Correcting the baseline drift without human knowledge

Yuanjie Liu (yjliu@cau.edu.cn)

Methodology

Keywords: baseline correction, parameter free, deep learning

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-33007/v1

Correcting the baseline drift without human knowledge

Yuanjie Liu

College of information and electrical engineering, China Agricultural University

Abstract: Baseline drift occurs universally in many sorts of analytical chemistry data, such as in MALDI-TOF, Raman, infrared and XRD spectra. In the era of big data, automatic correcting methods are eagerly demanded. However, traditional baseline correction methods are impossible to execute fully automatically. They always depend on some preset parameters. To build parameter-free methods, utilizing current intelligent algorithms is the best choice. However, it is a great challenge to provide a huge number of labeled samples which are required for effective training. In this article, a novel strategy has been developed to train a deep neural network successfully for baseline correction. The impossible mission of preparing millions of manually processed training samples was avoided. Under the new scheme, the power of deep learning was freely applied to achieve straightforward full automation in baseline recognition. Numerical experiments on authentic datasets indicated that the new intelligent model outperformed traditional methods.

Keywords: *baseline correction; parameter free; deep learning.*

INTRODUCTION

The spectrum (such as mass spectrum or Raman spectroscopy to which the baseline correction method of this article will be applied) usually contains peaks and noise superimposed on the background. This background or baseline is caused by many factors and is usually inevitable.

The baseline can be flat or straight lines with positive and negative slopes, or curves, or any combination of them. Its main feature is that the change is much slower than the signal peak. In other words, baseline drift is a global low-frequency noise. The baseline of the spectrum in analytical chemistry can severely impair the effectiveness of the signal, limit the available analytical methods, and affect sensitivity. Therefore, baseline correction is one of the most important data preprocessing steps in chemometric analysis. Subtracting the estimated background from the original spectrum results in a more easily interpretable signal, which determines the wavenumber of the signal peak and more accurately measures the area and amplitude of the peak. In order to carry out further qualitative or quantitative analysis, it is necessary to correct the background using an effective baseline recognition method.

Existing background estimation methods include (minimized) least squares [1], wavelet analysis [2], moving average [3], orthogonal basis decomposition [4], corner cutting [5] and other expertise [6] [7]. All these methods are developed based on

specific mathematical skills. However, a well-designed method usually encounters the same problem due to its corresponding classic mathematical technique: the more elegant the method, the more limited the scope of application. At least in the areas we discuss, these methods are classified as traditional ones and require fine-tuning of control parameters to ensure correct operation. Different processing examples require different preset parameters. Therefore, the automation level of these algorithms is severely restricted.

In the era of big data, analytical data is growing rapidly, and methodologies for complex analysis based on large-scale data sets have emerged. For example, cutting-edge technologies such as identifying the microbial species based on mass spectrometry data sets of biological macromolecules, or direct diagnosis of thalassemia based on analytical data. Since it is impossible to pre-process large amounts of raw data manually, full automation of baseline correction becomes increasingly important.

With the development of artificial intelligence, carefully designed learning systems have been able to automatically complete various difficult pattern recognition tasks. In some areas, such as image classification [8] and chess games [9], the performance of deep learning algorithms is even better than that of human. It is a natural idea to apply deep learning models to baseline detection. The main challenge to prevent this application comes from the preparation of training samples. As one of the most salient features, deep learning models require a large amount of training data to implement a kind of statistical regression with brute force. If trying to train a network model with sufficient size to achieve baseline detection, it is necessary to construct hundreds of thousands of spectra with known baselines to form a training data set. This requires extensive efforts to draw baselines for the spectra covering various situations, and its high cost is unaffordable. Early attempts in chemometrics, such as the application in Raman spectral data analysis for classification [10], have proved the advantages of deep learning methods in chemical information processing.

In this paper, by synthesizing training data, the author further shows the potential of deep learning in cross-type analytical data processing. An effective strategy is established to automatically generate a large number of labeled samples instead of generating them manually, thereby making effective training possible. The new scheme makes it possible to establish a completely autonomous process to obtain a fully automatic baseline corrector. The entire training process becomes unmanned: a special algorithm generates any number of samples with accurately known signals, noise, and baseline; use these samples to train a deep residual neural network, so that the model can directly output the processed result for the original signal trace; the trained model can be applied to the removal of baseline of any one-dimensional spectrum signal curve. The new algorithm is named Alpha Baseline Correction, and for convenience, it is referred to as AlphaBC for short.

To illustrate the advantages of the new scheme in terms of accuracy and efficiency, three important methods were selected for comparison: airPLS [11], wavelet method [12], and the background correction function in MestReNova [13]. The airPLS is a widely used method and has received a lot of citations since its publication. If the parameters are selected accurately, the results produced by this method are superior in most cases. The wavelet method has a long history and is often used for photoelectric

signal analysis. It can be considered that these two methods cover the two main technical routes of baseline correction: iteration and decomposition. MestReNova is a well-known commercial software focused on analytical chemistry. It is used by thousands of chemists in academic and industrial environments. Comparing AlphaBC with popular commercial products like MestReNova can better illustrate its advantages. The experiments in the following sections show that the AlphaBC method proposed in this article is more effective and robust.

METHODS

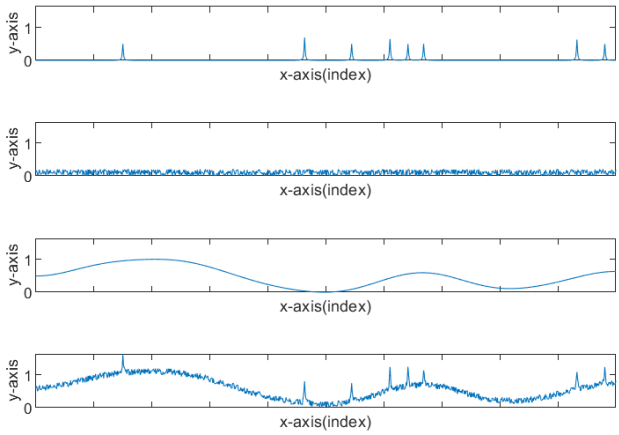


Figure 1. The procedure of generating a training sample. From top to bottom: the first subfigure presents randomly set signals generated according to Formula (2); the second one presents noises sampled from uniform distribution; the third one presents a baseline generated according to Formula (1); the last one presents the simulated training sample generated by adding the three portions together.

RANDOM SAMPLE GENERATION FOR TRAINING

The first part of the new method is to construct training samples that cover enough cases. This is achieved by using Fourier series. According to Fourier's theorem, any function can be represented by the sum of a series of trigonometric functions:

$$f(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos k\omega_0 x + b_k sin\ k\omega_0 x) \qquad (1)$$

Fourier analysis is widely used in signal processing scenarios for decomposition and frequency filtering. Here, it is used in reverse to construct an arbitrary curve to simulate a baseline. The specific method is to randomly select some items with a longer period in the summation formula, and then multiply with random coefficients to produce a slowly undulating waveform. Theoretically, this ensures that the simulated baseline is sampled from a sufficiently wide range.

A Gaussian-like peak is used to simulate the signal, which is a widely adopted standard model. The radial basis function (RBF for short) is selected as the specific implementation, defined as the following formula:

$$f(x) = \frac{h}{1 + w + (x - x_0)^2} \qquad (2)$$

Simulated peaks with different heights, widths and centers can be generated by randomly selecting the parameters h, w and x_0 in the RBF model. The noise is generated by uniformly distributed sampling. By adding the simulated baseline, signal peak, and noise together, an approximate profile that covers almost arbitrary instance is synthesized. The entire generating procedure is shown in Figure 1.

THE STRUCTURE OF THE DEEP NEURAL NETWORK

The baseline recognition task is performed by a 17-layer deep residual convolutional network [14] [8]. Compared with models with dozens or even hundreds of layers in typical deep learning applications such as computer vision, the size of the network is just a beginner. As far as the processing of one-dimensional data is concerned, such a network size is sufficient to achieve excellent performance. The complete training can be fulfilled by a portable computer within a few hours only through its CPU, and the processing of a single item using the trained model can be achieved within a few milliseconds. If the GPU is used, the speed can be increased many times. The test results show that the deep learning model is outstanding in removing baseline drift. In theory, deeper networks will perform better.

