Scilight

Smart Chemical Engineering

https://www.sciltp.com/journals/sce

Article

Prediction of Drug Solubility in Complex Micro-Environments: Machine Learning Study Based on Hansen Solubility Parameters

Cong Wang [1,†], Xiang Zhou [1,†], Huimin Gao [1], Yuanhui Ji [2] and Hongliang Qian [1,*]

[1] Department of Pharmaceutical Engineering, China Pharmaceutical University, Nanjing 211198, China
[2] Jiangsu Province Hi-Tech Key Laboratory for Biomedical Research, School of Chemistry and Chemical Engineering, Southeast University, Nanjing 211189, China
* Correspondence: hlqian@cpu.edu.cn
† These authors contributed equally to this work.

How To Cite: Wang, C.; Zhou, X.; Gao, H.; et al. Prediction of Drug Solubility in Complex Micro-Environments: Machine Learning Study Based on Hansen Solubility Parameters. *Smart Chemical Engineering* **2025**, *1*(1), 5. https://doi.org/10.53941/sce.2025.100005

Received: 4 August 2025
Revised: 5 September 2025
Accepted: 11 October 2025
Published: 24 October 2025

Abstract: Although machine learning (ML) has been widely applied to drug solubility prediction, most existing models focus on single-solvent systems, and accurate prediction in complex micro-environments (mixed solvents) remains challenging. Herein, we report the first predictive framework that integrates mechanism-driven microscopic variables (molecular descriptors and Hansen solubility parameters) into three ML algorithms: artificial neural networks (ANN), support vector machines (SVM), and random forests (RF). Among them, RF achieved the best performance, with the coefficient of determination (R^2) markedly improved from 0.830 to 0.988 when Hansen parameters were included as input features. Analysis of factor interactions revealed that solvent hydrogen-bonding capacity, polarity, mixing ratio, and temperature play key roles in modulating drug solubility, consistent with previous experimental studies. These results underscore the critical role of microscopic variables in capturing solubility behavior in mixed solvent systems. More broadly, this work demonstrates that integrating mechanism-driven descriptors with data-driven ML models offers a powerful and generalizable strategy for accurately predicting drug solubility in complex micro-environments.

Keywords: machine learning; complex micro-environments; mechanism-driven microscopic variables; drug solubility

1. Introduction

Since the vast majority of compounds in development suffer from poor aqueous solubility, this physicochemical property remains one of the most persistent and critical bottlenecks in modern drug discovery and development [1]. For example, Fluconazole belongs to the Biopharmaceutics Classification System (BCS) Class III class of drugs, which can be used to treat meningitis caused by cryptococcus neoformans, systemic candida infection and other diseases. However, the pure drug exhibited very low aqueous solubility (0.029 mg/mL), which resulted in incomplete absorption and poor bioavailability [2]. Fexofenadine hydrochloride, a non-sedating antihistamine, is prescribed to manage allergic disorders such as seasonal hay fever and chronic hives of unknown origin. However, fexofenadine was found to have a solubility of only 1.45 mg/mL in aqueous hydrochloric acid, so its oral absorption was poor [3].

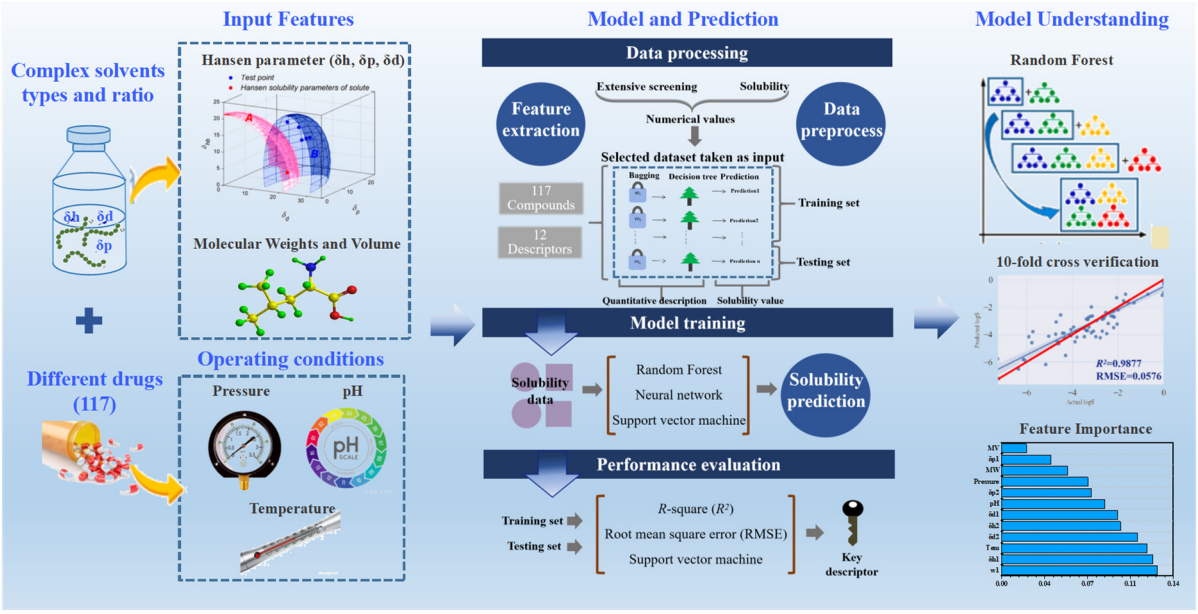
Extensive prior research has been devoted to addressing the myriad challenges arising from the poor solubility of pharmaceutical compounds [1,3,4]. Savjani et al. [5] reviewed a broad spectrum of strategies aimed

at improving the solubility of poorly soluble drugs. These encompass both physical and chemical modifications (such as particle-size reduction, crystal engineering, salt formation, solid dispersions, surfactant utilization, and complexation) as well as other complementary approaches. Among these methods, different solvents are very important for the drug preparation and changes in drug solubility. For example, in the process of drug preparation, it is necessary to screen single solvent or mixed solvents to improve drug solubility to increase the drug loading capacity and efficiency of carriers. At the same time, in clinical applications, drugs need to consider their solubility in the human micro-environments to achieve the highest bioavailability. In particular, the micro-environments in the human body are very complex and changeable. Owing to the complex micro-environments of drugs, their preparation is often difficult. Even when the same compound is processed under identical operating conditions, variations in solubility persist due to the heterogeneous composition of mixed dissolution media. In addition, the diversity of experimental parameters, including temperature, pH, and pressure, makes it unrealistic to experimentally measure the solubility of all compounds. Hence, the accurate prediction of drug solubility in mixed solvent systems remains a critical yet challenging task.

Nowadays, data-centric approaches in artificial intelligence (AI) have rapidly penetrated conventional scientific workflows, yielding impressive outcomes [6,7]. Central to these advances are machine learning algorithms that ingest large, heterogeneous datasets, distill salient molecular descriptors and patterns, and extrapolate reliable trends for predictive modeling. In parallel, an emerging body of literature is now exploring the use of such algorithms to forecast drug solubility with increasing accuracy [8]. Cysewski et al. [9] employed machine learning models in which intermolecular interactions served as molecular descriptors to estimate the solubility of dapsone in several neat solvents: namely N-methyl-2-pyrrolidone, dimethyl sulfoxide, 4-formylmorpholine, tetraethylenepentamine, and diethylene glycol bis (3-aminopropyl) ether. We developed a hybrid framework that fuses molecular-thermodynamic modeling with machine learning algorithms to deliver accurate solubility predictions for drugs dissolved in a set of pure solvents: ethyl acetate, acetone, ethanol, acetonitrile, and water [10]. Li et al. [11] effectively demonstrated the powerful potential of sigma profile (σ-profile) as an input feature to accurately predict the solubility of solids in supercritical carbon dioxide ($ScCO_2$) by building a machine learning model to study the solubility of 117 drugs in $ScCO_2$. Nevertheless, these studies only focus on modeling and predicting solubility in single solvents such as water, organic solvents, and supercritical CO_2 solvents [12,13]. It should be mentioned that the micro-environments in the process of drug preparation is often not a single solvent environment, which requires a variety of mixed solvents with different ratios to be prepared together. Therefore, greater attention should be given to the influence of mixed solvents on drug solubility. The scarcity of relevant studies on solubility changes in complex micro-environments largely arises from the fact that diverse drugs and mixed solvents cannot be adequately represented by macroscopic experimental conditions. Thus, the key question is how to identify an accurate yet simple descriptor that can capture the characteristics of different drugs and mixed solvents.

It is reported that the common descriptors used to represent the complex solvent preparation process in the micro-environments are mechanism-driven microscopic variables such as molecular descriptors or Hansen solubility parameters (HSP). Among them, Hansen, as an index to evaluate the affinity between substances, is divided into three parts, which represent the specific types of interaction, namely, dispersion interaction (δd), polar interaction (δp) and hydrogen bond interaction (δh) [14]. Molecular descriptors usually are used to represent physicochemical and structural properties of compounds. Three tiers of descriptors (one-, two-, and three-dimensional) map distinct molecular facets: bulk composition, connectivity patterns, and spatial architecture together with its functional implications [15]. Therefore, assuming that the Hansen solubility parameters or molecular descriptors are used to express the heterogeneity of different drugs and mixed solvents, it is expected to realize their prediction accuracy of drug solubility in complex micro-environments.

