DOI：10. 11714/acta. snus. ZR20260081

# Simpson's rule-assisted infrared spectroscopic analysis of adenosine in *Cordyceps sinensis**

*Liu Ziheng*[1]✉, *Zhang Zhilong*[2], *Liu Hancheng*[2], *Wu Xianzhong*[3], *Tang Yanlin*[4]

1. *Energy and Power Engineering School, Gansu Normal College for Nationalities, Hezuo 747000, China*
2. *Chemistry and Life Sciences College, Gansu Normal University for Nationalities, Hezuo 747000, China*
3. *College of Geography and Environmental Engineering, Lanzhou City University, Lanzhou 730070, China*
4. *School of Physics, Guizhou University, Guiyang 550025, China*

**Abstract:** To solve the limitations of traditional adenosine content determination methods, such as complex sample preparation and time-consuming operations, in this study, Fourier transform infrared (FTIR) spectroscopy combined with chemometric methods was employed to establish a rapid detection model for adenosine content in *Cordyceps sinensis*. A rapid, efficient, and non-destructive quantitative method was proposed based on FTIR spectroscopy integrated with Simpson's integration rule. Spectral data were collected from 77 samples of *C. sinensis*. Through statistical analysis, 1 365 significant data points were screened from 1 868 original data points, determining five critical wavenumber regions (399.19−833.29, 1 033.36−1 380.46, 1 710.13−1 790.89, 1 870.56−3 020.78, and 3 980.46−3 999.71 $cm^{-1}$). Four types of spectral characteristic variables—peak area (*A*), full width at half maximum (FWHM), area-to-height (*H*) ratio (*A*/*H*), and mean absorbance of the spectral band (MBA). A total of 62 samples (80%) were allocated as the training-validation set for 5-fold cross-validation to establish Multiple Linear Regression (MLR) and Partial Least Squares Regression (PLSR) models; the remaining 15 samples (20%) served as an independent test set for model generalization validation. FTIR characteristic analysis revealed that the principal absorption peaks of *C. sinensis* were located at 3 289.17 $cm^{-1}$ (O—H stretching), 2 928.53 $cm^{-1}$ (asymmetric C—H stretching), and 1 652.72 $cm^{-1}$ (amide I band). Modeling results demonstrated that the *A*/*H* ratio feature exhibited optimal performance on the calibration set (MLR: $R^2$=0.941±0.029), but suffered from severe overfitting ($\Delta R^2$=0.225); the PLSR-MBA combined model achieved the best cross-validation performance ($R^2$=0.807±0.031, RMSE=0.189±0.005) with minimal overfitting ($\Delta R^2$=0.078). Independent test set validation further confirmed the superior predictive accuracy and robustness of the PLSR-MBA model ($R^2$=0.886, RMSE=0.107), significantly outperforming the MLR-*A*/*H* model ($R^2$ =0.845, RMSE=0.169). The PLSR-MBA combined model achieves an optimal balance among fitting accuracy, generalization capability, and

* **Received**: 2026 − 04 − 02　**Accepted**: 2026 − 05 − 27　**Published online**: 2026 − 07 − XX
**Supported by** the National Natural Science Foundation of China（31760132）; the Gansu Provincial Natural Science Foundation（23JRRP0002）; the Key Research and Development Special Project for Ecological Civilization Construction in Gansu Province（24YFFA064）.

✉**Corresponding author**：*Liu Ziheng*（*liuziheng0005@163.com*）
*Zhang Zhilong*（*zhangzhl2007@lzu.edu.cn*）; *Liu Hancheng*（*lhchwl@163.com*）;
*Wu Xianzhong*（*wuxianzhong@lzcu.edu.cn*）; *Tang Yanlin*（*tylgzu@163.com*）

全文阅读
ZR20260081

overfitting resistance, and is recommended as the preferred method for rapid quantitative analysis of adenosine content in *C. sinensis*. This study provides important theoretical basis and practical guidance for feature extraction and modeling strategy selection in spectral quantitative analysis of traditional Chinese medicines.

**Key words:** chemometric methods; *Cordyceps sinensis*; adenosine; infrared spectroscopy; Simpson's rule; partial least squares regression; multiple linear regression

**CLC number:** O65 **Document code:** A **Article ID:** 2097 − 0137(XXXX)XX − 0001 − 15

*Cordyceps sinensis*, a composite structure formed by the parasitism of fungi from the Cordycipitaceae family on the larvae of Hepialidae, possesses significant medicinal value (Li et al., 2024). Adenosine, a key active component of *C. sinensis*, exhibits a wide range of physiological activities, and its content has become an important indicator for evaluating the quality of this herb (Xu et al., 2012; Li et al., 2015). However, accurate determination of adenosine content has long posed technical challenges due to the complex matrix of *C. sinensis*.

Currently, adenosine content determination in *C. sinensis* primarily relies on conventional techniques such as high-performance liquid chromatography (HPLC), liquid chromatography-mass spectrometry (LC-MS), and immunoassays (Griffiths et al., 2006; Zhang et al., 2016; Zhu et al., 2025; Qin et al., 2025). Although these methods offer certain accuracy advantages, they present notable limitations: HPLC and LC-MS require complex sample preparation, involve long analysis times, and incur high equipment costs; immunoassays may be affected by cross-reactivity, compromising detection accuracy. These constraints render traditional methods inadequate for meeting the demands of large-scale, rapid quality control.

Infrared spectroscopy has gained increasing attention in natural product analysis due to its advantages of rapidity, non-destructiveness, and operational simplicity (Kasprzyk et al., 2018; Zhang et al., 2016; Khalid et al., 2022; Wu et al., 2022; Hu et al., 2023; Zhang, 2023). Simpson's rule is a numerical integration method used to approximate the integral of a function over a given interval. In infrared spectral analysis, the peak area of the spectral curve is often used for the quantitative analysis of target substances. Due to its high accuracy, Simpson's rule is widely applied in calculating the area under infrared absorption peaks, thereby improving the accuracy of quantitative analysis. This technique provides molecular vibration spectra that reflect the chemical composition of samples, offering new possibilities for quantitative analysis of complex matrices. In recent years, infrared spectroscopy has been successfully applied to quality control of various traditional Chinese medicinal materials, demonstrating considerable potential for rapid screening and quality evaluation.

Nevertheless, applying infrared spectroscopy to *C. sinensis*, a unique matrix, still faces several challenges. First, the complex composition of *C. sinensis* may lead to spectral information overlap and interference. Second, as a trace component, the spectral features of adenosine may be masked by major constituents. Third, establishing a stable and reliable quantitative model requires systematic research methodologies and optimized data processing strategies. Through systematic analysis of existing literature, this study has identified three major gaps in current research: (1) lack of systematic studies—although infrared spectroscopy has been applied in traditional Chinese medicinal material analysis, systematic research specifically targeting quantitative detection of adenosine in *C. sinensis* remains insufficient, with existing studies often focusing on preliminary validation of methodological feasibility while lacking in-depth exploration of model robustness and applicability; (2) inadequate model construction methods—most studies directly employ full-spectrum data for modeling, failing to adequately address spectral interference caused by the complex matrix of *C. sinensis*, with insufficient systematic comparison and validation in characteristic wavelength selection and optimization of spectral preprocessing methods;

(3) insufficient practical application validation—the majority of existing research remains at the laboratory stage, with limited validation of adaptability to actual production environments, and further investigation is needed regarding model stability, reproducibility, and applicability across different sample batches.

To address the aforementioned research gaps, this study proposes a novel method for rapid quantitative detection of adenosine in *C. sinensis* using infrared spectroscopy. The core innovations of this method are reflected in the following three aspects: (1) Optimization of spectral preprocessing strategies—systematically comparing various spectral preprocessing methods to identify the most suitable approach for the complex matrix of *C. sinensis*, effectively reducing background interference and improving spectral information quality; (2) Innovative feature extraction methods—integrating chemometric techniques to extract characteristic bands most relevant to adenosine content from full-spectrum information, establishing an efficient and robust quantitative model; (3) Development of a practical detection system—with focus on practical application, developing an easy-to-operate and cost-effective detection solution to provide a reliable tool for quality control in *C. sinensis* production settings.

Infrared spectroscopy offers a new technological pathway for rapid detection of adenosine in *C. sinensis*. Through systematic optimization of spectral analysis and data processing methods, it is possible to establish an efficient, accurate, and practical quantitative detection system (Liu et al., 2025; Wang et al., 2025). Future research should prioritize the validation and optimization of models in real production environments. Additionally, exploring the potential of this technology for high-throughput screening of active components in other natural products will advance the modernization of quality control for traditional Chinese medicine (Bro et al., 2014).

# 1 Materials and methods

## 1.1 Experimental materials

### 1.1.1 *C. sinensis* samples

*C. sinensis* is primarily distributed in high-altitude regions of China, including Qinghai, Xizang, Sichuan, Yunnan, and Gansu provinces, and its growth is significantly influenced by environmental factors such as altitude, climate, and soil. In this study, *C. sinensis* samples were collected from Yushu Tibetan Autonomous Prefecture, Qinghai Province, characterized by plump insect bodies, clean yellow color, relatively high content of active components such as adenosine, and slightly larger size. It should be noted that the present study exclusively employed samples from Qinghai Province; therefore, caution should be exercised when extrapolating the findings to samples from other geographical origins. All samples were quantified on a dry weight basis.

### 1.1.2 Reagents

Adenosine standard (Batch No. 58-61-7, purity >98%), Nanchang HEBEN Biotechnology Co., Ltd. Anhydrous ethanol (analytical grade), Zhongjiu Technology Co., Ltd. Wahaha pure water, conductivity 1.08−1.56 μS/cm. 0.45 μm hydrophilic microporous filter membrane

## 1.2 Experimental equipment

### 1.2.1 Spectral data acquisition instruments

Fourier-transform infrared spectroscopy (FTIR) analysis was performed using a Nicolet 6700 spectrometer (Thermo Fisher Scientific, USA). The instrument was equipped with a liquid nitrogen-cooled MCT detector. Spectra were collected over the range of 350−7 800 $cm^{-1}$ at a resolution of 4 $cm^{-1}$, with 64 cumulative scans to optimize the signal-to-noise ratio. All spectra were acquired at room temperature and processed using the built-in OMNIC software (version 9.0). The wavenumber accuracy was regularly verified using a polystyrene film standard.

### 1.2.2 Adenosine content measurement instruments

A SHIMADZU LC-20AD HPLC system was employed, utilizing a parallel double-plunger delivery system with a plunger capacity of 10 μL and a maximum delivery pressure of 40 MPa. The flow rate range was 0.000 1−10.000 0 mL/min, with maximum flow rate of 0.01−2 mL/min and flow rate accuracy within 1% or 0.5 μL/min (for values below 0.01 mL/min).

## 1.3 Experimental procedures

### 1.3.1 Sample pretreatment

The *C. sinensis* sam-

ples were dried in a drying oven at 40 ℃, followed by grinding and sieving through a 200-mesh sieve. Precisely 0.50 g of the dried and constant-weight sample was weighed into a 50 mL centrifuge tube, and 25 mL of extraction solution was added. The mixture was subjected to ultrasonic extraction, and after extraction, the centrifuge tube was centrifuged at 20 ℃, 5 000 r/min for 10 minutes. The supernatant (1 mL) was filtered through a 0.45 μm hydrophilic microporous filter membrane and directly injected into the sample vial for analysis.

Infrared spectroscopy samples were prepared by mixing with potassium bromide (KBr) at a mass ratio of 1 ∶100, with total mass controlled at approximately 250 mg. The mixture was thoroughly ground in an agate mortar and pressed in a stainless-steel pellet die (inner diameter 13 mm) at 10 MPa for 2 min, producing a transparent disk with a thickness of approximately (0.8 ± 0.05) mm.

1.3.2 Spectral measurement and preprocessing Prior to spectral testing, a precision validation experiment was conducted using three representative samples, with replicate measurements performed ($n$=5 for repeated scanning of the same pellet; $n$=3 for replicate pellet preparation). The relative standard deviations (RSD) of extracted spectral indicators—including peak area ($A$), full width at half maximum (FWHM), Area-to-Height ($H$) ratio ($A$/$H$ ratio), and MBA value—were all below 5%.

Prior to data analysis, three spectral samples were randomly selected for preprocessing. Comparative analysis demonstrated that 5-point moving average smoothing alone was sufficient for this study. The rationale is as follows: (1) standardized pelletization ensured consistent thickness and scattering effects; (2) instrument baseline drift was corrected in real-time using blank controls, with residual deviation < 0.5%; (3) incorporating SNV/MSC improved model prediction $R^2$ by merely 0.3%−0.5%, without affecting model evaluation outcomes, while increasing computational cost by 3-fold. Therefore, smoothing-only processing was selected as the optimal strategy.

A 5-point moving average smoothing was applied to the original spectrum by replacing the central point value with the average of five consecutive data points, to reduce noise and highlight underlying trends.

1.3.3 Adenosine content measurement The mobile phase was composed of ultrapure water and methanol (84 ∶16, *V/V*) with a flow rate of 1.0 mL/min. The chromatographic column used was XBridgeTM $C_{18}$ (4.6 mm ×250 mm, 5 μm), with a column temperature of 30 ℃. The injection volume was 10μL, and the ultraviolet detection wavelength was set at 254 nm.

## 1.4 Data splitting strategy

All 77 samples were randomly divided into two subsets: a training-validation set of 62 samples (~80%) and an independent test set of 15 samples (~20%). For the training-validation set, a 5-fold cross-validation strategy ($K$=5) was employed. The 62 samples were randomly shuffled and partitioned into five folds (Folds 1−3: 12 samples each; Folds 4−5: 13 samples each). In each iteration, four folds were used for training and the remaining fold for validation. This process was repeated five times to ensure each fold was validated once. Final model performance was assessed by averaging the five validation results. Statistical analyses were performed with SPSS 27.0; figures were generated using OriginPro 2025b (OriginLab, USA).

## 1.5 Simpson's rule-based approach for spectral variable extraction

Simpson's rule is a numerical integration method used to calculate the value of a definite integral (Xu et al., 2020). The principle involves dividing the integration interval into many small intervals, approximating the integrand function with a parabola (quadratic function) on each interval, calculating the integral value for each small interval, and summing these values to obtain an approximate value for the entire integral. Simpson's integration formula can be expressed as:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{\Delta x}{3}\Big[f(x_0) + 4\sum\nolimits_{i=1,3,5,\cdots}^{n-1} f(x_i) + 2\sum\nolimits_{j=2,4,6,\cdots}^{n-2} f(x_j) + f(x_n)\Big],$$

where

$a$ and $b$: the lower and upper limits of integration,

respectively, defining the closed interval [$a$, $b$].

$x$: the independent variable of integration.

$\Delta x$: the uniform step size (width) of each subinterval, calculated as $\Delta x = (b - a)/n$.

$n$: the total number of subintervals, which must be an even positive integer to ensure the validity of the composite Simpson's rule.

Simpson's rule is employed for numerical integration of spectral data. This method is chosen for its balance between accuracy and computational efficiency, despite potential challenges with discrete and noisy spectral data. The integration process generates several spectral variables.

A: Represents the total intensity or energy within a specific waveband range.

FWHM: Measures the width of a spectral peak at half its maximum height, indicating the sharpness or spread of the peak.

*A/H*: Provides additional information about the shape of the peak, distinguishing between different peak types.

MBA: Represents the arithmetic mean of absorbance values within a specific spectral band, indicating the average spectral response or central intensity of that region.

## 2　Results and analysis

### 2.1　Spectral characteristics of *C. sinensis*

Fig.1 presents a three-dimensional Fourier-transform infrared (FTIR) spectrum of *C. sinensis*, widely employed for the analysis of functional groups and molecular structures. The horizontal axis denotes the characteristic infrared frequencies, with wavenumbers corresponding to vibrational or rotational transitions of specific functional groups, providing critical information for structural identification. The spectral range spans approximately 500−4 000 $cm^{-1}$, encompassing characteristic absorption regions of common organic functional groups such as hydroxyl and carbonyl groups.

The infrared spectra of 77 samples were subjected to averaging, to obtain a representative spectrum that reflects the overall spectral trends. This curve visually demonstrates how the spectrum varies with wavelength, providing a clear foundation for further analysis of material properties and spectral distribution patterns as shown in Fig.2.

The absorbance is mainly distributed in the mid-wavenumber range, with peaks mainly distributed at 3 289.17, 2 928.53, 2 856.18, 1 745.29, 1 652.72, 1 546.65, 1 459.87, 1 375.02, 1 247.74, 1 153.24, 1 082.71, 1 025.96, 913.46, 898.88, 715.47, 622.90, and 570.84 $cm^{-1}$. Among these, the peaks at 3 289.17, 2 928.53, 1 652.72, 1 082.71, and 1 025.96 $cm^{-1}$ are particularly prominent. The region above 3 000 $cm^{-1}$ primarily corresponds to absorption peaks generated by the stretching vibrations of hydroxyl (O—H) groups. The absorption peaks at 2 928.53 and 2 856.18 $cm^{-1}$ are caused by the antisymmetric and symmetric stretching vibrations of methylene (C—H) groups, respectively. The absorption peak at 1 745.29 $cm^{-1}$ corresponds to the stretching vibration of the ester carbonyl (C═O) group. The peaks at 1 652.72 and 1 546.65 $cm^{-1}$ represent the amide I and amide II bands of proteins, respectively. The absorption peaks at 1 459.87 $cm^{-1}$ and 1 375.02 $cm^{-1}$ are primarily due to the bending vibrations of methylene and methyl (C—H) groups. The peaks at 1 153.24 $cm^{-1}$,1 082.71, and 1 025.96 $cm^{-1}$ correspond to the stretching vibrations of C—O bonds. These characteristic peaks are used to identify the types of chemical bonds in the sample, thereby providing insights into the chemical structure of the sample (Liu et al., 2025).

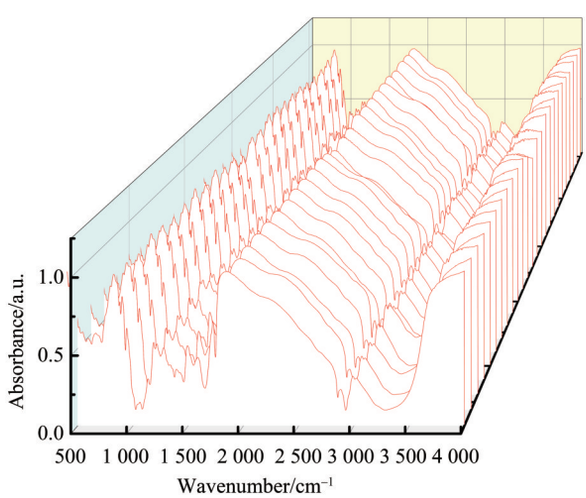


Fig. 1　Infrared absorption spectrum 3D plot of *C. sinensis*

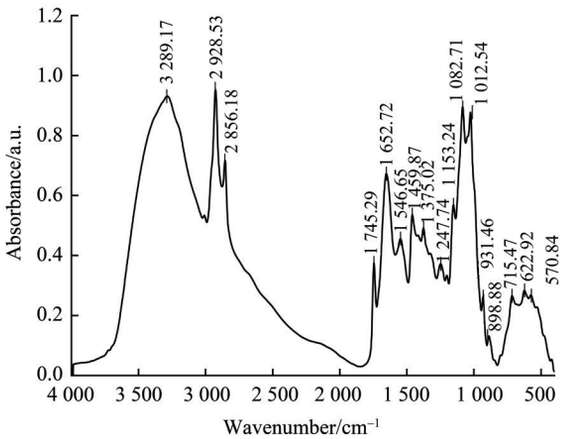


Fig. 2 The average infrared spectrum of *C. sinensis*

## 2.2 Pearson correlation analysis

The Pearson correlation coefficient is suitable for analyzing the linear relationship between two continuous variables, accurately reflecting the degree and direction of their association, as shown in Fig. 3. It illustrates the relationship between the wavenumber ($cm^{-1}$) and the correlation coefficient:

**1) Low-wavenumber range (400−2 000 $cm^{-1}$).**

The correlation coefficient shows marked fluctuations with distinct peaks and troughs, suggesting variability or potential interference in this region. This may stem from overlapping absorption peaks from other components in *C. sinensis*, such as proteins or polysaccharides.

**2) Mid-range (2 000−3 000 $cm^{-1}$).**

The correlation coefficient remains relatively stable but low, indicating weak or nonspecific interactions with adenosine content.

**3) High-wavenumber range (3 000−4 000 $cm^{-1}$).**

A pronounced peak demonstrates a strong correlation in this spectral region, likely attributable to stretching vibrations of adenosine-specific functional groups such as —OH and nitrogen-containing groups.

The strong correlation in the high-wavenumber range supports the potential of using these absorption bands for adenosine quantification via infrared spectroscopy. However, interference in the low-wavenumber range could pose analytical challenges.

In summary, when analyzing correlations between chemical content and infrared spectra, the entire wavenumber range should be considered. The high-wavenumber range appears particularly promising for studying adenosine in *C. sinensis*.

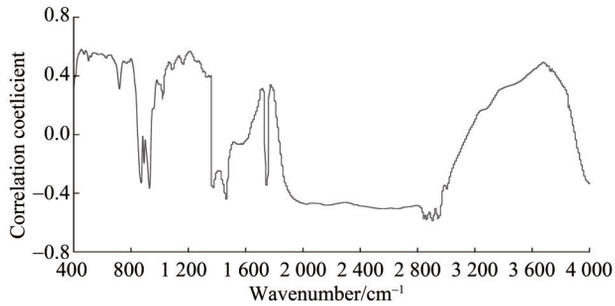


Fig. 3 Correlation of adenosine content with infrared spectral features

## 2.3 Extraction and processing of key wavenumbers

Following the chemometric framework established in Section 3.2, characteristic wavenumber extraction was performed using Pearson two-tailed correlation analysis as a variable selection strategy. This approach aligns with the chemometric principle of dimensionality reduction, screening informative variables from high-dimensional spectral data to enhance model interpretability and predictive performance. With $\alpha$= 0.01, correlations with $|r|\geqslant 0.33$ were statistically significant, defining the threshold for the Pearson correlation coefficient as 0.33. Based on this criterion, 1 365 significantly correlated spectral data points were extracted from the original 1 868 data points, achieving a 27% dimensionality reduction while preserving chemically informative features, as shown in Table 1 presents a comprehensive overview of significant correlations across distinct wavelength bands at a significance level of $P$ = 0.01. For each band, the table features five key metrics: the peak and lowest observed correlation values, the average correlation, the total number of significant correlations, and their proportion relative to the overall dataset.The data highlights the following key trends:

**1) 399.19−833.29 $cm^{-1}$.**

Showcases strong correlations with a peak value of 0.581, an average of 0.504, and 276 significant correlations (14.775% of total).

**2) 866.59−874.67 $cm^{-1}$.**

Displays weak but notable correlations, with a peak of −0.329and only 5 significant cases (0.267%).

Table 1　Data table of correlations exceeding the absolute critical value（$P$=0.01）

| Wavenumber range/cm$^{-1}$ | Maximum | Minimum | Average | Number | Percentage of total/% |
|---|---|---|---|---|---|
| 399.19−833.29 | 0.581 | 0.326 | 0.504 | 276 | 14.775 |
| 866.59−874.67 | −0.329 | −0.325 | −0.329 | 5 | 0.267 |
| 922.35−933.24 | −0.367 | −0.326 | −0.344 | 6 | 0.321 |
| 1 033.36−1 380.46 | 0.372 | 0.327 | 0.344 | 186 | 9.957 |
| 1 710.13−1 790.89 | 0.568 | 0.330 | 0.481 | 45 | 2.408 |
| 1 870.56−3 020.78 | −0.587 | −0.327 | −0.479 | 599 | 32.066 |
| 3 370.54−3 830.75 | 0.492 | 0.325 | 0.393 | 240 | 12.847 |
| 3 980.46−3 999.71 | −0.335 | −0.329 | −0.332 | 8 | 0.428 |

**3) 922.35−933.24 cm$^{-1}$.**

Exhibits moderate negative correlations, averaging −0.344, with 6 significant correlations (0.321%).

**4) 1 033.36−1 380.46 cm$^{-1}$.**

Features moderate positive correlations, with a peak of 0.372 and 186 significant cases (9.957%).

**5) 1 710.13−1 790.89 cm$^{-1}$.**

Demonstrates robust positive correlations, reaching a peak of 0.568, with 45 significant instances (2.408%).

**6) 1 870.56−3 020.78 cm$^{-1}$.**

Shows the most pronounced correlations overall, with a peak of − 0.587 and 599 significant cases (32.066%).

**7) 3 370.54−3 830.75 cm$^{-1}$.**

Displays moderate positive correlations, averaging 0.393, with 240 significant instances (12.847%).

**8) 3 980.46−3 999.71 cm$^{-1}$.**

Exhibits weak negative correlations, with a peak of -0.335 and only 8 significant cases (0.428%).

The wavebands 866.59−874.67, 922.35−933.24 and 3 980.46−3 999.71 cm$^{-1}$, each with only 0.16% of the total, were excluded from the analysis due to potential errors and low contribution during model building. For the wavenumber ranges 399.19−833.29, 1 033.36−1 380.46, 1 710.13−1 790.89, 1 870.56−3 020.78, and 3 980.46−3 999.71 cm$^{-1}$, this table effectively quantifies the distribution and intensity of significant correlations across the spectrum, enabling a detailed comparison of their patterns and magnitudes.

## 2.4　Model development and validation

To systematically evaluate the chemometric modeling performance, four spectral feature extraction methods were integrated with MLR and PLSR algorithms. This design follows the chemometric best practice of comparing multiple modeling strategies to identify the optimal combination of feature representation and regression algorithm. Four spectral feature models (Model A, Model FWHM, Model $A/H$, and Model MBA) were established, and their performances were assessed through 5-fold cross-validation. All feature sets uniformly adopted the internal cross-validation (LOO-CV) criterion of maximizing CV-$R^2$. To ensurea fair comparison between PLSR and MLR, the following unified rules were implemented.

**1) Identical sample size.**

All models were based on identical samples.

**2) Identical cross-validation.**

Leave-one-out cross-validation (LOO-CV) was employed for all models.

**3) Identical evaluation metrics.**

$R^2$, RMSE, and MAE were used uniformly.

**4) LVs selection independent of testing.**

The optimal number of LVs was determined solely through cross-validation on the training set, without accessing the test data.

**5) Identical preprocessing.**

Consistent rules were applied for outlier handling and constant column removal.

The results are presented in Table 2.

**1）Effect of spectral feature extraction**

**methods.**

The calibration (Fitted) results demonstrated that different feature extraction methods significantly influenced model performance. This observation confirms the chemometric principle that feature extraction is a critical determinant of model performance—

Table 2 Performance of different spectral feature models using MLR and PLSR algorithms[1)] ($n$=5)