The author has developed all computing and interface systems using python language, and the PyTorch deep learning framework is used for model construction and training. The specific model structure is shown in Table 1. The trained model has been deployed on a server, and any user can use this baseline correction function by accessing the address 119.3.245.82:5000. At present, the system is very simple, but it is also very easy to use. Users only need to store the intensity of the signal curve in a comma separated CSV file according to the arrangement of the columns, and then upload it. The processed results will be displayed directly on the webpage, as shown in Figure 2. The original intensity curve and the recognized baseline will be displayed in a three-dimensional rectangular coordinate system. If there are multiple input traces, they will be arranged in sequence along the y-axis, and the detection effect of a series of curves will be clear at a glance. In the actual web page, the 3D view can also be rotated, zoomed, and panned to facilitate observation from all angles. Due to different input resolutions, before uploading, users need to resample the original trace to ensure that the span of a single peak is less than 35 data points. Baseline recognition results can also be downloaded as CSV files to use freely. The intensity of each point of the results in the downloaded file is stored in the

corresponding column according to the original arrangement. Taking into account the users' general preferences, the results after the baseline removal, rather than the baseline itself, are provided for downloading.

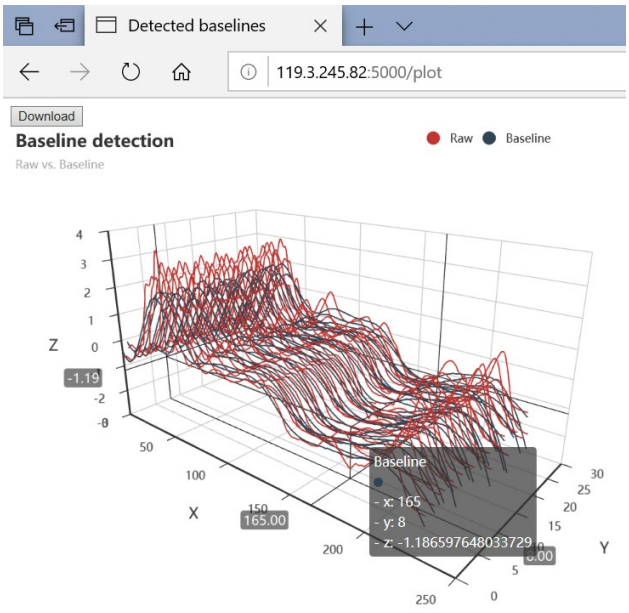
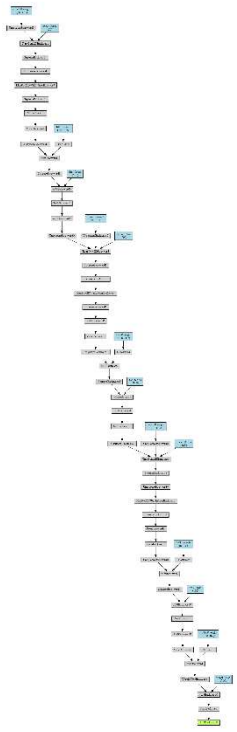


Figure 2. The baseline detection results presented on the webpage of 119.3.245.82:5000. The red curves are the input and the gray curves below the input are the corresponding baselines. Users can rotate, zoom, and pan freely to get a convenient viewpoint.

Since the input is only a sequence of intensity, and no additional information is needed, the specific meaning cannot be parsed. Moreover, the website is only used for the presentation of the new method, as well as offering convenience to peers, therefore it does not store any data. Thus, there is no risk of data leakage. If it is necessary for the community, in the future, the author is willing to develop programming interfaces invoked through the Internet.

The design and training mode of the system is described as follows. The setup of the artificial neural network is shown in Table 1. A total of 3 residual blocks with the same structure are used to form a 15-hidden-layer network (counting the last fully connected and regression layers, a total of 17 layers). There are 5 layers for each residual block: first a convolutional layer (20-channel output, convolutional kernel size 3); then a maximum pooling layer (width 5); a ReLU layer for activation; followed by a fully connected layer; after that, it is activated by a Tanh layer. The output of these 5 layers plus the input forms a residual block. When training the entire network, the input is the simulated spectrum and the target output is its baseline portion. For baseline detection using the trained network, the model will directly output the predicted baseline when inputting an authentic spectrum.

Considering the complexity of space and time, the input of the entire residual convolutional neural network is set to a 100-dimensional vector, so is the output. In other words, the trained model can process an intensity curve with a length of 100 each time. For the generating of training sample, considering the later stage of stitching, the simulating profile with a length of 200 is generated. Then a section with a length of 100 is intercepted from it as a training sample. This is a good simulation of the situation when a segment is taken from the complete original curve piece by piece. In order to facilitate users who only process a few pieces of data, the entire training process does not use batch processing strategies, which also means that there is no batch normalization step. Thus, each piece of training data is normalized separately: the intensity value of the vertical axis is linear transformed to between 0 and 1.



```
----------------------------------
Layer (type) Output Shape Param #
==================================
    Conv1d-1 [-1, 20, 98]      80
 MaxPool1d-2 [-1, 20, 19]       0
      ReLU-3 [-1, 20, 19]       0
    Linear-4         [-1]  38,100
      Tanh-5         [-1]       0
    Conv1d-6 [-1, 20, 98]      80
 MaxPool1d-7 [-1, 20, 19]       0
      ReLU-8 [-1, 20, 19]       0
    Linear-9         [-1]  38,100
     Tanh-10         [-1]       0
   Conv1d-11 [-1, 20, 98]      80
MaxPool1d-12 [-1, 20, 19]       0
     ReLU-13 [-1, 20, 19]       0
   Linear-14         [-1]  38,100
     Tanh-15         [-1]       0
   Linear-16 [-1, 1, 100]  10,100
     Tanh-17 [-1, 1, 100]       0
==================================
Total params: 124,480
Trainable params: 124,480
Non-trainable params: 0
----------------------------------
Forward/backward pass size (MB):
0.03
Params size (MB):
0.51
Estimated Total Size (MB):
0.53
----------------------------------
```

Table 1. The structure of the deep convolutional neural network used for baseline correction task. The table on the right is the specific sequence and parameter configuration of each layer in the network. The parameter capacity of the model exceeds 120 thousand.

After the model training is completed, the baseline of a spectrum can be identified segment by segment. In order to ensure smooth and continuous connection between sections, a length of 65 overlap between segments is retained. The values in the overlap interval between the two adjacent segments are merged according to the following formula (3):

$$M[i] = L[t+i] * \frac{s-i}{s-1} + R[i] * \frac{i}{s-1} \qquad (3)$$

where M represents the merged value, L represents the left one of the two adjacent segments, and R represents the right one. t is the index of the starting position of the overlapping part in L, which is 35 in the current experimental system. s is the length of the overlapping part, which is currently 65 in the system. i is the position index when calculating the overlapping part, and its value ranges from 0 to $s-1$. This formula ensures that on the leftmost side of the overlapped part, its value is $L[t]$, which is consistent with the predicted value of the left section, so that it is continuous on the left. The same is true on the right. The middle part gradually transits from the predicted value of the left section to the predicted value of the right section, thereby ensuring the overall continuity. At the same time, any value in the overlapping interval is a weighted sum of two baseline recognition results at the same position. This achieves an effective synthesis of the two identifications, and makes the final result more stable and accurate.

RESULTS

Experiments were carried out on both real spectra and synthetic profiles. The AlpahBC was compared with three widely used methods. The first is the highly cited method airPLS. It is a typical iterative algorithm with excellent performance. The other two are the wavelet method and the baseline correction function in MestReNova. The former represents another mainstream category: baseline recognition methods based on decomposition. The latter is a popular commercial software, and is regarded as one of the authoritative tools for data processing in analytical chemistry.

TESTING ON AUTHENTIC DATA

First, I used real data to visually verify the effectiveness of the new method. Baseline removal tests were implemented using XRD [15], Raman spectroscopy [15], MALDI-TOF and infrared spectroscopy. Figures 3 to 6 present the corresponding results in graphics, showing the performance of each method intuitively. The airPLS method relies on several manually selected parameters to control the quality of the results, including lambda, order, weight exception proportion, asymmetry parameter and maximum iteration times. Each parameter can significantly affect the performance. Similarly, using the classic multilevel 1-dimensional wavelet decomposition method, users need to set the type of wave and the number of decomposition levels; to use MestRenova, users need to select the type of built-in method and configure the corresponding parameters. In order to better explain the characteristics of each method, the author manually finetuned respective parameter configurations and selected the best results under comprehensive conditions for comparison. In contrast, the new method is completely automated without any human

intervention after a training session. The final comparison results are still favorable to the new method. From Figure 3 to Figure 6, such a reasonable conclusion can be drawn that the AlphaBC algorithm is more accurate and stable.

