Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rapid evaluation of the comprehensive quality of Paeonol/Cyclodextrin supramolecular complexes using CASSA based on near-infrared spectroscopy combined with artificial intelligence algorithm

Bo-yao Zhang
Nanjing University of Chinese Medicine
Xin-li Li
Nanjing University of Chinese Medicine
Cheng Qian
Nanjing University of Chinese Medicine
Si-min Xue
Nanjing University of Chinese Medicine
Zhi-tong Zhang
Nanjing University of Chinese Medicine
Si-yi Liu
Nanjing University of Chinese Medicine
Guo Zhang
Nanjing University of Chinese Medicine
Guo-ping Peng
Nanjing University of Chinese Medicine
Ning Ding
Nanjing University of Chinese Medicine
Yun-feng Zheng
zyunfeng@njucm.edu.cn

Nanjing University of Chinese Medicine

Research Article

Keywords: FT-NIR, Artificial Intelligence Algorithms, Cyclodextrin, CASSA, supramolecular complexes

Posted Date: April 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9167929/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Molecular Diversity on May 9th, 2026. See the published version at https://doi.org/10.1007/s11030-026-11579-2.

Abstract

Preparing volatile component/cyclodextrin supramolecular complexes is a common method for enhancing the stability of volatile components. However, the quality assessment of supramolecular complexes is highly complex. This study first prepared paeonol/cyclodextrin supramolecular complexes and evaluated their overall quality. Then, Fourier Transform Near-Infrared Spectroscopy (FT-NIR) was combined with artificial intelligence (AI) to build a Support Vector Machine Classification (SVM) model and a Partial Least Squares Regression (PLSR) model. The SVM classification model reached 100% accuracy, whereas the PLSR quantitative model demonstrated $R^2 > 0.90$ for both calibration and prediction sets. Results confirm that integrating FT-NIR with AI improves the accuracy and reliability of qualitative/quantitative models. Via the Comprehensive Analysis of Single Spectral Acquisition (CASSA) method, dual detection of the complexes' formation state and concentration was achieved, enabling rapid, comprehensive quality evaluation. This study demonstrates the excellent prospects of combining FT-NIR technology with the preparation of supramolecular complexes.

1. Introduction

Volatile components are an important active ingredient in traditional Chinese medicine. However, because its primary active ingredients are volatile constituents, this property often results in poor stability and susceptibility to oxidation after extraction when used as a medicinal product, making it difficult to control its pharmaceutical quality. To address these issues, an effective approach involves using β-cyclodextrin to package volatile components into supramolecular complexes (also known as inclusion complexes). This method ensures the stability of volatile components while enabling their absorption by the human body; moreover, it is widely used owing to its simple process and low cost. Examples include 2-hydroxypropyl-β-cyclodextrin (HP-β-CD) encapsulating nutmeg (Xi et al., 2023). Yet for questions including the contents of volatile components in supramolecular complexes and whether inclusion has been fully achieved, solutions primarily depend on techniques such as scanning electron microscopy, X-ray diffraction, differential scanning calorimetry (Xu et al., 2017), and HPLC—all of which are often cumbersome, time-consuming, and require both sophisticated equipment and highly skilled operators. Additionally, it remained difficult to determine whether free volatile monomer components were present during HPLC-based content analysis of the complex, thus leading to issues such as inconsistent and unreliable content results. Therefore, for the inclusion of volatile components today, there is an urgent need for a rapid and reliable technique to detect the inclusion state and content of supramolecular complexes.

Near-infrared spectroscopy (NIRS) is a technique that reflects absorption information from the harmonic and sum frequencies of molecular vibrations in chemical groups containing hydrogen elements (such as C-H, O-H, S-H, N-H, etc.). Consequently, this spectroscopic analysis covers nearly all organic compounds and mixtures (Kato et al., 2023; Liu et al., 2023). Furthermore, NIRS technology enables rapid scanning of samples with minimal pretreatment requirements, yielding stable, reliable, and accurate data. Currently, near-infrared technology is widely used in various fields such as rapid quantitative analysis of target components and identification of food and pharmaceutical products (Beć et al., 2022). For example, near-infrared technology can be used to rapidly detect adulteration in ginger powder and quantify its content (Yu et al., 2022). NIRS technology can also be used for rapid quantitative analysis of herbal medicine components. Artificial intelligence algorithms can establish models based on a series of data, gradually arriving at optimal solutions through repeated trial-and-error learning (Yang et al., 2024). Therefore, leveraging the vast data provided by near-infrared scanning technology and continuously optimizing models through artificial intelligence learning ultimately yields a relatively reliable and accurate mathematical model, which is a great solution.

Peony root bark is the dried root bark of *Paeonia suffruticosa Andr.*, a plant belonging to the Ranunculaceae family, possessing rich edible and medicinal value. Its primary component, paeonol, is a volatile compound exhibiting antibacterial, anti-inflammatory, antioxidant, and anticancer effects (Xu et al., 2024; Bai et al., 2022; Yu et al., 2018). Owing to its instability, this study formed a Paeonol/Cyclodextrin supramolecular complex (P-CD complex**)** by cyclodextrin inclusion of paeonol. The quality of the complex was evaluated using techniques such as X-ray diffraction and HPLC. Then, by integrating NIRS technology with artificial intelligence algorithms, a classification model was established capable of distinguishing three types: fully encapsulated paeonol, partially encapsulated paeonol, and free monomer groups (i.e., paeonol completely unencapsulated by cyclodextrin). Using preprocessing methods and feature wavelength extraction techniques, establish a PLSR quantitative model for the complexes in the fully encapsulated group. B = y combining NIRS technology with artificial intelligence algorithms, this approach enables rapid assessment of inclusion complexes formed between cyclodextrin and paeonol, as well as determination of the complex's content. Our institute has developed a method that is rapid and simple to operate, yielding accurate and reliable results with broad applicability. Also, it enables comprehensive quality evaluation through a single spectral acquisition, known as CASSA. This approach provides technical guidance and conceptual insights for comprehensive studies on the properties of supramolecular complexes.

2. Materials and Methods

2.1 Instruments and Reagents

Antaris II FT-NIR Spectrometer (Thermo Fisher Scientific, USA); Waters e2695 HPLC with a 2998PDA detector (Waters, USA) were utilized. Chromatography-grade methanol was purchased from Tedia, USA; X-ray diffractometer (Bruker Corporation, Germany, Model: D8 Advance); Field-emission scanning electron microscope (Hitachi High-Tech Corporation, Japan, Model: SU8600); Fourier transform infrared spectrometer (Thermo Fisher Scientific, U.S.A., Model: Nicolet iS5); Differential scanning calorimeter (TA Instruments, U.S.A., Model: DSC 2500); Thermogravimetric analyzer (TA Instruments, U.S.A., Model: TGA 55);ultrapure water was prepared using a Millipore Milli-Q ultrapure water system; paeonol-rich bark extract monomer (or: Paeonia suffruticosa bark extract monomer, paeonol-containing) was purchased from Ruimao Biotechnology Co., Ltd.; paeonol-rich bark extract reference standard (or: Paeonia suffruticosa bark extract reference standard) was obtained from the China National Institute for Food and Drug Control (Batch No.: 110708–202309, purity ≥ 99.9%); β-cyclodextrin was purchased from Anhui Shanhe Pharmaceutical Excipients Co., Ltd.; all other reagents were of analytical grade and purchased from Sinopharm Chemical Reagent Co., Ltd.

2.2 Sample Preparation

2.2.1 Preparation of completely encapsulated paeonol samples

Take an appropriate amount of paeonol extract, precisely weigh the mass, dissolve in an appropriate amount of anhydrous ethanol, add to a saturated β-cyclodextrin aqueous solution at 50°C, stir for 30 minutes, filter under vacuum after refrigerating for 24 hours, wash the filter cake with ethanol, dry at 60°C for 15 hours, and grind through a No. 4 sieve to obtain the paeonol complex. Add cyclodextrin passed through the No. 4 sieve and mix thoroughly. Prepare 160 portions of the complex (I). As shown in Fig. 1 (I).

2.2.2 Preparation of Partially Encapsulated paeonol Samples

Take the complex prepared in Section 2.2.1, add the ground paeonol monomer (passed through a No. 4 sieve), and mix thoroughly. Prepare 120 samples of partially encapsulated paeonol (II)—i.e., samples where the complex and

paeonol coexist—with the results shown in Fig. 1 (II).

2.2.3 Preparation of Paeonol Monomer and Cyclodextrin Monomer Samples

Take cyclodextrin passed through a No. 4 sieve and ground paeonol monomer passed through a No. 4 sieve separately, then mix them thoroughly. Prepare 40 portions of the mixture of paeonol monomer and cyclodextrin monomer (III). As shown in Fig. 1(III).

2.2.4 Solution preparation

Sample Solution Preparation

Approximately 0.02 g of the paeonol complex was accurately weighed and placed in a stoppered Erlenmeyer flask. Precisely 100 mL of methanol was added, the flask was sealed tightly, and its weight was recorded. After sonication for 20 minutes, the weight loss was replenished with methanol. The mixture was filtered, and 1 mL of the subsequent filtrate was precisely transferred to a 10 mL volumetric flask. Methanol was added to dilute the solution to the mark, and the mixture was thoroughly mixed to obtain the test solution.

