Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d2ay02072f

Development of a CH_2-dependent analytical method using near-infrared spectroscopy *via* the integration of two algorithms: non-dominated sorting genetic-II and competitive adaptive reweighted sampling (NSGAII-CARS)†

Xin He,[a] Huanyu E[b] and Guoyu Ding *[b]

In most of the near-infrared studies, near-infrared spectra (NIRS) were often mathematically treated. However, these algorithms selected a large number of variables and latent variables, and they caused the over-fitting phenomenon, which became very common. The large number of variables made it impossible to extract the "chemical information" directly from the NIRS. To build robust and interpretable mathematical models, the non-dominated sorting genetic-II-competitive adaptive reweighted sampling (NSGAII-CARS) algorithm was proposed to determine influential functional groups for quantitative analysis. In this research, data on a primary mixture of two amino acids (AAs), namely $NH_2(CH_2)_3COOH$ and $HOOC(NH_2)CH(CH_2)_2COOH$, was used to illustrate the algorithm. The principle of the algorithm was first to find out the different characteristic spectral regions of two amino acids by extreme points according to Non-dominated Sorting Genetic-II (NSGAII). Second, based on the absolute value of the regression coefficient, we found out [$\nu(CH_2)$ + $2\delta(CH_2)$] and [$2\nu(CH_2)$], where the wavenumber ranged from 6165 to 5683 cm^{-1}, were the influential functional groups for quantitative analysis. Finally, the CARS (competitive adaptive reweighted sampling) algorithm was combined with NSGAII to find the specific fingerprint points for the determination of two AAs. Compared with the previous results, the NSGAII-CARS algorithm not only pointed out the influential quantitative functional groups but also used only 6 points for $HOOC(NH_2)CH(CH_2)_2COOH$ and 18 points for $NH_2(CH_2)_3COOH$ to achieve the full-spectrum quantitative effect. The results proposed a general algorithm for the quantitative analysis of NIRS obtained in the binary or ternary mixed systems. The MATLAB codes of the NSGAII-CARS algorithm are available on the website: https://github.com/Mark1988NK/NSGAII-CARS-Algorithm.git.

Received 22nd December 2022
Accepted 6th February 2023

DOI: 10.1039/d2ay02072f

rsc.li/methods

1. Introduction

With its fast and non-destructive advantages, near-infrared spectroscopy (NIRS) is competent of real-time quantitative analysis.[1] NIRS spans from 12 000 cm^{-1} to 4000 cm^{-1} and reflects the overtones and combination absorption of functional group fundamental molecular vibrations including C–H, C=O, N–H, O–H, and S–H.[2] However, the anharmonic effects of combination or overtone from NIRS create a large number of overlapping vibrational patterns without explicit spectral dependencies. Therefore, it becomes very difficult to resolve the near-infrared spectra.[3] From the point of view of this research, it is most worth studying the molecular mechanism underlying the absorption in NIRS. The NIRS can be resolved by conventional experimental methods, chemometrics, two-dimensional correlation spectroscopy, or spectral simulation, a more recent research hotspot.[3,4] Conventional experimental methods for spectral analysis are isotope exchange and polarisation measurements. For example, the bands of CH in CH_3 or CH_2 groups were analyzed to understand how the CH/CH_2 vibrations occur by replacing the H atom with a halogen atom, and the principle behind this is that C–X (X = halogen) does not show absorption in the near-infrared region.[5,6] Chemometrics has also been used to study the molecular mechanism of NIRS.[7] Particularly, regression coefficients or loading plots are helpful for band assignments.[8,9] Recently, advances in anharmonic theories, when combined with ever-increasing computer technology, have made the theoretical analysis of NIRS possible.[10] A

[a]Department of Medical Oncology, The First Hospital of China Medical University, No. 210, Baita Street, Hunnan District, Shenyang 110001, China. E-mail: hexin@cmu.edu.cn

[b]Shenyang Medical College, Huanghe North Street 146, Shenyang 110034, China. E-mail: guoyuding@mail.nankai.edu.cn

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2ay02072f

growing number of spectral simulation investigations aimed at near-infrared regions has started.[3,11] For example, the application of fully anharmonic quantum chemical calculation for a better analysis of partial least squares regression (PLSR) models of the natural product rosmarinic acid (RA) in Rosmarini folium was reported. A good agreement between the theoretical and experimental NIR spectra was obtained, and the delicate band assignments of RA were analyzed in the 8000–4000 cm^{-1} wavenumber region.[12]

In this research, for a better analysis of PLSR models in the binary or ternary mixed systems, a new algorithm to select and analyze the wavenumber variables was proposed, and it was named NSGAII-CARS algorithm. The new algorithm combined the merits of two algorithms, namely, NSGAII and CARS. NSGAII is one of the most effective multi-objective genetic algorithms, which can supply many Pareto optimal solutions for researchers to help them "have their cake and eat it". Moreover, researchers can pick the extreme points from the Pareto front solution plane. For the NIRS analysis, the picked extreme point represents the most optimal wavenumber interval combination for the corresponding dependent variable Y_i. Researchers can complete the optimization process for 2–3 dependent variables at the same time. It has to be mentioned that, in this work, the adopted NSGAII algorithm was a variant of NSGA-II, which was a controlled, elitist genetic algorithm.[13] The controlled elitist genetic algorithm also retains the individuals that can help to increase the diversity of the population even if they have a larger root mean square error of cross validation (RMSECV). It has shown that the NSGA-II with controlled elitism was more likely to search for the best fit solution than the original NSGA-II. In this research, a binary mixed system consisting of two AAs, namely $NH_2(CH_2)_3COOH$ and $HOOC(NH_2)CH(CH_2)_2COOH$, was used to illustrate the algorithm. The principle of the algorithm was first to find out the different characteristic spectral regions of two amino acids by extreme points according to the Pareto optimal solution plane obtained by NSGAII. Two extreme points represented the lowest mean square error of two AAs in the PLSR model under the same genetic algorithm process. Because the molecular structure of two AAs differs by the –COOH and the $–CH_2$/CH sequence, the selected bands by NSGAII may provide a wealth of information regarding these functional groups. However, in the actual work, the NSGAII algorithm was time-consuming when applied to the NIRS analysis since the number of wavenumber variables was too large for the subsequent genetic algorithm. Therefore, the NIRS data set between 4000 and 12 000 cm^{-1} was subdivided equally into 10–50 intervals before the NSGAII algorithm. Finally, the computing time decreased significantly. However, the selected intervals by NSGAII were still coarse for two similar AAs.

To more finely distinguish the specific fingerprint points of two AAs, the CARS algorithm was used after the NSGAII algorithm. The CARS algorithm is a kind of variable selection method that eliminates the smaller PLS regression coefficients in a constantly iterative process, and is combined with Monte Carlo sampling.[14] During the iterative process, the number of variables was descended by an exponential function. With the number of variables descending, the final selected variables could be obtained from the iterative step corresponding to the minimum fitting error. In addition, the whole process of the CARS algorithm is like the "survival of the fittest" principle. In this research, CARS was used after NSGAII to dig out the sizeable absolute regression coefficients in the selected intervals, and this provided the specific fingerprint points for the determination of two AAs.

According to our results, NSGAII helped us to find that $[\nu(CH_2) + 2\delta(CH_2)]$ and $[2\nu(CH_2)]$, where the wavenumber ranged from 6165–5683 cm^{-1}, were the influential functional groups for the quantitative analysis of two AAs. It was a pity that the carboxyl group was not identified by NSGAII, and the reason behind this will be discussed in the following chapters. After NSGAII, the CARS algorithm found that only 6 points from CH/CH_2 in $NH_2(CH_2)_3COOH$ and 18 points from CH_2 in $HOOC(NH_2)CH(CH_2)_2COOH$ were enough to achieve the full-spectrum quantitative effect.

NSGAII is introduced in the NIRS analysis for the first time in this research. The most significant result of NSGAII is the Pareto front solution plane, and "this plane" will show the lowest RMSECVs of up to three objective functions simultaneously. In our view, this research results can propose a general paradigm for the quantitative analysis of NIRS obtained in a binary or ternary mixed system.

2. Materials and methods

2.1 Binary mixed NIRS dataset

The sample data of the binary mixed NIRS dataset was obtained from the early research of our group.[15] Now we share the data and NSGAII-CARS algorithm on the GitHub. The MATLAB codes for implementing NSGAII-CARS and binary mixed system dataset are freely available on the website: **https://github.com/Mark1988NK/NSGAII-CARS-Algorithm.git**. A TENSOR 37 FT-NIR spectrometer (Bruker Optik, Ettlingen, Germany) was used to collect the dataset, and the wavenumber range ranged from 12 000 to 4000 cm^{-1} with 8 cm^{-1} intervals by averaging over 32 scans. The transmission spectrum was acquired with a cuvette of 2 mm thickness. Unlike previous research, this research used water as the background. Therefore, the transmission spectrum looked different from other research studies. In this study, 71 samples collected in the first batch of biotransformation were used as the leave-one-out cross validation (LOOCV) dataset to build the PLSR model, and 70 samples collected in the second batch of biotransformation were used as the validation dataset. Fig. 1(A) displays the biotransformation reactor from L-glutamic acid ($HOOC(NH_2)CH(CH_2)_2COOH$, L-Glu) to γ-aminobutyric acid (NH2(CH2)3COOH, GABA) *via* the glutamate decarboxylase (GAD) high-expression strain of *E. coli* BL21. The binary mixed system is the reaction solution mainly containing the substrate L-Glu and the product GABA. Of course, the binary mixed system also includes catalytic enzyme GAD and trace amounts of coenzyme pyridoxal 5′-phosphate (PLP). Therefore, we made a homemade pre-filter system, which included a series of coarse filtration (20 μm) and fine filtration

(0.2 μm) to wipe off insoluble L-Glu and enzyme. From Fig. 1(B), the baseline shift is almost impossible to observe after pre-filtering when compared with other reaction solutions involving different concentrations of enzymes. In addition, the band at around 7000 cm^{-1} is the water band. It was not strange that the band appeared as a valley, not a peak, since we used water, not air, as the background.

2.2 NSGAII-CARS algorithm

Before building the calibration model, pre-processing of the spectra can clean the noise among the data, and reasonable pre-treatment methods can remarkably improve the model performance. The pre-processing methods used in this research were convolution smoothing, auto-scaling (auto), detrend, standard normal variable transformation (SNV), detrend + SNV, SNV + detrend, multiplicative scatter correction (MSC), Savitzky–Golay first derivative (Der1), Savitzky–Golay second derivative (Der2), Der1 + detrend, Der1 + SNV and Der1 + MSC.

Each spectrum in the dataset contained 2074 variables. However, the noise was observed in the range of 5300–4872 and 4224–4000 cm^{-1} because of low optical throughput caused by water absorption.[16] Finally, only 1906 variables were reserved after eliminating the noise parts.

The NSGAII-CARS algorithm is the integration of NSGAII and CARS algorithms. The essential components of the NSGAII algorithm include chromosome structure, fitness functions, genetic operators and the NSGAII process.

In this research, the candidate chromosome structure was represented by a vector of m dimensions, where m is the number of wavenumber variables. However, the 1906 variables were too large that would be time-consuming for subsequent evolutionary multi-objective optimization. Therefore, the data set between 4224 and 12 000 cm^{-1} was subdivided equally into 10–50 intervals as the gene in the chromosome, where the gene could take either "1" or "0", namely, the selected or unselected intervals, respectively.

The goodness of each chromosome was evaluated by the fitness functions. Therefore, choosing the appropriate fitness function is another important aspect of any genetic optimization procedure. In this research, the value of RMSECV from the PLSR model was computed to evaluate the performance. This meant that the smaller the RMSECV value was, the easier it would be retained during evolution.

Genetic operators promoted the mutation of each chromosome to find out the best individuals, namely, the individuals with the lowest RMSECV of the PLSR model. Genetic operators allowed us to create new wavenumber interval combination solutions based on the existing combination solutions in the population. There are three basic types of operators: selection, mutation and crossover. Selection operation was executed before mutation and crossover. Selection chose parents for the next generation based on the principle "the smaller the RMSECV, the easier they were chosen as the parents". The selection operator provided the competitive parents for mutation or crossover. Mutation changed a certain proportion of individuals to produce many new solutions, while crossover selected two individuals and created two new individuals. In this research, random binary tournament selection was performed: 80% of the population was used for crossover by the scatter function and 5% of the population was used for mutation by uniform function.

The parameters of NSGAII are listed in Table 1 and its process proceeded as follows:

Phase 1—initialization: the initial population included 100 chromosomes randomly generated. Here, these 100 chromosomes were the 100 random combinations of the 30 split wavenumber intervals.

Phase 2—produce offspring:

(a) Random binary tournament selection was used to select parents $P^{(t)}$ for the next generation.

(b) The parents implemented the basic types of operators (mutation and crossover) to create the offspring $Q^{(t)}$ which have the same size as their parents.

Phase 3—produce parents:

(a) Merged the current population and offspring $Q^{(t)}$ into one matrix $T^{(t)}$, to reserve elitism (the mechanism could ensure that all the best individuals were passed to the next generation). Calculated the rank and crowding distance for all individuals in the $T^{(t)}$, and sorted it into different ranks according to the nondominated sorting algorithm. The following two-objective optimization problem was used to illustrate the non-dominated sorting algorithm:

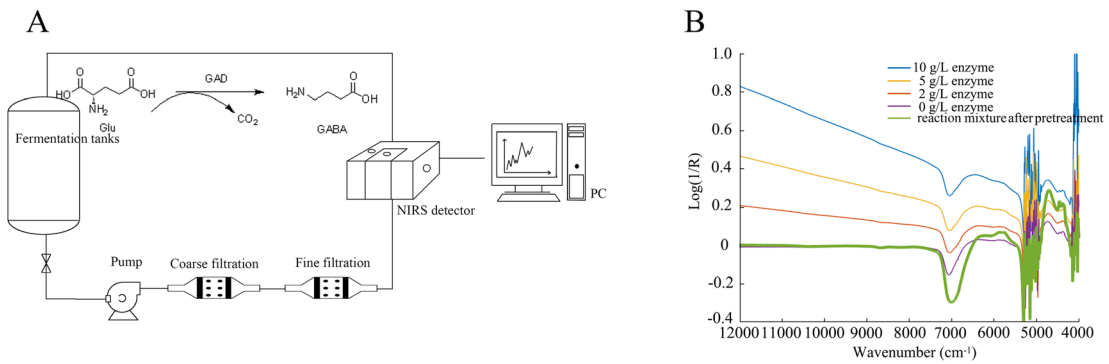

Fig. 1 (A) Schematic of a biotransformation reactor from L-Glu to GABA, (B) effect of the homemade pre-filter system.