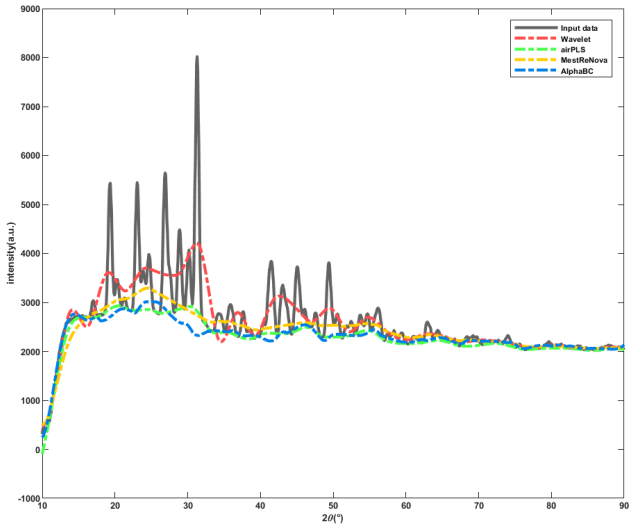


Figure 3. Results on X-ray Diffraction data (gray). The baseline detected by the proposed AlphaBC method (blue) captures the uplift at the beginning precisely and tracks the following basic trend stably. In contrast, the wavelet method (red) and the airPLS method (green) produce overcut baselines if the parameters were set to fit the uplift. MestReNova (yellow) generates less accurate result on both intervals.

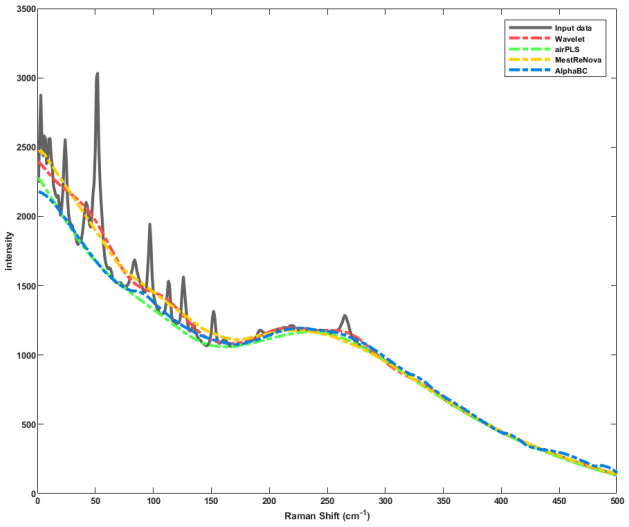


Figure 4. Results on Raman data (gray). The AlphaBC method (blue) approximates the background decline robustly from the beginning while the comparing methods (red, green, yellow) overcut the signals before being close to the visual background.

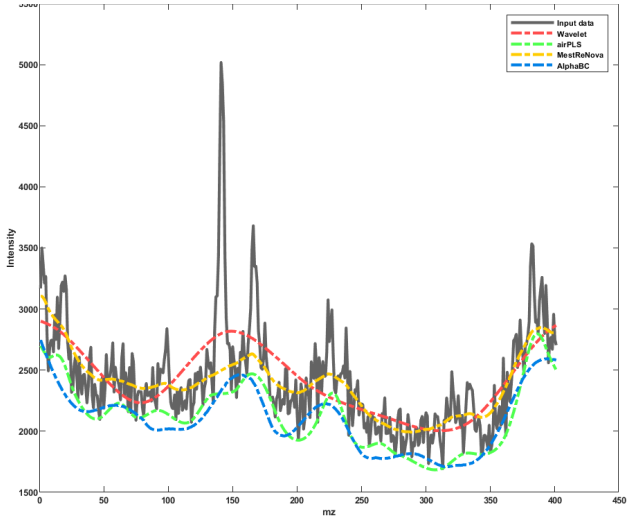


Figure 5. Results on MALDI-TOF data (gray). The proposed AlphaBC algorithm (blue) performs close to airPLS (green) and MestRenova (yellow). The three methods slightly outperform the wavelet method (red).

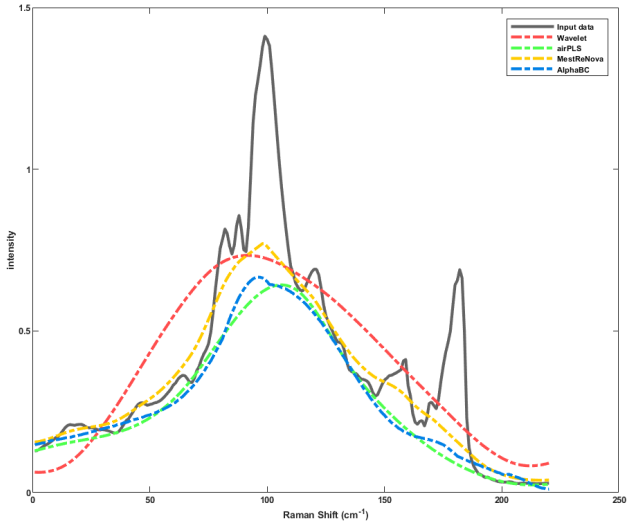


Figure 6. Results on infrared red data (gray). The AlphaBC algorithm (blue) works well.

QUANTITATIVE MEASURING

In order to quantitatively evaluate the performance of baseline correction algorithms, a set of artificial spectra with known baselines was created to measure the accuracy of recognition. For fairness, instead of using the Fourier series method for constructing the

training set of AlphaBC, the synthetic data for testing were constructed from real spectra in the following way. First, manually cut a spectrum with a sufficiently high threshold to ensure that the part above the threshold (denoted as $\sigma(x)$) only has signal peaks. Denote the part below the threshold as $\tau(x)$, and then use the following formula to generate a function $b(x)$ to simulate the baseline:

$$b(x) = \frac{1}{2w}\int_{-w}^{w} \tau(x+t)\,dt$$

The width for the integration w was chosen equal to the total width of $\sigma(x)$. Adding $b(x)$ to $\sigma(x)$ to construct a new trace $y(x) = \sigma(x) + b(x)$, then the baseline was accurately known and the accuracy of detection could be calculated immediately. The whole procedure is shown as a schematic in Figure 7. The average error under L^1 norm is used to measure the accuracy of the baseline detection. The experiments were carried out on a series of simulating spectra. For each method used for comparison, the required parameters had been carefully selected and fixed.

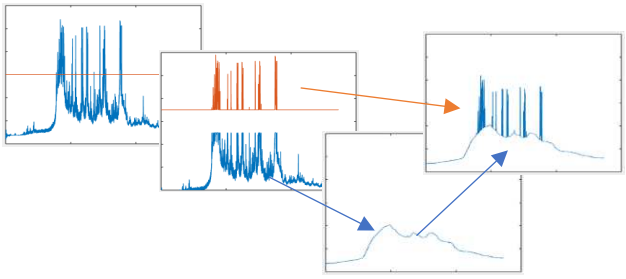


Figure 7. From left to right: a threshold is manually selected to extract a part of pure signals (the red part); the lower part is smoothed by averaging in a wide range to obtain a less fluctuating curve to simulate the baseline; add the two parts to form a synthesized spectrum with a precisely known baseline.

The test results are shown in Figure 8 and Table 2. As can be seen from Figure 8, the AlphaBC method is better than the wavelet method and the automatic baseline correction in MestReNova. Compared with the airPLS algorithm with preset parameters, except in few cases where the performance is similar, the AlphaBC clearly outperforms in other tests. More experiments directly counted

the relative error under the L^1 norm, the numerical results are listed in Table 2, and are graphically presented in Figure 9. These test results show that the AlphaBC performs best among these methods.

