Supplement of Atmos. Meas. Tech., 9, 3429–3454, 2016
http://www.atmos-meas-tech.net/9/3429/2016/
doi:10.5194/amt-9-3429-2016-supplement

Atmospheric Measurement Techniques
Open Access

Supplement of

Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: sparse methods for statistical selection of relevant absorption bands

Satoshi Takahama et al.

Correspondence to: Satoshi Takahama (satoshi.takahama@epfl.ch)

Contents

S1 Example spectra

All example spectra used in this work have been shown previously (Ruthenburg et al., 2014; Dillner and Takahama, 2015; Takahama and Dillner, 2015). In Figure S1, we include one example from each sample type (blank, laboratory, and ambient) in their raw and baseline corrected forms for illustration.

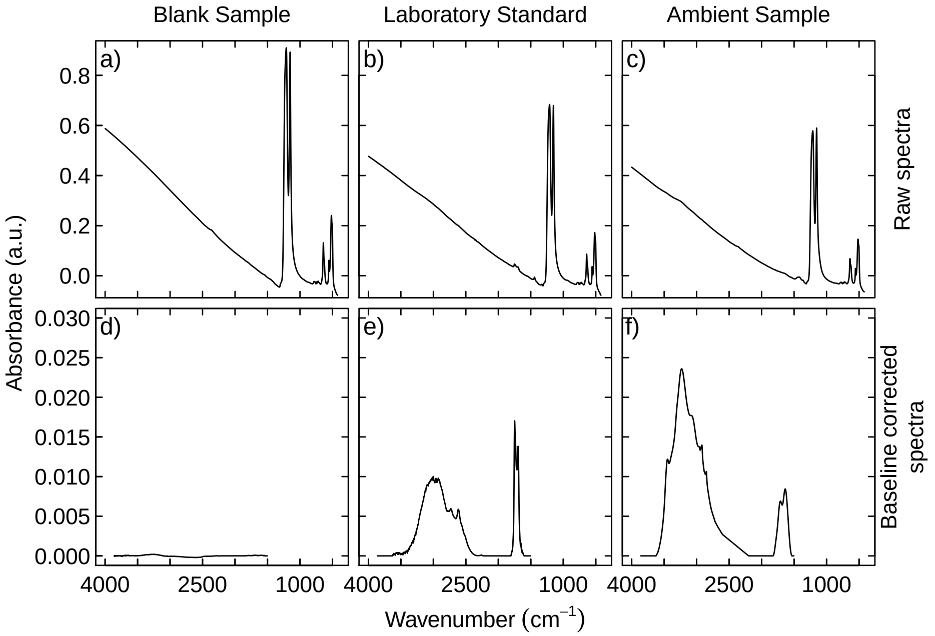

Figure S1. Top row: Examples of three types of raw spectra. Bottom row: Spectra corrected for PTFE contribution using polynomial fitting (wavenumbers below 1500 $\mathrm{cm}^{-1}$are not used in this spectra type).

S2 Model selection

As introduced in Section 2.3 of the main document, we consider several sparse models based on the RMSECV or penalized RMSECV. The main objective is to generate potential solutions that vary in number of NZVs and LVs using criteria based on the RMSECV and $\|\mathbf{b}\|$ values computed from the calibration set, and independently evaluate model performance on a test set that is not used during the calibration process. In this document, we describe the models considered (Section S2.1), their evaluation for FGs (Section S2.2) and TOR-equivalent values of OC and EC (Section S2.3), and a summary of selected models (Section S2.4).

S2.1 Models considered

Figure S2 qualitatively illustrates the concept of exploring solutions (and the calibration model that generated them) including the conventional minimum RMSECV solution, and those which seek additional parsimony with respect to the number of NZVs, LVs, or both. All candidate models are described in Table S1. We note that while reduction in NZVs is important to eliminating unnecessary wavenumbers for interpreting absorption bands, reduction in LVs is also important to reduce oscillations in adjacent regression coefficients (Gowen et al., 2011) to simplify our interpretation of how absorption bands contribute positively or negatively to the quantification of analytes (Section 3.3 and Figures 6 and 7 in main document).

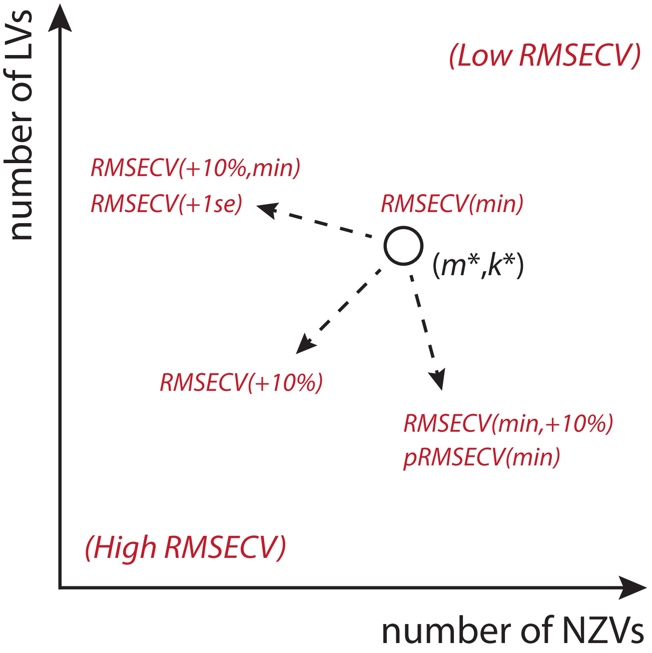

Figure S2. Illustration of solutions explored in the space of NZVs and LVs used by calibration models. The dotted arrows indicate general direction of parsimony targeted with respect to the minimum RMSECV solution (specified by m^* NZVs and k^* LVs) by various criteria. Arrows are approximately drawn and do not indicate strict positions of individual solutions. While the regions of low and high RMSECV denoted area also model-dependent, extreme reductions in NZVs and LVs toward zero will lead to unequivocal increase in RMSECV. Explanation of annotated elements (and additional criteria) are provided in Table S1.

Table S1. Description of models considered for each algorithm. “SPLS” includes both SPLSa and SPLSb.

Method	Solution	Description
SPLS	RMSECV(min)	minimum RMSECV solution
	RMSECV(min,+10%)	sparsity parameter fixed at minimum RMSECV value; number of LVs selected such that RMSECV is $\leq 10\%$ of minimum value
	RMSECV(+10%,min)	number of LVs selected according to minimum RMSECV for each sparsity parameter; sparsity parameter selected such that RMSECV is $\leq 10\%$ of minimum value
	pRMSECV(min)	sparsity parameter fixed at minimum RMSECV value; number of LVs selected as minimum pRMSECV solution
	RMSECV(+10%)	sparsity parameter and number of LVs selected such that RMSECV is $\leq 10\%$ of minimum value
EN	RMSECV(min)	minimum RMSECV solution
	RMSECV(+1se)	solution is selected such that RMSECV is $\leq$ one standard error of minimum value
EN-PLS	RMSECV(+1se,min)	EN solution for RMSECV(+1se) combined with minimum RMSECV solution for LV selection
	RMSECV(+1se,+10%)	EN solution for RMSECV(+1se) combined with LV selection such that RMSECV is $\leq 10\%$ of minimum value
	pRMSECV(+1se,min)	EN solution for RMSECV(+1se) combined with minimum pRMSECV solution for LV selection

S2.2 Evaluation: FGs

A challenge for FG quantification in ambient samples is that calibration models developed with laboratory standards must be extrapolated for application to aerosol mixtures which are necessarily more complex. As we presently do not have reference measurements against which to compare individual FG abundance, we base our evaluation on comparison of FG-OC with TOR OC. Solutions for FG abundance for the full model that agree well with TOR OC are obtained separately (Takahama and Dillner, 2015); a possible consideration is that we can find sparse solutions for each FG that maximizes correlation with the full solution. However, in this work we do not impose this restriction and compare differences in predicted FGs for solutions that collectively reproduce FG-OC (to the extent possible) as an exploratory measure, as shown in Figure 5 of the main document. We note that by selecting models which agree with TOR OC, there may be some bias or weight place on aCH as it comprises more than 50% of organic PM for this data set (Ruthenburg et al., 2014; Takahama and Dillner, 2015), and this may be the reason for stronger agreement among aCH predictions than aCOH observed in Figure 4 in the main document. In the following sections, we show comparison of predicted FGs in laboratory standards (Section S2.2.1), comparison of predicted FG-OC against TOR OC (Section S2.2.2), and the number of NZVs and LVs for each solution considered (Section S2.2.3) for raw and baseline corrected spectra calibration models.

