Deep Learning for Explainable Cancer Identification Spectroscopy

Thomas Hartigan

2025

Abstract

This is some stuff

1 Introduction

Cancer is one of the leading causes of death worldwide with around 20 million cases reported in 2022. This corresponds to one in five men or women developing cancer in their lifetime Bray et al. [2024]. The conventional method for cancer diagnosis, as outlined in Figure 1, involves taking a biopsy of suspect tissue from a patient during surgery, fixing this biopsy to a slide, and then staining the biopsy with haematoxylin and eosin (H&E). The haematoxylin stains nucleic acids a deep blue-purple colour, whilst the eosin is pink and non-specifically stains proteins Fischer et al. [2008]. These H&E stained samples can then be analysed by a pathologist to develop a diagnosis. This process is manual, error-prone, subjective, costly, and time-consuming; often requiring transport to a laboratory for processing by highly trained technicians Hollon et al. [2020]. The sample processing time alone is often 20 minutes or more Novis and Zarbo [1997], so the development of methods which forgo traditional lab-based preparation have significant potential to guide surgeries more effectively.

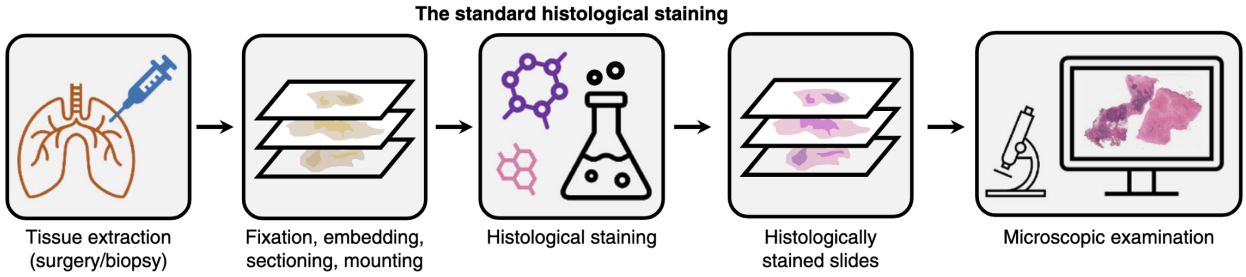
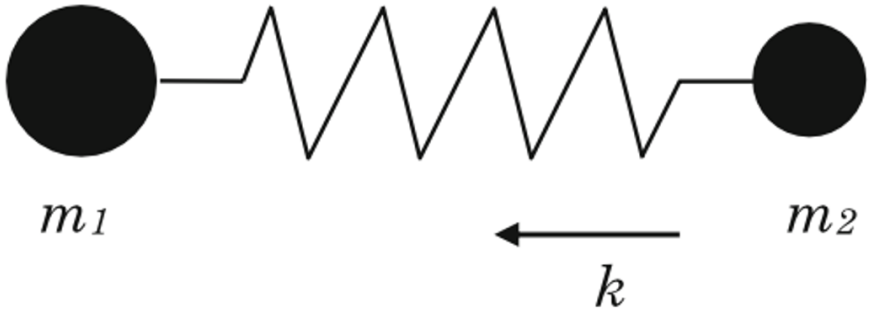


Figure 1: From Bai et al. [2023]

New methods are being developed to expedite the cancer diagnosis process by using machine learning in conjunction with a range of imaging technologies including stimulated Raman spectroscopy (SRS) Hollon et al. [2020], Sarri et al. [2019], Kondepudi et al. [2024], Jiang et al. [2022], second harmonic generation microscopy (SHG) Sarri et al. [2019], and Fourier transform infrared spectroscopy (FTIR) Tomas et al. [2022], Berisha et al. [2019]. The This project will focus on .....



Effect of chemical bonds :

C≡O	C=O	C—O
~ 2100 cm^{-1}	~ 1680-1630 cm^{-1}	~ 1200 cm^{-1}

Effect of H-bonding :

C=O	C=O·····H
~ 1680-1630 cm^{-1}	$\Delta\nu$ ~ - (10-20) cm^{-1}

Effect of mass :

^{13}C=O	^{13}C =^{18}O
$\Delta\nu$ ~ - (35-55) cm^{-1}	$\Delta\nu$ ~ - 65 cm^{-1}

Figure 2: From Berthomieu and Hienerwadel [2009]

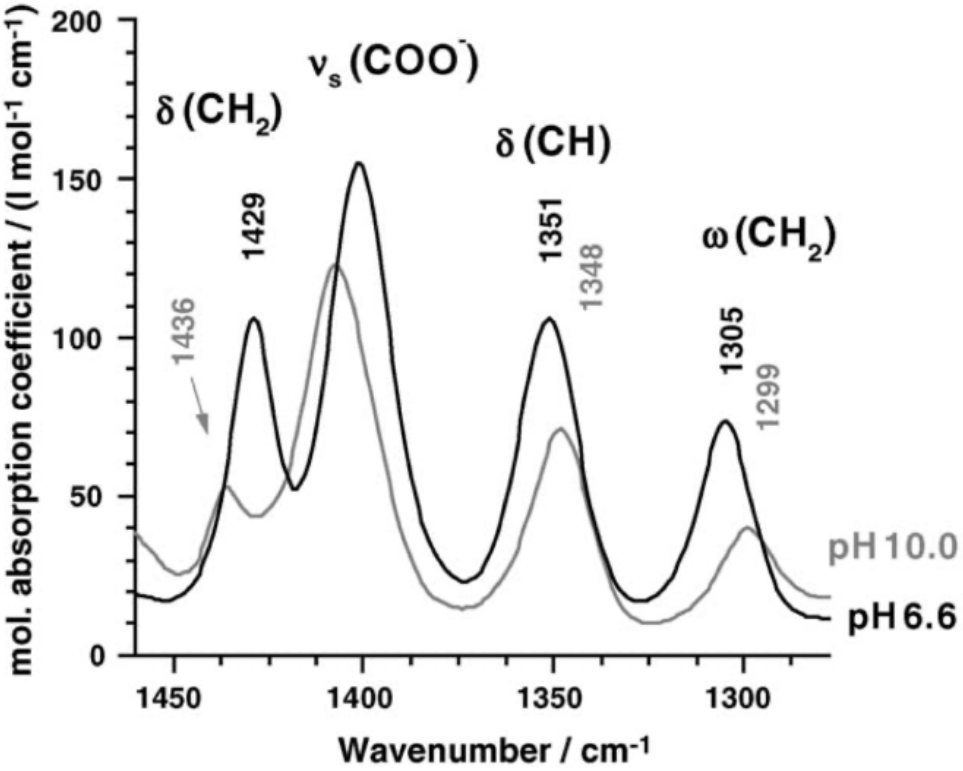


Figure 3: From Wolpert and Hellwig [2006]

1.1 Future Vision

1.2 Novelty of This Work

2 Methods and Results

2.1 Spectroscopy

2.1.1 IR Spectroscopy

2.1.2 Raman Spectroscopy

2.1.3 TPEF

2.1.4 SHG

2.2 Classical ML Methods for Spectral Classification

Table 1: Sensitivity and specificity (mean ± SD) for each model.

Model	Metric	Normal	Hyperplastic	Adenoma	Cancer	Accuracy
PCA > LDA	Sensitivity	42 ± 20.4	31 ± 17.7	84 ± 14.3	39 ± 26.3	61 ± 7.6
	Specificity	90 ± 11.1	94 ± 5.4	57 ± 19.9	95 ± 6.2	—
XGBoost	Sensitivity	58 ± 11.9	41 ± 14.7	80 ± 14.4	62 ± 33.6	68 ± 10.0
	Specificity	88 ± 6.6	91 ± 2.6	77 ± 10.2	96 ± 6.2	—
PCA > XGBoost	Sensitivity	43 ± 11.5	28 ± 11.9	55 ± 11.9	75 ± 20.9	51 ± 9.2
	Specificity	88 ± 10.3	90 ± 6.7	79 ± 7.7	77 ± 10.5	—