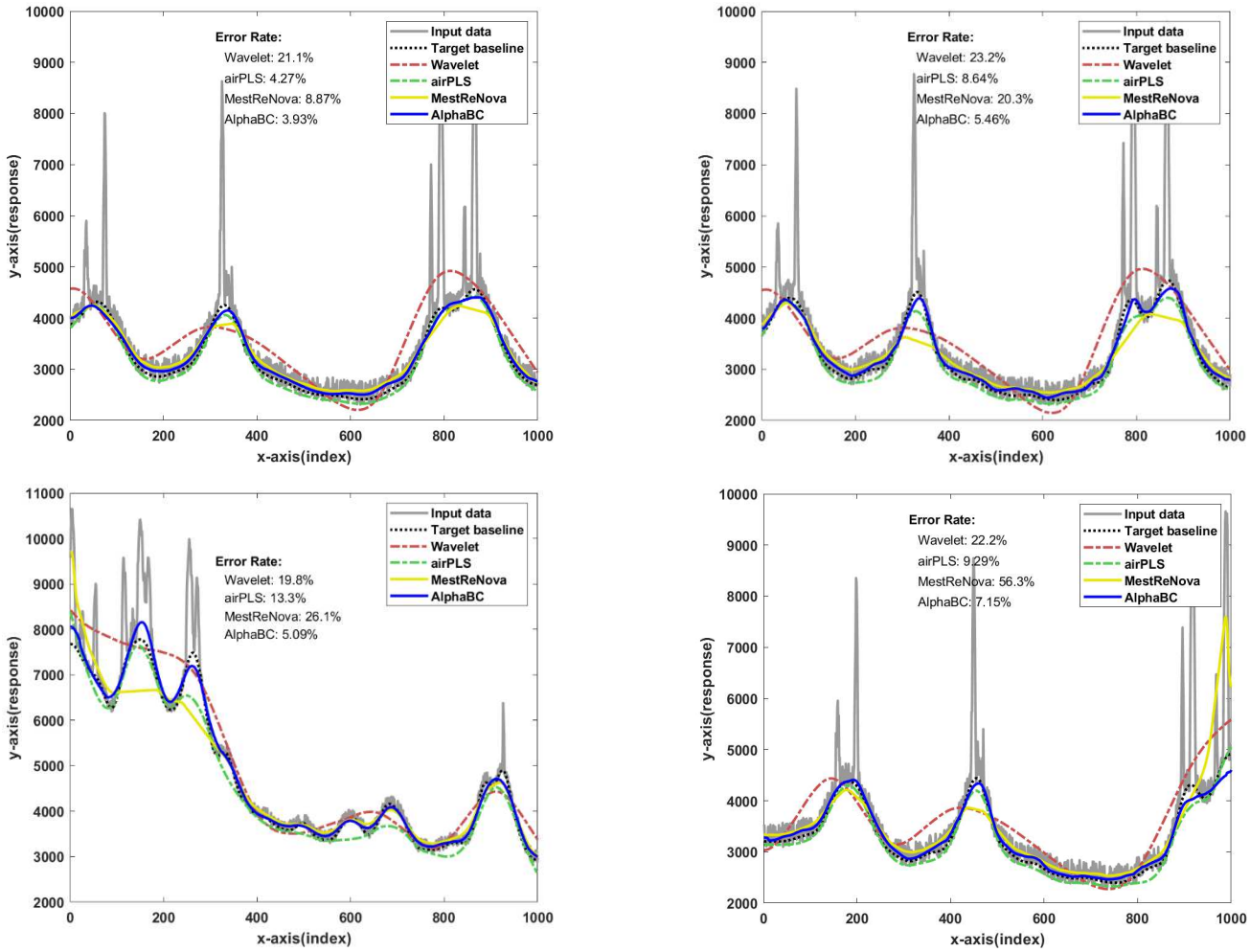


Figure 8. Numerical and graphical results on synthesized testing data (gray). The black dotted lines are the constructed known baselines. Corresponding errors are measured under L^1 norm of the discrepancies between the result produced by each method and the known target baseline. AlphaBC algorithm (blue) appears best in both visual effect and numerical measurement.

No.	wavelet	airPLS	MestReNova	AlphaBC
1	0.20990	0.12820	0.18518	0.05650
2	0.10125	0.06946	0.09623	0.03468
3	0.05938	0.01854	0.03326	0.01255
4	0.11454	0.02908	0.07281	0.02528
5	0.11619	0.06205	0.10290	0.03511
6	0.13788	0.08515	0.11122	0.03935
7	0.14983	0.10213	0.13655	0.05306
8	0.08544	0.04405	0.05170	0.02800
9	0.09468	0.04314	0.07151	0.03230
10	0.09471	0.02965	0.05418	0.02351
11	0.15881	0.11431	0.15682	0.04937
12	0.08787	0.02396	0.05967	0.02981
13	0.09253	0.03989	0.07480	0.02606
14	0.13248	0.10718	0.12889	0.03489
15	0.12510	0.12785	0.13786	0.04019
16	0.12882	0.09398	0.11960	0.03225
17	0.09813	0.05533	0.08240	0.02217
18	0.05772	0.01086	0.03840	0.00562

Table 2. The error rates of the four comparing algorithms.

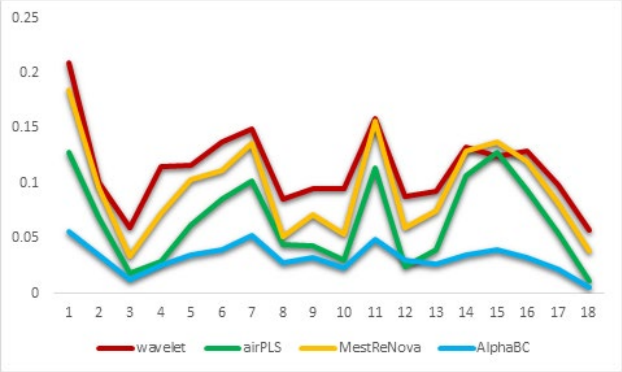


Figure 9. The graphical presentation of the error rates displayed in Table 2. It indicates that AlphaBC algorithm (blue) keeps lower error and is robust.

SUMMARY

Visually, the proposed AlphaBC method generated baselines closer to the trend of spectra's backgrounds than other compared algorithms when applied on the authentic data. It was further demonstrated by numeric testing on synthesized data with known

baselines. The AlphaBC method was able to keep the lowest relative error under L^1 norm among all tested algorithms. The experiments indicate that the new method is more accurate and effective. As a typical deep learning algorithm, although the training procedure of the AlphaBC takes a long time, once the training is completed, the baseline correction can be implemented very fast each time.

CONCLUSIONS

Baseline removal plays an important role in the digital signal processing of analytical data. It has a great influence on the accuracy of downstream processing such as peak detection and period identification.

The traditional methods for baseline removal have advantages in different aspects. The airPLS method is insensitive to noise; the wavelet method generates smooth baseline; the methods in MestReNova are relatively rich, providing users with a variety of options. The common disadvantage of these methods is that users need to carefully choose configuration parameters case by case to achieve effective results.

In contrast, the new scheme, the AlphaBC method, has no parameters and requires no human involvement throughout the entire process, and can be universally applied after one training session. The advantage of the new method is that it is not only parameter-free, but also more accurate.

In order to obtain good performance, the learning model needs to have a larger size. The larger the network scale, the more powerful the functions, and the more space required for training and storage. Space complexity is the main disadvantage of the AlphaBC method. However, following current popular solutions, this burden of space can be removed by providing online programming interfaces on the cloud.

DISCUSSION

Regarding the widths and positions of signal peaks in the simulated spectrum, the shape and the relative position of the baseline, etc., no matter how many samples are generated, it is impossible to cover all cases. As with all methods, baseline correction is always approximate. However, under the framework of the AlphaBC method, with the form of web applications, it is possible to further improve the accuracy of the intelligent model in combination with the help of the community.

It is planned to further add options of users' comment on whether the results are satisfactory in the currently published website. According to the completely voluntary principle, if the user is interested, he can also assign which part of the baseline is invalid and keep it separately. Using these precious manual labels, the system can be continuously trained to achieve self-update and evolution, and improve its accuracy every day.

PERSPECTIVES

The way of using crowdsourcing to help evolve intelligent model envisaged in the discussion part has been adopted and developed in the field of artificial intelligence for more than a decade. Although in analytical chemistry data processing, baseline identification is only a specific step, it may be a beginning of wider applications. The method and system in this article can be considered as an attempt to provide some reference for grand data-driven research projects such as the Materials Genome Initiative.

ABBREVIATIONS

AlphaBC: Alpha Baseline Correction;

airPLS: adaptive iteratively reweighted Penalized Least Squares;

XRD: X-ray Diffraction;

RBF: Radial Basis Function;

MALDI-TOF: Matrix-Assisted Laser Desorption/Ionization-Time Of Flight.

AUTHOR INFORMATION

Corresponding Author

* Corresponding author: Yuanjie Liu, E-mail: yjliu@cau.edu.cn.

DECLARATIONS

Acknowledgements

The author would like to thank Ms. Chong Wang, the High School Affiliated to Beijing Normal University, XuanWu, Beijing, China for her help with the language editing.

Authors' contributions

YL executed the design and the implementation of the overall procedure, its evaluation and drafted the manuscript. The author read and approved the final version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant no. 61807032) and the Fundamental Research Funds for the Central Universities of China (Grant no. 2019TC045).

These funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets used to generate the figures presenting results on Raman spectra, XRD profiles were downloaded from public databases which were cited within the manuscript. The data used to generate the figures presenting results on the energy curve of the audio, mass spectroscopy and infrared spectroscopy are available in the Additional file 1. The synthesized data for random testing were generated following the procedure explained within the manuscript.

Supporting Information

The trained model has been deployed on a cloud server. Anyone can visit the address of 119.3.245.82:5000 freely to test and use.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

[1] Y. Cai, C. Yang, D. Xu, and W. Gui, "Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation," Analytical Methods, 10.1039/C8AY00914G vol. 10, no. 28, pp. 3525-3533, 2018.

[2] F. Qian, Y. Wu, and P. Hao, "A fully automated algorithm of baseline correction based on wavelet feature points and segment interpolation," Optics & Laser Technology, vol. 96, pp. 202-207, 2017.