S2.2.1 Candidate models

Figures S3 and S4 show the number of NZVs and LVs for each of the solutions proposed in Table S1 for raw and baseline corrected spectra calibration models, respectively. These figures can be viewed concurrently with Figures 1 and 2 in the main document, which provides the range of RMSECVs encompassed by these models. The minimum RMSECV solution often has a large number of NZVs and LVs, but not always the largest in both. Comparing solutions corresponding to RMSECV(min,+10%), RMSECV(+10%,min), and so on, the degree of parsimony achieved with respect to RMSECV(min) varies greatly and not easily anticipated.

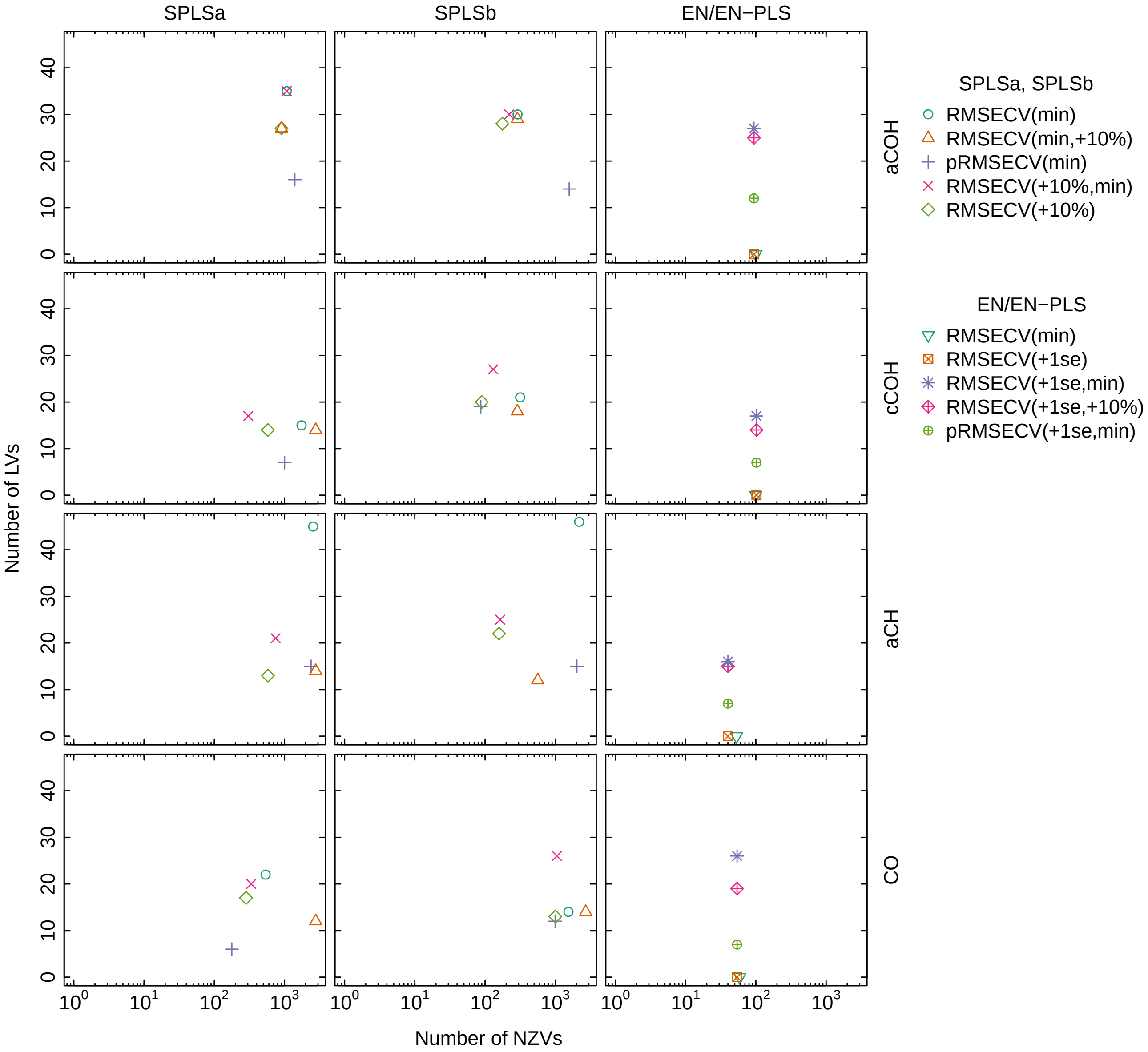

Figure S3. Candidate calibration models developed for FG prediction using raw spectra.

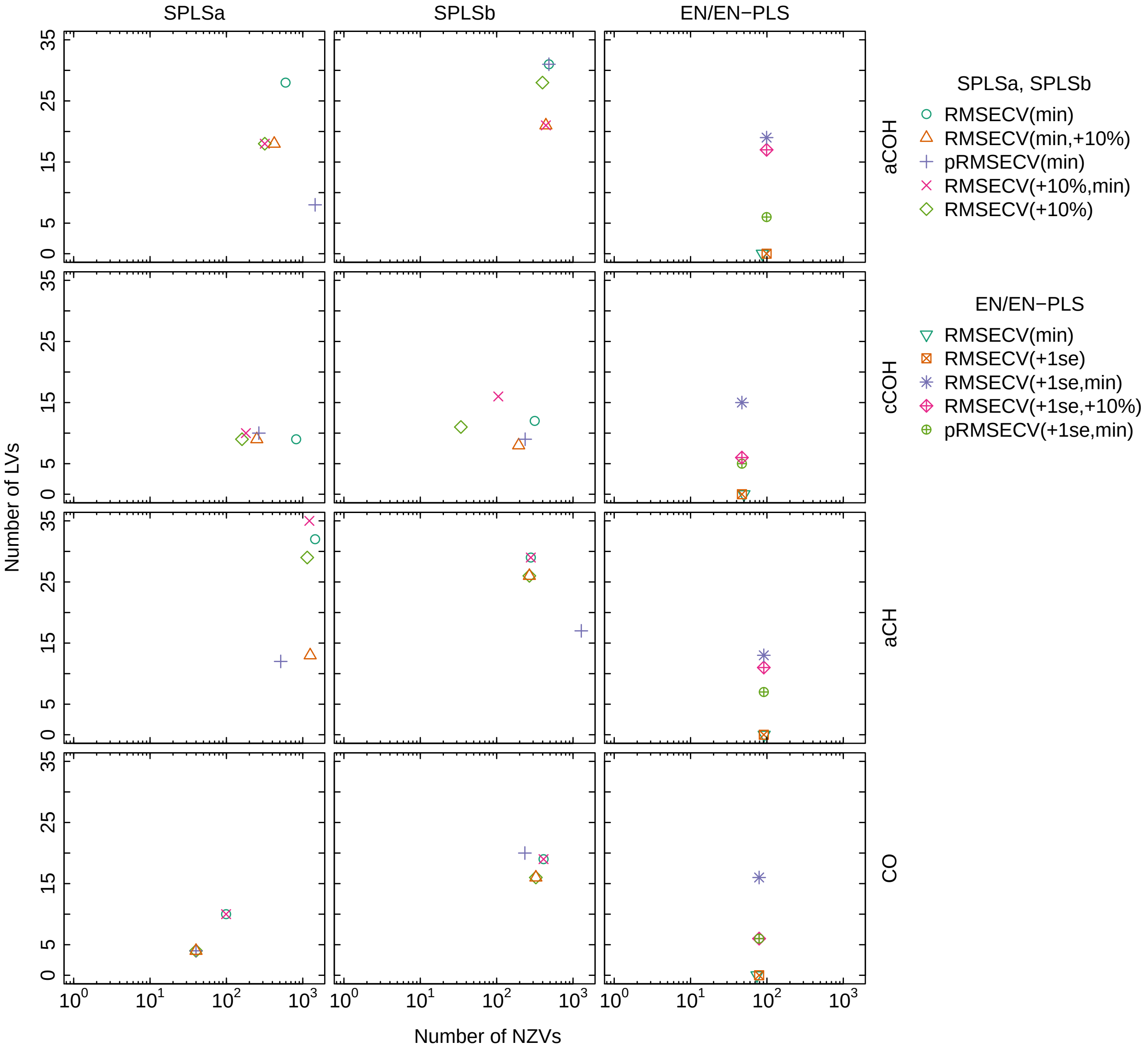

Figure S4. Candidate calibration models developed for FG prediction using baseline corrected spectra.