Table 1 NSGAII parameters for optimization

NSGAII	Parameters
Fitness function	PLS
Decision variables	1906
Population size	100
Selection method	Tournament
Mutation functions	Uniform
Mutation rate	0.05
Crossover function	Scattered
Crossover fraction	0.8
Distance measure function	Distance crowding
Pareto front population fraction	0.5
Intervals/number of iterations	10/50; 20/1000; 30/3000; 40/3000; 50/5000

$$\left.\begin{array}{l}\text{Minimize } f_1(x) = \text{RMSECV for GABA } (\text{PLSR}(x_1, x_2, \ldots, x_{30}))\\ \text{Minimize } f_2(x) = \text{RMSECV for Glu } (\text{PLSR}(x_1, x_2, \ldots, x_{30}))\\ \text{Subject to } x_{1-30} = 0, 1\end{array}\right\} \quad (1)$$

$$z^* = (f_1(x), f_2(x))^T \quad (2)$$

The ideal objective vector z^* is the minimal solution to all objective functions (f_1, f_2). However, in general, the ideal objective vector is a non-existent solution. Since it can't be ensured that the minimum objective solutions of f_1 and f_2 share the same independent variable x. Therefore, collecting the Pareto-optimal solutions z^* in each generation and sorting it into different ranks according to the nondominated sorting algorithm are the normal process of multi-objective optimization. In the Pareto-optimal solutions z^*, some solutions $x^{(p)}$ dominates other solutions $x^{(q)}$, and it means:

$$\text{For } \forall i \in \{1, 2\}, \text{ if } f_i(x^{(p)}) \leq f_i(x^{(q)}), \text{ then } x^{(p)} \text{ dominates } x^{(q)} \quad (3)$$

In the nondominated sorting algorithm, the solutions that no other solutions can dominate in the z^* are ranked as the first. Then, removing the first rank out from z^*, the left solutions repeat the previous process to get the second rank. The following ranks execute the same process until all solutions are allocated.

(b) Trimmed the $T^{(t)}$ to only keep 100 individuals by retaining the appropriate number of individuals in each rank. Actually, the pre-defined distribution of number of individuals in each rank followed a geometric distribution: $n_i = r \times n_{i-1}$; where n_i is the maximum number of allowed individuals in the i-th front and r (<1) was the reduction rate.

Phase 4—stop criterion:

If the stop criterion, maximum number of generations was not satisfied, set $t = t + 1$ and went back to phase 2(b).

After the NSGAII algorithm, the wavenumber variables still contained some irrelevant information or collinear variables. Therefore, the CARS algorithm was used next to search for those specific fingerprint points among the spectral intervals selected by NSGAII. These specific fingerprint points were the least points used for quantification. In the CARS, enforced variable reduction and adaptive reweighted sampling were used to search for these specific fingerprint points. The wavelengths with higher absolute values of regression coefficients were more likely to be retained in each iterative process.[17] In detail, enforced variable reduction followed an exponentially decreasing function (EDF). In this research, to find out these specific fingerprint points, totally 50 runs were set to iteratively filter the variables with noise or small absolute regression coefficients. In the i-th run of enforced variable reduction, the number of remaining variables was calculated as follows:

$$rv_i = \text{VNS} \times e^{-k \times i} \quad (4)$$

where VNS is the initial number of variables selected from the NSGAII method. The constant k controlled the EDF curve, which could be computed as follows:

$$k = \frac{\ln(0.5 \times \text{VNS})}{N} \quad (5)$$

The constant k is positively correlated with the speed of enforced variables reduction. When $i = 0$, all the VNS variables were used to build the PLSR model; when $i = N$, the number of variables was as low as two variables. Finally, the wavelengths with smaller RMSECV survived to predict the validation dataset.

All the algorithms in this work were carried out in the MATLAB 2019b environment, and the computer was equipped with a Win10 platform with an Intel® Core i5-2600 CPU @ 2.80 GHz × 32.

3. Results and discussion

3.1 NSGAII analysis used for the binary mixed system

Fig. 1(A) illustrates the binary mixed system, and the two defined objective functions in this research are minimum RMSECVs of L-Glu and GABA in the 71 samples collected in the first batch of biotransformation. In this research, NSGAII analysis was used for the binary mixed system. NSGAII shared the same seeds to begin the evolutionary algorithm for searching the wavelength variables with the minimum RMSECVs of L-Glu and GABA simultaneously. For evolutionary algorithms, it is time-consuming when dealing with a large number of variables. Therefore, to reduce the number of variables, the full-spectrum

region was first pre-treated with 12 different spectral pre-processing methods and then split into different equidistant spectral intervals to act as the chromosomes for the evolutionary algorithm. In this research, the spectra were split into 10, 20, 30, 40, and 50 equidistant subintervals, and the evolutionary algorithm was used to search the best combination of equidistant spectral intervals. The effects of the number of subintervals (10–50) on the calibration and validation prediction performance (RMSECV and RMSEP) were evaluated, and their results are listed in supplementary Tables S1–S5.† To our surprise, improving the numbers of spectral intervals did not remarkably improve the prediction accuracy of the PLSR models. Fig. 2(A) displays that when the spectra are split into 30 equidistant spectral intervals, it has the lowest RMSEP for both GABA and L-Glu. Therefore, the number of subintervals was set as 30 to execute the NSGAII algorithm. Then, NSGAII analysis searched the most suitable interval combination for the GABA and L-Glu simultaneously. In addition, the detailed NSGAII parameters for optimization are listed in Table 1. NSGAII lasted long, 3000 iterations for 30 subintervals. According to the controlled elitist approach by Deb in 2001,[13] the obtained 100 individuals at the last iteration were ranked by five ranks with geometric distribution, as shown in Fig. 2(B). Namely, the maximum allowable number of individuals in the first rank was the highest. After that, each rank was allowed to have an exponentially reducing number of individuals. In this research, the reduction rate was set to 0.5, which corresponded to the parameter of Pareto front population fraction, as shown in Table 1. It has been demonstrated by Deb that NSGAII with controlled elitism has much better convergence properties than the original NSGAII algorithm, as the uncontrolled elitism present in NSGAII produced a large selection pressure with a lack of diversity to push the search towards better regions of optimality. At a given number of iterations, the first Pareto front (rank 1) was obtained as shown by red circles in Fig. 2(B). Two triangular frames marked the extreme points of the Pareto front plane. At last, NSGAII intuitively displayed the lowest mean square error of two AAs with the same genetic algorithm seed, and it would be beneficial for finding out the specific fingerprint points of two structurally similar AAs.