In summary, it is difficult to accurately predict the solubility of various drugs in complex micro-environments by existing studies. Herein, to improve the prediction accuracy of drug solubility in complex micro-environments, mechanism-driven microscopic variables such as the Hansen solubility parameters or molecular descriptors, were used to represent different drugs and solvents as inputs in machine learning method for the first time. Among them, a comprehensive dataset spanning 1980–2024 was assembled by querying the Web of Science Core Collection with the keyword sets “Hansen + drug solubility” and “molecular descriptor + drug solubility”. Leveraging this corpus of over 10,000 data points, we employed three machine learning algorithms (ANN, SVM, and RF) to concurrently model and predict drug solubility within complex micro-environments, as outlined in Scheme 1.



Scheme 1. The framework for prediction of drug solubility in complex micro-environments: Machine learning study based on Hansen solubility parameters.

2. Materials and Methods

2.1. Data Acquisition and Curation

Since the vast majority of drug experiments are conducted under the condition of gas-liquid equilibrium, and Raoult's law for calculating solubility is defined in terms of mole fraction. Therefore, this paper takes mole fraction solubility (MFS) as the target variable to study drug solubility.

The experiment data of the Hansen solubility parameters for different solvents and drugs, mixed solvent ratio, drug compositions, operating conditions, and drug solubility were collected from the literature, and the total datasets (10,086 sample datasets) were listed in Database S1 of the Supplementary Materials. It should be specially explained that the mixed solvent ratio is defined as zero while the solvent is single, hence these datasets included all contents on the drug solubility in both single and mixed solvents. In addition, it was found through literature research that the data with Hansen solubility parameters as the characteristic variable had very few datasets (less than 1000 sample data) on solubility. Therefore, the analysis of drug solubility data was ignored. All drug samples were handled in diverse solvents and subjected to varying pH, pressure, and temperature, yet processed under conditions that preserved their intrinsic properties. Nevertheless, the compiled data still contain unavoidable experimental uncertainties.

2.2. Data Pre-Processing

Key determinants of drug solubility include the Hansen solubility parameters of the solute and each solvent, the mixed solvent ratio, drug composition, and operating conditions, all of which can be grouped into three principal categories: (I) The Hansen solubility parameters for different drugs and solvents, including drug and solvent polarity, dispersion and hydrogen; (II) The drug compositions, including the drug molecular weight and volume; (III) The operating conditions, including the temperature, pressure and pH. The target variable is the drug MFS.

The MFS at saturation (x_m, T) of drug in the various mixtures of the binary solvents at various temperatures were calculated by:

$$x_{m,T} = \frac{{}^{m_D}/_{MW_D}}{({}^{m_D}/_{MW_D} + \sum_{i=2}^{n}({}^{m_i}/_{MW_i}))} \quad (1)$$

where m_D and MW_D are the mass and the molar mass of drug, respectively, m_i and MW_i are the mass and molar mass of various solvent.

For mixed solvent systems, the HSP of the mixture were calculated using a volume-fraction-weighted average of the individual solvent components, following the commonly applied mixing rule [16]:

$$\delta_{i,mix} = \sum_j \varphi_j \cdot \delta_{i,j} \ (i = d, p, h) \tag{2}$$

where φ_j denotes the volume fraction of solvent j in the mixture, and $\delta_{i,j}$ represents the δd, δp, and δh. This approach has been widely used in practice to estimate HSP of solvent blends. This approach has been widely used in pharmaceutical and polymer research to estimate solute and solvent affinity in mixed solvent environments.

Typically, linear associations between any two variables are evaluated with the Pearson correlation coefficient (PCC, Equation (3)) and its corresponding *p*-value (Equation (4)).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

Here, $\bar{x}$ and $\bar{y}$ denote the mean values of variables x and y, and r is the PCC between any pair of inputs or between an input and the output. Bounded between −1 and +1, $|r|$ quantifies the strength of the linear association: values from 0 to +1 indicate positive correlation, whereas values from −1 to 0 indicate negative correlation. In both cases, the closer $|r|$ is to unity, the stronger the linear relationship.

$$P = \frac{r\sqrt{N\text{-}2}}{\sqrt{1 - r^2}} \tag{4}$$

Here, N denotes the number of samples and p is the two-tailed p-value calculated from the t-distribution with N-2 degrees of freedom. Together, the PCC and p-value reveal both the direction and statistical significance of each feature's influence on the target variable.

On the other hand, to evaluate potential multicollinearity among descriptors, we first calculated pairwise Pearson correlation coefficients and considered descriptors with $|r| > 0.85$ as highly correlated. In such cases, one of the redundant variables was removed to reduce feature redundancy. In addition, variance inflation factors (VIF) were computed for all descriptors, with a threshold of VIF > 10 indicating severe multicollinearity. This two-step screening ensured that the retained features provided complementary information and minimized the risk of overfitting in model development.

2.3. Artificial Neural Networks

ANN excel at uncovering intricate, latent relationships directly from experimental data, bypassing the need for explicit mathematical, physical, or chemical models, and are therefore particularly powerful for nonlinear, complex systems [17,18].

Practically, an ANN is nothing more than a computational routine that maps an m-dimensional input vector onto an n-dimensional output vector. Its basic building blocks (neurons) are arranged in layers and linked by weighted connections. Each neuron receives the outputs from every unit in the preceding layer, forms a weighted sum, and converts that aggregated signal into its own output. Specifically, the net input I_j to neuron j is computed as the sum of weighted contributions from all connected neurons i = 1, 2, …, n:

$$I_j = \theta_j + \sum_i w_{ij} o_i \tag{5}$$

Here, o_i denotes the output of neuron i, w_{ij} is the weight that quantifies the connection strength between neurons i and j, and θ_j serves as the bias term for neuron j. The net input I_j is passed through a transfer (activation) function to yield the output o_j, as defined in Equation (6):

$$o_j = \{1 + exp\,(-I_j)\}^{-1} \tag{6}$$

Training proceeds by iteratively adjusting the weights w_{ji} until the network reproduces the target outputs for the largest possible fraction of input patterns. For each training set, the overall error is quantified as follows:

$$E = \frac{1}{2}\sum_j (o_j^{out} - t_j)^2 \tag{7}$$

Here, o_j^{out} and t_j denote the predicted and target outputs for neuron k in the output layer. Throughout training, the weights are adjusted according to a learning rule; we employ Langevin-type error back-propagation with adaptive learning rate parameters [19]:

$$\Delta w_{ji}^{n+1} = -\eta \frac{\partial E}{n_{\partial w_{ji}}} + \alpha \Delta w_{ji}^{n} + \beta n \delta\,(1) \tag{8}$$

2.4. Support Vector Machine

Driven by advances in both technology and research, SVM (an AI paradigm distinct from neural networks) have gained prominence for modeling nonlinear, high-complexity systems without requiring any a priori knowledge of the underlying phenomena.

SVM is among the most widely adopted algorithms for both classification and regression. Ou et al. [20] pioneered their use for pattern recognition, transforming the datasets into learning vectors (each paired with a corresponding output) so that optimal solutions can be located within a nonlinear feature space.

In an SVM model, y is regressed on x through a function $f(x)$ subject to an ε-insensitive loss. Here, W denotes the weight vector, b is the regression bias, and $\emptyset$ represents a kernel function whose role is to define the functional form of $f(x)$.

$$f(x) = W^T.\emptyset(x) + b \tag{9}$$

$$y = f(x) + \varepsilon \tag{10}$$

SVMs address function-approximation tasks by learning from a training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$, where each x_i is an input vector and y_i its associated target value. SVMs encompass two closely related frameworks: support-vector classification (SVC) and support-vector regression (SVR) [21], and have been successfully applied across diverse domains, including trace-element analysis, phase-diagram evaluation, molecular and materials design, cancer diagnostics, and broad areas of chemistry, chemical engineering, and technology.

2.5. Random Forest Method

RF is an ensemble-learning method that aggregates the predictions of numerous decision trees. By averaging the outputs of individual trees for regression, or by majority voting for classification, RF delivers robust performance on both tasks.

2.5.1. Principles of the Random Forest Algorithm

Bootstrap aggregation (bagging) mitigates prediction error by repeatedly resampling the training data with replacement, fitting an independent learner to each replicate, and averaging their outputs. Yet the approach faces two practical hurdles: (I) The ideal number of bootstrap samples is not known in advance, so superfluous learners can erode both speed and efficiency; (II) The resulting models tend to be less diverse than those produced by alternative ensemble strategies.

2.5.2. Random Forest Regression

The datasets were first standardized with Equation (11) and then used to construct the analysis model.

$$z_i = \frac{(x_i - x)}{S_i} \tag{11}$$

Here, denotes the standardized value of variable x_i, x represents the actual variable value, while S_i and Z_i are its arithmetic mean and standard deviation, respectively.

