Supporting Information

Perfluoroalkylated substances effects in *Xenopus laevis* A6 kidney epithelial cells determined by ATR-FTIR spectroscopy and chemometric analysis

Eva Gorrochategui[1], Sílvia Lacorte[1], Romà Tauler[1] and Francis L. Martin[2,3*]

[1]Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, 08034, Catalonia, Spain; [2]Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK; [3]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston, UK

- Theory section
- Supplementary Figures
- Supplementary Table

Theory

A brief description of the chemometric methods used in this study is below.

Principal Component Analysis plus Linear Discriminant Analysis (PCA-LDA)

PCA-LDA is a method that exploits the benefit of linear discriminant analysis (or canonical variants analysis, LDA), after using principal component analysis.[1,2]

On the one hand, the central idea of PCA is to reduce the dimensionality of a dataset consisting of large number of interrelated variables, using a small number of PCA factors [*i.e.*, principal components (PCs)] to retain as much as possible of the variation present in the original data set (> 95%).[3] On the other hand, LDA is a data separation technique, which explicitly attempts to model the differences between the classes of the data set that were assigned *a priori*. New variables [linear discriminants (LD)] are found such that the ratio of the between-cluster variance to the within-cluster variance is maximized, and thus the clusters are visualized at maximum separation. Therefore, LDA, like regression methods such as partial least squares, is a "supervised" method, that requires some previous knowledge of the samples constituents (*i.e.*, classes).

PCA can be applied before LDA (thus "PCA-LDA") to reduce computational complexity, increase the recognition accuracy in different categories, and avoid LDA overfitting.[4]

ANOVA-simultaneous component analysis (ASCA)

The ASCA method can be understood as a direct generalization of the analysis of variance (ANOVA) for univariate data to the multivariate case.[5,6] Moreover, this method incorporates the information of the structure of datasets (*i.e.*, underlying factors such as time, dose or combinations thereof), enabling a better understanding of their biological information.

Thus, in ASCA, an ANOVA is initially performed on the raw data matrix, which is decomposed into the sum of different data matrices characterizing the variance caused by each one of the considered factors, plus a residual matrix containing the unexplained variance. The ANOVA equation valid for the three data sets of the present study, acquired following an experimental design of two factors (*i.e.*, chemical -*c*- and dose -*d*-) is shown in **Equation S1**:

$$X = \mu + \alpha_c + \beta_d + \alpha\beta_{(cd)} + E \quad \text{(S1)}$$

where **X** is a matrix containing the raw data acquired with the ATR-FTIR instrument, μ represents an overall offset, $\boldsymbol{\alpha_c}$ represents the effect of the factor "chemical", $\boldsymbol{\beta_d}$ represents the effect of the factor "dose", $\boldsymbol{\alpha\beta_{(cd)}}$ represents the interaction of "chemical" and "dose" and **E** is the residual matrix representing the natural variation among replicates. Thus, the performance of ANOVA allows the division of the variation of the distinct factors in orthogonal and independent parts, which is also one of the goals of the ASCA model.

Following ANOVA, a simultaneous component analysis (SCA) is applied individually to each of the ANOVA factor matrices. SCA is a generalization of PCA for the situation where the same variables have been measured in multiple conditions.

Thus, ASCA model combines the power of ANOVA to separate variance sources with the advantages of SCA to the modeling of the individual separate effect matrices. The mathematical basis of the resulting ASCA model for the dataset of the present study (*i.e.*, two-factor dataset) is shown in **Equation S2**:

$$X = \mu + T_c P_c^T + T_d P_d^T + T_{(cd)} P_{(cd)}^T + E \quad \text{(S2)}$$

where component scores of each sub-model are given by the matrices indicated by $\mathbf{T_c}$, $\mathbf{T_d}$, $\mathbf{T_{(cd)}}$, and the component loadings are given by matrices $\mathbf{P_c}$, $\mathbf{P_d}$, $\mathbf{P_{(cd)}}$. **E** is a matrix in which the residuals of all sub-models of the ASCA model are collected ($\mathbf{E} = \mathbf{E_c} + \mathbf{E_d} +$

$\mathbf{E}_{(cd)}$). Thus, this equation means that the matrix **X** is separated into contributions from the overall mean ($\boldsymbol{\mu}$), one SCA model ($\mathbf{T}_c\mathbf{P}_c^T$) describing the overall effect of the factor "chemical", one SCA model ($\mathbf{T}_d\mathbf{P}_d^T$) describing the overall effect of the factor "dose" and another SCA model ($\mathbf{T}_{(cd)}\mathbf{P}_{(cd)}^T$) describing the interaction of "dose" with "chemical". Hence, the ASCA model (**Equation S2**) is a direct multivariate generalization of the ANOVA model (**Equation S1)**.

In order to examine the statistical significance of the effects of the investigated factors and their interaction, ASCA performs a permutation test in which the null hypothesis ($\mathbf{H}_0$) assumes that there is no effect of the considered factor. However, such permutation test can only be performed under the assumption that raw data are well-balanced (*i.e.*, same number of observations for each factor level).

Figure S1. Visual effect of different pre-processing steps on a set of ATR-FTIR spectra, including feature selection, rubberband baseline correction, Amide I normalization, trained-mean centering and class definition.

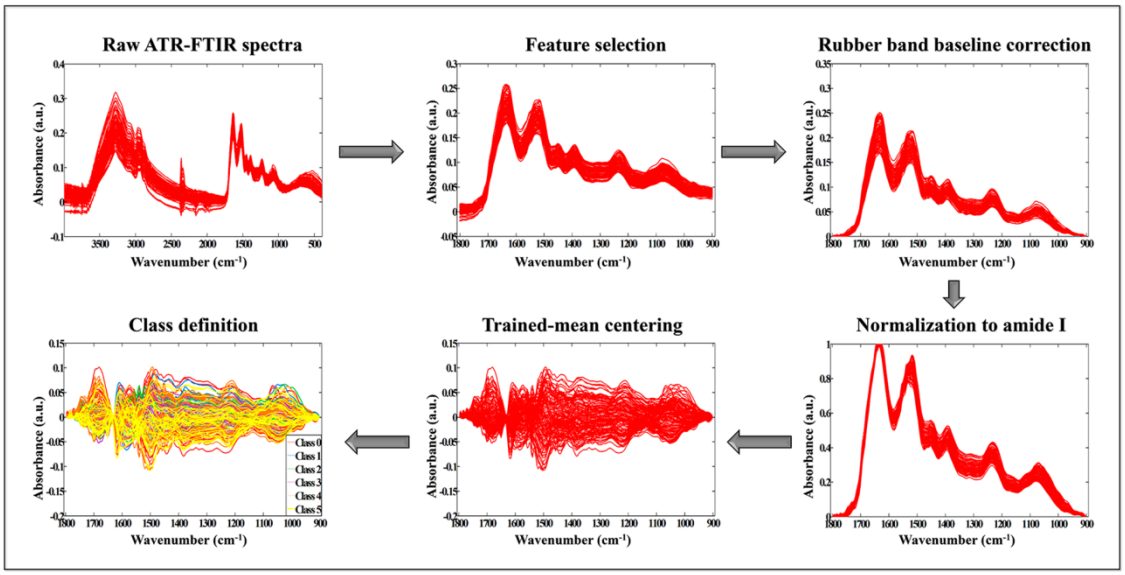

Figure S2. PCA-LDA loadings of PFBS, PFOA, PFOS and PFNA-treatments for the three experiments. Grey-shaded regions indicate spectral regions corresponding to particular biomolecular entities affected in all PFAS-treatments and thicker lines indicate PFAS-doses causing higher effects.

