Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Prediction of SO_2 solubility in ionic liquids based on machine learning and analysis of SHAP

Peng Jia [1,*], Yifan Zhang [1], Yuhan Yan [2], Qi Zhang [1], Yanhong Huang [3], Bin Zhao [1], Qianqian Xu [1], Enze Wang [1]

[1] School of Petrochemical Engineering, Shenyang University of Technology, Shenyang, Liaoning, China
[2] Department of General Foreign Language Teaching, Liaoyang Vocational College of Technology, Liaoyang, Liaoning, China
[3]Technical Research Department, Binzhou Special Equipment Inspection Institute, Binzhou, Shandong, China

*** Corresponding author:** Peng Jia (Email: jiapeng_1988@163.com)

Abstract: Sulfur dioxide (SO_2) is a major atmospheric pollutant, and ionic liquids (ILs) have shown great potential for SO_2 capture due to their unique physicochemical properties. However, the complex and designable structures of ILs make it challenging to establish accurate structure–solubility relationships. In this study, a prediction model based on molecular descriptors and a Long Short-Term Memory (LSTM) neural network was developed to predict SO_2 solubility in ILs. A dataset of 381 experimental solubility data points covering 48 ILs was collected and augmented to 1905 points using a physics-rule-based method. Molecular descriptors for anions and cations were computed from SMILES strings using RDKit, and an intelligent feature selection method reduced the feature dimensionality from 89 to 50. The proposed Z-Score-LSTM model achieved excellent predictive performance, with a coefficient of determination (R^2) of 0.9500, a root mean square error (RMSE) of 0.0521, and a mean absolute error (MAE) of 0.0252 on the test set. SHAP analysis revealed that temperature and cation partial charge distribution (Cation_PEOE_VSA13) are the most influential features, with anion polar surface area inhibiting solubility and cation shape indices promoting it. This work provides a high-precision, interpretable tool for the molecular design of novel SO_2-absorbing ionic liquids.

Keywords: SO_2 Absorption; Ionic Liquids; Machine Learning; LSTM; SHAP.

1. Introduction

Sulfur dioxide (SO_2) is a colorless gas with a pungent and irritating odor, primarily originating from the combustion of sulfur-containing fossil fuels and non-ferrous metal smelting activities[1]. Once released into the atmosphere, SO_2 is readily oxidized and dissolves in water to form sulfuric acid, making it a major precursor to acid rain[2]. Acid rain not only acidifies soils and water bodies, leading to the corrosion of buildings, but the resultant sulfate aerosols are also a primary culprit behind regional haze events and pose significant risks to human respiratory health[3]. Although the total SO_2 emissions in China have been preliminarily controlled in recent years due to the advancement of ultra-low emission retrofits in coal-fired power plants, the deep capture and resource utilization of SO_2 remain research challenges in non-electric industrial sectors such as steel and cement manufacturing, as well as in mitigating risks from accidental industrial source releases[4].

Ionic liquids (ILs) have demonstrated immense potential in the field of SO_2 capture due to their unique advantages, including negligible vapor pressure, high thermal stability, and structural designability. To accelerate the screening and design of ionic liquids, researchers have developed various theoretical prediction methods. Early studies primarily relied on equations of state (EOS), such as the Peng-Robinson (PR) and Redlich-Kwong (RK) equations[5]. However, traditional thermodynamic models exhibit limited prediction accuracy and involve complex, time-consuming calculations[6]. In recent years, data-driven machine learning methods have gradually become a research hotspot in this area.

For instance, Baghban et al.[7] established a predictive model for the solubility of SO_2 in various ionic liquids by combining the group contribution method with Least Squares Support Vector Machines (LSSVM). Using temperature, pressure, and 17 structural descriptors as inputs, they achieved a coefficient of determination (R^2) of 0.9978 and an average relative deviation of only 1.42%. Mokarizadeh et al.[8] employed the LSSVM-GA method to predict SO_2 solubility, with sensitivity analysis indicating that temperature and pressure exert the most significant influence. Mohammadi et al.[9] systematically compared four intelligent algorithms (DBN, GMDH, GP, KNN) and five equations of state (VPT, ZJ, PR, RK, SRK), finding that the Deep Belief Network (DBN) performed optimally, achieving an average absolute relative error of only 3.56%, which represents a precision improvement compared to the 4.6% error reported in previous studies.