[3] X. Shen et al., "Study on baseline correction methods for the Fourier transform infrared spectra with different signal-to-noise ratios," Appl Opt, vol. 57, no. 20, pp. 5794-5799, Jul 10 2018.

[4] Z. F. Xu, X. B. Sun, and P. D. Harrington, "Baseline Correction Method Using an Orthogonal Basis for Gas Chromatography/Mass Spectrometry Data," (in English), Analytical Chemistry, Article vol. 83, no. 19, pp. 7464-7471, Oct 2011.

[5] Y. Liu, X. Zhou, and Y. Yu, "A concise iterative method using the Bezier technique for baseline construction," Analyst, vol. 140, no. 23, pp. 7984-96, Dec 7 2015.

[6] Y. Liu and J. Lin, "A general-purpose signal processing algorithm for biological profiles using only first-order derivative information," BMC Bioinformatics, vol. 20, no. 1, p. 611, Nov 27 2019.

[7] Y. Liu and Y. Yu, "A survey of the baseline correction algorithms for real-time spectroscopy processing," in Photonics Asia, Beijing, China, 2016, vol. 10026, p. 100260Q: SPIE.

[8] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, and Ieee, "Deep Residual Learning for Image Recognition," in 2016 Ieee Conference on Computer Vision and Pattern Recognition(IEEE Conference on Computer Vision and Pattern Recognition, New York: Ieee, 2016, pp. 770-778.

[9] D. Silver et al., "Mastering the game of Go without human knowledge," Nature, vol. 550, no. 7676, pp. 354-359, Oct 18 2017.

[10] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, "Deep convolutional neural networks for Raman spectrum recognition: a unified solution," Analyst, vol. 142, no. 21, pp. 4067-4074, Oct 23 2017.

[11] A. Kanginejad and A. Mani-Varnosfaderani, "Chemometrics advances on the challenges of the gas chromatography–mass spectrometry metabolomics data: a review," Journal of the Iranian Chemical Society, vol. 15, no. 12, pp. 2733-2745, 2018.

[12] A. E. Cetin and M. Tofighi, "Projection-Based Wavelet Denoising [Lecture Notes]," IEEE Signal Processing Magazine, vol. 32, no. 5, pp. 120-124, 2015.

[13] K. Maloney, "Hands-on NMR experience without the NMR: Using MestreNova to teach an undergraduate organic structure elucidation course without an on-site high field NMR," (in English), Abstracts of Papers of the American Chemical Society, Meeting Abstract vol. 253, p. 1, Apr 2017.

[14] S. Gai and Z. Y. Bao, "New image denoising algorithm via improved deep convolutional neural network with perceptive loss," (in English), Expert Systems with Applications, Article vol. 138, p. 9, Dec 2019, Art. no. Unsp 112815.

[15] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, "The power of databases: The RRUFF project," in Highlights in Mineralogical Crystallography, T. Armbruster and R. M. Danisi, Eds. Berlin, Germany: W. De Gruyter, 2015, pp. 1–30.

Figures

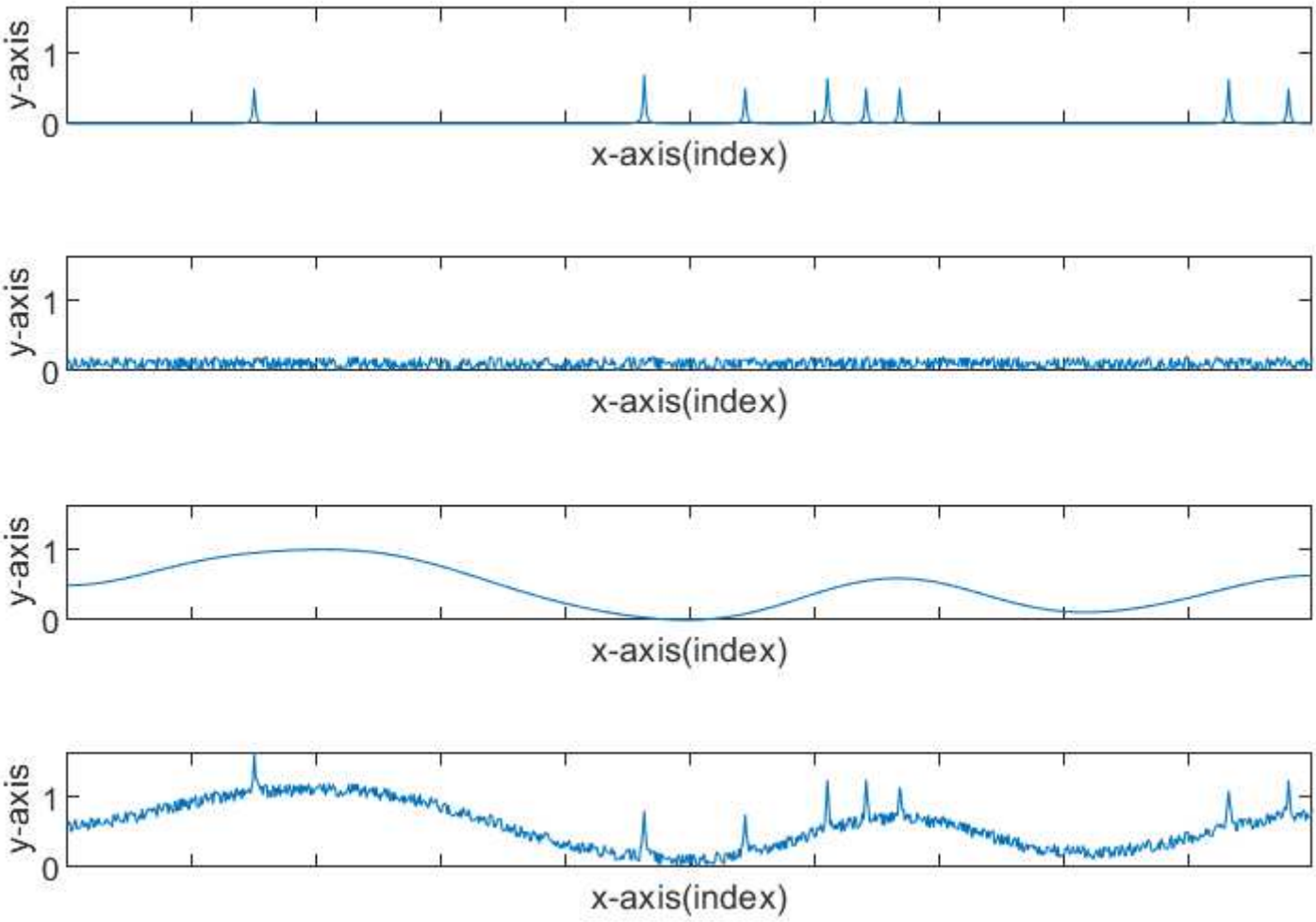


Figure 1

he procedure of generating a training sample. From top to bottom: the first subfigure presents randomly set signals generated according to Formula (2); the second one presents noises sampled from uniform distribution; the third one presents a baseline generated according to Formula (1); the last one presents the simulated training sample generated by adding the three portions together.