S2.2.2 Laboratory standards

FG abundance in laboratory standards are predicted well, as shown in Figures S5 and S6 (for raw and baseline corrected spectra, respectively). While there are a few samples which fall outside of this trend, in general the points are aligned with respect to reference values with a slope of 0.03 within unity; with a correlation coefficient of 0.99 or 1.00.

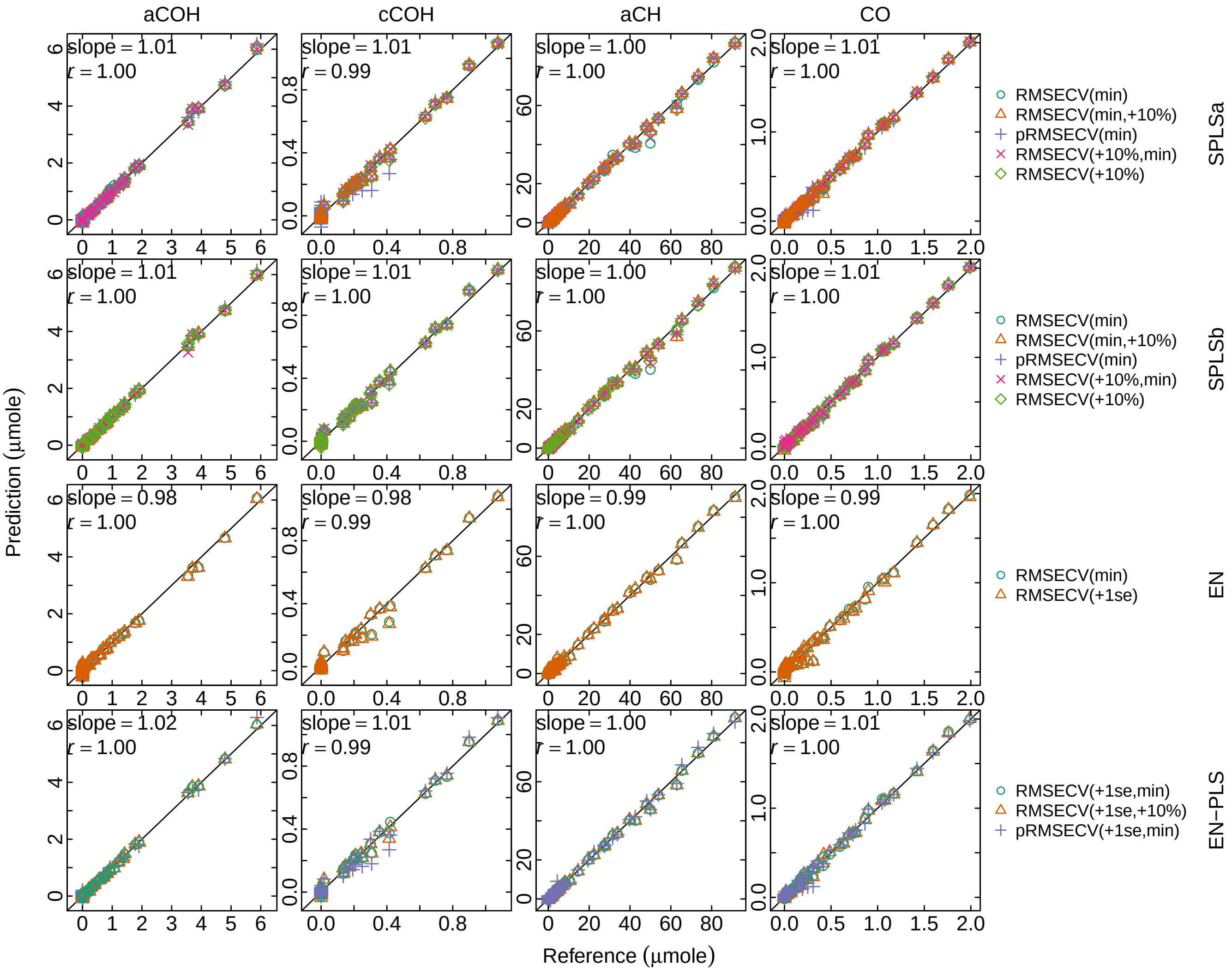

Figure S5. Scatter plots of FG abundances predicted by FT-IR using raw spectra calibration models and reference values obtained by gravimetric analysis. Various solutions are overlaid in each panel, and aggregate statistics are computed for all points and shown in upper left corner.

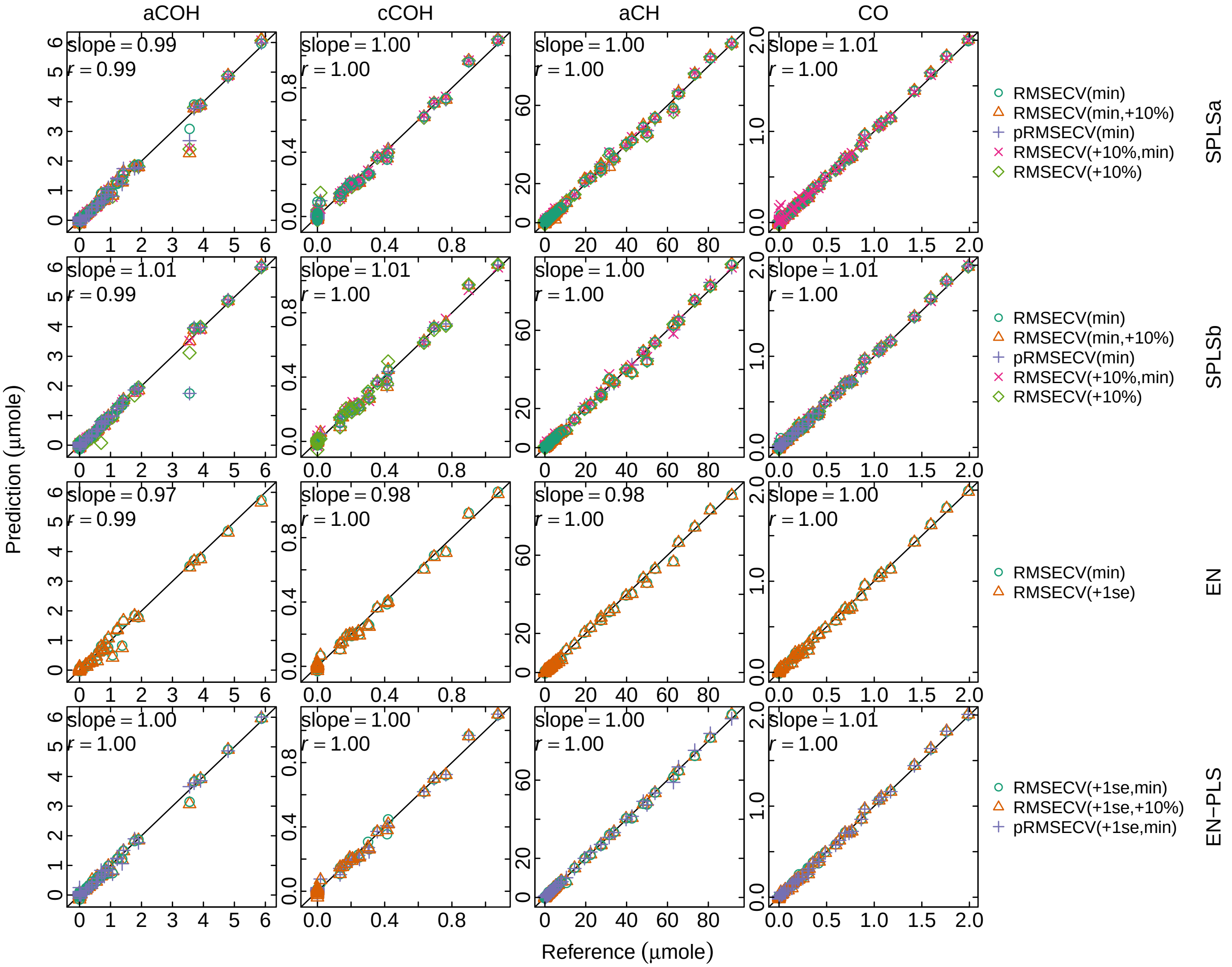

Figure S6. Same as Figure S5 but using baseline corrected spectra calibration models

S2.2.3 Ambient FG-OC

Predictions for FG-OC compared against TOR OC are shown in Figures S7 and S8 for raw and baseline corrected spectra calibration models, respectively. As observed in these figures, correlations can be near zero or negative with respect to TOR OC. In some cases, many of these metrics are heavily influenced by a subset of samples that are hyper-sensitive to model specification (i.e., number of NZVs and LVs, and spectra processing), which can appear at either extreme of predicted abundances depending the calibration model used. These samples have been marked by triangles in these figures, and correspond to clusters 4, 7, 13, 14, 16, 18 (comprising 29% of the samples) reported by Ruthenburg et al. (2014). (These samples have not been excluded from the calculation of regression slopes and correlation coefficients.) Smaller variations with cluster 12 in predicted abundances are described in Section 3.3.1. Since model differences lead to small changes in predicted concentration for laboratory standards (Section S2.2.2) and moderate changes in other ambient samples, these differences highlight the dissimilarity of this subset of sensitive samples to the calibration standards. We posit that this type of evaluation can be used to assess suitability of laboratory standards for particular samples, which will be the topic of future work.

There is no clear solution criteria which are suitable for all models. As stated in Section 2.3 of the main document, we use our selective judgement considering model parsimony and model performance, and selected models are summarized in Table S2. It is conceivable that additional metrics which explicitly weight the tradeoffs among external reference measurements and parsimony with respect to NZVs and LVs can be devised, but is not considered in this work.

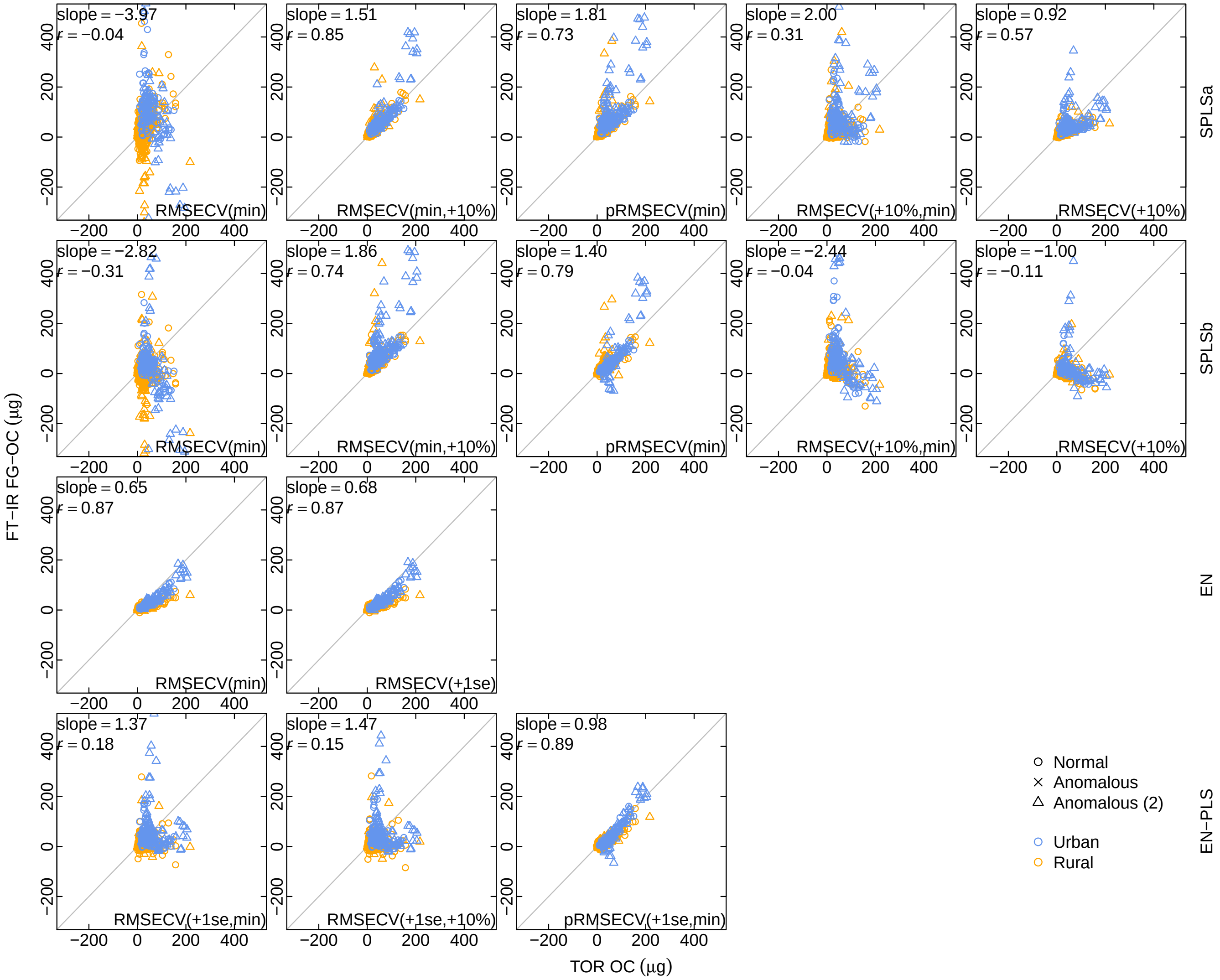

Figure S7. Predictions of FT-IR FG-OC (sum of carbon from individual FGs) predicted from calibration models developed with raw spectra of laboratory standards compared with TOR OC. Anomalous samples are excluded for calculation of performance metrics (i.e., orthogonal regression slope and correlation coefficient).