Preparation of Reference Solution

Take an appropriate amount of paeonol, precisely weigh the mass, place it in a volumetric flask, dilute to the mark with methanol, and mix thoroughly to prepare a solution containing 20 μg of Paeonol per 1 mL.

2.2.5 Chromatographic conditions

A Hedera ODS-2 column (4.6 mm × 250 mm, 5 μm) was employed. The mobile phase was composed of methanol and water (60:40, v/v). Detection was carried out at a wavelength of 274 nm. The column temperature was maintained at 30°C. The injection volume was set to 10 μL. The flow rate was adjusted to 0.8 mL/min.

2.2.6 Methodological examination

To verify the reliability of this method, validation tests including linearity, precision, repeatability, stability, and spiked recovery were performed in accordance with the guidelines of the International Conference on Harmonisation (ICH). A detailed description of the method is available in Section 1 of the Supplementary Materials.

2.3 Characterization of P-CD complex

The successful formation of the Paeonol/Cyclodextrin supramolecular complex (P-CD complex) was confirmed via scanning electron microscopy (SEM), X-ray diffraction (XRD), Fourier Transform Infrared Spectroscopy(FT-IR) ,and thermogravimetric analysis (TGA).

2.3.1 Differential Scanning Calorimetry

Approximately 5–10 mg of the sample was hermetically sealed in an aluminum pan and analyzed using a differential scanning calorimeter (DSC, TA Instruments, model: DSC 2500). The measurement was conducted from 30 to 295°C at a heating rate of 10°C/min under a constant flow of nitrogen gas (e.g., 50 mL/min) to prevent oxidation. An empty aluminum pan was used as the reference.

2.3.2 X-ray Diffraction

An X-ray diffractometer (Bruker, Germany; model: D8 Advance) was operated at room temperature under the following conditions: Cu Kα radiation, voltage of 40 kV, current of 40 mA, scanning speed of 2°/min, and a scanning range of 2–50°. The obtained patterns were baseline-corrected and smoothed using MDI Jade 6.5, generating the final XRD patterns.

2.3.3 Thermogravimetric Analysis

Thermogravimetric analysis (TGA) was performed on a thermogravimetric analyzer (TA Instruments, model: TGA 55). The measurements were carried out from 30 to 600°C at a heating rate of 20°C/min under a constant nitrogen purge (e.g., 50 mL/min) to prevent sample oxidation.

2.3.4 Fourier Transform Infrared Spectroscopy

An accurate amount (2 mg) of the sample was mixed thoroughly with pure KBr at a mass ratio of 1:100 (sample: KBr), and the mixture was placed in a mold and pressed into a transparent pellet under a hydraulic press. The Fourier transform infrared (FT-IR) spectrum was recorded on an infrared spectrometer (Thermo Fisher Scientific, Inc., Model: Nicolet iS5) over the range of 400–4000 cm^{-1} at a resolution of 4 cm^{-1}, with 32 scans.

2.3.5 Scanning Electron Microscope

A small amount of the sample was evenly spread on a piece of conductive adhesive tape, and any loosely attached particles were removed using a rubber bulb. The sample was then gold sputter-coated for approximately 60 seconds. SEM images were acquired using a field emission scanning electron microscope (Hitachi High-Technologies Corporation, model: SU8600) in the secondary electron (SE) mode at an acceleration voltage of 5 kV.

2.4 Near-infrared spectroscopy acquisition

The samples were scanned and collected using an Antaris II FT-NIR spectrometer (Thermo Fisher Scientific, USA) in diffuse reflection mode. Samples were loaded into Fisher Shell Type 1 Glass quartz sample cells (19 × 51 mm, 2 dram). Spectra were acquired over the wavelength range of 12,000–4,000 cm^{-1} with a resolution of 16 cm^{-1}. Each spectrum was scanned 32 times. For each sample, one background scan was performed followed by three parallel scans. The averaged spectrum was used as the experimental data for subsequent analysis.

2.5 Establishment of Classification Models

The average spectral data acquired via near-infrared spectroscopy (NIRS) exhibits high accuracy without the need for preprocessing or characteristic wavelength selection. Therefore, this study adopted a direct modeling approach that omits data processing, splitting the raw spectral data into a training set and a prediction set at a ratio of 7:3. The data were categorized into three groups: the fully encapsulated group (I), the partially encapsulated group (II), and the free monomer group (III). The optimal classification model was selected using training accuracy and prediction accuracy as evaluation metrics.

2.5.1 Classification Model Selection

(1) Bayesian Optimization is a highly effective global optimization algorithm aimed at finding the global optimum solution. It efficiently addresses the classic machine-intelligent problem in sequential decision theory: determining the next evaluation location based on information gathered about the unknown objective function f, thereby achieving the optimum solution most rapidly (Lorenz et al., 2017). The Bayesian optimization framework can obtain the optimal solution for complex objective functions with a minimal number of evaluations. Essentially, the framework first employs a surrogate model to approximate the true objective function. It then proactively selects the most

"promising" evaluation points for sampling based on this approximation, thereby avoiding unnecessary sampling (Dastin-van Rijn & Widge, 2025). Consequently, Bayesian Optimization is also termed active optimization. Additionally, the framework effectively leverages complete historical information to enhance search efficiency.

(2) Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) are primarily used for processing sequential data. Their defining feature is that the output of a neuron at one time step can be fed back as input to the same neuron in the next time step. This recurrent architecture is highly suited for time series data, as it preserves temporal dependencies within the data. When unfolding an RNN, repetitive structures emerge, and parameters in the architecture are shared—this significantly reduces the number of neural network parameters requiring training. On the other hand, shared parameters also enable the model to scale to data of varying lengths, allowing RNN inputs to be sequences of arbitrary length. For example, when training a fixed-length sentence: a feedforward neural network would assign a separate parameter to each input feature, whereas a recurrent neural network can share the same weight parameters across time steps. Although RNNs were originally designed to learn long-term dependencies, extensive practice has shown that standard RNNs often struggle to retain information over the long term. To address this long-term dependency problem, Hochreiter et al. proposed the LSTM network (Graves & Schmidhuber, 2005) to enhance traditional recurrent neural network models. Today, LSTM has become one of the most effective sequence models in practical applications. Compared to the hidden units in RNNs, those in LSTM have a more complex internal structure: as information flows through the network, the addition of linear interventions enables LSTM to selectively amplify or diminish information intensity (Malashin et al., 2024).

(3) Convolutional Neural Network (CNN). The basic structure of CNN consists of an input layer, convolutional layers, pooling layers (also called sampling layers), fully connected layers, and an output layer. Typically, multiple convolutional and pooling layers are employed, arranged alternately—i.e., one convolutional layer followed by one pooling layer, and so forth. In the convolutional layer, each neuron in the output feature map forms local connections with its local inputs from the previous layer. The neuron's input value is calculated by weighting and summing these local inputs using corresponding connection weights, then adding a bias term. This process is fundamentally equivalent to the convolution operation, from which CNN gets its name (Li et al., 2022; Zang et al., 2021; Sun et al., 2020).

(4) Extreme Learning Machine (ELM) randomly selects the input weights and hidden layer biases of the network, and derives output weights through analytical computation. This effectively overcomes the limitations of traditional SLFN learning algorithms and has been widely applied in various fields such as disease diagnosis, traffic sign recognition, and image quality assessment. Initially limited to single-hidden-layer feedforward neural networks, it was later extended to RBF neural networks, recurrent neural networks, generalized single-hidden-layer feedforward neural networks, and multi-hidden-layer feedforward neural networks. From a design perspective, ELM aims to unify research problems in machine learning—including regression, classification, clustering, compression, and feature extraction—within a single framework (Huang et al., 2015; Chen et al., 2014). In terms of learning efficiency, ELM is simple to implement, has an extremely fast learning speed, and requires minimal human intervention.

(5) Support Vector Machine (SVM) (Hsieh & Yeh, 2011) is a supervised learning algorithm based on statistical learning theory, widely applied in classification and regression tasks. Its core principle involves finding an optimal decision hyperplane that maximizes the margin (i.e., the distance between classes) between samples of different classes, thereby enhancing the model's generalization capability. The key to SVM lies in minimizing structural risk—not only minimizing low error rates on training data but also reducing model complexity by maximizing the classification margin to prevent overfitting. For linearly separable data, SVM employs hard margin optimization to directly determine a hyperplane that perfectly classifies all samples. In contrast, for linearly inseparable data, it

introduces slack variables and a penalty parameter (soft margin, to tolerate minor misclassifications), allowing some samples to be misclassified to enhance model robustness. The SVM optimization problem ultimately transforms into a convex quadratic programming problem, solvable via the Lagrange multiplier method. Its solution is determined by only a few support vectors, and thus exhibits sparsity and computational efficiency. Furthermore, SVM excels in scenarios with small sample sizes and high-dimensional data (e.g., text classification, image recognition), making it one of the classic algorithms in machine learning.

2.6 Establishment of Quantitative Models

Given the extremely high requirements for reliability and accuracy in quantitative models, such models require preprocessing methods—including spectral data preprocessing and characteristic wavelength extraction—to perform dimensionality reduction and optimization on sample data, thereby preventing overfitting. Using the content data of the complex in the fully encapsulated group as the independent variable, a quantitative model was established for the paeonol complex content in this group.

