sodeifian, G., Garlapati, C., Arbab Nooshabadi, M., Razmimanesh, F. & Tabibzadeh, A. Solubility measurement and modeling of hydroxychloroquine sulfate (antimalarial medication) in supercritical carbon dioxide. *Sci. Rep.* **13**, 8112. https://doi.org/10.1038/s41598-023-34900-7 (2023).
67. Sodeifian, G., Nasri, L., Razmimanesh, F. & Arbab Nooshabadi, M. Solubility of ibrutinib in supercritical carbon dioxide (Sc-CO2): data correlation and thermodynamic analysis. *J. Chem. Thermodyn.* **182**, 107050. https://doi.org/10.1016/j.jct.2023.107050 (2023).
68. Abadian, M., Sodeifian, G., Razmimanesh, F. & Zarei Mahmoudabadi, S. Experimental measurement and thermodynamic modeling of solubility of riluzole drug (neuroprotective agent) in supercritical carbon dioxide. *Fluid. Phase. Equilibria*. **567**, 113711. https://doi.org/10.1016/j.fluid.2022.113711 (2023).
69. Sodeifian, G., Hsieh, C. M., Tabibzadeh, A. & Wang, H. C. Arbab Nooshabadi, M. Solubility of Palbociclib in supercritical carbon dioxide from experimental measurement and Peng–Robinson equation of state. *Sci. Rep.* **13**, 2172. https://doi.org/10.1038/s41598-023-29228-1 (2023).
70. Sodeifian, G., Behvand Usefi, M. M., Razmimanesh, F. & Roshanghias, A. Determination of the solubility of Rivaroxaban (anticoagulant drug, for the treatment and prevention of blood clotting) in supercritical carbon dioxide: experimental data and correlations. *Arab. J. Chem.* **16**, 104421. https://doi.org/10.1016/j.arabjc.2022.104421 (2023).
71. Sodeifian, G., Garlapati, C. & Roshanghias, A. Experimental solubility and modeling of Crizotinib (anti-cancer medication) in supercritical carbon dioxide. *Sci. Rep.* **12**, 17494. https://doi.org/10.1038/s41598-022-22366-y (2022).
72. Sodeifian, G., Surya Alwi, R., Razmimanesh, F. & Sodeifian, F. Solubility of Prazosin hydrochloride (alpha blocker antihypertensive drug) in supercritical CO2: experimental and thermodynamic modelling. *J. Mol. Liq.* **362**, 119689. https://doi.org/10.1016/j.molliq.2022.119689 (2022).
73. Sodeifian, G., Alwi, R. S., Razmimanesh, F. & Roshanghias, A. Solubility of pazopanib hydrochloride (PZH, anticancer drug) in supercritical CO2: experimental and thermodynamic modeling. *J. Supercrit. Fluids*. **190**, 105759. https://doi.org/10.1016/j.supflu.2022.105759 (2022).
74. Sodeifian, G., Razmimanesh, F., Saadati Ardestani, N. & Sajadian, S. A. Experimental data and thermodynamic modeling of solubility of Azathioprine, as an immunosuppressive and anti-cancer drug, in supercritical carbon dioxide. *J. Mol. Liq.* **299**, 112179. https://doi.org/10.1016/j.molliq.2019.112179 (2020).
75. Sodeifian, G., Nasri, L., Razmimanesh, F. & Abadian, M. CO2 utilization for determining solubility of Teriflunomide (immunomodulatory agent) in supercritical carbon dioxide: experimental investigation and thermodynamic modeling. *J. CO2 Utilization*. **58**, 101931. https://doi.org/10.1016/j.jcou.2022.101931 (2022).
76. Sodeifian, G., Alwi, R. S. & Razmimanesh, F. Solubility of Pholcodine (antitussive drug) in supercritical carbon dioxide: experimental data and thermodynamic modeling. *Fluid. Phase. Equilibria*. **556**, 113396. https://doi.org/10.1016/j.fluid.2022.113396 (2022).
