

**RESEARCH ARTICLE**

# Quantifying the impact of sample, instrument, and data processing on biological signatures in modern and fossil tissues detected with Raman spectroscopy

Jasmina Wiemann<sup>1,2,3,4,5</sup>  | Philipp R. Heck<sup>1,2</sup>

<sup>1</sup>Robert A. Pritzker Center for Meteoritics and Polar Studies, Earth Science Section, Negaunee Integrative Research Center, Field Museum of Natural History, Chicago, Illinois, USA

<sup>2</sup>Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA

<sup>3</sup>Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, Maryland, USA

<sup>4</sup>Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California, USA

<sup>5</sup>Natural History Museum of Los Angeles County, Los Angeles, California, USA

**Correspondence**

Jasmina Wiemann, Robert A. Pritzker Center for Meteoritics and Polar Studies, Earth Science Section, Negaunee Integrative Research Center, Field Museum of Natural History, Chicago, IL, USA.

Email: [jwiemann@fieldmuseum.org](mailto:jwiemann@fieldmuseum.org)

**Funding information**

We acknowledge funding from the Agouron Institute (JW) and the TAWANI Foundation (PRH).

**Abstract**

Raman spectroscopy is a popular tool for characterizing complex biological materials and their geological remains. Ordination methods, such as principal component analysis (PCA), use spectral variance to create a compositional space, the ChemoSpace, grouping samples based on spectroscopic manifestations reflecting different biological properties or geological processes. PCA allows to reduce the dimensionality of complex spectroscopic data and facilitates the extraction of informative features into formats suitable for downstream statistical analyses, thus representing a first step in the development of diagnostic biosignatures from complex modern and fossil tissues. For such samples, however, there is presently no systematic and accessible survey of the impact of sample, instrument, and spectral processing on the occupation of the ChemoSpace. Here, the influence of sample count, unwanted signals and different signal-to-noise ratios, spectrometer decalibration, baseline subtraction, and spectral normalization on ChemoSpace grouping is investigated and exemplified using synthetic spectra. Increase in sample size improves the dissociation of groups in the ChemoSpace, and our sample yields a representative and mostly stable pattern in occupation with less than 10 samples per group. The impact of systemic interference of different amplitude and frequency, periodical or random features that can be introduced by instrument or sample, on compositional biological signatures is reduced by PCA and allows to extract biological information even when spectra of differing signal-to-noise ratios are compared. Routine offsets ( $\pm 1 \text{ cm}^{-1}$ ) in spectrometer calibration contribute in our sample to less than 0.1% of the total spectral variance captured in the ChemoSpace and do not obscure biological information. Standard adaptive baselining, together with normalization, increases spectral comparability and facilitates the extraction of informative features. The ChemoSpace approach to biosignatures represents a powerful tool for exploring, denoising, and integrating molecular information from modern and ancient organismal samples.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Journal of Raman Spectroscopy* published by John Wiley & Sons Ltd.

**KEYWORDS**

accessible science, biological tissues, biosignatures, ChemoSpace, dimensionality reduction, ordination methods

**1 | INTRODUCTION**

Raman spectroscopy allows non-destructive compositional fingerprinting of complex biological and geological materials.<sup>1–10</sup> Rapidly generated *in situ* spectra yield information on covalent, ionic, and non-covalent chemical interactions enabling a comparative search for informative heterogeneities across a diversity of samples,<sup>1</sup> such as modern organismal tissues and their fossilization products. Spectroscopic biosignatures, such as phylogenetic and metabolic signals, represent diagnostic tools in cancer research,<sup>3–7</sup> and a number of signatures present in fresh tissues preserve, occasionally altered but not unrecognizable, in fossilized carbonaceous tissues: in integrative data sets, spectroscopic signatures reflecting the relative abundance of different organic functional groups<sup>1</sup> and organo-mineral interactions<sup>2</sup> encode molecular manifestations of phylogenetic affinity,<sup>2–7,11–13</sup> physiology,<sup>2–7,11–17</sup> and degree and mode of environmental or diagenetic alteration.<sup>1,2,18</sup> These signals are relative and can only be analyzed in a comparative framework.<sup>1–7,11–18</sup>

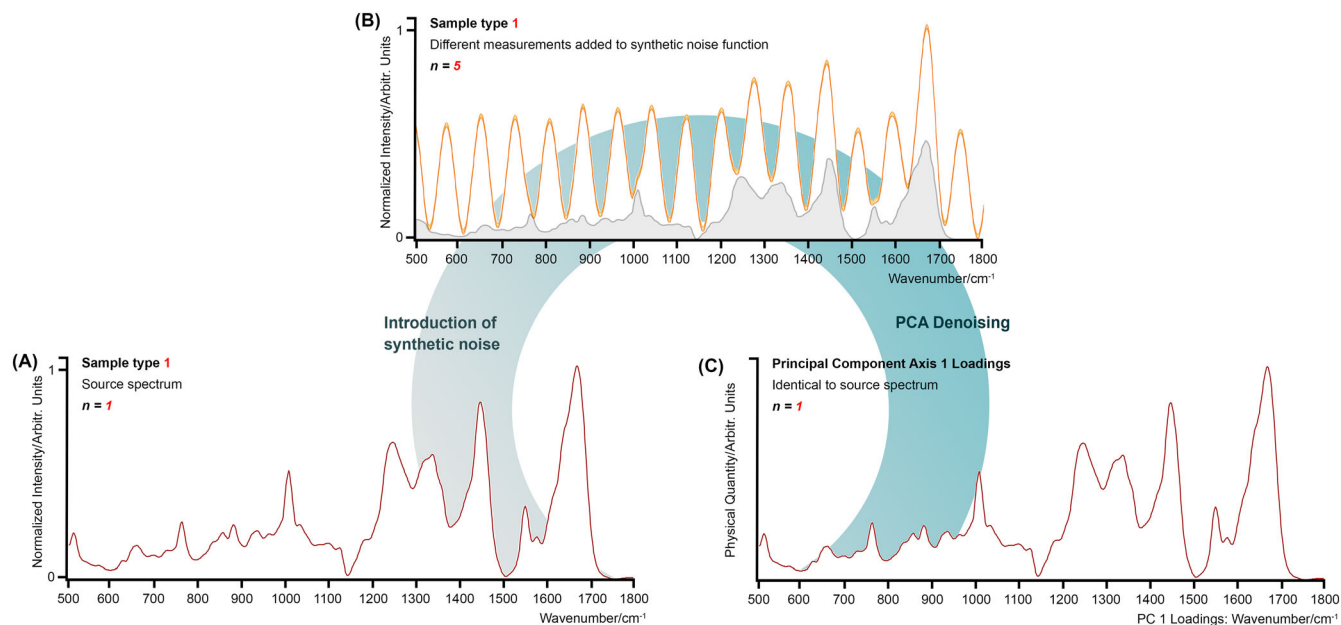
Spectra collected across a diversity of tissue samples may contain additional unwanted signals that reduce the signal-to-noise ratio<sup>1,2</sup> (“noise” representing here signal that is not of direct chemical nature). Examples include a nonlinear background based on sample fluorescence induced by the excitation source,<sup>1,19</sup> lower intensity counts due to diffusive scattering at rough sample surfaces,<sup>1,19</sup> and (quasi-)sinusoidal signals resulting from reflective scattering at layers with different optical properties within a tissue sample or introduced by certain instrument components (laser filters in combination with specific line gratings).<sup>19,20</sup> Most of these unwanted signals can be described as wave functions of different periodicity, amplitude, and frequency: fluorescence often times behaves like  $n = 1–1.5$  sine wave half cycles, diffusive scattering at tissue layers or filter materials is accurately represented by low-frequency periodical sine waves, and shot noise tends to behave like a random high-frequency, low-amplitude interference.<sup>1,19,20</sup> Noisy spectra containing a diversity of unwanted signals are a well-known challenge in biological tissue spectroscopy,<sup>1,7</sup> and processing routines beyond despiking and standard-based luminescence correction,<sup>21–23</sup> including adaptive baselining (background correction sensitive to the total spectral curve) and normalization (intensity scaling based on individual peaks or integrated spectral areas), are

employed to minimize the impact of unwanted signals on data interpretation.<sup>1,2,7,19</sup> Similarly, spectral phase shift introduced by temperature-based instrument decalibration can be traced and corrected across a series of analytical sessions,<sup>24</sup> but nonlinear decalibration rates render correction (up to  $\pm 1 \text{ cm}^{-1}$  wavenumber) during a single analytical session challenging.

