Materials and Structures (2026) 59:292
https://doi.org/10.1617/s11527-026-03170-w

RILEM TC REPORT

Investigating the differentiation of bitumen samples based on FTIR spectra using multivariate analysis (PCA-LDA) in the RILEM 295-FBB TG1 round robin test

Jan Unterbuchschachner · Sandra Weigel · Mohsen Motevalizadeh · Sadaf Khalighi · Johannes Mirwald · Aikaterini Varveri · Bernhard Hofko

Received: 10 December 2025 / Revised: 22 May 2026 / Accepted: 13 June 2026

Abstract Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy is a widely applicated method in bituminous binder analysis. However, differences in equipment, measurement routines, sample preparation, device and spectral evaluation make comparability of results between laboratories difficult. To demonstrate these difficulties, the task group 1 of the RILEM TC 295-FBB "Fingerprinting bituminous binders using physico-chemical analysis" has conducted an international round robin test. A total of 6,786 spectra were obtained from 21 laboratories. The laboratories used three different unmodified 70/100 penetration grade binders at three aging states and prepared them for measurement with the ATR-FTIR spectroscope using six different methods. Building on this dataset the present study applies multivariate analysis (MVA) to further investigate spectral variability and assess its potential for robust classification. Using min–max normalization and z-score filtering, the data was prepared for evaluation with Coefficient of Variation, Intraclass Correlation Coefficient and Principal Component-Linear Discriminant Analysis. Homogenized preparation methods achieved the highest

This manuscript was prepared by working group 1 within RILEM TC 295-FBB "Fingerprinting bituminous binders using physicochemical analysis". And further reviewed and approved by all members of the RILEM TC 295-FBB.

TC Membership: RILEM TC 295-FBB.

TC Chair: Bernhard Hofko.

TC Deputy Chair: Aikaterini Varveri.

Members: Dheeraj Adwani, Panos Apostolidis, Johan Blom, Johannes Büchner, Augusto Cannone-Falchetto, Xavier Carbonneau, Alan Carter, Emmanuel Chailleux, Davide Dalmazzo, Eshan Dave, Herve DiBenedetto, Michael Elwardany, Leidy Espinosa Ruiz, Yangming Gao, Victor Garcia Rabadan, Flavien Geisler, Andrea Graziani, Meng Guo, Ankit Gupta, Hmazeh Haghshenas, Bernhard Hofko, Patricija Kara De Maeijer, Sadaf Khalighi, Rita Kleiziene, Katarzyna Konieczna, Jan Król, Peng Lin, Yi Li, Bert Jan Lommerts, Maciej Maliszewski, Salvatore Mangiafico, David Mensching, Miomir Miljković, Johannes Mirwald, Lucas Mortier, Mohsen Motevalizadeh, Virginie Mouillet, Sayeda Nahar, Marko Oreskovic, Jorge Pais, Manfred Partl, Emiliano Pasquini, Nathalie Piérard, Georgios Pipintakos, Kristina Primerano, Jian Qiu, Ali Raman, Shisong Ren, Margarida Sa-Da-Costa, Cesare Sangiorgi, Aditi Sharma, Jo Sias, Hilde Soenen, Anand Sreeram, Mayank Sukhija, Yuxuan Sun, Pejoohan Tavassoti, Yannick Tchedele Langollo, Marjan Tusar, Jan Unterbuchschachner, Stefan Vansteenkiste, Aiaterini Varveri, Kamila Vasconcelos, Thibault Villette, Di Wang, Sandra Weigel, Stefan Werkovits, Jens Wetekam, Michael Wistuba, Yuan Zhang, Fan Zhang, Jiqing Zhu.

J. Unterbuchschachner (✉) · J. Mirwald · B. Hofko
Christian Doppler Laboratory for Chemo-Mechanical Analysis of Bituminous Materials, Institute of Transportation, TU Wien, Karlsplatz 13/E230-3, 1040 Vienna, Austria
e-mail: jan.unterbuchschachner@tuwien.ac.at

Published online: 03 July 2026

reproducibility and discrimination accuracy, while solvent based and non-homogenized methods introduced greater variability. Differences between ATR-FTIR devices and manufacturers were also evident, mainly through spectral intensity shifts. Although the differences between the binders and aging states are small, an accuracy of over 0.95 demonstrates that MVA methods can clearly distinguish between them. A feature importance analysis using a random forest highlighted the carbonyl, sulfoxide, and aliphatic aromatic bands as the main factors contributing to this discriminability. This study demonstrated that ATR-FTIR spectroscopy in combination with MVA can support the development of standardized protocols for bitumen analysis.

Keywords Bitumen · FTIR Spectroscopy · Multivariate analysis · Round Robin test · Material handling · Preparation routine

1 Introduction

Bitumen is one of the most important construction materials in road engineering, yet at the same time it is also one of the most chemically complex. It is not a single compound but rather a mixture of millions of molecules, ranging from lighter fractions with low molecular weight to heavy, polyaromatic asphaltenes. These molecules are usually described in terms of saturates, aromatics, resins, and asphaltenes, the so-called SARA fractions [1]. Together, they form a dynamic colloidal system whose structure and properties depend strongly on the crude oil origin, the refinery procedure, and any subsequent modification such as polymer addition, rejuvenation, or blending with other binders [2]. What makes bitumen especially challenging to study is that its composition is not constant over time. It continuously evolves during production, storage, and especially during service life on the pavement. Processes such as oxidation, and evaporation of lighter molecules alter the chemical balance of the material [3, 4]. The consequences of these changes are significant: the binder becomes stiffer and more brittle, its ability to relax stresses diminishes, and pavement deterioration accelerates [5]. For this reason, methods that can monitor the chemical composition and its evolution are central for both research and practice.

One of the most popular tools for this purpose is Fourier Transform Infrared (FTIR) spectroscopy. Its principle is based on the absorption of infrared light by molecular vibrations. When a molecule absorbs IR radiation, its dipole moment changes, leading to vibrations and rotations of specific bonds. These processes are reflected as characteristic absorption bands in the resulting spectrum [6]. In the case of bitumen, many functionally important groups are IR-active, including carbonyls, sulfoxides, carboxylic acids, anhydrides, aromatic rings, and various aliphatic hydrocarbon (CH_X) bonds. Each of these groups has a characteristic spectral position, which makes FTIR especially attractive because it can provide a "fingerprint" of the material. Despite the fact that bitumen is an extremely complex mixture, FTIR simplifies the picture by highlighting those chemical moieties that dominate in certain processes, such as oxidation or polymer modification [3, 5, 7].

A particularly relevant innovation in this context was the introduction of Attenuated Total Reflectance (ATR). Before ATR became popular, FTIR spectra were usually measured in transmission mode, which required dissolving the binder in a solvent, casting it on a salt crystal, and waiting for the solvent to evaporate before measurement [8]. This procedure was not only time-consuming but also risky, as it exposed the thin film to oxidation or contamination during evaporation. ATR made things much easier. In this mode, the solid binder is simply pressed against a crystal, most often diamond, and the spectrum is recorded directly within minutes. Diamond ATR crystals are robust, resistant to temperature and mechanical stress, and suitable for routine use. This development made

S. Weigel
Division 7.1 Building Materials, Federal Institute for Materials Research and Testing, Unter Den Eichen 87, 12205 Berlin, Germany

M. Motevalizadeh
Department of Engineering and Maintenance of Road Infrastructure, Faculty of Civil and Environmental Engineering, University of Kassel, Mönchebergstraße 7, 34125 Kassel, Germany

S. Khalighi · A. Varveri
Pavement Engineering Section, Engineering Structures Department, Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, the Netherlands

FTIR spectroscopy much more practical in bitumen research, leading to its broad adoption as a fast, cost-efficient, and non-destructive technique for evaluating chemical composition, detecting modifiers, and following the effects of aging [6, 9].

Although FTIR spectroscopy has many advantages, its application to bitumen is not without challenges. One of the most pressing issues is reproducibility, since there is no standardized procedure yet. When the same binder is analyzed in different laboratories, the spectra often differ in baseline position, intensity, or even peak shifts. This has been observed in several interlaboratory studies [6, 9, 10]. The reasons are manifold: Differences in sample preparation, such as heating procedures, homogenization, or the thickness of the film applied to the ATR crystal, can strongly influence absorbance values. Furthermore, variations in device hardware and operator practice introduce additional scatter. For example, ATR crystals can be made from diamond, germanium, or zinc selenide, and each material has slightly different refractive indices, penetration depths and sensitivities. Even the pressure applied when placing the binder on the crystal may alter the measurement. On top of these issues comes the challenge of data processing. Normalization strategies, baseline corrections, and integration ranges vary widely between laboratories, which can lead to substantial differences in reported indices [11, 12].

Several authors have addressed these problems. Mirwald et al. [6] investigated the influence of handling procedures and concluded that strict control of heating time, sample storage, and homogenization is crucial to minimize variability. Hofko et al. [9] carried out an interlaboratory testing within the RILEM TC 252-CMB with nine laboratories and showed that significant baseline differences exist, but also that preprocessing strategies can reduce variability to some extent. However, most of these studies were relatively limited in scope, either in the number of binders tested or in the number of participating laboratories. Broader studies with more comprehensive data are needed to address reproducibility in a systematic way.

When it comes to the actual interpretation of spectra, traditional analysis has usually been univariate. In most studies, researchers focused on a few well-established bands, especially the carbonyl absorption around 1700 cm^{-1} and the sulfoxide absorption around 1030 cm^{-1}, which are clear indicators for the incorporation of oxygen via formation of these oxygen containing functional groups [3, 5, 9]. Increases in these bands are often used to quantify aging. Other indices combine band areas with reference regions, such as aromatic signals, to correct for intensity differences. While these indices are useful, they only capture a small fraction of the available information. The extended fingerprint region between 1800 and 600 cm^{-1}, where a large number of chemically relevant signals occur, is particularly difficult to interpret because of overlapping peaks, low intensities, and baseline distortions [5]. As a result, important information is often overlooked. Moreover, simple indices are highly sensitive to operator choices, such as what kind of integration method was selected or where integration limits were set. All this makes univariate analysis prone to inconsistencies and less suitable for capturing the chemical complexity of bitumen.

To overcome these limitations, researchers increasingly apply multivariate analysis (MVA). The central idea of MVA is that the entire spectral profile is used, rather than focusing on selected individual spectral regions or bands. This approach captures subtle but systematic differences that would otherwise remain hidden. Techniques such as Principal Component Analysis (PCA) work without class labels. They reduce the dimensionality of the data while retaining the most relevant variance, thereby highlighting similarities and differences between samples in a lower-dimensional space. Supervised methods like Linear Discriminant Analysis (LDA) or Partial Least Squares Regression (PLSR) then build on this reduced representation to classify samples into predefined categories, such as different binder types or aging states. Other methods, such as Hierarchical Cluster Analysis (HCA) or Random Forest (RF) models, are used to group spectra according to similarity and to identify the spectral regions most important for discrimination [7, 11, 13–15].

The strength of MVA lies in its ability to address multiple problems at once. By analyzing the entire spectrum, overlapping peaks or baseline shifts are less problematic because the statistical model can still extract the relevant variance. By reducing dimensionality, noise is filtered out while chemically meaningful structures are retained. Furthermore, because many MVA methods are supervised, they can maximize separation between groups, such as binders from

different origins or different aging levels. Importantly, MVA is not only a classification or prediction tool but also provides chemical insights. Loadings, variable importance scores, or feature importance metrics point to the exact wavenumber regions that drive the classification or prediction, thereby linking statistical patterns back to chemical processes.

Several studies illustrate the benefits of MVA in bitumen research. Weigel and Stephan [16] applied Factor Analysis and LDA to classify binders according to refinery origin and aging stage. Their results showed that long-term aged binders can be clearly separated from unaged ones, although distinguishing unaged from short-term aged materials proved more difficult when multiple refinery sources were included. Follow-up work demonstrated that the same approach can also identify polymer and wax modifications [17] or detect the presence of polycyclic aromatic hydrocarbons (PAHs) in binders [18]. Ma et al. [3] used PCA combined with LDA and feature selection to investigate aging states, showing that both aliphatic and aromatic regions are important for classification. Motevalizadeh et al. [19] expanded the methodology by coupling PLSR with LDA to classify binders not only by aging but also by modification level, and they used Variable Importance in Projection (VIP) scores to explain which spectral regions contributed most to their models.

More recent studies integrated FTIR-based MVA with other types of data. Khalighi et al. [20] combined PCA with machine learning to link chemical and rheological information, showing that such approaches can improve the development of laboratory aging protocols that better represent field conditions. Other researchers applied HCA to group laboratory- and field-aged binders, revealing similarities and differences in their evolution [21]. Motevalizadeh and Mollenhauer [22] combined FTIR with micromechanical modeling to explore the behavior of mastics. PCA with Euclidean distance analysis has also been used to compare aging across different scales, from binder to mixture [23]. These studies underline that MVA is not limited to aging analysis but can also be applied to questions of binder modification, reclaimed asphalt content, or the relationship between chemical and mechanical properties.

Despite these promising applications, the limitations of existing literature are clear. Most datasets were relatively small, restricted to one or a few binder types, and focused on specific laboratory conditions. As a result, reproducibility across laboratories was rarely addressed, and transferability of methods remains uncertain. Moreover, preprocessing pipelines differ widely between studies, ranging from baseline correction to normalization and smoothing, making comparisons difficult. Finally, most studies concentrated on a single classification task, without developing a generalizable framework that could be applied more broadly. For these reasons, MVA has not yet become a standardized or widely accepted tool in bitumen research.

To address this gap, the RILEM TC 295-FBB TG1 launched an international round robin test on FTIR spectroscopy. The primary innovation of this study lies in the systematic evaluation of both inter- and intra-laboratory variability on an unprecedented scale, capturing both chemical diversity and procedural fluctuations. By collecting data from over 20 laboratories, this initiative generated a comprehensive dataset that provides a robust foundation for systematic, data-driven analysis.

2 Objective

The present study makes use of this dataset to assess the value of multivariate analysis (MVA) for FTIR-based bitumen characterization. Three objectives are central.

- The first is to examine reproducibility and variability across laboratories when spectra are analyzed with multivariate methods.
- The second is to identify which spectral regions are most relevant for differentiating binders by type and aging, linking statistical outcomes back to chemistry.
- The third is to test the robustness and transferability of MVA in this context, to provide a foundation for standardized procedures.

By addressing reproducibility and interpretability in an international framework, this work seeks to demonstrate that FTIR spectroscopy combined with multivariate techniques can serve as a reliable and broadly applicable tool. Ultimately, the findings aim to support the development of standardized protocols for chemical fingerprinting of bitumen, benefiting

both fundamental research and practical applications in pavement engineering.

3 Materials and methods

3.1 Round robin test

The data used in this study were generated as part of an international round robin test from the RILEM TC-295 FBB Task Group 1. 21 laboratories took part, each receiving the same homogenized and consistently aged binder samples to measure them using ATR-FTIR spectroscopy. In total, more than 6786 spectra were collected, spanning three binder types, three ageing levels (unaged, laboratory short- and long-term aged), and six different preparation techniques. The main goal of the round robin test was to assess how reproducible and consistent ATR-FTIR measurements are under real-world laboratory conditions. The resulting dataset offers a strong basis for examining how factors like equipment, sample handling, and preparation method can influence the measured spectra, using multivariate analysis.

3.2 Materials

This study focused on three unmodified bituminous binders, all classified as 70/100 penetration grade according to European specifications. Although they fall under the same specification class, the binders were sourced from different suppliers to reflect natural variations in raw materials. The binders selected were:

- B1158 (supplied by TU Wien): needle penetration of 84 (1/10 mm), softening point of 45.8 °C
- B1198 (supplied by TU Delft): needle penetration of 85 (1/10 mm), softening point of 47.0 °C
- B1199 (provided by Nynas): needle penetration of 84 (1/10 mm), softening point of 45.8 °C

Each binder was analyzed at three distinct ageing conditions to simulate its evolution over service life:

- A – unaged (original binder)
- B – short-term aged using the Rolling Thin Film Oven Test (RTFOT)
- C – long-term aged using a combination of RTFOT and the Pressure Ageing Vessel (PAV)

All ageing processes were carried out centrally at TU Wien to ensure consistent and reproducible conditions across all samples. After ageing, the binders were homogenized, flushed with nitrogen to prevent oxidation, and sealed in airtight containers before being distributed to the participating laboratories.

Short-term ageing followed the EN 12607–1 [24] protocol, in which the binders were exposed to 163 °C for 75 min in a RTFOT. Long-term ageing was performed according to EN 14769 [25], using a Pressure Ageing Vessel at 100 °C and 2.1 MPa for 20 h.

This resulted in a total of nine aged binder samples:

- B1158A, B1158B, B1158C
- B1198A, B1198B, B1198C
- B1199A, B1199B, B1199C

Further details on the ageing process, sample handling, and logistics can be found in [10].

3.3 Sample preparation methods

To evaluate the influence of sample handling on the spectral response of bituminous binders, six different sample preparation techniques were applied prior to ATR-FTIR spectroscopic analysis. These methods vary in terms of binder quantity, heating approach, homogenization effort, and use of solvents. All methods aimed to produce reproducible droplet-shaped samples suitable for application onto the ATR crystal, while minimizing alterations to the binder's chemical structure. Additional information on the sample preparation methods is provided in [10].

The Large Quantity – Heating Plate (LQ_HP) method involved heating approximately 5 g of binder in an open metal container on a heating plate at 180 °C for 5 to 10 min. Continuous stirring with a thermometer ensured thermal monitoring and sufficient sample homogeneity. Once a workable viscosity was achieved, sample droplets were formed using the thermometer tip and placed on silicone paper. The droplets were then covered to protect them from light exposure and environmental contamination and measured after cooling to room temperature.

In the Large Quantity – Oven (LQ_O) method, the binder was placed in a closed metal container and heated in a ventilated oven at 180 °C. After 5 to 10 min, the liquified binder was homogenized manually, and droplets were applied on silicone paper. After droplet formation, the samples were covered to prevent light exposure and environmental contamination before being measured at room temperature. While the enclosed heating setup minimizes oxidation, the lack of stirring during heating results in less uniform temperature distribution within the binder.

The Metal Can (MC) method represented the most direct approach, applying the binder straight from the shipped binder can onto the ATR crystal without reheating and homogenization. While extremely fast, this method can be susceptible to poor repeatability due to local inhomogeneities.

Two "small quantity" (> 1 g of binder) approaches were also tested: Small Quantity – Heating Plate (SQ_HP) and Small Quantity – Hot Gun (SQ_HG). Both methods used around 0.5–1 g of binder that was placed in a small metal spoon. In SQ_HP, the spoon was placed directly on a heating plate at 180 °C for one minute. In SQ_HG, the sample was heated using a hot air gun, with care taken not to exceed 180 °C. In both cases, stirring with a thermometer helped ensure uniform heating. These methods are energy-efficient and allow flexible, localized heating, but require careful temperature monitoring to avoid binder degradation. As with the previous heating-based methods, the prepared droplets were covered to limit light exposure and environmental contamination and were measured only after cooling to room temperature.

Finally, the Solvent (S_Tol) method involved dissolving 1 g of binder in 3 g of toluene. After full dissolution, droplets were applied directly to the ATR crystal. Once the solvent fully evaporated—verified by monitoring the preview spectrum—the measurement was initiated. While this approach avoids thermal stress, it introduces the possibility of solvent–binder interactions and relies on complete evaporation for accurate analysis.

Together, these six methods represent a comprehensive cross-section of commonly used binder preparation techniques for FTIR spectroscopy in laboratory environments. Their systematic comparison within the round robin test allows for a robust assessment of their impact on spectral quality, repeatability, and their suitability for quantitative or chemometric analysis.

3.4 Analysis methods

3.4.1 ATR-FTIR spectrometer and parameters

All devices used in the round robin test as well as their respective specifications can be found in Table 1. All devices were calibrated and operated according to protocols given to the laboratories before the round robin test to ensure inter-laboratory comparability.

The measurements of the spectra were performed over the mid-infrared range of 4000 to 600 cm^{-1} with a resolution of 4 cm^{-1}. For bituminous materials, this spectral range is particularly informative due to the characteristic absorption bands associated with aliphatic and aromatic C–H stretching vibrations (3050–2800 cm^{-1}), carbonyl (C = O) stretching from ketones and carboxylic acids (1750–1700 cm^{-1}), sulfoxide and sulfate ester functionalities (1030–1080 cm^{-1}), and aromatic ring out-of-plane vibrations (900–700 cm^{-1}),although bands in this region may also originate from aliphatic CH_2 bending around 722 cm^{-1}[9, 26, 27]. A majority of these functional group vibrations are sensitive to oxidative ageing and chemical modification processes, making them valuable chemical markers for assessing bitumen durability and structural changes over time. In particular, the formation of carbonyl and sulfoxide groups has been consistently linked to long-term ageing in bituminous binders [26]. The presence and evolution of these spectral features allow for both qualitative and quantitative assessment of degradation pathways, including polymer–bitumen interactions and binder hardening under environmental stress. However, potential noise near the 600 cm^{-1} cutoff requires caution when explicitly analyzing polystyrene degradation.

Each measurement was performed with an instrument related number of scans which were averaged to enhance the signal-to-noise ratio. Before the sample was applied on to the device a background spectrum of the clean ATR crystal was recorded. Immediately after the recording of the background (< 1 min) the sample was applied to minimize atmospheric interferences. After each measurement, the ATR crystal was cleaned using non-toxic bitumen solvent (e.g., limonene) followed by fast

Table 1 Information on the FTIR spectrometers used in the round robin test

Lab Code	Manufacturer	Model	Device Mode	ATR Crystal	Detector
1	2	A	ATR	Diamond/ZnSe	DTGS
2	3	A	ATR	Diamond	DLaTGS
3	4	A	ATR	Diamond	DLATGS
4	1	B	ATR	Diamond	DTGS
5	1	A	ATR	Germanium	DTGS
6	1	B	ATR	Diamond	DTGS
7	1	A	ATR	Diamond	DTGS
8	1	B	ATR	Diamond	DTGS
9	1	C	ATR	Diamond	MTC
10	3	A	ATR	Diamond	DTGS
11	3	A	ATR	Diamond	DTGS
12	2	C	ATR	Diamond	LiTaO3
13	2	A	ATR	Diamond/ZnSe	LiTaO3
14	2	A	ATR	Diamond/ZnSe	LiTaO3
15	3	A	ATR	Diamond	DTGS
16	3	A	ATR	Diamond	DTGS
17	1	B	ATR	Diamond	DTGS
18	1	B	ATR	Diamond	DTGS
19	2	B	ATR	Diamond	DTGS
20	3	B	ATR	Diamond	DTGS
21	3	C	ATR	Diamond	DTGS

evaporating solvent (e.g. isopropanol), to eliminate cross contamination. The measurement procedures were aligned with previously established protocols [6].

Four replicates were made for each of the four samples. These were measured for every combination of bitumen source (B1158, B1198, B1199), ageing state (A, B, C) and preparation method (LQ_HP, LQ_O, MC, SQ_HP, SQ_HG, S_Tol), resulting in 16 spectra per category. This structure is chosen to capture intra-sample variability and improve statistical power in downstream analysis.

3.4.2 Data pre-processing methods

To enhance robustness and spectral comparability all the FTIR spectral data underwent a structured data-pre-processing workflow, prior to the multivariate statistical analysis. Two steps of pre-processing were applied:

1. Min–Max normalization
2. Outlier filtering based on Z-score statistics

To preserve the integrity of the spectral data and avoid potential overfitting or distortion of genuine spectral features, no baseline correction or smoothing was applied. This decision aligns with recommendations in the literature [11], which suggest limiting preprocessing to essential steps to maintain the authenticity of the data.

Spectral intensity values across all samples were rescaled to a common range of 0 to 1 using Min–Max normalization. The normalization was performed using the local minimum and maximum within the aliphatic C-H stretching region (3200—2800 cm^{-1}), which is commonly recognized for its stability and high signal to noise ratio in bituminous materials [15]. With the choice of this spectral window, it was ensured that scaling was anchored in a chemically meaningful and consistent region across all spectra. The mathematical definition of this Min–Max normalization is detailed in Appendix A1.

This approach aligns with methodologies employed in previous studies on bitumen characterization using FTIR spectroscopy [28].

After normalization, outlier detection was applied using the Z-Score method. For each group

of spectra (defined by sample ID and preparation method), Z-Scores were calculated column wise across the wavenumber dimension (see Appendix A1 for the exact formulation).

If the Z-score was exceeding the threshold of 3 in at least one variable, the spectra was flagged and removed. With this approach anomalous intensities or noise artefacts, which may arise from inadequate ATR contact, contamination or instrument drift in spectra were excluded. In total, approximately 4% of the spectra (334 out of 6786) were removed as outliers according to this criterion. The remaining 6452 spectra were retained for further analysis, comprising 2409 spectra from LQ_HP, 1057 from LQ_O, 916 from MC, 770 from SQ_HG, 1040 from SQ_HP, and 260 from S_Tol.

This preprocessing of the data ensured that clean and comparable spectra were available for further analysis. The methods were implemented using custom Python routines and follow recommendations for spectral preprocessing in FTIR-based bitumen research [15].

3.4.3 Coefficient of variation and interclass correlation

Two statistical metrics were employed to assess the reproducibility and discrimination of the FTIR spectra across different sample preparation methods and binder identities:

1. Coefficient of Variation (CV)
2. Intraclass Correlation Coefficient (ICC)

They were calculated for every unique combination of preparation method and sample ID, allowing us to systematically compare the performance of various sample ID – preparation method pairings in terms of spectral reliability and differentiation.

3.4.3.1 Coefficient of variation The mathematical definition of the CV is provided in Appendix A2. In this study, CV values were calculated for each combination of preparation method and sample ID by determining the mean and standard deviation at each wavenumber across all replicates. The resulting values were then averaged across the entire spectrum to provide a single measure of variability within the groups. Low CV values indicate that the sample preparation is reliable and consistent, while high CV values suggest issues such as inadequate homogenization, contamination by dust, inconsistent film thickness (potential overuse of the pressure lever), or instability in the device.

In the realm of spectroscopic studies, the CV is a go-to method for checking how repeatable sample acquisition is. It's been instrumental in identifying strong experimental conditions [29, 30].

3.4.3.2 Intraclass correlation coefficient While the CV offers a way to measure internal variability, the ICC helps us to understand how much of the total variance is linked to differences between the groups (in this case, between different sample ID – preparation method pairings) rather than to random variation within groups. The ICC is calculated using a two-way random effects model with absolute agreement. The corresponding formula is listed in Appendix A2.

An ICC value close to 1 means that most of the variability comes from actual differences in preparation methods or samples. On the other hand, a low ICC indicates that the measurements are largely influenced by noise within the groups, making them less effective in distinguishing between them. This metric is particularly useful in research that includes repeated measurements under various circumstances and has been widely applied in multivariate spectroscopy to ensure measurement reliability [9, 31].

In this analysis, ICC values were calculated for the same groupings used in the CV. By comparing the CV and ICC for each group, a thorough understanding of sample quality and how well they can differentiate between chemical features in FTIR spectra was gained. The sample ID—preparation method combinations that showed both low CV and high ICC were deemed the most reliable for distinguishing these features.

The use of CV and ICC allows not only the identification of outlier preparation methods, but it also enables the detection of combinations that are suited for classification and clustering tasks. While high CV values can indicate reproducibility issues, a low ICC additionally confirms that the method fails to provide sufficient contrast between groups. In general, CV values below 0.3–0.4 are considered indicative of good repeatability, whereas ICC values above 0.9 denote excellent agreement, 0.75–0.9 good and below 0.75 moderate to poor reproducibility [31].

3.4.4 Principal component – linear discriminant analysis

To extract structure and reveal discriminatory patterns in high-dimensional FTIR spectral data, two complementary multivariate techniques were employed:

1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)

These methods provide both unsupervised and supervised insights into complex spectral datasets. Unsupervised methods (such as PCA) explore inherent structures in the data without using predefined labels, while supervised methods (such as LDA or Random Forest) rely on labelled data to train models for classification or prediction. Such approaches are commonly applied in chemometrics and material science for tasks like feature extraction, visualization and classification [32–34]. The mathematical principles and more theoretical background of these methods are detailed in Appendix A3.

3.4.4.1 Principal component analysis PCA is an unsupervised linear transformation technique used to reduce the number of dimensions in a dataset while still retaining as much of the original variance as possible. In the context of FTIR data, where each sample is represented by thousands of wavenumber-dependent absorbance values, PCA helps clarify the data by projecting it into a space defined by orthogonal components, referred to as principal components (PCs).

In this study, PCA was applied using the *sklearn.decomposition.PCA* class from the *scikit-learn* library [35]. The number of components was set to 50 PCs in the dimensionality reduction step to maintain a high level of the original spectral variance while also cutting down the number of wavenumber variables. This setup allowed a balance between noise reduction and computational manageability, ensuring there is still enough information for accurate classification later. The selection of 50 PCs was based on the cumulative explained variance, which exceeded 95% for all the spectra.

3.4.4.2 Linear discriminant analysis LDA is a supervised classification and dimensionality reduction method. It projects high-dimensional dimensional data onto a low-dimensional space where predefined classes are maximally separated. This allows for the assessment of whether different preparation methods, instruments, or sample identifiers produce distinguishable spectral patterns. LDA tries to find linear combinations of features (in this case the PCs from the before PCA) that maximize the separation between predefined classes.

In this work, LDA was implemented using the *LinearDisciminantAnalysis* class from *scikit-learn* [35]. The number of components was typically set to 2 for visualization. The target labels for LDA included preparation method, sample ID and the devices. To evaluate the model performance and avoid overfitting, a five-fold cross-validation (XVal) was applied. In k-fold XVal, the dataset is randomly partitioned into k subsets (folds). For each iteration one fold is used as the test set while the remaining k-1 folds are used for training. The process is repeated k times, and the overall classification accuracy is computed as the average of all folds (see Appendix A3 for the exact equation).

Given the uneven distribution of samples among the classes, a stratified k-fold cross validation approach was adopted to ensure that each fold maintained approximately the same class proportions as the full dataset. This procedure reduces potential bias and provides a more robust estimate of the model's generalization performance [36].

LDA is a great tool for separating classes, but it does tend to overfit, especially in high-dimensional situations where the number of features exceeds the number of samples. Additionally, LDA assumes that predictor variables are linearly independent, and that within-class covariance matrices are equal. This assumption is often violated in high-resolution spectral data due to strong collinearity between neighboring wavenumbers [37]. This collinearity can lead to unstable covariance matrix estimations, resulting in poor generalization or numerical issues. PCA mitigates these issues by reducing dimensionality beforehand, which orthogonalizes the features and eliminates redundant information. This process stabilizes LDA, minimizes noise amplification and improves classification performance. For this reason, PCA was used in combination with LDA in this analysis to first extract meaningful low-dimensional data and then apply LDA on the resulting scores [38].

3.4.5 Random forest

To validate and enhance the outcomes from the LDA, an alternative machine learning approach based on Random Forest (RF) classification was applied. Unlike LDA, which is a linear and parametric method, RF offers a non-linear, ensemble-based solution that can handle complex variable interactions and high-dimensional datasets without the need of reducing dimensions or following strict statistical guidelines. This made RF especially well-suited for confirming the robustness and generalizability of the class distinctions derived from the FTIR spectral data. The formal statistical background including the RF definition and the Gini impurity criterion used for node splitting is provided in Appendix A4.

In this study, RF was applied to the same dataset that has been used for the PCA-LDA. The aim was to see if this non-parametric classifier (a classifier that doesn't rely on assumptions like linearity or equal variance) could provide similar insights into class separability among different preparation methods, sample IDs or the devices. Furthermore, RF's feature importance metrics allowed independent identification of the wavenumber regions that were crucial for classification.

The Random Forest classifier was set up using the *RandomForestClassifier* class from the *scikit-learn* Python package [35]. The model was configured with 100 decision trees (*n_estimators = 100*), which provided a sufficient ensemble size to ensure the classification performance remains stable. For each node split, a random subset of features equal to the square root of the total number of predictors (*max_features = 'sqrt'*) was considered. This is a standard setup that enhances model diversity and reduces overfitting. A fixed random seed (*random_state = 42*) was set to guarantee the reproducibility across different runs. The model was trained using bootstrap, where each tree was fitted on a bootstrapped sample taken with replacement from the original dataset. To enhance computational efficiency, all available processor cores were utilized at the same time. The selection of hyperparameters is in line with established practices in high-dimensional spectral analysis, where the feature spaces are often large and the interactions between variables can be quite complex [39]. Similar to LDA, k-fold cross-validation was implemented with k = 5 to evaluate how well the model performs.

3.4.6 Feature importance

One of the key advantages of the RF algorithm (beyond robustness and classification accuracy) is its ability to measure the relative importance of input features for prediction. In FTIR spectroscopy, where each feature corresponds to the absorbance at a specific wavenumber, this ability allows a straightforward and clear evaluation of which spectral regions are most significant for distinguishing between classes. The exact equation and mechanics for computing this feature importance are detailed in Appendix A5.

This metric was implemented as the *feature_importances_* attribute in the *RandomForestClassifier* class of the *scikit-learn* Python package [35].

In the present study, feature importance was calculated for each wavenumber after the RF classifier was trained on the Min–Max normalized spectral data. The resulting importance values were then normalized to a range of [0,1] and plotted against the corresponding wavenumber axis. This method serves as a valuable counterpart to the coefficient-based interpretation found in LDA. While LDA is all about linear separation and provides global discriminant directions, RF feature importance reveals non-linear decision boundaries and captures local interactions between variables.

Although impurity-based feature importance in RF is widely used and intuitive it has some limitations. It often favours features that show more variability or have a larger number of distinct values. This fact is less problematic in working with FTIR data, where all features are continuous and similarly scaled after the normalization [40]. Moreover, the importance values are relative and specific to the model, which means they should be considered alongside cross-validation and repeated training to ensure stability. Even with these limitations, RF feature importance is still one of the most powerful and interpretable methods for feature selection in spectroscopy [41]. In this study it helped to identify not only which spectral regions were discriminative, but also whether those regions were robust across different sample subsets and class definitions.

4 Results and discussion

4.1 Reproducibility of preparation methods (CV and ICC analysis)

The reproducibility of the various sample preparation methods was initially assessed using the CV and ICC analyses. The results of the CV analysis are presented in a heatmap, shown in Fig. 1.

In the CV heatmap, blue fields indicate lower variation coefficients, thus representing higher reproducibility of FTIR measurements across the different laboratories. As the variation between laboratories increases, the coefficient increases, which is reflected by a gradual shift from blue to red in the color gradient. The size of the rectangles corresponds to the number of available data for each combination. The preparation method showing the least variation was SQ_HG, with CV values ranging from 0 to 0.23. This method therefore offered the highest average reproducibility of the FTIR spectra in the participating laboratories.

The methods LQ_O and S_Tol exhibited the next best reproducibility. In contrast, the method SQ_HP showed the greatest variation, with a maximum CV of 1 for the sample B1158B.

However, it should be noted in this evaluation that the CV values in the diagrams represent the average variability across the entire spectrum. This includes regions that do not contain significant information related to the bitumen structure. It is, therefore, insightful to examine the CV profiles more closely. Figure 2 shows the mean CV values of the nine samples analyzed for the different preparation methods. Additionally, the spectrum of sample B1158 (Lab 1, LQ_HP) is presented for better interpretation.

The figure reveals that the largest variations occur in regions where no prominent bands are present in

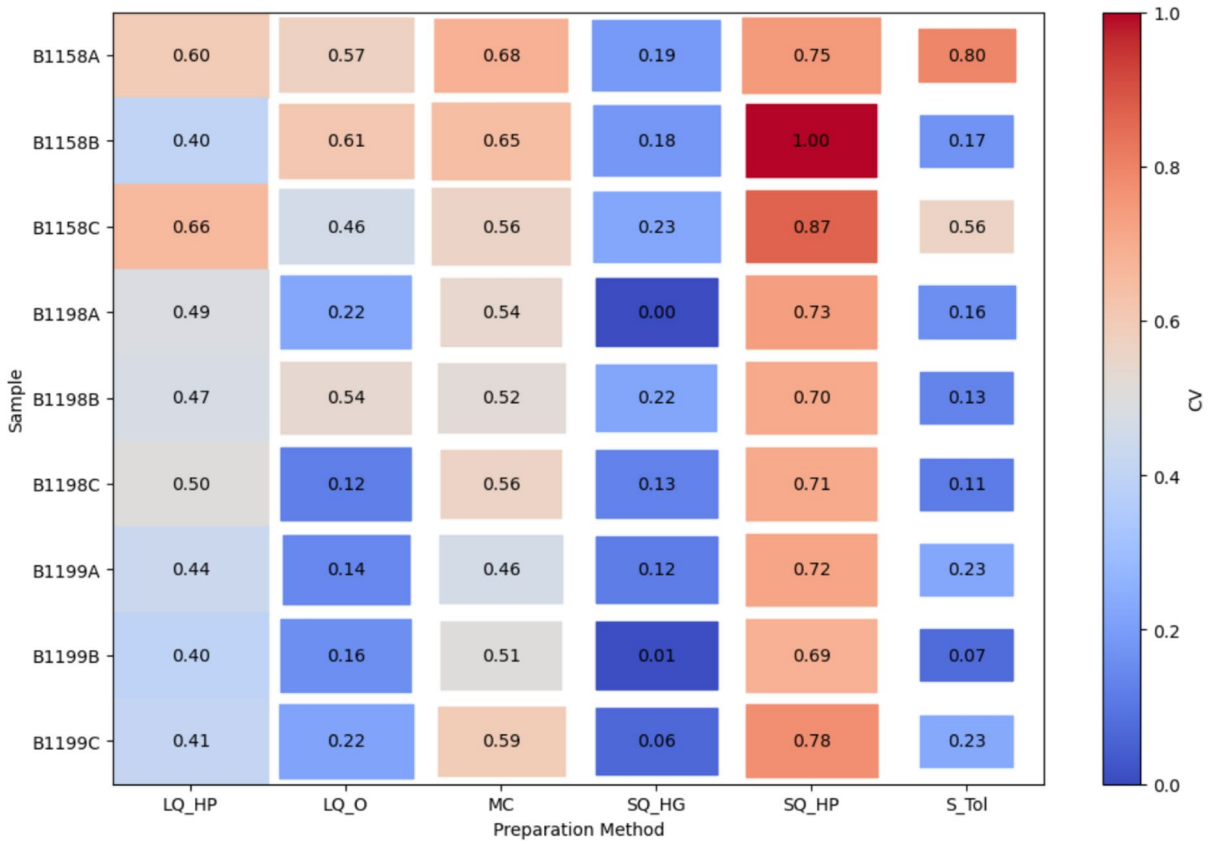

Fig. 1 Heatmap of CV for each combination of binder type and sample preparation method with lower CV values (dark blue) for higher repeatability and higher values (red) for higher variability

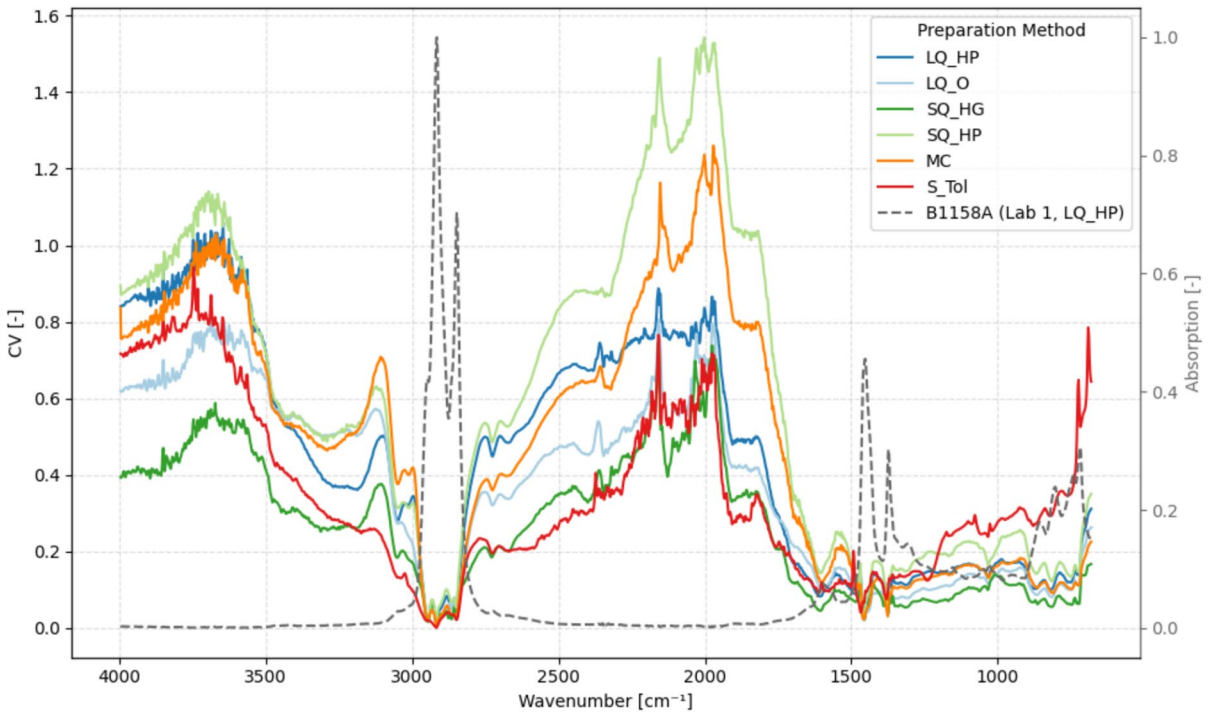

Fig. 2 Mean values of the CV for various preparation methods over the entire investigated wavenumber range, supplemented by the spectrum of the sample B1158A (Lab 1, LQ_HP) for comparison

the bitumen spectrum, such as between approximately 4000–3000 cm^{-1} and 2700–1800 cm^{-1}. In these regions, even small deviations in absorbance lead to significant increases in the CV. In the region around 2000 cm^{-1}, atmospheric influences, such as carbon dioxide (CO_2) and water (H_2O), or the own absorption of the ATR crystal affect the measurements. To eliminate these influences, atmospheric corrections are typically applied by the instrument's software. These corrections can introduce additional variations, which could further contribute to the high CV values in this range. However, these regions are not relevant for the FTIR analysis of bitumen and therefore are not of major concern.

