Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Check for updates

Cite this: *Analyst*, 2025, **150**, 3198

Differentiation and identification of commensal and pathogenic oral bacteria at strain level using ATR-FTIR spectroscopy†

Katharina Anna Frings, *[a,b] Rumjhum Mukherjee, [b,c] Vivien Schulze,[a,b] Nils Heine,[b,c] Nicolas Debener,[b,d] Janina Bahnemann, [e,f] Szymon Piotr Szafrański, [b,c] Meike Stiesch,[b,c] Katharina Doll-Nikutta,*[b,c] Maria Leilani Torres-Mapa *[a,b] and Alexander Heisterkamp[a,b]

The correct identification of different bacteria is a critical task in clinical applications and basic research especially in the oral cavity which has a complex bacterial community. Complementary to a variety of phenotyping and genotyping methods, we propose FTIR spectroscopy as a fast and non-destructive technique for accurate bacterial identification. This technique can be used to investigate the chemical makeup of a given sample and also allows for bacterial classification at strain level. In this work, we investigate the ability of ATR-FTIR spectroscopy to identify different oral bacteria from known laboratory stains as well as strains from patient-derived samples. Using this technique, six measured species could be classified with high accuracy (>97%) using chemometric models. Furthermore, the model which was only trained with laboratory strains could still correctly identify the patient-derived strains at the genus level. These results open the possibility of constructing a simplified tailored classification model based only on a target species and few other representative species, while still being able to distinguish the target species from a much larger number of other bacterial species for application to oral microbial communities.

Received 13th February 2025,
Accepted 14th June 2025

DOI: 10.1039/d5an00165j

rsc.li/analyst

Introduction

The bacterial community found in the human oral cavity is extremely diverse, typically composed of more than 700 different species.[1] These bacteria tend to form multispecies biofilms on both soft and hard tissues within the oral cavity including the oral mucosa, teeth as well as periodontal pockets, potentially leading to severe dental diseases such as caries or periodontitis.[2,3] Inflammatory conditions have been shown to be induced by specific bacterial species eliciting the host's immune response.[4] In order to form biofilms, early colonisers, mainly *Streptococcus* species, are equipped to adhere to the salivary pellicle and other surfaces in the oral cavity. Adherence of other bacterial species to these early colonisers then promotes biofilm maturation.[5] Soon after initial adhesion, the bacteria embed themselves in a matrix of extracellular polysaccharides (EPS) which promotes stronger adhesion and shields them from anti-microbial agents, the immune system and bacteriophages.[6] Biofilm formation also facilitates simplified gene transfer and fosters antibiotic resistance.[7]

Due to the makeup of this biofilm, pathogenic oral multispecies biofilms are often detected at a later stage and early intervention remains challenging. Developing methods for early detection of pathogenic disease-forming biofilms can provide a means for targeted therapy, minimizing the loss of healthy tissue and/or implants. Certain bacterial species, such as *Porphyromonas gingivalis* are known to be prevalent for disease-related biofilms.[8–10] Hence, classification of the patient specific bacteria can be used to identify the presence of pathogens such as *Porphyromonas gingivalis* and allows for tailored therapeutic interventions, especially in the critical early stages of the disease.

There are currently several approaches for bacterial typing, ranging from phenotyping techniques like serotyping or antibiogram to genotyping methods including PCR and whole

[a]*Institute of Quantum Optics, Leibniz University Hannover, Hannover, Germany. E-mail: frings@iqo.uni-hannover.de, torres@iqo.uni-hannover.de*
[b]*Lower Saxony Center for Biomedical Engineering, Implant Research and Development (NIFE), Hannover, Germany*
[c]*Department of Prosthetic Dentistry and Biomedical Materials Science, Hannover Medical School, Hannover, Germany*
[d]*Institute of Technical Chemistry, Leibniz University Hannover, Hannover, Germany*
[e]*Institute of Physics, University of Augsburg, Augsburg, Germany*
[f]*Centre for Advanced Analytics and Predictive Sciences (CAAPS), University of Augsburg, Augsburg, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5an00165j

This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

genome sequencing (WGS) as well as indirect measures of genetic sequences like pulsed-field gel electrophoresis.[11] The latter has been considered the gold standard of bacterial typing and was used in many epidemiological studies. However, it requires specific protocols for each bacterial species and is difficult to standardize across different laboratory settings.[12,13] In recent years, WGS and similar techniques have become more important in bacterial typing due to reduced cost and operation time.[14] Although real-time sequencing can provide comprehensive genetic information and enable the analysis of mixed samples, the technique is labour-intensive and rarely used in clinical settings due to time constraints especially in critical cases.[14] But development in this direction is ongoing.[15] In the case of early detection of specific pathogens in dental biofilms, a fast and cost-effective method is needed, that can be used for routine examination and would allow for early and targeted intervention.

Vibrational spectroscopy can be a useful tool in bacterial classification. Both Fourier transform infrared spectroscopy (FTIR) as well as Raman spectroscopy have been successfully used to identify different bacteria at the strain-level.[16–19] Both methods are non-destructive and investigate the chemical makeup of a sample by detecting the excitation of vibrational modes within molecules. Since the selection criteria of the excited modes are technique-specific, FTIR and Raman can be viewed as complementary. Our work focuses on the use of FTIR spectroscopy, however similar approaches using Raman spectroscopy can be found elsewhere.[20,21]

FTIR spectroscopy relies on mid-IR light with wavelengths ranging between 2.5 μm and 25 μm (wavenumber: 4000 cm^{-1}–400 cm^{-1}). Vibrational modes of molecular bonds are excited by the infrared light with resonant wavelengths.[22] The resulting spectrum of a sample shows an overlay of certain peaks which can exhibit a very complex structure. The specific spectral pattern is unique to the chemical makeup of a bacteria and as such, it can be used to accurately and consistently identify and differentiate bacterial species and strains. Attenuated total internal reflection (ATR) spectroscopy is often employed for this purpose, since it requires little to no sample preparation. Previous studies have shown the usefulness of FTIR to identify pathogens of food born and clinical illnesses such as *Listeria monocytogenes*,[23] *Staphylococcus aureus*,[24] and *Salmonella enterica*.[25] FTIR has also been applied to detect antibiotic resistances in bacteria and distinguish between antibiotic resistant and susceptive strains.[26–28]

In combination with chemometric models, FTIR spectroscopy offers a non-destructive and fast alternative to common techniques. The combined information on the peak position, overall band-shape and the ratio of peak intensities enables the classification of different spectra. Although the increase in computational power over the last decades has popularized the use of complex machine and deep learning approaches, classical chemometric models and multivariate analysis are still primarily employed for bacterial classification yielding high accuracy. For accurate classification of unknown bacterial species, however, the system must first be trained on a spectrum of the species in question. This requires acquisition of thousands of spectra from hundreds of bacterial species and strains that especially present in the oral cavity. To our knowledge, a comprehensive and publicly available spectral data bank for bacterial species does not exist at the moment. Although the establishment of a data bank has been proposed, its implementation remains difficult to achieve in practice because differences in sample preparation and instrumentation can yield altered spectra. An alternative approach would be to have a reduced spectral library, including only a few species with representative strains and evaluate if related strains of the same genus can still be identified.

In this work, we cultivated and measured six different oral bacteria species that play an important role in biofilm development using ATR-FTIR spectroscopy. Chemometric models were trained to classify these spectra and then used to distinguish between the different species. In addition, 15 patient-derived strains of oral bacteria were cultivated and measured. These strains were also classified using the same chemometric models. Lastly, we investigated the ability of the chemometric models to identify the FTIR spectra of patient-derived bacterial species that have been previously unknown to the model based on the training data of six explementary species. This study opens future applications of FTIR spectroscopy and chemometric models for bacterial typing, as part of a workflow for biobanking and clinical diagnosis.

Experimental methods

Sample preparation

Six bacterial species were used, in the following referred to as laboratory strains. *Actinomyces naeslundii* (DSM 43013), *Veillonella dispar* (DSM 20735), *Porphyromonas gingivalis* (DSM 20709) and *Fusobacterium nucleatum* (DSM 15643) were purchased from the German Collection of Microorganisms and Cell Cultures (DSMZ; Braunschweig, Germany). *Streptococcus oralis* (ATCC 9811) and *Aggregatibacter actinomycetemcomitans* (MCCM 2474) were obtained from the American Type Culture Collection (ATCC; Manassas, VA, USA) and the Microbial Culture Collection Marburg (MCCM; Marburg, Germany), respectively. All species were kept as glycerol stocks at −80 °C prior to experiments. *S. oralis*, *A. naeslundii* and *V. dispar* were individually incubated in 10 ml of brain heart infusion medium (BHI; CM1135B, OXOID, Hampshire, United Kingdom) supplemented with 10 μg ml^{-1} Vitamin K for 24 hours at 37 °C in anaerobic conditions (air tight containment with Thermo Scientific™ Oxoid™ AnaeroGen™ 2.5 L bags (Fisher Scientific GmbH, Schwerte, Germany)). *A. actinomycetemcomitans* was incubated in 10 ml Todd-Hewitt broth medium (THB; CM0189, OXOID) supplemented with 10% yeast extract at 37 °C in a 5% CO_2 incubator for 48 hours. *F. nucleatum* and *P. gingivalis* were first streaked on fastidious anaerobe agar plates (NCM2020A; Neogen, Lansing, MI, USA) supplemented with 5% defibrinated sheep blood (SR0051X; OXOID) for 72 hours in anaerobic conditions at 37 °C prior to

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

liquid culture. The bacteria were collected from the agar plate and cultured in 10 ml of sterile fastidious anaerobe broth medium (FAB; LAB071, Neogen). *P. gingivalis* and *F. nucleatum* were incubated at 37 °C in anaerobic conditions for 48 and 24 hours, respectively. Three individual cultures from each bacteria species were prepared according to the protocol above. After their respective incubation times, all samples were centrifuged in an ultracentrifuge (Avanti JXN30; Beckman Coulter, Brea, CA, USA) at 4000*g* for 15 minutes at 4 °C. Afterwards, the supernatant was removed and the pellet was resuspended in 8 ml of ultrapure water. All samples were washed three times by centrifugation for 5 minutes at 2000*g* and resuspended in ultrapure water to ensure the removal of any excess medium prior to fixation. All samples were then fixed using 4% freshly prepared paraformaldehyde (PFA) solution. After the last washing step, the pellet was resuspended in 4% PFA instead of ultrapure water and left at room-temperature for 20 minutes. Afterwards, the samples were centrifuged again, the PFA was removed and the pellet was resuspended in ultrapure water. Once again, three washing steps, with a 15 minutes long pause in between, were performed to clear the sample of excess PFA. After washing and fixation, all samples were stored at 4 °C prior to measurement.

A total of 15 patient-derived strains were obtained from the CCUG and BiobankBIT culture collections, further information on the strains can be found in Table 1.[29] The bacterial strains used in this study were either previously published or obtained from established clinical culture collections (see Table S1†). No strains were isolated directly from patient samples by the authors; therefore, no additional ethical approval was required. All species were streaked on fastidious anaerobe agar, supplemented with 5% sheep blood (MSPS_029) [46 g L^{-1} fastidious anaerobe agar (FAA, LAB090/NCM2020A; LabM/Neogen), 5% sheep blood (SR0051E; Thermo Scientific)] from glycerol stocks, incubated for 72 hours or 96 hours and in case of *P. gingivalis* strains, cultured at 37 °C in an anaerobic chamber with 5% CO_2. All bacteria grown on plates were scraped and mixed in 1 ml of ultrapure water. The bacteria were washed and fixed as described above. For all samples, preparation was repeated three times, cultured on different days.

Table 1 Overview of patient-derived samples, the corresponding reference numbers from other culture collections can be found in the ESI S1†

SPS number	Genus	Species
475	*Streptococcus*	*Oralis* subsp. *oralis*
472	*Streptococcus*	*Mitis*
476	*Streptococcus*	*Oralis*
883	*Streptococcus*	*Oralis*
527	*Fusobacterium*	*Nucleatum* subsp. *polymorphum*
804	*Fusobacterium*	*Periodonticum*
805	*Fusobacterium*	*Nucleatum* subsp. *vincentii*
806	*Fusobacterium*	*Simiae*
807	*Fusobacterium*	*Nucleatum* subsp. *fusiforme*
808	*Fusobacterium*	*Nucleatum* subsp. *animalis*
452	*Porphyromonas*	*Gingivalis*
454	*Porphyromonas*	*Gingivalis*
455	*Porphyromonas*	*Gingivalis*
789	*Porphyromonas*	*Gingivalis*
791	*Porphyromonas*	*Gingivalis*

FTIR spectroscopy

All measurements were performed using a FTIR spectrometer (Spectrum Two, L160000E, PerkinElmer, Waltham, MA, USA) with a DTGS detector. The spectrometer was equipped with an UATR-Accessory (L1600129, PerkinElmer) including a single-bounce diamond ATR-crystal. The spectral acquisition was done using the Spectrum 10 software (PerkinElmer) within the wavenumber range between 400 cm^{-1}–4000 cm^{-1}, with a resolution of 4 cm^{-1} and 10 spectra recorded and averaged per measurement. Prior to measurements being taken, a background spectrum was acquired automatically when starting up the program. All samples were vortexed before measurement and 1 µl sample was pipetted onto the ATR crystal. The spectrum was acquired after 5 minutes or after complete evaporation of the residual water. After the measurements, the ATR crystal was cleaned using 70% ethanol before the next sample was placed onto the crystal. For the laboratory strains, 40 spectra per stock were acquired or a total of 120 spectra per species. For the patient-derived samples, 10 spectra per stock with a total of 30 spectra per species were recorded.

Data pre-processing and analysis

All spectra were pre-processed using OriginPro (Version 2022; OriginLab Corporation, Northampton, MA, USA) following the same protocol. After loading in OriginPro, all spectra were vector normalized and scaled to [0,1]. Principle component analysis (PCA) was performed for first evaluation of the data. Prior to PCA analysis, the data was mean centred using MATLAB R2021b (MathWorks, Inc., Portola Valley, CA, USA). For further analysis and classification, the classification toolbox for MATLAB was used.[30] For all classification, the data set was split into training and test set. In order to avoid bias in the classification, several techniques can be used to split the data and achieve nearly random distribution.[31] The MLM algorithm proposed by Santos *et al.* is freely available for MATLAB at https://doi.org/10.6084/m9.figshare.7393517.v1.[31] This algorithm was used to split the laboratory species as well as the patient species data into training and test set at a ratio of 80 : 20, respectively. This resulted in a laboratory species training set containing 572 spectra and a test set containing 144 spectra. The patient data set contains 359 spectra in the training set and 90 spectra in the test set. Column and row pre-processing can also potentially exert a strong impact on the classification results.

The classification toolbox allows for automatic column and row pre-processing of each data set before training the models. The ideal pre-processing was chosen, based on exploratory data analysis of each data set. For row pre-processing, standard normal variate, multiplicative scatter correction, first and second derivative were tested. For column pre-processing mean centering, auto-, variance- and range scaling were tested. For the laboratory species, the first derivative was chosen as

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

row pre-processing. Although the second derivative is known to better resolve overlapping peaks and is therefore widely used in secondary protein analysis,[32,33] the first derivative performed better in classification for the given data set. As column pre-processing, variance scaling was chosen and applied across all wavenumbers. For the patient-derived spectra, standard normal variate was chosen as row pre-processing and mean centering as column pre-processing. Only the spectral region between 800–1800 cm^{-1} was used for classification.

Results

To develop a fast and non-destructive method for accurate oral bacteria identification in clinical dentistry and basic research, the present study used ATR-FTIR spectroscopy to characterize six oral bacterial strains. The generated spectra were then used to train a chemometric model that allowed for the successful identification of patient-derived bacterial species.

FTIR spectra of six different oral bacteria

In order to use FTIR spectroscopy to assess a biofilm's state and development, the technique must be able to distinguish between the spectra of different species. In general, bacterial spectra can be divided into different regions, based on signals from different types of macromolecules. Fig. 1 shows an example spectrum with these different parts highlighted. The definition on these parts was first proposed by Naumann *et al.*[34,35] and has been widely adopted by many researchers in the field. The most relevant biological signals are between 3600–2800 cm^{-1} and 1800–400 cm^{-1}. The region in between is considered to be biologically silent because there are no significant peaks from biological components.[36] Although the signal of Amide I and II in region C (Fig. 1) are the most prominent in biological samples, they cannot be used to classify specific bacterial species, because they arise from all proteins which are present in the sample. For classification purposes, regions F and G are of particular interest, since they provide insights into specific molecules present in the sample. In region F, also called the polysaccharide region, the absorption bands of symmetric stretching of PO_2 groups and stretching vibrations of C–O–C and C–O–P can be seen, mostly resulting from carbohydrates.[34,37,38] Region G which is also referred to as the 'real' fingerprint region shows multiple weak and very specific peaks from nucleotides and amino acids.

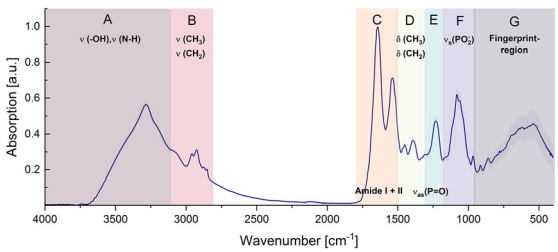

Fig. 1 Exemplary FTIR spectrum of bacteria with important functional regions assigned according to Naumann *et al.*[34,35] (A) 4000–3100 cm^{-1} –OH and N–H stretching modes, (B) 3100–2800 cm^{-1} C–H stretching vibrations of $-CH_3$ and $>CH_2$ corresponding to fatty-acid chains of membrane amphiphiles, (C) 1800–1500 cm^{-1} amide I and II arising from the amide groups of proteins, (D) 1500–1300 cm^{-1} bending modes of $>CH_2$ and $>CH_3$ from lipids and proteins[40] (E) 1230 cm^{-1} >P═O asymmetric stretching from phospholipids, (F) 1200–900 cm^{-1} PO_2^- symmetric stretching from nucleic acids, (G) 900–600 cm^{-1} fingerprint region.

The laboratory strains used in this study are representative members of certain groups of bacteria that are crucial in oral biofilm development. *S. oralis* and *A. naeslundii* were chosen as common early colonisers, with *V. dispar* and *F. nucleatum* acting as bridging bacteria and paving the way for attachment of pathogenic bacteria like *P. gingivalis* and *A. ac* as seen in mature biofilms.[39] Spectra of the six different laboratory strains are shown in Fig. 2A. The overall spectral band shape showed similarity across all species, due to the presence of common biological macromolecules as described before. However, slight differences could be observed among the band shapes. Fig. 2B shows the biologically most relevant region between 400–1800 cm^{-1}, that was also used for classification. Most prominent differences between the spectra could be observed at lower wavenumbers, especially in the region between 950–1200 cm^{-1} (Fig. 2C). This corresponds to the PO_2^- symmetric stretching from nucleic acids. Clear differences between the species could be observed in this region due to the differences in the bacterial genome GC content.

An important aspect of species phenotyping using FTIR spectroscopy is the practical need to minimize the variation in sample preparation. Some studies have shown that the growth medium, culture conditions and sample handling can change the spectra of the species in some cases.[40] This effect seems to be species dependent as other studies have, in contrast, found little to no effect.[41] In the present study, different bacterial species were cultured in different growth media, due to their auxotrophic nature towards specific nutrients. Extra washing steps were implemented to ensure complete removal of medium before measurement. Nevertheless, this did not seem to influence the overall capability of FTIR to differentiate between the species. A PCA of all spectra can be seen in Fig. 2D, which clearly showed clustering of the individual species. Fig. 2E shows the PCA of each species, each symbol representing spectra from a different sample run. It could be observed that for spectra of some species, such as *S. oralis* and *V. dispar*, samples from individual culture batches cluster together whereas spectra of the other species are more homogenously spread. This observation supports the hypothesis of species dependency, which implies that some species may be more susceptible to small variations in sample preparation.[42]

Multivariate classification and differentiation of bacterial species

In order to classify the data and demonstrate the ability of multivariate methods to correctly identify bacterial species by their FTIR spectrum, as a first step, linear discriminant ana-

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

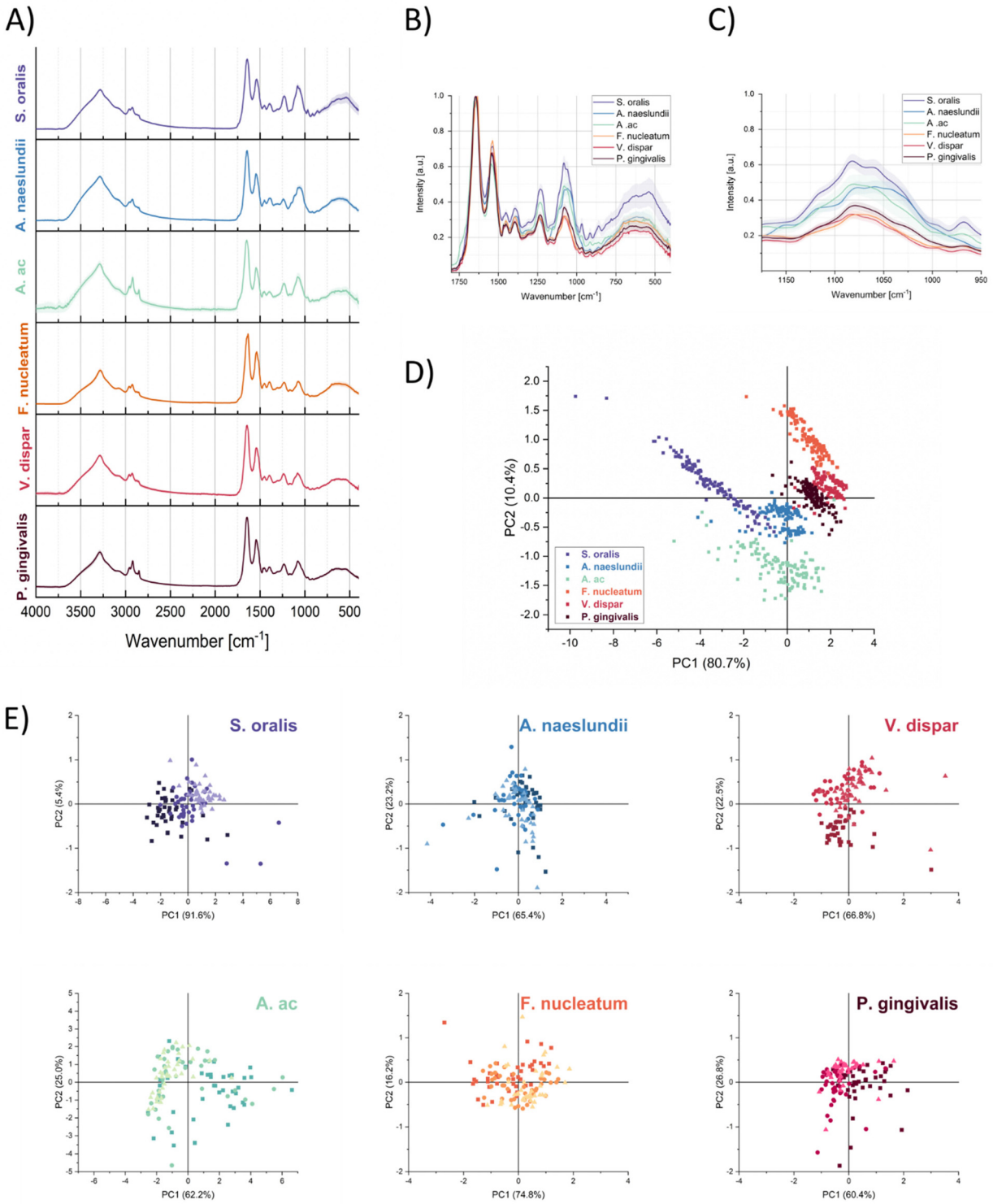

Fig. 2 FTIR spectra of six different bacterial species, (A) mean spectra with standard deviation shown, (B) zoomed in view between 1800–400 cm^{-1}, (C) zoomed in view between 1200–950 cm^{-1}, (D) PCA of all spectra, (E) PCA of individual species, markers show each sample run.

lysis was used. Linear discriminant analysis (LDA), is often combined with dimension reduction methods like PCA, since the number of samples needs to be larger than the number of variables. Fig. 3 shows the scores on first and second canonical variable of LDA performed on the first 10 principal components. The first 10 PCs explained 97% of the variance of the data. The six laboratory strains clustered accordingly and could be well distinguished from each other.

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

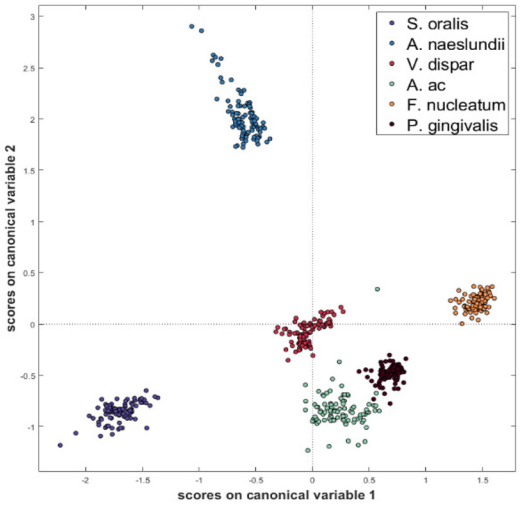

Fig. 3 Score plot on first two canonical scores of PCA-LDA of the six laboratory species data.

The results shown in Fig. 3 are based on the training data, and correspond to 80% of all spectra per species. In order to assess the classification ability of this technique, the remaining 20% of spectra were used as a test set which had not been seen by the network previously. Table 2 shows the prediction results from the test set on the model, previously trained only on the training set. As illustrated, the model was nevertheless able to accurately predict the correct class for all spectra within the test set.

The sensitivity, specificity and accuracy of the predictions made by the model for each species were calculated during cross-validation (CV), and the prediction of the test set (Test) can be found in the ESI S2.† PCA-LDA was found to be able to classify an unknown test set with 100% accuracy. These results strongly underscore the fact that FTIR spectroscopy is an excellent and highly specific method to distinguish and identify oral bacteria at species level.

FTIR measurement of patient-derived oral bacteria

In order to test the capability of FTIR within clinical applications, patient-derived samples of varying strains and subspecies were measured. Different strains of *F. nucleatum* and *P. gingivalis* were chosen, due to their clinical relevance in pathogenic biofilms within the human oral cavity. These strains are known to be an indicator of a shift in species distribution towards a more pathogenic situation.[43] In addition, *S. oralis* was chosen as a control since it is present in most biofilms and should not interfere with the identification of pathogens. Fig. 4 shows the mean spectra of the different patient-derived strains. In every graph, the corresponding laboratory strain is shown as a reference. For most species, the patient-derived strains showed similar band shapes with only minor differences. A clear exception is *P. gingivalis* sample SPS_791 which showed very different band shape from all other *Porphyromonas* species as well as all other species measured, even though Saenger sequencing of this patient-derived strain has confirmed its taxonomy as *P. gingivalis*. Thus, we identified a *P. gingivalis* strain with a unique chemical fingerprint of an unknown origin. To exclude contamination during cultivation, the spectra were collected from three independent cultures on MSPS_029. However, they still showed consistent spectral patterns with little to no differences. Hence, further microbiological investigations are needed to analyse potential differences in cell wall composition or metabolism of this particular strain in the future.

Table 2 Result of classification test set of laboratory strains. The abbreviations are as follows: *S. oralis* (SO), *A. naeslundii* (AN), *V. dispar* (VD), *A. actinomycetemcomitans* (Aac), *F. nucleatum* (FN) and *P. gingivalis* (PG)

Real/predicted	SO	AN	VD	Aac	FN	PG
S. oralis	24	0	0	0	0	0
A. naeslundii	0	24	0	0	0	0
V. dispar	0	0	24	0	0	0
A. ac	0	0	0	24	0	0
F. nucleatum	0	0	0	0	24	0
P. gingivalis	0	0	0	0	0	24

Multivariate classification of patient-derived oral bacteria

Similar to the laboratory strains, PCA-LDA was performed on the patient-derived species. The patient-derived species of each genus showed very similar band shapes. The score plot in Fig. 5 illustrates the clear clusters that were observed between the three genera, as well as between the individual strains. This pattern was further confirmed, when testing the trained model with the patient-derived spectra as test data. The corresponding accuracy, sensitivity and specificity was calculated and can be found in the ESI S3.† All samples could be classified with high accuracy. Since the test set only contained six spectra per strain, the analysis was performed multiple times with varying spectra in the training and test set with similar results. S4† presents the results of a representative classification run.

Multivariate model based on laboratory species for classification of patient-derived samples

Since the number of bacterial species in the oral cavity is vast, and incorporation of all known species in a PCA-LDA model would accordingly be extremely complex and time consuming, we additionally tested, if the model trained with only the laboratory data could nevertheless accurately identify the patient-derived strains. With this approach, it would be possible to create a reduced model and limit it to a representative strain of one species or even genus, which would still facilitate accurate classification of all similar strains of the respective species.

In addition to PCA-LDA, two additional multivariate classification models were tested for this purpose, namely partial least squares discriminant analysis (PLS-DA) and k-nearest neighbour classification (k-NN). These models have all been

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

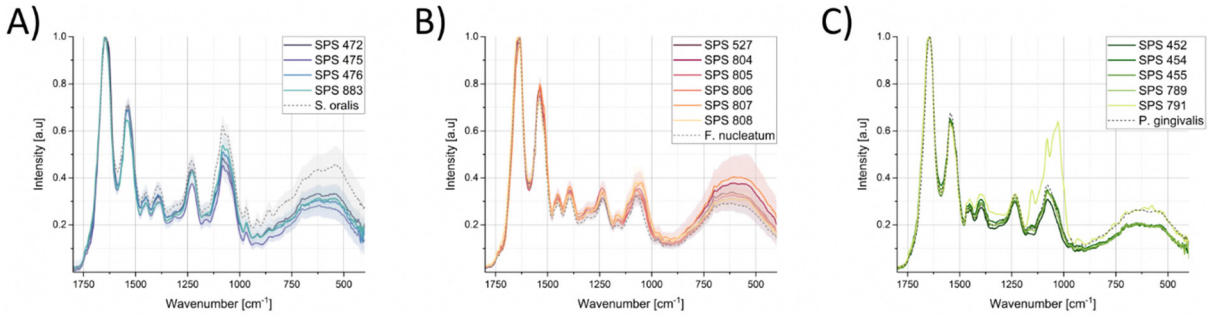

Fig. 4 FTIR spectra of the different patient-derived samples, (A) *Streptococcus* strains with *S. oralis* laboratory strain shown in dotted line, (B) *Fusobacterium* strains with *F. nucleatum* laboratory strain shown in dotted line, (C) *Porphyromonas* strains with *P. gingivalis* laboratory strain shown in dotted line. All spectra are mean spectra with standard deviation shown in shaded area.

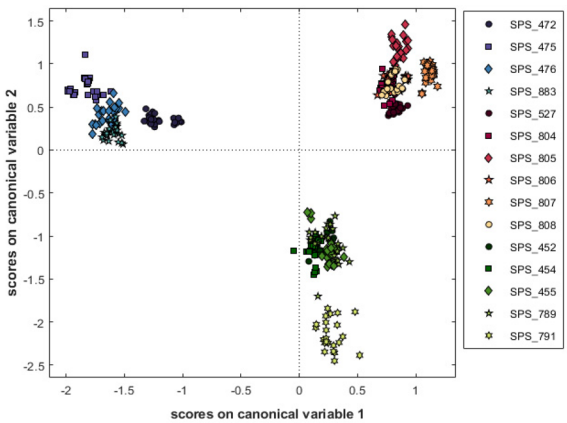

Fig. 5 PCA-LDA score plot of patient-derived species, LDA performed on the first 10 PCs of the patient species training set.

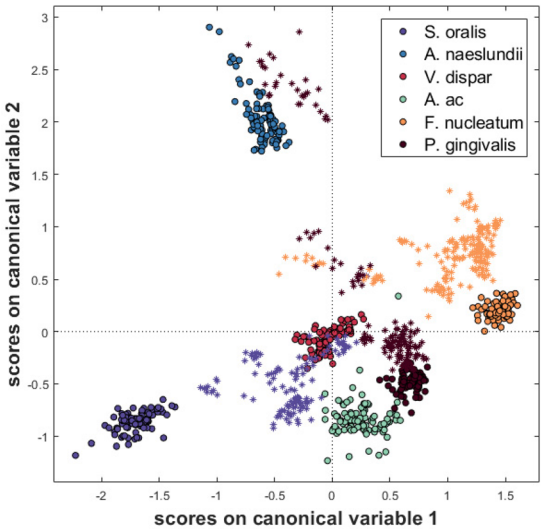

Fig. 6 Score plot of PCA-LDA of lab species as shown before with the predicted patient samples shown as (*) coloured according to the correct class.

selected because they have been shown to be able to classify similar data sets with high accuracy.[44,45]

All models were trained using the same laboratory species data set and initially tested with the same laboratory species testing set, with all three showing similar results to the PCA-LDA as shown previously. The results can be found in the ESI (S5 and S6).† In the next step, the performance of the trained models on predicting the unknown patient-derived species was tested. Fig. 6 shows the score plot of the previous PCA-LDA analysis including the scores of the predicted patient

Table 3 Results of the classification of patient-derived samples by different models only trained on laboratory species data

Real/predicted	SO	AN	VD	AAC	FN	PG	NA	Accuracy	Sensitivity	Specificity
PCA-LDA PC = 10										
Streptococcus	15	0	96	9	0	0	0	0.708	0.125	0.825
Fusobacterium	0	5	16	0	158	1	0	0.959	0.878	0.976
Porphyromonas	0	35	0	1	0	113	0	0.919	0.758	0.952
PLS-DA LV = 9										
Streptococcus	22	0	0	1	0	0	97	0.767	0.183	0.864
Fusobacterium	0	3	0	0	128	0	49	0.917	0.711	0.952
Porphyromonas	0	13	0	19	0	9	108	0.732	0.06	0.843
k-NN k = 8										
Streptococcus	95	0	0	11	8	6	0	0.931	0.792	0.958
Fusobacterium	0	7	8	0	165	0	0	0.972	0.917	0.983
Porphyromonas	0	35	0	0	0	114	0	0.922	0.765	0.953

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

data. *Fusobacterium* and *Porphyromonas* species were found to cluster closely to the corresponding laboratory species. In contrast, the *Streptococcus* data was comparatively widespread. This behaviour can also be seen in Table 3, where the classification results of all models are displayed.

For the patient-derived *Fusobacterium* species, classification showed accurate results (>92% accuracy) across all models, with k-NN showing the highest accuracy as well as specificity and sensitivity (>97%).

By contrast, the accuracy for predicting the *Porphyromonas* species was found to be highly dependent on the model used. The results for PCA-LDA and k-NN were very similar, whereas PLS-DA was unable to classify *Porphyromonas*, and instead categorized it as 'unknown'. For *Streptococcus*, only k-NN performed well, whereas PCA-DA and PLS-DA misclassified it as *V. dispar* or did not assign any class at all, respectively. Overall, k-NN showed the highest classification accuracy for all patient-derived samples and confirmed that it is in general possible to correctly classify patient-derived species based on FITR spectra when using a model solely trained with laboratory strains.

Discussion

We have successfully demonstrated that FTIR spectroscopy has the ability to differentiate closely related oral bacteria strains, in agreement with the literature on this topic.[40] However, in order to fully capitalize upon the potential of FTIR spectroscopy as a diagnostic tool for rapid oral bacteria species identification, a spectral database containing the bacterial species of interest would also be needed. Several studies have attempted to build spectral databases for identification of specific bacteria.[46–48] All of these databases contain spectra of a few hundred different species, that could all be correctly identified. This approach could also be implemented within the context of the bacterial communities commonly found in the human oral cavity; however, the spectra database would need to be extended to contain oral bacteria. Since the oral bacterial community can be very diverse, we demonstrated that it would be possible to identify bacterial strains that are not part of the spectral data bank as the right genus. This opens up the possibility to use the data bank for genera identification while still increasing the number of strains included.

Our results showed that patient-derived strains could be correctly classified based on FTIR spectra using a k-NN model, without losing the ability to also differentiate between strains. It should also be noted, that k-NN is a straightforward classification technique directly comparing the similarity between samples. This simplified approach could potentially lead to misclassification especially when measured spectra have only minimal differences.[49] Nevertheless, with the bacterial strains measured in this study, k-NN could correctly classify the individual strains with high accuracy – indicating that this technique is suitable for the task at hand.

We readily acknowledge that the reduced model shown here only contained a small number of bacterial species. Although different *Streptococcus* species have already been included as a reference to see if it interferes with the classification, it remains to be investigated, how the performance of the model changes when other species are measured and more closely related species are incorporated in the model. When increasing the number of samples, it should also be kept in mind, that the k-NN classification technique is a memory-based classification model which can become computationally costly when dealing with large datasets.[49] Several approaches can be used to speed up the classification process such as condensed nearest neighbour (CNN) or the KD-tree method.[50] However, when increasing the complexity of the data set, the exploration of more complex algorithms like neural networks might become necessary.

Conclusion

This work demonstrated the ability of FTIR spectroscopy to distinguish between different oral bacteria species that play key roles in the development of (pathogenic) oral multispecies biofilms. In addition to its classification ability at the strain-level, FTIR spectroscopy was also used to correctly classify the genus when the model was trained only using laboratory species spectra. When patient-derived samples belonging to the genus *Streptococcus*, *Fusobacterium* and *Porphyromonas* were additionally cultivated and measured, these different strains could be correctly classified with >97% accuracy. Furthermore, we showed that 15 different patient-derived strains could be identified correctly at genus-level with a spectral library containing spectra of a single representative species per genus. Our results highlight the prospect of constructing a spectral library containing only a few species while still enabling identification of a larger number of different strains.

Author contributions

K. F.: writing – original draft, data curation, formal analysis, investigation, methodology, validation, visualization, writing – review and editing R. M.: writing – review and editing, resources, validation V. S.: writing – review and editing, investigation, formal analysis, validation N. H.: writing – review and editing, resources N. D.: writing – review and editing J. B.: writing – review and editing, conceptualization, funding acquisition, project administration S. P. S.: writing – review and editing, resources M. S.: writing – review and editing, resources K. D.-N.: writing – review and editing, conceptualization, funding acquisition, project administration, resources M. L. T.-M.: writing – original draft, conceptualization, funding acquisition, project administration, resources, supervision, writing – review and editing A. H.: writing – review and editing, conceptualization, funding acquisition, project administration, resources.

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including the raw spectral data and the results of the Matlab classification are available at zenodo at **https://doi.org/10.5281/zenodo.14856442**.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB/TRR-298-SIIRI – Project-ID 426335750.

References

1 B. J. Paster, S. K. Boches, J. L. Galvin, R. E. Ericson, C. N. Lau, V. A. Levanos, *et al.*, Bacterial diversity in human subgingival plaque, *J. Bacteriol.*, 2001, **183**(12), 3770–3783.
2 J. A. Aas, B. J. Paster, L. N. Stokes, I. Olsen and F. E. Dewhirst, Defining the normal bacterial flora of the oral cavity, *J. Clin. Microbiol.*, 2005, **43**(11), 5721–5732.
3 S. P. Szafranski, M. L. Wos-Oxley, R. Vilchez-Vargas, R. Jáuregui, I. Plumeier, F. Klawonn, *et al.*, High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis, *Appl. Environ. Microbiol.*, 2015, **81**(3), 1047–1058.
4 M. Kilian, I. L. C. Chapple, M. Hannig, P. D. Marsh, V. Meuric, A. M. L. Pedersen, *et al.*, The oral microbiome - an update for oral healthcare professionals, *Br. Dent. J.*, 2016, **221**(10), 657–666.
5 N. S. Jakubovics and P. E. Kolenbrander, The road to ruin: the formation of disease-associated oral biofilms, *Oral Dis.*, 2010, **16**(8), 729–739.
6 B. Vu, M. Chen, R. J. Crawford and E. P. Ivanova, Bacterial extracellular polysaccharides involved in biofilm formation, *Molecules*, 2009, **14**(7), 2535–2554.
7 B. Y. Wang, B. Chi and H. K. Kuramitsu, Genetic exchange between Treponema denticola and Streptococcus gordonii in biofilms, *Oral Microbiol. Immunol.*, 2002, **17**(2), 108–112.
8 J. Mysak, S. Podzimek, P. Sommerova, Y. Lyuya-Mi, J. Bartova, T. Janatova, *et al.*, Porphyromonas gingivalis: major periodontopathic pathogen overview, *J. Immunol. Res.*, 2014, **2014**, 476068.
9 A. L. Griffen, M. R. Becker, S. R. Lyons, M. L. Moeschberger and E. J. Leys, Prevalence of Porphyromonas gingivalis and periodontal health status, *J. Clin. Microbiol.*, 1998, **36**(11), 3239–3242.
10 C. Ébs, M. Romandini, S. Sadilina, A. C. P. Sant'Ana and M. Sanz, Microbiota associated with peri-implantitis-A systematic review with meta-analyses, *Clin. Oral Implants Res.*, 2023, **34**(11), 1176–1187.
11 N. Almasian-Tehrani, M. Alebouyeh, S. Armin, N. Soleimani, L. Azimi and R. Shaker-Darabad, Overview of typing techniques as molecular epidemiology tools for bacterial characterization, *Cell. Mol. Biomed. Rep.*, 2021, **1**(2), 69–77.
12 H.-M. Neoh, X.-E. Tan, H. F. Sapri and T. L. Tan, Pulsed-field gel electrophoresis (PFGE): A review of the "gold standard" for bacteria typing and current alternatives, *Infect., Genet. Evol.*, 2019, **74**, 103935.
13 S. J. Salipante, D. J. SenGupta, L. A. Cummings, T. A. Land, D. R. Hoogestraat and B. T. Cookson, Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology, *J. Clin. Microbiol.*, 2015, **53**(4), 1072–1079.
14 S. R. Simar, B. M. Hanson and C. A. Arias, Techniques in bacterial strain typing: past, present, and future, *Curr. Opin. Infect. Dis.*, 2021, **34**(4), 339–345.
15 F. Tagini and G. Greub, Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review, *Eur. J. Clin. Microbiol. Infect. Dis.*, 2017, **36**(11), 2007–2020.
16 C. Quintelas, E. C. Ferreira, J. A. Lopes and C. Sousa, An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing, *Biotechnol. J.*, 2018, **13**(1), 1700449.
17 D. F. M. Willemse-Erix, M. J. Scholtes-Timmerman, J.-W. Jachtenberg, W. B. van Leeuwen, D. Horst-Kreft, T. C. Bakker Schut, *et al.*, Optical fingerprinting in bacterial epidemiology: Raman spectroscopy as a real-time typing method, *J. Clin. Microbiol.*, 2009, **47**(3), 652–659.
18 L. Mariey, J. P. Signolle, C. Amiel and J. Travert, Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics, *Vib. Spectrosc.*, 2001, **26**(2), 151–159.
19 D. Martak, B. Valot, M. Sauget, P. Cholley, M. Thouverez, X. Bertrand, *et al.*, Fourier-Transform InfraRed Spectroscopy Can Quickly Type Gram-negative Bacilli Responsible for Hospital Outbreaks, *Front. Microbiol.*, 2019, **10**, 1440.
20 L. S. Kriem, K. Wright, R. A. Ccahuana-Vasquez and S. Rupp, Confocal Raman microscopy to identify bacteria in oral subgingival biofilm models, *PLoS One*, 2020, **15**(5), e0232912.
21 J. Zhang, Y. Liu, H. Li, S. Cao, X. Li, H. Yin, *et al.*, Discrimination of periodontal pathogens using Raman spectroscopy combined with machine learning algorithms, *J. Innovative Opt. Health Sci.*, 2022, **15**(03), 2240001.
22 P. R. Griffiths and J. A. de Haseth, Chemical analysis, in *Fourier transform infrared spectrometry*, Wiley-Interscience, Hoboken, NJ, 2nd edn, 2007, vol. 171. Available from: URL: **https://www.loc.gov/catdir/enhancements/fy0653/2006022115-d.html**.
23 E. B. Nyarko, K. A. Puzey and C. W. Donnelly, Rapid differentiation of Listeria monocytogenes epidemic clones III and IV and their intact compared with heat-killed populations using Fourier transform infrared spectroscopy and chemometrics, *J. Food Sci.*, 2014, **79**(6), M1189–M1196.
24 M. A. Al-Holy, M. Lin, H. Al-Qadiri, A. G. Cavinato and B. A. Rasco, Classification of foodborne pathogens by Fourier transform infrared spectroscopy and pattern reco-

 This journal is © The Royal Society of Chemistry 2025

Open Access Article. Published on 17 June 2025. Downloaded on 1/22/2026 10:50:50 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

gnition techniques, *Rapid Methods Autom. Microbiol.*, 2006, **14**(2), 189–200.

25 O. E. Preisner, J. C. Menezes, R. Guiomar, J. Machado and J. A. Lopes, Discrimination of Salmonella enterica serotypes by Fourier transform infrared spectroscopy, *Food Res. Int.*, 2012, **45**(2), 1058–1064.

26 Y. G. Marangoni-Ghoreyshi, T. Franca, J. Esteves, A. Maranni, K. D. Pereira Portes, C. Cena, *et al.*, Multiresistant diarrheagenic Escherichia coli identified by FTIR and machine learning: a feasible strategy to improve the group classification, *RSC Adv.*, 2023, **13**(36), 24909–24917.

27 K. Kochan, C. Nethercott, J. Taghavimoghaddam, Z. Richardson, L. Lai, S. A. Crawford, *et al.*, Rapid Approach for Detection of Antibiotic Resistance in Bacteria Using Vibrational Spectroscopy, *Anal. Chem.*, 2020, **92**(12), 8235–8243.

28 A. Salman, U. Sharaha, E. Rodriguez-Diaz, E. Shufan, K. Riesenberg, I. J. Bigio, *et al.*, Detection of antibiotic resistant Escherichia Coli bacteria using infrared microscopy and advanced multivariate analysis, *Analyst*, 2017, **142**(12), 2136–2144.

29 N. S. Stumpp, J. Eberhard, N. C. Gellrich, W. Geurtsen, H. Windhagen, A. Haverich, *et al.*, Die Biobank für Biofilme, Implantate und assoziirte Gewebe (BIT), *Dtsch. Zahnaerztl. Z.*, 2012, **67**(4), 260–264.

30 D. Ballabio and V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, *Anal. Methods*, 2013, **5**(16), 3790.

31 C. L. M. Morais, M. C. D. Santos, K. M. G. Lima and F. L. Martin, Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach, *Bioinformatics*, 2019, **35**(24), 5257–5263.

32 J. Wu, M.-J. Wang and J. P. Stark, Evaluation of band structure and concentration of ionic liquid in molecular mixtures by using second derivatives of FTIR spectra, *J. Quant. Spectrosc. Radiat. Transfer*, 2006, **102**(2), 228–235.

33 D. Usoltsev, V. Sitnikova, A. Kajava and M. Uspenskaya, Systematic FTIR Spectroscopy Study of the Secondary Structure Changes in Human Serum Albumin under Various Denaturation Conditions, *Biomolecules*, 2019, **9**(8), 359.

34 D. Helm, H. Labischinski, G. Schallehn and D. Naumann, Classification and identification of bacteria by Fourier-transform infrared spectroscopy, *J. Gen. Microbiol.*, 1991, **137**(1), 69–79.

35 D. Naumann, V. Fijala, H. Labischinski and P. Giesbrecht, The rapid differentiation and identification of pathogenic bacteria using Fourier transform infrared spectroscopic and multivariate statistical analysis, *J. Mol. Struct.*, 1988, **174**, 165–170.

36 S. Mittal and R. Bhargava, A comparison of mid-infrared spectral regions on accuracy of tissue classification, *Analyst*, 2019, **144**(8), 2635–2642.

37 P. Lasch and D. Naumann, *Infrared Spectroscopy in Microbiology, in Encyclopedia of Analytical Chemistry*, 2007, pp. 1–32.

38 C. Yu and J. Irudayaraj, Spectroscopic characterization of microorganisms by Fourier transform infrared microspectroscopy, *Biopolymers*, 2005, **77**(6), 368–377.

39 D. Verma, P. K. Garg and A. K. Dubey, Insights into the human oral microbiome, *Arch. Microbiol.*, 2018, **200**(4), 525–540.

40 H. C. van der Mei, D. Naumann and H. J. Busscher, Grouping of streptococcus mitis strains grown on different growth media by FT-IR, *Infrared Phys. Technol.*, 1996, **37**(4), 561–564.

41 D. Lefier, D. Hirst, C. Holt and A. G. Williams, Effect of sampling procedure and strain variation in Listeria monocytogenes on the discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variates analysis, *FEMS Microbiol. Lett.*, 1997, **147**(1), 45–50.

42 A. Oust, T. Møretrø, C. Kirschner, J. A. Narvhus and A. Kohler, Evaluation of the robustness of FT-IR spectra of lactobacilli towards changes in the bacterial growth conditions, *FEMS Microbiol. Lett.*, 2004, **239**(1), 111–116.

43 V. Murugaiyan, S. Utreja, K. M. Hovey, Y. Sun, M. J. LaMonte, J. Wactawski-Wende, *et al.*, Defining Porphyromonas gingivalis strains associated with periodontal disease, *Sci. Rep.*, 2024, **14**(1), 6222.

44 N. M. Amiali, G. R. Golding, J. Sedman, A. E. Simor and A. A. Ismail, Rapid identification of community-associated methicillin-resistant Staphylococcus aureus by Fourier transform infrared spectroscopy, *Diagn. Microbiol. Infect. Dis.*, 2011, **70**(2), 157–166.

45 B. Feng, H. Shen, F. Yang, J. Yan, S. Yang, N. Gan, *et al.*, Efficient classification of Escherichia coli and Shigella using FT-IR spectroscopy and multivariate analysis, *Spectrochim. Acta, Part A*, 2022, **279**, 121369.

46 M. Kümmerle, S. Scherer and H. Seiler, Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy, *Appl. Environ. Microbiol.*, 1998, **64**(6), 2207–2214.

47 H. Oberreuter, H. Seiler and S. Scherer, Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy, *Int. J. Syst. Evol. Microbiol.*, 2002, **52**(Pt 1), 91–100.

48 M. Wenning, N. R. Büchl and S. Scherer, Species and strain identification of lactic acid bacteria using FTIR spectroscopy and artificial neural networks, *J. Biophotonics*, 2010, **3**(8–9), 493–505.

49 P. Cunningham and S. J. Delany, k-Nearest Neighbour Classifiers - A Tutorial, *ACM Comput. Surv.*, 2022, **54**(6), 1–25.

50 A. Mucherino, P. J. Papajorgji and P. M. Pardalos, k-Nearest Neighbor Classification, in *Data Mining in Agriculture*, ed. A. Mucherino, P. J. Papajorgji and P. M. Pardalos, Springer New York, New York, NY, 2009, pp. 83–106. (Springer Optimization and Its Applications).