In every tree of a random forest, node splits are selected to minimize the Gini impurity (Equation (12)); and lower Gini values correspond to cleaner data partitions. Splitting proceeds recursively until either a leaf node becomes indivisible or all its samples share the same class, completing the tree's training [22].

$$Gini_{\mathrm{a}}(D, x_j) = \left(\frac{|D_1|}{|D|}\right)\left(1 - \sum_k^K \left(\frac{|C_k|}{|D_1|}\right)^2\right) + \left(\frac{|D_2|}{|D|}\right)\left(1 - \sum_k^K \left(\frac{|C_k|}{|D_2|}\right)^2\right) \tag{12}$$

Let K be the number of classes in the datasets, with C_k denoting the number of samples in class k. Given a training subset D on a certain value according to feature x_j, a node a is created by splitting D on a specific value of feature x, producing the disjoint subsets D_1 and D_2.

RF consist of many deep decision trees, making it impractical to trace the overall decision path by inspecting each tree individually. Since examining individual trees is feasible only when they are simple (with limited depth and features), the overall ensemble is typically treated as a black box. A common way to gain insight is to rank feature importance using the Gini impurity, which is calculated as shown in Equation (13):

$$Im_{ja} = Gini_a - Gini_b - Gini_c \tag{13}$$

Let $Gini_b$ and $Gini_c$ denote the Gini impurities of the two child nodes created by the split, and the importance of feature x_j in the random forest is then given by Equation (14):

$$Im_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in R} Im_{ja} \tag{14}$$

Here, n is the total number of trees and R denotes the set of nodes that split on feature x_j within tree t. Yet Gini-based importance remains a coarse, global metric, offering scant insight into individual predictions. To address this limitation, we incorporate the SHapley Additive exPlanations (SHAP) framework as a finer-grained tool for knowledge extraction, detailed in the following section [22].

In this study, RF models were built with the scikit-learn library (Python 3.7). Training proceeded in three stages [23]: (I) Data were randomly split 70:30 into training and test sets; (II) From the training set, B bootstrap subsamples were drawn with replacement. Each subsample was used to grow an unpruned decision tree; at every node a random subset of m features ($1 \le m \le K$) was examined, and the split that minimized the Gini impurity was retained. After the forest was assembled, the final prediction was obtained by majority vote (classification) or averaging (regression); (III) Hyper-parameters (number of trees, max_depth, min_samples_split, min_samples_leaf) were tuned by 10-fold cross-validation on the training set, and the best configuration was retrained on the full training set and evaluated on the untouched 30% test set. The hyperparameter optimization results are summarized in the Supplementary Materials (Table S1), with the optimal configuration yielding an average cross-validation R^2 of 0.982.

In this study, RF models were built with the scikit-learn library (Python 3.7). Training proceeded in three stages [23]: (I) Data were randomly split 70:30 into training and test sets; (II) From the training set, B bootstrap subsamples were drawn with replacement. Each subsample was used to grow an unpruned decision tree; at every node a random subset of m features ($1 \le m \le K$) was examined, and the split that minimized the Gini impurity was retained. After the forest was assembled, the final prediction was obtained by majority vote (classification) or averaging (regression); (III) Hyper-parameters (number of trees, max_depth, min_samples_split, min_samples_leaf) were tuned by 10-fold cross-validation on the training set, and the best configuration was retrained on the full training set and evaluated on the untouched 30% test set.

2.6. Model Performance Evaluation Metrics

Model performance is routinely quantified by the coefficient of determination (R^2) and the root-mean-square error (RMSE), calculated as shown in Equations (15) and (16).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(Y_i^{exp} - \bar{Y}^{exp}\right)^2} \tag{15}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2} \tag{16}$$

Here, Y_i^{exp} and Y_i^{pred} denote the observed and predicted values, respectively, and $\bar{Y}^{exp}$ is their arithmetic mean.

Three algorithms (ANN, SVM, and RF) were benchmarked on 10,086 data points using scikit-learn 0.21.3. Hyperparameters were optimized with GridSearchCV employing three-fold cross-validation. The optimized models were then further evaluated using 10-fold cross-validation. To compare model performance, paired t-tests were conducted on RMSE and R^2 values, which indicated that RF performed significantly better than the other algorithms. Consequently, RF was selected for subsequent experiments (Scheme 1).

2.7. SHAP Value Principles

The SHAP adopts additive feature attribution: the model's output is expressed as a linear sum of individual feature contributions. By assigning each input feature a SHAP value, the method quantifies its marginal contribution toward the final prediction [24]. The formal definition is as follows:

$$g(Z') = \varphi_0 + \sum_{i=1}^{M}\left(\varphi_i Z_i'\right) \tag{17}$$

Here, g denotes an interpretable model, M is the total number of features, and $Z' \in M$ is the binary representation of a simplified feature vector. Each component Z_i' ($i \in \{1, \ldots, M\}$) indicates whether feature i is included in the computation: $Z_i' = 1$ signals the feature is observed, whereas $Z_i' = 0$ indicates it is omitted. Consequently, the effect of including or excluding a feature is assessed by toggling its corresponding Z_i' between 0 and 1.

The SHAP interaction values (grounded in the SHapley interaction index from cooperative game theory) quantify how pairs of features jointly influence model predictions [24]. Under the data-generating process defined in equation (10), SHAP summary plots highlight influential features and measure their individual contributions [23]. We examined these contributions in two ways: (I) Instance-level interpretation: each SHAP value represents the impact of a single feature on one specific prediction; (II) Global computation: SHapley values were calculated with the Python SHAP library using the optimized models trained on all 10,086 training instances. Besides, the interaction values further expose local dependencies between features. In the interaction plots, the horizontal axis represents the signed SHAP values, where positive values indicate that a feature increases the prediction, negative values indicate a decrease, and the absolute magnitude reflects the effect size. The vertical axis depicts the interaction value, showing how the combined effect of the feature and another covariate varies across their joint range [25].

2.8. Feature Importance Analysis

Individual conditional expectation (ICE) plots extend partial dependence plots (PDPs) by tracing how the predicted outcome changes as a single feature is varied while all others are held fixed for each observation. By displaying these per-instance curves, ICE plots reveal heterogeneous effects that can be masked in the averaged PDP. Specifically, if half the observations in the PDP exhibit a positive slope and the other half a negative one, the PDP would appear flat, misleadingly suggesting no influence [26]. Consequently, ICE is indispensable for assessing true feature impact.

3. Results and Discussion

3.1. Comparison of Molecular Descriptors and Hansen Solubility Parameters

To compare the solubility prediction accuracy of Hansen solubility parameters and molecular descriptors, a number of samples (≥1000 datasets, Database S1 in Supplementary Materials) were used to model the molecular descriptors and Hansen solubility parameters as characteristic variables respectively, and the decision coefficient R^2 and RMSE were calculated under RF algorithm. The results showed that the decision coefficient (R^2) obtained by molecular description modeling was 0.932, and the R^2 obtained by Hansen solubility parameters was 0.988. Although the predicted results of the two models were basically consistent, the molecular descriptors include various operational variables (including but not limited to 10 parameters, which are lipophilicity, hydrogen bond donors atoms number, hydrogen bond acceptor atoms number, hydrogen bonding capacity, pKa, Topological surface area, S-dipolarity/polarizability descriptor, Crystal density, Ionization energy and Gibbs free energy) while Hansen solubility parameters only includes three indexes (δh, δp, δd). Meanwhile, Huang et al. [27] proved that the Hansen solubility parameters could well explain the changing trend of drug solubility. In consequence, considering the complexity of different mixed solvents and drugs, the prediction accuracy of the model and the possibility of reproducibility of the experimental operation and other reasons, Hansen solubility parameters were selected to represent complex micro-environments with different solvents and drugs in different ratios (0.1, … , 0.9) in all subsequent computation.

3.2. Description and Statistical Analysis of Datasets

In the complex micro-environments of different ratios of mixed solvents, a total of 10,086 datasets with 117 drugs were collected and sorted. The violin plot in Figure 1 showed the distribution and central tendency of the data, and simultaneously provided the upper and lower quartiles of the Hansen solubility parameters, drug compositions, operating conditions and mole fraction solubility of different solvents and drugs, which reflects the degree of dispersion of the data. δh1, δp1, and δd1 are the hydrogen bond, polarity, and dispersion of the solvent, respectively. The variable w1 is defined as the ratio of the solvent mixtures. The parameters δh2, δp2 and δd2 describe the hydrogen-bonding, polarity, and dispersion characteristics of the drug, respectively. Meanwhile, a larger box in the Figure 1 indicates a wider distribution of the data, while a smaller one indicates a concentration of the data. Here, the polarity of the solvent ranges from 0 to 26.6 $MPa^{1/2}$, the dispersion values are low and clustered between 15 and 20 $MPa^{1/2}$. Similarly, most of the data points for hydrogen bonds are concentrated between 6.78 and 23.6 $MPa^{1/2}$. The range of action of the different mixtures is uniformly distributed between 0 and 1. The polarity, dispersion, and hydrogen bonding of the drugs ranged from 0–29.18 $MPa^{1/2}$, 1.78–36.6 $MPa^{1/2}$, and 0–65.94 $MPa^{1/2}$. The molecular weight and molecular volume of the drug were concentrated at 280.94 g/mL and 374.41 cm^3/mol, respectively. The temperature, pressure, and pH range from 273.15–338.15 K, 85–270 kPa, and 1.2–7.4, respectively (Figure 1).

