




# Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling

Shuxia Guo<sup>1,2,3</sup>, Jürgen Popp<sup>2,3</sup> and Thomas Bocklitz<sup>2,3</sup> 

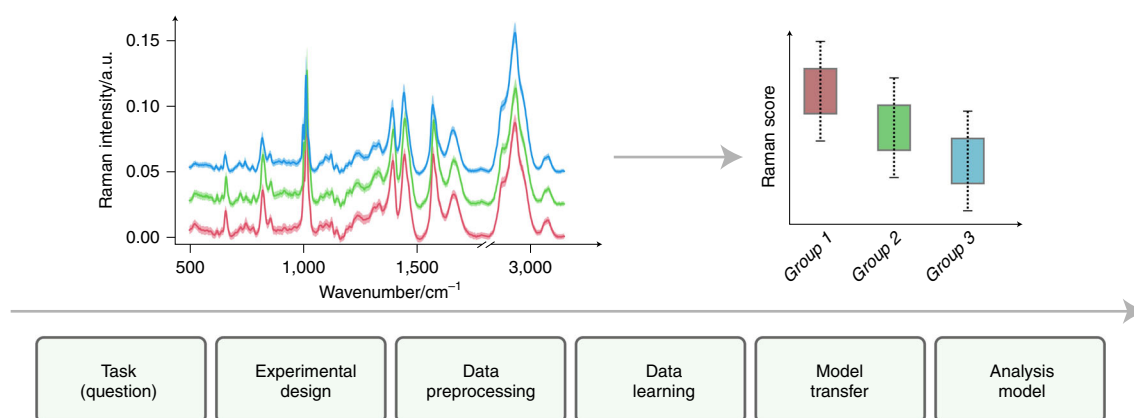
Raman spectroscopy is increasingly being used in biology, forensics, diagnostics, pharmaceuticals and food science applications. This growth is triggered not only by improvements in the computational and experimental setups but also by the development of chemometric techniques. Chemometric techniques are the analytical processes used to detect and extract information from subtle differences in Raman spectra obtained from related samples. This information could be used to find out, for example, whether a mixture of bacterial cells contains different species, or whether a mammalian cell is healthy or not. Chemometric techniques include spectral processing (ensuring that the spectra used for the subsequent computational processes are as clean as possible) as well as the statistical analysis of the data required for finding the spectral differences that are most useful for differentiation between, for example, different cell types. For Raman spectra, this analysis process is not yet standardized, and there are many confounding pitfalls. This protocol provides guidance on how to perform a Raman spectral analysis: how to avoid these pitfalls, and strategies to circumvent problematic issues. The protocol is divided into four parts: experimental design, data preprocessing, data learning and model transfer. We exemplify our workflow using three example datasets where the spectra from individual cells were collected in single-cell mode, and one dataset where the data were collected from a raster scanning-based Raman spectral imaging experiment of mice tissue. Our aim is to help move Raman-based technologies from proof-of-concept studies toward real-world applications.

## Introduction

Raman spectroscopy is a photonic technique that is based on vibrational Raman scattering. Raman scattering is only observed in a tiny proportion ( $<1$  in  $10^6$ ) of the photons scattered by a molecule, and the Raman scattered photons have a different frequency to that of the incident photons used for excitation. The energy exchange between the incident photons and the molecule leads to an energy change of the photons, either gain or loss, and the energy change depends on the energy states of the molecule. In most cases, Raman spectroscopy is measuring the different vibrational states in the molecular components of the sample. Looking at the spectra, the  $x$  axis (i.e., the spectral axis) shows the frequency differences (i.e., the energy change) between the incident and the Raman scattered photons, which are represented as wavenumbers in the unit of inverse centimeters ( $\text{cm}^{-1}$ ). The  $y$  axis denotes the intensities of the Raman scattered light at different wavenumber positions. The position of any Raman band represents the energy of a molecular vibration, while its intensity is proportional to the number of the corresponding vibrations in the sample. These two facts form the basis of Raman spectroscopy applied in molecular identification and quantitative analysis, respectively<sup>1,2</sup>, because these facts suggest that a Raman spectrum is the linear combination of the Raman spectra of the compounds. This quasi Beer law assumes a linear relationship between the concentration of each compound and its contribution to the spectrum. This assumption allows a simple, useful model if the sample consists of a few well-defined compounds that are present at concentrations within their linear response range, but it is not easily applicable if complex samples are studied, e.g., biological or medical samples.

Nevertheless, a Raman spectrum of a biological sample still represents its vibrational information and can be considered as a unique fingerprint of the sample. Raman spectroscopy has, therefore,

<sup>1</sup>Institute for Brain and Intelligence, Southeast University, Nanjing, China. <sup>2</sup>Leibniz Institute of Photonic Technology Jena (IPHT Jena), Member of Leibniz Health Technologies, Jena, Germany. <sup>3</sup>Institute of Physical Chemistry and Abbe Centre of Photonics, Friedrich Schiller University of Jena, Jena, Germany. ✉e-mail: [thomas.bocklitz@uni-jena.de](mailto:thomas.bocklitz@uni-jena.de)



**Fig. 1 | Overview of the Raman spectroscopic analysis protocol.** The protocol is divided into four blocks: experimental design, data preprocessing, data learning and model transfer. The result of the pipeline is the transfer into higher-level information.

proven versatile in many biological and medical investigations<sup>3,4</sup> such as microbial identification<sup>5–9</sup>, characterization of metabolisms<sup>10–12</sup>, medical diagnostics<sup>13–15</sup>, and forensics<sup>16–18</sup>. Medical samples being studied by Raman spectroscopy include subcellular structures, cells and tissues. In most cases, Raman spectroscopy is used to find out the group membership of a sample or the concentration of a certain substance within the sample. The underlying assumption is that the differences between the samples manifest themselves as the position shifts and/or the intensity variations of certain Raman bands within Raman spectra.

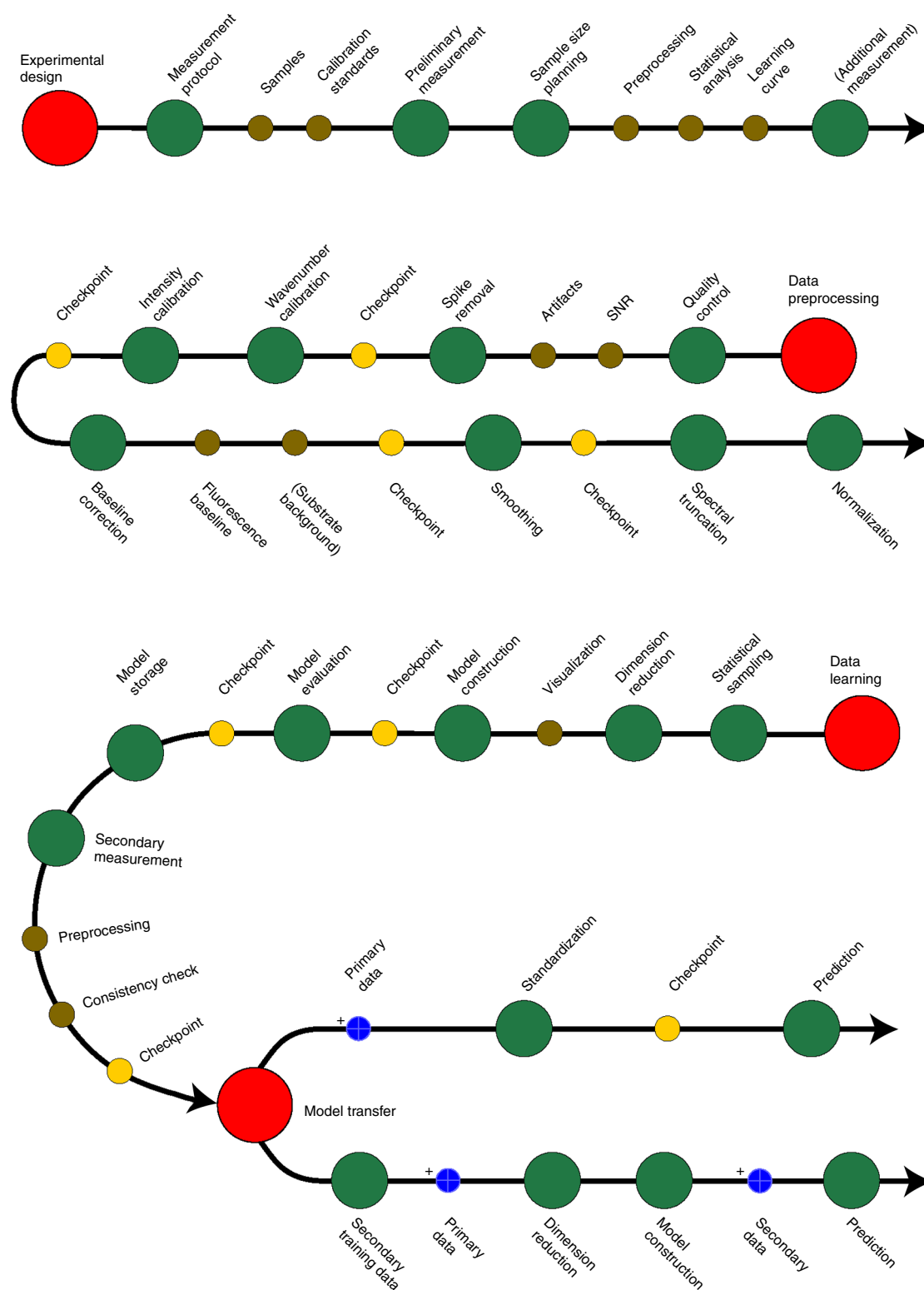
However, such spectral variations are often very subtle and easily masked by instrumental drifts, measurement errors or other artifacts such as the fluorescence background<sup>19–21</sup>. Spectral variations may also exist among different replicates of a sample or the same sample at different time points, because of unavoidable biological changes, sample degradation, variations in sample preparation, or any other uncontrollable experimental factor. All these aspects make Raman spectra hard to interpret without the support of chemometrics, which has become indispensable to remove the corrupting effects and to extract subtle spectral variations of interest from measured Raman spectra. Both are required for enhancing the sensitivity of Raman spectroscopy, which is essential for biological investigations.

Chemometrics techniques have been widely applied in the fields of chemistry, biology and biochemistry<sup>22,23</sup>. Many chemometric tasks can be achieved with machine learning methods, for instance, to group samples via clustering, to classify samples using a classification method or to derive the concentration of a substance with a regression model. However, chemometrics is not only about clustering, classification or regression. Rather, it is a sequence of procedures including experimental design, data processing, data learning and data interpretation. To conduct all these steps is not an easy task but requires intensive support from disciplines of machine learning, statistics, data management and natural science like chemistry, biology and physics. This is especially relevant for Raman-based biological applications, due to both the complexity of Raman spectroscopy and the properties of biological samples. There are many challenges and pitfalls worth taking a close look at in terms of chemometrics in Raman-based biological investigations. To address these points, we present a protocol composed of four sections: experimental design, data preprocessing, data learning and model transfer (Figs. 1 and 2)<sup>24</sup>.

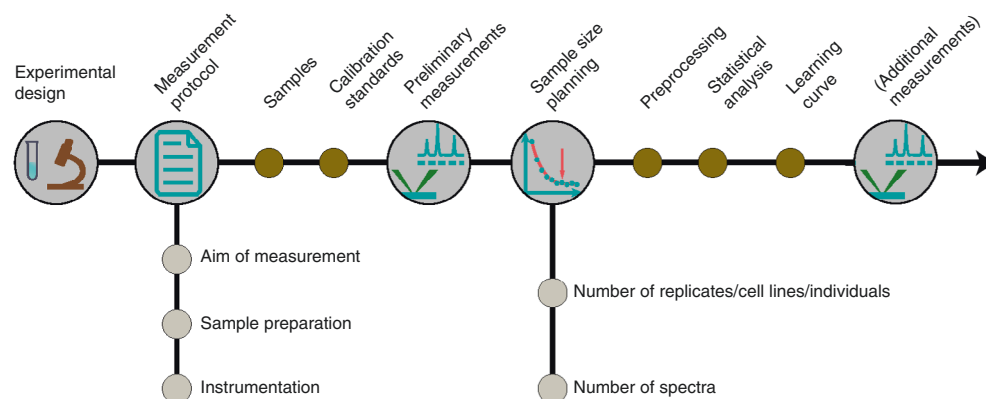
In Fig. 2, the milestone steps are marked by dark green spheres, while the key procedures under each step are given by brown spheres. The blue spheres along with a plus sign represent data to be input. We have also inserted checkpoints (yellow spheres) where a quality control step is necessary to avoid errors and ensure a meaningful data analysis.

The details on each of the four blocks are discussed in the next subsections, including their respective requirements and cautions, challenges and solutions. These discussions, however, are not restricted to any specific methods. Rather, they aim to provide general rules of each milestone step that should not be violated in Raman spectral analysis.

As long as these guidelines are followed, the users are free to choose how they perform each step according to their own requirements, preference and convenience. In the ‘Anticipated results’ section, we will show results obtained using four example datasets: (1) Raman spectra of tumor cells;



**Fig. 2 | Detailed overview of the Raman spectroscopic analysis protocol.** The protocol is divided into four blocks: experimental design, data preprocessing, data learning and model transfer. The experimental design covers the topics of measurement strategy and SSP. Spectral preprocessing includes all steps that are needed to remove corrupting effects in the measured Raman spectra and standardize the data. Data learning aims to build a model based on given data, while model transfer is included to deal with the failed prediction of a model built on newly measured data. SNR, signal-to-noise ratio.



**Fig. 3 | Workflow of the experimental design.** The experimental design includes two major parts. The measurement protocol describes the aim(s) of the measurements/study, the way of sample preparation and the instrument to be used. The SSP determines the minimal number of samples required to ensure a statistically meaningful outcome of the measurements/study.

(2) Raman spectral data of bacterial spores; (3) Raman spectra of vegetative bacteria; (4) Raman spectral data of colon tissue.

These examples represent typical scenarios in biological applications and were selected to demonstrate different issues and their solutions, including:

- Model training and evaluation
- Spectrometer calibration and model transfer

The spectrometer calibration tries to make Raman data independent of a Raman device used to measure it, while model transfer tries to make a model insensitive to the changes from device to device. Both of these methods are important for comparing Raman spectra obtained from different instruments, laboratories and experiments. The whole protocol focuses on the experimental design and the data analysis; for methods relating to sample preparation and data collection, please refer to the literature (for example, refs. <sup>25,26</sup>), including other protocol articles.

### Experimental design

Proper design of experiments (DOE) is essential for successful measurements and studies, because well-designed studies and measurements are the prerequisites to clearly answer questions under investigation<sup>27</sup>. From the experimental point of view, this means a spectroscopic model of the sample needs to be developed, which suggests whether the spectra can (potentially) answer the question under investigation. This can be generated using preliminary data or via theoretical investigations. From the chemometrics point of view, DOE is mainly composed of two parts as shown in Fig. 3: the measurement protocol and the sample size planning (SSP). The details will be given in the following.

### Measurement protocol

The measurement protocol is a clear statement about the question/aim/subject of an investigation, the procedures of sample preparation, the instruments to be used, the strategies of sampling and other aspects<sup>26</sup>. The question/aim/subject of a study has to be defined as clearly as possible from the very beginning. From a data analysis point of view, this definition means that the following sequence of questions must be answered during the development phase:

- What is the task (exploration to find biomarkers of a given problem, un-mixing to resolve different biochemical components from the samples under investigation, regression to derive the concentration of a substance, or clustering/classification to group samples into different categories)?
- What are the properties of the sample (e.g., levels of fluorescent emission, homogeneity, Raman features of interest)?
- What is the basic unit of one sample, i.e., the highest hierarchy of the sampling (one patient, one cell line, one replicate)? These aspects dictate basically all aspects of the experiment and data analysis, as will be discussed later on.

### Sample preparation

Sample preparation is not always needed in Raman spectroscopy. Yet special issues should be taken into account, especially in biological applications. This includes but is not limited to the need for deactivation of harmful bacteria and the cultivation protocol in a study of cell lines. Nonetheless, sample preparation is mostly determined by the properties of the sample and the sample holders available for the measurements. Details for this aspect are beyond the scope of this protocol and can be found in ref. <sup>25</sup>.

### Choice of instrument

The selection of instruments is largely limited by the availability in the laboratory. There are, however, a few issues worth considering. A near-infrared (NIR) laser source is preferred for highly fluorescent samples. A UV light can also be an option to suppress fluorescence as long as the sample is not degraded because of phototoxicity. Shifted excitation Raman difference spectroscopy may be preferred if extremely intense fluorescence is expected<sup>28–30</sup>, and the issue cannot be tackled by simply switching to NIR or UV excitation. Moreover, the penetration depth may be an important factor if signals from deep layers of the sample are needed. Special techniques like spatially offset Raman spectroscopy can be applied for larger penetration depth if necessary<sup>31</sup>.

### Calibration standards

One particularly critical point in Raman spectroscopy is related to the standard materials that can be used for spectrometer calibration, including the wavenumber and intensity calibration<sup>2,32,33</sup>. Together with a list of potential standards, the measurement of the standards should be described in the measurement protocol as well, including when and how to perform the standard measurements. In particular, the standard material should be selected taking into consideration the properties of samples to be measured and the instrument to be used.

Generally speaking, the Raman bands (i.e., calibration lines) of a wavenumber standard are required to be well defined and separated clearly from each other. While one single calibration line is sufficient for Fourier transform–Raman measurements, multiple lines are needed for the calibration of Raman setups with dispersive elements, e.g., gratings. In the latter case, the calibration lines need to spread the whole wavenumber region of interest of the samples under investigation.

The intensity standard needs to be homogeneous and emit reproducibly within the wavenumber region of interest. Moreover, both wavenumber and intensity standards should be chosen such that it is possible for them to be measured under the same conditions and optical geometry as for the samples. Given all these considerations, the available standard materials can be found in literature or databases<sup>34–36</sup>.

After determining the standards to be used, a regular measurement for the standards is highly recommended as long as this does not interrupt the measurement of a sample. The standards have to be measured again if any changes happen to the instrument. A good practice is to measure the standards every day before moving on to real samples. We recommend measuring multiple standard spectra every time, allowing for quality control or averaging during future use. Typically, the same standard material is measured at the beginning of every measurement day or measurement campaign. The measured standard spectra are used for spectral calibration, which will be described in the section on spectral preprocessing. Importantly, these standard spectra can also serve as samples for quality control. For instance, the peak width of Raman bands of cyclohexane or 4-acetamidophenol, which is used for wavenumber calibration, can be used to benchmark the spectral resolution of a measurement as well.

### Storing data

Another issue that is often overlooked is the data management, regarding, e.g., the management of the metadata and the data structure. While this does not directly influence the measurements and the analysis, it is extremely important in terms of the findability, accessibility, interoperability and reusability of the data. It is important that this process be established before starting the experiment. This point has been well stressed as the ‘FAIR’ requirements of data<sup>37</sup>. On this basis, we suggest constructing a database accessible within the same research group, which can structure the whole Raman spectra within a research group. This is a good starting point for Raman spectral repositories shared by all experts.

