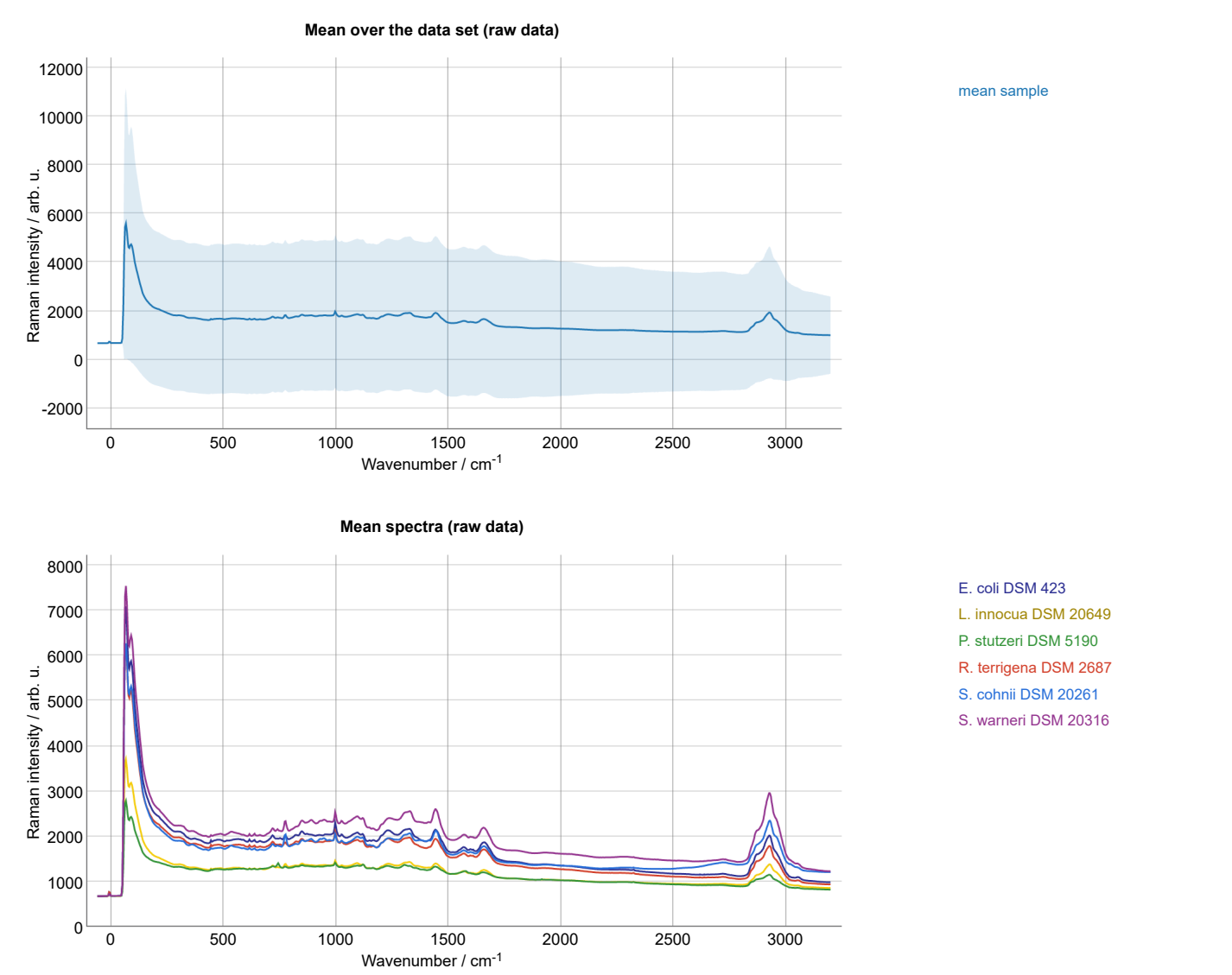


Mean sample spectra



Metadata

Batches (7)

- 1..... 1_Batch
- 2..... 2_Batch
- 3..... 3_Batch
- 4..... 4_Batch
- 5..... 5_Batch
- 6..... 6_Batch
- 7..... 7_Batch

Categorical responses (11)

- 1..... batch
- 2..... type
- 3..... filename
- 4..... date_device
- 5..... device
- 6..... genus
- 7..... path
- 8..... metadata_source
- 9..... strain
- 10..... species
- 11..... date

Dark spectra (0)

Dates (13)