To complement these graphical representations, descriptive statistics including the minimum, maximum, mean, median, skewness, and kurtosis of all descriptors were calculated and are reported in the Supplementary Materials (Table S2). These values provide quantitative confirmation of the observed distributions and dispersion patterns, enhancing interpretability while confirming that no extreme irregularities exist that could bias the modeling results. Together, the violin plots and descriptive statistics offer a comprehensive overview of the dataset's scale, variability, and central tendency, thereby strengthening the reliability of subsequent machine learning analysis.

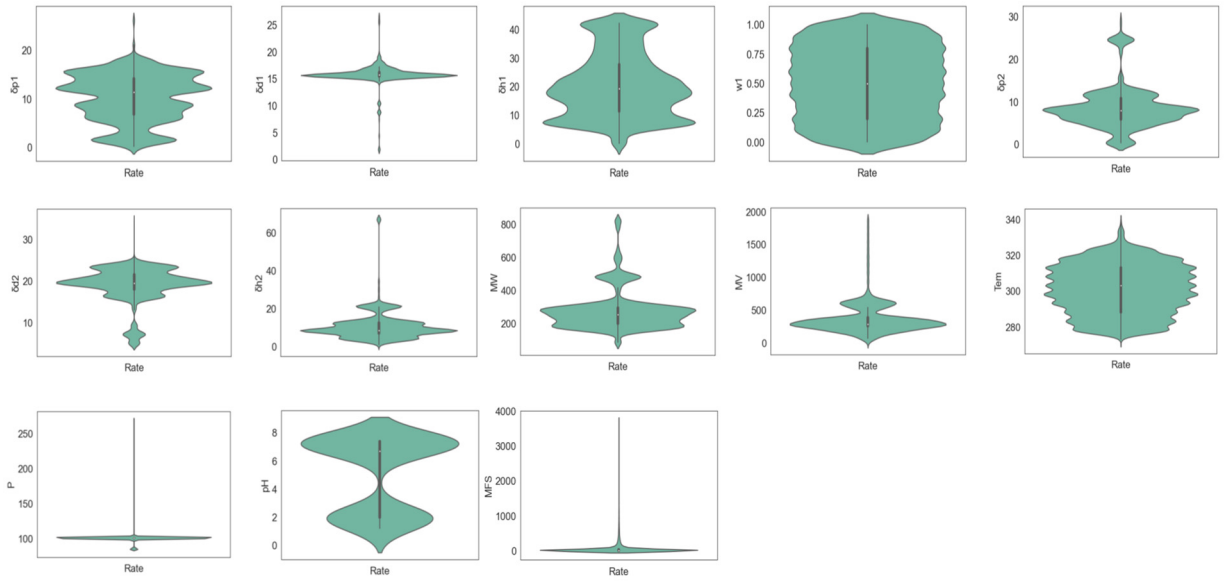


Figure 1. Violin Plot of the total datasets of the various feature variables and target variable.

3.3. Comparison of Different Machine Learning Methods

This study employed twelve feature variables, encompassing HSPs, to construct predictive models for the target variable (MFS) using three machine learning algorithms: ANN, SVM, and RF. Model performance was assessed by R^2 and RMSE (Table 1), with the RF algorithm achieving the highest predictive accuracy, as reflected by its superior training and test R^2 values (0.99 and 0.97, respectively), the lowest RMSE values (0.03 and 0.07, respectively), and notably, the smallest Mean Relative Error (Mean RE, 5.20%) and Standard Deviation of Relative Error (SD of RE, 3.10%) on the test set. These results indicate that RF not only provided the most accurate predictions but also exhibited the greatest stability among the tested models. In contrast, although the ANN model attained a reasonable test ($R^2 = 0.91$), its higher MRE (18.30%) and larger error fluctuation (SD of RE = 10.50%) suggest limitations in its precision and practical utility for this specific prediction task. Besides, RF delivered the best generalization on tabular physicochemical descriptors, reflecting a favorable bias-variance trade-off. In contrast, the ANN (despite regularization and early stopping) exhibited larger train-validation gaps, consistent with overfitting risk on a moderate-sized, heterogeneous tabular dataset. Kernel SVMs were competitive but ultimately below RF, likely due to sensitivity to kernel bandwidth and limited ability to model localized high-order interactions. 10-fold cross-validation (Table 1) corroborate these trends, and SHAP attributions from the RF model align with known physicochemical principles, supporting its superior suitability for this problem. Consequently, all subsequent analyses of feature importance and drug solubility were based on the RF model.

The difference observed between training and test errors reflects the complexity and heterogeneity of the dataset rather than model instability. All models were developed using 10-fold cross-validation, and the average cross-validation ($R^2 = 0.987$) was closely aligned with the independent test ($R^2 = 0.97$), confirming the robustness of the model's generalization ability. The slightly higher training performance can be attributed to the flexibility of Random Forest in capturing nonlinear relationships, as well as the limited dataset size, where the test split may include under-represented or more challenging compounds. Importantly, the feature importance patterns derived from SHAP analysis were consistent with established physicochemical principles, indicating that the RF model captured meaningful structure-property relationships rather than overfitting to noise.

Table 1. Comparison of prediction results obtained by different machine learning methods.

Index Methods	RF	SVM	ANN
Train R^2	0.99	0.96	0.93
Test R^2	0.97	0.91	0.91
Train RMSE	0.03	0.16	1.0
Test RMSE	0.07	0.31	1.1
Mean of RE (%)	5.2	12.5	18.3
SD of RE (%)	3.1	7.8	10.5

3.4. Principal Component Analysis for Datasets Assessment

Figure 2 displays the PCC and their corresponding *p*-values for all variable pairs. The absolute magnitude of each coefficient reflects the strength of the linear relationship, while the *p*-values indicate the statistical significance of correlations among Hansen solubility parameters, drug composition variables, process conditions, and observed solubility in complex micro-environments. As shown in Figure 2, significant linear correlations ($p < 0.01$) were identified between MFS and several features: δd1, δh1, w1, δd2, pH, and temperature. These strong associations suggest that hydrogen bonding exerts a considerable influence on MFS variation. Specifically, solvent hydrogen bonding capacity, solvent mixing ratio, and temperature all showed substantial correlations with drug solubility, with correlation coefficients of 0.64, 0.78, and 0.57, respectively. This finding is consistent with previous reports. Yang et al. [28] highlighted the significant effect of mixed solvent ratio on drug solubility. Similarly, Gao et al. [29] demonstrated that dissolution temperature considerably influences solubility, which may be attributed to differences in temperature sensitivity among drugs and solvents, leading to variations in the rate and extent of solubility change. Besides, the correlations among descriptors were moderate ($|r| = 0.16–0.78$), well below the threshold for severe multicollinearity ($|r| > 0.85$). VIF (all < 10) further confirmed that no problematic redundancy existed. These results indicate that the models were not adversely affected by descriptor correlations, and the feature importance patterns identified by SHAP remained consistent with established physicochemical principles.

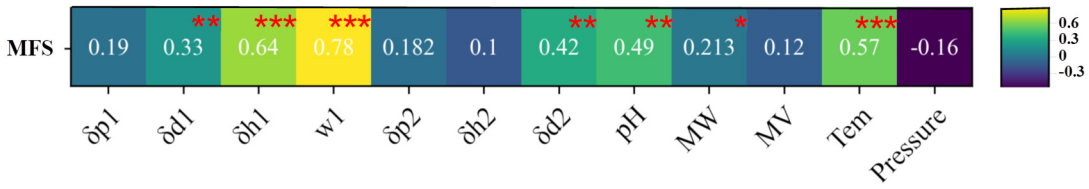


Figure 2. Pearson correlation heatmap of Hansen solubility parameters, drug compositions, operating conditions, and solubility across the entire datasets (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

3.5. Performance Assessment of the RF Model

The hyperparameters of the RF model were obtained through ten cross-validation, as shown in Figure 3A and 3B, respectively, representing the retraining of the solubility prediction model in complex micro-environments without Hansen solubility parameters (only macro variables include drug compositions + operating conditions) and with Hansen solubility parameters (micro variables (δh, δp, δd) + macro variables) under different mixed solvent ratios. 30% of the actual data are used as indicators to evaluate the predictive performance of the developed RF model. Among them, the R^2 score of the RF optimal parameter model obtained through the training set is only 0.830 without Hansen solubility parameters conditions (Figure 3A), while the R^2 score of model with the Hansen solubility parameters is 0.988 (Figure 3B). The results showed that under the condition of complex micro-environments represented by Hansen solubility parameters, the results of RF model established by training and test set data can better evaluate the drug solubility.

The prediction accuracy of the RF model for drug solubility by visualizing the predicted results (Figure 4). As shown in Figure 4A,B, based on without Hansen solubility parameters and with Hansen solubility parameters in Figure 3A,B respectively, 30% of the datasets were randomly selected from the training set to predict the actual values and predicted values. The actual drug mole fraction solubility values were 0–0.789 (Figure 4A) and 0–0.985 (Figure 4B), respectively. The results showed that the predicted values are in good agreement with the actual values. This further proves that the RF algorithm is suitable for predicting the drug solubility under different conditions, and the accuracy of the target analysis results is obtained.

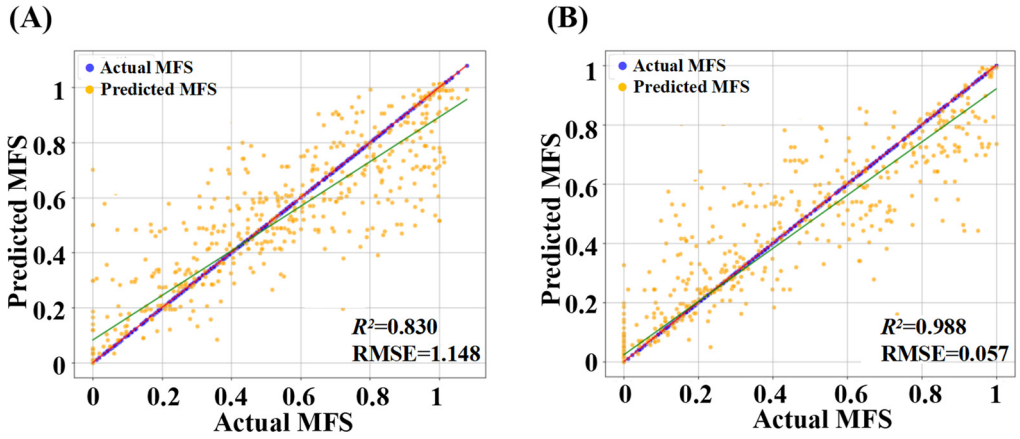


Figure 3. Comparison of the actual and predicted values of drug solubility under different input variables: (**A**) Without Hansen solubility parameters; (**B**) With Hansen solubility parameters. The blue line is the result of a linear relationship based on actual data. The red line indicates the predicted values = the actual values line.

On the other hand, with the purpose of the further evaluate the stability of the prediction model and obtain the specific difference between the actual value and the predicted value, the relative error of Figure 4A,B was calculated. The results showed that for models without Hansen solubility parameters, most of the relative errors are within 20% (Figure 4C). The prediction performance based on with Hansen solubility parameters were better, and most of its relative error results are within 4% (Figure 4D), which further proves that selecting Hansen solubility parameters as a microscopic descriptor to represent the change of drug solubility in complex micro-environments can significantly improve the prediction accuracy of the model.

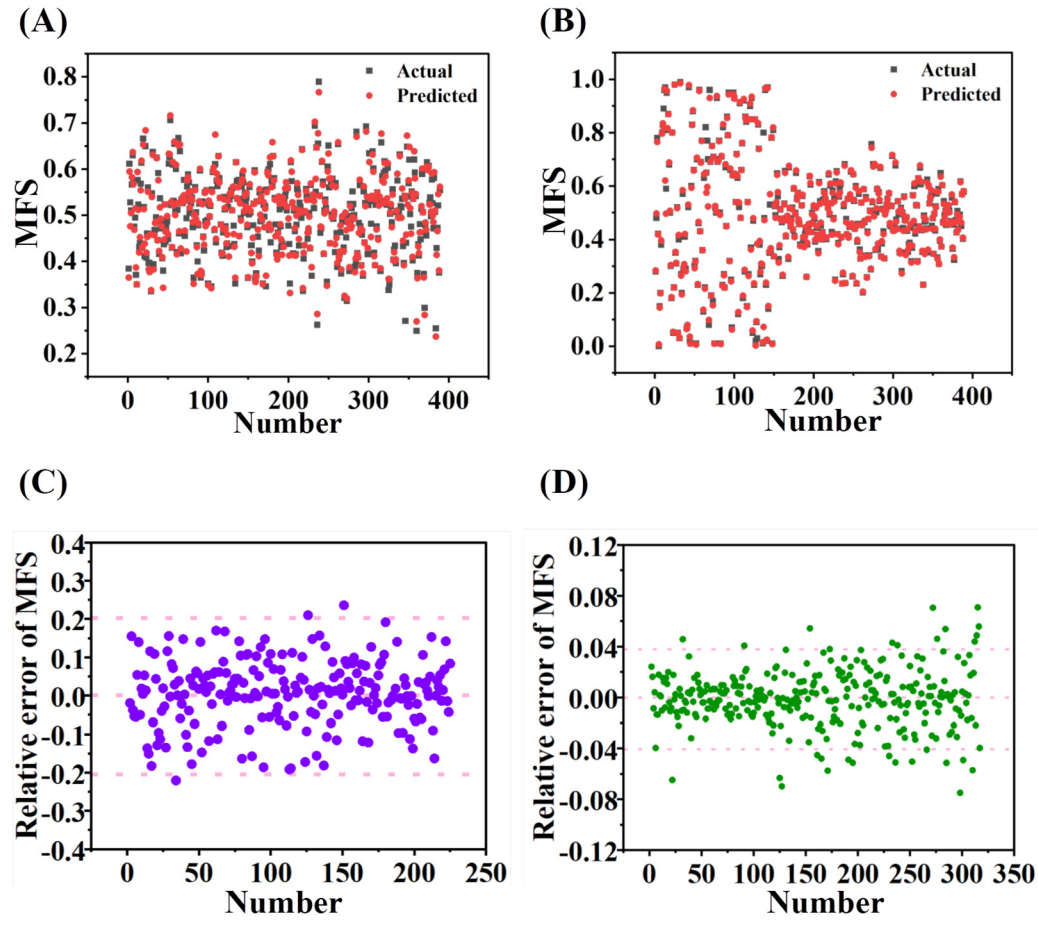


Figure 4. Correlations (**A**,**B**) and relative error (**C**,**D**) between drug solubility datasets based on whether the predicted and actual values of Hansen solubility parameters were contained were obtained from the trained RF model. Without Hansen solubility parameters (**A**,**C**), and with Hansen solubility parameters (**B**,**D**).

3.6. Analysis of Feature Importance

Figure 5A presents the ranked feature importance, based on the tuned RF model, revealing how drug identity, Hansen solubility parameters, composition, and operating conditions contribute (either linearly or nonlinearly) to predicted solubility. As shown in Figure 5A, w1 is the most important, which account for 0.128. Notably, it can be seen that the most important influencing factors are almost all Hansen solubility parameters of different drugs and solvents under the mixed ratios, which indicates that the complex micro-environments is crucial to the influence of drug solubility. Khorsandi et al. [30] demonstrated that Hansen solubility parameters can be used to select suitable cosolvents for drugs. The calculation showed that polarity, dispersion and hydrogen bond were the main reasons for the change of drug solubility in solution. In addition, Gao et al. [29] experimentally found that solvent polarity and hydrogen bonding increased with the organic solvent mass fraction, proving that solvent hydrogen bonding plays a crucial role in drug solubility.

Figure 5B demonstrates that microscopic conditions exert a markedly stronger influence on drug solubility than macroscopic variables, suggesting that alterations in complex micro-environments may substantially improve solubility. Consistent with this, feature attribution in our dataset indicated that descriptors of microscopic solute-solvent affinity (most notably the HSP components (δh, δp, δd)) accounted for a larger share of explained variance in solubility predictions (65%) compared with macroscopic descriptors such as temperature, or bulk medium related properties (35%). This observation aligns with recent reports showing that HSP-based compatibility metrics correlate strongly with drug-polymer miscibility, amorphous solid dispersion performance, and solvent screening [31–33], underscoring the critical role of intermolecular interactions (δh, δp, δd) in governing dissolution and solvation at the molecular scale. Nevertheless, we acknowledge that this conclusion should not be overgeneralized. In systems where polymorphic form, precipitation kinetics, glass-transition-controlled mobility, or pronounced cosolvency and temperature effects prevail, macroscopic conditions may exert comparable or even dominant influences [34–36].

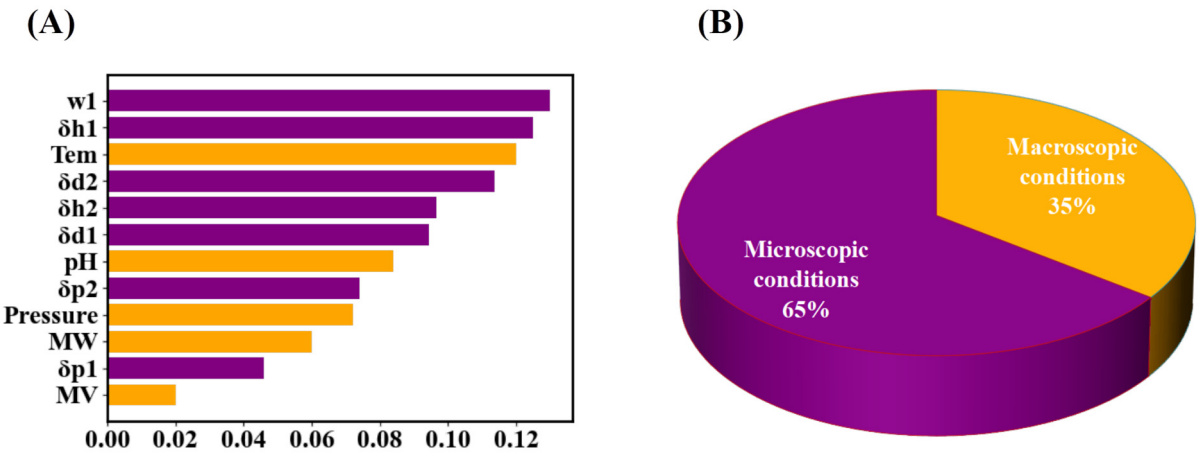


Figure 5. Feature importance analysis of input variables for drug solubility using the RF model: (**A**) Ranked feature importance of individual variables; (**B**) Overall contribution analysis of variable categories. Purple and yellow colors denote microscopic and macroscopic variables, respectively.

3.7. PDP and ICE of Drug Solubility

The PDP and ICE diagrams summarized the influence trend of Hansen solubility parameters of different drugs and solvents, drug compositions and operating conditions on the drug solubility in the RF model under the condition of different ratio of mixed solvents. Figure 6 illustrates univariate partial-dependence relationships for predicting drug solubility. In each panel, one input feature is varied across its observed range while all remaining features are fixed at their mean values. The ICE curves show the resulting prediction for every individual observation, and the PDP (thick line) gives their average, summarizing the overall trend [37]. Bear in mind that the thick curve reflects the mean of model-derived predictions, not the scatter of experimental observations. Because the input feature is evaluated at only a limited set of points, the displayed trend is approximate and can be easily dismissed if viewed superficially.

Figure 6A showed that the solvent with a hydrogen bond value less than 8.12 $MPa^{1/2}$ had a better effect on the drug solubility (>50 mg/mL), which is consistent with the experimental results [28]. And the solvent with a hydrogen bond value more than 8.12 $MPa^{1/2}$ had no significant effect on drug solubility. Khorsandi et al. [30] experimentally demonstrated that drug solubility increased with the mixture solvent ratio, which may be due to

the destruction of local molecular interaction in the non-polar part of the drug by constantly changing the type and mixing ratio of the mixed solvents, thereby reducing crystallization ability and increasing dissolution ability [38]. As shown in Figure 6C, drug solubility increased progressively with rising temperature above 250.15 K. Khorsandi et al. [30] experimentally confirmed this trend, reporting that the solubility of diethylstilbestrol increased more than 80-fold across the examined temperature range. Figure 6D further shows that when the δd2 was 15 $MPa^{1/2}$, solubility initially decreased, then increased to 58 mg/mL, and subsequently remained constant. Li et al. [38] experimentally demonstrated that dispersion forces make a major contribution to solubility, showing that the dispersion force of dapsone crystals was approximately three times more influential than polarity. Especially, the non-monotonic profile observed in Figure 6D can be rationalized by the competing balance between solute-solute and solute-solvent dispersion interactions. At relatively low δd2 values, the solvent's dispersion capacity is insufficient to disrupt hydrophobic aggregation among drug molecules, leading to reduced solubility. As δd2 approaches 15 $MPa^{1/2}$, optimal van der Waals and hydrophobic matching between solute and solvent molecules enhances molecular dispersion and produces a solubility peak. Beyond this threshold, further increases in δd2 mainly strengthen solvent-solvent dispersion interactions rather than improving solute-solvent stabilization, giving rise to the observed plateau. Similar non-linear solubility trends have also been reported in mixed solvent systems, where preferential solvation and competition among δh, δp, and δd forces generate composition-dependent maxima [39,40].

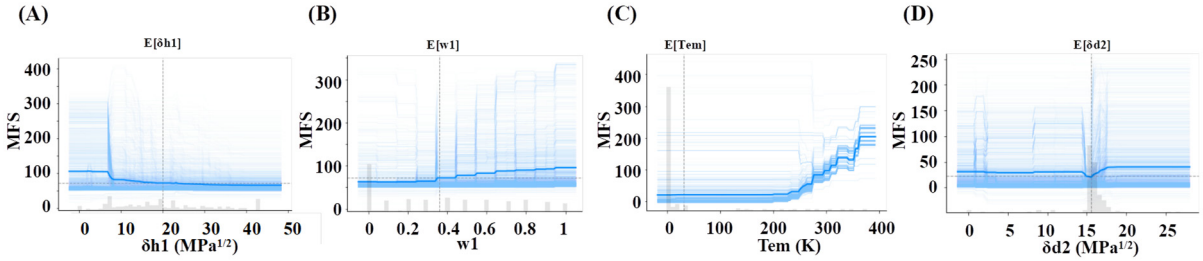


Figure 6. PDP (thick line) and ICE (fine lines) curves for drug solubility. Each PDP shows the mean prediction derived from the ICE curves of a random 30% subset, with the selected input feature on the *x*-axis and predicted solubility on the *y*-axis.

3.8. SHAP Values of the Drug Solubility

Further confirmation of the feature-importance ranking is provided in Figure 7, where SHAP values reveal both the marginal influence of each feature and the joint effects of feature pairs on drug solubility under complex micro-environmental conditions. On the one hand, to better interpret the predicted results of drug solubility, we randomly sampled 30% of the training data to explore how these features affect the model output and observed the prediction scores of the feature variable and the target variable separately, which can evaluate how these values are globally attributed to each feature [41]. Figure 7A,C illustrate the respective contributions of each feature. The base value represents the average mole fraction solubility of the drug, calculated to be 67.09 across all data. In contrast, the final value corresponds to the optimal solubility value of 95.80 observed for the sample under investigation. Red bands signify that the corresponding feature values drive the prediction upward; blue bands drive it downward. The broader the band, the larger the absolute contribution. Cumulatively, these increments shift the target from its baseline to the final output. Whether a feature exerts a positive or negative influence is dictated by its value relative to the threshold learned during training. For a comprehensive evaluation of the constructed model, both its predictive accuracy and the interpretability of the model mechanism must be considered [42].

On the other hand, Figure 7B displays every training instance as a dot, with color (blue to red) encoding the feature value (low to high) and the horizontal position representing its SHAP value. Features are stacked vertically by descending mean |SHAP|, clarifying their relative importance. Each point thus quantifies a feature's contribution to the model's prediction for that sample. Note that interactions are plotted twice: once for each interacting feature, and are flagged with an asterisk (e.g., "w1*-Tem" denotes the interaction between w1 and Tem, colored by the value of Tem).

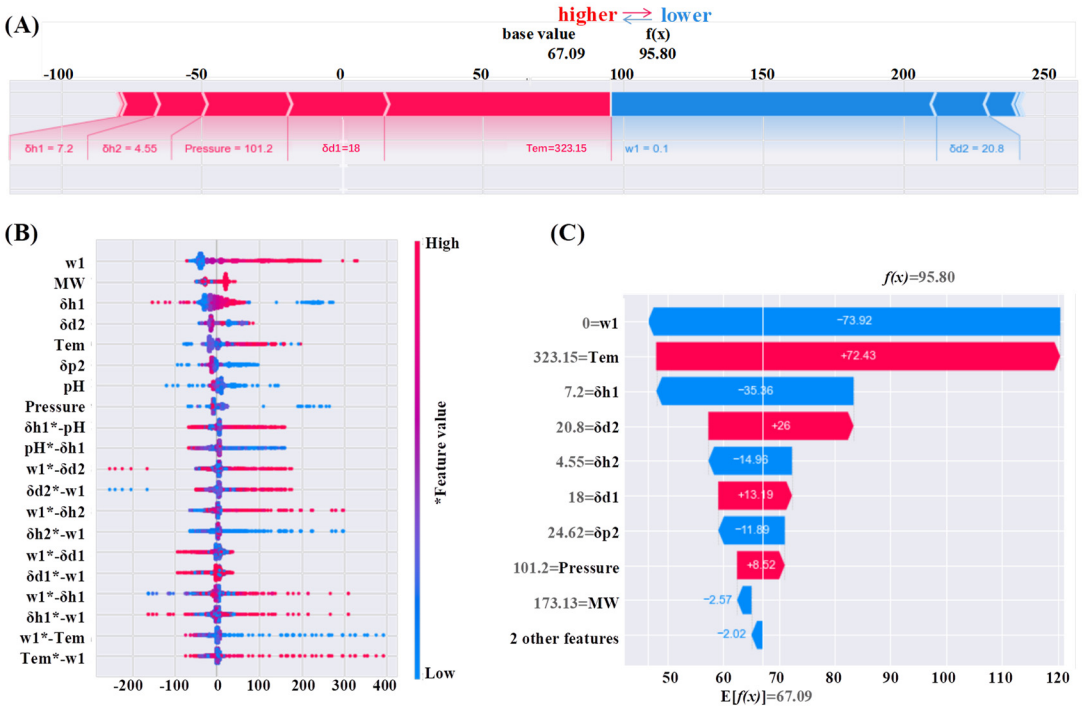


Figure 7. SHAP was used to analyze the effects of drugs in the complex micro-environments based on the RF model. (**A**) The SHAP explanation for a single prediction; (**B**) The SHAP summary plot; (**C**) The SHAP values of Waterfall plot.

3.8.1. Global Interpretability of Model Predictions

As Figure 7A illustrates, δh1, δh2, Pressure, Tem, and δd1 (red bands) most strongly elevate the predicted probability, whereas w1 and δd2 (blue bands) exert the greatest downward influence on the baseline value. Among them, the reason why the scores of solvent and drug hydrogen bonds (δh1 and δh2) are higher than baseline is that temperature (323.15 K), solvent dispersion (4.55 $MPa^{1/2}$) and pressure (101.2 kPa) provide better support for the increase of drug solubility, despite the relatively unfavorable effects of different mixed solvent ratios (0.1) and drug dispersion forces (18 $MPa^{1/2}$) on drug solubility, which is consistent with the experimental results of Yang et al. [28]. It is worth noting that if the length of the pink bar is subtracted from the length of the blue bar, the difference is equal to the distance between the benchmark value and the predicted value.

Figure 7B evidences a positive solubility response to increasing solvent ratio (w1), drug molar mass, and temperature, while δh1 (especially at high values) exerts a marked negative influence; similar adverse trends are observed for elevated δp2, pH, and pressure. These patterns corroborate the univariate importance metrics in Figure 5. Among pairwise effects, the w1 and temperature, w1 and δh1, and w1 and δh2 interactions are the most influential, whereas the w1 and δd1, and solubility and molar-mass couplings are negligible.

Figure 7C traces the SHAP-based attribution pathway from the baseline value (EV) to the final prediction *f(x)*, revealing the relative contribution of each feature along the way [43]. Thus, for the analysis results of different characteristics shown in Figure 7C, its w1 and δh1 have a large negative impact on the drug solubility. However, temperature and δd2 are the characteristics that most contribute to the increased of drug solubility. Li et al. [44] experimentally demonstrated that, while the temperature is constant, for crystalline flutamide, the main contribution of the solubility increase comes from the dispersion force.

3.8.2. SHAP Values Associated with the Feature Importance

Different features show different importance in influencing drug solubility and may involve complex interactions related to the molecular structure, energy changes and other aspects of different solvents and drugs in the micro-environments. Volkova et al. [45] demonstrated that the thermodynamic parameters of the intermolecular forces of drugs and the transfer between different solvents have a crucial effect on the drug solubility. The SHAP decision plot in Figure 8 showed a nonlinear trend of drug solubility with increasing w1, suggesting that increasing w1 in the higher range may lead to a rapid increase in drug solubility, which was consistent with the results in Figure 7B. Meanwhile, the SHAP decision plots revealed a nonlinear relationship between solvent polarity and drug solubility. At lower solvent contents, increasing polarity was associated with a

decrease in solubility. Beyond a certain threshold, however, further increases exerted little to no additional effect. Gao et al. [29] demonstrated that solvent polarity has a nonlinear effect on drug solubility, and the possible reason is that stronger interactions between mixed solvent molecules and higher cohesive energy density values have a greater effect on solvent polarity and thus on solubility.

The SHAP decision plot (Figure 8) revealed that some parameters contributed negatively to solubility predictions, such as the solvent hydrogen-bonding component (δh1). According to HSP theory, this occurs when strong solvent-solvent hydrogen bonding dominates over solvent-drug interactions, particularly for drugs with intrinsically low hydrogen-bonding capacity (low δh2). In such cases, the energetic penalty of breaking solvent-solvent associations is not compensated by weak solvent-drug bonds, leading to reduced solubility and negative SHAP values. Importantly, the three HSP components (δd, δp δh) act cooperatively rather than independently, and their combined effect can be captured by the relative energy difference (RED). A close match between solvent and drug parameters leads to synergistic effects and positive SHAP contributions, whereas mismatches (especially along the hydrogen-bonding axis) result in antagonism and negative contributions. Consistently positive SHAP values for parameters such as w1 further illustrate how increasing the fraction of a favorable solvent enhances solubility. Similar synergistic and antagonistic behaviors of HSP components have been reported in mixed solvent systems, supporting our interpretation [39,40,46].

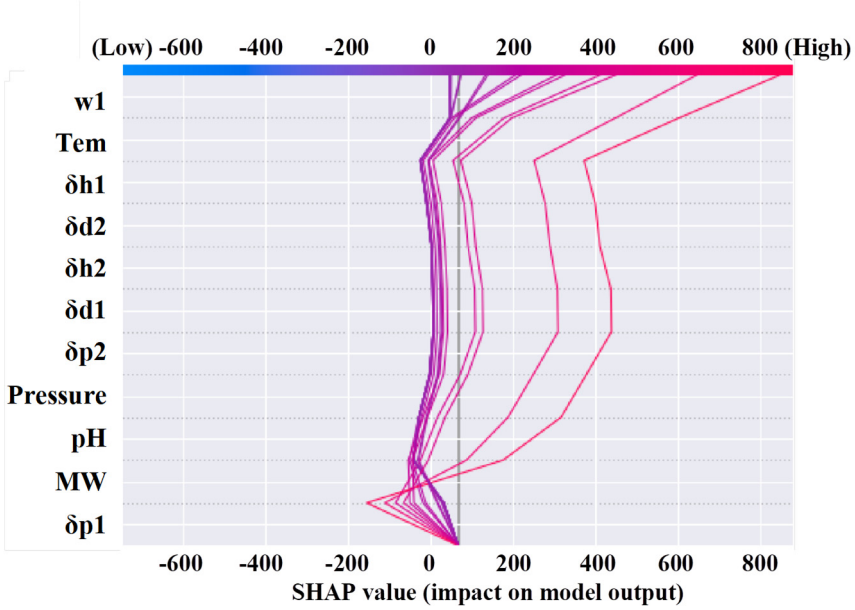


Figure 8. The influence of drug solubility in the complex micro-environments based on the RF model was analyzed by shape decision plot.

4. Conclusions

In summary, this study integrated mechanism-driven microscopic variables (Hansen solubility parameters or molecular descriptors) as inputs in data-driven machine learning methods, drug solubility data in complex micro-environments were collected (more than 10,000 datasets) and predicted successfully by ANN, SVM, and RF methods for the first time. The results showed that the model based on RF algorithm with Hansen solubility parameters had the highest prediction accuracy for drug solubility (R^2 = 0.988). The mechanism-driven microscopic variables (Hansen solubility parameters) showed especially important for drug solubility prediction with the help of SHAP and PDP analysis. Adding mechanism-driven microscopic variables as inputs in data-driven machine learning methods provides new ideas for exploring the influence of complex micro-environments on drug solubility.

Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/2510231652579723/SCE-2508000042-Supplementary-Materials.zip. Table S1: Hyperparameter search space and optimal configuration of the RF model. Table S2: Descriptive statistics of the descriptors used in this study. Database S1: Hansen solubility parameters & Molecular descriptors. Supplementary Information-Codes of Hansen solubility parameters.

Author Contributions

C.W.: Investigation, Conceptualization, Methodology, Data curation, Writing—original draft; X.Z.: Data curation; H.G.: Data curation; Y.J.: Funding acquisition, Supervision; H.Q.: Conceptualization, Funding acquisition, Writing—review & editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded by the National Natural Science Foundation of China (22178391) and the Fundamental Research Funds for the Central Universities (2632025TD08).

Data Availability Statement

All predicted and experimental data (Excel files), together with the Python codes (PDF files) used for model construction and validation of drug solubility, are available in a public GitHub repository at https://github.com/17314455019/Hansen-solubility-parameters.

Conflicts of Interest

The authors affirm that no competing financial interests or personal relationships exist that could have influenced the reported work.

Use of AI and AI-assisted Technologies

During the preparation of this work, the author used Python 3.7 for data processing and modeling predictions. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Pignatello, R.; Corsaro, R.; Bonaccorso, A.; et al. Soluplus ((R)) Polymeric Nanomicelles Improve Solubility of BCS-class II Drugs. *Drug Deliv. Transl. Res*. **2022**, *12*, 1991–2006.
2. Matsui, K.; Tsume, Y.; Amidon, G.E.; et al. In Vitro Dissolution of Fluconazole and Dipyridamole in Gastrointestinal Simulator (GIS), Predicting in Vivo Dissolution and Drug–drug Interaction Caused by Acid-reducing Agents. *Mol. Pharm.* **2015**, *12*, 2418–2428.
3. Eedara, B.B.; Nyavanandi, D.; Narala, S.; et al. Improved Dissolution Rate and Intestinal Absorption of Fexofenadine Hydrochloride by the Preparation of Solid Dispersions: In Vitro and In Situ Evaluation. *Pharmaceutics* **2021**, *13*, 310.
4. Liu, X.W.; Zhao, L.M.; Wu, B.J.; et al. Improving Solubility of Poorly Watersoluble Drugs by Protein-based Strategy: A review. *Int. J. Pharm*. **2023**, *634*, 122704.
5. Savjani, K.T.; Gajjar, A.K.; Savjani, J.K. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm*. **2012**, *1*, 195727.
6. Calmet, H.; Dosimont, D.; Oks, D.; et al. Machine Learning and Sensitivity Analysis for Predicting Nasal Drug Delivery for Targeted Deposition. *Int. J. Pharm*. **2023**, *642*, 123098.
7. Galata, D.L.; Konyves, Z.; Nagy, B.; et al. Real-time Release Testing of Dissolution based on Surrogate Models Developed by Machine Learning Algorithms Using NIR Spectra, Compression Force and Particle Size Distribution as Input Data. *Int. J. Pharm*. **2021**, *597*, 120338.
8. Song, Y.; Ding, Y.; Su, J.; et al. Unlocking the Potential of Machine Learning in Co-crystal Prediction by a Novel Approach Integrating Molecular Thermodynamics. *Angew. Chem. Int. Ed*. **2025**, *64*, e202502410.
9. Cysewski, P.; Przybyłek, M.; Jeliński, T. Intermolecular Interactions as a Measure of Dapsone Solubility in Neat Solvents and Binary Solvent Mixtures. *Materials* **2023**, *16*, 6336.
10. Ge, K.; Ji, Y. Novel Computational Approach by Combining Machine Learning with Molecular Thermodynamics for Predicting Drug Solubility in Solvents. *Ind. Eng. Chem. Res.* **2021**, *60*, 9259–9268.
11. Li, J.E.; Chien, S.C.; Hsieh, C.M. Modeling Solid Solute Solubility in Supercritical Carbon Dioxide by Machine Learning Algorithms Using Molecular Sigma Profiles. *J. Mol. Liq*. **2024**, *395*, 123884.
12. Boobier, S.; Hose, D.R.; Blacker, A.J.; et al. Machine Learning with Physicochemical Relationships: Solubility Prediction in Organic Solvents and Water. *Nat. Commun*. **2020**, *11*, 5753.
13. Wang, T.; Su, C.H. Medium Gaussian SVM, Wide Neural Network and Stepwise Linear Method in Estimation of Lornoxicam Pharmaceutical Solubility in Supercritical Solvent. *J. Mol. Liq*. **2022**, *349*, 118120.