2.6.1 Preprocessing Method

This study used The Unscrambler X 10.4 software (CAMO, Inc., Texas, USA) to preprocess raw spectra. Nine preprocessing methods were applied to the spectra, including first derivative (1st derivative, 1D), second derivative (2nd derivative, 2D), Savitzky-Golay (SG), standard normal variate (SNV), multiplicative scatter correction (MSC), median filter (MF), SNV+1D, MSC+1D, and SG + 1D. These nine methods were evaluated to select the one that yielded the highest model fitting accuracy.

2.6.2 Characteristic Wavelength Selection

(1) Competitive Adaptive Reweighted Sampling (CARS) is a feature selection method based on full-spectrum partial least squares (PLS) regression. By mimicking Darwinian evolution algorithms, it selects the optimal combination of effective variables in the spectrum: specifically, it retains wavelength points with larger absolute regression coefficients in the PLS model while removing those with smaller weights in the same model. Through iterative and competitive operations, it selects N wavelength subsets in each iteration and identifies the subset with the lowest root mean square error of cross-validation (RMSECV) value, which corresponds to the optimal number of variables (Li et al., 2009).

(2) The Iterative Retaining Informative Variables (IRIV) method (Yun et al., 2014) is a feature selection algorithm based on the Binary Matrix Shuffle Filter (BMSF). It comprehensively accounts for the importance of each variable, and after optimization, further screens feature wavelength variables to determine the optimal number of features. As an effective method for selecting optimal feature wavelengths in near-infrared spectroscopy (NIRS), IRIV adopts a strategy of randomly combining variables to account for potential interactions between them, thereby classifying all variables into four categories: strong informative variables, weak informative variables, non-informative variables, and interfering variables. Through multiple iterations—each aimed at retaining strong and weak informative variables while eliminating non-informative and interfering variables—the optimal variable set is ultimately obtained via backward elimination.

(3) Interval Combination Optimization (ICO) is one of the commonly used interval variable selection methods, as it can provide more reasonable wavelength intervals and enhance the predictive capability of models. Proposed within the framework of Model Portfolio Analysis (MPA) combined with Weighted Bootstrap Sampling (WBS), ICO leverages the strengths of both approaches. As a general framework for variable selection and evaluation, MPA offers comprehensive insights from a large number of submodels; its core concept is to establish data analysis methods by

statistically analyzing the distribution of these submodels, making interval selection methods more advantageous than single-wavelength selection approaches. Weighted Bootstrap Sampling (WBS) is a random sampling technique in which different objects are assigned weights based on their occurrence frequencies. As an interval selection method, ICO searches through the interval weights across the entire variable space and selects useful subintervals (Song et al., 2016).

(4) The Successive Projections Algorithm (SPA) is a forward selection method—an iterative forward technique that minimizes multicollinearity—traditionally used for variable selection in multivariate calibration. It starts with a single wavelength and adds a new wavelength in each iteration until a specified number (N) of wavelengths is reached. The optimal number of variables is determined by calculating the root mean square error of cross-validation (RMSECV) for the N-wavelength subset using multiple linear regression (MLR) (Canova et al., 2023; Xiaobo et al., 2010).

2.6.3 Establishment of the PLSR Quantitative Model

Partial Least Squares Regression (PLSR) analysis was performed using MATLAB 2024a software (MathWorks Inc., Natick, MA, USA) to enable quantitative prediction of the compound. The Kennard and Stone (KS) algorithm was employed to divide all samples into a training set and a test set at a 7:3 ratio, establishing the PLSR model for performance prediction.

$$R^2 = 1 - \frac{\sum \left(Yi - \widehat{Yi}\right)^2}{\sum (Yi - Yn)}$$

$$RMSE = \sqrt{\frac{\sum \left(\widehat{Yi} - Yi\right)^2}{n}}$$

$$RPD = \frac{SD}{RMSEP}$$

In the above formulas, *Yi* represents the ith reference value, $\widehat{Yi}$ represents the ith predicted value, and *Yn* represents the mean of n referencevalue.

3. Results and discussion

3.1 Characterization of P-CD complex

3.1.1Differential Scanning Calorimetry

In Fig. 2(A), an endothermic peak for β-cyclodextrin at 92°C, as well as two endothermic peaks for paeonol at 50°C and 224°C, are observed in the free monomer group, indicating no changes after physical mixing. When guest molecules are incorporated into the β-CD cavity, their melting and sublimation points may shift to different temperatures or disappear (Nerome et al., 2013). In the DSC curve of the inclusion complex, the endothermic peak of paeonol at 50°C vanishes, confirming P-CD complex formation.

3.1.2 X-ray Diffraction

As shown in Fig. 2(B), paeonol exhibits strong diffraction peaks in the range of 10° < 2θ < 30°, with notably high intensity at 11.72°, 16.41°, 20.95°, 23.58°, and 25.54°. These peaks are characteristic of paeonol, indicating that it is a crystalline compound. The XRD pattern of the free monomer group (physical mixture of paeonol and β-cyclodextrin)

represents the superposition of the patterns of paeonol monomer and β-cyclodextrin, confirming that it is merely a simple physical mixture without chemical reactions or structural modifications. The absence of paeonol's characteristic peaks in the inclusion complex confirms that paeonol is encapsulated within the β-cyclodextrin cavity to form an inclusion complex, and consequently loses its original crystalline structure (Cao et al., 2023; Betlejewska-Kielak et al., 2021; Guo et al., 2011).

3.1.3 Thermogravimetric Analysis

The TGA curves for each substance are shown in Fig. 2(C). Paeonol exhibits poor thermal stability, with a total mass loss of over 85% at relatively low temperatures (50–177°C). The mass loss of the P-CD complex occurs in three stages: the first stage accounts for approximately 12% mass loss, attributable to the evaporation of surface and internal water; the second stage occurs between 300–328°C, with a mass loss of approximately 77%, representing the main weight loss phase, which is associated with the thermal decomposition of the P-CD complex during this stage. The thermal stability of paeonol was significantly enhanced due to interactions between the guest molecules and the inner cavity of β-cyclodextrin, indicating the formation of the inclusion complex (Li et al., 2016).

3.1.4 Fourier Transform Infrared Spectroscopy

As shown in Fig. 2(D), the infrared spectrum of β-cyclodextrin exhibits four characteristic peaks: a broad hydroxyl (-OH) stretching vibration peak near 3,440 cm^{-1}; methylene (-CH_2) and methyl (-CH_3) stretching vibration peaks near 2,920 cm^{-1}; a peak near 1,642 cm^{-1}, which is attributed to the bending vibration of adsorbed water or hydroxyl groups (not a carbonyl stretch, as β-cyclodextrin contains no carbonyl groups); and an ether bond (-C-O-C) stretching vibration peak near 1,020 cm^{-1}. In the infrared spectrum of paeonol, the broad peak at 2,976 cm^{-1} corresponds to the stretching vibration of the phenolic hydroxyl (-OH) group; the high-intensity absorption peak at 1,620 cm^{-1} represents the stretching vibration of the aromatic ketone carbonyl group, which undergoes a significant bathochromic shift to lower wavenumbers due to conjugation between the benzene ring and the carbonyl group; the strong absorption peak at 1,207 cm^{-1} corresponds to the stretching vibration of the carbon-oxygen (C-O) bond in the aromatic ether. The aromatic ring skeletal stretching vibrations observed in the 1,650–1,430 cm^{-1} region further confirm the presence of the benzene ring. These analytical results are consistent with the theoretical structure of paeonol. The infrared spectrum of the physical mixture exhibits characteristic peaks similar to those of β-cyclodextrin and paeonol, representing a simple superposition of both compounds. This indicates no significant intermolecular interactions during physical mixing (Cao et al., 2023). The disappearance of the characteristic peaks of paeonol in the infrared spectrum of the paeonol/β-cyclodextrin supramolecular complexes indicates the formation of the inclusion complex (Mansi et al., 2025).

3.1.5 Scanning Electron Microscope

As shown in Fig. 3(A-D), β-cyclodextrin (A) appears as irregularly shaped lumps, while paeonol (B) consists of fine crystalline particles. The physical mixture (free monomer group) exhibits the characteristic features of both β-cyclodextrin and paeonol (C). In contrast, the P-CD complex (D) appears as polygonal aggregates with smooth surfaces and a layered structure. Its morphology differs from that of both β-cyclodextrin and paeonol, indicating the formation of the P-CD complex (Zhu et al., 2020).

3.2 Classification of Raw Near-Infrared Spectral Analysis

As shown in Fig. 4(A), no significant differences were observed among the 320 sets of raw near-infrared spectral data—including 160 fully encapsulated samples, 120 incompletely encapsulated samples, and 40 mixed monomer samples—making it difficult to distinguish between fully encapsulated and non-fully encapsulated samples. A

principal component analysis (PCA) model was established using all raw spectral data, and the results are presented in Fig. 4(B). Although the three sample categories could be preliminarily classified—with Principal Component 1 (PC1) contributing 88.8% and the total cumulative contribution reaching 99.1%—complete differentiation of the three categories remained unattainable. Furthermore, raw near-infrared spectra often contain extensive sample information, including irrelevant data that may interfere with quantitative spectral analysis. Therefore, we preprocessed the raw spectral data and combined it with artificial intelligence algorithms for characteristic wavelength selection, ultimately establishing a partial least squares regression (PLSR) model to detect the content of the P-CD complex.

3.3 Classification Model Development

The experiment employed LSTM-Bayesian optimization, CNN-Bayesian optimization, ELM, and an optimizable SVM model to establish classification models using raw spectral data. The results are shown in Table 1:

Table 1
Classification Model Accuracy

Model	**Training set accuracy**			**Prediction Set Accuracy**		
	Class □	**Class □**	**Class □**	**Class □**	**Class □**	**Class □**
LSTM	100%	98.9%	92.9%	100%	100%	87.5%
CNN	98.1%	98.9%	92.9%	100%	100%	93.3%
ELM	100%	97.9%	100%	96.4%	100%	100%
SVM	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
Note: Bold text indicates the best model for each category.						

Among these, the optimal SVM model obtained after 30 iterations achieved a confusion matrix for the three-class classification shown in Fig. 5. Each class attained 100% accuracy, with identical accuracy rates for both the training and prediction sets, indicating no overfitting. Among the four selected classification models, the optimizable SVM model demonstrated the highest accuracy and good reliability. The other models failed to completely and correctly classify the three categories. Therefore, this study proposes to adopt the SVM model as the classification model.

3.4 Preprocessing of raw spectral data

3.4.1 Preprocessing of Fully Enclosed Group Spectral Data

By nine preprocessing methods, PLSR models were established. Among them, the four methods—1D, 2D, MSC, and SNV—exhibited higher R^2 values as shown in Table-2, demonstrating more pronounced effects. The preprocessed spectra are depicted in Fig. 6.

Table 2
Complete Enclosure Group Preprocessing

Preprocessing	A	R^2_T	R^2_P	RPD
1D	**13**	**1.00**	**0.94**	**4.08**
2D	**12**	**0.99**	**0.95**	**4.54**
MF	13	1.00	0.90	3.67
MSC	**9**	**0.97**	**0.92**	**3.71**
MSC+1D	39	1.00	0.42	1.45
SG	14	1.00	0.93	3.94
SG+1D	25	1.00	0.38	1.46
SNV	**23**	**1.00**	**0.91**	**3.52**
SNV+1D	39	1.00	-0.13	1.10

Note: Bold text indicates the selected preprocessing method.

3.5 Characteristic Wavelength Selection

After preprocessing near-infrared hyperspectral data, feature wavelengths are selected using multi-intelligence algorithms to extract relevant spectral variables. Models are constructed based on CARS-I RIV, CARS-SPA, and ICO feature wavelength algorithms, and their predictive capabilities are evaluated.

3.5.1 Fully Enclosed Group Characteristic Wavelength

For the fully encapsulated sample group, the results of three feature wavelength extraction methods—CARS-SPA, CARS-IRIV, and ICO—are presented in Table 4. Among these methods, the combination of ICO (feature wavelength extraction) and SNV (spectral preprocessing) yielded the following model parameters: R^2_T=0.9757, R^2_P=0.9476, RMSEP/RMSEC = 1.04, and RPD = 4.43. These parameter results indicate that the model is accurate and reliable, with excellent predictive performance and a reasonable fitting level.

Preliminary dimensionality reduction of the raw spectral data via preprocessing simplified data complexity. Subsequently, during the iterations of the ICO algorithm, the weight coefficients of each wavelength band changed as the number of iterations increased. A yellower color indicates a weight coefficient closer to 1, while a bluer color indicates a weight coefficient closer to 0; if the color is between blue and yellow, the weight coefficient falls between 0 and 1. The width of the finally selected wavelength intervals (Fig. 7I-A) was automatically optimized using a local search strategy integrated into the ICO algorithm. A total of 465 feature wavelengths were selected (Fig. 7I-C), which significantly reduced information complexity compared to the original 1557 wavelengths and further reduced data dimensionality. This indicates that the ICO algorithm not only rapidly extracts spectral data from complex compound samples but also refines spectral information based on this foundation.

In summary, selecting the ICO + SNV algorithm as the PLSR quantitative model for establishing a complete complex is the most effective and reliable approach.

Table 3
Evaluation of PLSR Parameters for Complete Inclusion Complexes

Method	Preprocessing	A	Variables	Calibration		Prediction		RPD	RMSEP/RMSEC
				RSMEC	R^2_T	RMSEP	R^2_P		
ICO	1D	7	310	4.18	0.9648	4.84	0.9021	3.20	1.16
	2D	37	366	0.00	1.0000	4.40	0.9167	3.53	82309.25
	MSC	26	456	0.01	1.0000	5.48	0.8734	2.83	743.23
	SNV	**11**	**465**	**3.46**	**0.9757**	**3.59**	**0.9476**	**4.43**	**1.04**
IRIV	1D	3	4	7.80	0.8801	9.12	0.6123	1.61	1.17
	2D	6	7	7.04	0.9023	8.71	0.6466	1.68	1.24
	MSC	22	38	1.86	0.9931	2.74	0.9658	5.46	1.48
	SNV	5	7	3.66	0.9733	5.85	0.8415	2.52	1.60
SPA	1D	4	9	7.03	0.9025	8.45	0.6488	1.71	1.20
	2D	6	9	7.00	0.9034	9.40	0.5547	1.52	1.34
	MSC	20	29	2.64	0.9862	3.37	0.9471	4.35	1.28
	SNV	6	9	4.09	0.9670	4.76	0.8943	3.12	1.16

Note: Bold text indicates the optimal PLSR model results.

3.6 Visualization of PLSR Quantitative Models

3.6.1 Complete Enclosure Group PLSR Analysis

As shown in Fig. 8, the scatter plots of the PLSR models for the P-CD complex (Fig. 8A&B) reveal that all samples cluster closely around the regression lines—indicating the successful establishment of the regression models and strong predictive capability. Additionally, the high degree of overlap between the red and blue lines demonstrates the close agreement between the predicted values and reference values of the samples (Fig. 8D&E). Collectively, these results confirm that the model established for the fully encapsulated group samples (following preprocessing and characteristic wavelength selection) is reliable and possesses robust predictive capability. Therefore, this study ultimately adopted the SNV + ICO-PLSR method to establish a quantitative model for the complex content in the fully encapsulated complex group.

4. Conclusion

In this study, the combination of spectral data obtained via FT-NIR rapid scanning and deep learning multi-intelligence algorithms not only addressed the difficulty in evaluating the inclusion state of Paeonol after cyclodextrin encapsulation but also resolved inaccuracies in complex content detection. This approach enabled comprehensive analysis of single spectral acquisition (CASSA) while improving both reliability and accuracy. First of all, categorical analysis of the raw NIR spectral data successfully achieved the classification of three inclusion states: complete inclusion, incomplete inclusion, and free monomer. Second, the raw spectral data were subjected to dimensionality reduction and optimization using preprocessing and characteristic wavelength selection methods, which enhanced

the predictive capability and accuracy of the regression model. A PLSR quantitative model for P-CD complex content was established for the fully encapsulated group. Based on the model parameters, this established model exhibits good linearity and predictive capability, demonstrating high reliability.

In conclusion, the FT-NIR technology combined with artificial intelligence algorithms employed in this study establishes predictive models for cyclodextrin-included volatile components. Compared to traditional methods for complex characterization and content detection, this approach enables comprehensive, rapid, and accurate multi-parameter evaluation of complex quality in a single measurement. It facilitates rapid tracking of complexes, ensures their pharmaceutical stability, and provides a more efficient solution for establishing quality standards for these compounds.

CRediT authorship contribution statement

Bo-yao Zhang: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft. **Xin-li Li**: Data curation, Formal analysis, Software, Project administration. **Cheng Qian**: Methodology, Validation, Visualization. **Si-min Xue**: Conceptualization, Investigation, Data curation. **Zhi-tong Zhang**: Project administration, Methodology. **Si-yi Liu**: Visualization. **Guo Zhang**: Data curation. **Guo-ping Peng**: Supervision, Writing – review and editing. **Ning Ding**: Supervision, Formal analysis. **Yun-feng Zheng**: Writing – review and editing, Methodology, Funding acquisition.

Declarations

Declaration of competing interest

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

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research study was supported by the National Natural Science Foundation of China (No. 81473390),and Jiangsu “333” project.

Author Contribution

Bo-yao Zhang: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft. Xin-li Li: Data curation, Formal analysis, Software, Project administration. Cheng Qian: Methodology, Validation, Visualization. Si-min Xue: Conceptualization, Investigation, Data curation. Zhi-tong Zhang: Project administration, Methodology. Si-yi Liu: Visualization. Guo Zhang: Data curation. Guo-ping Peng: Supervision, Writing – review and editing. Ning Ding: Supervision, Formal analysis. Yun-feng Zheng: Writing – review and editing, Methodology, Funding acquisition.

Data availability

The data that has been used is confidential.