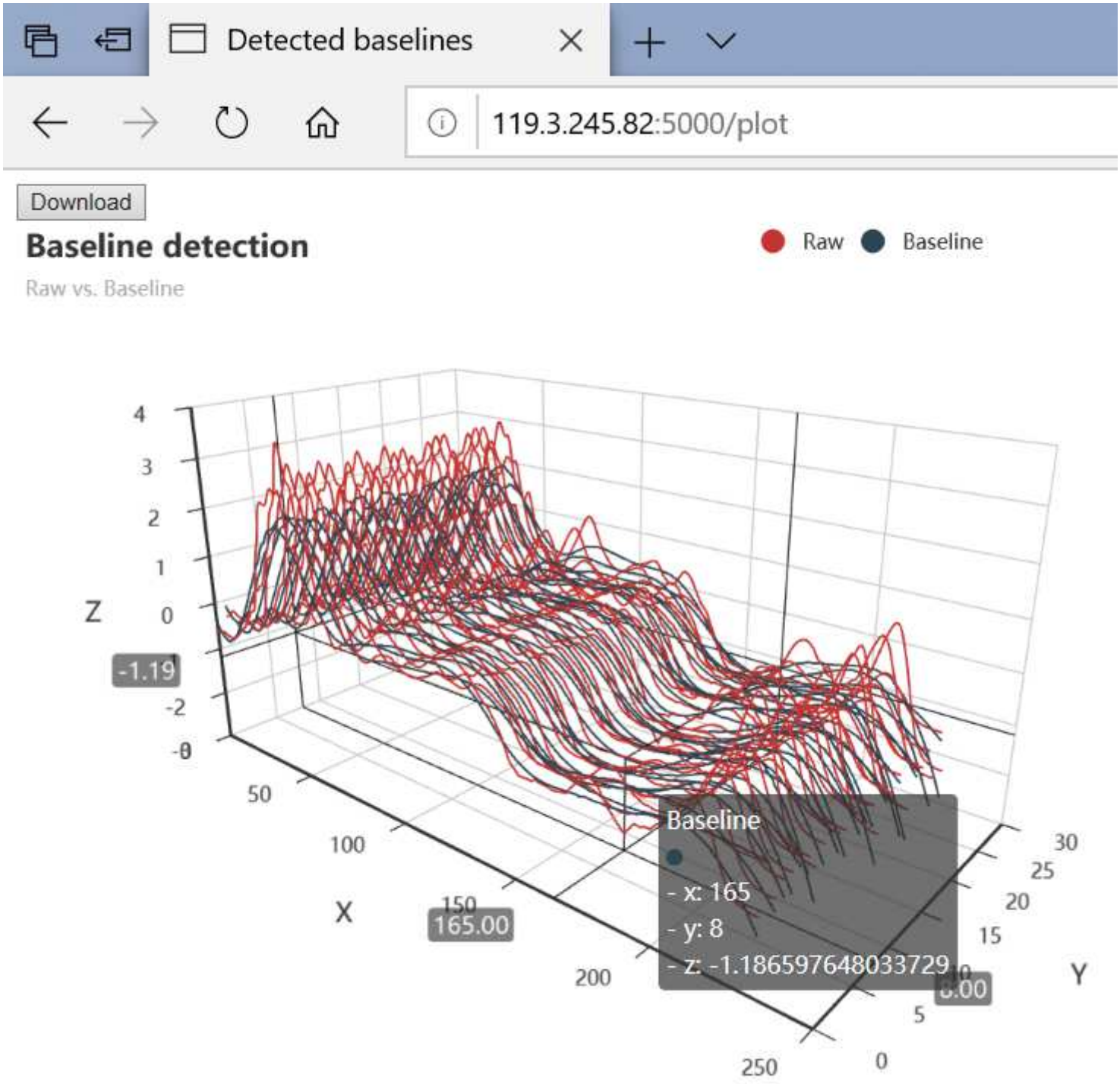


Figure 2

The baseline detection results presented on the webpage of 119.3.245.82:5000. The red curves are the input and the gray curves below the input are the corresponding baselines. Users can rotate, zoom, and pan freely to get a convenient viewpoint.

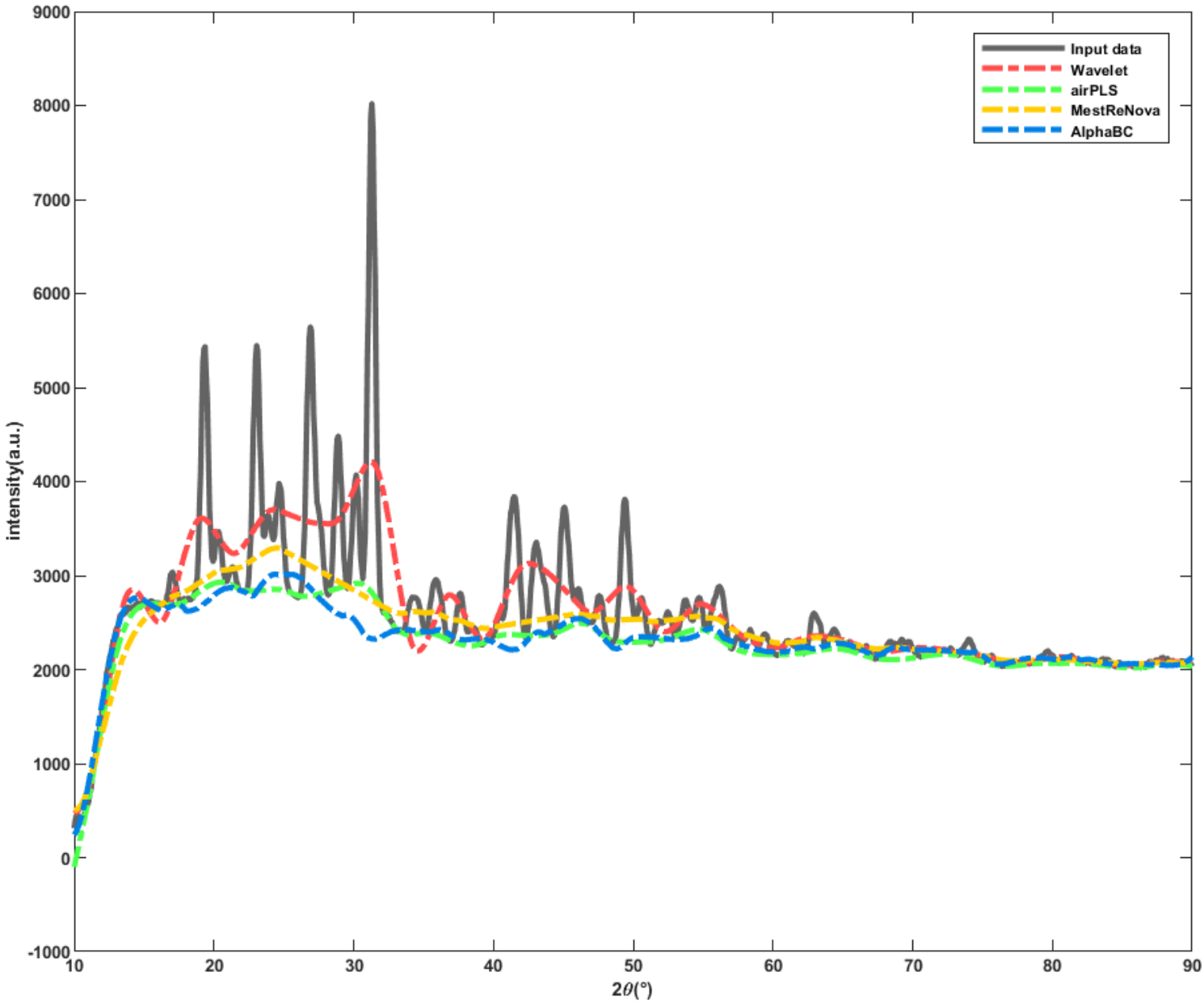


Figure 3

Results on X-ray Diffraction data (gray). The baseline detected by the proposed AlphaBC method (blue) captures the uplift at the beginning precisely and tracks the following basic trend stably. In contrast, the wavelet method (red) and the airPLS method (green) produce overcut baselines if the parameters were set to fit the uplift. MestReNova (yellow) generates less accurate result on both intervals.

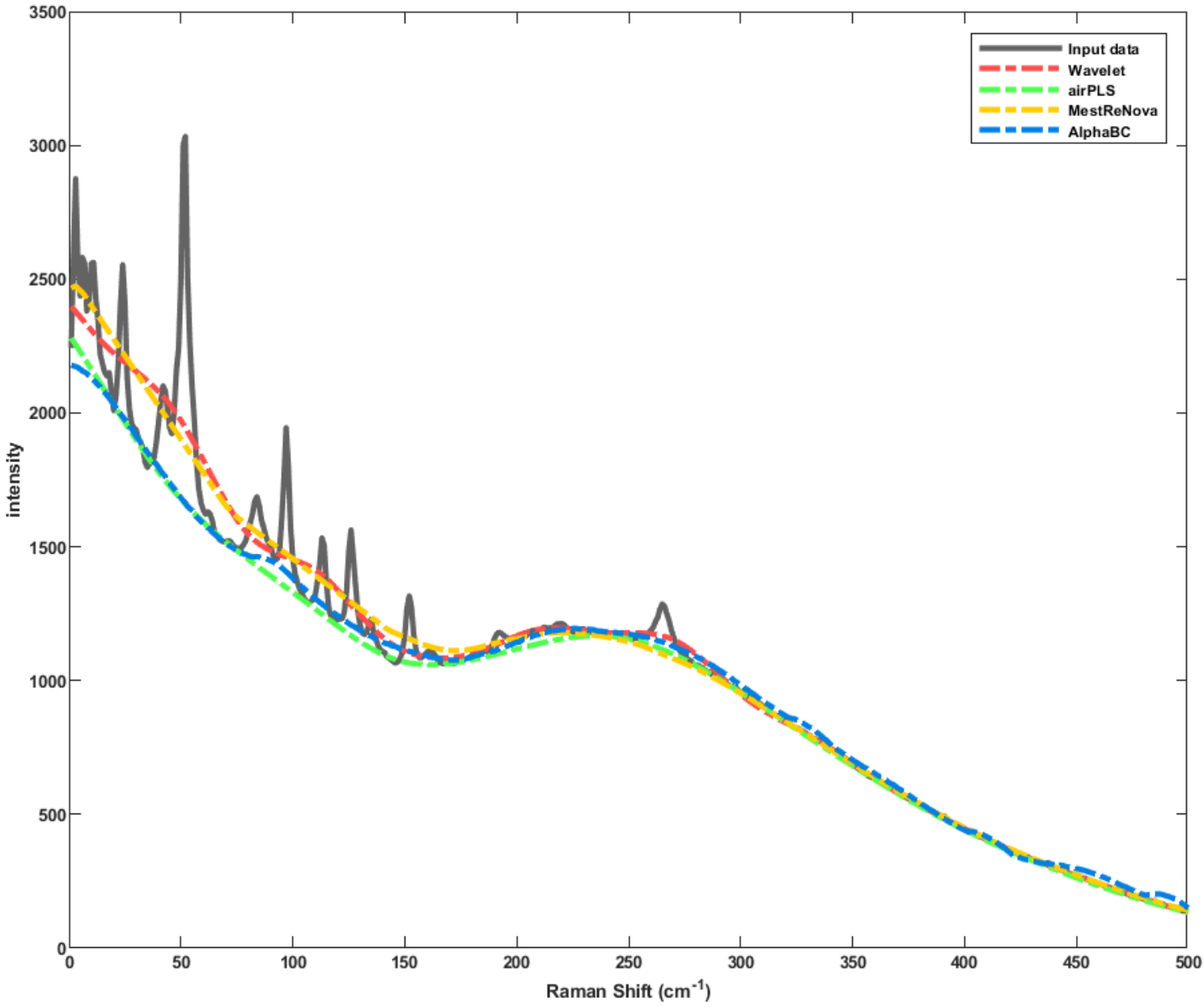


Figure 4

Results on Raman data (gray). The AlphaBC method (blue) approximates the background decline robustly from the beginning while the comparing methods (red, green, yellow) overcut the signals before being close to the visual background.

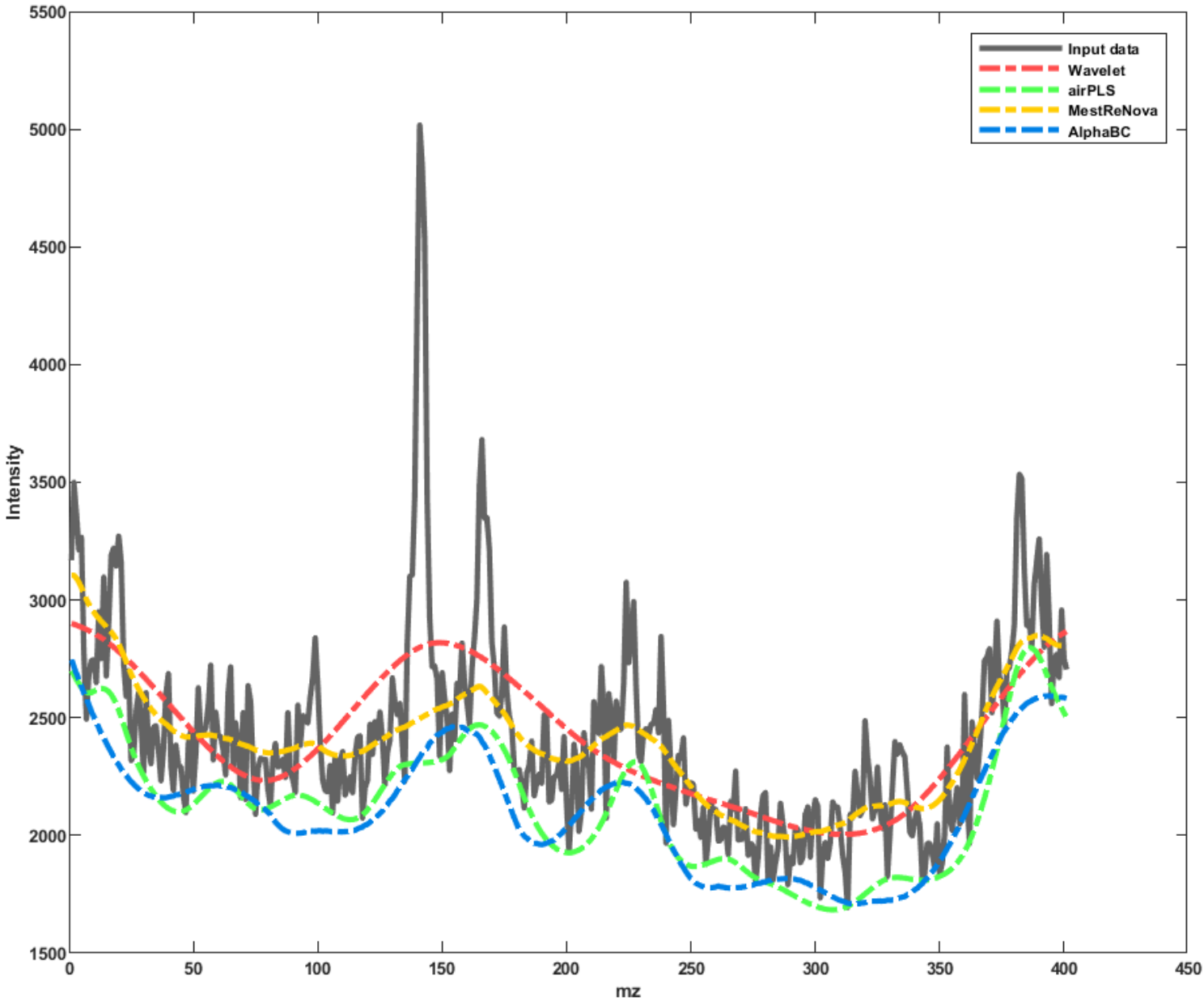


Figure 5

Results on MALDI-TOF data (gray). The proposed AlphaBC algorithm (blue) performs close to airPLS (green) and MestRenova (yellow). The three methods slightly outperform the wavelet method (red).

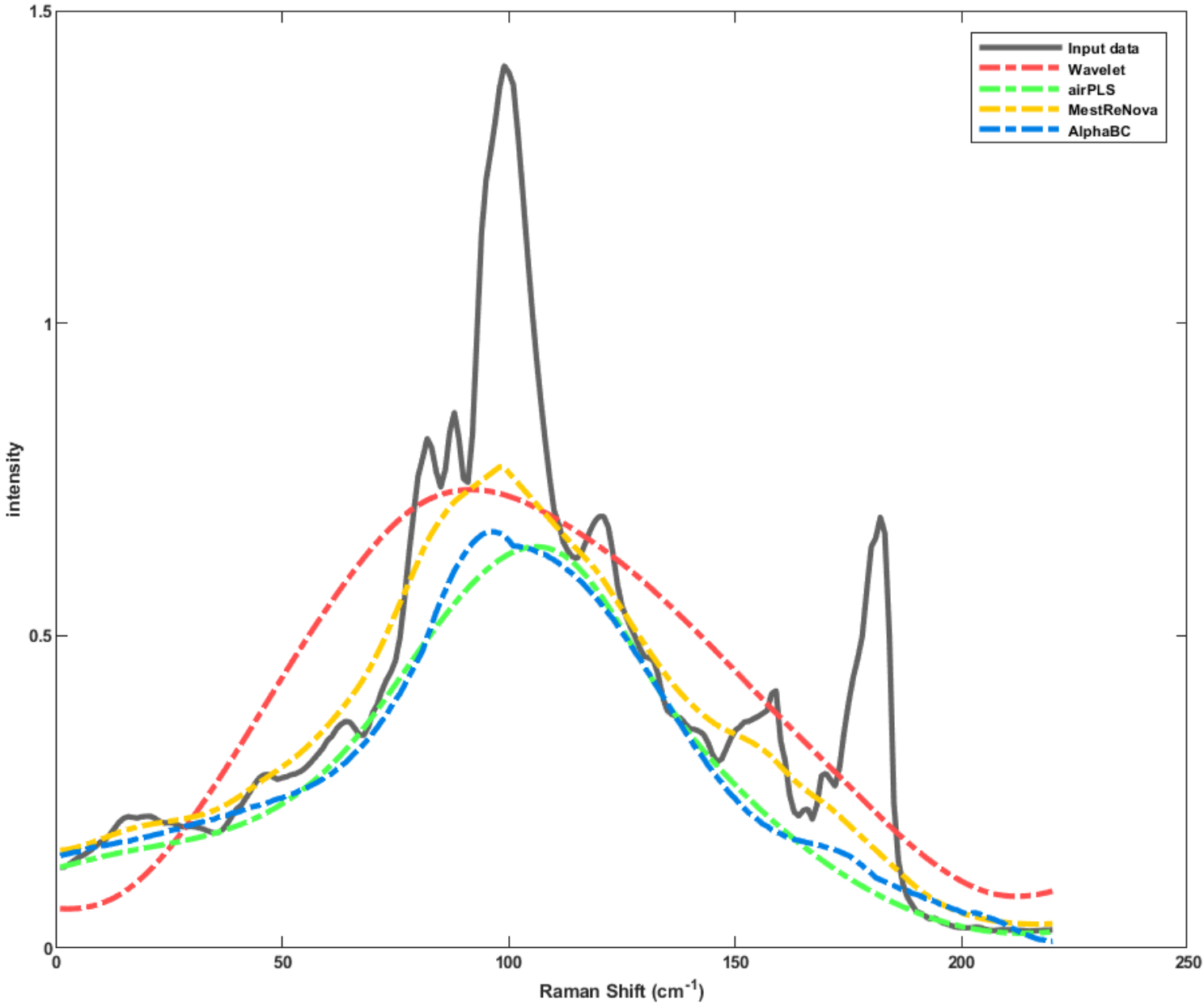


Figure 6

Results on infrared red data (gray). The AlphaBC algorithm (blue) works well.