14. Zhao, L.; Wang, Q.; Ma, K. Solubility Parameter of Ionic Liquids: A Comparative Study of Inverse Gas Chromatography and Hansen Solubility Sphere. *Acs. Sustain. Chem. Eng*. **2019**, *7*, 10544–10551.
15. Yap, C.W. PaDEL-descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem*. **2011**, *32*, 1466–1474.
16. Hansen, C.M. *Hansen Solubility Parameters: A User's Handbook*; CRC Press: Boca Raton, FL, USA, 2007.
17. Prasad, V.; Bequette, B.W. Nonlinear System Identification and Model Reduction Using Artificial Neural Networks. *Comput. Chem. Eng*. **2003**, *27*, 1741–1754.
18. Noor, R.M.; Ahmad, Z.; Don, M.M.; et al. Modelling and Control of Different Types of Polymerization Processes Using Neural Networks Technique: A Review. *Can. J. Chem. Eng*. **2010**, *88*, 1065–1084.
19. Kiss, I.Z.; Gaspar, V. Controlling Chaos with Artificial Neural Network: Numerical Studies and Experiments. *J. Phys. Chem. C*. **2000**, *104*, 8033–8037.
20. Ou, J.; Luo, X.; Liu, J.; et al. Predicting Microbial Extracellular Electron Transfer Activity in Paddy Soils with Soil Physicochemical Properties Using Machine Learning. *Sci. China Technol*. Sc. **2024**, *67*, 259–270.
21. Were, K.; Bui, D.T.; Dick, Ø.B.; et al. A Comparative Assessment of Support Vector Regression, Artificial Neural Networks, and Random Forests for Predicting and Mapping Soil Organic Carbon Stocks Across an Afromontane Landscape. *Ecol. Ind*. **2015**, *52*, 394–403.
22. Gao, Y.; Han, H.; Lu, H.; et al. Knowledge Mining for Chiller Faults Based on Explanation of Data-driven Diagnosis. *Appl. Therm. Eng*. **2022**, *205*, 118032.
23. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; et al. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. **2020**, *13*, 6308–6325.
24. Li, Z. Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost. *Comput. Environ. Urban Syst*. **2022**, *96*, 101845.
25. Yang, Y.; Yuan, Y.; Han, Z.; et al. Interpretability Analysis for Thermal Sensation Machine Learning Models: An Exploration Based on the SHAP Approach. *Indoor Air*. **2022**, *32*, e12984.
26. Goldstein, A.; Kapelner, A.; Bleich, J.; et al. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat*. **2015**, *24*, 44–65.
27. Huang, Z.; Sha, J.; Chang, Y.; et al. Solubility Measurement, Model Evaluation and Hansen Solubility Parameter of Ipriflavone in Three Binary Solvents. *J. Chem. Thermodyn*. **2021**, *152*, 106285.
28. Yang, Z.; Shao, D.; Zhou, G. Solubility Parameter of lenalidomide for Predicting the Type of Solubility Profile and Application of Thermodynamic Model. *J. Chem. Thermodyn*. **2019**, *132*, 268–275.
29. Gao, Z.; Li, Z.; Li, M.; et al. Solubility Measurement and Thermodynamic Model Correlation of Propyl Gallate in Pure and Binary Solvents from T = (293.15 to 333.15) K. *J. Mol. Liq.* **2020**, *318*, 114035.
30. Khorsandi, M.; Shekaari, H.; Mokhtarpour, M.; et al. Cytotoxicity of Some Choline-based Deep Eutectic Solvents and Their Effect on Solubility of Coumarin Drug. *European J. Pharm. Sci.* **2021**, *167*, 106022.
31. Patel, K.G.; Maynard, R.K.; Ferguson IV, L.S.; et al. Experimentally Determined Hansen Solubility Parameters of Biobased and Biodegradable Polyesters. *ACS Sustain. Chem. Eng*. **2024**, *12*, 2386–2393.
32. Petříková, E.; Patera, J.; Gorlová, O. Influence of Active Pharmaceutical Ingredient Structures on Hansen Solubility Parameters. *Eur. J. Pharm. Sci*. **2021**, *167*, 106016.
33. Oktay, A.N.; Polli, J.E. Screening of Polymers for Oral Ritonavir Amorphous Solid Dispersions by Film Casting. *Pharm*. **2024**, *16*, 1373.
34. Braga, D.; Casali, L.; Grepioni, F. The Relevance of Crystal Forms in the Pharmaceutical Field: Sword of Damocles or Innovation Tools? *Int. J. Mol. Sci.* **2022**, *23*, 9013.
35. Huang, Z.; Staufenbiel, S.; Bodmeier, R. Kinetic Solubility Improvement and Influence of Polymers on Controlled Supersaturation of Itraconazole-succinic Acid Nano-co-crystals. *Int. J. Pharm*. **2022**, *616*, 121536.
36. Shakeel, F.; Al-Shdefat, R.; Ali, M.; et al. Temperature-dependent Solubilization and Thermodynamic Characteristics of Ribociclib in varied {PEG 400 + Water} Combinations. *BMC Chem.* **2025**, *19*, 79.
37. Chang, S.C.; Chu, C.L.; Chen, C.K.; et al. The Comparison and Interpretation of Machine-Learning Models in Post-Stroke Functional Outcome Prediction. *Diagnostics* **2021**, *11*, 1784.
38. Li, H.X.; Xie, Y.; Xue, Y.; et al. Comprehensive Insight into Solubility, Dissolution Properties and Solvation Behaviour of Dapsone in Co-solvent Solutions. *J. Mole. Liq*. **2021**, *341*, 117403.
39. Rivas-Ozuna, D.A.; Medina, R.; Sánchez, M.; et al. Solubility of Pyrazinamide in 1,4-Dioxane + Water and Ethanol + Water Mixtures: Preferential Solvation and Non-Linear Trends. *J. Solution Chem*. **2024**, *53*, 2027–2041.
40. El Hamd, M.A.; El-Toukhy, M.S. Hydrotropy and Solubility Maxima in Pharmaceutical Systems: Insights into Hydrogen-Bond and Dispersion Network Modulation. *Sustain. Chem. Pharm*. **2024**, *25*, 100593.

41. Xu, C.; Li, H.; Yang, J.; et al. Interpretable Prediction of 3-year All-cause Mortality in Patients with Chronic heart Failure Based on Machine Learning. *BMC Med. Inform. Decis*. **2023**, *23*, 267.
42. Wang, C.; Cheng, Y.; Ma, Y.; et al. Prediction of Enhanced Drug Solubility Related to Clathrate Compositions and Operating Conditions: Machine Learning Study. *Int. J. Pharm*. **2023**, *646*, 123458.
43. Sun, J.; Sun, C.K.; Tang, Y.X.; et al. Application of SHAP for Explainable Machine Learning on Age-Based Subgrou Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification. *Healthcare* **2023**, *11*, 2000.
44. Li, Y.; Li, C.; Gao, X.; et al. Equilibrium Solubility, Inter-and Intra-molecular Interactions and Solvation Performance of Flutamide in Some Aqueous Blended Co-solvents. *J. Chem. Thermodyn.* **2021**, *163*, 106611.
45. Volkova, T.V.; Simonova, O.R.; Levshin, I.B.; et al. Physicochemical Profile of New Antifungal Compound: pH-dependent Solubility, Distribution, Permeability and Ionization Assay. *J. Mole. Liq*. **2021**, *336*, 116535.
46. Aydi, A.; Achoura, K.; Bellakhal, N.; et al. Solubility Profile of Amygdalin in Aqueous Ethanol Mixtures: Cooperative Hydrogen-Bonding and Dispersion Contributions. *Crystals* **2023**, *13*, 112.