References

1. Bai D, Li X, Wang S, Zhang T, Wei Y, Wang Q, Dong W, Song J, Gao P, Li Y, Wang S, Dai L (2022) Advances in extraction methods, chemical constituents, pharmacological activities, molecular targets and toxicology of volatile oil from Acorus calamus var. angustatus Besser. Front Pharmacol 13:1004529
2. Beć KB, Grabska J, Huck CW (2022) Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives. Foods, *11*(10)
3. Betlejewska-Kielak K, Bednarek E, Budzianowski A, Michalska K, Maurin JK (2021) Comprehensive Characterisation of the Ketoprofen-β-Cyclodextrin Inclusion Complex Using X-ray Techniques and NMR Spectroscopy. Molecules, *26*(13)
4. Canova LDS, Vallese FD, Pistonesi MF, de Araújo Gomes A (2023) An improved successive projections algorithm version to variable selection in multiple linear regression. Anal Chim Acta 1274:341560
5. Cao C, Deng C, Xuan F, Zhou Y (2023) Structural characterization and molecular dynamics simulation of inclusion complex of large ring cyclodextrin and alpha-tocopherol. China Oils Fats 48(5):49–55 Article 1003–7969(2023)48:5
6. Chen H, Peng J, Zhou Y, Li L, Pan Z (2014) Extreme learning machine for ranking: generalization analysis and applications. Neural Netw 53:119–126
7. Dastin-van Rijn EM, Widge AS (2025) Failure modes and mitigations for Bayesian optimization of neuromodulation parameters. J Neural Eng, *22*(3)
8. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610
9. Guo P, Su Y, Cheng Q, Pan Q, Li H (2011) Crystal structure determination of the β-cyclodextrin-p-aminobenzoic acid inclusion complex from powder X-ray diffraction data. Carbohydr Res 346(7):986–990
10. Hsieh TJ, Yeh WC (2011) Knowledge discovery employing grid scheme least squares support vector machines based on orthogonal design bee colony algorithm. IEEE Trans Syst Man Cybern B Cybern 41(5):1198–1212
11. Huang G, Huang GB, Song S, You K (2015) Trends in extreme learning machines: a review. Neural Netw 61:32–48
12. Kato M, Kim J, Oh J, Shimizu D, Fukui N, Shinokubo H (2023) Near-Infrared-Responsive Hydrocarbons Designed by π-Extension of Indeno[1,2,3,4-pgra] perylene at the 1,2,12-Positions. Chemistry, 29(23), e202300249
13. Li H, Liang Y, Xu Q, Cao D (2009) Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal Chim Acta 648(1):77–84
14. Liu CH, He Z, Ruchlin C, Che Y, Somers K, Perepichka DF (2023) Thiele's Fluorocarbons: Stable Diradicaloids with Efficient Visible-to-Near-Infrared Fluorescence from a Zwitterionic Excited State. J Am Chem Soc 145(29):15702–15707
15. Li Y, Li F, Cai HY, Chen X, Sun W, Shen WY (2016) Structural characterization of inclusion complex of arbutin and hydroxypropyl-β-cyclodextrin. Trop J Pharm Res 15(10):2227–2233
16. Li Z, Liu F, Yang W, Peng S, Zhou J (2022) A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Trans Neural Netw Learn Syst 33(12):6999–7019
17. Lomarat P, Phechkrajang C, Sunghad P, Anantachoke N (2024) Raman spectroscopy coupled with the PLSR model: A rapid method for analyzing gamma-oryzanol content in rice bran oil. Food Chem X 24:101923
18. Lorenz R, Hampshire A, Leech R (2017) Neuroadaptive Bayesian Optimization and Hypothesis Testing. Trends Cogn Sci 21(3):155–167
19. Malashin I, Tynchenko V, Gantimurov A, Nelyub V, Borodulin A (2024) Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. Polym (Basel), *16*(18)

20. Mansi, Khanna P, Yadav S, Singh A, Khanna L (2025) Inclusion complexes of novel formyl chromone Schiff bases with β-Cyclodextrin: Synthesis, characterization, DNA binding studies and in-vitro release study. Carbohydr Polym, *347*, 17, Article 122667.
21. Nerome H, Machmudah S, Wahyudiono, Fukuzato R, Higashiura T, Youn YS, Lee YW, Goto M (2013) Nanoparticle formation of lycopene/β-cyclodextrin inclusion complex using supercritical antisolvent precipitation. J Supercrit Fluids 83:97–103
22. Sankom A, Mahakarnchanakul W, Rittiron R, Sajjaanantakul T, Thongket T (2021) Detection of Profenofos in Chinese Kale, Cabbage, and Chili Spur Pepper Using Fourier Transform Near-Infrared and Fourier Transform Mid-Infrared Spectroscopies. ACS Omega 6(40):26404–26415
23. Song X, Huang Y, Yan H, Xiong Y, Min S (2016) A novel algorithm for spectral interval combination optimization. Anal Chim Acta 948:19–29
24. Sun Y, Xue B, Zhang M, Yen GG, Lv J (2020) Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. IEEE Trans Cybern 50(9):3840–3854
25. Wu K, Ma F, Wei C, Gan F, Du C (2023) Rapid Determination of Nitrate Nitrogen Isotope in Water Using Fourier Transform Infrared Attenuated Total Reflectance Spectroscopy (FTIR-ATR) Coupled with Deconvolution Algorithm. Molecules, *28*(2)
26. Xiaobo Z, Jiewen Z, Povey MJ, Holmes M, Hanpin M (2010) Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 667(1–2):14–32
27. Xi X, Huang J, Zhang S, Lu Q, Fang Z, Li C, Zhang Q, Liu Y, Chen H, Liu A, Liu S, Wang C, Li S, Hu B (2023) Preparation and characterization of inclusion complex of Myristica fragrans Houtt. (nutmeg) essential oil with 2-hydroxypropyl-β-cyclodextrin. Food Chem 423:136316
28. Xu F, Yang Q, Wu L, Qi R, Wu Y, Li Y, Tang L, Guo DA, Liu B (2017) Investigation of Inclusion Complex of Patchouli Alcohol with β-Cyclodextrin. PLoS ONE, 12(1), e0169578
29. Xu Y, Bian S, Shang L, Wang X, Bai X, Zhang W (2024) Phytochemistry, pharmacological effects and mechanism of action of volatile oil from Panax ginseng C.A.Mey: a review. Front Pharmacol 15:1436624
30. Xu Y, Liu J, Sun Y, Chen S, Miao X (2023) Fast detection of volatile fatty acids in biogas slurry using NIR spectroscopy combined with feature wavelength selection. Sci Total Environ 857(Pt 1):159282
31. Yang J, Ma X, Guan H, Yang C, Zhang Y, Li G, Li Z, Lu Y (2024) A quality detection method of corn based on spectral technology and deep learning model. Spectrochim Acta Mol Biomol Spectrosc 305:123472
32. Yu DX, Guo S, Zhang X, Yan H, Zhang ZY, Chen X, Chen JY, Jin SJ, Yang J, Duan JA (2022) Rapid detection of adulteration in powder of ginger (Zingiber officinale Roscoe) by FT-NIR spectroscopy combined with chemometrics. Food Chem X 15:100450
33. Yu X, Sun S, Guo Y, Liu Y, Yang D, Li G, Lü S (2018) Citri Reticulatae Pericarpium (Chenpi): Botany, ethnopharmacology, phytochemistry, and pharmacology of a frequently used traditional Chinese medicine. J Ethnopharmacol 220:265–282
34. Yun YH, Wang WT, Tan ML, Liang YZ, Li HD, Cao DS, Lu HM, Xu QS (2014) A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. Anal Chim Acta 807:36–43
35. Zang B, Ding L, Feng Z, Zhu M, Lei T, Xing M, Zhou X (2021) CNN-LRP: Understanding Convolutional Neural Networks Performance for Target Recognition in SAR Images. *Sensors (Basel)*, *21*(13)
36. Zhu W, Wu J, Guo X, Sun X, Li Q, Wang J, Chen L (2020) Development and physicochemical characterization of chitosan hydrochloride/sulfobutyl ether-β-cyclodextrin nanoparticles for cinnamaldehyde entrapment. J Food Biochem, 44(6), e13197