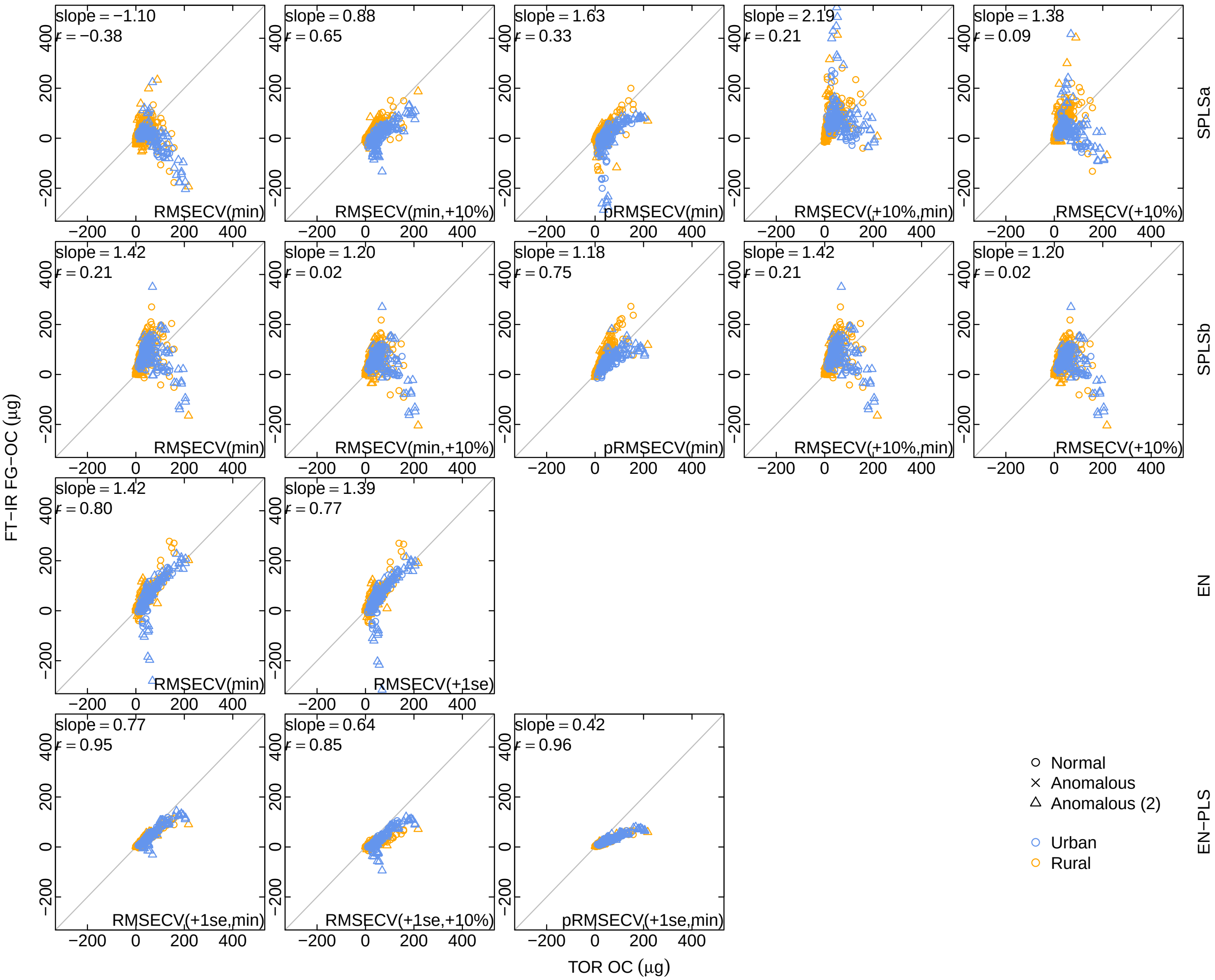

Figure S8. Same as Figure S7, but using models developed with baseline corrected spectra.

S2.3 Evaluation: TOR

In the following sections, we further describe the candidate models (Section S2.3.1) and evaluation for prediction of TOR OC and EC (Section S2.3.2).

S2.3.1 Candidate models

5 Figures S9 and S10 show the number of NZVs and LVs for each of the solutions proposed in Table S1 for raw and baseline corrected spectra calibration models, respectively. These figures can be viewed concurrently with Figures 1 and 2 in the main document, which provides the range of RMSECVs encompassed by these models. As with FGs, minimum RMSECV models have a large number of NZVs and LVs, but are often not at the extreme in both.

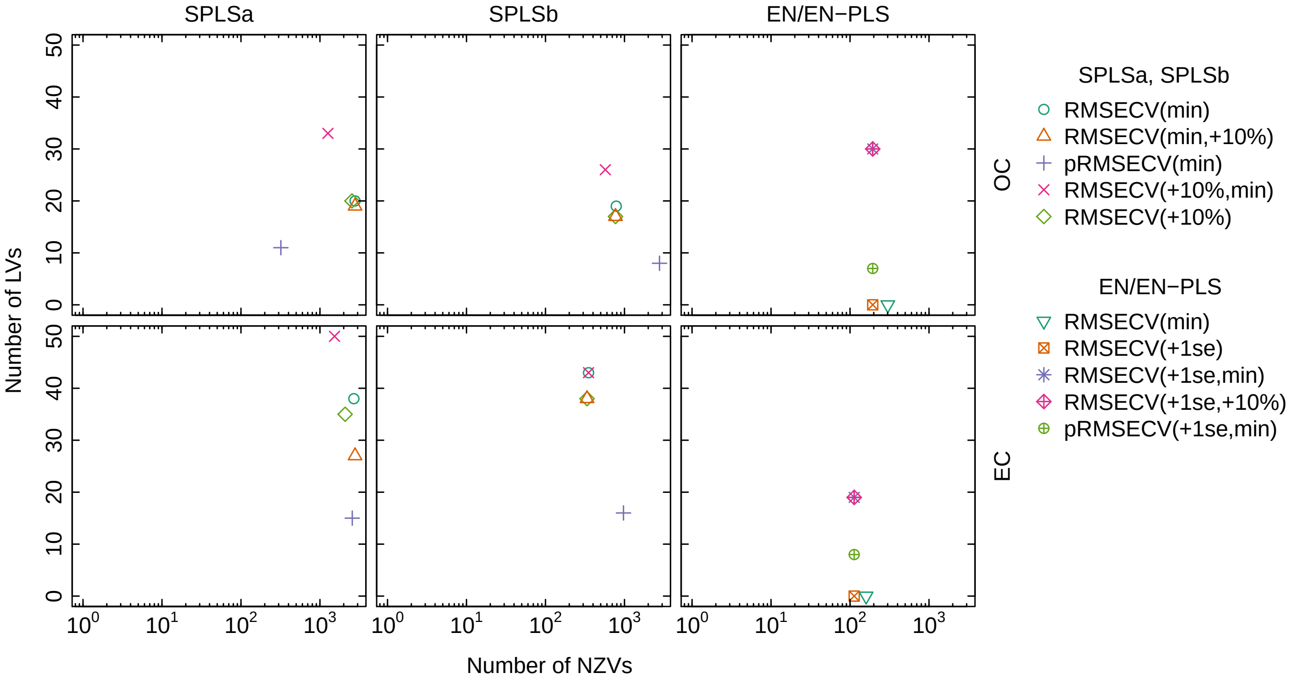

Figure S9. Candidate calibration models developed for TOR OC and EC prediction using raw spectra.

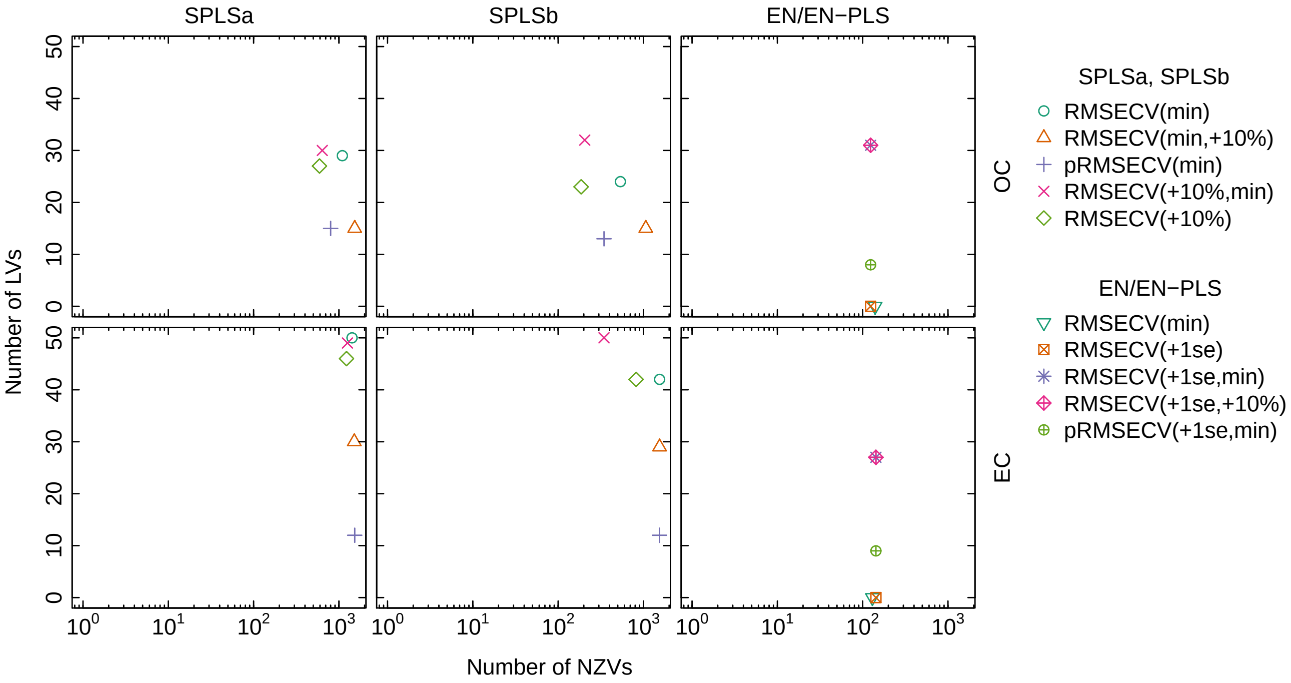

Figure S10. Candidate calibration models developed for TOR OC and EC prediction using baseline corrected spectra.