77. Sodeifian, G., Sajadian, S. A. & Derakhsheshpour, R. Experimental measurement and thermodynamic modeling of Lansoprazole solubility in supercritical carbon dioxide: application of SAFT-VR EoS. *Fluid. Phase. Equilibria*. **507**, 112422. https://doi.org/10.1016/j.fluid.2019.112422 (2020).
78. Sodeifian, G., Saadati Ardestani, N., Sajadian, S. A., Golmohammadi, M. R. & Fazlali, A. Prediction of solubility of sodium valproate in supercritical carbon dioxide: experimental study and thermodynamic modeling. *J. Chem. Eng. Data*. **65**, 1747–1760. https://doi.org/10.1021/acs.jced.9b01069 (2020).
79. Sodeifian, G., Garlapati, C., Hazaveie, S. M. & Sodeifian, F. Solubility of 2,4,7-Triamino-6-phenylpteridine (Triamterene, diuretic Drug) in supercritical carbon dioxide: experimental data and modeling. *J. Chem. Eng. Data*. **65**, 4406–4416. https://doi.org/10.1021/acs.jced.0c00268 (2020).
80. Hazaveie, S. M., Sodeifian, G. & Sajadian, S. A. Measurement and thermodynamic modeling of solubility of Tamsulosin drug (anti cancer and anti-prostatic tumor activity) in supercritical carbon dioxide. *J. Supercrit. Fluids*. **163**, 104875. https://doi.org/10.1016/j.supflu.2020.104875 (2020).
81. Sodeifian, G., Saadati Ardestani, N., Razmimanesh, F. & Sajadian, S. A. Experimental and thermodynamic analyses of supercritical CO2-Solubility of Minoxidil as an antihypertensive drug. *Fluid. Phase. Equilibria*. **522**, 112745. https://doi.org/10.1016/j.fluid.2020.112745 (2020).
82. Sodeifian, G., Garlapati, C., Razmimanesh, F. & Sodeifian, F. Solubility of amlodipine besylate (Calcium channel blocker Drug) in supercritical carbon dioxide: measurement and correlations. *J. Chem. Eng. Data*. **66**, 1119–1131. https://doi.org/10.1021/acs.jced.0c00913 (2021).
83. Sodeifian, G., Hazaveie, S. M. & Sodeifian, F. Determination of galantamine solubility (an anti-alzheimer drug) in supercritical carbon dioxide (CO2): experimental correlation and thermodynamic modeling. *J. Mol. Liq.* **330**, 115695. https://doi.org/10.1016/j.molliq.2021.115695 (2021).
84. Sodeifian, G., Alwi, R. S., Razmimanesh, F. & Tamura, K. Solubility of quetiapine hemifumarate (antipsychotic drug) in supercritical carbon dioxide: Experimental, modeling and Hansen solubility parameter application. *Fluid. Phase. Equilibria*. **537**, 113003. https://doi.org/10.1016/j.fluid.2021.113003 (2021).
85. Sodeifian, G., Garlapati, C., Razmimanesh, F. & Sodeifian, F. The solubility of Sulfabenzamide (an antibacterial drug) in supercritical carbon dioxide: evaluation of a new thermodynamic model. *J. Mol. Liq.* **335**, 116446. https://doi.org/10.1016/j.molliq.2021.116446 (2021).
86. Sodeifian, G., Garlapati, C., Razmimanesh, F. & Ghanaat-Ghamsari, M. Measurement and modeling of clemastine fumarate (antihistamine drug) solubility in supercritical carbon dioxide. *Sci. Rep.* **11**, 24344. https://doi.org/10.1038/s41598-021-03596-y (2021).
87. Sodeifian, G., Surya Alwi, R., Razmimanesh, F. & Abadian, M. Solubility of dasatinib monohydrate (anticancer drug) in supercritical CO2: experimental and thermodynamic modeling. *J. Mol. Liq.* **346**, 117899. https://doi.org/10.1016/j.molliq.2021.117899 (2022).
88. Sodeifian, G., Razmimanesh, F. & Sajadian, S. A. Prediction of solubility of Sunitinib malate (an anti-cancer drug) in supercritical carbon dioxide (SC–CO2): experimental correlations and thermodynamic modeling. *J. Mol. Liq.* **297**, 111740. https://doi.org/10.1016/j.molliq.2019.111740 (2020).

89. Sodeifian, G. & Sajadian, S. A. Solubility measurement and Preparation of nanoparticles of an anticancer drug (Letrozole) using rapid expansion of supercritical solutions with solid cosolvent (RESS-SC). *J. Supercrit. Fluids*. **133**, 239–252. https://doi.org/10.1016/j.supflu.2017.10.015 (2018).
90. Majrashi, M. et al. Experimental measurement and thermodynamic modeling of Chlorothiazide solubility in supercritical carbon dioxide. *Case Stud. Therm. Eng.* **41**, 102621. https://doi.org/10.1016/j.csite.2022.102621 (2023).
91. Chim, R. B., de Matos, M. B. C., Braga, M. E. M., Dias, A. M. A. & de Sousa, H. C. Solubility of dexamethasone in supercritical carbon dioxide. *J. Chem. Eng. Data*. **57**, 3756–3760. https://doi.org/10.1021/je301065f (2012).
92. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
93. Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Routledge, 2017).
94. Sutariya, B., Sarkar, P., Indurkar, P. D. & Karan, S. Machine learning-assisted performance prediction from the synthesis conditions of nanofiltration membranes. *Sep. Purif. Technol.* **354**, 128960. https://doi.org/10.1016/j.seppur.2024.128960 (2025).
95. Sheikhshoaei, A. H., Sanati, A. & Khoshsima, A. Deep learning models to predict CO2 solubility in imidazolium-based ionic liquids. *Sci. Rep.* **15**, 26445. https://doi.org/10.1038/s41598-025-12004-8 (2025).
96. Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785–794.
97. Wang, C., Deng, C. & Wang, S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit. Lett.* **136**, 190–197 (2020).
98. Liu, H. et al. A generic machine learning model for CO2 equilibrium solubility into blended amine solutions. *Sep. Purif. Technol.* **334**, 126100. https://doi.org/10.1016/j.seppur.2023.126100 (2024).
99. Sheikhshoaei, A. H., Khoshsima, A., Salehi, A., Sanati, A. & Hemmati-Sarapardeh, A. Predicting ammonia solubility in ionic liquids using machine learning models based on critical properties. *Results Eng.* **27**, 106951. https://doi.org/10.1016/j.rineng.2025.106951 (2025).
100. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems (NeurIPS), 31, 6639–6649. (2018).
101. Bo, Y., Liu, Q., Huang, X. & Pan, Y. Real-time hard-rock tunnel prediction model for rock mass classification using catboost integrated with sequential Model-Based optimization. *Tunn. Undergr. Space Technol.* **124**, 104448 (2022).
102. Sheikhshoaei, A. H. & Sanati, A. New insight into viscosity prediction of imidazolium-based ionic liquids and their mixtures with machine learning models. *Sci. Rep.* **15**, 22672. https://doi.org/10.1038/s41598-025-08947-7 (2025).
103. Sheikhshoaei, A. H. & Sanati, A. Interfacial tension modeling of aqueous ionic liquids via machine and deep learning. *Energy Fuels*. **39**, 17506–17521. https://doi.org/10.1021/acs.energyfuels.5c02984 (2025).
104. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (NeurIPS), 30, 3149–3157. (2017).
105. Yang, X., Dindoruk, B. & Lu, L. A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. *J. Petrol. Sci. Eng.* **185**, 106598 (2020).
106. Machado, M. R., Karray, S. & De Sousa, I. T. in *14th international conference on computer science & education (ICCSE)* 1111–1116 (IEEE). (2019).
107. Sheikhshoaei, A. H. & Khoshsima, A. Machine and deep learning models for predicting high pressure density of heterocyclic thiophenic compounds based on critical properties. *Sci. Rep.* **15**, 25465. https://doi.org/10.1038/s41598-025-09600-z (2025).
108. Wang, S. W., Chang, S. Y. & Hsieh, C. M. Measurement and modeling of solubility of Gliclazide (hypoglycemic drug) and Captopril (antihypertension drug) in supercritical carbon dioxide. *J. Supercrit. Fluids*. **174**, 105244. https://doi.org/10.1016/j.supflu.2021.105244 (2021).
109. Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26**, 694–701 (2007).
110. Rousseeuw, P. J. & Leroy, A. M. *Robust regression and outlier detection* (Wiley, 2003).
111. Sheikhshoaei, A. H., Khoshsima, A. & Zabihzadeh, D. Predicting the heat capacity of strontium-praseodymium oxysilicate SrPr4(SiO4)3O using machine learning, deep learning, and hybrid models. *Chem. Thermodyn. Therm. Anal.* **17**, 100154. https://doi.org/10.1016/j.ctta.2024.100154 (2025).

Acknowledgements

The authors sincerely would like to thank the deputy of research, University of Kashan for supporting this valuable and fruitful project.

Author contributions

A.S. wrote main manuscript, and implemented modeling and G.S. Reviewed and edited the manuscript, provided experimental data. All authors reviewed the manuscript.

Funding

This work is based upon research funded by Iran National Science Foundation (INSF) under project No. 4015683. The authors are also grateful to the Research Deputy of Kashan University for the financial support of the present work under Grant number Pajoohaneh # 1404-18.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24006-7.

Correspondence and requests for materials should be addressed to G.S.