1..... 2017-03-01 (Bacteria_Spectra_SSP)
2..... 2017-03-02 (Bacteria_Spectra_SSP)
3..... 2017-03-03 (Bacteria_Spectra_SSP)
4..... 2017-03-07 (Bacteria_Spectra_SSP)
5..... 2017-03-08 (Bacteria_Spectra_SSP)
6..... 2017-03-09 (Bacteria_Spectra_SSP)
7..... 2017-03-10 (Bacteria_Spectra_SSP)
8..... 2017-04-04 (Bacteria_Spectra_SSP)
9..... 2017-04-11 (Bacteria_Spectra_SSP)
10..... 2017-04-12 (Bacteria_Spectra_SSP)
11..... 2017-04-13 (Bacteria_Spectra_SSP)
12..... 2017-05-09 (Bacteria_Spectra_SSP)
13..... 2017-05-10 (Bacteria_Spectra_SSP)

Intensity standard (0)

Metadata tables (1)

1..... metadata_train_bacteria_species.csv (4404 spectra)

Number of spectra

1..... Total: 4404
2..... Sample: 4214
3..... Wavenumber standard: 190
4..... Intensity standard: 0
5..... Dark background: 0
6..... Reference: 0

Reference sample spectra (0)

Responses (0)

Types (6)

1..... E. coli DSM 423
2..... L. innocua DSM 20649
3..... P. stutzeri DSM 5190
4..... R. terrigena DSM 2687
5..... S. cohnii DSM 20261
6..... S. warneri DSM 20316

Wavenumber standard data available (13)

1..... 2017-03-01 (Bacteria_Spectra_SSP)
2..... 2017-03-02 (Bacteria_Spectra_SSP)
3..... 2017-03-03 (Bacteria_Spectra_SSP)
4..... 2017-03-07 (Bacteria_Spectra_SSP)
5..... 2017-03-08 (Bacteria_Spectra_SSP)
6..... 2017-03-09 (Bacteria_Spectra_SSP)
7..... 2017-03-10 (Bacteria_Spectra_SSP)
8..... 2017-04-04 (Bacteria_Spectra_SSP)
9..... 2017-04-11 (Bacteria_Spectra_SSP)
10..... 2017-04-12 (Bacteria_Spectra_SSP)
11..... 2017-04-13 (Bacteria_Spectra_SSP)
12..... 2017-05-09 (Bacteria_Spectra_SSP)
13..... 2017-05-10 (Bacteria_Spectra_SSP)

Wavenumber standard data not available (0)

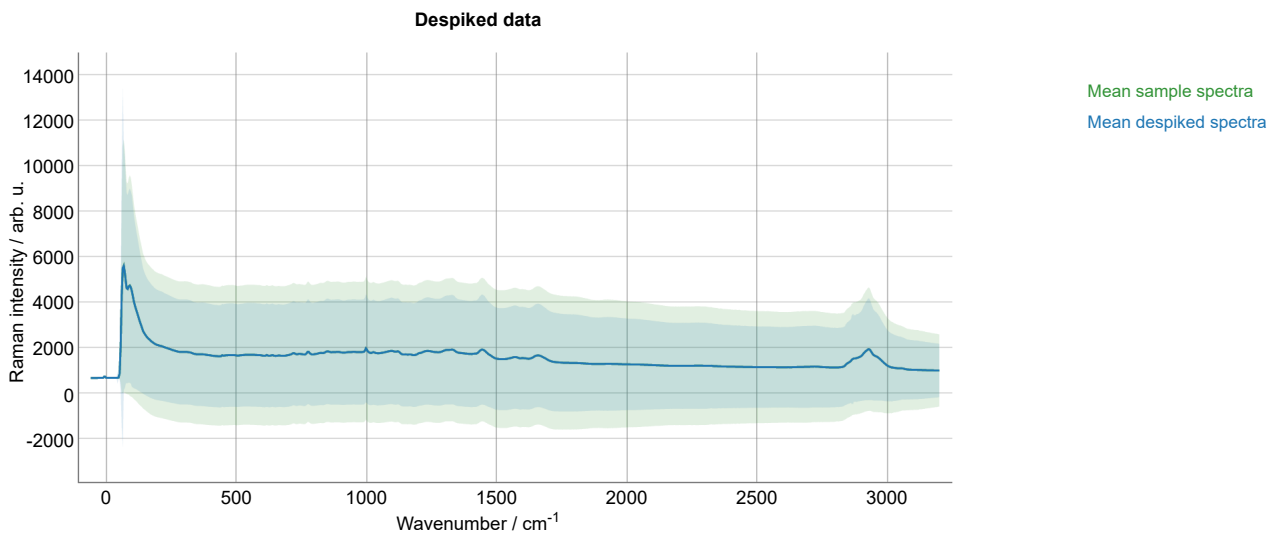
Data Pre-treatment: Despiking

Despiking approach

two_spec

Despiking threshold

10



Data Pre-treatment: Calibration

Peak fitting method

spline

Standard peaks

[213.3, 329.2, 390.9, 465.1, 504, 651.6, 710.8, 797.2, 857.9, 968.7, 1105.5, 1168.5, 1236.8, 1278.5, 1323.9, 1371.5, 1515.1, 1561.5, 1648.4, 2931.1, 3064.6, 3102.4]
cm⁻¹

Degree of polynomial

3

Resolution

3 cm⁻¹

Wavenumber area

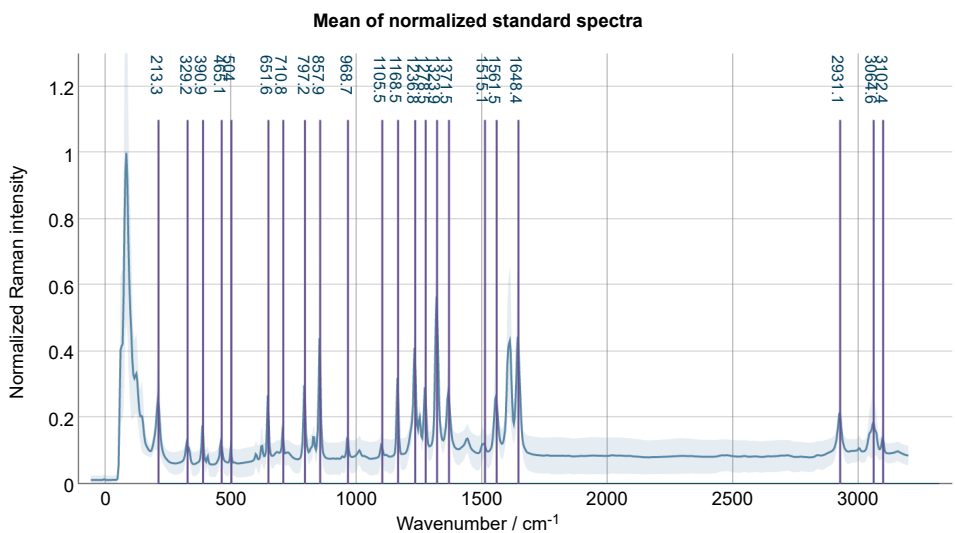
[200, 3100] cm⁻¹

Intensity calibration

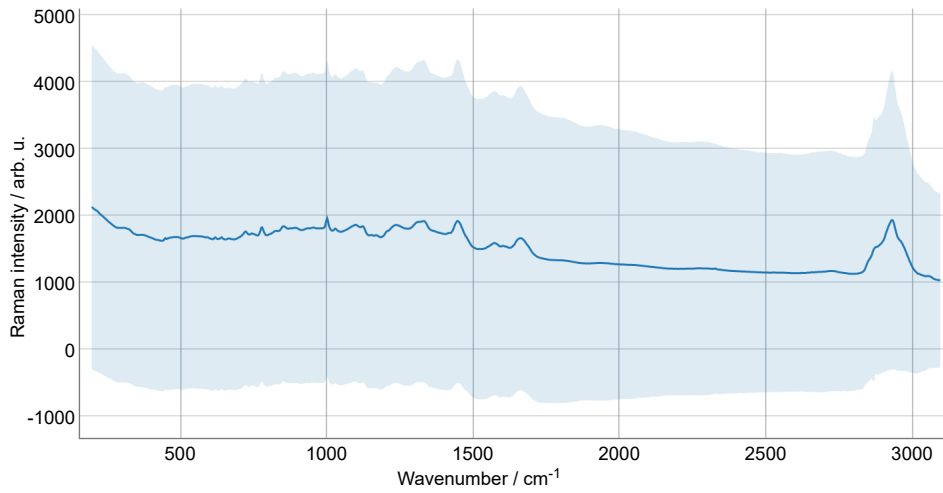
device

Dark background

device

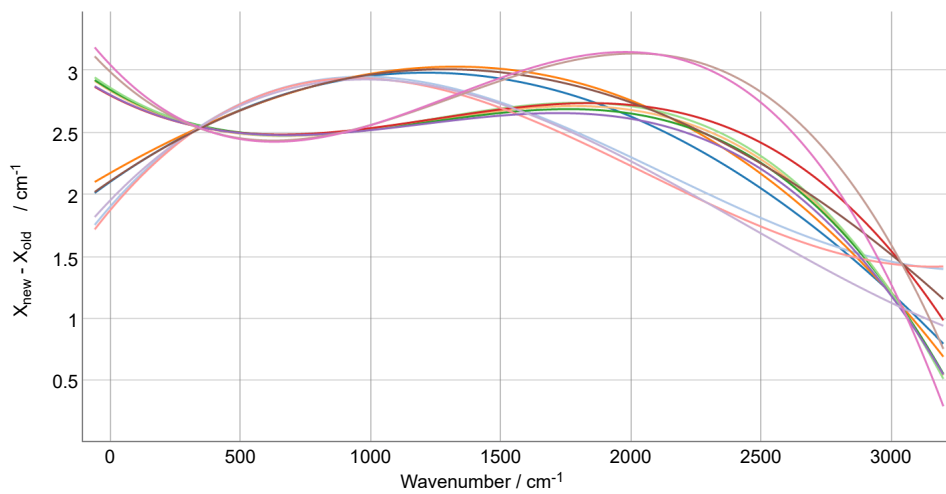


Calibrated data



Mean calibrated spectra

Difference between original and calibrated wavenumber axis



Bacteria_Spectra_SSP | 2017-03-01
Bacteria_Spectra_SSP | 2017-03-02
Bacteria_Spectra_SSP | 2017-03-03
Bacteria_Spectra_SSP | 2017-03-07
Bacteria_Spectra_SSP | 2017-03-08
Bacteria_Spectra_SSP | 2017-03-09
Bacteria_Spectra_SSP | 2017-03-10
Bacteria_Spectra_SSP | 2017-04-04
Bacteria_Spectra_SSP | 2017-04-11
Bacteria_Spectra_SSP | 2017-04-12
Bacteria_Spectra_SSP | 2017-04-13
Bacteria_Spectra_SSP | 2017-05-09
Bacteria_Spectra_SSP | 2017-05-10

Data Pre-processing: Background corrections

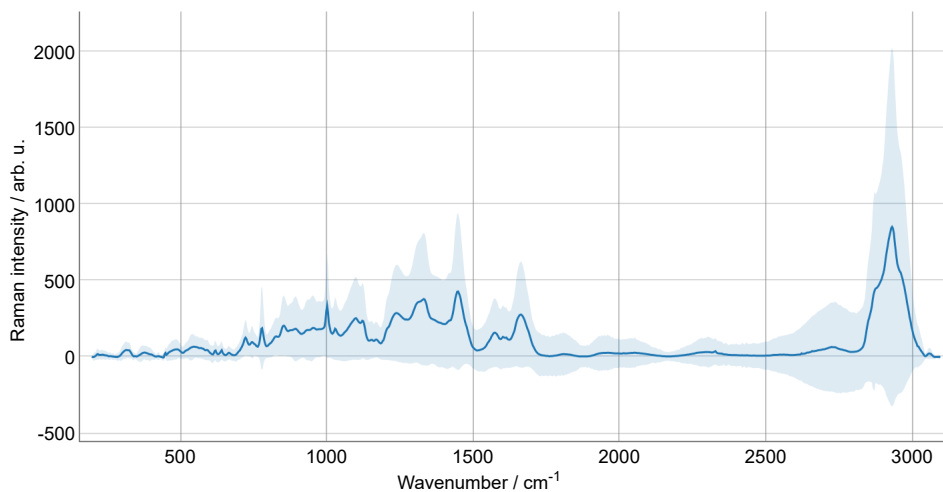
Baseline correction algorithm

snip

Baseline iteration

40

Baseline corrected



Background corrected spectra

Data Pre-processing: Normalization

Normalization type

vector

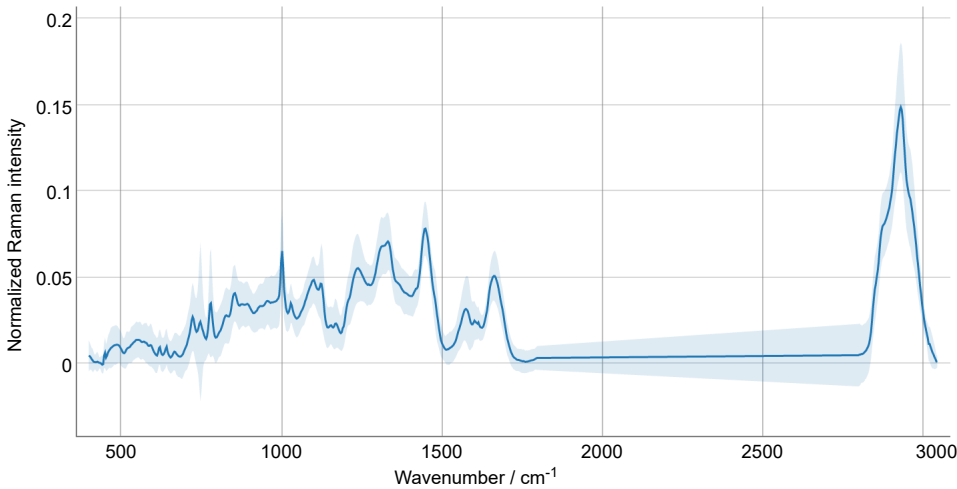
Used area

[400, 3050] cm⁻¹

Excluded area

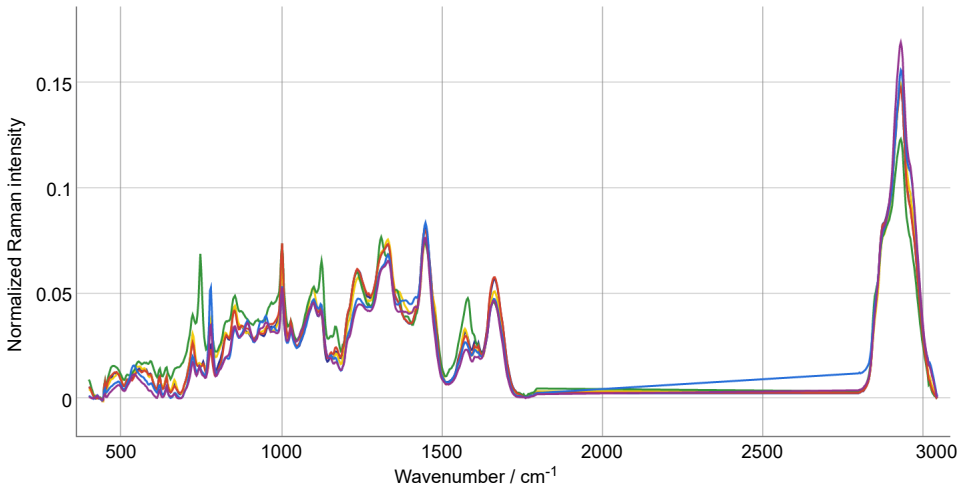
[1800, 2800] cm⁻¹

Mean normalized spectrum over the data set



Mean normalized spectra

Mean preprocessed spectra

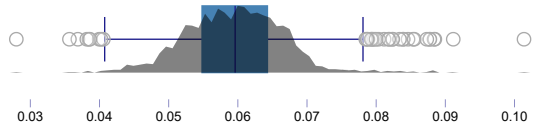


- E. coli DSM 423
- L. innocua DSM 20649
- P. stutzeri DSM 5190
- R. terrigena DSM 2687
- S. cohnii DSM 20261
- S. warneri DSM 20316

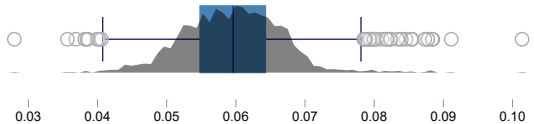
Quality filters

The result of applying filters: 1928 out of 2107 spectra (91.50%) passed the quality check

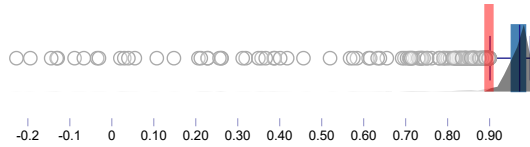
Min intensity of selected quality peak
Filter disabled



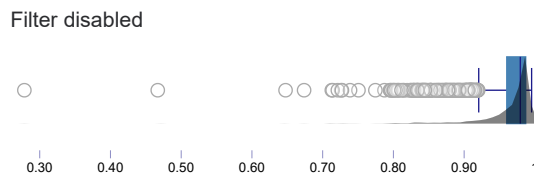
Max intensity of selected quality peak
Filter disabled