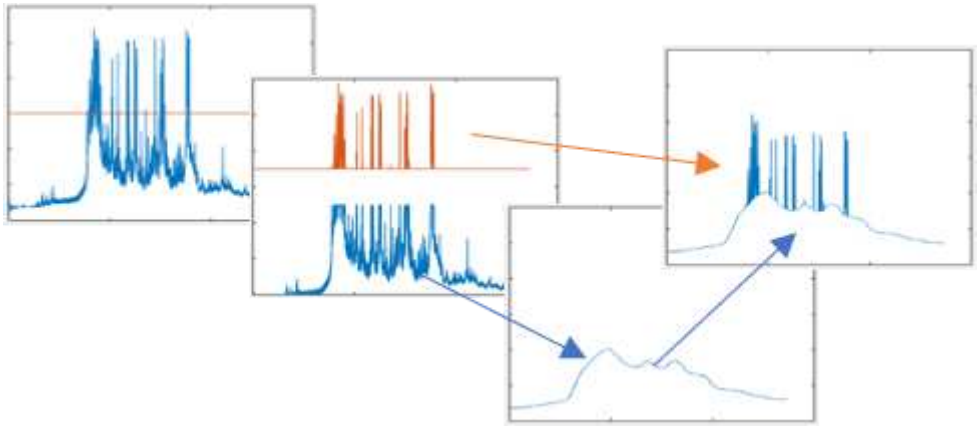


Figure 7

From left to right: a threshold is manually selected to extract a part of pure signals (the red part); the lower part is smoothed by averaging in a wide range to obtain a less fluctuating curve to simulate the baseline; add the two parts to form a synthesized spectrum with a precisely known baseline.

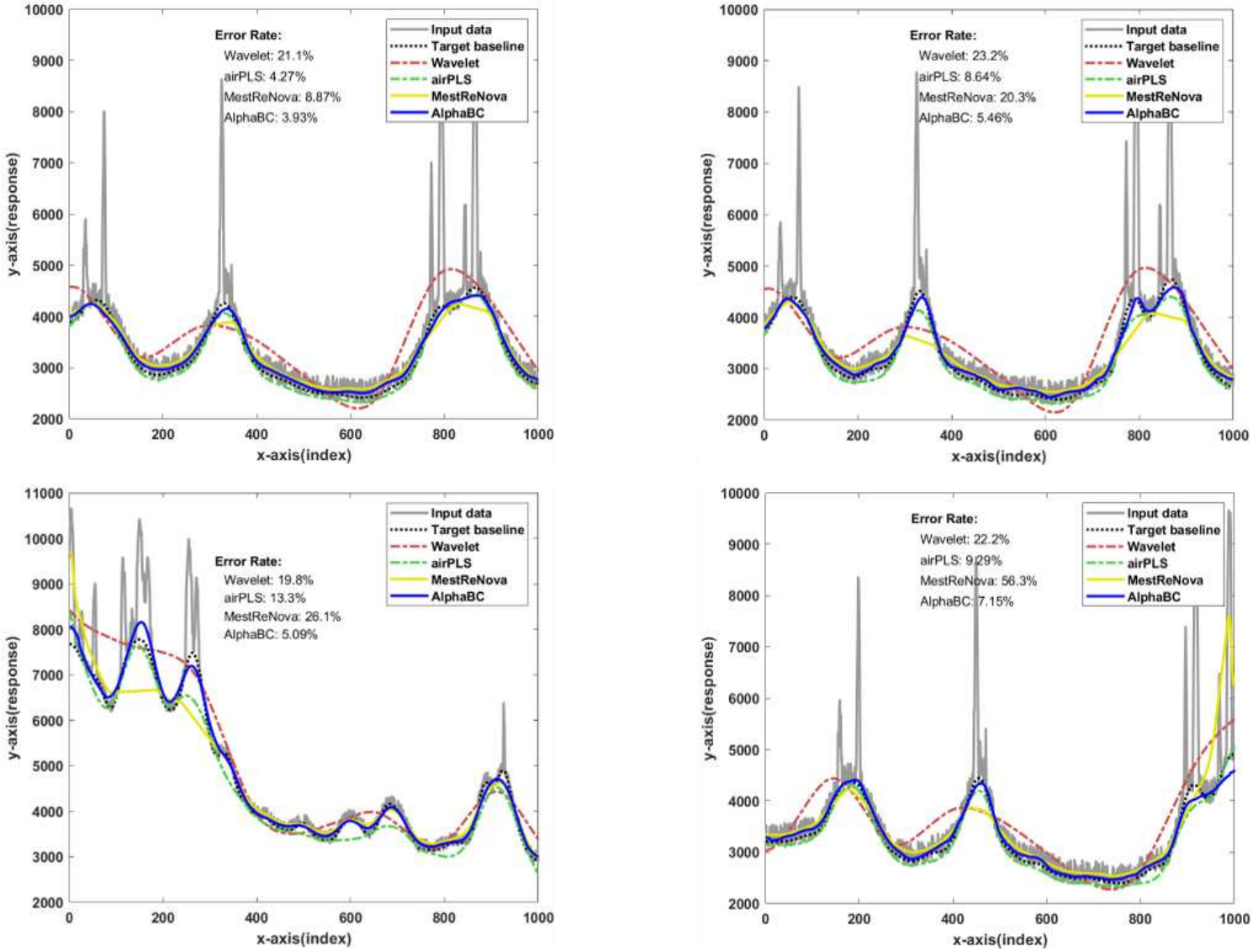


Figure 8

Numerical and graphical results on synthesized testing data (gray). The black dotted lines are the constructed known baselines. Corresponding errors are measured under L^1 norm of the discrepancies between the result produced by each method and the known target baseline. AlphaBC algorithm (blue) appears best in both visual effect and numerical measurement.

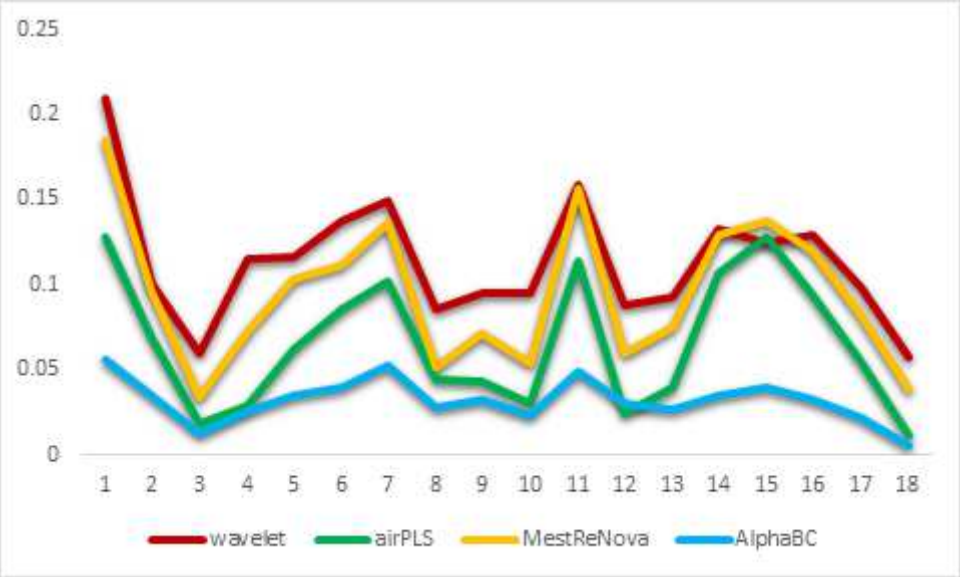


Figure 9

The graphical presentation of the error rates displayed in Table 2. It indicates that AlphaBC algorithm (blue) keeps lower error and is robust.

Supplementary Files

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

- supplement15.csv