Figures

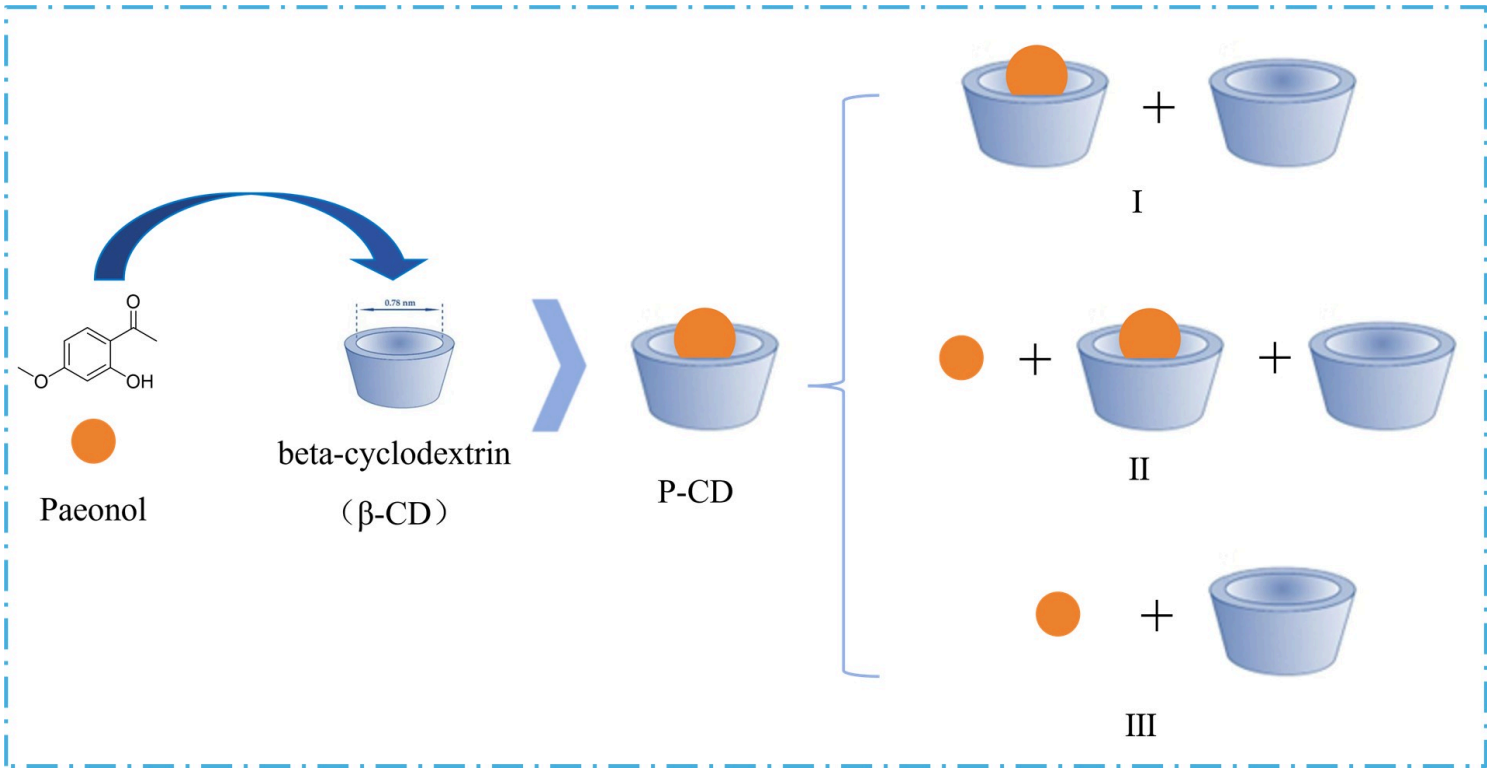


Figure 1

Inclusion Complex Preparation

:Complete Inclusion Group : Incomplete Enclosure Group : Free Monomer Group

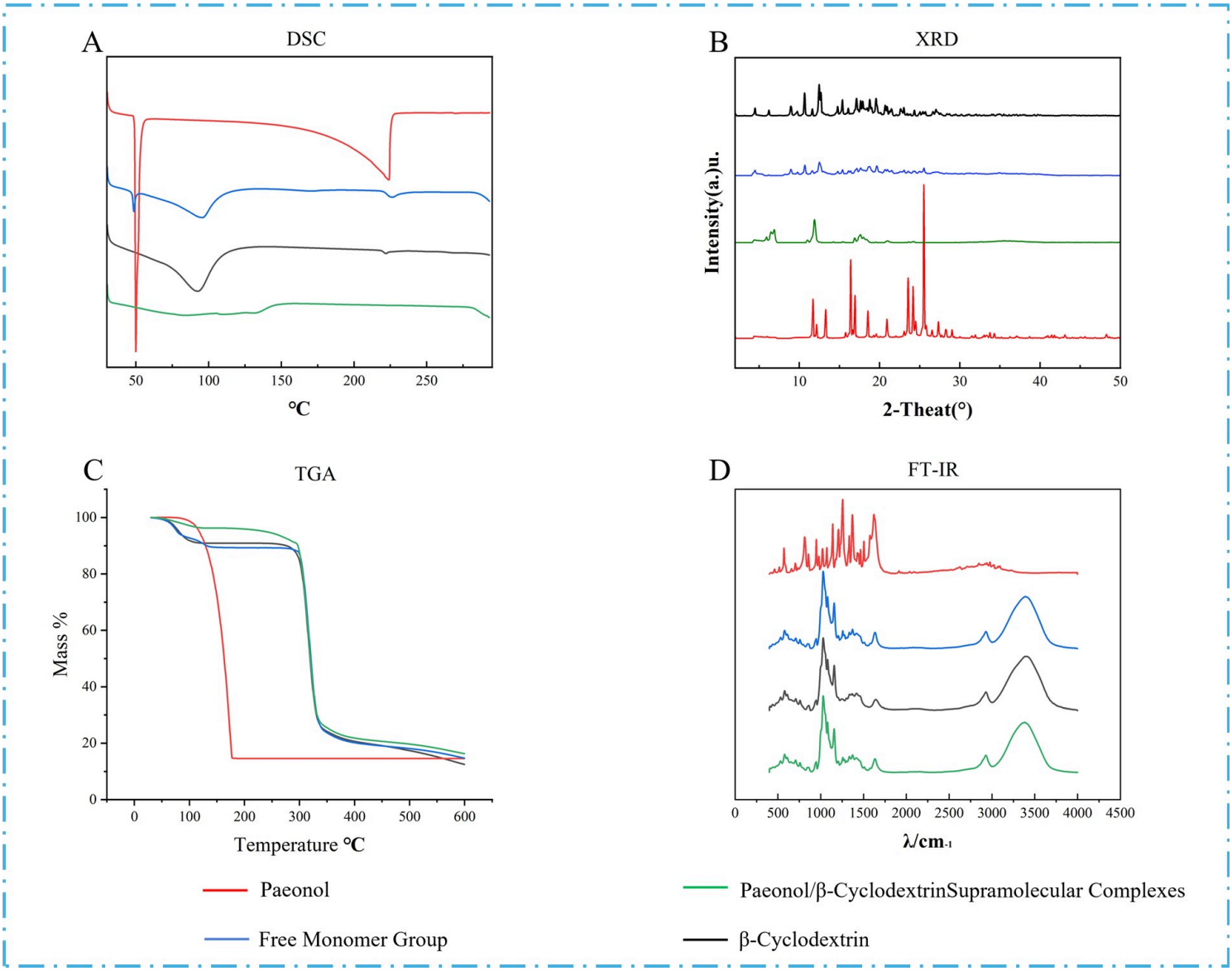


Figure 2

Characterization of Supramolecular Complexes

A: Differential Scanning Calorimetry B: X-ray diffraction