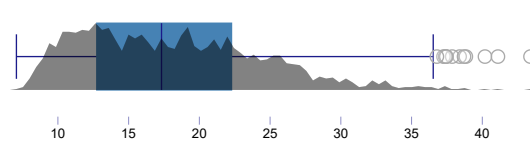
Correlation with mean (reference) calibrated spectrum
Threshold: 0.9



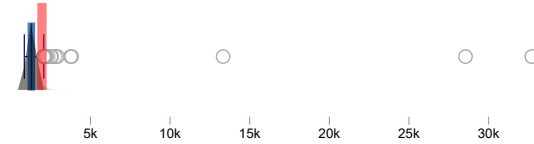
Correlation with mean (reference) preprocessed spectrum
Filter disabled



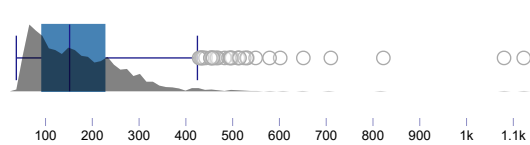
Minimal signal-to-noise ratio (SNR)
Filter disabled



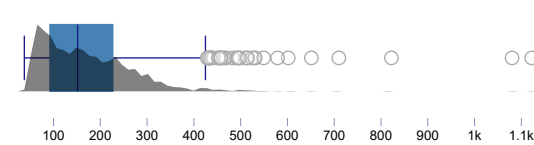
Max integrated background intensity
Threshold: 2000



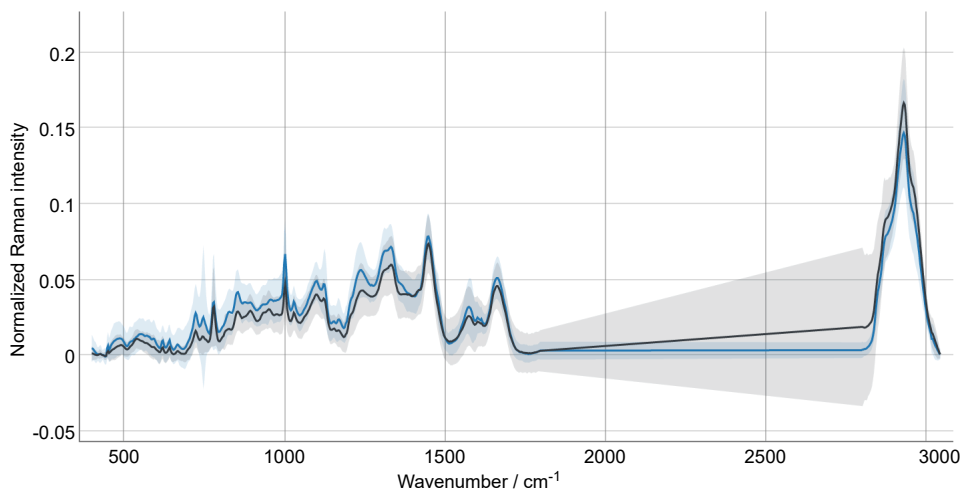
Min integrated intensity of baseline corrected spectrum
Filter disabled



Max integrated intensity of baseline corrected spectrum
Filter disabled



Included and excluded spectra after applying filters



Passed quality check
Failed quality check

Model

Model type

pcalda

Evaluation type

batch

Number of features

8

Included labels

Classes:

["type"]

Aggregate spectra

1

Accuracy: **0.777** Balanced accuracy: **0.782** Cohen's kappa: **0.732**

Optimal number of features: **8**

Number of features in the inner validation: **9, 8, 8, 7, 9, 8, 8**

Confusion table

True \ Predicted		[1]	[2]	[3]	[4]	[5]	[6]	sens	spec
E. coli DSM 423	[1]	220	4	1	111	0	0	0.655	0.909
L. innocua DSM 20649	[2]	4	302	0	1	1	41	0.865	0.962
P. stutzeri DSM 5190	[3]	19	0	323	0	0	0	0.944	0.999
R. terrigena DSM 2687	[4]	117	49	0	185	0	0	0.527	0.918
S. cohnii DSM 20261	[5]	5	7	0	11	207	39	0.77	0.992
S. warneri DSM 20316	[6]	0	0	0	6	13	262	0.932	0.951