The data management rules should cover:

- Data structures
- (File) Naming conventions
- Metadata to be provided (e.g., measurement parameters, sample properties)

To make the description easier, a simple data structure is shown as an example in Extended Data Fig. 1 based on the hierarchy ‘device—replicate—group’. Such data hierarchy is flexible and ‘device—group—replicate’ works just as well. In particular, the ‘Info’ files contain the metadata, such as key remarks on the measurement, setup settings, substrate, sample preparation and (possible) pretreatment of the data (e.g., spectrometer calibration by the in-built programs of a setup).

Standard spectra used for the spectrometer calibration, including dark current spectra, spectra from wavenumber standards like 4-AAP and intensity standard like the NIST fluorescence standard (SRM2242), are stored and linked to the spectra of the real samples using timestamps or similar techniques. Such a time stamp might be stored in the file names, e.g., ‘ddmmyy\_hhmmss’, which are often generated automatically by the instrument upon custom preferences. This time information helps to automatically match the standard spectra with the measurement of real samples during the spectrometer calibration. In general, it requires massive effort to build a data structure/database that can be used by all researchers, yet it is well worth it for multiple reasons: a well-defined data structure saves the efforts of tailoring the data importation for each measurement. More importantly, a good database enhances the possibility of broader cooperation collectively contributed by researchers from all over the world, e.g., researchers from different laboratories and in inter-laboratory collaboration.

### Sample size planning

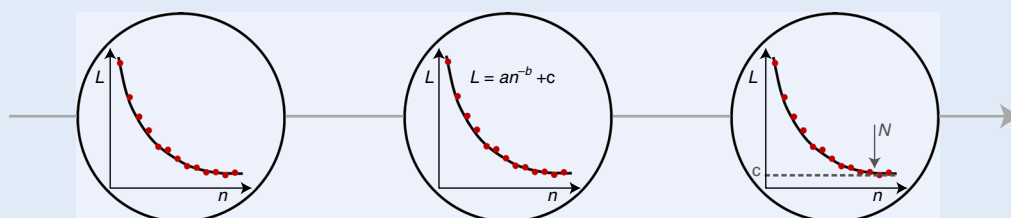
While a good measurement protocol helps to suppress the (negative) influence from the measurements, instruments and other corrupting effects, the SSP determines the minimal number of samples required to reach a statistically meaningful conclusion or to build a model with acceptable performance<sup>38</sup>. In other words, it determines the minimal number of samples required to capture the essential characteristics of the population under investigation. SSP is required in principle for all model-based steps, be it preprocessing, dimension reduction or modeling.

SSP can be achieved in different ways based on certain assumptions about the problem under investigation. For example, ref. <sup>36</sup> proposed a sequence of formulae to estimate the sample size, corresponding to different distributions of the metrics that are used to evaluate the prediction on testing data (e.g., sensitivity, specificity)<sup>36</sup>. These SSP methods yield wrong sample size estimates if the assumption on the distribution does not hold. Another commonly applied SSP approach is the power analysis<sup>39,40</sup>. The term ‘power’ is defined in a statistical test as the probability of correctly rejecting the null hypothesis  $H_0$  given the sample size, the type I error  $\alpha$  (i.e., the probability of wrongly rejecting  $H_0$ ) and the effect size  $d$  (i.e., the deviation of the population from the  $H_0$ )<sup>41</sup>. The power analysis aims to derive the minimal number of samples to reach a predefined power for a specific statistical procedure given the type I error  $\alpha$  and the effect size  $d$ . In practice, the effect size is often estimated based on a priori knowledge of the study. It is well defined and easily interpreted for univariate data. Yet, to extend the definition to multivariate data (e.g., Raman spectra) is not straightforward because of the hidden intervariable relations<sup>39,41</sup>. This fact makes the conventional power analysis less appealing for the SSP with multivariate data. Alternatively, SSP is being conducted based on the learning curve (LC) that shows the change of a predefined metric over the increasing sample size. This has been shown in recent studies based on different metrics such as the Tucker’s  $\phi$ <sup>39</sup>, root mean square error (RMSE) and accuracy. In particular, we proposed to conduct SSP for statistical modeling by fitting the LC according to an inverse power law<sup>42,43</sup>. The details of this method are given in Box 1. Despite SSP being dependent on the statistical modeling and the metrics to evaluate the model prediction, the result is adequate as long as the metric is properly calculated. By ‘properly’ we mean that the metric is calculated based on data that are independent to the training data. Otherwise, the metric is overestimated and the estimated sample size is lower than it is supposed to be.

Regardless of the approaches used for SSP, it is critical to be aware that the ‘sample’ should refer to the highest level of the data hierarchy<sup>43</sup>. In Raman-based biological investigations, for example, this can be a patient, a batch or a replicate. SSP at a lower sampling level, e.g., spectral level, is not sufficient unless each spectrum is from a different patient/batch/replicate. A real-world study may benefit from a two-level SSP, i.e., to combine the sample- and spectrum-level SSP. Multiple samples represent the properties of the population; hence, the sample-level SSP tells how many samples are



# Box 1 | Sample size planning



In the workflow for SSP shown above, the LC is generated by a chosen model metric ( $L$ ) derived with different sample sizes ( $n$ ) (ref. <sup>42</sup>). This curve is subsequently fitted according to the inverse power law. The fitted curve is used to determine the minimum sample size required to get statistically meaningful results ( $N$ ). This could either be where the LC converges (i.e., when the performance metric is close to the Bayes error  $c$ ) or where the performance metric is higher than a predefined threshold.

The core of SSP is to generate a LC, which is a plot between the sample size ( $n$ ) and a predefined performance characteristic/merit ( $L$ ). In our method, the merit is defined as the benchmarks of a statistical model, for example, mean sensitivity or accuracy for a classification and RMSE for a regression model. These benchmarks can be calculated in the same way of model evaluation (external validation), i.e., to predict independent testing data with a model built on  $n$  training samples. In addition, the model training can be conducted in combination with an internal validation to optimize the model parameters. The LC is generated by repeating the model training and evaluation with different values of  $n$ .

$$L(n) = a \cdot n^{-b} + c \quad (\text{B1.1})$$

Subsequently, the LC is fitted according to the inverse power law given by Eq. (B1.1). Therein, the parameters  $a$  and  $b$  refer to the learning rate and the decay rate, respectively. Parameter  $c$  is the so-called Bayes error rate, which is the error of the model trained with an infinite number of samples. To a certain extent, the Bayes error indicates the limit of a statistical/machine learning model, or the difficulty of the problem under investigation.

The minimal sample size  $N$  can be determined where the LC converges toward the Bayes error rate or where the model performance meets a predefined requirement.

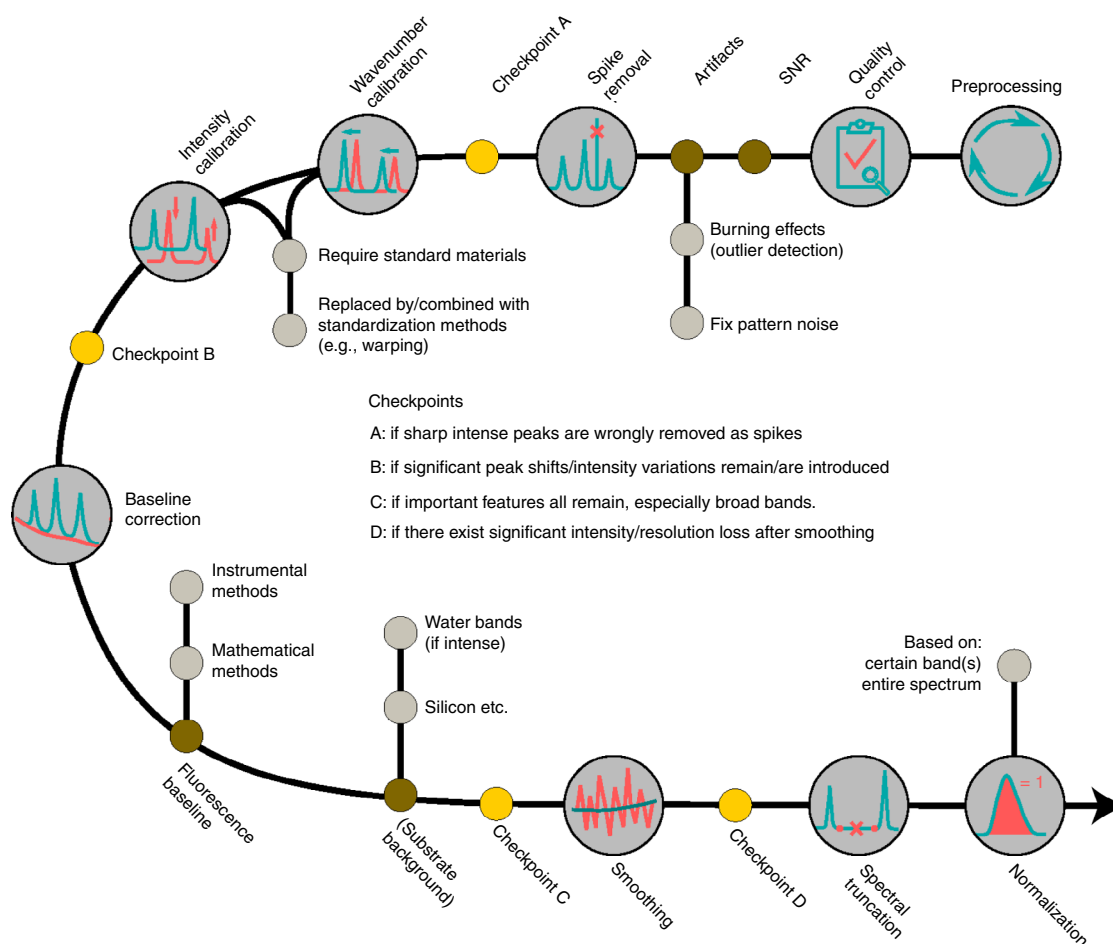
required as a good population representative, whereas multiple spectra from the same sample can bring information regarding the heterogeneity of one sample. Hence the spectrum-level SSP helps to determine the minimal number of spectra required for each sample to adequately capture its heterogeneity.

## Spectral preprocessing

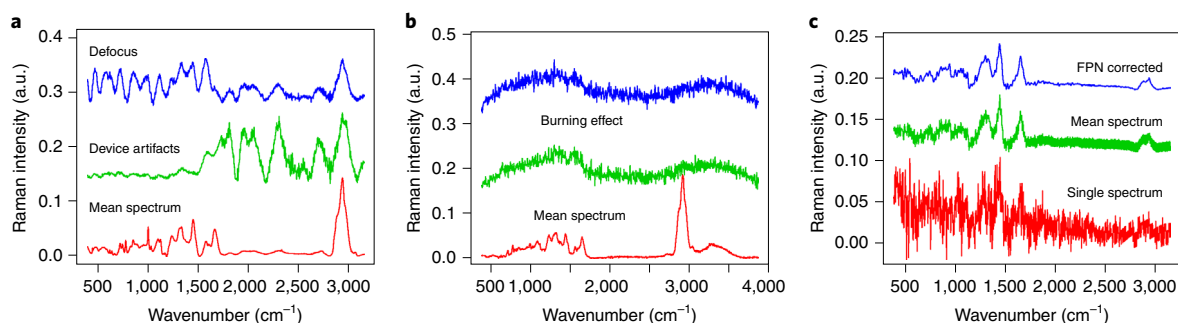
Even though a rigorous DOE helps to suppress the influence of unwanted effects on the Raman spectroscopic measurements, the measured spectra will still contain multiple unwanted contributions originating from the instrument or the sample itself. Data cleaning is necessary to remove these corrupting effects and to recover the ‘pure’ Raman signal of interest<sup>44,45</sup>. This is the major focus of the spectral preprocessing (workflow is illustrated in Fig. 4). In the following text, we will discuss the pitfalls and cautions of each step; these general guidelines should be applied regardless of the methods used. Methods that are suitable for each step are overviewed briefly; more comprehensive descriptions can be found elsewhere<sup>44</sup>.

## Quality control

This step aims to determine the spectral quality and to remove major measurement errors/anomalies. Spectra that show severe burning effects, device artifacts, defocusing and/or substantial signal from contaminations are normally considered as outliers and excluded from further analysis. Examples of such outliers are given in Fig. 5a,b. Spectra can also be regarded as outliers and excluded if the noise level is substantially higher than average and the Raman signal is hardly visible. In practice, different types of outliers can be detected altogether with one single method regardless of the specific patterns or the origins of the artifacts. The simplest way is thresholding based on a predefined maker, such as Hotelling’s  $t$ -squared, Mahalanobis distance or  $Q$  residuals<sup>46</sup>. Another method was reported lately to deal with the difficulties in parameter tuning and to handle the masking and swamping problems<sup>47</sup>. The critical point of outlier removal is that a spectrum can only be regarded as an outlier and removed if strong evidence exists. It cannot be excluded merely for the sake of better results. Removing spectra merely because of, e.g., an inferior prediction by a statistical model is unacceptable.



**Fig. 4 | Workflow of spectral preprocessing.** The main preprocessing steps are quality control, spikes removal, calibration, baseline correction and normalization. Smoothing and spectral truncation are optional but often applied. The necessary checkpoints are marked to ensure that the preprocessing does not introduce artifacts or distort the spectral features.



**Fig. 5 | Illustrative examples of spectra corrupted by different effects/artifacts.** **a**, Examples from a Raman dataset of bacterial cells (dataset 3). The red, green and blue spectra represent the mean spectrum of the dataset (dataset 3), a spectrum corrupted by device artifacts and a spectrum measured in defocusing condition, respectively. **b**, The blue and green spectra show severe burning effects in comparison with the mean spectrum (red) of the dataset. **c**, The FPN is seen obviously in the Raman spectrum (red) owing to low Raman intensity and is not decreased in the mean Raman spectrum (green). The mean spectrum after FPN correction by Fourier filtering is plotted as blue line, which shows that the FPN correction is needed for correction. a.u., arbitrary units.

### Spike removal

Cosmic spikes in Raman spectroscopy originate from electrons generated on the CCD or complementary metal-oxide semiconductor detector by high-energy cosmic particles. They appear randomly in a Raman spectrum and manifest themselves as very narrow but extremely intense spectral features. Due to their high intensities, spikes make data analysis difficult. If there are interfering spikes, the outcome for normalization and feature extraction are not meaningful. Therefore, the



spikes are removed in the step of spike removal. The first and most crucial step of spike removal is spike detection. There are different ways to do so. In the most straightforward case, one can record two spectra from the same sample and detect spikes as the abnormal intensity changes between them, assuming that the two spectra are identical and any sharp intensity differences can only be introduced by spikes. This works upon the knowledge that the spikes rarely occur at the same wavenumber position in two successive measurements. In many cases, however, acquiring two spectra for each sample is too time-consuming, especially for Raman imaging, which measures Raman spectra in point scanning mode. A possible solution is to detect the spikes as abnormal intensity changes between two spectra from the successive measurements of two sampling points or samples, provided the Raman features are almost the same for these two measurements. This method can fail if spectral variations exist between the two successively measured samples, for example, Raman scanning of a heterogeneous sample. In this case, a better solution is to compare each spectrum with the mean spectrum of its neighbors given a spatial window size. If multiple spectra are inaccessible, spikes can also be detected within a single spectrum. Therein, spikes are regarded as abnormally sharp intensity changes along the wavenumber axis. Yet these methods often fail to distinguish spikes from very narrow Raman bands. Among all approaches, the optimal way of spike detection is perhaps to jointly inspect the sharp Raman intensity changes along the wavenumber axis and between the successive measurements. This has been shown in ref. <sup>48</sup>, where spikes were enhanced and easily detected after 2D Laplacian filtering on the spectral data cube.

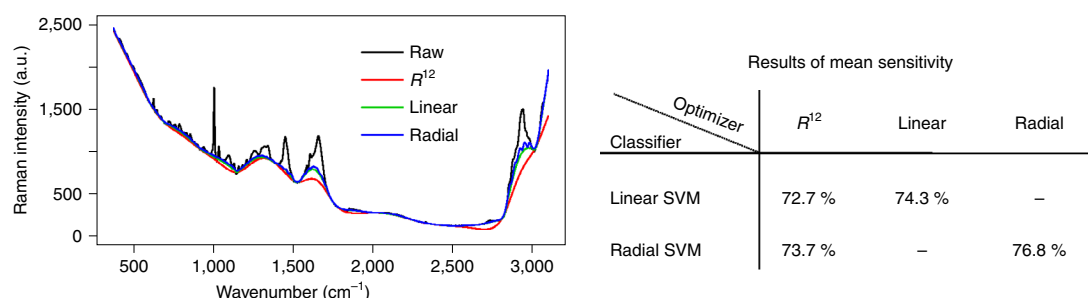
After being detected, a spike can be removed by linear interpolation based on the two boundary points of the spike. Alternatively, a spike can also be replaced by the values of its successive measurement at the same wavenumber positions of this spike. In this case, the fluorescence difference and the intensity variations between the two measurements have to be taken into account.

### Wavenumber and intensity calibration

A successful application of Raman spectroscopy requires that the recorded Raman spectra are independent of the measurement instrument or any other sample-unrelated factors<sup>2,21</sup>. That is to say, a sample is supposed to give identical Raman spectra even if measured by different setups or under different measurement conditions. This is hardly the case in practice. Rather, spectral variations are observed among devices, over time, under changes of measurement conditions, etc. A well-designed standardization method is required to remove these unwanted spectral changes and standardize all measured Raman spectra against the same reference. One of the most basic approaches for such standardization in Raman spectroscopy is spectrometer calibration, composed of wavenumber and intensity calibration<sup>2</sup>. This helps to remove the instrumental influence on the measurements and makes the spectra comparable among different instruments. Spectrometer calibration is typically performed based on standard materials or standard measurements. The selection of such standard materials or standard measurements has been discussed previously in the subsection on DOE. On this basis, the wavenumber axis is calibrated by fitting a (polynomial) function between the measured and theoretical positions of the well-defined Raman bands of a wavenumber standard. The intensity axis is calibrated by dividing the measured Raman intensities by the intensity response function of the instrument, which is derived as the ratio between measured and theoretical emission of an intensity standard over the wavenumber range of interest<sup>24</sup>. Noteworthy, the spikes on the standard spectra, if there are any, should be removed before performing the calibration. A baseline correction might also be necessary, if the baseline is too intense in the wavenumber standard spectra and affects the peak positions.

It should be remembered, however, that the spectrometer calibration cannot remove all undesired spectral variations. On the one hand, the instrument-related spectral variations may still remain after the calibration, as discussed in ref. <sup>24</sup>. On the other hand, undesired spectral variations may originate from not only the instrument- and measurement-related factors but also factors such as interreplicate changes and sample degradation. Such sample-related variations cannot be removed by spectrometer calibration either. The remaining unwanted spectral changes are still substantial in Raman spectroscopy and can degrade the prediction of a statistical model on newly measured data, which will be discussed in detail later on<sup>49</sup>.

Warping can be used as an alternative or complementary procedure to calibration<sup>50,51</sup>. Without the need of spectra of standard materials, warping does not necessarily bring Raman spectra any closer to their true spectra. Rather, it aligns all Raman spectra against a reference spectrum, e.g., the mean spectrum of the dataset. This reduces spectral variations between the Raman spectra. In fact,



**Fig. 6 | Results of baseline correction and the corresponding mean sensitivities from a three-group classification.** The baseline estimates that give the best classification of the linear- or radial-kernel SVM tend to be unphysical, e.g., the intensities of the Raman bands are altered. Alternatively, the baseline correction optimized by the numerical merit of baseline correction leads to both a reasonable baseline correction and a satisfying classification. Data adapted from ref. <sup>64</sup>.

warping techniques were shown to remove unwanted variations regardless of their sources, i.e., instruments or samples. For instance, Raman spectra of time-varying batches were successfully synchronized using correlation optimized warping approach in ref. <sup>52</sup>. Nevertheless, the warping techniques might reduce spectral variations of interest and have to be applied with caution. In addition, the results are reference dependent and may differ from laboratory to laboratory, unless the same reference is used.

In many cases, an interpolation on the wavenumber axis is required to get a unique wavenumber axis for multiple measurements bearing different spectral resolution (i.e., pixel size).

### Baseline correction

Baseline correction is one of the best-recognized steps in Raman spectral preprocessing. It refers to two meanings in literature: the removal of spectra with only substrate information or fluorescence baseline removal. The former is used to remove the Raman signals of the substrate from a measured Raman spectrum; the latter aims to remove the fluorescence of the sample manifested in a Raman spectrum as a slowly changing baseline under Raman bands. The substrate contribution needs to be removed if the substrate features substantial Raman bands, especially if these Raman bands overlap with those of the samples<sup>53,54</sup>. To do so, a spectrum of the substrate is often required as a reference to estimate the contribution of the substrate in the recorded Raman spectra. This task is difficult for heterogeneous substrates as it is often seen in forensic scenarios, e.g., to detect bloodstain on jeans material. A statistical approach can be useful in this situation; for instance, multivariate curve resolution can be utilized to deal with such heterogeneous substrate contributions<sup>55</sup>. Fluorescence baseline removal is often more complicated than substrate correction as the fluorescence baseline is both sample and setup dependent. Such fluorescence baseline is mostly removed with mathematical approaches, such as calculating derivative spectra, sensitive nonlinear iterative peak clipping<sup>56</sup> algorithm, asymmetric least squares (ALS) smoothing<sup>57</sup>, modified polynomial fitting<sup>58</sup>, standard normal variate, multiplicative scattering correction and extended multiplicative signal correction (EMSC)<sup>45,59</sup>. These methods are flexible, easy to use without the need of instrumental modifications, and perform adequately in most cases. Nonetheless, approaches based on instrumental modifications can be necessary if the fluorescence is too intense to be mathematically corrected. Techniques of this category include time-gated Raman spectroscopy<sup>60</sup>, modulated Raman spectroscopy<sup>61</sup> and shifted excitation Raman difference spectroscopy<sup>28–30</sup>.

Despite the broad recognition and extensive investigation of fluorescence removal, it is not easy to automatically select the best-performing method and the optimal parameters for a specific dataset. The importance of such selection is often overlooked, even though it is well admitted that no approach always outperforms another. Among the existent strategies, a commonly used one is to do the optimization for each dataset according to the results of a subsequent analysis, e.g., a regression or a classification model<sup>44,62,63</sup>. This method, however, is model-dependent and requires a large number of training samples to construct the model. Moreover, the optimization is dependent on the results of the subsequent model. An underlying risk is that a physically unreasonable baseline correction can contribute to the classification and lead to a better prediction<sup>64</sup>. Therefore, the correction leading to the best performance of the subsequent model can be unreasonable. This is illustrated in Fig. 6, in which the baselines giving the best performance of the linear- or radial-kernel support vector machine (SVM) tend to alter the intensities of the Raman bands. Alternatively, the fluorescence removal can be

**Box 2 | Automatically optimize fluorescence baseline correction**

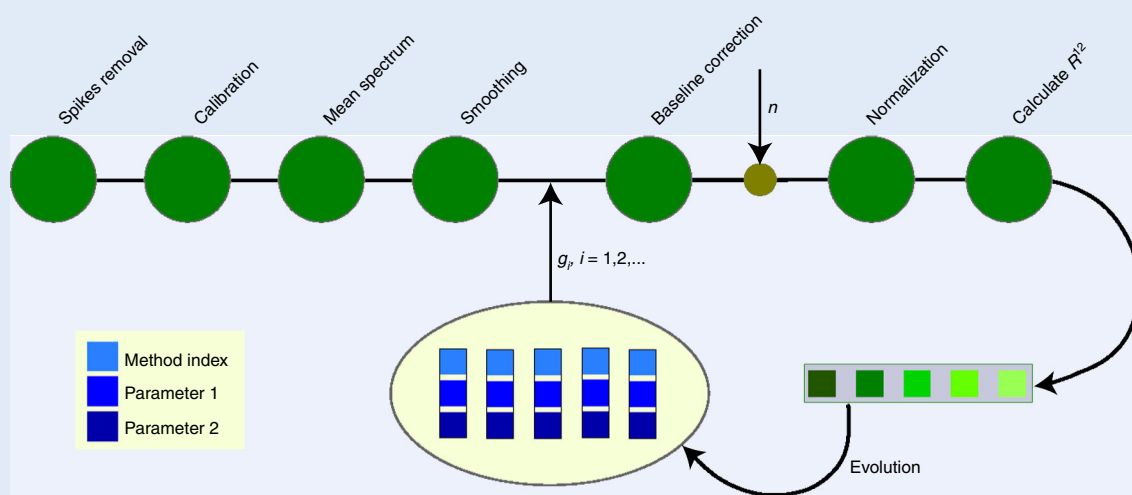
$$m_1 = \frac{\ln(N_p)}{A_p} + \frac{A_s t}{\ln(N_s)} \quad (\text{B2.1a})$$

$$m_2 = \frac{A_p}{A_s t + A_p} \quad (\text{B2.1b})$$

$$t = \frac{\max(I_{i \in \mathbf{s}}) - \min(I_{i \in \mathbf{s}})}{(\sum_{i \in \mathbf{n}} I_i) / N_n} \quad (\text{B2.1c})$$

$$R^{12} = \frac{m_1}{m_2} \quad (\text{B2.1d})$$

The numerical marker  $R^{12}$ , which benchmarks the goodness of a baseline correction, is given by Eq. (B2.1) (ref. <sup>64</sup>). The definition is based on three regions of a baseline corrected spectrum: silent region (**s**), peaks region (**p**) and region for normalization (**n**). These regions are indicated by the subscripts in the equations. The number of points contained in each specific region is denoted by  $N$ . The terms  $A$  and  $I$  represent the area and intensity, respectively. A better baseline correction features a smaller value of  $R^{12}$ . In particular, the three regions **s**, **p** and **n** can be assigned manually or determined automatically. A manual assignment allows a priori knowledge to be utilized for the baseline optimization. The workflow for  $R^{12}$ -based optimization, shown below, is based on a genetic algorithm. The genes correspond to the index of a baseline correction and the parameters of this method. For each generation, the baseline correction is performed using the method and parameters given in each chromosome, and the  $R^{12}$  is calculated. The population is evolved to minimize the  $R^{12}$  value. The chromosome giving the minimal  $R^{12}$  in the last generation is considered to yield the optimal baseline correction.



In detail, the optimization starts with calculating the mean spectra (of each group) after spike removal and calibration, followed by an optional step of smoothing. The mean spectra were used for the genetic algorithm to optimize the baseline correction method and parameters. Therein a baseline correction is performed using different methods and parameters. The resulting spectrum is used to calculate the maker  $R^{12}$  after being normalized against the spectral region **n**. If the mean spectrum is calculated for each group separately, the values of  $R^{12}$  for all mean spectra of different groups are averaged. The optimal baseline correction is the one giving the minimal  $R^{12}$ . In particular, the workflow takes a genetic algorithm as an example, in which the population evolves to minimize the value of  $R^{12}$ . Other optimization procedures such as simulated annealing can also be used.

optimized based on a numerical marker that represents the goodness of a baseline correction. This is the basic idea of the method described in Box 2 (ref. <sup>64</sup>). The numerical marker  $R^{12}$  is calculated from the spectra themselves without the need for any modeling procedure. Therefore, the optimization is more objective and not biased by pursuing the best prediction of the optimization. As shown in Fig. 6, the intensities of Raman peaks are better preserved by the  $R^{12}$ -based method in comparison with the classification-based method. In addition, the mean sensitivities of the classification using the baseline correction optimized by  $R^{12}$ - and model-based methods are comparable for both the linear- and radial-base-kernel SVMs. It is, hence, fair to say that the  $R^{12}$ -based optimization is able to balance the goodness of baseline correction and the performance of the subsequent classification. In addition, the  $R^{12}$ -based optimization is less costly and faster without the need to build statistical models.

### Smoothing

Smoothing or filtering is optional in the analysis of Raman spectra and can be done by spectral and/or spatial filtering. Spectral filtering removes noise with a low-pass filter along the wavenumber axis. As shown in Eq. (1a), the intensity at a given wavenumber  $I(\tilde{\nu}_i)$  is replaced with the output of the filter based on its spectral neighborhood  $I(\tilde{\nu}_{j \in \{i-k, \dots, i+k\}})$ . The filter can be the mean, median, Gaussian, polynomial function, etc. Spatial filtering bears a similar idea as spectral filtering, but it applies the low-pass filter to the spatial domain and is suitable for Raman imaging data (i.e., hyperspectral data cubes). The procedure is given in Eq. (1b). A Raman spectrum at a given spatial coordinate  $I(x_0, y_0, \cdot)$  is replaced with the results of the filter from the spectra on its spatial neighborhood  $I(x, y, \cdot)$ . Both approaches have advantages and disadvantages. Spectral filtering degrades the spectral resolution but preserves the spatial resolution, and vice versa for spatial filtering. The users can choose either of them according to the characteristics of their data and the aim of the analysis. For hyperspectral data, spectral filtering is preferred if the spatial resolution is of more importance, and vice versa. In cases of single spectral measurements where spatial filtering is impossible, smoothing (i.e., spectral filtering) is recommended only for highly noisy data to preserve the spectral resolution as far as possible. This is especially important if closely located or narrow Raman bands are of interest.

$$I(\tilde{\nu}_i) = (I(\tilde{\nu}_{j \in \{i-k, \dots, i+k\}})) \quad (1a)$$

$$I(x_0, y_0, \cdot) = (I(x \in \{x_0 - k, \dots, x_0 + k\}, y \in \{y_0 - k, \dots, y_0 + k\}, \cdot)) \quad (1b)$$

A special case of smoothing is related to the fixed pattern noise (FPN)<sup>65</sup>. FPN might originate from the etaloning effect of CCD pixels, and it becomes obvious in fast measurements with extremely low intensities. Unlike white (Gaussian) or shot (Poisson) noise, FPN cannot be decreased by averaging over multiple measurements (Fig. 5c). However, it can often be distinguished from Raman spectral features in the frequency domain and can be removed by approaches like Fourier and wavelet filtering. An example of FPN correction by Fourier filtering is shown as the blue line in Fig. 5c.

It must be noted here that the effect of smoothing is most times negligible and sometimes even detrimental to the performance of the subsequent spectral analysis. This is particularly true for smoothing approaches based on moving-window strategies, as these procedures introduce correlations into the noise structure. Most multivariate techniques could perform worse after smoothing in comparison with raw data as they do not take correlated noise into account<sup>66</sup>.

### Spectral truncation and normalization

Spectral truncation is used to exclude wavenumber regions absent of substantial Raman signals (silent region) or regions showing either strong contributions of the substrate, water or artifacts. This step can be regarded as a rough feature selection.

After spectral truncation, a normalization follows as the last step of preprocessing<sup>24</sup>, aiming to remove the influence from the fluctuation of excitation intensity or changes in the focusing. To do so, the spectral intensities are divided by the area, maximum or  $l_2$  norm of a selected spectral region ( $R$ ), as given in Eq. (2a–c), respectively. The region ( $R$ ) used for the calculation is task dependent. It can be the whole spectrum, a single wavenumber position or a Raman band where the Raman intensity is supposed to be constant among the samples.

$$I(\tilde{\nu}) = I(\tilde{\nu}) / \sum_i I(\tilde{\nu}_i), i \in R \quad (2a)$$

$$I(\tilde{\nu}) = I(\tilde{\nu}) / \max_i I(\tilde{\nu}_i), i \in R \quad (2b)$$

$$I(\tilde{\nu}) = I(\tilde{\nu}) / \sum_i I(\tilde{\nu}_i)^2, i \in R \quad (2c)$$

A slightly different idea to normalization is scaling<sup>67</sup>, which includes min–max scaling, mean scaling and z-score scaling, as defined in Eq. (3a–c), respectively. The terms ‘min’, ‘max’, ‘mean’ and ‘sd’ represent minimum, maximum, average and standard deviation of the spectral data, respectively. The  $l_2$  norm-based normalization (Eq. (2c)) is also named unit-length scaling. In particular, z-score scaling normalizes the variance of each variable to 1 in order to equalize the importance of all

variables. This is not recommended in spectral data in our experience, as it may destroy spectral patterns that are useful for further analysis.

$$I(\tilde{\nu}) = \frac{I(\tilde{\nu}) - \min(I(\tilde{\nu}))}{\max(I(\tilde{\nu})) - \min(I(\tilde{\nu}))} \quad (3a)$$

$$I(\tilde{\nu}) = \frac{I(\tilde{\nu}) - \text{mean}(I(\tilde{\nu}))}{\max(I(\tilde{\nu})) - \min(I(\tilde{\nu}))} \quad (3b)$$

$$I(\tilde{\nu}) = \frac{I(\tilde{\nu}) - \text{mean}(I(\tilde{\nu}))}{\text{sd}(I(\tilde{\nu}))} \quad (3c)$$

### Remarks on preprocessing

The order of performing the above-mentioned preprocessing steps is not fixed. Nonetheless, some general guidelines are worth keeping in mind and can be summarized as follows.

- 1 Quality control (outlier detection) does not have to come first. To be on the safe side, we suggest a quality control as both the first and last step of preprocessing. Preliminary data sorting at the first place is recommended to remove severe outliers or errors in measurements. This is particularly useful if a mean spectrum is required in any preprocessing steps, such as a model-based preprocessing<sup>59</sup>. A second quality control/check after preprocessing is helpful to detect outliers that have been masked by the corrupting effects in the raw data.
- 2 Interpolation on the wavenumber axis should be done after the spike removal so that the spikes do not get broadened, which would make them harder to remove.
- 3 Smoothing and baseline correction are sometimes entangled. Severe noise can lead to obvious baseline offset in baseline correction approaches where the Raman intensities are constrained to be nonnegative. It is thus better to smooth the spectra before baseline correction. Alternatively, some baseline correction approaches handle this issue by allowing negative values in the corrected spectrum, such as ALS correction<sup>57</sup>.
- 4 Truncation of an inner part of a spectrum is better applied as the last step before normalization to allow a comprehensive treatment of the spectra without breaks in it.
- 5 The calculation of derivative spectra can correct for baseline contributions, but it may introduce a smoothing effect as the intensity values in a (spectral) region near a (spectral) point are recalculated for the correction.
- 6 Model-based preprocessing such as multiplicative scattering correction and EMSC can act as baseline correction as well, but these methods often perform an additional normalization step.

Above all, the preprocessing steps are not independent but impact each other. It is suggested to conduct the preprocessing as a whole and, if necessary, adjust a previous step according to the feedback of a later step. In the very end, all spectra in the same dataset should be preprocessed with the same procedures and parameters. It makes no sense to preprocess spectra differently to obtain better results of a subsequent analysis. Moreover, proper spectral preprocessing is important if Raman spectra should be analyzed using database searches. These database searches compare a spectrum from a sample of interest with spectra of known substances. Any ‘artifacts’ resulting from semi-optimal preprocessing could lead to misleading matches in such database searches.

### Data learning

Data learning attempts to translate the Raman signals into high-level information. An easy way to do so is to identify substances according to the similarity between the measured spectra and the spectra from known substances within spectral databases. However, this is often not feasible in scenarios like biological applications in which the samples are too complicated to be identified with database indexing. This is particularly true if contributions from heterogeneous substrates are contained in the measured spectra<sup>55</sup>. Thus, more advanced approaches based on machine learning<sup>68</sup> are required. In these cases, usually a model is trained with a certain number of samples of which the high-level information is known (i.e., training data). Then, this model is used to predict unknown samples to obtain the high-level information directly. Provided the training data represent the population under investigation well, it is possible to build a model that learns the essential characteristics of the population and generalizes well to unknown samples (i.e., similar prediction error on training and unknown samples). Unfortunately, this is often not the case in real-world situations. Rather, the



training data are mostly a (small) statistical sampling of the population and fail to span a complete data space of this population. The unknown samples to be predicted can bear substantial differences to the training data and hence are often predicted by the model with more errors than for the training data. To ensure the usability of a model in the future phase, it is necessary to rigorously check its performance on predicting unknown samples independent of the training data and keep the prediction error under control. To do so, we built up a workflow of data learning shown in the upper part of Fig. 7 (ref. <sup>69</sup>). The major steps of the workflow are statistical sampling, statistical modeling and model evaluation, which will be explained in the next subsections.

### Statistical sampling

As the first phase of data learning, statistical sampling is conducted to prepare data for the statistical modeling. It is a particularly useful strategy to reach a valid statistical model with a limited sample size, which is often seen in practice. Herein the accessible dataset is split into three subsets: training, validation, and testing datasets. The three subsets are used for model training, model optimization and model evaluation, respectively. The workflow will be described in the next subsections in detail. The data split is in many cases repeated multiple times, e.g., in the methods of cross-validation (CV) or bootstrapping<sup>70,71</sup>. Each repetition generates a different separation of the three subsets, and the statistical modeling is conducted multiple times. This gives the chance to verify the stability and reproducibility of the model, e.g., according to the mean and standard deviation of the accuracies or RMSEs calculated from the multiple predictions on the different testing datasets.

### Statistical modeling and machine learning

Statistical modeling and machine learning follow the statistical sampling. This refers to the procedures of model training and model validation, based on the training and validation datasets generated by the statistical sampling, respectively. To be specific, a series of models are built based on the training data with varying parameters and/or methods. The best model is then selected according to the predictions on the validation data. The details to do so are given in the following paragraph.

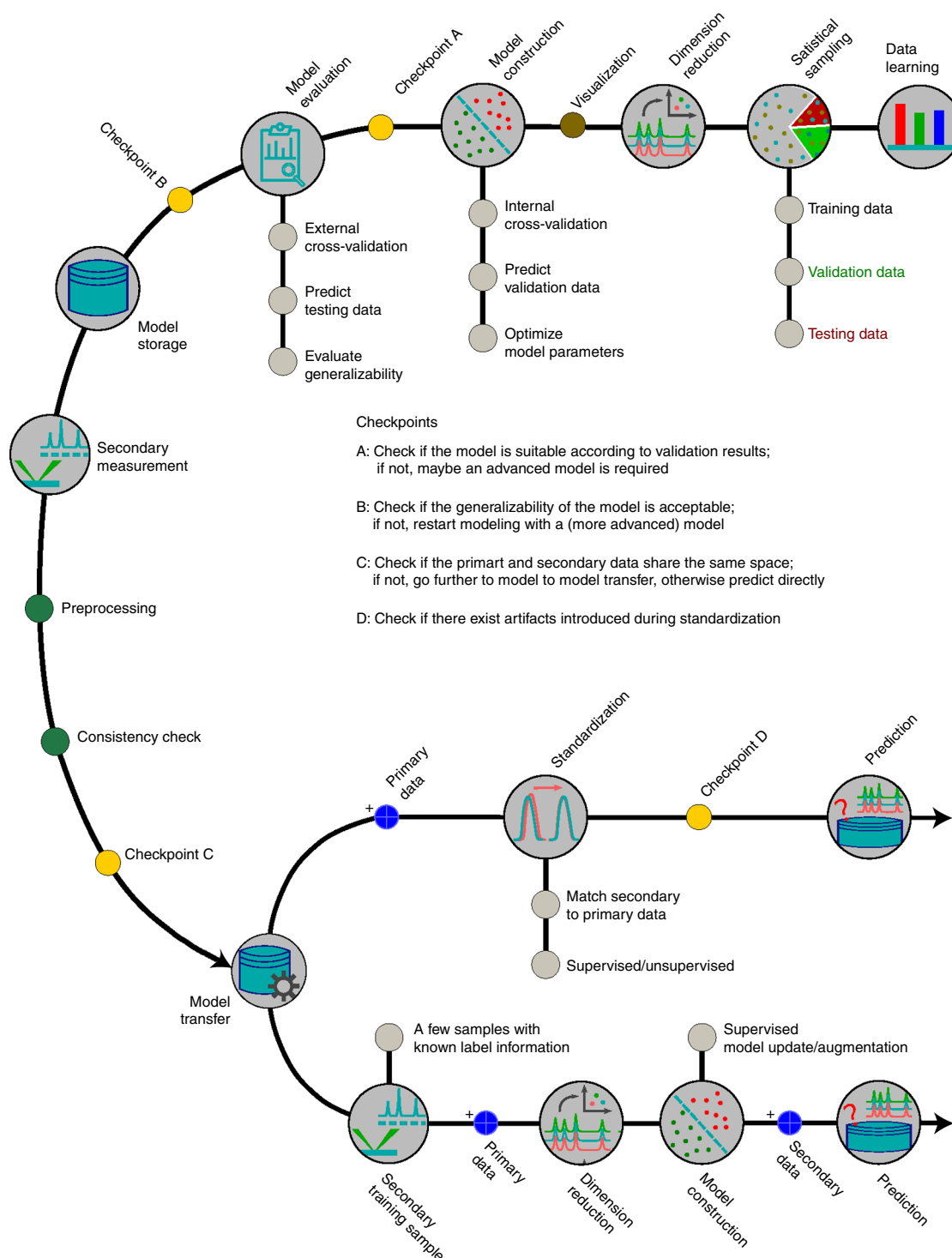
Statistical modeling and machine learning start in most cases with a dimension reduction. This is particularly important for Raman spectroscopy, where the dataset is composed of a large number of correlated features and the sample size is limited<sup>72</sup>. An exception occurs for deep learning, in which the dimension reduction can be omitted thanks to the feature engineering of the deep neural networks<sup>73</sup>. The benefits of dimension reduction are twofold:

- First, it makes the visualization simpler and clearer and, hence, helps to derive a better overview about the characteristics of the datasets
- Second, it can improve and speed up the subsequent modeling by removing redundant information and extracting useful features from the data

Existent dimension reduction approaches can be categorized in different ways. Depending on whether or not the response information is used during the calculation, dimension reduction can be considered as a supervised or unsupervised procedure. Typical examples are principal component analysis (PCA) and partial least squares (PLS), respectively. Dimension reduction approaches can also be categorized as feature selection or feature extraction, with the resulting variables (of lower dimensionality than the original dataset) in the original or a transformed data space, respectively. In particular, feature selection methods pick the variables that are the most significant according to a predefined benchmark. Feature extraction, in contrast, works by transforming the spectral data into a new space with a lower dimensionality spanned by a set of basis vectors, and the original spectra are represented by their coordinates in this new space, namely the scores. The basis vectors are supposed to catch major information in the data, for instance, the source of variances in terms of PCA. A similar idea to feature extraction is un-mixing. The basis vectors are the endmembers, i.e., spectra of distinct components in a sample under investigation<sup>74</sup>. The resulting scores represent the abundance information of each component. Typical approaches include vertex component analysis<sup>75</sup>, N-FINDER<sup>74</sup>, independent component analysis<sup>76</sup> and multivariate curve resolution alternative least squares<sup>11</sup>. Besides all these linear methods mentioned so far, nonlinear dimension reduction methods exist as well, such as Isomap<sup>77</sup>, locally linear embedding<sup>78,79</sup> and feature engineering by neural networks<sup>80</sup>.

In the second phase of data learning, the output of the dimension reduction is fed into a subsequent model, which might be a clustering, classification or regression model. A model to do so can be categorized as linear or nonlinear, parametric or nonparametric, supervised or unsupervised<sup>68</sup>.





**Fig. 7 | Workflow of data learning.** It starts from statistical sampling, which splits the whole dataset into training, validation and testing data. Model building, composed of dimension reduction and statistical modeling, is complemented by an internal and external validation. Internal validation is based on the training and validation data and aims to find the optimal model parameters. External validation is to evaluate the optimal model according to its prediction on the testing data. The model and the evaluation results are then stored and used to predict new data in the future. All data to be predicted are preprocessed in the same way as the training data. The new data may differ substantially from the training data owing to interreplicate variations or instrumental changes and fail to be predicted properly. In this case, a model transfer is needed, and the related approaches can be categorized as data-based or model-based methods, as shown in the upper and lower branches, respectively.

**Table 1 |** Confusion table and metrics to evaluate a classification/clustering model

True				Metrics
Predicted	P	P	¬P	accuracy = $\frac{a+d}{a+b+c+d}$
	P	<b>a</b>	c	sensitivity = $\frac{a}{a+b}$
	¬P	b	<b>d</b>	$\kappa = \frac{\text{accuracy} - p_e}{1 - p_e}$

The term  $p_e$  refers to the hypothetical probability of a random agreement by chance alone, e.g., the probabilities when grouping occurs randomly into each category

While the selection of a model to be used is data dependent, it should be kept in mind that the generalizability of a model is likely to decrease with the increased complexity of the model. Without sacrificing performance, a model should be as parsimonious as possible. This means that a linear and parametric model is preferred in terms of the generalizability compared with a nonlinear and non-parametric model. By ‘classification’ we refer to both ‘one-class’ and ‘multi-class’ classification tasks. In the ‘one-class’ task, only one class is well defined and is used to train a prediction rule to decide whether a future sample belongs to this specific class or not. These one-class models are utilized as abnormality detection with respect to the given class<sup>81</sup>. In the ‘multi-class’ task, also referred to as ‘discrimination’, a model is trained based on two or more well-defined classes to find a discrimination rule to assign future samples to any of these well-defined groups<sup>82</sup>.

Another important part of model construction is the variable importance or significance of variables. These coefficients are calculated from the trained model and indicate how significant each variable is for the model and the task. Variables that correspond to coefficients of larger variable importance are considered more important for the model, and this interpretation should be made in combination with the spectroscopic model of the data at hand. These values can be further utilized for feature selection leading to a more parsimonious model. However, it should be recognized that variables with too high or too noisy model coefficients should be better removed from the modeling as they are very likely to be unreliable.

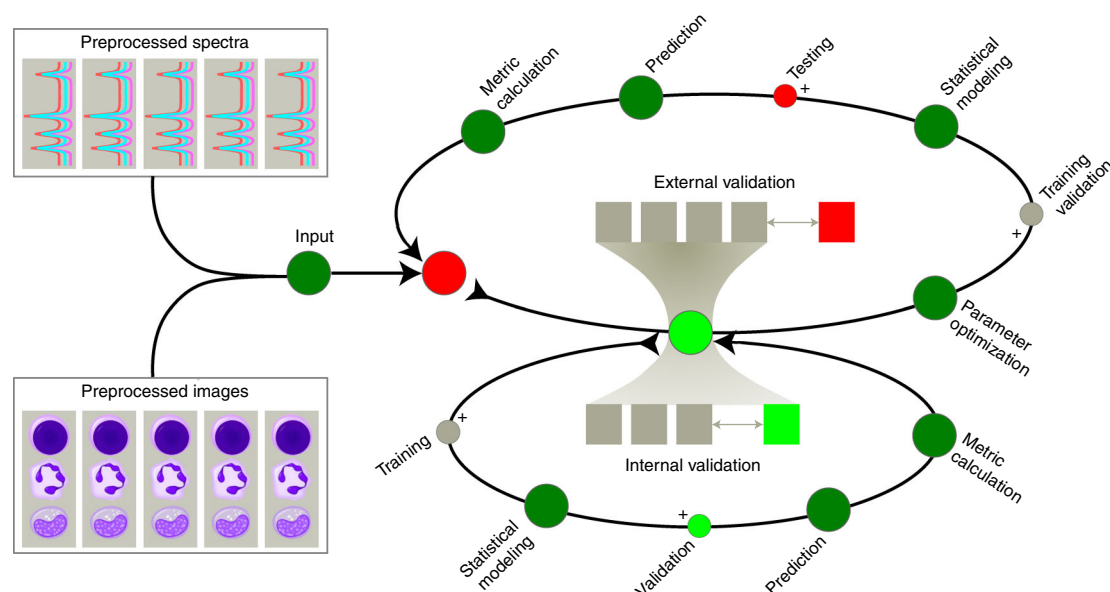
### Model evaluation

As mentioned above, a model usually predicts unknown samples worse than it predicts the training/validation data. This is known as the phenomenon of model shrink<sup>83</sup>. In extreme cases, a statistical model can predict the training/validation data perfectly but completely fails to predict unknown samples owing to overfitting, i.e., the model fits the training data too perfectly and loses its generalizability. It is thus important to check the prediction on unknown samples and keep the error rate under control to make sure a statistical model is usable in practice. For this reason, an additional procedure is conducted after model construction, namely model evaluation. Herein the model constructed from the previous steps is used to predict the testing data generated from the statistical sampling. If the prediction error is too large given a predefined threshold, the statistical modeling should be reconducted with modifications, e.g., different parameters or even a different method.

The model evaluation tries to estimate the prediction performance of a given model, which can be benchmarked via different figures of merit. For regression models, the deviation of the prediction to the true values can be calculated and is most often used. For example, the root mean squared error of prediction (RMSEP) defined in Eq. (4) is one of these performance markers for regression models. Classification or clustering models are often benchmarked by a confusion matrix formed by the prediction and the ground truth. Such a confusion matrix is presented in Table 1. A variety of characteristics can be calculated based on this confusion matrix, including accuracy, sensitivity, specificity, selectivity and Cohen’s kappa  $\kappa$ . As a comprehensive list of the metrics is beyond the scope of this tutorial, more details on this perspective can be found elsewhere<sup>84–86</sup>.

$$RMSEP = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4)$$

The above-mentioned steps, including the statistical sampling, statistical modeling and model evaluation, can be conducted using a two-layer CV<sup>69</sup>. A typical workflow is shown in Fig. 8. It is



**Fig. 8 | Workflow of data learning based on two-layer validation.** A part of the whole dataset is separated as the testing data (red block) in each iteration of the external validation. The rest of the data are used for the internal validation, in which they are split into training (gray blocks) and validation (green block) data. The model is built on the training data, and the optimal parameters are determined according to the prediction on the validation data. Finally, the optimal model is used to predict the testing data to evaluate its performance.

composed of two layers: an internal layer and an external one. The internal validation is responsible for the statistical modeling and machine learning. It aims to find the optimal parameters and/or the optimal model according to the prediction on the validation dataset. The external validation is responsible for the model evaluation, where the generalization performance of the model is benchmarked by its prediction on the testing data. The statistical sampling is conducted for each iteration of the internal and external validation. For each iteration of the external loop, a certain amount of data is taken out as testing data (red block). The rest of the data (gray blocks) are fed into the internal validation, where they are further split into training (gray blocks) and validation (green block) data. The training data are used to build the statistical model, and the validation data are predicted. The performance of the prediction is benchmarked by a predefined metric. The values of the metric are collected through all the internal iterations and used to select the best model. Thereafter, a statistical model is rebuilt with the optimal parameters/method based on the training and validation data altogether. The model is then used to predict the testing data in order to benchmark its generalization performance. The model along with the results of the evaluation is stored for future use, typically to predict new data measured from unknown samples. Noteworthy, the dimension reduction and model construction are represented as a single block ‘statistical modeling’ in Fig. 8. The idea behind is that it is good practice to treat the dimension reduction and model construction as one single step and involve them altogether in the loop of internal and external validation.

As you may recognize, the multiple iterations of the external validation lead to a number of models. These models differ in two aspects. First, the data used for the model training are different. Second, the parameters/method of the model may differ from each other as well since they are optimized with different training/validation data. These multiple models provide a possibility of data fusion, i.e., to merge their predictions on unknown samples via voting schemes, which may result in more stable results.

### Remarks on data learning

Before concluding this part, it is worth noting two crucial issues for proper data learning, especially for a reliable model evaluation. Foremost, the prediction of a model on training data is normally better than the prediction on (testing) samples different to the training data. Thus, a reliable model evaluation can only be ensured if the testing data are independent from the training/validation data. This leads to the independence requirement between the training/validation data and the testing data<sup>87,88</sup>. The testing data must be from samples independent of the training/validation data, which

can be different patients or batches, according to the highest level of the sampling hierarchy. The independence requirement between training and validation data is relatively less important because the prediction on the validation data is used to optimize the model, and a violation of the independence between the training and validation data leads to similar overestimation of the models to be compared. It is thus still very likely to select the optimal model.

In addition to the independent requirement, it is beneficial and recommended to conduct the statistical sampling as the first step of the data learning so that all components (e.g., dimension reduction and classification) of a statistical model can be optimized and evaluated thoroughly. Or, at least, the statistical sampling should be conducted prior to all supervised procedures, be it dimension reduction, classification or regression, so that all supervised procedures are reliably evaluated. An example to do so can be found in the ‘inside-CV’ in ref. <sup>69</sup>. Therein the data split was performed before the dimension reduction, and all steps of the data learning were enclosed within the CV loop. A data split after a data learning step, especially a supervised step, can implicitly violate the independence requirement and hence lead to an overestimation of the model performance. This point is well illustrated in the section ‘expected results’.

### Model transfer

Supposing all procedures of the data learning are performed properly, a model is expected to be able to predict new data well in the future. However, this is not always true in reality, and a model is likely to fail to predict new data. This is dramatically severe in Raman spectroscopy, because this technique is very sensitive and tiny variations in instruments, measurement conditions or sample preparation can be reflected as substantial shifts of Raman bands or changes in Raman intensities. These unwanted spectral variations are impossible to avoid in practice, and they reduce the performance of an existent model while predicting the new data. Building a new model for these new data is not always possible because of, for instance, the limit access to new training data. A more feasible alternative would be model transfer, which aims to improve the predictive performance of an existing model on new data without remeasuring a large number of training samples<sup>89</sup>. In this context, the training and new datasets are denoted as primary and secondary datasets, respectively. Model transfer can be categorized as model-based or data-based, according to the subject that is modified. The basic idea of these methods is illustrated at the bottom of Fig. 8. Briefly, the data-based methods attempt to remove the difference between the primary and secondary data, for instance, by matching the secondary to the primary data. The model-based methods do not modify the datasets; rather, they adjust the model parameters by taking the features of the secondary data into account. The prediction on the secondary data is thus improved. Existent approaches for model transfer are Procrustes analysis, orthogonal signal correction, parametric time warping and piecewise direct standardization<sup>90–92</sup>. In addition, methods developed in our group and their workflows are given in Boxes 3 and 4. The presented method corresponds to the two model transfer mechanisms<sup>49,93,94</sup>. For more details about model transfer approaches in biological investigations, the readers are kindly referred to a recent publication<sup>95</sup>.

The capability of model transfer to improve the model prediction in the presence of substantial unwanted spectral variations has been well demonstrated in previous publications. However, this does not mean that model transfer is always necessary. It is not needed, for example, if the unwanted intragroup spectral variations are not substantial in comparison with the intergroup differences of interest. In practice, we recommend an additional step named ‘consistency check’ after receiving the secondary data to be predicted (Fig. 7). This means to compare the two types of variations: the variations between the primary and secondary data, as well as the variations of interest. Model transfer is necessary if the differences between the two datasets are larger than the intergroup variations of interest. A very simple way is to build a PCA model on the primary data and use it to predict primary as well as secondary data. The two different variations can be compared via visual inspection on the score plots. Alternatively, we proposed a numeric marker for an automatic consistency check, namely relative Fisher’s discriminant ratio<sup>94</sup>, which quantifies the significance of the difference between the primary and secondary data with respect to the intergroup changes of interest. A value >1 indicates the necessity of model transfer and vice versa. The corresponding workflow of this approach is given in Box 5.

The descriptions and discussions above provide a thorough overview of the important chemometric steps as well as the approaches dealing with the open issues encountered in Raman spectral analysis. In the next sections, we will turn to practical applications and provide a standard operating

### Box 3 | Model-based model transfer

#### Tikhonov regularization

Prerequisite: a few spectra from the secondary device with known response information<sup>49,89</sup>.

$$\begin{pmatrix} X \\ \eta I \\ \lambda L \end{pmatrix} b = \begin{pmatrix} y \\ 0 \\ \lambda y^* \end{pmatrix} \quad (\text{B3.1})$$

- 1 Take secondary spectra with known response information, and split them into training, validation and testing subsets.
- 2 Rebuild the model following the procedure of statistical modeling shown in Fig. 8 after augmenting the training dataset  $(\mathbf{X}, \mathbf{y})$  with the secondary training data  $(\mathbf{L}, \mathbf{y}^*)$  according to Eq. (B3.1), where  $\mathbf{I}$  is the identity matrix.
- 3 The parameters  $\lambda$  and  $\eta$  are optimized in a similar way as the model parameters based on the prediction on the secondary validation data.
- 4 The constructed model can be evaluated with the secondary testing data.

#### Score movement

Prerequisite: the secondary data follow a distribution similar to the primary data, which can be verified according to their score distribution based on PCA<sup>93</sup>.

The workflow is in general the same to the normal statistical modeling described in Fig. 8 except the dimension reduction step. This can be explained taking PCA as an example.

- 1 Build a PCA model on the primary training data, and use it to predict the secondary data.
- 2 Calculate the distance between the primary and secondary score clusters for each principal component.
- 3 The PCA scores of the primary data, be they from the training, validation or evaluation datasets, are moved based on the calculated distance to match the corresponding secondary scores.

Construct the statistical model on the primary score space after movement.

procedure (SOP) to walk through the chemometric procedures in Raman spectroscopy. We start this SOP from the point the data have been acquired and skip the experimental design. This allows us to focus more on the data analysis issues. We demonstrate the methods on a few example datasets measured from different biological samples. Their brief information is given in the next section. The results from the four datasets will be shown before concluding this protocol to illustrate the essential issues in Raman spectroscopy.

## Materials

### Reagents

*Note:* rather than give the details of sample preparation and Raman spectroscopy, we focus on the spectral analysis in this protocol. Interested readers are directed to refs.<sup>42,97–99</sup>.

- *Tumor cells:* this dataset was composed of Raman spectra of breast carcinoma derived tumor cells (MCF-7, BT-20) and acute myeloid leukemia cells (OCI-AML3). The measurement was done in single-cell mode on an upright Raman microscope (Kaiser Optical Systems). The samples of each group (cell line) contained nine replicates (batches), comprising in total 1,553 cells/spectrum (MCF-7: 558, BT-20: 477 and OCI-AML3: 518)
- *Bacterial spores:* herein we investigated spores of three *Bacillus* species: *Bacillus mycoides*, *B. thuringiensis* and *B. subtilis*. The Raman spectra were acquired for single bacteria using four micro-Raman devices (Bio Particle Explore, rap.ID Particle Systems). Quartz and nickel foil were used as the substrate for the measurements on the first and the other three devices, respectively. There were respectively 1,592, 624, 654 and 848 spectra measured by the four devices, with almost equal sample size for the three species
- *Vegetative bacteria:* Raman spectra were measured in single-cell mode from six bacterial species: *Escherichia coli*, *Klebsiella terrigena*, *Pseudomonas stutzeri*, *Listeria innocua*, *Staphylococcus warneri* and *S. cohnii*. All species were cultivated independently in nine replicates. The measurements were done on a Bio Particle Explore device similar as for the bacterial spore data. We measured in total 2,708 bacterial cells, almost equally distributed among the six groups
- *Colon tissue of mice:* Raman spectroscopy was performed on colon and rectum tissues of mice in raster-scanning mode. From all the spectra measured, we are particularly interested in those from epithelium and aimed to distinguish normal epithelium from adenoma and carcinoma. The groups of adenoma and carcinoma, which correspond to the states of the adenoma–carcinoma sequence, were combined into one single group named ‘abnormal’. We averaged the spectra belonging to the same group for each scan and used the mean spectra after all preprocessing steps. This led to 219 normal and 266 abnormal spectra from prepared samples belonging to 47 individuals

**Box 4 | Data-based model transfer****Spectral augmentation**

Prerequisite: there are separable Raman bands in the sample's spectra<sup>93</sup>.

- 1 Estimate peak shifts and intensity variation between the primary and secondary mean spectra at predefined Raman bands.
- 2 Augment primary dataset.

For each primary spectrum:

- Generate values of peak shift  $s(\tilde{\nu})$  according to Eq. (B4.1), where the coefficients  $a_i \in [-1, 1]$  are random values

$$s(\tilde{\nu}) = a_0 + a_1\tilde{\nu} + a_2\tilde{\nu}^2 + a_3\tilde{\nu}^3 + a_4\tilde{\nu}^4 \quad (\text{B4.1})$$

- Generate intensity variations  $r(\tilde{\nu})$  according to Eq. (B4.2). To do so, 20 wavenumber positions  $\tilde{\nu}_i$  are randomly picked and used to fit a three-order polynomial relating 20 random numbers  $b_i \in [0.5, 2]$  to these selected wavenumbers (Eq. (B4.2a)). The resulting coefficients are substituted into Eq. (B4.2b) to obtain the intensity variations over the whole spectral region

$$b_i = c_0 + c_1\tilde{\nu}_i + c_2\tilde{\nu}_i^2 + c_3\tilde{\nu}_i^3 \quad (\text{B4.2a})$$

$$r(\tilde{\nu}) = c_0 + c_1\tilde{\nu} + c_2\tilde{\nu}^2 + c_3\tilde{\nu}^3 \quad (\text{B4.2b})$$

- Normalize  $s(\tilde{\nu})$  and  $r(\tilde{\nu})$  so that they fall into the range of the spectral differences estimated in Step 1
  - Enforce these changes into the primary spectrum, and keep the respective group information unchanged
- 3 Build the model based on the augmented primary data.  
Remarks: instead of performing feature extraction using PCA, PLS, etc., on the augmented primary data, we suggest to build these models with the original primary data and then predict the augmented primary data. This helps to obtain more reliable loading vectors. The statistical modeling is performed on the scores from the augmented data.

**Extended multiplicative signal correction**

- 1 Calculate the mean spectrum of primary and secondary data, and collect them in one matrix<sup>94</sup>.

Remarks: we suggest calculating the mean spectrum for each sampling unit separately. The 'sampling unit' in this context refers to the measurements between which the variations need to be removed, for instance, measurements on each device or from each replicate.

- 2 Perform PCA on the mean spectral matrix with column centering.
- 3 Fit each spectrum in both primary and secondary datasets with Eq. (B4.3), and obtain its correction by Eq. (B4.4). Herein the term  $m(\tilde{\nu})$  represents reference spectrum and is usually the mean spectrum of the complete dataset. The vectors  $p_k(\tilde{\nu})$  are given by the loadings from the previous step. The polynomials parameterized by coefficients  $d_i$  are used to contain any slowly changing baseline contributions in each spectrum, and  $n = 2$  or  $n = 3$  is sufficient in most cases.

$$l(\tilde{\nu}) = a + b \cdot m(\tilde{\nu}) + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + \dots + d_n\tilde{\nu}^n + \sum_{k=1}^N g_k \cdot p_k(\tilde{\nu}) + e(\tilde{\nu}) \quad (\text{B4.3})$$

$$l_c(\tilde{\nu}) = \left( l(\tilde{\nu}) - a - d_1\tilde{\nu} - d_2\tilde{\nu}^2 - \dots - d_n\tilde{\nu}^n - \sum_{k=1}^N g_k \cdot p_k(\tilde{\nu}) \right) / b \quad (\text{B4.4})$$

- 4 Rebuild the model based on the corrected primary data.

**Box 5 | Verify the necessity of model transfer**

- 1 (Optional) Build PCA on primary data, and predict secondary data.
- 2 Calculate a numeric marker quantifying the spectral difference between the primary and secondary data with respect to the group differences.
- 3 Compare the calculated marker with a predefined threshold.

In Eq. (B5.1), an example of a numeric marker, namely the relative Fisher's discriminant ratio, is shown. Therein,  $S_m^p$  and  $S_n^p$  represent the spectral matrix of group  $m$ ,  $n$  in the primary data, respectively, while  $S^s$  gives the matrix of the secondary dataset. The term  $d(\cdot, \cdot)$  denotes the Fisher's discriminant ratio defined in ref. <sup>96</sup>.

$$d_r = \sum \frac{d(S_m^p, S_n^p)}{\sqrt{d(S_m^p, S^s) \cdot d(S_n^p, S^s)}}, m, n = 1, 2, \dots, k, m \neq n. \quad (\text{B5.1})$$

The users can replace the Fisher's discriminant ratio with other distance benchmarks, such as Pearson's correlation coefficients<sup>94</sup>.

**Software**

- R Studio (based on Microsoft R Open 3.5.1), open source R packages: we focused our analysis mainly on the R language based on in-house written algorithms and open source R packages. However, it can be performed on other software platforms as well, such as Python and MATLAB. All computations in



this study were performed in R Studio installed on a laptop (OS Win-10, intel Core i7 CPU and 8 GB random-access memory)

## Procedure

### Data importing

- 1 Import spectra and metadata of both samples and the calibration standards based on the predefined data/folder structure.

### Spectral preprocessing

#### Quality control

- 2 Choose a numerical marker to benchmark the data quality, and do the necessary calculations for measuring this. The numerical marker could be:
  - The numerical difference between each spectrum and the mean spectrum of the whole dataset
  - The correlation coefficient between each spectrum and the mean spectrum of the whole dataset
 Instead of using the entire spectrum, the marker can be calculated based on a certain wavenumber region that is more robust to sample variations. For example, it is easier to detect sample degradation like the burning effect in the silent wavenumber region than in the fingerprint region because the silent region is less influenced by the sample variations than the fingerprint region.
- 3 Use a threshold for the calculated marker to detect outliers. The threshold can be  $\mu + t \cdot \sigma$ , where  $\mu$  and  $\sigma$  denote the mean (or median) and standard deviation of the marker values from all spectra, respectively. The value of  $t$  determines the strictness of the outlier detection. A smaller  $t$  identifies more spectra as outliers. We suggest to use a larger  $t$  to avoid false alarms and conduct outlier detection iteratively to avoid masking effects (see Step 5).
- 4 Exclude the outliers from the dataset.
- 5 Repeat Steps 2–4 to remove outliers that were not evident the first time, due to the presence of severe errors in the data.
- 6 Remove spikes by using one of the following approaches based on the property of the spectra. Use the comparison-based method (option A) if you have two measurements for the same sample. Let these two spectra be  $s_1$  and  $s_2$ . Option A can also be employed if the sample is measured in a mapping or time series mode and there is no dramatic change between two successive measurements. In this case, the comparison is performed between the  $i$ th and  $(i+1)$ th spectra, and the output is the corrected spectra without averaging in Step 6A(iv).

The filtering-based methods are suitable for single spectral measurements (option B).

#### (A) Comparison-based method

- (i) Calculate the second-order derivative ( $s_1''$  and  $s_2''$ ) for each of the two spectra and the absolute values of their differences  $d_s = |s_1'' - s_2''|$ .
- (ii) Identify the spikes. These can be identified as regions where  $d_s > \mu + t \cdot \sigma$ , i.e., abnormally large absolute difference between the two derivative spectra. Terms  $\mu$  and  $\sigma$  are the mean and standard variance of  $d_s$ , excluding the values that are below ~5–10% or above ~90–95% percentile. The threshold  $t$  is tunable and normally a value between 5 and 10.
- (iii) Replace each region containing spikes ( $p$ ) occurring on one of the spectra with the corresponding values of the other spectrum:  $(s_i^p - b_i^p) = f \cdot (s_j^p - b_j^p)$ . Herein  $b$  denotes the baseline of the spectrum, and  $f = \frac{\text{mean}(s_i^p - b_i^p)}{\text{mean}(s_j^p - b_j^p)}$  represents the intensity ratio of the two spectra.
- (iv) Calculate the mean spectrum of the two corrected spectra as the output in the case of two repeated measurements.

#### ? TROUBLESHOOTING

#### (B) Filtering-based method

- (i) Calculate the second-order derivative ( $s''$ ) for the spectrum to be corrected.
- (ii) Taking the first five local maxima ( $L_{\max}$ ) with the largest intensity in  $s''$  and their neighboring local minima ( $L_{\min}$ ), making sure  $L_{\max}^i < L_{\min}^i$ ,  $i = 1, 2, \dots, 5$ . Each pair of  $(L_{\max}^i, L_{\min}^i)$  corresponds to one spike candidate.
- (iii) Exclude the candidates from further correction unless they meet both of the two criteria: peak width smaller than a predefined threshold  $t$  (i.e.,  $L_{\min}^i - L_{\max}^i < t$ ); absolute values of  $s''[L_{\min}^i]$  and  $s''[L_{\max}^i]$  both larger than a predefined threshold (i.e., the peak features a sharp increase following a sharp decrease).

- (iv) Correct each of the detected spikes by a linear interpolation based on the two boundary points of the spike.

#### Wavenumber calibration

- 7 Choose the spectrum of the wavenumber standard that is acquired at the time point closest to the measurement of the samples.  
**▲ CRITICAL STEP** Do the calibration separately for spectra that are measured at different time points using the most appropriately timed standard spectra, especially if the samples are measured over a long period.
- 8 If multiple spectra are measured for the wavenumber standard, remove those of poor quality (e.g., high fluorescence background, low signal-to-noise ratio) and take the average from the good measurements.
- 9 Search the peak positions  $\tilde{\nu}_m$  of the well-defined Raman bands  $\tilde{\nu}_t$  on the standard spectrum. To do so, a neighborhood of each Raman band is defined according to a priori knowledge of the position and width of this band. The peak position of the Raman band can be the center of a Gaussian peak fitted in this predefined neighborhood.
- 10 Fit a function (e.g., polynomial, spline) relating the peak positions on the measured standard spectrum  $\tilde{\nu}_m$  to their theoretical values  $\tilde{\nu}_t$ . A three-order polynomial works well in most cases ( $\tilde{\nu}_t(i) = a_0 + a_1 \cdot \tilde{\nu}_m(i) + a_2 \cdot \tilde{\nu}_m(i)^2 + a_3 \cdot \tilde{\nu}_m(i)^3$ ).
- 11 Calculate the new wavenumber axis from the fitted function according to  $\tilde{\nu}_{new}(i) = a_0 + a_1 \cdot \tilde{\nu}_{old}(i) + a_2 \cdot \tilde{\nu}_{old}(i)^2 + a_3 \cdot \tilde{\nu}_{old}(i)^3$ .
- 12 Interpolate the samples spectra onto the new wavenumber axis based on the following equations:  

$$f_s = \text{spline}(\tilde{\nu}_{old}, I_{old})$$

$$I_{new} = f_s(\tilde{\nu}_{new})$$

#### Intensity calibration

**▲ CRITICAL** Similar to wavenumber calibration, the standard spectra have to be those acquired at a time point closest to the measurement time of the sample spectra.

**▲ CRITICAL** Intensity calibration is, in general, not necessary in qualitative tasks. However, it becomes necessary in one of the following scenarios: (a) peak ratio is important for the analysis; (b) the intensity response function is not smooth and introduces fake bands/features into measured Raman spectra or changes the shape of Raman bands; (c) cross-instrumental comparison/analysis is required.

#### ? TROUBLESHOOTING

- 13 Conduct wavenumber calibration on the emission spectra  $E^m(\tilde{\nu})$  of the intensity standard as well as the dark current spectrum  $D^m(\tilde{\nu})$ .
- 14 Normalize the measured and theoretical emission spectra against their respective maximum.
- 15 Divide the theoretical by the measured emission spectra to get the response function:  $R(\tilde{\nu}) = E^t(\tilde{\nu})/E^m(\tilde{\nu})$ .
- 16 Correct each measured Raman spectrum following the equation below:  

$$I^c(\tilde{\nu}) = (I^m(\tilde{\nu}) - D^m(\tilde{\nu})) \cdot R(\tilde{\nu})$$

#### ? TROUBLESHOOTING

#### Baseline correction

**▲ CRITICAL** Steps 17 and 18 can be merged into one step in some approaches like EMSC by inserting the reference spectrum of substrate to be removed into the EMSC model.

- 17 Remove fluorescence baseline with one of the mathematical approaches, such as ALS<sup>57</sup>, polynomial fitting<sup>58</sup> and EMSC<sup>59</sup>. Alternatively, we direct readers to Box 2 showing a procedure of automatically choosing the optimal baseline correction method and parameters.
- 18 Remove contributions from substrate or water, if they are substantial and impact the analysis or interpretation of the Raman signals.

#### Spectral truncation

- 19 Cut off wavenumber regions where no substantial Raman signals are present. In biological applications, the region 1,800–2,800  $\text{cm}^{-1}$  is usually excluded as the silent region. However, this is not the case, for instance, if isotope labeling is used and Raman bands appear within this region<sup>100</sup>.

### Normalization

- 20 Divide the Raman intensities with the maximum, area, or  $l_2$  norm of certain Raman regions or the whole spectral region of interest (see Eqs. (2,3)).

### Visualization

- 21 **Spectral overview:** visualize the mean spectra, and note the standard deviation of each group (see Fig. 1 as example). Also look at the difference spectra obtained by subtracting mean spectra obtained from different groups. As a rough guide, the wavenumber positions where the difference is larger than the standard deviation are considered as substantial features. This can give a practical knowledge such as which chemical composition changes contribute to the group differences.
- 22 **Correlation coefficients:** calculate the correlation coefficients between mean spectra of different groups as well as the correlation coefficients between mean spectra of different replicates belonging to the same group. Visualize the two sets of values, for instance, in a heat map. This provides a general idea of the intergroup and interreplicate variations of the dataset (see Fig. 14 as an example).
- 23 **Loading and score plots:** perform a PCA on the dataset, and plot the loading vectors and the scores. Scatter plots generated from the scores of two or three components can give an overview of how the sample spectra cluster together (see Fig. 11 as an example). This also shows whether the groups are easily separable (their separability). The loading vectors reveal the features (chemical composition) that contribute to each principal component.

### Data learning

#### Statistical sampling

- 24 Split the whole dataset into  $N$  folds.  $N$  is typically a small number between 5 and 25 and depends on the data measured. Three is the minimal number of folds.  
**▲ CRITICAL STEP** Two issues are important for this split. First, spectra from the same sampling unit should be arranged exclusively into one fold. The ‘sampling unit’ here refers to the highest level in data hierarchy, such as replicate, batch or individual (patient, mouse, etc.). Second, the distribution of different groups or concentrations in the whole dataset should be preserved in each fold. That means all the different groups and concentrations present in the whole dataset should be represented in each fold.

#### Statistical modeling

**▲ CRITICAL** The selection of a statistical model to be used is data dependent. The most employed model is PCA-linear discriminant analysis (LDA). It is easy to use, but it is worth noting that LDA assumes the same covariance for all different groups. It is better to use SVM with a linear kernel if this assumption does not hold.

- 25 Construct a statistical model following the procedure of two-layer CV as described below. Select one of the folds, the  $i$ th fold, to be testing data ( $i \leq N$ ). This will be used for external validation.  
**? TROUBLESHOOTING**
- 26 **Model construction and internal validation** Choose one of the folds, the  $j$ th fold, as validation data ( $j \leq N, j \neq i$ ). This will be used for internal validation. All folds other than the  $v_i, j^{th}$  are used as training data.
- 27 **Dimension reduction** Input training data into one of the following methods: PCA, PLS, forward/backward feature selection, etc.
- 28 Depending on the used method, either the PCA/PLS model or the selected wavenumber indices of the feature selection method are saved.
- 29 (Optional) Repeat Steps 27–28 with different dimensionalities  $M$  to test the influence of the size of the new feature space. An example would be to test different numbers of principal components if a PCA is utilized.
- 30 **Modeling** Take each output of dimension reduction on the training data, and feed these outputs into a statistical model, such as LDA, SVM or artificial neural network. Choose parameters that you will use in the training of the statistical model, which are called hyper-parameters.
- 31 Train this model using the training data and its ground truth, e.g., minimal error on training data.
- 32 Save the trained model for further use on the computer or in a file system.

- 33 Repeat Steps 30–32 to test multiple combinations of the dimensionality  $M$  (of the feature extraction) and different (hyper-)parameters of the employed model.
- 34 **Internal validation** Predict the validation data with the dimension reduction model built on the training data.
- 35 Feed the output into the trained models, and record the prediction.
- 36 Calculate the predefined benchmark based on the prediction and the ground truth, for example, accuracy, sensitivity, specificity, RMSE.
- 37 Collect the benchmarks of all models for this validation dataset.
- 38 Repeat Steps 34–37, for each model based on different combinations of the dimensionality  $N$  and model hyper-parameters.
- 39 Collect the benchmarks of all built models for this validation dataset. Repeat Steps 26–39 using a different fold as validation data.
- 40 **Optimization** Fuse the benchmarks over all validation datasets for each built model. Different strategies can be used for this fusion: average, median, or sum of ranking difference<sup>47,101</sup>.
- 41 Select the combination of the dimensionality  $N$  and model hyper-parameters that lead to the best prediction on the validation data according to the fused benchmarks.
- 42 Repeat Steps 40–41 with the optimal dimensionality  $M$  and model hyper-parameters (based on Step 39) to train the final model.
- 43 **External validation** Predict the testing data based on the dimension reduction and statistical model.
- 44 Calculate the benchmark of the prediction.
- 45 Store the rebuilt model and the benchmark value. Repeat Steps 25–45 using a different fold as testing data.

#### Predict new data

- 46 Preprocess the new data using approaches exactly the same as for the training data.
- 47 Check whether it is necessary to perform model transfer by following the workflow described in Box 5. If necessary, turn to model transfer (Steps 50–51); otherwise continue.
- 48 Predict the new data with the model built in statistical modeling (Step 25).

#### ? TROUBLESHOOTING

#### Model transfer

- 49 There might be more than one model built with the procedure given in Box 3, each corresponding to one iteration of the external validation. The prediction of these multiple models can be fused, taking into account their performance benchmarked by statistical modeling (Step 25).
- 50 Choose one of the methods given in Boxes 3 and 4 depending on the properties of the dataset, and conduct the model transfer following the corresponding steps.
- 51 Predict the secondary data using the rebuilt model in the case of model-based model transfer; or conduct the prediction for the transformed secondary data in the case of data-based model transfer.

### Troubleshooting

**Step 6A** (comparison-based spikes removal): if the fluorescence baseline varies dramatically for the two successive measurements, we suggest to perform a baseline correction prior to Step 6A(i) and then add the estimated baseline to the spikes-removed spectrum after Step 6A(iv).

**Steps 7–16** (wavenumber and intensity calibration): in very rare cases, the wavenumber axis may be different for the calibration standards and the samples. If this happens, the spectra of the samples should be resampled onto the wavenumber axis of the calibration standards before performing the spectrometer calibration.

**Steps 24–25** (sampling and statistical modeling): if the number of sampling units is limited (e.g., fewer than three), it may become impossible to form a sufficient number of independent folds for the two-layer CV. In this case, the internal validation can be replaced with a normal  $k$ -fold CV, i.e., to split the dataset excluding the independent testing data randomly into multiple folds ignoring the information of the sampling unit. The statistical modeling based on this procedure is still valid as long as the testing data are from an independent sampling unit<sup>69</sup>.

**Steps 46–49** (predict new data): the wavenumber axis for the new data to be predicted may differ from the training data. In this case, it is important to interpolate the new data onto the wavenumber axis of the training data after the wavenumber calibration.

## Timing

In this section, we describe the time required for the analysis procedures that are substantially time-consuming: baseline correction and data learning.

**Steps 17–18** (baseline correction): the  $R^{12}$ -based optimization using the workflow of Eq. (B2.1a) in Box 2 takes ~3–4 min, in which the optimization is based on genetic algorithm (150 generations) using the mean spectra of different groups (three groups in our study). The speed of baseline correction varies greatly with the employed method, e.g., 3.5 min and 2 s for 1,553 spectra using ALS and sensitive nonlinear iterative peak method, respectively

**Steps 24–25** (data learning): 8 min in case of three-group classification using PCA-LDA for a dataset composed of nine replicates (1,553 spectra). The number of principal components was optimized over 49 values, to equal 2:50

## Anticipated results

In this section, we describe the analysis of our four benchmark datasets measured from different biological samples. We will focus particularly on the application of the approaches introduced above to deal with two important challenges of the analysis in Raman spectra: proper model evaluation, and model transfer.

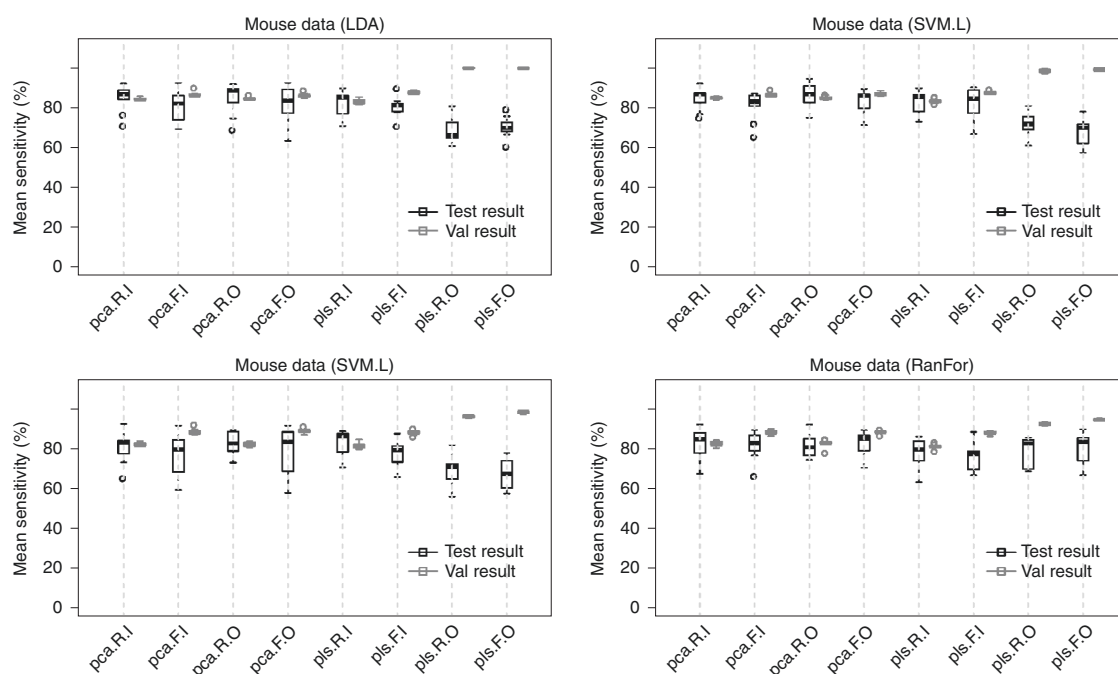
### Statistical sampling and model evaluation

The analysis in this subsection was conducted on the datasets measured from tumor cells, vegetative bacteria and mice colon tissue. The basic independent unit of these studies are the cell replicate, bacteria replicate and individual mouse, respectively. To make the description easier, we will denote these different basic units all as ‘replicate’. Spectral preprocessing was performed for all datasets based on the SOP given in the previous section. The resulting data were used for data learning under the framework of statistical modeling (Fig. 8). The procedure was the same for all datasets and can be summarized as follows.

- 1 To start, the whole dataset was split into nine folds.
- 2 Data from the same replicate were included exclusively in one fold.
- 3 The external validation iterates using each fold once as the testing data.
- 4 The remaining eight folds were utilized in the internal validation, where a CV was conducted using two mechanisms: a random CV and a replicate CV.
- 5 For random CV, the input data were mixed and redistributed randomly into eight folds.
- 6 In replicate CV, the redistribution was skipped i.e., the data from the same replicate were kept exclusively in one fold.
- 7 For both random and replicate CV, each fold was used once as validation data and predicted with the model trained on the remaining folds (training data).
- 8 The trained model (from the eight folds) was used to predict the testing data (left-out fold in Step 3), i.e., all spectra of the testing data.
- 9 We choose the mean sensitivity of the prediction as marker. When calculated based on the validation data, it is called validation results. It is calculated within the inner loop of the two-layer CV (internal validation).
- 10 The mean sensitivity of the prediction of the testing data is called testing results, and it is calculated based on the outer loop of the two-layer CV (external validation).

Given the strategy described so far, we can assume that the testing data are independent from the training/validation data since they are from different replicates. Therefore, we will treat the testing results as a unbiased model evaluation and compare them with the validation results. The internal validation will be considered unbiased only if it gives validation results comparable to the testing results. To make the conclusion more general, we based the classification on two dimension-reduction methods (PCA, PLS) and four classifiers (LDA, linear SVM, radial SVM and random forest). Particularly, we performed the dimension reduction either outside or inside the loop of the internal validation, that is, involving or excluding the validation data during the dimension reduction, respectively.

The results of the mouse data are shown in Fig. 9, in which the validation and testing results are represented by ‘Val result’ and ‘Test result’, respectively. Using PCA for dimension reduction, we could say the internal validation was an unbiased evaluation as long as the replicate CV was used. That was proven by the agreement between the validation and testing results. The way PCA was



**Fig. 9 | Validation and testing results of the mice data.** Each box was generated out of nine values representing the mean sensitivity of the validation and testing results coming from the nine folds of the external CV loop. The external validation is a ninefold or a nine-replicate CV. Each fold was used once as the test data, while the remaining eight folds/replicate were used for eightfold/replicate internal validation. Therefore, we have nine iterations for the external validation, within each of the eight iterations of internal validation. Accordingly, the 'test result' corresponds to the mean sensitivities from the nine external iterations. The 'validation result' gives the best mean sensitivity of the internal validation within each external iteration. The classification was performed using two dimension reduction methods and four classifiers in the framework of different sample sampling. The internal validation is considered unbiased if the testing and validation results are comparable; otherwise it is biased.

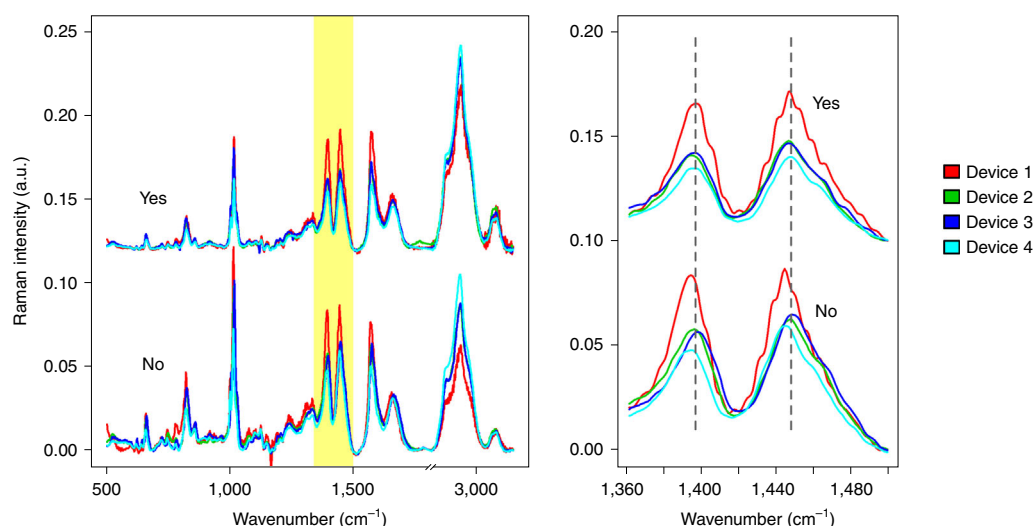
performed, inside or outside the validation loop, did not seem to matter. In the case of PLS, however, the internal validation was observed to be unbiased only if the replicate CV was applied with PLS inside the validation loop. This conclusion could be drawn for all the classification models employed. Explaining this phenomenon takes us back to the independence requirement for a reliable model evaluation: (i) the data to be predicted have to be independent from those used for model training; (ii) the model training has to be done excluding completely the data to be predicted. Obviously, the random CV violates the first requirement, rendering the internal validation biased. In this case, the model is highly overestimated, which manifested itself by much higher validation results than the testing results. The second requirement does not hold by performing dimension reduction outside the validation loop. However, its influence on the model evaluation partly depends on the methods of dimension reduction. In the case of an unsupervised approach, such as PCA, the information to be predicted is not considered during the calculation. The usage of validation data does not necessarily lead to a better prediction of the model on these data. Supervised dimension reduction methods are different, as they utilize the group information to be predicted and are greedy in pursuing the best prediction. If the validation data are involved during the calculation, the validation data can substantially modulate the model to reach as good a prediction as possible for both training and validation data. Evaluating the model with the validation data is hence strongly biased in this case.

The results of the other two datasets are shown in Extended Data Figs. 2 and 3, and they lead to the same conclusion as discussed above. These results indicated two important guidelines of statistical modeling: the model has to be evaluated with data that are independent from the training data; the data used for such evaluation have to be excluded from the modeling (i.e., dimension reduction and classification). The second point deserves more caution for supervised steps in which the information to be predicted is used to guide the calculation.

### Instrumental variations and spectrometer calibration

In this section, we focus on the spectral variations originated from the instruments and the influence of the spectrometer calibration on the data analysis. The investigation was performed on the dataset





**Fig. 10 | Mean spectra of single cell spectra from the group *B. mycoides* measured using the four devices.** The spectra with and without spectrometer calibration are marked as ‘Yes’ and ‘No’, respectively. The right plot shows a zoom-in version of the spectral region highlighted in yellow. It is clearly observed that the interdevice spectral variations decreased substantially after the calibration but are not completely removed. The remaining difference is still obvious, especially between the first and the other three devices (adjusted from the plot in ref. <sup>49</sup>). The mean spectra are generated from 508, 254, 230 and 206 single-cell spectra for the four devices, respectively.

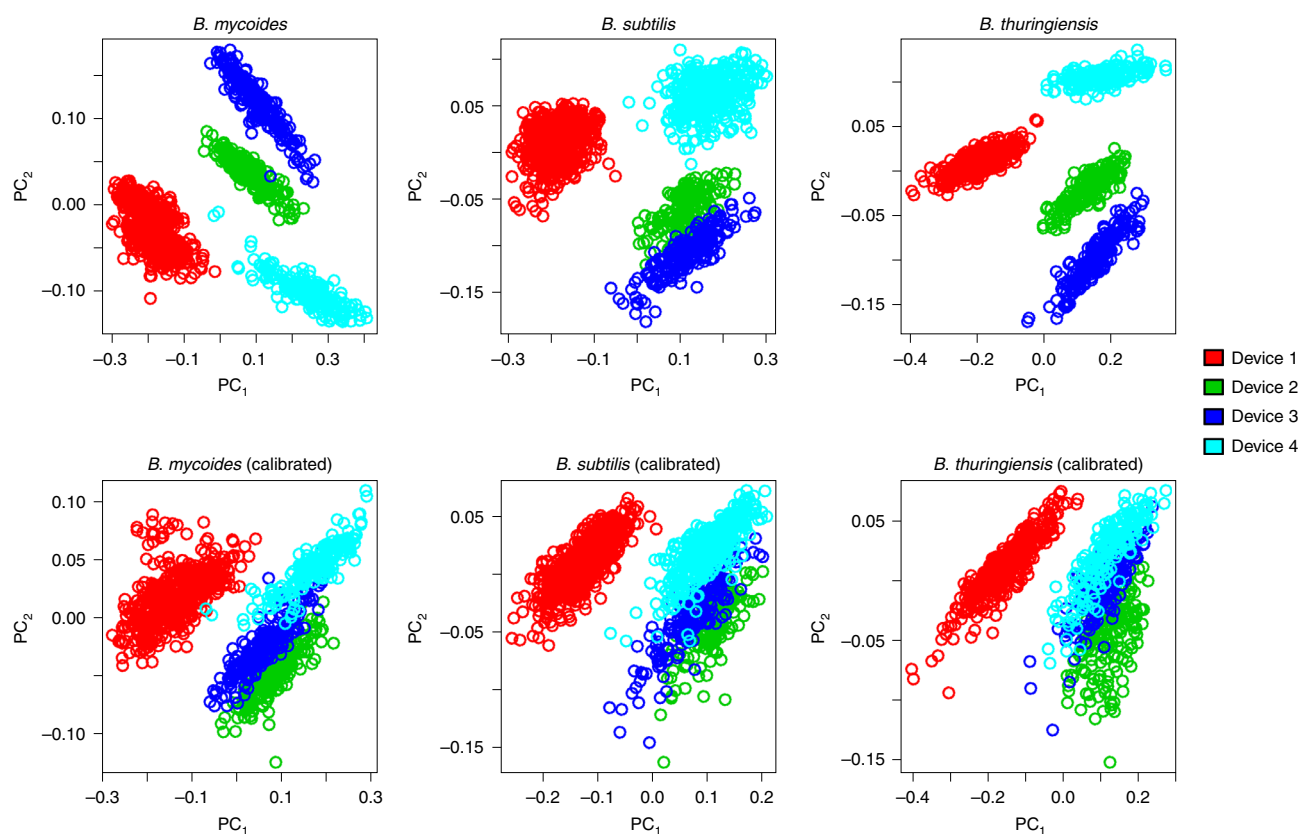
of bacterial spores measured with four devices. Shown in Fig. 10 are the mean spectra after data preprocessing for the group *B. mycoides* from the four devices. The data preprocessing was done with or without the spectrometer calibration, marked in the figure by ‘Yes’ and ‘No’, respectively. The impact of the calibration can be seen as the decrease of the interdevice spectral differences. However, the decrease is limited, as the calibrated spectra still vary substantially from device to device. This is clearly shown in the plot on the right side, which is a zoomed-in version of the yellow highlighted region. On one hand, the spectrometer calibration is intended to remove device-related variations, and it cannot correct spectral changes originating from other factors. Therefore, the spectral differences between data from the first and the other three devices, originating mainly from the substrate, were still present after spectrometer calibration. On the other hand, the performance of the spectrometer calibration, even for the device-related variations, is limited. This can be seen from the remaining spectral variations between the data of devices 2–4, which were measured from identical samples.

To make the conclusion clearer, we additionally constructed PCA models for the spectra of the same group but from different devices. The PCA scores of the first two components are visualized as scatter plots in Fig. 11. The scores before and after calibration are given in the first and second row, respectively. A clear separation can be seen between clusters formed by different devices without the spectrometer calibration, and this was common for all three groups. Such between-cluster distance decreased substantially but was still present after applying a spectrometer calibration. Spectra can still be easily distinguished according to the device they were measured on. Additionally, the distance between the first and the other three devices was more obvious, which is mainly due to the substrate difference.

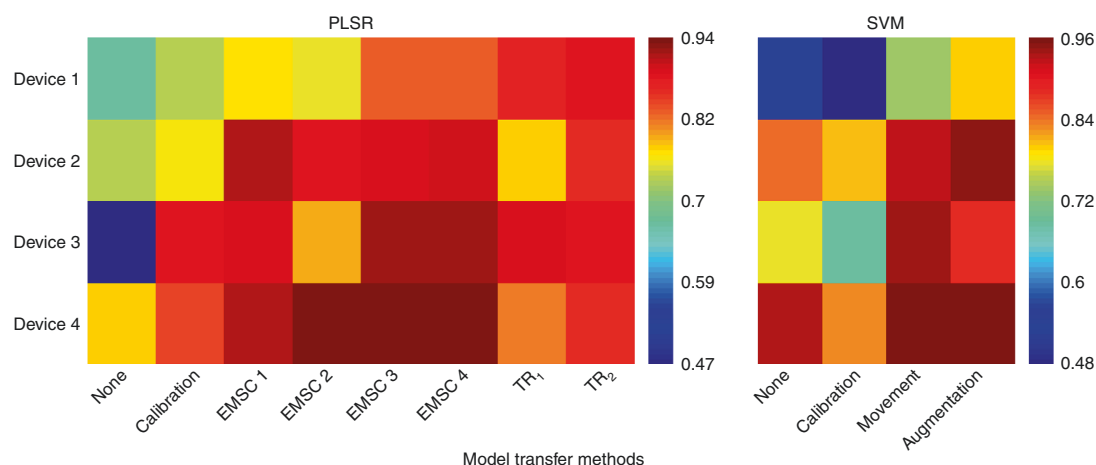
Based on above results in both the spectral and scores space, it was demonstrated that the spectrometer calibration is limited in removing device-related spectral differences and cannot handle sample-related spectral variations. It is thus very common in Raman spectroscopy to see testing data substantially differ from the training data, which often leads to a failed prediction of a trained statistical model on this testing data. Approaches related to this issue, namely model transfer, are described in the next subsection.

### Model transfer

The data analysis in this subsection is based on the spore’s dataset measured on four devices. We start with the results of the different model transfer approaches introduced in Boxes 3 and 4. A three-group classification was performed to distinguish the three bacterial species using either a partial least squared regression (PLSR) or a SVM. This classification was conducted separately for each of the

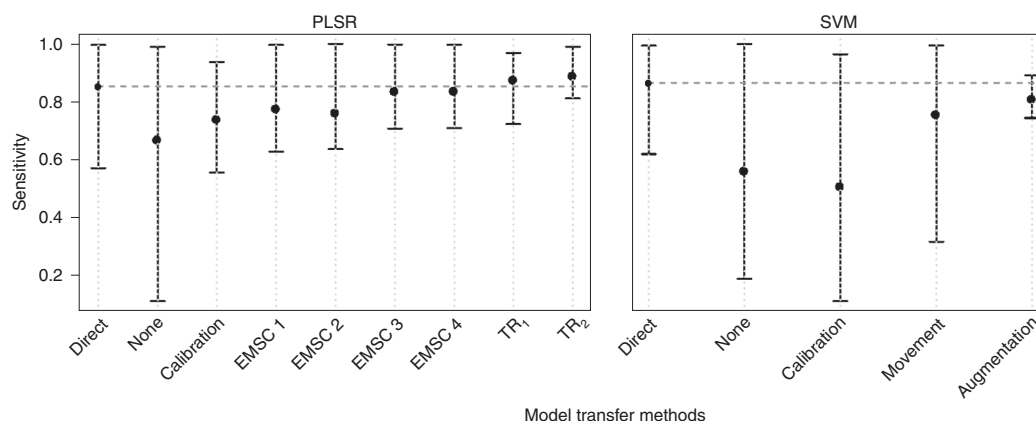


**Fig. 11 | PCA score plot of bacteria spore spectra.** The PCA models were built on the spectra from the same group but measured with different devices, with and without calibration. The spectra from different devices became closer after calibration for all the three groups, shown as the smaller distance between the clusters. However, we can still see a clear separation between the devices, especially for device 1, where a different substrate was used.



**Fig. 12 | Prediction of data from different devices in a leave-one-device-out CV with different model transfer methods.** The mean sensitivities of the leave-one-device-out CV is shown, which should be as high as possible and ideally 1. The analysis was based on two different classifiers: PLS regression and SVM. The performance of the spectrometer calibration was limited, shown by the low prediction. The term 'direct' refers to the prediction of one device with a model built directly on the same device. The label 'none' means prediction of one device with a model built on the other devices but with no model transfer. The results were substantially improved after employing the model transfer methods. Data adapted from ref. <sup>94</sup>.

model transfer methods. The classifier was trained always on three out of the four devices and subsequently used to predict the other device (i.e., interdevice prediction). The resulting mean sensitivities are shown in Fig. 12, which are visualized separately for the prediction on each device.



**Fig. 13 | Prediction on the data of the first device based on the PLS regression or SVM classifiers.** In the figure, the sensitivities of the three groups are visualized, which should be as high as possible and ideally 1. The column 'direct' represents sensitivities of the three groups resulting from a leave-one-replicate-out CV on the first device. The other columns represent sensitivities of the three groups, if the data of the first device were predicted by the model built on the other three devices. The employed model transfer methods are noted on the x axis. In particular, the 'direct' term represents the prediction by the model built on the first device using a leave-one-replicate CV. Data adapted from ref. <sup>94</sup>.

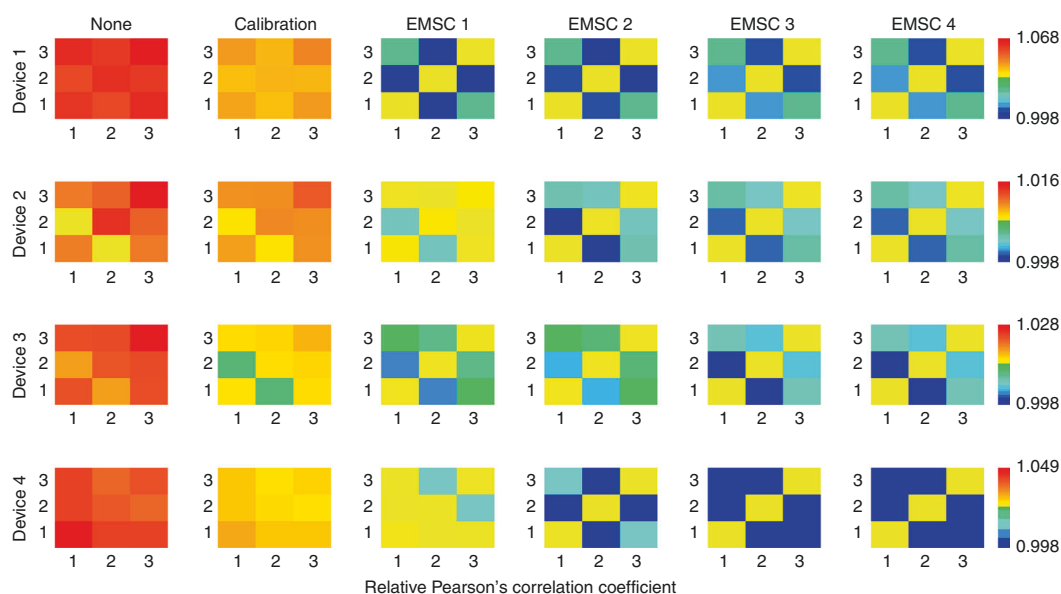
The *x* axis indicates the different model transfer techniques. In particular, 'none' represents the case with neither spectrometer calibration nor model transfer, while 'calibration' represents the situation where the data were subject to spectrometer calibration but none of the model transfer methods was applied. The other terms denote the model transfer approaches employed along with spectrometer calibration. The numbers 1, 2, 3, 4 following 'EMSC' give the number of principle components used in the replicate EMSC model. As can be seen, the prediction was rarely adequate without any model transfer. The limit of spectrometer calibration in removing the unwanted spectral variations from the devices was proven by the relatively poor interdevice prediction, especially in the case of SVM. The prediction was substantially improved by applying additional model transfer methods, showing the necessity for model transfer.

As a further verification, we conducted additionally a leave-one-replicate-out CV with the data from the first device, using both of the two classifiers. The prediction was named 'direct prediction' and compared with the interdevice prediction in different cases of model transfer (i.e., to predict the first device with the model built on devices 2–4). The resulting sensitivities of the three groups are shown in Fig. 13. The upper and bottom ends of the bars represent the maximal and minimal sensitivity from the three groups, respectively. The dots represent the mean sensitivity. The interdevice prediction is shown to be comparable or even superior to the direct prediction with the help of model transfer methods like replicate EMSC- and Tikhonov regularization-based methods. The score movement-based method is observed to be less adequate, most probably due to the large spectral variations between the Raman spectra from the first and other devices.

Along with the model transfer approaches, we proposed to check the necessity of model transfer based on the marker defined in Eq. (5.1) (Box 5). In general, a model transfer is necessary if the marker is above 1; i.e., the interdevice variations are larger than the intergroup difference. As a validation, we calculated the marker based on the data processed in different ways: without spectrometer calibration, with spectrometer calibration, with spectrometer calibration and replicate EMSC. The replicate EMSC was performed including one, two, three or four principal components in the model. The results of the marker are shown in Fig. 14, in which the values >1, equal to 1, or <1 are coded as red/yellow, transparent and green/cyan/blue, respectively. Accordingly, we could conclude that model transfer is needed in the case of the first two columns. This aligns with the previous results shown in Fig. 13, in which the interdevice prediction was inadequate and substantially worse than direct prediction. In contrast, the marker was mostly <1 for the data subject to replicate EMSC, denying the necessity of model transfer. As a proof, we could see that in these cases the results for interdevice predictions were comparable to the direct predictions in Fig. 13.

## Summary

We provide a Raman spectral data analysis protocol in this contribution. We start with a general overview of the chemometric procedures in Raman spectroscopy, including the experimental design, data preprocessing, data learning and model transfer. The possible pitfalls are discussed together with



**Fig. 14 | Results of relative Pearson's correlation coefficients.** The calculation was done from data produced by different model transfer methods. Values  $>1$ , equal to 1, or  $<1$  are coded as red/yellow, transparent and green/cyan/blue, respectively. A value  $>1$  means the interdevice variations are larger than the intergroup difference, i.e., a model transfer is needed for satisfying interdevice predictions. Accordingly, the model transfer is needed in the case of the first two columns. This aligns with the previous results shown in Fig. 13, in which the interdevice prediction was inadequate and substantially worse than direct prediction. With the marker generally  $<1$ , in contrast, a satisfying interdevice prediction is possible without model transfer for the data subject to replicate EMSC. Data adapted from ref. <sup>94</sup>.

the potential solutions. We particularly described the methods dealing with special issues in Raman spectroscopy: the SSP, the optimization of baseline correction, and the model transfer. This is followed by a SOP that walks through the chemometric steps and data learning in Raman spectroscopy. Thereafter, we presented the results on four biological datasets to pinpoint two important issues in Raman spectral analysis: model evaluation and model transfer. After the model is finished, a model interpretation by means or variable importance should be performed and the variables should be linked to the spectroscopic data. With this contribution, we hope to help the standardization of Raman spectral analysis and hence to push Raman-based technologies from proof-of-concept studies further to real-world applications.

### Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

One example dataset used to demonstrate the protocol has been made openly accessible within the GitHub repository: <https://github.com/Bocklitz-Lab/Example-Raman-spectral-analysis>. Other data can be found with this protocol and the supporting primary research papers. Source data are provided with this paper.

### Code availability

Code can be found in various open-source packages. One example analysis code can be found in GitHub: <https://github.com/Bocklitz-Lab/Example-Raman-spectral-analysis>.

## References

1. Popp, J. et al. *Handbook of Biophotonics* Vol. 1 (Wiley-VCH, 2011).
2. McCreery, R. L. *Raman Spectroscopy for Chemical Analysis* Vol. 225 (John Wiley & Sons, 2005).
3. Cheng, J.-X. & Xie, X. S. Vibrational spectroscopic imaging of living systems: an emerging platform for biology and medicine. *Science* **350**, aaa8870 (2015).

4. Bocklitz, T. W. et al. Raman based molecular imaging and analytics: a magic bullet for biomedical applications? *Anal. Chem.* **88**, 133–151 (2016).
5. Lorenz, B. et al. Cultivation-free Raman spectroscopic investigations of bacteria. *Trends Microbiol.* **25**, 413–424 (2017).
6. Liu, C.-Y. et al. Rapid bacterial antibiotic susceptibility test based on simple surface-enhanced Raman spectroscopic biomarkers. *Sci. Rep.* **6**, 23375 (2016).
7. Prochazka, D. et al. Combination of laser-induced breakdown spectroscopy and Raman spectroscopy for multivariate classification of bacteria. *Spectrochim. Acta B. Spectrosc.* **139**, 6–12 (2018).
8. Silge, A. et al. The application of UV resonance Raman spectroscopy for the differentiation of clinically relevant *Candida* species. *Anal. Bioanal. Chem.* **410**, 5839–5847 (2018).
9. Hanson, C. et al. Simultaneous isolation and label-free identification of bacteria using contactless dielectrophoresis and Raman spectroscopy. *Electrophoresis* **40**, 1446–1456 (2019).
10. Van Nest, S. J. et al. Raman spectroscopy detects metabolic signatures of radiation response and hypoxic fluctuations in non-small cell lung cancer. *BMC Cancer* **19**, 474 (2019).
11. Marro, M. et al. Unravelling the metabolic progression of breast cancer cells to bone metastasis by coupling Raman spectroscopy and a novel use of MCR-ALS algorithm. *Anal. Chem.* **90**, 5594–5602 (2018).
12. Aljakouch, K. et al. Raman microspectroscopic evidence for the metabolism of a tyrosine kinase inhibitor, neratinib, in cancer cells. *Angew. Chem. Int. Ed.* **57**, 7250–7254 (2018).
13. Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman spectroscopy. *Chem. Soc. Rev.* **45**, 1958–1979 (2016).
14. Kong, K. et al. Raman spectroscopy for medical diagnostics—from in-vitro biofluid assays to in-vivo cancer detection. *Adv. Drug Deliv. Rev.* **89**, 121–134 (2015).
15. Koo, K. M. et al. Design and clinical verification of surface-enhanced Raman spectroscopy diagnostic technology for individual cancer risk prediction. *ACS Nano* **12**, 8362–8371 (2018).
16. Doty, K. C. & Lednev, I. K. Raman spectroscopy for forensic purposes: recent applications for serology and gunshot residue analysis. *TrAC Trends Anal. Chem.* **103**, 215–222 (2018).
17. Khandasammy, S. R. et al. Bloodstains, paintings, and drugs: Raman spectroscopy applications in forensic science. *Forensic Chem.* **8**, 111–133 (2018).
18. de Oliveira Penido, C. A. F. et al. Raman spectroscopy in forensic analysis: identification of cocaine and other illegal drugs of abuse. *J. Raman Spectrosc.* **47**, 28–38 (2016).
19. Guo, S., Ryabchykov, O., Ali, N., Houhou, R. & Bocklitz, T. Comprehensive chemometrics. in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (eds Brown, S. D. et al.) 333–360 (Elsevier, 2020).
20. Ryabchykov, O., Guo, S. & Bocklitz, T. Analyzing Raman spectroscopic data. in *Micro-Raman Spectroscopy: Theory and Application* (eds Popp, J. & Mayerhöfer, T.) 81–106 (De Gruyter, 2020).
21. Guo, S. et al. Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study. *Anal. Chem.* **92**, 15745–15756 (2020).
22. Morais, C. L. et al. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* **15**, 2143–2162 (2020).
23. Baker, M. J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771 (2014).
24. Ryabchykov, O., Guo, S. & Bocklitz, T. Analyzing Raman spectroscopic data. *Phys. Sci. Rev.* <https://doi.org/10.1515/psr-2017-0043> (2019).
25. Butler, H. J. et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **11**, 664 (2016).
26. Smith, E. & Dent, G. *Modern Raman Spectroscopy: A Practical Approach* (Wiley, 2019).
27. Quinn, G. P. & Keough, M. J. *Experimental Design and Data Analysis for Biologists* (Cambridge University Press, 2002).
28. Shreve, A. P., Cherepy, N. J. & Mathies, R. A. Effective rejection of fluorescence interference in Raman spectroscopy using a shifted excitation difference technique. *Appl. Spectrosc.* **46**, 707–711 (1992).
29. Zhao, J., Carrabba, M. M. & Allen, F. S. Automated fluorescence rejection using shifted excitation Raman difference spectroscopy. *Appl. Spectrosc.* **56**, 834–845 (2002).
30. Guo, S. et al. Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS). *Talanta* **186**, 372–380 (2018).
31. Matousek, P. et al. Subsurface probing in diffusely scattering media using spatially offset Raman spectroscopy. *Appl. Spectrosc.* **59**, 393–400 (2005).
32. Bocklitz, T. et al. Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **149**, 544–549 (2015).
33. Dörfer, T. et al. Checking and improving calibration of Raman spectra using chemometric approaches. *Z. Phys. Chem.* **225**, 753–764 (2011).
34. ASTM E1840–96(2014): *Standard Guide for Raman Shift Standards for Spectrometer Calibration* (ASTM International, 2014).
35. Carrabba, M. M. Wavenumber standards for Raman Spectrometry. in *Handbook of Vibrational Spectroscopy* Vol 1 (Wiley, 2006).
36. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *J. Biomed. Inform.* **48**, 193–204 (2014).



37. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
38. Gy, P. *Sampling for Analytical Purposes* (John Wiley & Sons, 1998).
39. Saccenti, E. & Timmerman, M. E. Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. *J. Proteome Res.* **15**, 2379–2393 (2016).
40. Cohen, J. Statistical power analysis. *Curr. Dir. Psychol. Sci.* **1**, 98–101 (1992).
41. Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**, 591–605 (2007).
42. Ali, N. et al. Sample-size planning for multivariate data: a Raman-spectroscopy-based example. *Anal. Chem.* **90**, 12485–12492 (2018).
43. Beleites, C. et al. Sample size planning for classification models. *Anal. Chim. Acta* **760**, 25–33 (2013).
44. Bocklitz, T. et al. How to pre-process Raman spectra for reliable and stable models? *Anal. Chim. Acta* **704**, 47–56 (2011).
45. Heraud, P. et al. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. *J. Chemom.* **20**, 193–197 (2006).
46. Penny, K. I. & Jolliffe, I. T. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *J. R. Stat. Soc. D.* **50**, 295–307 (2001).
47. Brownfield, B. & Kalivas, J. H. Consensus outlier detection using sum of ranking differences of common and new outlier measures without tuning parameter selections. *Anal. Chem.* **89**, 5087–5094 (2017).
48. Ryabchykov, O. et al. Automatization of spike correction in Raman spectra of biological samples. *Chemom. Intell. Lab. Syst.* **155**, 1–6 (2016).
49. Guo, S. et al. Towards an improvement of model transferability for Raman spectroscopy in biological applications. *Vib. Spectrosc.* **91**, 111–118 (2017).
50. Bloemberg, T. G. et al. Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. *Anal. Chim. Acta* **781**, 14–32 (2013).
51. Tomasi, G., Van Den Berg, F. & Andersson, C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemom.* **18**, 231–241 (2004).
52. Liu, Y.-J. et al. Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW). *Anal. Chim. Acta* **952**, 9–17 (2017).
53. Beier, B. D. & Berger, A. J. Method for automated background subtraction from Raman spectra containing known contaminants. *Analyst* **134**, 1198–1202 (2009).
54. McLaughlin, G., Sikirzhyski, V. & Lednev, I. K. Circumventing substrate interference in the Raman spectroscopic identification of blood stains. *Forensic Sci. Int.* **231**, 157–166 (2013).
55. McLaughlin, G. et al. Universal detection of body fluid traces in situ with Raman hyperspectroscopy for forensic purposes: evaluation of a new detection algorithm (HAMAND) using semen samples. *J. Raman Spectrosc.* **50**, 1147–1153 (2019).
56. Ryan, C. et al. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrum. Methods Phys. Res. B* **34**, 396–402 (1988).
57. Eilers, P. H. & Boelens, H. F. Baseline correction with asymmetric least squares smoothing. *Leiden. Univ. Med. Cent. Rep.* **1**, 5 (2005).
58. Lieber, C. A. & Mahadevan-Jansen, A. Automated method for subtraction of fluorescence from biological Raman spectra. *Appl. Spectrosc.* **57**, 1363–1367 (2003).
59. Afseth, N. K. & Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **117**, 92–99 (2012).
60. Knorr, F., Smith, Z. J. & Wachsmann-Hogiu, S. Development of a time-gated system for Raman spectroscopy of biological samples. *Opt. Express* **18**, 20049–20058 (2010).
61. Praveen, B. B. et al. Fluorescence suppression using wavelength modulated Raman spectroscopy in fiber-probe-based tissue analysis. *J. Biomed. Opt.* **17**, 077006 (2012).
62. Engel, J. et al. Breaking with trends in pre-processing? *TrAC Trends Anal. Chem.* **50**, 96–106 (2013).
63. Gerretzen, J. et al. Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Anal. Chim. Acta* **938**, 44–52 (2016).
64. Guo, S., Bocklitz, T. & Popp, J. Optimization of Raman-spectrum baseline correction in biological application. *Analyst* **141**, 2396–2404 (2016).
65. Morishita, A., Imaging device and image processing program for estimating fixed pattern noise from partial noise output of available pixel area. Google Patents (2012).
66. Brown, C. D. & Wentzell, P. D. Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. *J. Chemom.* **13**, 133–152 (1999).
67. Theodoridis, S. and Koutroumbas, K. *Pattern Recognition* 4th edn (Academic Press, 2008).
68. Hastie, T. et al. The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **27**, 83–85 (2005).
69. Guo, S. et al. Common mistakes in cross-validating classification models. *Anal. Methods* **9**, 4410–4417 (2017).
70. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* Vol. 2, 1137–1145 (1995).



71. de Boves Harrington, P. Statistical validation of classification and calibration models using bootstrapped Latin partitions. *TrAC Trends Anal. Chem.* **25**, 1112–1124 (2006).
72. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
73. Liu, J. et al. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst* **142**, 4067–4074 (2017).
74. Hedegaard, M. et al. Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging. *Theor. Chem. Acc.* **130**, 1249–1260 (2011).
75. Nascimento, J. M. & Dias, J. M. Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **43**, 898–910 (2005).
76. Li, R. & Wang, X. Dimension reduction of process dynamic trends using independent component analysis. *Comput. Chem. Eng.* **26**, 467–473 (2002).
77. Zhang, Z., Chow, T. W. & Zhao, M. M-Isomap: orthogonal constrained marginal isomap for nonlinear dimensionality reduction. *IEEE Trans. Cybern.* **43**, 180–191 (2012).
78. de Silva, V. & Tenenbaum, J. B. Global versus local methods in nonlinear dimensionality reduction. in *Advances in Neural Information Processing Systems* (2003).
79. Shan, R., Cai, W. & Shao, X. Variable selection based on locally linear embedding mapping for near-infrared spectral analysis. *Chemom. Intell. Lab. Syst.* **131**, 31–36 (2014).
80. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
81. Wold, S. Pattern recognition by means of disjoint principal components models. *Pattern Recognit.* **8**, 127–139 (1976).
82. Barker, M. & Rayens, W. Partial least squares for discrimination. *J. Chemom.* **17**, 166–173 (2003).
83. Copas, J. B. Regression, prediction and shrinkage. *J. R. Stat. Soc. B Methodol.* **45**, 311–335 (1983).
84. Szymańska, E. et al. Chemometrics and qualitative analysis have a vibrant relationship. *TrAC Trends Anal. Chem.* **69**, 34–51 (2015).
85. Ballabio, D., Grisoni, F. & Todeschini, R. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.* **174**, 33–44 (2018).
86. Olivieri, A. C. Analytical figures of merit: from univariate to multiway calibration. *Chem. Rev.* **114**, 5358–5378 (2014).
87. Petersen, L., Minkinen, P. & Esbensen, K. H. Representative sampling for reliable data analysis: theory of sampling. *Chemom. Intell. Lab. Syst.* **77**, 261–277 (2005).
88. Esbensen, K. H. & Geladi, P. Principles of proper validation: use and abuse of re-sampling for validation. *J. Chemom.* **24**, 168–187 (2010).
89. Kalivas, J. H. et al. Calibration maintenance and transfer using Tikhonov regularization approaches. *Appl. Spectrosc.* **63**, 800–809 (2009).
90. Fernández Pierna, J. et al. Standardization of NIR microscopy spectra obtained from inter-laboratory studies by using a standardization cell. *Biotechnol. Agron. Soc. Environ.* **17**, 547–555 (2013).
91. Sjöblom, J. et al. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemom. Intell. Lab. Syst.* **44**, 229–244 (1998).
92. Wang, Y., Veltkamp, D. J. & Kowalski, B. R. Multivariate instrument standardization. *Anal. Chem.* **63**, 2750–2756 (1991).
93. Guo, S. et al. Model transfer for Raman-spectroscopy-based bacterial classification. *J. Raman Spectrosc.* **49**, 627–637 (2018).
94. Guo, S. et al. Extended multiplicative signal correction based model transfer for Raman spectroscopy in biological applications. *Anal. Chem.* **90**, 9787–9795 (2018).
95. Morais, C. L. et al. Standardization of complex biologically derived spectrochemical datasets. *Nat. Protoc.* **14**, 1546–1577 (2019).
96. Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936).
97. Neugebauer, U. et al. Towards detection and identification of circulating tumour cells using Raman spectroscopy. *Analyst* **135**, 3178–3182 (2010).
98. Stöckel, S. et al. Identification of *Bacillus anthracis* via Raman spectroscopy and chemometric approaches. *Anal. Chem.* **84**, 9873–9880 (2012).
99. Vogler, N. et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. *J. Biophotonics* **9**, 533–541 (2016).
100. Kumar, B. N. V. et al. Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing. *Anal. Chem.* **88**, 7574–7582 (2016).
101. Héberger, K. & Kollár-Hunek, K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *J. Chemom.* **25**, 151–158 (2011).

## Acknowledgements

The research in this contribution was supported by the Free State of Thuringia under the number 2019 FGR 0083 and cofinanced by European Union funds within the framework of the European Social Fund (ESF) via the TAB-FG MorphoTox. The authors highly acknowledge the financial support from the BMBF for the project LPI-BT1 (FKZ 13N15466) and the scholarship from China Scholarship Council (CSS) for SG. Part of the protocol relates to the NFDI4Chem project (441958208) funded by the German Research Foundation (DFG).

### Author contributions

T.B. conceived the project. S.G., T.B. and J.P. performed the conception and design of the protocol. T.B. and J.P. oversaw the overall planning of the project. J.P. supervised the experimental part, while T.B. supervised the computational part. S.G. performed the computations and the data analysis. S.G. and T.B. wrote the first draft of the protocol. All authors discussed the results and contributed to the manuscript review.

### Competing interests

The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41596-021-00620-3>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41596-021-00620-3>.

**Correspondence and requests for materials** should be addressed to Thomas Bocklitz.

**Peer review information** *Nature Protocols* thanks Luiz Fernando Cappa De Oliveira, Igor Lednev and Alejandro Olivieri for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 23 March 2021; Accepted: 19 August 2021;

Published online: 5 November 2021

### Related links

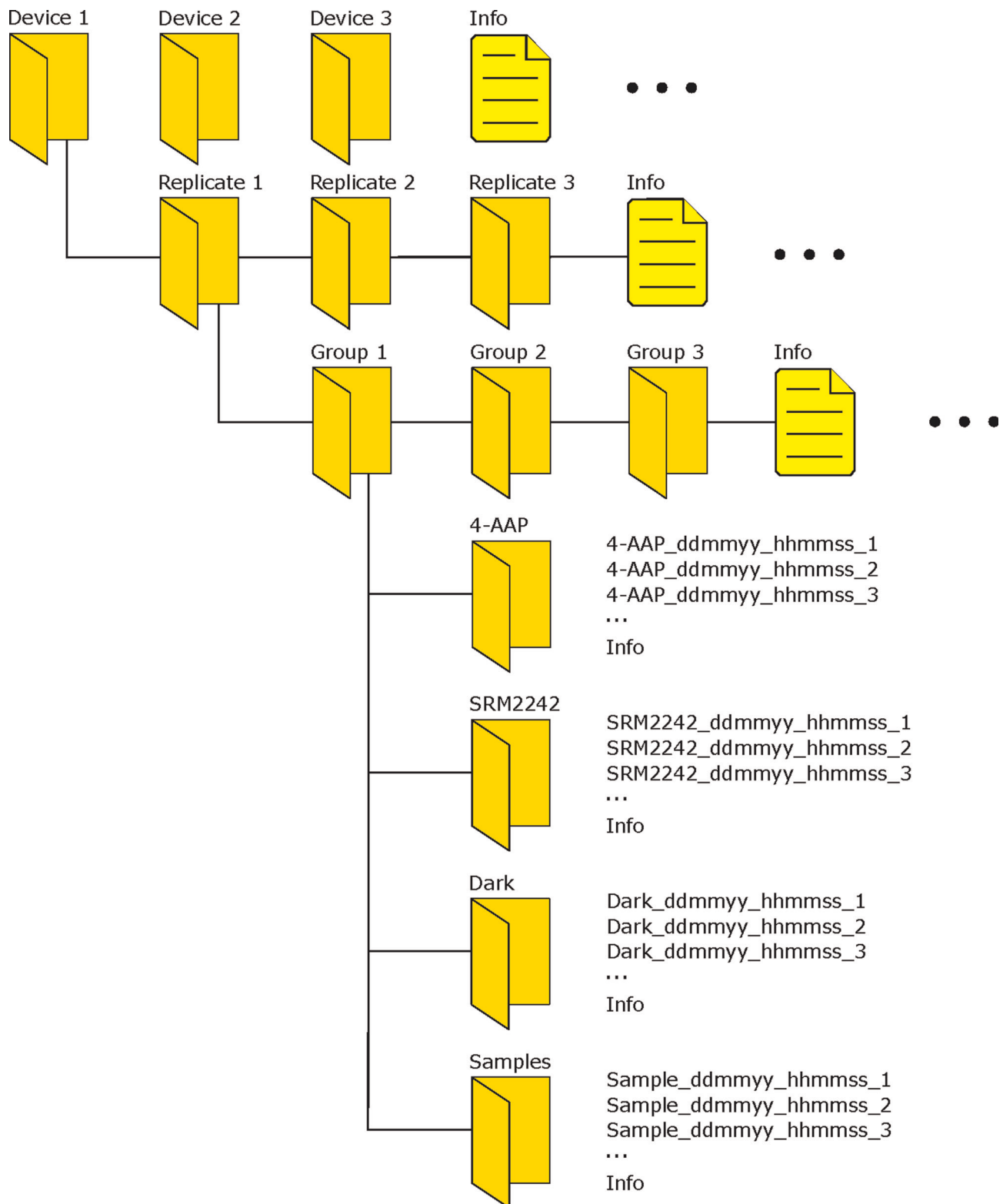
#### Key references using this protocol

Ali, N. et al. *Anal. Chem.* **90**, 12485–12492 (2018): <https://pubs.acs.org/doi/10.1021/acs.analchem.8b02167>  
Neugebauer, U. et al. *Analyst* **135**, 3178–3182 (2010): <https://pubs.rsc.org/en/content/articlelanding/2010/AN/c0an00608d>

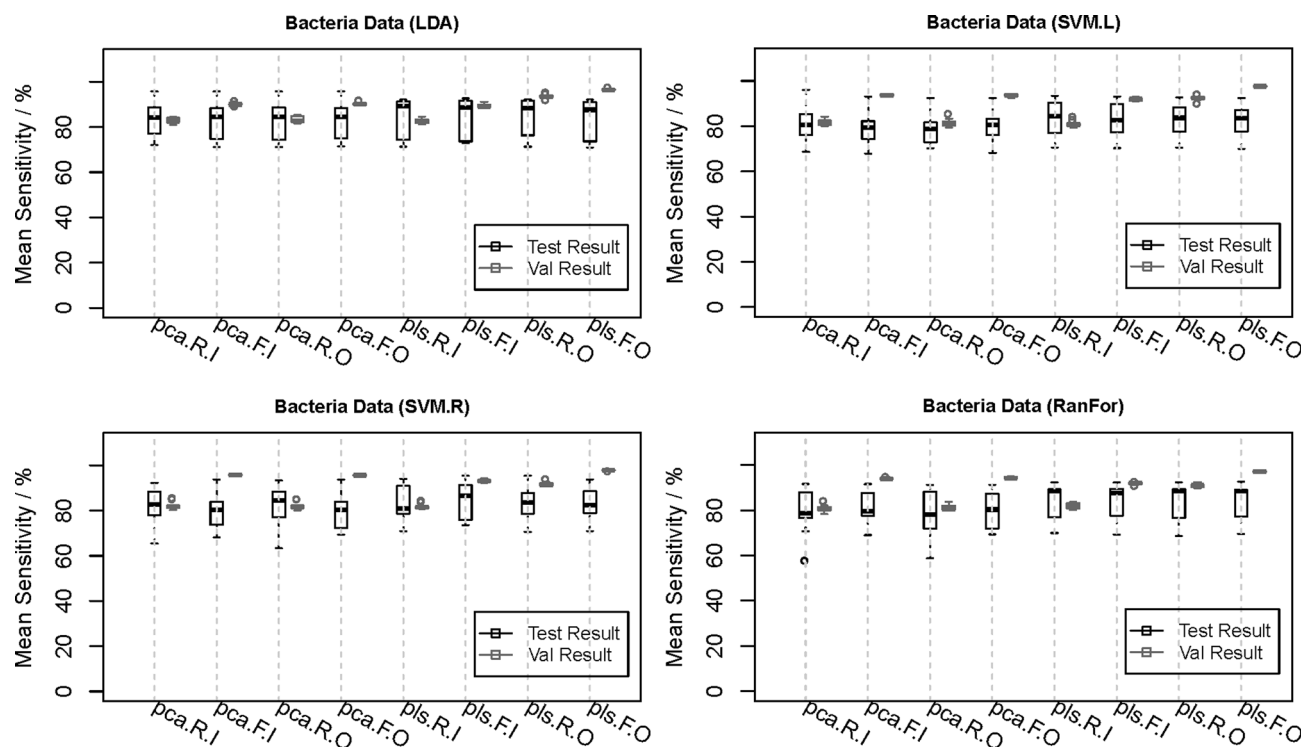
Stöckel, S. et al. *Anal. Chem.* **84**, 9873–9880 (2012): <https://pubs.acs.org/doi/abs/10.1021/ac302250t>

Vogler, N. et al. *J. Biophoton.* **9**, 533–541 (2016): <https://onlinelibrary.wiley.com/doi/10.1002/jbio.201500237>

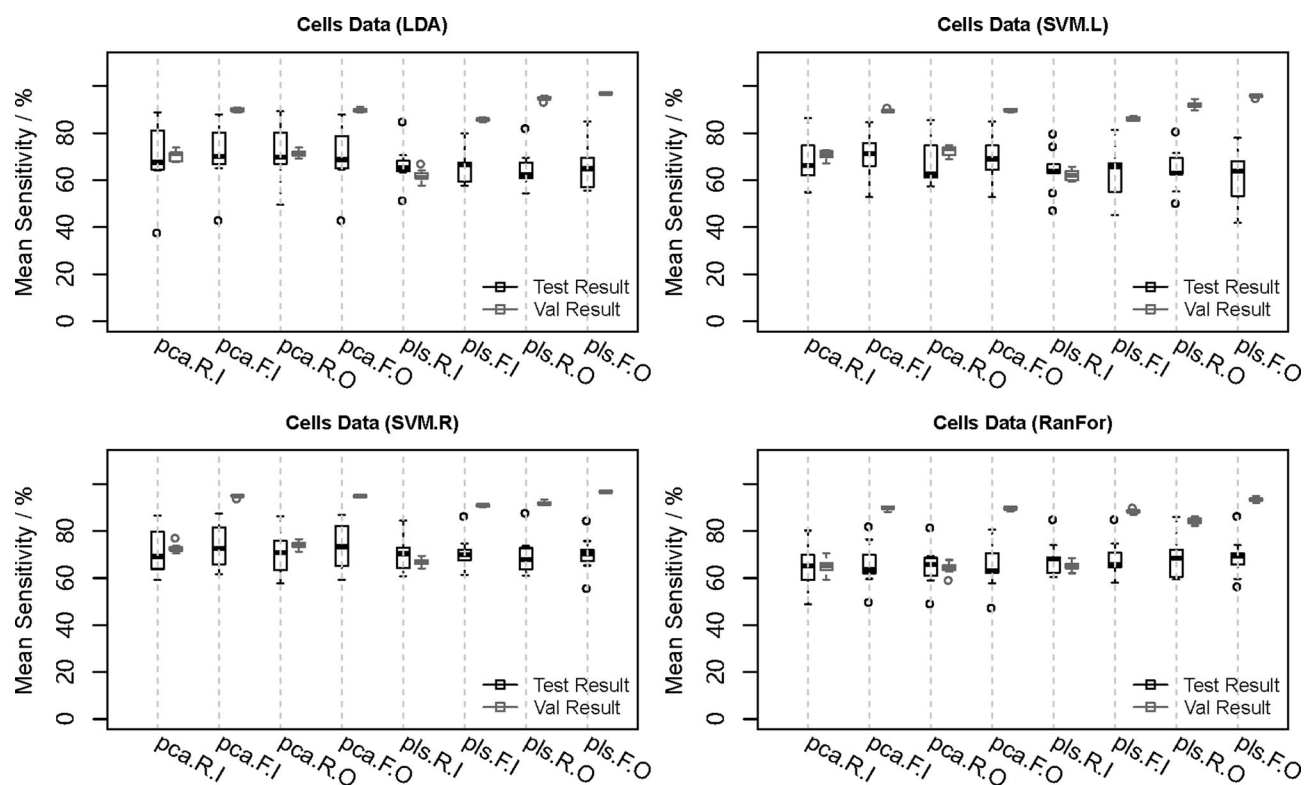
Butler, H. J. et al. *Nat. Prot.* **11**, 664–687 (2016): <https://doi.org/10.1038/nprot.2016.036>



**Extended Data Fig. 1 | An example of data structure.** The data is structured hierarchically following device-replicate-group. The calibration files are saved along with the sample spectra under the folder each group. The date and time information of the measurement is marked in file names in a format 'ddmmyy\_hhmmss'. The 'Info' files in each folder contain necessary records of the measurement.



**Extended Data Fig. 2 | Results of model validation and evaluation based on two dimensional reduction methods and different mechanisms of sampling for the bacterial dataset (Dataset 2).** The classification was performed using two dimension reduction methods and four classifiers in the framework of different sample sampling. Each box contains 9 values representing the mean sensitivity of the validation and testing results produced during the 9 iterations of the 9-fold/9-replicate external validation. The internal validation is considered unbiased if the testing and validation results are comparable, otherwise it is biased.



**Extended Data Fig. 3 | Results of model validation and evaluation based on two dimensional reduction methods and different mechanisms of sampling for the cell's dataset (Dataset 1).** Each box contains 9 values representing the mean sensitivity of the validation and testing results produced during the 9 iterations of the 9-fold/9-replicate external validation. The internal validation is considered unbiased if the testing and validation results are comparable, otherwise it is biased.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection The measurmenet software oif the Raman mcirciscopes was used

Data analysis Code was made avalaiabel: <https://github.com/Bocklitz-Lab/Example-Raman-spectral-analysis>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We provided one of teh datasets as open soure rpositorxy: <https://github.com/Bocklitz-Lab/Example-Raman-spectral-analysis>



## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Published dataset were used to present the protocol.
Data exclusions	We dont excluded data from the 4 published datasets.
Replication	No replications was peerformed.
Randomization	Within the publsied datasets single spectra of different specimen were utilized.
Blinding	The evaluation was done using cross valdiation appaches, which doesnt work with blinded samples/data.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging