

# Product Biomass Samples IR Batch Spectroscopy

[Tags >](#)

[REPLACER](#)

## Goals

Twelve samples of biomass were provided which were harvested from a test reactor and subsequently freeze-dried. The primary objective of the work was to **identify chemical similarities or disparities between the samples** in order to assess whether the reactor output showed evidence of chemical drift over time. This addresses the practical question of whether the reactor can deliver consistent biomass quality across different batches.

A secondary objective was more methodological in nature: to explore whether **photothermal infrared (PTIR) spectroscopy and microscopy** can be used to establish spatio-spectral correlations within the samples. Such correlations, if reliably measurable, could allow the discrimination of different chemical classes (proteins, lipids, carbohydrates, nucleic acids, etc.) at micrometer-scale spatial resolution.

## Sample Preparation

The provided biomass consisted of freeze-dried *A. platensis*, which takes the form a green powder with grain size estimated between 0.1 and 1 mm.

From each sample of dried biomass, I first produced a **re-hydrated sample**. To that end, I placed one grain of dried biomass into a test tube on a scale. Recording the weight of the freeze-dried biomass, I added 10x that mass of MilliQ water, by pipette, directly onto the biomass grain. The rehydrated samples were *not* stirred in order to not destroy the structure of the biomass.

To prepare consistent samples for PTIR microscopy / spectroscopy, I started with an optically polished CaF<sub>2</sub> window. This was cleaned by

- rinsing it with a sequence of solvents (MilliQ water → Ethanol → Isopropanol)
- mechanically cleaning it with a microfibre cloth soaked in Isopropanol
- rinsing it with the same solvents as above, but in reverse order
- blowing off excess water with pressurized nitrogen gas

A stainless steel spacer ring ( $R_{\text{inner}} = 5\text{mm}$ ,  $d = 10\mu\text{m}$ ) was placed atop the CaF<sub>2</sub> window. A volume of  $2\mu\text{l}$  of the rehydrated sample was pipetted onto the middle of the CaF<sub>2</sub> window, in the centre of the spacer ring.

A second CaF<sub>2</sub> window, cleaned in the same manner as described above, was placed on top to close the sandwich. Gentle pressure was applied to compress the sample to the thickness of the spacer ring.

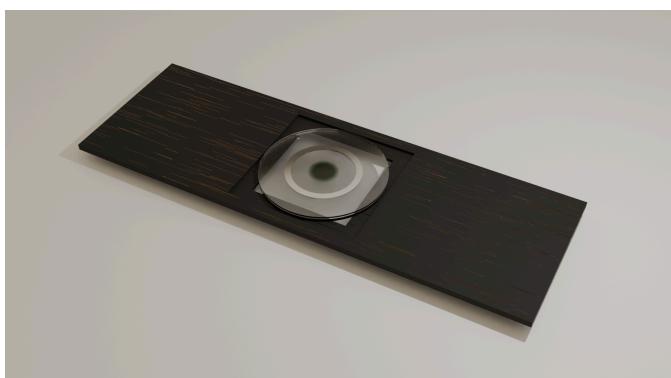


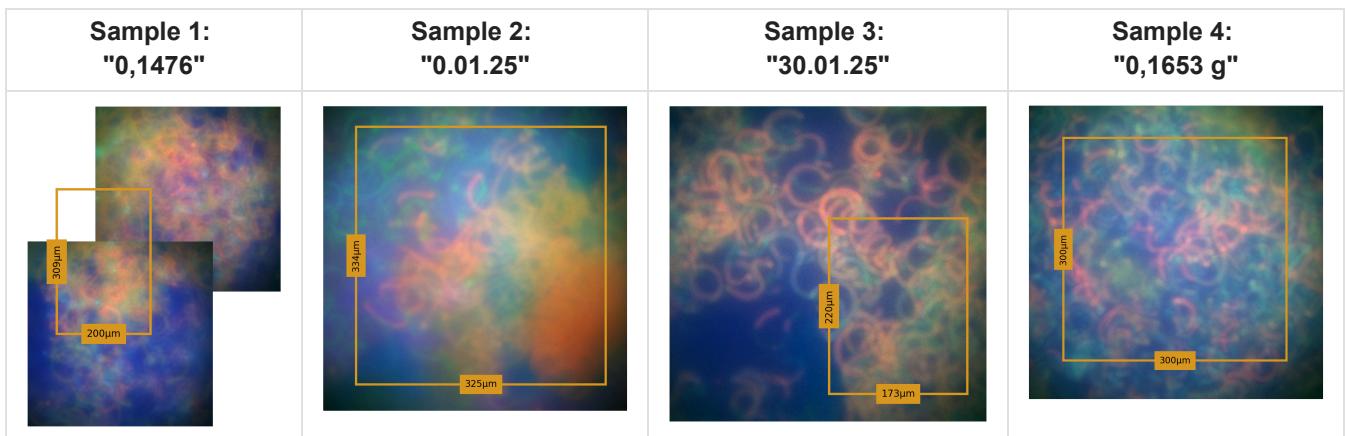
Illustration: Sample fixed to the [custom sample holder](#), ready to be placed into the microscope.

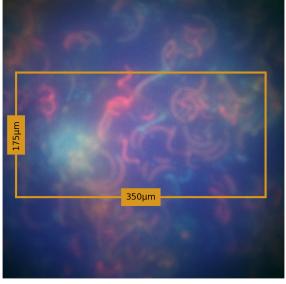
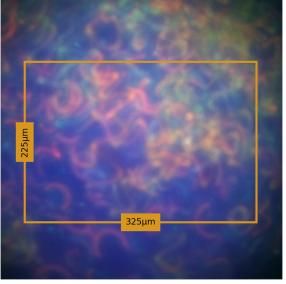
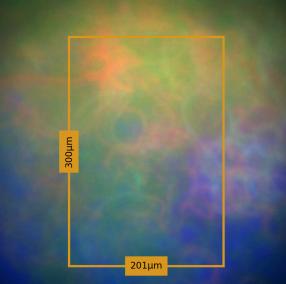
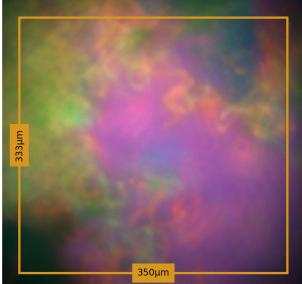
Sample	Labelled	Focus [ $\mu\text{m}$ ]	Detector Gain
1	<b>0,1476</b>	2660.7	10x
2	illegible (cap blue, paper sticker: <b>0.01.25</b> )	2484.5	50x
3	<b>30.01.25</b> A. platensis 0,15 g	2649.9	50x
4	<b>0,1653 g</b>	2592.0	50x
5	<b>98,6 mg</b>	2657.2	50x
6	03.02.25 A. platensis <b>0,19 g</b>	2569.8	50x
7	<b>A. platensis</b>	2641.9	50x
8	A. platensis 0,20 g <b>26.01.25</b>	2578.8	50x
9	<b>A. platensis 0,23 g</b>		
10			
11			
12			

After production of the ninth sample, a [crash of the setup](#) was noticed. The crash was unfortunately determined to be irrecoverable. No further measurements could be made.

## Measuring

Before spectral measurements, autofluorescence micrographs were captured to select a suitable region for spectroscopy:



Sample 5: "98,6 mg"	Sample 6: "0,19 g"	Sample 7: "A. platensis"	Sample 8: "26.01.25"
			

The Auto-Background procedure was executed once, before the first measurement. The PTIR setup was used in **co-propagation** mode, i.e. the IR pump and visible probe laser beams were both directed into the sample through the top **40x Cassegrain** objective. For each sampled point, the **average of 3 spectra** recorded at the same position was recorded.

Recorded Channels were

- [0] **OPTIR**: The amplitude of the PT signal, corrected for the background determined by the mIRage.
- [1] **Phase**: Phase of the PT signal
- [2] **X**: The real part (or in-phase component) of the PT signal
- [3] **Y**: The imaginary part (or quarter-phase delayed component) of the PT signal

The probe laser was modulated at a frequency of 100 kHz with a duty cycle of 1%.