S2.3.2 Ambient TOR OC and EC

Comparisons of predicted and reference TOR OC concentrations are shown in Figures S11 and S12 for raw and baseline corrected spectra models, respectively; corresponding comparisons for TOR EC are shown in Figures S13 and S14. Predictions are generally robust with respect to model variations in parameters. For TOR OC, correlation coefficients are equal to or greater than 0.97, and slopes are within 0.05 of unity, except for the EN model developed with raw spectra. For TOR EC, correlation coefficients are equal to or greater than 0.94 and slopes within 0.03, again with the exception of the EN model (developed with both the raw and baseline corrected spectra). While many of these models produce suitable predictions, we select models with RMSECV(+10%) for SPLS methods to consider parsimony in both NZV and LVs. RMSECV(+1se) is selected for EN, and pRMSECV(min) for EN-PLS as they lead to smaller NZVs and LVs. These selections are summarized in Table S2.

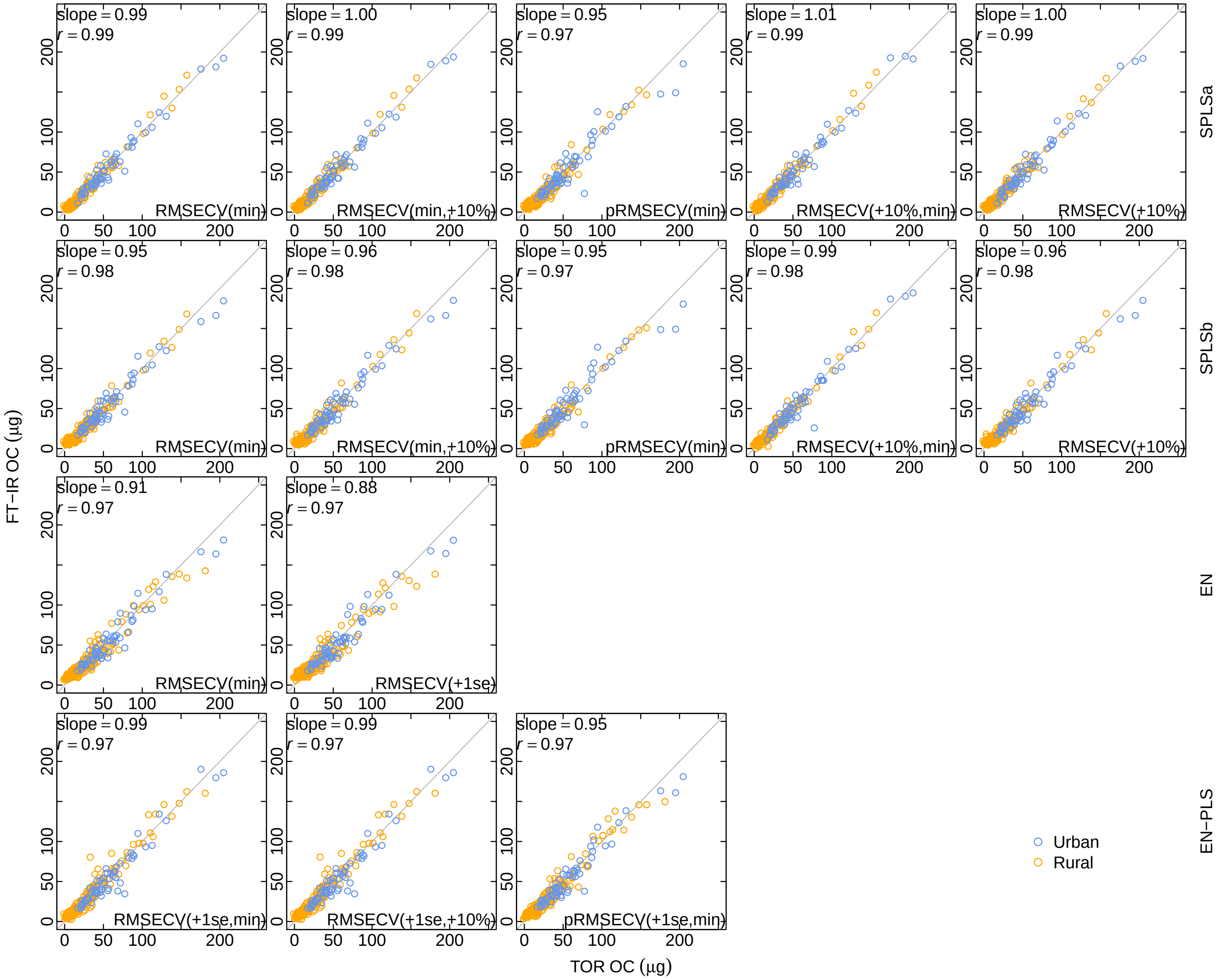

Figure S11. Predictions of FT-IR OC estimated from direct calibration of raw spectra to ambient samples compared with TOR OC. All samples are used for calculation of performance metrics (i.e., orthogonal regression slope and correlation coefficient), in contrast to Figure S7, where some points were excluded.

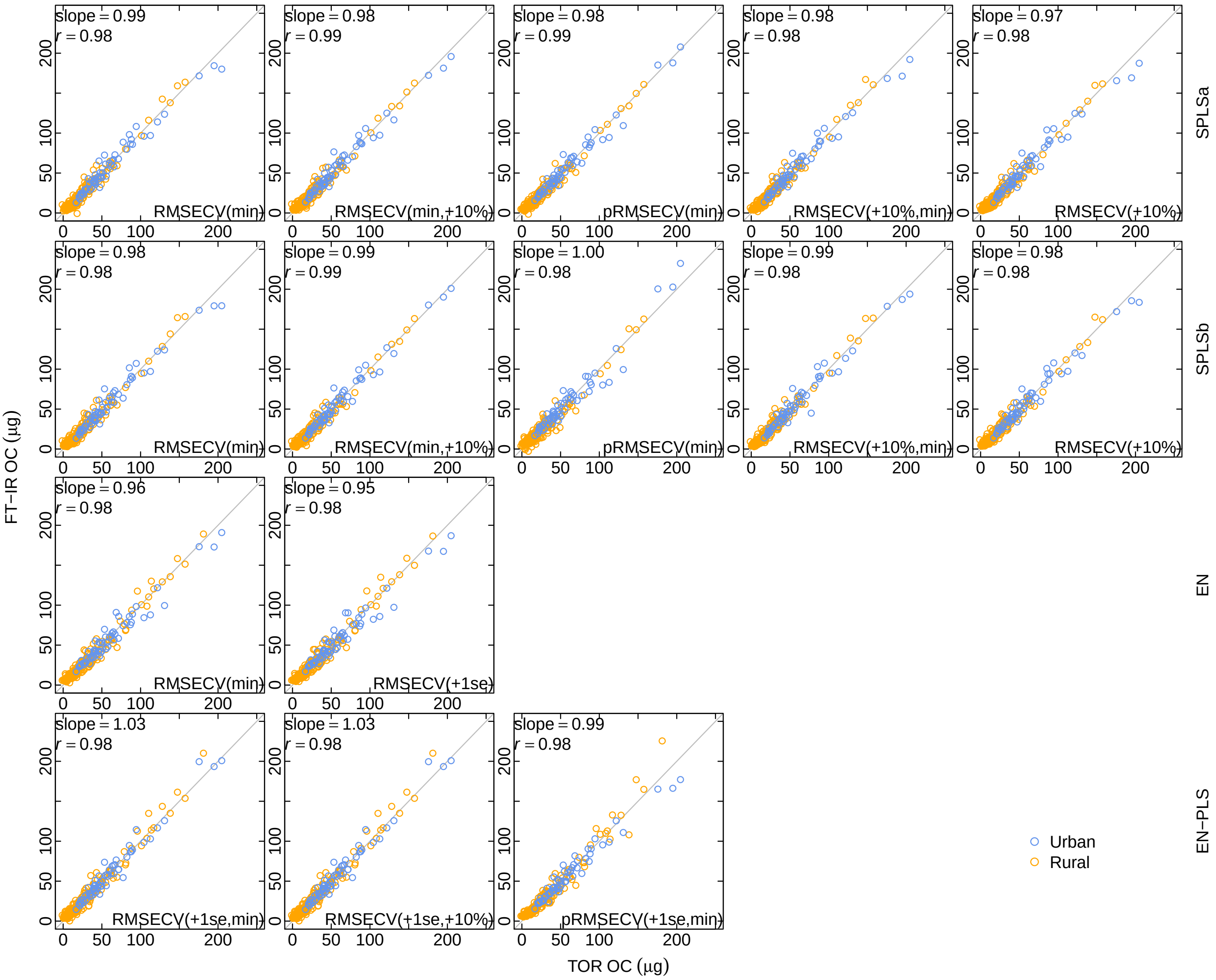

Figure S12. Same as Figure S11, but using models developed with baseline corrected spectra.

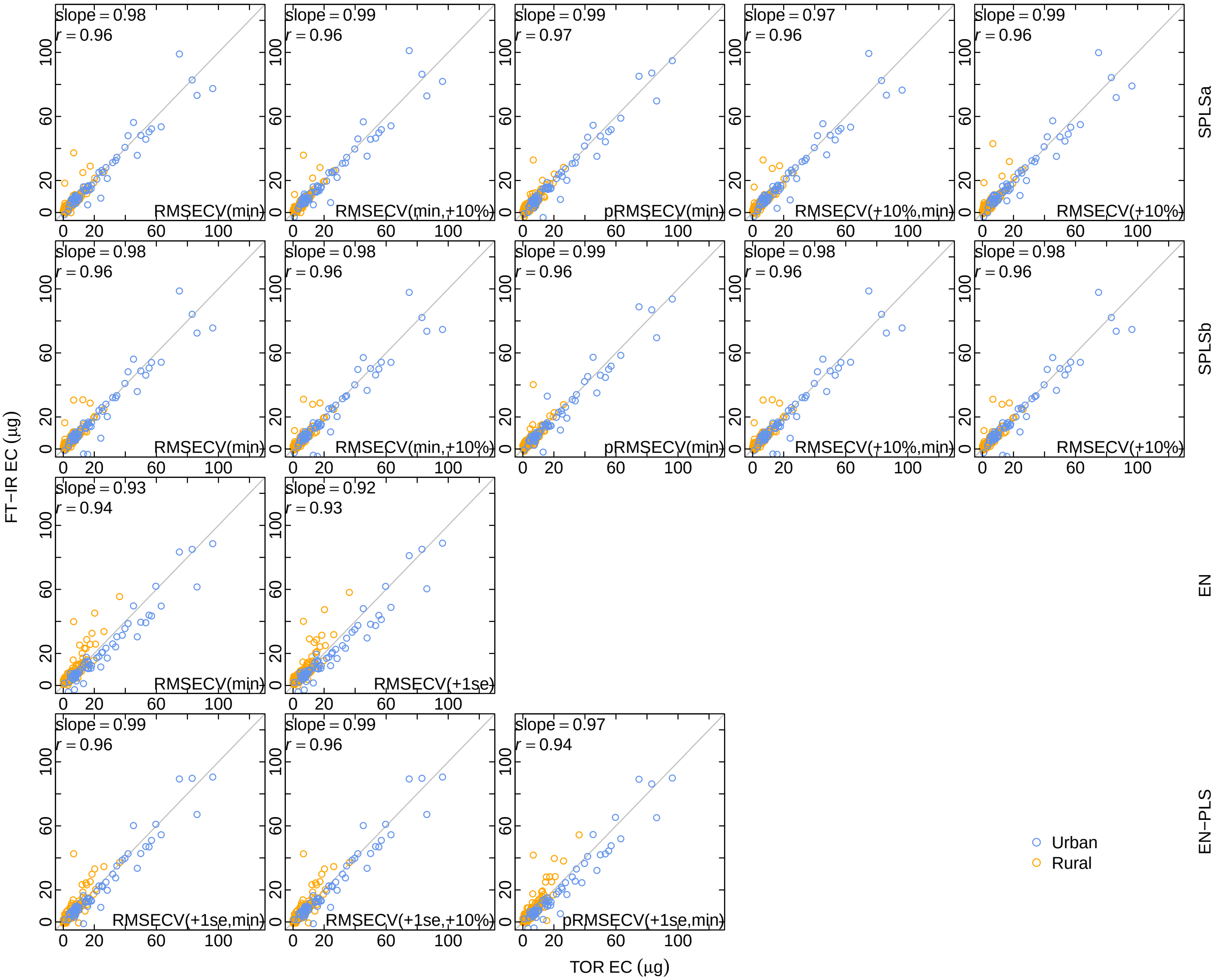

Figure S13. Predictions of FT-IR EC estimated from direct calibration of raw spectra to ambient samples compared with TOR EC. All samples are used for calculation of performance metrics (i.e., orthogonal regression slope and correlation coefficient).

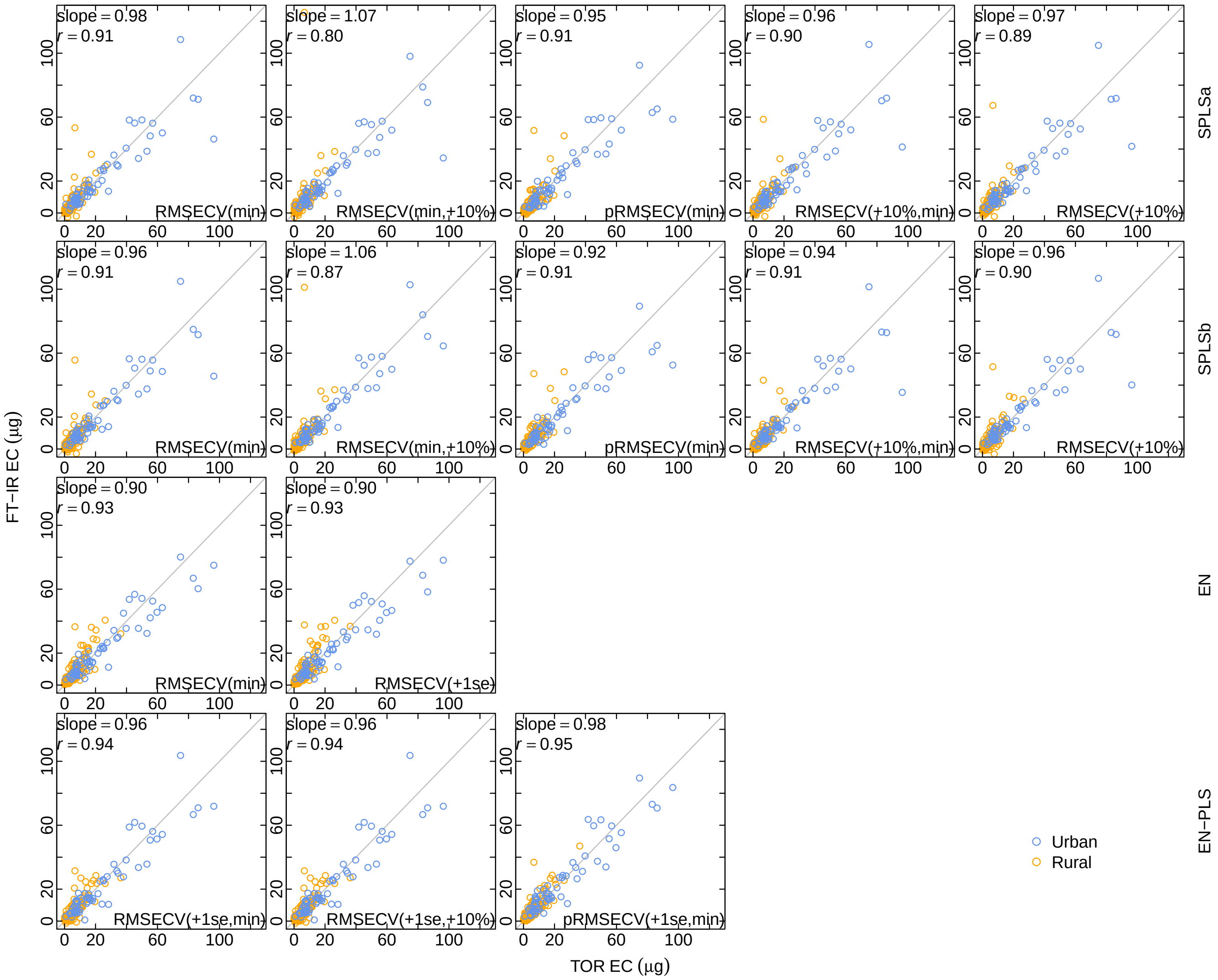

Figure S14. Same as Figure S13, but using models developed with baseline corrected spectra.

S2.4 Models selected

The solutions selected are shown in Table S2, with parameters included in Table 1 of the main document. FGs all share the same solution method as the evaluation is based on an aggregated measure (FG-OC) compared against an external reference (TOR OC). TOR OC and EC are evaluated individually without consideration for the other, but in this work we have selected the same models for a given algorithm and spectral preparation.

Table S2. Models selected for final FG and TOR calibration models presented in main document.

Analyte	Method	Spectra type	Solution
Functional groups	SPLSa	raw	pRMSECV(min)
Functional groups	SPLSa	baseline corrected	RMSECV(min,+10%)
Functional groups	SPLSb	raw	pRMSECV(min)
Functional groups	SPLSb	baseline corrected	pRMSECV(min)
Functional groups	EN	raw	RMSECV(+1se)
Functional groups	EN	baseline corrected	RMSECV(+1se)
Functional groups	EN-PLS	raw	pRMSECV(+1se,min)
Functional groups	EN-PLS	baseline corrected	RMSECV(+1se,min)
TOR OC and EC	SPLSa	raw	RMSECV(+10%)
TOR OC and EC	SPLSa	baseline corrected	RMSECV(+10%)
TOR OC and EC	SPLSb	raw	RMSECV(+10%)
TOR OC and EC	SPLSb	baseline corrected	RMSECV(+10%)
TOR OC and EC	EN	raw	RMSECV(+1se)
TOR OC and EC	EN	baseline corrected	RMSECV(+1se)
TOR OC and EC	EN-PLS	raw	pRMSECV(+1se,min)
TOR OC and EC	EN-PLS	baseline corrected	pRMSECV(+1se,min)

S3 VIP interpretation

Wavenumbers and associated vibrational modes extracted from Shurvell (2006) and Pavia et al. (2008) are shown in Tables S3–S6. Sulfones and sulfonates have been excluded from consideration in this analysis.

Table S3. Range of wavenumbers associated with VIP scores higher than 0.5 for the TOR-OC calibration model using EN-PLS with raw spectra. The vibrational modes associated with the selected wavenumbers and the sign of the coefficients are also listed.

Wavenumber (cm^{-1})	Vibrational mode	Coefficient
501–503	C-O-C bend in ethers C-N-C bend in amines C-C=O bend in carboxylic groups NO$_2$ bend in nitro compounds Alkyl chain deformation modes (weak) Ring deformation in benzene derivatives	< 0
602–605	C=O out of plane bend in amides N-C=O bend in amides C-CO-C bend in ketones O-C-O bend in esters O-C=O bend in carboxylic group NO$_2$ bend in aliphatic nitro groups C=C-H bend in alkynes	< 0
638–640	=CH$_2$ bend in vinyl groups C≡C-H bend in alkynes Ring deformation in naphthalenes NO$_2$ bend in aliphatic nitro groups C-O-H bend in alcohols C-C-CHO bend in aldehydes O-C=O bend in carboxylic group OH out of plane bend in phenols	> 0
1141–1145	C-O stretch in alcohols, ethers, esters, carboxylic groups and anhydrides C-C-N bend in amines C-F stretch	< 0
1191–1198	C-O stretch in alcohols, ethers, esters, carboxylic groups and anhydrides C-C-N bend in amines C-F stretch	< 0
1505–1517	Benzene ring stretch NO$_2$ antisym. stretch in aromatic compounds N-H deformation in secondary amides CH$_2$ and CH$_3$ bend in aliphatic compounds	> 0
1702–1722	C=O stretch in ketones, aldehydes, carboxylic groups and esters C-O stretch in anhydrides Conjugated ester C=O and phenyl or alkene Substituted benzene rings overtones	> 0
1898–1983	C=C=C antisymmetric stretch in allenes Substituted benzene rings overtones (weak)	< 0
3963–3998	(PTFE scattering)	> 0

Table S4. Range of wavenumbers associated with VIP scores higher than 0.5 for the TOR-OC calibration model using EN-PLS with baseline corrected spectra. The vibrational modes associated with the selected wavenumbers and the sign of the coefficients are also listed.

Wavenumber (cm^{-1})	**Vibrational mode**	**Coefficient**
1589–1607	Conjugates ketone, aldehyde and ester C=O and phenyl C=C stretch in alkene N-H bend in primary and secondary amine and amide NH_2 deformation in primary amines (weak) Benzene ring stretch NH_2 deformation in primary amide	> 0
1717–1738	C=O stretch in ketones, aldehydes, carboxylic group and esters Conjugated ester C=O and phenyl or alkene Substituted benzene rings overtones (weak)	> 0
2913–2921	C-H stretch in alkanes O-H stretch in carboxylic group C-H stretch in aldehydes (weak)	> 0
2985–3018	C-H stretch in alkanes =C-H stretch in alkenes and aromatics O-H stretch in carboxylic groups	< 0
3354–3363	NH_2 antisym. stretch in primary amines O-H stretch in alcohol and phenols	> 0

Table S5. Range of wavenumbers associated with VIP scores higher than 0.5 for the TOR-EC calibration model using EN-PLS with raw spectra. The vibrational modes associated with the selected wavenumbers and the sign of the coefficients are also listed.

Wavenumber (cm^{-1})	**Vibrational mode**	**Coefficient**
511–516	C-O-C bend in ethers C-N-C bend in amines	< 0
1058–1063	C-O stretch in alcohols C-N stretch in primary amines C-O-C antisymm. stretch in ethers	< 0
1135–1142	C-O stretch in alcohols, ethers, esters, carboxylic groups and anhydrides C-C-N bend in amines C-F stretch	< 0
1275–1279	C-O stretch in alkyl aryl ethers C-N stretch in aromatic amines C-O-C antisym. stretch in esters C-F stretch	> 0
1497–1531	N-H bend in secondary amides Benzene ring stretch NO_2 antisym. stretch in aromatic compounds	> 0
2256–2309	O-H stretch in carboxylic group C=C=C in allenes	< 0
3990–3998	(PTFE scattering)	> 0

Table S6. Range of wavenumbers associated with VIP scores higher than 0.5 for the TOR-EC calibration model using EN-PLS with baseline corrected spectra. The vibrational modes associated with the selected wavenumbers and the sign of the coefficients are also listed.

Wavenumber (cm^{-1})	**Vibrational mode**	**Coefficient**
1587–1601	Conjugated ketone, aldehyde and ester C=O and phenyl N-H bend in primary and secondary amides and amines C=C stretch in alkenes C=C stretch in aromatic rings Benzene ring stretch	> 0
1722–1739	C=O stretch in ketones, aldehydes, carboxylic groups and esters Conjugated ester C=O and phenyl or alkene	> 0
1783–1789	C=O stretch in lactones C=O stretch in carbonyl compounds Substituted benzene rings overtones (weak)	> 0
2670–2711	C-H stretch when -CH_3 attached to O or N C-H bend overtones in aldehydes	< 0
2917–2927	C-H stretch in alkanes	> 0
2978–3000	O-H stretch in carboxylic group =C-H in aromatics or alkenes C-H stretch in alkanes	< 0
3686–3691	(PTFE scattering or residual water vapor interference)	< 0

References

Dillner, A. M. and Takahama, S.: Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: organic carbon, Atmospheric Measurement Techniques, 8, 1097–1109, doi:10.5194/amt-8-1097-2015, 2015.

Gowen, A. A., Downey, G., Esquerre, C., and O'Donnell, C. P.: Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, Journal of Chemometrics, 25, 375–381, doi:10.1002/cem.1349, 2011.

Pavia, D., Lampman, G., and Kriz, G.: Introduction to Spectroscopy, Brooks/Cole Pub Co., Belmont, CA, 2008.

Ruthenburg, T. C., Perlin, P. C., Liu, V., McDade, C. E., and Dillner, A. M.: Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the IMPROVE network, Atmospheric Environment, 86, 47–57, doi:10.1016/j.atmosenv.2013.12.034, 2014.

Shurvell, H.: Spectra–Structure Correlations in the Mid- and Far-Infrared, John Wiley & Sons, Ltd, doi:10.1002/0470027320.s4101, 2006.

Takahama, S. and Dillner, A. M.: Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy, Journal of Chemometrics, 29, 659–668, doi:10.1002/cem.2761, 2015.