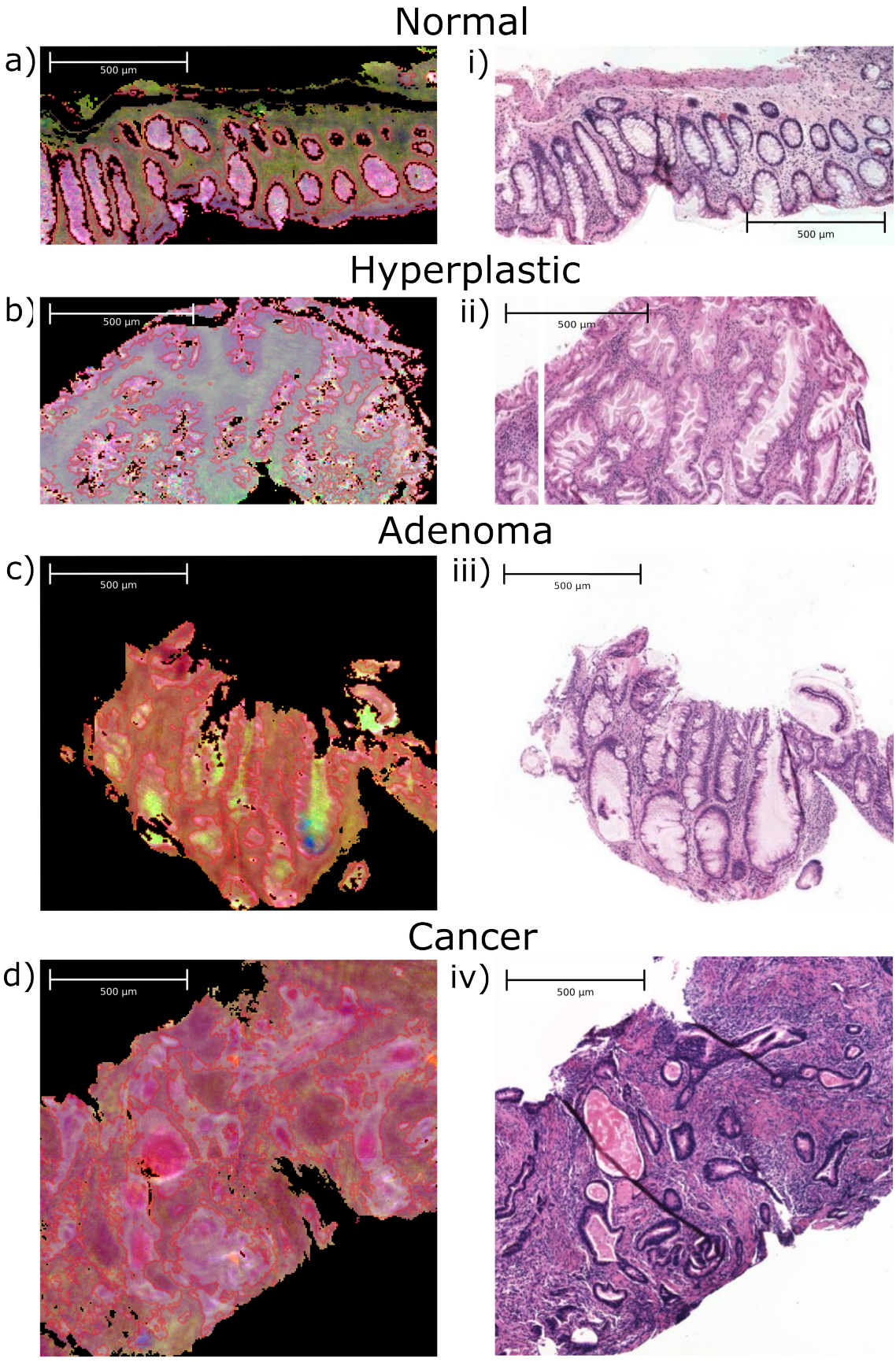


Figure 4: Caption

2.2.1 XGBoost

2.2.2 Spectral Matching

2.3 Deep Learning Methods for Spectral Classification

2.3.1 Dataset Challenges

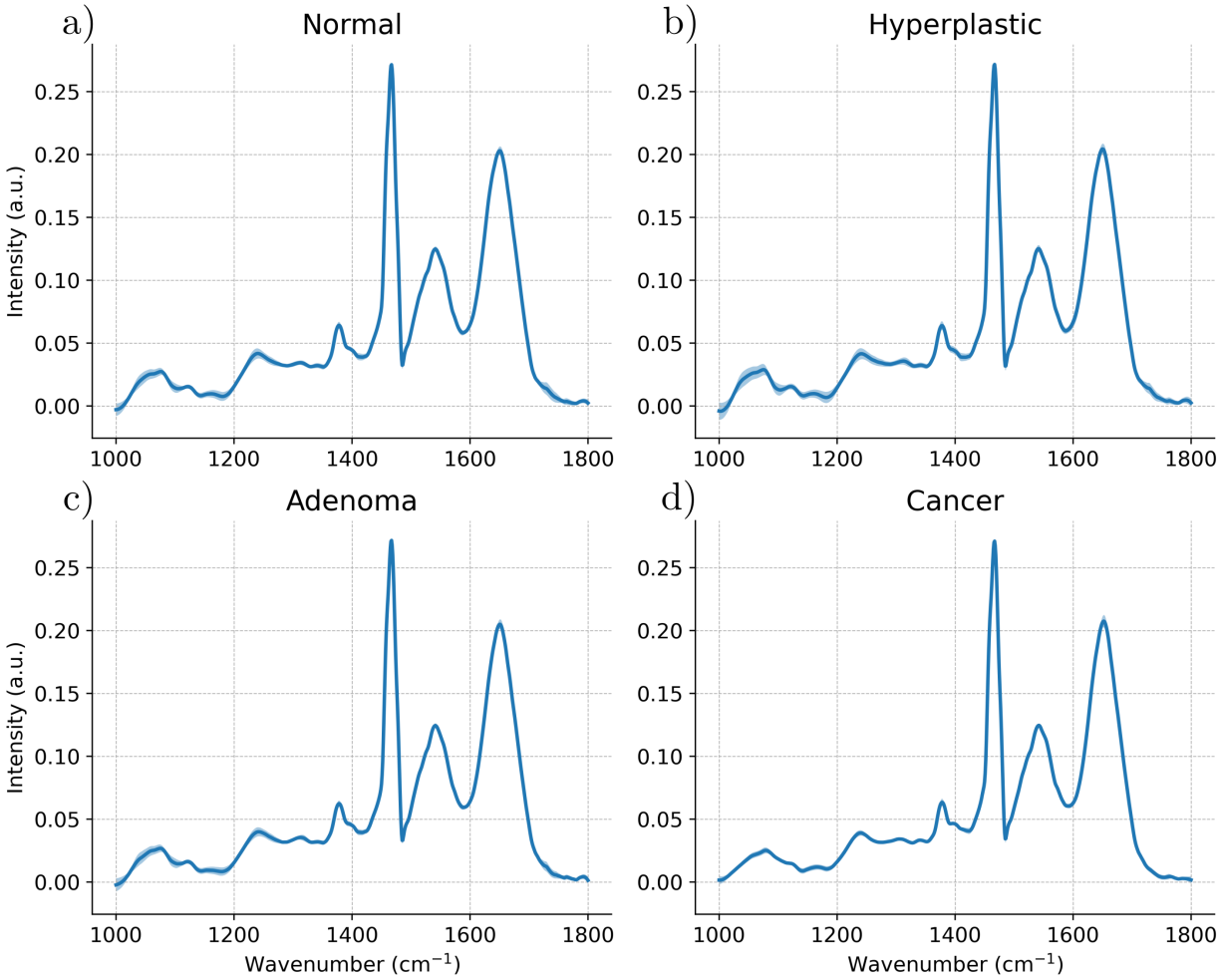


Figure 5: Caption

2.3.2 De-noising Methods

2.3.3 Evaluation Metrics

2.3.4 Model Architecture

2.3.5 Regularisation Techniques

2.3.6 Model Performance

2.4 Explainability

2.4.1 Local Interpretable Model-Agnostic Explanations

2.4.2 CRIME

2.4.3 VAE Methods

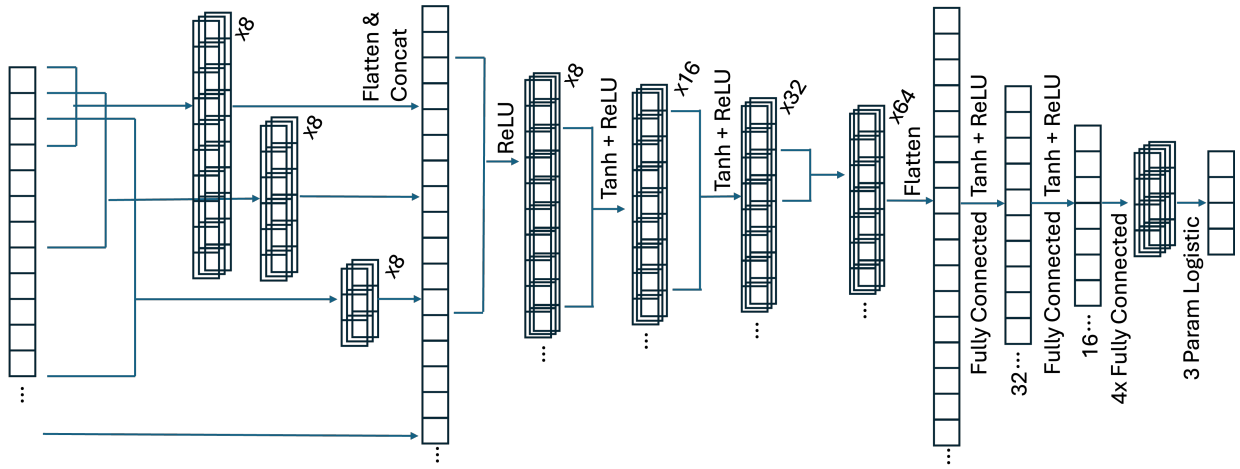


Figure 6: Caption

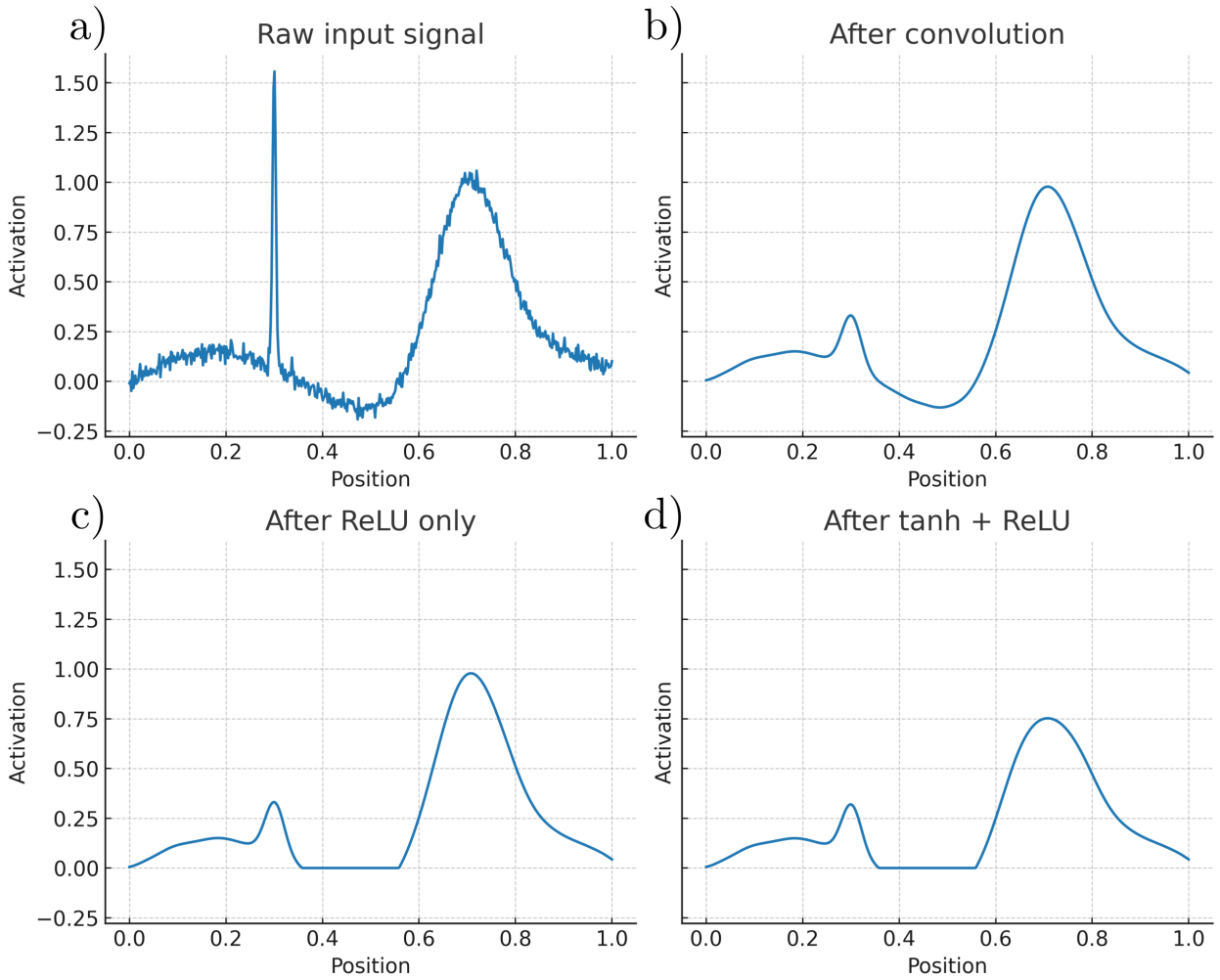


Figure 7: Caption

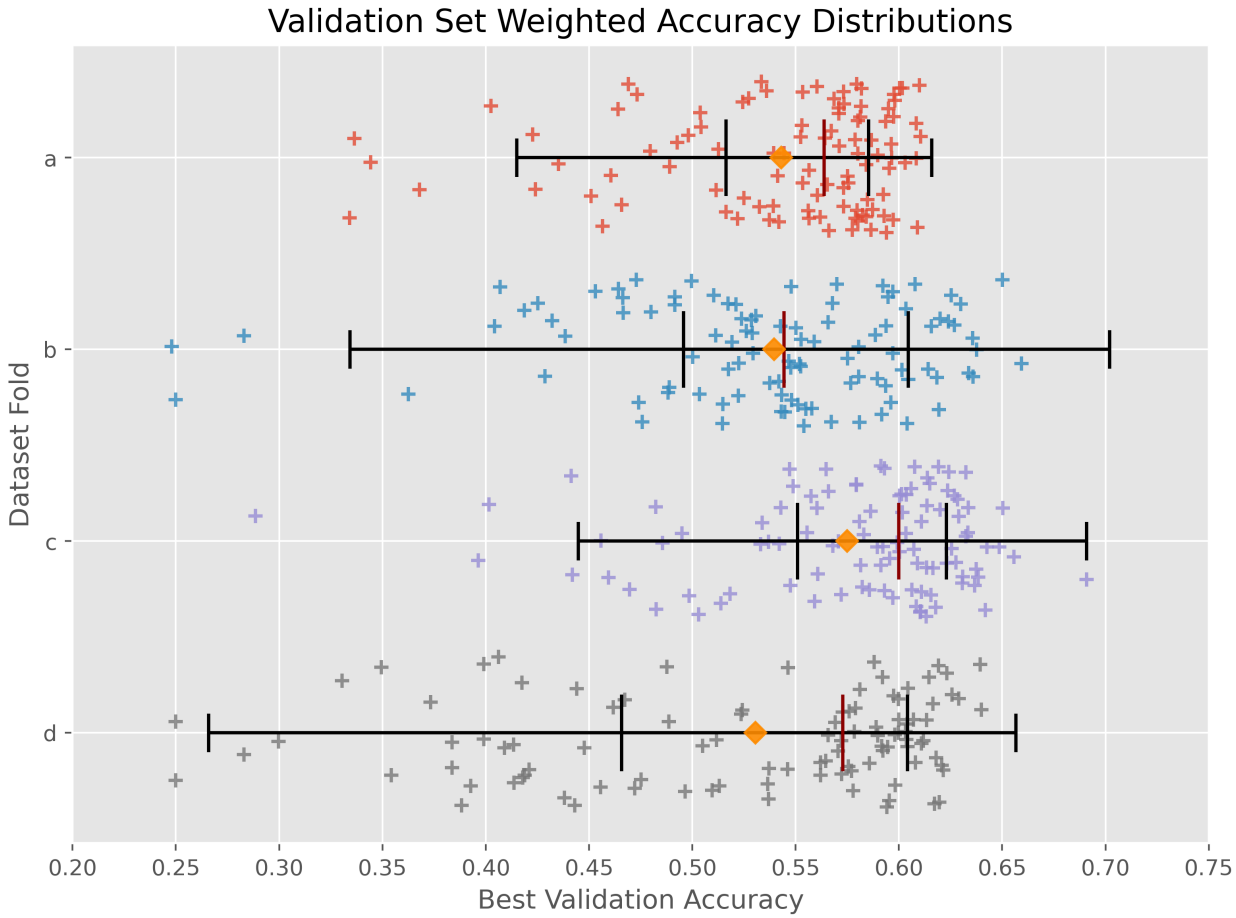


Figure 8: Caption

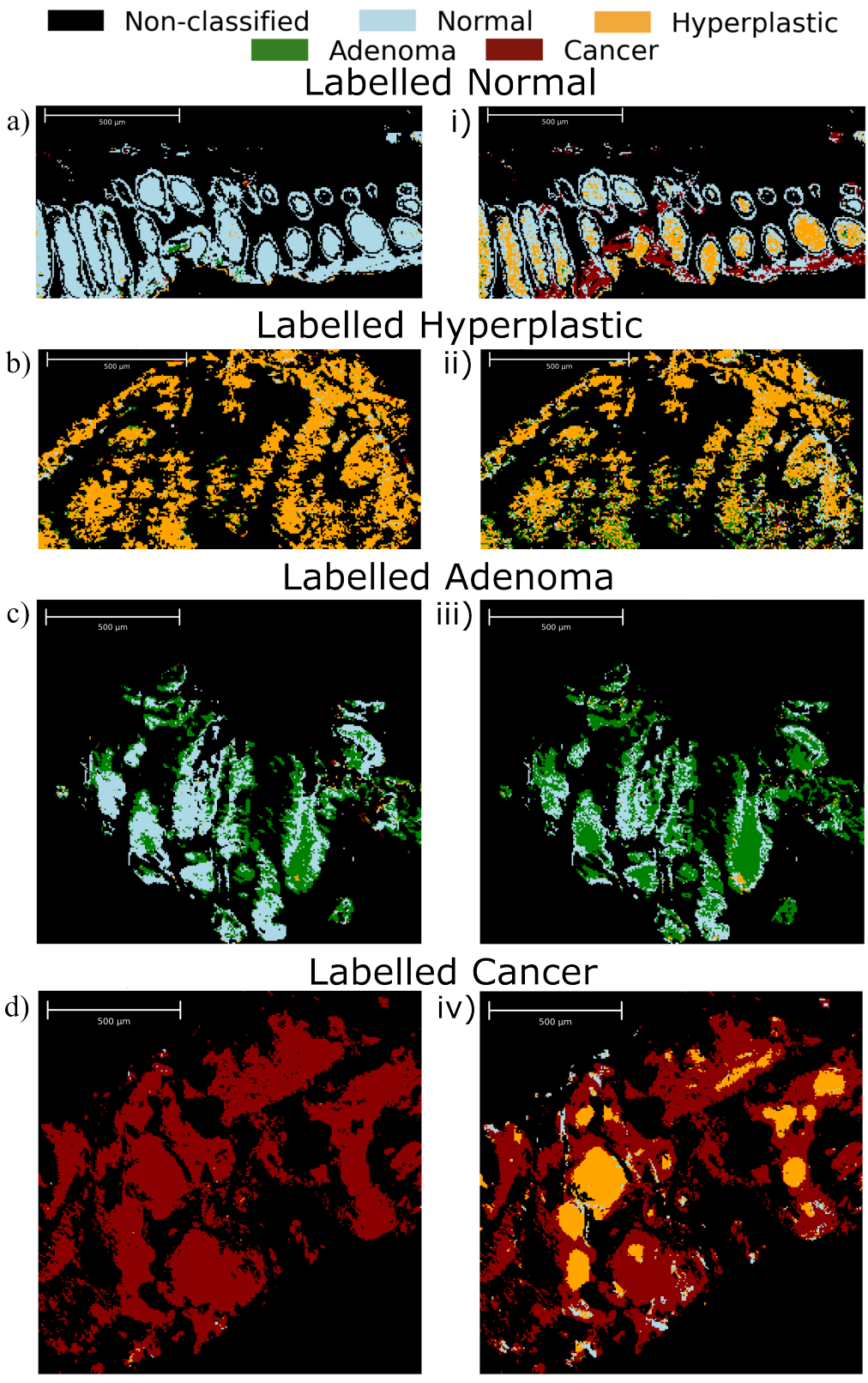


Figure 9: Caption

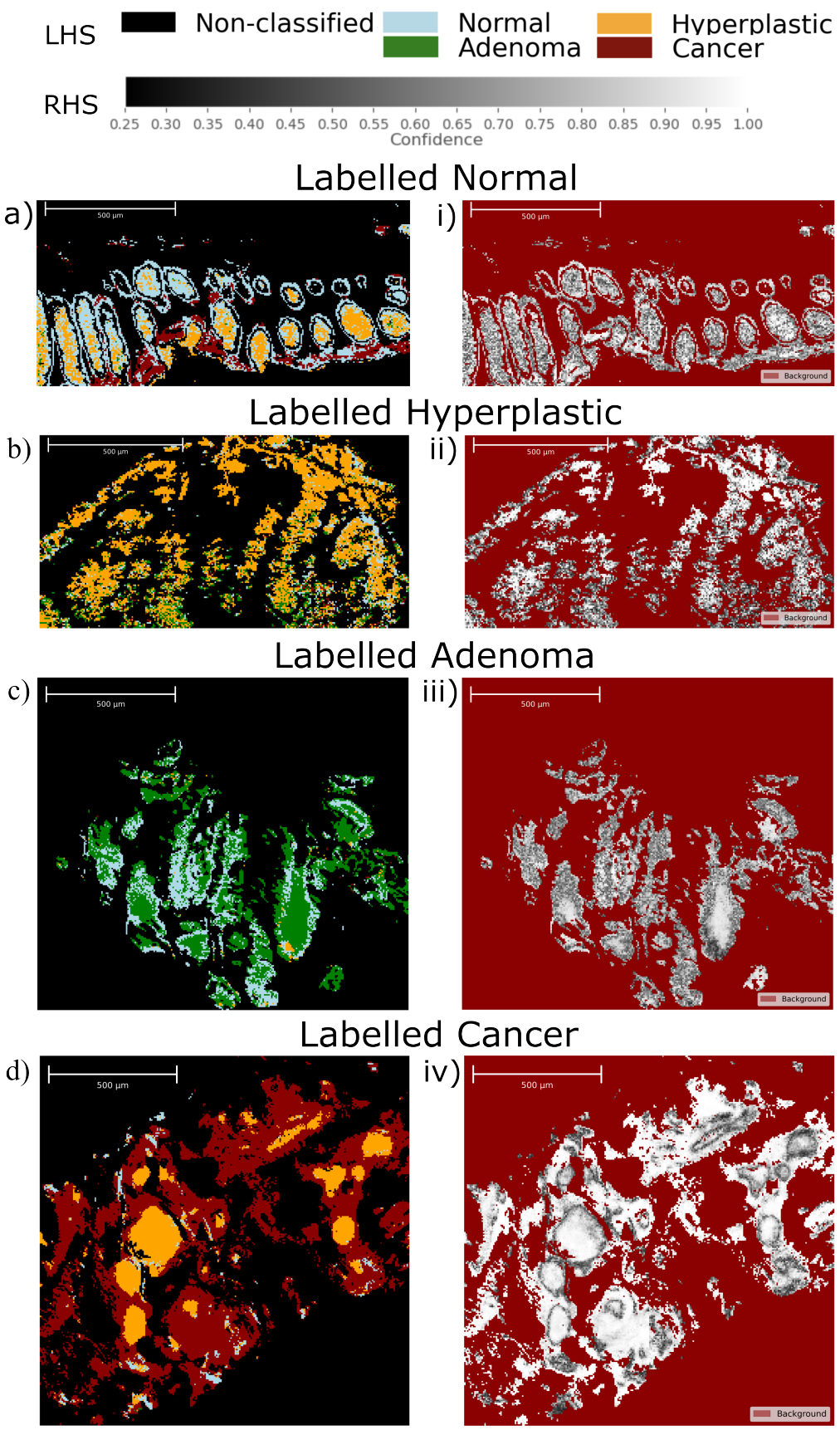


Figure 10: Caption

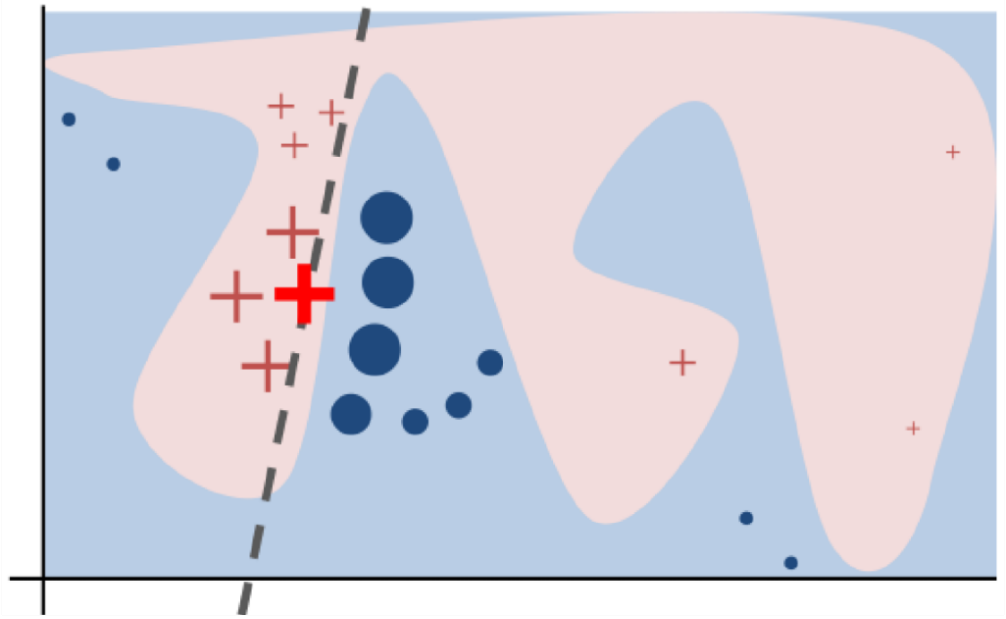


Figure 11: From Ribeiro et al. [2016]