In the last 30 years, spectroscopy has shifted from the exclusively qualitative interpretation of Raman spectra<sup>8–10</sup> toward a comparative approach<sup>1,2,5–7</sup> that relies, as an essential first step in the data analysis, on ordination methods (dimensionality reduction), such as principal component analysis (PCA). PCA allows to explore, denoise (Figure 1), identify, and extract informative heterogeneities (effectively “latent variables”) from sets of inherently complex spectra, each characterized by a very large number of data points collected over the wavenumber range.<sup>1,2,7,11–18</sup> PCA captures the covariance of spectral features in an  $n$ -dimensional compositional space, the ChemoSpace, where  $n$  equals the number of features considered.<sup>25,26</sup> The ChemoSpace is based on a variance–covariance matrix ( $[number\ of\ spectra] \times [number\ of\ features]$ ).<sup>25,26</sup>

The general order of magnitude of the minimum number of spectra required to achieve a stable Raman ChemoSpace occupation remains yet to be determined: the data point distribution across the ChemoSpace changes with the number and type of spectra or selected peaks included in the analysis; increasing the number of considered spectra increases the statistical power of sample group separation. Once the number of included Raman spectra allows for an accurate representation of the compositional diversity in a sample set, a stable pattern in ChemoSpace occupation is reached.

The number of considered features and thus dimensions of the ChemoSpace ( $n$ ) can encompass all the data points that contribute to a spectrum or, more commonly, selected peaks of interest.<sup>1,2</sup> PCA benefits from the subsampling of peaks in spectra, an approach that prevents overweighting broad signals and spectral regions uninformative for a given question.<sup>27</sup> Because the normalized intensities of Raman peaks represent the relative abundances of molecular features in the sampled area, the ChemoSpace can be thought of as a multivariate application of Lambert–Beer’s law, which defines that the spectroscopic signal of a compound is proportional to its abundance in the sample. When spectra of different biological sample types are analyzed by means of PCA,



**FIGURE 1** Schematic drawing showcasing the denoising potential of principal component analysis (PCA). (A)  $n = 1$  Raman spectrum of the sample type 1 plotted over the organic fingerprint region. A set ( $n = 5$ ) of synthetic technical replicates based on the spectrum plotted in (A) were summed with a medium-frequency synthetic interference function, in order to generate the  $n = 5$  spectra shown in (B). (B) The artificially noised  $n = 5$  varieties of the source spectrum shown in (A). The source spectrum in (A) is plotted under the summed spectral curves and is shaded in gray. PCA is applied to the  $n = 5$  summed spectra. (C) Resulting PC 1 axis loadings plotted over the organic fingerprint region match the original source spectrum shown in (A). Detectable synthetic interference has been efficiently removed by PCA. PCA allows for the robust denoising of spectroscopic data collected for biological or paleontological tissue samples. Source spectra and interference functions can be found in Table S1.

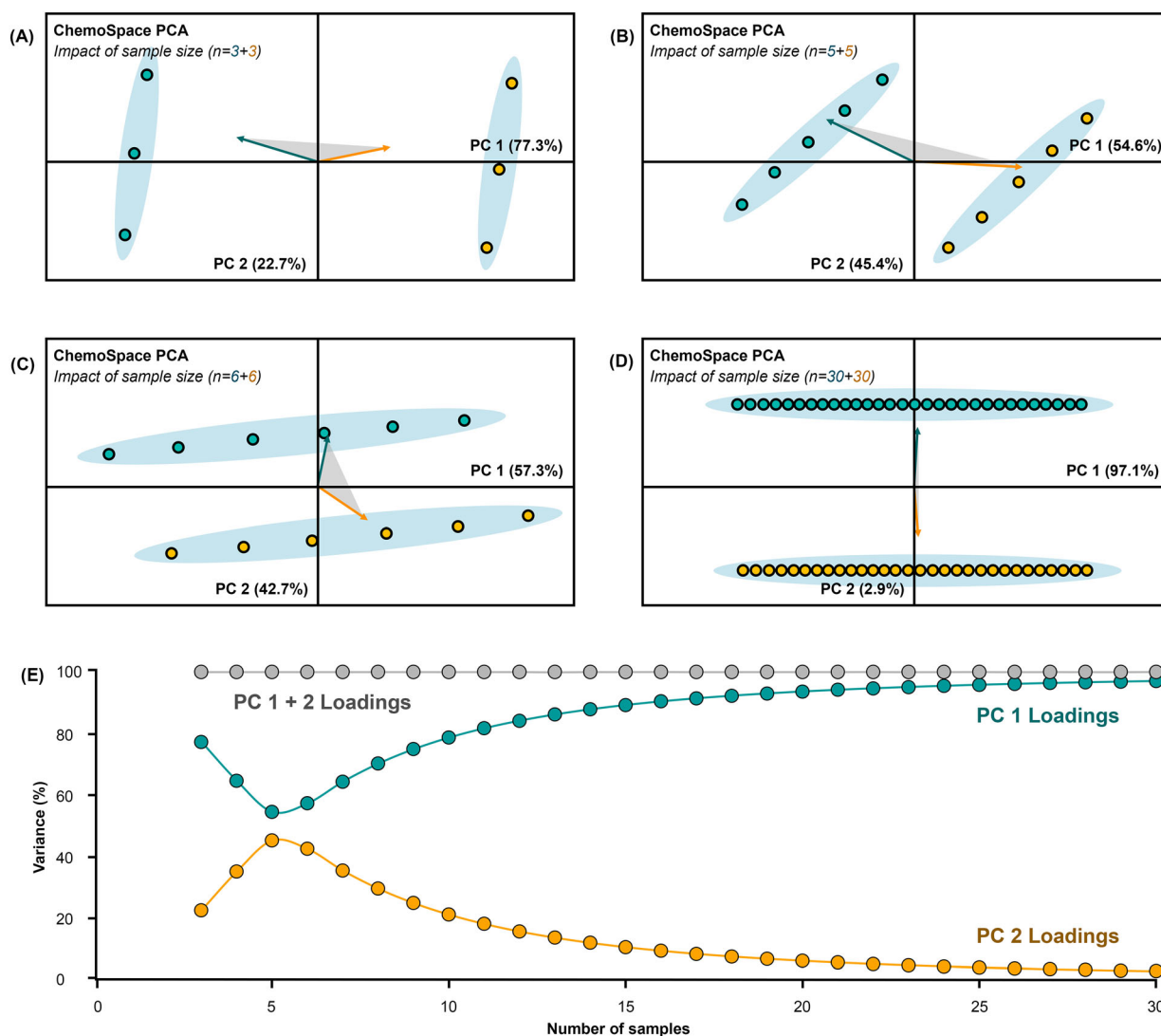
variance corresponding to different biosignatures is commonly expressed by the first two or three principal components (PCs)—the axes of the ChemoSpace displaying different aspects of variance in the data, sorted by descending contributions to the total variance.<sup>1,2,26</sup> Based on the information represented along the PCs and the distribution of eigenvectors that illustrate the impact of individual peaks on the placing of a sample in the ChemoSpace, PCA allows for the exploration and identification of features that are particularly informative for sample grouping and thus represents an essential tool for the subsampling of spectral data points required toward downstream classification or cluster analyses. Co-dependence of individual or overlapping Raman peaks based on molecular connectivity, as well as (spatial) covariance of certain compounds in biological systems, has previously posed an additional challenge (also coined the “cage of covariance”) to the stand-alone interpretation of modern biological ChemoSpaces<sup>28,29</sup>—however, cross-interrogation of complementary spectroscopic data (i.e., Raman and Fourier-Transform Infrared Spectroscopy [FT-IR]) and experimental chemical alteration of individual reference samples offer suitable controls when interpreting compositional spaces.

The impact of analytical variables and different types of unwanted spectral features on classification

approaches to spectroscopic biosignatures in modern and fossil tissues, such as linear discriminant analysis (LDA)<sup>30</sup> and its corresponding machine-learning tools (i.e., support vector machines, SVM),<sup>31,32</sup> is known and has led to a number of end user recommendations,<sup>30–32</sup> but it is only incompletely characterized for the PCA ChemoSpace. Given the potential of the ChemoSpace to address questions in modern biology<sup>1,3,4</sup> and clinical diagnostics,<sup>5–7</sup> and the recent peak in interest by the paleontological,<sup>2,11–18</sup> geological,<sup>33</sup> and astrobiological<sup>33</sup> research communities, a systematic survey of the impact of sample size (Figure 2), spectral signal-to-noise ratios (Figures 3, 4A,C, 5, and 6), spectrometer decalibration (Figure 7), baseline subtraction routines (Figure 8), and normalization procedures (Figure 4B,D) on informative ChemoSpace grouping, accessible to non-specialists from different disciplines, is overdue. In this study, we utilize simplified models of representative tissue spectra to quantify and explain trends in the impact of sample, instrument, and data processing on ChemoSpace occupation and the detectability of compositional biosignatures.

## 2 | METHODS

In order to illustrate the effects of sample size, instrument features, and spectral processing on ChemoSpace



**FIGURE 2** The impact of sample size on the ChemoSpace occupation. The selection of plots aims to showcase the initial ChemoSpace occupation with only  $n = 3$  samples per group (A), the key steps in cluster rotation resulting from an increase in sample number (B, C), the stable ChemoSpace occupation based on  $n = 30$  samples per group (D), and the relationship between the number of samples and the amount of variance represented in the ChemoSpace for this example. Arrows and the shaded area in between them represent eigenvectors in the biplot. (A) ChemoSpace plot resulting from  $n = 3$  varieties (synthetic technical replicates) of two sample types (1: teal; 2: orange). (B) ChemoSpace plot resulting from  $n = 5$  varieties (synthetic technical replicates) of the two sample types (1: teal; 2: orange). (C) ChemoSpace plot resulting from  $n = 6$  varieties (synthetic technical replicates) of the two sample types (1: teal; 2: orange). (D) ChemoSpace plot resulting from  $n = 30$  varieties (synthetic technical replicates) of the two sample types (1: teal; 2: orange). All PC loadings are listed in the ChemoSpace plots. All source data can be found in spreadsheet Table S2. (E) Graph showing the relationship between the number of samples and the amount of variance explained on principal component axis (PC) 1 (teal), PC 2 (orange), and both combined (gray). All source data can be found in Table S2.

occupation, we have selected spectra of two different biological tissues with a simple composition (avian eggshell membrane and avian eggshell [*Gallus domesticus*]), labeled as sample types 1 and 2. For the purpose of experimentation without major signal distortion, these spectra were modified in a number of ways: (1) scaling whole-spectra to generate varieties (5–30, depending on the specific analysis) of the same source signal,

(2) superimposing synthetic sinusoidal wave functions (as simplified representatives of effects related to reflective scattering at tissue layers and features introduced by edge filters) with low and medium frequencies (the latter equal the average Raman band width in the source spectra), (3) superimposing measurements of high-frequency random shot noise (which is common in spectra of biological materials), (4) shifting spectra along the  $x$ -axis

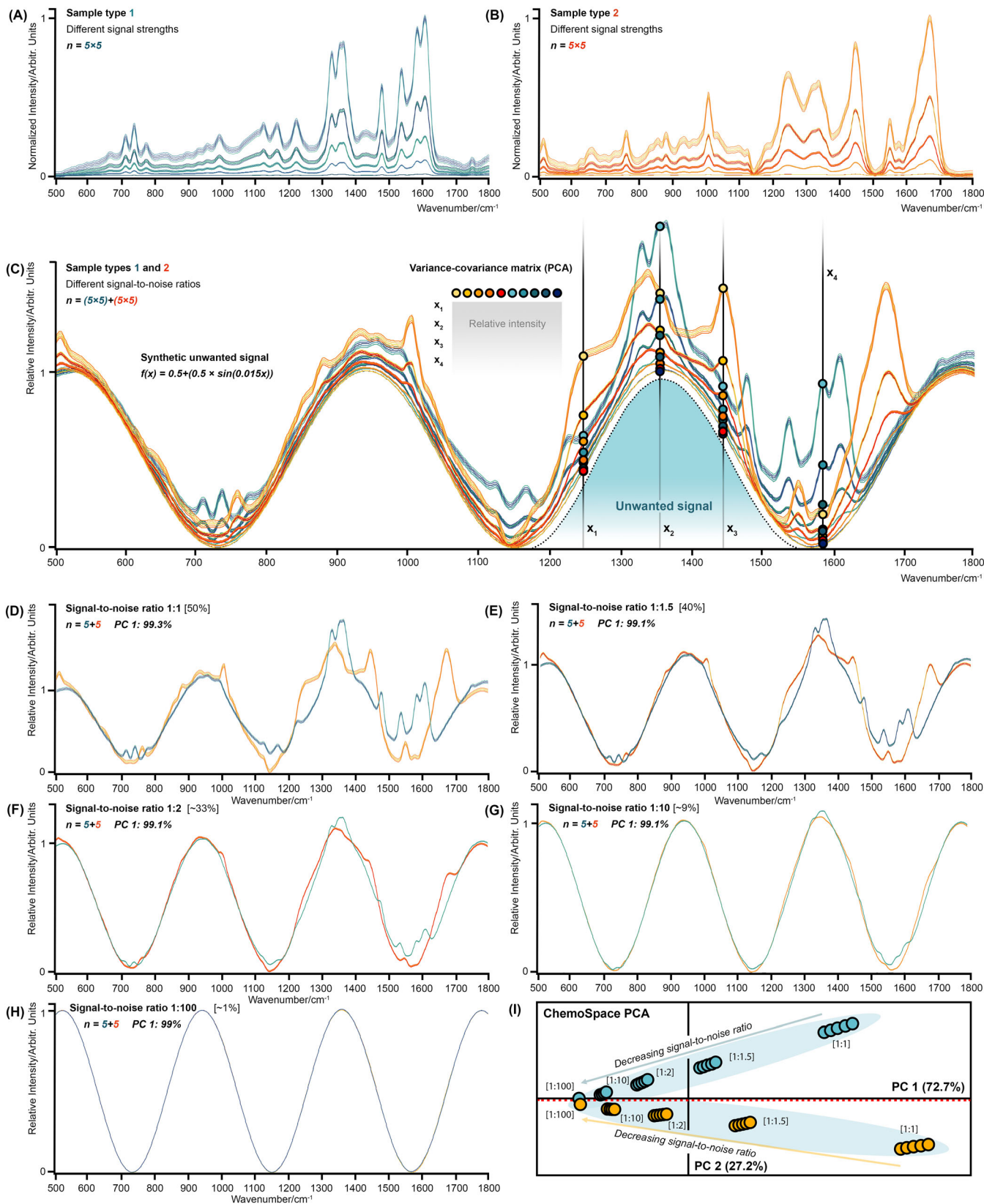


FIGURE 3 Legend on next page.

**FIGURE 3** The impact of systemic low-frequency sinusoidal interference and different signal-to-noise ratios on ChemoSpace occupation. (A) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  differently scaled sets of spectra corresponding to the sample type 1 over the organic fingerprint region. (B) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  differently scaled sets of spectra corresponding to the sample type 2 over the organic fingerprint region. (C) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  differently scaled sets of spectra corresponding to the two sample types (1: teal hues; 2: orange hues) added to the normalized synthetic interference function (for details, see figure or Section 2) over the organic fingerprint region; the interference function represents unwanted spectral features introduced by edge filter ripples, refraction at optical layers within a stratified biological tissues sample, or Mie-ripples. Four Raman band positions are indicated ( $x_1 - x_4$ ), and the colored data points label the mean average intensity of the individual sets of spectra, in order to visually explain how the variance–covariance matrix is built. Signal-to-noise ratios range from  $\sim 1\%$  to 50%. Sets of spectra matching in their signal-to-noise ratio are extracted in (D)–(H). (D) Set of spectra extracted from (C) with a signal-to-noise ratio of 1:1. (E) Set of spectra extracted from (C) with a signal-to-noise ratio of 1:1.5. (F) Set of spectra extracted from (C) with a signal-to-noise ratio of 1:2. (G) Set of spectra extracted from (C) with a signal-to-noise ratio of 1:10. (H) Set of spectra extracted from (C) with a signal-to-noise ratio of 1:100. (I) ChemoSpace across principal components (PCs) 1 and 2 based on a variance–covariance matrix including select relative intensities (see Section 2) extracted from the plot in (C). Data point fill colors correspond to the sample type (1: teal hues; 2: orange hues; compare C). The different signal-to-noise ratios are shown for groups of data points, and the labeled arrows indicate trends in the data distribution across the ChemoSpace. All source data can be found in Tables S1 and S3.

(+1, +0.5, 0,  $-0.5$ ,  $-1 \text{ cm}^{-1}$  offsets), (5) baselining spectra with linear and adaptive approaches (as performed with the SpectraGryph freeware<sup>34</sup>: linear [no offset]; 50%, 30%, 20%, 10% baseline adaptivity options), and (6) normalizing spectra relative to the highest peak. All source data are available in Tables S1–S9.

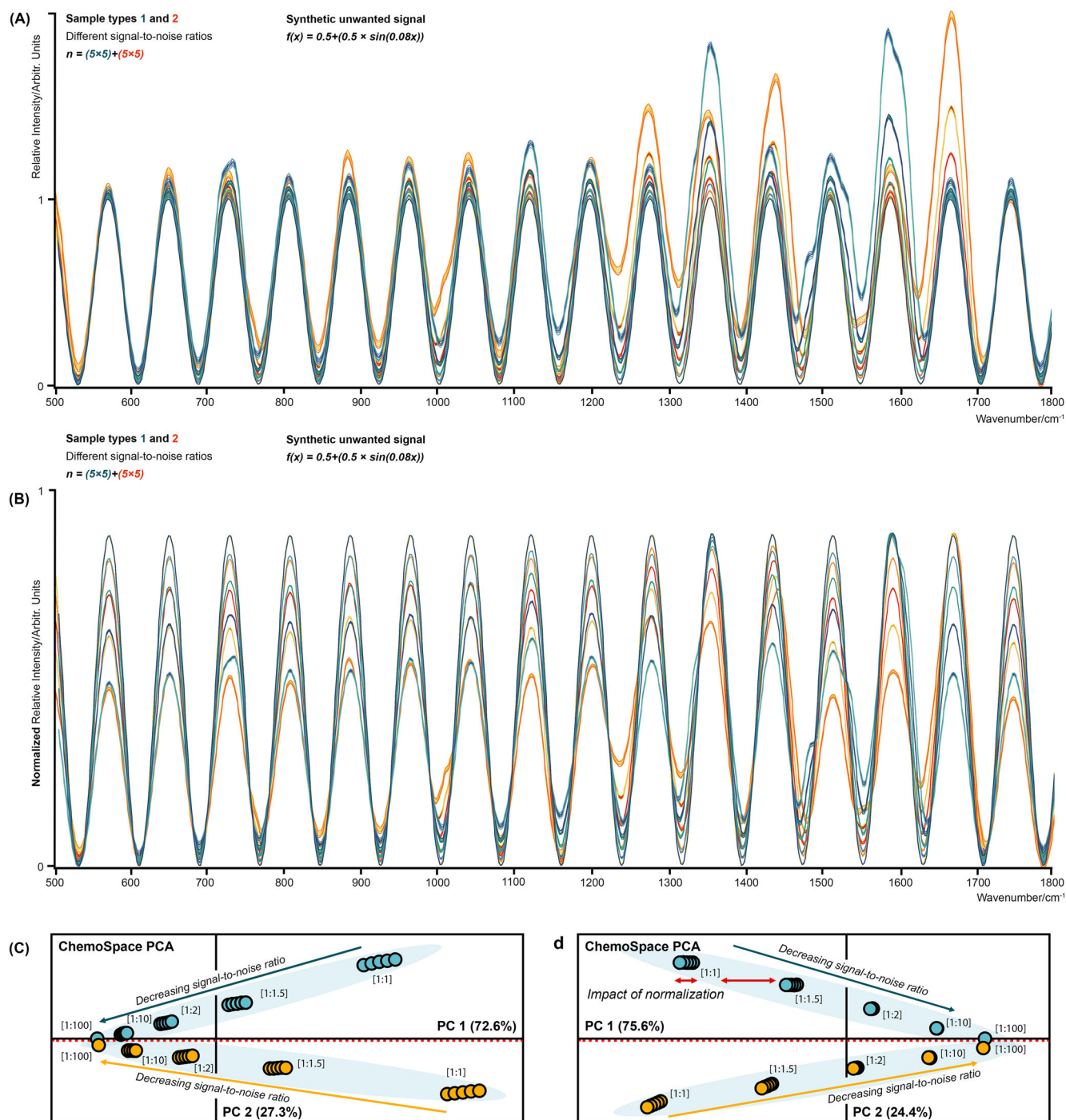
## 2.1 | Impact of the sample: Sample size

Adding samples to a small initial data set is expected to result in rotation of the axis separating the two sample groups as the amount of variation within the groups increases.<sup>25,26</sup> Stable ChemoSpace occupation requires a representative sample, the sample size varying depending on the amount of spectral variance captured in the data set. Various experimental and computational tools can be employed to aid the determination of ideal sample sizes and the critical evaluation of PCA model stability<sup>35–37</sup>; however, here, we aim to showcase the impact of an increasing number of spectra per sample group on ChemoSpace occupation as it is representative for complex biological tissues: 30 scaled varieties of the two source spectra (sample types 1 and 2; Table S1), that is, a total of 60, were generated. The ChemoSpaces resulting from the individual sets of 3, 5, and 6 samples per sample type were plotted to showcase key steps in the axis rotation of groups, as well as the terminal ChemoSpace occupation (set of 30 samples). To do so, all spectra were plotted in SpectraGryph.<sup>34</sup> Relative intensities were extracted from all spectra at 39 Raman band positions: 510, 536, 577, 644, 667, 698, 711, 725, 739, 753, 761, 778, 811, 839, 856, 880, 931, 959, 993, 1005, 1031, 1124, 1165, 1186, 1229, 1249, 1330, 1344, 1356, 1363, 1418, 1445, 1478, 1535, 1550, 1586, 1609, 1676,  $1751 \text{ cm}^{-1}$ . These data

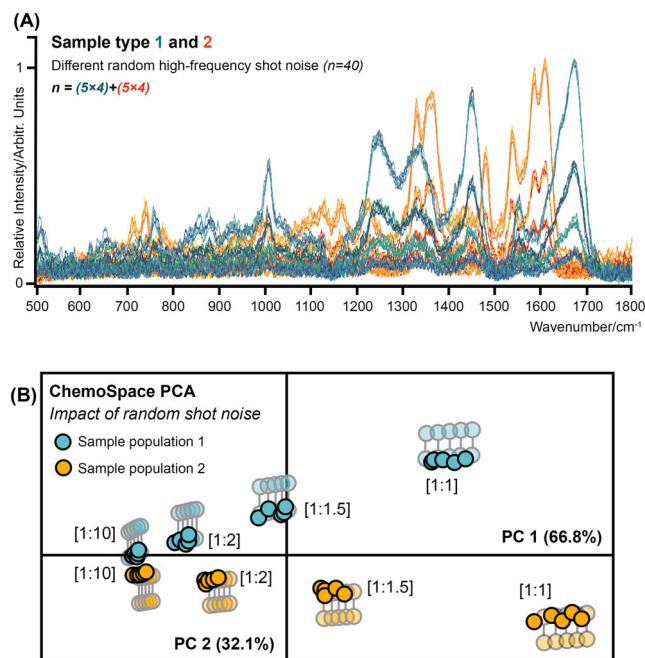
resulted in  $[3 \text{ to } 30] \times [39]$  variance–covariance matrices (Table S2). 2D-ChemoSpaces (Figure 2A–D) and variance captured along the PC axes (PC loadings, Figure 2E) were graphed in PAST 3.0<sup>38</sup> and are shown in Figure 2.

## 2.2 | Impact of the sample and instrument: Systemic unwanted signals

To determine the impact of simplified systemic unwanted signals,<sup>19,20</sup> such as sinusoidal features resulting from reflective scattering at tissue layers or instrument optics (laser-cancelling filters) on ChemoSpace occupation, 5 individual varieties of the two source spectra (sample types 1 and 2) were scaled to 100%,  $\sim 66\%$ , 50%, 10%, 0.1% of their normalized intensity (the highest peak scaled to the value 1) and the results plotted in SpectraGryph (Figure 3A,B). Two sinusoidal interference functions (unwanted signals) were computed, one with a low frequency ( $f(x) = 0.5 + (0.5 \times \sin(0.015x))$ ) (Table S3) and the other one with a medium frequency ( $f(x) = 0.5 + (0.5 \times \sin(0.08x))$ ) matching the average Raman band width in the source spectra (Tables S4 and S5). In addition, high-frequency random shot noise was collected from spectroscopic measurements (40 random noise signals collected over the organic fingerprint region; Table S6). The sets of scaled spectra for sample types 1 and 2 were added to these interference functions, resulting in different signal-to-noise ratios: 1:1 (informative signal content: 50%), 1:1.5 (informative signal content: 40%), 1:2 (informative signal content:  $\sim 33\%$ ), 1:10 (informative signal content:  $\sim 9\%$ ), 1:100 (informative signal content:  $\sim 1\%$ ; not included in the analysis of random shot noise for visualization purposes). The resulting combined signals for low (all spectral varieties in Figure 3C,



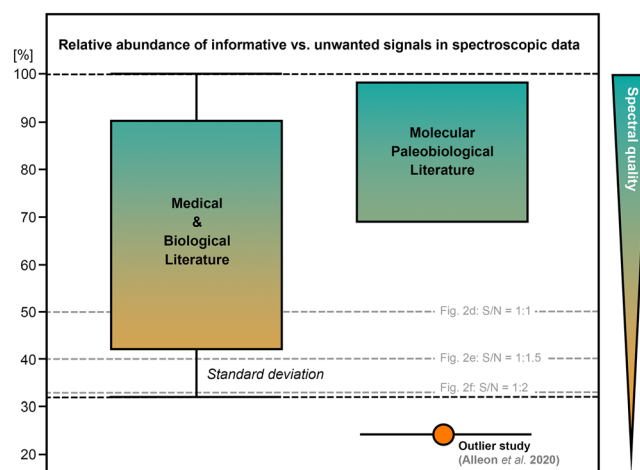
**FIGURE 4** The impact of medium-frequency sinusoidal interference and different signal-to-noise ratios on ChemoSpace occupation. (A) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  differently scaled sets of spectra corresponding to the two sample types (1: teal hues; 2: orange hues), added to the normalized synthetic interference function (for details, see figure) over the organic fingerprint region. (B) The same spectra as in (A), normalized (standard normalization) to the highest peak in each spectrum. (C) ChemoSpace plot across principal components (PCs) 1 and 2 based on a variance–covariance matrix including select relative intensities (see Section 2) extracted from the plot in (A). Data point fill colors correspond to the sample type (1: teal hues; 2: orange hues; compare A). The different signal-to-noise ratios are shown for groups of data points, and the labeled arrows indicate general patterns in the data distribution across the ChemoSpace. (D) ChemoSpace plot across PCs 1 and 2 based on a variance–covariance matrix including select relative intensities (see Section 2) extracted from the plot in (B). Data point fill colors correspond to the sample type (1: teal hues; 2: orange hues; compare B). The different signal-to-noise ratios are highlighted for groups of data points, and the labeled arrows indicate trends in the data distribution across the ChemoSpace. All source data can be found in Tables S4 and S5. The spectral denoising process is showcased in Figure 1.



**FIGURE 5** The impact of high-frequency random shot noise and different signal-to-noise ratios on ChemoSpace occupation. (A) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 4$  differently scaled sets of spectra corresponding to the two sample types (1: teal hues; 2: orange hues), added to the measured high-frequency random shot noise over the organic fingerprint region. (B) ChemoSpace across principal components (PCs) 1 and 2 based on a variance–covariance matrix including select relative intensities (see Section 2) extracted from the plot in (A). Data point fill colors correspond to the sample type (1;2), and the values in the parentheses correspond to the signal-to-noise ratio (compare A). Semi-transparent data points of spectra without added random shot noise have been plotted in the background (and were projected upwards or downwards to reveal the arrangement of data points in the compositional space) for direct comparison with the opaque data points (which are plotted in the foreground) containing random shot noise. All source data can be found in Table S6.

individually scaled subsamples in Figure 3D–H), medium (Figure 4A), and high (Figure 5A) frequency interference were plotted separately in SpectraGryph,<sup>34</sup> and intensities at the selected 39 band positions (listed above) were extracted. The resulting variance–covariance matrices containing low (Figure 3I), medium (Figure 4C), and high (Figure 5B) frequency interference were subjected to PCAs in PAST 3.0,<sup>38</sup> and the resulting PC loadings and 2D-ChemoSpaces were graphed.

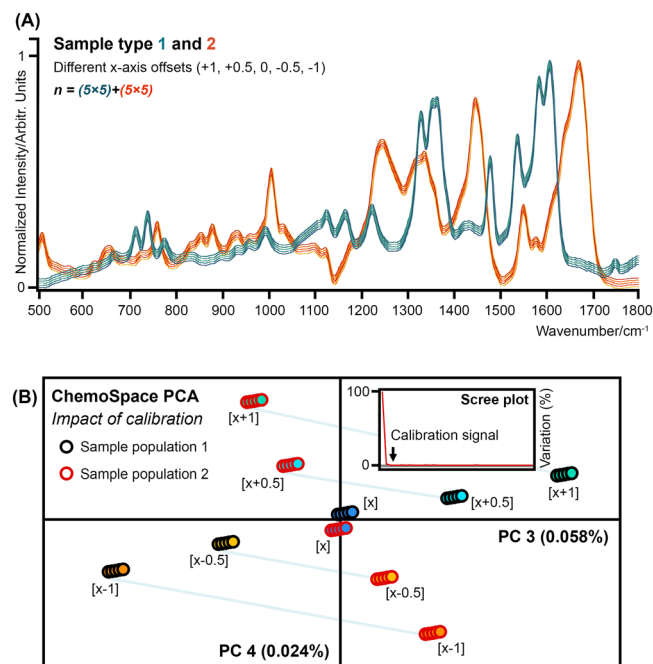
To contextualize and constrain the signal-to-noise ratios in Raman spectra of modern and fossil biological samples, between 3 and 5 (as available in the individual studies) technical replicates of organic Raman spectra published in the fields of medicine, biology, and paleobiology were compiled from the literature (Table S7). Technical replicates (Figure 1A) were plotted



**FIGURE 6** Trends in the relative abundance of informative versus unwanted signals (compare with the signal-to-noise ratio) in spectroscopic data published in the molecular medical and biological literature ( $n = 8$  data sets,  $n = 3$ –5 replicates were analyzed; see Table S7 and the molecular paleobiological literature [ $n = 6$  data sets,  $n = 3$ –5 replicates were analyzed; Table S6]: categories are separated along the x-axis of the plot). The percentage of true compositional signal relative to the total amount of spectroscopic signal, which includes both compositional and unwanted signals, in the published sets of spectra is shown on the y-axis of the plot. The bars associated with the percentage of informative signal in spectra from medical and biological publications represent the standard deviation based on the analyzed spectral sample ( $\pm 1\sigma$ ). For molecular paleobiological studies with sufficient spectral data published alongside the article,<sup>2,11,12,16,18</sup> one outlier study was identified.<sup>20</sup> Signal-to-noise ratios (S/N) corresponding to the listed percentage of informative spectral signals in Figure 3D–F are plotted in form of gray, dashed lines (labeled in the figure). The color gradient in the data bars corresponds to trends in the spectral quality.

in SpectraGryph<sup>34</sup> and whole-spectral data were exported (resolution varies across published data sets) to create corresponding variance–covariance matrices. PC 1 axis loading functions were extracted (Figure 1C), plotted, and normalized together with one of the 3–5 source spectra in SpectraGryph.<sup>34</sup> Integrals of each spectrum and the corresponding PC 1 axis loading function, which represents the true compositional signal, were calculated over the whole spectral range (resolution differs across published data sets). The area under the PC 1 axis loading function was compared with that under the source spectrum containing potential unwanted signals and expressed as the percentage of coverage (Figure 1B). Percentage ranges capturing the relationship between the total spectral signal and the true compositional signal were plotted in PAST 3.0<sup>38</sup> (Figure 5). Figure 1 illustrates the process of denoising the biological tissue spectra through PCA.

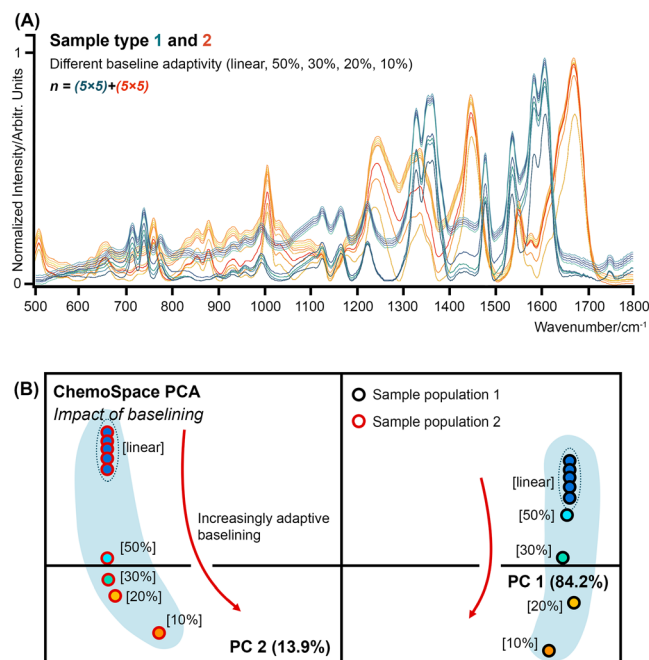




**FIGURE 7** The impact of spectrometer decalibration on the occupation of the ChemoSpace. (A) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  different  $x$ -axis offsets applied to the two sample types (1: teal; 2: orange) over the organic fingerprint region. (B) ChemoSpace across principal components (PCs) 3 and 4 based on a variance–covariance matrix including select relative intensities (see Section 2) extracted from the plot in (A). Data point outline colors correspond to the sample type (1;2), and fill colors correspond to the  $x$ -axis offset (compare A). The scree plot of PC loadings indicates the placement of the decalibration signal. All source data can be found in Table S8.

### 2.3 | Impact of the instrument: Spectrometer decalibration

To characterize how ChemoSpace occupation is impacted by minute spectrometer decalibration that occurs routinely during longer analytical sessions in response to changes in room temperature,<sup>24</sup> the 5 scaled varieties of the two source spectra (sample types 1 and 2) were shifted along the  $x$ -axis as follows: +1, +0.5, +0, -0.5, -1  $\text{cm}^{-1}$ , resulting in a total of  $n = (5 \times 5) + (5 \times 5) = 50$  spectral varieties. All resulting spectra were plotted in SpectraGryph<sup>34</sup> (Figure 6A). A variance–covariance matrix (Table S8) was built based on the extracted intensities of major peaks at the previously introduced 39 band positions. The resulting variance–covariance matrix ( $50 \times 39$ ) was subjected to PCA in PAST 3.0<sup>38</sup> and the (1) variance explained by the calibration signal, (2) sample separation based on calibration differences captured along PCs 3 and 4 in the ChemoSpace, and (3) corresponding scree plot are illustrated in Figure 6B.



**FIGURE 8** The impact of baseline subtraction on the occupation of the ChemoSpace. (A) Plot of  $n = 5$  varieties (synthetic technical replicates) of  $n = 5$  different baseline subtraction approaches (in SpectraGryph<sup>26</sup>: linear, 50%, 30%, 20%, 10%) applied to the two sample types (1: teal hues; 2: orange hues) over the organic fingerprint region. (B) ChemoSpace plot across principal components (PCs) 1 and 2 based on variance–covariance matrix including relative intensities (see Section 2) extracted from the spectra in (A). Data point outline colors correspond to the sample type (1: black; 2: red), and the fill colors correspond to the different baseline subtraction routines (labeled in the figure). Red arrows point toward increasingly adaptive baselines. All source data can be found in Table S9.

### 2.4 | Impact of spectral processing: Spectral baselining

Baseline subtraction is an established approach<sup>1</sup> employed to increase the comparability of spectra when background signals differ across samples. Background shapes differ substantially in sets of spectra collected from, i.e., modern and fossil biological tissues. To capture the influence of baselining on ChemoSpace occupation, 5 varieties of the two source spectra (sample types 1 and 2) were subjected to the linear option and the 50%, 30%, 20%, and 10% adaptive baselining options (no  $y$ -axis offset in either case) in SpectraGryph. All  $n = (5 \times 5) + (5 \times 5) = 50$  resulting spectra were plotted in SpectraGryph (Figure 7A). Excessive ( $\leq 20\%$  in SpectraGryph) baseline adaptivity leads to partial subtraction of signal associated with the highest peaks in the spectra and alters the ratio of normalized signal intensities that encode biosignatures. Relative intensities at the same

39 band positions (introduced above) were extracted from all spectra and incorporated into a [50] x [39] variance-covariance matrix (Table S9). PCA was performed in PAST 3.0<sup>38</sup> to capture the impact of different baselines on ChemoSpace occupation reflected in PC loadings and sample position in the ChemoSpace plot based on PCs 1 and 2 (Figure 7B).

## 2.5 | Impact of spectral processing: Spectral normalization

Normalization scales a spectrum based on the highest peak, a particular selected peak, or the area under the spectral curve.<sup>1,34</sup> It is commonly applied prior to any quantitative analysis<sup>1</sup> to increase comparability across spectra given the variability of absolute Raman intensities among diverse samples. The combined set of 50 varieties of spectra containing the synthetic, medium-frequency interference (introduced above) was plotted (Figure 3A) to capture the impact of normalization on ChemoSpace occupation. The highest peak of each spectrum was scaled to a value of 1 (a common approach) using the SpectraGryph<sup>34</sup> normalization option (Figure 3B). Relative intensities were extracted from all spectra at the 39 wavenumber positions generating a [50] × [39] variance-covariance matrix. Figure 3D shows the resulting PC loadings and ChemoSpace plot based on PCs 1 and 2.

## 3 | RESULTS AND DISCUSSION

The effects of sample size, instrument decalibration, and spectral processing on ChemoSpace occupation were simulated and showcased in six distinct experiments. Minute changes in spectrometer calibration, the systemic presence of unwanted signal, differences in the spectral signal-to-noise ratios, linear and standard adaptive baseline subtraction, and spectral normalization do not overprint the biologically informative grouping of tissue samples in the ChemoSpace. Spectral processing, including baseline correction and normalization prior to PCA, improved data comparability and biosignature separation. A near-stable pattern in ChemoSpace occupation is, in this example, reached with as few as 6 spectra per sample type.

### 3.1 | Stable ChemoSpace occupation can be reached with less than 10 samples

The number of samples required to achieve a stable ChemoSpace occupation is as few as 6 per sample type in

this data set representing biological tissues (Figure 2E). With 12 samples, the two clusters are separated across PC 2 which accounts for 42.7% of the variance in the data set, whereas intra-group variance accounts for 57.3% of the total and is captured on PC 1. In contrast to PC loadings, eigenvectors in the ChemoSpace biplot (teal and orange arrows in Figure 2A–D) allow the sources of variance in the data, including biological signals within and across tissues, to be differentiated even when cluster separation occurs diagonally in the ChemoSpace. Such eigenvector trajectories allowed us to infer that rotation of the axis separating sample clusters in this ChemoSpace results from an increase in the contribution of intra-group variance to the total variance as spectra are added: intra-group variance becomes the primary source of variance and is displayed along PC 1. This experiment suggests that the sampling strategy should reflect the scientific question of interest: in integrative data sets including modern and fossil tissues, ChemoSpace grouping will account more accurately for variation in different modes of (diagenetic) alteration of a biological tissue, for example, when an increasing number of fossil samples from different depositional settings is considered. The sample set analyzed here is not supposed to provide a generalizable model that can be directly transferred to other data sets, but rather aims to showcase and explain trends in the relationship between sampling strategy and ChemoSpace occupation.

### 3.2 | PCA allows biosignatures to be detected in a ChemoSpace even if systemic unwanted signals are present in spectra

PCA as employed here is based on a variance-covariance matrix. The focus on variance rather than qualitative comparisons of absolute spectral differences (Figure 3C) facilitates the detection of biologically meaningful sample grouping, even when prominent unwanted signals, such as sample- or instrument-related spectral features,<sup>19,20</sup> are present. In addition, the extraction of relative intensities at informative wavenumber positions allows features relevant to a given question to be emphasized.<sup>27</sup> A mostly stable, omnipresent interference signal is unlikely to become the primary source of variance in a diverse data set, such as a sample of different tissues. PCA reliably separates the two clusters corresponding to signals 1 and 2, regardless of the frequency of a periodic, systemically present, unwanted signal, or the total spectral signal-to-noise ratio (Figures 3I and 4C). An omnipresent and invariant unwanted signal will not overprint compositional biosignatures in a ChemoSpace PCA, even if it includes more complex spectral features, such as,

i.e., signals associated with quartz glass slides. Random high-frequency shot noise at realistic intensity, as modeled in Figure 5A, is shown to lead to minor displacements of individual data points in the compositional space; however, it does not overprint compositional biosignatures separating the two sample groups (Figure 5B).

Although a decrease in the signal-to-noise ratio of spectral data results in increased convergence of the two sample clusters in the ChemoSpace (regardless of the nature of unwanted signal present), clusters are separated even for spectra with a signal-to-noise ratio of 1:100 (Figures 3H–I and 4C). The spectra modeled here in Figures 4 and 5 (with signal-to-noise ratios ranging from 1:100 to 1:1) include a higher amount of unwanted signal than most published Raman spectra. Informative spectral content ranges from ~42% to 90% ( $\pm 1\sigma$ ) in the biological and medical literature (Figure 6, based on 8 spectral data sets, 3–5 replicates were analyzed) and ~69% to 98% in the molecular paleobiological literature<sup>2,11–14,16–19</sup> excluding one statistical outlier<sup>20</sup> (Figure 6, based on 6 spectral data sets, 3–5 replicates were analyzed). Field-specific ranges of compositional signal content in published spectra mostly overlap. Comparatively high signal-to-noise ratios in carbonaceous fossilization products of biological tissues are the result of smoother textures following dehydration and compaction, as well as reduced fluorescence (Figure 6). Regardless of the type of unwanted signal present in biological or geological organic Raman spectra, PCA reliably extracts informative features (see Figure 1 for denoising).

### 3.3 | Minor in-session spectrometer decalibration does not overprint ChemoSpace biosignatures

Spectrometer decalibration accounting for  $\pm 1 \text{ cm}^{-1}$  wavenumber is only evident across PCs 3 and 4 (Figure 7) and explains less than 0.1% variance in this data set. Thus, any type of biosignature accounting for more than 0.06% variance (loading PC 3) in the data set will outweigh the decalibration signal in the ChemoSpace. All previously published spectroscopic biosignatures<sup>5–7,11–18</sup> exceed the amount of variance resulting from decalibration by at least two orders of magnitude.

### 3.4 | Standard adaptive baselining increases comparability and ChemoSpace signal extraction

It is essential to subtract spectral backgrounds without affecting informative bands, in order to prevent

differences in background shape from appearing as a major source of variance which could potentially overprint biosignatures.<sup>1</sup> Linear baseline subtraction (Figure 8A) does not completely remove nonlinear background signals, which are common in spectra of heterogeneous and stratified biological tissues, and may introduce or amplify spectral incomparability (see linear baseline subtraction applied to sample spectrum 1 in Figure 8A). Adaptive baselining, in contrast, eliminates all types of background signal (Figure 8A) regardless of shape. Baselining may result in minor spatial convergence of informative clusters (sample types 1 and 2) in the ChemoSpace (Figure 8B), if adaptivity exceeds the standard (less of the original spectral signal remains; threshold determined here:  $< 30\%$  in SpectraGryph<sup>34</sup>): as baseline adaptivity increases beyond the standard, broad Raman bands of high intensity lose comparatively more signal than narrow bands with relatively low intensity (Figure 8A). This loss of the biological signal encoded in informative band ratios decreases the separation of groups in the ChemoSpace (indicated by the red arrows in Figure 8B). Standard adaptive baselines (threshold:  $\geq 30\%$  in SpectraGryph<sup>34</sup>) increase intra-group comparability without cluster convergence, resulting in the collapse of individual spectral data points within a sample type in the ChemoSpace (Figure 7B).

### 3.5 | Normalization increases comparability and improves signal

Spectral normalization emphasizes key differences within a sample set by amplifying differences in relative spectral intensities (Figure 4B). The ChemoSpace PCA shows how normalization increases direct comparability (all normalized spectra share a highest peak scaled to 1; they do not range in intensity counts over orders of magnitude like in the raw spectra) across synthetic replicates, as demonstrated by the closer grouping of data points within a subsample (Figure 4D). Normalization also homogenizes the distribution of data within clusters associated with signals 1 and 2—an inference based on the resulting uniform data point spacing (Figure 4D) compared to the non-uniform data point spacing observed among non-normalized spectra (Figure 4C). Most biosignatures are encoded in the relative abundance of different functional groups,<sup>1,2</sup> so spectral normalization (based on the highest peak in the spectrum, a different informative peak, or a spectral area) facilitates the extraction of meaningful signal. The suitability of different modes of normalization for individual data sets depends on the specific question and the nature and comparability of spectral intensities in the sample set.

## 4 | CONCLUSIONS

Quantification of the impact of sample size, instrument features, and spectral processing on the occupation of the ChemoSpace provides an analytical framework for the extraction of molecular biosignatures from spectroscopic fingerprints of tissues from extant and extinct organisms: Minor instrument decalibration during an analytical session does not overprint major biological signatures in a ChemoSpace PCA. Spectral processing routines, such as standard adaptive baseline subtraction, as well as normalization prior to statistical analysis of spectra, increase data comparability and facilitate the extraction of informative features. Stable ChemoSpace occupation can be achieved with fewer than 10 spectra per sample group when analyzing biosignatures. PCA facilitates the distinction of informative compositional and systemic unwanted signals, regardless of the waveform, periodicity, frequency, and amplitude of a spectral interference, even at relatively low signal-to-noise ratios. The ChemoSpace approach to biosignatures represents a powerful tool for exploring, denoising, and integrating information from modern and ancient organismal samples.

## ACKNOWLEDGMENTS

The authors thank D. Briggs for helpful comments and edits and J. Eiler, M. Brown, and G. Rossman for helpful conversations.


## CONFLICT OF INTEREST STATEMENT

We declare no competing interests.

## DATA AVAILABILITY STATEMENT

All source data, synthetic interference functions, and corresponding variance–covariance matrices are available in the Supporting Information spreadsheet (Tables S1–S9) and are intended to be published alongside our article.

## ORCID

Jasmina Wiemann  <https://orcid.org/0000-0003-3581-1711>

## REFERENCES

- [1] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, *Nat. Protoc.* **2016**, *11*(4), 664.
- [2] J. Wiemann, J. M. Crawford, D. E. Briggs, *Sci. Adv.* **2020**, *6*(28), eaba6883.
- [3] A. C. S. Talari, Z. Movasaghi, S. Rehman, I. U. Rehman, *Appl. Spectr. Rev.* **2015**, *50*(1), 46.
- [4] R. Manoharan, Y. Wang, M. S. Feld, *Spectrochim. Acta, Part a* **1996**, *52*(2), 215.
- [5] A. Mahadevan-Jansen, R. R. Richards-Kortum, *J. Biomed. Opt.* **1996**, *1*(1), 31.
- [6] A. C. S. Talari, C. A. Evans, I. Holen, R. E. Coleman, I. U. Rehman, *J. Raman Spectrosc.* **2015**, *46*(5), 421.
- [7] X. Li, T. Yang, S. Li, D. Wang, Y. Song, S. Zhang, *Laser Phys.* **2016**, *26*(3), 035702.
- [8] J. D. Pasteris, O. Beyssac, *Elements* **2020**, *16*(2), 87.
- [9] A. Olcott Marshall, C. P. Marshall, *Palaeontology* **2015**, *58*(2), 201.
- [10] J. W. Schopf, A. B. Kudryavtsev, D. G. Agresti, T. J. Wdowiak, A. D. Czaja, *Nature* **2002**, *416*(6876), 73.
- [11] M. A. Norell, J. Wiemann, M. Fabbri, C. Yu, C. A. Marsicano, A. Moore-Nall, D. J. Varricchio, D. Pol, D. K. Zelenitsky, *Nature* **2020**, *583*(7816), 406.
- [12] V. E. McCoy, J. Wiemann, J. C. Lamsdell, C. D. Whalen, S. Lidgard, P. Mayer, H. Petermann, D. E. Briggs, *Geobiology* **2020**, *18*(5), 560.
- [13] M. Tripp, J. Wiemann, J. Brocks, P. Mayer, L. Schwark, K. Grice, *Biology* **2022**, *11*(9), 1289.
- [14] M. Fabbri, J. Wiemann, F. Manucci, D. E. Briggs, *Palaeontology* **2020**, *63*(2), 185.
- [15] C. C. Loron, E. Rodriguez Dzul, P. J. Orr, A. V. Gromov, N. C. Fraser, S. McMahon, *Nat. Commun.* **2023**, *14*(1), 1387.
- [16] J. Wiemann, T. R. Yang, M. A. Norell, *Nature* **2018**, *563*(7732), 555.
- [17] J. Wiemann, I. Menéndez, J. M. Crawford, M. Fabbri, J. A. Gauthier, P. M. Hull, M. A. Norell, D. E. Briggs, *Nature* **2022**, *606*(7914), 522.
- [18] J. Wiemann, M. Fabbri, T. R. Yang, K. Stein, P. M. Sander, M. A. Norell, D. E. Briggs, *Nat. Commun.* **2018**, *9*(1), 4741.
- [19] J. Wiemann, D. E. Briggs, *Bioessays* **2022**, *44*(2), 2100070.
- [20] J. Alleon, G. Montagnac, B. Reynard, T. Brulé, M. Thoury, P. Gueriau, *Bioessays* **2021**, *43*(4), 2000295.
- [21] S. J. Choquette, E. S. Etz, W. S. Hurst, D. H. Blackburn, S. D. Leigh, *Appl. Spectrosc.* **2007**, *61*(2), 117.
- [22] J. D. Rodriguez, B. J. Westenberger, L. F. Buhse, J. F. Kauffman, *Analyst* **2011**, *136*(20), 4232.
- [23] T. Dörfer, T. Bocklitz, N. Tarcea, M. Schmitt, J. Popp, *Zeitschrift für Physikalische Chemie* **2011**, *225*(6–7), 753.
- [24] D. Hutsebaut, P. Vandenabeele, L. Moens, *Analyst* **2005**, *130*(8), 1204.
- [25] H. Abdi, L. J. Williams, *Wiley Interdisc. Rev. Comput. Stat.* **2010**, *2*(4), 433.
- [26] M. L. Zelditch, D. L. Swiderski, H. D. Sheets, *Geometric morphometrics for biologists: a primer*, Elsevier Academic Press, New York and London, **2012**.
- [27] P. D. Lewis, G. E. Menzies, *Vib. Spectrosc.* **2015**, *81*, 62.
- [28] C. E. Eskildsen, T. Næs, P. B. Skou, L. E. Solberg, K. R. Dankel, S. A. Basmoen, J. P. Wold, S. S. Horn, B. Hillestad, N. A. Poulsen, M. Christensen, *Chemom. Intel. Lab. Syst.* **2021**, *213*, 104311.
- [29] D. T. Berhe, C. E. Eskildsen, R. Lametsch, M. S. Hviid, F. van den Berg, S. B. Engelsen, *Meat Sci.* **2016**, *111*, 18.
- [30] W. Lee, A. T. Lenferink, C. Otto, H. L. Offerhaus, *J. Raman Spectrosc.* **2020**, *51*(2), 293.
- [31] M. Sattlecker, N. Stone, J. Smith, C. Bessant, *J. Raman Spectrosc.* **2011**, *42*(5), 897.
- [32] C. L. Morais, K. M. Lima, M. Singh, F. L. Martin, *Nat. Protoc.* **2020**, *15*(7), 2143.

- [33] O. Beyssac, *Elements Int. Mag. Mineral. Geochem. Petrol.* **2020**, 16(2), 117.
- [34] F. Menges, **2020**. Spectragryph Optical Spectroscopy Software, Version 1.2. 14. Oberstdorf, Germany 2019. Available online: <https://www.ffmpeg2.de/spectragryph/>
- [35] N. Ali, S. Girnus, P. Rösch, J. Popp, T. Bocklitz, *Anal. Chem.* **2018**, 90(21), 12485.
- [36] S. Guo, P. Rösch, J. Popp, T. Bocklitz, *J. Chemometr.* **2020**, 34(4), e3202.
- [37] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp, *Anal. Chim. Acta* **2011**, 704(1–2), 47.
- [38] Ø. Hammer, D. A. Harper, P. D. Ryan, *Palaeontologia Electronica* **2001**, 4(1), 9.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** J. Wiemann, P. R. Heck, *J Raman Spectrosc* **2024**, 1. <https://doi.org/10.1002/jrs.6669>