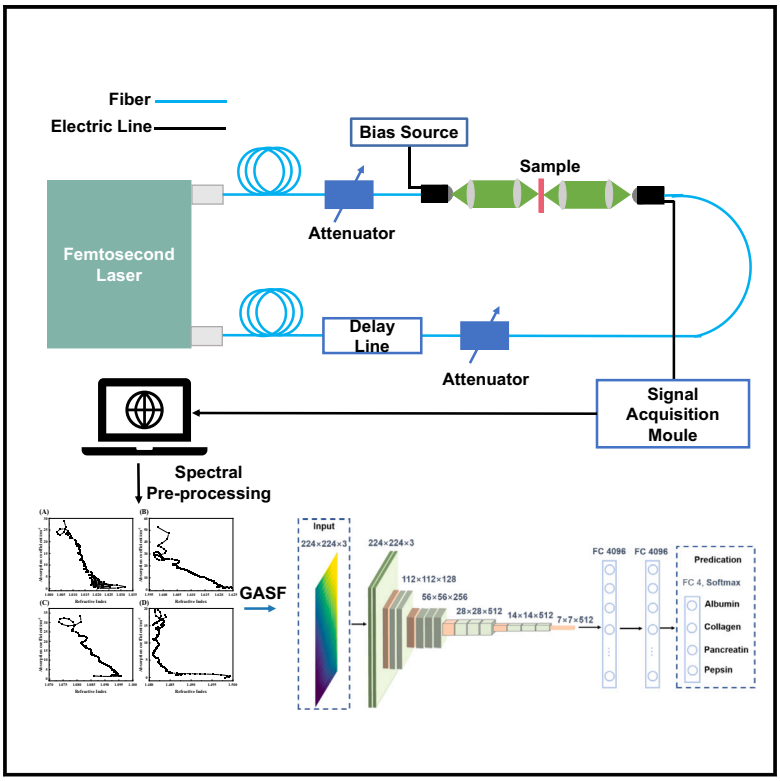
iScience

Article

Distinguishing different proteins based on terahertz spectra by visual geometry group 16 neural network

Graphical abstract



Highlights

- 2D-image data were obtained from 1D-spectra data using GASF
- VGG-16 model based on transfer learning was developed to distinguish proteins
- SVM, GPC, BiGRU, and CNN-BiGRU models were also used to distinguish proteins
- Good performances were obtained by VGG-16 model

Authors

Yusa Chen, Xiwen Huang, Meizhang Wu, ..., Wengang Wu, Guozhong Zhao, Tianhua Meng

Correspondence

wuwg@pku.edu.cn

In brief

Applied computing in medical science; Artificial intelligence applications; Artificial intelligence programming language

Chen et al., 2025, iScience *28*, 112148
April 18, 2025
https://doi.org/10.1016/j.isci.2025.112148

CellPress

iScience

CellPress
OPEN ACCESS

Article

Distinguishing different proteins based on terahertz spectra by visual geometry group 16 neural network

Yusa Chen,[1,2] Xiwen Huang,[3] Meizhang Wu,[4,5] Jixuan Hao,[3] Yunhao Cao,[1,2] Hongshun Sun,[1,2] Lijun Ma,[1,2] Liye Li,[1,2] Wengang Wu,[1,2,7,*] Guozhong Zhao,[3] and Tianhua Meng[6]

[1]National Key Laboratory of Advanced Micro and Nano Manufacture Technology, Beijing 100871, P.R. China
[2]School of Integrated Circuits, Peking University, Beijing 100871, P.R. China
[3]Department of Physics, Capital Normal University, Beijing 100048, China
[4]School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing 100096, China
[5]School of Automation, University of Science and Technology Beijing, Beijing 100083, P.R. China
[6]Institute of Solid State Physics, Shanxi Provincial Key Laboratory of Microstructure Electromagnetic Functional Materials, Shanxi Datong University, Datong 037009, China
[7]Lead contact
*Correspondence: wuwg@pku.edu.cn
https://doi.org/10.1016/j.isci.2025.112148

SUMMARY

Detecting different kinds of proteins is of great significance for medical diagnosis, biological research, and other fields. We combine both terahertz (THz) absorption and refractive index spectra with the visual geometry group 16 (VGG-16) neural network to intelligently identify four proteins, namely albumin, collagen, pepsin, and pancreatin in this study. The THz absorption-refractive index spectra of the proteins were converted to two-dimensional image features by the Grassia angular summation field (GASF) method and used as a dataset, which enabled the VGG-16 model to achieve 98.8% accuracy in distinguishing the four proteins. We also compared the VGG-16 model with other machine learning models, which demonstrate that it has better performance. Overall, the VGG-16 neural network transfer learning technique proposed in this study can quickly and accurately achieve the identification of different kinds of proteins. This research might have potentially important applications in biotechnology fields, such as biosensors, biopharmaceuticals, and medicine.

INTRODUCTION

The identification and analysis of proteins are of critical importance in various fields, including biochemistry, medicine, and material science, due to the essential roles that proteins play in biological systems. Albumin is synthesized primarily by the liver and has a variety of functions, including maintenance of plasma colloid osmotic pressure, transport of nutrients and drugs, and regulation of plasma pH. In addition, albumin is involved in the immune response, antioxidant response, and anticoagulant process.[1,2] Collagen is a major component of connective tissue and plays a key role in the structure and function of tissues such as skin, bones, tendons, ligaments, and blood vessels. Its involvement in the formation and repair of muscle tissue helps to promote muscle growth and repair damaged muscle tissue.[3,4] Pepsin works primarily in the stomach and is responsible for breaking down proteins into smaller peptides and amino acids, thus helping the body to digest proteins.[5,6] Pancreatin is an enzyme secreted by the pancreas that is primarily used to help the body digest proteins, starches, and fats.[7,8] The identification and study of the aforementioned proteins can help to reveal the biological mechanisms and expand their applications in medicine and biomedicine.[9,10] On the other hand, the markers of certain diseases are the presence or abnormal expression of specific proteins, and the identification of proteins can help in early diagnosis.

Currently, the main methods for identifying proteins are mass spectrometry,[11] X-ray crystallography,[12] nuclear magnetic resonance spectroscopy,[13] and enzyme-linked immunosorbent assay,[14] etc. Mass spectrometry is widely used for protein identification and quantification, with high sensitivity and specificity. It can provide detailed structural information and is particularly effective for complex samples.[15] However, mass spectrometry often requires sample preparation that can be time-consuming and costly. X-ray crystallography is the gold standard for determining the 3D structure of proteins at atomic resolution, providing invaluable insights into protein function and interactions.[16] But the need for high-quality protein crystals makes this method challenging for many proteins, especially those that are difficult to crystallize or for large protein complexes. Nuclear magnetic resonance spectroscopy is usually non-destructive,[17] but in some cases, prolonged exposure to high-intensity

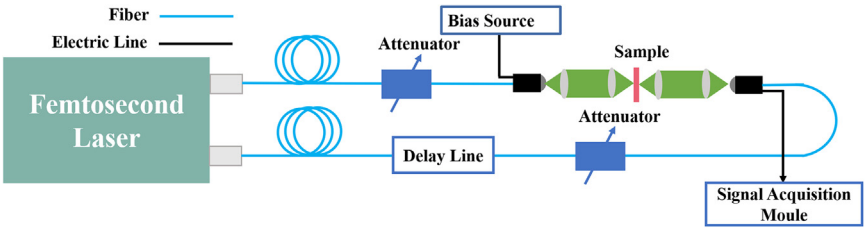


Figure 1. Schematic diagram of the THz-TDS

magnetic fields or the use of highly concentrated solvents may affect samples, especially complex biomolecules. Enzyme-linked immunosorbent assays, require the labeling of specific antigens or antibodies in the sample and involve multiple steps that may be affected by non-specific binding.[18] At the same time, because it is a chemical reaction, the sample is usually consumed, and although the sample is generally not completely destroyed, the number of tests is limited. Overall, the aforementioned methods suffer from the problems of being destructive to the sample, complexity of sample preparation, and time-consuming testing.

Terahertz (THz) waves are electromagnetic waves with frequencies between 0.1 and 10 THz.[19] Many of the vibrational leaps and rotational kinetic energy levels of biomolecules are in the THz range, which makes THz spectroscopy very specific to the spatial arrangement and structure of biomolecules.[20–22] In recent years, many researchers have carried out THz spectroscopic studies of biomolecules and reported THz absorption spectra of various biomolecules.[23–27] The key advantages of THz technology for protein detection include its non-destructive nature, rapid analysis, ability to operate without labels, and the detailed molecular information it provides.[28,29] These characteristics make it a promising tool in various fields, including biomedical research, clinical diagnostics, and biotechnology, especially for quick, high-sensitivity protein detection studies.[30–32]

With the development of artificial intelligence, deep learning technology has been widely applied to analyze THz spectroscopy data. Various neural networks, such as recurrent neural networks (RNN), convolutional neural networks (CNN),[33,34] CNN bi-directional gated recurrent network patterns (BiGRU),[35] and effective channel attention networks (ECA)-CNN have been used to classify amino acids.[36] A neural network approach has also been proposed to extract the THz band parameters of materials instead of the iterative root-finding method.[37,38] Loahavilai et al.[39] examined various compositions of ternary mixtures containing caffeine, quinic acid, and niacin using THz spectroscopy alongside a CNN model. They assessed the CNN model's effectiveness in predicting the mixture compositions. Liang et al.[40] created a quantitative model of THz spectroscopy measurement and a CNN model to accurately detect each active ingredient in a fixed-dose combinatorial anti-tuberculosis formulation. These studies show that THz spectroscopy combined with deep learning technology is widely used for intelligent detection, but little work has reported inputting both THz absorption spectra and refractive index spectra into neural network models as co-learned features. Meanwhile, the above models need to be trained on large-scale data, which consumes a lot of time and computational resources, and it is difficult to get a fast and efficient solution in practical applications.

Therefore, in the current study, a method for distinguishing different kinds of proteins using the THz absorption-refractive index spectra combined with the transfer learning technique of the visual geometry group 16 (VGG-16) deep neural network is proposed. The THz-TDS system used in this study is schematically shown in Figure 1. The THz absorption spectra and THz absorption-refractive index spectra of four proteins, namely, albumin, collagen, pepsin, and pancreatic, are converted into

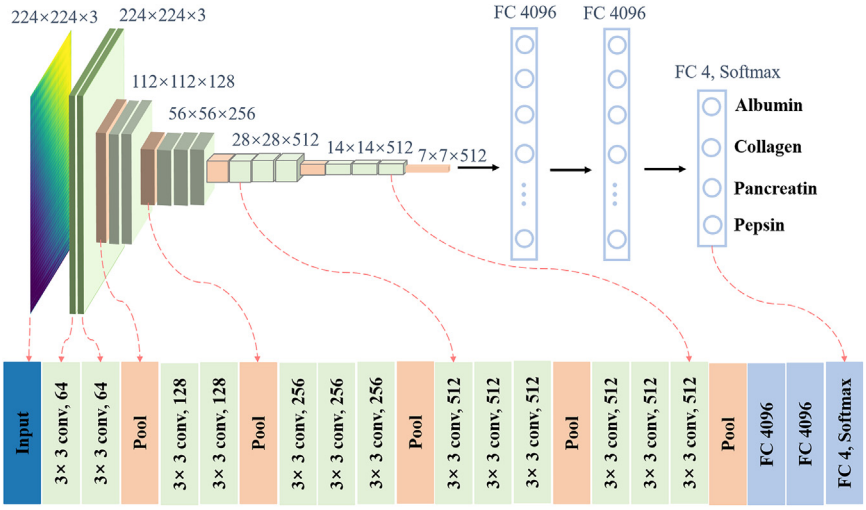


Figure 2. Diagram of the VGG-16 model architecture based on transfer learning

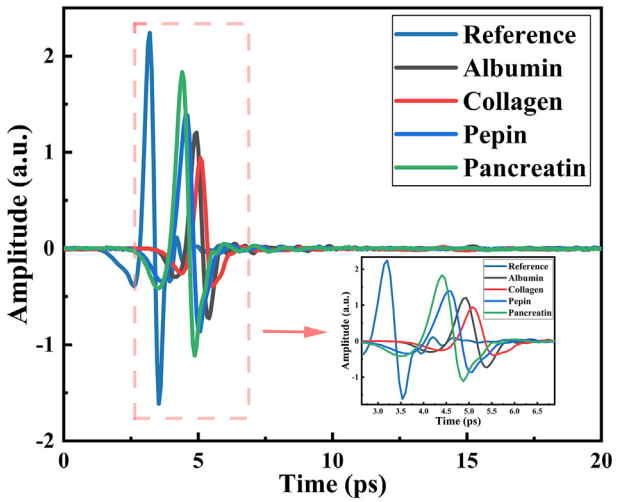


Figure 3. THz time-domain spectra of the reference and the four protein samples

two-dimensional images as a deep learning dataset using the Grassia angular summation field (GASF) method, respectively. The VGG-16 model (Figure 2) for detecting the four proteins is constructed utilizing the transfer learning technique, and the results show that by taking the THz absorption-refractive index spectral data as the training and testing dataset, the VGG-16 model has a faster distinction speed and a higher distinction accuracy (98.8%) for the four proteins. This study provides an accurate and fast method for distinguishing different kinds of proteins. The method can be extended to different frequency regions, types of spectroscopy, or other spectral identification problems, provided suitable training datasets exist or can be obtained.

RESULTS

Spectroscopic analysis

The time-domain spectra of each sample were measured by the THz-TDS (Figure 3). From the time-domain waveform, compared with the reference signal, the maximum peaks of the four proteins show attenuation to varying degrees. The more time the scanning takes, the delay time of different proteins will be different, and the longer the time delay of the protein signal, the worse the signal of peak attenuation. Such phenomena occur due to the absorption, reflection, and dispersion of the THz pulse on the sample surface.[41]

The absorption and refractive index spectra of the four proteins can be obtained through Fourier transform of the THz time-domain spectral signal (Figures 4A and 4B). Figure 4A indicates no obvious characteristic absorption peaks in the range of 0.2–1.6 THz and only the trend of the absorption spectra can be seen. The absorption curves of different proteins are different, but the overall trend is similar, i.e., the protein absorption increases with the increase of frequency. Figure S1 shows the ten time-domain absorption spectra of each protein obtained after 1,000 scans and averaging. From the Figure S1, it can be seen that the THz spectral data of the four proteins are in good agreement, which also confirms the reliability of the data. Figure 4C shows the maximum, minimum, and average absorptions for different types of proteins. It can be seen that the maximum absorption coefficient, the minimum absorption coefficient, and the average absorption coefficient of collagen are the largest among the four proteins, which implies that the chemical bonds in collagen have stronger vibrational or rotational modes in the terahertz band, leading to stronger absorption. In contrast, pancreatin has the smallest maximum absorption coefficient, and the smallest average absorption coefficient, implying that the molecular vibrational or rotational modes associated with terahertz spectroscopy in pancreatin are likely to be weaker, and therefore less absorbed in this frequency band. Figure 4B shows that different kinds of proteins have different refractive indices. It indicates that the transmission of terahertz waves varies for different proteins. Figure 4D shows the maximum, minimum, and average refractive index of different types of proteins. It can be seen that the average refractive index of collagen is the maximum, which indicates that the transmission of the THz wave in collagen is slow. The average refractive index of albumin is the minimum, which indicates that the transmission of the THz wave in albumin is fast. This is the properties of the proteins. The results showed that different protein samples have different characteristics and different refractive indices.

Distinguishing by using THz absorption spectra

After the THz absorption spectra of the four proteins are converted into two-dimensional images using the GASF method, they are employed as the dataset for the transfer learning-based VGG-16 neural network. The dataset is divided into 80% for training and 20% for testing. The optimizer called Adam with a learning rate of 0.00015 was used to compile the model and the categorical cross-entropy, as a loss function, was used for the loss value in the optimization process of the model. VGG-16 typically involves deep architectures that can extract features of abstract and invariant data, leading to better predictive performance compared to traditional machine learning algorithms.[42,43] The input two-dimensional image is subjected to a convolution operation through multiple convolution kernels, each of which learns a different feature from the input image. The first convolutional layer may learn only simple edge and texture features, while subsequent convolutional layers learn more complex structures such as details, local morphology in the image. The visualization of feature maps can provide a better understanding of the process of feature extraction and the performance of the VGG-16 model. Figures 5A–5M shows the feature maps of the two-dimensional image of the THz absorption spectrum of albumin after thirteen convolutional layers of the VGG-16 network model, respectively. The total number of feature maps output after each convolutional layer is indicated, with only 9 feature maps shown for each convolutional layer in the figure. The pooling layer follows the convolutional layer, and VGG-16 uses maximum pooling to reduce the spatial resolution of the feature map. After convolution and pooling operations, the features in the image are flattened and fed into the fully connected layer. The role of the fully connected layer is to integrate the extracted features and make identification decisions based on these features. VGG-16 has three fully connected layers, where the last fully connected layer is responsible for outputting the final

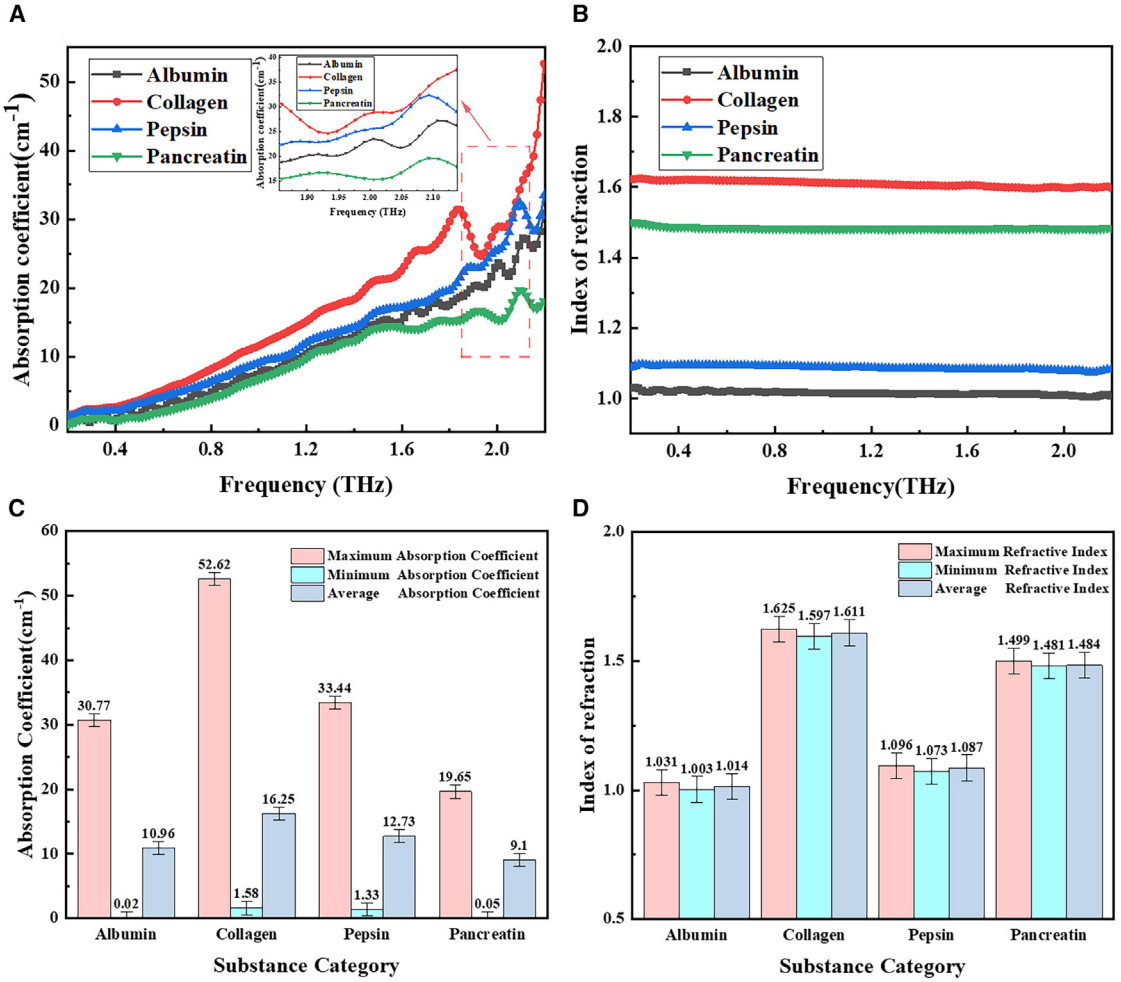


Figure 4. Absorption spectrum and refraction spectrum information

(A) The THz absorption coefficient spectra.

(B) Refractive index spectra.

(C) The absorption coefficient and average absorption coefficient.

(D) The refractive indices and average refractive indices, for different proteins.

protein recognition result. Figures 5N–5P shows the distribution of the weight parameters of the three fully connected layers.

For the THz absorption spectra of the four proteins, the loss and accuracy curves of the VGG-16 model based on transfer learning are shown in Figures 6A and 6B. Because the convolutional layer of the VGG-16 model uses weight parameters that have already been trained, the model begins to converge after 12 rounds of training, at which point the accuracy on the training and test datasets reaches 89.9% and 90.9%, respectively. Figure 6C calculates the confusion matrix for the four proteins,[44] which is mainly used to compare objective results with actual measurements to characterize the classification performance of each category. The overall accuracy of the model for the four proteins was calculated to be 90.90%. It can be seen that the model has a relatively high probability of misidentifying pancreatin and does not accurately recognize pepsin and pancreatin. Therefore, the VGG-16 model based on transfer learning does not have a high accuracy in recognizing the THz absorption spectra of the four proteins. To improve the recognition accuracy of the four proteins, the refractive indices, and absorptions of the four proteins are jointly used as learning features to input the model for training and recognition.

Distinguishing by using THz absorption-refractive index spectra

By taking the refractive index of the proteins as the horizontal coordinate while the absorption as the vertical coordinate, the absorption-refractive index spectra are obtained, as shown in Figure 6. It can be seen in Figure 7 that the absorption-refractive index spectra of the four proteins are significantly different compared to the absorption spectra, and therefore, they can be used as the learning features of the VGG-16 model.

All the absorption-refractive index spectra of the four proteins were converted into two-dimensional graphs by using the GASF

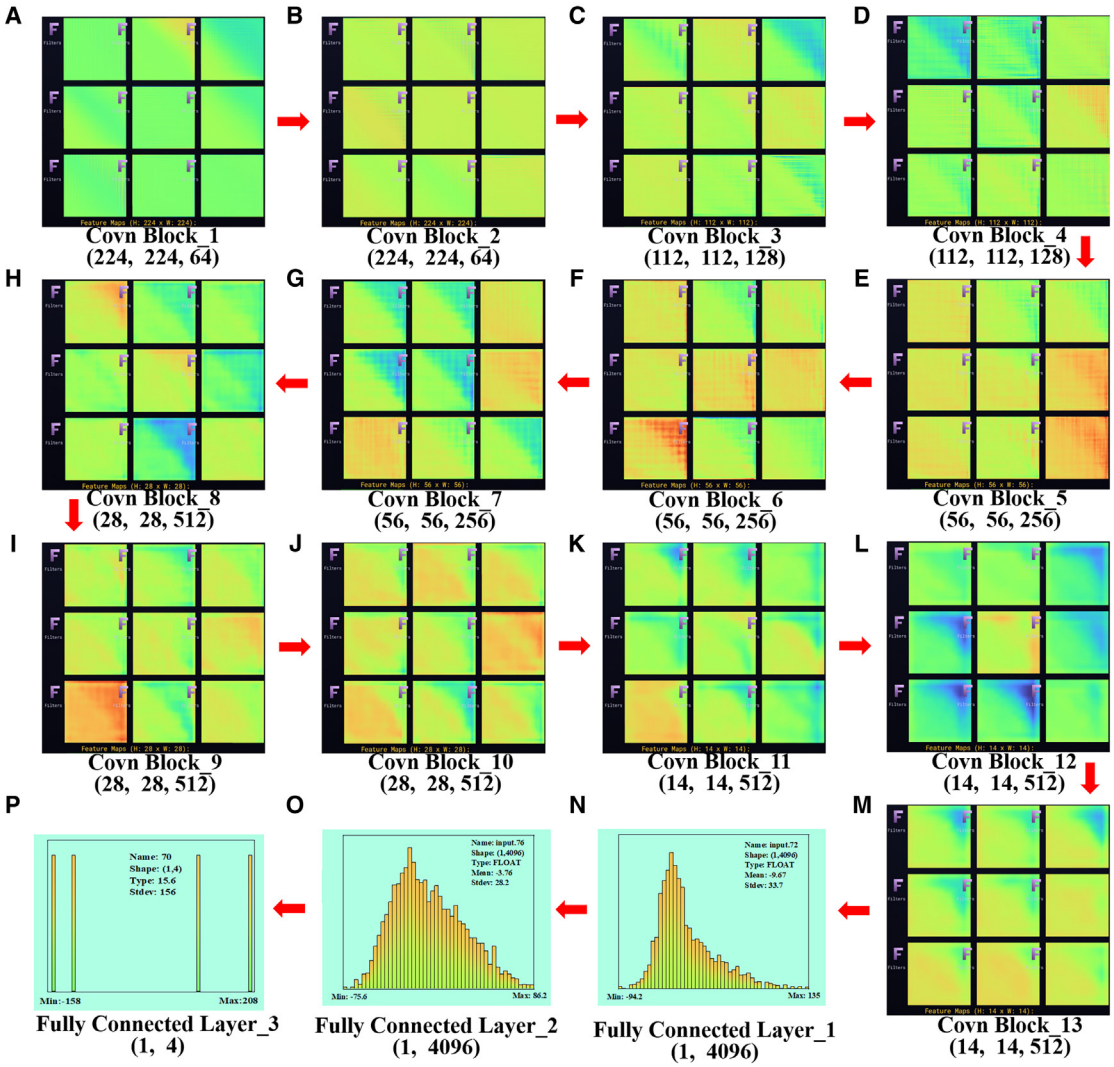


Figure 5. The visualization of feature maps

(A–M) The output feature maps of the two-dimensional image of the THz absorption spectrum of albumin after each convolutional layer, respectively.
(N–P) Distribution of weight parameters for three fully connected layers.

method. It was divided into a training set (80%) and a testing set (20%). Figures 8A–8M are the feature maps of the two-dimensional image of the absorption-refractive index spectra after thirteen convolutional layers of the VGG-16 model, respectively. Figures 8N–8P shows the distribution of weight parameters of the three fully connected layers.

Figures 9A and 9B show the loss and accuracy curves of the VGG-16 model when the THz absorption-refractive index spectra are used as the dataset. Figure 9A shows that the loss value of the model decreases with the number of iterations. The model converges when the number of iterations reaches 10. Figure 9B indicates that the accuracy curves of the training and test datasets increase with the number of iterations, and the accuracies on the training and test sets reach 98.6% and 98.8%, respectively, indicating that the VGG-16 model based on transfer learning has excellent recognition ability for the absorption-refractive index spectra of the four proteins. Meanwhile, the VGG-16 model based on transfer learning utilizes the already learned model weights and does not need to retrain the model, making the model training time much shorter. The confusion matrix of the four proteins was calculated in Figure 9C, which shows that after using absorption-refractive index spectra as the dataset, the model's recognition accuracy for the three proteins collagen, pepsin, and pancreatin reaches 100% and that for albumin reaches 95%, and the overall accuracy of the VGG-16 model for the four proteins reaches 98.80%.

DISCUSSION

Figure 10 shows the identification results of the VGG-16 model based on transfer learning for the THz absorption spectra and the THz absorption-refractive index spectra of the four

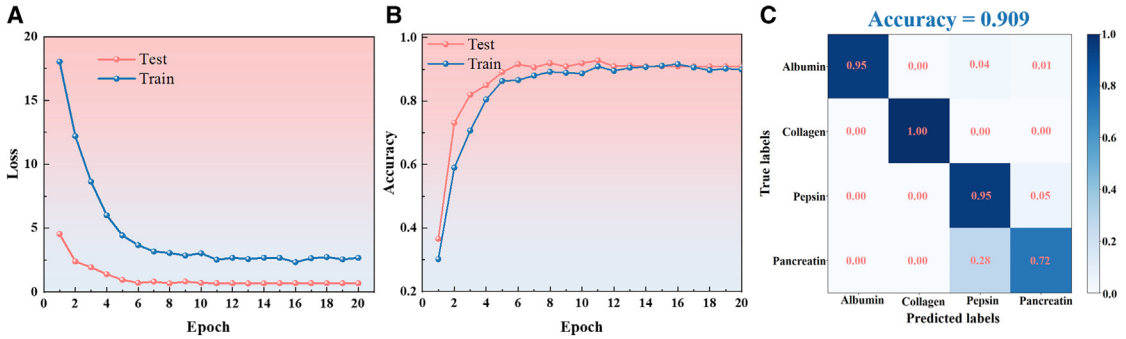


Figure 6. Classification results of different kinds of proteins in the range of 0.2–2.2 THz based on the VGG-16 model

(A) Loss curves of absorption spectra on training and test sets.

(B) Accuracy curves of absorption spectra on training and test sets.

(C) Confusion matrix for detailed exploring of the classification results of the absorption spectra.

proteins, respectively. It can be observed that using the THz absorption-refractive index spectra as learning features; the model has less loss and higher recognition accuracy.

To verify the effectiveness of the VGG-16 model based on transfer learning in this article, we use absorption-refractive spectra which contain more features as a dataset, and compare them

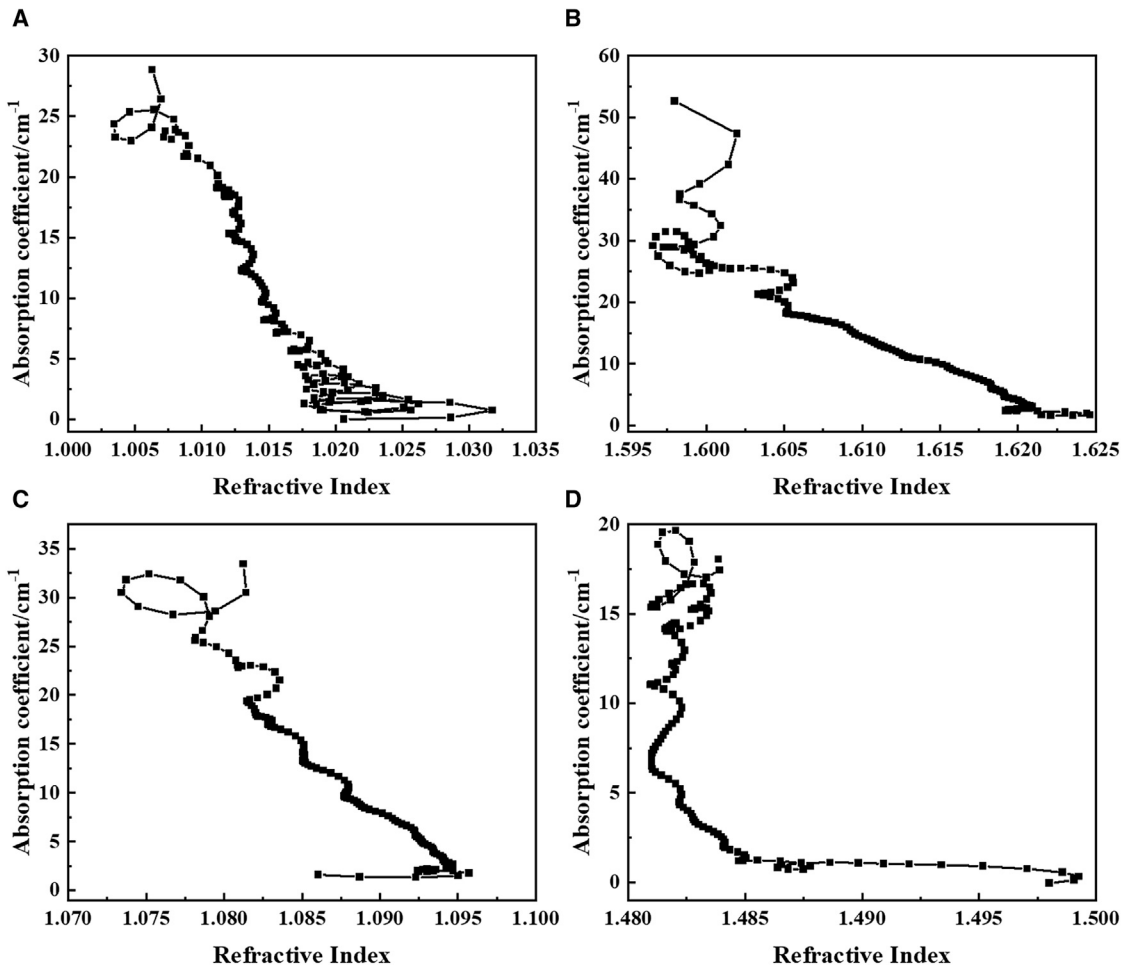


Figure 7. The absorption-refractive index spectra

(A) Albumin.

(B) Collagen.

(C) Pepsin.

(D) Pancreatin.