Despite the positive progress made in the studies cited above, existing models for predicting SO_2 solubility still suffer from the following shortcomings: (1) Most model input features are limited to macroscopic thermodynamic parameters such as temperature and pressure, lacking the molecular descriptors capable of finely characterizing the structural differences between cations and anions; (2) The inherent correlation mechanisms between the structural features of different ionic liquids and SO_2 solubility lack systematic interpretability studies; (3) The generalizability of current models is limited, making it difficult to extrapolate predictions to ionic liquids with entirely novel structures. In light of this, this paper proposes to develop a high-precision prediction model based on the molecular structural descriptors of ionic liquids combined with a Long Short-Term Memory (LSTM) neural network algorithm, aiming to

provide theoretical guidance for the molecular design of novel SO_2 absorbents.

2. Data Collection

In this study, a total of 381 SO_2 solubility data points in ionic liquids (ILs) were collected, covering 48 distinct ILs. The data span a temperature range of 282.4 K to 413.15 K and a pressure range of 0.01 bar to 3.08 bar, primarily sourced from literature surveys and publicly available databases[10-15]. Representative IL families include quaternary ammonium salts (with acetate as the anion) and imidazolium-based ILs (with tetrafluoroborate as the anion).

To compensate for the scarcity of experimental data, a physics-rule-based data augmentation technique was employed to expand the original dataset. Specifically, random perturbations within a range of ±5% were applied to the temperature and pressure of each original data point. The corresponding solubility values were then assigned based on the thermodynamic principle that gas solubility increases with pressure and decreases with temperature. Following this augmentation, the SO_2 dataset was expanded from 381 to 1905 data points, achieving a fivefold increase.

A stratified random sampling method was subsequently applied to partition the augmented dataset into training and test sets at a ratio of 80% to 20%. This partitioning process was independently repeated five times. The resulting partitions demonstrated sound statistical stability, with coefficients of variation for the mean solubility values in the training and test sets being 0.20% and 0.82%, respectively, both well below the 5% threshold.

Based on the SMILES strings of the ILs, molecular descriptors were computed using the RDKit cheminformatics toolkit. Descriptors were calculated separately for the anion and cation of each IL, encompassing four major categories: physicochemical properties (e.g., molecular weight, LogP, TPSA), charge distribution (e.g., maximum absolute charge), surface area/volume (e.g., LabuteASA, PEOE_VSA series), and topological structure (e.g., BalabanJ, BertzCT). Additionally, nine combined descriptors, such as total molecular weight, sum of LogP, and maximum charge difference, were constructed to capture the synergistic effects between the anion and cation.

Following variance thresholding (threshold = 0.01) and correlation analysis (threshold = 0.98), 87 high-value descriptors were retained from the 111 initially calculated for the SO_2 system. Further feature selection was performed using an intelligent algorithm, with criteria of target correlation ≥ 0.03 and inter-feature correlation ≤ 0.85, ultimately reducing the feature count to 50, representing a dimensionality reduction of 43.8%. A comparative analysis of the three feature selection methods is presented in Table 1, wherein the intelligent selection method consistently outperformed the others in terms of R^2, RMSE, and cross-validated R^2.

Table 1. Performance comparison of three feature selection methods

Feature Set	Number of Features	R^2	RMSE	MAE	Cross-Validation R^2
Original	89	0.8263	0.0975	0.0711	0.7978±0.0370
Correlation-based screening	60	0.8264	0.0975	0.0719	0.7997±0.0308
Intelligent selection	50	0.8279	0.0971	0.0714	0.8023±0.0326

3. Modeling

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) specifically designed to address the long-term dependency problem inherent in conventional RNNs, aiming to resolve the vanishing gradient problem when traditional RNNs process long sequences[16]. The architecture of the LSTM model is illustrated in Figure 1. The core structure of an LSTM is formed by stacking LSTM units, each containing a cell state and three gated structures. The cell state acts as a "conveyor belt" running through the entire time series, responsible for transmitting long-term information. The input gate determines which new information will be stored in the cell state, the forget gate controls which old information will be discarded, and the output gate regulates how the information from the cell state is output. These gates are processed by S-shaped activation functions (e.g., sigmoid) with output values ranging between 0 and 1, enabling a soft switching function. The working principle of the gating mechanism is that at each time step, the input data, combined with the hidden state from the previous time step, dynamically updates the cell state and hidden state through computations involving the input gate, forget gate, and output gate. The LSTM model proposed in this paper consists of three core components: an LSTM layer, an attention layer, and a fully connected layer.

The LSTM layer serves as the primary backbone for extracting temporal dependencies from the input sequence. By recurrently propagating the cell state and hidden state forward along the time axis, this layer captures both short-term fluctuations and long-term trends in the input features[17]. Subsequently, the attention layer is introduced to adaptively assign different importance weights to the hidden states at different time steps. Unlike a standard LSTM that treats all time steps equally, the attention mechanism enables the model to focus on the most informative time steps (e.g., those corresponding to critical changes in operating conditions) while suppressing irrelevant or noisy information [18]. This is typically achieved by computing a context vector as a weighted sum of the hidden states, where the weights are learned through a small feed-forward network conditioned on the hidden states[19]. Finally, the fully connected layer maps the output of the attention layer—typically the weighted context vector—to the target variable (e.g., SO_2 solubility). One or more hidden dense layers with nonlinear activation functions (e.g., ReLU) can be stacked prior to the output neuron to capture complex nonlinear relationships between the learned temporal features and the prediction target[20]. The overall architecture is trained end-to-end using backpropagation through time (BPTT) with an appropriate loss function (e.g., mean squared error for regression tasks). This combination of LSTM, attention, and fully connected layers has been demonstrated to effectively balance model flexibility and interpretability in various chemical engineering applications involving time-series or pseudo-time-series data[21].

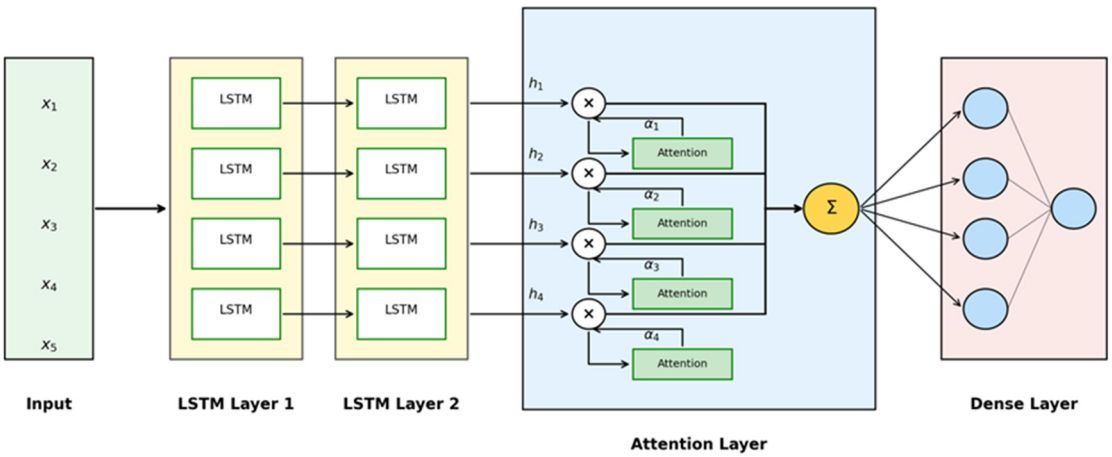

Fig 1. Structure of the LSTM model

4. Results and Discussion

4.1. Model Diagnostic Plots

In the study of SO_2 solubility prediction in ionic liquids, molecular descriptors have been normalized and standardized. Specifically, an intelligent feature selection method reduced the number of features from 89 to 50 while maintaining predictive performance. This method significantly decreased the computational complexity of subsequent modeling, simplified the model structure, and effectively retained key information.

Regarding the optimization strategy for model hyperparameters, the hyperparameter optimization framework adopted is the Bayesian optimization algorithm based on Tree-structured Parzen Estimator (TPE). The core concept of this optimization method is to guide the hyperparameter search by constructing a probabilistic model, thereby improving search efficiency while ensuring optimization quality. The optimized model hyperparameters are presented in Table 2.

Figure 2 shows the scatter plot comparing the predicted SO_2 solubility in ionic liquids obtained using the Z-Score-LSTM model with the corresponding experimental values. As can be seen from the figure, the scatter points are tightly distributed around the diagonal line, indicating that the Z-Score-LSTM model has a good predictive capability for SO_2 solubility in ionic liquids. Within the solubility range of 0 to 1.0, the vast majority of data points lie near the diagonal without obvious systematic deviation, suggesting that the model can accurately capture the variation of solubility with input features. A small number of outlier points deviating from the diagonal can be observed in the figure, which are mainly distributed in the medium-to-high concentration region. These outliers may originate from incidental errors in experimental measurements or the specific structural characteristics of certain ionic liquids that cause their dissolution behavior to deviate from general patterns. Overall, the model exhibits good prediction accuracy across the entire solubility range. For the training set, the R^2 value is 0.9773, the MAE is 0.0188, the RMSE is 0.0352, and the MSE is 0.0012; for the testing set, the R^2 value is 0.9500, the MAE is 0.252, the RMSE is 0.0521, and the MSE is 0.0027.

Table 2. Hyperparameter search space of the Z-Score-LSTM model

Parameter dimension	Search range	Optimal hyperparameter	Meaning
Hidden layer dimension	[64,256]	192	Controls model capacity
Number of LSTM layers	[1,3]	2	Adjusts network depth
Dropout rate	[0.1,0.4]	0.15288648829860468	Regularization strength
Learning rate	[1e-4,1e-2]	0.003944582189078852	Optimization step size
Batch size	[16,32,64]	16	Gradient stability

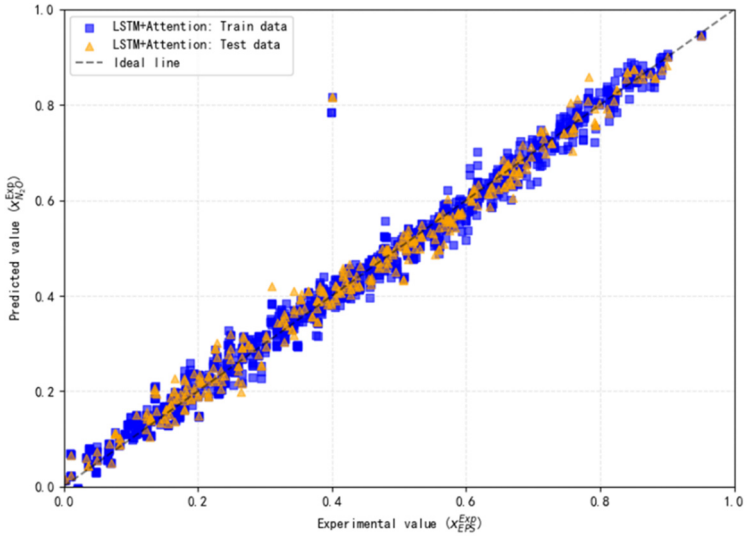

Fig 2. Experimental vs Predicted SO_2 Solubility in Ionic Liquids by Z-Score-LSTM Model

The distribution of absolute prediction errors of the Z-Score-LSTM model on the training set and the test set is shown in Figure 3. From the perspective of overall distribution characteristics, the absolute errors for both the training set and the test set exhibit a right-skewed distribution, with the peak concentrated in the low-error range, indicating that the model has good overall prediction accuracy. In terms of prediction reliability, 90% of the samples in the test set have absolute errors below 0.05, and extreme errors (>0.08) account for less than 3% of the samples, demonstrating that the model has high reliability in practical applications and can provide reliable predictions for the vast majority of samples.

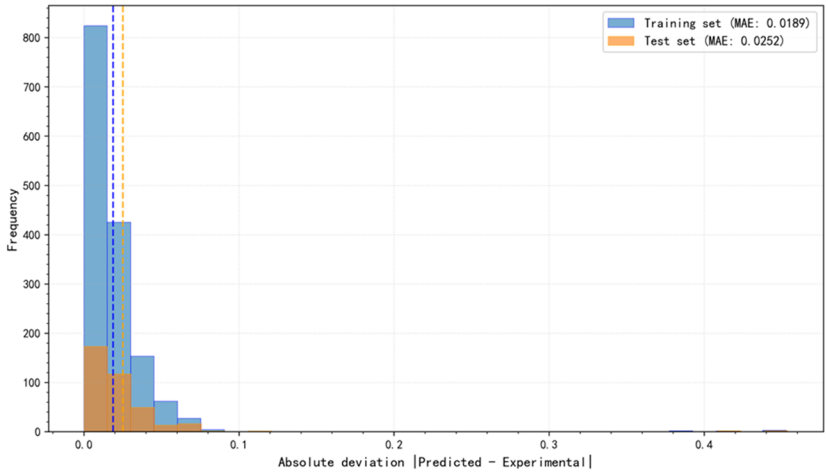

Fig 3. Distribution of absolute deviations for SO_2 solubility in ionic liquids predicted by the Z-Score-LSTM model

The distribution of prediction residuals on the training set and the test set is shown in Figure 4, which specifically includes two subfigures. From the overall residual distribution, the residual points for both the training set and the test set are symmetrically distributed around the zero line, without any obvious trend toward positive or negative sides, indicating that the model predictions have no systematic bias. The balanced distribution of positive and negative residuals suggests that the model neither systematically overestimates nor systematically underestimates the solubility values of SO_2. The residual points are randomly scattered on both sides of the zero line, without showing any obvious regular patterns (e.g., funnel shapes, curved shapes), satisfying the basic assumption in regression analysis that residuals should be random variables and validating the rationality of the model structure. In terms of concentration, the vast majority of residual points are concentrated within the range of ±0.10, indicating that the model has high prediction accuracy. Regarding generalization capability, the residual distribution patterns of the training set and the test set are similar, and although the residual range of the test set is slightly larger, it remains within an acceptable range, confirming that the model possesses good generalization capability.

4.2. Model Evaluation Metrics

There were several evaluation metrics that could be referenced to assess the relationship between predicted and actual values in a model. In this study, the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were used to quantitatively analyze the performance and accuracy of the fit of model. These metrics provided a comprehensive understanding for the predictive capabilities of model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\rho_{exp}-\rho_{pre})_i^{\,2}}{\sum_{i=1}^{N}(\rho_{exp}-\overline{\rho_{exp}})_i^{\,2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\rho_{exp}-\rho_{pre})_i^2} \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\rho_{exp}-\rho_{pre}\right|_i \quad (3)$$

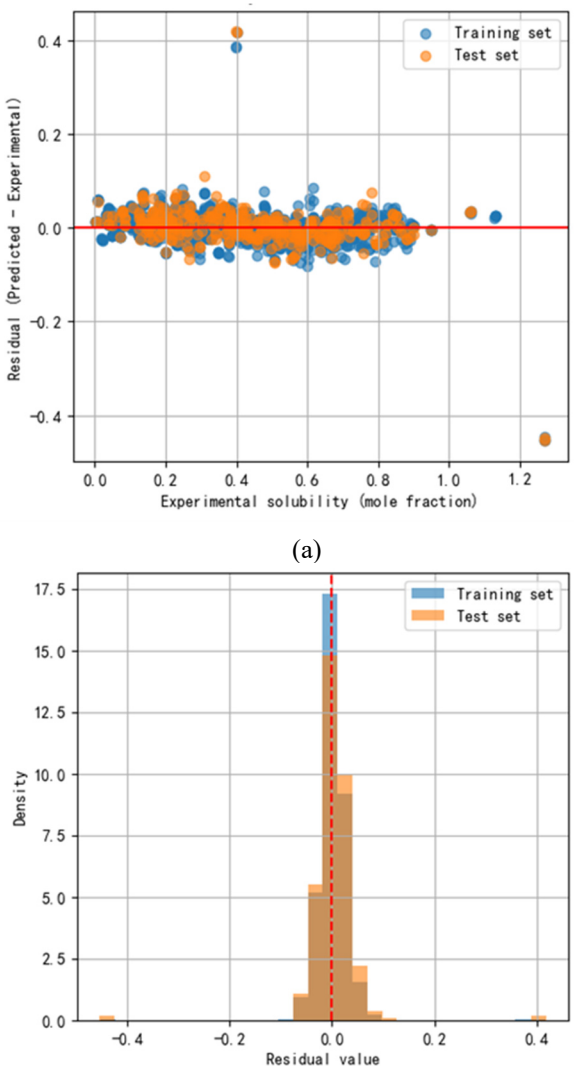

Fig 4. Residual distribution of SO_2 solubility in ionic liquids predicted by the Z-Score-LSTM model: (a) scatter plot, (b) histogram

In the formulas above , N represented the number of data points, and ρ_{exp}, ρ_{pre}, $\overline{\rho_{exp}}$, and $\overline{\rho_{pre}}$ corresponded to the experimental values, predicted values, the mean of the experimental values for the solubility of SO_2 in ILs, and the mean value of the predicted data respectively. Both RMSE and MAE measured the prediction error, where smaller values indicated lower discrepancies between the predicted and experimental values, signifying better predictive accuracy of the model. The R^2 reflected the degree of model fit, with larger values indicating a higher level of fit. Table 3 presents the statistical parameters of the model for predicting SO_2 solubility in ILs.

Table 3. Statistical parameters for models in predicting SO_2 solubility in ILs.

Model	R^2	RMSE	MAE
Z-Score-LSTM	0.9500	0.0521	0.0252

5. Model Explanation of SHAP Method

A visual interpretation of the prediction behavior of the Z-Score-LSTM model was conducted, and the resulting SHAP swarm plot is shown in Figure 5, which reflects the direction and magnitude of the contribution of different features to the model output. The figure lists the top 20 features ranked by SHAP values. On the left side are the feature names, and on the right side are the corresponding positive (red region) and negative (blue region) impacts on the model output. The length of the bars represents the mean absolute SHAP value, indicating the average influence of each feature on the prediction results.

From the overall trend, Cation_PEOE_VSA13 and T are the two most significant features affecting SO_2 solubility, exhibiting relatively high positive and negative contributions, respectively. Specifically, the SHAP values of Cation_PEOE_VSA13 are concentrated around 0.25, showing a strong positive correlation, indicating that the partial charge distribution region of the cation described by this feature promotes SO_2 dissolution. In contrast, T predominantly shows a negative influence, with SHAP values of approximately -0.30, suggesting that in the model predictions, an increase in temperature generally leads to a decrease in SO_2 solubility, which is consistent with the exothermic nature of the gas dissolution process. Furthermore, anion-related descriptors such as Anion_TPSA and Anion_PEOE_VSA5 also exhibit negative contributions, indicating that the polar surface area and charge distribution of the anion inhibit SO_2 dissolution. In contrast, cation shape indices such as Cation_Kappa2 and Cation_Kappa3 show a certain degree of positive influence, suggesting that the molecular flexibility and topological structure of the cation have a promoting effect on the dissolution behavior.

Finally, TPSA_Ratio has a relatively small impact on the model output, with SHAP values close to zero, indicating that its contribution to predicting SO_2 solubility in the Z-Score-LSTM model is weak. Overall, the SHAP analysis results reveal the dominant roles of key factors such as cation charge distribution, temperature, and anionic polar surface area in predicting SO_2 solubility, providing an interpretable basis for understanding the influence mechanism of ionic liquid structure on gas dissolution behavior.