3.2 Spectroscopic study of GABA and L-Glu in the near-infrared region

Fig. 3(A) depicts the different characteristic spectral regions of two structurally similar AAs, which were searched out using NSGAII. For further analyses of the different characteristic spectral regions, 10 mg mL^{-1} aqueous GABA solution and L-Glu solution were prepared. Fig. 3(B) depicts the GABA NIR absorbance spectrum pre-treated by Der1 + detrend in the 12 000–4000 cm^{-1} region with water as the background. With the same method, L-Glu was obtained, as shown in Fig. 3(C). It should be noted that the noise part located within interval 28 was eliminated in the range of 5300–4872 cm^{-1}. In addition, the noise outside interval 30 was also eliminated in the range of 4224–4000 cm^{-1}. The noise is caused by two large water absorbance bands in the center of 5200 and 3680 cm^{-1}.[16] In detail, the noise part of 4224–4000 cm^{-1} comes from the fundamental frequency vibration of water molecules, including the ν_1 symmetric stretching of H_2O at ~3640 cm^{-1} and ν_3 antisymmetric stretching of H_2O at ~3725 cm^{-1}. In addition, the noise part of 5300–4872 cm^{-1} comes from the combination mode of water molecules, including the $\nu_2 + \nu_3$ (bending and antisymmetric stretching vibration) mode of H_2O at ~5360 cm^{-1} and $\nu_2 + \nu_1$ (bending and symmetric stretching vibration) mode of H_2O at ~5275 cm^{-1}.[18,19] The other noise part from water, for example, the first overtone of the O–H stretch centered at 6894 cm^{-1} (corresponding to no. 2 peak in Fig. 3(B)) was not excluded since the intensity of the first overtone was much lower than its fundamental or combination frequency. Therefore, the huge disturbance in intervals 21, 22, from Fig. 3(A–C), was a normal phenomenon.

For R–NH_3^+, the calculated combination overtone of $\nu_{as}(NH_3^+)$ and $\delta(NH_3^+)$ locates at ~4702 cm^{-1}.[20] However, the peak from the

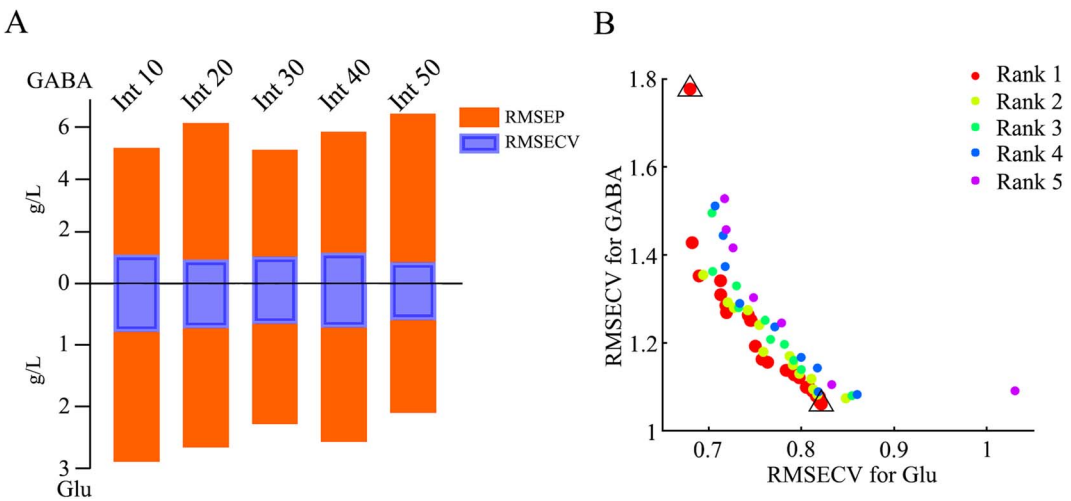

Fig. 2 (A) RMSEP and RMSECV of the final and optimal solutions to the NSGAII algorithm under different equidistant intervals. (B) Pareto front obtained by NSGAII in the maximum number of generations.

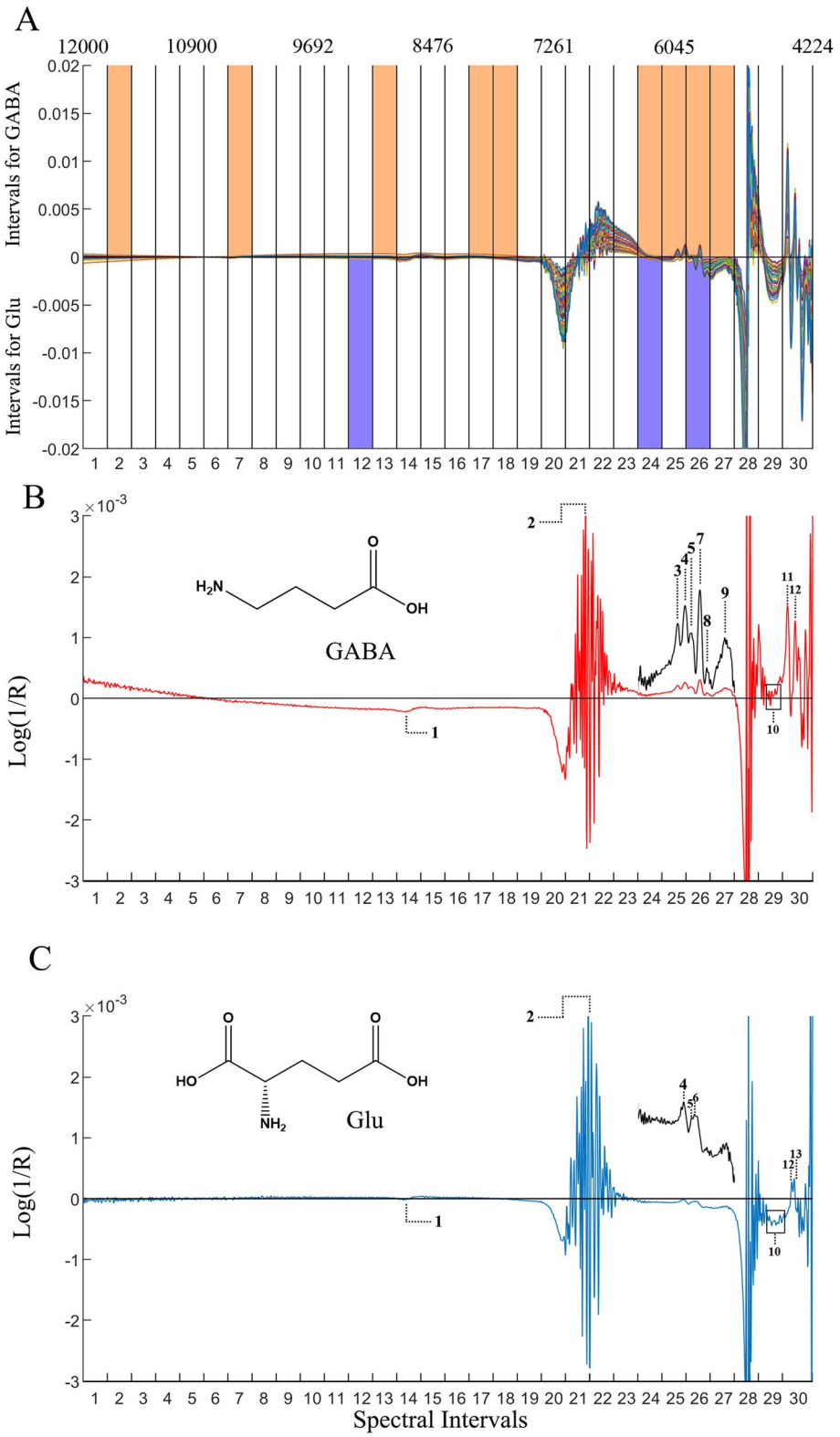

Fig. 3 (A) Selected intervals of GABA and L-Glu by NSGAII. (B) Band assignments in the NIR spectrum of GABA. (C) Band assignments in the NIR spectrum of L-Glu (the scale of the upper black spectrum was expanded by ten times against the original spectrum).

combination overtone of R–NH_3^+ was seriously disturbed by the combination overtone from water molecules. The phenomenon was also found in the first overtone of the N–H stretch (6548 cm^{-1}), which was also seriously disturbed by the first overtone of the O–H stretch (6894 cm^{-1}). Hence, the R–NH_3^+ group was not a wise choice for quantitative analysis in water.

For R–COOH, the first overtone of the O–H stretch from the carboxyl group cannot be distinguished from the O–H stretch of water. The combination of C=O stretch and O–H stretch locates at ~5300 cm^{-1}, which was also seriously disturbed by the combination overtone from water molecules.[21,22] Considering the cyclic dimers formed by AAs, the region of 4500–4700 cm^{-1} (corresponding to no. 10 peak of interval 29 shown in Fig. 3(B and C)) may be caused by a number of heavily overlapping bands of the cyclic dimers. It effectively formed a single spectral feature due to combination bands involving ν(C=O), δ_{ip}(OH) and $\nu_s(CH_2)$, $\nu_{as}(CH_2)$ modes (Table 2).[21] Although interval 29 was an effective wavelength range to quantify the AAs in theory, and it indeed displayed a concentration gradient change, as you can see in Fig. 3(A), it was not selected by NSGAII. Maybe interval 29 lacked specificity for determination, since GABA and L-Glu both share the R–CH_2COOH structure.

For the –CH_2– sequence, the region from 12 000 to 4000 cm^{-1} in the near-infrared absorption was assigned to one of the five different overtone or combination intervals, in the order of decreasing frequency, including the [3ν(CH)], {[2ν(CH)] + [δ(CH)]}, [2ν(CH)], {[ν(CH)] + [2δ(CH)]} and {[ν(CH)] + [δ(CH)]} types.[6] These five absorption groups, each of which not only located in its characteristic frequency regions but also consisted of the absorptions of a characteristic intensity level. In detail, the absorption group in the 4000–4500 cm^{-1} region (corresponding to interval 30) was assigned to the first-order combination of the [ν(CH)] + [δ(CH)] type. The absorption group in the 4500–5850 cm^{-1} region (corresponding to intervals 26–29) was assigned to the second-order combination of the [ν(CH)] + [2δ(CH)] type. The absorption group in the 5850–6100 cm^{-1} region (corresponding to intervals 25–26) was assigned to the first-order overtone of CH stretching vibrations or [2ν(CH)] type. The absorption group in the 6800–7500 cm^{-1} region (corresponding to intervals 19–21) was assigned to the second-order combination of [2ν(CH)] + [δ(CH)] type. The absorption group in the 8500–9000 cm^{-1} region (corresponding to intervals 13–15) was assigned to the second-order overtone of CH stretching vibrations or [3ν(CH)] type.

Table 2 Band assignments in the NIRS of GABA and L-Glu

Band number	Wavenumber (cm^{-1})	Band assignment
1	8747	$3\nu(CH_2)$
2	6894	2ν(OH)
3	6015	$2\nu_a(CH_2)$
4	5941	$\nu_a(CH_2) + \nu_s(CH_2)$
5	5876	$2\nu_s(CH_2)$
6	5845	2ν(CH)
7	5783	$\nu(CH_2) + 2\delta(CH_2)^a$
8	5714	$\nu(CH_2) + 2\delta(CH_2)$
9	5532	$\nu(CH_2) + 2\delta(CH_2)$
10	4673–4569	(νC=O, δ_{ip}OH) + ($\nu_{as}CH_2$, ν_sCH_2)
11	4479	$\nu_a(CH_2) + \delta_b(CH_2)$
12	4436	$\nu_s(CH_2) + \delta_b(CH_2)$
13	4401	ν(CH) + δ_{ip}(CH)

[a] The band was assigned to the second-order combination of $\nu(CH_2)$ + $2\delta(CH_2)$ and was enhanced by the Fermi resonance.

In the absorption region of [ν(CH)] + [δ(CH)] type, or interval 30, AAs showed a large number of overlapped absorptions. The deformation modes of the CH_2 group include bending, twisting, wagging and rocking, which are specified as δ_b(CH), δ_t(CH), δ_w(CH) and δ_r(CH) respectively. The stretching modes of the CH_2 group include symmetric and antisymmetric CH, which are specified as ν_s(CH) and ν_a(CH), respectively. Bands could be assigned for CH_2X_2, however for CH_2XCHX_2, $CH_3(CH_2)_5CH_3$, or more complex molecules, it should only be assigned in principle.[6] Here, we deduced that peaks 11 and 12 may be the first-order combination bands of $\nu_a(CH_2)$ + $\delta_b(CH_2)$ and $\nu_s(CH_2)$ + $\delta_b(CH_2)$. In addition, peak 13 was specific for L-Glu. It may be assigned to the first-order combination bands of ν(CH) + δ_{ip}(CH). The detailed information is listed in Table 2. In addition, interval 30 included the interference of combination of ν_s HB bridge and δ_{ip} HB bridge at 4282 cm^{-1} (combination mode of symmetric stretching and in-plane deformation for the double hydrogen bonded O–H···O bridge), which in theory contributed mainly to the increase in NIRS baseline.[21] Hence, the [ν(CH)] + [δ(CH)] type was also not a wise choice for quantitative analysis.

In the absorption region of [ν(CH)] + [2δ(CH)] type, or the intervals 26–29, they are much weaker than that of the bands in the first-order overtones or combinations, as observed from the spectra in Fig. 3(B and C). It was difficult to assign these bands to relevant ν(CH) and δ(CH) modes, since the second-order combination bands of the [ν(CH)] + [2δ(CH)] type may be influenced by a delicate structural difference in CH_2.[6] Here, peaks 8 and 9 were roughly assigned to [$\nu(CH_2)$ + $2\delta(CH_2)$] in Table 2. However, the absorption group in the 4500–5850 cm^{-1} region was very complex and diversified. In this region, it included not only the [ν(CH)] + [2δ(CH)] type, but also the combination mode of water molecules (interval 28), combination overtone of $\nu_{as}(NH_3^+)$ and $\delta(NH_3^+)$ (interval 29) and a series of combination overtone from R–COOH, as discussed above. Only part of interval 26 was undoubtedly attributable to the combination of [ν(CH)] + [2δ(CH)].

In the absorption region of [2ν(CH)] type, or intervals 25–26, whose intensity is the second strongest among C–H vibration absorption. For CH_2X_2, CH_2XCHX_2, and $CH_3(CH_2)_5CH_3$, they both share the normal modes including $2\nu_a(CH_2)$, $\nu_a(CH_2)$ + $\nu_s(CH_2)$ and $2\nu_s(CH_2)$. The distinct bands 6015, 5941 and 5876 cm^{-1} in Fig. 3(B and C) should be assigned to the above-mentioned three normal modes. Peak 7, or 5783 cm^{-1}, has the biggest intensity with the increasing separation from the adjacent peak. From this behavior, peak 7 was assigned to the second-order combination of $\nu(CH_2)$ + $2\delta(CH_2)$ and was enhanced by the Fermi resonance.[6] Peak 6 or 5845 cm^{-1} was unique to L-Glu, and it was assigned to 2ν(CH). In general, the region of [2ν(CH)] type was less affected by other functional groups in wavenumber and it has sufficient sensitivity and hydrophobic properties, which could act as the effective functional group for quantitative analysis in water.

For the second-order combination of [2ν(CH)] + [δ(CH)] type or second-order overtone of CH stretching vibrations, the occurrence probabilities of triple frequency are very low, so it has a very low signal-to-noise ratio. For example, you can see the

$3\nu(CH_2)$ peak 1 at 8747 cm^{-1} in Fig. 3(B and C), and it is not suitable for quantitative analysis in theory. Hence, the intervals selected by NSGAII were redundant.

For further analyses of the cause of redundant wavenumber, in Table S3,† it was found that the latent variables (LVs) selected by NSGAII are 15 for GABA and 14 for L-Glu. Latent variables of PLSR model are other important influencing factors besides the spectral pre-treatment method. Too many latent variables will cause the risk of over-fitting. Here, 14 or 15 LVs are too large. Therefore, the RMSEC of PLSR *vs.* optimal LV plot is displayed in Fig. 4(A). From Fig. 4(A), it can be found that four LVs are enough to build the calibration model. More variables could not further improve the performance of the model. In contrast, it would cause the risk of over-fitting. Therefore, for the NSGAII algorithm, the LVs will be set to 4 in the future research. Fig. 4(B and C) shows the regression coefficients for GABA and L-Glu under four LVs. It can be found that some intervals are indeed redundant. Only intervals 25, 26 were suitable for the quantitative analysis of GABA and interval 26 was suitable for L-Glu. Interestingly, on interval 26, the regression coefficients of GABA and L-Glu are the opposite numbers. This phenomenon can also be understood that when L-Glu is transformed into GABA, the content of L-Glu is decreased and that of GABA is increased.

Finally, it can be concluded that $[\nu(CH_2) + 2\delta(CH_2)]$ and $[2\nu(CH_2)]$, where the wavenumber ranged from 6165 to 5683 cm^{-1}, or intervals 25, 26 were the influential functional groups for the quantitative analysis of L-Glu and GABA.

3.3 Wavenumber point quantification using the NSGAII-CARS algorithm

Building a simple and efficient PLSR model for multicomponent determination is necessary to make the miniaturized

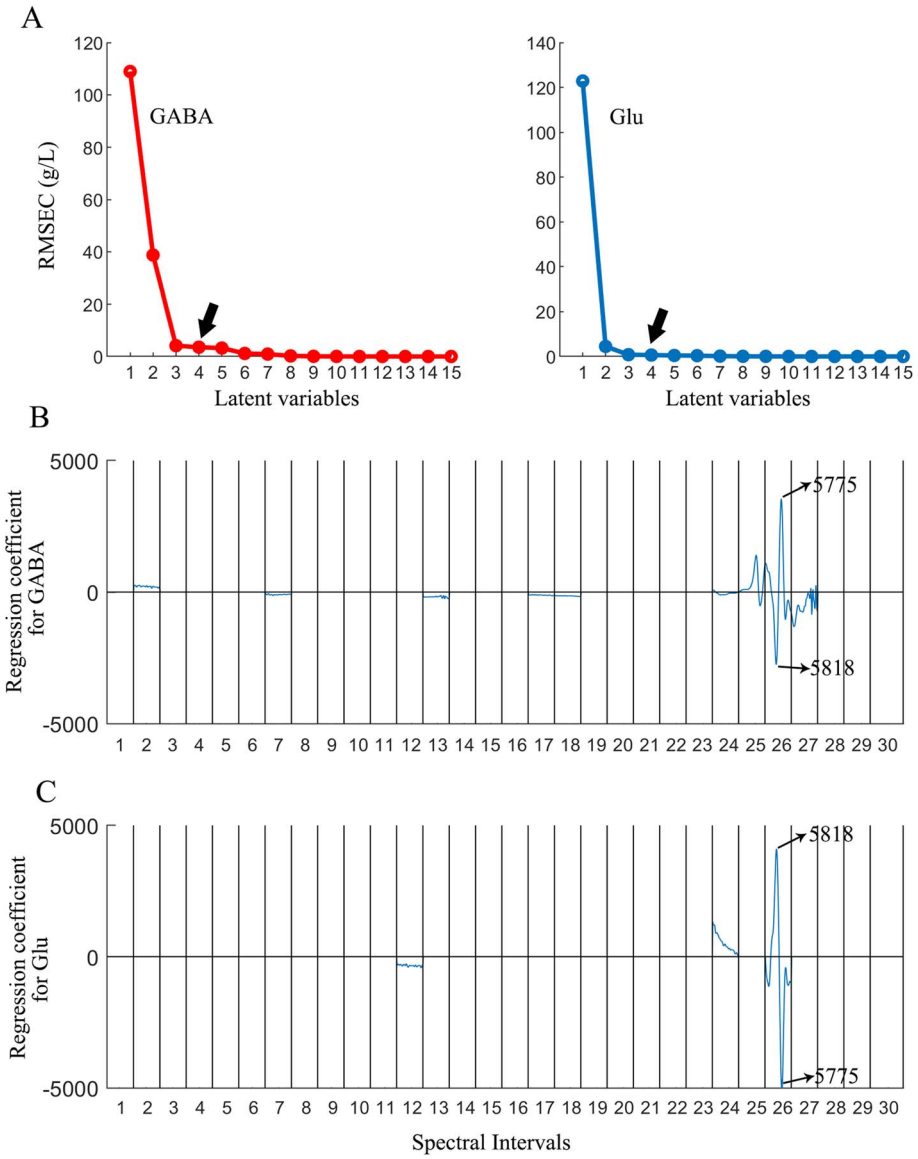

Fig. 4 (A) RMSEC *versus* the numbers of latent variables of PLS regression. (B) Regression coefficients for the calibration of GABA. (C) Regression coefficients for the calibration of L-Glu.

Table 3 Comparison of three different wavenumber selection algorithms

Compound	Method	Pre-treatment	Intervals	n_{Var}	LVs	RMSECV	RMSEP	Mean	R^2(cal)	R^2(val)
GABA	NSGA-II	Der1 + detrend	2, 7, 3, 17, 18, 24–27	567	4	2.03	4.52	79.49	0.9980	0.9900
GABA	CARS	Der1 + detrend	21, 29, 30	15	**5**	1.24	5.18	79.49	0.9994	0.9868
GABA	NSGA-II-CARS	Der1 + detrend	25, 26	18	4	1.09	3.31	79.49	0.9996	0.9946
L-Glu	NSGA-II	Der1 + detrend	12, 24, 26	189	4	0.87	1.09	19.00	0.9959	0.9947
L-Glu	CARS	Der1 + detrend	21, 22, 26, 28, 29, 30	64	8	0.53	1.64	19.00	0.9992	0.9880
L-Glu	NSGA-II-CARS	Der1 + detrend	26	6	2	0.70	1.66	19.00	0.9976	0.9877

instrument, and this decreases the equipment costs. Given this, some innovative wavelength selection algorithms have been developed to search for the specific fingerprint points.[23,24] NSGAII and regression coefficient analysis has made us realize that intervals 25, 26 were the influential functional groups for quantitative analysis. However, these large wavelength variables (including 126 variables in two intervals and 63 variables in one interval) made the spectral resolution still complicated, and it has not achieved the aim of fine distinction between CH_2 in GABA and CH in L-Glu. Therefore, to search for the specific fingerprint points from the selected intervals by NSGAII, the CARS algorithm was adopted with its advantage of enforced variable reduction. Different from the NSGAII algorithm, which is based on the principle of genetic evolution, the CARS algorithm retains the wavelengths with large absolute regression coefficients in the PLSR model. It should be noted that the disadvantage of CARS is that the serious over-fitting will occur when some uninformative variables (or no chemical dependent information) but with large absolute regression coefficients are introduced.[25] When the CARS algorithm was combined with

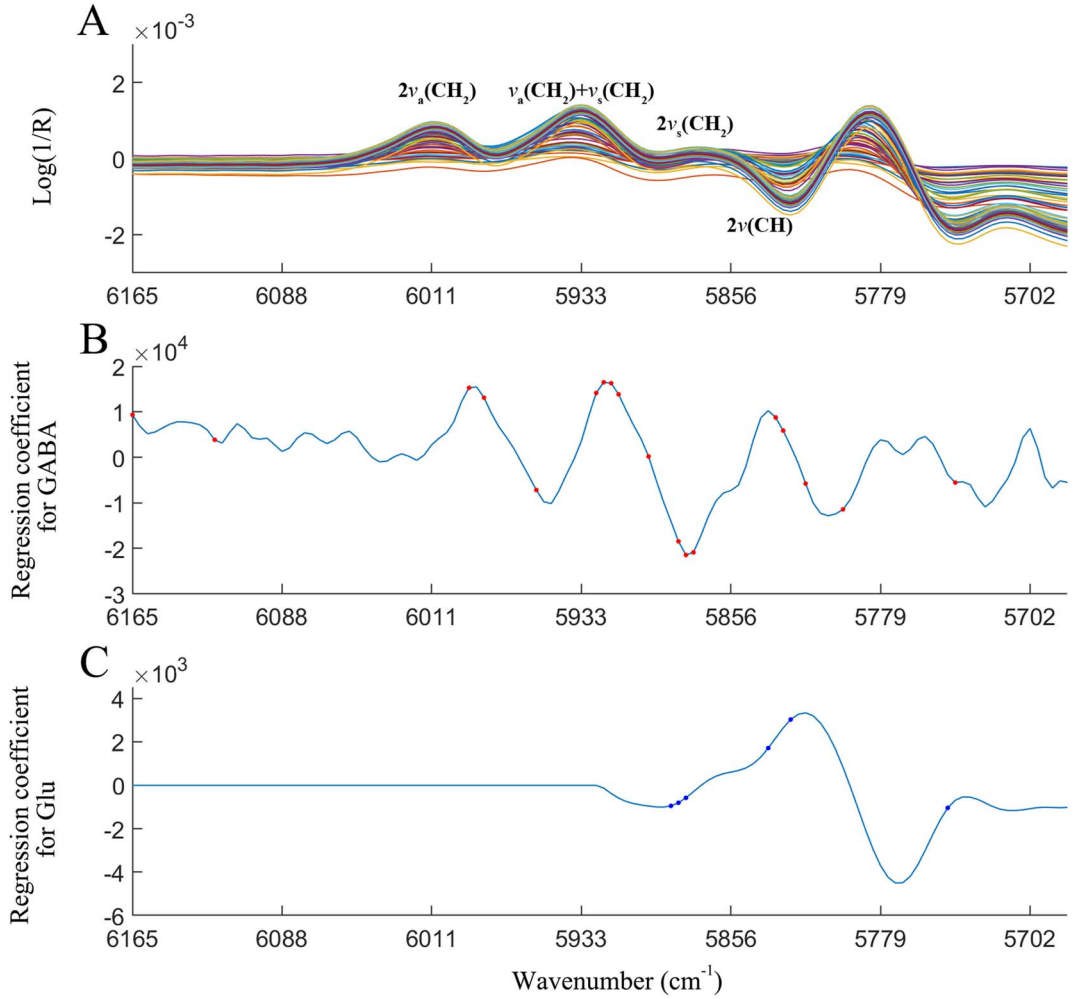

Fig. 5 (A) Near-infrared spectra of the binary mixed system in the 6165–5683 cm^{-1} region. (B) Regression coefficients for the calibration of GABA. (C) Regression coefficients for the calibration of L-Glu.

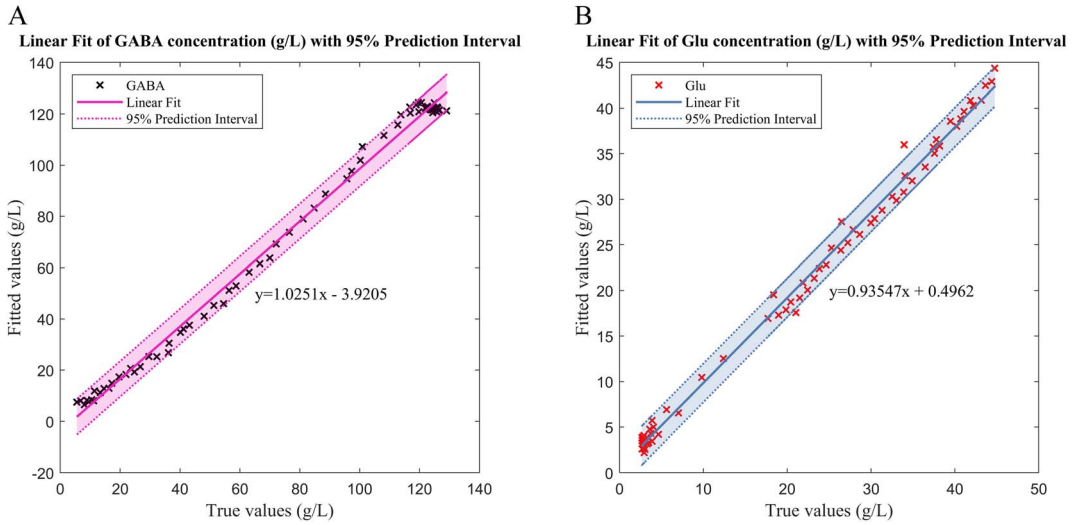

Fig. 6 (A) Fitted values *vs.* true values for GABA in the validation set. (B) Fitted values *vs.* true values for L-Glu in the validation set.

NSGAII, NSGAII remedied the disadvantages of CARS. NSGAII first located the optimal informative quantitative regions, under which CARS searched for its specific fingerprint points. In Table 3, the NSGAII algorithm is compared with the NSGAII-CARS algorithm, and it is observed that similar fitting results (RMSEP, R^2) are achieved by these two kinds of algorithms. However, NSGAII-CARS requires even fewer points to fulfill the task of excellent prediction. The CARS algorithm is also compared with the NSGAII-CARS algorithm, and the CARS method also displays a good fitting effect. For GABA, CARS used only 15 points to fulfill the full-spectrum quantitative effect, but the fitting effect is worse than that of NSGA-II-CARS. For L-Glu, the number of points used by CARS is a little more than that used by NSGA-II-CARS, but the fitting effect is better than that of NSGA-II-CARS. The selected points by CARS include intervals 21, 22, 29, 30 (from 7137–6655 and 4343–4224 cm^{-1}); 7137–6655 cm^{-1} is the zone of first overtone of the O–H stretch caused by water, and 4343–4224 cm^{-1} is the zone of $[\nu(CH)] + [\delta(CH)]$ combination overtone, however, this zone also included the interference of combination of ν_s HB bridge and δ_{ip} HB bridge at 4282 cm^{-1}. Hence, intervals 21, 22, 29, 30 selected by CARS are not a wise choice for the quantitative analysis of the molecular mechanism underlying absorption in NIRS. NSGA-II does not recommend these intervals, but advises intervals 25, 26. In a word, CARS and NSGA-II-CARS both give a good fitting effect, but they select different intervals according to their own algorithm theory.

Fig. 5(A) depicts the NIRS outline from 6165 to 5683 cm^{-1} and C–H band assignments in the region of $[2\nu(CH_2)]$ type. As you can see from Fig. 5(B), CARS selects the points from $2\nu_a(CH_2)$, $\nu_a(CH_2) + \nu_s(CH_2)$ and $2\nu_s(CH_2)$ for GABA. From Fig. 5(C), CARS selects the points from $2\nu_s(CH_2)$ and $2\nu(CH)$ for L-Glu. Compared with the previous results, the NSGAII-CARS algorithm not only pointed out the effective quantitative functional groups but also used only 6 points for L-Glu and 18 points for GABA to achieve the full-spectrum quantitative effect. The models were built based on the 71 samples collected in the first batch of biotransformation with these wavenumber points obtained by the NSGAII-CARS algorithm. In addition, 70 samples collected in the second batch of biotransformation were used as the validation. Fig. 6(A) and (B) shows the fitting effects for GABA and L-Glu in the validation. Excellent fitting effects confirm the feasibility of the NSGAII-CARS algorithm.

4. Conclusion

Because of the large number of overlapping vibrational variables from NIRS, in most cases, its applications were almost based on the mathematical treatments of the data but not on analytical knowledge or chemical information that the NIRS could give. In the present work, the NSGAII-CARS algorithm was developed to understand the band assignments of CH_2 and CH groups. NSGAII is capable of searching for the different characteristic spectral regions of the binary or ternary mixed systems by extreme points according to the Pareto front plane. CARS is capable of providing the subtle wavenumber point quantification. Since NSGAII belongs to the population evolutionary algorithm, it will be time-consuming when facing a huge number of variables. In this research, the spectral range was divided into equidistant intervals before NSGAII. However, in future research, it is better to divide the wavelength range "personalized" according to these different functional groups (–COOH, $–NH_2$, $–CH_2$, –CONH–, *etc.*) appearing in the compounds. The MATLAB codes for implementing NSGAII-CARS and binary mixed system dataset are freely available on the website: **https://github.com/Mark1988NK/NSGAII-CARS-Algorithm.git**.

Author contributions

Xin He, provided the research goals and aims, and presentation, preparation and creation of the published work. Huanyu E is responsible to maintain research data for initial use and later

re-use. Guoyu Ding is responsible for programming, designing computer programs; running of the computer code and algorithms supporting.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by 2019 Scientific Research Fund of Shenyang Medical College (Grant Number 20191018), 2019 Shenyang Medical College Doctoral Research launch fund (Grant Number 20195072) and 2021 General Project of Liaoning Provincial Department of Education (Grant Number LJKZ1138).

References

1 D. Biagi, P. Nencioni, M. Valleri, N. Calamassi and P. Mura, *J. Pharm. Biomed. Anal.*, 2021, **204**, 114277.
2 J. J. Workman, *Appl. Spectrosc. Rev.*, 1996, **31**, 251–320.
3 K. B. Bec and C. W. Huck, *Front. Chem.*, 2019, **7**, 48.
4 Y. Ozaki, S. Sasic and H. H. Jiang, *J. Near Infrared Spectrosc.*, 2001, **9**, 63–95.
5 R. R. Iwamoto, A. A. Nara and T. T. Matsuda, *Appl. Spectrosc.*, 2005, **59**, 1393–1398.
6 R. Iwamoto, A. Nara and T. Matsuda, *Appl. Spectrosc.*, 2006, **60**, 450–458.
7 Y. Dong, B. Xiang, Y. Geng and W. Yuan, *Chemom. Intell. Lab. Syst.*, 2013, **126**, 21–29.
8 T. Furukawa, M. Watari, H. W. Siesler and Y. Ozaki, *J. Appl. Polym. Sci.*, 2003, **87**, 616–625.
9 S. Saranwong and S. Kawano, *J. Near Infrared Spectrosc.*, 2008, **16**, 497–504.
10 K. B. Bec, J. Grabska and C. W. Huck, *Spectrochim. Acta, Part A*, 2022, **279**, 121438.
11 K. B. Bec, J. Grabska, C. G. Kirchler and C. W. Huck, *J. Mol. Liq.*, 2018, **268**, 895–902.
12 C. G. Kirchler, C. K. Pezzei, K. B. Bec, S. Mayr, M. Ishigaki, Y. Ozaki and C. W. Huck, *Analyst*, 2017, **142**, 455–464.
13 K. Deb and T. Goel, presented at *the International Conference on Evolutionary Multi-Criterion Optimization*, Berlin, Heidelberg, 2001.
14 H. Li, Y. Liang, Q. Xu and D. Cao, *Anal. Chim. Acta*, 2009, **648**, 77–84.
15 G. Ding, Y. Hou, J. Peng, Y. Shen, M. Jiang and G. Bai, *J. Pharm. Anal.*, 2016, **6**, 171–178.
16 H. Chung, M. A. Arnold, M. Rhiel and D. W. Murhammer, *Appl. Spectrosc.*, 1996, **50**, 270–276.
17 Q. Luo, Y. Yun, W. Fan, J. Huang, L. Zhang, B. Deng and H. Lu, *RSC Adv.*, 2015, **5**, 5046–5052.
18 W. A. P. Luck and W. Ditter, *J. Phys. Chem.*, 1970, **74**, 3687–3695.
19 G. Della Ventura, F. Radica, F. Bellatreccia, A. Cavallo, F. Capitelli and S. Harley, *Contrib. Mineral. Petrol.*, 2012, **164**, 881–894.
20 S. Holly, O. Egyed and G. Jalsovszky, *Spectrochim. Acta, Part A*, 1992, **48**, 101–109.
21 J. Grabska, K. B. Bec, M. Ishigaki, M. J. Wojcik and Y. Ozaki, *Spectrochim. Acta, Part A*, 2017, **185**, 35–44.
22 K. B. Bec, Y. Futami, M. J. Wojcik, T. Nakajima and Y. Ozaki, *J. Phys. Chem. A*, 2016, **120**, 6170–6183.
23 B. C. Deng, Y. H. Yun, P. Ma, C. C. Lin, D. B. Ren and Y. Z. Liang, *Analyst*, 2015, **140**, 1876–1885.
24 Y. H. Yun, W. T. Wang, M. L. Tan, Y. Z. Liang, H. D. Li, D. S. Cao, H. M. Lu and Q. S. Xu, *Anal. Chim. Acta*, 2014, **807**, 36–43.
25 G. Ding, Y. Wang, A. Liu, Y. Hou, T. Zhang, G. Bai and C. Liu, *RSC Adv.*, 2017, **7**, 22034–22044.