In contrast, the regions associated with bitumen bands exhibit noticeably lower CV values for all methods, indicating greater consistency in these areas. However, the ranges between 3500–3000 cm^{-1} and 1800–1600 cm^{-1} stand out, as the CV values and thus the variations remain relatively high, although broad and additional smaller bands are present in the bitumen spectra. This can be attributed to the low band intensities, where even minor absolute deviations lead to relatively high CV values.

To further interpret the reproducibility analysis, absolute deviations of the spectra were assessed by measuring the span between the maximum and minimum absorbance values for each wavenumber. These deviations were first calculated for each preparation method and sample (across the different laboratories) and subsequently averaged over all samples of a preparation method. The results are presented in Fig. 3.

The comparison clearly shows that the deviations between laboratories for the different preparation methods are of comparable magnitude. In most of the spectral range, these deviations are below 0.1, indicating minimal variability. Only in regions corresponding to larger bands the deviations increase, with values reaching up to 0.3.

An interesting observation in terms of very high spans was made for the S_Tol method in the range of approximately 700–750 cm^{-1}, where characteristic toluene bands occur. This suggests that the solvent may not have fully evaporated during the measurement. For accurate and comparable spectra, it is

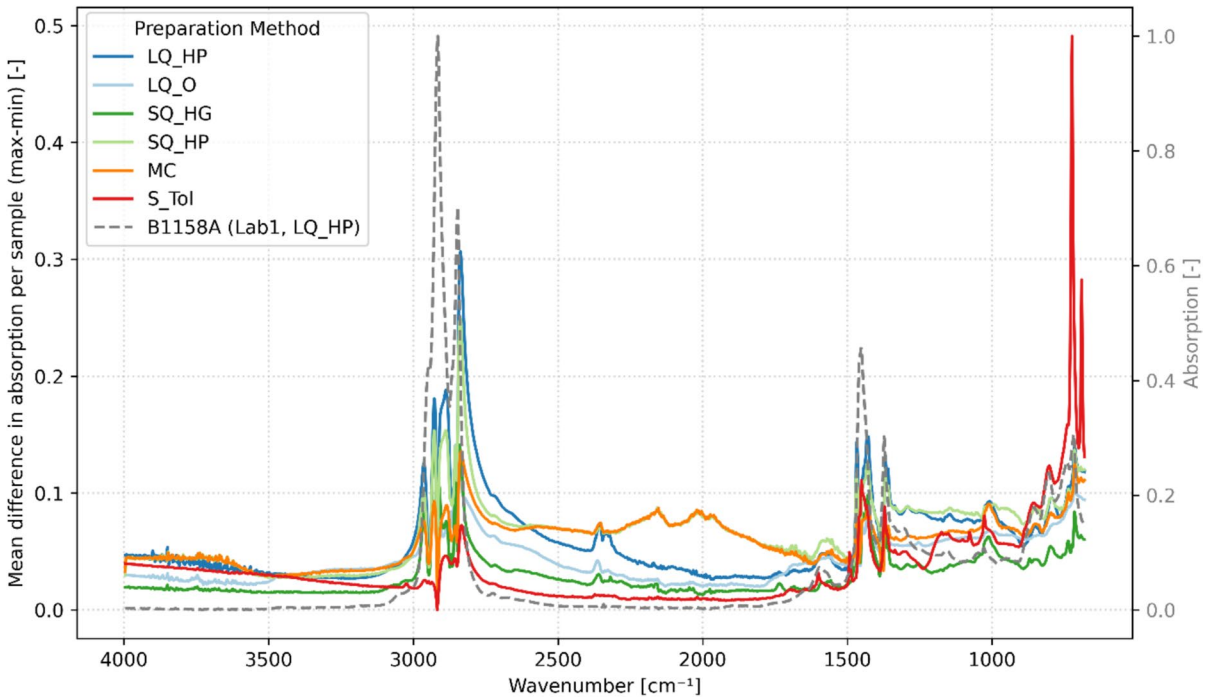

Fig. 3 Mean difference of the spectral values of the nine samples (across the different laboratories) per preparation method, supplemented by the spectrum of the sample B1158A (Lab 1, LQ_HP) for comparison

crucial that the solvent is completely evaporated. Adequate evaporation times must be ensured and adhered to.

Thus, the most significant variability was observed in the S_Tol method, which is attributed to insufficient evaporation time. For the other methods, the results, in terms of absolute deviations, were within comparable ranges.

The results of the ICC analysis are shown as a further heatmap in Fig. 4. High ICC values indicate that the variation between the different groups is much larger than within the groups, meaning that the variation is largely explained by the differences between the groups and thus the bitumen samples. These high values are marked by red fields, while increasing noise (unexplained variance within the groups) leads to lighter or bluer areas. High ICC values therefore demonstrate that the differences between the samples are well captured in the FTIR spectra.

The diagram shows very good ICC values for the methods LQ_HP, LQ_O, MC, and SQ_HG, with the highest values observed for the SQ_HG method. In contrast, the methods SQ_HP and especially S_Tol displayed relatively lower ICC values, indicating that distinguishing between the samples is less effective for these methods compared to the others.

Overall, these results show differences between the various preparation methods with regard to the reliability of the FTIR spectra. While the reproducibility is relatively comparable, the SQ_HG and S_Tol methods lead to poorer distinction between different bitumen samples. However, it should be noted that the S_Tol method was only used by two laboratories and the basis is therefore very small.

4.2 Multivariate classification: PCA-LDA results

The distinction of the various binder groups was examined using PCA-LDA. Prior to distinguishing between the binder samples themselves, the influence of preparation methods and the devices used on the FTIR spectra was first considered to identify potential impacts from the measurement conditions.

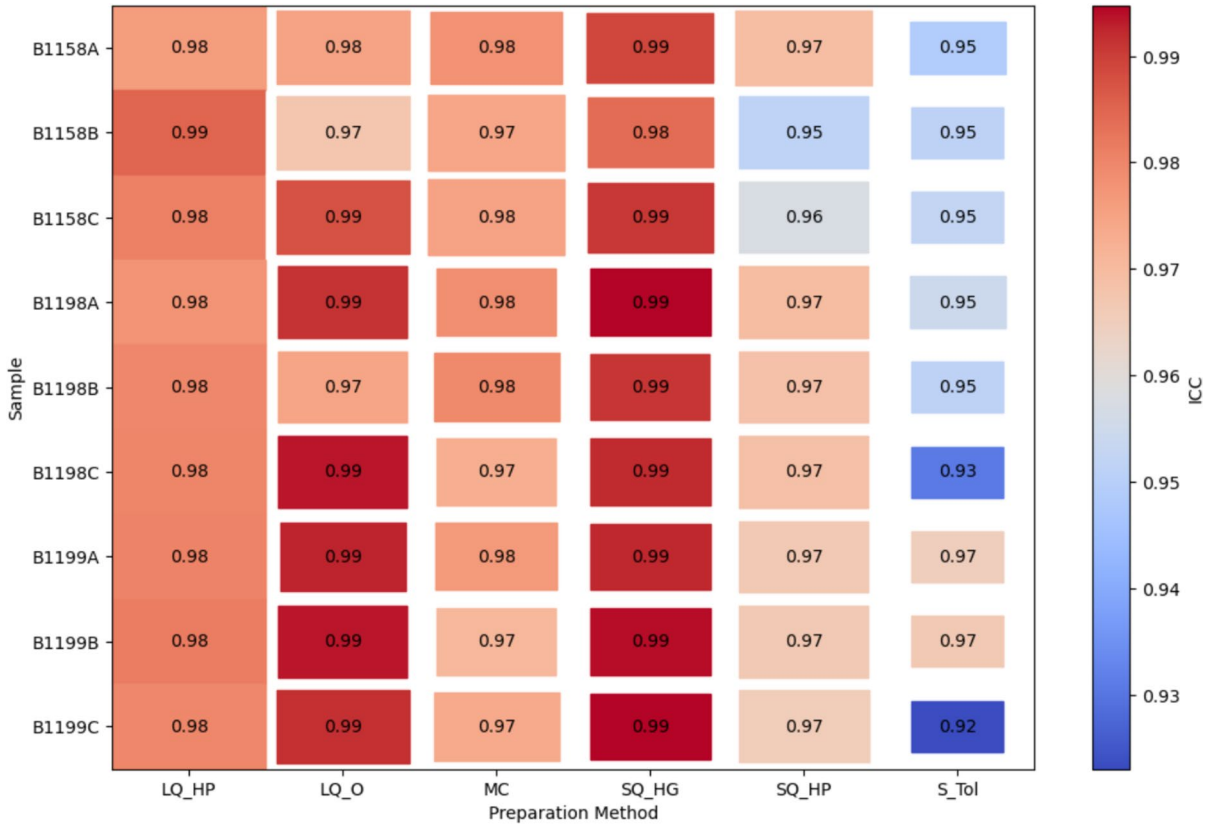

Fig. 4 Heatmap of ICC for each combination of binder type and sample preparation method with higher ICC values (dark blue) for better reproducibility within the corresponding sample group

4.2.1 *Effect of preparation methods on the FTIR spectra*

Initially, the focus was on the different preparation methods. Figure 5 presents a comparison of the first two LDA components.

The analysis clearly shows that spectra from preparation methods involving homogenization (LQ_HP, LQ_O, SQ_HG, SQ_HP) group together, while spectra from the MC method form a separate group that is adjacent to the homogenized spectra. Spectra from the S_Tol method form a distinct, highly scattered group, completely separating from the other spectra. The spectra of samples prepared and measured using the S_Tol method thus exhibit significant differences when compared to those prepared using the other methods.

Spectra recorded after homogenization show no systematic differences in the spectra, regardless of sample quantity (LQ, SQ) or the heating method (HG, HP, O). Spectra from the MC method, however, differ from those of homogenized samples, whereby they exhibit smaller differences than those of the S_Tol method. A detailed analysis of these differences and the relevant wavenumber ranges is presented in Sect. 4.3.

4.2.2 *Device manufacturers and device type effect on FTIR spectra*

In the next step, it was examined whether it is possible to differentiate between the various FTIR spectrometer manufacturers and device types based on the spectra. The results in Fig. 6 show the differentiation of manufacturers and device types used in this study.

Three distinct groups are evident: devices from the manufacturer 1, 2, and a combined group for 3 and 4. Group 2, however, exhibits partly higher scatter and some outliers.

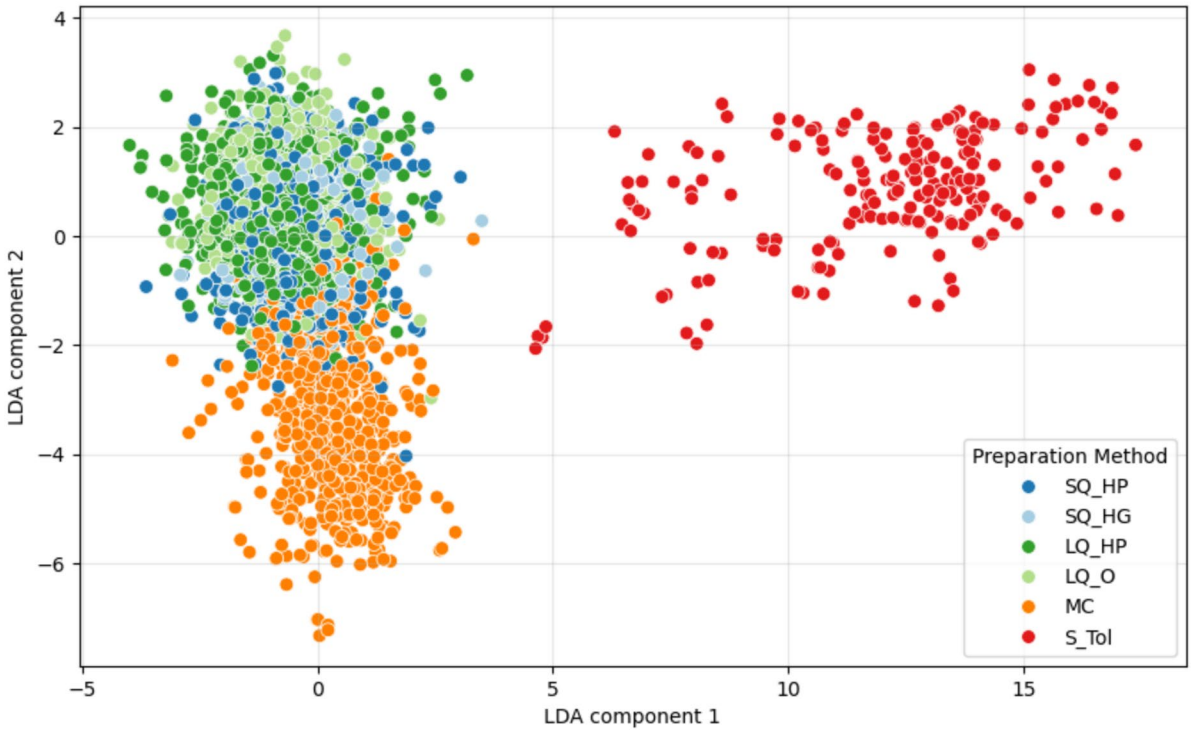

Fig. 5 LDA plot of the spectral data clustered by preparation method

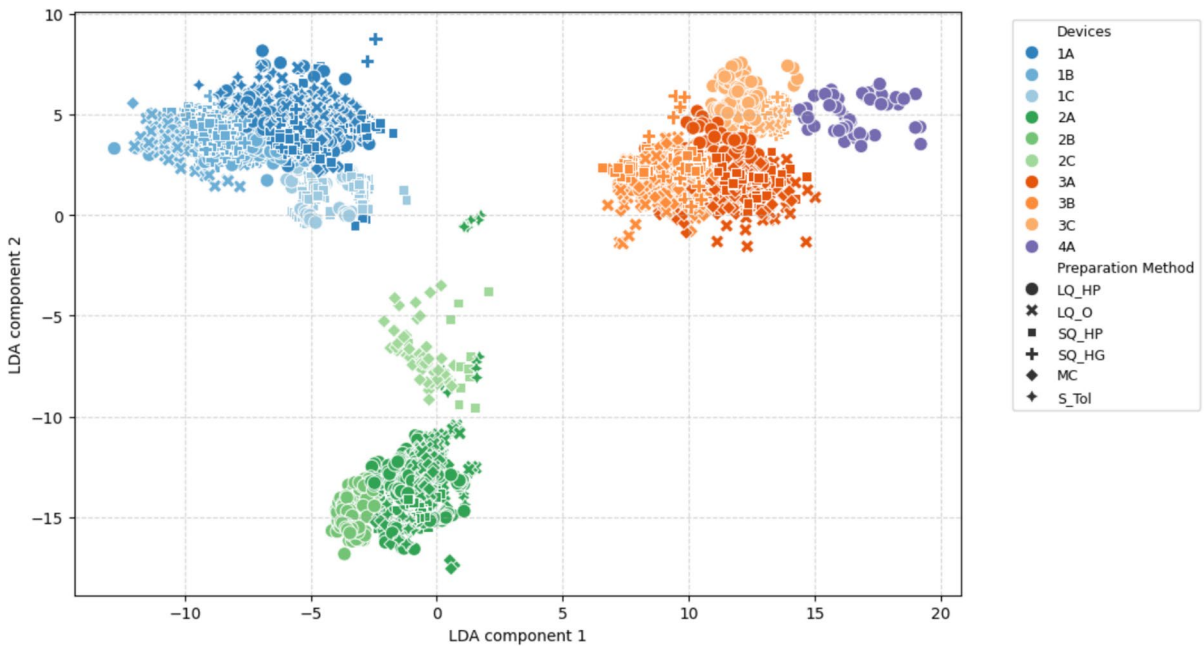

Fig. 6 LDA plot of manufacturer and device specific separation

For a more detailed analysis, a distinction was also made within the manufacturers according to the respective device type. Across all manufacturers, it is evident that there is no complete overlap between devices; rather, the device groups are positioned alongside each other. This suggests that devices from the same manufacturer produce spectra that are not identical but rather exhibit slight differences which can likely be attributed to variations in optical alignment, detector sensitivity, light source intensity, calibration routines, or gradual detector aging.

An interesting observation was made within group 2. Here, the measurements with device 2C show significant deviations compared to the other devices 2A and 2C. Further outliers are not attributable to specific devices, but rather to the preparation method used. These outliers correspond to dissolved samples, pointing to a potential issue with solvent evaporation.

The results of these analyses suggest that different devices, particularly those from different manufacturers, lead to varying spectra. This highlights the need for standardized calibration procedures or cross instrument normalization in future studies. Section 4.3 analyses these differences in more detail.

4.2.3 Effect of binder type on the FTIR-spectrum

Finally, it was assessed whether it is possible to identify differences between the bitumen samples or their provenance regardless of the preparation method and the device used. The results in the form of the projection of the first and second LDA components is shown in Fig. 7.

The results in the figure demonstrate that all preparation methods allow for a clear separation of the three binder samples and thus their provenance. Once again, the S_Tol method stands out, as it exhibits the poorest separation. This is evident in the higher variation within the bitumen groups (B1158, B1198, B1199), particularly for B1198A, and the smaller distances between the different bitumen groups. However, it should be noted that these results are based solely on data from two laboratories. For homogenized samples and those prepared via the MC method, no significant differences are observed.

For sample B1199, Fig. 7 suggests that further differentiation based on the aging state may be possible. Therefore, further analysis was conducted to determine whether differentiation by aging state could be achieved using various preparation methods. For that, bitumen from different refineries should be considered separately for this analysis [16]. Accordingly,

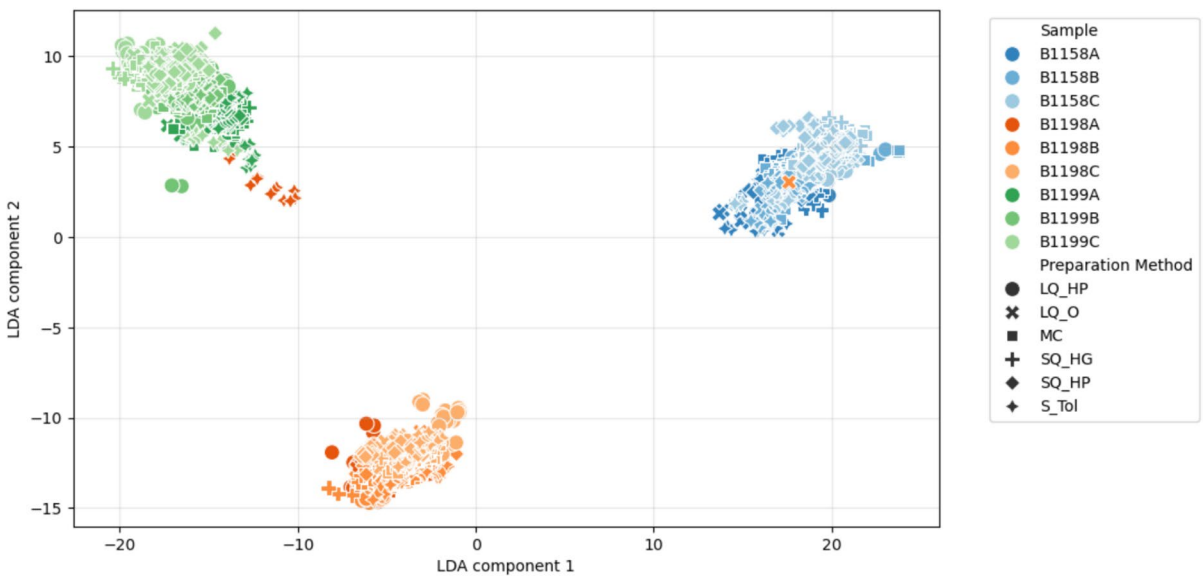

Fig. 7 LDA plot of different binder samples B1158, B1198 and B1199 in different ageing states (A – unaged, B – short-term aged, C – long-term aged)

PCA-LDA was applied to each bitumen sample (B1158, B1198, B1199) based on the preparation method, and the results in terms of the accuracy of the cross-validation (k-fold with k = 5) of each analysis are summarized in the heatmap in Fig. 8.

This heatmap shows that cross-validated accuracies of over 0.97 were achieved for almost all analyses, which represents a very good differentiation between the different ageing states of the binder samples under consideration. Excellent separation was found for the methods LQ_O, SQ_HG, and SQ_HP. The lowest separability was determined with the accuracy of 0.94 for the S_Tol method for sample B1198, which is presumably due to the influence of the solvent residues.

Concerning the samples, for B1199, it is possible to separate the samples based on aging state across all preparation methods. Furthermore, the long-term aged samples could be differentiated for all samples regardless of the preparation method. The analyses with values below 1.00 showed problems in differentiating between the unaged and the short-term aged samples.

The results indicate that all preparation methods enable differentiation of binder samples based on their provenance. An exception may be the S_Tol method, whereby only two laboratories applied this method in the round robin test. Furthermore, all methods allow for the clear distinction of the long-term aged state. However, challenges arise for certain methods in differentiating between the short-term aged and unaged states of the samples (LQ_HP, MC, S_Tol). This may be due either to the method itself or to influencing factors that have not yet been identified. Further investigations are necessary to address these challenges.

4.3 Feature importance with random Forest

After the differentiation of preparation methods, manufacturers and devices, as well as binder samples using PCA-LDA, the feature importance of the spectra was examined with Random Forest to identify which wavenumber regions are relevant for these distinctions.

4.3.1 Feature importance for different sample preparation methods

Figure 9 presents the feature importance of scores for the differentiation of the preparation methods. The previous PCA-LDA analysis indicated that the spectra from the S_Tol method deviates the most from the other methods, while the homogenized methods tend to result in comparable spectra. The MC method also exhibits differences from the homogenized methods, though these differences are less pronounced.

According to Fig. 9, the greatest differences between the preparation methods occur in the aliphatic regions (approx. 3000–2750 cm^{-1}, 1490–1400 cm^{-1}, 730–700 cm^{-1}) and the aromatic CH-bands of the spectra (900–700 cm^{-1}). For a more detailed comparison, Fig. 10 shows the spectra from Lab 5 for the preparation methods for sample B1158A.

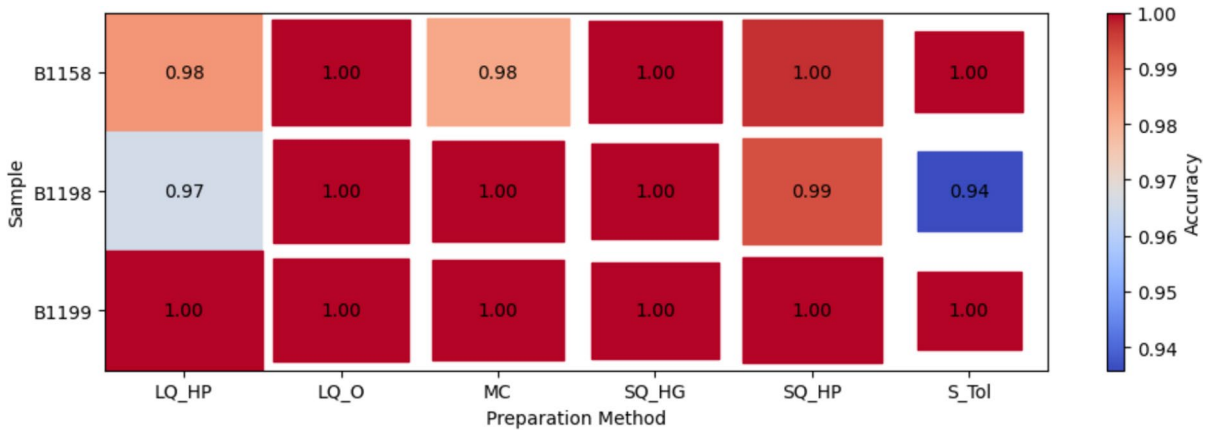

Fig. 8 The heatmap shows the PCA-LDA accuracies of the cross validation for each preparation method and sample, with red values indicating high accuracy and blue values indicating lower accuracy and thus lower differentiability

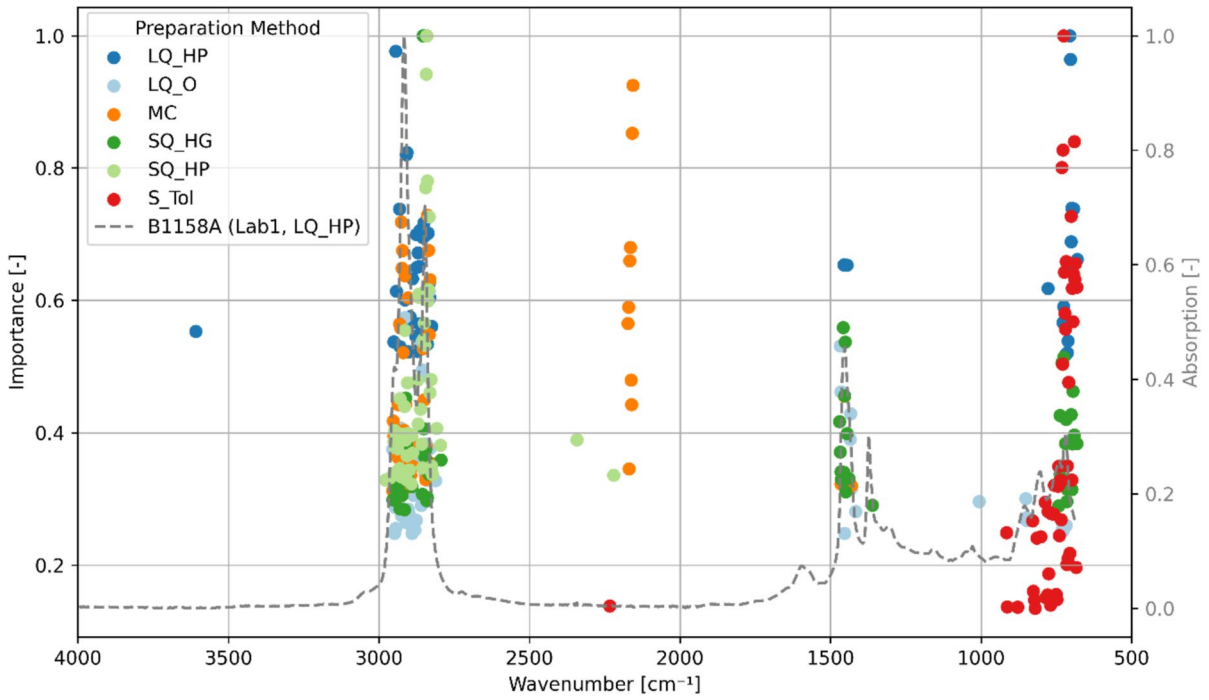

Fig. 9 Feature importance scores (Y-axis, left) obtained from the Random Forest classification model for the separation of the preparation methods plotted across wavenumber range (x-Axis) supplemented by the spectrum of the sample B1158A (Lab 1, LQ_HP)I for comparison

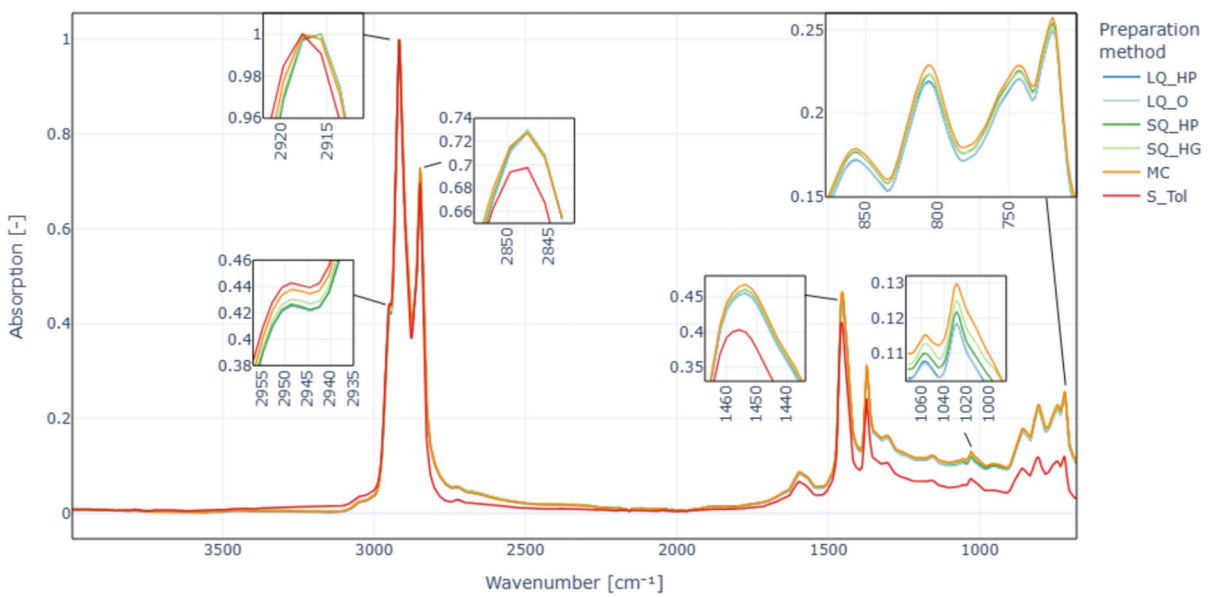

Fig. 10 Mean spectra of the bitumen sample B1158A from Lab 5 for the preparation methods with homogenization (LQ_HP, LQ_O, SQ_HG, SQ_HP), the metal can method (MC) and the method with samples dissolved in toluene (S_Tol) with enlargement of relevant wavenumber ranges

For the homogenized binder samples, it is evident that the spectra are nearly identical, which is consistent with the PCA-LDA results. For the MC method, slightly more intense bands can be seen for the aliphatic bands at approx. 2950 cm^{-1} and 1455 cm^{-1} and the aromatic bands between 890 and 700 cm^{-1}, as indicated by the feature importance scores. Furthermore, the MC method exhibits slightly higher intensities in the fingerprint region compared to spectra from homogenized methods. This observation was made for most of the participating laboratories and samples. For some samples and laboratories, higher intensities of sulfoxide bands were also measured with the MC method, which were not present in spectra from homogenized preparations. In the case of a very strong intensity, it can be assumed that the sample was not taken from the bulk, but too close to the surface, so that ageing effects of the surface were recorded. Therefore, it is crucial, when using the MC method, to scrape off the top millimeters of the binder surface and to take the sample from the bulk material.

In contrast, the method S_Tol shows noticeably lower intensities compared to the other methods (see Fig. 10). Not only are the intensities lower, but the ratios of the intensities of the different bands also appear to change. However, this result should be verified with a larger number of laboratories. Only two laboratories conducted these analyses, leading to significant scatter, but suggesting the presence of additional influencing factors. These should be identified and needed further considered in the measurement protocol. Further, there is a suspicion that interactions between the solvent and the bitumen could lead to changes in the spectra. However, further studies on dissolved samples show that comprehensive information on the respective bitumen samples can also be obtained from the dissolved spectra [16, 42].

In summary, it can be concluded that all the considered preparation methods lead to spectra exhibiting the same bands. However, the intensities of these bands differ depending on the method used. Methods involving prior homogenization (LQ_HP, LQ_O, SQ_HG, SQ_HP) produce very similar spectra, while cold application methods (MC) show slight deviations, particularly in the fingerprint region, with generally higher intensities. Spectra from the S_Tol method are characterized by noticeably lower intensities compared to those from other preparation methods, whereby these findings should be further verified by increasing the number of laboratories involved.

4.3.2 Feature importance for different devices

In the next step, the relevant wavenumber ranges were identified, which are critical for distinguishing between device manufacturers and types. According to Fig. 11, the important scores are primarily located in the aliphatic regions of the spectra. For devices of manufacturer 2 and 3, in addition, the range between approx. 4000–3000 cm^{-1} as well as around 1600 cm^{-1} appears to be relevant. For devices of manufacturer 4, making clear statements is difficult, as only one device was present in the study.

For a more in-depth interpretation of the results, Fig. 12 shows the spectra for the various devices, grouped by manufacturers.

This figure shows that spectra from different device types within a manufacturer already exhibit differences. However, it should be considered that these devices are located in different laboratories and operated by different personnel. Therefore, the variance observed between the individual spectra is not solely attributed to the device differences but also includes the influence of the operators. Nevertheless, significant differences are visible, some of which exceed the variances caused by the preparation method. For devices of manufacturer 1, deviations are evident in both the higher wavenumber region between 3100 and 2100 cm^{-1} and in the fingerprint region. These deviations mainly occur between device types, while measurements from the same device type are in good agreement. Here, an exception must be made for the type 1C, because this device was only present once and thus no conclusions can be drawn about the comparability of this device.

For devices of manufacturer 2, deviations are most prominent in the fingerprint region between approximately 1350 and 900 cm^{-1}, and these differences also appear within the same device type. However, for device 2B, no differentiated statements can be made, as it was only represented once.

Spectra from devices of manufacturer 3 show deviations across nearly the entire wavenumber range between and within device types, whereby the devices 3A and 3C being represented only once each.

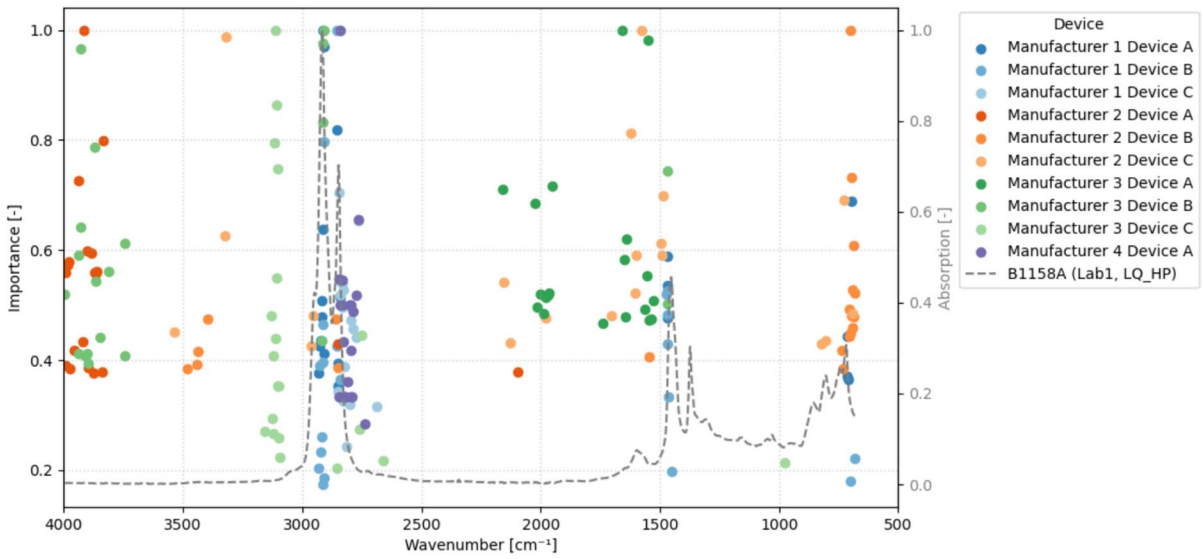

Fig. 11 Feature importance scores (Y-axis, left) obtained from the Random Forest classification model for the separation of the FTIR device type plotted across wavenumber range (x-Axis) supplemented by the spectrum of the sample B1158A (Lab 1, LQ_HP)I for comparison

For manufacturer 4, only one device was included in the round robin test, making it impossible to compare its spectra.

The scattering within the manufacturer groups also complicates the comparison between manufacturers. To simplify the comparison, mean spectra for each manufacturer group were calculated and compared. Figure 13 presents these results for sample B1158A, and comparable findings were obtained for all samples.

According to this comparison, it becomes evident that the spectra from devices of manufacturers 1 and 2 exhibit the greatest similarity on average, while spectra from devices from manufacturers 3 and 4 tend to show more intense spectra over larger wavenumber ranges.

According to the importance scores, the relevant differences are mainly observed in the aliphatic band range between approximately 3000 and 2600 cm^{-1}, around 1450 cm^{-1}, around 1600 cm^{-1} and at around 720 cm^{-1}. A detailed examination of these wavenumber ranges does indeed reveal differences between the spectra, as slight shifts in these bands become apparent, illustrated for some of the wavenumber ranges in the enlargements of Fig. 13. Slight shifts in wavenumbers can also be observed in other bands, although these appear less relevant for distinguishing between the devices. This is likely due to the fact that the spectra in these other regions exhibit stronger deviations depending on other factors, such as the preparation method.

In conclusion, both the device manufacturer and the device type have an impact on the measured spectra. Although devices from the same manufacturer produce similar spectra, these can also differ considerably. However, the differences between the various manufacturers are even more significant, with the observed shift in the bands being particularly conspicuous/noticeable.

4.3.3 Feature importance for different binder types

Finally, the relevant spectral differences for differentiating the binder samples are considered. The relevant features for this differentiation are shown in Fig. 14.

For this differentiation, oxygen-containing compounds are especially important, which are reflected in the OH (approx. 3200 cm^{-1}), carbonyl (approx. 1700 cm^{-1}), and sulfoxide group (approx. 1030 cm^{-1}). These bands are known to increase due to aging and the associated oxygen uptake [5, 43–45]. These effects of aging are exemplified for sample B1158 in Fig. 15.

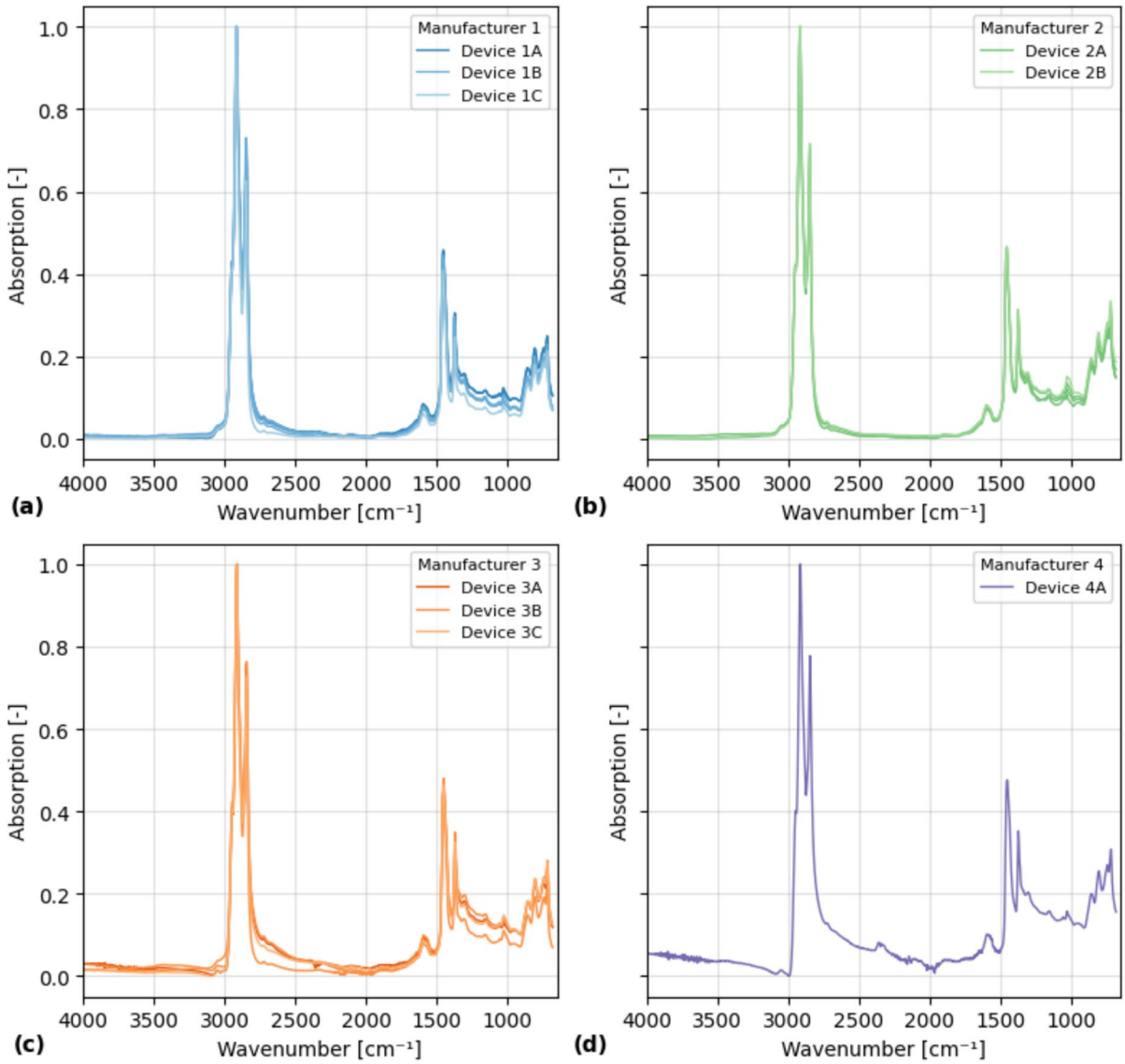

Fig. 12 Mean spectra of the sample B1158A (LQ_HP) from different devices for the manufacturers **a** 1, **b** 2, **c** 3 and **d** 4

In addition to the oxygen-containing compounds, various aliphatic structures (approx. 3000–2750 cm^{-1}, 1500–1340 cm^{-1}) and aromatic compounds (approx. 1620–1550 cm^{-1}, 900–770 cm^{-1}) also play a significant role in differentiating the different binders. This relevance becomes clear when comparing binders from different provenances. Figure 16 demonstrates this for the unaged samples of the respective binders. In other work by RILEM TC 295-FBB TG1 these differences between the binders have also been demonstrated.

From this figure, it is evident that the spectra of the three binders considered differ significantly, suggesting distinct chemical structures among the samples. The differences are most apparent in the aliphatic structures (approx. 3000–2750 cm^{-1}, 1500–1340 cm^{-1}, 722 cm^{-1}) and the aromatic compounds (approx. 1620–1550 cm^{-1}, 900–730 cm^{-1}, 700 cm^{-1}). These variations in aliphatic and aromatic profiles are typical indicators of the specific crude oil origin and refining processes, such as distillation type or crude acidity.

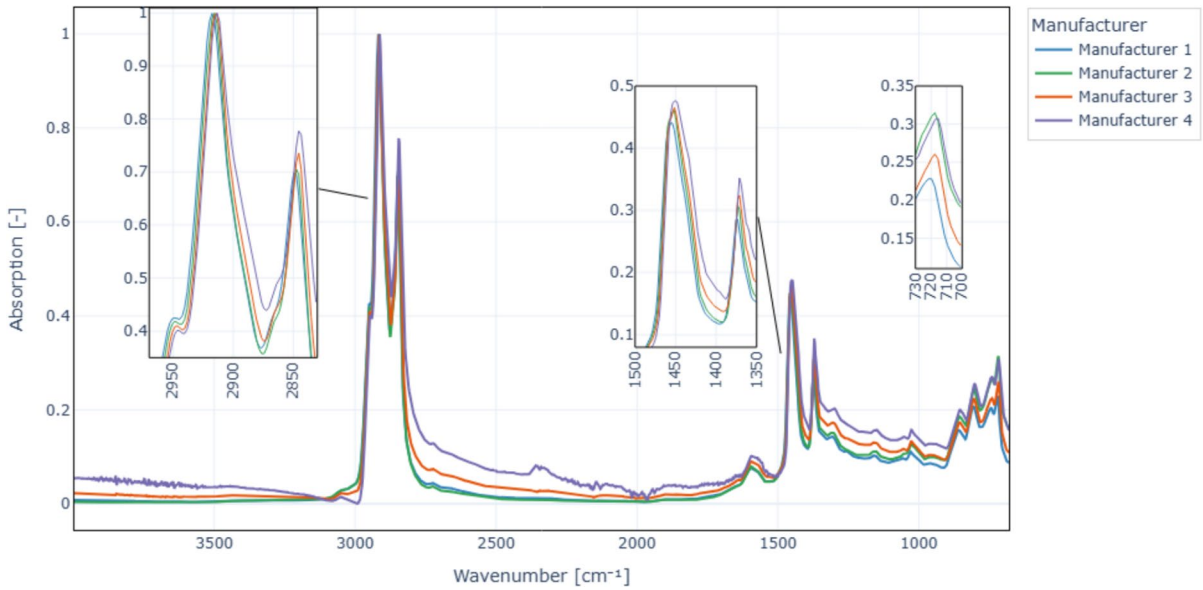

Fig. 13 Mean spectra of the different devices of each manufacturer for sample B1158A (LQ_HP) with enlargement of the areas 3000–2800 cm^{-1}, 1500–1300 cm^{-1} and 730–700 cm^{-1} for the comparison of the mean spectra of the different devices of each manufacturer

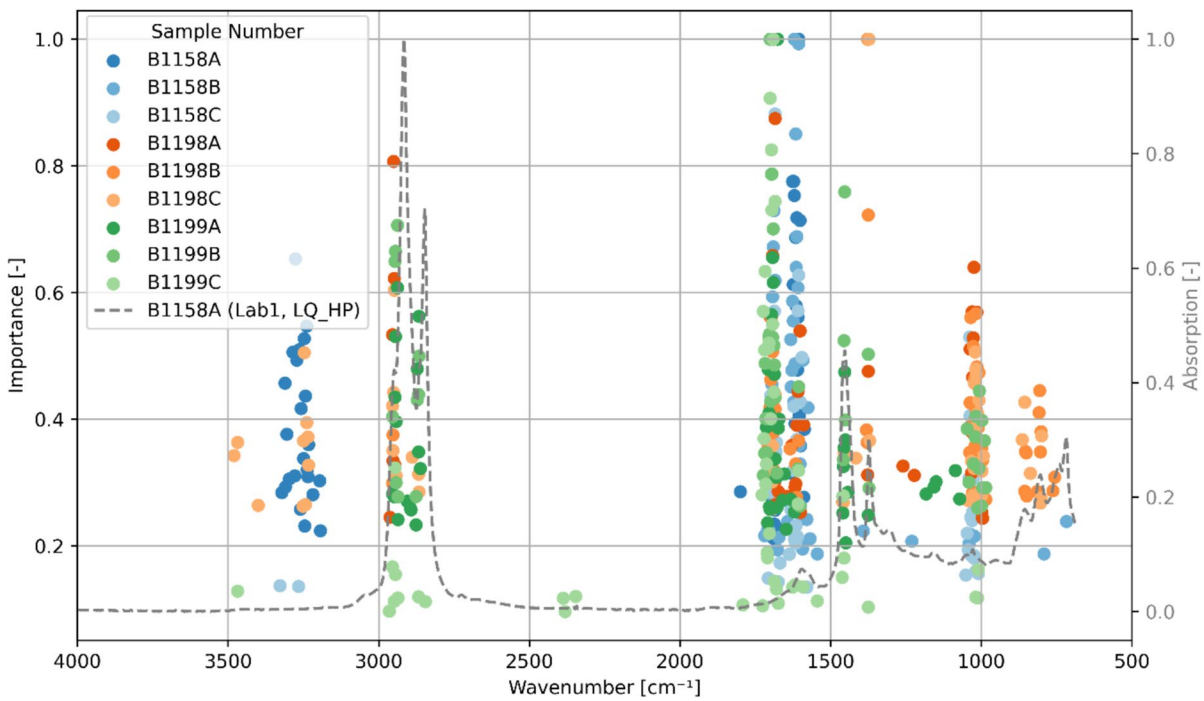

Fig. 14 Feature importance scores (Y-axis, left) obtained from the Random Forest classification model for the separation of the different binder samples plotted across wavenumber range (x-Axis) supplemented by the spectrum of the sample B1158A (Lab 1, LQ_HP)I for comparison

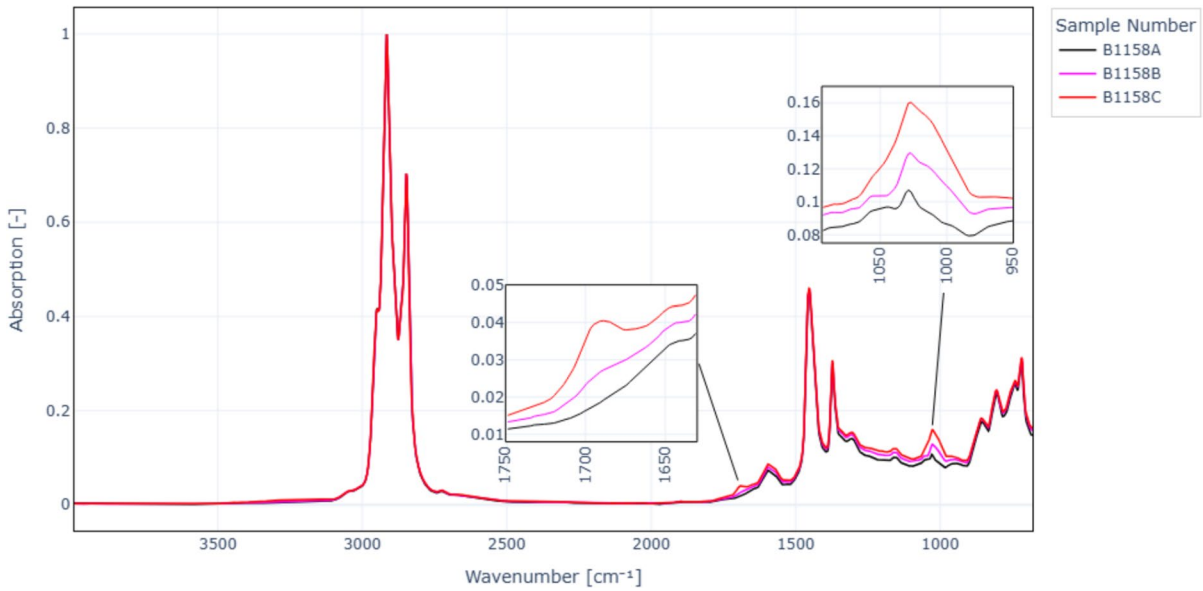

Fig. 15 Mean spectra of the sample B1158 (Lab 1, LQ_HP) in the unaged **A**, short-term aged **B** and long-term aged **C** state

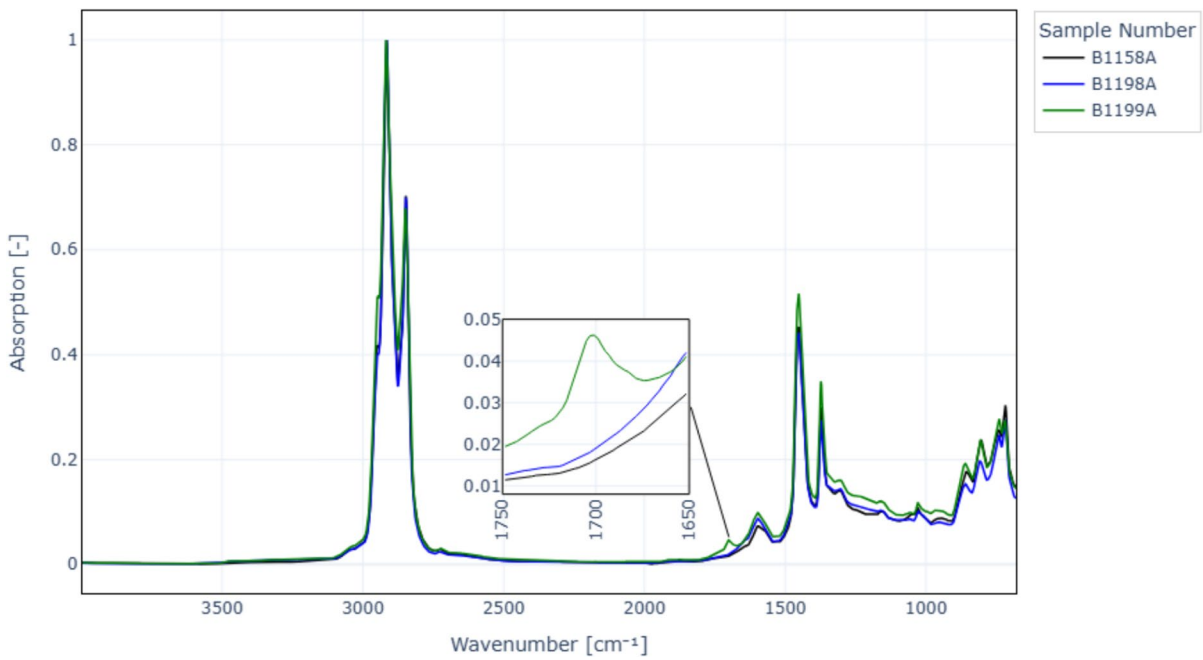

Fig. 16 Mean Spectra of the sample B1158A, B1198A, and B1199A (Lab 1, LQ_HP) from different provenience

Another significant difference is observed for the bitumen sample B1199, which shows a pronounced carbonyl band even in its unaged state compared to the other two samples. These differences are consistent across all investigated preparation methods and laboratories. Deviations were only noted for the S_Tol method, although other publications have also successfully differentiated various samples using dissolved samples [16]. Therefore, these findings should be further verified.

To summarize, it can be noted that the differentiation of the binder samples is based on different wavenumber ranges. To differentiate the ageing condition, the oxygen-containing compounds are particularly important, while differentiation according to provenance is based primarily on the aliphatic and aromatic structures. This finding is consistent with previous research [16].

5 Summary and conclusion

The results from this round robin test within the RILEM TC 295-FBB TG1 indicate that both the sample preparation methods and the manufacturers and device types influence the measured FTIR spectra. Despite the diversity of methods and instruments, spectra with consistent bands were obtained across all combinations, enabling reliable qualitative analysis. The primary differences were found in the intensity of bands, which can significantly affect (semi-)quantitative evaluations and minor wavenumber shifts, particularly between different device manufacturers.

Preparation methods emerged as a major source of variability. In terms of these, binders that underwent homogenization, regardless of the specific method used (LQ_HP, LQ_O, SQ_HP, SQ_HG), produced the most comparable spectra across all laboratories. These methods also showed stable reproducibility and, in most cases, strong discriminatory power. Cold application methods (MC), while convenient, resulted in smaller deviations in intensity and, in some cases, alterations in the sulfoxide bands. This observation shows that it is critical that for cold applications, a representative sample is taken from the bulk material to avoid capturing aging effects from the surface of the sample. The strongest deviations in the FTIR spectra were found for the measurements on the samples dissolved in toluene (S_Tol), whereby the deviations could be seen both in a significantly lower intensity and in deviating ratios of the band intensities. Due to the additional steps involved in this preparation method, there are presumably even more influencing factors to consider that can affect the measurement. Additionally, solvent-based methods may introduce artefacts or instability into the spectra. As previous studies have also achieved meaningful results with dissolved samples, the findings from this round robin test should be verified further due to the small number of laboratories involved using the S_Tol method.

Beyond these observations, the preparation methods differed not only in the mean spectra but also in their information value. Some preparation methods, such as SQ_HG and LQ_O, were particularly effective in distinguishing different bitumen samples. While satisfactory results were also observed with LQ_HP and MC, the weakest differentiation was seen with the SQ_HP and S_Tol methods.

Regarding the impact of device types, FTIR devices from different manufacturers resulted in partly significant spectral differences, which was mainly reflected in a wavenumber shift of the spectra. When comparing devices from the same manufacturer, the spectra were more similar, but nevertheless differences became visible. Despite these differences, spectra from all devices allowed for clear differentiation of the three binder samples from different provenances. Additionally, the long-term aging state of the samples was distinguishable across all devices.

Lessons learned from this systematic analysis highlight that strict homogenization protocols and the use of multivariate techniques are key to overcoming the inherent variability of inter-laboratory FTIR data. However, some limitations should be acknowledged. The study focused on three unmodified 70/100 grade binders, and the limited database for the solvent-based method restricts the generalizability of those specific results. Additionally, certain methods showed challenges in distinguishing between unaged and short-term aged states, suggesting that further research is needed to identify all influencing factors.

Future perspectives should focus on the development of standardized cross-instrument calibration protocols to harmonize spectral shifts across different manufacturers. Furthermore, expanding this multivariate framework to polymer-modified binders and recycled materials will be essential to further validate the robustness and broad applicability of the proposed chemical fingerprinting approach.

Overall, these findings highlight that both preparation and instruments influence FTIR results in a measurable way. However, the study also has indicated that under good harmonization, accurate and consistent differentiation by binder type and ageing condition can be achieved across different laboratory environments. For preparation, homogenization methods should be prioritized, with clear guidance on sample handling for cold

application and solvent-based approaches. For instrumentation, cross-laboratory calibration and verification routines could help alleviate device-related shifts and intensity differences.

Acknowledgements The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. Furthermore, the authors would also like to express their gratitude to the CD laboratories company partners BMI Group, OMV Downstream and Pittel+Brausewetter for their financial support. Furthermore, the authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Funding Open access funding provided by TU Wien (TUW). Funding provided by Christian Doppler Forschungsgesellschaft, 1836999, Bernhard Hofko.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors used ChatGPT to enhance the readability and language of the manuscript. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final version of the article.

Appendix A: Mathematical Foundations andStatistical Methods

This appendix contains the detailed mathematical definitions and supplementary background information for the statistical methods and machine learning algorithms used in the main manuscript.

Appendix A1: Data pre-processing

Min–max normalization To scale the spectral intensities to a common range of 0 to 1, Min–Max normalization was applied. The mathematical transformation is given by the equation:

$$x'_{i,j} = \frac{x_{i,j} - \min(X_i^{2800-3200})}{\max\left(X_i^{2800-3200}\right) - \min(X_i^{2800-3200})} \tag{1}$$

here $x_{i,j}$′ is the absorbance at wavenumber j for sample i, and $x_i^{2800-3200}$ represents the spectral values of sample i within the selected window.

Outlier filtering (Z-score) Outlier detection was performed column-wise across the wavenumber dimension:

$$Z_i = \frac{x_i - \bar{x}}{s} \tag{2}$$

where x_i spectral intensity at a given wavenumber, $\bar{x}$ the mean and s the standard deviation within the respective group. A Z-score exceeding a threshold of 3 in at least one variable led to the exclusion of the spectrum.

Appendix A2: Reproducibility metrics

Coefficient of variation (CV) The CV serves as a normalized measure of dispersion for spectra with different magnitude levels:

$$CV = \frac{\sigma}{\mu} \tag{3}$$

where σ is the standard deviation and μ is the mean of spectral intensity values across replicates for a given wavenumber.

Intraclass correlation coefficient The ICC evaluates the proportion of the total variance attributed to differences between the groups. It was calculated using a two-way random effects model with absolute agreement:

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}} \tag{4}$$

where $\sigma^2_{between}$ is the variance between groups and σ^2_{within} is the variance within groups.

Appendix A3: PCA-LDA classification

Principal component analysis (PCA) PCA takes the centered data matrix X and decomposes it into:

$$X = TP^T + E \tag{5}$$

where T is the score matrix (projections of the samples onto the principal components), P is the loading matrix (directions of maximum variance) and E is the residual matrix.

Linear discriminant analysis (LDA) LDA maximizes the separability of classes by solving the generalized eigenvalue problem:

$$S_B * w = \lambda S_W * w \tag{6}$$

where S_B is the between-class scatter matrix, S_W is the within-class matrix, w is the linear discriminant vector and λ is the corresponding eigenvalue indicating the achieved class separation.

Cross-validation The classification accuracy was determined via a k-fold cross-validation:

$$Acc_{XVal} = \frac{1}{k}\sum_{i=1}^{k}\frac{correct\ predictions\ in\ fold\ i}{total\ samples\ in\ fold\ i} \tag{7}$$

Appendix A4: Random forest

A Random Forest is an ensemble algorithm that constructs a multitude of decision trees. Formally, given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^p$ are input features (absorbance values at different wavenumbers) and y_i are categorical class labels (e.g., preparation method), a random forest consists of T decision trees $h_t(x)$, each trained on a bootstrap subset of $\mathcal{D}$. The final prediction for a sample x is given by:

$$\hat{y} = mode[h_1(x), h_2(x), \ldots, h_T(x)] \tag{8}$$

For each tree a subset $m \ll p$ of features is considered at each node, and the split is chosen to maximize a splitting criterion such as the Gini impurity (G):

$$G = \sum_{k=1}^{K} p_k(1 - p_k) \tag{9}$$

where p_k is the proportion of samples of class k at a given node. The feature contributing most to impurity reduction over all trees can be interpreted as important (see also chapter 3.4.6).

Appendix A5: Feature importance

The importance of a feature f, tis calculated by the average reduction of the node impurity across the entire ensemble:

$$I(f) = \frac{1}{T} * \sum_{t=1}^{T}\sum_{n \in nodes(t):\, f_{n=f}} \Delta i(n) \tag{10}$$

where T is the number of trees, n are the nodes where feature f is used, and $\Delta i(n)$ is the impurity decrease at node n.

References

1. Lesueur D (2009) The colloidal structure of bitumen: consequences on the rheology and on the mechanisms of bitumen modification. Adv Coll Interface Sci 145(1):42–82
2. Behnood A (2019) Application of rejuvenators to improve the rheological and mechanical properties of asphalt binders and mixtures: a review. J Clean Prod 231:171–182
3. Ma L et al (2023) Chemical characterisation of bitumen type and ageing state based on FTIR spectroscopy and discriminant analysis integrated with variable selection methods. Road Mater Pavement Des 24(sup1):506–520
4. Primerano K et al (2022) Influence of selected reactive oxygen species on the long-term aging of bitumen. Mater Struct 55(5):133
5. Mirwald J et al (2020) Investigating bitumen long-term-ageing in the laboratory by spectroscopic analysis of the SARA fractions. Constr Build Mater 258:119577
6. Mirwald J, Nura D, Hofko B (2022) Recommendations for handling bitumen prior to FTIR spectroscopy. Mater Struct 55(2):26
7. Weigel S, Stephan D (2017) The prediction of bitumen properties based on FTIR and multivariate analysis methods. Fuel 208:655–661
8. Lamontagne J et al (2001) Comparison by Fourier transform infrared (FTIR) spectroscopy of different ageing techniques: application to road bitumens. Fuel 80(4):483–488

9. Hofko B et al (2018) FTIR spectral analysis of bituminous binders: Reproducibility and impact of ageing temperature. Mater Struct 51:1–16
10. Mirwald J et al (2025) Evaluating the reproducibility and consistency of different sample preparation techniques used for ATR-FTIR spectroscopy from the RILEM 295-FBB TG1 round robin test. Mater Struct 58(8):255
11. Khalighi S et al (2024) Evaluating the impact of data pre-processing methods on classification of ATR-FTIR spectra of bituminous binders. Fuel 376:132701
12. Lee LC, Liong C-Y, Jemain AA (2017) A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum. Chemom Intell Lab Syst 163:64–75
13. Motevalizadeh SM, Mollenhauer K (2024) Exploration of chemical changes in bituminous mastics induced by aging: insights from FTIR spectroscopy, DSR measurements, and machine learning. Int J Pavement Eng 25(1):2418927
14. Motevalizadeh SM, Mollenhauer K, Wetekam J (2024) FTIR spectroscopy and multivariate discriminant analysis for classifying bituminous mastics: exploring aging states and mastic composition. Constr Build Mater 438:137188
15. Primerano K et al (2023) Characterization of long-term aged bitumen with FTIR spectroscopy and multivariate analysis methods. Constr Build Mater 409:133956
16. Weigel S, Stephan D (2018) Differentiation of bitumen according to the refinery and ageing state based on FTIR spectroscopy and multivariate analysis methods. Mater Struct 51(5):130
17. Weigel S et al (2021) Identification and quantification of additives in bituminous binders based on FTIR spectroscopy and multivariate analysis methods. Mater Struct 54(4):171
18. Weigel S, Wetekam J, Mollenhauer K (2023) Identification and classification of PAH in asphalt binders with FTIR spectroscopy and multivariate analysis methods. Fuel 337:126845
19. Motevalizadeh M, Mollenhauer K, Wetekam J (2025) Characterisation of bitumen grade and modification using a multilevel multivariate discriminant analysis approach and FTIR spectroscopy. Road Mater Pavement Des 26(sup1):214–234
20. Khalighi S et al (2024) Multivariate chemo-rheological framework for optimizing laboratory aging protocols of paving binders. Mater Des 248:113520
21. Khalighi S, Erkens S, Varveri A (2025) Exploring the impact of humidity and water on bituminous binder aging: a multivariate analysis approach (TI CAB). Road Mater Pavement Des 26(4):753–777
22. Motevalizadeh SM, Mollenhauer K (2024) Use of multivariate clustering analysis to investigate the physicochemical interactions in bitumen mastics using micromechanical modeling and FTIR spectroscopy. Constr Build Mater 448:138230
23. Khalighi S et al (2025) Multi-scale analysis of ageing behaviour in bituminous materials. Road Mater Pavement Des 26(sup1):654–679
24. EN, C., 12607–1:(2014) Bitumen and Bituminous Binders—Determination of the Resistance to Hardening under Influence of Heat and Air—Part 1: RTFOT Method. European Committee for Standardization: Brussels, Belgium.
25. En, C., 14769: (2012) Bitumen and Bituminous Binders—Accelerated Long-Term Ageing Conditioning by a Pressure Ageing Vessel (PAV). European Committee for Standardization: Brussels, Belgium.
26. Petersen J (2009) A Review of the Fundamentals of Asphalt Oxidation (Transportation Research Circular No. E-C140). Transportation Research Board, Washington, DC, USA.
27. Škultecké J, Bitarytė S, Liubinas K (2023) FTIR-ATR spectroscopy to identify and quantify the SBS in modified bitumen.
28. Khalighi S et al. (2025) The impact of reactive oxygen species coupled with moisture on bitumen long-term aging. Road Materials and Pavement Design. p. 1–20.
29. Fearn T (2002) Assessing calibrations: sep, rpd, rer and r 2. NIR news 13(6):12–13
30. Camargo IGdN et al (2020) Effect of thermal and oxidative aging on asphalt binders rheology and chemical composition. Materials (Basel) 13(19):4438
31. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163
32. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52
33. Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. Anal Methods 5(16):3790–3798
34. Xanthopoulos P, et al. (2013) Linear discriminant analysis. Robust data mining, p. 27–33.
35. Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
36. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. in Ijcai. Montreal, Canada.
37. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
38. Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci 374(2065):20150202
39. Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Mining Knowl Discov 9(3):e1301
40. Strobl C et al (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinform 8(1):25
41. Gromski PS et al (2015) A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Anal Chim Acta 879:10–23
42. Weigel S (2023) Estimating aging sensitivity of bitumen using FTIR spectroscopy and multivariate evaluation. Eng Res Express 5:025067
43. Pipintakos G et al (2021) Exploring the oxidative mechanisms of bitumen after laboratory short- and long-term ageing. Constr Build Mater 289:123182

44. Lu X, Isacsson U (2002) Effect of ageing on bitumen chemistry and rheology. Constr Build Mater 16(1):15–22
45. Van den Bergh W (2011) The effect of ageing on the fatigue and healing properties of bituminous mortars.