---

## Results

### Spectra

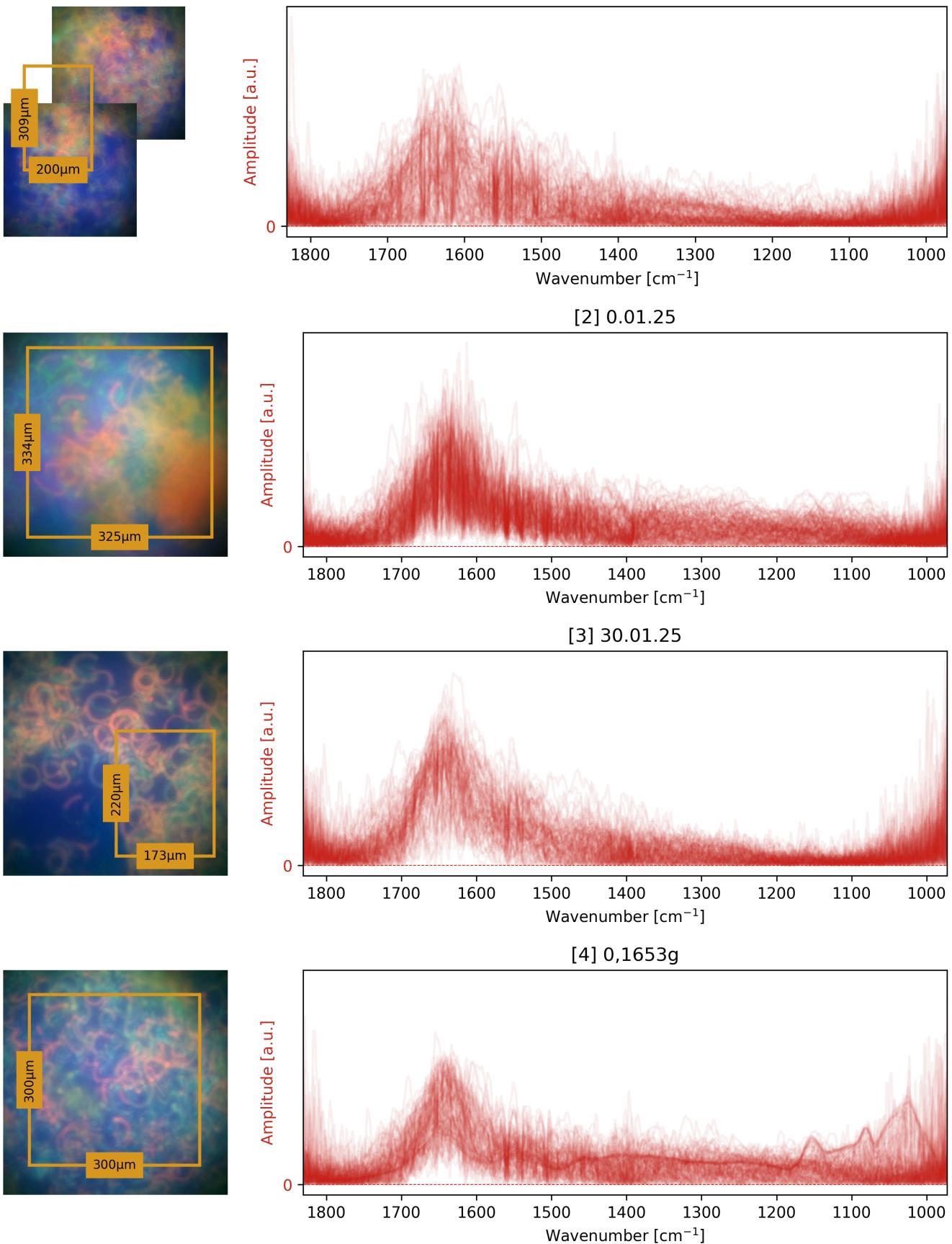
#### Definitions

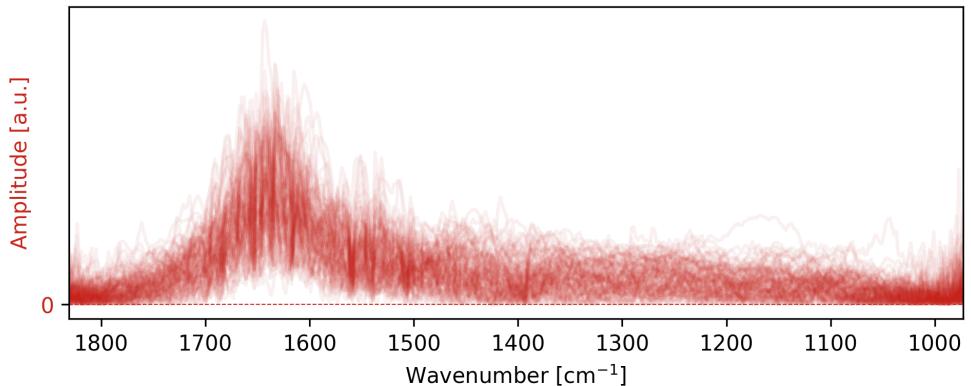
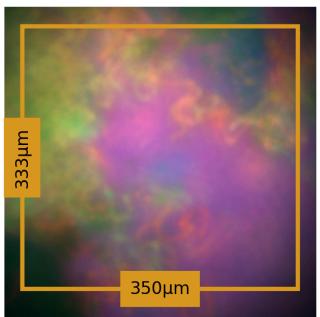
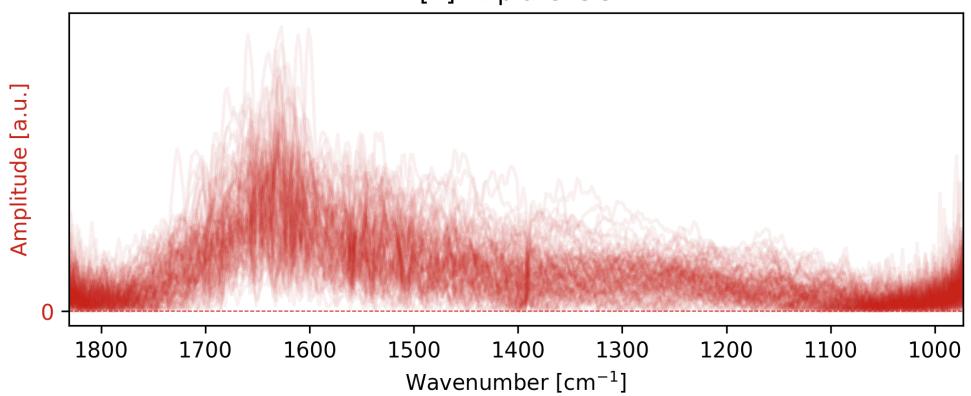
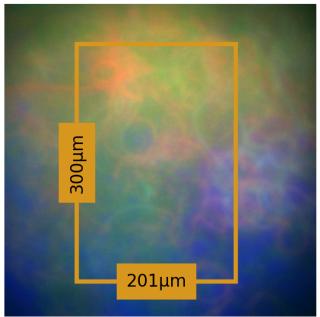
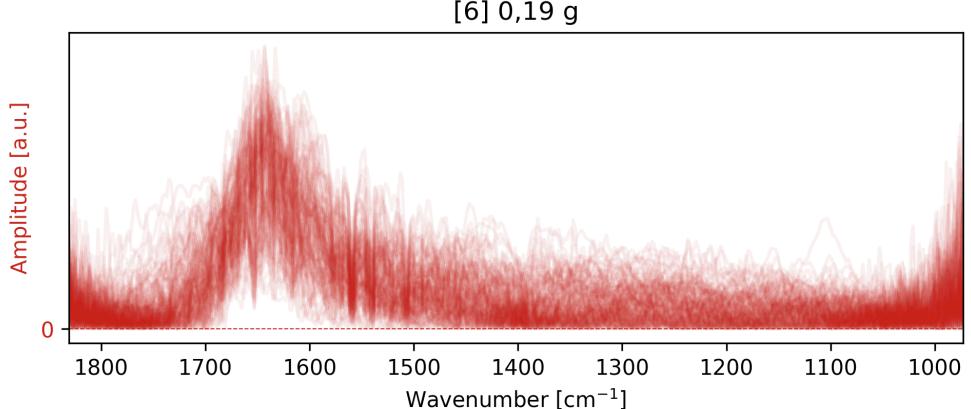
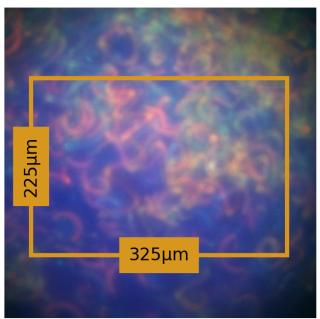
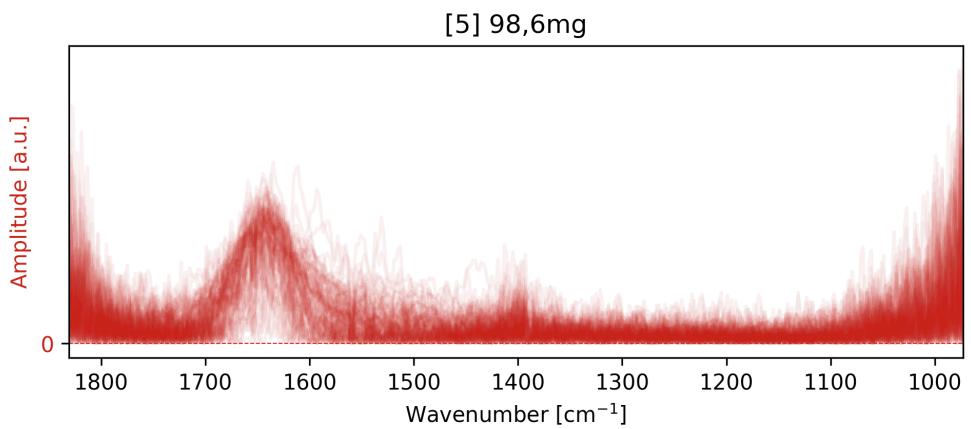
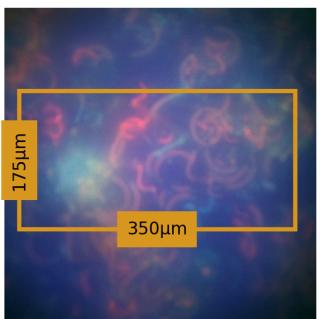
We assume that the "true" photothermal signal is encoded by  $X + iY$ . We thus calculate the **amplitude** of the signal as  $\sqrt{X^2 + Y^2}$  and the **phase** as  $\text{atan2}(Y, X)$ , instead of taking the phase as recorded by the software. We do this, because the phase of the average signal is not the same as the average phase of the signal:

$$\text{atan2}(\langle Y_i \rangle, \langle X_i \rangle) \neq \langle \text{atan2}(Y_i, X_i) \rangle$$

### Individual Spectra

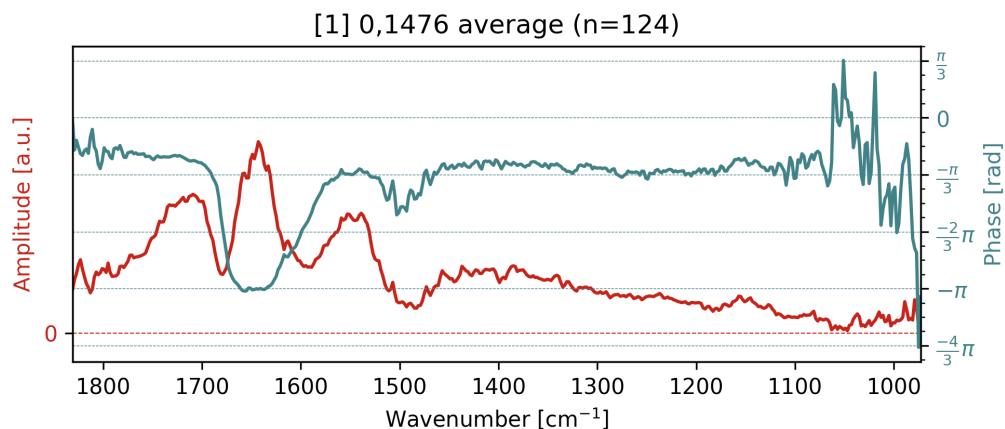
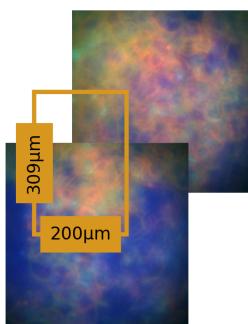
...were very noisy. Shown below are the  $L^2$ -normalized amplitude spectra from each sample. For each sample, an autofluorescence composite micrograph is given. The indicated areas are those, wherein spectra were recorded.



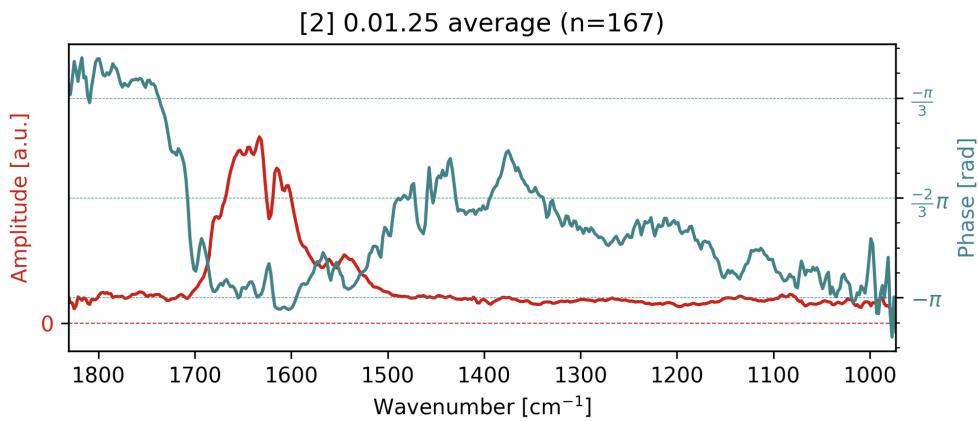
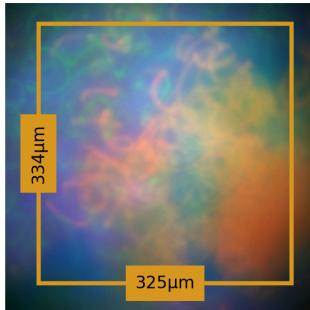


## Average Spectra

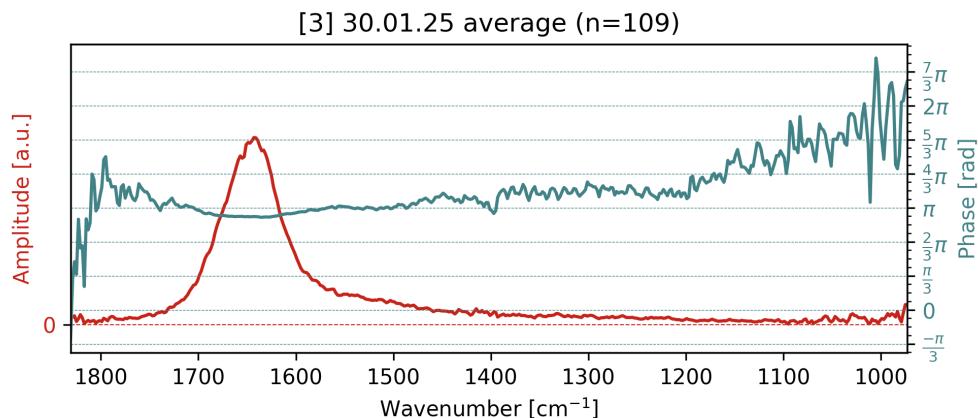
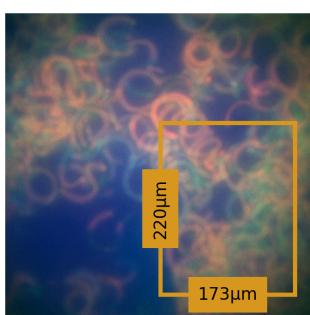
Here, average spectra are shown with amplitude and phase.



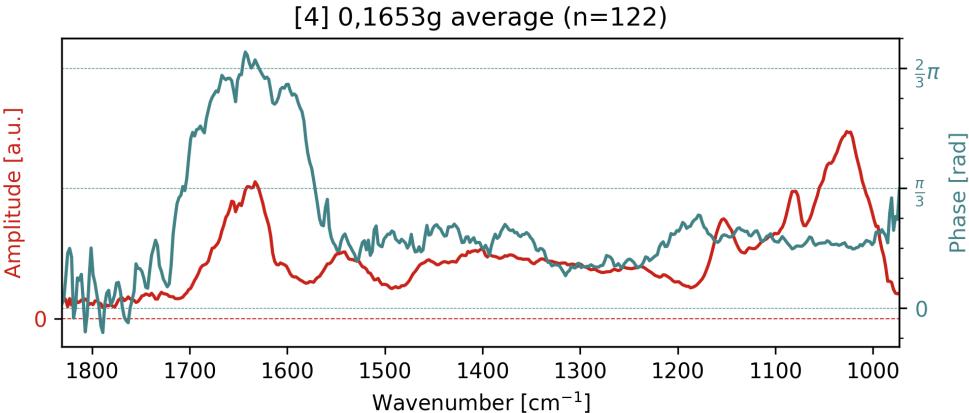
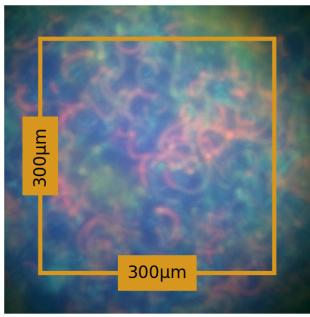
[1] Note how the main peak has a different phase from the secondary peaks. This might hint at the main peak corresponding to heating of a large volume (water background), taking a longer time for the temperature profile to relax and the secondary peaks corresponding to actual biomolecules, making up much less volume and therefore relaxing more quickly.



[2] Here, we similarly see a lagging phase in the region around  $1650\text{cm}^{-1}$ .

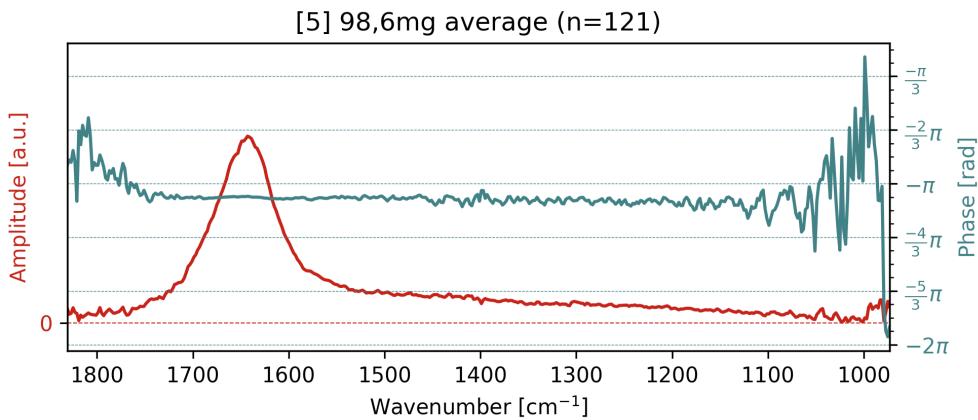
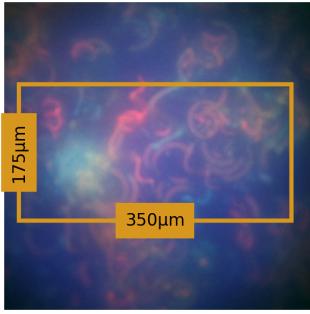


[3] Here, the phase seems to be invariant. The amplitude looks very smooth with only one visible peak and no secondaries. It doesn't look like any biomass is detected here, for some reason.

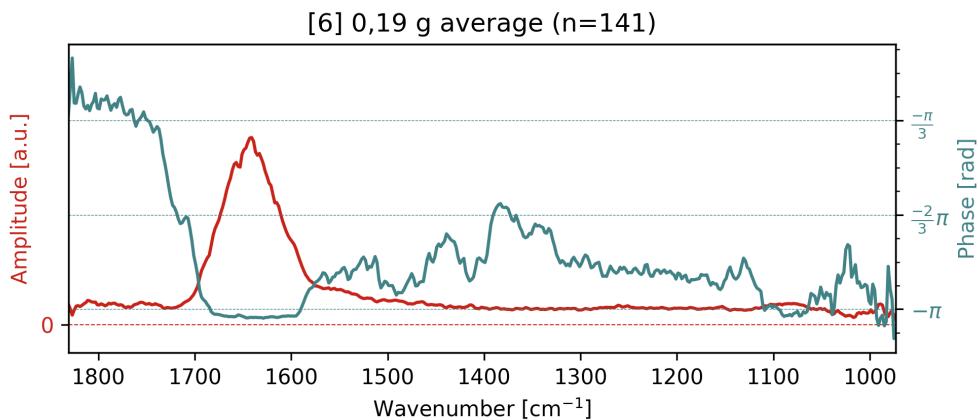
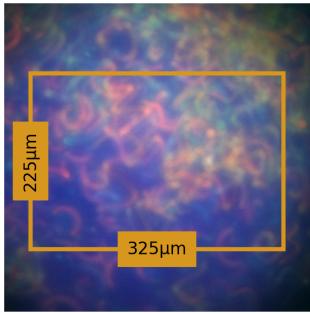


[4] Here, the phase doesn't lag around the water absorption but instead seems to advance. Reason unknown. Many additional peaks are visible in the amplitude of the signal. Their locations are consistent with usual bio-signatures, e.g.

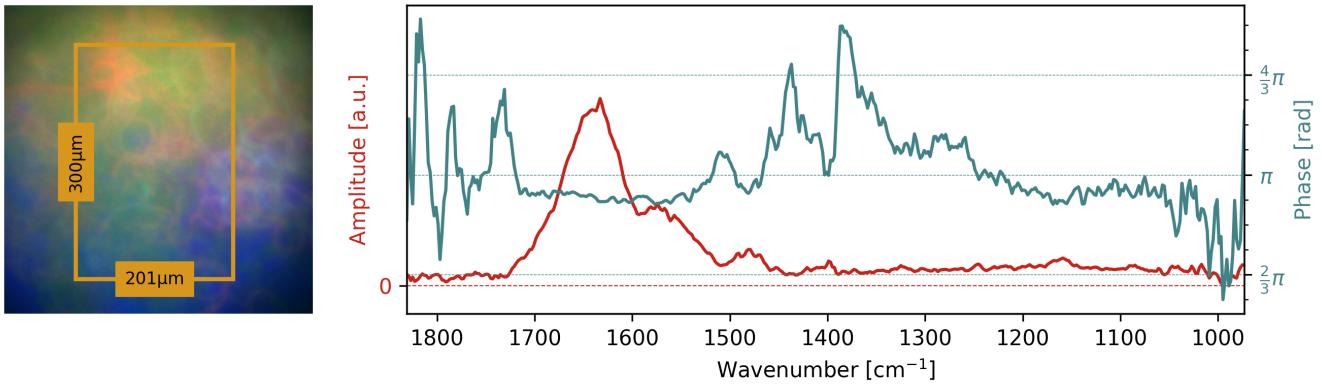
$\sim 1550 \text{ cm}^{-1}$	$\sim 1460 \text{ cm}^{-1}$	$\sim 1150 \text{ cm}^{-1}$	$\sim 1080 \text{ cm}^{-1}$	$\sim 1020 \text{ cm}^{-1}$
Amide II band	$\text{CH}_2$ scissor	C—O—C stretch	$\text{PO}_4$ stretch	carbohydrate C—O stretch



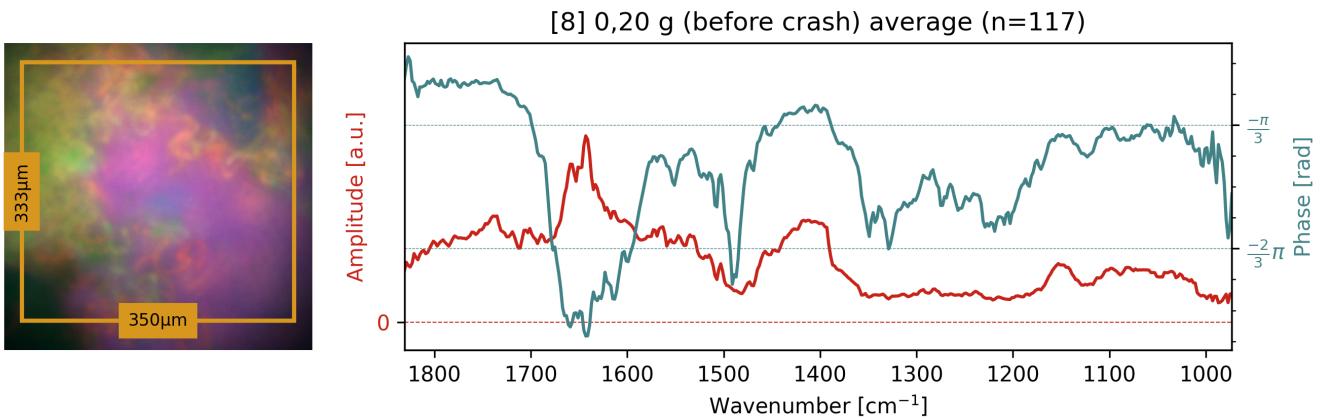
[5] as in [3], the phase is remarkably constant and no secondary peaks are visible.



[6] Secondary peaks are almost invisible. However, the phase clearly lags around the water absorption range.



[7] A peak in the high 1500s appears more like a shoulder to the side of the main peak at 1650 cm<sup>-1</sup>. Smaller peaks are discernible in the amplitude at lower wavenumbers, and coincide with clear local phase lags.



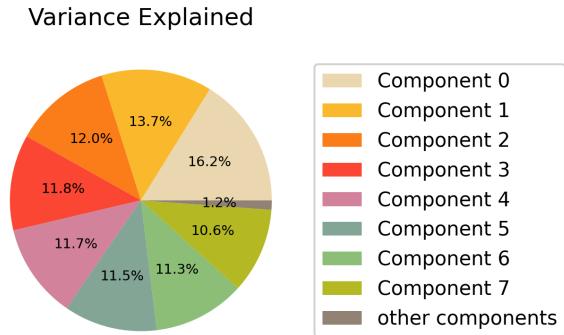
[8] Multiple peaks with irregular shapes are discernible in the amplitude. Many features (moth peaks and valleys) of the amplitude spectrum coincide with features in the phase spectrum.

## Decomposition of Spectra

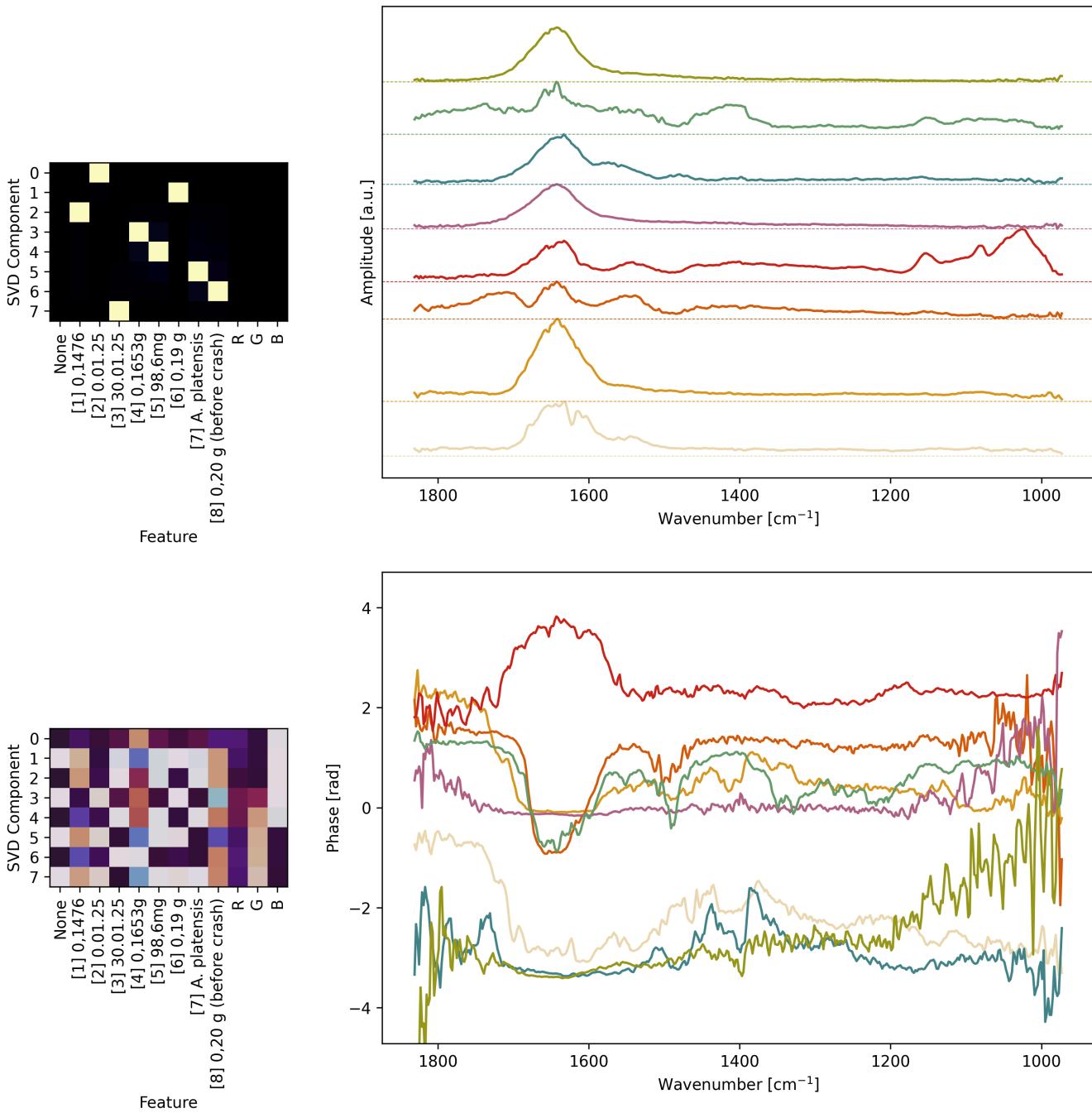
As usual, we cannot extract comprehensible information directly from the complete hyperspectral datasets and therefore must resort to some manner of dimension reduction. In this, it appears prudent to analyse the complex-valued spectra and to not disregard the phase.

## Singular Value Decomposition

I performed PCA (or SVD, because `sklearn.decompose.PCA` doesn't support complex-valued input) The explained variance per SVD component looked like so:



This means that the first eight components were significant and the rest could be disregarded, essentially. Inspecting the components gave the following:

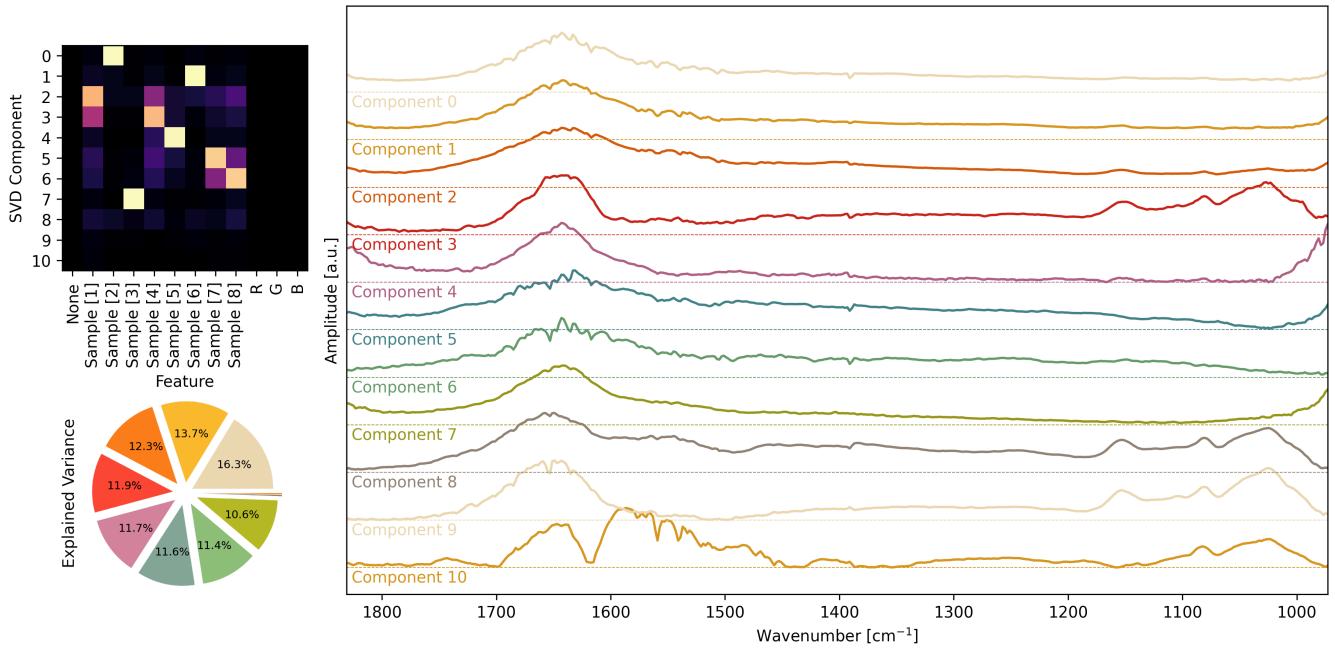


- **top left:** magnitude of the mapping between SVD components and features (sample indices)
- **top right:** magnitude of the SVD components vs wavenumbers
- **bottom left:** phase of the mapping between SVD components and features (sample indices)
- **bottom right:** phase of the SVD components vs wavenumbers

The interpretation is as straight-forward as it is disappointing:

**The eight relevant SVD components were nothing but the average spectra of the eight samples.**

So, counter to the earlier idea, I attempted the decomposition disregarding the phase. When running the **SVD** only on the **amplitude** of the signal, we unfortunately get similar results:



- **top left:** magnitude of the mapping between SVD components and features (sample indices)
- **bottom left:** explained variance per SVD component. only 0.6% of the total variance of the dataset are explained by components beyond 7.
- **right:** SVD components vs wavenumbers

However, there is some overlap between measurements now. Still, it ought to be very clear that we should attempt to model the individual peaks and try decomposing into those.

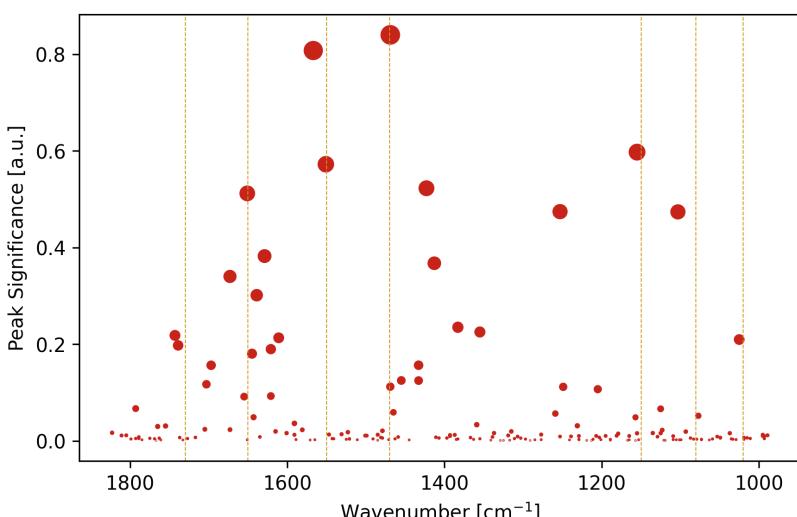
## Attempts to Isolate Spectral Peaks

Let's accumulate information about all spectral peaks using `scipy.signal.find_peaks`. Obviously important is the position (in terms of wavenumbers) of each peak, but also, we require a measure for total absorption that a peak is associated with. Given the values of "prominence" and "width" that the `find_peaks` algorithm computes, we define a peak's

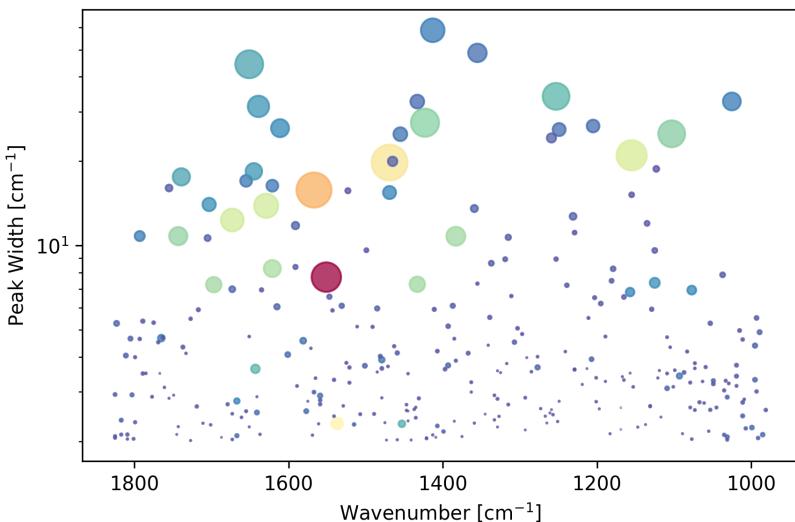
$$\text{significance} := \text{prominence} \cdot \text{width} .$$

Then, we can visualize all peaks like so:

**A**



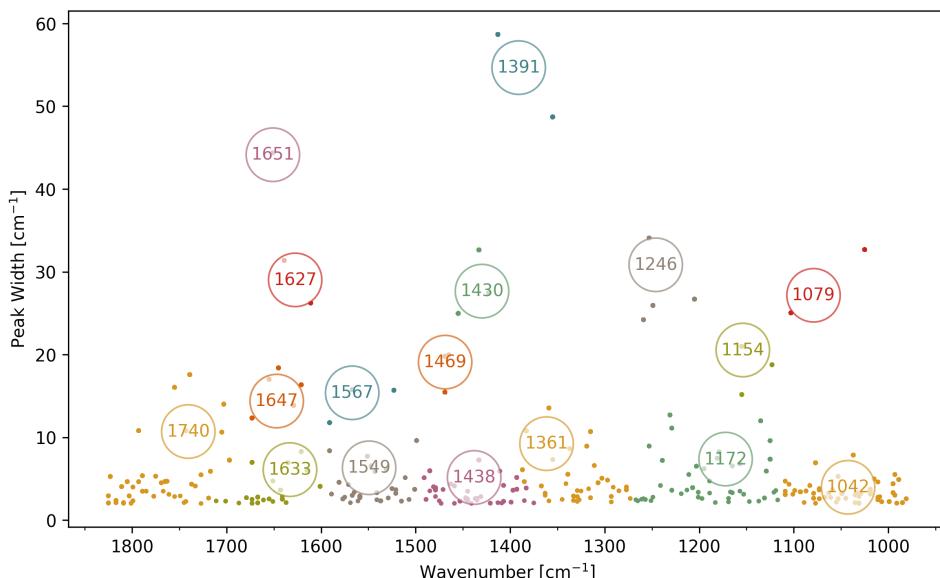
Significance of peaks vs. position. Prominence was normalized to `max(spectrum)`. The vertical lines highlight a selection of wavenumbers where we expect biosignatures.

**B**

Here, peak width (y coordinate) and prominence (colour) are separated. (The spot size continues to signify significance). When searching for clusters, we should be aware that two fundamental peaks may lie at the same position but have different widths (like Amide-I and water absorption)

Next, we require a way to group these peaks into clusters.

I implemented weighted k-means clustering, tagging peaks by wavenumber and width, normalising width to 10x w.r.t. the wavelength dimension and duplicating datapoints by weight (width x prominence).



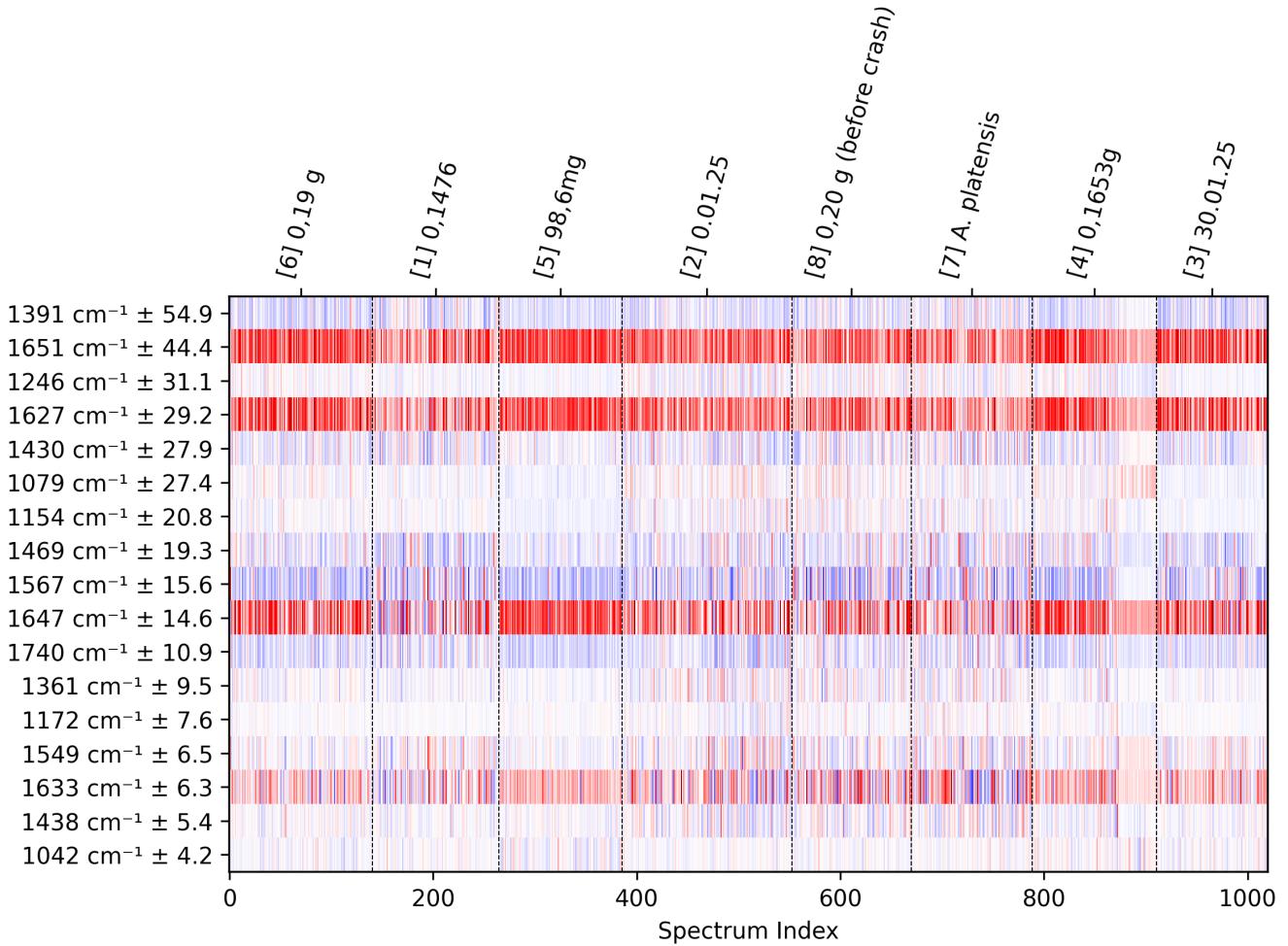
The following clusters of peaks are identified:

- $1391\text{ cm}^{-1}$  (width: 54.9) : **carboxylate symmetric stretch**
- $1651\text{ cm}^{-1}$  (width: 44.4) : **Amide I / Water** (tend more towards water)
- $1246\text{ cm}^{-1}$  (width: 31.1) : **symmetric PO<sub>4</sub> stretch or C–N stretch**
- $1627\text{ cm}^{-1}$  (width: 29.2) : **Amide I / Water**
- $1430\text{ cm}^{-1}$  (width: 27.9) : **CH<sub>2/3</sub> scissoring/bending**
- $1079\text{ cm}^{-1}$  (width: 27.4) : **symmetric PO<sub>4</sub> stretch**
- $1154\text{ cm}^{-1}$  (width: 20.8) : **C–O–C stretch**
- $1469\text{ cm}^{-1}$  (width: 19.3) : **CH<sub>2/3</sub> scissoring/bending**
- $1567\text{ cm}^{-1}$  (width: 15.6) : **Amide II**
- $1647\text{ cm}^{-1}$  (width: 14.6) : **Amide I / Water** (tend more towards Amide I)
- $1740\text{ cm}^{-1}$  (width: 10.9) : **C=O stretch**
- $1361\text{ cm}^{-1}$  (width: 9.5)

- $1172 \text{ cm}^{-1}$  (width: 7.6)
- $1549 \text{ cm}^{-1}$  (width: 6.5) : **Amide II**
- $1633 \text{ cm}^{-1}$  (width: 6.3) : **Amide I / Water**
- $1438 \text{ cm}^{-1}$  (width: 5.4)
- $1042 \text{ cm}^{-1}$  (width: 4.2)

Note: Position and width were given to the k-means algorithm without any transformation applied. Significance was used as a pseudo-weight, i.e. multiplicity of points. To achieve more "relevant" clusters, one could construct a transformation that virtually amplifies the distance between wider peaks. Then again, it should actually be more likely that a pair of peaks with a given distance in position should belong to the same cluster, if they are themselves wider.

As a measure of how well any of these peaks coincides with an actual spectrum, I calculated the scalar product of the spectrum with certain basis functions corresponding to each peak cluster's center. We get...



Red means positive, blue means negative.

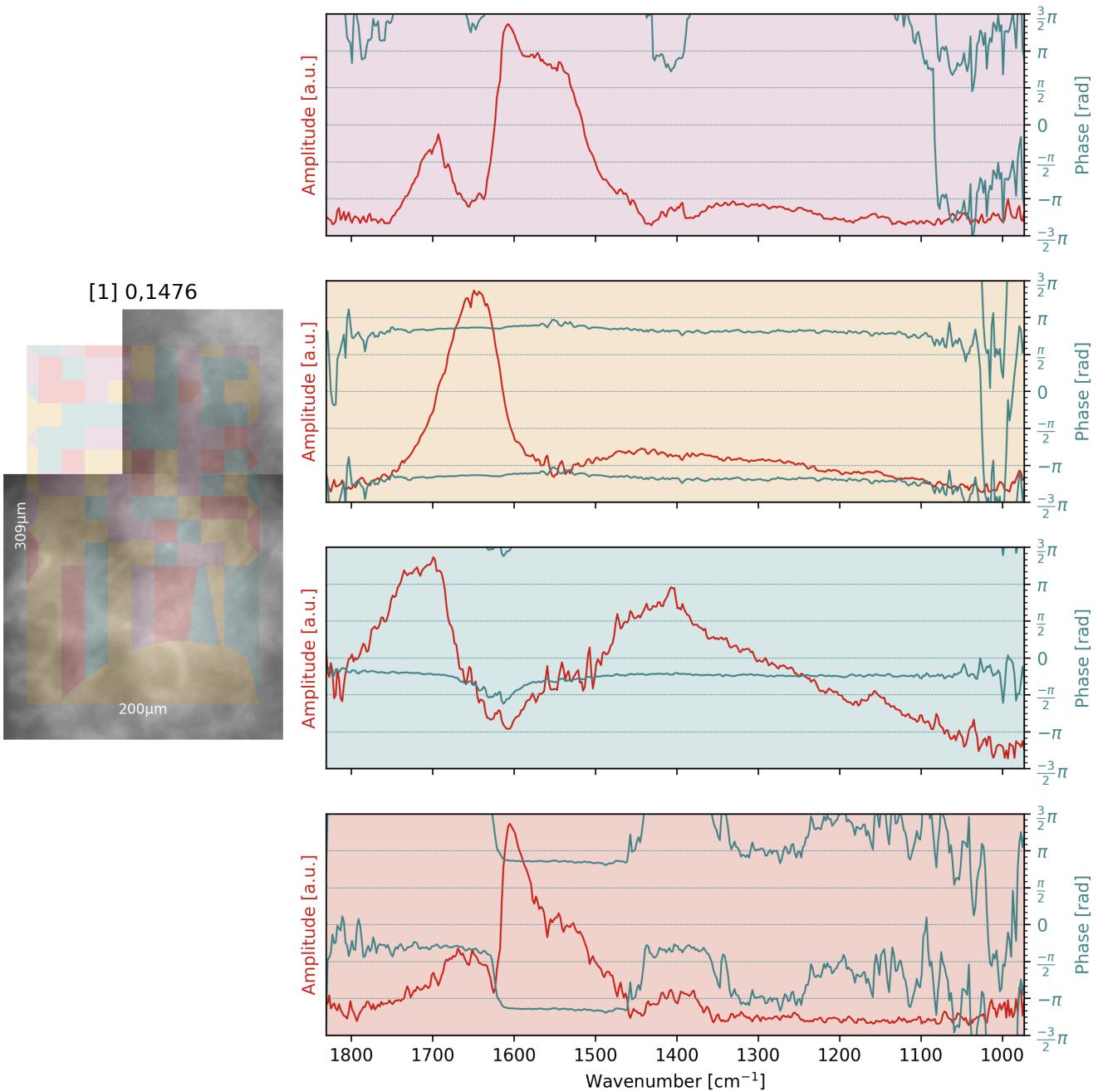
Unfortunately, decomposing spectra into a basis of these peaks did not uncover any correlation between samples either.

## Hyperspectral Image Segmentation

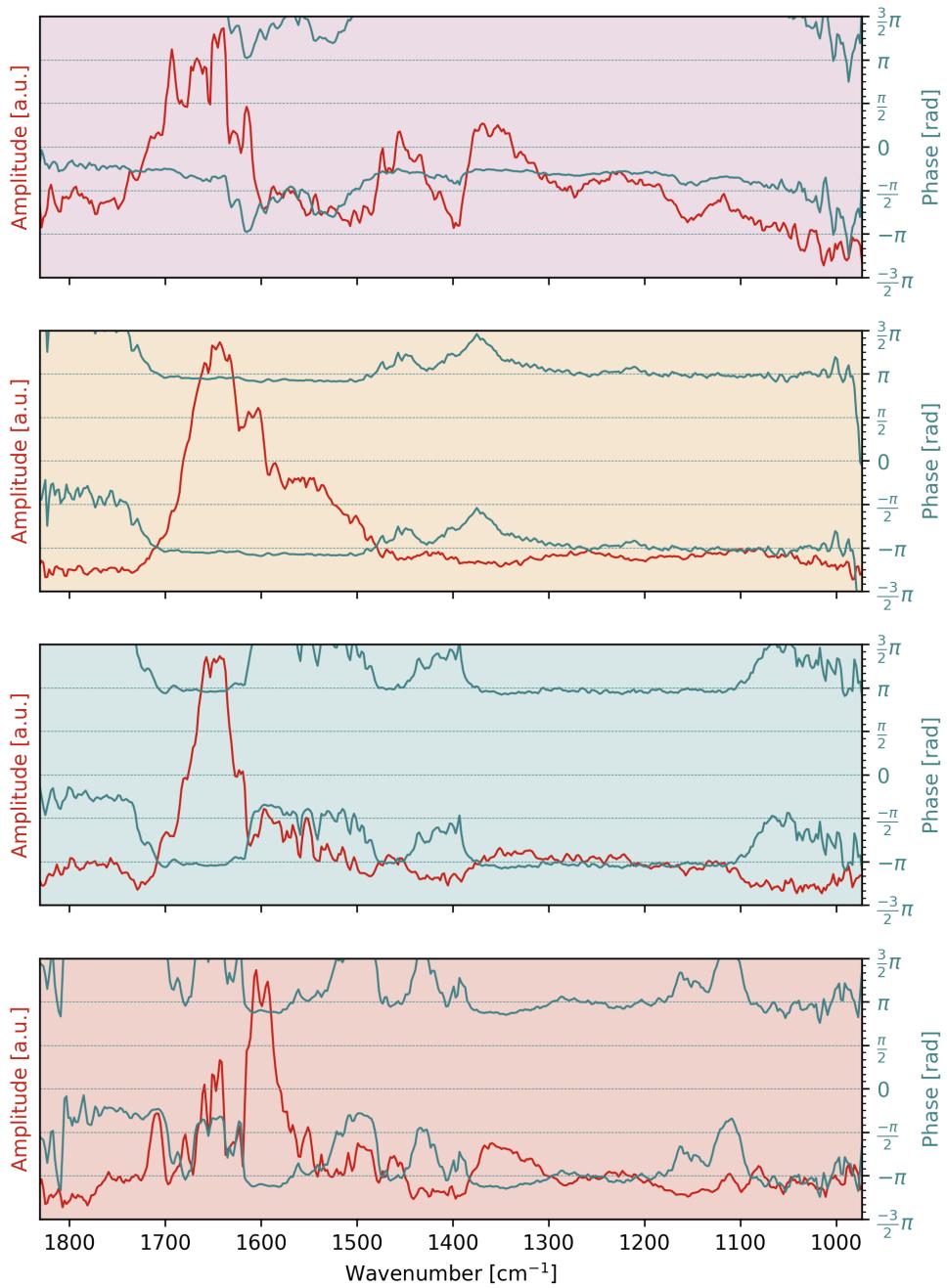
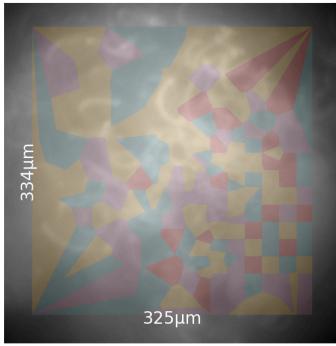
Instead of trying to model and track hypothetical spectral features through all measurements, let's attempt to classify spectra and see if these classifications are spatially correlated.

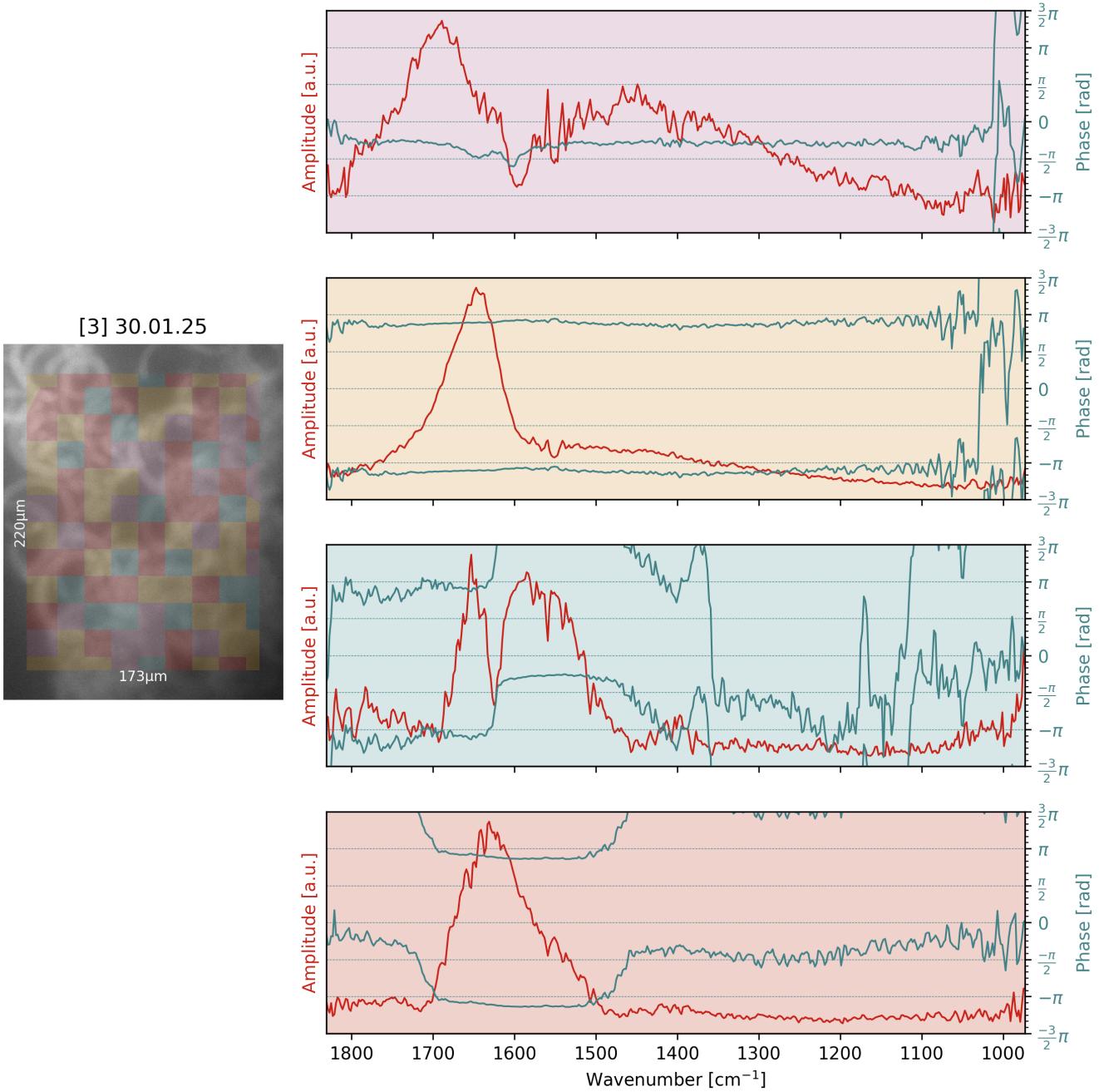
I calculated k-means clusters of the spectra in each sample (separately). I prescribed the number of clusters to be 4. We consider the X and Y channels.

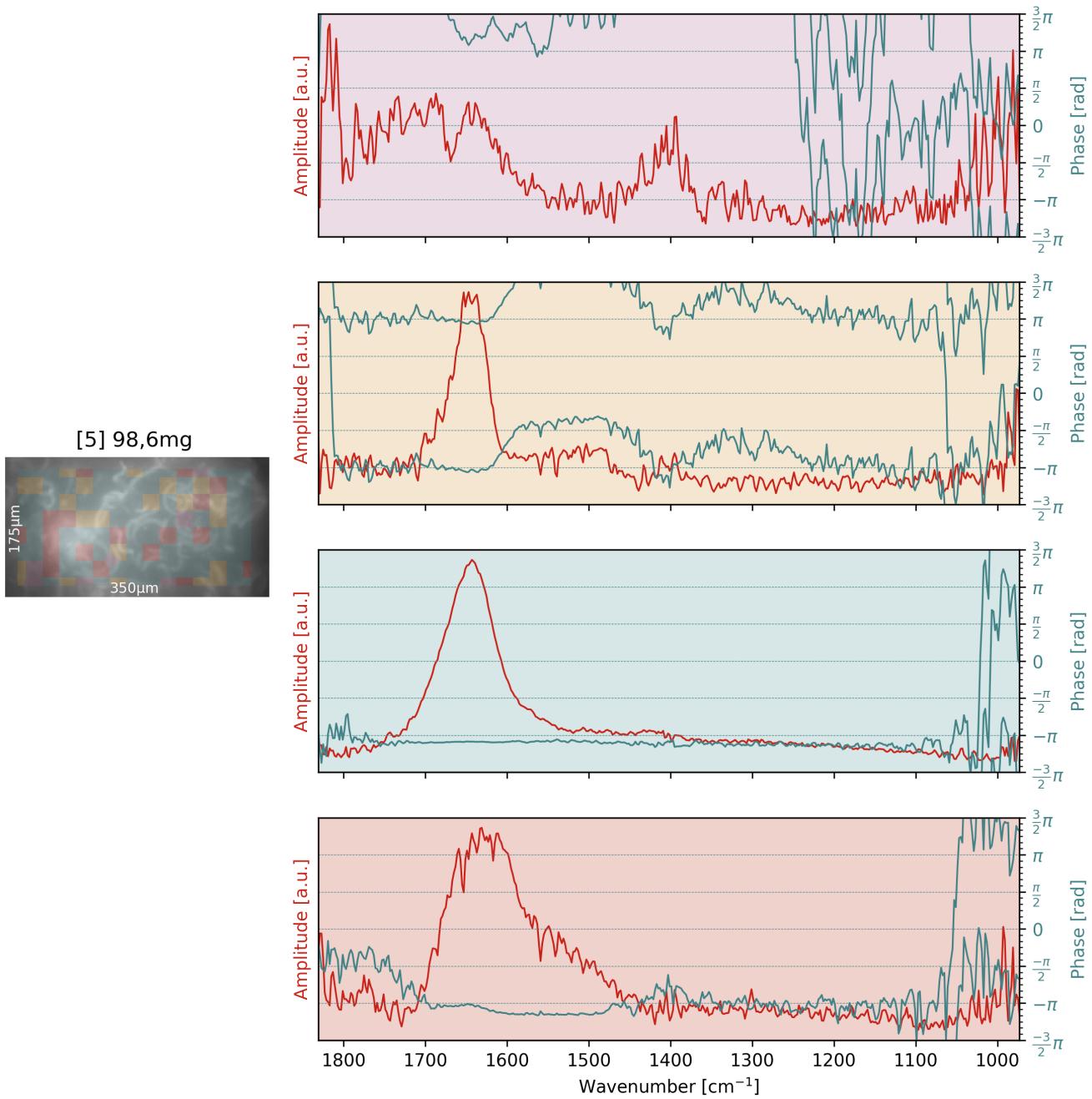
Each region of the image is then labelled with the cluster that its spatially closest spectrum belongs to. In the following, labelled images are presented on the left and representative spectra of the 4 clusters (the cluster centers) are shown on the right.

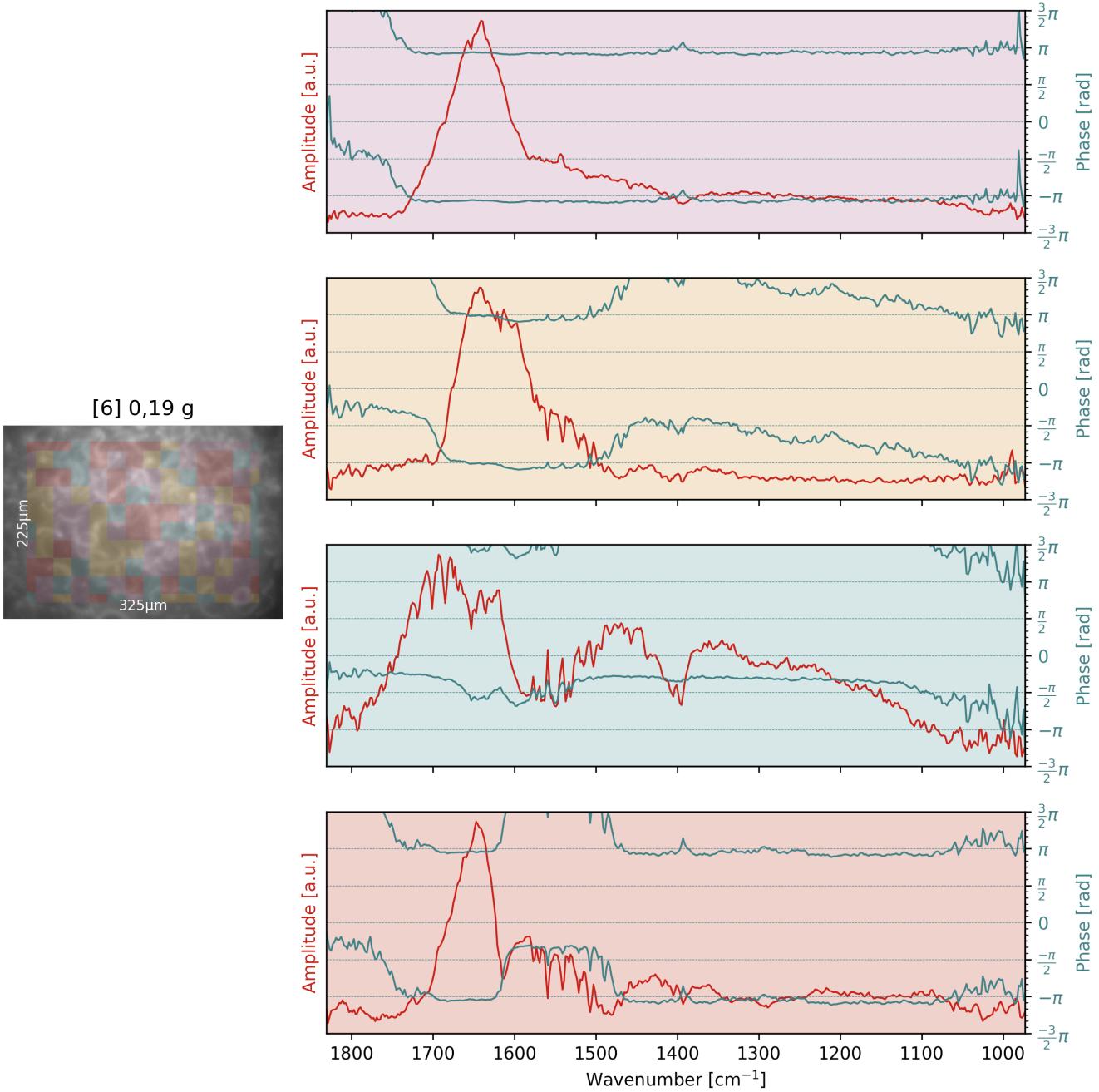


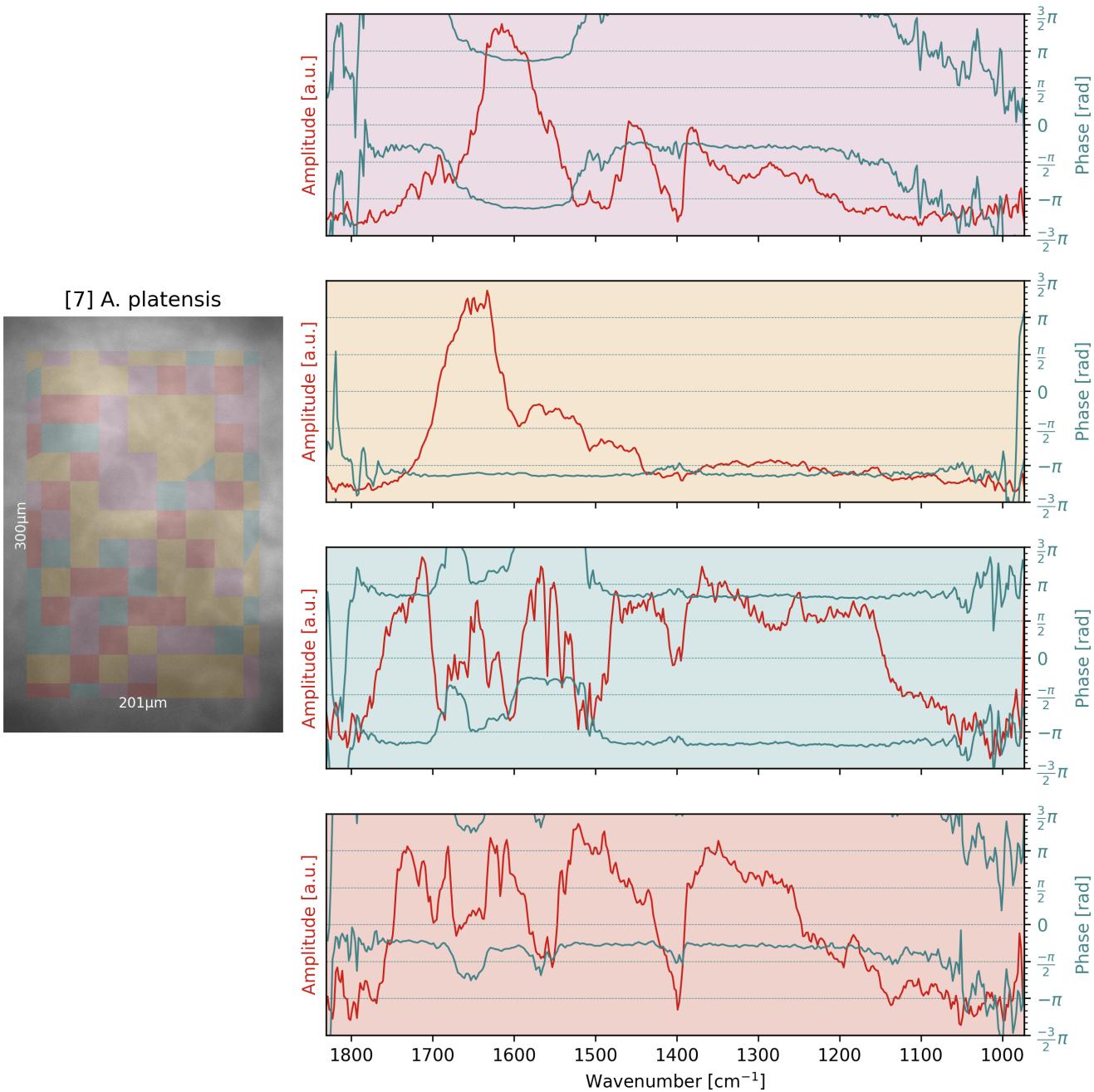
[2] 0.01.25



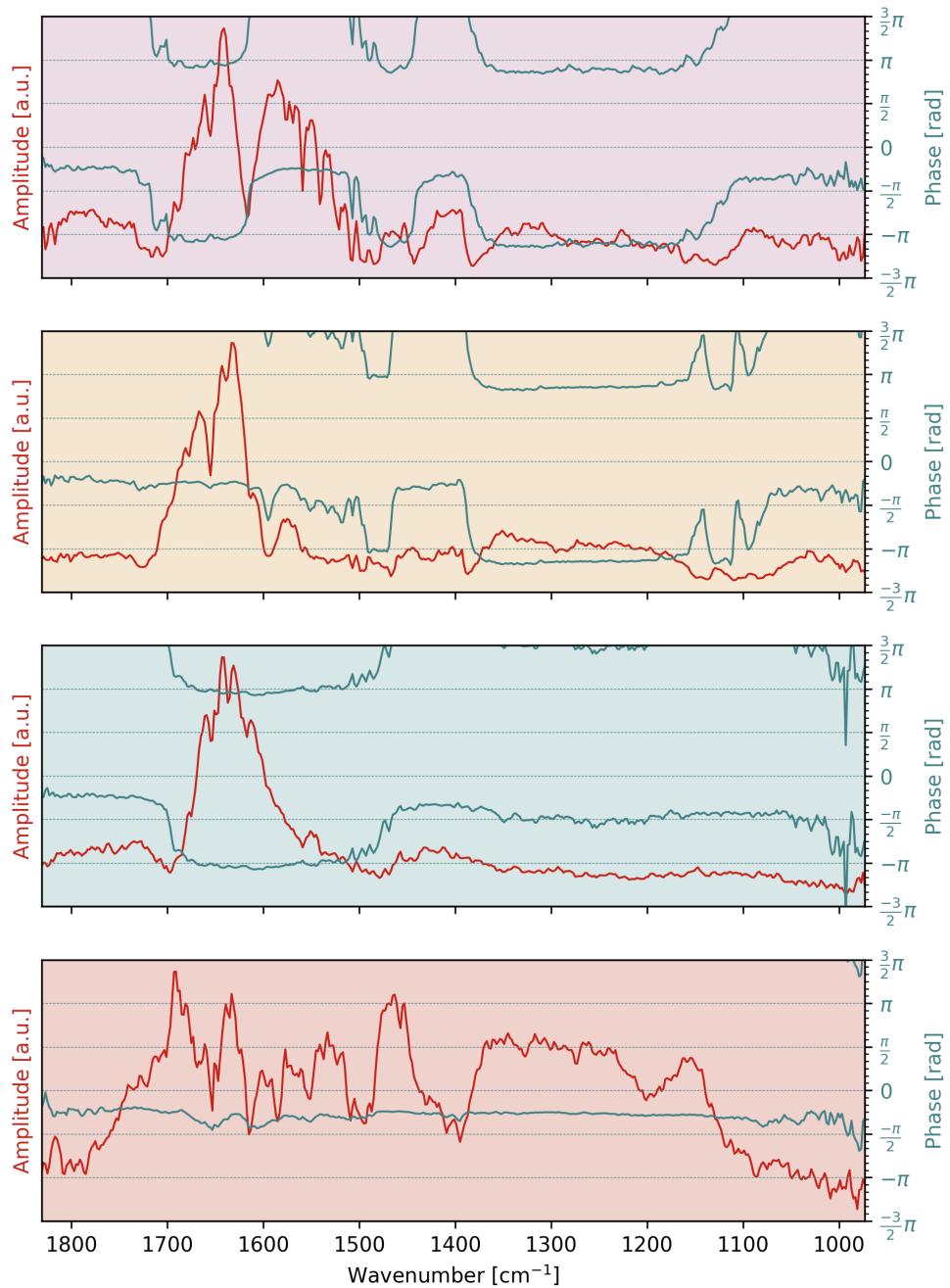
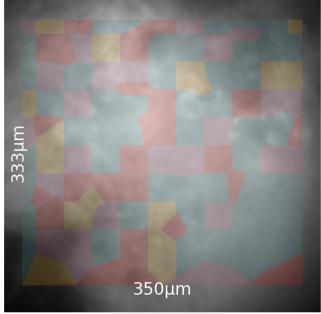




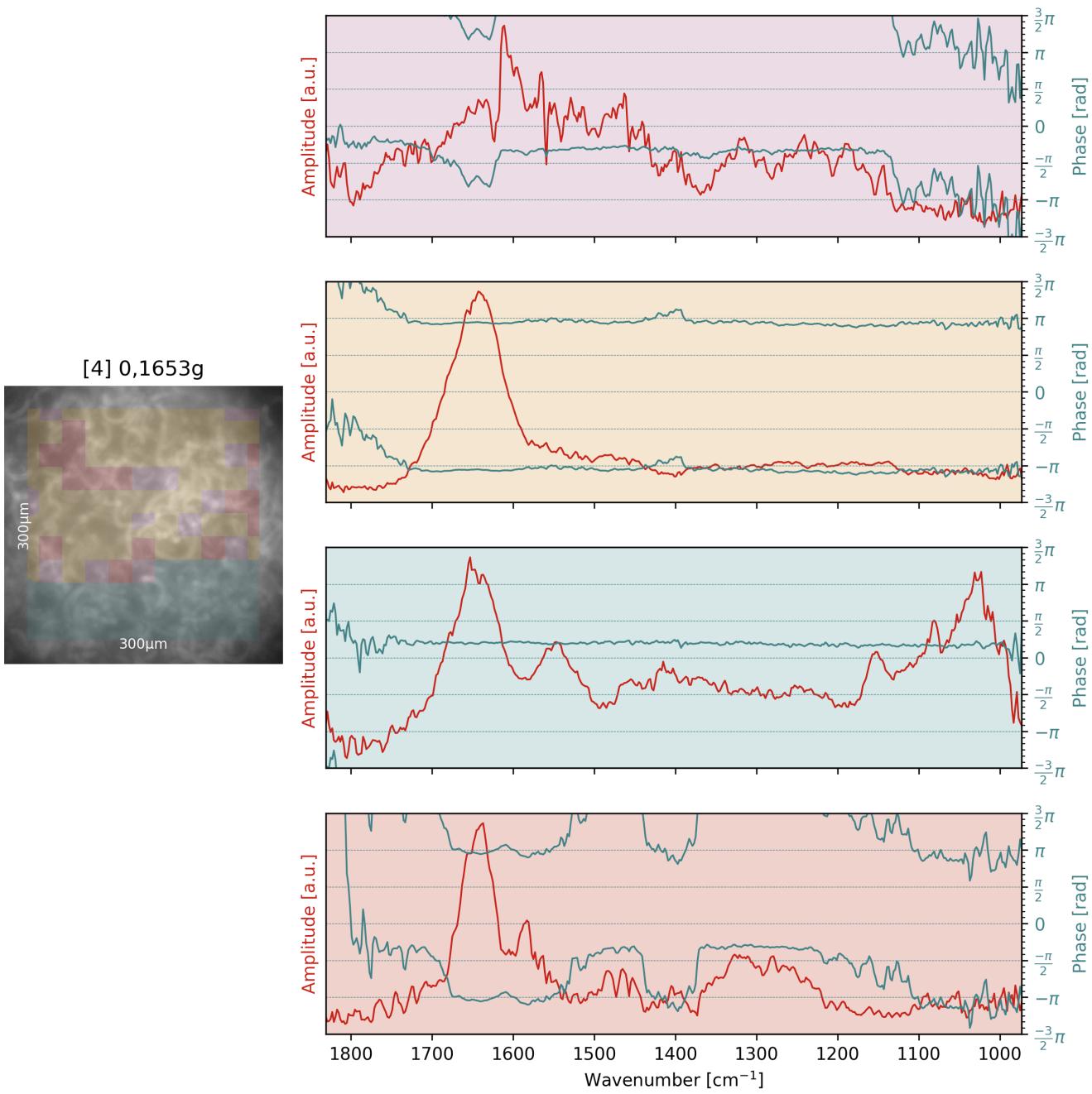




[8] 0,20 g (before crash)



In most of these, there seems to be no apparent spatio-spectral correlation.



Here, however, a clear spatial correlation is apparent. The upper two thirds of the image is labelled mostly as cluster 3 (yellow) with some clusters 1 (red) and 4 (purple) in between. Meanwhile, (something like) the lower third is exclusively labelled as cluster 2 (blue).

Both clusters 2 and 3 have, at their centres, comparatively low-noise spectra. Note that only cluster 2 shows the secondary peaks in the region between 1000 and 1200 cm<sup>-1</sup> and that its Amide-II peak is the most prominent one and the one placed most closely to 1550 cm<sup>-1</sup>.

One might surmise that in the blue-labelled region, and, to a lesser extent, in the red-labelled region, both proteins and lipids are being detected, while in the yellow-coloured region, only water is being detected. This conclusion is supported by the phase of the blue and yellow cluster centres: The phase of the yellow spectrum is close to constant at a value of  $\pm\pi$ . In the blue spectrum, the phase is close to constant as well, but around a value of approximately  $+\frac{\pi}{6}$ .

This can be assumed to be caused by different relaxation times for the IR-induced temperature profile, which, in turn, suggest that, in one case, a large volume of the solvent (MilliQ water) is being heated and in the other, a smaller volume of a biological structure being heated by the IR.

# Conclusions and Next Steps

From the measurements carried out so far, the mere presence of distinct biological signatures could be confirmed only in a subset of the samples (notably samples 1, 4, and 7). This indicates that, at the current stage, the available experimental and analytical methods are insufficient to make quantitative statements about the relative contributions of proteins, lipids, and carbohydrates in the biomass. Both the experimental setup and the subsequent analysis workflows therefore require further refinement before reliable quantitative conclusions can be drawn.

In relation to the reactor validation itself, conventional FTIR spectroscopy would likely provide faster and more robust information on overall chemical composition than the PTIR approach pursued here. The high-resolution chemical imaging capabilities of PTIR, however, remain of interest in a more fundamental context, particularly for exploring spatio-chemical correlations or, in a longer-term perspective, for tracking metabolic processes during biomass growth. It is important to recognize that this may not directly align with the current priorities of the reactor development. Nevertheless, it represents an important methodological component of the broader project and is of high relevance for future research directions.

Understanding the phase of the signal, in particular, appears promising as a means of distinguishing between background solvent contributions and genuine biomolecular features. Further development of this idea, however, will require systematic measurements on well-defined and reproducible model systems (e.g. polymer beads, lipid vesicles), rather than complex, variable biological samples. Only with such controlled systems can the fundamental behaviour of the PTIR signal be properly understood and analytical methods reliably developed.

Unfortunately, progress is currently limited by the inconsistent performance of the experimental setup. At present, malfunctions and unstable behaviour make reproducible measurements essentially impossible. Manufacturer support has so far not led to lasting solutions. It will be necessary to establish a more systematic and internally managed approach to diagnosing and correcting these issues. In the long term, building the expertise to maintain and stabilise the setup independently will be essential for advancing this line of research.

The immediate next steps are twofold:

1. the development of robust protocols on reproducible model systems, and
2. the stabilisation of the experimental platform.

These will form the necessary foundation for any future application to complex biological materials.