C: Thermogravimetric Analysis D: Fourier Transform Infrared Spectroscopy

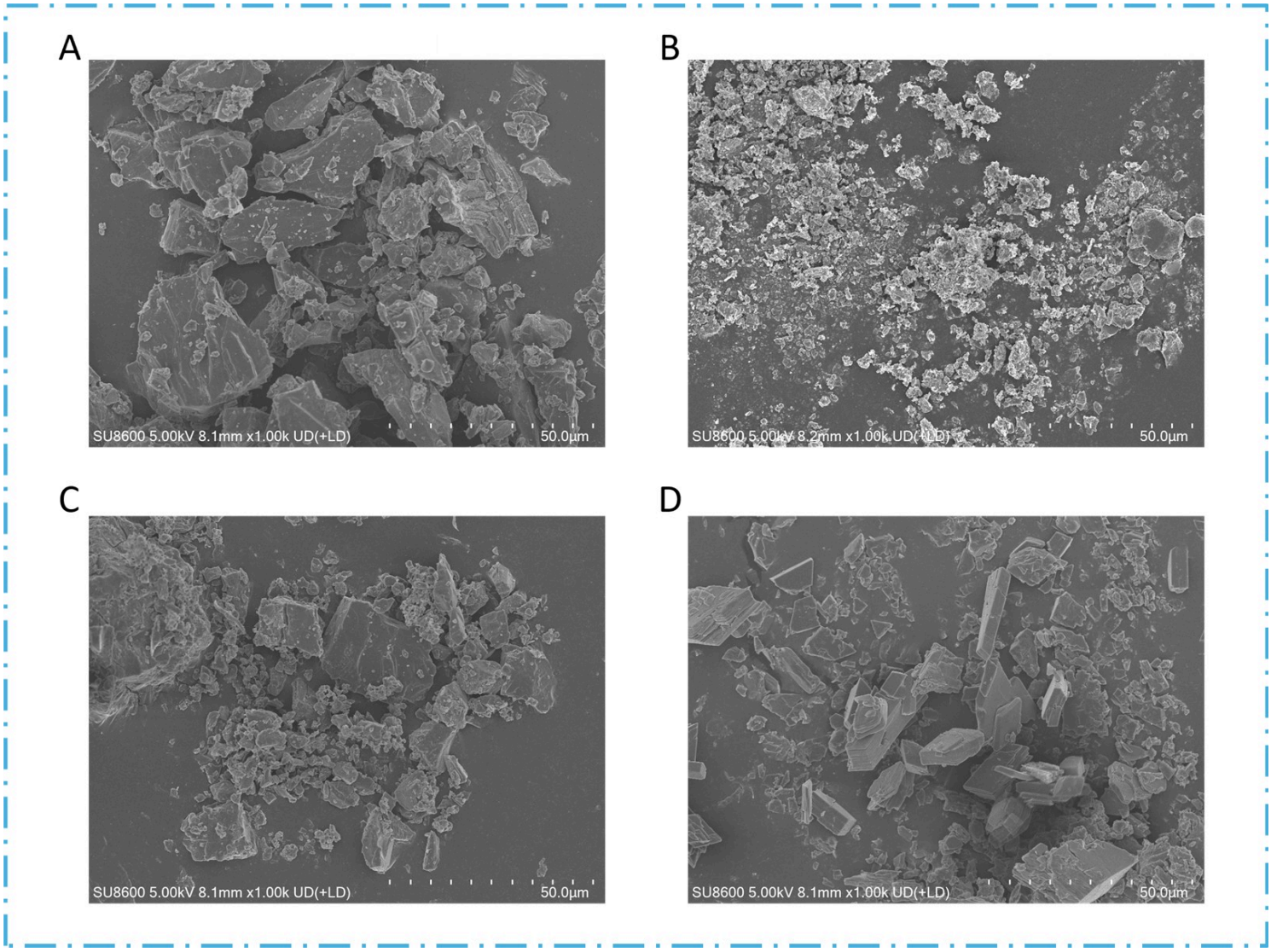


Figure 3

Scanning Electron Microscope

A: β-Cyclodextrin B:Paeonol C:Free Monomer Group D:Supramolecular Complex

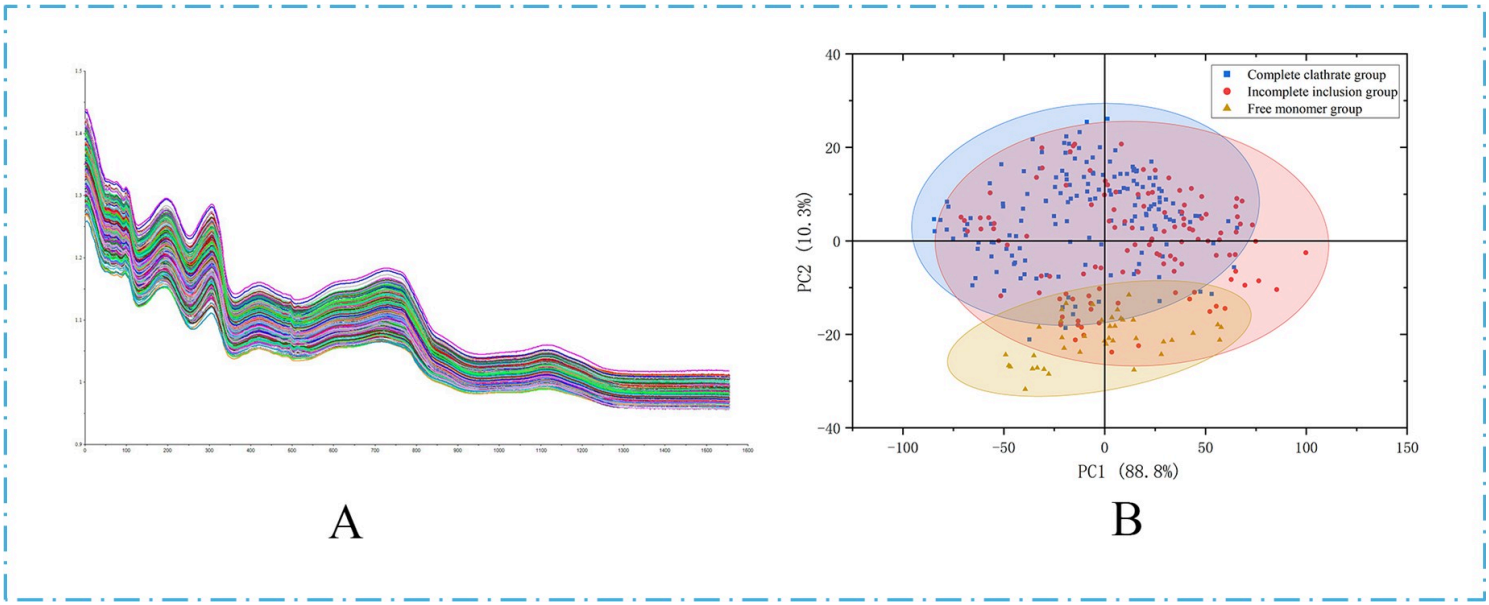


Figure 4

Raw Spectral Analysis

(A): Raw Spectra

(B): Raw Spectra Principal Component Analysis (PCA)