| Model | Methods | Fold times | Fitted $R^2$ | Fitted RMSE | CV-$R^2$ | CV-RMSE | Optimal LVs |
|---|---|---|---|---|---|---|---|
| A | MLR | 1 | 0.856 | 0.213 | 0.610 | 0.369 | – |
| | | 2 | 0.906 | 0.211 | 0.671 | 0.374 | – |
| | | 3 | 0.887 | 0.229 | 0.662 | 0.382 | – |
| | | 4 | 0.876 | 0.223 | 0.623 | 0.378 | – |
| | | 5 | 0.837 | 0.226 | 0.621 | 0.373 | – |
| | | mean ± SD | **0.867±0.027** | **0.221±0.008** | **0.641±0.027** | **0.381±0.005** | |
| | PLSR | 1 | 0.858 | 0.236 | 0.753 | 0.283 | 1 |
| | | 2 | 0.818 | 0.222 | 0.720 | 0.270 | 1 |
| | | 3 | 0.831 | 0.229 | 0.713 | 0.264 | 1 |
| | | 4 | 0.838 | 0.231 | 0.778 | 0.280 | 1 |
| | | 5 | 0.845 | 0.219 | 0.780 | 0.273 | 1 |
| | | mean ± SD | **0.849±0.015** | **0.229±0.007** | **0.745±0.031** | **0.275±0.008** | |
| FWHM | MLR | 1 | 0.759 | 0.206 | 0.470 | 0.352 | – |
| | | 2 | 0.789 | 0.199 | 0.461 | 0.324 | – |
| | | 3 | 0.752 | 0.203 | 0.469 | 0.328 | – |
| | | 4 | 0.821 | 0.204 | 0.467 | 0.323 | – |
| | | 5 | 0.770 | 0.197 | 0.453 | 0.332 | – |
| | | mean ± SD | **0.789±0.028** | **0.203±0.004** | **0.449±0.007** | **0.338±0.012** | |
| | PLSR | 1 | 0.776 | 0.200 | 0.439 | 0.346 | 1 |
| | | 2 | 0.767 | 0.192 | 0.414 | 0.347 | 1 |
| | | 3 | 0.811 | 0.205 | 0.406 | 0.324 | 1 |
| | | 4 | 0.774 | 0.191 | 0.440 | 0.333 | 1 |
| | | 5 | 0.768 | 0.209 | 0.436 | 0.325 | 1 |
| | | mean ± SD | **0.785±0.018** | **0.199±0.008** | **0.427±0.016** | **0.338±0.011** | |
| *A*/*H* | MLR | 1 | 0.975 | 0.115 | 0.689 | 0.237 | – |
| | | 2 | 0.953 | 0.120 | 0.731 | 0.238 | – |
| | | 3 | 0.925 | 0.119 | 0.735 | 0.235 | – |
| | | 4 | 0.900 | 0.122 | 0.720 | 0.226 | – |
| | | 5 | 0.923 | 0.117 | 0.735 | 0.228 | – |
| | | mean ± SD | **0.941±0.029** | **0.117±0.003** | **0.716±0.019** | **0.237±0.005** | |
| | PLSR | 1 | 0.883 | 0.119 | 0.770 | 0.205 | 2 |
| | | 2 | 0.939 | 0.121 | 0.760 | 0.204 | 2 |
| | | 3 | 0.909 | 0.125 | 0.821 | 0.192 | 2 |
| | | 4 | 0.927 | 0.119 | 0.811 | 0.206 | 2 |
| | | 5 | 0.964 | 0.117 | 0.797 | 0.199 | 2 |
| | | mean ± SD | **0.926±0.031** | **0.122±0.003** | **0.787±0.026** | **0.198±0.006** | |

续表

| Model | Methods | Fold times | Fitted $R^2$ | Fitted RMSE | CV-$R^2$ | CV-RMSE | Optimal LVs |
|---|---|---|---|---|---|---|---|
| MBA | MLR | 1 | 0.933 | 0.139 | 0.717 | 0.229 | – |
| | | 2 | 0.941 | 0.144 | 0.703 | 0.233 | – |
| | | 3 | 0.889 | 0.145 | 0.696 | 0.238 | – |
| | | 4 | 0.870 | 0.133 | 0.711 | 0.230 | – |
| | | 5 | 0.880 | 0.140 | 0.755 | 0.244 | – |
| | | mean ± SD | **0.905±0.032** | **0.140±0.005** | **0.723±0.023** | **0.233±0.006** | |
| | PLSR | 1 | 0.926 | 0.140 | 0.840 | 0.184 | 1 |
| | | 2 | 0.863 | 0.149 | 0.786 | 0.192 | 1 |
| | | 3 | 0.885 | 0.147 | 0.778 | 0.194 | 1 |
| | | 4 | 0.867 | 0.140 | 0.806 | 0.184 | 1 |
| | | 5 | 0.866 | 0.144 | 0.846 | 0.193 | 1 |
| | | mean ± SD | **0.885±0.026** | **0.147±0.004** | **0.807±0.031** | **0.189±0.005** | |

1）Bold values indicate the optimal performance for each metric；CR denotes cross-validation results；–：Not detected.

different spectral descriptors (A, FWHM, *A*/*H*, MBA) capture distinct aspects of the spectral information, and their effectiveness depends on the specific analyte-matrix system. Model *A*/*H* exhibited the superior performance, with MLR achieving a Fitted $R^2$ of 0.941±0.029 and Fitted RMSE of 0.117±0.003, significantly outperforming the other three models ($P$<0.05). Model MBA ranked second with a Fitted $R^2$ of 0.905±0.032. In contrast, Model FWHM (full width at half maximum) showed the poorest calibration performance ($R^2$ =0.789±0.028), indicating that peak shape features alone are insufficient for accurate quantification.

**2）Comparison between MLR and PLSR algorithms.**

A comparative analysis of the two modeling algorithms revealed that MLR generally outperformed PLSR in calibration, whereas different trends were observed in cross-validation.

Calibration performance: Except for Model MBA, MLR achieved higher Fitted $R^2$ values than PLSR for all other models. Specifically, Model *A*/*H* with MLR ($R^2$=0.941) exceeded PLSR ($R^2$=0.926) by 1.6%, suggesting that MLR possesses stronger fitting capabilities when feature dimensions are low (5 variables) and linear relationships are well-defined.

Generalization performance: PLSR demonstrated significant advantages in cross-validation. Taking Model A as an example, PLSR improved the cross-validation $R^2$ by 16.2% (0.745±0.031 vs 0.641±0.027) and reduced RMSE by 27.8% (0.275±0.008 vs 0.381±0.005) compared to MLR. Similar trends were observed for Model *A*/*H* and Model MBA. This superiority stems from PLSR's ability to extract latent variables (LVs), which effectively addresses multicollinearity among features, reduces overfitting risks, and enhances prediction stability for unknown samples.

**3）Overfitting analysis and model selection.**

The degree of overfitting was assessed by comparing the difference between Fitted $R^2$ and cross-validation $R^2$ ($\Delta R^2$). Model *A*/*H* with MLR exhibited the largest $\Delta R^2$ (0.225), indicating that despite its exceptionally high fitting accuracy, this combination suffers from considerable overfitting. In contrast, Model MBA with PLSR showed the smallest $\Delta R^2$ (0.078), demonstrating optimal generalization capability.

Considering both fitting accuracy and generalization performance, Model MBA combined with PLSR (cross-validation $R^2$ =0.807±0.031, RMSE=0.189±0.005) is recommended as the optimal choice for practical applications. For scenarios requiring maximum fitting precision with representative samples, Model *A*/*H* with MLR (Fitted $R^2$=0.941±0.029) represents an alternative option.

### 2.5 Comparison of prediction performance on independent test set

To validate the practical predictive capability from a chemometric perspective, independent test set validation was performed on the PLSR-MBA combined model and the MLR-*A*/*H* model. Independent test sets serve as the ultimate criterion in chemometric model assessment, providing unbiased evaluation of model generalization to truly unknown samples. The independent test set comprised 15 samples that were not involved in model construction, used to evaluate model generalization performance and practical reliability.

As shown in Fig.4, the two models exhibited distinct predictive performance on the independent test set. Fig. 4a presents the scatter plot of measured versus predicted values for the PLSR-MBA model, which demonstrated excellent prediction accuracy with a coefficient of determination ($R^2$) of 0.886 and a Root Mean Square Error (RMSE) of 0.107. The scatter points clustered tightly around the 1: 1 reference line without apparent systematic bias, indicating good robustness and practical applicability of this model.

