Validity and Redundancy of Spectral Data in the Detection Algorithm of Sucrose-Doped Content in Tea

LIU Meng-xuan[1,2,3,4], WU Qiong[5], WANG Xu-quan[1,2,4], CHEN Qi[5], ZHANG Yong-gang[1,2], HUANG Song-lei[1,2*], FANG Jia-xiong[1,2*]

1. State Key Laboratories of Transducer Technology, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
2. Key Laboratory of Infrared Imaging Materials and Detectors, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
3. ShanghaiTech University, Shanghai 201210, China
4. University of Chinese Academy of Sciences, Beijing 100049, China
5. Technology Center of Hefei Customs District, Hefei 245000, China

Abstract Near-infrared spectroscopy (NIRS) technology integrated with Genetic Algorithm-Back Propagation (GA-BP) neural network was used to spectral sucrose-doped content in 162 tea samples in the NIR wavelength range of 1～2.5 μm. The parameters of the GA and BP neural network were optimized by the sample set to analyze the validity and redundancy of spectral bands. The raw data in the range of 1～2.5 μm was divided into 1～1.7, 1～1.3, 1.3～1.7, 1.7～2.5 and 2～2.2 μm sets. The established quantitative detection model was used to conduct model training on different wavelength bands at the same resolution. The prediction results show that, for the target content, data redundancy appears in both 1～1.7 and 1～2.5 μm bands. The model could be effectively extracted using only 1.3～1.7 or 1.7～2.5 μm band. The prediction model was also conducted using different spectral resolutions from 2 to 20 nm in the same band. In the wavelength range of 1～2.5 μm, the R was between 0.9 and 0.95 when the RMSEP ranged from 1.7 to 2.1. While in the wavelength range of 1～1.7 μm, the R was in the range of 0.9 to 0.93 when the RMSEP was between 1.95 and 2.25. The results indicate that, for the target content, redundancy exists in the 1～2.5 and 1～1.7 μm bands on both wavelength range and spectral resolution. Through the analysis of spectral features and modeling of the algorithm, the effectiveness of spectral data acquirement could be improved dramatically; for the detection of sucrose-doped content in tea, a much narrower wavelength range and lower spectral resolution could be adopted.

Keywords Genetic algorithm; BP neural network; Near-infrared spectroscopy; Validity; Tea
中图分类号: O657.3 **文献标识码**: A **DOI**: 10.3964/j.issn.1000-0593(2022)11-3647-06

Received: 2021-10-26; **accepted**: 2022-02-23
Foundation item: Supported by the National Natural Science Foundation of China (62175250), Science and Technology Major Project of the Ministry of Science and Technology of Anhui Province (s202003a0620001), and Open project of State Key Laboratories of Transducer Technology(SKT1907)
Biography: LIU Meng-xuan, (1997—), female, master, Research on Spectral Analysis Algorithms e-mail: m15800963373@163.com
* Corresponding authors e-mail: huangsl@mail.sitp.ac.cn; jxfang@mail.sitp.ac.cn

Introduction

Tea is one of the world's most popular beverages, with a special flavor and high nutritional value. For hundreds of years, tea consumption has been expanding worldwide, leading to the frequent occurrence of adulteration[1]. In particular, there is artificial adulteration of sucrose in exported green tea. Thus, it is necessary to detect the sucrose-doped content in tea. Up to the present, sensory evaluation and wet chemical analysis are still commonly used for judging whether the tea is artificially mixed with sucrose. However, both methods have disadvantages. Sensory evaluation is easily affected by many factors, such as environmental variation and personal subjectivity. It lacks reproducibility and fairness[2]. The wet chemical analysis relies on precision instruments, such as liquid chromatography-mass spectrometry and high-performance liquid chromatography[3]. Nevertheless, these methods are high cost, time-consuming and labor-intensive. Hence, developing and implementing rapid and low-cost methods would be highly beneficial to tea industries and regulatory bodies.

NIRS is a rapid, nondestructive and large scale inspection method as a green analysis technology. Combining it with suitable chemometrics methods has been used to establish prediction models for tea categories, grades, and content of different ingredients[4-11]. Until now, there are few studies on the detection of sucrose-doped content in tea. The GA-BP neural network has the advantages of strong linear learning ability, strong feature extraction ability and strong model expression ability. The feature information of the target component in the NIR spectrum of multi-component substances can be extracted by this algorithm. Therefore, the objective of the current study is to explore the application of NIRS and the GA-BP neural network in detecting sucrose-doped content in tea.

The NIR spectrometer with high resolution and wide wavelength range contain more information and noise. To avoid losing characteristic information, all spectral data is used to build a predictive model. However, this method may introduce more noise and have data redundancy, which cannot make the model prediction effect the best[12]. To optimize the prediction results and reduce detection cost, it is necessary to study the redundancy of the full spectral range modeling in terms of wavelength range and resolution.

The methods that can be used to study spectral band redundancy has two ways. One is to divide the spectral band based on the wavelength range of the portable NIR spectrometer. Then, study the spectral band redundancy and resolution effects. The other is to select the band according to the characteristic band interval of the target substance with less interference from other components to build the model. Both methods were adopted to build a high predictability detection model to explore the validity and redundancy of spectral data.

This paper used162 samples of tea mixed with sucrose, whose spectrum was collected by an FT-NIR spectrometer. A GA-BP neural network model was applied to analyze the validity and redundancy at different spectral bands and resolutions. Moreover, a further study about whether a NIR spectrometer with a narrow wavelength range and lower resolution has the potential to detect the sucrose-doped content in tea is also qualified.

1 Experiment

1.1 Experimental Materials

A total of 162 experimental samples were from Huangshan export green tea, which was prepared by GB/T 8302—2013[13] to ensure consistency, while NIR spectra were measured by FT-NIR. The measurement model of diffuse reflectance absorbance was adopted. The scanning wavenumber range is 4 000～10 000 cm^{-1}. The wavenumber interval is set to be 0.48 cm^{-1}. The standard sucrose-doped content of the samples was in the range of 0.91%～22.6%, which was measured by high-performance liquid chromatography, according to the GB 5009.8—2016[14] standard.

1.2 Research methods

The original spectrum in the 1 ～ 2.5 μm was preprocessed by multivariate scattering correction[1]. Considering that the wavelength range of common handheld spectrometers adopting InGaAs device is normally in about 1～1.7 or 1.7～2.5 μm range with lower resolution, the raw data was divided into 1～1.7 and 1.7～2.5 μm bands. It could be seen from the NIR spectrum of the sample that there was a characteristic peak around 1.4 μm, respectively. However, there may also be the influence of moisture around 1.4 μm. In order to easily distinguish whether the characteristic peak was the interference of the target content or other components and to further study the redundancy of the 1～1.7 μm band, the 1～1.7 μm band was divided into 1～1.3 and 1.3～1.7 μm. The spectrum of tea samples contains other mixed substances, especially moisture. For the second band, the characteristic range band of 2～2.2 μm was selected by relative value analysis. The entire research schematic diagram is shown in Fig. 1. The investigation includes the effectiveness of each spectral band under the same resolution and spectral resolution effects of the same

wavelength range by the GA-BP neural network quantitative detection model. The 162 experimental samples were divided into the training set and prediction set approximately at the ratio of 3∶1, of which 120 samples were randomly used for model training, and the remaining 42 samples were only used to evaluate the model prediction results. Predictability evaluations of the detection model were based on correlation coefficient (R), and root mean square error of prediction (RMSEP).

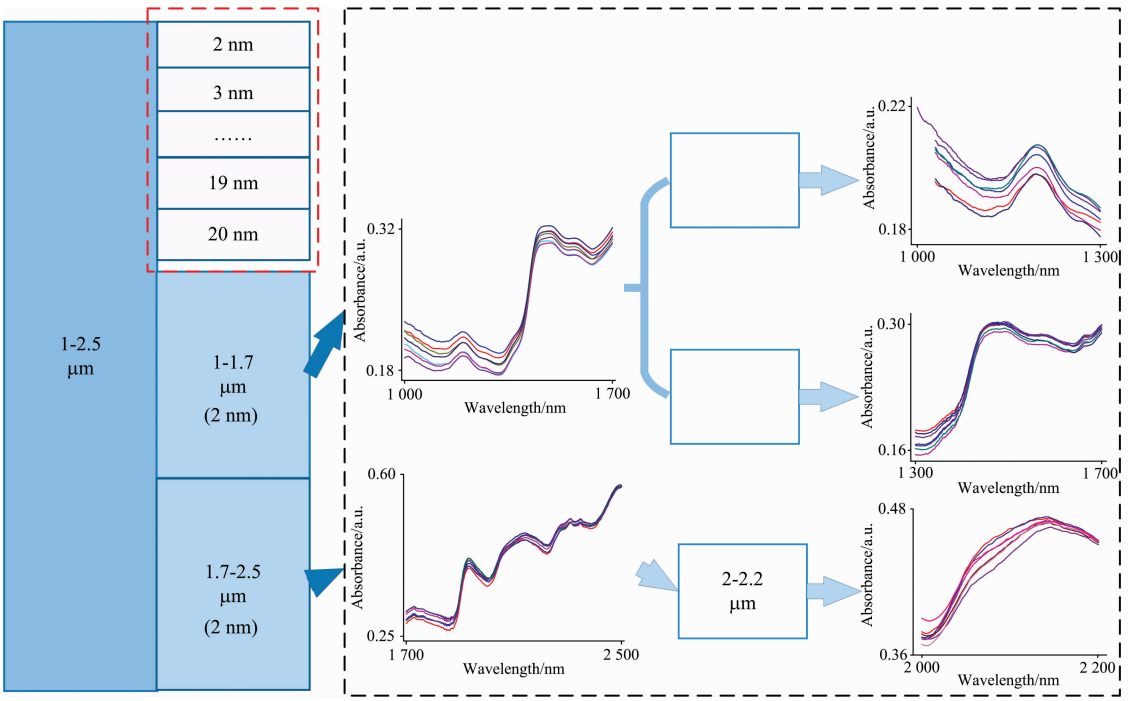

Fig. 1 Schematic diagram of analysis of different spectral bands and resolution

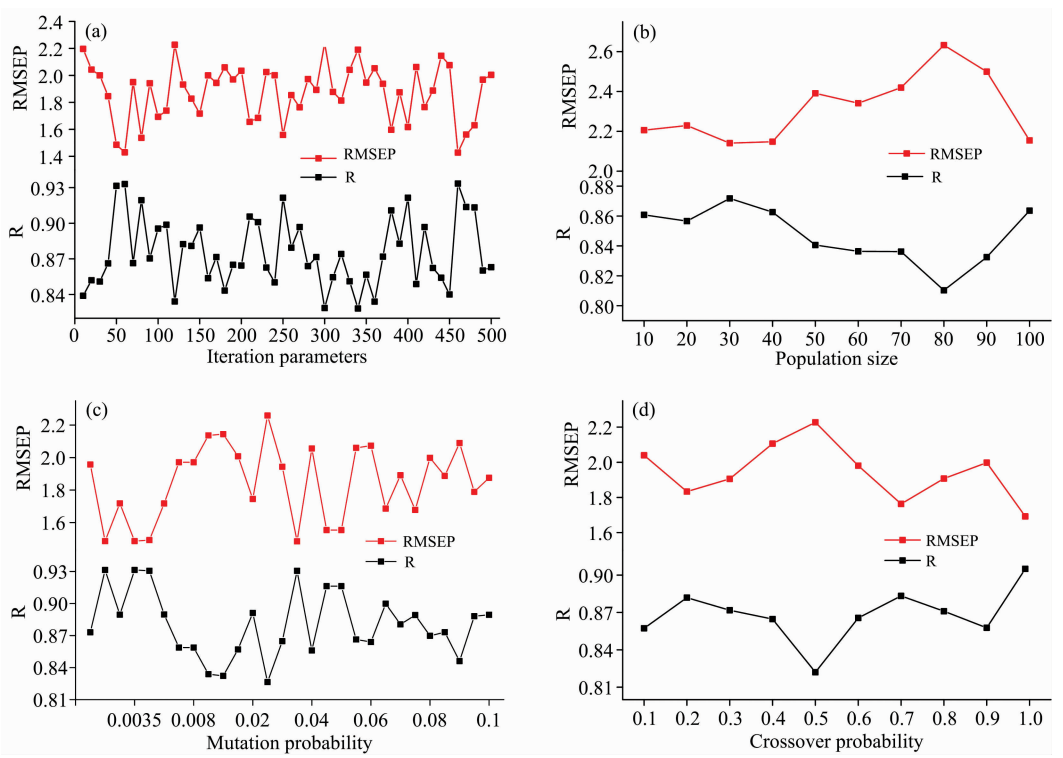

Fig. 2 Parameter selection of genetic algorithm

(a): Iteration parameters; (b): Population size; (c): Mutation probability; (d): Crossover probability

1.3 Model parameters setting

This study adopted the GA-BP neural network algorithm[16] to establish a quantitative model of sucrose-doped content in tea. Some parameters of the algorithm would affect the prediction effect of the model, which needed to be determined based on the sample set, including iteration parameters, population size, crossover probability, mutation probability in GA, as well as epochs, training target error, learning rate, training function, and node transfer function in the BP neural network.

2 Results and discussions

2.1 Parameters optimization selection

The parameter selection criterion was the R and RMSEP between the predicted value and the standard value of the 42 prediction samples which did not participate in the training. Fora better prediction effect of the model, larger R and smaller RMSEP were needed. When one parameter was changed, the other parameters and the sample set remained the same. The final result is shown in Fig. 2.

It could be seen from Fig. 2 that the iteration parameter of GA was 60. The population size was 30. The mutation probability was 0.003 5. The crossover probability was 0.99.

The BP neural network training algorithm mainly included the gradient descent method, quasi-Newton algorithm, L-M algorithm, and Bayesian regularization algorithm, which was related to the training set, the complexity of the research object, and the size of the network. Some representative training algorithms were selected to test.

Table 1 Test results of different training functions

Training algorithm	Training function	RMSEP	Time /s
Bayesian regularization algorithm	trainbr	1.62	130
Quasi-Newton algorithm	trainbfg	1.64	94
Levenberg-Marquardt	trainlm	3.93	5
One-step secant algorithm	trainoss	9.78	2
Quantitative Conjugate Gradient	trainscg	43.13	2

As shown in Table 1, the training function was trainbr.

The BP neural network had different node transfer functions, which contains three main types: logsig, tansig, and purelin. The different combination of the hidden layer and output layer node transfer functions would affect the model prediction result. The test results are shown in Table 2.

The R, RMSEP, and the model training time were comprehensively considered. The node transfer function of the hidden layer and the output layer of the neural network were purelin and tansig.

Other parameters in the BP neural network were determined by model training. The number of neurons in the hidden layer was 16. The epochs were 100. The learning rate was 0.01. The training target error was 0.000 001.

Table 2 Test results of different node transfer function

Hidden layer function	Output layer function	R	RMSEP	training time/s
logsig	logsig	0	9.99	116
purelin	logsig	0	9.99	124
tansig	logsig	0	10	168
logsig	purelin	0.68	3.2	170
tansig	purelin	0.69	3.16	171
logsig	tansig	0.78	2.61	167
tansig	tansig	0.78	2.61	170
purelin	purelin	0.92	1.63	130
purelin	tansig	0.92	1.55	130

2.2 Spectral bands analysis results

A GA-BP neural network model was established after optimizing the parameters. The resolution of different spectral bands was 2 nm. The training set and prediction set were randomly selected for multiple testing. Record the R and RMSEP respectively. The average value of 100 times was used as the final value. The prediction results of each spectral band are shown in Fig. 3.

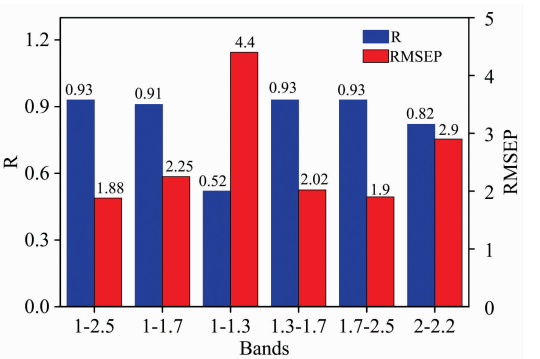

Fig. 3 Model prediction results of different spectral bands at the same resolution

Fig. 3 showed that the R of the 1～1.3 μm spectral band was the smallest, and RMSEP was the largest. It demonstrated that the 1～1.3 μm spectral band could not be used for quantitative detecting sucrose-doped content in tea. The prediction results of the models in the 1～2.5, 1～1.7, 1.3～1.7 and 1.7～2.5 μm sets indicated that these sets could be used alone to establish the sucrose-doped content detection model. The differences between the model prediction results of the four sets were minimal. Through analyzing the difference among the model prediction results in the 1～1.3, 1.3～1.7 and 1～1.7 μm sets, the 1～1.7 μm spectral band was less effective and had redundancy. The 1～

1.3 μm spectral band was invalid, which negatively impacted the model and decreased model accuracy. Comparing the results in the 1.7～2.5 and 2～2.2 μm bands, the 2～2.2 μm band could be used to establish the model, while its accuracy needed to be improved.

2.3 Spectral resolution analysis results

By averaging point values, change the resolution of the spectral band in 1～2.5 μm from 2 to 20 nm by averaging point values. The model was trained 100 times at each resolution and recorded the average value, as shown in Fig. 4.

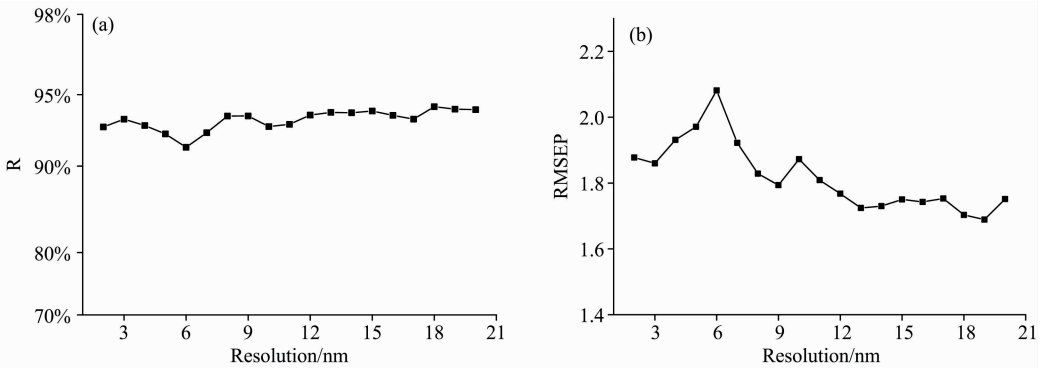

Fig. 4 Model prediction results of different resolution at 1～2.5 μm spectral band

Fig. 4 showed that R ranged from 0.9 to 0.95, and RMSEP was 1.7 to 2.1 at the different spectral resolutions. To further study the application of the portable NIR spectrometer, the 1～1.7 μm spectral band data was also used for the resolution experiment. Similarly, change the spectral resolution from 2 to 10 nm. The experimental results showed that the R was in the interval of 0.9～0.93, and RMSEP was between 1.95 and 2.25. The model prediction results of the two spectral bands indicated that the resolution had little effect on the detection model of sucrose-doped content in tea. However, the model results at low resolution were better than those at high resolution. The phenomenon may be differences in tea samples, spectral acquisition error, excessive noise in the raw data, neural network over fitting, and the method of altering the spectral resolution.

3 Conclusions

This paper mainly analyzed the validity and redundancy of spectral data by the GA-BP neural network detection algorithm of sucrose-doped content in tea. Analyze the different spectral bands at the same resolution. The prediction results showed that 1～2.5, 1～1.7, 1.3～1.7, 1.7～2.5 and 2～2.2 μm spectral bands could be used to establish a detection model, and the modeling effect of 1.3～1.7 μm was better, which conformed to the wavelength range of the portable NIR spectrometer. Analyze the different spectral resolutions at the same band. The prediction results indicated that the resolution had little effect on the model and the spectral resolution of 10～20 nm was enough for the portable NIR spectrometer. Through the analysis of different spectral bands and resolutions, the redundancy exists in 1.3～1.7 and 1～2.5 μm on both wavelength range and spectral resolution. It is of great significance to further explore the application of low spectral resolution portable NIR spectrometer in tea.