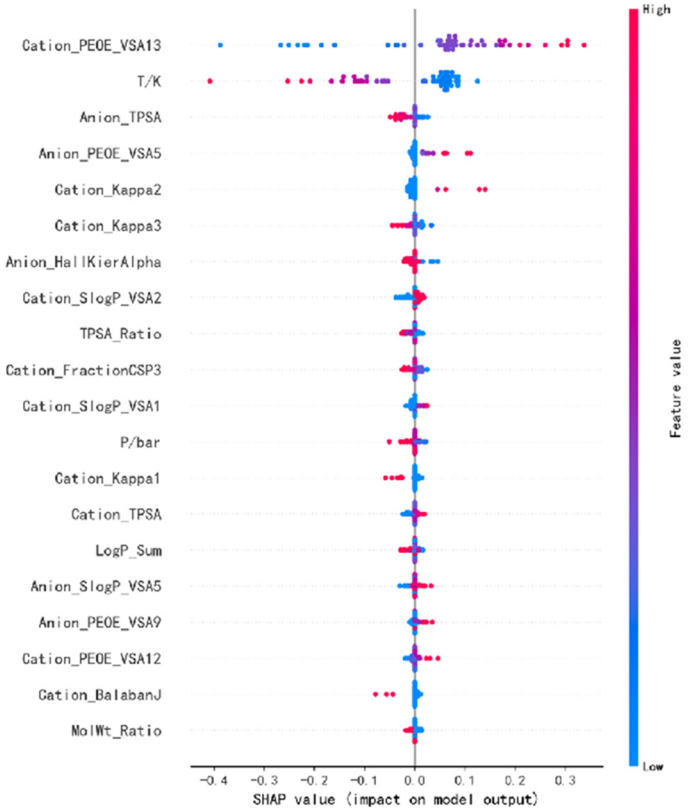

Fig 5. The SHAP values for Z-Score-LSTM

6. Conclusion

Sulfur dioxide (SO_2) is a major atmospheric pollutant, and ionic liquids (ILs) have shown great potential for SO_2 capture due to their unique physicochemical properties. However, the complex and designable structures of ILs make it challenging to establish accurate structure–solubility relationships. In this study, a prediction model based on molecular descriptors and a Long Short-Term Memory (LSTM) neural network was developed to predict SO_2 solubility in ILs. A dataset of 381

experimental solubility data points covering 48 ILs was collected and augmented to 1905 points using a physics-rule-based method. Molecular descriptors for anions and cations were computed from SMILES strings using RDKit, and an intelligent feature selection method reduced the feature dimensionality from 89 to 50. The proposed Z-Score-LSTM model achieved excellent predictive performance, with a coefficient of determination (R^2) of 0.9500, a root mean square error (RMSE) of 0.0521, and a mean absolute error (MAE) of 0.0252 on the test set. SHAP analysis revealed that temperature and cation partial charge distribution (Cation_PEOE_VSA13) are the most influential features, with anion polar surface area inhibiting solubility and cation shape indices promoting it. This work provides a high-precision, interpretable tool for the molecular design of novel SO_2-absorbing ionic liquids.

Acknowledgments

This work was supported by the 2024 Liaoning Provincial Department of Education Basic Research Projects for Colleges and Universities [grant number LJ212410142155] and Liaoning Province Science and Technology Joint Plan [2024-BSLH-201].

References

[1] Yao F G, Li L, Zhong S. Sulfur dioxide emissions curbing effects and influencing mechanisms of China's emission trading system. Plos One, 2022, 17(11): 1-32.

[2] Xue Y M, Cui X L, Li K X, et al. Statistical source analysis of recurring sulfur dioxide pollution events in a chemical industrial park. Atmospheric Environment, 2023, 295: 119564.

[3] Wei J C, Wang Z, Zhang Y X, et al. Revising the ambient air quality standards for SO_2 in China: A comprehensive analysis of air quality trends, health impacts, and global alignment. Ecotoxicology and Environmental Safety, 2025, 293: 118021.

[4] Yan Z K, Lai S Y, Ngan C L, et al. Recent advances in energy-efficient and regenerative SO_2 absorption over deep eutectic solvents. Journal of Environmental Chemical Engineering, 2022, 10(6): 108967.

[5] Yokozeki A, Shiflett M B. Gas solubilities in ionic liquids using a generic van der Waals equation of state. Journal of Supercritical Fluids, 2010, 55(2): 846-851.

[6] Ghobadi A F, Taghikhani V, Elliott J R. Investigation on the solubility of SO_2 and CO_2 in imidazolium-based ionic liquids using NPT Monte Carlo simulation. The Journal of Physical Chemistry B, 2011, 115(46): 13599-13607.

[7] Baghban A, Sasanipour J, Habibzadeh S, et al. Sulfur dioxide solubility prediction in ionic liquids by a group contribution — LSSVM model. Chemical Engineering Research and Design, 2019, 142: 180-191.

[8] Mokarizadeh A M A. Comparison of LSSVM model results with artificial neural network model for determination of the solubility of SO_2 in ionic liquids. Journal of Molecular Liquids, 2020, 304: 112726.

[9] Mohammadi M R, Hadavimoghaddam F, Atashrouz S, et al. Toward predicting SO_2 solubility in ionic liquids utilizing soft computing approaches and equations of state. Journal of the Taiwan Institute of Chemical Engineers, 2022, 133: 104220.

[10] Zhao F, Liu Q H, Song M H, et al. Ionic liquid for simultaneous desulfurization and dehydration of flue gas: data-driven thermodynamics and mechanisms. AIChE Journal, 2024, 70(8): e18456.

[11] Han G Q, Jiang Y T, Deng D S, et al. Absorption of SO_2 by renewable ionic liquid/polyethylene glycol binary mixture and thermodynamic analysis. RSC Advances, 2015, 5(107): 87750-87757.

[12] Taylor S F R, Mcclung M, Mcreynolds C, et al. Understanding the competitive gas absorption of CO_2 and SO_2 in superbase ionic liquids. Industrial & Engineering Chemistry Research, 2018, 57(50): 17033-17042.

[13] Xu Q, Jiang W, Xiao J, et al. Solubility of sulfur dioxide in tetraglyme-NH_4SCN ionic liquid: high absorption efficiency. RSC Advances, 2018, 8(73): 42116-42122.

[14] Zeng S, He H, Gao H, et al. Improving SO_2 capture by tuning functional groups on the cation of pyridinium-based ionic liquids. RSC Advances, 2015, 5(4): 2470-2478.

[15] Huang J, Riisager A, W R, et al. Tuning ionic liquids for high gas solubility and reversible gas sorption. Journal of Molecular Catalysis A: Chemical, 2008, 279(2): 170-176.

[16] Alomari F A, Aljamal N, Banna M A, et al. Long short-term memory networks: a comprehensive survey. AI, 2025, 6(9): 215.

[17] Han S, Dong H B. A temporal window attention-based window-dependent long short-term memory network for multivariate time series prediction. Entropy, 2022, 25(1): 10.

[18] Tang M A, Bai Z C, Qiu J D, et al. Prediction of SO_2 concentration in WFGD system based on GWO optimized CNN-BiLSTM-attention. The Canadian Journal of Chemical Engineering, 2025, 103(7).

[19] Attention-enhanced multi-time scale LSTM for soft sensor modeling of corn starch liquefaction. Chinese Journal of Chemical Engineering, 2025.

[20] Efficient prediction framework for large-scale nonlinear petrochemical process based on feature selection and temporal-attention LSTM: Applied to fluid catalytic cracking. Chemical Engineering Science, 2024.

[21] Adaptive deep learning modeling of green ammonia production process based on two-layer attention mechanism LSTM. Processes, 2025, 13(5): 1480.