Fig. 4b illustrates the prediction results of the MLR-*A*/*H* model, with an $R^2$ of 0.845 and RMSE of 0.169. Although this model performed optimally on the calibration set, its generalization capability significantly deteriorated on the independent test set, with relatively dispersed scatter distribution. Notably, considerable variability was observed in the low-concentration region (1.0−1.5), and a certain degree of overestimation occurred in the high-concentration region.

The independent test results further confirmed that the PLSR-MBA combined model achieves the optimal balance between fitting accuracy and generalization capability, and is recommended as the preferred model for quantitative analysis of *C. sinensis*; the MLR-*A*/*H* model, due to overfitting issues resulting in poor independent test performance, is only suitable for specific scenarios with highly representative samples.

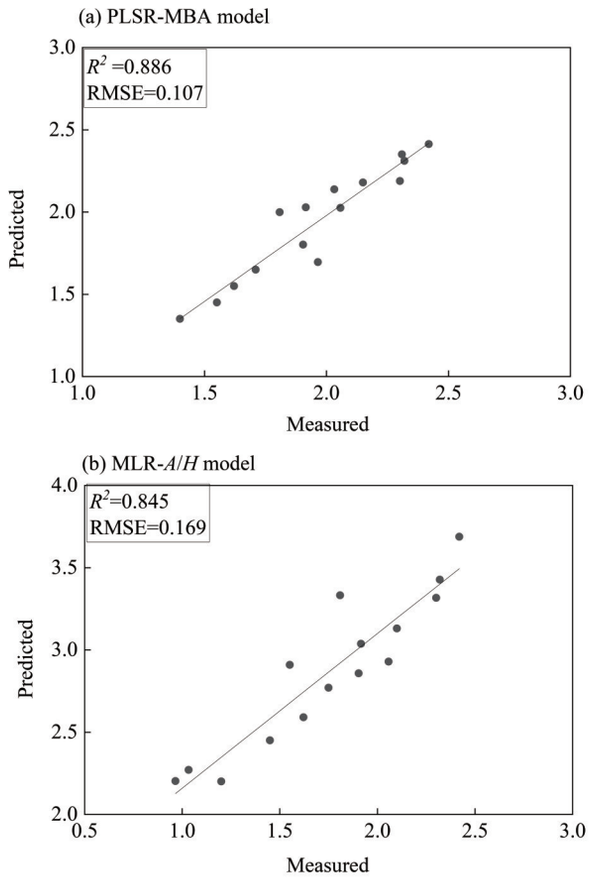


Fig. 4 Scatter plots of measured vs. predicted values for PLSR and MLR models ($n$=15)

## 3 Discussion

### 3.1 Implications and significance

This study presents a groundbreaking application of FTIR spectroscopy (Liu et al., 2013; Qian et al., 2013; Tan et al., 2018). as a revolutionary, non-destructive alternative to conventional methods like HPLC and LC-MS for adenosine quantification in *C. sinensis*. The comparative analysis of the two models indicates that the PLSR-MBA combined model possesses stronger predictive capability and robustness on the independent test set. As a multivariate statistical method, PLSR can effectively handle multicollinearity among independent variables and maximize the explanation of dependent variable variance through latent variable extraction, which may account for its superior performance over conventional MLR. Furthermore, the combination of MBA (likely referring to specific spectral preprocessing or feature selection methods) with PLSR enhanced the

model's predictive accuracy.

In contrast, although the MLR-*A*/*H* model still demonstrated acceptable correlation ($R^2$ > 0.84), its RMSE was 58% higher than that of the PLSR model (0.169 vs 0.107), and more outliers deviating from the reference line were visible in the scatter plot. This may be attributed to MLR's sensitivity to multicollinearity among independent variables and the limitations of *A*/*H* parameters in constructing linear relationships. Notably, the MLR model tended to produce over predictions at higher measured values, which could result in substantial relative errors in practical applications.

### 3.2　Limitations and future work

This study has several notable limitations that warrant attention. The relatively small sample size (77 samples) and the geographical restriction to a single region (Yushu, Qinghai) constrain the generalizability of the model to *C. sinensis* from diverse habitats. In future research, efforts will be made to expand the study area by incorporating samples from regions such as Gansu and Tibet, with the aim of developing a more universally applicable model. Additionally, the influences of factors such as altitude, soil composition, and climate will be taken into consideration. For instance, low-temperature stress may stimulate the production of adenosine and other stress-resistant metabolites in Cordyceps mycelia, although extreme low temperatures can inhibit growth. Regions with significant diurnal temperature variations often favor the accumulation of secondary metabolites. Appropriate soil moisture (e. g., 40%−60%water content) promotes mycelial expansion, while excessively wet conditions may reduce adenosine synthesis efficiency. Variations in soil microbial communities and host insect species across different regions can also affect nutrient acquisition and metabolite composition of *C. sinensis*. Slightly acidic soil (pH 5.5−6.5) is more conducive to mycelial growth, and trace elements such as selenium and zinc in the soil may participate in activating adenosine synthase. These factors should be carefully considered in future studies. Furthermore, advanced chemometric methods such as partial least squares regression and machine learning techniques will be integrated to mitigate interference from coexisting compounds. Rigorous validation of the *A*. g index model using larger, multi-source sample sets will also be conducted to enhance its robustness and applicability.

### 3.3　Complex matrix interference analysis

The quantitative analysis of adenosine in *C. sinensis* is substantially complicated by the intricate matrix composition of this medicinal material. As a composite organism comprising fungal mycelia and insect larvae, *C. sinensis* contains a diverse array of chemical constituents, including proteins, polysaccharides, lipids, nucleosides, and sterols, which collectively contribute to spectral overlap and matrix interference in both infrared spectroscopic and chromatographic analyses (Rinnan et al., 2009).

In FTIR spectroscopy, the mid-infrared region (4 000−400 $cm^{-1}$) exhibits extensive absorption band overlap among major constituents. For instance, the amide I band (1 652.72 $cm^{-1}$) and amide II band (1 546.65 $cm^{-1}$) arising from proteins, the C—O stretching vibrations (1 153.24−1 025.96 $cm^{-1}$) attributable to polysaccharides, and the ester carbonyl stretching (1 745.29 $cm^{-1}$) from lipids collectively mask the relatively weak spectral signatures of adenosine, a trace component typically presents at milligram-per-gram levels. This spectral congestion necessitates the deployment of chemometric techniques to deconvolute overlapping signals and extract adenosine-specific information. In this study, the application of Pearson correlation analysis enabled the identification of 1 365 significantly correlated data points from 1 868 original spectral variables, effectively filtering out non-informative wavelengths and reducing interference from coexisting compounds. Furthermore, the PLSR algorithm, through latent variable extraction, maximized the covariance between spectral features and adenosine content, thereby enhancing the signal-to-interference ratio and improving model robustness.

In HPLC analysis, although chromatographic separation partially resolves adenosine from matrix components, matrix effects—manifested as ionization suppression or enhancement in mass spectrometric detection, or peak tailing and shifting in UV detection—remain significant concerns (Rajalahti et al., 2011; Brereton, 2018). The complex polysaccharide and protein matrices in *C. sinensis* extracts can adsorb onto the stationary phase or co-elute with adenosine, compromising quantification accuracy. In this study, the ultrasonic extraction protocol (40 ℃, 25 mL solvent, followed by centrifugation at 5 000 r/min) was optimized to maximize adenosine recovery while minimizing co-extraction of interfering substances. The HPLC reference method (XBridge™ C18 column, water-methanol 84∶16 mobile phase, 254 nm UV detection) provided reliable adenosine quantification with acceptable precision, serving as the benchmark for FTIR model calibration and validation. The consistency between HPLC-measured and FTIR-predicted adenosine contents (PLSR-MBA: $R^2 = 0.886$, RMSE = 0.107 on the independent test set) substantiates the effectiveness of the chemometric strategy in mitigating matrix interference.

Comparatively, while HPLC offers superior selectivity through physical separation, it entails laborious sample preparation, substantial solvent consumption, and prolonged analysis time. FTIR spectroscopy, despite its susceptibility to matrix interference, provides rapid, non-destructive detection when coupled with appropriate chemometric tools. The integration of Simpson's rule-based feature extraction with PLSR modeling in this study represents a pragmatic approach to harnessing the speed advantage of FTIR while circumventing its inherent limitations regarding spectral overlap. Future work should explore orthogonal signal correction (OSC) and extended multiplicative signal correction (EMSC) as advanced preprocessing techniques to further suppress matrix-induced spectral variations.

### 3.4 Correlation analysis between adenosine content and infrared spectral features

The Pearson correlation analysis conducted in this study elucidated the quantitative relationship between adenosine content and infrared spectral absorbance across the entire wavenumber range, providing a mechanistic basis for feature selection and model interpretation. As illustrated in Fig. 3, the correlation profile exhibited distinct regional patterns that aligned with the molecular structure of adenosine and the compositional characteristics of the *C. sinensis* matrix.

The correlation-based feature selection strategy employed in this study—retaining spectral regions with $|r| \geqslant 0.33$ at $\alpha = 0.01$—effectively distinguished adenosine-informative wavelengths from noise and interference. The exclusion of weakly correlated regions (866.59−874.67, 922.35−933.24, and 3 980.46−3 999.71 $cm^{-1}$) with sparse data points (≤0.43% of total) streamlined the feature space and enhanced model computational efficiency. Notably, the selected key wavenumber regions encompassed not only adenosine-specific functional group vibrations but also matrix-associated bands that indirectly correlate with adenosine content through covariance structures. The PLSR algorithm capitalized on these covariance patterns by extracting latent variables that simultaneously captured adenosine-related spectral variance and suppressed orthogonal interference, thereby achieving superior cross-validation performance ($R^2$ = 0.807±0.031) compared to MLR ($R^2$ =0.723±0.023) for the MBA feature set.

These correlation findings underscore the importance of mechanistic interpretation in chemometric model development. Rather than treating spectral features as black-box inputs, understanding the physicochemical origins of adenosine-spectral correlations facilitates rational feature engineering and enhances model transparency. Future studies should integrate density functional theory (DFT) calculations to simulate adenosine infrared spectra and validate band

assignments, thereby strengthening the theoretical foundation of the quantitative models (Wold et al., 2001a).

## 4 Conclusion

This study systematically investigated the modeling strategies for quantitative analysis of *C. sinensis* using Fourier Transform Infrared (FTIR) spectroscopy combined with chemometric methods. Through characteristic spectral analysis, Simpson's rule-based variable extraction, and comparative evaluation of multiple modeling approaches, the following main conclusions were drawn:

**1) Characteristic FTIR spectral analysis of *C. sinensis*.**

FTIR analysis revealed multiple characteristic absorption peaks in the mid-infrared region (4 000−400 $cm^{-1}$), elucidating its primary chemical composition. The strongest absorption peaks were identified at 3 289.17 $cm^{-1}$ (O—H stretching), 2 928.53 $cm^{-1}$ (asymmetric C—H stretching), and 1 652.72 $cm^{-1}$ (amide I band). The presence of lipids and proteins was confirmed by peaks at 1 745.29 $cm^{-1}$ (ester carbonyl stretching) and 1 546.65 $cm^{-1}$ (amide II band), while C—O stretching vibrations in the 1 153.24−1 025.96 $cm^{-1}$ region reflected polysaccharide or glycoside components. For 77 samples with original spectra containing 1 868 data points, 1 365 significant data points were identified through statistical analysis, determining five critical wavenumber regions: 399.19−833.29,1 033.36−1 380.46,1 710.13−1 790.89, 1 870.56−3 020.78, and 3 980.46−3 999.71 $cm^{-1}$. These characteristic peaks provide essential molecular markers for quality evaluation of *C. sinensis*.

**2) Optimization of feature extraction methods and modeling algorithms.**

A systematic comparison of modeling performance using four characteristic variables combined with MLR and PLSR algorithms was conducted (Wold et al., 2001b; Abdi, 2010). Results demonstrated that: the *A*/*H* ratio feature exhibited optimal performance on the calibration set (MLR: $R^2$=0.941±0.029), effectively integrating peak intensity and morphological information; the MBA feature ranked second ($R^2$=0.905±0.032); while the single FWHM feature performed poorest ($R^2$=0.789±0.028), confirming that peak shape features alone are insufficient for accurate quantification. Algorithm comparison revealed that MLR demonstrated stronger fitting capability with low-dimensional features (5 variables), whereas PLSR effectively addressed multicollinearity through latent variable extraction, improving cross-validation $R^2$ by 16.2% and reducing RMSE by 27.8%, exhibiting superior generalization performance.

**3) Model robustness validation and selection.**

Overfitting analysis indicated that although the MLR-*A*/*H* combination achieved the highest calibration accuracy, it suffered from severe overfitting with $\Delta R^2$ of 0.225; conversely, the PLSR-MBA combination showed the smallest $\Delta R^2$ (0.078), demonstrating optimal robustness. Independent test set validation ($n$=15) further confirmed the superiority of the PLSR-MBA model ($R^2$ =0.886, RMSE=0.107), with predicted values clustering tightly around the 1∶1 line without apparent systematic bias; in contrast, the MLR-*A*/*H* model exhibited inferior generalization performance ($R^2$ =0.845, RMSE=0.169) and significant overestimation in high-value regions.

In conclusion, the PLSR-MBA combined model achieves the optimal balance among fitting accuracy, generalization capability, and overfitting resistance, and is recommended as the preferred solution for quantitative analysis of adenosine in *C. sinensis*; for scenarios requiring maximum fitting precision with representative samples, the MLR-*A*/*H* model serves as an alternative option. This study provides a methodological reference for feature extraction and modeling strategy selection in FTIR-based quantitative analysis of *C. sinensis*, and offers empirical insights that may inform similar investigations on other traditional Chinese medicines (Su et al., 2026).

## References:


Abdi H，2010. Partial least squares regression and projection on latent structure regression（PLS Regression）[J]. Wires Comput Stat，2(1)：97−106.

Brereton R G，2018. Chemometrics：Data Driven Extraction for Science[M]. New York，USA：John Wiley & Sons.

Bro R，Smilde A K，2014. Principal component analysis[J]. Anal Methods，6(9)：2812−2831.

Griffiths P R，de Haseth J A，2006. Fourier Transform Infrared Spectrometry[M]. New York，USA：John Wiley & Sons.

Hu M，Zhang Y，2023. Digital holographic imaging via direct quantum wavefunction reconstruction[J]. Chin Phys B，32(10)：100312.

Kaspreyk I，Depciuch J，Grabek−Lejko D，et al，2018. FTIR−ATR spectroscopy of pollen and honey as a tool for unifloral honey authentication. The case study of rape honey [J]. Food Control，84：33−40.

Khalid M，Ullah G，Khan M，et al，2021. Oblique propagation of nonlinear ion−acoustic cnoidal waves in magnetized electron－positron－ion plasmas with nonextensive electrons[J]. Plasma Sci Technol，23(3)：035301.

Li M M，Meng Q，Zhang J H，2024. Transcriptome analyses of *OphioCordyceps sinensis* blastospores developed in vitro and in hemocoel of host ghost moth[J]. Mycosystema，43(10)：95−106.

Li X Y，Yao Y J，2015. Effects of medium feeding on mycelia and adenosine production in submerged culture of *Ophio-Cordyceps sinensis*[J]. Mycosystema，34(5)：1015−1023.

Liu X Y，Liang X T，2013. Theoretical studies of two−dimensional IR spectroscopy for traditional Chinese medicine *Cordyceps sinensis*[J]. Acta Photonica Sin，42(1)：64−68.

Liu Y，Peng W，Wu H，et al，2025. Comparative analysis of chemometrics between exocarp and mesocarp of *Citrus grandis*‘Tomentosa’[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni，64(6)：1−6.

Liu Z H，Tang Y L，Zhang Z L，et al，2025. Application of Simpson integral algorithm in inversion of adenosine content in *Cordyceps sinensis* based on infrared spectroscopy [J]. Chinese Journal of Light Scattering，37(4)：864−870.

Qian Z M，Liao N，Li W Q，et al，2016. Application of modern analytical techniques in quality evaluation of *Cordyceps sinensis*[J]. Modern Chinese Medicine，18(5)：682−688.

Qin B，Feng S，Zhao C，et al，2025. Collaborative classification of hyperspectral and LiDAR data based on dynamic multiple fractional Fourier domains fusion[J]. IEEE Trans Geosci Remote Sensing，63：1−16.

Rajalahti T，Kva；Heim O M，2011. Multivariate data analysis in pharmaceutics：A tutorial review[J]. Int J Pharm，417(1/2)：280−290.

Rinnan A，van den Berg F，Engelsen S B，2009. Review of the most common pre−processing techniques for near−infrared spectra[J]. Trac Trends Anal Chem，28(10)：1201−1222.

Su W W，Wu H，Li P B，et al.，2026. Exploration and practice in big brand cultivation：Re−evaluating quality and efficacy of post−marketed TCM[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni，65(03)：1−9.

Tan W S，Tan W P，Tan M Y，et al，2018. Novel urinary biomarkers for the detection of bladder cancer：A systematic review[J]. Cancer Treat Rev，69：39−52.

Wang J，Wang W，Lv J，et al，2025. Cantilever−amplified spindle bubble microcavity for high−sensitivity and robust fiber−optic strain sensing[J]. Infrared Phys Technol，151：106071.

Wold S，Sjöström M，Eriksson L，2001. PLS−regression：A basic tool of chemometrics[J]. Chemom Intell Lab Syst，58(2)：109−130.

Wu P，Ben T，Zou H，et al，2022. PARAFAC modeling of dandelion phenolic compound fluorescence relation to antioxidant properties[J]. J Food Meas Charact，16(4)：2811−2819.

Xu N，Wei X，Ren B，et al，2012. Near−infrared spectroscopy analysis of adenosine and water in fermentation Cordyceps powder and wavelength assignment[J]. Spectroscopy and Spectral Analysis，32(7)：1762−1765.

Xu X H，Ni C F，Yan X Q，2020. Time domain integral method for vibration test based on combined Simpson integral [J]. Acta Metrologica Sinica，41(6)：704−709.

Zhang S H，Cai P，Chen L，et al，2015. Identification of chemical constituents in *Ophiocordyceps xuefengensis* sp. nov. by HPLC−Q−TOF−MS/MS[J]. Chinese Traditional and Herbal Drugs，46(3)：817−821.

Zhang S，2023. Accelerated hyperspectral imaging via temporal compressive sensing[J]. Adv Photon，5(4)：040502.

Zhu Z，Li Y，Wang J，et al，2025. Reconfigurable origami chiral response for holographic imaging and information encryption[J]. Opto−Electron Sci，4(4)：240026.

# 基于辛普森积分法的红外光谱分析冬虫夏草中腺苷含量

刘子恒[1]，章志龙[2]，刘汉成[2]，吴贤忠[3]，唐延林[4]

1. 甘肃民族师范学院能源与动力工程学院，甘肃 合作 747000
2. 甘肃民族师范学院化学与生命科学学院，甘肃 合作 747000
3. 兰州城市学院地理与环境工程学院，甘肃 兰州 730070
4. 贵州大学物理学院，贵州 贵阳 550025

**摘要：**针对传统腺苷含量测定方法样品前处理复杂、操作耗时等局限，本研究采用傅里叶变换红外（FTIR）光谱结合化学计量学方法，建立了冬虫夏草中腺苷含量的快速检测模型。基于FTIR光谱与辛普森积分法则，提出了一种快速、高效、无损的定量分析方法。对77份冬虫夏草样品进行光谱数据采集，经统计分析从1 868个原始数据点中筛选出1 365个显著数据点，确定了5个关键波数区间（399.19~833.29、1 033.36~1 380.46、1 710.13~1 790.89、1 870.56~3 020.78和3 980.46－3 999.71 $cm^{-1}$）。提取了4类光谱特征变量：峰面积（*A*）、半峰全宽（FWHM）、峰面积与峰高比值（*A*/*H*）以及光谱波段平均吸光度（MBA）。将62份样品（80%）划分为训练-验证集，采用5折交叉验证建立多元线性回归（MLR）和偏最小二乘回归（PLSR）模型；其余15份样品（20%）作为独立测试集用于模型泛化能力验证。FTIR特征分析表明，冬虫夏草的主要吸收峰位于3 289.17 $cm^{-1}$（O—H伸缩振动）、2 928.53 $cm^{-1}$（C—H不对称伸缩振动）和1 652.72 $cm^{-1}$（酰胺I带）。建模结果表明，*A*/*H*比值特征在校正集上表现最优（MLR：$R^2$=0.941±0.029），但存在严重过拟合现象（Δ$R^2$=0.225）；PLSR-MBA组合模型获得了最佳的交叉验证性能（$R^2$=0.807±0.031，RMSE=0.189±0.005），且过拟合程度最小（Δ$R^2$=0.078）。独立测试集验证进一步证实了PLSR-MBA模型具有更优的预测精度与稳健性（$R^2$=0.886，RMSE=0.107），显著优于MLR-*A*/*H*模型（$R^2$=0.845，RMSE=0.169）。PLSR-MBA组合模型在拟合精度、泛化能力与抗过拟合性之间实现了最佳平衡，推荐作为冬虫夏草腺苷含量快速定量分析的首选方法。本研究为中药光谱定量分析中的特征提取与建模策略选择提供了重要的理论依据与实践指导。

**关键词：**化学计量学方法；冬虫夏草；腺苷；红外光谱；辛普森法则；偏最小二乘回归；多元线性回归

（责任编辑　张　冰）