References

[1] Huang Yifeng, Dong Wentao, Alireza Sanaeifar, et al. Computers and Electronics in Agriculture, 2020, 173.
[2] Ren Guangxin, Wang Yujie, Ning Jingming, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 237.
[3] Ren Guangxin, Wang Shengpeng, Ning Jingming, et al. Food Research International, 2013, 53(2).
[4] Li Wencui, Zhou Xinqi, Fan Qiye, et al. Modern Food Science and Technology, 2021, 37(5): 303.
[5] Wang Mengdong, Wang Shengpeng. Journal of Huazhong AgriculturalUniversity, 2015, 34(1): 123.
[6] Wang Jiahua, Wang Yifang, Cheng Jingjing, et al. LWT, 2018, 96.
[7] Wang Man, Zhang Zhengzhu, Ning Jingming, et al. Science and Technology of Food Industry, 2014, 35(22): 57.
[8] Li Chunlin, Guo Haowei, Zong Bangzheng, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 206.
[9] Lee Min-Ki, Kim Heon-Woong, Lee Seon-Hye, et al. European Food Research and Technology, 2019, 245(5).
[10] Chen Suming, Wang Ching-yin, Tsai Chao-yin, et al. Vibrational Spectroscopy, 2021, 115.
[11] Wang Yuxia, Xu Rongrong, Ren Guangxin, et al. Journal of Tea Science, 2011, 31(4): 355.

[12] LIU Jing, SHANG Li-ping, QU Wei-wei, et al. Spectroscopy and Spectral Analysis, 2010, 30(10): 2685.
[13] National Standard of the People's Republic of China. GB/T 8302—2013. Tea Sampling.
[14] National Standard of the People's Republic of China. GB5009.8—2016. Determination of Fructose, Glucose, Sucrose, Maltose and Lactose in Food.
[15] Ke Pengyu, Liu Mengxuan, Wang Xuquan, et al. Journal of Infrared and Millimeter Waves, 2021, 40(5): 582.
[16] Wang Xiaochuan, Shi Feng. Analysis of 43 Cases of MATLAB Neural Network. Beijing: Beijing University of Aeronautics and Astronautics Press, 2011. 20.

茶叶掺糖含量检测算法中光谱数据有效性及冗余度研究

刘梦璇[1,2,3,4]，吴 琼[5]，王绪泉[1,2,4]，陈 琦[5]，张永刚[1,2]，黄松垒[1,2*]，方家熊[1,2*]

1. 中国科学院上海技术物理研究所，传感技术联合国家重点实验室，上海 200083
2. 中国科学院上海技术物理研究所，中国科学院红外成像材料与器件重点实验室，上海 200083
3. 上海科技大学，上海 201210
4. 中国科学院大学，北京 100049
5. 合肥海关技术中心，安徽 合肥 245000

摘 要 基于近红外光谱(NIRS)技术和遗传算法-反向传播(GA-BP)神经网络建立模型，分析茶叶掺蔗糖样品的 1～2.5 μm 原始光谱数据的有效性及冗余度。固定样本数据，对模型的参数优化选择后建立茶叶蔗糖含量定量检测模型。将 1～2.5 μm 原始数据分 1～1.7，1～1.3，1.3～1.7，1.7～2.5 和 2～2.2 μm。利用建立的模型对同一分辨率下的不同波段进行模型训练。预测结果表明，1～1.7 和 1～2.5 μm 波段存在数据冗余。仅使用 1.3～1.7 或 1.7～2.5 μm 波段即可有效建立模型。预测模型对同一波段下的不同分辨率进行研究，从 2 nm 到 20 nm 改变分辨率，当波段范围为 1～2.5 μm 时，模型的 R 均介于 0.9 和 0.95 之间，且 RMSEP 也在 1.7 和 2.1 之间。当波段范围为 1～1.7 μm 时，模型的 R 均在 0.9 和 0.93 之间，且 RMSEP 也在 1.95 和 2.25 之间。结果表明，1～2.5 μm 原始数据中确实存在波长范围和光谱分辨率的冗余。通过光谱特征分析和算法建模，可以显著提高光谱数据获取的有效性；对于茶叶中蔗糖含量的检测，可以采用更窄的波长范围和更低的光谱分辨率。

关键词 遗传算法；BP 神经网络；近红外光谱分析；有效性；茶叶

(收稿日期：2021-10-26，修订日期：2022-02-23)

* 通讯作者