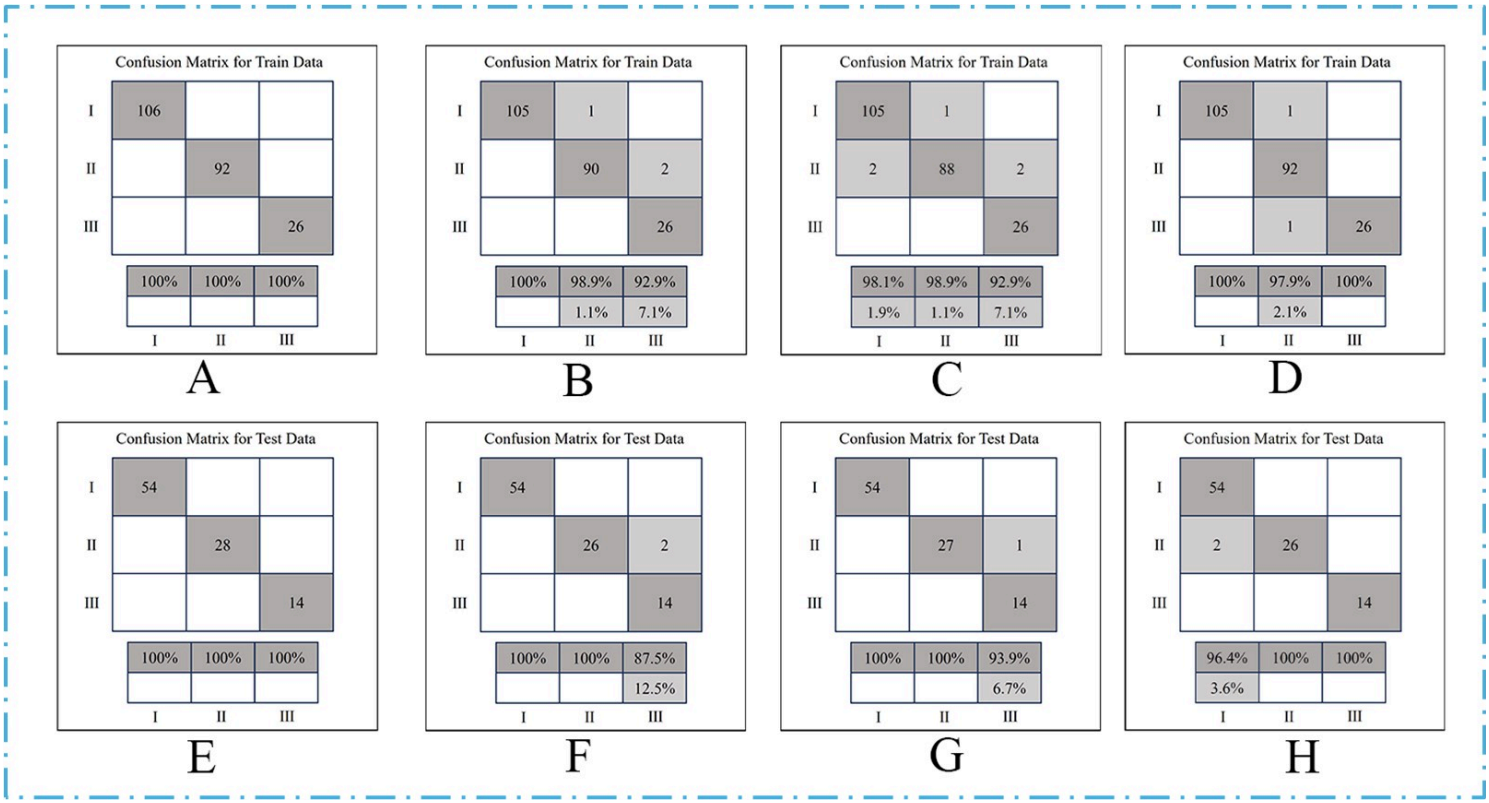


Figure 5

Classification Confusion Matrix Diagram

A□E SVM Model B□F LSTM- Bayesian Model

C□G CNN- Bayesian Model D□H ELM Model

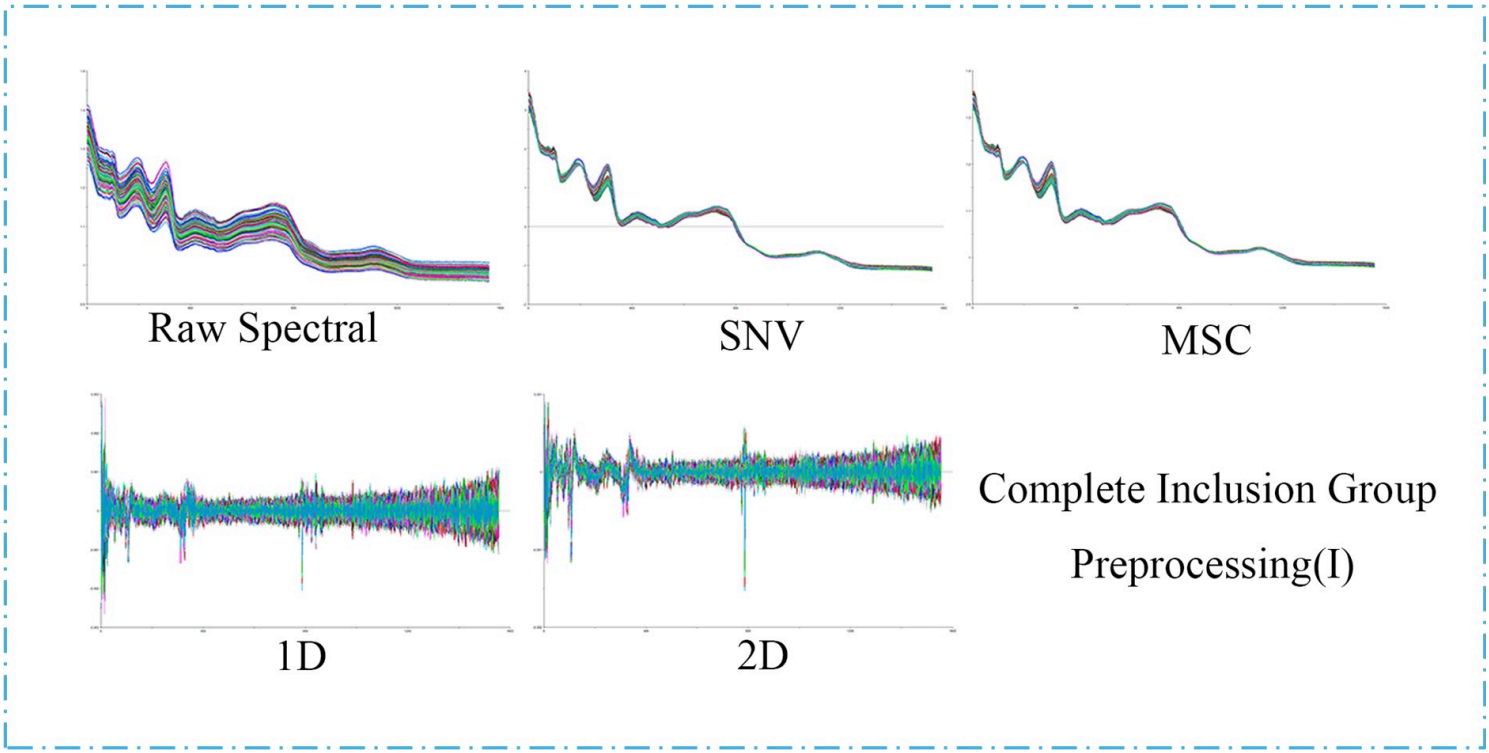


Figure 6

Spectral Data Preprocessing

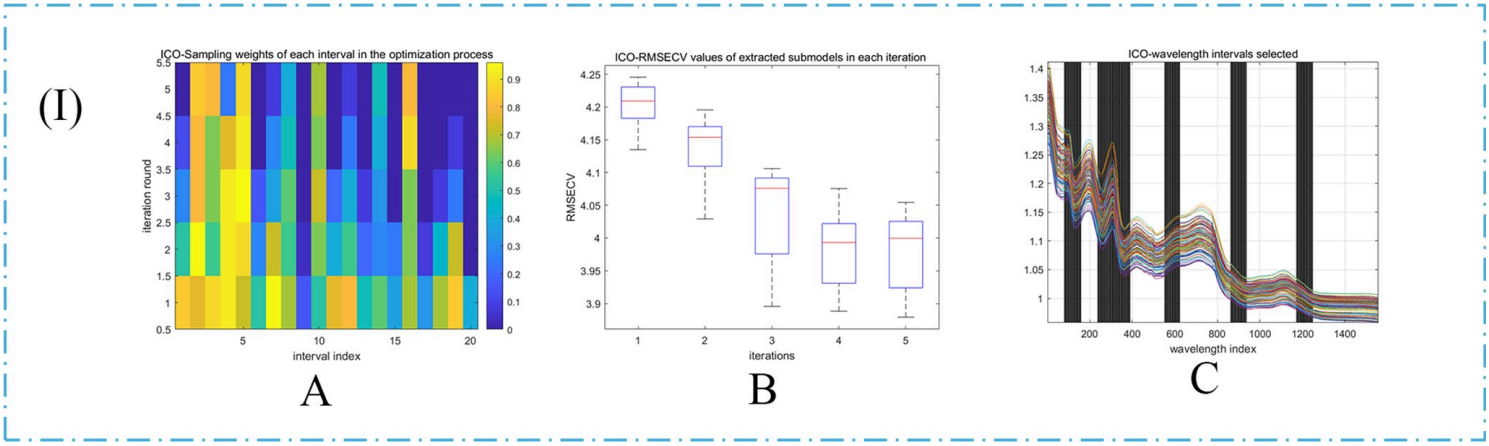


Figure 7

(I) ICO+SNV Characteristic Wavelength Selection Results: A: Sampling weights for each feature interval during optimization. B: Feature intervals selected by the ICO algorithm.

C: Variation in RMSECV.

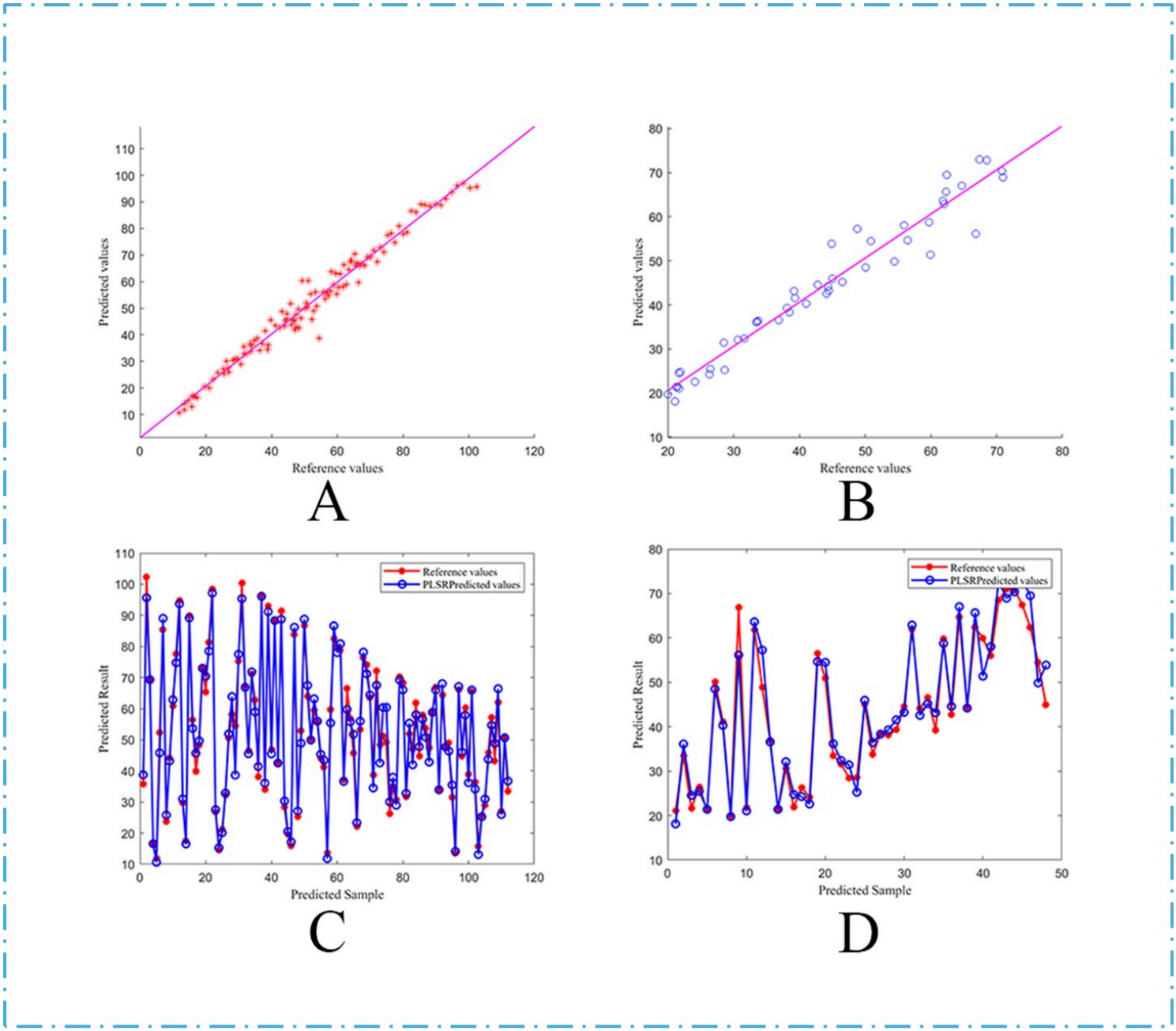


Figure 8

Complete Inclusion Group Inclusion Complex ICO+SNV PLSR Model

A: Training Set Visualization B: Prediction Set Visualization

C: Training Set Results Comparison D: Prediction Set Results Comparison Chart

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx