STRAIN-LEVEL IDENTIFICATION AND ANALYSIS OF AVIAN CORONAVIRUS USING RAMAN SPECTROSCOPY AND INTERPRETABLE MACHINE LEARNING

Peng Jin[1] *Yin-Ting Yeh*[2] *Jiarong Ye*[1] *Ziyang Wang*[3] *Yuan Xue*[4] *Na Zhang*[2] *Shengxi Huang*[3]
Elodie Ghedin[5] *Huaguang Lu*[6] *Anthony Schmitt*[6] *Sharon X. Huang*[1] *Mauricio Terrones*[2]

[1]College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
[2]Department of Physics, The Pennsylvania State University, University Park, PA, USA
[3]Department of Electrical and Computer Engineering, Rice University, TX, USA
[4]Department of Electrical and Computer Engineering, Johns Hopkins University, MD, USA
[5]Systems Genomics Section, National Institute of Allergy and Infectious Diseases, NIH, MD, USA
[6]Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA

ABSTRACT

Strain-level identification of viruses is important for decision making in public health management. Recently, Raman spectroscopy has attained great attention in virus identification since it enables rapid and label-free analysis. In this paper, we present an interpretable machine learning approach for strain-level identification of avian coronaviruses based on Raman spectra. Specifically, we design a spectral transformer to classify the Raman spectra of 32 avian coronavirus strains. After training, relevance maps can be generated through gradient and relevance propagation to further understand the contribution of each wavenumber to the identification. Experimental results show that the proposed method outperforms several machine learning and deep learning baseline models, and achieves 72.72% accuracy in the 32-class identification problem. The relevance maps generated reveal some wavenumber ranges that are important for the identification of almost all strains, and these ranges correlate with Raman peak ranges for lipids, nucleic acids, and proteins.

***Index Terms*—** Virus Identification, Raman Spectroscopy, Interpretable Machine Learning

1. INTRODUCTION

Viruses can evolve and spread rapidly, thus periodically causing disease outbreaks and threatening public health, such as the Ebola outbreaks [1], the ongoing COVID-19 pandemic [2] and the recent Monkeypox outbreak [3]. Different strains of a virus can lead to different levels of severity in illness. For instance, the Monkeypox strain currently circulating outside of Africa is less severe than the one circulating in the Congo basin [3]. Hence, strain-level virus identification is important for the deployment of a proper public health response [4].

Raman spectroscopy is a non-destructive, label-free technique which provides information about chemical composition and structure in samples. Its potential has been demonstrated in various applications, such as biomarker detection [5], bacterial pathogens identification [6] and cancer diagnosis [7]. By incorporating machine learning techniques, Raman spectroscopy is also applied for virus identification. However, most of the prior work focuses on the detection of single virus strains [8, 9] or the identification of two to four virus species [10, 11, 12]. More recently, one study explored the identification of nine respiratory virus species, where each contains up to two strains [13]. Another similar study also classified nine respiratory virus species and enteroviruses, and up to two strains are included for each species [14]. However, it is still unclear to what extent Raman spectroscopy together with machine learning can be used for the identification of multiple virus strains within a species.

In this paper, we present a study to classify 32 strains of avian coronavirus[1], which is commonly named infectious bronchitis virus (IBV) in avian diseases, using interpretable machine learning and Raman spectroscopy. In particular, we collected 96,802 Raman spectra and developed a spectral transformer for classification as illustrated in Fig. 1. Compared to several conventional machine learning methods and a 1D convolutional neural network (CNN) based deep learning baseline [15], the proposed spectral transformer yielded the best performance in terms of all evaluation metrics and achieved an accuracy of 72.72%. In addition, we employ an improved Layer-wise Relevance Propagation (LRP) method [16] for the interpretation of the spectral transformer. A weighted attention relevance map is generated for each IBV strain. We find there are some Raman wavenumber ranges that are important for the identification of almost all the strains, and those ranges correlate well with Raman peak ranges of important biomolecules existing in viruses.

2. METHODS

In this section, we first describe the network architecture of our spectral transformer and then discuss the details of how

[1]Details of the 32 IBV strains are available at `https://github.com/PengJin95/Raman_IBV`

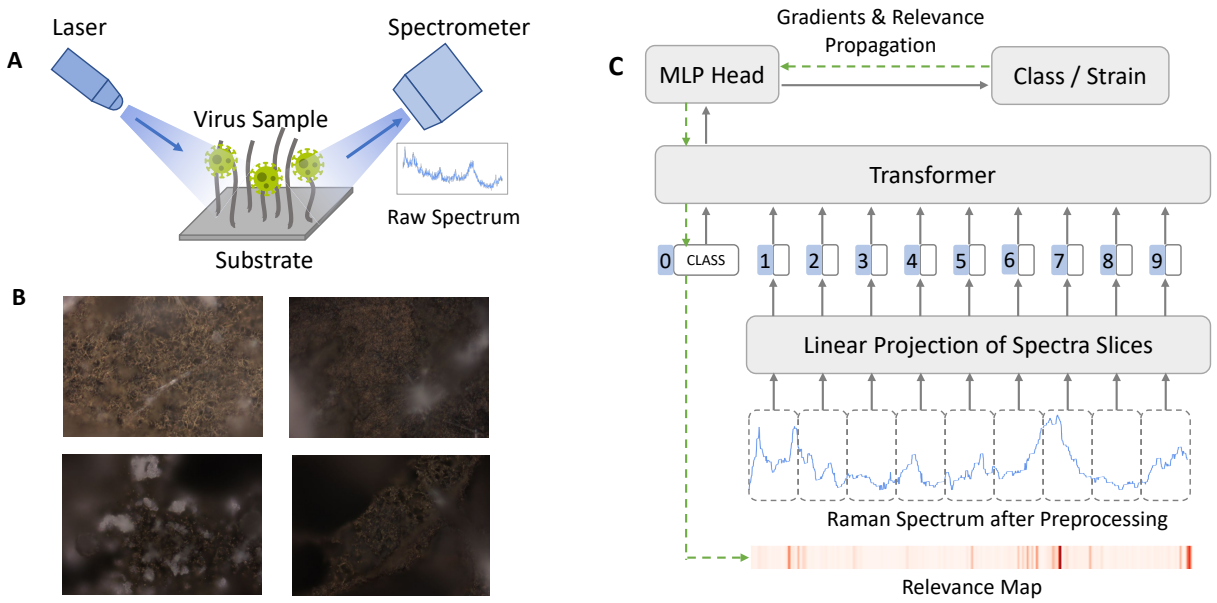

Fig. 1. **(A) Schematic illustration of Raman spectra acquisition from virus samples.** Raman spectra are obtained from virus samples enriched in gold-nanostructure-coated carbon nanotubes. **(B) Example optical images of virus samples.** Each image contains one of the IBV strains. **(C) Overall architecture of our interpretable spectral transformer.** After training, relevance maps can be generated by propagating the gradients of attention maps and relevance with respect to the target strain.

interpretation works.

2.1. Network Architecture

The overall network architecture is shown in Fig. 1. Inspired by Vision Transformer [17], we first divide a 1D spectrum $\mathbf{x} \in \mathbb{R}^L$ into a series of spectrum slices $\mathbf{x} \in \mathbb{R}^{N \times S}$, where L and S are the number of wavenumbers covered in a spectrum and in a spectrum slice respectively, and $N = \lceil L/S \rceil$ is the number of slices. To ensure the spectrum is evenly separated, we use replication padding at the end of the spectrum when necessary. This sequence of spectrum slices is then fed into the linear projection layer to generate slice tokens with D dimensions. After projection, a learnable `[CLASS]` token is prepended to aggregate the information from all other tokens. The order of each spectrum slice is preserved by adding additional position embeddings to the slice embeddings. The resulting vectors are then fed into the conventional Transformer encoder [18] which consists of multi-head self-attention (MSA) blocks, multi-layer perceptron (MLP) blocks, LayerNorm (LN) [19], residual connections [20] and GELU as the activation function. At the end of the network, an MLP head and a softmax activation is appended to the first output vector of the transformer encoder to predict the class label, which, in our case, corresponds to a specific IBV strain.

2.2. Interpretation with Relevance and Gradients

In computer vision, 2D heatmaps are commonly visualized to intuitively interpret the predictions of a classifier. Similarly, we aim to generate 1D heatmaps for spectra to highlight the important wavenumbers that are relevant to the identification of certain IBV strains, as shown in Fig. 1. For our spectral transformer, one straightforward approach is to directly utilize the attention maps to indicate the relevance scores. However, since our transformer contains multiple MSA blocks, simply extracting the attention maps of a certain block can lead to missing contribution of other blocks. Therefore, we employ an improved Layer-wise Relevance Propagation (LRP) method [16] which integrates the weighted attention relevance for each MSA block.

Given a target class t, we compute the element-wise multiplication between the output of the network and the one-hot encoding vector $\mathbb{1}_t$ which indicates the target class. We can then calculate the gradients of each attention map $\mathbf{A}^{(b)} \in \mathbb{R}^{H \times N' \times N'}$ in the network with respect to the target class, where H is the number of heads in an MSA block b, and N' is the number of tokens including the `[CLASS]` token.

For relevance propagation, we use $R_j^{(n)}$ to denote the relevance of the jth input tensor $\mathbf{X}_j^{(n)}$ to the layer n, where $n \in [1 \dots M]$, and layer M represents the first layer. According to [16], the relevance rule can be described as

$$R_j^{(n)} = \sum_i \mathbf{X}_j^{(n)} \cdot \frac{\partial \mathbf{X}_i^{(n-1)}}{\partial \mathbf{X}_j^{(n)}} \cdot \frac{R_i^{(n-1)}}{\mathbf{X}_i^{(n-1)}}. \tag{1}$$

$R^{(0)}$ is set to be the aforementioned one-hot encoding vector $\mathbb{1}_t$ to initialize the relevance propagation. In this way, we can calculate the relevance of the input tensor to the softmax layer in the MSA block b which shares the same dimension with attention maps. We use $R^{(n_b)}$ to denote this relevance.

By aggregating both gradients and relevance, the relevance maps of the input tokens $\mathbf{C}$ are given by

$$\bar{\mathbf{A}}^{(b)} = I + \mathbb{E}(\nabla \mathbf{A}^{(b)} \odot R^{(n_b)})^{+}, \quad (2)$$

$$\mathbf{C} = \bar{\mathbf{A}}^{(1)} \cdot \bar{\mathbf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}^{(B)}, \quad (3)$$

where $\odot$ is the element-wise product, B is the total number of MSA blocks, $\mathbb{E}(\cdot)^{+}$ is the average operation across MSA heads which includes positive values only. To generate the final relevance map that covers the same wavenumber range as the spectrum, we take the relevance map of the `[CLASS]` token and use linear interpolation for upsampling.

3. EXPERIMENTAL RESULTS

In this section, we first describe our dataset including virus sample preparation, the spectra data acquisition process and data preprocessing. We then present the experiment results of our identification and interpretation.

3.1. Dataset Description

Virus sample preparation: Each IBV strain was propagated in specific-pathogen-free (SPF) embryonating chicken eggs (ECE). A stock reference IBV strain or field isolate was prepared at a 1:5 dilution with viral transport medium (VTM) and then was inoculated into 9-to-11-day-old ECE via the chorioallantoic cavity route, 0.2 ml per egg, 3-5 eggs per sample. Chorioallantoic fluid (CAF) was harvested after a 48-72 hrs incubation in a 37 °C egg incubator. The CAF was tested for the presence of IBV by RT-PCR test.

Spectra data acquisition: Raman spectra were recorded by Raman microscopy (Horiba XploRA Plus Raman microscopy) using 532 nm lasers for 30 seconds under 50X magnification with 10 μW laser power. A total of 96,802 Raman spectra across 32 IBV strains were collected in our dataset. The wavenumber of each spectrum ranges from 600 cm^{-1} to 1600 cm^{-1}.

Data preprocessing: We first applied a median filter with window size 9 to remove the spike noise in the spectra. We then adopted asymmetric least squares (ALS) smoothing for baseline correction. Machine learning models may leverage the baseline bias from fluorescence and other sources for virus identification, which is not the expected behavior. Therefore, we used ALS to estimate a polynomial baseline and remove it from the spectra. After baseline correction, the intensity of each spectrum was normalized to [0, 1].

3.2. Implementation Details

Training Details: We selected 80% of the spectra for training and 20% for testing. Stratified sampling was used for splitting so that the percentage of the spectra in each strain was preserved for both the training set and the test set. We applied 5-fold cross validation for hyperparameter tuning. We eventually selected the model with 6 layers of MSA blocks, and each block consists of 6 headers with 64 hidden size and 1024 MLP size. Each spectrum slice size is 5. To train our spectral transformer, we employed the Adam [21] optimizer with momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 1×10^{-4}. The initial learning rate was set to 0.001, and the learning rate was decayed by a factor of 0.2 every time the validation loss stops decreasing after 3 epochs. The minimum learning rate was set to 1×10^{-6}. The size of a mini-batch was set to 512. The spectral transformer was implemented in Pytorch and trained on a single NVIDIA RTX 3090 GPU, and the parameters were randomly initialized.

Evaluation Metrics: We considered four metrics to measure the virus identification performance of each model, including top-1 accuracy, area under the receiver operating characteristic curve (AUC), sensitivity and specificity. For AUC, we used the one-vs-rest configuration. All the metrics were averaged across each strain except for the accuracy which is computed as the ratio between the total number of correct predictions and the total number of samples.

Comparison: We compared our spectral transformer to conventional machine learning algorithms including logistic regression, random forest and XGBoost [22]. We also implemented a six-layer 1D CNN as the deep learning baseline, following prior work [15].

Model	Accuracy	AUC	Sensitivity	Specificity
Logistic Regression	63.37	96.95	61.19	98.80
Random Forest	63.81	97.24	59.80	98.81
XGBoost	63.33	97.49	60.17	98.79
1D-CNN [15]	70.45	98.56	68.64	99.04
Spectral Transformer	**72.72**	**98.83**	**72.34**	**99.11**

Table 1. **Classification performance averaged across the 32 IBV strains (except the overall accuracy).** Our proposed spectral transformer outperforms the conventional machine learning methods in terms of all metrics.

3.3. Strain-level Virus Identification

Table 1 shows the identification performance of different methods on all of the 32 IBV strains. Compared to the conventional machine learning algorithms, our spectral transformer yielded the best results in terms of all four metrics. We further found that for some of the strains, there exist relatively large gaps between the performance of different models.

As illustrated in Fig. 2, the spectral transformer yielded much higher accuracy than the conventional methods on some strains. For instance, for both IBV1 and IBV3, the accuracy scores given by the conventional models were all below 50%, and the accuracy given by the 1D-CNN is also below 60% for IBV1, while the spectral transformer yielded 75.73% accuracy for IBV1 and 62.24% accuracy for IBV3. Even for

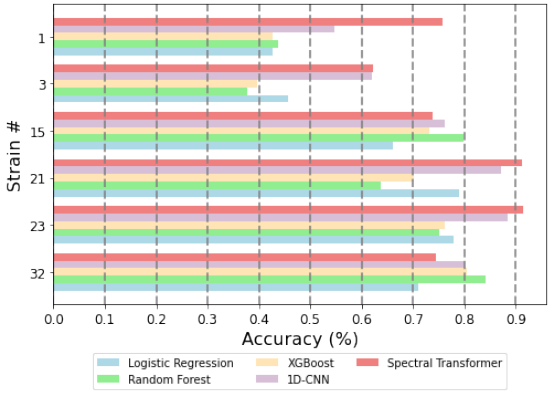

Fig. 2. **Classification accuracy of six selected IBV strains given by different models.** For some strains (e.g. IBV1), the spectral transformer yields much higher accuracy.

strains such as IBV21 and IBV23 where conventional methods achieved higher accuracy, the spectral transformer still outperformed the baselines with 91.23% accuracy for IBV21 and 91.33% accuracy for IBV23. Overall, the spectral transformer yielded the highest accuracy on 26 of 32 IBV strains. We also include the accuracy of IBV15 and IBV32 in the figure as examples of several strains where the baseline methods performed slightly better.

3.4. Interpretation

To further understand which wavenumber the spectral transformer considers more important for classification, we computed the relevance map for each spectra in the training set with respect to the predicted class and generated a mean relevance map for each of the 32 IBV strains as shown in Fig. 3. We observed that there are some wavenumber ranges which have high relevance for almost all the strains, including 650 cm^{-1} to 750 cm^{-1}, 1200 cm^{-1} to 1300 cm^{-1} and 1550 cm^{-1} to 1600 cm^{-1}. These wavenumber ranges have been shown in the literature [23, 24] to correlate with Raman peak ranges of lipids, nucleic acids, and proteins. Differences in these biomolecules are associated to differences in virus surface proteins and lipid bilayers.

We further compared the relevance map of the spectral transformer with the feature importance map of XGBoost generated using SHAP [25]. Fig 4 shows an example of comparison on IBV1. We observed that both methods yield relatively high importance around 1200 cm^{-1}, 1300 cm^{-1} and 1600 cm^{-1}. However, for XGBoost, the feature importance is much higher around 1600 cm^{-1}m, while it almost ignores the wavenumbers below 1100 cm^{-1}. By contrast, the spectral transformer captures some low wavenumber ranges, especially around 700 cm^{-1}. This may explain why it achieved higher accuracy for IBV1 as shown in Fig. 2.

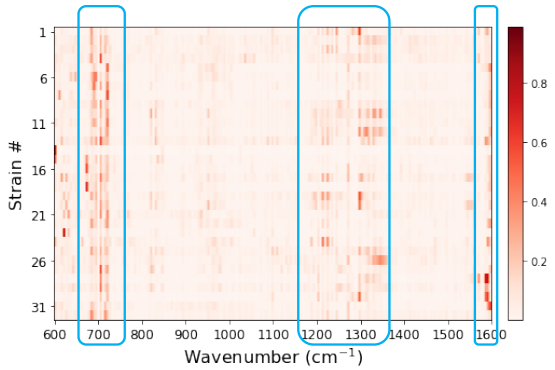

Fig. 3. **Mean relevance map for each IBV strain.** There are some wavenumber ranges that are important for the identification of almost all the strains (highlighted in blue square).

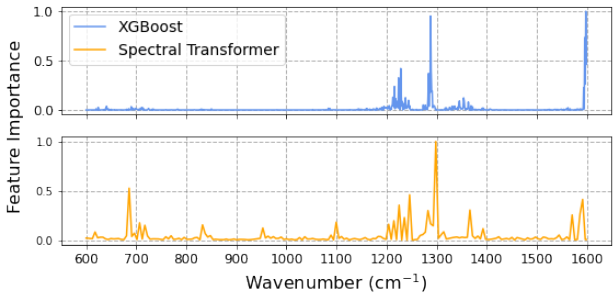

Fig. 4. **Comparison between the feature importance of XGBoost and the spectral transformer on IBV1.**

4. DISCUSSION AND CONCLUSION

In this study, we introduced a spectral transformer to classify the Raman spectra of 32 IBV strains. The relevance maps generated through gradient and relevance propagation were used to facilitate the interpretation of the proposed spectral transformer. We demonstrated through experiments that the spectral transformer outperformed the baseline methods and achieved 72.72% accuracy. We further highlighted the wavenumber ranges picked by the spectral transformer that were more important for the identification of IBV strains.

Despite the success of our proposed spectral transformer in strain-level virus identification, there are two major limitations to our study. First, for each IBV strain in our dataset, the Raman spectra were collected from the same virus sample. Although the spectral transformer yielded relatively accurate classification results on our dataset, it remains unclear if it can achieve similar performance on the Raman spectra collected from other virus samples of the same strain. Future research may include more samples to verify if the model is robust to sample-specific noise. Another limitation is that the current method is not designed to identify the virus strains that are not included in our training set which can be a scenario in virus identification. Therefore, transferring the current method to the zero-shot scenario can also be a future research direction.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in accordance with approval by the Pennsylvania State University, the Institutional Animal Care and Use Committee (IACUC) and Institutional Biosafety Committee (IBC).

6. ACKNOWLEDGMENTS

We thank the National Science Foundation (NSF)'s Growing Convergence Research Big Idea (under Grant ECCS-1934977) and NSF's Early-concept Grants for Exploratory Research (under Grant OIA-2030857). Z.W. and S.H. were partially supported by NSF under grant numbers ECCS-2246564 and ECCS-1943895. This work was also supported in part by the Division of Intramural Research of the NIAID/NIH (E.G.), research grant C940000844 from the Pennsylvania Department of Agriculture to H.L. and A.S., award R2022-03 from the Pennsylvania Soybean Board to H.L. and A.S., and USDA NIFA awards PEN04748 to H.L. and PEN04771 to A.S.

7. REFERENCES

[1] D. Malvy, A. K. McElroy, H. de Clerck, S. Günther, and J. van Griensven, "Ebola virus disease," *The Lancet*, vol. 393, no. 10174, pp. 936–948, 2019.

[2] T. Singhal, "A review of coronavirus disease-2019 (covid-19)," *Indian J. Pediatr.*, vol. 87, no. 4, pp. 281–286, 2020.

[3] F. S. Minhaj, Y. P. Ogale, F. Whitehill, J. Schultz, M. Foote, W. Davidson, C. M. Hughes, K. Wilkins, L. Bachmann, R. Chatelain, et al., "Monkeypox outbreak—nine states, may 2022," *Morbidity and Mortality Weekly Report*, vol. 71, no. 23, pp. 764, 2022.

[4] World Health Organization et al., *Infection prevention and control of epidemic-and pandemic-prone acute respiratory infections in health care*, WHO, 2014.

[5] Z. Wang, J. Ye, K. Zhang, L. Ding, T. Granzier-Nakajima, J. C. Ranasinghe, Y. Xue, S. Sharma, I. Biase, M. Terrones, et al., "Rapid biomarker screening of alzheimer's disease by interpretable machine learning and graphene-assisted raman spectroscopy," *ACS nano*, vol. 16, no. 4, pp. 6426–6436, 2022.

[6] C. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. AE Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[7] S. Cui, S. Zhang, and S. Yue, "Raman spectroscopy and imaging for cancer diagnosis," *J Healthc Eng*, vol. 2018, 2018.

[8] S. Khan, R. Ullah, A. Khan, R. Ashraf, Hina Ali, M. Bilal, and M. Saleem, "Analysis of hepatitis b virus infection in blood sera using raman spectroscopy and machine learning," *Photodiagnosis and photodynamic therapy*, vol. 23, pp. 89–93, 2018.

[9] D. Zhang, X. Zhang, R. Ma, S. Deng, X. Wang, X. Wang, X. Zhang, X. Huang, Y. Liu, G. Li, et al., "Ultra-fast and onsite interrogation of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) in waters via surface enhanced raman scattering (sers)," *Water research*, vol. 200, pp. 117243, 2021.

[10] Y. Yeh, K. Gulino, Y. Zhang, A. Sabestien, T. Chou, B. Zhou, Z. Lin, I. Albert, H. Lu, V. Swaminathan, et al., "A rapid and label-free platform for virus capture and identification from clinical samples," *PNAS*, vol. 117, no. 2, pp. 895–901, 2020.

[11] D. Paria, K. S. Kwok, P. Raj, P. Zheng, D. H. Gracias, and I. Barman, "Label-free spectroscopic sars-cov-2 detection on versatile nanoimprinted substrates," *Nano letters*, vol. 22, no. 9, pp. 3620–3627, 2022.

[12] Z. Zhang, D. Li, X. Wang, Y. Wang, J. Lin, S. Jiang, Z. Wu, Y. He, X. Gao, Z. Zhu, et al., "Rapid detection of viruses: based on silver nanoparticles modified with bromine ions and acetonitrile," *J. Chem. Eng.*, vol. 438, pp. 135589, 2022.

[13] Y. Yang, B. Xu, J. Murray, J. Haverstick, X. Chen, R. A. Tripp, and Y. Zhao, "Rapid and quantitative detection of respiratory viruses using surface-enhanced raman spectroscopy and machine learning," *Biosens. Bioelectron.*, vol. 217, pp. 114721, 2022.

[14] J. Ye, Y. Yeh, Y. Xue, Z. Wang, N. Zhang, H. Liu, K. Zhang, R. Ricker, Z. Yu, A. Roder, et al., "Accurate virus identification with interpretable raman signatures by machine learning," *PNAS*, vol. 119, no. 23, pp. e2118836119, 2022.

[15] D. Ma, L. Shang, J. Tang, Y. Bao, J. Fu, and J. Yin, "Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network," *Spectrochim. Acta A*, vol. 256, pp. 119732, 2021.

[16] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *CVPR*, 2021, pp. 782–791.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[19] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *stat*, vol. 1050, pp. 21, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[22] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[23] Krzysztof Czamara, Katarzyna Majzner, Marta Z Pacia, K Kochan, Agnieszka Kaczor, and M Baranska, "Raman spectroscopy of lipids: a review," *Journal of Raman Spectroscopy*, vol. 46, no. 1, pp. 4–20, 2015.

[24] Daniel Němeček and George J Thomas Jr, "Raman spectroscopy of viruses and viral proteins," in *Frontiers of Molecular Spectroscopy*, pp. 553–595. Elsevier, 2009.

[25] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.