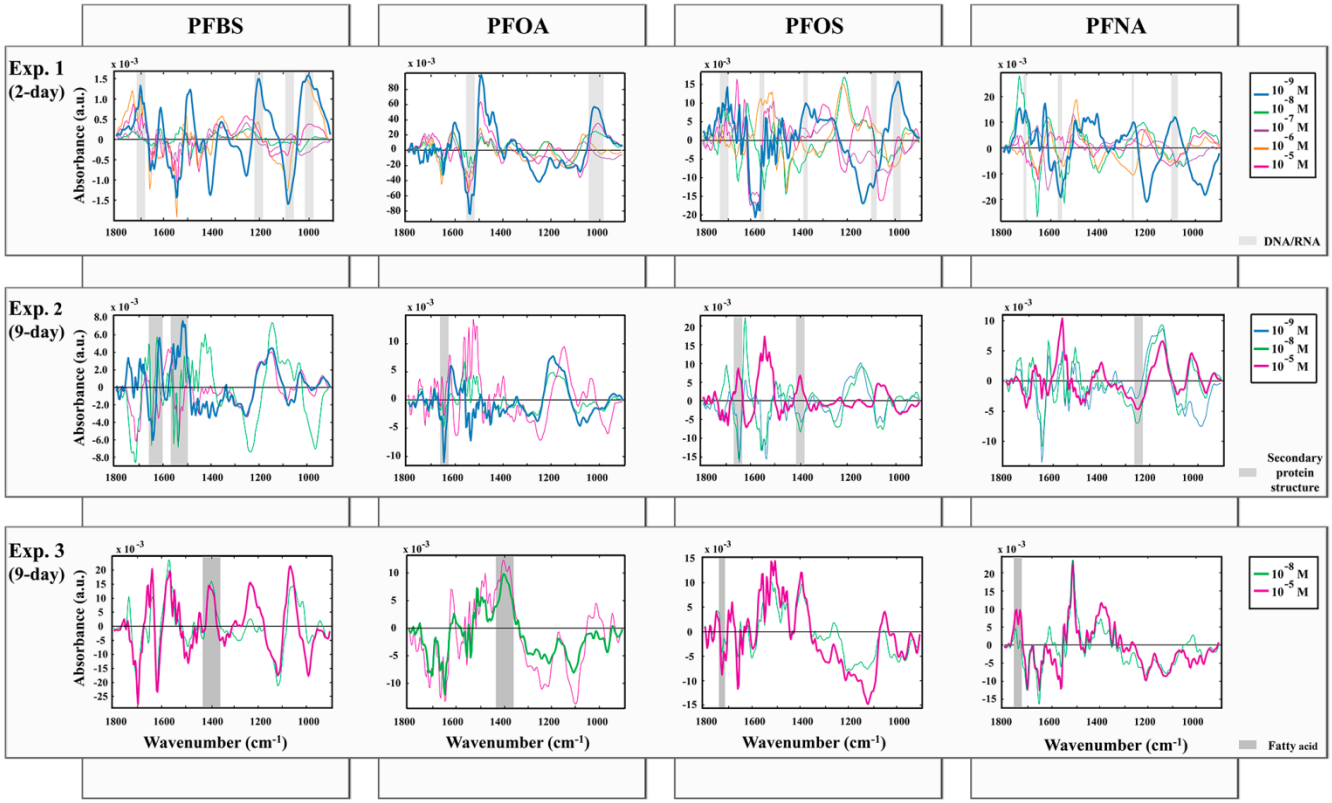

Figure S3. Structure of the data sets arranged for experiments 1, 2 and 3 to perform subsequent ASCA analyses. In the left part of the figures, each rectangle represents an ATR-FTIR spectrum, and the total of them is further arranged into a matrix, as indicated in the righthand side of the figures. In all cases, the following indices are used: w=1,…, W for the wavenumbers; d=1,…, D for the doses tested (d_1=0 M, d_2=10^{-9} M, d_3=, 10^{-8} M, d_4=10^{-7} M, d_5=10^{-6} M and d_6=10^{-5} M); c=1,…, C for the chemicals tested (c_1= PFBS, c_2=PFOS, c_3=PFOA and c_4=PFNA); and, i=1,…, I for the number of replicates of each category.

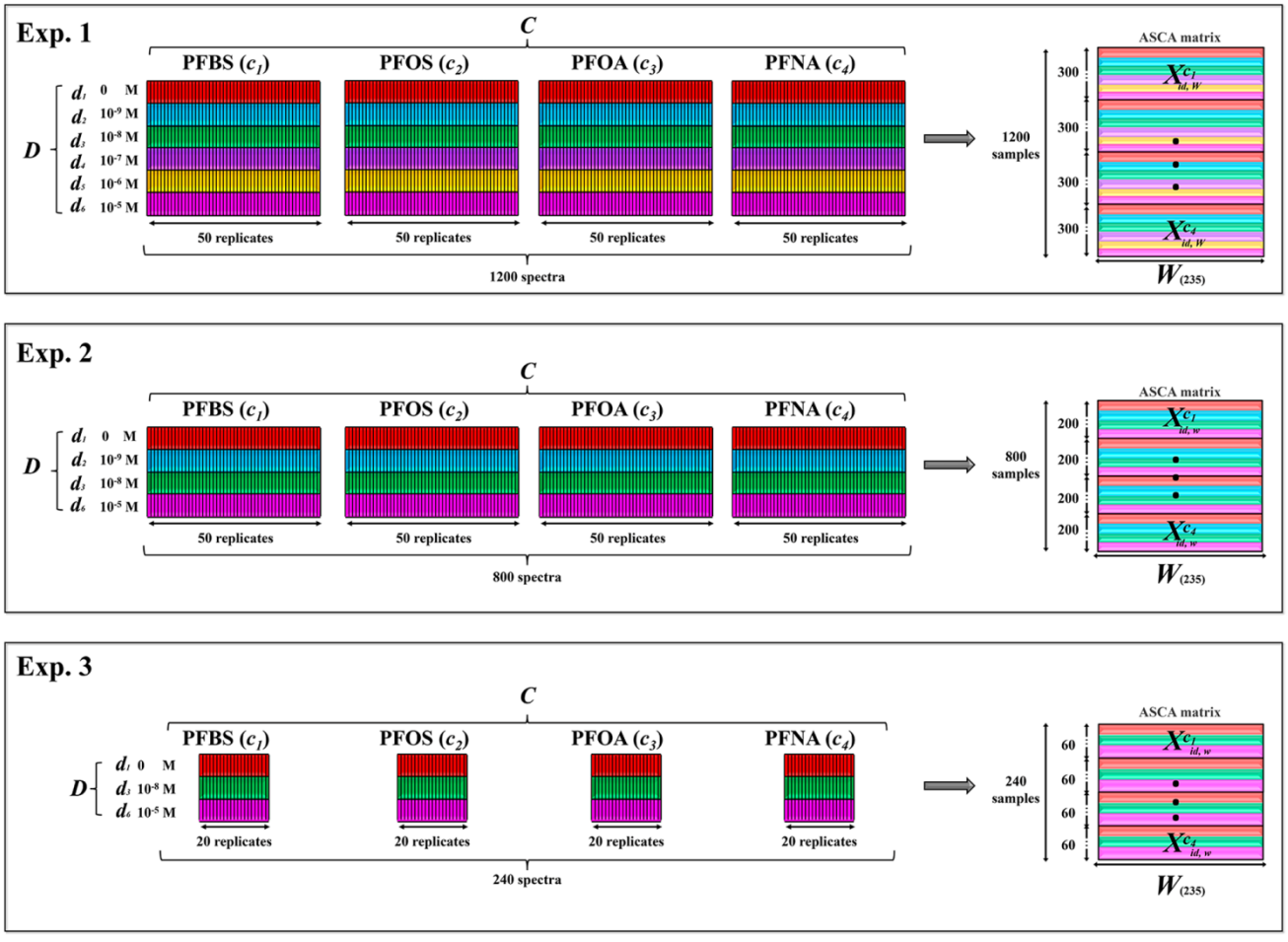

Table S1. Principal segregating wavenumbers and associated biomolecular entities[7] derived from cluster vectors plots of **PCA-LDA**. The indicated wavenumbers (ranked from higher to lower priority) are shown for the distinct experiments and chemicals (*i.e.*, PFBS, PFOS, PFOA and PFNA) at the dose presenting highest effects. Grey-shaded raws indicate spectral bands associated to particular entities affected by all chemicals: DNA/RNA (light grey, exp. 1), secondary protein structures (medium grey, exp. 2) and fatty acids (dark grey, exp. 3).

Exp.	Chemical treatment	Dose of highest effects	Wavenumber (cm^{-1})	Biological fingerprint
1	**PFBS**	**10^{-9} M**	1080	Stretching PO_2^- symmetric vibrations
			994	C-O ribose, C-C
			1201	PO_2^- asymmetric (phosphate I)
			1403	Symmetric CH_3 bending modes of the methyl groups of proteins
			1549	Amide II
			1698/9	C_2=O guanine/ N-H thymine
			1489	In-plane CH bending vibration
	PFOS	**10^{-9} M**	1581	Ring C-C stretching of phenyl
			1567	Ring base
			1137	Oligosaccharide C-OH stretching band
			996	C-O ribose, C-C
			1698/9	C_2=O guanine/ N-H thymine
			1095	Stretching PO_2^- symmetric vibrations
			1717	C=O thymine; C=O stretching vibration of DNA and RNA; C=O stretching vibration of purine base
			1373	Stretching C-N cytosine, guanine
			1736	C=O stretching (lipids)
	PFOA	**10^{-9} M**	1494	In-plane CH bending vibration
			1543	Amide II
			1020	DNA
			1524	Stretching C=N, C=C
			1555	Ring base
			1444	$\delta(CH_2)$, lipids, fatty acids
			1250	Amide III
	PFNA	**10^{-9} M**	1204	Vibrational modes of collagen proteins-Amide III
			1558	Ring base
			1624	Unassigned band
			1728	C=O band
			1089	Stretching PO_2^- symmetric in RNA
			1705	C=O thymine
			1408	Unassigned band

			1258	PO_2^- asymmetric (phosphate I)
2	**PFBS**	**10^{-9} M**	1517	Amide II
			1510	In-plane CH bending vibration from the phenyl rings
			1616	Ring C-C stretching of phenyl
			1643	Amide I band (arises from C=O stretching vibrations)
			1559	Ring base
			1540	Protein Amide II absorption-predominately β-sheet of Amide II
			1750	ν(C=C) lipids, fatty acids
	PFOS	**10^{-5} M**	1543	Amide II
			1524	Stretching C=N, C=C
			1559	Ring base
			1652	Amide I
			1396	Symmetric CH_3 bending of the methyl groups of proteins
			1620	Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode
	PFOA	**10^{-9} M**	1647	Amide I in normal tissues
			1192	Unassigned band
			1535	Stretching C=N, C=C
			1620	Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode
	PFNA	**10^{-5} M**	1562	Unassigned band
			1146	Phosphate and oligosaccharides
			1026	Carbohydrates peak for solutions; vibrational frequency of CH_2OH groups of carbohydrates (including glucose, fructose, glycogen, etc.); glycogen
			1254	Amide III
			1400	Symmetric stretching vibration of COO^- group of fatty acids and amino acids
			1713	C=O thymine
3	**PFBS**	**10^{-5} M**	1698/9	C_2=O guanine/ N-H thymine
			1620	Peak of nucleic acids due to the base carbonyl stretching and ring breathing mode
			1068	Stretching C-O ribose
			1639	Amide I
			1567	Ring base
			994	C-O ribose, C-C
			1119	Symmetric stretching P-O-C; C-O stretching mode
			1234	Composed of Amide III and phosphate vibration of nucleic acids

		1400	Symmetric stretching vibration of COO^- group of fatty acids and amino acids
PFOS	**10^{-5} M**	1119	Symmetric stretching P-O-C; C-O stretching mode
		1504	In-plane CH bending from the phenyl rings
		1517	Amide II
		1396	Symmetric CH_3 bending of the methyl groups of proteins
		1659	Amide I
		1559	Ring base
		1539	Protein Amide II absorption-predominately β–sheet of Amide II
		1724	C=O stretching band mode of the fatty acid ester
PFOA	**10^{-8} M**	1400	Symmetric stretching vibration of COO^- group of fatty acids and amino acids
		1647	Amide I in normal tissues- for cancer is in lower frequencies
		1512	In-plane CH bending vibration from the phenyl rings
		1113	Symmetric stretching P-O-C
PFNA	**10^{-5} M**	1512	In-plane CH bending vibration from the phenyl rings
		1701	C=O guanine
		1651	Amide I
		1562	Unassigned band
		1393	Unassigned band
		1528	C=N guanine
		1748	ν(C=C) lipids, fatty acids
		1732	Absorption band of fatty acid ester; Fatty acid ester band

References

(1) Fearn, T. (2002) Discriminant analysis, in *Handbook of Vibrational Spectroscopy* (Griffiths, J. M. C. and P. R., Ed.), pp 2086–2093. New York.

(2) Walsh, M. J., Singh, M. N., Pollock, H. M., Cooper, L. J., German, M. J., Stringfellow, H. F., Fullwood, N. J., Paraskevaidis, E., Martin-Hirsch, P. L., and Martin, F. L. (2007) ATR microspectroscopy with multivariate analysis segregates grades of exfoliative cervical cytology. *Biochem. Biophys. Res. Commun. 352*, 213–219.

(3) Llabjani, V., Jones, K. C., Thomas, G. O., Walker, L. A., Shore, R. F., and Martin, F. L. (2009) Polybrominated diphenyl ether-associated alterations in cell biochemistry as determined by attenuated total reflection Fourier-transform infrared spectroscopy: a comparison with DNA-reactive and/or endocrine-disrupting agents. *Environ. Sci. Technol. 43*, 3356-3364.

(4) Llabjani, V., Trevisan, J., Jones, K. C., Shore, R. F., and Martin, F. L. (2010) Binary mixture effects by PBDE congeners (47, 153, 183, or 209) and PCB congeners (126 or 153) in MCF-7 cells: biochemical alterations assessed by IR spectroscopy and multivariate analysis. *Environ. Sci. Technol. 44*, 3992–3998.

(5) Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., and Timmerman, M. E. (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics 21*, 3043–3048.

(6) Farrés, M., Villagrasa, M., Eljarrat, E., Barceló, D., and Tauler, R. (2012) Chemometric evaluation of different experimental conditions on wheat (Triticum aestivum L.) development using liquid chromatography mass spectrometry (LC-MS) profiles of benzoxazinone derivatives. *Anal. Chim. Acta 731*, 24–31.

(7) Movasaghi, Z., Rehman, S., and ur Rehman, D. I. (2008) Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev. 43*, 134–179.