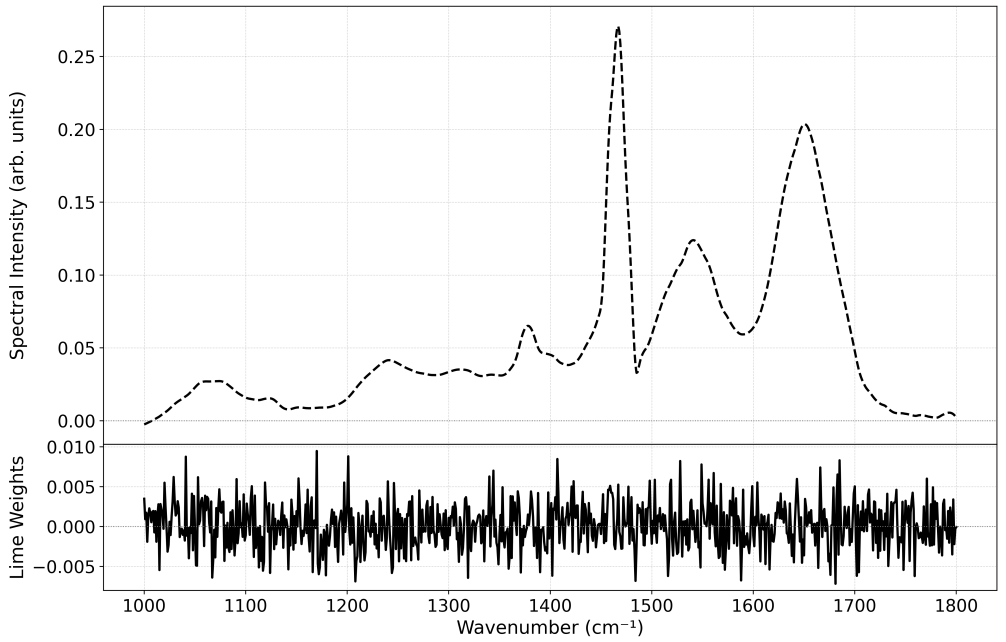


Figure 12: Caption

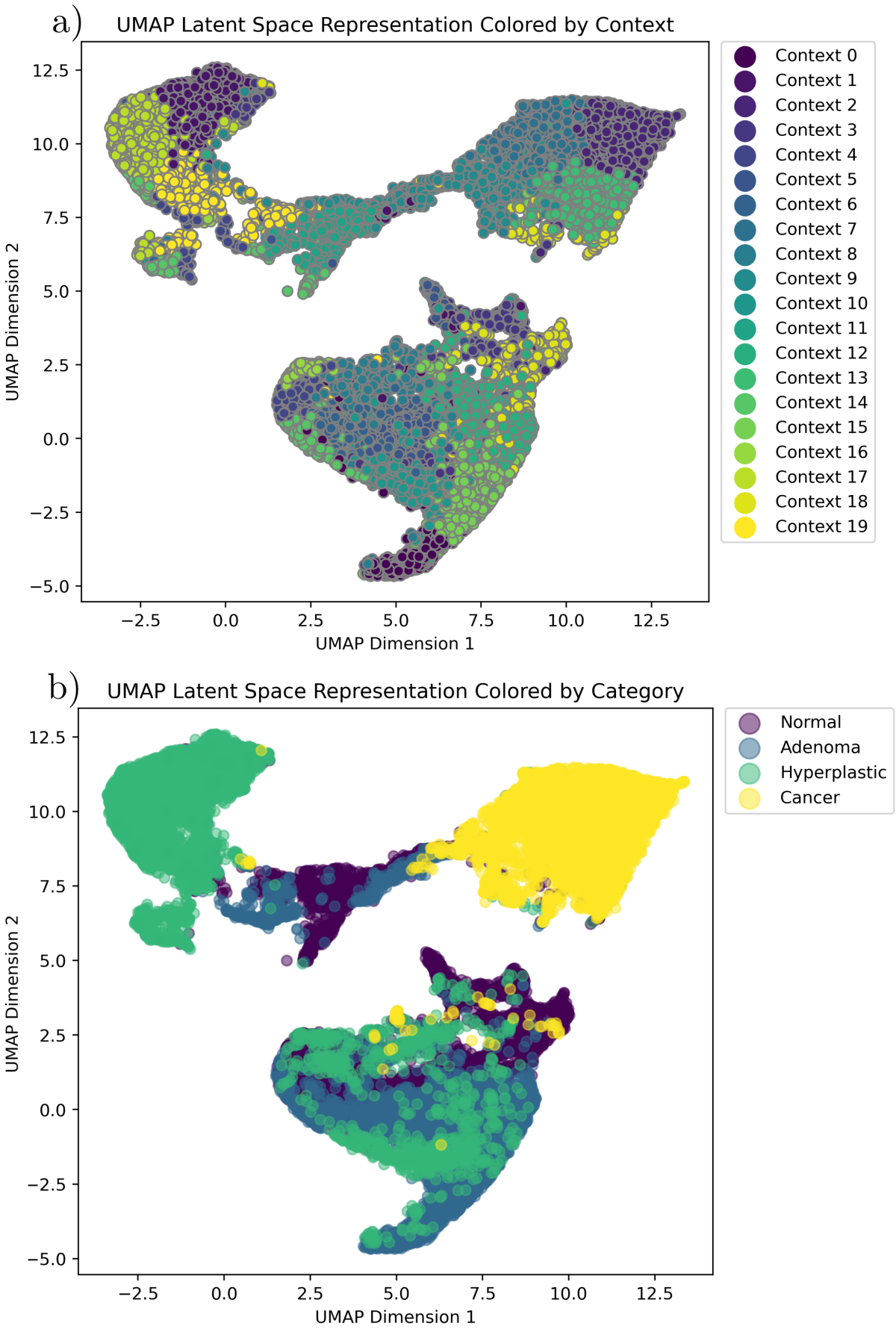


Figure 13: Caption

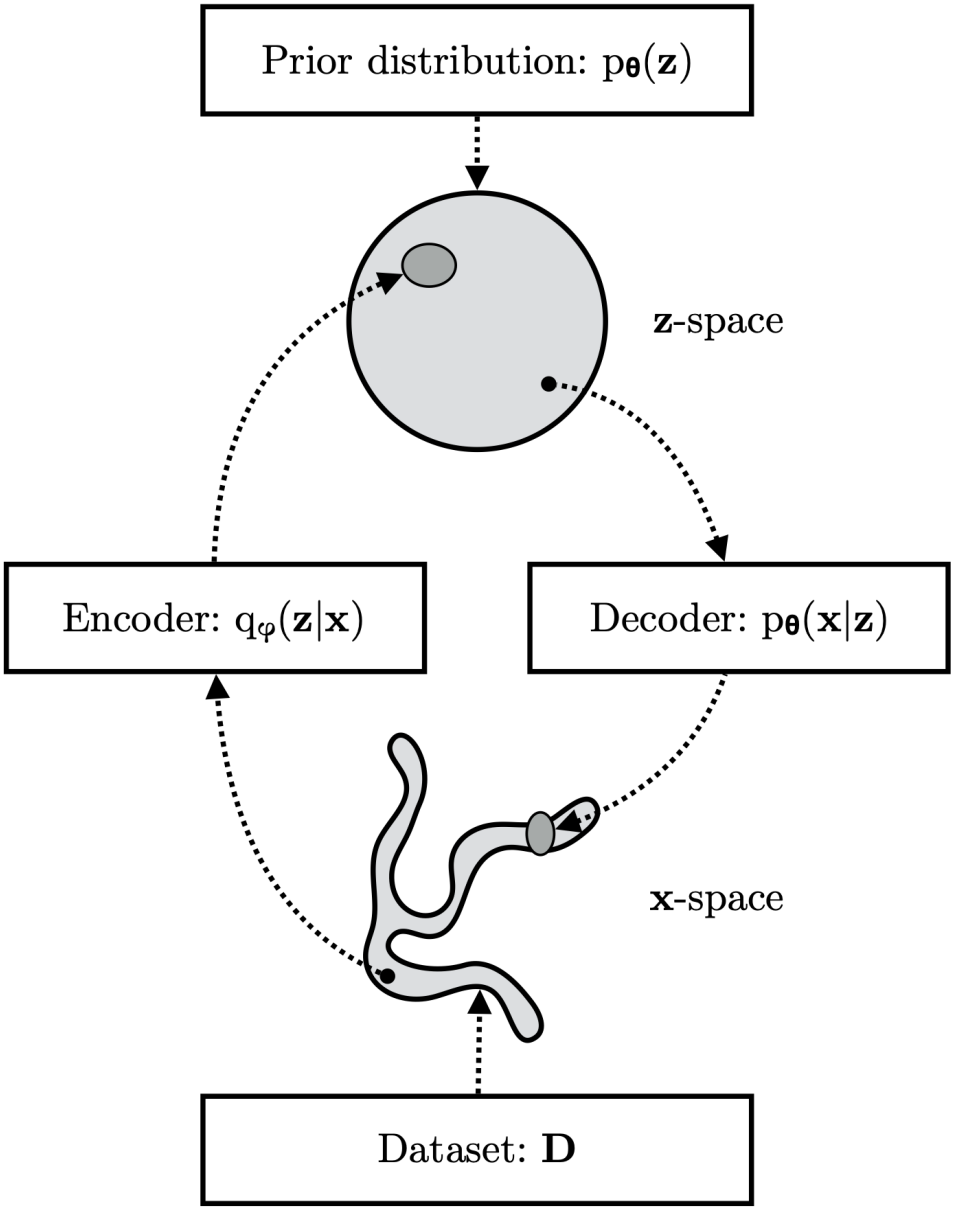


Figure 14: From Kingma and Welling

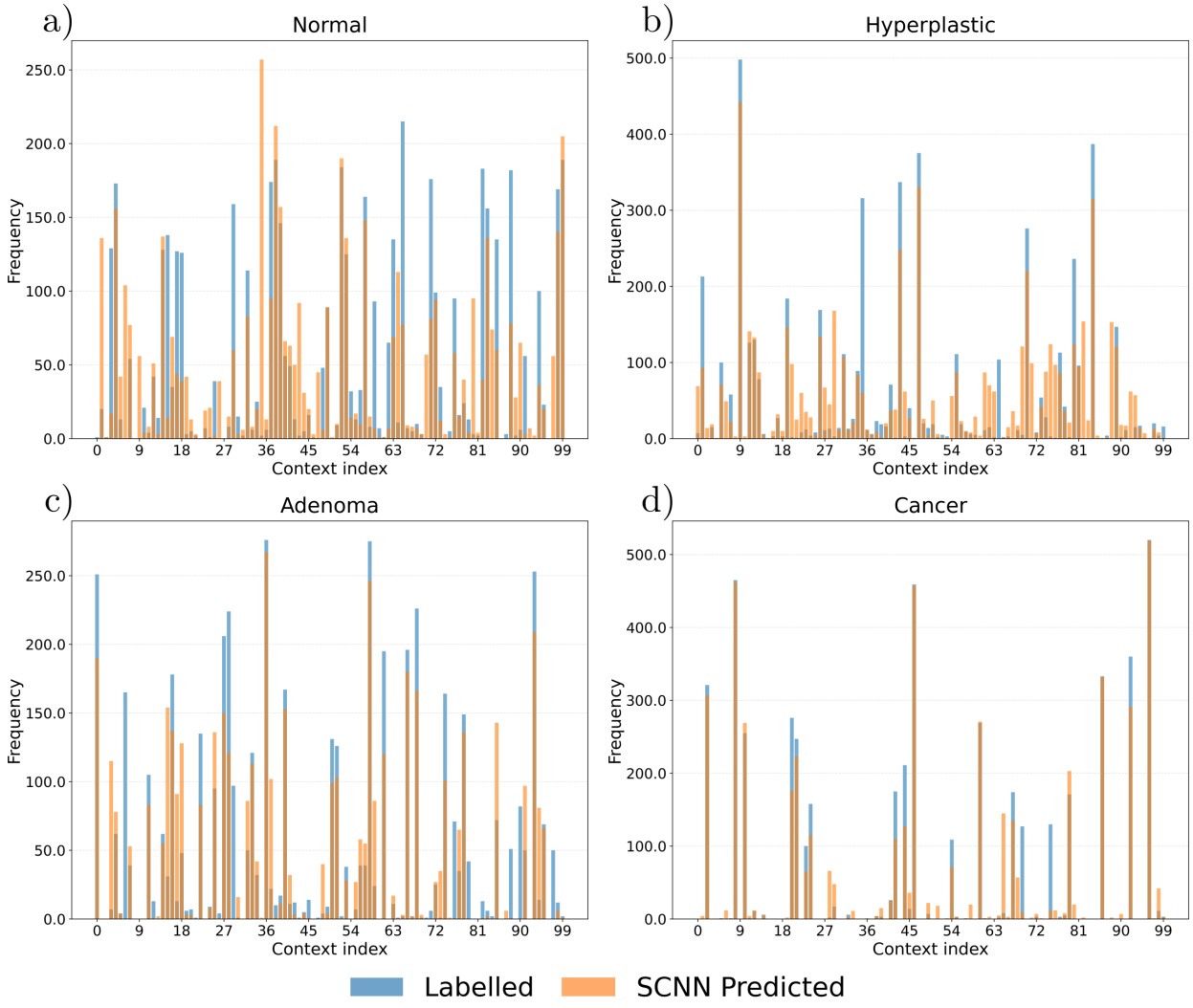


Figure 15: Caption

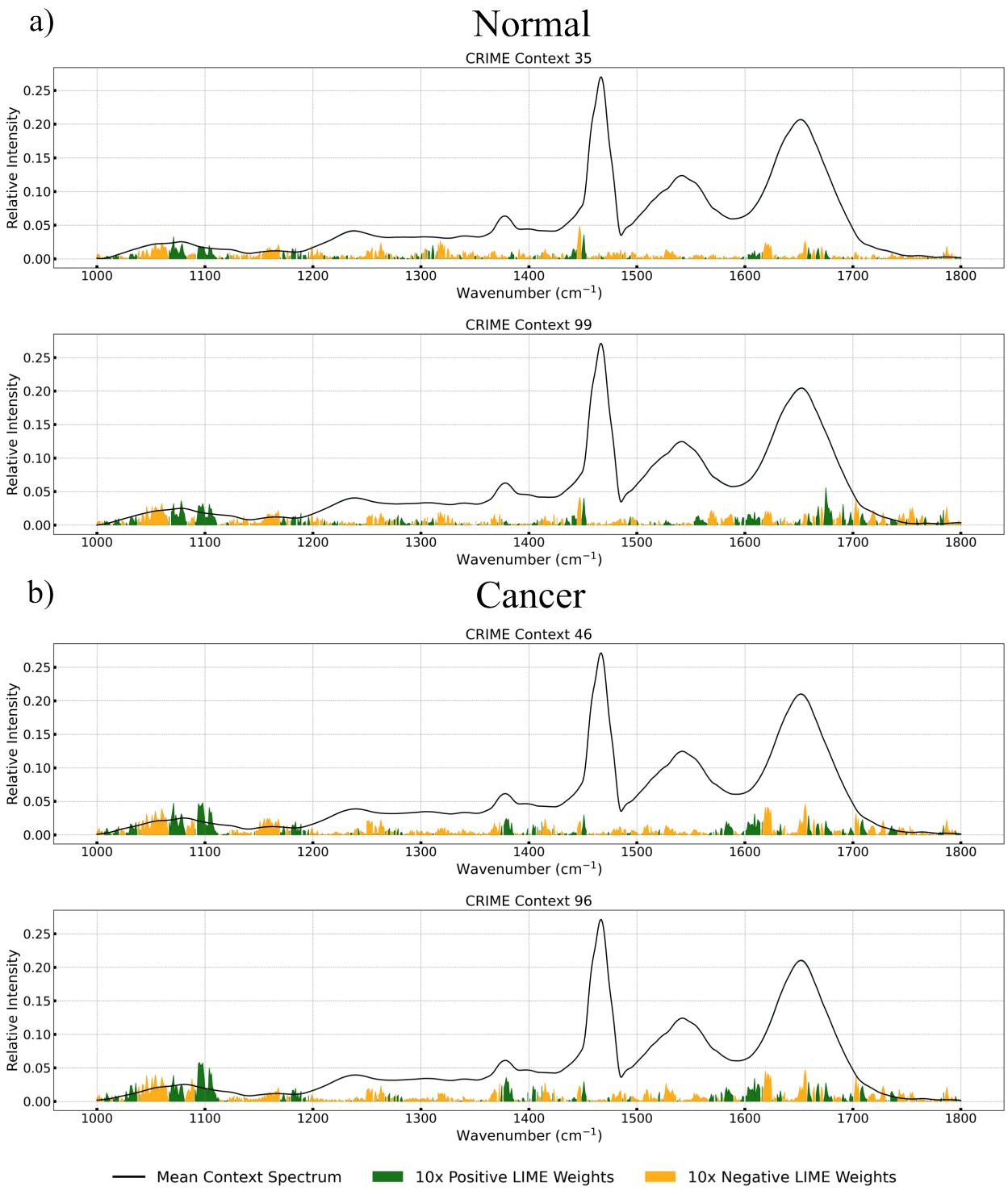


Figure 16: Caption

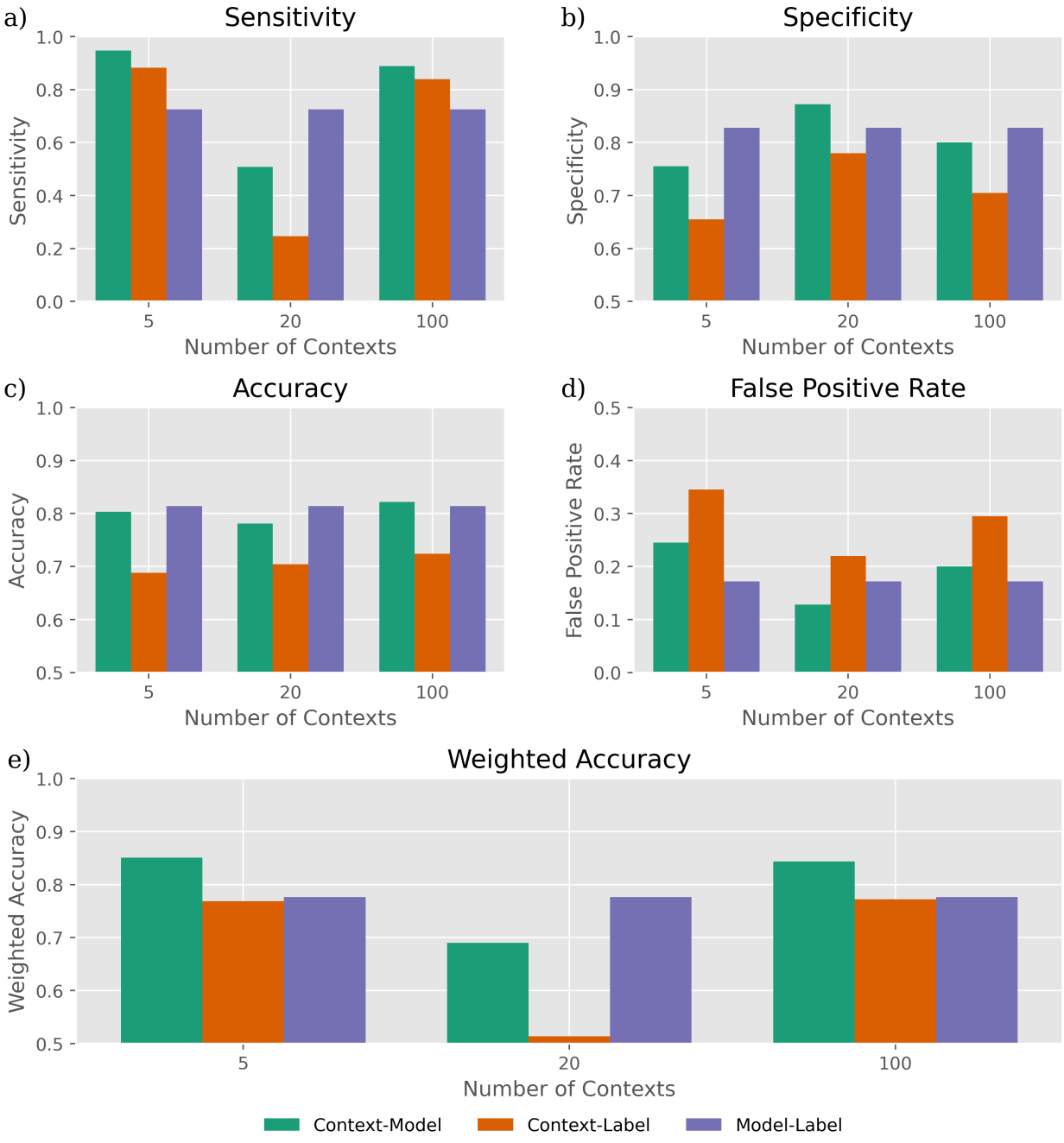


Figure 17: Caption

Table 2: Metrics for single fold test set with 50% voting threshold and no label smoothing

Metrics	Normal	Hyperplastic	Adenoma	Cancer
Accuracy	0.699	0.680	0.699	0.945
Sensitivity	0.662	0.258	0.452	0.783
Specificity	0.708	0.762	0.936	0.972
False Positive Rate	0.292	0.238	0.064	0.028
Overall Weighted Accuracy: 0.5387				

Table 3: Metrics for single fold test set with 50% voting threshold and 10% label smoothing

Metrics	Normal	Hyperplastic	Adenoma	Cancer
Accuracy	0.702	0.674	0.700	0.944
Sensitivity	0.643	0.257	0.458	0.782
Specificity	0.716	0.755	0.933	0.971
False Positive Rate	0.284	0.245	0.067	0.029
Overall Weighted Accuracy: 0.5349				

Clustering VAE Approaches

Logic Explained Networks

3 Results

4 Discussion

5 Conclusions

Table 4: Metrics for single fold test set with 50% voting threshold and 20% label smoothing

Metrics	Normal	Hyperplastic	Adenoma	Cancer
Accuracy	0.700	0.672	0.692	0.941
Sensitivity	0.650	0.266	0.436	0.786
Specificity	0.713	0.751	0.937	0.967
False Positive Rate	0.287	0.249	0.063	0.033
Overall Weighted Accuracy: 0.5345				

Table 5: XGBoost

Metrics	Normal	Hyperplastic	Adenoma	Cancer
Accuracy	0.809 ± 0.026	0.826 ± 0.011	0.749 ± 0.073	0.888 ± 0.055
Sensitivity	0.490 ± 0.115	0.386 ± 0.133	0.781 ± 0.055	0.603 ± 0.091
Specificity	0.886 ± 0.029	0.911 ± 0.036	0.711 ± 0.104	0.937 ± 0.057
False Positive Rate	0.115 ± 0.029	0.090 ± 0.036	0.289 ± 0.104	0.063 ± 0.057
Overall Weighted Accuracy: 0.565 ± 0.058				

6 Introduction

Cancer is one of the leading causes of death worldwide with around 20 million cases reported in 2022 - corresponding to around one in five men or women developing cancer in their lifetime Bray et al. [2024]. The conventional method for cancer diagnosis includes taking a biopsy of suspect tissue from a patient during surgery, and then subsequent pathological analysis. This pathological analysis generally involves staining of the biopsy with haematoxylin and eosin (HE). The haematoxylin stains nucleic acids a deep blue-purple colour, whilst the eosin is pink and non-specifically stains proteins Fischer et al. [2008]. These HE stained samples can then be used by a pathologist to perform diagnosis. This process is manual, error-prone, subjective, costly, and time-consuming; often requiring transport to a laboratory for processing by highly trained technicians Hollon et al. [2020]. The time between pathologists receiving samples and a diagnosis being returned to a surgeon is often around 20 minutes Novis and Zarbo [1997] - any reduction in this processing time would therefore help to guide surgeries more effectively.

New methods are being developed to expedite the cancer diagnosis process by using machine learning in conjunction with a range of imaging technologies including stimulated Raman spectroscopy (SRS) Hollon et al. [2020], Sarri et al. [2019], Kondepudi et al. [2024], Jiang et al. [2022], second harmonic generation microscopy (SHG) Sarri et al. [2019], and Fourier transform infrared spectroscopy (FTIR) Tomas et al. [2022], Berisha et al. [2019]. This project will focus on performing cancer detection and possibly classification using graph neural networks applied to SRS and FTIR spectral data provided by collaborators (see §6.1).

6.1 Collaborators

The Raman imaging section of this project is being undertaken with support from Cambridge Raman Imaging (CRI) - a company founded by Professor Ferrari of the Graphene Centre, University of Cambridge, and Professor Cerullo of Politecnico di Milano. CRI have developed a prototype broadband coherent Raman scattering microscope using patented technology which enables much faster image acquisition than standard coherent Raman spectroscopy at a much lower cost. CRI are coordinating the EU funded Chemometric HistopAthology via coherent Raman imaging for precision Medicine (CHARM) project noa [2022a], which is using the CRI technology to produce SRS spectra of head and neck cancer biopsies. After the SRS spectra are produced, the samples undergo HE staining, enabling a pathologist to then annotate each sample. The annotations, HE images and SRS spectra are then shared using a cloud service.

The FTIR section of this project is being supported by the ulTRafast hOlograPHic FTIR microscopY (TROPHY) project noa [2022b], coordinated by Politecnico di Milano. TROPHY is providing FTIR spectra for samples from 90 patients in the fingerprint region of $1000 - 1800$ cm^{-1} along with four annotation classes:

Table 6: SCNN

Metrics	Normal	Hyperplastic	Adenoma	Cancer
Accuracy	0.756 ± 0.071	0.801 ± 0.083	0.707 ± 0.017	0.881 ± 0.070
Sensitivity	0.739 ± 0.111	0.323 ± 0.069	0.527 ± 0.071	0.777 ± 0.082
Specificity	0.760 ± 0.101	0.894 ± 0.095	0.883 ± 0.089	0.899 ± 0.072
False Positive Rate	0.240 ± 0.102	0.107 ± 0.095	0.118 ± 0.089	0.102 ± 0.072
Overall Weighted Accuracy: 0.591 ± 0.003				

normal, adenoma, hyperplastic, and cancer as described by Nallala et al. Nallala et al. [2016]. This data is also being shared using a cloud service.

6.2 Previous Works

This work is a continuation of two part III projects run last year, one by Moe Vali within the department of physics Vali [2024] and another by Ivo Petrov within the department of Computer Science and Technology Petrov [2024]. These works utilised a previous iteration of the CHARM data which contained SRS spectra at two frequencies, taken from freshly frozen biopsy samples. They then implemented UNET and UNETR techniques both to construct an estimated HE image for each sample, and to perform cancer / non-cancer segmentation, achieving good results in both tasks. The codebase used for these projects, and subsequently updated over summer by Tiago Azevedo, will be used as a starting point for this project.

6.3 This Work

This work will expand on those described in §6.2 by using graph neural networks (GNNs) to extract more relevant information from the data, and to enable more efficient learning of the connections between key features. Furthermore, this project will be utilising a new set of CHARM data which includes spectra from SHG, two-photon excitation microscopy (TPEF) and 31 SRS frequencies with wavenumbers between 2800 cm^{-1} and 3100 cm^{-1}. The hope is that similar GNN methods can be used to analyse both the FTIR data from TROPHY and this new CHARM data.

The use of this new CHARM data also introduces new challenges, as it was taken using samples preserved by Formalin-fixed, Paraffin-embedding (FFPE). This process has led to the introduction of new artifacts in SRS channels when using the CRI machine, so require pre-processing to improve data quality. Some of this pre-processing has been attempted by members of the CHARM collaboration already. However, other methods may need to be examined throughout this project.

The core objectives of this work are to assess the ability for graph representations to encode the information provided in the CHARM and TROPHY datasets; and to investigate the effectiveness of different GNN methods at extracting relevant information from this encoding to contribute to cancer detection ability. Once these objectives are thoroughly explored, it may also be possible to investigate extensions, such as combining the ability of this method with the UNET and UNETR methods explored last year, providing granular classification into normal, adenoma, hyperplastic, and cancer for the TROPHY data, and using the graph information to assist virtual staining tasks. The steps to achieve these objectives are outlined in §8.1 and Figure 22. Furthermore, more details on possible extensions to the project are outlined in §8.3.

7 Background

7.1 Spectroscopy

7.1.1 Raman Spectroscopy

During spontaneous Raman spectroscopy, a sample is irradiated with radiation of a single frequency ω_l, and radiation scattered from the sample is recorded. A change in the energy of the scattered photons, $\hbar\omega_\nu$ then corresponds to the energy difference between the lowest and first excited vibrational levels of the molecules

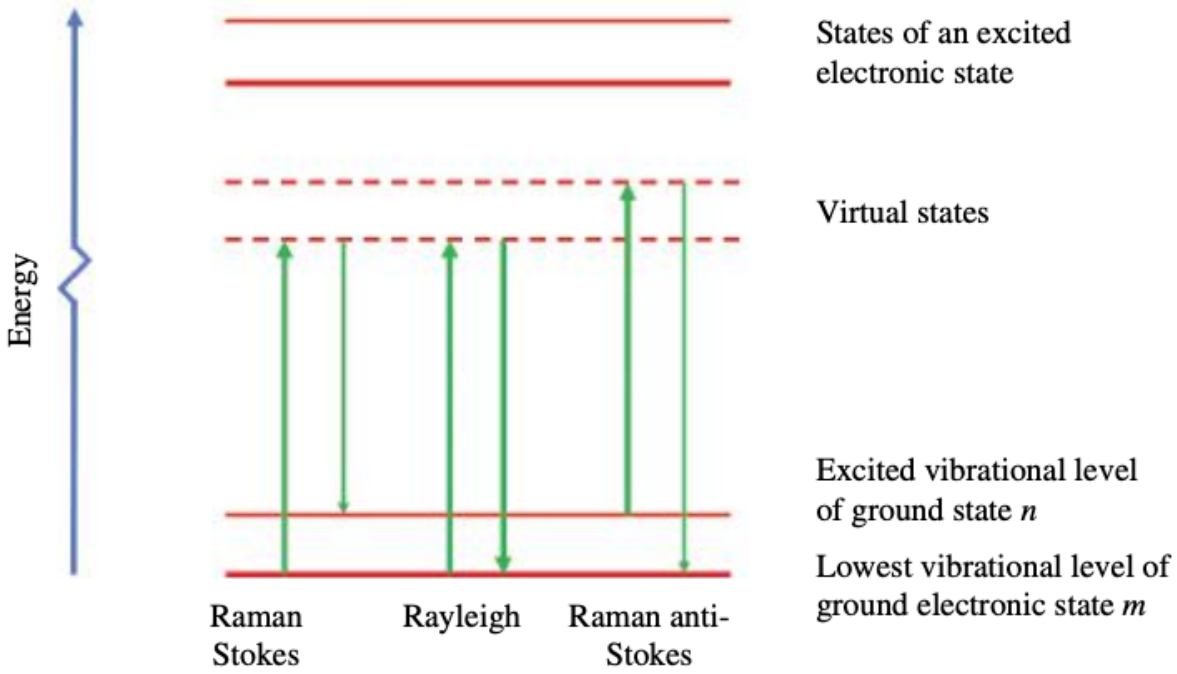


Figure 18: The energy levels involved in Raman Stokes and Raman Anti-Stokes processes, from Smith and Dent [2019].

the light is incident on. For processes where the lowest vibrational level of the ground electronic state m absorb and re-emit a photon, ending up in the first excited vibrational level of the ground state n, the frequency of the re-emitted photon is $\omega_l - \omega_\nu$, corresponding to a Raman Stokes process. Similarly, for a process where the state n absorbs and re-emits a photon, ending up in state m, the re-emitted photon has a frequency $\omega_l + \omega_\nu$, corresponding to a Raman Anti-Stokes process.

The majority of the incident photons undergo Rayleigh scattering. Furthermore, considering the Boltzmann distribution, we expect a majority of systems at room temperature to be in a non-excited vibrational state. This causes Stokes scattering to be more intense than anti-Stokes scattering. This process produces a very faint signal with only around one in 10^7 scattered photons undergoing Raman scattering Smith and Dent [2019]. Spontaneous Raman scattering therefore requires long acquisition times for signal integration, making this process impractical for time-critical implementations or spatial data acquisition.

Using the process above, Raman spectroscopy offers a method for probing changes in the polarisability of molecules, which correspond to changes in the vibrational modes of those molecules. It can therefore only detect molecules where the polarisability tensor α changes between different vibrational modes. It should also be noted that this polarisability is related to the intensity of Raman scattering by equation (1), where K is a constant, l is the laser power, ω is the frequency of incident radiation, and α is the polarisability of the electrons in the molecule. Note that in tensor form, the polarisability is related to the dipole of a molecule μ and the electric field from an incident photon E by $\mu = \alpha E$.

$$I = Kl\alpha^2\omega^4 \tag{1}$$

Raman spectroscopy presents many advantages within biological imaging due to its ability to offer chemical insights into the composition of nucleic acids, proteins and lipids Shen et al. [2021] whilst being less affected by water than infrared methods. It also offers narrow excitation bands, making signals relatively easy to classify, and is capable of producing high-resolution images due to the relatively short wavelengths used Xu et al. [2025].

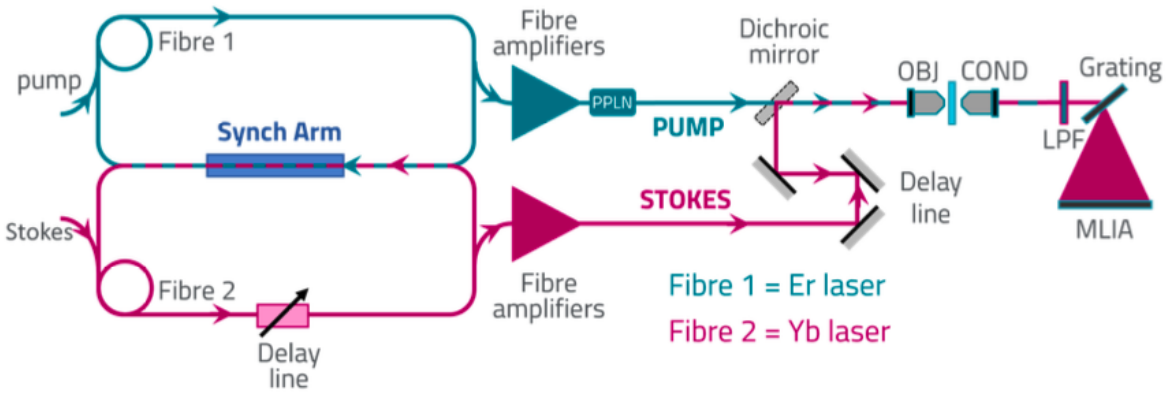


Figure 19: A schematic overview of the CRI SRS implementation. From CRI.

The Raman spectroscopy data for this project has been collected using a patented broadband coherent Raman spectroscopy device, which utilises stimulated Raman spectroscopy (SRS) to pump photons from the excited vibrational level to the ground state, increasing photon emission rates by up to a factor of 10^6, enabling much faster data acquisition at microscopic resolutions Prince et al. [2017]. By using picosecond laser pulses which disperse to have a wide frequency spectrum, it is also possible to conduct these measurements across multiple frequencies simultaneously, further increasing acquisition speed. A basic overview of the CRI system is shown in Figure 19.

7.1.2 Second Harmonic Generation and Two-Photon Excitation Microscopies

The polarisation of a material interacting with light is given by equation (2) where $\chi^{(n)}$ is the nth order non-linear susceptibility tensor and $\mathbf{E}$ is the electric field vector Shen [2003]. The second term in equation (2) corresponds to second harmonic generation (SHG), whereas the third term corresponds to two- and three-photon absorption and SRS processes.

$$\mathbf{P}(\mathbf{k},\omega) = \mathbf{P}^{(1)}(\mathbf{k},\omega) + \mathbf{P}^{(2)}(\mathbf{k},\omega) + \mathbf{P}^{(3)}(\mathbf{k},\omega) + \dots \tag{2}$$

With

$$\begin{aligned}
\mathbf{P}^{(1)}(\mathbf{k},\omega) &= \chi^{(1)}(\mathbf{k},\omega)\cdot\mathbf{E}(\mathbf{k},\omega),\\
\mathbf{P}^{(2)}(\mathbf{k},\omega) &= \chi^{(2)}(\mathbf{k}=\mathbf{k}_i+\mathbf{k}_j,\ \omega=\omega_i+\omega_j)\\
&\quad : \mathbf{E}(\mathbf{k}_i,\omega_i)\mathbf{E}(\mathbf{k}_j,\omega_j),\\
\mathbf{P}^{(3)}(\mathbf{k},\omega) &= \chi^{(3)}(\mathbf{k}=\mathbf{k}_i+\mathbf{k}_j+\mathbf{k}_l,\ \omega=\omega_i+\omega_j+\omega_l)\\
&\quad : \mathbf{E}(\mathbf{k}_i,\omega_i)\mathbf{E}(\mathbf{k}_j,\omega_j)\mathbf{E}(\mathbf{k}_l,\omega_l).
\end{aligned}$$

When incident on a highly polarisable and non-centrosymmetric molecule, two photons can combine upon scattering, causing SHG emission with double the frequency of the incident radiation PERRY et al. [2012]. Notably, the requirement for non-centrosymmetric environments is a consequence of the second-order symmetry of SHG term in equation (2). As a consequence, incidence on centrosymmetric molecules causes the signal to vanish Chen et al. [2012]. This makes SHG especially sensitive to fibrillar collagen in a range of tissues, with changes due to cancerous growths often detectable.

The use of two incident photon frequencies from the pump and Stokes beams combined with femptosecond laser pulses for SRS also enable two-photon excitation microscopy to be conducted, where both incident photons simultaneously induce excitations within the incident molecule, which can then be detected Denk et al. [1990]. This method is important for cancer detection, as it provides increased imaging depth, can elucidate metabolism within living cells, and can also be conducted using lower energy photons, reducing the

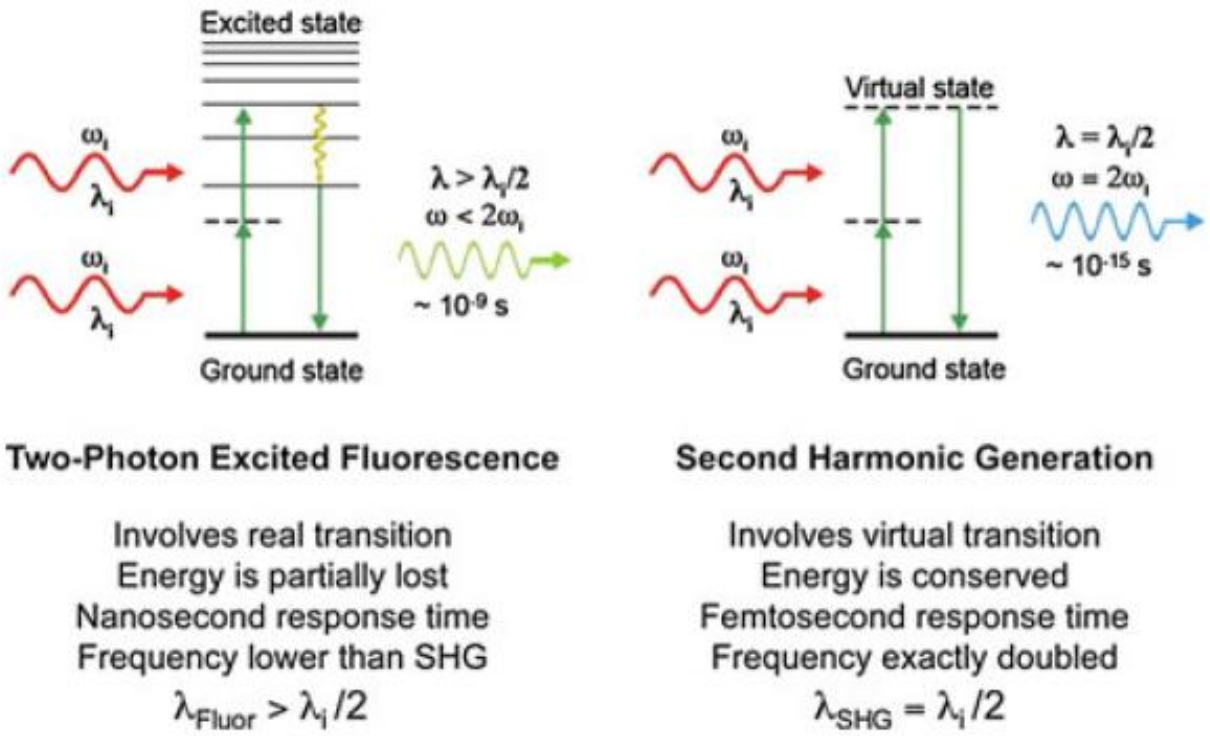


Figure 20: A comparison of the energy processes in TPEF and SHG PERRY et al. [2012].

risk of photo-damage to a sample PERRY et al. [2012]. A summary of the energy processes within TPEF and SGH emissions are summarised in Figure 20.

7.1.3 Fourier Transform Infrared (FTIR) Spectroscopy

Infrared spectroscopy works by directing a broadband spectrum with wavenumbers ranging from around $400 - 4000$ cm^{-1} (for mid-IR spectroscopy) Su and Lee [2020] onto a sample. This coherent spectrum then excites resonant vibrational modes within molecular bonds inside the sample, causing absorption of specific frequencies depending on the composition of the incident sample. This decreases the intensity relative to the baseline in the absorbed frequencies, therefore allowing a characteristic spectrum for each bond to be established. Modern devices use a Fourier transform-based (FTIR) interferogram approach Griffiths [1983], enabling high signal-to-noise (SNR) ratios to be achieved.

FTIR can detect biochemical compositions including nucleic acids, proteins, lipids and carbohydrates within biological samples by matching the observed spectra with known forms for different bonding types, functional groups and intermolecular interactions Su and Lee [2020]. Typical absorption groups within biological samples are detailed in Table 7.

7.2 Graph Neural Networks

Here, we define a graph like in Broadwater [2025] as "data structures with elements expressed as nodes or vertices, and relationships between those elements, expressed as edges". The intrinsic links within a graph are excellent for exploiting relationships between data. Traditionally, they are used in the context of systems with intrinsic connections, such as; the relationships between people on social media, where individuals can be encoded as vertices and the relationships between them encoded by edges; or maps, where street intersections can be stored as vertices and the kind of connections between each intersection (what type of road, speed limit etc.) can be encoded by the edges. By encoding a graph with a list of connected vertices and the weights of each connection, it is immediately seen that the ordering of such a list has no impact. Graph neural networks leverage this permutation invariance to encode and exchange information across a graph

Wavenumber (cm^{-1})	Assignment
3080–2800	Anti-symmetric and symmetric C–H stretches from proteins and lipids
1745–1725	Ester carbonyl of lipids
1700–1500	Amide I and II groups in peptide linkages of proteins
1270–1080	Anti-symmetric and symmetric C–O and P–O areas in DNA, RNA and phospholipids
1200–900	Carbohydrate vibrations of glucose, fructose and glycogen

Table 7: Assignments of Various Functional Groups Based on Wavenumber

structure during the learning process Broadwater [2025].

By including relational information within a graph, models can develop physical intuition Sanchez-Gonzalez et al. [2018, 2020] and harness the topological structures in Biological imaging. It is thought that this approach can offer advantages over vision transformer models by learning context-aware representations between biologically relevant areas of spectra rather than arbitrary patches of these spectra Brussee et al. [2024]. Furthermore, due to the large dataset sizes involved in histopathological imaging, images are usually split into patches for transformer model applications, producing bias. By modelling the whole image as a graph, memory requirements can be significantly reduced, allowing a GNN model to learn using the global structure of an image. Utilising the whole image representation in conjunction with graphs encoding smaller-scale features then allows for hierarchical modelling of each sample Brussee et al. [2024]. Due to the entity-attachment qualities graphs encode, GNNs can also provide advantages in terms of explainability, although this work will not explore explainability in detail.

The data for this project shows distinctive characteristics, such as cell nuclei, so it is thought that by extracting relevant information from each spectra and encoding that information within a graph with sensible relational information, it should be possible to predict which regions are cancerous and which are not. The task of feature extraction, however, is sensitive, as we wish to encode all relevant information from the spectra within the graph. Therefore, multiple methods for feature extraction will be explored, beginning with basic segmentation approaches, tissue graphs and superpixels, as outlined in Brussee et al. [2024] and depicted in Figure 21. To extract the features for each vertex in these graphs, morphological features, CNNs or Vision transformers can be used. There has also been progress recently in using self-supervised learning for such representations, and this could also be investigated Tendle and Hasan [2021]. One particularly interesting method which automates this feature extraction to utilise GNNs is vision graph U-NETs Jiang et al. [2024], which this work will try to explore.

8 Project Status

8.1 Plan

The primary goal of this project is to assess the feasibility of applying GNNs to the CHARM and TROPHY datasets for automated cancer annotation. The steps towards this outcome are outlined in Figure 22. It is hoped that this method can be combined with the UNETR and UNET methods developed by projects in the previous academic year to achieve even stronger performance.

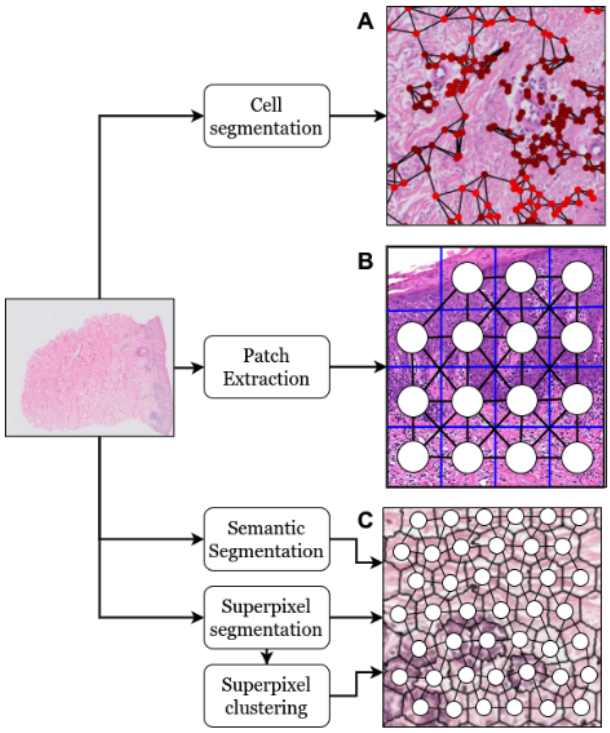


Figure 21: Commonly used graph types in histopathology A) showing a cell graph, B) a patch graph and C) a tissue graph. From Brussee et al. [2024].

Month	October		November				December					January				February				March					April				May				June
Week Beginning	21	28	4	11	18	25	2	9	16	23	30	6	13	20	27	3	10	17	24	3	10	17	24	31	7	14	21	28	5	12	19	26	2
Literature Review																																	
Weekly Meeting With Collaborators																																	
Working Environment Setup																																	
Data Pre-Processing and Validation																																	
Exams																																	
Initial Graph Representation Methods																																	
Initial GNN Implementation																																	
Investigate Vision Graph UNET																																	
Investigate Project Extensions																																	
Writeup																																	
Submission Deadline																																	

Figure 22: Project timeline. Blue highlighting represents preparatory work done by members of the CHARM collaboration.

8.2 Progress

Initial progress with the project has been steady, with a good review of the literature undertaken. It is hoped that by the fourth week of December, all necessary literature to begin on GNN implementation will have been examined. Furthermore, access has been granted to the Kiiara machine owned by the Computational Biology group. This machine contains four Titan XP graphics cards, which should be sufficient for initial implementation and experimentation, but could be limiting when attempting training using the full datasets.

Good progress has also been made in data pre-processing, with members of the CHARM collaboration delivering aligned HE images and pre-processed SRS signals around mid-November. Notably, they used a different pre-processing method for the SRS signal than that detailed in the reports from last year Vali [2024], Petrov [2024]. It has been noticed in data pre-processing for this work that implementing and expanding on the Fourier domain methods used last year gives similar, but notably different results - see Figure 23. This is therefore an area which warrants further consideration. Potential avenues to explore include using the logarithm of intensity data, as they often cover many orders of magnitude, and taking ratios of SRS channel intensities in an attempt to remove focus-induced systematic errors.

There have, however, been some delays in data acquisition for the project. Although CHARM has provided SRS, TPEF, SHG and HE images, they have yet to begin the annotation process. This is due to a range of logistical issues, but they hope annotations can be completed in January. That should give plenty of time for this project. Furthermore, a subset of the TROPHY data seems to contain large processing artifacts not present in the original data. A revised version of this data has been requested, but there should be enough correct data to begin with.

8.3 Extensions

There are many exciting possibilities for extending this project, many of which have been mentioned in the above text, such as self-supervised learning for feature extraction during graph creation. The CHARM project is also trying to generate virtual HE images using the SRS, TPEF and SHG data, so it would be possible to explore whether GNNs can contribute positively to this task.

9 Conclusion

This project is an exciting fusion of state-of-the-art microscopy technology and machine learning applications to help improve the reliability and efficiency of cancer detection. Good progress has been made throughout Michaelmas term towards the objectives set out in the project timeline, and the project seems set on a good trajectory to begin GNN experimentation in the new year.

References

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. ISSN 1542-4863. doi: 10.3322/caac.21834.

Andrew H. Fischer, Kenneth A. Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *CSH protocols*, 2008:pdb.prot4986, May 2008. doi: 10.1101/pdb.prot4986.

Todd C. Hollon, Balaji Pandian, Arjun R. Adapa, Esteban Urias, Akshay V. Save, Siri Sahib S. Khalsa, Daniel G. Eichberg, Randy S. D'Amico, Zia U. Farooq, Spencer Lewis, Petros D. Petridis, Tamara Marie, Ashish H. Shah, Hugh J. L. Garton, Cormac O. Maher, Jason A. Heth, Erin L. McKean, Stephen E. Sullivan, Shawn L. Hervey-Jumper, Parag G. Patil, B. Gregory Thompson, Oren Sagher, Guy M. McKhann, Ricardo J. Komotar, Michael E. Ivan, Matija Snuderl, Marc L. Otten, Timothy D. Johnson, Michael B. Sisti, Jeffrey N. Bruce, Karin M. Muraszko, Jay Trautman, Christian W. Freudiger, Peter Canoll, Honglak Lee, Sandra Camelo-Piragua, and Daniel A. Orringer. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine*, 26(1):52–58, January 2020. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0715-9.

D. A. Novis and R. J. Zarbo. Interinstitutional comparison of frozen section turnaround time. A College of American Pathologists Q-Probes study of 32868 frozen sections in 700 hospitals. *Archives of Pathology & Laboratory Medicine*, 121(6):559–567, June 1997. ISSN 0003-9985.

Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, March 2023. ISSN 2047-7538. doi: 10.1038/s41377-023-01104-7.

Barbara Sarri, Rafaël Canonge, Xavier Audier, Emma Simon, Julien Wojak, Fabrice Caillol, Cécile Cador, Didier Marguet, Flora Poizat, Marc Giovannini, and Hervé Rigneault. Fast stimulated Raman and second harmonic generation imaging for intraoperative gastro-intestinal cancer detection. *Scientific Reports*, 9(1): 10052, July 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-46489-x.

Akhil Kondepudi, Melike Pekmezci, Xinhai Hou, Katie Scotford, Cheng Jiang, Akshay Rao, Edward S. Harake, Asadur Chowdury, Wajd Al-Holou, Lin Wang, Aditya Pandey, Pedro R. Lowenstein, Maria G. Castro, Lisa Irina Koerner, Thomas Roetzer-Pejrimovsky, Georg Widhalm, Sandra Camelo-Piragua, Misha

Movahed-Ezazi, Daniel A. Orringer, Honglak Lee, Christian Freudiger, Mitchel Berger, Shawn Hervey-Jumper, and Todd Hollon. Foundation models for fast, label-free detection of glioma infiltration. *Nature*, pages 1–7, November 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08169-3.

Cheng Jiang, Asadur Chowdury, Xinhai Hou, Akhil Kondepudi, Christian W. Freudiger, Kyle Conway, Sandra Camelo-Piragua, Daniel A. Orringer, Honglak Lee, and Todd C. Hollon. OpenSRH: Optimizing brain tumor surgery using intraoperative stimulated Raman histology. *Advances in neural information processing systems*, 35(DB):28502–28516, December 2022. ISSN 1049-5258.

Rock Christian Tomas, Anthony Jay Sayat, Andrea Nicole Atienza, Jannah Lianne Danganan, Ma Rollene Ramos, Allan Fellizar, Kin Israel Notarte, Lara Mae Angeles, Ruth Bangaoil, Abegail Santillan, and Pia Marie Albano. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. *PLOS ONE*, 17(1):e0262489, January 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0262489.

Sebastian Berisha, Mahsa Lotfollahi, Jahandar Jahanipour, Ilker Gurcan, Michael Walsh, Rohit Bhargava, Hien Van Nguyen, and David Mayerich. Deep learning for FTIR histology: Leveraging spatial and spectral features with convolutional neural networks. *Analyst*, 144(5):1642–1653, February 2019. ISSN 1364-5528. doi: 10.1039/C8AN01495G.

Catherine Berthomieu and Rainer Hienerwadel. Fourier transform infrared (FTIR) spectroscopy. *Photosynthesis Research*, 101(2-3):157–170, September 2009. ISSN 0166-8595, 1573-5079. doi: 10.1007/s11120-009-9439-x.

Martina Wolpert and Petra Hellwig. Infrared spectra and molar absorption coefficients of the 20 alpha amino acids in aqueous solutions in the spectral range from 1800 to 500cm-1. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 64(4):987–1001, July 2006. ISSN 13861425. doi: 10.1016/j.saa.2005.08.025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016.

Diederik P Kingma and Max Welling. An Introduction to Variational Autoencoders.

Home - CHARM. https://charm-eic.eu/, June 2022a.

ulTRafast hOlograPHic FTIR microscopY. https://www.trophy-project.eu/, June 2022b.

Jayakrupakar Nallala, Gavin Rhys Lloyd, Neil Shepherd, and Nick Stone. High-resolution FTIR imaging of colon tissues for elucidation of individual cellular and histopathological features. *Analyst*, 141(2):630–639, January 2016. ISSN 1364-5528. doi: 10.1039/C5AN01871D.

Moe Vali. Deep Learning Approaches using Raman Spectroscopy for Label Free Tumour Image Segmentation. 2024.

Ivo Vladislavov Petrov. Diagnosing cancer using AI on Raman Spectroscopic Signals. 2024.

Ewen Smith and Geoffrey Dent. *Modern Raman Spectroscopy: A Practical Approach*. John Wiley & Sons Ltd, Hoboken, NJ, second ed edition, 2019. ISBN 978-1-119-44058-1 978-1-119-44059-8 978-1-119-44054-3.

Yanting Shen, Jing Yue, Weiqing Xu, and Shuping Xu. Recent progress of surface-enhanced Raman spectroscopy for subcellular compartment analysis. *Theranostics*, 11(10):4872–4893, March 2021. ISSN 1838-7640. doi: 10.7150/thno.56409.

Wenjing Xu, Wei Zhu, Yukang Xia, Shun Hu, Guangfu Liao, Zushun Xu, Aiguo Shen, and Jiming Hu. Raman spectroscopy for cell analysis: Retrospect and prospect. *Talanta*, 285:127283, April 2025. ISSN 0039-9140. doi: 10.1016/j.talanta.2024.127283.

Richard C. Prince, Renee R. Frontiera, and Eric O. Potma. Stimulated Raman Scattering: From Bulk to Nano. *Chemical Reviews*, 117(7):5070–5094, April 2017. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00545.

Yuen R. Shen. *Principles of Nonlinear Optics*. Wiley Classics Library. Wiley-Interscience, Hoboken, NJ, wiley classics library ed edition, 2003. ISBN 978-0-471-43080-3.

SETH W. PERRY, RYAN M. BURKE, and EDWARD B. BROWN. Two-Photon and Second Harmonic Microscopy in Clinical and Translational Cancer Research. *Annals of Biomedical Engineering*, 40(2):277–291, February 2012. ISSN 0090-6964. doi: 10.1007/s10439-012-0512-9.

Xiyi Chen, Oleg Nadiarynkh, Sergey Plotnikov, and Paul J. Campagnola. Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. *Nature Protocols*, 7(4):654–669, April 2012. ISSN 1750-2799. doi: 10.1038/nprot.2012.009.

W. Denk, J. H. Strickler, and W. W. Webb. Two-photon laser scanning fluorescence microscopy. *Science (New York, N.Y.)*, 248(4951):73–76, April 1990. ISSN 0036-8075. doi: 10.1126/science.2321027.

Kar-Yan Su and Wai-Leng Lee. Fourier Transform Infrared Spectroscopy as a Cancer Screening and Diagnostic Tool: A Review and Prospects. *Cancers*, 12(1):115, January 2020. ISSN 2072-6694. doi: 10.3390/cancers12010115.

Peter R. Griffiths. Fourier Transform Infrared Spectrometry. *Science*, 222(4621):297–302, October 1983. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.6623077.

Siemen Brussee, Giorgio Buzzanca, Anne M. R. Schrader, and Jesper Kers. Graph Neural Networks in Histopathology: Emerging Trends and Future Directions, June 2024.

Keita Broadwater. Graph Neural Networks in Action. *Manning*, 2025.

Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph Networks as Learnable Physics Engines for Inference and Control. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4470–4479. PMLR, July 2018.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to Simulate Complex Physics with Graph Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8459–8468. PMLR, November 2020.

Atharva Tendle and Mohammad Rashedul Hasan. A study of the generalizability of self-supervised representations. *Machine Learning with Applications*, 6:100124, December 2021. ISSN 2666-8270. doi: 10.1016/j.mlwa.2021.100124.

Yuanhong Jiang, Qiaoqiao Ding, Yu Guang Wang, Pietro Liò, and Xiaoqun Zhang. Vision graph U-Net: Geometric learning enhanced encoder for medical image segmentation and restoration. *Inverse Problems and Imaging*, 18(3):672–689, 2024. ISSN 1930-8337, 1930-8345. doi: 10.3934/ipi.2023049.

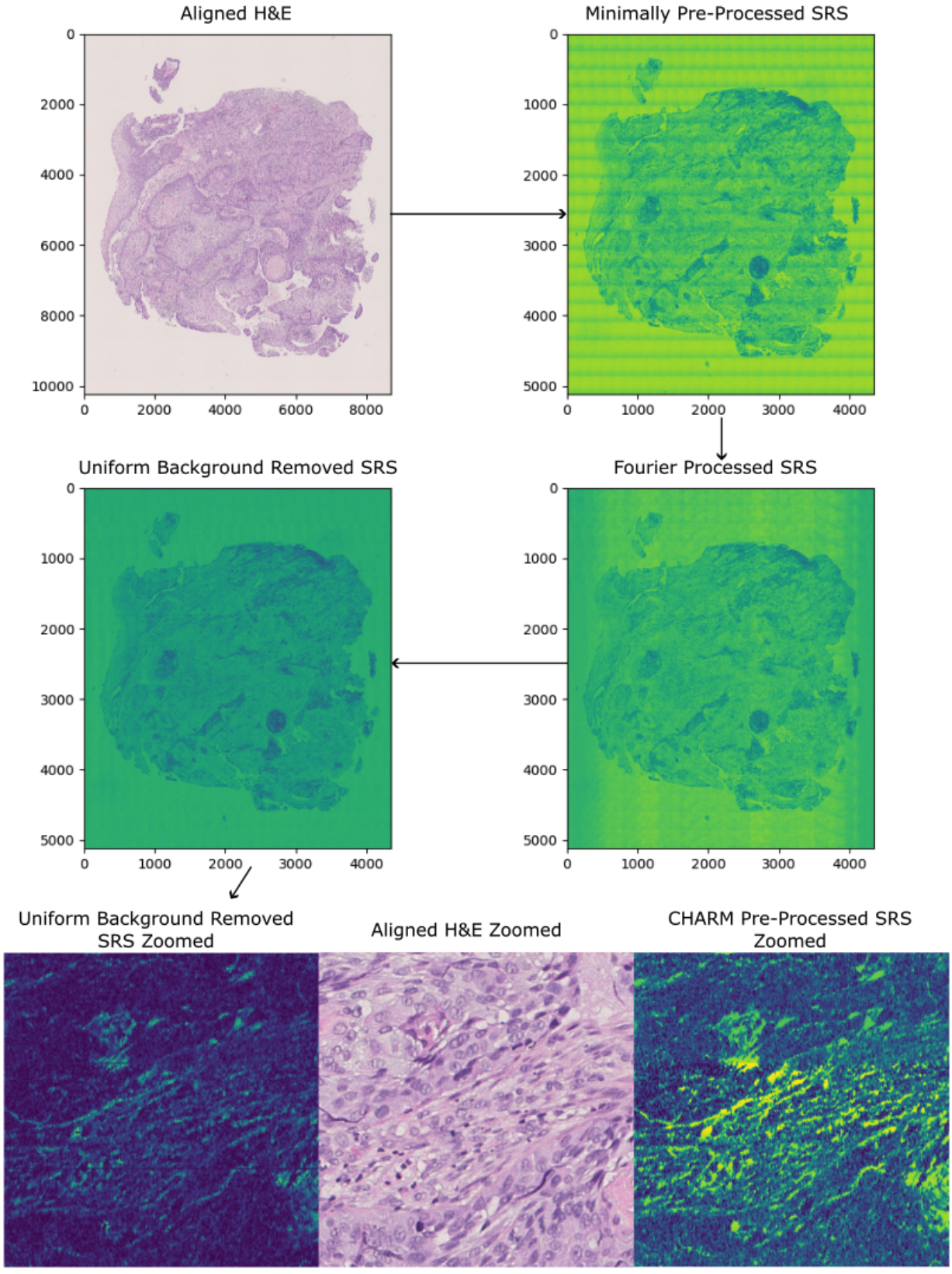


Figure 23: Plots showing a comparison between the Fourier domain noise removal strategy and that used by the CHARM group. The HE images for the respective regions are also provided. Note that the large differences between the SRS signal and the HE image are expected, as we are only showing one of the 31 recorded SRS spectra for the sample.