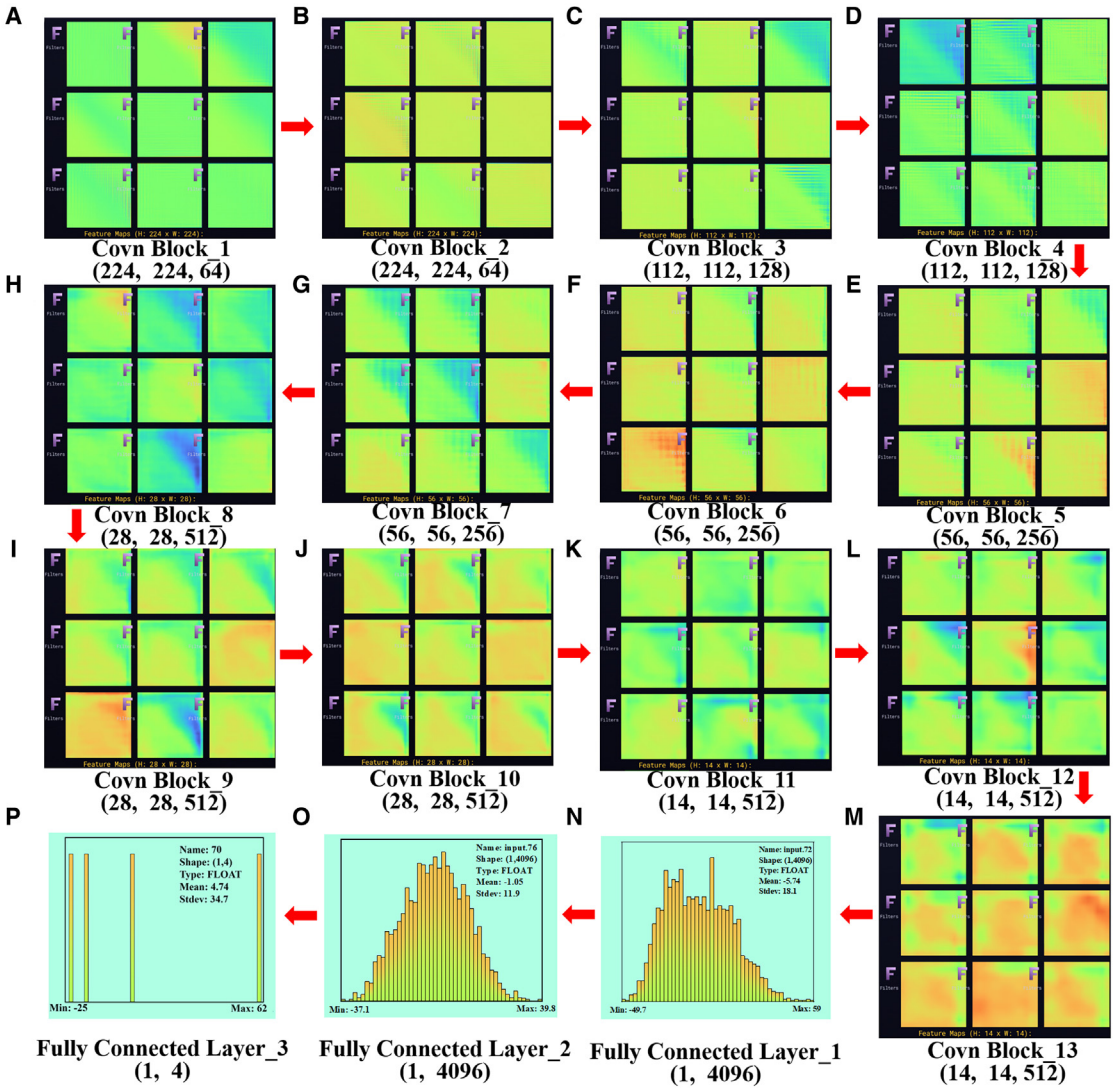


Figure 8. The visualization of feature maps

(A–M) The output feature maps of the two-dimensional image of the absorption-refractive index spectra of albumin after each convolutional layer, respectively. (N–P) Distribution of weight parameters for three fully connected layers.

with some other established methods, such as the support vector machine (SVM),[45] Gaussian processes classifier (GPC),[46,47] BiGRU,[48] and convolutional neural network and bidirectional gated recurrent (CNN-BiGRU).[35] As can be seen in Figure 11, the identification accuracy of our proposed VGG-16 model is 19.74% and 16.61% higher than that of SVM and GPC, and 8.8%, 1.61% higher than that of CNN-BiGRU and BiGRU. Therefore, the constructed VGG-16 model can identify four proteins well.

Limitations of the study

The method was tested on only four specific proteins (albumin, collagen, pepsin, and pancreatin), and its generalizability to a complex biological mixtures remains to be validated. While the VGG-16 model achieved high accuracy, its computational complexity and resource requirements could pose challenges for real-time or large-scale applications.

Conclusion

This study proposes a method of employing the VGG-16 neural network to accurately identify the THz absorption-refractive index spectra for detecting the four kinds of proteins, namely albumin, collagen, pepsin, and pancreatin. The VGG-16 model based on transfer learning saves training time and has high recognition accuracy. On the other hand, the THz spectra are converted into two-dimensional images as the dataset using the GASF algorithm. If only the THz absorption spectra of the proteins, whose features are very similar, are used as feature information for training and predicting, the recognition accuracy of the model for the four proteins is 90.9%, resulting in a lower identification accuracy of the model. Further research finds that, when the THz absorption-refractive index spectra of the proteins are used as feature information to train and fine-tune the model, the model's accuracy in identifying the four types of proteins

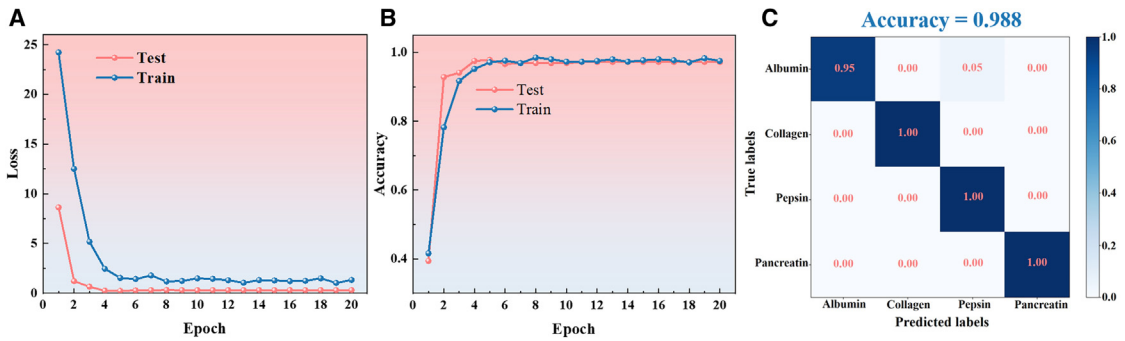


Figure 9. Classification results of different kinds of proteins in the range of 0.2–2.2 THz based on the VGG-16 model

(A) Loss curves of absorption-refractive index spectra on training and test sets.

(B) Accuracy curves of absorption-refractive index spectra on training and test sets.

(C) Confusion matrix for detailed exploring of the classification results of the absorption-refractive index spectra.

improves by up to 98.8%. Meanwhile, the VGG-16 model also achieved better performance as compared with other deep learning and machine learning models. Therefore, for the recognition of biochemical substances with THz spectroscopy, the absorption and refractive index spectra can be considered together as feature information. In conclusion, our proposed method can achieve rapid and accurate identification of the four proteins and could be extended to the identification of other biochemical substances, such as amino acids, sugars, nucleic acids, etc.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Wengang Wu (wuwg@pku.edu.cn).

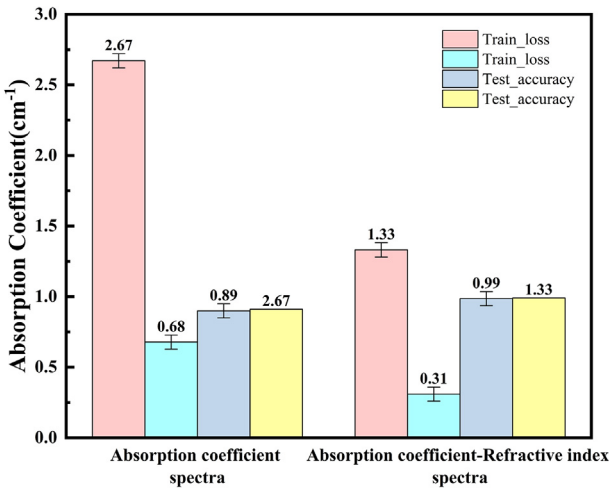


Figure 10. Comparison of THz absorption spectra and THz absorption-refractive index spectra identification results of the four proteins by VGG-16 model

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All code necessary to reproduce the results is detailed in the key resources table. (https://doi.org/10.5281/zenodo.14931084)
- The datasets analyzed in this paper are taken from existing publicly available datasets. The DOIs are listed in the key resources table. (https://doi.org/10.5281/zenodo.14931094)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (Grant No. 2021YFB3200100). The authors thank the Peking Nanofab for providing excellent fabrication conditions.

AUTHOR CONTRIBUTIONS

Y.C., conceptualization, data curation, validation, and writing the original draft; X.H., visualization, resources, and methodology; M.W., formal analysis and

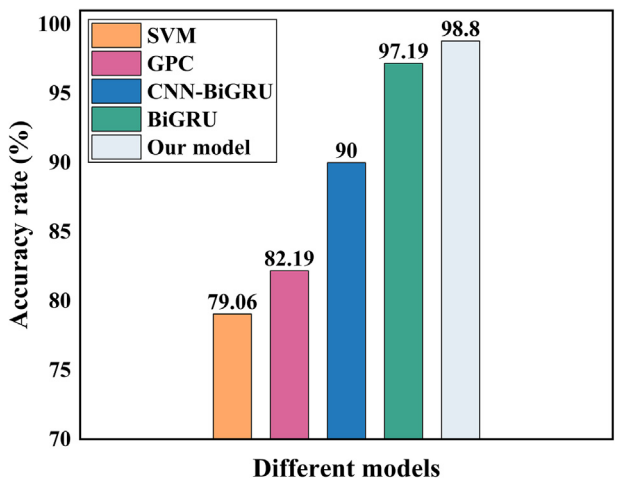


Figure 11. The accuracy rate comparison of each model

writing the original draft; J.H., methodology and formal analysis; Y.C., methodology and formal analysis; H.S., investigation and resources; L.M., conceptualization; L.L., investigation; W.W., methodology, supervision, and writing – review and editing; G.Z., investigation; T.M., resources.

DECLARATION OF INTERESTS

The authors declare no conflicts of interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Experimental equipment
  - Sample preparation
  - Spectral acquisition and pretreatment
  - Acquisition of two-dimensional image data
  - Transfer learning
  - VGG-16 neural network
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Metrics
  - Software tools
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112148.

Received: October 4, 2024
Revised: January 17, 2025
Accepted: February 27, 2025
Published: March 3, 2025

REFERENCES

1. Jagdish, R.K., Maras, J.S., and Sarin, S.K. (2021). Albumin in advanced liver diseases: the good and bad of a drug. Hepatology *74*, 2848–2862.
2. Ronit, A., Kirkegaard-Klitbo, D.M., Dohlmann, T.L., Lundgren, J., Sabin, C.A., Phillips, A.N., Nordestgaard, B.G., and Afzal, S. (2020). Plasma albumin and incident cardiovascular disease: results from the CGPS and an updated meta-analysis. Arterioscler. Thromb. Vasc. Biol. *40*, 473–482.
3. Rezvani Ghomi, E., Nourbakhsh, N., Akbari Kenari, M., Zare, M., and Ramakrishna, S. (2021). Collagen-based biomaterials for biomedical applications. J. Biomed. Mater. Res. B Appl. Biomater. *109*, 1986–1999.
4. Li, W., Zhao, W., Cheng, S., Zhang, H., Yi, Z., Sun, T., Wu, P., Zeng, Q., and Raza, R. (2024). Tunable metamaterial absorption device based on Fabry–Perot resonance as temperature and refractive index sensing. Opt. Lasers Eng. *181*, 108368.
5. Bandorski, D., Tello, K., Erdal, H., Sommerlad, J., Wilhelm, J., Vadasz, I., Hecker, M., Walmrath, D., Seeger, W., Krauss, E., and Kuhnert, S. (2023). Clinical Utility of Pepsin and Bile Acid in Tracheal Secretions for Accurate Diagnosis of Aspiration in ICU Patients. J. Clin. Med. *12*, 5466.
6. Li, Y., Xu, G., Zhou, B., Tang, Y., Liu, X., Wu, Y., Wang, Y., Kong, J., Xu, T., He, C., et al. (2022). Effects of acids, pepsin, bile acids, and trypsin on laryngopharyngeal reflux diseases: physiopathology and therapeutic targets. Eur. Arch. Otorhinolaryngol. *279*, 2743–2752.
7. Arutla, M., Sarkar, S., Unnisa, M., Sarkar, P., Raj, M.A., Mrudula, M.R., Deepika, G., Pasham, S., Jakkampudi, A., Prasanna, A., et al. (2021). Malnutrition after pancreatic enzyme replacement therapy in chronic pancreatitis: risk factors in real world practice. Pancreatology *21*, 34–41.
8. Layer, P., Kashirskaya, N., and Gubergrits, N. (2019). Contribution of pancreatic enzyme replacement therapy to survival and quality of life in patients with pancreatic exocrine insufficiency. World J. Gastroenterol. *25*, 2430–2441.
9. Wang, T., Zhao, L., Huang, P., Zhang, X., and Xu, J. (2021). Haze concentration adaptive network for image dehazing. Neurocomputing *439*, 75–85.
10. Zhao, H., Qiu, X., Lu, W., Huang, H., and Jin, X. (2020). High-quality retinal vessel segmentation using generative adversarial network with a large receptive field. Int. J. Imaging Syst. Technol. *30*, 828–842.
11. Kelly, R.T. (2020). Single-cell proteomics: progress and prospects. Mol. Cell. Proteomics *19*, 1739–1748.
12. Ye, Q., Lu, S., and Corbett, K.D. (2021). Structural basis for SARS-CoV-2 nucleocapsid protein recognition by single-domain antibodies. Front. Immunol. *12*, 719037.
13. Letertre, M.P.M., Giraudeau, P., and De Tullio, P. (2021). Nuclear magnetic resonance spectroscopy in clinical metabolomics and personalized medicine: current challenges and perspectives. Front. Mol. Biosci. *8*, 698337.
14. Lim, H.J., Saha, T., Tey, B.T., Tan, W.S., and Ooi, C.W. (2020). Quartz crystal microbalance-based biosensors as rapid diagnostic devices for infectious diseases. Biosens. Bioelectron. *168*, 112513.
15. Mann, M., Hendrickson, R.C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. *70*, 437–473.
16. Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. Cell *159*, 995–1014.
17. Galvan, D., de Aguiar, L.M., Bona, E., Marini, F., and Killner, M.H.M. (2023). Successful combination of benchtop nuclear magnetic resonance spectroscopy and chemometric tools: A review. Anal. Chim. Acta *1273*, 341495.
18. Hayrapetyan, H., Tran, T., Tellez-Corrales, E., and Madiraju, C. (2023). Enzyme-linked immunosorbent assay: types and applications. Methods Mol. Biol. *2612*, 1–17.
19. Qu, F., Lin, L., Cai, C., Chu, B., Wang, Y., He, Y., and Nie, P. (2021). Terahertz fingerprint characterization of 2, 4-dichlorophenoxyacetic acid and its enhanced detection in food matrices combined with spectral baseline correction. Food Chem. *334*, 127474.
20. Cheng, S., Li, W., Zhang, H., Akhtar, M.N., Yi, Z., Zeng, Q., Ma, C., Sun, T., Wu, P., and Ahmad, S. (2024). High sensitivity five band tunable metamaterial absorption device based on block like Dirac semimetals. Opt. Commun. *569*, 130816.
21. Liu, H., Hu, D.J.J., Sun, Q., Wei, L., Li, K., Liao, C., Li, B., Zhao, C., Dong, X., Tang, Y., et al. (2023). Specialty optical fibers for advanced sensing applications. Opto-Electronic Science *2*, 220025.
22. Nie, P., Cai, C., Qu, F., Lin, L., Dong, T., and He, Y. (2019). Study of 2, 4-d spectral characteristics and its detection in Zizania latifolia using terahertz time-domain spectroscopy. Appl. Sci. *9*, 2248.
23. Zhou, X., Pu, H., and Sun, D.-W. (2021). DNA functionalized metal and metal oxide nanoparticles: Principles and recent advances in food safety detection. Crit. Rev. Food Sci. Nutr. *61*, 2277–2296.
24. Ye, P., Wang, G., Yang, Y., Meng, Q., Wang, J., Su, B., and Zhang, C. (2021). Terahertz absorption properties of two solid amino acids and their aqueous solutions. Int. J. Opt. *2021*, 1–7.
25. Wang, Y., Zhao, Z., Qin, J., Liu, H., Liu, A., and Xu, M. (2020). Rapid in situ analysis of l-histidine and α-lactose in dietary supplements by fingerprint peaks using terahertz frequency-domain spectroscopy. Talanta *208*, 120469.
26. Tych, K.M., Burnett, A.D., Wood, C.D., Cunningham, J.E., Pearson, A.R., Davies, A.G., and Linfield, E.H. (2011). Applying broadband terahertz time-domain spectroscopy to the analysis of crystalline proteins: a dehydration study. J. Appl. Crystallogr. *44*, 129–133.
27. Wen-ai, W., and Wei, L. (2021). Terahertz Spectroscopy Characteristics of Sugar Compounds. Spectrosc. Spectr. Anal. *41*, 2391–2396.

28. Li, Q., Lei, T., and Sun, D.-W. (2023). Analysis and detection using novel terahertz spectroscopy technique in dietary carbohydrate-related research: Principles and application advances. Crit. Rev. Food Sci. Nutr. *63*, 1793–1805.

29. Jing, J., Liu, K., Jiang, J., Xu, T., Wang, S., and Liu, T. (2023). Highly sensitive and stable probe refractometer based on configurable plasmonic resonance with nano-modified fiber core. Opto-Electronic Advances *6*, 220072.

30. Wei, L., Yu, L., Jiaoqi, H., Guorong, H., Yang, Z., and Weiling, F. (2018). Application of terahertz spectroscopy in biomolecule detection. Frontiers in Laboratory Medicine *2*, 127–133.

31. Pu, H., Yu, J., Sun, D.-W., Wei, Q., and Li, Q. (2023). Distinguishing pericarpium citri reticulatae of different origins using terahertz time-domain spectroscopy combined with convolutional neural networks. Spectrochim. Acta, Part A *299*, 122771.

32. Liu, Y., Pu, H., Li, Q., and Sun, D.-W. (2023). Discrimination of Pericarpium Citri Reticulatae in different years using Terahertz Time-Domain spectroscopy combined with convolutional neural network. Spectrochim. Acta, Part A *286*, 122035.

33. Qi-feng, H., and Jian, C. (2021). Research of terahertz time-domain spectral identification based on deep learning. Spectrosc. Spectr. Anal. *41*, 94–99.

34. Wang, S., and Xiang, J. (2020). A minimum entropy deconvolution-enhanced convolutional neural networks for fault diagnosis of axial piston pumps. Soft Comput. *24*, 2983–2997.

35. Li, T., Xu, Y., Luo, J., He, J., and Lin, S. (2021). A method of amino acid terahertz spectrum recognition based on the convolutional neural network and bidirectional gated recurrent network model. Sci. Program. *2021*, 1–7.

36. Wang, B., Qin, X., Meng, K., Zhu, L., and Li, Z. (2022). Classification of Amino Acids Using Hybrid Terahertz Spectrum and an Efficient Channel Attention Convolutional Neural Network. Nanomaterials *12*, 2114.

37. Klokkou, N., Gorecki, J., Wilkinson, J.S., and Apostolopoulos, V. (2022). Artificial neural networks for material parameter extraction in terahertz time-domain spectroscopy. Opt. Express *30*, 15583–15595.

38. Zhou, Z., Jia, S., and Cao, L. (2022). A general neural network model for complex refractive index extraction of low-loss materials in the transmission-mode thz-tds. Sensors *22*, 7877.

39. Loahavilai, P., Datta, S., Prasertsuk, K., Jintamethasawat, R., Rattanawan, P., Chia, J.Y., Kingkan, C., Thanapirom, C., and Limpanuparb, T. (2022). Chemometric analysis of a ternary mixture of caffeine, quinic acid, and nicotinic acid by terahertz spectroscopy. ACS Omega *7*, 35783–35791.

40. Liang, J., Lu, X., Chang, T., and Cui, H.-L. (2022). Deep learning aided quantitative analysis of anti-tuberculosis fixed-dose combinatorial formulation by terahertz spectroscopy. Spectrochim. Acta, Part A *269*, 120746.

41. Shen, Y., Yin, Y., Li, B., Zhao, C., and Li, G. (2021). Detection of impurities in wheat using terahertz spectral imaging and convolutional neural networks. Comput. Electron. Agric. *181*, 105931.

42. Zhou, L., Zhang, C., Liu, F., Qiu, Z., and He, Y. (2019). Application of deep learning in food: a review. Compr. Rev. Food Sci. Food Saf. *18*, 1793–1811.

43. Lin, Y., Jin, X., Chen, J., Sodhro, A.H., and Pan, Z. (2019). An analytic computation-driven algorithm for Decentralized Multicore Systems. Future Gener. Comput. Syst. *96*, 101–110.

44. Huang, P., Cao, Y., Chen, J., Ge, W., Hou, D., and Zhang, G. (2019). Analysis and inspection techniques for mouse liver injury based on terahertz spectroscopy. Opt. Express *27*, 26014–26026.

45. Li, K., Chen, X., Zhang, R., and Pickwell-MacPherson, E. (2020). Classification for glucose and lactose terahertz spectrums based on SVM and DNN methods. IEEE Trans. Terahertz Sci. Technol. *10*, 617–623.

46. Liao, J., Wang, B., Wang, Z., and Zhu, L. (2023). Amino-acid classification based on terahertz absorption spectroscopy with Gaussian process and maximum likelihood. Sens. Actuators, B *388*, 133806.

47. Seeger, M. (2004). Gaussian processes for machine learning. Int. J. Neural Syst. *14*, 69–106.

48. Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., and Elmaghraby, A. (2020). Ensemble deep learning models for heart disease classification: A case study from Mexico. Information *11*, 207.

49. Schecklman, S., Zurk, L.M., Henry, S., and Kniffin, G.P. (2011). Terahertz material detection from diffuse surface scattering. J. Appl. Phys. *109*, 094902.

50. Wang, H., Shi, W., Hou, L., Wang, Z., Wu, M., Li, C., and Li, C. (2022). Effect of THz spectra of L-Arginine molecules by the combination of water molecules. iScience *25*, 103788.

51. Rasekh, P., Safari, A., Yildirim, M., Bhardwaj, R., Ménard, J.-M., Dolgaleva, K., and Boyd, R.W. (2021). Terahertz nonlinear spectroscopy of water vapor. ACS Photonics *8*, 1683–1688.

52. Tang, M., Xia, L., Wei, D., Yan, S., Zhang, M., Yang, Z., Wang, H., Du, C., and Cui, H.-L. (2020). Rapid and label-free metamaterial-based biosensor for fatty acid detection with terahertz time-domain spectroscopy. Spectrochim. Acta, Part A *228*, 117736.

53. Dorney, T.D., Baraniuk, R.G., and Mittleman, D.M. (2001). Material parameter estimation with terahertz time-domain spectroscopy. JOSA A *18*, 1562–1571.

54. Duvillaret, L., Garet, F., and Coutaz, J.-L. (1996). A reliable method for extraction of material parameters in terahertz time-domain spectroscopy. IEEE J. Sel. Top. Quantum Electron. *2*, 739–746.

55. Zhou, Y., Long, X., Sun, M., and Chen, Z. (2022). Bearing fault diagnosis based on Gramian angular field and DenseNet. Math. Biosci. Eng. *19*, 14086–14101.

56. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. Proc. IEEE *109*, 43–76.

57. Olivas, E.S., Guerrero, J. D.M., Sober, M.M., Benedito, J.R.M., and López, A.J.S. (2010). In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (IGI global), pp. 242–264.

58. Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.1556.

59. Zhang, J., Yang, Y., Feng, X., Xu, H., Chen, J., and He, Y. (2020). Identification of bacterial blight resistant rice seeds using terahertz imaging and hyperspectral imaging combined with convolutional neural network. Front. Plant Sci. *11*, 821.

60. Andreieva, V., and Shvai, N. (2020). Generalization of cross-entropy loss function for image. https://doi.org/10.18523/2617-7080320203-10.

61. Zhang, Y., Gao, J., Cen, H., Lu, Y., Yu, X., He, Y., and Pieters, J.G. (2019). Automated spectral feature extraction from hyperspectral images to differentiate weedy rice and barnyard grass from a rice crop. Comput. Electron. Agric. *159*, 42–49.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Albumin	Shanghai Mai Lin Biochemical Technology Co.	A924305
Collagen	Shanghai Mai Lin Biochemical Technology Co.	C875812
Pepsin	Shanghai Mai Lin Biochemical Technology Co.	P916045
Pancreatin	Shanghai Mai Lin Biochemical Technology Co.	P885908
Deposited data		
The datasets	This study	https://doi.org/10.5281/zenodo.14931094
Software and algorithms		
Python	Version 3.8.0	https://www.python.org/
PyTorch	Version 1.8.0	https://pytorch.org/
Numpy	Version 1.24.3	https://numpy.org/
Origin	Version 2021	https://www.originlab.com/
Our code	This study	https://doi.org/10.5281/zenodo.14931084

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study is computational science research and does not utilize experimental models typical of the life sciences.

METHOD DETAILS

Experimental equipment

THz-TDS is one of the most typical THz spectroscopy measurement technologies, which exhibits the advantage of obtaining both the THz absorption spectra and refractive index spectra of biomolecules at the same time. The THz-TDS system used in this study is schematically shown in Figure 1. It consists of an 800 nm femtosecond laser, a time-delay control system, a Lock-in, a THz radiation emitter, and a detector to obtain terahertz time-domain spectroscopy signals. To facilitate the analysis of THz spectral data, the time-domain signals are Fourier transformed into frequency-domain signals with a resolution of 14.00 GHz, ranging from 0.1 THz to 3.0 THz. We used 138 frequency points from 0.2-2.2 THz to identify different kinds of proteins due to the low signal-to-noise ratio of 0.1-0.2 THz and 2.2-3 THz. To avoid the influence of water vapor in the air, we built a vacuum system and placed the THz-TDS in this system to ensure that the samples were tested without the influence of the surrounding environment and to maximize the authenticity of the experiment.

Sample preparation

The samples of albumin, collagen, pepsin, pancreatin, and high-density polyethylene used in the experiment were purchased from Shanghai Macklin Biochemical Technology Co. The purity of both albumin and pepsin was $\geq$ 98.5%. The moisture content of collagen is $\leq$ 8% and the heavy metal content is $\leq$ 10 ppm. Pancreatin has a white to yellow to tan powder appearance and an enzymatic activity of $\geq$ 130 μ/mg, and moisture content $\leq$ 7%. The four protein samples were stored at a low temperature below 8°C before the experiment, and the preparation and handling methods of the samples during the experiment had an important influence on the experimental data. Accurate THz spectral information will not be obtained if the test samples are not prepared properly. Weighed 125 mg of polyethylene as well as 25 mg of the samples at a concentration of 1:5, respectively, the mixtures were placed in a mortar and slowly ground with a pestle and pestle for about 3 min to reduce the Mie scattering.[49] Then they were transferred to a mold with a diameter of 15 mm, respectively, and pressurized at a pressure of 6 MPa for about 3 min. Round solid tablets with a range of thicknesses in the range of 1.0 ± 0.1 mm were obtained. The front and rear surfaces of the tested samples were carefully prepared to be smooth and parallel, ensuring no damage. The thickness of each sample was measured using a helical micrometer and documented. The samples were then stored in a sample bag.

Spectral acquisition and pretreatment

Since THz radiation is very sensitive to water molecules,[50,51] the vacuum function was utilized before each test to remove all the air from the system, so that the testing process was completely in a vacuum environment, thus avoiding the influence of water vapor in the air on the test results. The temperature of the laboratory was controlled at 26°C (±1°C), and the terahertz spectra when the sample was not placed were first collected as a reference, and then the protein solid tablets were put into the measurement chamber. To obtain protein THz time-domain spectra, each spectrum was obtained by scanning 100 times to minimize the effect of random interferences, and the average spectrum was recorded. The recorded time-domain spectra are converted to THz magnitude and phase by a Fourier transform. In addition, the absorption $\alpha(\omega)$ and the refractive index $n(\omega)$ can be obtained directly by normalizing the frequency-domain spectrum to the corresponding reference spectrum using the following equations[52–54]:

$$n(\omega) = 1 + \varphi(\omega)\cdot\frac{c}{\omega d} \quad \text{(Equation 1)}$$

$$\alpha(\omega) = \frac{2}{d}\ln\left[\frac{4n(\omega)}{\rho(\omega)\cdot[1+n(\omega)]^2}\right] \quad \text{(Equation 2)}$$

Where c is the speed of light in vacuum, d is the thickness of the sample, $\rho(\omega)$ is the amplitude, ω is the angular frequency, and $\varphi(\omega)$ is the phase difference between the sample signal and the reference signal.

Acquisition of two-dimensional image data

THz spectra are characterized by time-series signals because the delay control system regulates the time delay between two pulse trains emitted by a femtosecond laser. The Gramian Angular Field (GAF) transforms one-dimensional spectral data from a Cartesian to a polar coordinate system. This conversion enables the identification of temporal correlations between various frequency points by computing the GASF and the Gramian Angular Difference Field (GADF), as illustrated below[55]:

$$GASF = \begin{bmatrix} \cos(\varphi_1+\varphi_1) & \cos(\varphi_1+\varphi_2) & \cdots & \cos(\varphi_1+\varphi_m) \\ \cos(\varphi_2+\varphi_1) & \cos(\varphi_2+\varphi_2) & \cdots & \cos(\varphi_2+\varphi_m) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\varphi_m+\varphi_1) & \cos(\varphi_m+\varphi_2) & \cdots & \cos(\varphi_m+\varphi_m) \end{bmatrix} \quad \text{(Equation 3)}$$

$$GADF = \begin{bmatrix} \cos(\varphi_1-\varphi_1) & \cos(\varphi_1-\varphi_2) & \cdots & \cos(\varphi_1-\varphi_m) \\ \cos(\varphi_2-\varphi_1) & \cos(\varphi_2-\varphi_2) & \cdots & \cos(\varphi_2-\varphi_m) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\varphi_m-\varphi_1) & \cos(\varphi_m-\varphi_2) & \cdots & \cos(\varphi_m-\varphi_m) \end{bmatrix} \quad \text{(Equation 4)}$$

Where *GASF* represents the cosine of the sum of angles, while *GADF* denotes the cosine of the angular difference. m refers to the number of frequency points and φ_m indicates the angle value at the mth frequency point. In this paper, to maintain the time dependence of the THz spectral data as well as to exploit the stability and periodicity of the THz spectra, the GASF is utilized to convert the THz absorption spectra and THz absorption-refractive index spectra into a two-dimensional image.

Transfer learning

Transfer learning[56,57] is a machine learning method that improves the learning effectiveness and efficiency of a new task by applying a model or part of its knowledge that has already been trained on one task to other related tasks. The core idea of transfer learning is to utilize already acquired knowledge to accelerate and improve the learning process of a new task, reducing the dependence on large-scale data and computational resources. In this study, the discrimination of four different kinds of proteins was achieved using the VGG-16 model based on the transfer learning technique.

VGG-16 neural network

VGG-16 is a widely used convolutional neural network model which is mainly used for image classification and recognition tasks. In this study, this model is used for the discrimination of four proteins using the transfer learning technique. The VGG-16 model consists of thirteen convolutional layers and three fully connected layers as shown in Figure 2.

The main feature of the convolutional layers of VGG-16 is the use of several 3x3 convolutional kernels that are stacked layer by layer to increase the depth of the network, the convolutional layer weight parameters are obtained by a transfer learning technique, and only the weight parameters of the last three fully connected layers are trained.[58] Each convolutional operation is followed by a rectified linear units (ReLU) activation function, which is followed by a 2×2 maximum pooling layer for downsampling. After the convolutional layers, the network further processes the features through three fully connected layers, the last of which is connected to a Softmax classifier for generating probability distributions for the four classes.

ReLU is used as an interlayer activation function to mitigate the problem of vanishing gradients and to improve the robustness of the model when dealing with nonlinear data. The expression for ReLU is shown below:

$$ReLU(x) = \begin{cases} x & if\ x > 0 \\ 0 & if\ x \leq 0 \end{cases} \quad \text{(Equation 5)}$$

Where x denotes the eigenvalue of the neurons.

QUANTIFICATION AND STATISTICAL ANALYSIS

Metrics

The VGG-16 neural network model based on the transfer learning technique uses the cross-entropy function as the loss function. The cross-entropy value is non-negative, and the smaller the value of cross-entropy, the better the model prediction. The equation is as follows[59,60]:

$$Loss = -\sum_{i=1}^{u} y_i \log \hat{y}_i \quad \text{(Equation 6)}$$

Where the true distribution is y, the network output distribution is $\hat{y}$, and the total number of categories is u. The performances of the classification models were evaluated by using the ratio of the accuracy (Acc), which could be defined as follows:

$$Acc = \frac{Correct\ \text{classification of samples}}{Total\quad samples} \times 100 \quad \text{(Equation 7)}$$

In addition, the confusion matrix also provides a detailed description of the errors produced by the classification model and shows the true label and the predicted label. In this confusion matrix, the recognition rate is used to evaluate the predictive power of each category, which is defined as follows[61]:

$$Recognitionrate_i = \frac{E_{ii}}{\sum jE_{ij}} \quad \text{(Equation 8)}$$

Where E_{ii} denotes diagonal elements of the ith class, and $\sum_j E_{ij}$ represents the total numbers of ith class samples.

Software tools

Python 3.9 was used to preprocess the THz spectra. Graphical work was performed in Origin 2021 software (Origin Lab Corporation, Northampton, MA, USA). The VGG-16 model based on transfer learning was constructed in the PyTorch framework, and all software work was run on a win10 64-bit computer equipped with a CPU (12th Gen Intel (R) i7-12700K) and GPU (NVIDIA GeForce RTX 3090).

ADDITIONAL RESOURCES

This study did not create or expand any websites or resources, and it does not involve clinical experiments.