Signal processing and analysis of PTR-TOF-MS data from exhaled breath for biomarker discovery

Camille Roquencourt

université PARIS-SACLAY

Signal processing and analysis of PTR-TOF-MS data from exhaled breath for biomarker discovery

Traitement du signal et analyse des données PTR-TOF-MS de l'air expiré pour la découverte de biomarqueurs

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Mathématiques et Informatique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée à l'Institut LIST (Université Paris-Saclay, CEA), sous la direction de **Stanislas Grassin Delyle**, professeur des universités - praticien hospitalier, et le co-encadrement d'**Etienne Thévenot**, ingénieur chercheur.

Thèse soutenue à Paris-Saclay, le 25 mars 2022, par

Camille ROQUENCOURT

Composition du jury

Karine Bennis Zeitouni Professeure des universités, Université de Versailles-Saint-Quentin-en-Yvelines	Présidente
Frédéric Bertrand Professeur des universités, Université de Technologie de Troyes	Rapporteur & Examinateur
Thomas Burger Directeur de recherche, HDR, CNRS, Université Grenoble Alpes	Rapporteur & Examinateur
Wolfram Miekisch Senior researcher, University of Rostock	Examinateur
Stanislas Grassin-Delyle Professeur des universités - praticien hospitalier , Université de Versailles-Saint-Quentin-en-Yvelines	Directeur de thèse

THESE DE DOCTORAT
NNT : 2022UPASG024

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Traitement du signal et analyse des données PTR-TOF-MS à partir de l'expiration pour la découverte de biomarqueurs

Mots clés: Traitement du signal, logiciel, air expiré, PTR-TOF-MS

Résumé: L'analyse des composés organiques volatils (COVs) dans l'air expiré est une méthode non invasive prometteuse en médecine pour le diagnostic précoce, le phénotypage, le suivi de la maladie et du traitement et le dépistage à grande échelle. La spectrométrie de masse à temps de vol par réaction de transfert de protons (PTR-TOF-MS) présente un intérêt majeur pour l'analyse en temps réel des COVs et la découverte de nouveaux biomarqueurs. Le manque de méthodes et d'outils logiciels pour le traitement des données PTR-TOF-MS provenant de cohortes représente actuellement un verrou pour le développement de ces approches.
Nous avons ainsi développé une suite d'algorithmes permettant le traitement des données brutes jusqu'au tableau des intensités des molécules détectées, grâce à la détection des expirations et des pics dans les spectres de masse, la quantification dans la dimension temporelle, l'alignement entre les échantillons et l'imputation des valeurs manquantes. Nous avons notamment mis au point un modèle innovant de déconvolution des pics en 2 dimensions reposant sur une régression du signal par splines pénalisées, ainsi qu'une méthode permettant de sélectionner spécifiquement les COVs dans l'air expiré. L'ensemble du processus est implémenté dans le paquet R/Bioconductor ptairMS, disponible en ligne. Nous avons validé notre approche à la fois sur des données expérimentales (mélange de COVs à des concentrations standardisées) et par simulation. Les résultats montrent que l'identification des COVs provenant de l'air expiré à partir du modèle proposé atteint une sensibilité de 99 %. Une interface graphique a également été développée pour faciliter l'analyse des données et l'interprétation des résultats par les expérimentateurs (les cliniciens notamment).
Nous avons appliqué notre méthodologie à la caractérisation de l'air expiré d'adultes sous ventilation mécanique atteints de l'infection COVID-19. Les analyses de l'air expiré de 40 patients atteints d'un syndrome de détresse respiratoire aiguë (SDRA) ont été effectuées quotidiennement, de l'entrée à la sortie de l'hôpital. Nous avons d'abord réalisé un modèle de classification pour prédire le statut de l'infection, en utilisant l'acquisition disponible la plus proche de l'admission à l'hôpital. Ce modèle permet de prédire le statut de l'infection avec une précision de 93%. Ensuite, nous avons utilisé toutes les données disponibles pour une analyse longitudinale de l'évolution des COVs en fonction de la durée de l'hospitalisation, en utilisant un modèle à effets mixtes. Après sélection de variables, quatre biomarqueurs de l'infection par le COVID-19 ont pu être identifiés. Ces résultats soulignent la valeur des données PTR-TOF-MS et du logiciel ptairMS pour la découverte de biomarqueurs dans l'air expiré.

Title: Signal processing and analysis of PTR-TOF-MS data from exhaled breath for biomarker discovery

Keywords: Signal processing, Sotfware, Exhaled breath, PTR-TOF-MS

Abstract: The analysis of Volatile Organic Compounds (VOCs) in exhaled breath is a promising non-invasive approach in medicine for early diagnosis, phenotyping, disease and treatment monitoring and large-scale screening. Proton Transfer Reaction Time-Of-Flight Mass Spectrometry (PTR-TOF-MS) is of major interest for the real time analysis of VOCs and the discovery of new biomarkers in the clinics. However, there is currently a lack of methods and software tools for the processing of PTR-TOF-MS data from cohorts.
We therefore developed a suite of algorithms that process raw data from the patient acquisitions, and build the table of feature intensities, through expiration and peak detection, quantification, alignment between samples, and missing value imputation. Notably, we developed an innovative 2D peak deconvolution model based on penalized splines signal regression, and a method to specifically select the VOCs from exhaled breath. The full workflow is implemented in the freely available ptairMS R/Bioconductor package. Our approach was validated both on experimental data (mixture of VOCs at standardized concentrations) and simulations, which showed that the sensitivity for the identification of VOCs from exhaled breath reached 99 %. A graphical interface was also developed to facilitate data analysis and result interpretation by experimenters (e.g., clinicians).
We applied our methodology to the characterization of exhaled breath from mechanically ventilated adults with COVID-19 infection. Analysis of exhaled breath from 28 patients with an acute respiratory distress syndrome (ARDS) and COVID-19 infection, and 12 patients with non-COVID-19 ARDS were performed daily from the hospital admission to the discharge. First, classification models were built to predict the status of the infection, using the closest available acquisition to the entry into hospital, and achieved high prediction accuracies (93 %). Then, all the available data acquired during the hospital stay were used for the longitudinal analysis of the VOCs evolution as a function of the hospitalization time by mixed-effects modeling. Following feature ranking and selection, four biomarkers of COVID-19 infection were identified. Altogether, these results highlight the value of the PTR-TOF-MS data and the ptairMS software for biomarker discovery in exhaled breath.

Acknowledgements

First of all I would like to express my sincere gratitude to my two supervisors, Mr Etienne Thévenot and Mr Stanislas Grassin Delyle, for proposing this thesis subject, trusting me and providing the best supervision possible, both on the human and scientific levels. Many thanks to Etienne for his benevolence, kindness and involvement throughout the PhD, as well as during the review of this manuscript, and to Stanislas for his availability, reactivity and his confidence for a future collaboration.

I would like to thank my rapporteurs Mr Thomas Burger and Mr Frédéric Bertrand for kindly agreeing to evaluate this thesis, as well as the examiners Mrs Karine Bennis Zeitouni and Mr Wolfram Miekisch.

I thank the CEA and more particularly the Data Intelligence Unit (SID), for the optimal working conditions. I would also like to thank all the bioinformatics team, Alyssa, Camilo, Sylvain, Eric, Pierrick and Krystyna, for all the constructive discussions and advice, and all the other colleagues of the SID that I met during those three years, for the good atmosphere and the unfailing support.

I would also like to thank the Exhalomics team and the pneumology department at the Foch Hospital, Prof. Philippe Devillier, Prof. Antoine Magnan, Prof. Louis-Jean Couderc, Dr. Hélène Salvator, Dr. Emmanuel Naline, for their excellent welcome and scientific expertise, it is a pleasure to work with them.

Finally, I would like to thank my parents and my brother for their constant help and support, my boyfriend Imad and my friends who have been my crutch for these years, Manel, Ines, Margot, Karim, Ruana, Nabila, and the two best flatmates ever during the lockdown: Joana, who helped me prepare my English orals, and Célia, with whom I have shared every step of our lives for 26 years.

Preface

Using exhaled breath in the clinics as a tool for diagnosis, disease monitoring or therapeutic drug monitoring is very attractive, as sampling is easy and non-invasive and since the analysis can be performed in real-time at the point-of-care. The olfactory signature of illness is also supported by studies using trained dogs, able to detect specific diseases from the patient's exhaled breath, and therefore urges for technological approaches which would be be more reproducible and comprehensive. The Exhalomics® platform from the Hôpital Foch (Suresnes,France) is equipped with a Proton Transfer Reaction Time Of Flight Mass Spectrometer (PTR-TOF-MS) used in clinical research for on-line analysis of exhaled breath. In a close collaboration between the CEA LIST and the Exhalomics team, this thesis aims to provide innovative mathematical methods and bioinformatic tools for biomarker discovery in exhaled breath. The first part of the work was dedicated to the development of algorithms and software tools for the pre-processing of raw data provided by the instrument, whereas the second part focused on the longitudinal analysis of these data from clinical trials conducted at the Hôpital Foch. This thesis was founded by the Agence Nationale de la Recherche (SoftwAiR project, ANR-18-CE45-0017).

At the beginning of the second year of this research (December 2019), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) appeared, affecting the whole world. The developed method for data pre-processing was ready, and gave us the opportunity to apply our methodology to the analysis of exhaled breath from patients suffering from severe COVID-19 infection. Our longitudinal analysis and machine learning approaches led to the identification of a breath signature of the infection (which has been patented), and to a first publication (Grassin-Delyle et al., 2021). Then, our pre-processing workflow and corresponding ptairMS R/Bioconductor package were published (Roquencourt et al., 2022).

This manuscript contains three parts: first the introduction (Part I, chapters 1-3), including all the necessary elements for the understanding of the context, the challenges and the mathematical methods used; then our detailed contributions (Part II), with the development and validation of the pre-processing workflow for PTR-TOF-MS data from exhaled breath (chapters 4-5) and its application to biomarker discovery of COVID-19 infection in

intubated, mechanically ventilated patients (chapter 6); and finally the conclusion and perspectives (Part III). The two articles are shown in appendix B.

Notation

Bold-face, lower-case letters refer to vectors $\boldsymbol{x}$; italic lower-case letters refer to vector elements x_i or scalars a. Bold-face, capital letters refer to matrices $\boldsymbol{X}$, and special front `param` to software parameters.

Valorisation and teaching activities

Publications

- Stanislas Grassin-Delyle, Camille Roquencourt, Pierre Moine, Gabriel Saffroy, Stanislas Carn, Nicholas Heming, Jérôme Fleuriet, Hélène Salvator, Emmanuel Naline, Louis-Jean Couderc, Philippe Devillier, Etienne A. Thévenot, and Djillali Annane (2021). Metabolomics of exhaledbreath in critically ill COVID-19 patients: A pilot study. *EBioMedicine*, 63:103154, doi:10.1016/j.ebiom.2020.103154.

- Camille Roquencourt, Stanislas Grassin Delyle, and Etienne A. Thévenot (2022). ptairMS: real-time proces sing and analysis of PTR-TOF-MS data for biomarker discovery in-exhaled breath. *Bioinformatics*, doi:10.1093/bioinformatics/btac031.

Patent

- EP20306170.0 European patent

Oral communications (online)

- Metabolomics Society 2021; *Signal processing and data analysis of mass spectrometry data from exhaled breath for biomarker discovery*; Early career member best abstract award

- JOBIM 2021 (Journées Ouvertes en Biologie, Informatique et Mathématiques); *Processing of Proton Transfer Reaction Time-of Flight Mass Spectrometry (PTR-TOF-MS) data for untargeted biomarker discovery in exhaled breath: application to COVID-19 intubated ventilated patient*; Best presentation award

- ECCB 2021 (European Conference on Computational Biology); *ptairMS: processing and analysis of PTR-TOF-MS data for biomarker discovery in exhaled breath*

- ICFSP 2021 (International Conference Frontier Signal Processing); *Innovative methods based on 2D penalized regression for the processing and analysis of mass spectrometry*

data from exhaled breath

- ASMS 2021 (American Society for Mass Spectrometry); *Processing and analysis of PTR-TOF mass spectrometry data for biomarker discovery in exhaled breath*; switched to a poster presentation as US borders were closed due to health restrictions

Monitorat (1st and 2nd year of the PhD; 64h/year)

- Introduction to databases; Licence 2; Université Paris-Saclay
- Introduction to databases; 2nd year; Polytech Paris-Saclay
- Introduction to the C++ language; Licence 1; Université Paris-Saclay

Résumé

La 'volatolomique', analyse globale des composés organiques volatils (COV) dans l'air expiré, est une approche prometteuse pour la médecine personnalisée. En effet, l'air que nous expirons est composé à 1% de ces COVs, qui proviennent directement du métabolisme. Des signatures volatolomiques caractéristiques de maladies (biomarqueurs) pourraient donc être identifiées dans l'air expiré. De récents travaux ont ainsi mis en avant l'étude des COVs pour la détection de plusieurs pathologies, dont le cancer, l'asthme, la cirrhose, ou la mucoviscidose (Einoch Amor et al., 2019; Pereira et al., 2015; Feil et al., 2021; Guirao et al., 2019).

L'avantage majeur de la volatolomique par rapport aux examens biologiques classiques est que le prélèvement est complètement non-invasif, simple et rapide. De plus, certains instruments permettent une analyse en temps réel de l'air expiré, tel que la spectrométrie de masse par réaction de transfert de protons (PTR-TOF-MS, Jordan et al. 2009). L'analyse se fait par introduction directe, l'ionisation des COVs a lieu en temps réel, par transfert d'un proton à partir d'un ion primaire (généralement H_3O^+). Les ions ainsi formés (COV + H^+) sont ensuite analysés par un spectromètre de masse à temps de vol.

Le traitement des données brutes issues des instruments PTR-TOF-MS représente un enjeu majeur pour la recherche de biomarqueurs dans l'air expiré. Les principaux défis sont la détection et la déconvolution des pics dans la dimension de masse, ainsi que l'estimation de leurs intensités tout au long de l'acquisition (dans l'échelle temporelle), afin d'identifier les molécules provenant uniquement de l'air expiré. Au démarrage de cette thèse, deux logiciels existaient pour le traitement des données PTR-TOF-MS (Holzinger, 2015; Müller et al., 2013), dont l'un seulement était en libre accès. Ces logiciels sont généralement utilisés pour l'analyse de l'air atmosphérique, et se focalisent sur la détection des pics dans la dimension de masse. Ils ne prennent pas en compte les expirations pour filtrer les variables provenant explicitement de l'air expiré, et ne sont pas adaptés à l'analyse de cohortes (e.g. traitement des fichiers en parallèle).

Nous avons ainsi développé une suite d'algorithmes permettant le traitement des données brutes jusqu'au tableau des intensités des molécules détectées, grâce à la détection des expirations et des pics dans les spectres de masse, la quantification dans la di-

mension temporelle, l'alignement entre les échantillons et l'imputation des valeurs manquantes (Roquencourt et al., 2022). Nous avons notamment mis au point un modèle innovant de déconvolution des pics en 2 dimensions reposant sur une régression du signal par splines pénalisées, ainsi qu'une méthode permettant de sélectionner spécifiquement les COVs dans l'air expiré. L'ensemble du traitement est implémenté dans le paquet R/Bioconductor ptairMS, disponible en ligne. Une interface graphique a également été développée pour faciliter l'analyse des données et l'interprétation des résultats par les expérimentateurs (les cliniciens notamment).

Nous avons d'abord validé notre approche sur des données expérimentales (mélange de COVs à des concentrations standardisées). Après traitement des fichiers par ptairMS, tous les composés attendus ont été détectés, ainsi que leurs isotopes, avec une erreur en masse inférieure à 20 ppm, et une erreur de quantification inférieure à 8%.

Afin de comparer les performances de ptairMS aux deux logiciels existants, nous avons développé un algorithme de simulation de données PTR-TOF-MS issus de l'air expiré, disponible en ligne dans le paquet R ptairData. ptairMS a obtenu la meilleure précision de détection des pics parmi les trois logiciels (99.99%). L'erreur absolue moyenne (MAPE) entre l'évolution temporelle estimée et l'entrée de la simulation est de 4,96% pour ptairMS, contre 14,65% et 5,38% pour les deux autres logiciels. Enfin, nous avons comparé la capacité à discriminer les composés spécifiques de l'air expiré, en utilisant deux t-tests unilatéraux comparant les intensités entre les phases d'expiration et d'air ambiant. ptairMS s'est avéré capable de détecter l'origine des VOCs avec une précision de 99%.

Nous avons ensuite appliqué notre méthodologie à la caractérisation de l'air expiré d'adultes sous ventilation mécanique atteints de l'infection COVID-19. Les analyses de l'air expiré de 40 patients atteints d'un syndrome de détresse respiratoire aiguë (SDRA) ont été effectuées quotidiennement, de l'entrée à la sortie de l'hôpital. Nous avons d'abord réalisé un modèle de classification pour prédire le statut de l'infection, en utilisant l'acquisition disponible la plus proche de l'admission à l'hôpital. Ce modèle permet de prédire le statut de l'infection avec une précision de 93%. Ensuite, nous avons utilisé toutes les données disponibles pour une analyse longitudinale de l'évolution des COVs en fonction de la durée de l'hospitalisation, en utilisant un modèle à effets mixtes. Après sélection de variables, quatre biomarqueurs de l'infection par le COVID-19 ont pu être identifiés (Grassin-Delyle et al., 2021).

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Context

1.1 Biomarker discorvery in exhaled breath

1.1.1 Metabolomic biomarkers

Metabolomics is the study of chemical processes involving small molecule (metabolites, with a molecular weight $<$1,500 Dalton (Da)) that are intermediates and products of life-sustaining chemical reactions in organisms (Oliver et al., 1998). These metabolites, which are the end products of regulatory processes in the organism (Figure 1.1), are important indicators of physiological or pathological states (Wishart, 2019). **Targeted metabolomics** refers to the (usually absolute) quantification of *known* metabolites in a biological sample (saliva, urina, blood; Roberts et al. 2012). In contrast, **untargeted metabolomics** aims at detecting and providing a (usually relative) quantification of *all* metabolites present in the sample. Since the majority of the detected compounds are not known *a priori* in an untargeted metabolomics experiment, additional experiments are usually required for the structural characterisation and identification of the compounds of interest (*e.g.* those highlighted by the statistical analysis).

Biomakers are indicators of a specific biological state, particularly one relevant to the risk of the contraction, the presence or the stage of a disease, or the response to therapeutics (Rifai et al., 2006; Johnson et al., 2016). The full validation of a biomarker usually involves three mains steps:

- **Discovery**: Using an untargeted metabolomics approach, samples from a cohort of patients are collected and analysed. Thanks to statistical learning methods, candidate metabolites providing classification models with a high prediction performance (e.g. for diagnosis, prognosis, or response to treatment are then identified) are detected.

- **Identification**: Chemical identification of the selected metabolites is then necessary for further clinical validation, through both computational (e.g., matching with in-house or public databases), and additional experimental approaches (e.g. tandem mass spectrometry).

- **Validation**: The key step to confirm or refute the candidates metabolites utility in clinical diagnostics is their validation with a second, usually larger, independent cohort.

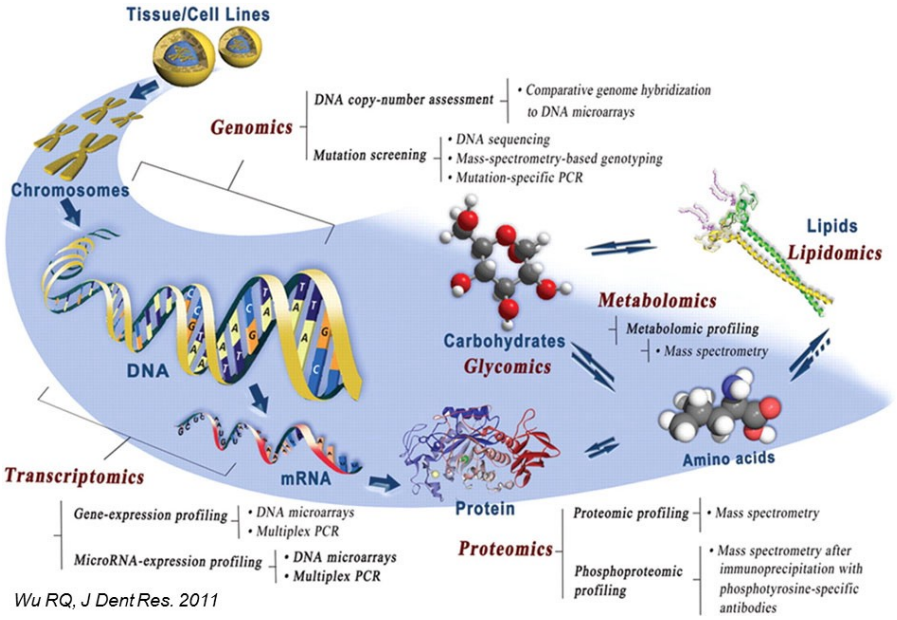


Figure 1.1: Metabolomics among the main omics approaches

1.1.2 Volatolomics: analysis of exhaled breath for personalised medicine

Volatolomics is the analysis of Volatile Organic Compounds (**VOCs**), which can be found in several human matrices such as saliva, urine, skin, blood, and exhaled breath (Amann et al., 2014). More specially, breathomics (breath-based metabolomics) focuses on the capture, identification, and quantification of VOCs in human breath, and their use as tools in medicine (Rattray et al., 2014). Over the past few years, a thousand of individual VOCs have been detected and identified in the human body (Drabińska et al., 2021; de Lacy Costello et al., 2014; Kuo et al., 2020). VOCs may be directly derived from pulmonary metabolism (and thus reflect the metabolic state of the lungs), but they may also be derived from all other organs by being transported through the bloodstream to the lungs, and then into the exhaled breath (Figure 1.2).

Recently, many studies have highlighted the potential of VOC analysis from exhaled breath for early diagnosis and phenotyping of several diseases, such as lung diseases (asthma, cancer, acute respiratory distress syndrome), cardiovascular diseases, cancer (breast, ovarian and liver; Einoch Amor et al. 2019; Pereira et al. 2015), therapeutic drug

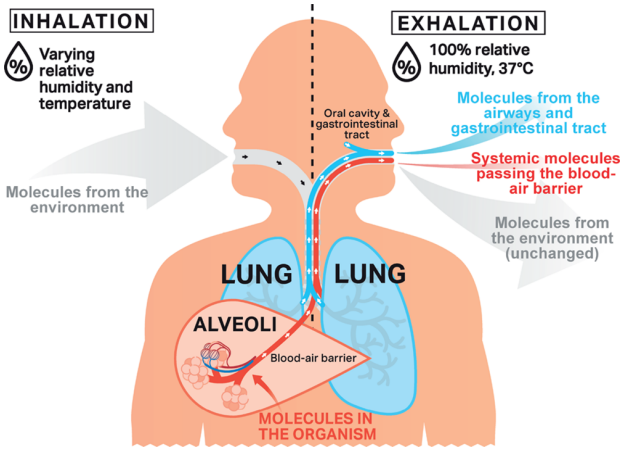


Figure 1.2: Pathways of exhaled molecules in the human body (Bruderer et al., 2019). Endogenous VOCs are excreted through the red and blue pathways.

monitoring (Chen et al., 2021; Boots et al., 2015), and infectious diseases, as tuberculosis, bacterial colonisation of the airways (Koo et al., 2014; Nakhleh et al., 2014; Suarez-Cuartin et al., 2018), ventilator-associated pneumonia in intensive care patients (Schnabel et al., 2015; Bos et al., 2014a), or viral infections (Traxler et al., 2018). In the infectious diseases context, the detected "breathprint" is a mixture of metabolites from microbial origin (i.e. direct biomarkers of the presence of pathogens), and metabolites generated by the host in response to the infection. The existence of the olfactory fingerprints is corroborated by works with dogs, showing the remarkable ability of canine olfaction to identify patients with specific cancer or infectious diseases based on the sniffing of exhaled breath or sweat samples (Feil et al. 2021; Guirao et al. 2019; Vesga et al. 2021; ten Hagen et al. 2021; see also the KDOG project from the Curie Institute).

Breath analysis offers several advantages, the most important being its non-invasive nature and the simplicity of collection, in contrast to biopsy or nasopharyngeal swabs, the current gold standard for the diagnosis of cancer and COVID-19 respectively, which are highly invasive and not risk free. Secondly, recent analytical technologies enable real time analysis and sample collection at the point of care, which is a major asset for large populating screening and **personalised (or precision) medicine** (which refers to the tailoring of medical treatments to the individual characteristics of each patient; Devillier et al. 2017; Martinez-Lozano Sinues et al. 2013). Finally, breath is available in nearly unlimited quantities.

While the discovery of VOC biomarkers is a very promising approach, their detection and identification remain challenging. First, VOCs present in the exhaled breath may be either endogenous (internal metabolic production), or exogenous (current or previous environmental exposures), as illustrated on Figure 1.2, where the exogenous VOCs outnumber the endogenous ones (de Lacy Costello et al., 2014). Second, measurement of

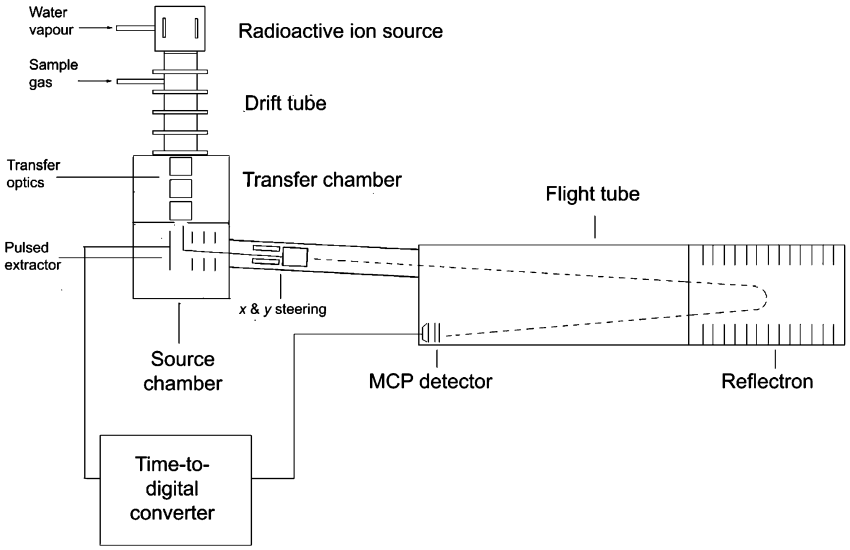


Figure 1.3: Schematic representation of the PTR-TOF-MS instrument (from Blake et al. 2004).

exhaled-breath VOCs requires analytical methodologies that capture metabolites of interest in a reproducible manner, while minimising interference from the sample matrix (Pleil et al., 2013). In this regard, the development of mass spectrometry approaches for volatolomics during the last decade offers exciting opportunities, as detailed below.

1.1.3 Mass spectrometry approaches for VOC analysis

Several analytical methods are available for exhaled breath analysis (Figure 1.4; Rattray et al. 2014). **Mass Spectrometry** (MS) is the method of choice for untargeted VOC analysis due to its sensitivity and selectivity. It measures the mass-to-charge ratio (m/z) of all molecules present in a sample, provided that they can be ionized. Various kinds of MS instruments are used to detect, quantify, and identify the chemical compounds (de Hoffmann and Stroobant, 2007).

Gas chromatography coupled to mass spectrometry (GC-MS) has been applied successfully to VOC biomarker identification from breath (Phillips et al., 2007; Horvath et al., 2009; Löser et al., 2020). However, GC-MS is a time-consuming offline analysis which requires the storage of breath samples in plastic Tedlar bags or sorbent tubes (resulting in several analytical biases; Miekisch et al. 2008; Beauchamp 2011), as well as a suitable laboratory environment, and qualified chemists.

On-line technologies, where the patient blows directly into the mass spectrometer, have emerged as promising approaches for the real-time analysis, since they do not require sample storage, and since results are available on the fly (Bruderer et al., 2019; López-Lorente et al., 2021). The most important factors for on-line monitoring are sensitivity,

Technique	Detection limit	Advantages	Disadvantages	Potential for point-of-care use?
Spectrometric based				
GC-MS	ppb	Current gold standard: can identify unknowns; quantitative; automated	Expensive; time consuming; not currently portable; sensitivity not improved by preconcentration; requires dry samples	No
		Highly sensitive; can preconcentrate samples to detect lower levels, automated	Complicated data deconvolution and compound identification processes	
		VOCs can be captured on different absorbent beds, such as SPME, TD, and Monotrap		
SIFT-MS	ppb	Real-time analysis; can achieve absolute quantification	Expensive; not ideal for broad profiling	No
DMS	ppt	Robust, compact, sensitive	Confident identification needs to be carried out on a MS system	Yes
PTR-MS/PTR-ToFMS	ppt	Has high specificity and can detect very low mass compounds	Cross-signal interference; expensive	Yes
ESI-MS	ppb	Minimal need for adaptive sampling technology, rapid	Requires subject to be beside analytical platform for analysis; relatively expensive	No
FAIMS	ppb	Can be miniaturized; (+)ve and (−)ve ions can be detected simultaneously (Owlstone)	Requires preprogramming; not applicable to unknown compounds; reduced sensitivity in complex matrices; can suffer from signal suppression	Yes
Sensor based				
eNose	ppt	Clinical PoC; data available in real time; ease of use; programmable; handheld	Requires preprogramming; calibration and signal needs to be compared with MS signal; database of disease signals needs to be created (Cyranose)	Yes
		Different sensor design, such as quartz microbalance and conducting polymers (Cyranose), allows for large range of compound coverage		
Gold Nano-Biosensor	ppt	Rapid; no need for preconcentration; highly sensitive; disease specific	In development: requires significant research for PoC	Potentially
Surface Plasmon Resonance	ppt	Highly selective; high throughput	Selective recognition needs to be preprogrammed on an appropriate chip surface (aqueous media)	Potentially
Piezoelectric Cantilever	ppt	Can be specifically tailored to individual compounds, not just classes	Possible issues with poisoning of binding ligands	Yes
		As lithographic techniques improve, more sensors can be applied to smaller chips	Sensitive to vibration	

[a] http://www.hichrom.com/product_range/existing_products/GLS/Monotrap.htm.

Figure 1.4: Current and emerging analytical platforms for the detection and quantification of breath VOCs (from Rattray et al. 2014).

selectivity, scan speed, and robustness. Different variants of MS techniques enable direct sampling and ionisation, including Selected Ion Flow Tube Mass Spectrometry (SIFT-MS, Španěl and Smith 2011), which provides absolute quantification but with low resolution, Secondary Electrospray Ionization (SESI -MS, Wu et al. 2000), which achieves the highest mass resolution reported to date (>140,000) but requires laboratory analytical platform for analysis, and Proton Transfer Reaction (**PTR-MS**; Ellis and Mayhew 2014), which provides both high specificity and the possibility to collect breath at the point of care.

When coupled to Time-of-Flight (TOF) Mass Spectrometry, PTR-TOF-MS (Blake et al., 2004; Herbig et al., 2009; Jordan et al., 2009) has emerged as a promising approach with high sensitivity and specificity for VOC analysis in a wide range of applications (including environment, food quality, biology). Ionisation is based on proton transfer from a reagent ion, most commonly H_3O^+:

$$VOC + H_3O^+ \rightarrow (VOC)H^+ + H_2O$$

As a result, only molecules with a relatively higher proton affinity than water are ionised, excluding the major components of air (N_2, O_2, and CO_2). Furthermore, fragmentation is minimal since proton transfer is a relatively soft ionisation technique. Protonated VOCs are then focused by a lens system and detected in a high resolution reflectron time-of-flight mass spectrometer, according to their mass/charge (m/z) ratio (Figure 1.3). Finally, real-time quantification of VOCs is achieved by ion counting and normalisations based on reaction rates and transmission factors (Cappellin et al., 2012b).

In the area of health and care, PTR-TOF-MS opens up unique opportunities for real-time analysis at the point of care (Smith et al., 2014). Its potential for bio-medicine has been shown in applications such as emphysema, liver cirrhosis, chronic kidney disease and diabetes (Cristescu et al., 2011; Fernández del Río et al., 2015; Obermeier et al., 2017; Pleil et al., 2019; Trefz et al., 2013). However, there is currently a lack of numerical methods and efficient, user-friendly software tools for the processing of PTR-TOF-MS data in the clinics.

1.2 Signal processing of mass spectrometry-based data

The processing of mass spectrometry (MS)-based data consists in transforming the raw data files generated by the mass spectrometer instrument into a representation that facilitates access to characteristics of each observed ion (Katajamaa and Orešič, 2007). It includes the pre-processing of each file (one file per biological sample), by listing the m/z value and quantity of all detected ions (peak picking), followed by the alignment between the samples to generate the sample by variable table of intensities (i.e. the *peak table*). Finally, additional information about the ions is added (such as the isotope distribution

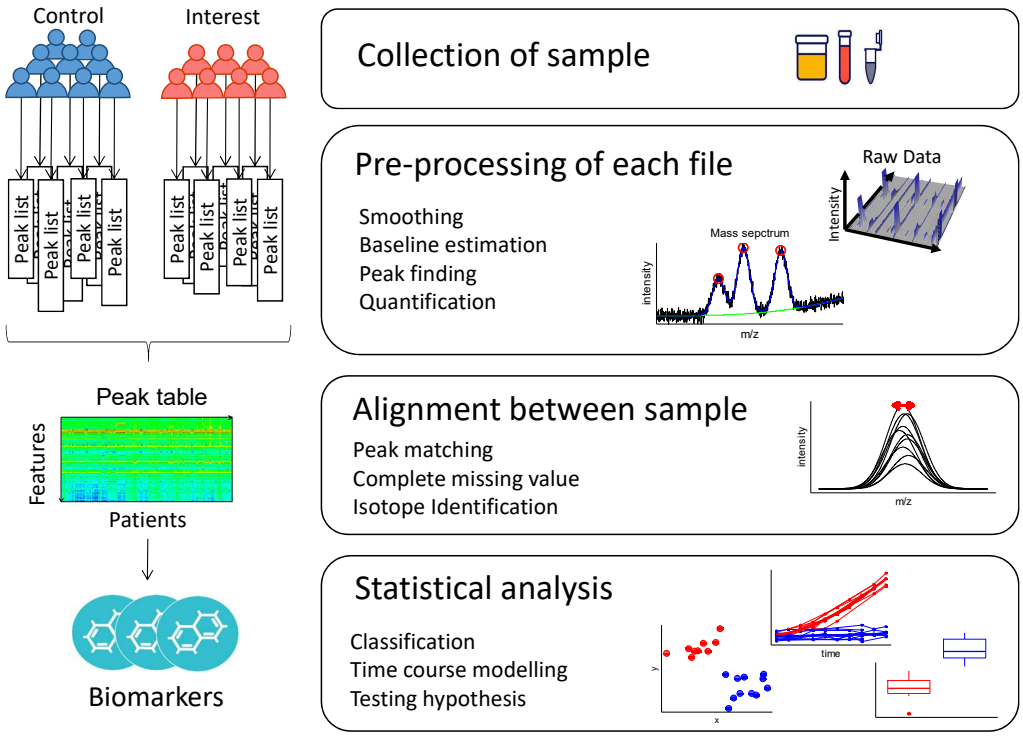


Figure 1.5: Processing workflow for biomarker discovery with MS

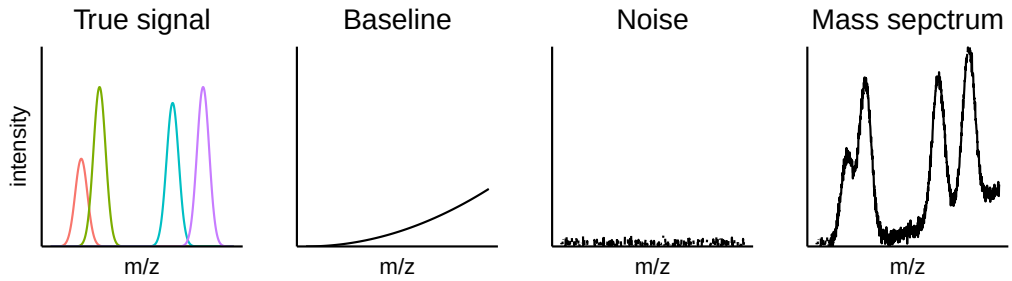


Figure 1.6: Example of a true signal, with additional noise and baseline interferences, which result in the observed mass spectrum.

or annotation obtained by matching the m/z value (and retention time) against in-house or external databases. An overview of the processing workflow applied to biomarker discovery with MS data is described in Figure 1.5.

1.2.1 Peak detection and quantification

Peak detection and quantification is a critical step for MS data processing. A peak is a localised maximum signal produced by the detector around the m/z value of the detected ion. The aim of peak picking is therefore to identify the exact peak location m/z from the raw signal as well as the total ion count.

A mass spectrum contains the ion intensities recorded as a function of the mass to charge ratios m/z, in most cases, instrumental noise and a baseline are present. Figure 1.6, represents the modelling of a mass spectrum by the addition of peak signals, baseline, and noise. Raw data filtering is therefore needed to facilitate the subsequent peak detection.

We describe hereafter the peak picking procedure as a sequence of three steps (Yang et al., 2009): smoothing, baseline correction and peak finding.

Smoothing

Smoothing methods consist in reducing the noise contained in the measured spectrum. Several methods has been described in the literature, as Gaussian filtering (Yang et al., 2009), Kaiser window (Kaiser, 1977), or more recently wavelet transform (WT). In WT approaches, mass spectra are transformed into the wavelet domain and represented in terms of wavelet coefficients in multiple scales. Du et al. 2006 proposed the Undecimated Discrete Wavelet Transformation (UDWT), which is shift-invariant, for spectra denoising, and simultaneous removal of the baseline.

However, the Savitzky-Golay (SG) filter (Savitzky and Golay, 1964) is one of the most popular smoothing algorithm, since it enables to compute the exact first and second deriva-

tive at each point of the signal, which is very useful to detect local maxima. It consists in a moving average filter that performs independent polynomial regression of degree d on a subset of consecutive data points of odd size $2m + 1$ (windows), and takes the central point of the fitted polynomial curve as output.

The choice of the windows size and degree d is then important, since too large windows (or small degree) leads to underestimation and too small windows (or large degree) leads to over fitting and doesn't smooth enough (bias-variance trade-off). An optimal windows selection algorithm has been proposed by Vivo Truyols and Schoenmakers (2006), which minimise the difference between auto-correlation of the fitting residuals (i.e., the differences between the input signal and the smoothed signal) and the auto-correlation of blank signal. More recently, John et al. (2021) also proposed an adaptive method for both degree and windows size choice, based on minimising a generalised unbiased estimation of Mean Squares Error (GUE-MSE) between the true signal and the smoothed output, without any specific distributional assumption on noise.

Baseline correction

After denoising, the baseline needs to be removed from the spectrum before proceeding to peak finding. It classically consists in estimating the baseline before subtracting it.

Many iterative algorithms have been proposed in the literature for baseline correction, including polynomial fitting (the signal is iteratively cut off above the fitted curve; Gan et al. 2006), reweighted penalized least squares (at each iteration, the signal above the fitted curve is assigned a lower weight than signal below; Zhang et al. 2010; Baek et al. 2015; Ruckstuhl et al. 2001), quantile regression (the 0.01 quantile of the signal is estimated instead of the mean Komsta 2011), mixture probabilistic modeling (by computing at each iteration the probability of each point to belong to the baseline; de Rooi and Eilers 2012), and the sensitive nonlinear iterative peak algorithm (SNIP), based on a low statistics digital filter (Ryan et al., 1988; Morháč and Matoušek, 2008). All of these algorithms depend on parameters (such as the degree of the polynomial regression), and require a convergence criterion. The choice of algorithm depends mainly on the type of baseline observed in the data, and thus the type of MS instrument.

Peak finding

The main step of peak picking is the determination of peak locations. A peak can be defined by the *m/z* centre μ, width σ, and height h or area under the curve $\mathcal{A}$. The width σ is usually defined as the full width of the peak at half maximum (FWHM) (Figure 1.7). The resolution of an instrument corresponds to the separation capability between two peaks, and is defined as $R = \frac{m}{\Delta_m}$, where Δ_m is usually the FWHM. Several methods of peak finding are available:

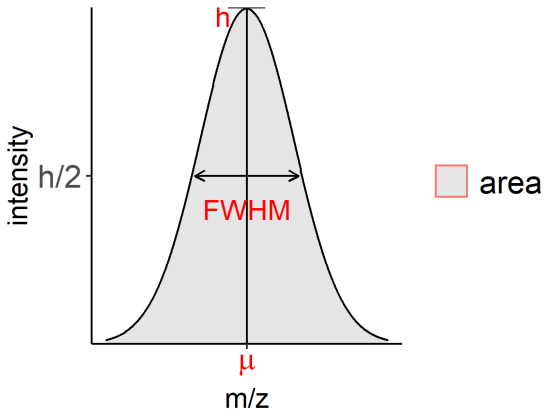


Figure 1.7: Parameters that characterise a peak: the centre μ, the height h, the width at half maximum (FWHM) and the area.

- **Local maxima**

Here, peak locations are defined as the local maxima on the denoised and baseline corrected signal. A function f is said to have a local maximum at the point x^* if there exists some neighbourhood V of x^* such that $f(x^*) \geq f(x)$ for all $x \in V$. Furthermore, if f is differentiable:

$$\frac{\partial f}{\partial x}(x^*) = 0 \quad and \quad \frac{\partial^2 f}{\partial x}(x^*) < 0 \tag{1.1}$$

Based on this definition, most local maxima detection algorithms use the first differences between successive points, and thus list the x such that $f(x_{i-1}) < f(x_i)$ and $f(x_i) > f(x_{i+1})$. However, if the signal is not perfectly denoised, this step may result in false positive peaks, corresponding to little bumps caused by noise. A peak quality control step must therefore be added, such as an intensity threshold (signal to noise ratio; e.g., the 95th percentile of the denoised signal; Du et al. 2006), a threshold on the difference between successive points of the peak (e.g. the median absolute deviation; Coombes et al. 2003), or a minimum distance between two consecutive peaks (Coombes et al., 2003).

Another intuitive option is to use the value of the first and second derivatives of the Savitzky-Golay filter to select points which satisfy Equation 1.1 (Yang et al., 2009).

- **Ridge lines on continuous wavelet transform (CWT)**

Du et al. (2006) proposed to detect peaks by identifying ridge lines on the Continuous Wavelet Transform along different scales. These ridges characterise the regularity of the signal and can be used to detect peaks (Mallat and Zhong, 1992). One advantage of this method is to avoid the need for baseline removal or peak smoothing before peak detection.

- **Deconvolution with a model peak function**

Medium or low resolution mass analysers generate overlapping peaks. In such a case, a deconvolution method must be used to separate and quantify each peak. An approach is to use a model of the peak function, denoted $p_\theta(t)$ with $\theta = (\mu, \sigma, h)$, and to apply a regression algorithm minimising a loss function between the denoised and baseline corrected observed signal $\tilde{\mathbf{y}}$, and the mixture of peak functions:

$$\underset{\boldsymbol{\theta}}{min}||\tilde{\mathbf{y}} - \sum_{i=1}^{P} p_{\theta_i}(\mathbf{m})||^2 \tag{1.2}$$

with $\mathbf{m}$ the vector of m/z values and P the number of overlapping peaks. This can be achieved with standard nonlinear optimisation algorithms, such as the Levenberg-Marquardt algorithm (Lange et al., 2006), particle swarm optimization (PSO; Wijetunge et al. 2015) or Expectation-Maximization (EM; Yu and Peng 2010). The number of peaks P and the initial values of θ must be defined, e.g. by using local maximum detection methods as described before.

Asymmetric peak functions are usually needed, due to imperfections of the mass analyser. Several asymmetric peak shapes have been described in the literature, including bi-gaussian (Yu and Peng, 2010), mixture of gaussians (Leptos et al., 2006), Lorentzian, sech2 (Lange et al., 2006; Stancik and Brauns, 2008), or combination of these (Wijetunge et al., 2015).

To select the best number of peaks P and the best fit function, model selection based on the Bayesian Information Criterion (BIC) or R_2 criteria are used (Lange et al., 2006; Yu and Peng, 2010):

$$BIC = -2\ln\left((\sum_{i}^{n} \tilde{y}_i - \hat{y}_i)^2\right) + P \cdot \ln(n) \tag{1.3}$$

$$R_2 = 1 - \frac{\sum_i^n (\tilde{y}_i - \hat{y}_i)^2}{\sum_i^n (\tilde{y}_i - \bar{y})^2} \tag{1.4}$$

with $\hat{y}_i = \sum_{j=1}^{P} p_{\hat{\theta}_j}(m_i)$, $\hat{\theta}_j$ the solution of equation 1.2 and $\bar{y}$ the average of the denoised and baseline corrected observed signal .

The last step of peak picking is to provide the total ion count for each detected peak. It is usually computed as the area under the curve of the fitted peak shape, or the sum of the raw signal between the peak boundaries if there was no peak deconvolution.

1.2.2 Alignment

Once the peaks have been detected in the individual sample files, a matching (i.e. alignment of m/z values) across the samples is required to generate a single matrix of intensities for the whole experiment, where each row corresponds to one ion and each column contains the quantities of these ions (e.g., peak area) in one sample. Regarding the PTR-MS instrument, the internal mass calibration (section 2.1) enables to perform an initial alignment between the mass spectra by using reference peaks (Jeffries, 2005; Frenzel et al., 2003). Then, since the mass shift error is non-linear, additional methods are required to group masses corresponding to the same ion. Instead of using fixed interval matching (i.e. binning), Smith et al. (2006); Delabrière et al. (2017) proposed a kernel density estimator to compute the overall distribution of peaks m/z, and to dynamically identify boundaries of regions where many peaks have similar m/z.

1.2.3 Identification

At that stage, the detected features (ions) are defined by their mass. Two kinds of additional information are sought to provide further chemical insight. First, the identification of isotope pairs among the features can be detected by looking for mass differences corresponding to one neutron and by checking the correlations between the intensity profiles among the samples (Treutler and Neumann, 2016). Second, the mass can be matched to databases of metabolites, or the chemical formula, to suggest candidates. The key parameters for a successful match are the mass accuracy of the instrument, its resolution (i.e. its ability to separate neighbouring peaks), and the content of available databases. For further characterisation (e.g. distinction between isomers), complementary analytical approaches are required, such as one- or two-dimensional gas chromatography (Phillips et al. 2013; see the discussion in section III).

1.3 Online exhaled breath data processing

The principal challenge of exhaled breath analysis is to differentiate between VOCs coming from the body and the external environment (endogenous vs exogenous). Indeed, real time analysis method continuously records spectra during the acquisition, the ambient air of the room is analysed during the intervals between two expirations (e.g. when the patient inhales). Furthermore, Miekisch et al. (2008) showed that alveolar samples (which correspond to the end tidal of expiration, coming from alveoli, see Figure 1.2) showed the highest concentrations of endogenous and lowest concentration of exogenous substances. It is therefore important to detect the alveolar expiration phases and discard compounds that do not originate from exhaled breath.

1.3.1 Expiration phases detection

During real-time breath acquisitions, several exhalations are usually recorded. As explained in section 1.1.2, gas exchanges take place in the alveoli: as a result, VOCs produced by the metabolism are present in the alveolar air, which represent the end tidal of expiration. Herbig et al. (2009) therefore suggested to identify breath phases by using the signal of tracer compounds that originate from the blood–gas exchange in the alveoli and are present in high concentration in a breath sample, such as acetone (m/z 59.049), CO_2 (m/z 44.997) or humidity with the water dimer isotope (m/z 39.033).

Schwoebel et al. (2011) used the water dimer signal (m/z 37.028) to distinguish between inspiratory and alveolar air. An algorithm was designed to automatically detect those phases using the signal trace around *m/z* 37, with a threshold on the intensity and the stability of cycle. Points greater (respectively, lower) than the mean of the whole trace are considered as expirations (respectively inspirations), and the gradient signals (difference between successive points) from the same expiration cycle (respectively inspiration) has to be less than a fixed value (2.5%).

This method was further generalised by Trefz et al. (2013), who set two percentage thresholds t_{exp} and t_{inh}, and defined expiration (respectively inhalation) as the part of the trace where the intensity is higher (respectively lower) than t_{exp}% (respectively t_{inh}%) of the signal trace maximum. This approach was used in several studies from the same group: using isoprene as breath tracer (Sukul et al., 2014) , on ventilated patients (Brock et al., 2017), or using acetone (Trefz et al., 2019b; Sukul et al., 2021).

1.3.2 Ambient inhaled air

During online acquisition of exhaled breath, the ambient air of the room is both analysed by the instrument and inhaled by the patient. Compound from ambient air can thus be a significant source of confounding variables. It has been demonstrated that for the compounds present in ambient air, their concentration in exhaled breath is related to their concentration in the ambient inhaled air (Phillips, 1997; Beauchamp, 2011; Filipiak et al., 2012; Španěl et al., 2013; Pleil et al., 2013; Smith et al., 2014). Phillips (1997) therefore introduced the concept of "alveolar gradient", which corresponds to the concentration in breath minus the concentration in inhaled air. If the gradient is positive, the VOCs is considered from exhaled breath, and if it is negative or close to zero, it is considered as an ambient air pollutant. This method assumes that the subject is in equilibrium with room air before the sampling (in practice, the patient is allowed to breath quietly in the room for a few minutes).

The quantitative analysis of seven VOCs present in ambient air showed that all these compounds were partially retained in the exhaled breath, and that there was a linear relationship between the exhaled and inhaled air concentrations (Španěl et al., 2013). A

correction which is specific to each compound may therefore be applied for targeted studies.

In the more general case of untargeted approaches, the ambient air intensity is usually subtracted from the averaged expiration intensity (van den Velde et al., 2007; Zhou et al., 2017). Alternatively, breath-specific compounds are selected by thresholding the expiration intensity as a function of ambient air (Bajtarevic et al., 2009; Wehinger et al., 2007).

Chapter 2

Current processing of PTR-TOF-MS data

2.1 Data acquisition

MS instruments consist of an ion source, a mass analyser, and a detector (Gross, 2011). In PTR-TOF-MS instruments (section 1.1.3), chemical ionisation is achieved by proton transfer, usually from a source of hydronium ions (H_3O^+, called *primary* ions). In addition, the TOF analyser provides high sensitivity and resolving power (Jordan et al., 2009). Finally, ion counting is performed by using a microchannel plate detector.

During data acquisition, which is very fast, the instrument continuously analyses the air flowing through a buffer tube (i.e. ambient air by default) and the patient is asked to expire a few times into the tube. A buffered-end tidal system may be used to prolong the end of expirations and to achieve efficient breath capture (Herbig et al., 2008), shown in Figure 2.1.

Raw data are provided in the form of a numerical matrix, where the indices of the rows are the TOF bins (which will be converted into m/z values during the calibration step of data processing), and the column indices are the acquisition times (in seconds). A bin j is a time interval of duration t_{bin} (in ns), during which the ions arriving between $](j-1) \times t_{bin}, j \times t_{bin}]$ are counted by the detector. The resulting intensities for all bins form an extraction, or spectrum: spectra may be averaged by the processing algorithms to reduce the signal/noise ratio (see the nomenclature on Figure 2.2).

Raw files may be large ($\sim$ 50 MB); they are generally stored in the HDF5 open format (Koziol, 2011), which allows direct access to specific blocks of data of interest if necessary (e.g. during imputation of missing values, when a refined analysis of the raw data within the region of interest is required). The raw files also contain the metadata collected during

Figure 2.1: PTR-Qi-TOF MS with a buffered-end tidal device (BET med, Herbig et al. 2008), Exhalomics, Foch Hospital.

the acquisition (date, drift temperature and pressure, etc.).

2.2 Data pre-processing

There are few pre-processing algorithms for untargeted peak detection of PTR-TOF-MS data described in the literature: Cappellin et al. (2011a), Müller et al. (2011) and Holzinger (2015). These three algorithms (and particularly the last two) follow the same steps: internal calibration of the mass axis with reference peaks, dead time correction, peak detection on the average mass spectrum, and quantification of peaks along the acquisition time. Online exhaled breath data analysis has also been the subject of several developments, such as the detection of expiration phases (Herbig et al., 2009; Schwoebel et al., 2011; Trefz et al., 2013), and the correction of the inhaled ambient air concentration (Phillips et al., 1994; Beauchamp, 2011; Španěl et al., 2013). We present in this chapter the state of the art and the remaining challenges for the processing of PTR-TOF-MS data in the context of online exhaled breath analysis.

2.2.1 Calibration of the mass axis

Related formula

TOF-MS analysers separate ions of different mass to charge ratios (m/z) based on their specific velocities. As all ions are accelerated with an equal kinetic energy, the lower the m/z, the faster the ions reach the detector. Their flight times t are then recorded by the detector. Brown and Gilfrich (1991) demonstrate that the following equation describes the relationship between mass and flight time:

$$m/z = \left(\frac{t-a}{b}\right)^2 \tag{2.1}$$

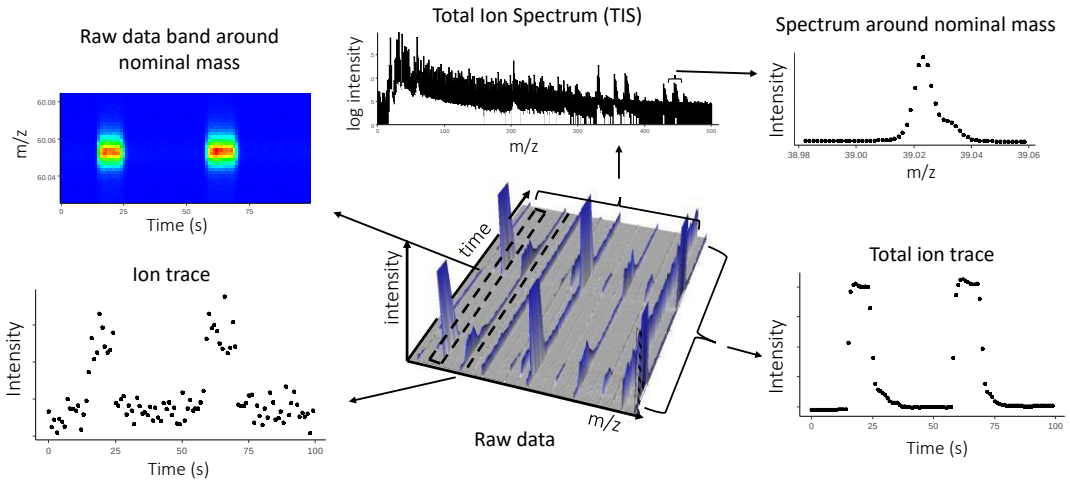


Figure 2.2: Extract of PTR-TOF-MS raw data from exhaled breath, and nomenclature used. The raw data are stored as a matrix of intensities, whose dimensions are the TOF values (converted to m/z values during calibration), and the acquisition time. A (mass) **spectrum** contains all points at the same acquisition time. The sum of spectra across all times is called the **total ion spectrum (TIS)**. An **ion trace** is the sum of all intensities recorded around an m/z value (similarly to the extracted ion chromatogram in liquid chromatography coupled to mass spectrometry), while the **total ion trace** is the ion trace across all m/z values (similar to the total ion chromatogram, TIC). Since VOCs have low weights ($<$500 Da), the signal is concentrated around nominal masses: data processing algorithms therefore process **raw data bands** around the nominal masses independently.

where (a, b) are calibration constants which depend on the distance travelled to the detector and the the accelerating voltage, and can be determined from the flight times of at least two ions of known m/z.

Experimentally, however, the linear relationship between t and $\sqrt{m/z}$ does not cross the origin (Guilhaus et al., 2000). This is a relatively minor effect that can be only observed on high resolution TOF instruments or at very low m/z. Cappellin et al. (2010) therefore proposed to add a third coefficient to Equation 2.1:

$$m/z = a + bt + ct^2 \tag{2.2}$$

Alternatively, Holzinger (2015) suggests to improve the mass accuracy by optimising the exponent parameter q:

$$m/z = (\frac{t-a}{b})^q \tag{2.3}$$

In practice, the Formula 2.1 remains the most used, especially by manufacturers for external calibration.

Choice of the reference peak

The first external calibration used to convert the TOF axis to m/z values usually does not provide sufficient accuracy. The parameters of the previous equation are then updated by selecting reference peaks with known m/z, called *calibration peaks*. The mass accuracy therefore depends on the choice of those peaks: they should be i) well distributed along the whole axis, ii) without neighbours at the same nominal mass, iii) present in all scans, and iv) not saturated. The following optimisation problem is then solved with non-linear optimisation algorithms:

$$\min_{\theta} \sum_{i} \left((m/z)_i - f_\theta(t_i)\right)^2$$

where f_θ is one of the equations linking m/z and time of flight tof (*i.e.* Equation 2.1, 2.2, or 2.3), θ the two or three parameters to be estimated, $(m/z)_i$ the exact mass to charge ratio of the calibration peaks, and t_i the observed tof of this compound in the mass spectrum. Note that a precise determination of the calibration peak centroids t_i is therefore critical to achieve a good mass accuracy (see the *peak detection* section).

The most often used reference peak in the literature is the isotope of the primary ion (since the primary ion itself is saturated), i.e. $H_3^{18}O^+$ at m/z 21.022 when the reagent ion

is H_3O^+. Additional ions have been used as calibration references, depending on the sample analysed (exhaled breath, atmospheric air, food etc.), including nitric oxide NO^+ (m/z 29.998), dioxygen O_2^+ (m/z 31.999), the isotope of water cluster (m/z 39.0326), or acetone $H_7C_3O^+$ (m/z 59.0491) (Müller et al., 2013; Cappellin et al., 2011a; Herbig et al., 2009; Trefz et al., 2018). Finally, the instrument itself continuously produces external ions (generally with high m/z) aimed at improving the calibration accuracy.

An alternative strategy avoiding the need for calibration peaks was proposed by Holzinger (2015), by determining an autonomous mass scale calibration based on the matching to a library of compound formulae generated *in silico*. The full calibration procedure relies on three evaluations: a first calibration is performed by combining any two of the largest 16 peaks and assumes that these peaks correspond to a pair of primary ions. For the standard operation mode based on proton transfer from H_3O^+, the pair of primary ions is H_3O^+ (m/z 19.018) and $H_2O.H_3O^+$ (m/z 37.028). The second step performs a variation of constants on the previous parameter values, by maximising the number of matches to the compound library. The third calibration computes parameters according to the "classical" user-specified m/z values. Among the three sets of parameters obtained, the one which maximises the matches to the library compounds is selected.

Calibration shift

Due to low changes of temperature and small variations of the PTR-TOF instrumental parameters, a drift of the mass accuracy over the acquisition time is observed (Cappellin et al., 2010; Müller et al., 2011; Holzinger, 2015). To correct this effect (especially for long term acquisitions), several calibrations are performed periodically to update the parameters values. Then, an interpolation of the time bins of each mass spectrum is performed.

2.2.2 Dead time correction

Instrumental dead times are caused by the finite time response of the multi-channel plate (MCP) detector and the amplifier–discriminator, when two or more ions arrive at the detector within a single data acquisition time bin of the time-to-digital converter (TDC; Müller et al. 2013). It can therefore lead to an underestimation of high-intensity ion signals and limit the dynamic range of the measurements. Titzmann et al. (2010) proposed to use a Poisson counting to correct this effect.

2.2.3 Peak detection on the mass spectra

The main challenge of peak detection for PTR-TOF-MS data is the presence of several peaks at one nominal mass (multiple peaks), as well as the asymmetric shape of the peaks. Based on the work by Titzmann et al. (2010), Müller et al. (2011) proposed a cumulative peak shape function computed from the data (which was later improved by Holzinger 2015) as

well as an iterative residual analysis algorithm. We will now present these peak detection algorithms according to the three steps described in Section 1.2.1: smoothing, baseline correction and peak finding.

Denoising and baseline correction

PTR-TOF-MS spectra are affected by two main sources of error: electronic random noise and saturation effects. The first issue is addressed by detecting the peaks on the TIS. Denoising is therefore not a critical point of PTR-TOF-MS processing: Holzinger (2015) and Cappellin et al. (2011a) use a smoothing filter (Savitzky-Golay filtering and Wavelet denoising, respectively; 1.2.1) to facilitate the subsequent peak detection.

The baseline in PTR-TOF-MS spectra especially affects those peaks that are close to saturation. Müller et al. (2013) and Cappellin et al. (2011a) thus used a local baseline correction at each unit m/z interval, by subtracting a linear fit (respectively, polynomial) computed between the upstream and downstream signal points. Holzinger (2015) also applied a local baseline algorithm, but on each 90 ns partition (1 bin corresponding generally to 0.2 ns). The algorithm consists of a 7-fold iteration of the two steps: localise the position of the highest signal and remove the signal located $\pm$ 9 ns around this position. The baseline is finally set to the mean of the remaining data.

Peak finding

Since several peaks may be present at one nominal mass, and since the mass resolution of the instrument is not always sufficient to separate them, a deconvolution step is required (see the section 2.2.3). The three algorithms described in the literature all use the same method: 1) detect local maxima on the average spectrum (i.e. average of all spectra acquired during the acquisition), and 2) use a peak model function to separate and quantify the peaks.

- Detection of local maxima

Since volatile organic compounds are molecules with a low weight ($<$ 500 Da), the signal is expected to be close to nominal masses (note that since most of the ions detected by the PTR-TOF-MS technology carry a single charge z=1, the measured m/z value therefore corresponds to the ion mass). Müller et al. (2011) thus propose to reduce the peak search to windows around each nominal mass $m \pm 0.3\ Da$. Peak detection then relies on a classical algorithm for the detection of local maxima (2.2.3), with two additional quality controls: a minimum distance of 1000 ppm between 2 peaks, and an adaptive noise threshold corresponding to the maximum of the signal around the nominal mass (i.e. in $[(m-1)+0.3; m-0.3[\cup]m+0.3; (m+1)-0.3]$).

Holzinger (2015) used the first and second derivatives of the Savitzky-Golay filter to de-

tect local maxima. Two quality controls were also added: the signal at each local maximum must exceed the noise by 8 times the variability of the noise, and the ratio between the maximum and the full width of the peak ($end - start$) must be in $[20; 10000]$. The noise is defined here as the median of the signal around the nominal mass $\pm 1\ Da$.

- Peak separation with the model peak function

In the case of signals based on counting, Titzmann et al. (2010) proposed to improve the peak analysis by using the cumulative signal for the fit. Indeed, for TOF data, the i^{th} data point of the spectrum represents the number of ions which arrived *within* the i^{th} time bin, and not exactly *at* this time bin. Thus the cumulative signal, corresponding to the cumulative distribution function (CDF), is more appropriate than the probability distribution function (PDF) of the ion (the latter corresponding to the classical peak shape).

Furthermore, Müller et al. (2011) proposed a peak shape function computed from the data for each single acquisition. It first derives a $referencePeak$, which corresponds to the average normalised cumulative signal of the calibration peaks (section 2.2.1) after baseline correction, and a normalised TOF range $TOF_\Delta = \frac{tof - t}{\Delta}$, where tof is the TOF axis obtained by converting the mass axis with the external calibration coefficient, t is the peak centre and Δ is the FWHM. The peak function is then defined as:

$$peak_i(tof, \Delta_i, t_i, A_i) = interpolation(tof, TOF_\Delta \times \Delta_i + t_i, referencePeak \times A_i) \quad (2.4)$$

with Δ_i the peak width, t_i the peak centre, A_i the area, and $interpolation$ the cubic interpolation function. Müller et al. (2013) also proposed the following initial values and boundaries for the parameters:

- $\Delta_i = a(\frac{t_i^2}{b^2} + 1)^{0.5}$, where (a, b) are the calibration coefficients determined in section 2.2.1 (this formula was introduced by Coles and Guilhaus 1994). The fitting constraints are empirically set to $\pm$ 6% of Δ_i
- A_i: sum of all data bins within an interval of $10 \times \Delta_i$
- t_i: apex of the detected local maximum, with boundaries set to $\pm\Delta_i/5$

Holzinger (2015) further improved the reference peak shape, by using all the peaks (i.e. not only the calibration peaks) with a maximum signal in the following range: (a) a predefined minimum (the default value is 800 counts), and (b) a maximum which is the larger of either 10 times the minimum signal or 1% of the maximum signal of the entire spectrum. This allows for a better generalisation of the peak function. The final peak shape function is obtained by computing the 10% quantile of all the selected peaks, which is further

smoothed by using a Savitzky–Golay filter.

Of note, a Gaussian peak function has been proposed by Cappellin et al. (2011a): despite its agreement with the data on the top part of the peak and its reduced computing time, Gaussian models cannot fit well the asymmetric tails of the peaks.

- Iterative residual analysis

To improve the peak separation of PTR-TOF-MS spectra, Müller et al. (2010) proposed an iterative residual analysis. Following the detection of the first local maxima and the fitting of the peak model, the smoothed fit residual is analysed for additional peak maxima. The procedure for the detection of local maxima is identical to the previous one, except that softer thresholds are applied: intensity higher than 8 times the standard deviation of the residual for intensity threshold and $\Delta_i/3$ for the minimum peak separation. If maxima are detected in the residual, they are added to the sum of model peaks for a second fit on the spectrum. This step is repeated until one of the following criteria is satisfied: the R2 criteria of the residuals is greater that 0.995, the total number of peaks reaches 5, the number of iterations reaches 3, or there is no new peak in the residuals.

2.2.4 Temporal estimation

PTR-TOF-MS instruments not only record the mass of the compounds, but also their evolution with time. In fact, during an acquisition, PTR-TOF-MS instruments continuously record mass spectra along time (e.g. 1 mass spectrum per second). Consequently, PTR-TOF-MS data from one acquisition (i.e. in one file) consists in a matrix of TOF counts with mass and time as dimensions.

In the previous sections, we reviewed the processing along the mass axis (applied to the sum of the mass spectra). Existing software further perform the global compound quantification during the whole acquisition by integrating the signal between the m/z boundaries along the time dimension.

More precisely, Holzinger (2015) integrates the raw signal contained within 2 standard deviations around the peak m/z apex. In case of overlapping integration boundaries, the common boundary is set at an equal distance between the neighbouring peak apexes, and a correction factor is applied to take into account the overlap (Holzinger, 2015).

Alternatively, Müller et al. (2011) computes specific intervals for each peak, and a superposition of the "model peaks" is fitted with the tight fitting constraints described above for each single spectrum. Finally, peak areas are TOF-MS duty corrected and saved together with additional peak information. Ultimately, signal counts in the resulting temporal evolution are scaled-up with a correction factor to match the intensities in the integrated spectrum.

2.2.5 Normalisation and quantification

An interesting property of the ionisation by PTR is the conversion of ion intensities into absolute quantities. This is achieved by normalising the ion intensities by the reagent ion (e.g. H_3O^+) intensities (Vlasenko et al., 2010; warneke et al., 2001), the reaction rate coefficient k between the VOC and the reagent ion (Hartungen et al., 2004; Cappellin et al., 2011b), and the residence time of the primary ions in the drift tube (Cappellin et al., 2012b). When a k for a specific VOC was not available, a standard value of $2 \times 10^{-9} cm^3 s^{-1}$ is used (Sekimoto et al., 2017). The final normalisation by the density of the air in the reaction chamber gives the absolute concentration of the VOC, expressed in part per billion (ppb).

2.3 Software

Two processing software tools for PTR-TOF-MS data have been described, the open-source **PTRwid** tool, developed by Holzinger (2015) in IDL language (Figure 2.3), and the commercial **Ionicon Data Analyzer (IDA)** released in 2020, based on the algorithms developed by Müller et al. (2013) (Figure 2.4). These software tools allow for the analysis of single files from high-resolution, TOF-MS acquisition, with the automatic calibration for PTRwid (see section 2.2.1). Both propose a csv file output, with the list of the peak m/z centres and their quantification in ppb or cps at each time point from the acquisition. They also suggest a putative chemical formula for each detected peak, by generating all possible chemical formulae $C_a C_b^{13} H_c O_d N_e$, with $a \in [1, 40]$, b $\in \{0, 1\}$, c $\in [max(1, a-9), a]$, $d \in [0, 5]$.

To address the issue of the analysis of multiple files, Holzinger (2015) proposed the "unified mass list" tool, that enables to align peaks from different samples: peaks detected in each individual file are first counted by bin, of width equal to the maximum between 1 mDa and 8 ppm; the corresponding histogram is then analysed for each nominal mass (smoothing with a running mean of 5 points, detection of local maxima and Gaussian fitting of 11 data points). The peak centre estimated on this histogram provides the so-called "unified peak list". For IDA, the analysis of multiple files consists in merging the files and analysing the total spectrum.

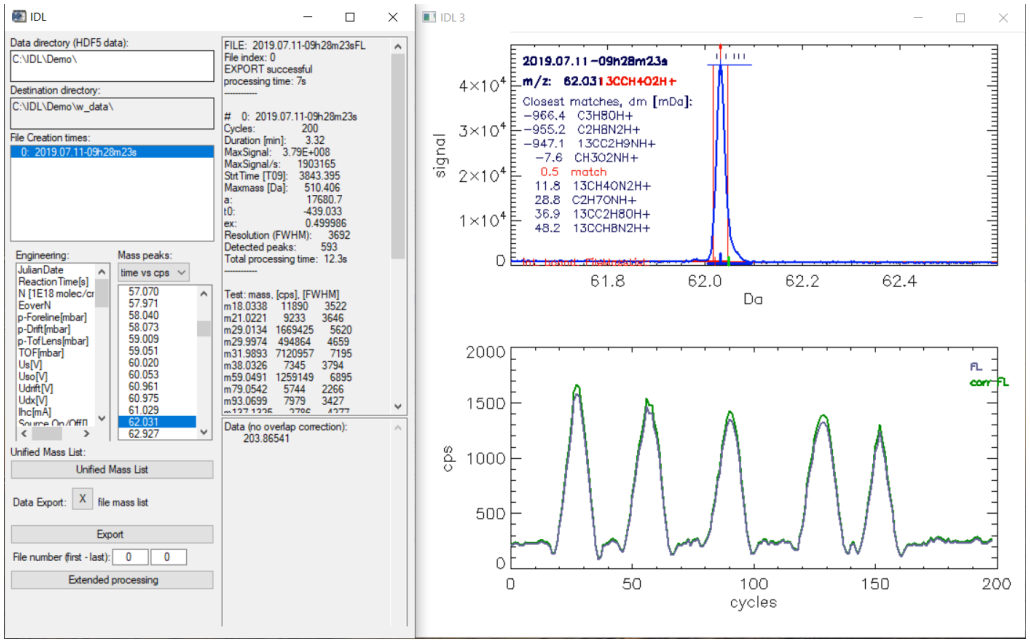


Figure 2.3: Screenshot from PTRwid, with opned simulated data (see the section 5.2.1)

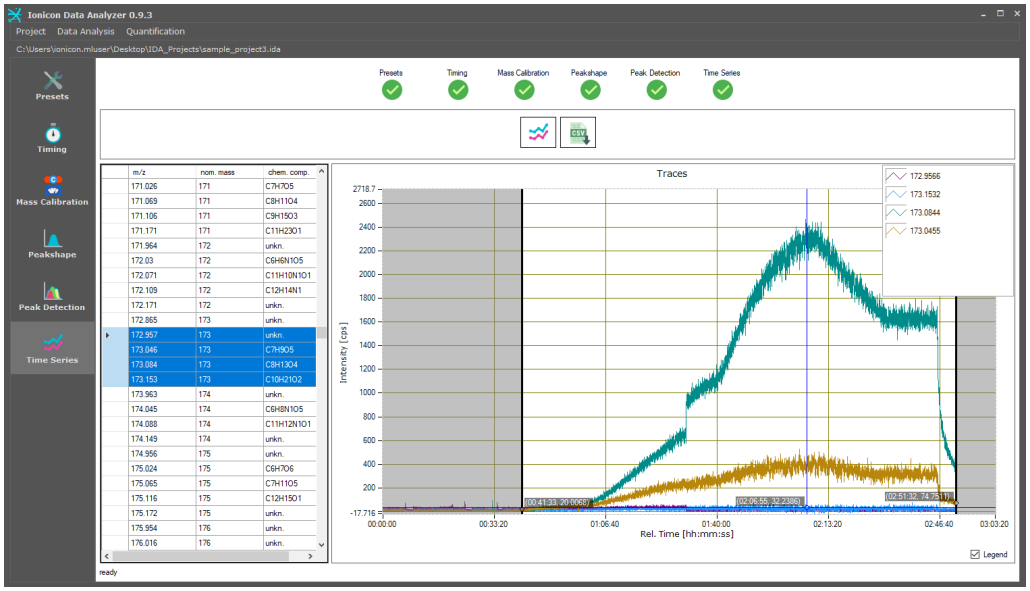


Figure 2.4: IDA software (screenshot from https://www.ionicon.com/accessories/details/ionicon-data-analyzer-ida).

Chapter 3

Mathematical approaches for classification and longitudinal analysis

3.1 Penalised spline regression

In this section, we present penalised spline regression, (Marx and Eilers, 2005; Wood, 2006; Bollaerts et al., 2006; Ruppert et al., 2009), which permits to estimate any shape of function without parametric assumptions and generalising well to multi-dimensions. In particular, the P-spline approach introduced by Eilers and Marx (1996) is very powerful to model any profile without *a priori* knowledge on the data and to provide interpretive coefficients and penalisation.

3.1.1 Penalised smooth regression

Let some data be $(y_i, x_i)_{i=1}^n$, we want to estimate a smooth function f, without any parametric assumption such that:

$$y_i = f(x_i) \quad \forall i \in 1, ..., n$$

f may then be expressed as a linear combination of K basis functions $(b_1(x), ..., b_K(x))$: $f(x) = \sum_{j=1}^{K} \beta_j b_j(x)$. We thus come back to a parametric linear model:

$$y_i = \sum_{j=1}^{k} \beta_j b_j(x_i) + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3.1}$$

The parameter β is then estimated by the least squares method:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{argmin} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k}\beta_j b_j(x_i))^2 \\ &= \underset{\beta}{argmin} \quad ||\mathbf{y} - \mathbf{X}\beta||^2 \\ &= (\mathbf{X^T X})^{-1}\mathbf{X^T y}\end{aligned}$$

where the i^{th} row of the $(n \times k)$ matrix $\mathbf{X}$, $\mathbf{X}_i = (b_1(x_i), b_2(x_i), ..., b_k(x_i))$, and $\mathbf{y} = (y_1, ..., y_n)$.

It then remains to choose the basis and the dimension k. Several basis of smooth functions could be used, for instance polynomial, Gaussian or spline. Grenn and Silverman (1994) demonstrate that splines are the best interpolators function in the sense of minimising the integrated squared second derivative of f on $[x_1, x_n]$. So, in order to estimate any shape of function, splines present themselves as ideal candidates.

Then to select the best dimension k, one way is to start with a large dimension, and then use hypothesis testing methods or AIC criteria to select K by backward selection. However such an approach is problematic since the fit of the model tends to depend on the basis function locations. An alternative to controlling smoothness is to add a "wiggliness" penalty to the least squares optimisation problem (Ramsay et al., 1996):

$$\underset{\beta}{min}||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \int_{x_1}^{x_n} f''(x)^2 dx$$

where λ is the smooth coefficient. Because f is linear in the parameters β, the penalty can always be written as a quadratic form in β : $\beta^T S\beta$, where S is a matrix of known coefficients (Wood, 2006). The optimisation problem (Equation 3.3) become:

$$\begin{aligned}\hat{\beta} =& \underset{\beta}{argmin}||Y - X\beta||^2 + \lambda\beta^T S\beta \\ \hat{\beta} =& (X^T X + \lambda S)^{-1} X^T Y\end{aligned} \tag{3.2}$$

If $\lambda \to \infty$, the fitted curve f approaches the standard linear regression to the observed data. In contrary, where $\lambda \to 0$ the curve will tend to become more and more variable, and at 0, f will approach an interpolant to the data, satisfying $f(x_i) = y_i \; \forall i$.

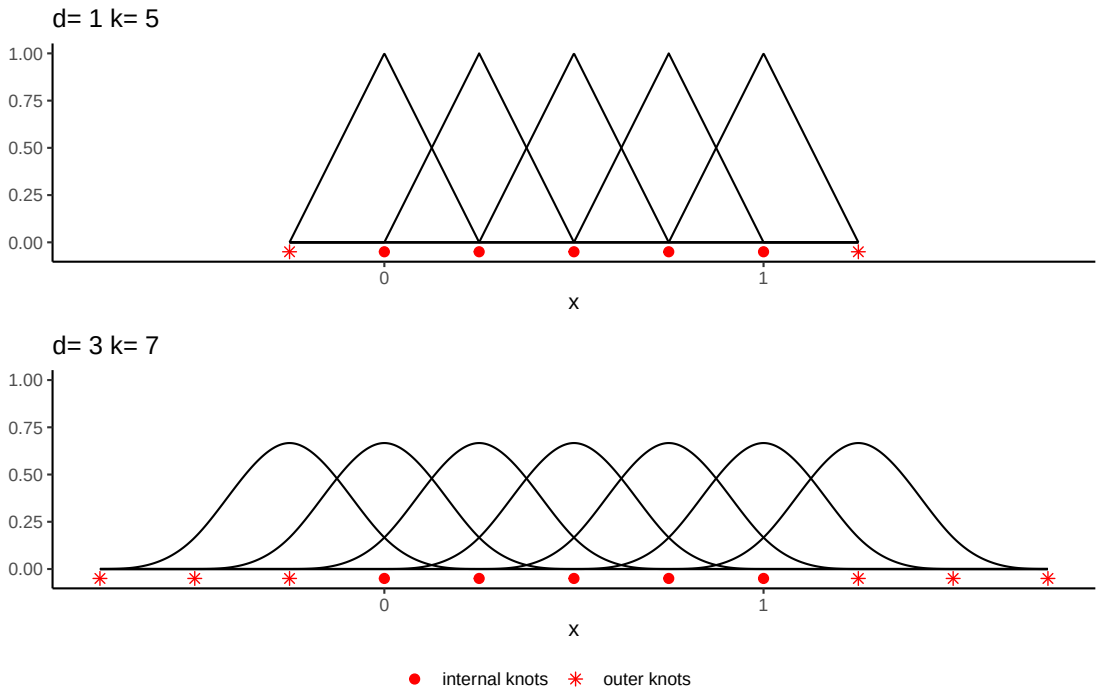


Figure 3.1: B-spline functions spread over a set of 5 equidistant knots for data between 0 and 1, with degree $d = 1$ (top panel) and $d = 3$ (bottom panel). This results in a basis of dimension 5 for $d = 1$ and 7 for $d = 3$.

3.1.2 P-splines

P-splines (penalised B-splines) are B-splines to which the penalty added corresponds to the difference between successive parameters β_i, to control the smoothness of the estimated function f. It has been used in many applications and theoretical works (Eilers et al., 2015) such as data smoothing (Currie and Durban, 2002), Bayesian statistics (Gressani and Lambert, 2021), and machine learning with generalised additive models (GAM; Brezger and Lang 2006; Wood 2006). Let us first introduce B-splines.

Basis splines (B-splines) are polynomial functions with a minimal compact support, introduced by de Boor (1978). To construct a B-spline basis of dimension k, $(b_1, ..., b_k)$, we first define a degree d and a set of knots $k_1, ..., k_q$ such that $k_1 < k_2 < ... < k_q$, and $q = k+d+1$. The first and last d knots are called outer knots, and the $k + 1 - d$ central knots are internal knots and must be located within $[x_{min}, x_{max}]$. Only the position of the internal knots impacts the estimation of the function f. Then each element b_i^d is a polynomial function a of degree d over the interval $[k_i, k_{i+d+1}]$, and zero otherwise:

$$b_i^d(x) = \frac{x - k_i}{k_{i+d} - k_i} b_i^{d-1}(x) + \frac{k_{i+d+1} - x}{k_{i+d+1} - k_{i+1}} b_{i+1}^{d-1}(x)$$

$$\forall i \in 1, ..., K \quad \text{and } b_i^0(x) = \begin{cases} 1 \text{ if } k_i < x < k_{i+1} \\ 0 \text{ otherwise} \end{cases}$$

A examples of B-splines basis with fives knots distributed between $[0, 1]$ are shown in Figure 3.1. B-splines were developed as a very stable basis for large scale spline interpolation (Unser et al., 1993), but the real statistical interest in B-splines has resulted from the work of Eilers and Marx (1996), by using them to develop P-splines. We add to the least squares regression a penalty on the difference between successive parameters β_i of order b:

$$\min_{\beta} \quad ||Y - X\beta||^2 + \lambda \sum_{j=b+1}^{k} \Delta^b(\beta_j)^2 \tag{3.3}$$

where $\Delta^b(\beta_j) = \Delta(\Delta^{(b-1)}(\beta_j))$ and $\Delta(\beta_j) = \beta_j - \beta_{j-1}$. For instance, if $b = 1$ and $k = 3$, the penalty could be written as follow:

$$\begin{aligned}
\mathcal{P} =& \sum_{j=2}^{3} (\beta_j - \beta_{j-1})^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + \beta_3^2 \\
\mathcal{P} =& \beta^T \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \beta \\
\mathcal{P} =& \beta^T \begin{bmatrix} 1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \beta \\
\mathcal{P} =& \beta^T D^T D \beta
\end{aligned}$$

where the $(k - b) \times k$ matrix D corresponds to the difference of successive row of the identity matrix of dimension k: $\Delta^b I_k$. For practical computation, the problem can be reformulated as follows (Eilers and Marx, 1996):

$$||Y - X\beta||^2 + \lambda\beta^T S^T S\beta = \left\| \begin{bmatrix} Y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} S \end{bmatrix} \beta \right\|^2$$

It simply corresponds to the unpenalised least squares problem, hence the model can be fitted by standard linear regression. An example of spline regression with B-spline and penalised spline regression with P-spline are shown on Figure 3.2. P-splines are extremely easy to set up and use, and allow a good deal of flexibility, in that any order of penalty can be combined with any order of B-spline basis.

3.1.3 Penalty, knots location and basis dimension

We now discuss the choice of the penalty parameter λ, the knot location and the dimension K.

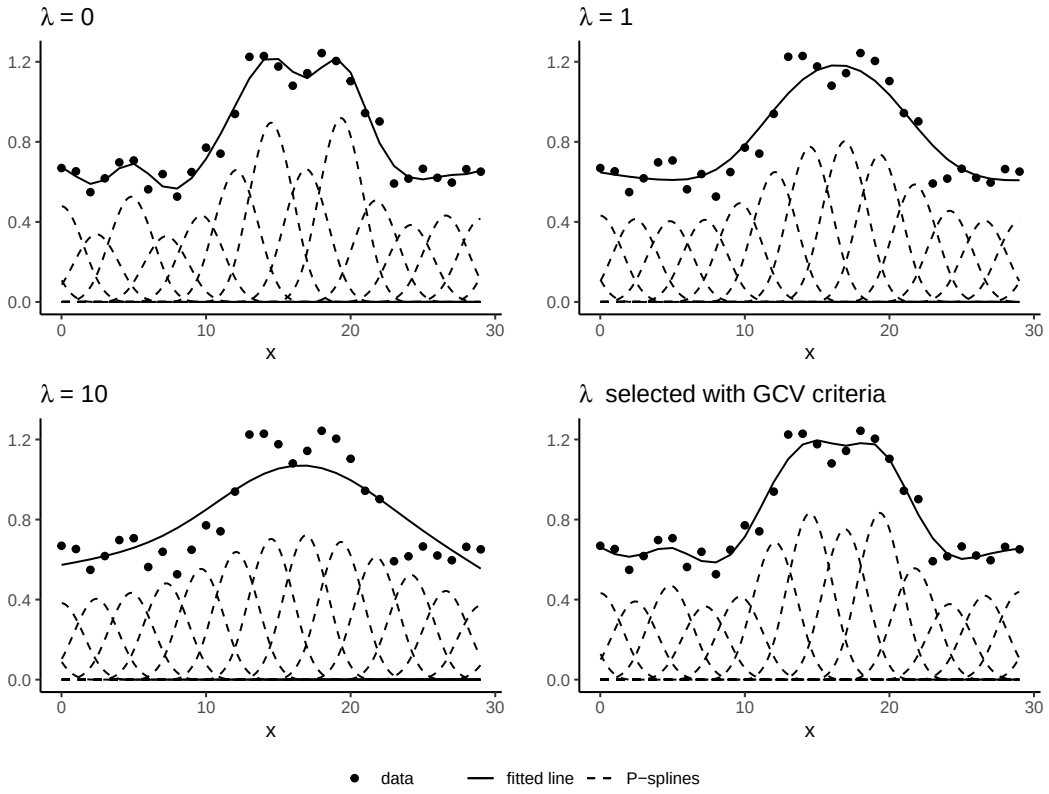


Figure 3.2: Example of penalised spline regression with a P-spline basis of dimension 15 with a penalty difference of order 2, for different values of the penalisation parameter λ. We observe that too high penalties lead to underfitting, and the contrary to overfitting.

The knots can be distributed equidistantly over the data interval, or concentrated around the information in the case of unequal distribution (by using quantiles for instance). For P-splines, the choice of the knots location is generally equidistant in order to make the penalty interpretive, since the penalty parameter is the same for the difference of all successive knots.

Regarding penalty and dimension, a too high penalty and a too small basis dimension result to underfitting, whereas a too small penalty and a too large basis lead to overfitting. It is therefore important to find a trade-off between bias and variance. Kim and Gu (2004) showed that the basis size should scale as $10n^{\frac{2}{9}}$, where n is the number of observation. However, it is important to note that the exact size of the basis dimension is not really critical, since the smoothing parameter controls the actual effective degrees of freedom. The basis dimension is a mere upper bound to the flexibility of the function.

Generalized Cross Validation (GCV)

The smooth penalty parameter λ can be estimated by cross validation. This method consists in separating the data into k sub-parts, and successively using one part for testing and the others for training. The error metric of each sub-model on the test data are then averaged to get the Cross Validation (CV) criterion. In case k is equal to the number of data n, the Leave One Out Cross Validation (LOOCV) is computed as:

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f_{\hat{\beta}}^{-i}(x_i))$$

where $f_{\hat{\beta}}^{-i}$ is the function estimated on all the data excepted x_i and $\mathcal{L}$ is a loss function (e.g. squared error loss, absolute error, indicator function). To get the best parameter λ, the simplest way is to compute the $LOOCV$ on a grid of λ values, and choose the one which gives the lowest $LOOCV$ error. But this method can be very time consuming, since it requires the training of $n \times$ the size of the grid search models. However, in case of penalised regression, calculating CV by performing n model fits is unnecessary. The GCV criterion can be used, which is approximately equivalent to $LOOCV$ and can be derived from the model fit and the whole data set (proof in Golub et al. 1979) :

$$GCV = \frac{n||Y - X\hat{\beta}||^2}{(n - tr(A))^2}$$

where A is the influence matrix (or hat matrix) of the model (Equation 3.2): $A = X(X^TX + \lambda S)^{-1}X^T$. An example of the influence of the λ parameter is show in Figure 3.2.

3.1.4 Multidimensional penalised regression

The penalised spline regression theory can be generalised to multidimensional smoothing. Here, we illustrate the two-dimensional case (Marx and Eilers, 2005; Wood, 2006; Dierckx, 1995; Durban et al., 2002). The objective now is to estimate a function f such that:

$$y_{ij} = f(x_i, z_j) \quad \forall i \in 1, ..., n_1 \quad and \quad \forall j \in 1, ..., n_2$$

To do so, we choose a basis for each axis (not necessary the same) $(b_1(x), .., b_{K_x}(x))$ and $(a_1(z), .., a_{K_z}(z))$:

$$f_\delta(x) = \sum_{i=1}^{K_x} \delta_i b_i(x) \quad f_\alpha(z) = \sum_{j=1}^{K_z} \alpha_j a_j(z)$$

A two dimensional function is then obtained by multiplying each coefficient of each basis term to term (tensor product):

$$f_\beta(x, z) = \sum_{i=1}^{K_x} \sum_{j=1}^{K_z} \beta_{ij} b_i(x) a_j(z) \quad \text{with } \beta_{ij} = \delta_i \times \alpha_j$$

This function can be written in a matrix format: $f_\beta((x_1, ..., x_{n_1}), (z_1, ..., z_{n_2})) = X\beta$

with: $\beta \in \mathbb{R}^{(K_x \times K_z)}, X \in \mathbb{R}^{(n_1 \times n_2) \times (K_x \times K_z)}$

$X_i = X_{xi} \otimes X_{zi}$, where $\otimes$ represent the kronecker product, and X_i the i^{th} row of X

$X_{xi} = (b_1(x_i), ..., b_{K_x}(x_i)),\ X_{zi} = (a_1(z_i), ..., a_{K_z}(z_i))$

A specific penalty may be applied to each axis:

$$\min_\beta ||Y - X\beta||^2 + \lambda_x \beta^T P_x \beta + \lambda_z \beta^T P_z \beta$$

$$\min_\beta ||Y - X\beta||^2 + \beta^T S \beta \quad \text{with } S = \lambda_x \beta^T \tilde{P}_x \beta + \lambda_z \beta^T \tilde{P}_z \beta$$

and $\tilde{P}_x = P_x \otimes I_{K_z} \quad \tilde{P}_z = I_{K_x} \otimes P_z$, where I_K is the identity matrix of dimension K. This brings us back to a similar 1D optimisation problem as in Equation 3.2. Examples of 2-dimensional B-spline basis built with the tensor product are shown in Figure 3.3.

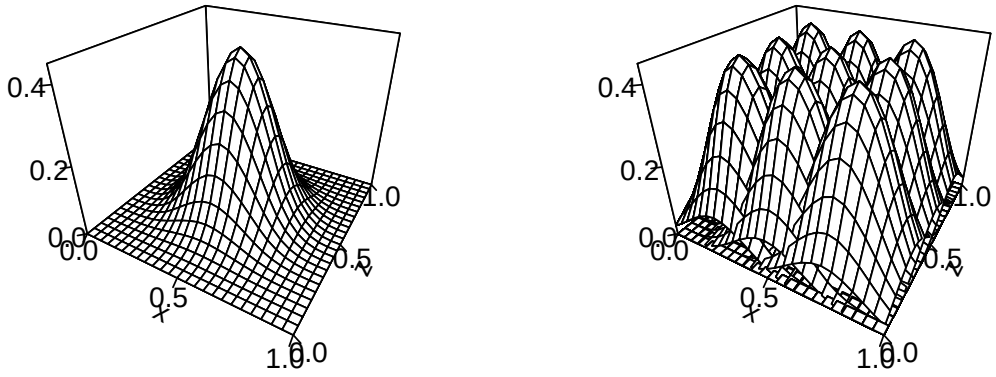


Figure 3.3: 2D B-spline consisting of a single (left) or 9 (right) basis functions, with equidistant knots between 0 and 1.

3.2 Statistical learning for biomarker discovery

The objective of biomarker discovery is to find a significant feature subset with optimal predictive properties (see the section 1.1.1). The pre-processing of data from volatolomics, or more generally metabolomics, usually provides hundreds of features for one sample, resulting in a data table with a high feature over sample ratio, and incomplete, noisy, and collinear data structures (Trygg et al., 2007a). Classification methods with feature selection and reduction of dimension are therefore needed to avoid over-fitting and prediction variability. In addition, clinical studies are often designed as longitudinal data, with multiple measurements of the same individual over time, to increase the statistical power.

We therefore present in this section state of the art methods to address these three issues: classification models in the case of dimensions greater than the number of samples, feature selection methods and time course modelling.

3.2.1 Classification

We denote the random variables $Y \in \{0, 1\}$ and $X = (X^1, ..., X^p)$, and their observations $\mathbf{y} = (y_1, ..., y_n)^T$ a $n \times 1$ vector and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ a $n \times p$ matrix with $\mathbf{x}_i = (x_i^1, ..., x_i^n)^T$, where $\mathbf{X}$ are the predictor variables and $\mathbf{y}$ the response. We suppose that $p > n$. We want to construct a decision rule from the observations that enables to predict Y. To do so, we search for a prediction function $\hat{Y}(x)$ that minimises the risk function for classification $P(\widehat{Y}(X) \neq Y)$. We describe hereafter four reference supervised machine learning approaches adapted to high dimension, namely Elastic Net, Random Forest, Support Vector Machine (SVM) and Partial Least Squares - Discriminant Analysis (PLS-DA). In addition, we present associated feature selection methods, either based on the ranking of features according to their contribution to the prediction (Random Forest, PLS-DA, SVM), or based on integrated sparse constraints (Elastic-Net). Furthermore, hypothesis testing to discrim-

inate two classes are introduced.

Elastic Net

The Elastic Net model, developed by Zou and Hastie (2005), is a regularised logistic regression that linearly combines the L1 and L2 penalties of the lasso (Tibshirani, 1996) and ridge (Hoerl and Kennard, 1988) regression. It is particularly useful when the number of predictors is much larger than the number of observations, because it includes variable selection within the model-building procedure by setting the smallest coefficients to zero (unlike ridge regression). In addition, if a group of variables is highly correlated, and one of the variables is selected, the whole group is automatically included (see below), which is not the case in the lasso approach.

The model assumes that Y follows a Bernoulli distribution conditional to X, and that:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + X\beta)}} \tag{3.4}$$

The $p + 1$ coefficients (β_0, β) are then estimated by maximising the (or minimising the negative) penalised log-likelihood function of the model $\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{X})$ with the LARS-EN algorithm proposed by Zou and Hastie (2005) based on the Least Angle Regression, which is similar to forward stepwise regression (see the section 3.2.2), but instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual (Efron et al., 2004):

$$\hat{\beta} = \underset{\beta}{argmin} \left\{ -\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{X}) + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||^2 \right\} \tag{3.5}$$

Features which are not selected by the model get their coefficients set to zero. Interestingly, Zou and Hastie (2005) demonstrate that the two penalties L1 and L2 lead to selecting 'grouped' correlated variables, since if two variables i and j have a correlation ρ close to 1, the difference between the coefficient $\hat{\beta}_i$ and $\hat{\beta}_j$ is bounded by $(1 - \rho)$.

The parameters (λ_1, λ_2) are usually tuned by cross validation, in order to find the right balance between the bias and variance, and to minimise the miss-classification error.

The Elastic Net produces a sparse model with a good prediction accuracy, while encouraging a grouping effect.

Random Forest

Random forests, introduced by Breiman (2001), are a combination of decision tree predictors, such that each tree is built with a bootstrap sampling of observations and a random

Predictor	Regularisation parameters	Ranking feature metrics
$\widehat{Y}(\mathbf{X}) = \begin{cases} 1 \text{ if } \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}\mathbf{X})}} > 0.5 \\ 0 \text{ otherwise} \end{cases}$	penalty coefficients λ_1 and λ_2	absolute values of non-zero $\hat{\beta}$ coefficients

Table 3.1: Elastic net summary

subset of features, independently and with the same distribution for all trees. The predicted class is then assigned by a majority vote: each tree provides a class according to its own classifier, and then the most frequent class from the ensemble of trees is returned.

Random forest is reported as an excellent and fast classifier, with simple theory. Overfitting in the case of a higher number of features than observation is prevented by using different subsets of the training data and different subsets of features for training the individual trees (Biau and Scornet, 2015). It has already been applied to many metabolomics data in clinical studies aiming at biomarker discovery (Touw et al., 2012; Chen et al., 2013).

Two methods have been proposed to measure the feature importance in the model, namely the Gini importance (Breiman, 2001) and the permutation accuracy importance (Seoane et al., 2014). The latter estimates the decrease of the prediction performance when the values of that variable are randomly permuted within the out-of-bag observations.

Predictor	Regularisation parameters	Ranking feature metrics
$\hat{Y}(\mathbf{X}) = \begin{cases} 1 \text{ if } \sum_{i=1}^{K} \mathbb{1}_{\widehat{Y}^{T_i}=1} > \sum_{i=1}^{K} \mathbb{1}_{\widehat{Y}^{T_i}=0} \\ 0 \text{ otherwise} \end{cases}$ where $\widehat{Y}^{T_i}$ is the prediction of the i^{th} tree of the forest and K is the number of trees	maximum depth of a tree in the forest; maximum number of features in the leaf (last node of the tree) and number of variables to be randomly drawn for each individual tree	Feature importance

Table 3.2: Random forest summary

Support Vector Machine

Support Vector Machine (SVM), introduced by Boser et al. (1996); Vapnik (1995), maximise the distance between the training data set and the decision boundary between two different classes. The underlying hypothesis is that the larger this marge is, the better the generalisation error of the classifier will be.

SVM works well both in situations when the separation between classes is linear or not, and is effective in cases where the number of dimensions is greater than the number of samples. Many studies have already demonstrated the potential of SVM for biomarker discovery in metabolomics and mass spectrometry data (Marchiori et al., 2006; Mahadevan et al., 2008; Heinemann et al., 2014).

In this section, the observed responses y_i will be assumed to be in $\{-1, 1\}$. The separating hyperplane (or decision function) $D(x)$ is formulated as a function of a Kernel K:

$$D(x) = \sum_{j=1}^{n} \alpha_j K(x_j, x) + b \tag{3.6}$$

Kernel functions represent dot products in the feature space. They enable the algorithms to be used in a feature space without explicitly carrying out computations within that space. According to Aronszajn (1950), kernels can be written $K(x, x') = \sum_{i=1}^{N} \psi_i(x)\psi_i(x')$, with ψ any function of an Hilbert space and N the kernel dimension. Equation 3.6 is thus equivalent to:

$$\begin{aligned} D(x) &= \sum_{j=1}^{n} \sum_{i=1}^{N} \alpha_j \psi_i(x_j)\psi_i(x) + b \\ D(x) &= \sum_{i=1}^{N} w_i \psi_i(x) + b \quad \text{with } w_i = \sum_{j=1}^{n} \alpha_j \psi_i(x_j) \end{aligned} \tag{3.7}$$

α are called dual parameters, w direct parameters, and b the bias. To estimate them, the margin between the class boundary and the training points is formulated in the direct space of Equation 3.7, by maximising the normalised distance of any training point x to the hyperspace: $d(x, D) = \frac{D(x)}{||w||}$. This is equivalent to minimising $||w||$ under the constraint that the observation $\mathbf{x}$ is assigned to the good class, i.e. $sign(D(\mathbf{x_i})) = sign(\mathbf{y_i})$:

$$\begin{aligned} min \quad & ||w||^2 \\ s.t. \quad & y_i D(x_i) \geq 1 \quad \forall i \end{aligned} \tag{3.8}$$

This problem is then transformed into the dual space by using the Lagrange multiplier method. This results in a quadratic optimisation problem with linear constraints. The latter can be easily resolved by numeric optimisation algorithms such as descent methods.

A regularisation parameter λ may be added to the dual optimisation problem on the α coefficient. This approach, called Kernel Ridge Regression (Saunders et al., 1998), was designed to reduce over-fitting resulting from the "curse of dimensionality".

Predictor	Regularisation parameters	Ranking feature metrics
$\hat{Y}(\mathbf{X}) = sign(\hat{\mathbf{w}}^T \psi(\mathbf{X}) + \hat{b})$	regularisation parameter λ	values of w in case of a linear kernel

Table 3.3: SVM summary

Partial Least Square - Discriminant Analysis (PLS-DA)

Partial Least Squares regression (PLS) is a dimension reduction technique based on latent variables that maximises the covariance with the response (in contrast to principle component analysis, which maximises the variance of the components; Wold et al. 2001). It was developed in the late 60s by Herman Wold, and later applied by his son Svante Wold to high dimension and multi-collinear datasets (Wold et al., 1984; Brereton and Lloyd, 2014; Fordellone et al., 2018). It finds a linear regression model by projecting the predicted variables and the observed variables into a new space. PLS was later extended to classification problems by using a dummy matrix Y as the response (Barker and Rayens, 2003).

In parallel, Trygg and Wold (2002) proposed to include within the PLS algorithm an orthogonal signal correction filter to remove systematic variation in the predictors (i.e. variation from X that is not correlated to Y). The resulting model, called Orthogonal Partial Least Squares (OPLS), has similar performances compared to PLS, but facilitates interpretation. In particular, OPLS models of a 1-dimension **y** response have a single predictive component (Trygg and Wold, 2002). Due to its ability to perform well with high dimension and multi-collinear datasets, the PLS approach is very popular in metabolomics (Trygg et al., 2007b), e.g. for mass spectrometry, nuclear magnetic resonance, or near-infrared data.

In the classical multi-linear modelling $Y = X\beta + \epsilon$, the least square solution is $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. In cases where the number of predictors is larger than the number of observations, or when multi-collinearity is present, the $\mathbf{X}^T\mathbf{X}$ matrix becomes singular.

PLS solves this problem by decomposing the data matrix $\mathbf{X}$ and the response $\mathbf{y}$ into k orthogonal scores (components) in the form of a $(n \times k)$ matrix $\mathbf{T}$, and two loading matrices $\mathbf{P}$ and $\mathbf{Q}$, of respective dimensions $(p \times k)$ and $(1 \times k)$:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \epsilon \\ \mathbf{y} &= \mathbf{T}\mathbf{Q}^T + \nu \end{aligned} \tag{3.9}$$

where ϵ and ν are error vectors of independent and identically distributed random normal variables. A weight matrix W $(p \times k)$ is then defined as:

$$\mathbf{T} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}$$

and by substitution into Equation 3.9, the model becomes:

$$\mathbf{y} = \mathbf{XW}(\mathbf{P^T W})^{-1}\mathbf{Q^T} + \nu$$

The majority of the PLS algorithms (such as NIPALS for *non-linear iterative partial least squares*; Wold et al. 2001) estimate the components iteratively and are numerically stable (Andersson, 2009).

To measure the role of the original variables in the model, Wold et al. (2001) proposed the Variable Influence on Projection metric (VIP):

$$VIP_m^2 = p \times \frac{\sum_{j=1}^{k} w_{mj}^2 \times SSY_j}{\sum_{j=1}^{k} SSY_j} \tag{3.10}$$

for the m^{th} variable, where SSY_j is the sum of squares of $\mathbf{y}$ explained by component j, and k and p are the number of components and features, respectively.

To overcome the problem of high dimension, Sparse PLS (sPLS; Cao et al. 2011) perform simultaneous variable selection, including the LASSO penalisation on loading vectors of the matrix $\mathbf{P}$ and $\mathbf{Q}$ to reduce the number of original variables.

Finally, the optimal number of components, k, is generally selected by cross validation based on Q^2, AIC, or BIC criteria (Wold, 1978; Kvalheim et al., 2018; Nengsih et al., 2019). Szymańska et al. (2012) also proposed permutation test in addition to cross validation, based on the random permutation of the response. The model obtained are expected to be less efficient than with original (non permuted) data.

Predictor	Regularisation parameters	Ranking feature metrics
$\hat{Y}(\mathbf{X}) = \mathbf{X}\hat{W}(\hat{P}^T\hat{W})^{-1}\hat{Q}^T$	Number of components	Variable Influence on Projection (VIP)

Table 3.4: PLS summary

Hypothesis testing

Hypothesis testing is commonly used either for dimension reduction (Saccenti et al., 2013), or for binary classification Li and Tong (2020), as it provides a mathematical framework to infer the difference of behaviour of each feature X^k between several groups. We first

present Hypothesis testing as a classification method.

Let $X_0^k = X^k \cap (Y = 0)$ (respectively, X_1^k) the k^{th} explanatory feature when $Y = 0$ (respectively, $Y = 1$). The test is called parametric if a probabilistic law is used for X_1^k and X_0^k, and non-parametric if there is no assumption of such a law. A statistical test is defined by two opposite hypotheses H_0 (null hypothesis) and H_1 (alternative hypothesis). The critical region $\mathcal{R}_\alpha(x)$, a sub-set of observations leading to the H_0 to be rejected, is defined for a significance level α such that: $P_{H_0}(X \in \mathcal{R}_\alpha(x)) \leq \alpha$, where $P_{H_0}(X \in \mathcal{R}_\alpha)$ is the probability to reject H_0 under the H_0 hypothesis (Type 1 error; Taeger and Kuhnt 2014).

The *p*-value is then defined as the minimum probability to reject wrongly H_0:

$$p = inf\{\alpha; \quad s.t \quad x \in \mathcal{R}_\alpha\}$$

The smaller the *p*-value, the higher the evidence against H_0 (conversely, however, a *p*-value close to 1 does not mean that there is strong evidence in favour of H_0). The *p*-value thus reflects only the non-matching to H_0 (Thiese et al., 2016). We describe here two tests:

- Student's *t*-test:

  The *t*-test (Student, 1908) is a parametric test, with assumption of normality: $X_0^k \sim \mathcal{N}(\mu_0, \sigma^2)$ and $X_1^k \sim \mathcal{N}(\mu_1, \sigma^2)$. The test is then formulated as follows :

  $$H_0 : \mu_0 = \mu_1 \quad vs \quad H1 : \begin{cases} \mu_0 \neq \mu_1 \text{ bilateral} \\ \mu_0 > \mu_1 \text{ unilateral} \\ \mu_0 < \mu_1 \end{cases}$$

  Let us denote $\bar{X}$ the empiric mean of the X_i^k under H_0: we have $\sqrt{n}\frac{\bar{X}-\mu_0}{\sigma} \sim \mathcal{N}(0,1)$, and, according to Cochran theorem, $(n-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$ with the estimator of variance $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. We end up comparing the *t*-statistic to the $\mathcal{T}(n-1)$ Student's law with $n-1$ degree of freedom:

  $$Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}}} \sim \mathcal{T}(n-1) \tag{3.11}$$

- Wilcoxon-Mann-Whitney test

  A commonly used non-parametric counterpart of the *t*-test is the Wilcoxon - Mann-Whitney test (or Mann – Whitney *U* test; Wilcoxon 1945), which is based on ranks.

Under the null hypothesis H_0, the distributions of both populations are equal:

$$H_0 : P(X_0^k < X_1^k) = P(X_1^k < X_0^k) = \frac{1}{2} \quad vs$$

$$H1 : \begin{cases} P(X_0^k < X_1^k) \neq \frac{1}{2} \text{ bilateral} \\ P(X_0^k < X_1^k) > \frac{1}{2} \text{ unilateral: } X_1^k \text{ greater than } X_0^k \\ P(X_0^k < X_1^k) < \frac{1}{2} \text{ unilateral: } X_1^k \text{ lower than } X_0^k \end{cases}$$

No assumption of the probabilistic law for X_i^k is required: only the ordered values of the observations $(x_1^k, ..., x_n^k)$ are used. Let us note the statistic $U_0 = \sum r_i$ the sum of ranks of the x_i^k for $y_i = 0$, and U_1 the sum of ranks for $y_i = 1$. Under H_0, X_0^k and X_1^k have the same distribution, as well as U_0 and U_1. The distribution of the sum of ranks can then be asymptotically approximated by a normal distribution (Mann and Whitney, 1947; Iman, 1974). This test can be used even if the observations of the two classes are of different sizes, and it can be also adapted to paired data (Wilcoxon, 1945).

To select features with a significant difference in means (or medians) between the two groups, all features are first tested independently, resulting in a vector of *p*-values $(p_1, ..., p_p)$. A correction must then be applied to the results of these multiple tests (Burger, 2017), since using the α threshold directly for each p_j would result in a global increase of false positives: in the case of p independent comparisons, the number of false positives, or Family-Wise Error Rate (FWER) is $1 - (1 - \alpha)^p$. A mean of controlling the FWER (i.e. of controlling the probability of at least one Type 1 error) is to use the α/p threshold for each test (Bonferroni correction).

A less stringent criterion is usually applied in omics studies, which focuses on controlling the False Discovery Rate (FDR), i.e. the expected proportion of "discoveries" (rejected null hypotheses) that are false. In particular, Benjamini and Hochberg (1995) demonstrated that selecting features such that $p_{(i)} < \frac{i}{p}\alpha^*$ controls the FDR to α^* (where $p_{(i)}$ are the ordered *p*-values, i the rank, and p the number of features).

3.2.2 Feature selection

Feature selection consists in selecting a subset of relevant features used for a predictive model. A high number of features in a data set, larger than the number of samples, leads to model over fitting. Furthermore, selecting the most promising candidates between the first untargeted step and the subsequent validation phases is critical in the biomarker pipeline (section 1.1.1).

Methods for features selection can be classified in three categories (Jović et al., 2015):

- Filter methods: select variables before building the model, based on a criterion independent from the performance of the classification algorithm, such as the *p*-value from statistical hypothesis tests (correlation, Chi-square), or multivariate metrics (e.g., Variable Importance), and a fixed threshold

- Embedded methods: select features during the building of the model; examples include regularisation models (Lasso, Elastic Net), or variant algorithms from SVM (Weston et al., 2000), and RF (Genuer et al., 2010).

- Wrapper methods: train a model iteratively with several subsets of features, and select the one that gives the best predictive performance. The most known subsetting strategies are: (i) forward selection, starting with an empty feature set, and then adding one or more features at each iteration, (ii) backward elimination, starting with the whole feature set, and removing one or more features, (iii) bidirectional selection (stepwise), from an empty set or from the whole set, simultaneously considering larger and smaller feature subsets, or (iv) heuristic selection, that generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further.

On the one hand, filter methods are the fastest and the simplest approaches. However, since the filtering is performed before the training of the model, the selected features may not be optimal for the classifier performance. In addition, the choice of the threshold is arbitrary. On the other hand, embedded methods usually achieve good prediction performances, while still being computationally efficient. They are, however, specific to a single type of classifier, which may be a limitation when one wants to compare several approaches with distinct mathematical backgrounds (to maximise the chance of finding an optimal classification).

The wrapper feature selection methods are then a good trade-off. One of them is recursive feature elimination (RFE). It has been applied successfully to several machine learning algorithms, including SVM (Guyon et al., 2002; Sanz et al., 2018), RF (Granitto et al., 2006), and PLS (Sahran et al., 2018). It is a backward recursive process, which iteratively ranks features according to a measure of their importance (related to the algorithm used) and removes the weakest one(s). There are several possible stopping criteria, such as: run until the feature subset is empty, or until the model performance reaches a threshold, or until the performance does not improve from one iteration to the next.

A limitation of RFE is that the selection criterion is based on the classifier performance only: the added-value of including a particular group of features instead of noise into the model (i.e. the statistical significance of the selection for the model performance) is not evaluated. Rinaudo et al. (2016) therefore proposed a wrapper algorithm based on random permutation of feature intensities within the test subsets (obtained by resampling),

to assess the significance of the features on the model performance.

Performance metrics for binary classification models

- **Confusion matrix**: summary table of the prediction results; correct and incorrect predictions are highlighted and divided by class.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- **Accuracy**: global prediction of the model. It is a valid choice of evaluation for classification problems which are well balanced:

$$Accuracy = \frac{TP + TN}{\text{number of samples}} \tag{3.12}$$

- **Sensitivity (or recall, or true positive rate, or power)**: percentage of true positives which are well predicted; it reflects the ability to detect the disease among ill patients

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.13}$$

- **Specificity (or true negative rate)**: percentage of the true negative which are well predicted; it reflects the ability reject the disease status among healthy individuals

$$Specificity = \frac{TN}{TN + FP} \tag{3.14}$$

- **AUC, area under the Receiver Operating Characteristic (ROC) curve**: the ROC curve displays the sensitivity against (1 − specificity); the AUC indicates how well the probabilities from the positive class are separated from the negative class. Thus, AUC values above 0.5 indicate better prediction performances than a random guess (a value of 1 corresponding to a perfect classifier)

- **Log Loss (or logistic loss, or cross-entropy loss)**: when the output of a classifier is a prediction probability p, it measures the uncertainty of the model. In the case of two models with equal accuracies, it will favour the model that predicted probabilities which distinguish more strongly the classes. It is useful to compare models on the basis of their probabilistic outcome (the lower the Log Loss value, the better the prediction).

$$Log\ loss = -\sum_{i}^{n}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{3.15}$$

3.2.3 Time-course modelling

In this section we seek for modelling longitudinal data, where individuals are measured repeatedly thought time, in contrast to cross-sectional data where only a single response is available for each person. This is a common problem in clinical analysis, when we want to study the effect of a pathology or a treatment over time.

Let $y(t_{ij})$ be a response variable obtained for an individual i at different time levels j and possibly under changing experimental conditions (eg. treatment administered from a date, change of climatic conditions), and c_i be a categorical variable (we restrict ourselves here to the number of 2 classes of patients). We want i) to model y as a function of time by taking into account the individual effects, and ii) to test if there is a different evolution between the two groups in time. This could results in unbalanced data sets, and general multivariate models are not suitable for this analysis due to the covariance structure between individual measurement (all observations are not independent).

Mixed-effect models (Harville, 1977; Laird and Ware, 1982; Demidenko, 2004; Galecki and Burzykowski, 2013; Pinheiro and Bates, 2000) are well suited for the analysis of longitudinal data, because they include multilevel random effects (which allow data from the same individual to be combined) and explicit modelling and analysis of between and within individual variation. These models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors. Examples of such grouped data include longitudinal data, repeated measures data, multilevel data, and block designs.

Mixed effect model

Mixed-effects model are defined as follows:

$$y(t_{ij}) = \underbrace{f(t_{ij})}_{\text{fixed effect}} + \underbrace{g_i(t_{ij})}_{\text{random effect}} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad iid$$

This means that there is a common mean longitudinal response across all individuals (fixed effect f) and an individual-specific deviation from this mean curve (random effect g_i). To illustrate and demonstrate the proprieties of mixed-effects models, let us start with the linear case.

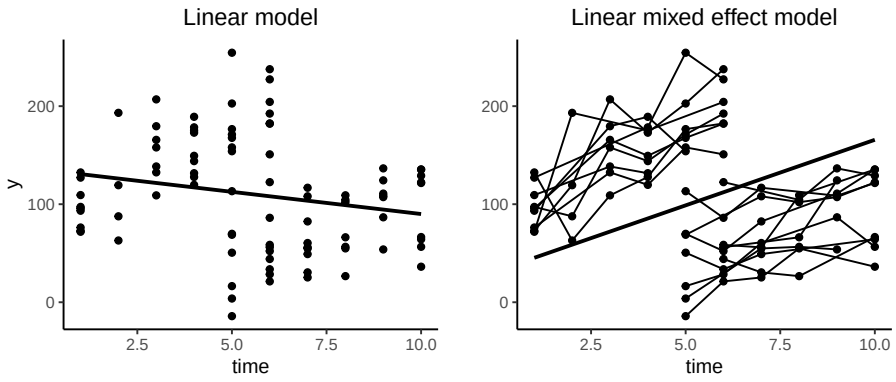


Figure 3.4: Comparison between classical linear modelling (left panel; all points are considered independent), and linear mixed-effects modelling as in Equation 3.17 (right panel; points of the same individual are connected). As evidenced with these simulated data, the two models lead to opposite conclusions (Simpson's paradox).

Linear Mixed Effect model (LME)

Laird and Ware (1982) define linear mixed-effects (LME) as:

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_\mathbf{i} \quad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I_n}) \quad \mathbf{b_i} \sim \mathcal{N}(0, \mathbf{D}) \quad \mathbf{b} \perp\!\!\!\perp \epsilon \tag{3.16}$$

- $\mathbf{y_i}$: $n_i \times 1$ vector of responses of the i^{th} individual
- $\mathbf{X_i}\boldsymbol{\beta}$: fixed effects of p covariates, with $\boldsymbol{\beta}$ the $p \times 1$ vector of unknown parameters
- $\mathbf{Z_i}\mathbf{b_i}$: random effects of k factor with $\mathbf{Z_i}$ the $n_i \times k$ design matrix between the factors and the n_i observations, and $\mathbf{b_i}$ the $k \times 1$ random vector of covariance $\mathbf{D}$
- $\boldsymbol{\epsilon_i}$: within-individual error term with independent and identically distributed components

The $\mathbf{b_i}$ are supposed to be independent from each other and to ϵ_i. If all n_i are not equal, the model is called *unbalanced*.

A comparison between classical linear modelling and linear mixed-effects modelling is illustrated on Figure 3.4, with one factor random constant effect:

$$y(t_{ij}) = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{\mathbf{ij}} \quad b_i \sim \mathcal{N}(0, \tau^2) \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{3.17}$$

To make the analogy with Equation 3.16, $\mathbf{Z_i}$ is equal to the $n_i \times 1$ vector of repeated 1 values, b_i correspond to the random deviation from the fixed effect for the i^{th} individual and $\mathbf{D}$ is the variance parameter noted τ^2. The hypothesis here is that the random effect is constant in time, and τ^2 represents the variation between individuals.

Estimation of parameters

Parameters of Equation 3.16 (β, b, θ) with $\theta = (\sigma^2, D)$ could then be estimated by maximising the likelihood $L(\beta, b, \theta)$ of the model 3.16, which can be written as follows:

$$\mathbf{y}_i = \mathbf{X_i}\beta + \nu_i \quad \nu_i = \mathbf{Z}\mathbf{b}_i + \epsilon_i \sim \mathcal{N}(0, \Sigma_\theta) \quad \Sigma_\theta = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{I}\sigma^2 \tag{3.18}$$

Parameters are then be estimated iteratively with either the EM (Expectation - Maximisation) or the Newton–Raphson algorithm. But since the ML variance estimator $\hat{\sigma}^2$ is biased, Patterson and Thompson (1971) and Harville (1977) have proposed to reduce this bias with the Restricted Maximum Likelihood (REML) estimators. The latter is obtained by maximising the likelihood, not of all the data, but rather by the average of the likelihood over all the possible values of β : $\int L(\beta, b, \theta) d\beta$. This method could also have a Bayesian interpretation, corresponding to assuming a locally uniform prior distribution for the fixed effects β (Laird and Ware, 1982). This method provides estimations for the $\hat{\theta}$ and $\hat{\beta}(\hat{\theta}), \hat{b}(\hat{\theta})$. The derivation of ML and REML estimators are detailed and discussed by Laird and Ware (1982).

Testing hypotheses for the fixed effects

To assess the significance of longitudinal evolution between two groups of individuals, a binary variable c is added to the fixed effects, and we test if the related β coefficient(s) are different from zero. Using the formulation of model in Equation 3.16, it is equivalent to testing:

$$H_0 : \boldsymbol{G\beta} = 0 \quad H_1 : \boldsymbol{G\beta} \neq 0 \tag{3.19}$$

where G is a $r \times p$ matrix and r is the number of coefficients tested as different from zero.

The Likelihood Ratio test compares the log likelihood of the both models (H_0 and H_1). However, as explained by Pinheiro and Bates (2000), likelihood ratio tests are not valid when comparing LME models with different fixed effects fitted using REML, since there is a term in the REML criterion that changes with the change in the fixed-effects specification.

In contrast, the F-test, which is similar to likelihood ratio test, relies on a single model for comparison (assuming that the variables not common to both models are zero). In the classical linear model, we have the following results which follows from the Wald statistic (Galecki and Burzykowski, 2013; Scheipl et al., 2008):

$$F = \frac{(G\hat{\beta})^T (G\hat{\mathsf{Var}}(\beta)G^T)^{-1} G\hat{\beta}}{r} \quad \sim \quad \mathcal{F}(r, df)$$

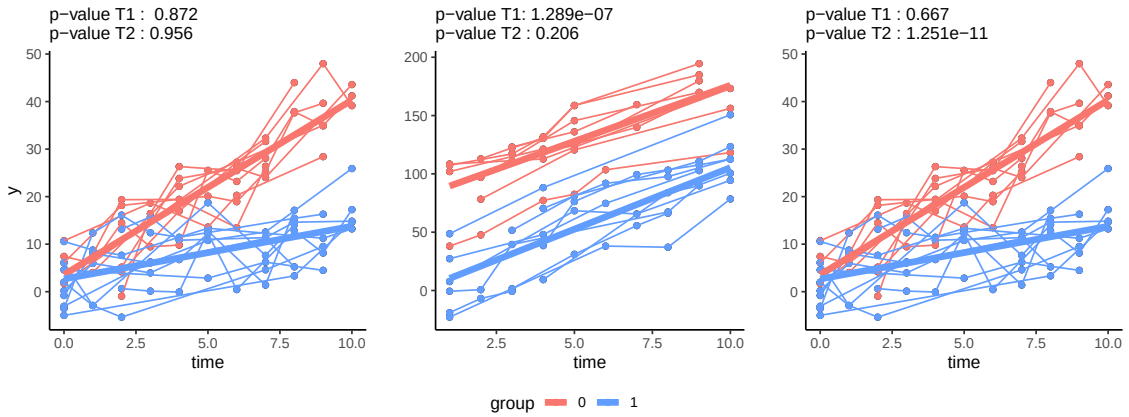


Figure 3.5: Results of the two F-tests (T1) and (T2) on three simulated datasets. Points belonging to the same subject are connected with lines, and the straight lines without dots represent the fixed effects estimated for each group. The *p*-value of the tests are indicated.

where df is the degree of freedom of the model ($n - p$ in the classical linear case) and $\mathcal{F}$ represents the Fisher law. In LME models, we have $\hat{\text{Var}}(\beta) = (X^T \Sigma^{-1}_{\hat{\theta}} X)^{-1}$, where $\hat{\text{Var}}(\beta)$ is the estimated variance of the fixed effects parameter β, conditional to the estimated ML or REML random effect variance-covariance parameters $\hat{\theta}$. Satterthwaite (1946) and Kenward and Roger (1997) have proposed methods for the computation of df and the approximation of the Fisher distribution.

Three examples are shown in Figure 3.5, with the following test applied:

$$y(t_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 c_i + \beta_3 t_{ij} c_i + b_i + \epsilon_{ij}$$

- $H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$ tests only if there is a difference of value at $t = 0$ (intercept) between the two groups (T1) ($G = (0, 0, 1, 0)$)
- $H_0 : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$ tests a difference of slopes between the two groups (T2) ($G = (0, 0, 0, 1)$)
- $H_0 : (\beta_2, \beta_3) = 0 \quad H_1 : (\beta_2, \beta_3) \neq (0, 0)$ performs a multiple test for both slope and intercept with $G = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Non-linear mixed-effects models with penalised spline regression

If there is no *a priori* regarding the shape of the response y with time, the penalised spline regression presented in section 3.1 can be conveniently used for mixed-effects

modelling, thereby connecting non parametric mixed-effects modelling and linear mixed-effects modelling. Let us note $(B_1, ..., B_K)$ a base of smooth functions of dimension K; the following model is equivalent to model 3.16:

$$\boldsymbol{y_i} = \sum_{k=1}^{K} \boldsymbol{\beta_k} B_k(\boldsymbol{t_i}) + \boldsymbol{b_i} + \epsilon_i$$
$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{b_i} + \epsilon_i$$

where $\boldsymbol{X_i}$ is a $n_i \times K$ matrix corresponding to $(B_1(t_i), ..., B_K(t_i))$, and $\boldsymbol{t_i}$ the $n_i \times 1$ vector of time points for individual i.

A penalisation may be applied to the smooth fixed coefficient β:

$$\mathcal{P} = \lambda \beta^T S \beta$$

with S a positive semi-definite matrix. The least square problem then becomes:

$$||y - X\beta - Zb|| + \lambda \beta^T S \beta$$

The S for P-splines is explained in section 3.1.2. By using the eigen-decomposition, $S = UDU^T$, where U is an orthogonal matrix of eigenvectors, and D is a diagonal matrix, the model can be re-parameterised (Wood, 2006; Lee et al., 2013) to get to a model of the form at equation 3.16. Estimation of parameters is then equivalent.

Testing hypotheses on the fixed effect can then be performed using a Fisher test as described above (an example of a non-linear mixed model is shown on Figure 3.6):

$$y(t_{ij}) = \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + (\alpha_0 + \sum_{k=1}^{K} \alpha_k b_k(t_{ij})) \times c_i + b_i + \epsilon_{ij}$$

- $H_0 : (\alpha_1, ..., \alpha_K) = 0 \; vs. \; H_1 : (\alpha_1, ..., \alpha_K) \neq 0$ tests if the two groups have the same trend in time or not, without any *a priori* knowledge on this trend (T3)

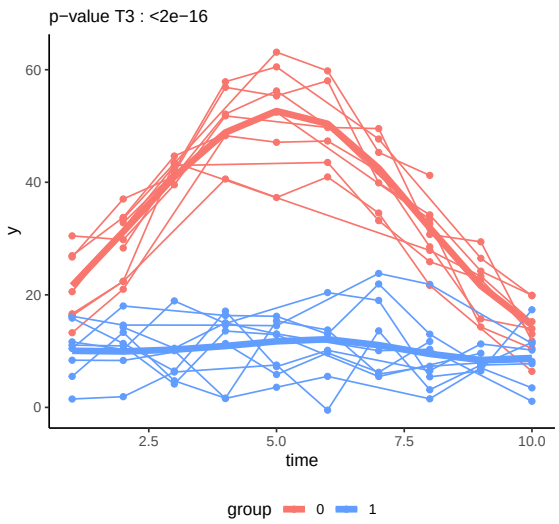


Figure 3.6: Example of non linear mixed-effects modelling, using a sum of 5 P-splines for the fixed effect. The *p*-value from the F-test (T3) is shown, and the fixed effect for each group is represented as a straight line.

Part II

Results

Chapter 4

Design and implementation of innovative methods for the processing of PTR-TOF-MS data: ptairMS

Existing PTR-TOF-MS pre-processing tools are particularly suited for the analysis of very large files resulting from continuous environmental monitoring, including robust pic picking methods in the m/z dimension, but poor temporal signal treatment as described in section 2.2.4. Furthermore, there are specific needs for breath research in patient cohorts which have to be covered, for instance, the simultaneous analysis of multiple samples requires that peak lists from different samples may be aligned; in addition, the parallel processing of several files would be a time-sparing capability; furthermore, a correct distinction of the signals coming from the background and the expiratory phases is needed; finally, implementing a background correction of the ambient air composition as a function of time would be an asset for accurate peak detection and quantification.

We therefore developed a suite of algorithms for the processing of PTR-TOF-MS data from exhaled breath, based on an innovative 2D model based on P-splines regression that enables a precise estimation of the peak evolution over the acquisition time, and integrating several tools for cohort management in an R package, called ptairMS. It takes as input the name of the directory containing the raw files in HDF5 format and ultimately generates the samples by variables table of peak intensities. The main steps of the workflow are summarised below (Algorithm 1) and detailed in the following section:

1. Processing of each file

1.1 Internal calibration of the m/z axis

1.2 Determination of expiration limits

1.3 Detecting peaks on the TIS

1.4 Estimating the temporal evolution for each peak

1.5 Quantifying

1.6 Statistical testing of intensity differences between ambient air and expiration phases

2. Alignment between samples followed by quality control

2.1 Peak matching between samples

2.2 Filtering features based on reproducibility within the whole cohort or sample classes and on the p-value from the test in 1.6

3. Imputation of missing values

4. Putative annotation (including isotopes)

5. Peak table update when new files are included in the input directory

4.1 Pre-processing for each file

4.1.1 Calibration

Calibration converts the Time-of-Flight (TOF) values recorded by the mass spectrometer into m/z values (see section 2.2.1). We used the formula proposed by Brown and Gilfrich (1991):

$$m/z = \frac{(tof - b)^2}{a}$$

To estimate the parameters (a, b), the Levenberg-Marquardt algorithm is used, with couples $(tof, m/z)$ of reference peaks without overlap. For exhaled breath, we suggest to use the following peaks: the primary ion isotope (m/z 21.022), nitrogen (m/z 29.013), the acetone isotope (m/z 60.053), and the two external calibration ions from the instrument: (iodobenzene m/z 203.943, and diiodobenzene m/z 330.850).

As a drift over time is observed due to low changes of temperature, calibration is performed periodically (e.g. every minute) to update the (a, b) values. The shift is subsequently estimated for each m/z as a function of time by linear interpolation, and corrected locally (for each nominal mass ± 0.4) before peak detection.

Algorithm 1: ptairMS workflow

Data: Directory of HDF5 file (optional: sample metadata csv file)
forall *file not processed in the directory* **do**
  (1.1) Multiple internal calibration along time
  (1.2) Expiration and inhalation phases detection
  Compute peak shape, resolution on calibration peaks, and amount of primary ions
end
Manual check
forall *file checked in the directory* **do**
  forall *nominal mass* m **do**
    Reduce raw data around $m \pm 0.4$ and correct the calibration shift
    (1.3) Peak detection on the TIS
    forall *peak detected* **do**
      (1.4) Estimating the temporal evolution of each peak with 2D model
      (1.5) Statistical testing between exhaled breath and ambient air intensities
      (1.6) Quantifying in exhaled breath phases
    end
  end
  Result: Individual peak list
end
(2.1) Peak matching between samples
(2.2) Filtering features
forall *missing features* m/z **do**
  (3) Imputation by returning back to raw data
end
(4) Putative annotation (including isotopes)
Result: Peak table, sample and features metadata

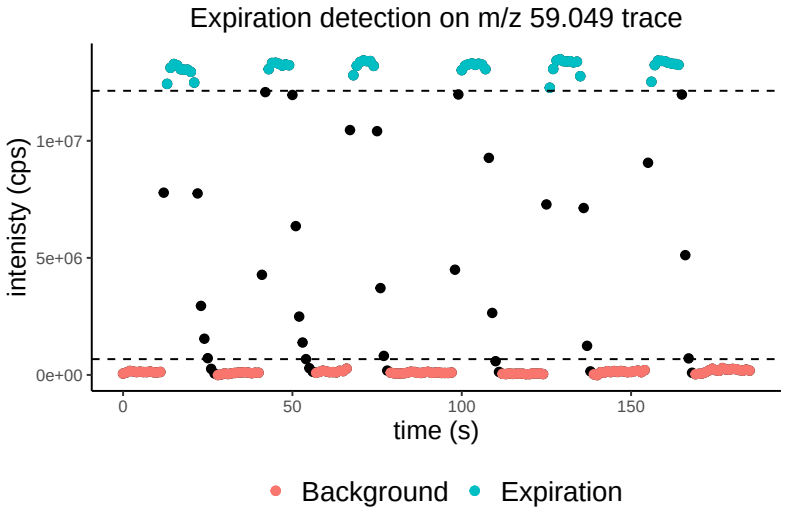


Figure 4.1: Expiration phases and ambient air detection on ion trace intensities in count per second (cps).

4.1.2 Expiration detection

Determination of expiration limits and background (ambient air) is a very important step for the analysis, as boundaries will be used for quantification and for the statistical test for feature selection in section 4.1.5. A raw data ion trace is used to automatically detect expirations, using the same method as described by Schwoebel et al. (2011); Trefz et al. (2013): after a polynomial baseline removal, the signal above `fracMaxTIC` $\times max(trace)$ is considered as expiration (Figure 4.1). Conversely, the signal below `fracMaxTICBg` $\times max(trace)$ is considered as background (inhalation phases). In addition, differences between successive points in expiration (respectively, inhalation) phases must be lower than `derivThresholdExp` (respectively, `derivThresholdBg`). All parameters from the ptairMS software are described in Table 4.2.

Finally, to assist the user in this important step, we have designed a specific panel from the graphical interface of our software tool to the visualisation (and possible manual modification) of the expiration limits (Figure 4.8).

4.1.3 Peak detection and quantification on the Total Ion Spectrum (TIS)

The peak picking (section 1.2.1) algorithm in the m/z dimension is mainly based on Müller et al. 2013. Since VOCs have low weights ($<$ 500 Da), we detect peaks on the TIS around nominal masses ($\pm 0.4\ Da$). The successive steps are (see Figure 4.2):

1. Baseline removal using the SNIP algorithm (Ryan et al., 1988)
2. Estimation of the noise threshold and auto-correlation within the "off-peak" interval $[m - 0.6, m - 0.4] \cup [m + 0.4, m + 0.6]$ (Müller et al., 2011)
3. Savitzky Golay (SG) signal filtering by using optimal windows (Vivo Truyols and Schoenmakers, 2006; Savitzky and Golay, 1964)
4. Detection of local maxima by using the first and second derivatives of SG smoothing, followed by quality control on peak separation (in ppm) and intensity threshold set to the maximum between a) the noise threshold, b) min intensity parameter, and c) 1% of the signal maximum
5. Peak deconvolution, by using a peak function depending on the vector parameters μ (peak centre), σ^l, σ^r (peak widths at half maximum, left and right from the peak centre), and h (peak height):

$$\underset{\boldsymbol{\mu},\boldsymbol{\sigma^l},\boldsymbol{\sigma^r},\boldsymbol{h}}{min} ||\tilde{\mathbf{y}} - \sum_{i=1}^{P} h_i \times peak_{\mu_i,\sigma_i^l,\sigma_i^r}(m)||^2 \tag{4.1}$$

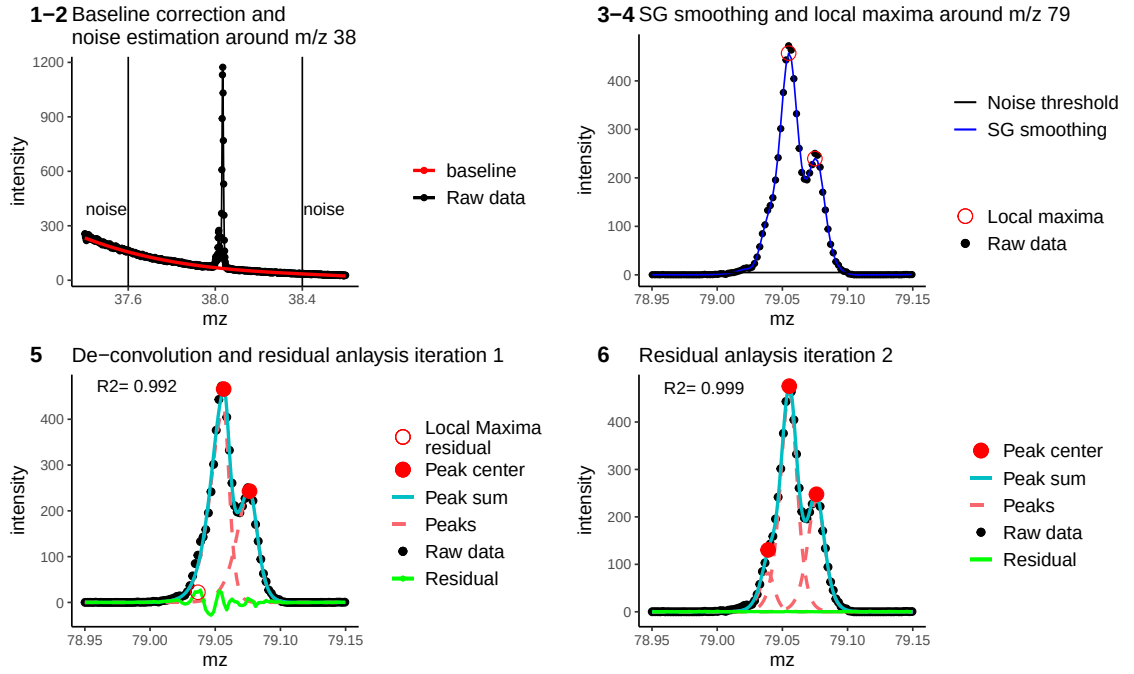


Figure 4.2: Peak detection on the Total Ion Spectrum (TIS) around nominal masses. For illustrative purposes, we focus on m/z 38 for baseline correction (1-2) and m/z 79 for deconvolution (3-6). We used the $sech_2$ function for deconvolution at steps 5 and 6.

with $\tilde{\mathbf{y}}$ the baseline corrected and smoothed signal, and P the number of detected local maxima. The optimisation is done with the Levenberg-Marquardt algorithm, with the following initialisation and constraints: $\boldsymbol{\sigma^l} = \boldsymbol{\sigma^r} = \boldsymbol{\sigma}/2$ with $\sigma = m/res_{mean} \in [m/res_{max}; m/res_{min}]$, $\boldsymbol{\mu}$ the local maxima detected at the step 4 above $\pm\sigma \times 4\ Da$, and **h** the values of the spectrum at mass(es) $\boldsymbol{\mu}$ (always positive).

6. Iterative peak detection on the residuals (Müller et al., 2011), using the same method as described above, with a decreased noise threshold of 20%. Iterations stop as soon as one of the following criteria is met: $R2 > \texttt{R2}_{min}$ (default: 0.995), noise autocorrelation $<$ `autocorMax` , the maximum number of iterations is reached (default: 4), the maximum number of detected peaks is reached (default: 7).

Peak shape

To find the most suitable peak shape, four asymmetric functions are evaluated on the calibration peaks, and the one providing the best R2 value is selected: average calibration peak shape used by Müller et al. (2010) and Holzinger (2015) (see section 2.2.3), the $sech_2$ function (Equation 4.2; Lange et al. 2007), Bi-Gaussian (Equation 4.3), and Lorentzian functions (Equation 4.4; Lange et al. 2007). An example of the peak shape selection is shown in Figure 4.3.

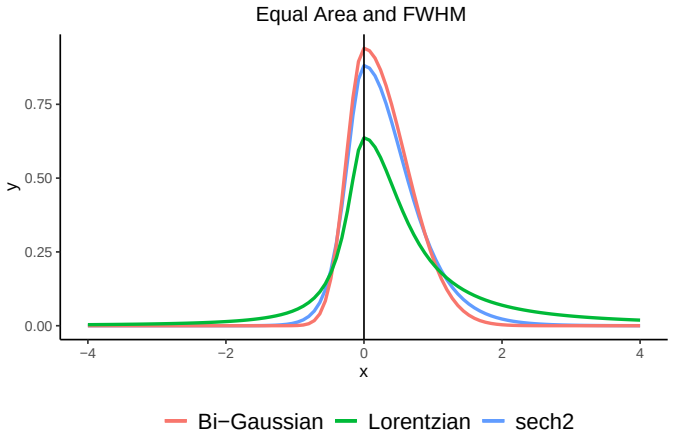


Figure 4.3: Asymmetric peak shape functions included in ptairMS with $\sigma^l = 0.3$ and $\sigma^r = 0.7$, in addition to the average peak function obtained from raw data.

$$sech_2(m) = h \times \frac{1}{cosh(\lambda(x-\mu))^2} \quad \text{with} \quad \lambda = \lambda_1 \quad \text{if} \quad x < \mu \quad \lambda = \lambda_2 \quad \text{if} \quad x \geq \mu \tag{4.2}$$

$$g(m) = h \times exp^{-\frac{(x-\mu)^2}{2\sigma}} \quad \text{with} \quad \sigma = \sigma_1 \quad \text{if} \quad x < \mu \quad \sigma = \sigma_2 \quad \text{if} \quad x \geq \mu \tag{4.3}$$

$$L(m) = h \times \frac{1}{1 + \lambda^2(x-\mu)^2} \quad \text{with} \quad \lambda = \lambda_1 \quad \text{if} \quad x < \mu \quad \lambda = \lambda_2 \quad \text{if} \quad x \geq \mu \tag{4.4}$$

4.1.4 Estimating the temporal evolution for each peak

After peak detection on the TIS, the next step aims at estimating the evolution of the peak intensity over the acquisition time. Current methods, which consist in summing raw data around detected peak centres (see section 2.2.4) may be biased when there are two or more overlapping peaks with different temporal evolutions. We therefore proposed a 2D regression approach, using a tensor product between P-splines (section 3.1) and the previously estimated m/z peak functions.

The P-spline approach is very powerful to model any profile without *a priori* knowledge of the data and to provide interpretive coefficients (Eilers and Marx, 2021; Wood, 2006). It has been used in many applications and theoretical works (Eilers et al., 2015) such as data smoothing (Currie and Durban, 2002), Bayesian statistics (Gressani and Lambert, 2021), and machine learning with generalised additive models (GAM; Brezger and Lang 2006; Wood 2006). To model interactions in multiple dimensions, the tensor product provides

a straightforward generalisation of the P-spline basis (Sidiropoulos et al., 2017). Here, we thus used tensor product modelling to achieve a fast deconvolution of peaks in both m/z and time dimensions simultaneously, as described below (see Figure 4.4):

1. Raw data are processed sequentially within bands around the detected peaks (the 1% quantile of the estimated mixture of peak functions is used to define the m/z bounds) and covering the full acquisition time

2. The baseline in the m/z dimension is estimated at each time point by linear regression between the two m/z boundaries and is subsequently removed

3. The calibration shift estimated in section 4.1.1 is corrected by linear interpolation

4. The modelling of peak evolution with time is performed by using a two dimensional model:

   Let us denote the functions representing the acquisition time $g(t)$ and the m/z profile $h(m)$, respectively:

$$g(t) = \sum_{j=1}^{K} \alpha_j b_j(t) \qquad h(m) = \sum_{i=1}^{n_{peak}} h_i peak_i(m) \tag{4.5}$$

   with $peak_i(m)$ being the function of peak i estimated in the previous step (equation 4.1), n_{peak} the number of detected peaks, and $(b_1, \dots b_K)$ the cubic B-spline functions for the set of K knots. The 2D model is obtained by writing each peak coefficient h_i in the B-spline basis (tensor product, section 3.1.4):

$$f_\beta(t, m) = \sum_{i=1}^{n_{peak}} \sum_{j=1}^{K} \beta_{ij} b_j(t) \times peak_i(m) \tag{4.6}$$

   The β_{ij} coefficients are estimated according to the P-spline theory by minimising the following penalised regression, where the penalty λ is applied only to the time dimension:

$$\min_{\beta} \sum_{t=1}^{T} \sum_{m=1}^{M} (Y_{mt} - f_\beta(m, t))^2 + \lambda \sum_{i=1}^{n_{peak}} \sum_{j=3}^{K} (\Delta^2 \beta_{ij})^2 \tag{4.7}$$

   where $\Delta^2 \beta_{ij} = \beta_{(i,j)} - 2\beta_{(i,j-1)} + \beta_{(i,j-2)}$ is the second order difference between successive coefficients in the time dimension, i (respectively, j) represents the knots location on the mass (respectively, time) axis, m (respectively, t) represents the index on the mass (respectively, time) axis, and $\boldsymbol{Y}$ is the raw data matrix of dimensions $M \times T$ after baseline removal and correction of the calibration shift.

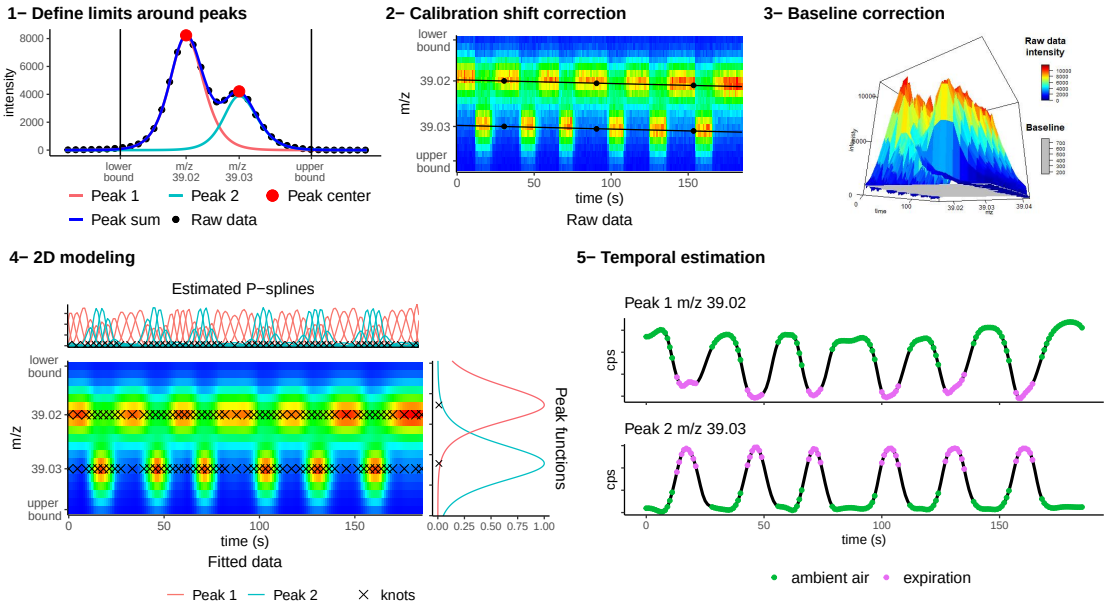


Figure 4.4: Modelling of the VOC temporal evolution, starting from the peaks detected in the total ion spectrum (1), and resulting in the estimation of a temporal series for each of them (5).

5. Quantification (in counts per spectrum) is then performed at each time point t by summing the previous estimated 2D model along the m/z dimension

$$c_t^i = \sum_{m=1}^{M} \sum_{j=1}^{K} \beta_{ij} \times s_j(t) \times peak_i(m)$$

This results in a temporal series of intensities for each peak $(c_1^i, ..., c_T^i)$.

The choice of the knot locations and the penalty coefficient λ are very important, since too many knots may lead to over fitting and too few knots may result in under fitting.

Knots location

Classically, knots are uniformly distributed over the data range in order to facilitate the interpretation of the penalty applied to the successive knot differences (Eilers and Marx, 1996). In our case, however, i) exhaled breath phases are the main focus of our quantification, and ii) inhaled air phases are generally constant. We therefore propose to target the knot locations mainly around the expiration phases: knots are spread uniformly every 3 seconds, except for inhalation phases longer than 3 seconds, where only the first, middle and last points of the phase are used as knots (Figure 4.5). This allows to reduce the dimension of the model, and thus the computational time, while maintaining a good fit (Table 4.1). Alternatively, a uniform distribution of the knots along the time axis is also available in the ptairMS software, in case the user has no a priori knowledge about the temporal profile of the compound.

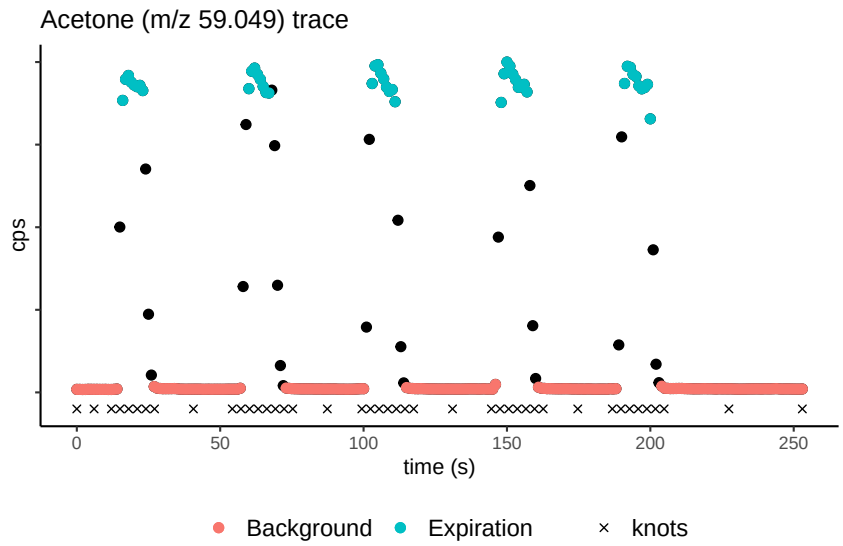


Figure 4.5: Proposed knots location around expiration

Knots location	MPAE global (%)	MAPE expiration only (%)	Computational Time per peak cluster (s)	Dimension (number of knots)
Focused on expirations	4.22	4.3	0.16	52
Uniform, every 3s	3.62	4.2	0.49	103
Uniform, every 5s	5.42	6.2	0.22	62
Uniform, every 10s	8.24	9.5	0.09	32

Table 4.1: Comparison of knots location: Our "expiration targeted" strategy was compared to uniform knot locations on 4,930 simulated peaks (please see the section 5.2.1 for the description of our simulation methodology). The following metrics are computed: Mean Absolute Percentage Error on the temporal profile estimation (MAPE; computed either on the full acquisition or on the expiration phases only), computational time, and total number of knots, correlated with the previous one. The results displayed above show that the proposed knot locations focused on expirations is a good trade-off between fit quality and computational time.

Penalty coefficient

The optimal penalty coefficient λ is selected with a grid search using the generalised cross-validation criterion (GCV; section 3.1.3). Since the penalty is set only to the time dimension, the GCV criterion is not computed from the 2D model, but from the 1D spline regression on the total time trace of the raw data band (containing the mixture of peaks), to reduce computational time.

4.1.5 Quantification

As presented in the introduction, PTR ionisation enables to compute "absolute" quantities. Here, we thus describe how the times series for each peak i, $(c_1^i, ..., c_T^i)$, are normalised and converted to absolute quantities $(Q_1^i, ..., Q_T^i)$. First, since the intensities provided by the instrument at each time point are in fact the sum of a fixed number of internal acquisitions, the c_t^i are normalised (as counts of ions per second; cps) by dividing by the integrated internal time period and by multiplying by the single ion pulse voltage (Müller et al., 2014). To obtain the concentration, the latter values are then normalised by the reagent ion (H_3O^+) intensities, the reaction rate coefficient between the VOC and H_3O^+, and the residence time of the primary ions in the drift tube (normalized cps, ncps) (Cappellin et al., 2012a). The final normalisation by the density of the air in the reaction chamber (ideal gas law) gives the absolute concentration of the VOC, expressed in part per billion (ppb) (section 2.2.5).

The final absolute concentration of each VOC i in exhaled breath is obtained by averaging absolute quantities in the expiration phases:

$$Q^i = \frac{\sum_{t \in exp} Q_t^i}{|exp|}$$

where $|exp|$ corresponds to the number of expiration points.

Ambient inhaled air correction

As discussed in section 1.3.2, inhaled air may impact exhaled breath concentrations. To correct the ambient inhaled air level in exhaled breath, we thus propose to subtract the ambient air baseline from the temporal profile $(Q_1^i, ..., Q_T^i)$ of each detected VOC before averaging in exhaled breath phases: to do so, a polynomial fit of default degree 3 computed on the ambient air time points is used. Note that the subtraction step may be omitted in particular cases, as detailed in the discussion.

4.1.6 Statistical testing of intensity differences between expiration and ambient air phases

Two unilateral statistical tests (*t*-tests) are used to compare intensities within and between expirations on the estimated temporal evolution (Figure 4.4, panel 5; see the statistical test section 3.2.1). Compounds with intensities that are significantly higher (respectively, lower) within expiration phases are considered to be from exhaled breath (respectively, from ambient air). If none of the tests is significant, the compound is labeled as "constant" (e.g. in the case of internal ions generated by the instrument). The significance is evaluated with the *p*-value.

$$H_0 : Q^i_{t\in exp} = Q^i_{t\in amb} \quad \begin{array}{ll} \text{if } H_1 : Q^i_{t\in exp} > Q^i_{t\in amb} \text{ is significant :} & \textbf{exhaled breath} \\ \text{if } H_1 : Q^i_{t\in exp} < Q^i_{t\in amb} \text{ is significant :} & \textbf{ambient air} \\ \text{if none is significant:} & \textbf{constant} \end{array} \tag{4.8}$$

4.2 Alignment between samples followed by quality control

4.2.1 Peak matching

Once the peak lists have been extracted from each file, alignment of the features between the samples (section 4.6) is performed by using a kernel Gaussian density estimation (Smith et al., 2006). For each nominal mass, we estimate a kernel Gaussian density from all the detected m/z peak centres $(m/z_1, ..., m/z_n)$. Then, the peaks from the estimated density and their boundaries are detected as follows (Figure 4.6): starting from the first point, a new peak starts when the density increases, reaches its centre when it starts to decrease, and ends when it increases again. All individual peaks contained within the same boundaries are considered as belonging to the same feature. The new $\overline{m/z}$ of this feature corresponds to the median of the $(m/z)_i$ values in this group (Figure 4.6).

The standard deviation of the smoothing kernel is set with the parameter (`ppm`), and corresponds to the maximum deviation authorised between acquisitions on the m/z axis. The higher the value of this parameter, the lower the number of groups (Figure 4.6).

4.2.2 Quality control

Two quality control steps may be further applied to select the features:

1. with a high reproducibility between samples (alternatively, between classes of samples), by keeping features present in at least `fracGroup` percent of the samples (or of one class)

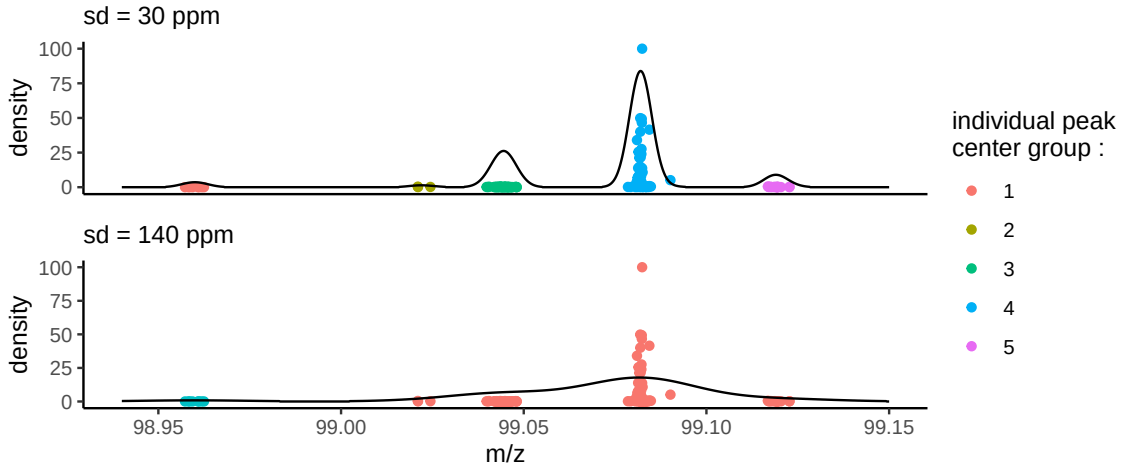


Figure 4.6: Alignment of 253 detected peaks around the nominal mass 99 for 166 individual. The individual peaks are plotted as dots at their detected m/z value in each sample, and in relative intensities on the y axis (i.e. the maximum intensity is set to 100). The kernel density is shown as a solid line, for standard deviations (sd) of 30 ppm (top) and 140 ppm (bottom). The detected groups are coloured: five features are detected with sd = 30 ppm, but only two with sd = 140 ppm.

2. labelled as "exhaled breath" in the majority of the samples, by thresholding the *p*-values of the statistical tests described above (4.1.6) in at least `fracExp` percent of the samples

4.3 Imputation of missing values

Following the alignment step, missing values occur for peaks that have not been detected in the total spectrum at step 4.1.3 for several reasons including: intensity under the limit of detection (LOD; missing not at random: MNAR) or peaks that could not be deconvolved (missing at random: MAR; Wei et al. 2018). Since the raw data are available, the ptairMS software was designed to re-rerun the peak detection algorithm 4.1.3 on the raw data, and take into account the already detected neighbouring peaks, with a restricted m/z width for the peak centre ($\pm$ 30 ppm), and without any minimum intensity threshold. This may allow the recovery of peaks that have been missed during the peak detection (e.g. too convoluted with the neighbouring peak, or slightly below the limit of detection). If the peak is indeed missing, this is equivalent to integrating the noise of the instrument.

4.4 Putative annotation (including isotopes)

Putative annotations are computed by matching the measured ion masses to an internal table extracted from the Human Breathomics Database (Kuo et al., 2020). In addition, isotopes (i.e. molecules that have the same number of protons and electrons but different

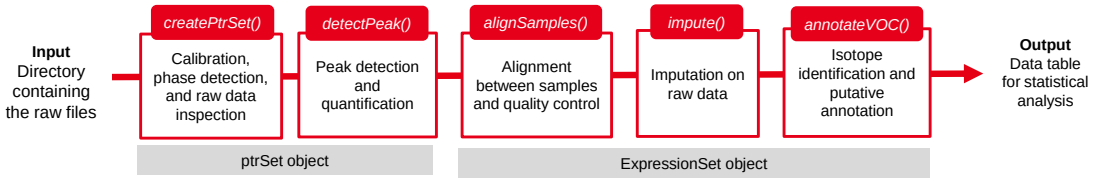


Figure 4.7: The ptairMS workflow

number of neutrons, and therefore different physico-chemical properties) are suggested on the basis of three criteria (Kuhl et al., 2009):

- m/z difference value for isotope ^{2}H, ^{13}C ^{15}N, ^{17}O, and ^{18}O, with a interval of ± 50 ppm
- Two Pearson's correlation test, one between temporal profiles within the sample and an other with intensities between the samples, using a *p*-value threshold at 1%.

4.5 ptairMS software

All algorithms were written in R (R Core Team, 2021) and implemented as the ptairMS package freely available on the Bioconductor platform (Gentleman et al., 2004). The companion ptairData experiment package, also available on Bioconductor, contains the raw files from two data sets from exhaled breath and bacteria culture head space, respectively, as well as the simulated raw data file described in the following Results section.

The workflow consists of five steps (Figure 4.7):

1. *createPtrSet*: A ptrSet object is generated by taking as input the name of the directory containing the raw files (in HDF5 format), possibly grouped into subfolders according to classes of samples. This object is then completed at each step of the processing. In addition, the ptrSet may be updated by adding new raw files to the directory, or by providing new sample metadata
2. *detectPeak*: peak detection and quantification are performed within each file and the ptrSet object is updated with the sample metadata, the peak list for each sample, and several quality metrics
3. *alignSamples*: The peak lists are aligned between samples, and an `ExpressionSet` object is returned, containing the table of peak intensities, the sample metadata, and the feature metadata (which can be accessed with the `exprs`, `pData` and `fData` methods from the Biobase package, respectively)
4. *impute*: Missing values in the table of intensities may be replaced by the integrated signal in the expected raw data region

Parameter	Description	Default value(s)
`mzCalibRef`	Reference mass values for calibration of the mass axis	m/z 21.022, 29.01, 41.03, 60.05, 203.94, 330.84
`calibrationPeriod`	Time duration of each calibration	60 seconds
`mzBreathTracer`	Nominal mass of the ion trace used to compute the expiration time limits	Acetone
`fracMaxTIC`	Percentage of the maximum of the ion trace used to determine the expiration time limits	80%
`ppm`	Minimum peak proximity	130 ppm (part per million)
`minIntensity`	Minimum peak intensity	auto tuned
`resolutionRange`	Minimum, mean, and maximum resolution (m/Δ_m)	auto tuned
`fctFit`	Parametric peak shape function to be used	auto tuned
`knotsPeriod`	Time period between two knots for the 2D modelling	3 s
`smoothPenalty`	Value of the smoothing coefficient λ	auto tuned
`ppmGroup`	Maximal width for an m/z group	70 ppm
`fracGroup` / `fracExp`	Detection robustness between the samples/expiration phases	0.8/0.3
`pValGreaterThres` / `pValLessThres`	*p*-value threshold for the unilateral testing that the quantification (in cps) of expiration points is higher/lower than the intensities in the background	0.0001

Table 4.2: Parameters from the ptairMS software

5. *annotateVOC*: Suggestions of feature annotations may be provided, based on the Human Breathomics Database (https://hbdb.cmdm.tw; Kuo et al. 2020)

All parameters are described in Table 4.2. The auto tuned parameters `minIntensity`, `resolutionRange`, and `fctFit` are determined from the calibration peaks (selected with the `mzCalibRef` parameter). The `smoothPenalty` is selected by cross validation as explained in Section 4.1.4.

Eventually, the output contains the table of peak intensities as well as the sample and variable metadata, which can be exported as three tabular files, or as a single ExpressionSet object for subsequent statistical analysis. The detailed tutorial of the ptairMS package is available on the Bioconductor repository (`https://bioconductor.org/packages/release/bioc/vignettes/ptairMS/inst/doc/ptairMS.html`).

The whole workflow can be run interactively through a graphical user interface, which provides visualisations (expiration phases, peaks in the raw data, peak table, individual VOCs), quality controls (calibration, resolution, peak shape, and evolution of reagent ions with time), and exploratory data analysis (Figure 4.8).

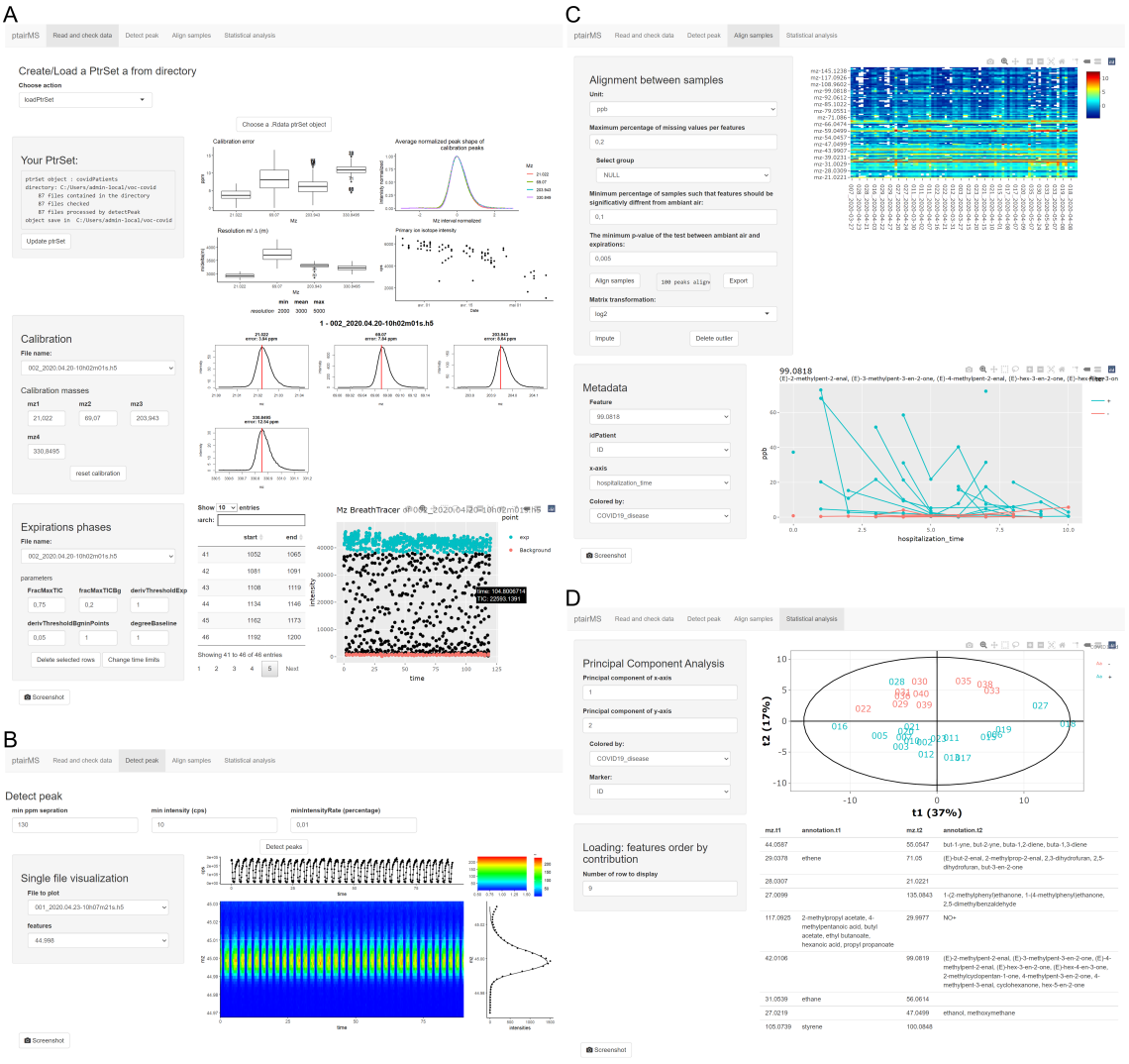


Figure 4.8: The ptairMS graphical interface, as illustrated with the COVID-19 data described in Chapter 6. (A) The *Read and check data* tab enables to open the data (either from a new study or to update an existing one), to perform the calibration and the detection of expirations, and provides optimal parameter values for the peak shape and the resolution. (B) The *Detect peak* tab provides single file visualisations of the raw data and of the detected peaks and temporal profiles. (C) The *Align samples* tab displays the final peak table as well as the individual features coloured according to the sample metadata. (D) The *Statistical Analysis* tab displays the score plot from the Principal Component Analysis of the peak table (only the first time point of each patient is shown here, as in Grassin-Delyle et al. 2021, and the list of features with their putative annotations, in decreasing order of loading values.

Chapter 5

Application to simulated and real datasets

5.1 Quantification and detection in a standardised gas mixture

5.1.1 Standardised gas mixture data set

To validate the algorithmic performance of ptairMS, a reference gas containing a mixture of VOCs in known amounts was first used: the TO-14 standard gas mixture (Restek) contains 14 compounds (Figure 5.1) which results in 26 spectral peaks (8 distinct masses and 18 isotopes).

Ten dilutions of the gas mixture were measured, with or without applying an activated charcoal filter (Supelpure HC hydrocarbon trap, Sigma-Aldrich, Saint-Quentin-Fallavier, France) on the ambient air input (three replicates each), resulting in 60 raw files. During each acquisition, the aspiration of the reference gas was switched on and off three times to mimic "expiration" profiles. Sample analysis was performed with a PTR-Qi-TOF (Ionicon, Innsbrück, Austria) at the Foch Hospital.

5.1.2 Results

The 60 raw files were pre-processed by ptairMS in less than 15 min (on a quad-core laptop). Calibration was performed using m/z 21.022, 203.943, and 330.849. The default values were used for the peak detection (see the section 4.1.3). A total of 314 (respectively, 180) compounds were detected in the absence (respectively, presence) of the charcoal filter.

For the alignment steps, filters were set to keep features with at least 90% of one dilution factor (`fracGroup` = 0.9), and in the simulated "expiration" phases of at least 90% of

CERTIFIED CONCENTRATIONS			
Component	**Requested Concentration**	**Reported Mole %**	**Accuracy**
1,2 DICHLOROBENZENE	1.000 PPM	1.010 PPM	+/- 5%
1,2,4 TRICHLOROBENZENE	1.000 PPM	0.9700 PPM	+/- 5%
1,2,4 TRIMETHYLBENZENE	1.000 PPM	1.030 PPM	+/- 5%
1,3 DICHLORO BENZENE	1.000 PPM	1.030 PPM	+/- 5%
1,3,5 TRIMETHYL BENZENE	1.000 PPM	1.020 PPM	+/- 5%
1,4 DICHLOROBENZENE	1.000 PPM	1.010 PPM	+/- 5%
BENZENE	1.000 PPM	1.000 PPM	+/- 5%
CHLOROBENZENE	1.000 PPM	1.010 PPM	+/- 5%
ETHYL BENZENE	1.000 PPM	0.9700 PPM	+/- 5%
M XYLENE	1.000 PPM	1.000 PPM	+/- 5%
O XYLENE	1.000 PPM	1.040 PPM	+/- 5%
P XYLENE	1.000 PPM	1.000 PPM	+/- 5%
STYRENE	1.000 PPM	1.050 PPM	+/- 5%
TOLUENE	1.000 PPM	0.9800 PPM	+/- 5%
NITROGEN	100.0 %	99.998588 %	+/- 2%

Figure 5.1: List of the compounds and their absolute concentrations in the TO-14 gas mixture, as provided by the manufacturer (Restek)

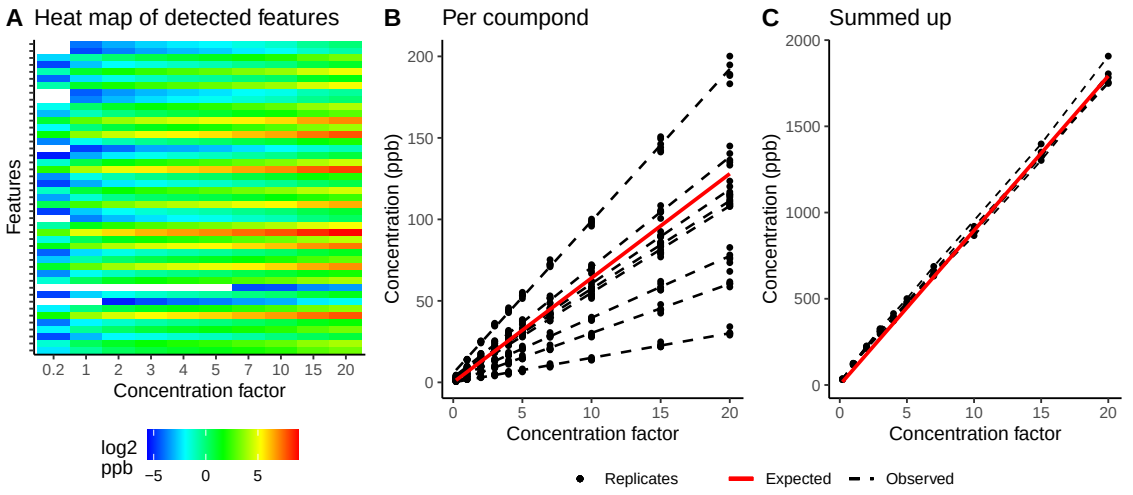


Figure 5.2: ptairMS analysis of a reference VOC mixture. (A) Heatmap of the log2 concentrations in ppb of the 45 selected VOCs before the imputation step. (B) Amounts (in ppb) of the 45 compounds (dots), as well as the regression line for each of them (dashed black lines). The expected quantity (according to the manufacturer) is shown as a red solid line. (C) Observed replicates (dashed, black) and expected (solid, red) quantities for the sum of the 45 compounds for each replicate as a function of the concentration factor.

Expected ppb per compound	Mean absolute error (%)	CV (%)
$[1.3; 13]$	47.9	4.3
$[19; 32]$	8.1	3.4
$[44; 128]$	2.5	2.8

Table 5.1: Mean absolute percentage error (MAPE) and coefficient of variation (CV) between replicates of the ptairMS processed data from the reference gas mixture acquisitions.

all samples (`fracExp` = 0.9). This resulted in 45 compounds (Figure 5.2A).

Importantly, all the expected compounds were detected, as well as their isotopes, with an m/z error less than 20 ppm, and an average coefficient of linearity R2 with the concentration factor of 0.999. The 19 additional detected features most likely correspond to fragments from these VOCs, since some are below the expected quantity (Figure 5.2B). To evaluate the quantification, we compared the sum of the measured 45 compound quantities and the total amount of compounds predicted by the manufacturer values from Figure 5.1: the error was less than 8.1% for the quantities above 19 ppb (Table 5.1 and Figure 5.2C). Furthermore, the coefficient of variation (CV) between replicates was less than 5% (Table 5.1), even in the absence of charcoal filter.

5.2 Temporal profile classification and comparison to existing software on simulated data

We then compared ptairMS to the two existing PTR-MS software tools, namely PTRwid (Holzinger 2015; publicly available) and IDA (based on Müller et al. 2013; commercial), as introduced in the section 2.3. To do so, we first simulated exhaled breath data files (based on real expiration profiles), and we then computed the list of peaks and their temporal estimation with each software tool. In the following sections, the simulation method is described (available in the package ptairData), and the results of the comparison are presented.

5.2.1 Simulated data

The simulation algorithm to generate a raw PTR-TOF-MS data file from exhaled breath is described in Algorithm 2, and the successive steps are detailed below. The general idea is to simulate raw data bands around nominal $[m \pm 0.5Da]$ masses, and then paste them into a created HDF5 file. We first simulate peaks in the mass dimension, and then the evolution of these peaks at each time:

(1) Temporal profiles were exacted from an in-house database of 200 acquisitions of exhaled breath from patients. After pre-processing with ptairMS, we selected 1) ex-

Algorithm 2: Simulation of PTR-TOF MS data from exhaled breath

Data: Library of exhaled breath temporal profiles extracted from several raw files, and then smoothed and normalised (1)
Data: List of chemical formulae used by PTRwid for calibration matches (2)
Randomly draw a file j from the library, and extract the mass and time axes.
forall *nominal mass* m **do**
  Generate random background noise (3)
  if *there is a compound in the chemical formula list of nominal mass* m **then**
    Random draw of the parameters from the (mixture of) m/z peaks (4)
    Given a temporal profile $(q_1, ..., q_T)$ from the file j:
    Compute the number of ions n (area of the peak) for a given peak height h (5)
    forall t *in time axis* $[1, .., T]$ **do**
      Draw $n \times ratio \times q_t$ random variable of $sech_2$ law
      Write the histogram as a spectrum in the data matrix
    end
    Result: Simulated peak(s) + background noise
  else
    Result: Background noise
  end
end

piration profiles (with intensities significantly greater in expirations than ambient air; $p-value < 2 \times 10^{-20}$), and 2) ambient air profiles (with the *p*-value of the opposite test $< 10^{-10}$) . The profiles were then normalised (mean set to 1) and smoothed with the Savitzky-Golay filter. This resulted in approximately 12,800 expiration profiles and 11,000 ambient air profiles.

(2) m/z values were generated from the library of compound formulae used by PTRwid for calibration matching $C_a C_b^{13} H_c O_d N_e$, with $a \in [1, 40]$, b $\in \{0, 1\}$, c $\in [max(1, a - 9), a]$, $d \in [0, 5]$ and $e \in [0, 2]$. We included only masses between 15 and 400 Da (corresponding to a total of 9,114 formulae).

(3) The Background noise was generated using a Poisson stochastic process (Gundlach-Graham et al., 2018), with λ=0.1 and a Gaussian distribution $\mathcal{N}(11, 3)$ to model the single ion Pulse-Height.

(4) Random drawing of the m/z peak parameters for nominal masses distinct from 19, 21, 29, 30, 32, 59, 204, and 331:

- number of overlapping peaks set to 1, 2, and 3, with probabilities $(0.4,\ 0.5,\ 0.1)$
- class of temporal profile set to "expiration", "ambient air", or "constant", with probabilities $(0.4,\ 0.4,\ 0.2)$
- intensity of the highest peak on the average total spectrum: uniform distribu-

tion $\mathcal{U}(50, 500)$ for multiple peaks, and Gaussian distribution $\mathcal{N}(1500, 100)$ for single peaks (most of the peaks used to compute the peak shape are single peaks; see Figure 5.3)

- intensity ratio of the neighbouring peaks: $\mathcal{U}(0.5, 1)$
- peak width set to m/*resolution*, where *resolution* is drawn from $\mathcal{N}(5000, 500)$
- $sech_2$ skewness: $\mathcal{U}(-0.3, 0.3)$
- asymmetry coefficient: $\mathcal{U}(0.4, 0.6)$
- peak proximity (in ppm): $\mathcal{U}(190, 230)$

(5) The relation between the peak height h and the peak area of the $sech_2$ function is detailed in appendix A.2

To get closer to exhaled breath raw data, and enables PTRwid to finds peaks for its calibration, the following peaks were simulated without overlap: primary ion and water dimer m/z 19.017 and m/z 37.0284 with the highest intensity (200,000), their isotopes m/z 21.022, m/z 38.033 with the corresponding ratio of intensities, m/z 59.049 with intensity set to 20,000, and the calibration peaks m/z 203.943 and m/z 330.84 with intensity set to 1,000.

Parameters have been selected based on raw data observation, especially for the peak width, asymmetry coefficient, and number of overlapping peaks. Then, peak proximity and ratio were set on a reasonable range, in order to challenge the peak deconvolution.

No baseline nor calibration shift were added in the mass and time dimensions, the focus of the simulation is the estimation of the temporal evolution. The code used for the simulation, as well as a representative simulated data file in the HDF5 format are included in the ptairData companion package, also available in Bioconductor. An example of simulation is shown in Figure 5.4.

5.2.2 Software parameters

The simulated files were processed with ptairMS (version 0.1), PTRwid (version 002 IDL), and IDA (version beta 0.9.4.8). Mass calibration was performed using the features at m/z 21.022, 203.943 and 330.84 for the three software, intentionally simulated without overlap at the exact m/z. The calibration stability period (*calibrationPeriod* for ptairMS and *Timming* window for IDA) was set to the acquisition duration, since no calibration shift was added. To ensure a good estimation of the peak shape for the three software, we simulated more single peaks in the intensity range used for the calculation of the peak shape : between 20 and 200 cps for IDA, and *minSig* set to 300 for PTRwid (see Figure 5.3). Note that, for

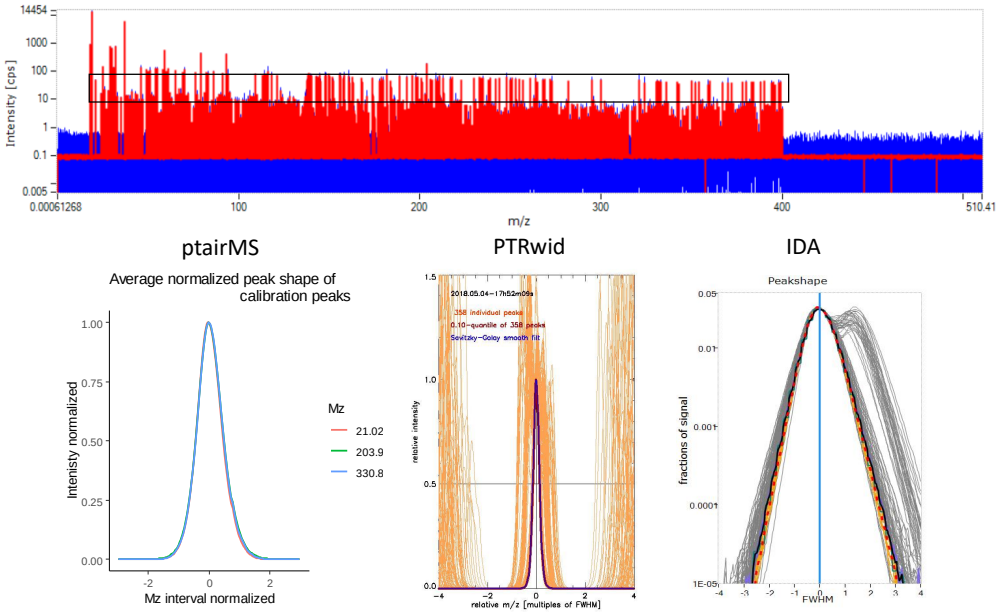


Figure 5.3: Peak shape computation on a simulated file for the three software. A simulated TIS is shown in the top chart of the graph, opened in IDA interface. The boxed signal corresponds to the peak used for the computation of the peak shape, corresponding mostly to peaks without overlap. The resulting peak shape visualisation output of each software is then shown on the bottom chart.

the interested reader wishing to reproduce the results with IDA, the *cps* with this software tool are normalised by the single ion signal (mV $\times$ ns) and multiplied by the bin interval (ns) (which results in an $\approx$ 14.5 factor between the cps values provided by the two other software tools, when a bin period of 0.2 ns is used). Finally, the `sensitivity` parameter for the IDA peak detection was set to 25 %, in order to limit the number of false positives. Other parameters from each software tool were kept to default values.

5.2.3 Results

The three software were compared on ten simulated files, containing a total of 7,028 peaks (Table 5.2). The best precision of peak detection and mass accuracy were obtained with ptairMS, and the peak detection recall was slightly lower than IDA (98.40% vs 98.49%). Of note, the mass accuracy depends only on peak detection, since no mass deviation was included in the simulation. In addition, the reported mass accuracy for PTRwid was computed before calibration: indeed, the masses from the simulated multiple peaks may not match with the internal library of chemical formulae used by PTRwid for calibration, especially for masses $>$ 300 Da (for information, the mass accuracy for PTRwid after calibration is 20 ppm).

Quantification was further evaluated on the peaks which were well detected by all software. The mean absolute percentage error (MAPE) between the estimated temporal evo-

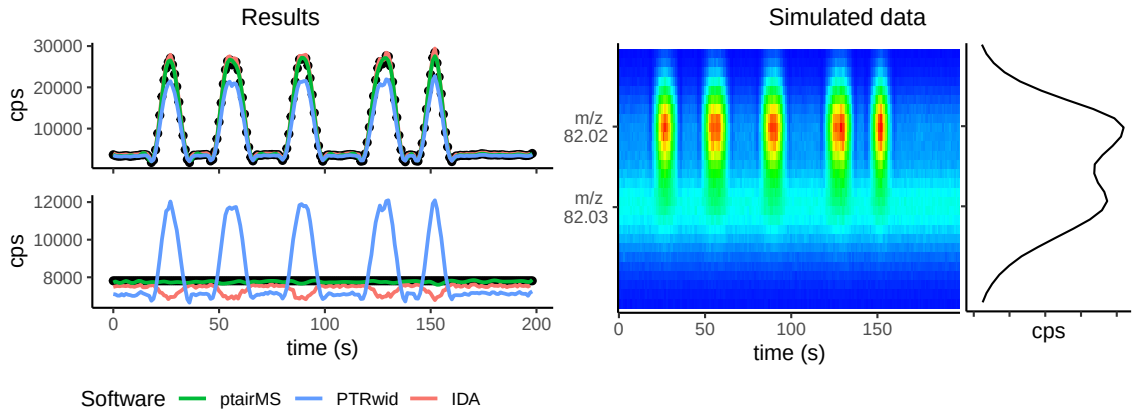


Figure 5.4: Estimation of the temporal profile by ptairMS, compared to the PTRwid and IDA software on simulated data. Right: raw simulated data of two overlapping peaks (as shown in 2D), and the corresponding total mass spectrum. In this particular example, the VOC at m/z 82.02 (respectively, m/z 82.03) was simulated by using an “expiration” (respectively, a “constant”) temporal profile. Left: temporal profiles estimated by the three software tools (solid coloured lines), compared to the simulated profile (ground truth shown as black dots), for the two peaks (top: m/z 82.02, and bottom: m/z 82.03). As observed with the peak at m/z 82.03, the temporal estimations from PTRwid and IDA lead to an erroneous classification of this particular VOC as "expiration" or "ambient air".

lution and the input of the simulation was 4.96 % for ptairMS, 14.65% for PTRwid, and 5.38% for IDA. Finally, we compared the ability to discriminate the compounds from exhaled breath and ambient air, based on two unilateral *t*-tests comparing the intensities in the two acquisition phases (see section 4.1.5). ptairMS was shown to detect the expiration profiles with the highest sensitivity, and with a global accuracy of 99% (compared to 86% and 95% for PTRwid and IDA; Table 5.2). As illustrated in Figure 5.4 with two simulated peaks of close m/z values, an exogenous VOC (i.e., with a constant profile) was classified as “expiration” by both PTRwid and IDA (m/z 82.034), as a result of a less precise temporal estimation. The ptairMS software is therefore well suited for biomarker research with breath analysis.

5.3 Application to real datasets

The ptairMS software has been designed for biomarker discovery in large clinical cohorts. First, it is fast (<1 min for a 3-5 min acquisition) and files can be processed with parallel computing and in a batch mode. Second, studies can be readily incremented with new files (e.g. if new patients are included): only the processing of these new files and the final alignment between samples are performed to update the peak table of the whole cohort.

The ptairMS software is well adapted for exhaled breath analysis, but it can also be used for head space analysis, as we did on publicly available data from truffle (Vita et al., 2015) analyzed with a PTR-TOF 8000 instrument (Ionicon; Figure 5.5 from the Appendix).

Software	ptairMS	PTRwid	IDA
Mass accuracy (ppm)	**3**	12*	5
Peak detection precision (%)	**99.99**	98.87	97.30
Peak detection recall (%)	98.40	87.19	**98.49**
MAPE (%)	**4.96**	14.65	5.38
Expiration sensitivity (%)	**98.53**	91.45	94.52
Expiration specificity (%)	**99.01**	86.31	97.03
Global accuracy (%)	**99.12**	86.73	95.31

Table 5.2: Comparison of peak detection and quantification by ptairMS, PTRwid, and IDA on 10 simulated files (7,028 peaks). The precision (respectively, recall) of peak detection is the proportion of detected peaks which correspond to actual simulated peaks (respectively, the proportion of actual simulated peaks which were detected by the software tools). * The reported mass accuracy for PTRwid was computed before calibration as explained in the text. The Mean Absolute Percentage Error (MAPE) is used to assess the quality of the temporal profile estimation. Expiration sensitivity, specificity, and accuracy refer to the classification of VOC origin as exhaled breath (vs. ambient air). For each metric, the best performance is shown in bold.

These results highlight the ability of the algorithms to adapt to various resolutions, time bin periods, peak shapes, and temporal profiles.

5.4 Discussion

We have developed an innovative workflow for the fast processing of PTR-TOF-MS data from exhaled breath. The suite of algorithms includes untargeted peak detection and deconvolution in the mass dimension, expiration phases detection, modeling of the temporal evolution of the peak intensity during the acquisition, and quantification. Compared to the two existing software, it provides for the first time the required features enabling the analysis of clinical cohorts, with multiple parallel file processing, incremental addition of new patient files, quality control of acquisitions along clinical trials, alignment between the samples, and final quality control to discard exogenous VOCs. The full workflow was implemented in the R package ptairMS, which is publicly available on the Bioconductor platform and includes a detailed tutorial. Raw files from two experimental data sets, as well as one simulated file, are provided in the companion ptairData package. The public availability of all data and source code is of high value for the reproducibility of the analyses and the benchmark of software tools (Wilkinson et al., 2016).

The quality of the untargeted peak detection and absolute quantification was assessed by using a standardised gas mixture: all compounds were detected by ptairMS with an m/z precision inferior to 20 ppm, an intensity error below 8.1% (for compounds with concentrations greater than 19 ppb), an average R2 coefficient with the concentration factor of 0.999, and a CV less than 5%, thus demonstrating the performance of the detection and quantification. However, it is important to note that the standardised gas used does not

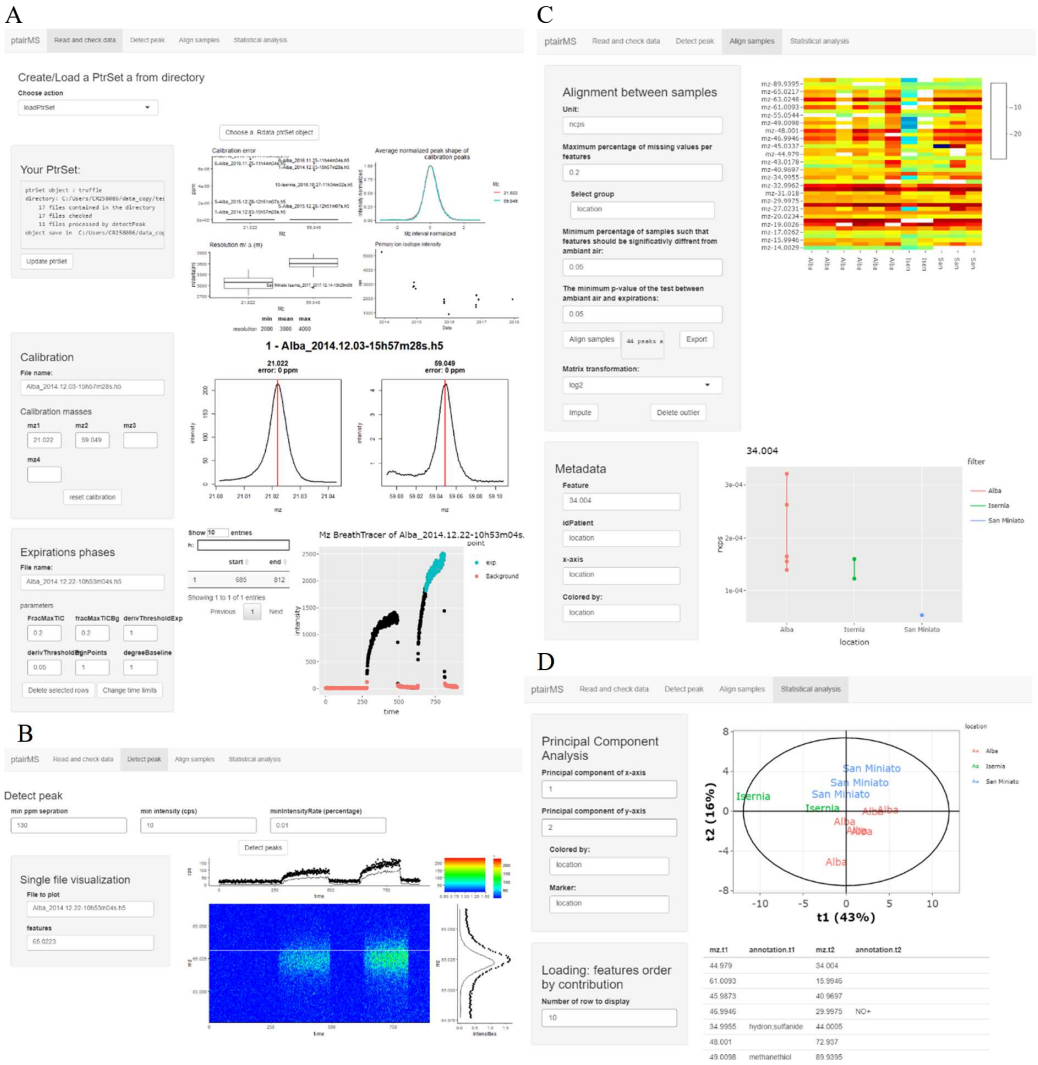


Figure 5.5: Application to the truffle biological matrix (Vita et al., 2015). The publicly available truffle dataset (https://figshare.com/articles/dataset/PTR-ToF-MS_data/7467746) has been generated by head space analysis of truffles, collected at three distinct Italian areas (Alba, Isernia, and San Miniato), with a PTR-TOF-MS 8000 instrument (IoniconAnalytik GmbH, Innsbruck, Austria). Eleven files from the datasets were processed by ptairMS: the temporal phases were defined using the trace at m/z 63.02, corresponding to a known VOCs emitted from truffle fruiting bodies in the literature, namely dimethyl sulphide (Vita et al., 2020). (A) Two head space phases were observed in the "check data" window: empty jar and truffle (first and second temporal peaks, respectively). Only the truffle phase was selected for the subsequent processing. After peak detection (B), and alignment between the samples (C), 44 features were quantified. Clusters corresponding to the origin of the truffles are visible on the score plot from the Principal Component Analysis of their head space profiles (D).

reflect breath matrices. In practice, humidity saturation of exhaled breath biases the VOC quantification in PTR-MS instruments, with divergent behaviour for different substance classes (Trefz et al., 2018). This effect also impacts the proposed correction of the ambient air level (which consists in subtracting the ambient air baseline from the temporal profile estimated for each VOC). Since the exhaled breath and ambient air have different concentrations of humidity, O_2, and CO_2, the direct subtraction should not therefore be considered as an absolute quantification, but rather as a relative concentration, which can be used to compare patients. To further compute accurate concentration differences between inspiratory and expiratory phases, adequate humidity-adapted calibrations are required (Trefz et al., 2018).

A simulation algorithm of PTR-TOF-MS data has been developed for the software performance comparison. It used both real data for the temporal evolution of exhaled breath VOCs, and theoretical modelling for peak shape and noise. It is available on the ptairData Bioconductor package, making possible further comparison and bench marking for exhaled breath PTR-TOF-MS data processing. Parameters of this simulation have been chosen to challenge more the temporal estimation of each peak than the peak picking or the mass axis calibration, which is quite similar for the three software (e.g. stable peak shape, peak separation of at least 150 ppm, no mass deviation and baseline). However, the simulation code may be easily modified to extend the focus of the benchmark.

Since the estimation of temporal profile is a key aspect of breath analysis to determine the VOC origin (i.e. exhaled breath vs. ambient air), we have developed a 2D model based on P-splines regression. Compared to the existing software IDA (Müller et al., 2013) and PTRwid (Holzinger, 2015), which are well suited for single-file, large data from environmental monitoring, we demonstrate that ptairMS is very convenient for breath analysis, achieving highest sensitivity and accurate quantification with an accuracy up to 99%. It should be noted that the temporal estimation of the peak intensities relies on the m/z values previously computed on the total ion spectrum (i.e. these m/z values are not re-evaluated at each time point) which allows a fast computation, but the time deconvolution depends then crucially on the mass detection. The ptairMS algorithms provides precise m/z and intensity estimations, in a computation time ($<$ 1 min) which is compatible with the real-time patient analysis.

Since the resolution of the PTR-TOF-MS does not always allow complete peak separation in the mass dimension, the peak picking algorithm relies on the subsequently steps: denoising, baseline correction, detection of local maxima, and finally deconvolution using parametric peak shape. Since peak shapes observed in TOF analysers are asymmetric, and may change according to the resolution (Müller et al., 2011), we proposed to test both theoretic model functions and estimated shapes from the raw data, and select the most appropriate for each dataset according to the R_2 criterion. This method yielded good estimations and facilitates visualisation and interpretation of the signals. Interestingly,

some recently described algorithms simultaneously perform the three processing steps (denoising, baseline correction and detection of local maxima; Picaud et al. 2018), with the aim to reduce the potential unrecoverable artefacts introduced by the sequential approach.

To impute missing values, ptairMS returns back to the raw data, and re-runs the processing algorithm with flexible parameter settings to extract the raw signals. This method is relatively fast since subsets of the raw data are easily accessible with the HDF5 hierarchical format, and is assumed to be as close as possible to the raw signals. Alternatively, methods based on the table of intensities are also used in metabolomics and in other omics data (Wei et al., 2018). These methods borrow information from features with similar profiles, assuming that values are missing (completely) at random, and include k-nearest neighbours (kNN; Troyanskaya et al. 2001), random forest (RF; Stekhoven and Bühlmann 2011), or singular value decomposition (SVD; Hastie et al. 2001) imputations. To take into account the stochastic process underlying missingness and imputation, multiple imputation approaches are also of interest: such methods perform repeated imputation to generate multiple datasets, which are subsequently used to estimate the mean and the variance of the parameter of interest (e.g. the test statistic; Chion et al. 2021).

Putative annotation is finally performed, based on the matching with a database of 1,488 exhaled breath compound ($\sim$ 400 isotopic masses; Kuo et al. 2020; Drabińska et al. 2021). We observed that about 60% of the detected VOCs have a suggested annotation by ptairMS. The database may be easily updated by the user with the `annotateVOC` function. An alternative approach based on the generation and matching of elemental formulae is used by the PTRwid software (Holzinger, 2015). However, the formula database includes several endogenous compounds that are not found in the exhaled air, and many of the corresponding masses are too close to be distinguished by PTR-TOF-MS. Beyond mass library search and isotope detection, complementary experiments with hyphenated techniques such as GC-MS are required to achieve higher structural identification levels (Sumner et al., 2007) for the most interesting VOCs (Ibrahim et al., 2019; Nardi-Agmon et al., 2016; Wilde et al., 2019).

Importantly, ptairMS automatically suggests optimal values for the parameters, such as the resolution and the peak shape (as evaluated on the calibration peaks), but also the location of spline knots (at higher densities within the expiration phases) and the penalisation for the 2D regression (based on generalised cross validation). This enables to adapt the processing to specific instruments (e.g. with distinct resolutions) but also to various biological matrices (e.g. with different time profiles). As an example, ptairMS was used to process files from both PTR-TOF 8000 and PTR-Qi-TOF instruments (Ionicon). Files from other vendors (e.g. Tofwerk) should be processed accordingly, since they are in the same open source HDF5 format, which is a data storage format of choice within the MS community (Askenazi et al., 2017). Beyond exhaled breath, ptairMS was successfully ap-

plied to atmospheric air data (hospital room and corridor air), headspace analysis from mycobacteria (see the package tutorial) and truffles (Vita et al. 2015; Figure 5.5).

A graphical interface was developed to facilitate data analysis and result interpretation by experimenters (e.g. clinicians). It covers the processing of raw data up to the exploratory data analysis of the cohort, with interactive tables and graphics. Since clinical studies may last several months, or even years, the interface includes a dedicated panel for the real time control of instrument parameters to avoid unwanted effects resulting from drift in temperature, pressure, or variations in the amount of reagent ion. Incremental addition of new patient files is also possible without the need to reprocess all of the previous acquisitions. New features in future implementations will include visualisations (such as the superposition of multiple temporal profiles for several patients), and statistical testing of clinical metadata for each detected VOC.

Altogether, these results demonstrate the value of the ptairMS software as a key resource in breathomics for real-time analysis at the point of care and in biomarker discovery studies, with a high clinical potential for the phenotyping of health and disease, therapeutic drug monitoring, toxicological studies and precision medicine (Fernández del Río et al., 2015; Ibrahim et al., 2021; Löser et al., 2020; Zhou et al., 2017).

Chapter 6

Application to biomarker discovery in the clinic: intubated, mechanically ventilated COVID-19 patients

As of December 2021, about 280 million of people worldwide had been infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and about 5 million had died from coronavirus disease 2019 (COVID-19; coronavirus.jhu.edu). Approximately 5% of patients with COVID-19 will develop acute respiratory distress syndrome (ARDS), septic shock, or multiple organ dysfunction. Around the world, unprecedented research efforts are being focused on the prevention, early detection, diagnosis and management of this lethal disease.

In this context, our breathomics approach with the PTR-TOF-MS technology and ptairMS software tool was ideally suited for the rapid and non-invasive diagnosis of COVID-19 infection. The suitability of this approach was investigated as a part of a research project devoted to severe infections (Rapid rEcognition of Corticostroid sensitive or resistant Sepsis, RHU RECORDS, lead investigator: Prof. Djillali Annane, Intensive Care Unit, Raymond Poincaré hospital). Exhaled breath from mechanically ventilated adults was analysed during the first wave of COVID-19 to assess diagnosic performance in this patient population (Grassin-Delyle et al., 2021).

6.1 Study participants

Forty adults with ARDS were included between March 25^{th} and June 25^{th}, 2020, of whom 28 had proven COVID-19. This prospective study was part of the observational phase of the ongoing RECORDS trial (NCT04280497).

Sample metadata (Table 6.1) included patient demographics (sex, age, body weight, height, and body mass index), clinical and laboratory data (body temperature, Simplified Acute Physiology Score (SAPS II) and SOFA scores (Le Gall et al., 1993; Force, 2012), and serum CRP and creatinine levels), comorbidities (high blood pressure, chronic obstructive pulmonary disease, ischemic cardiac disease, and cancer), ventilation parameters (respiratory rate, positive end-expiratory pressure, and tidal volume) and treatments unrelated to COVID-19 (catecholamines, renal replacement, glucocorticoids, and fludrocortisone).

6.2 Data collection and processing

Each patient's exhaled breath was analysed daily in the morning, from the hospital entry until discharge (death or recover). Measurements were made with a PTR-TOF-MS (Ionicon Analytik GmbH, Innsbruck, Austria) placed outside the patient room, and samples were obtained via a heated transfer line (length: 1.6 m) connected directly to the end of the endotracheal tube (i.e., without disconnection from the mechanical ventilator), with an air flow of 50 mL/min.

Importantly, ventilated patient acquisition presents some differences from classical exhaled breath obtained by direct introduction. First, the air phases between the expiration phases are from medical air (not the usual ambient air), composed of oxygen, in a percentage defined according to the patient's health status (FiO_2, from 22% to 100%), and nitrogen. The room or corridor air should not impact the exhaled breath. Second, previous studies have shown that exhaled VOC concentrations determined with online PTR-TOF-MS may be influenced through distribution of pulmonary ventilation, as the positive end expiratory pressure (PEEP), and fraction of inspired dioxygen (Brock et al., 2017; Trefz et al., 2019a).

To eliminate the dependency on the oxygen concentration in the sample matrix, recordings were performed in patients with a fraction of inspired oxygen set to 100% for at least 3 min, and the acquisition duration was set to 2 min, with an acquisition time unit of 0.1 s. H_3O^+ was used as the primary ion and the instrument settings were as follows: source voltage 120 V; drift tube pressure, 3.8 mbar.

Data were processed daily with the ptairMS software, with a calibration every minute based on the peaks at m/z 21.022, 60.05, 203.94 and 330.8495. The expiration phases where detected using the ions trace of CO_2H^+ (m/z 44.99). Peak detection was performed

	COVID-19 ARDS	Non-COVID-19 ARDS	*p*-value
Number of patients (n)	28	12	-
Males/Females (n)	20/8	6/6	0.28
Age (years)	61 [55-72]	72 [54-79]	0.75
Body weight (kg)	80.0 [66.6-87.6]	86.5 [65.3-94.1]	0.71
Height (cm)	170 [164-175]	173 [169-175]	0.55
Body mass index (kg/m2)	26.3 [23.7-32.4]	28.9 [23.0-30.9]	0.79
SAPS II score	62 [49-68]	46 [40-57]	0.05
SOFA score	11 [7-12]	8 [5-12]	0.37
Body temperature (°C)	37.4 [36.5-38.3]	37.3 [36.8-37.8]	0.84
Respiratory rate (breaths per min)	26 [25-28]	20 [18-23]	$>$ **0.001**
Tidal volume (mL)	420 [400-475]	438 [400-490]	0.99
Fraction of inspired dioxygen (%)	80 [50-100]	48 [31-68]	**0.007**
Positive end-expiratory pressure (PEEP)	10 [8-13]	5.5 [5-8]	$>$ **0.001**
Serum creatinine (mM)	74 [56-137]	67 [44-86]	0.30
Serum C-reactive protein (mg/L)	195 [175-268]	76 [23-119]	**0.002**
Comorbidities: n (%)			
- high blood pressure	11 (39)	6 (50)	0.73
- chronic obstructive pulmonary disease	2 (7)	1 (8)	0.99
- ischemic cardiac disease	5 (18)	3 (25)	0.68
- cancer	2 (7)	3 (25)	0.15
Treatments before admission: n (%)			
- glucocorticoids	1 (4)	3 (25)	0.07
- conversion enzyme inhibitors	5 (18)	1 (8)	0.54
- angiotensin antagonists	2 (7)	2 (16)	0.57
Interventions after admission: n (%)			
- catecholamines	17 (61)	4 (33)	0.17
- renal replacement therapy	9 (32)	0 (0)	**0.038**
Treatments after admission: n (%)			
- hydroxychloroquine	27 (96)	1 (8)	$>$ **0.001**
- remdesivir	2 (7)	0 (0)	0.99
- lopinavir/ritonavir	7 (25)	0 (0)	0.08
- glucocorticoids	11 (39)	6 (50)	0.73
- fludrocortisone	1 (4)	4 (33)	**0.022**
- eculizumab	12 (43)	4 (33)	0.73

Table 6.1: Patient characteristics and treatments, by infection status. The *p*-values result either from a Wilcoxon–Mann–Whitney statistical test in case of a quantitative covariate or from a chi-squared test for qualitative covariates, and from a correction for multiple testing (in bold if < 0.05).

with default values (section 4.1.3), and knots where placed every 0.5 second. An overview of the data with the ptairMS graphical interface is shown in Figure 4.8. After alignment (with a standard deviation of the kernel density set to 40 ppm; section 4.2.1) the following steps were applied:

- Only ions detected in more than 70% of at least one group (COVID vs. non-COVID-19 ARDS) and significantly greater in the expiration phases of at least 5% of the samples were kept, resulting in 81 features
- Missing values were imputed with the ptairMS package (section 4.3)
- Data were log2-transformed
- Outliers (patients with a z-score >3 for at least five features) were discarded
- Saturated ions (acetone, H_3O^+, $H_2O - H_3O^+$, oxygen) and isotopes were removed, resulting in a final table of 65 features

6.3 Data analysis

In the context of biomarker discovery, we first used an untargeted metabolomic strategy to discover the signature associated with COVID-19 ARDS, using the first breath sample available after the admission (section 6.3.1). We then investigated the evolution of each VOC during the hospitalisation time, to validate the results obtained at the first day and to further analyse the difference of VOC concentration evolution between each groups (COVID +/-; section 6.3.2).

6.3.1 Classification for early diagnosis

To build a predictive model for early diagnosis of COVID-19 infection, we used the first breath sample collected for each patient after admission. Ten of the 40 participants had been hospitalised for more than 10 days at the start of the sampling period, and were thus excluded from this first part of the study. Then two outliers where excluded (one COVID-19 negative admitted to the ICU for attempted suicide with medication, and the other COVID-19 positive with very saturated peaks for unknown reasons). This resulted in a subset of 28 patients (18 COVID-19 positive and 10 COVID-19 negative patients), and 65 features. In such a case where the number of features exceeds the number of samples, a particular attention should be paid to the multivariate modelling and to the feature selection methods to avoid over fitting. A summary of all the methods and software tools used in this section is provided in Table 6.2.

Method	Tuned parameter(s)	Feature selection method	Metric used to rank the features	R package
Wilcoxon-Mann-Whitney test	-	*p*-value threshold	*p*-value	phenomis
Principal component analysis (PCA)	-	-	Loadings	ropls
Orthogonal partial least squares (O-PLS-DA)	Number of components	Recursive Feature Elimination	Variable importance in projection	ropls/ caret
Random forest (RF)	Maximum number of nodes in the tree	Recursive Feature Elimination	Variable importance	caret
Elastic net (EN)	Penalisation parameters	L1 - L2 penalisation	Coefficient values	caret
Support vector machine (SVM)	Degree and constant value of the polynomial hyperplanes	Recursive Feature Elimination	Coefficient values	e1071 / sigFeature

Table 6.2: Summary of the statistical methods used for the prediction of the COVID-19 status. Further description of the R packages are provided in the corresponding publications: phenomis (Imbert et al., 2021); ropls (Thévenot et al., 2015); caret (tutorial); e1071 (tutorial); sigFeature (Das et al., 2020).

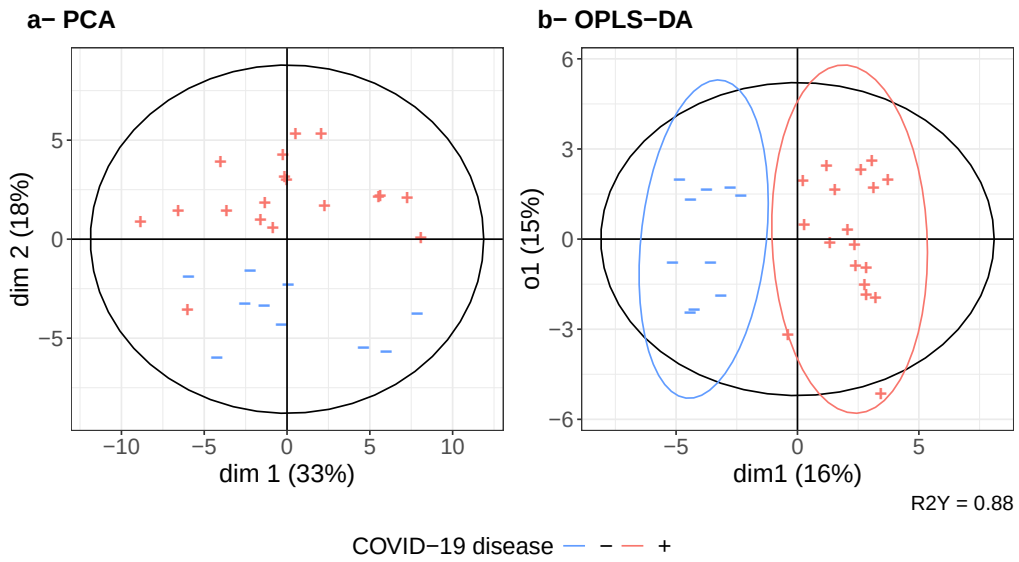


Figure 6.1: (a) Score plot from the principal component analysis (PCA) showing the two first components. (b) Score plot from the Orthogonal Partial Least Squares - Discriminant Analysis according to the predictive component (abscissa) and the first orthogonal component (ordinate).

Methods

Learning approaches: We first used principal component analysis (PCA) to project and observe the data in a lower dimension, and see if there are components that discriminate the two groups of patients.

We then performed univariate analysis, using a Wilcoxon-Mann-Whitney test (section 3.2.1), to detect significant features individually, without taking into account the correlations between the VOCs at this stage. The *p*-values were adjusted to control the false discovery rate (FDR; Benjamini and Hochberg 1995) at a 5% threshold.

Finally, we tested four multivariate machine learning models (section 3.2), with different mathematical backgrounds: Orthogonal Partial Least-Squares discriminate analysis (O-PLS DA), Random Forest (RF), Elastic Net (EN), and Support Vector Machine (SVM) with a polynomial kernel. Parameters from each model were tuned using a grid search (Table 6.2). These classification methods are widely used within the omics community, including metabolomics (Guo et al., 2010; Rinaudo et al., 2016). Multivariate analysis is complementary to univariate hypothesis testing since it enables to build predictive models, and since it takes into account interactions between features. Benchmarking several machine learning approaches aims at increasing the robustness of the results and at improving the predictions. Indeed, depending on the data structure, some models may perform better than others. The prediction performances were compared using the $Log\ Loss$ and AUC complementary metrics (equation 3.15), computed with a stratified 10-fold cross validation, repeated four times.

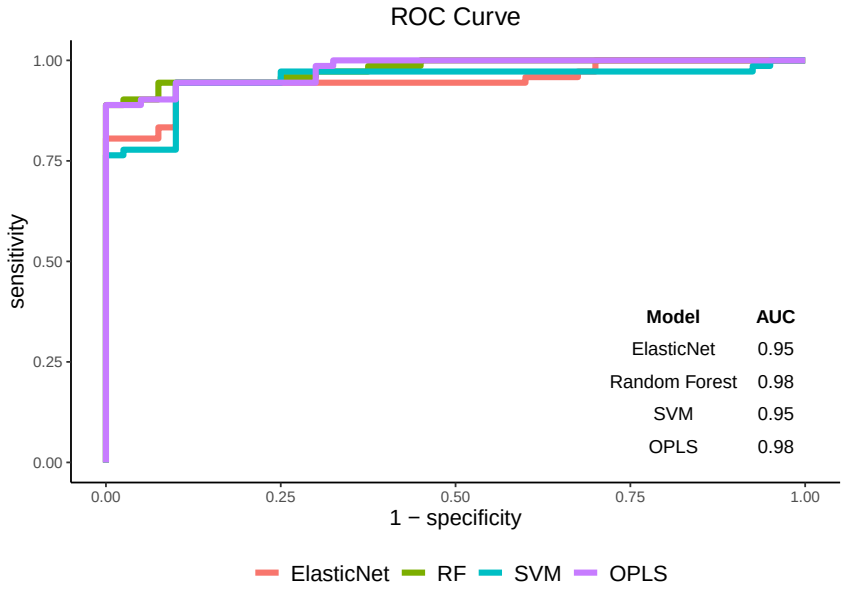


Figure 6.2: ROC curves from the the four complementary machine learning approaches (EN, RF, SVM, and OPLS-DA).

Model	Accuracy	Sensitivity	Specificity	AUC	Log Loss mean	Log Loss sd	Number of features
Elastic Net	0.93	0.90	0.94	0.95	7.45	0.54	22
Random Forest	0.93	0.90	0.94	**0.98**	**7.38**	0.21	16
SVM	0.93	0.90	0.94	0.95	7.68	0.48	22
O-PLS-DA	0.93	0.90	0.94	**0.98**	11.77	**0.10**	**12**

Table 6.3: Comparison of model performances. The accuracy, sensitivity, specificity, AUC and Log Loss were computed using a 10-fold cross-validation, repeated 4 times (metrics defined in the section 3.2.2). The standard deviation (sd) across the cross-validation folds is also indicated, and provides information about the prediction variance. RF proved to be the best performing model, according to the Log Loss metric, with the same AUC value as O-PLS-DA. O-PLS-DA selected the lowest subset of features (12), which provided high accuracy (93%), with the lowest variance of prediction (Log Loss sd of 0.1).

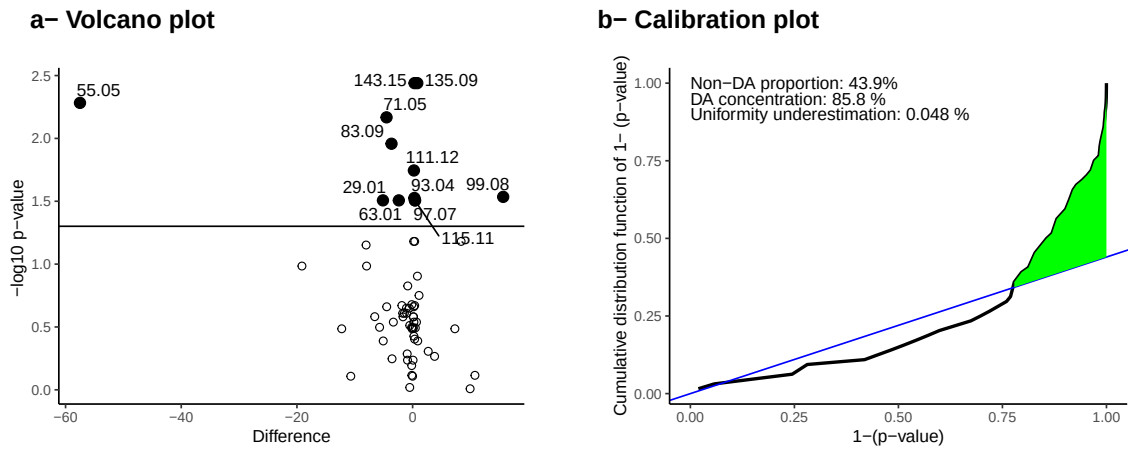


Figure 6.3: Quality plots for the *p*-values from the Wilcoxon-Mann-Whitney test. (a) The volcano plot shows the $-log10$ corrected *p*-value as a function of the difference between the group medians (a negative value indicates that the VOC median concentration is greater in the COVID- group compared to the COVID+ group). The FDR threshold (0.05) is shown as an horizontal line. The selected features are labelled with their m/z value. Ions at m/z 55.05 shows the greatest difference in concentration between the two groups (+ 60 ppb for covid negative patients) and ions 99.08 the greatest difference in the opposite direction (+ 20 ppb for the positive group) (b) Calibration plot of the raw ordered *p*-values, provided by the cp4p R package (Gianetto et al., 2015). This plot checks the assumption underlying the FDR correction: the *p*-values of non differentially abundant (DA) features are uniformly distributed on the [0,1] interval (the corresponding cumulative distribution is displayed as the blue line), while the remaining *p*-values (corresponding to DA features) are concentrated nearby zero (green area).

Feature selection and ranking: To select the most relevant features for the COVID-19 diagnosis among the 65 VOCs from exhaled breath, we used the following feature selection and ranking methods (Table 6.2):

- EN models internally perform feature selection during training through the L1 and L2 penalisation (section 3.2.1), and features which are not selected by the model get their coefficients set to zero. The selected features were ranked by significance using the ordered absolute values of the coefficients
- SVM, RF and O-PLS-DA feature selection was performed using Recursive Feature Elimination (RFE; section 3.2.2). At each iteration, features were ranked using, respectively, the estimated coefficient values, the feature importance, and the Variable Importance in Projection metrics (VIP; equation 3.10)

Results

The second dimension of the PCA (18% of the total variance) was shown to provide a discrimination of the ARDS patients according to their COVID-19 status (Figure 6.1a), suggesting that COVID-19 was associated with a specific signature in the expired breath.

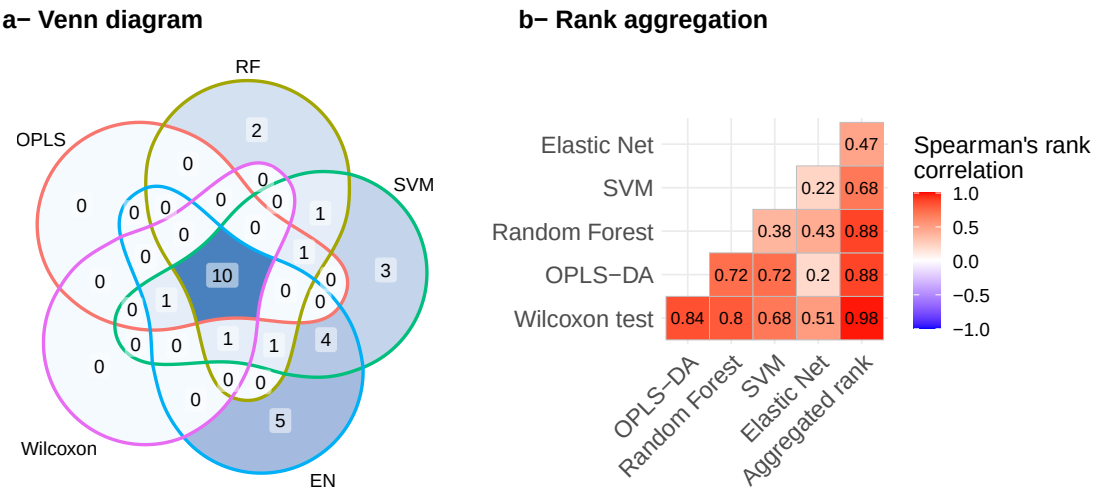


Figure 6.4: Comparison of feature selection methods. (a) Venn diagram comparing the features selected by each method: Wilcoxon test, RF, SVM, EN and O-PLS-DA. Ten features were common to all selections, and 29 were selected in at least one model. (b) Matrix of the Spearman correlations between the ranks of the 10 common features selected by each classifier (Table 6.2). The ranks were then aggregated by maximising the sum of the Spearman correlations with each classifier ranking (RankAggreg R package; Pihur et al. 2009). RF, OPLS-DA and Wilcoxon test metrics proved to be the most correlated.

The univariate analysis (Wilcoxon-Mann-Whitney test) highlighted 12 significant VOCs at an FDR threshold of 5% (quality plots showed that the *p*-values were moderately well calibrated for the FDR correction; Figure 6.3).

The use of four complementary machine learning algorithms enabled to achieve an accuracy of 93% for all four classifiers, based on the selection of 22, 16, 22 and 12 features for the EN, RF, SVM, and OPLS-DA, respectively (Figure 6.2 and Table 6.3). Although the models gave quite similar performances, RF was shown to be the best performing model according to the Log Loss criterion. Ten VOCs were common to all selection approaches (Figure 6.4a), and 29 were selected by at least one method. Of note, the OPLS-DA modelling, which is very popular in metabolomics, yielded a robust (1 orthogonal component only) and significant model (as assessed by permutation testing of the response labels; Szymańska et al. 2012) with good predictive performances (Q_2= 0.69; Figure 6.1b).

Finally, the ranks of the 10 features selected by the five methods were aggregated according to their rankings by the specific metrics (Table 6.2), by maximising the sum of the Spearman correlations with each of the model rankings (RankAggreg R package; Pihur et al. 2009). The correlation matrix is shown in Figure 6.4b, and the putative VOC annotations provided by ptairMS are shown in Table 6.4.

m/z	matched m/z	matched formula	putative annotations	*p*-values	
				mean	trend
135.09	135.093	$C_7H_{15}Cl + H^+$	1-chloroheptane	**0.008**	**0.036**
143.15	143.143	$C_9H_{18}O + H^+$	nonanal	**0.002**	0.200
71.05	71.049	$C_4H_6O + H^+$	but-2-enal and 4 other matches	0.100	0.474
83.09	83.086	$C_6H_{10} + H^+$	hexa-2,4-diene and 14 other matches	0.655	0.825
55.05	55.054	$C_4H_6 + H^+$	but-1-yne and 3 other matches	0.425	0.232
111.12	111.117	$C_8H_{14}+H^+$	octa-2,4-diene and 17 other matches	**0.020**	**0.040**
99.08	99.080	$C_6H_{10}O + H^+$	2-methylpent-2-enal and 9 other matches	**0.007**	**<0.001**
93.04	-	-	-	0.404	0.334
29.01	29.013	$N_2 + H^+$	nitrogen	0.139	0.518
115.11	115.112	$C_7H_{14}O + H^+$	heptanal and 6 other matches	0.229	0.144

Table 6.4: Putative annotation of features selected with the five classifiers, ordered by the aggregated rank. Only the first putative VOC annotation is shown. The *p*-values from the two longitudinal tests (section 6.3.2) are also indicated.

6.3.2 Time course modelling

We then modelled each VOCs concentration, noted Y as a function of the hospitalisation time t, to both validate the previous selected VOCs, and to further characterise the evolution of exhaled breath VOC concentrations of ventilated patients along hospitalisation time. Patients with only one acquisition were deleted, resulting in 25 positive against 11 negative patients.

Methods

We used a nonlinear mixed-effects model, as introduced in section 3.2.3. The fixed effect is the evolution of the VOC concentration as a function of the period of mechanical ventilation t, and the random effect the individual-specific deviation from this fixed effect.

As we have no *a priori* knowledge on the trend of VOC concentrations over days, we used a semi-parametric modelling using splines (section 3.1.2), with knots placed approximately every 3 days: for instance, for $t \in [1, 10]$, we used $K = 4$ B-splines. It results in the following final model for each patient i and each time j:

$$Y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + b_i + \epsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad b_i \sim \mathcal{N}(0, \tau^2) \quad \mathbf{b} \perp\!\!\!\perp \boldsymbol{\epsilon} \tag{6.1}$$

where $\sum_{k=1}^{K} \beta_k b_k$ is the fixed effect with b_k the B-spline basis, and the intercept b_i is the random effect per patient with variance τ^2.

To identify features with significantly different (respectively, means and trends) between the two groups (COVID +/-), we introduced a second fixed effect with the binary variable z_i (where $z_i = 1$ if the patient i is positive to the COVID-19 infection, and 0 otherwise), and performed an F-test using (see section 3.2.3) each of the following two models:

$$Y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + \left(\alpha_0 + \sum_{k=1}^{K} \alpha_k b_k(t_{ij})\right) \times z_i + b_i + \epsilon_{ij} \tag{6.2}$$

1. $H_0 : \alpha_{\mathbf{0}} = 0 \; vs \; H_1 : \alpha_0 \neq 0$, tests if there is a difference of value for the intercept (mean)

2. $H_0 : (\alpha_1, ...\alpha_K) = \mathbf{0} \; vs \; H_1 : (\alpha_1, ...\alpha_K) \neq \mathbf{0}$, multiple test of length K to identify differences of trend

The first test will identify features with a difference in concentration means at t_0 between the two groups, whereas the second test will identify VOCs with a different trend (e.g. increase or decrease). A multiple test could be performed to test both hypotheses at the same time, as described in section 3.2.3; however, it is interesting to specifically distinguish VOCs with a higher concentration in one group but with the same evolution, from those with a different trend between the two groups. The *p*-values were adjusted for the false discovery rate (Benjamini and Hochberg, 1995) for each test. We also test several time limits for the hospitalisation time t : 10 to 60 days (the maximum) with a step of 5 days.

Results

Features with a *p*-value $<$ 0.05 after correction for at least one test were selected (False Discovery Rate threshold of 5%). Four VOCs, also selected by the models in the first *diagnostic* approach, were identified (m/z 99.08, 111.12, 135.09, and 143.15; Figure 6.5), and putatively identified as methylpent-2-enal, 2,4-octadiene, 1-chloroheptane, and nonanal (Table 6.4). The VOC concentrations of all of these candidate biomarkers were significantly higher in the breath of patients with COVID-19 ARDS, and tended to decrease over the first 10 days of hospitalisation, except for m/z 145.15 which remains quite stable during this period (Figure 6.5). After 10 days, the evolution remains relatively constant for all the four VOCs.

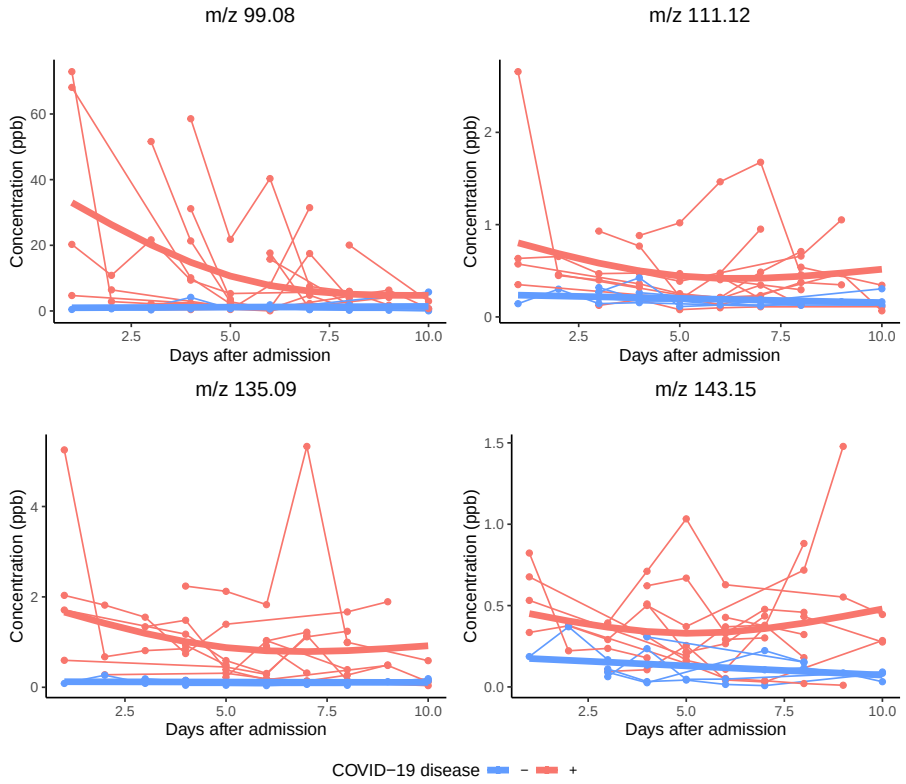


Figure 6.5: Longitudinal analysis of VOCs in expired breath along the first 10 days. The four features (m/z 99.08, 111.12, 135.09, and 143.15) contributing the most to the longitudinal analysis of the intubated, mechanically ventilated patients with COVID-19 ARDS (in red, n = 12) or non-COVID-19 ARDS (in blue, n = 6) are shown. All the points for a given patient are connected, and the bold lines correspond to the fixed effect for each group.

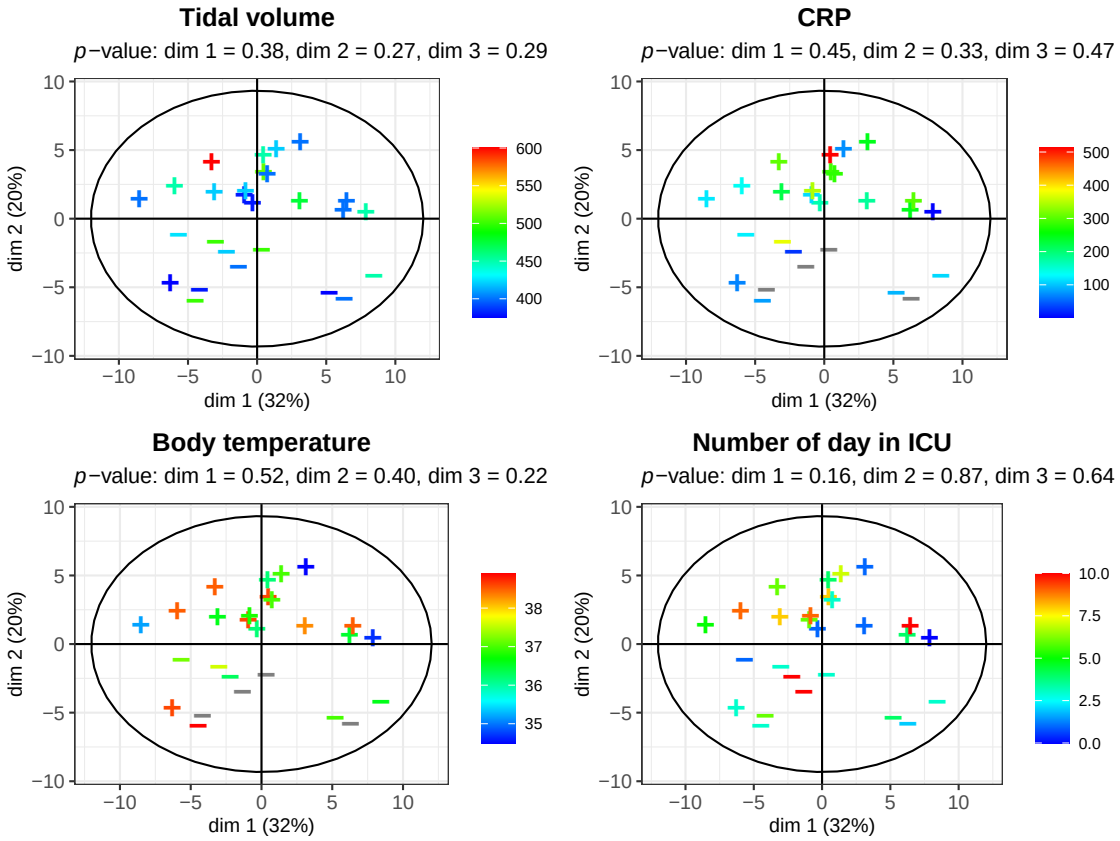


Figure 6.6: Analysis of potential relationships between clinical covariates and the COVID-19 status. The results for four covariates, namely the tidal volume, the serum C-reactive protein level (CRP), the body temperature, and the number of days in the ICU, are shown as a score plot from the principal component analysis, coloured according to the covariate values. The *p*-values from the Pearson test of the correlation between the covariate and the three principal components is shown at the top of each plot. +: positive COVID-19 status; -: negative COVID-19 status.

6.4 Evaluation of potential interfering factors

When building our statistical classifier for the diagnostic of the COVID-19 status, we needed to check that none of the other external covariates (e.g., clinical and demographic variables) had an impact on the VOC concentrations that would interfere with the model's predictions (e.g., underestimating or masking the differences between groups).

We therefore investigated the potential associations between the VOC concentrations and all the available covariates listed in Table 6.1, including patient demographics, clinical and laboratory data, comorbidities, ventilation parameters (respiratory rate, positive end-expiratory pressure, and tidal volume) and treatments unrelated to COVID-19. We studied separately the covariates according to whether they were associated with the COVID-19 status or not, as determined by a Wilcoxon-Mann-Whitney test (respectively, a chi-squared test) for quantitative (respectively, qualitative) covariates, and a correction for multiple testing (Table 6.1). Note that the fraction of inspired oxygen was not considered hereafter since its value was set to 100% before the acquisitions.

Covariates with no correlation to the COVID-19 status

For the covariates that were not significantly related to the COVID-19 status, the association between all detected VOC concentrations and the covariate was first tested by using a univariate analysis (Pearson correlation test for quantitative covariates and Wilcoxon-Mann-Whitney test for categorical covariates, followed by a correction for multiple testing). No significant association was detected at the 5% threshold.

We also applied a multivariate analysis to the correlations between the covariate and each of the first three components from the PCA of the VOC dataset. Again, no significant correlation was observed (Figure 6.6).

Covariates correlated to the COVID-19 status

For the continous covariates significantly related to the COVID-19 status, namely the positive end-expiratory pressure (PEEP), the respiratory rate, and the serum C-reactive protein (CRP), we further checked for associations within each of the two COVID-19 groups separately. This was necessary because the VOCs of interest were related to COVID-19 status, and hence were also correlated to PEEP, the respiratory rate and CRP when the whole cohort was considered (Figure 6.7a). In contrast, the correlation coefficient was low ($r < 0.4$) and the associated *p*-value was not significant when the correlation was assessed within each group separately (i.e., when the COVID-19 status was matched; Figure 6.7b-c-d).

For the three qualitative covariates significantly related to the COVID-19 infection (i.e. corresponding to treatments or intervention after admission), the number of samples

does not enable to test a possible link with the VOC biomarkers, either because the treatment perfectly matches the COVID-19 status (e.g. in the case of hydroxychloroquine which was administrated to the COVID positive patients only), or because two few patients were treated (fludrocortisone: 5 patients, and renal replacement therapy: 9 patients).

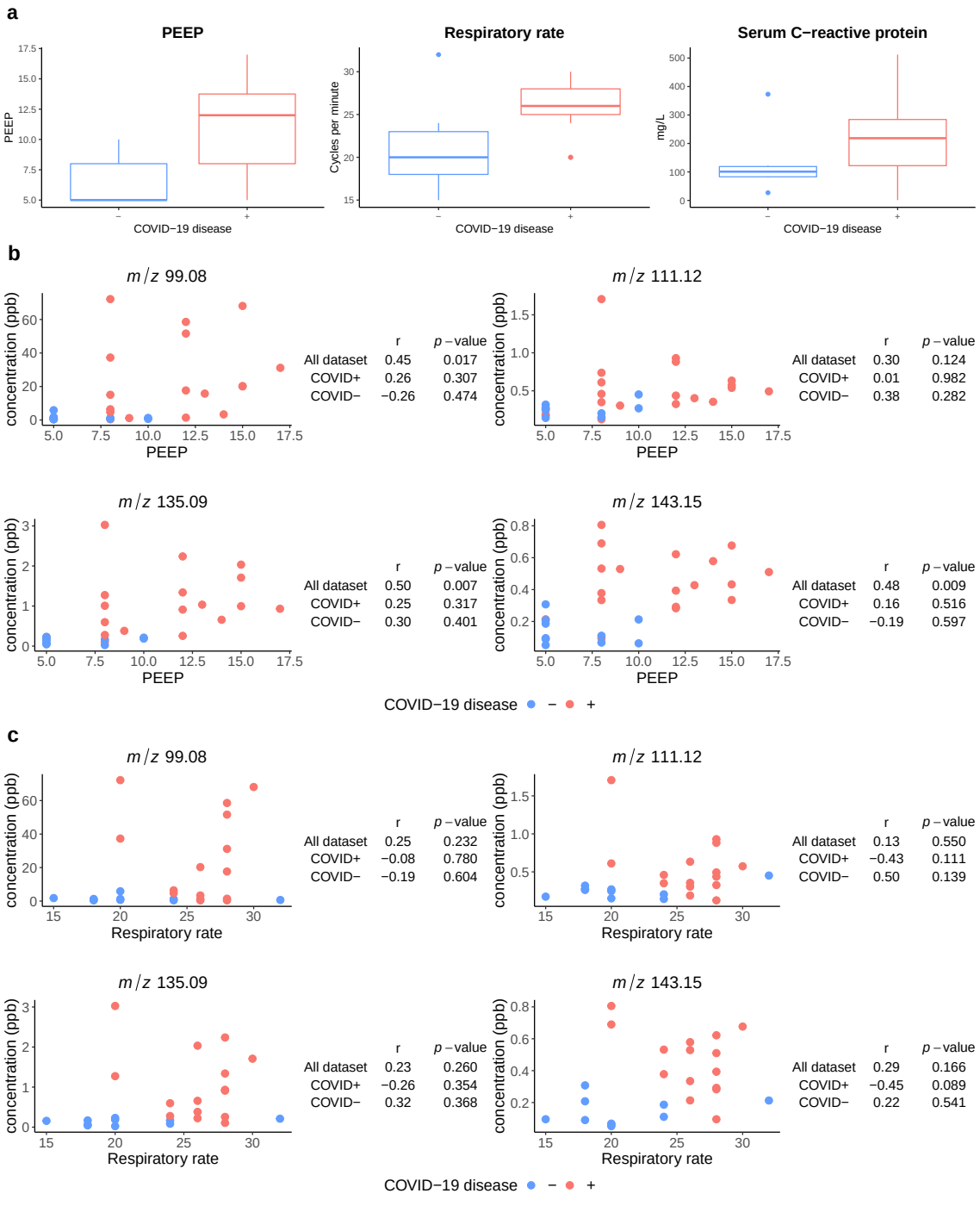
6.5 Discussion

We applied the ptairMS software to analyse the exhaled breath from mechanically ventilated adults with acute respiratory distress syndrome (ARDS). Our data processing and analysis workflow, including both classification and time course modelling, resulted in the selection of four VOC candidate biomarkers for the diagnosis of COVID-19 infection. This study thus provides a proof of concept for the measurement of VOCs and the determination of a specific VOC breathprint in the exhaled breath from patients with COVID-19-related ARDS requiring invasive mechanical ventilation in the ICU (Grassin-Delyle et al., 2021).

Four distinct supervised machine learning models were compared, namely Orthogonal Partial Least Squares - Discriminant Analysis (OPLS-DA), Support Vector Machine (SVM), Random Forest (RF), and Elastic Net (EN). An accuracy of 93% for the prediction of the COVID-19 infection was achieved (90% sensitivity and 94% specificity) for all classifiers, based on a 16 VOC signature selected by the RF algorithm. According to the *Log loss* metric criteria, RF provided the most confident prediction. The OPLS-DA classifier also achieved good performances (AUC 0.98) with only 12 selected features, yet with lower confidence.

Popular classifiers such as XGBoost (de Clercq et al., 2020; Stamate et al., 2019), or Artificial Neural Network (Pomyen et al., 2020) were not applied to this study, due to the limited number of samples (28) and the resulting high risk of overfitting of these complex algorithms (which rely on a large number of parameters). Also due to the low number of observations, we did not divide the data into Training-Test-Validation subsets. Our approach thus takes advantage of all the information available, but may result in biased (overoptimistic) predictions. Our observations thus require to be validated on an external and larger cohort.

Feature selection methods, including recursive feature elimination (RFE), were applied to each model. Ten of the 65 initial VOCs were selected by all classifiers, and 29 by at least one of them. Interestingly, application of the "statistical" RFE approach proposed by Rinaudo et al. (2016) resulted in an accuracy of 90% with the RF algorithm and a five feature signature (m/z 143.1451, 135.089, 55.05, 71.05, and 83.09), which is included in the selection provided by the classical RFE method. The biosigner approach differs from the classical RFE in two aspects: 1) the significance of a feature subset is estimated by comparing the model predictions before and after random permutation of the intensities of


a
PEEP
Respiratory rate
Serum C−reactive protein
PEEP
Cycles per minute
mg/L
COVID−19 disease
b
m/z 99.08
m/z 111.12
m/z 135.09
m/z 143.15
concentration (ppb)
PEEP
r p − value
All dataset 0.45 0.017
COVID+ 0.26 0.307
COVID− −0.26 0.474
All dataset 0.30 0.124
COVID+ 0.01 0.982
COVID− 0.38 0.282
All dataset 0.50 0.007
COVID+ 0.25 0.317
COVID− 0.30 0.401
All dataset 0.48 0.009
COVID+ 0.16 0.516
COVID− −0.19 0.597
COVID−19 disease − +
c
m/z 99.08
m/z 111.12
m/z 135.09
m/z 143.15
concentration (ppb)
Respiratory rate
r p − value
All dataset 0.25 0.232
COVID+ −0.08 0.780
COVID− −0.19 0.604
All dataset 0.13 0.550
COVID+ −0.43 0.111
COVID− 0.50 0.139
All dataset 0.23 0.260
COVID+ −0.26 0.354
COVID− 0.32 0.368
All dataset 0.29 0.166
COVID+ −0.45 0.089
COVID− 0.22 0.541
COVID−19 disease − +

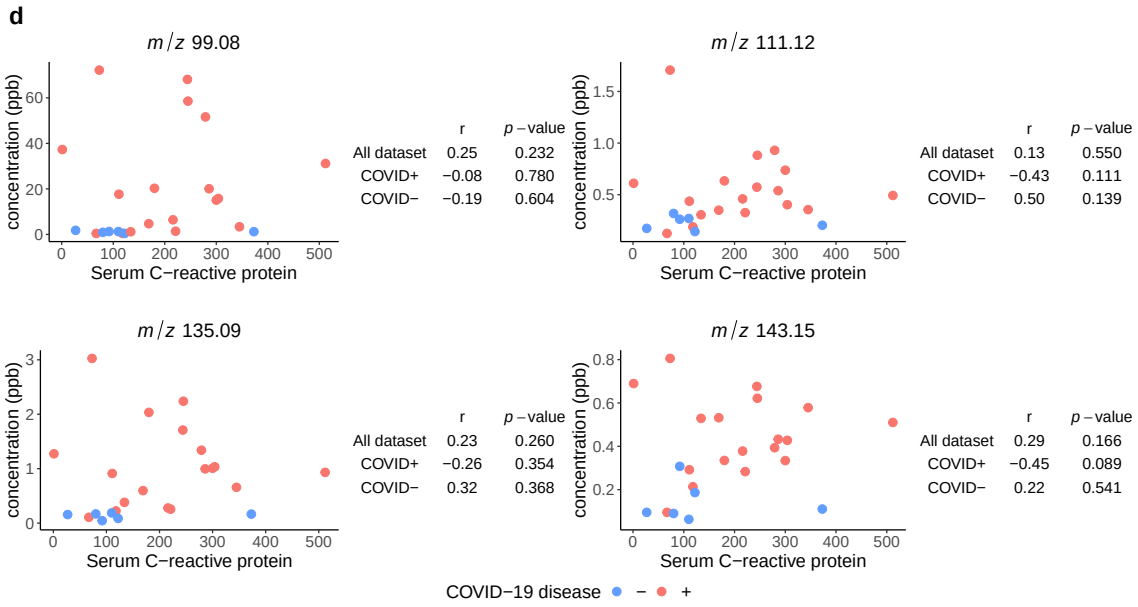


Figure 6.7: Study of the impact of the positive end-expiratory pressure (PEEP), the respiratory rate, and the serum C-reactive protein (CRP), on the relationship between each of the four VOC biomarkers and the COVID-19 status. a. PEEP, respiratory rate, and CRP values according to the COVID-19 status. b-d. VOC concentrations as a function of PEEP (b), the respiratory rate (c), and CRP (d). The Pearson correlation coefficients (r) and the *p*-values from the correlation tests (either computed on the whole dataset or for each COVID-19 subset separately) are indicated.

these features in test subsets generated by resampling, and 2) the whole feature selection procedure is repeated recursively until all features of the selected subset are found significant, or until there is no feature left to be tested (Rinaudo et al., 2016). Its application to the COVID-19 study highlighted very short signatures which still provide high prediction performances.

A time course analysis, using mixed effect models across the hospitalisation time and a Fisher test, was also used. It confirmed that four of the selected features had a significantly different behaviour between the two groups (m/z 99.08, 111.12, 135.09, and 143.15), with a significantly higher concentration in the breath of patients with COVID-19 infection, and a tendency to decrease over the first 10 days of hospitalisation. The fact that some features were selected only by the classification approach at t_0 but not by the longitudinal analysis may be explained by the fact that these features are only observed at the beginning of the infection. Interestingly, the majority of these VOCs have higher concentrations in the breath of COVID-19 *negative* patients (m/z 71.05, 83.09, 55.05 and 29.01), in contrast to the four biomarkers confirmed by the longitudinal analysis.

The time course methodology used may be easily extended to more than two classes (e.g. in the case of multiple levels of infection severity), by transforming the categorical response z with Q levels in $(Q-1)$ binary (dummy) variables : $z^q = 1$ for the presence of

the category q. Then, the variables are include in the mixed effect model :

$$Y_{ij} = \underbrace{f_{\boldsymbol{\beta}}(t_{ij})}_{\text{effect of the first class}} + \sum_{q=2}^{Q} \underbrace{f_{\boldsymbol{\alpha}^q}(t_{ij}) \times z_i^q}_{\text{shift for each other class q}} + b_i + \epsilon_{ij}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $b_i \sim \mathcal{N}(0, \tau^2)$, $\mathbf{b} \perp\!\!\!\perp \boldsymbol{\epsilon}$, for patient i and time point j. Then to test if there is at least one category that differ from the first, we perform multiple Fisher tests of dimension $(q-1)$ for intercept test, and $(q-1)\times$ (number of parameters of the function f) for trend test (e.g number of knots) on multiplying coefficients $\boldsymbol{\alpha}$: $H_0 : (\boldsymbol{\alpha^2}...\boldsymbol{\alpha^Q}) = \mathbf{0}$ vs $H_1 : (\boldsymbol{\alpha^2}...\boldsymbol{\alpha^Q}) \neq 0$.

We investigated potential interfering factors (i.e. covariates with significant median differences between the COVID-19 and non-COVID-19 subgroups). In particular, COVID-19 infected patients had a higher respiratory rate, fraction of inspired oxygen (FiO_2), PEEP, and CRP values on admission. The respiratory rate, PEEP and CRP, however, were not found to interfere with the VOC predictive signature. Furthermore, FiO_2 was set to 100% during all acquisitions to avoid any impact of dioxygen variations on PTR ionisation (Trefz et al., 2019a). The hydroxychloroquine treatment may also be a potential confounder, since it was administrated specifically to the patients with COVID-19 ARDS. However, no correspondence was observed between the VOCs described in the present study and the molecular masses of the known metabolites of hydroxychloroquine. In addition, the observed concentrations of the VOC biomarkers decreased with time, whereas the hydroxychloroquine dosage was constant during hospitalisation.

In line with a previous report on ARDS analysis by GC-MS (Bos et al., 2014b), the VOC concentrations described in our study were not correlated with the severity of illness (as judged by the SAPS II and the SOFA scores). This finding suggests that the exhaled breath signature is a marker of COVID-19 *per se*, rather than a marker of the severity of illness. Likewise, the VOC concentrations were not correlated with the viral load (as independently determined by Polymerase Chain Reaction, PCR), suggesting that this signature may be a marker of the disease related to SARS-CoV-2 rather than of the virus carriage.

Two of the four prominent VOCs (with putative annotation: methylpent-2-enal and nonanal) are aldehydes, while 2,4-octadiene is an alkadiene. These three compounds are known to be expressed in breath (van de Kant et al., 2013; Corradi et al., 2004). Nonanal is a sub-product of the destruction of the cell membrane as a result of oxidative stress; reactive oxygen species may be generated by various types of inflammatory, immune and structural cell in the airways (Rahman, 2003).

A critical issue in breath analysis is the standardisation, to make results from independent studies comparable (Miekisch et al., 2012; Herbig and Beauchamp, 2014; Bruderer

et al., 2019; Henderson et al., 2020). Indeed, breath composition is influenced not only by ambient inhaled air, but also by many external factors, as shown in several studies of the Rostock Medical Breath Research Analytics and Technologies (ROMBAT) team: the body position (Sukul et al., 2015), exhalation strength (Sukul et al., 2016), upper-airway restrictions (Sukul et al., 2017), menstrual cycles (Sukul et al., 2018), medication, specific dietary, or even sampling procedures (e.g. the use of Tedlar bags; Miekisch et al. 2008). Importantly, a particular attention should be paid during the design of the study to the matching of patients and sampling conditions between the groups of interest. Finally, a validation study using similar sampling methods and processing parameters is of critical importance for the clinical use of the candidate biomarkers.

Since the end of 2020, other studies were performed for the early diagnosis of COVID-19 infection from exhaled breath on non ARDS patients, with different MS methods, including GC coupled to ion mobility (GC-IMS; 98 patients, sensitivity and specificity: 82.4% and 75%; Ruszkiewicz et al. 2020), PTR-TOF-MS (340 patients, accuracy: 81.2%; Liangou et al. 2021), GC-MS (81 patients, sensitivity and specificity: 68% and 85%; Ibrahim et al. 2021); GCxGC-TOF-MS on exhaled breath condensate (EBC; 37 patients, AUC=0.98, accuracy: 100%; Barberis et al. 2021). The VOCs selected by these studies differ from our 4 biomarkers, which may be explained by the fact that the progression of the physiologic response of non ARDS patients is different compared to severely ill and mechanically ventilated patients. The differences between the results may also result from distinct sampling procedures, or specific ionisation selectivity and sensitivity from the MS instruments. Altogether, these candidates provide a broader picture of the COVID-19 physio-pathology, and the results from these studies highlight the potential of exhaled breath for early and non invasive diagnosis.

Part III

Conclusion and perspectives

Conclusion

In this thesis, we have developed innovative tools and methods for biomarker discovery in exhaled breath by means of Proton Transfer Reaction Time-of-Flight Mass Spectrometry (PTR-TOF-MS), from the raw data processing up to the statistical analysis for clinical applications.

We have developed the first freely available workflow for the PTR-TOF-MS data preprocessing from exhaled breath (Chapter 4), starting from the raw data files, and providing as output the sample by variable table of intensities. Compared to existing software, it provides new features for the monitoring of cohorts from exhaled breath. Especially, an innovative 2D model based on P-splines regression enables a precise estimation of the peak evolution over the acquisition time. The comparison on simulated data showed that the developed methods clearly improve the classification of the VOC origin (exhaled breath or ambient air), which is of critical interest for biomarker discovery. The developed workflow has been implemented in the R package ptairMS, which is publicly available on the Bioconductor platform, and includes a detailed tutorial and a graphical interface, which makes it easy for clinicians to use. Our software is already used in routine at the Exhalomics platform located within the pneumology department from the Hôpital Foch (Suresnes, France), to process the acquisitions from breathing patients. Several clinical studies are currently underway, mainly in pneumology, infectious diseases and oncology.

Our methodology then allowed the longitudinal analysis of intubated, mechanically ventilated patients in record time (less than 6 months between the inclusion of the first patient and the submission of the manuscript), and enabled to discover a biomarker signature of four VOCs for the diagnosis of COVID-19 infection (Chapter 6; Grassin-Delyle et al. 2021). The currently most used method for the diagnosis of COVID-19 is nasopharyngeal swab collection followed by reverse transcriptase-polymerase chain reaction analysis (RT-PCR): this approach is invasive, requires instrumentation in laboratories, and has very high specificity but moderate sensitivity (Zitek, 2020). Diagnosis with exhaled breath analysis is thereby of major interest for high-throughput population testing, since it is totally non-invasive, painless, and gives the diagnosis in real-time. The design and commercialisation of a breath test for COVID-19 infection is currently a very competitive field around the world. Our VOCs signature is therefore the subject of a European patent, and a validation on a larger and independent cohort is in progress.

Our work therefore provides the scientific community with the computational methods and tools to conduct clinical studies on exhaled breath through the PTR-TOF-MS technology. It paves the way for new rapid and non-invasive tests at the patient bedside for diagnostic purposes, monitoring of treatment response, or high-throughput population screening.

Perspectives

Bayesian deconvolution

Our proposed method for 2D peak deconvolution involves sequential steps, starting with the detection of peaks in the mass dimension and followed by the building of the 2D model. This is therefore an approximation since we perform the peak detection in a single dimension once all mass spectra have been summed. Bayesian deconvolution methods may therefore be a valuable alternative, since they include i) combined estimation of peak locations and intensities, ii) denoising, and iii) baseline removal in 2D. In particular, non-parametric Bayesian approaches allow to separate the baseline component from the set of peaks, without using a parametric model for the baseline, and to deconvolute the peaks without imposing a total number of peaks *a priori*.

The Bayesian approach considers each of the unknown quantities that we want to estimate as random variables, with a prior probability law. These quantities are then updated from the observations, through the Bayes' rule (Gelman et al., 2004). This approach has the advantage to provide knowledge of uncertainties and credibles intervals. In our case, the model for peak detection would thus be written as:

$$(x_i, y_i)|(P, B, w_i) \sim w_i P + (1 - w_i)B$$

with (x_i, y_i) the observed spectrum in 2D (i.e. m/z and time dimensions), P the mixture of peaks, B the baseline, and w_i the probability of belonging to peak or baseline. The posterior law $\boldsymbol{y}|(P, B, \boldsymbol{w})$ would then estimated by Markov Chain Monte Carlo methods (Grenn, 1995).

Barat et al. (2007a,b) proposed non parametric prior laws for P and B, respectively the Dirichlet Process Mixture (DPM; Antoniak 1974) and 2D Polya Trees (PT; Mauldin et al. 1992), and a Beta distribution for w. Their SINBAD algorithm shows great performances for peak location and quantification on gamma-ray spectra, especially for highly convoluted peaks (Rohée et al., 2015; Rohée et al., 2016). Applying such an approach to PTR-TOF-MS data would require new developments to adapt the priors and to ensure that the MCMC framework converges rapidly.

Deep learning

Deep Learning (DL) has become one of the most active fields in artificial intelligence, with high performances in a broad area of applications, especially for image classification using Convolutional Neural Networks (CNN; Rawat and Wang 2017).

Applying DL to the 2D MS data considered as images, is therefore appealing: this would eliminate the need for feature-engineering (peak detection and deconvolution). DL meth-

ods have already be applied to Imaging Mass Spectrometry (IMS; Behrmann et al. 2017) and tandem mass spectrometry (Data Independent Analysis; Tran et al. 2019; Cadow et al. 2021. In case of PTR-TOF-MS data, a pixel would correspond to the count of ions within an m/z bin and a time acquisition period. Since the duration of acquisition is different between patients, the images would have to be resized (Siu and Hung, 2012).

DL offers a fast and accurate prediction, having a more global view of the data than classical feature extraction methods. However, it presents two main limitations: the limited amount of labelled data for training and the lack of interpretability (e.g the m/z value of the discriminant metabolites). To overcome the former issue, Cadow et al. (2021) used a collection of publicly available DL models already trained for the task of natural image classification. To address the interpretability of DL models, Behrmann et al. (2017) proposed a strategy to interpret the learned model in the spectral domain, based on a sensitivity analysis between the predicted class probabilities and each spectrum input. Nevertheless, interpretability remains an open challenge for clinical applications (Ching et al., 2018).

Instead of using DL on the whole raw data for direct prediction, the learning may be restricted to the pre-processing workflow. Kantz et al. (2019) decreased the number of false positive peak detection by 90% by training CNN models on manually labelled LC-MS raw data subsets around detected peaks in the m/z and RT dimensions. In the context of exhaled breath analysis, one could think to learn the VOC origin (exhaled breath or external contamination) by training DL models on the raw data bands obtained after peak detection in the mass dimension (instead of relying on statistical tests to discriminate between exhaled and ambient phases).

Varying-coefficient models using P-splines

P-splines were used in both parts of this thesis, as they are particularly interesting flexible tools for nonlinear smooth modelling without any parametric assumption. In the pre-processing part, P-splines were used with a tensor product for 2D signal regression to model the evolution of peaks during the acquisition time. In the longitudinal analysis part, they were used within a mixed-effect modelling of the evolution of VOC concentrations during hospitalisation time for each patient. In the latter case, however, we only performed univariate time course modelling analysis. An alternative multivariate approach, to take into account the interactions between the VOCs, is provided by varying coefficient regression.

Varying-coefficient models (VCM; Hastie and Tibshirani 1993) are predictive models where coefficients are allowed to change smoothly with the value of other variables, called "effect modifiers" (e.g. time, age). Let us denote t_{ij} the time-points at which the measurements for the ith patient were recorded, y_{ij} the response, and $\boldsymbol{X}_{ij} = (x_{ij}^1, ...x_{ij}^p)$ the p

predictor values. In the case of a generalised linear model, the model is written as follow (Hoover et al., 1998):

$$g(y_{ij}) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon(t_{ij})$$

where g is a link function (e.g. *logit* for binary variable or identity for regression), $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_p(t))$ are smooth functions, and $\epsilon(t)$ is a zero-mean stochastic process. The $\beta(t)$ smooth functions would then be estimated with P-splines: $\beta(t) = \sum_{k=1}^{K} \gamma_k b_k(t)$, where $(b_1(t), ..., b_K(t))$ are B-splines function, and a difference penalisation is applied to the optimisation least squares problem, according to the P-spline theory (Marx, 2010; Li and Zhang, 2010). This model has been used in several applications including economics, spatial modelling and epidemiology, and has been generalised to the Bayesian framework (Franco-Villoria et al., 2019; Heuclin et al., 2021).

Electronic noses

The Exhalomics platform is also equipped with several electronic noses (eNose; Gardner and Bartlett (1994); Devillier et al. (2017)). eNose technologies have already been applied to clinical applications (Di Natale et al., 2014; Farraia et al., 2019). It is a portable and low cost technology, using an array of sensors that are relatively selective for different families of VOCs, and is compatible with online acquisitions (Bruderer et al., 2019). PTR-TOF-MS and eNoses are therefore two complementary technologies which are evaluated in parallel for each patient at the Hôpital Foch. In particular, coupling eNose with PTR-TOF-MS may be useful to support the development of optimised sensors, as recently explored by other teams for malaria transmissible stage prediction (Capuano et al., 2019).

VOCs identification with GCxGC MS

We have shown that the PTR-TOF-MS instrument is a method of choice for biomarker discovery, due to its fast response time, its high sensitivity (limits of detection in the pptv range), and since it can be operated readily at the point of care. However, this approach only provides information on the mass/charge ratio of the compound, which is limiting for the structural identification of the metabolite (and hence for the characterisation of its biological role). Additional MS technologies offering higher mass resolution, chromatographic separation, and fragmentation, are thus required for metabolite identification. The Exhalomics platform recently acquired a two-dimensional gas chromatography TOF mass spectrometer (GCxGC-TOF-MS; Liu and Phillips 1991; Phillips et al. 2013), a powerful tool for multidimensional analysis of complex samples with the potential to identify a greater number of VOCs.

Comprehensive two-dimensional GC (GCxGC) extends the chromatographic separation

by pairing two columns with complementary stationary phases. Therefore, compounds that would co-elute in conventional GC may be separated by the GCxGC system. The resulting data contain three dimensions: two retention times (one from each chromatographic separation) and a mass spectrum that is relatively unique to each compound. Very few open source tools for the pre-processing of such data already exist (Ramaker et al., 2017; Quiroz-Moreno et al., 2020; Wilde et al., 2020), and focus on baseline correction, denoising, peak alignment using correlation optimised warping with a reference chromatogram (Zhang et al., 2008), and identification by matching mass spectral signatures to a library of mass spectra. Peak detection and deconvolution in 2D, however, remain to be developed, and will benefit from the rich datasets currently analysed on the Exhalomics platform.

Appendix A

Characteristic of sech2 functions

$$f(x) = \frac{h}{cosh(\lambda(x-p))^2} \quad \text{with} \quad \lambda = \lambda_1 \quad \text{if} \quad x < p \quad \lambda = \lambda_2 \quad \text{if} \quad x \geq p$$

Peak width

$$\lambda_i = \frac{log(\sqrt{2}+1)}{\Delta_i} \tag{A.1}$$

with Δ_i left/right width at half maximum

Peak area

$$\begin{aligned} A &= \int_{-\infty}^{+\infty} f(x)dx \\ &= \int_{-\infty}^{p} \frac{h}{cosh(\lambda_1(x-p))^2}dx + \int_{p}^{+\infty} \frac{h}{cosh(\lambda_2(x-p))^2}dx \\ &= \int_{-\infty}^{0} \frac{h}{cosh(u)^2}\frac{1}{\lambda_1}du + \int_{0}^{+\infty} \frac{h}{cosh(u)^2}\frac{1}{\lambda_2}du \\ &= \frac{h}{\lambda_1}\left[tanh\right]_{-\infty}^{0} + \frac{h}{\lambda_2}\left[tanh\right]_{0}^{+\infty} \\ &= h(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}) \end{aligned} \tag{A.2}$$

Appendix B

Articles

Bioinformatics, 2022, 1–8
https://doi.org/10.1093/bioinformatics/btac031
Advance Access Publication Date: 19 January 2022
Original Paper

Gene expression

ptairMS: real-time processing and analysis of PTR-TOF-MS data for biomarker discovery in exhaled breath

Camille Roquencourt [1,*], Stanislas Grassin-Delyle[2,3,4] and Etienne A. Thévenot [5]

[1]Département Métrologie Instrumentation & Information (DM2I), CEA, LIST, Laboratoire Sciences des Données et de la Décision, F-91191 Gif-Sur-Yvette, France, [2]Département des maladies des voies respiratoires, Hôpital Foch, Exhalomics, Suresnes 92150, France, [3]Département de Biotechnologie de la Santé, Université Paris-Saclay, UVSQ, INSERM, Infection et inflammation, Montigny le Bretonneux 78180, France, [4]FHU SEPSIS (Saclay and Paris Seine Nord Endeavour to PerSonalize Interventions for Sepsis), Garches 92380, France and [5]Département Médicaments et Technologies pour la Santé (MTS), Université Paris-Saclay, CEA, INRAE, MetaboHUB, F-91191 Gif sur Yvette, France

*To whom correspondence should be addressed.

Associate Editor: Olga Vitek

Received on September 29, 2021; revised on December 24, 2021; editorial decision on December 27, 2021; accepted on January 16, 2022

Abstract

Motivation: Analysis of volatile organic compounds (VOCs) in exhaled breath by proton transfer reaction time-of-flight mass spectrometry (PTR-TOF-MS) is of increasing interest for real-time, non-invasive diagnosis, phenotyping and therapeutic drug monitoring in the clinics. However, there is currently a lack of methods and software tools for the processing of PTR-TOF-MS data from cohorts and suited for biomarker discovery studies.

Results: We developed a comprehensive suite of algorithms that process raw data from patient acquisitions and generate the table of feature intensities. Notably, we included an innovative two-dimensional peak deconvolution model based on penalized splines signal regression for accurate estimation of the temporal profile and feature quantification, as well as a method to specifically select the VOCs from exhaled breath. The workflow was implemented as the ptairMS software, which contains a graphical interface to facilitate cohort management and data analysis. The approach was validated on both simulated and experimental datasets, and we showed that the sensitivity and specificity of the VOC detection reached 99% and 98.4%, respectively, and that the error of quantification was below 8.1% for concentrations down to 19 ppb.

Availability and implementation: The ptairMS software is publicly available as an R package on Bioconductor (doi: 10.18129/B9.bioc.ptairMS), as well as its companion experiment package ptairData (doi: 10.18129/B9.bioc.ptairData).

Contact: camille.roquencourt@hotmail.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Volatolomics is the study of volatile organic compounds (VOCs) emitted by a biological system (Amann *et al.*, 2014), which can be found in several human matrices such as saliva, urine, skin, blood and exhaled breath. Recently, many studies have highlighted the potential of VOC analysis from exhaled breath for early diagnosis, disease phenotyping, therapeutic drug monitoring or toxicological analysis (Boots *et al.*, 2015; Bruderer *et al.*, 2019; Einoch Amor *et al.*, 2019; Pereira *et al.*, 2015; Rattray *et al.*, 2014). One of the main advantages of breath analysis is its non-invasive nature (Devillier *et al.*, 2017).

Mass spectrometry is a powerful method for the study of small volatile molecules (Rattray *et al.*, 2014). Recently, 'on-line' technologies, where the patient blows directly into the mass spectrometer, have emerged as promising approaches for the real-time analysis at the point of care (Bruderer *et al.*, 2019; Devillier *et al.*, 2017). Such strategies are of major interest for the screening and monitoring of individual patients or cohorts (Trefz *et al.*, 2013). The potential of proton transfer reaction coupled to time-of-flight mass spectrometry (PTR-TOF-MS; Blake *et al.*, 2009; Herbig *et al.*, 2009; Jordan *et al.*, 2009) for biomedicine has been shown in applications such as emphysema, liver cirrhosis, chronic kidney disease and diabetes (Cristescu *et al.*, 2011; Fernández del Río *et al.*, 2015; Obermeier *et al.*, 2017; Pleil *et al.*, 2019). PTR-TOF-MS spectrometers provide limits of detection in the parts per billion by volume (ppbv) range and rely on VOCs ionization with a transfer of proton from a reagent ion (usually H_3O^+), then subsequent detection of the

resulting ions with time-of-flight (TOF)-MS. During data acquisition, which is very fast, the instrument continuously analyzes the air flowing through a buffer tube (i.e. ambient air by default) and the patient is asked to expire a few times into the tube. Each data file (in the HDF5 open format; Koziol, 2011) contains the ion intensities stored as a numerical matrix whose dimensions are the TOF bins (which can be converted to *m/z* values) and the acquisition time.

Two processing software are currently available for PTR-TOF-MS data, the commercial Ionicon Data Analyzer (IDA) released in 2020 based on the algorithms by Müller *et al.* (2013) and the open-source PTRwid (Holzinger, 2015). These software tools allow the analysis of high-resolution, TOF-MS data with the following characteristics: (i) single (or multiple for PTRwid) file analysis, (ii) internal *m/z* calibration, (iii) untargeted peak detection and deconvolution and (iv) quantification and suggestion of elemental composition. They are particularly suited for the analysis of very large files resulting from continuous environmental monitoring. However, there are specific needs for breath research in patient cohorts which have to be covered. For instance, the simultaneous analysis of multiple samples requires that peak lists from different samples may be aligned; in addition, the parallel processing of several files would be a time-sparing capability; furthermore, a correct distinction of the signals coming from the background and the expiratory phases is needed; finally, implementing a background correction of the ambient air composition as a function of time would be an asset for accurate peak detection and quantification (Beauchamp, 2011; Filipiak *et al.*, 2012; Španěl *et al.*, 2013).

We have therefore developed a suite of algorithms for the processing and analysis of PTR-TOF-MS data for untargeted breath analysis and biomarker discovery in patient cohorts. In particular, the penalized regression on a B-spline basis (P-splines) was used for adaptive temporal modeling (Eilers and Marx, 1996), and the coefficients in both *m/z* and time dimensions were jointly estimated with a two-dimensional (2D) tensor product. This approach enables to estimate all temporal trends without any parametric hypothesis, and to precisely separate peaks in the *m/z* dimension at each time. The temporal profiles are then used to correct the external contamination, using linear ambient air baseline removal and statistical testing of mean intensity in ambient air versus exhaled breath.

The whole workflow from the raw data files up to the table of peak intensities is implemented as the ptairMS package (doi: 10.18129/B9.bioc.ptairMS) available on Bioconductor (Gentleman *et al.*, 2004). It includes specific features to facilitate routine clinical analysis (e.g. graphical user interface, quality control checks, sample metadata management, iterative inclusion of new acquisitions). In the following, we will first describe the methods used for each step of the workflow, and then present the results obtained with simulated, experimental, and clinical datasets.

2 Materials and methods

The suite of algorithms developed for the processing of PTR-TOF-MS data from exhaled breath, and implemented in the ptairMS R package, takes as input the name of the directory containing the raw files in HDF5 format, and ultimately generates the samples by variables table of peak intensities. The main steps of the workflow are summarized below and detailed in the following of Section 2. This workflow proposes innovative developments for the breathomics analysis of cohorts, including 2D processing and ambient air quantification and correction methods, which were implemented to previous literature on breath analysis.

1. Processing of each file
   a. Internal calibration of the *m/z* axis
   b. Determination of expiration limits
   c. Untargeted peak detection and quantification in exhaled breath
      - Detecting peaks on the average total ion spectrum
      - Estimating the temporal evolution for each peak
      - Quantifying
      - Ambient inhaled air correction
      - Statistical testing of intensity differences between ambient air and expiration phases
2. Alignment between samples followed by quality control
   - Aligning features between samples
   - Filtering features based on reproducibility within the whole cohort or sample classes
   - Filtering features based on the *P*-value from the test in (1.c)
3. Imputation of missing values
4. Putative annotation (including isotopes)
5. Export of the peak table and metadata
6. Peak table update when new files are included in the input directory

2.1 Processing of each file

2.1.1 Calibration

Calibration converts the TOF values recorded by the mass spectrometer into *m/z* values: $m/z = \frac{(tof-b)^2}{a}$ (Brown and Gilfrich, 1991). To estimate the parameters (a, b), the Levenberg–Marquardt algorithm is used, with couples (tof, *m/z*) of reference peaks without overlap (Cappellin *et al.*, 2010; Müller *et al.*, 2013). For exhaled breath, we suggest using the following peaks: the primary ion isotope (*m/z* 21.022), dinitrogen (*m/z* 29.013) and the acetone isotope (*m/z* 60.053). External calibration ions such as iodobenzene (*m/z* 203.943), and diiodobenzene (*m/z* 330.850) can also be used for calibration in instruments with internal permeation devices. As a drift over time is observed due to low changes of temperature, calibration is performed periodically (e.g. every minute) to update the (a, b) values. The shift is subsequently estimated for each *m/z* as a function of time by linear interpolation.

2.1.2 Expiration detection

Determination of expiration limits and background (ambient air) is a very important step for the analysis, as boundaries will be used for quantification and for the statistical test for features selection in Section 2.1.3. Classically, a raw data ion trace is used to automatically detect expiration. Herbig *et al.* (2009) propose to use acetone (*m/z* 59.049), CO_2 (*m/z* 44.997) or humidity with the water cluster isotope (*m/z* 39.033) as ion traces. We used the same method as described by Schwoebel *et al.* (2011) and Trefz *et al.* (2013), to automatically detect expiration and inhalation phases on an ion trace. In addition, we designed a specific panel from our graphical interface to the visualization (and possible manual modification) of the expiration limits (as described in Section 3.3 below).

2.1.3 Untargeted peak detection and quantification in exhaled breath

Raw data consist in a numerical matrix of TOF counts, whose dimensions are $\sim 10^5$ bins (*m/z* between 0 and 500 Da), and $\sim 10^2$ s (depending on the acquisition time). After *m/z* calibration, data are processed sequentially within bands centered at each nominal mass within an interval of ± 0.6 Da (since VOCs are of low molecular weight, <500 Da, peak *m/z* are clustered around nominal masses; Cappellin *et al.*, 2011; Müller *et al.*, 2011), and covering the full time range. The following steps are then applied: (i) peaks are detected in the mass axis on the sum spectrum, (ii) their temporal evolution is estimated by a tensor product with P-splines, (iii) statistical tests are performed to identify if VOCs come from exhaled breath or ambient air and (iv) their average intensity in expiration phases are quantified in ppb (Fig. 1).

***Peak detection on the average spectrum in 1D*:** The peak picking algorithm in the *m/z* dimension is mainly based on Müller *et al.* (2013). Due to the medium resolution of the instrument (5000 to 10 000), a parametric peak function is required for peak separation. The described estimation of the peak shape starts from the 10% envelop quantile of the normalized and filtered raw spectrum between

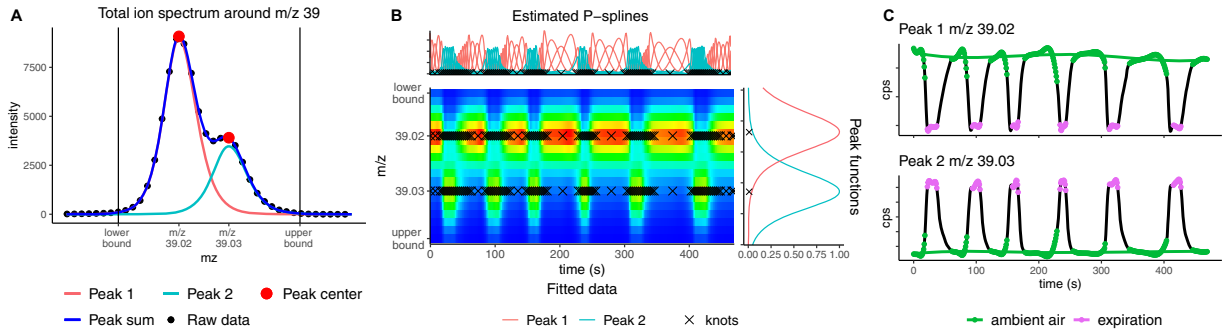


Fig. 1. Main steps of the pre-processing algorithms for a single PTR-TOF-MS raw file containing six expirations. (**A**) Peak detection in the *m/z* dimension with a parametric peak shape after baseline correction. (**B**) Two-dimensional penalized regression, with a tensor product between the mixture of peak functions from Step (A) and a P-spline basis. The penalty parameter for the time axis is estimated by the generalized cross-validation criterion. Crosses indicate knot locations (i.e. where the coefficients are estimated). The fitted splines for Peak 1 at *m/z* 39.02 (respectively, Peak 2 at *m/z* 39.03) are shown in red (respectively blue). (**C**) Estimation of the temporal evolutions by summing each modeled peak from Step (B) along the time dimension. Two unilateral *t*-tests are applied to compare expiration and ambient air intensities. If expiration values are significantly greater (respectively, lower) than ambient air, as for Peak 2 (respectively, Peak 1), the feature is considered as originating from 'expiration' (respectively, 'ambient air')

a given intensity range, and performs an iterative peak detection on the residuals to deconvolve the peaks (Holzinger, 2015; Müller *et al.*, 2013). We also included three alternative parametric functions which may be useful for TOF peak shapes, namely the asymmetric sech2, gaussian and lorentzian functions (Lange *et al.*, 2007). The best peak function is selected automatically according to the R2 criterion on the calibration peaks. To sum up, the different steps of the peak detection on the average ion spectrum around each nominal mass are (Fig. 1A):

1. Baseline removal (Ryan *et al.*, 1988)
2. Estimation of the noise threshold and autocorrelation within the 'off-peak' interval $[m-0.6, m-0.4] \cup [m+0.4, m+0.6]$ (Müller *et al.*, 2011)
3. Savitzky–Golay signal filtering by using optimal windows, followed by detection of local maxima by using the first and second derivatives (Savitzky and Golay, 1964; Vivo Truyols and Schoenmakers, 2006)
4. Peak deconvolution, by using a peak function of the mass m and depending on the parameters μ (peak center), σ (peak width) and h (peak height): $h \times peak_{(\mu,\sigma)}(m)$
5. Iterative residual analysis, which stops as soon as one of the following criteria is met: R2 $>$ R2$_{min}$ (default: 0.995), noise autocorrelation $<$ autocorMax (default: 0.3), the maximum number of iterations is reached (default: 4), the maximum number of detected peaks is reached (default: 7) (Müller *et al.*, 2013)

***Estimation of the temporal evolution with penalized signal regression using P-splines in 2D*:** To estimate the temporal evolution of each peak, we used a 2D regression approach (Marx and Eilers, 2005), which consists of a tensor product between P-splines and the previously estimated *m/z* peak functions (Fig. 1B). B-splines (basis splines) are polynomial basis functions spread all over a set of knots (de Boor, 1978; Dierckx, 1995). P-splines (penalized B-splines) are B-splines with a difference penalty applied to the coefficients to control the smoothness, and thus overfitting (Eilers and Marx, 1996). The P-spline approach is very powerful to model any profile without a priori knowledge of the data and to provide interpretable coefficients (Eilers and Marx, 2021; Wood, 2006). It has been used in many applications and theoretical works (Eilers *et al.*, 2015), such as data smoothing (Currie and Durban, 2002), Bayesian statistics (Gressani and Lambert, 2021) and machine learning with generalized additive models (Brezger and Lang, 2006; Wood, 2006). To model interactions in multiple dimensions, the tensor product provides a straightforward generalization of this basis (Sidiropoulos *et al.*, 2017). Here, we therefore used tensor product modeling to achieve a fast deconvolution of peaks in both *m/z* and time dimensions simultaneously, as described below. Raw data are processed sequentially within bands around detected peaks (the 1% quantile of the estimated mixture of peak functions is used to define the *m/z* bounds), and covering the full acquisition time. In a preliminary step, the baseline in the *m/z* dimension is estimated at each time point by linear regression between the two *m/z* boundaries and is subsequently removed, and the calibration shift estimated in Section 2.1.1 is corrected by linear interpolation. Let us then denote $f(t) = \sum_{j=1}^{K} \alpha_j s_j(t)$, and $g(m) = \sum_{i=1}^{n_{peak}} h_i peak_{\hat{\mu}_i,\hat{\sigma}_i}(m)$, the functions representing the acquisition time and the *m/z* profiles, respectively, with $peak_{\hat{\mu}_i,\hat{\sigma}_i}(m)$ being the function of peak i estimated in the previous section, and with $(s_1, \ldots, s_K)$ being cubic B-spline functions for the set of knots $(k_1, \ldots, k_K)$. The 2D model is obtained by writing each peak coefficient h_i in the B-spline basis: $f_\beta(t,m) = \sum_{i=1}^{n_{peak}} \sum_{j=1}^{K} \beta_{ij} s_j(t) \times peak_{\hat{\mu}_i,\hat{\sigma}_i}(m)$, with $\beta_{ij} = h_i \times \alpha_j$.

The β_{ij} coefficients are estimated according to the P-splines theory, by minimizing the following penalized regression, where the penalty is applied only to the time dimension:

$$\min_{\beta} \sum_{t=1}^{T} \sum_{m=1}^{M} (\mathrm{Y_{mt}} - \mathrm{f}_\beta(m,t))^2 + \lambda \sum_{i=1}^{n_{peak}} \sum_{j=3}^{K} (\Delta^2 \beta_{ij})^2 \qquad (1)$$

where $\Delta^2 \beta_{ij} = \beta_{i,j} - 2\beta_{i,j-1} + \beta_{i,j-2}$ is the second order difference, i (resp. j) represents the knots location of mass (respectively, time) axis, m (respectively, t) represents the index of mass (respectively, time) axis, and Y is the raw data matrix of dimensions $M \times T$ after baseline removal and calibration shift correction.

The choice of the knot locations and the penalty coefficient λ are very important, since too many knots may lead to over fitting, and too few knots may result in under fitting. Classically, knots are uniformly distributed over the data range in order to facilitate the interpretation of the penalty applied to the successive knot differences (Eilers and Marx, 1996). In our case, however, (i) exhaled breath phases are the main focus of our quantification and (ii) inhaled air phases are generally constant. We therefore propose to target the knot locations mainly around the expiration phases (Supplementary Fig. S1). This allows to reduce the dimension of the model, and thus the computational time, while maintaining a good fit (Supplementary Table S1). Alternatively, a uniform distribution of the knots along the time axis may be selected, in case the user has no a priori knowledge about the temporal profile of the compound. The optimal λ value is selected with grid search using the generalized cross-validation criterion (Eilers and Marx, 2010).

***Quantification*:** For each peak i, quantification (in counts per extraction) is first performed at each time point t by summing

the 2D model along the m/z dimension: $c_t^i = \sum_{m=1}^{M}\sum_{j=1}^{K}\hat{\beta}_{ij}\times s_j(t)\times peak_{\hat{\mu}_i,\hat{\sigma}_i}(m)$. This results in a temporal series $(c_1^i,\ldots,c_T^i)$, with T being the acquisition duration (Fig. 1C).

These amounts of VOC i at each time point are then normalized and converted to absolute quantities Q_t^i as follows. First, since the intensities provided by the instrument at each time point are in fact the sum of a fixed number of internal acquisitions, the c_t^i are normalized (as counts of ions per second; cps) by dividing by the integrated internal time period and by multiplying by the single ion pulse voltage (Müller *et al.*, 2014). To obtain the concentration, the latter values are then normalized by the reagent ion (H_3O^+) intensities, the reaction rate coefficient between the VOC and H_3O^+, and the residence time of the primary ions in the drift tube (normalized cps, ncps; Cappellin *et al.*, 2012). The final normalization by the density of the air in the reaction chamber gives the absolute concentration of the VOC, expressed in part per billion (ppb).

The absolute concentration of VOC i in exhaled breath is obtained by averaging all Q_t^i corresponding to the time points t within the expiration phases.

***Ambient inhaled air correction*:** To correct the ambient inhaled air level in exhaled breath, we propose to subtract the ambient air baseline of the temporal profile of each detected VOC, using a polynomial fit (default degree 3) computed on the ambient air time points. This method is based on the concept of 'alveolar gradient', introduced by Phillips (1997). Note that the subtraction step may be omitted in particular cases, as detailed in the discussion (a specific parameter is included in the software tool).

***Statistical testing of intensity differences between expiration and ambient air phases*:** Two unilateral statistical tests (t-tests) are used to compare intensities within and between expirations (i.e. exhaled breath and ambient air). Compounds with intensities that are significantly higher (respectively, lower) within expiration phases are considered to be from exhaled breath (respectively, from ambient air). If none of the tests is significant, the compound is labeled as 'constant' (e.g. in the case of internal ions generated by the instrument).

2.2 Alignment

Once the peak lists have been extracted from each file, alignment of the features between the samples is performed by using a kernel Gaussian density (Delabrière *et al.*, 2017; Smith *et al.*, 2006). Two quality control steps may then be applied to select features (i) with a high reproducibility between samples (alternatively between classes of samples), and/or (ii) labeled as 'exhaled breath' in the majority of samples (by thresholding the P-value of the statistical tests described above).

2.3 Imputation

Imputation of missing values is performed by re-rerunning the peak detection algorithm on the raw data with updated constraints in the m/z dimension, namely without any minimum intensity threshold and with a restricted m/z width for the peak center.

2.4 Annotation and isotope detection

Putative annotations are computed by matching the measured ion masses to an internal table extracted from the Human Breathomics Database (Kuo *et al.*, 2020). Isotope annotations are suggested on the basis of three criteria: m/z difference value, correlation of the temporal profiles within the sample, and correlation of the intensities between the samples.

2.5 Software implementation

All algorithms were written in R (R Core Team, 2021), and implemented as the ptairMS package (https://doi.org/10.18129/B9.bioc.ptairMS), freely available on the Bioconductor platform (Gentleman *et al.*, 2004). The companion ptairData experiment package (https://doi.org/10.18129/B9.bioc.ptairData), also available on Bioconductor, contains the raw files from two datasets from exhaled breath and bacteria culture head space, respectively, as well as the simulated raw data file described in the following Section 3.

The main ptairMS methods are described in Supplementary Figure S2. Briefly, a ptrSet object is built by providing the name of the directory containing the HDF5 raw files. This object is then completed at each step of the processing. In addition, the ptrSet may be updated by adding new raw files to the directory, or by providing new sample metadata. The ptairMS output contains the table of peak intensities as well as the sample and variable metadata, which can be exported as three tabular files, or as a single ExpressionSet object, for subsequent statistical analysis.

3 Results

We developed a suite of algorithms for the preprocessing of PTR-TOF-MS data files and the untargeted analysis of exhaled breath from cohorts. Our workflow consists of the following main modules: peak detection, expiratory phases detection, temporal estimation, VOCs quantification and alignment between samples (Supplementary Fig. S2). It has been implemented in R as the ptairMS package, which is freely available on the Bioconductor repository. The package includes a Shiny graphical interface to facilitate data management and analysis by the end-user.

3.1 Quantification and untargeted VOCs detection in a standardized gas mixture

The quality of VOC detection and absolute quantification by ptairMS was first assessed with the analysis of a reference gas containing a mixture of VOCs in known amounts: 14 compounds with 8 distinct masses and 18 isotopes (TO-14 standard gas mixture, Restek; see the detailed list of expected molecules in the Supplementary Table S3). Ten dilutions of the gas mixture were measured in six replicates, with or without applying an activated charcoal filter (Supelpure HC hydrocarbon trap, Sigma-Aldrich, Saint-Quentin-Fallavier, France) on the ambient air input (three replicates each). During each acquisition, the aspiration of the reference gas was switched on and off three times to mimic 'expiration' profiles. Sample analysis was performed with a PTR-Qi-TOF (Ionicon, Innsbrück, Austria).

The 60 raw files were pre-processed by ptairMS in less than 15 min (on a quad-core laptop). A total of 314 (respectively, 180) compounds were detected in the absence (respectively, presence) of the charcoal filter. In particular, 45 compounds were selected after sample alignment in at least 90% of one dilution factor, and in the simulated 'expiration' phases of at least 90% of all samples (Fig. 2A), according to the statistical test implemented in ptairMS to compare intensities between simulated expiration and ambient air phases (see Section 2).

Importantly, all the expected compounds were detected, as well as their isotopes, with an m/z error inferior to 20 ppm, and an average coefficient of linearity R2 with the concentration factor of 0.999. The 19 additional detected features most likely correspond to

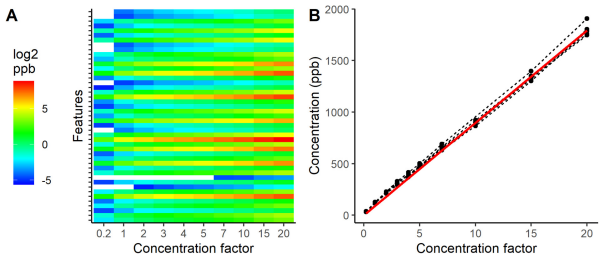


Fig. 2. ptairMS analysis of a reference VOC mixture. (**A**) Heatmap of the log2 concentrations in ppb of the 45 selected VOCs before the imputation step. (**B**) The sum of the 45 compounds concentrations for each replicate (dashed line) as a function of the concentration factor. The expected total concentration is shown as a red line

Table 1. Mean absolute percentage error (MAPE) and coefficient of variation (CV) between replicates of the ptairMS processed data from the reference gas mixture acquisitions

Expected ppb per compound	MAPE (%)	CV (%)
[1.3; 13]	47.9	4.3
[19; 32]	8.1	3.4
[44; 128]	2.5	2.8

fragments from these VOCs, since some are below the expected concentration (Supplementary Fig. S3). To evaluate the quantification, we computed the difference between the sum of the 45 compound concentrations and the expected concentration, which was less than 8.1% for the concentrations above 19 ppb (Table 1 and Fig. 2B). The coefficient of variation (CV) between replicates was <5% (Table 1), even in the absence of charcoal filter, which demonstrates that the ambient air intensity is well subtracted from the exhaled breath signal in ptairMS.

3.2 VOCs temporal profile classification and comparison to the state of the art on simulated data

The performance of the present and previously described software (Holzinger, 2015; Müller *et al.*, 2013) were compared using simulated data from PTR-TOF-MS exhaled breath analysis. First, temporal evolutions were extracted from a large in-house database of patient acquisitions (>10 000 expiration and ambient air profiles), after normalization and Savitzky–Golay smoothing. Second, peak clusters were generated around nominal masses 21 to 400, with an asymmetric sech2 peak shape distribution. Peaks parameters were randomly selected for each nominal mass: i.e. the asymmetry coefficient the peak width, the number of overlapping peaks (1 to 3), the peak proximity, the intensity of the highest peak, the ratio of neighboring peaks, and the class of temporal profile ('expiration', 'ambient air' or 'constant'). The exact *m/z* value of the first peak was selected from the formula library $C_xH_yO_zN_t$ used by PTRwid (Holzinger, 2015). Finally, background noise was added by using a Poisson stochastic process (Gundlach-Graham *et al.*, 2018), with a Gaussian distribution to model the single ion Pulse-Height. The random drawing of each parameters is detailed in the Supplementary Table S2, and the code used for the simulation, as well as a representative simulated data file in the HDF5 format, are included in the ptairData R/Bioconductor companion package.

Ten simulated files, containing a total of 7028 peaks, were processed with ptairMS (version 0.1), PTRwid (version 002 IDL) and IDA (version beta 0.9.4.8). ptairMS, which is the only software allowing simultaneous multiple file processing, enabled to process the 10 files in <10 min. Mass calibration was performed using the peaks at *m/z* 21.022, 203.943 and 330.84 for the three software, intentionally simulated without overlap at the exact masses. The calibration stability period was set to the acquisition duration, since no calibration shift was added. To ensure a good estimation of the peak shape for the three software, we simulated more single peaks in the intensity range set for the calculation of the peak shape. Finally, the 'sensitivity' parameter for IDA peak detection was decreased to 25%, in order to limit the number of false positives. The other parameters from each software tool were kept to default values.

Results of the comparison are shown in Table 2. The best precision of peak detection and mass accuracy were obtained with ptairMS, and the peak detection recall was slightly lower than IDA (98.40% versus 98.49%). The mass accuracy depends only on peak detection, since no mass deviation was included in the simulation. Of note, the reported mass accuracy for PTRwid was computed before calibration: indeed, the masses from the simulated multiple peaks may not match with the internal chemical formula library used by PTRwid for calibration, especially for masses >300 Da (the mass accuracy for PTRwid after calibration was 20 ppm).

Quantification was further evaluated on the peaks which were well detected by all software. The mean absolute percentage error between the estimated temporal evolution and the input of the simulation was 4.96% for ptairMS and 14.65% (respectively, 5.38%) for PTRwid (respectively, IDA). Finally, we compared the ability to discriminate the compounds from exhaled breath and ambient air, based on two unilateral *t*-tests comparing the intensities in the two acquisition phases (see Section 2.1.3). ptairMS was shown to detect the expiration profiles with the highest sensitivity and specificity, with a global accuracy of 99% (compared to 87% and 95% for PTRwid and IDA; Table 2). As illustrated in Figure 3 on two simulated peaks with close *m/z* values, an exogenous VOC (i.e. with a constant profile) at *m/z* 82.034 was erroneously classified as 'expiration' by PTRwid and IDA but not by ptairMS, as a result of a less precise temporal estimation of the two first software tools. Altogether, these results demonstrate that ptairMS is well suited for biomarker research by breath analysis.

Table 2. Comparison of peak detection and quantification by ptairMS, PTRwid and IDA on 10 simulated files (7028 peaks)

Software	ptairMS	PTRwid	IDA
Mass accuracy (ppm)	**3**	12[a]	5
Peak detection precision (%)	**99.99**	98.87	97.30
Peak detection recall (%)	98.40	87.19	**98.49**
MAPE (%)	**4.96**	14.65	5.38
Expiration sensitivity (%)	**98.53**	91.45	94.52
Expiration specificity (%)	**99.01**	86.31	97.03
Global accuracy (%)	**99.12**	86.73	95.31

Note: The precision (respectively, recall) of peak detection is the proportion of detected peaks which correspond to actual simulated peaks (respectively, the proportion of actual simulated peaks which were detected by the software tools). The mean absolute percentage error (MAPE) is used to assess the quality of the temporal profile estimation. Expiration sensitivity, specificity and accuracy refer to the classification of VOC origin as exhaled breath (vs. ambient air). For each metric, the best performance is shown in bold.

[a]The reported mass accuracy for PTRwid was computed before calibration as explained in the text.

3.3 Application to real datasets

The ptairMS software has been designed for biomarker discovery in large clinical cohorts. First, it is fast (<1 min for a 3–5 min acquisition), and files can be processed with parallel computing and in a batch mode. Second, studies can be readily incremented with new files (e.g. if new patients are included): only the processing of these new files and the final alignment between samples are performed to update the peak table of the whole cohort. Third, the whole workflow can be run interactively through a graphical user interface, which provides visualizations (expiration phases, peaks in the raw data, peak table, individual VOCs), quality controls (calibration, resolution, peak shape and evolution of the reagent ions with time), and exploratory data analysis (Fig. 4). A detailed documentation including several use cases is included in the package.

ptairMS is already used in routine in the clinic to process the acquisitions from freely breathing patients in some breath research centers using PTR-Qi-TOF MS. Files from a distinct PTR-TOF 8000 instrument (Ionicon) (Trefz *et al.*, 2013; Vita *et al.*, 2015) were also successfully processed with ptairMS (Supplementary Figs S4 and S5). These results highlight the ability of the algorithms to adapt to various resolutions, time bin periods, peak shapes and temporal profiles.

4 Discussion

We have developed an innovative workflow for the fast processing of PTR-TOF-MS data from exhaled breath. The suite of algorithms includes untargeted peak detection and deconvolution in the mass dimension, expiration phases detection, estimation of the temporal evolution of the peak intensity during the acquisition and quantification. Compared to the two existing software, it enables for the first time to

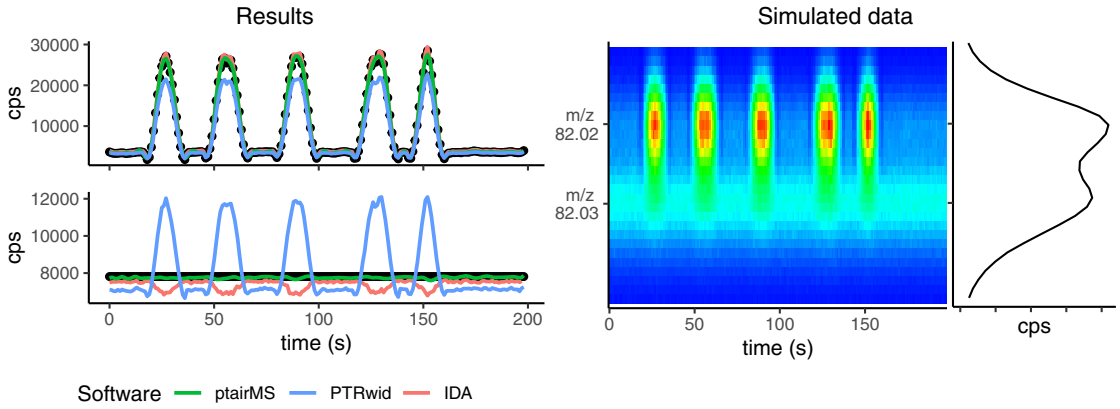


Fig. 3. Estimation of the temporal profile by ptairMS, compared to the PTRwid and IDA software on simulated data. Right: raw simulated data of two overlapping peaks (as shown in 2D), and the corresponding total mass spectrum. In this particular example, the VOC at *m/z* 82.02 (respectively, *m/z* 82.03) was simulated by using an 'expiration' (respectively, a 'constant') temporal profile. Left: temporal profiles estimated by the three software (solid colored lines), compared to the simulated profile (ground truth shown as black dots), for the two peaks (top: *m/z* 82.02 and bottom: *m/z* 82.03). As observed with the peak at *m/z* 82.03, the temporal estimations from PTRwid and IDA lead to an erroneous classification of the VOC as expiration or ambient air

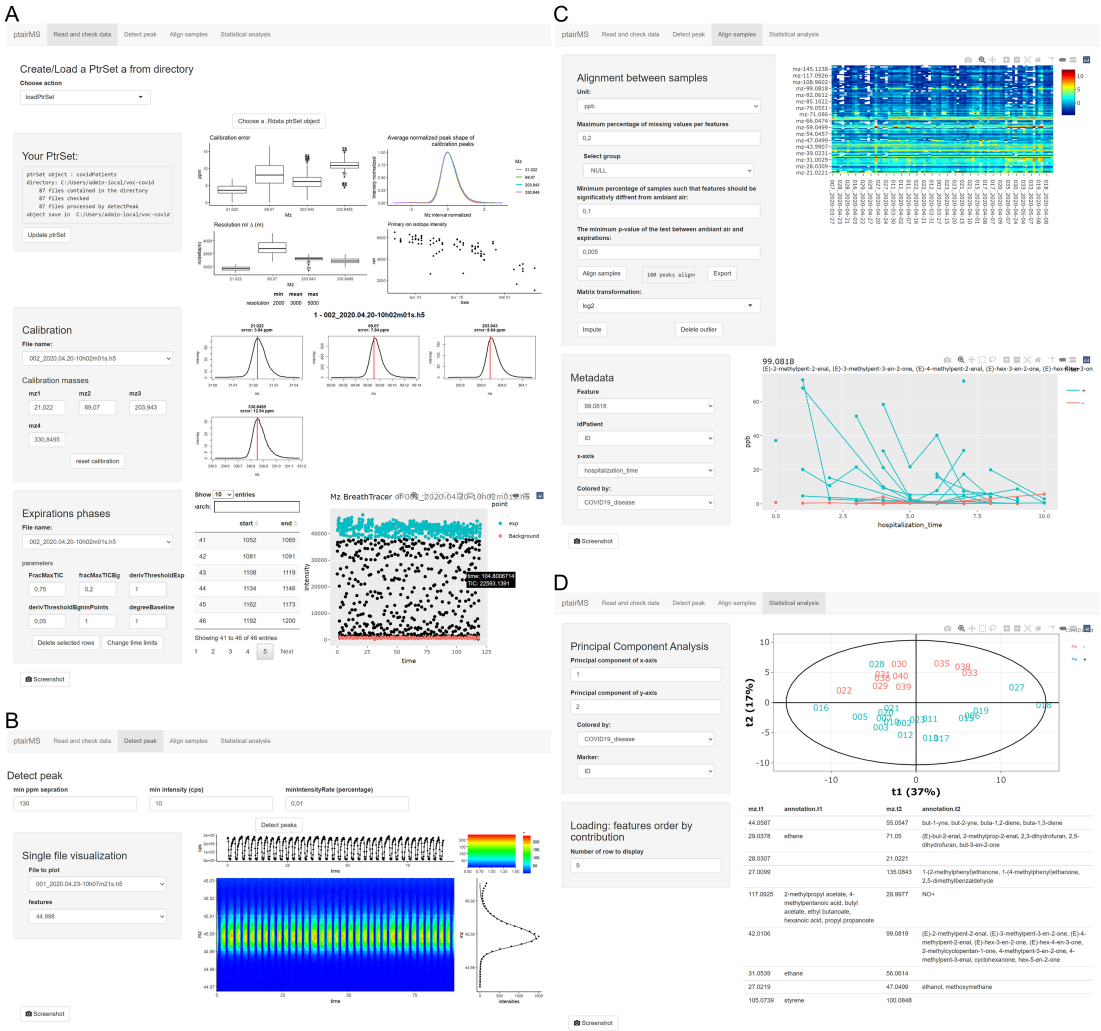


Fig. 4. The ptairMS graphical user interface to monitor the processing and exploratory analysis of cohorts, as illustrated with the COVID-19 dataset (Grassin-Delyle *et al.*, 2021). (**A**) The 'Read and check data' tab enables to open the data (either from a new study or to update an existing one), and to perform the calibration and the detection of expirations, and provides optimal parameter values for the peak shape and the resolution. (**B**) The 'Detect peak' tab provides single file visualizations of the raw data, of the detected peaks, and of the temporal profiles. (**C**) The 'Align samples' tab displays the final peak table as well as the individual features colored according to the sample metadata. (**D**) The 'Statistical Analysis' tab displays the score plot from the Principal Component Analysis of the peak table [only the first time point of each patient is shown here, as in Grassin-Delyle *et al.* (2021)], and the list of features with their putative annotations, in decreasing order of loading values

conduct the analysis of clinical cohorts, with parallel file processing, incremental addition of new patient files, quality control of the acquisitions along clinical trials, alignment between the samples, and final statistical tests to discard exogenous VOCs. The full workflow was implemented in the R package ptairMS which is publicly available on the Bioconductor repository and includes a detailed tutorial. Raw files from two experimental datasets, as well as one simulated file, are provided in the companion ptairData package. The public availability of all data and source code will therefore be of high value for the reproducibility of the analyzes, and the benchmark of software tools (Wilkinson *et al.*, 2016).

The quality of the untargeted peak detection and absolute quantification was assessed by using a standardized gas mixture: all compounds were detected by ptairMS with an *m/z* precision lower than 20 ppm, an intensity error below 8.1% (for compounds with concentrations >19 ppb), an average R2 coefficient with the concentration factor of 0.999, and a CV <5%, thus demonstrating the performance of the detection and quantification. However, it is important to note that the standardized gas used does not reflect breath matrices. In practice, humidity saturation of exhaled breath biases the VOC quantification in PTR-MS instruments, with divergent behavior for different substance classes (Trefz *et al.*, 2018). This effect also impacts the proposed correction of the ambient air level (which consists in subtracting the ambient air baseline from the temporal profile estimated for each VOC). Since the exhaled breath and ambient air have different concentrations of humidity, O_2, and CO_2, the direct subtraction should not therefore be considered as an absolute quantification, but rather as a relative concentration, which can be used to compare patients. To further compute accurate concentration differences between inspiratory and expiratory phases, adequate humidity-adapted calibrations are required (Trefz *et al.*, 2018).

Since the estimation of the temporal profiles is a key aspect of breath analysis, we have developed a 2D model based on P-spline regression. Compared to the existing software which are well suited for single-file, large data from environmental monitoring, we demonstrate that ptairMS is very convenient for breath analysis, achieving highest sensitivity and accurate quantification. It should be noted that the temporal estimation of the peak intensities relies on the *m/z* values previously computed on the total ion spectrum (i.e. these *m/z* values are not re-evaluated at each time point) which allows a fast computation. While alternative approaches may be considered for the combined estimation of location and intensity of the peaks in 2D (such as Bayesian methods or non-linear optimization; Barat *et al.*, 2007; Binette *et al.*, 2020; He *et al.*, 2014), the ptairMS algorithms already provides precise *m/z* and intensity estimations, in a computation time (<1 min) which is compatible with the real-time patient analysis.

The classification of the VOC origin between exhaled breath and ambient air was shown to be improved with ptairMS (due to the 2D modeling), with an accuracy up to 99%. The control of external factors such as the ambient air (Trefz *et al.*, 2013), but also the dioxygen concentration (Trefz *et al.*, 2019), the patient medication, or specific diets, is of critical importance in breath analysis (Hanna *et al.*, 2019). ptairMS therefore checks the sample reproducibility after alignment to avoid some of these unwanted variations. In all cases, attention should be paid during the design of the study to the matching of patients and sampling conditions between the groups of interest.

Importantly, ptairMS automatically suggests optimal values for the parameters, such as the resolution and the peak shape (as evaluated on the calibration peaks), but also the location of spline knots (at higher densities within the expiration phases) and the penalization for the 2D regression (based on generalized cross-validation). This enables to adapt the processing to specific instruments (e.g. with distinct resolutions) but also to various biological matrices (e.g. with different time profiles). As an example, ptairMS was used to process files from both PTR-TOF 8000 and PTR-Qi-TOF instruments (Ionicon Analytik). Files from other vendors (e.g. Tofwerk) should be processed accordingly, since they are in the same open source HDF5 format, which is a data storage format of choice within the MS community (Askenazi *et al.*, 2017). Beyond exhaled breath, ptairMS was successfully applied to atmospheric air data (hospital room and corridor air), headspace analysis from mycobacteria (see the package tutorial) and truffles (Vita *et al.*, 2015; Supplementary Fig. S5).

A graphical interface was developed to facilitate data analysis and result interpretation by experimenters (e.g. clinicians). It covers the processing of raw data up to the exploratory data analysis of the cohort, with interactive tables and graphics. Since clinical studies may last several months, or even years, the interface includes a dedicated panel for the real-time control of instrument parameters to avoid unwanted effects resulting from drift in temperature, pressure, or variations in the amount of reagent ion. Incremental addition of new patient files is also possible without the need to reprocess all of the previous acquisitions. New features in future implementations will include visualizations (such as the superposition of multiple temporal profiles for several patients), and statistical testing of clinical metadata for each detected VOC. Finally, a putative annotation of the compounds and their isotopes based on the *m/z* values is provided to facilitate interpretation. To achieve higher confidence levels of 2 or 1 for the most interesting VOCs, complementary experiments with hyphenated techniques such as GC-MS are required (Ibrahim *et al.*, 2019; Nardi-Agmon *et al.*, 2016; Wilde *et al.*, 2019).

Recently, ptairMS was successfully applied to intubated, mechanically ventilated patients, and enabled to discover a biomarker signature of four VOCs for the diagnosis of coronavirus disease-19 infection (Grassin-Delyle *et al.*, 2021). In addition, it is routinely used for clinical trials in centers performing exhaled breath research, not only for online patient analysis, but also for the off-line analysis of breath collected in sampling bags, allowing the analysis of samples from multisite patients.

Altogether, these results demonstrate the value of the ptairMS software as a key resource in breathomics for real-time analysis at the point of care and in biomarker discovery studies, with a high clinical potential for the phenotyping of health and disease, therapeutic drug monitoring, toxicological studies and precision medicine (Fernández del Río *et al.*, 2015; Ibrahim *et al.*, 2019; Jung *et al.*, 2021; Löser *et al.*, 2020; Zhou *et al.*, 2017).

Acknowledgements

We thank all the staff from the Exhalomics platform (Foch Hospital, Suresnes) for providing PTR-MS data. We thank Pierrick Roger and Paul Zheng for their help with the volatolomics database, the raw data format and the package design. We thank Dr Phillip Trefz for kindly providing acquisition files for the PTR-TOF-MS 8000 instrument. We also thanks Dr Vincent Le Moine and Pr Jean-Louis Herrmann for providing bacteria cultures.

Funding

The study was funded by the Agence Nationale de la Recherche (SoftwAiR project, ANR-18-CE45-0017).

Conflict of Interest: none declared.

References

Amann,A. *et al.* (2014) The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *J. Breath Res.*, **8**, 034001.

Askenazi,M. *et al.* (2017) The arc of Mass Spectrometry Exchange Formats is long, but it bends toward HDF5: plain HDF5 as a mass spectrometry exchange format. *Mass Spectrom. Rev.*, **36**, 668–673.

Barat,E. *et al.* (2007). A nonparametric bayesian approach for PET reconstruction. In: *2007 IEEE Nuclear Science Symposium Conference Record*. IEEE, Honolulu, HI, USA, pp. 4155–4162.

Beauchamp,J. (2011) Inhaled today, not gone tomorrow: pharmacokinetics and environmental exposure of volatiles in exhaled breath. *J. Breath Res.*, **5**, 037103.

Binette,O. *et al.* (2020) *Bayesian Closed Surface Fitting Through Tensor Products. J. Mach. Learn. Res.*, 26.

Blake,R.S. *et al.* (2009) Proton-transfer reaction mass spectrometry. *Chem. Rev.*, **109**, 861–896.

Boots,A.W. *et al.* (2015) Exhaled molecular fingerprinting in diagnosis and monitoring: validating volatile promises. *Trends Mol. Med.*, **21**, 633–644.
Brezger,A. and Lang,S. (2006) Generalized structured additive regression based on Bayesian P-splines. *Comput. Stat. Data Anal.*, **50**, 967–991.
Brown,R. and Gilfrich,N. (1991) Design and performance of a matrix-assisted laser desorption time-of-flight mass spectrometer utilizing a pulsed nitrogen laser. *Anal. Chim. Acta*, **248**, 541–552.
Bruderer,T. *et al.* (2019) On-line analysis of exhaled breath: focus review. *Chem. Rev.*, **119**, 10803–10828.
Cappellin,L. *et al.* (2010) Improved mass accuracy in PTR-TOF-MS: another step towards better compound identification in PTR-MS. *Int. J. Mass Spectrom.*, **290**, 60–63.
Cappellin,L. *et al.* (2011) On data analysis in PTR-TOF-MS: from raw spectra to data mining. *Sens. Actuators B Chem.*, **155**, 183–190.
Cappellin,L. *et al.* (2012) On quantitative determination of volatile organic compound concentrations using proton transfer reaction time-of-flight mass spectrometry. *Environ. Sci. Technol.*, **46**, 2283–2290.
Cristescu,S.M. *et al.* (2011) Screening for emphysema via exhaled volatile organic compounds. *J. Breath Res.*, **5**, 046009.
Currie,I.D. and Durban,M. (2002) Flexible smoothing with P-splines: a unified approach. *Stat. Modelling*, **2**, 333–349.
de Boor,C. (1978) *A Practical Guide to Splines*. Applied Mathematical Sciences, Springer, New York.
Delabrière,A. *et al.* (2017) proFIA: a data preprocessing workflow for flow injection analysis coupled to high-resolution mass spectrometry. *Bioinformatics*, **33**, 3767–3775.
Devillier,P. *et al.* (2017) Metabolomics in the diagnosis and pharmacotherapy of lung diseases. *Curr. Pharm. Des.*, **23**,
Dierckx,P. (1995) *Curve and Surface Fitting with Splines*. Monographs on numerical analysis p. 5.
Eilers,P. and Marx,B. (2021). *Practical Smoothing: The Joys of P-Splines*. Cambridge University Press, Cambridge.
Eilers,P.H.C. and Marx,B.D. (1996) Flexible smoothing with B-splines and penalties. *Stat. Sci.*, **11**, 89–121.
Eilers,P.H.C. and Marx,B.D. (2010) Splines, knots, and penalties. *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 637–653.
Eilers,P.H.C. *et al.* (2015) *Twenty Years of P-Splines*. SORT (Statistics and Operations Research Transactions). 39. 149–186.
Einoch Amor,R. *et al.* (2019) Breath analysis of cancer in the present and the future. *Eur. Respir. Rev.*, **28**, 190002.
Fernández del Río,R. *et al.* (2015) Volatile biomarkers in breath associated with liver cirrhosis—comparisons of pre- and post-liver transplant breath samples. *EBioMedicine*, **2**, 1243–1250.
Filipiak,W. *et al.* (2012) Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. *J Breath Res*, **6**, 036008.
Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
Grassin-Delyle,S. *et al.* (2021) Metabolomics of exhaled breath in critically ill COVID-19 patients: a pilot study. *EBioMedicine*, **63**, 103154.
Gressani,O. and Lambert,P. (2021) Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Comput. Stat. Data Anal.*, **154**, 107088.
Gundlach-Graham,A. *et al.* (2018) Monte Carlo simulation of low-count signals in time-of-flight mass spectrometry and its application to single-particle detection. *Anal. Chem.*, **90**, 11847–11855.
Hanna,G.B. *et al.* (2019) Accuracy and methodologic challenges of volatile organic compound-based exhaled breath tests for cancer diagnosis: a systematic review and meta-analysis. *JAMA Oncol.*, **5**, e182815.
He,X. *et al.* (2014) A spline filter for multidimensional nonlinear state estimation. *Signal Process.*, **102**, 282–295.
Herbig,J. *et al.* (2009) On-line breath analysis with PTR-TOF. *J. Breath Res.*, **3**, 027004.
Holzinger,R. (2015) PTRwid: a new widget tool for processing PTR-TOF-MS data. *Atmos. Meas. Tech.*, **8**, 3903–3922.
Ibrahim,W. *et al.* (2019) Assessment of breath volatile organic compounds in acute cardiorespiratory breathlessness: a protocol describing a prospective real-world observational study. *BMJ Open*, **9**, e025486.
Jordan,A. *et al.* (2009) A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *Int. J. Mass Spectrom.*, **286**, 122–128.
Jung,Y.J. *et al.* (2021) Advanced diagnostic technology of volatile organic compounds real time analysis from exhaled breath of gastric cancer patients using proton-transfer-reaction time-of-flight mass spectrometry. *Front. Oncol.*, **11**, 560591.
Koziol,Q. (2011). HDF5. In: Padua D. (ed.) *Encyclopedia of Parallel Computing*. Springer US, Boston, MA, pp. 827–833.
Kuo,T.-C. *et al.* (2020) Human breathomics database. *Database (Oxford)*, **2020**, baz139.
Lange,E. *et al.* (2007) A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics*, **23**, i273–i281.
Löser,B. *et al.* (2020) Changes of exhaled volatile organic compounds in postoperative patients undergoing analgesic treatment: a prospective observational study. *Metabolites*, **10**, 321.
Marx,B.D. and Eilers,P.H. (2005) Multidimensional penalized signal regression. *Technometrics*, **47**, 13–22.
Müller,M. *et al.* (2011) Enhanced spectral analysis of C-TOF aerosol mass spectrometer data: iterative residual analysis and cumulative peak fitting. *Int. J. Mass Spectrom.*, **306**, 1–8.
Müller,M. *et al.* (2013) A new software tool for the analysis of high resolution PTR-TOF mass spectra. *Chemom. Intell. Lab. Syst.*, **127**, 158–165.
Müller,M. *et al.* (2014) Detector aging induced mass discrimination and non-linearity effects in PTR-TOF-MS. *Int. J. Mass Spectrom.*, **365–366**, 93–97.
Nardi-Agmon,I. *et al.* (2016) Exhaled breath analysis for monitoring response to treatment in advanced lung cancer. *J. Thorac. Oncol.*, **11**, 827–837.
Obermeier,J. *et al.* (2017) Exhaled volatile substances mirror clinical conditions in pediatric chronic kidney disease. *PLoS One*, **12**, e0178745.
Pereira,J. *et al.* (2015) Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview. *Metabolites*, **5**, 3–55.
Phillips,M. (1997) Method for the collection and assay of volatile organic compounds in breath. *Anal. Biochem.*, **247**, 272–278.
Pleil,J.D. *et al.* (2019) Advances in proton transfer reaction mass spectrometry (PTR-MS): applications in exhaled breath analysis, food science, and atmospheric chemistry. *J. Breath Res.*, **13**, 039002.
R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *URL* https://www.R-project.org/.
Rattray,N.J. *et al.* (2014) Taking your breath away: metabolomics breathes life in to personalized medicine. *Trends Biotechnol.*, **32**, 538–548.
Ryan,C.G. *et al.* (1988) SNIP, A Statistics Sensitive Background Treatment for the Quantitative Analysis of the Pixe Spectra in Geoscience Application. *Nucl. Instrum. Methods. Phys. Res. B.*, **34**, 396–402.
Savitzky,A. and Golay,M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627–1639.
Schwoebel,H. *et al.* (2011) Phase-resolved real-time breath analysis during exercise by means of smart processing of PTR-MS data. *Anal. Bioanal. Chem.*, **401**, 2079–2091.
Sidiropoulos,N.D. *et al.* (2017) Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.*, **65**, 3551–3582.
Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.
Španěl,P. *et al.* (2013) A quantitative study of the influence of inhaled compounds on their concentrations in exhaled breath. *J. Breath Res.*, **7**, 017106.
Trefz,P. *et al.* (2013) Continuous real time breath gas monitoring in the clinical environment by proton-transfer-reaction-time-of-flight-mass spectrometry. *Anal. Chem.*, **85**, 10321–10329.
Trefz,P. *et al.* (2018) Effects of humidity, CO_2 and O_2 on real-time quantitation of breath biomarkers by means of PTR-ToF-MS. *J. Breath Res.*, **12**, 026016.
Trefz,P. *et al.* (2019) Effects of elevated oxygen levels on VOC analysis by means of PTR-ToF-MS. *J. Breath Res.*, **13**, 046004.
Vita,F. *et al.* (2015) Volatile organic compounds in truffle (Tuber magnatum Pico): comparison of samples from different regions of Italy and from different seasons. *Sci. Rep.*, **5**, 12629.
Vivo Truyols,G. and Schoenmakers,P.J. (2006) Automatic selection of optimal savitzky golay smoothing. *Anal. Chem.*, **78**, 4598–4608.
Wilde,M.J. *et al.* (2019) Breath analysis by two-dimensional gas chromatography with dual flame ionisation and mass spectrometric detection—method optimisation and integration within a large-scale clinical study. *J. Chromatogr. A*, **1594**, 160–172.
Wilkinson,M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
Wood,S.N. (2006) *Generalized Additive Models: an introduction with R. Chapman & Hall/CRC, Boca Raton, FL*. p. 397.
Zhou,W. *et al.* (2017) Exhaled breath online measurement for cervical cancer patients and healthy subjects by proton transfer reaction mass spectrometry. *Anal. Bioanal. Chem.*, **409**, 5603–5612.

EBioMedicine 63 (2021) 103154

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

Metabolomics of exhaled breath in critically ill COVID-19 patients: A pilot study

Stanislas Grassin-Delyle, Ph.D.[a,b,f,#,*], Camille Roquencourt, M.S.[c,#], Pierre Moine, M.D.[b,e], Gabriel Saffroy[e], Stanislas Carn[e], Nicholas Heming, M.D.[b,e], Jérôme Fleuriet, Ph.D.[e], Hélène Salvator, M.D.[a,f], Emmanuel Naline, Ph.D.[a,f], Louis-Jean Couderc, M.D.[a,f], Philippe Devillier, M.D.[a,f], Etienne A. Thévenot, Ph.D.[d,f], Djillali Annane, M.D.[b,e,f], for the Garches COVID-19 Collaborative Group RECORDS Collaborators and Exhalomics® Collaborators

[a] *Hôpital Foch, Exhalomics®, Département des maladies des voies respiratoires, Suresnes, France (S.G.D., H.S., E.N., L-J.C., P.D.)*
[b] *Université Paris-Saclay, UVSQ, INSERM, Infection et inflammation, Montigny le Bretonneux, France (S.G.D., P.M., N.H., D.A.)*
[c] *CEA, LIST, Laboratoire Sciences des Données et de la Décision, Gif-sur-Yvette, France (C.R.)*
[d] *Département Médicaments et Technologies pour la Santé (DMTS), Université Paris-Saclay, CEA, INRAE, MetaboHUB, Gif-sur-Yvette, France (E.T.)*
[e] *Intensive Care Unit, Raymond Poincaré Hospital, Assistance Publique-Hôpitaux de Paris, Garches, France (P.M., G.S., S.C., N.H., J.F., D.A.)*
[f] *FHU SEPSIS (Saclay and Paris Seine Nord Endeavour to PerSonalize Interventions for Sepsis) (S.G.D., H.S., E.N., L-J.C., P.D., E.T., D.A.).*

ARTICLE INFO

Article History:
Received 14 September 2020
Revised 6 November 2020
Accepted 17 November 2020
Available online 4 December 2020

Keywords:
COVID-19
Intensive care
Mechanical ventilation
Breath analysis
Metabolomics

ABSTRACT

Background: Early diagnosis of coronavirus disease 2019 (COVID-19) is of the utmost importance but remains challenging. The objective of the current study was to characterize exhaled breath from mechanically ventilated adults with COVID-19.

Methods: In this prospective observational study, we used real-time, online, proton transfer reaction time-of-flight mass spectrometry to perform a metabolomic analysis of expired air from adults undergoing invasive mechanical ventilation in the intensive care unit due to severe COVID-19 or non-COVID-19 acute respiratory distress syndrome (ARDS).

Findings: Between March 25th and June 25th, 2020, we included 40 patients with ARDS, of whom 28 had proven COVID-19. In a multivariate analysis, we identified a characteristic breathprint for COVID-19. We could differentiate between COVID-19 and non-COVID-19 ARDS with accuracy of 93% (sensitivity: 90%, specificity: 94%, area under the receiver operating characteristic curve: 0·94-0·98, after cross-validation). The four most prominent volatile compounds in COVID-19 patients were methylpent-2-enal, 2,4-octadiene 1-chloroheptane, and nonanal.

Interpretation: The real-time, non-invasive detection of methylpent-2-enal, 2,4-octadiene 1-chloroheptane, and nonanal in exhaled breath may identify ARDS patients with COVID-19.

Funding: The study was funded by Agence Nationale de la Recherche (SoftwAiR, ANR-18-CE45-0017 and RHU4 RECORDS, Programme d'Investissements d'Avenir, ANR-18-RHUS-0004), Région Île de France (SESAME 2016), and Fondation Foch.

Introduction

As of November 21st, 2020, about 57 million of people worldwide had been infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and about 1·4 million had died from coronavirus disease 2019 (COVID-19) [1]. Approximately 5% of patients with COVID-19 will develop acute respiratory distress syndrome (ARDS), septic shock, or multiple organ dysfunction [2]. Around the world, unprecedented research efforts are being focused on the prevention, early detection, diagnosis and management of this lethal disease. To date, only one antiviral drug (remdesivir) has been approved for the treatment of patients hospitalized for COVID-19 [3]. More recently, a large trial showed that dexamethasone at a daily dose of 6 mg for 10 days substantially reduced the risk of 28 day death (age-adjusted rate ratio [95% confidence interval (CI)]: 0·83 [0·75 to 0·93], particularly in patients with severe disease requiring invasive mechanical ventilation (rate ratio: 0·64 [0·51 to 0·81]) [4]. Although the early

* Corresponding author.
E-mail address: stanislas.grassin-delyle@uvsq.fr (S. Grassin-Delyle).
# contributed equally to this work

https://doi.org/10.1016/j.ebiom.2020.103154

Research in context

Evidence before this study

Early diagnosis of coronavirus disease 2019 (COVID-19) is of the utmost importance but remains challenging. Around 5% of patients with COVID-19 will develop acute respiratory distress syndrome (ARDS), septic shock and/or multiple organ failure; ideally, these patients should be identified as soon as possible. Breath analysis is an innovative, non-invasive, real-time, point-of-care technique for detecting volatile organic compounds (VOCs) in expired breath. It has potential for use in diagnosis and large-scale screening. However, it was not previously known whether patients with COVID-19 have a breath "signature" (also known as a "breathprint").

Added value of this study

Here, we show that breath analysis can discriminate between COVID-19 ARDS and non-COVID-19 ARDS. We characterized a VOC breathprint that was able to identify COVID-19 ARDS patients requiring invasive mechanical ventilation with high sensitivity and specificity. The four most prominent volatile compounds in the patients' breath were methylpent-2-enal, 2,4-octadiene 1-chloroheptane, and nonanal. The COVID-19 breathprint did not depend on the severity of the ARDS or the patient's viral load.

Implications of all the available evidence

All the available evidence suggest that real-time, non-invasive breath analysis could enable the large-scale screening and thus earlier treatment of patients likely to develop severe forms of COVID-19.

immune response may not depend on the severity of the illness, the most severely ill patients show persistent elevations of blood inflammatory markers (such as IL-1α, IL-1β, IL-6, IL-10, IL-18 and TNF-α) 10 or so days after SARS-CoV-2 infection, with a very high risk of subsequent organ injury [5–7]. Proteomic and metabolomic studies of serum have described a COVID-19-specific molecular signature; severe and non-severe forms of COVID-19 differ with regard to amino acid metabolism and the expression of acute phase proteins [8]. Breath analysis is an innovative, non-invasive, real-time point-of-care technique for detecting volatile organic compounds (VOCs) with potential for use in diagnosis and large-scale screening [9,10]. Thousands of VOCs have been identified in human breath following infectious, inflammatory or pathological events [11,12]. It has been suggested that the analysis of exhaled breath can be used to diagnose tuberculosis, invasive fungal infections, and bacterial colonization of the respiratory tract [13–16], together with ARDS and ventilator-associated pneumonia in patients in the intensive care unit (ICU) [17–22]. Likewise, previous studies have suggested that VOC analysis is of value in the diagnosis of viral infections in patients with chronic obstructive pulmonary disease and of influenza infections in a swine model [23,24]. The airway and lung damage caused by SARS-CoV-2 [25] might conceivably result in the release of characteristic VOCs in the exhaled breath. To test this hypothesis, we determined the metabolomic breath signature in a group of ARDS patients with or without COVID-19 and who required invasive mechanical ventilation.

Methods

Study design and oversight

This prospective study was part of the observational phase of the ongoing RECORDS trial (NCT04280497) and was conducted at the ICU of Raymond Poincaré Hospital (Garches, France). The RECORDS study protocol was approved by an ethics commitee (*Comite de Protection des Personnes EST I*, Dijon, France; reference 20.03.10.51415) and the French National Agency for Healthcare Product Safety (ANSM, Paris, France). The study was registered with the European Union Drug Regulating Authorities Clinical Trials Database (EudraCT 2020-000296-21). Whenever possible, participants or their legally authorized next of kin provided written, informed consent before inclusion. In the remaining cases, patients provided their deferred, written, informed consent. This investigator-led study was publicly funded. All the authors had full and independent access to all data and vouch for the integrity, accuracy, and completeness of the data and analysis and for the adherence of the trial to the protocol.

Study participants

Adult patients (aged 18 or over) in ICUs were eligible for inclusion if they had ARDS and required invasive mechanical ventilation. ARDS was defined as all of the following: (i) acute onset, *i.e.,* within one week of an apparent clinical insult, followed by progression of the respiratory syndrome, (ii) bilateral opacities on chest imaging not explained by another lung disease (*e.g.,* pleural effusion, atelectasis, nodules etc.), (iii) no evidence of heart failure or volume overload, and (iv) $PaO_2/FiO_2 \leq 300$ mm Hg, and positive end expiratory pressure (PEEP) ≥ 5 cm H_2O [26]. The main exclusion criteria were pregnancy, an expectation of death within 48 h, and the withholding or withdrawal of treatment.

Study measurements and procedures

Variables recorded at baseline were patient demographics and anthropometrics, the source of infection, and the severity of illness (according to the Simplified Acute Physiology Score (SAPS) II and the Sequential Organ Failure Assessment (SOFA)) [27,28]. The following variables were recorded at baseline and daily during the hospital stay: core body temperature, vital signs, central hemodynamic data, standard laboratory data, microbiological and virologic data. Samples for routine surveillance of lower respiratory tract colonization were obtained every 72 h until the patient had been weaned off mechanical ventilation or had died. A nonbronchoscopic bronchoalveolar lavage was performed with three 20 mL aliquots of sterile 0·9% saline solution, with a view to collect at least 5–10 mL of effluent per sample. Samples of blood and nasopharyngeal, bronchial or bronchoalveolar lavage fluids were assayed for SARS-CoV-2 and other respiratory viruses with a PCR test, as described by the French National Reference Center for Respiratory Viruses (Institut Pasteur, Paris, France). We also recorded life-supportive therapies including mechanical ventilation, renal replacement therapy, intravenous fluids bolus and the administration of vasopressors, and adjunct therapies including corticosteroids, thiamine, vitamin C, other vitamins, nutritional supplements, blood products, anticoagulants, sedatives, stress ulcer prophylaxis, and anti-infective drugs.

Breath analysis

Each patient's expired air was analyzed daily in the morning until discharge. Measurements were made with a proton-transfer-reaction quadrupole time-of-flight mass spectrometer (Ionicon Analytik GmbH, Innsbruck, Austria) placed outside the patient room. Samples were obtained via a heated transfer line (length: 1.6 m) connected directly to the end of the endotracheal tube (*i.e.,* without disconnection from the mechanical ventilator) and with an air flow of 50 mL/min. To eliminate the dependency on the oxygen concentration in the sample matrix, recordings were performed in patients with a fraction of inspired oxygen of 100% for at least 3 min [29]. The acquisition took 2 min. H_3O^+ was used as the primary ion and the instrument settings were as follows: source voltage, 120 V; drift

tube pressure, 3·8 mbar; drift tube temperature, 60 °C; and drift tube voltage, 959 V. The mass spectrum was acquired up to *m/z* = 392, with a time resolution of 0·1 s.

Data and statistical analysis

Patient characteristics were expressed as the median [interquartile range (IQR)] for continuous variables and the frequency (percentage) for categorical variables. Patients with and without COVID-19 were compared using Fisher's exact test for categorical variables, and a t-test or the Mann-Whitney test for normally and non-normally distributed continuous variables (as evaluated with the d'Agostino-Pearson test), respectively.

Mass spectrometry data were processed with the *ptairMS* R package (https://github.com/camilleroquencourt/ptairMS) and included mass calibration, expiratory phase detection on the CO_2 extracted ion chromatogram, peak detection and quantification with background subtraction, normalization, alignment, isotope identification, and imputation of missing values. All concentration values were quoted in ppb [30]. After aligning each individual peak, ions detected in more than 70% of at least one group (COVID *vs*. non-COVID-19 ARDS) were kept; this resulted in 81 features. Missing values (corresponding to ions in exhaled breath that were not detected by the preprocessing algorithm) were imputed with the *ptairMS* package, which returns to the raw data and integrates the noise at the exact missing *m/z*. Data were then log2-transformed and standardized. Outliers (patients with a z-score >3 for at least five features) were deleted. In the remaining patients, saturated ions (acetone, H_3O^+, H_2O-H_3O^+, oxygen) and isotopes were deleted to leave a final table of 65 features. For the univariate analysis, a Wilcoxon test was performed and *p*-values were adjusted to control for the false discovery rate [31]. For multivariate analysis, data were analyzed first with principal component analysis and then with machine learning algorithms with different mathematical backgrounds (orthogonal partial least-squares discriminant analysis (OPLS-DA), linear support vector machine (SVM), elastic net, and random forest (RF); summarized in Table S1) with the R packages *ropls, e1071*, and caret [32–35]. A 10-fold, stratified cross-validation was repeated four times (in order to avoid overfitting the small number of data points), and features were selected with the elastic net and RF approaches. The models' parameters were tuned to optimize the accuracy of cross-validation. Features were ranked according to the specific metrics of each modeling method (*p*-values from the Wilcoxon test, absolute loading values from PCA, the variable importance in projection from OPLS-DA, the coefficient values from the elastic net and SVM models, and the feature importance from the RF model). An aggregated ranking was then computed by maximizing the sum of the Spearman correlation with each of the metric rankings (*RankAggreg* R package) [36]. The correlations between the metric rankings and the aggregated rank are shown in Fig. S3. To limit the risk of overfitting, we aggregated several metrics from statistical models with different mathematical backgrounds. The effects of tidal volume, serum C-reactive protein (CRP) level, body temperature, and the number of days spent in the ICU were investigated in a correlation test with the three first components of the PCA (using a Pearson's test for continuous variables and a chi-squared test for categorical variables) to detect putative factors with a strong impact on the VOC concentrations which may interfere with the prediction of the COVID-19 status (Fig. S1). For the positive end-expiratory pressure (PEEP) and respiratory rate (the median levels of which differed for each COVID-19 status), we performed a Pearson correlation test within each group (as described in the Supplementary Material and Fig. S2). No significant correlations were detected by any of these tests.

A longitudinal univariate analysis of the most important features was performed with a mixed effects model. The fixed effect represents the change in the VOC concentration as a function of the period of mechanical ventilation, with only one measurement per patient per day. We chose a spline function (sum of four b-spline functions basis of degree three uniformly distributed over time) for the fixed effect and an intercept per patient for the random effect. Intergroup differences in trends and means were assessed with an F-test (*p*-value <0·05) adjusted for the false discovery rate. The test compares the residuals of models with and without COVID status as a predictor. Correlations between VOC concentrations, the SAPS II, the SOFA score, and the viral load were analyzed using Pearson's correlation test, after adjustment for the false discovery rate.

Role of the funding source

The funding source had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication. The corresponding author confirms that he had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Patients

Between March 25th and June 25th, 2020, 40 patients (of whom 28 had confirmed COVID-19-related ARDS) were included in the study and a total of 303 measurements were made. Compared with the patients with non-COVID-19 ARDS, the patients with COVID-19 ARDS had (i) a higher respiratory rate, FiO_2, PEEP and CRP on admission, (ii) a higher incidence of treatment with hydroxychloroquine and a lower incidence of treatment with fludrocortisone after admission, and (iii) a greater likelihood of renal replacement therapy (Table 1).

Metabolomic analysis of exhaled breath

We first used an untargeted metabolomic strategy to discover the signature associated with COVID-19 ARDS. To this end, we used the first breath sample collected after admission. Twelve of the 40 participants had been hospitalized for more than 10 days at the start of the sampling period and so were excluded from this first part of the study. Hence, we analyzed 18 patients with COVID-19 ARDS and 10 with non-COVID-19 ARDS. The study groups' demographic characteristics are summarized in Table S2. A principal component analysis and an orthogonal partial least-squares discriminant analysis showed that COVID-19 was associated with a specific signature in the expired air, *i.e.,* the breathprint could discriminate between COVID-19 ARDS and non-COVID-19 ARDS cases (Fig. 1). The use of three machine learning algorithms yielded an accuracy of 93% for all three classifiers, based on the selection of 19, 16 or 65 features for the elastic net, random forest, and support vector machine algorithms, respectively (in a 10-fold stratified cross-validation, repeated four times). The corresponding receiver operating characteristic curves are shown in Fig. 2a. A Wilcoxon test with *p*-value correction for the false discovery rate highlighted VOCs that significantly distinguished between the two groups ($p<0·05$). We checked that none of the other external covariates impacted the VOC concentrations and interfered with the model's predictions (see the Supplementary Material). To determine which VOCs were most discriminant for COVID-19 ARDS, we performed a rank aggregation based on the various metrics from the previously mentioned models and the hypothesis tests. The four most relevant features in the rank aggregation were at *m/z* 99·08, 111·12, 135·09, and 143·15 (Fig. 3a). Using these four features only, the elastic net, random forest, and support vector machine algorithms yielded an accuracy of between 89% and 93% (Fig. 2b). We therefore investigated the expression of these VOCs in the whole study population throughout the period of mechanical ventilation (Fig. 3b). We observed that the VOC concentrations (i) were significantly higher in

Table 1
Patient characteristics and treatments

	COVID-19 ARDS	Non-COVID-19 ARDS	*p* value
Number of patients *(n)*	28	12	-
Males/females *(n)*	20/8	6/6	0·28
Age (years)	61 [55-72]	72 [54-79]	0·75
Body weight (kg)	80·0 [66·6-87·6]	86·5 [65·3-94·1]	0·71
Height (cm)	170 [164-175]	173 [169-175]	0·55
Body mass index (kg/m^2)	26·3 [23·7-32·4]	28·9 [23·0-30·9]	0·79
SAPS II score in the first 24 hours	62 [49-68]	46 [40-57]	0·051
SOFA score in the first 24 hours	11 [7-12]	8 [5-12]	0·37
Comorbidities: (*n* (%))			
high blood pressure	11 (39)	6 (50)	0·73
chronic obstructive pulmonary disease	2 (7)	1 (8)	>0·99
ischemic cardiac disease	5 (18)	3 (25)	0·68
cancer	2 (7)	3 (25)	0·15
Treatments before admission: (*n* (%))			
glucocorticoids	1 (4)	3 (25)	0·073
conversion enzyme inhibitors	5 (18)	1 (8)	0·54
angiotensin antagonists	2 (7)	2 (16)	0·57
Interventions after admission: (*n* (%))			
catecholamines	17 (61)	4 (33)	0·17
renal replacement therapy	9 (32)	0 (0)	0·038
Treatments after admission: (*n* (%))			
hydroxychloroquine	27 (96)	1 (8)	<0·0001
remdesivir	2 (7)	0 (0)	>0·99
lopinavir/ritonavir	7 (25)	0 (0)	0·081
glucocorticoids	11 (39)	6 (50)	0·73
fludrocortisone	1 (4)	4 (33)	0·022
eculizumab	12 (43)	4 (33)	0·73
Body temperature at first sample (°C)	37·4 [36·5-38·3]	37·3 [36·8-37·8]	0·84
Respiratory rate at first sample (breaths per min)	26 [25-28]	20 [18-23]	<0·0001
Tidal volume at first sample (mL)	420 [400-475]	438 [400-490]	0·99
Fraction of inspired oxygen at first sample (%)	80 [50-100]	48 [31-68]	0·007
Positive end-expiratory pressure at first sample (cm H_2O)	10 [8-13]	5·5 [5-8]	0·0002
Serum creatinine at first sample (μM)	74 [56-137]	67 [44-86]	0·30
Serum C-reactive protein at first sample (mg/L)	195 [175-268]	76 [23-119]	0·002

Continuous data are presented as the median [IQR].

the breath of patients with COVID-19 ARDS than in the breath of patients with non-COVID-19 ARDS, and (ii) tended to decrease over the first 10 days of hospitalization. The putative annotations for the four compounds at *m/z* 99·08, 111·12, 135·09, and 143·15 were respectively methylpent-2-enal, 2,4-octadiene 1-chloroheptane, and nonanal.

Correlation with viral load and severity scores

The viral load in bronchoalveolar fluid was measured for 18 patients. The median [IQR] value in the first sample was 7·2 [6·2–8·4] log eq. copies/mL. The VOC concentrations in the first sample were not significantly correlated with the bronchoalveolar fluid viral load or

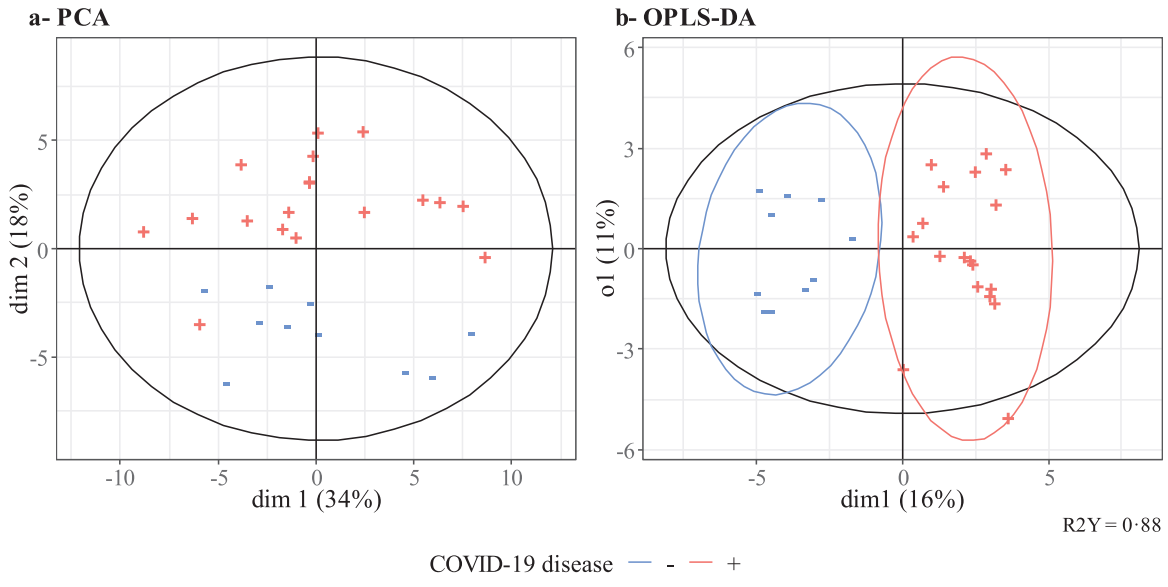


Fig. 1. Multivariate analysis. Principal component analysis (left) and orthogonal partial least squares - discriminant analysis (right) of the breath signature in intubated, mechanically ventilated ICU patients with a positive (red) or negative (blue) PCR test for SARS-CoV-2.

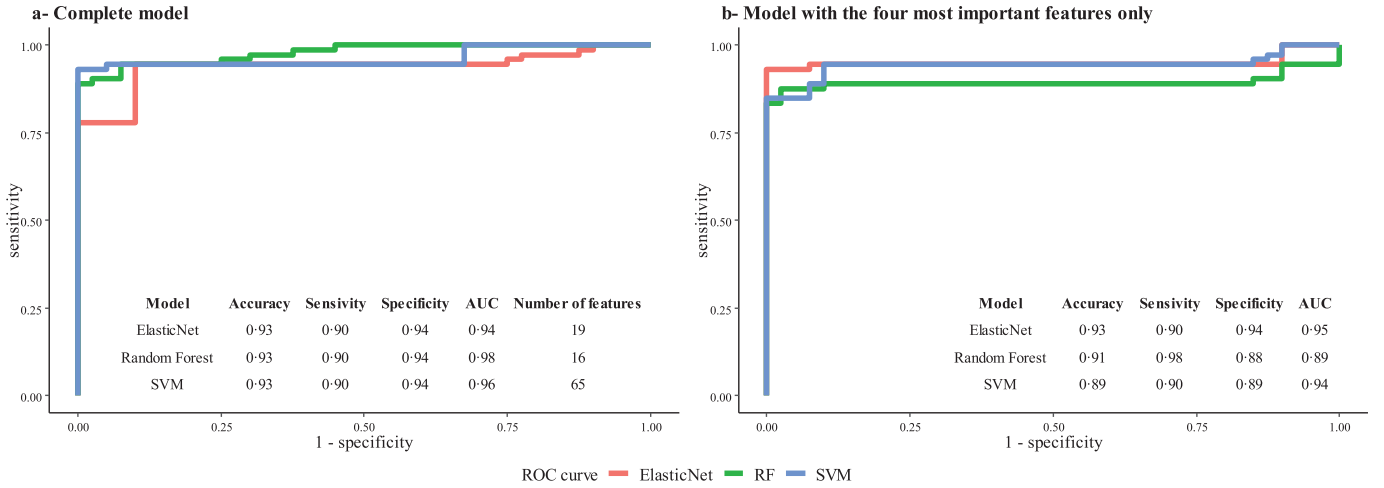


Fig. 2. Receiver operating characteristic curves for models classifying patients with COVID-19 *vs.* non-COVID-19 ARDS. a. Complete model. The use of three machine learning algorithms (elastic net, support vector machine (SVM), and random forest (RF)) yielded an accuracy of up to 93%, with a 10-fold cross validation repeated four times and based on the selection of 19 features (elastic net), 16 features (random forest) or all 65 features (support vector machine) from the full dataset. After internal cross-validation, the sensitivity was 90% and the specificity was 94%. b. Model with the four most important features only. After internal cross-validation, the sensitivity ranged from 90% to 98% and the specificity ranged from 88% to 94%.

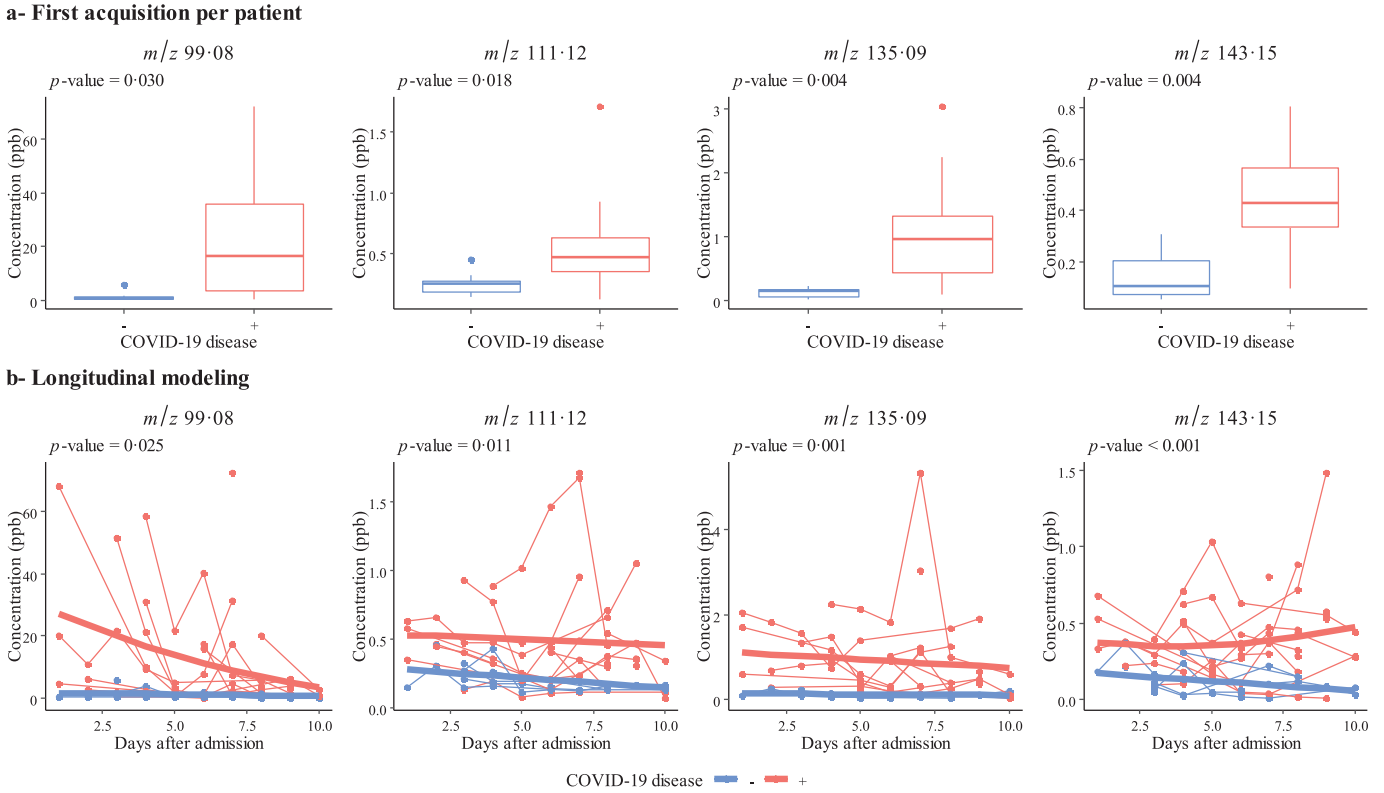


Fig. 3. Longitudinal analysis of VOCs in expired breath. The four features (*m/z* 99·08, 111·12, 135·09, and 143·15) contributing the most to the models were assessed in the first sample available for each patient (a) and over time (b) during the ICU stay for intubated, mechanically ventilated patients with COVID-19 ARDS (in red, $n = 28$) or non-COVID-19 ARDS (in blue, $n = 12$). All the points for a given patient are connected, and the bold lines correspond to the fixed effect of the mixed model for each group. *p*-values come from a Wilcoxon test (a) and an F-test (b).

with the severity of illness (*i.e.*, the SAPS II and SOFA score) [27,28] measured during the first 24 h in the ICU (Table 2, $|r| < 0{\cdot}4$).

Discussion

This study provided proof of concept for the measurement of VOCs and the determination of a specific VOC breathprint in the exhaled breath from patients with COVID-19-related ARDS requiring invasive mechanical ventilation in the ICU. This breathprint was independent of the severity of illness and the viral load. Four VOCs (methylpent-2-enal, 2,4-octadiene 1-chloroheptane, and nonanal) may discriminate between COVID-19 and non-COVID-19 ARDS.

We applied a highly sensitive, rapid, non-invasive, real-time mass spectrometry breath analysis [37,38]. This contrasts with offline technologies, which require a sampling step and remote, time-consuming analytical steps [21,22]. Implementation of a non-

Table 2
Correlations between VOC concentrations and the SAPS II, SOFA score and viral load.

VOC (*m/z*)	SAPS II score *r*	*p*-value	SOFA score *r*	*p*-value	Viral load *r*	*p*-value
99·08	0·04	0·88	0·36	0·13	0·08	0·70
111·12	0·02	0·93	0·28	0·25	-0·14	0·48
135·09	0·05	0·85	0·35	0·14	-0·0004	1·00
143·15	0·12	0·62	0·27	0·25	-0·23	0·24

r: Pearson's correlation coefficient.

targeted strategy (as described here) is mandatory for the discovery of novel biomarkers. The subsequent diagnostic validation and clinical implementation can be based on less cumbersome technologies, such as mass spectrometers dedicated to targeted analyses or portable "electronic noses" with a set of sensors that are relatively selective for different families of VOCs (as previously used in patients with ARDS) [20].

The first (cross-sectional) part of the present study enabled us to identify a specific signature. We then performed a longitudinal analysis of expired air in ARDS patients, which allowed us to confirm the VOC signature and to track the changes over time in the VOC concentrations. Two of the four prominent VOCs (methylpent-2-enal and nonanal) are aldehydes, while 2,4-octadiene is an alkadiene. These three compounds are known to be expressed in breath [39,40], while 1-chloroheptane is probably not endogenous. Nonanal is a sub-product of the destruction of the cell membrane as a result of oxidative stress; reactive oxygen species may be generated by various type of inflammatory, immune and structural cell in the airways [41]. In studies of expired air from patients with ARDS, Schubert *et al.* found abnormally low isoprene concentrations and Bos *et al.* reported abnormally high concentrations of octane, acetaldehyde and 3-methylheptane [21,22]. Differences in study populations (non-COVID-19 *vs.* COVID-19 ARDS) and analytical methods (offline *vs.* online) might explain the differences between the VOCs identified in the present study and those identified in previous studies of ARDS [21,22]. Although there may be an association between VOCs and disease, the underlying biochemistry has not been fully characterized.

In line with previous reports, the VOC concentrations measured here were not correlated with the severity of illness (as judged by the SAPS II and the SOFA score) [21]. This finding suggest that the exhaled breath signature is a marker of COVID-19 *per se*, rather than of the severity of illness. Likewise, the VOC concentrations were not correlated with viral load, suggesting that this signature may be a marker of the disease related to SARS-CoV-2 rather than of virus carriage.

Our interpretation of the present data may have been limited by differences between the COVID-19 and non- COVID-19 ARDS subgroups. Patients with COVID-19 ARDS cohort had higher respiratory rate, FiO_2, PEEP and CRP values on admission. The respiratory rate, PEEP and CRP were not found to interfere with the VOC predictive signature, and all the patients were sampled when breathing 100% FiO_2 (to avoid mass spectrometry interference by oxygen) [29]. Similarly, patients with COVID-19 ARDS were more likely to have been treated with hydroxychloroquine. However, this drug was administered to the patients after their first sample had been analyzed. Although the VOC concentrations decreased over time, the treatments did not change, and there was no correspondence between the VOCs described in the present study and the molecular masses of the known metabolites of hydroxychloroquine. Lastly, the sample size of this pilot study was limited and these observations will require confirmation with an external validation cohort.

In conclusion, we determined a COVID-19-specific breath metabolomic signature in patients with ARDS requiring invasive mechanical ventilation. Knowledge of this specific breathprint might enable the development of rapid, non-invasive, point-of-care tests for large-scale COVID-19 screening.

Acknowledgements

The authors would like to thank all the staff members at the intensive care unit at Raymond Poincaré Hospital for their collaboration, Professor Marie-Anne Welti and Professor Elyanne Gault for providing virological data, and Dr David Fraser (Biotech Communication SARL, Ploudalmézeau, France) for copy-editing assistance. The study was funded by Agence Nationale de la Recherche (SoftwAiR, ANR-18-CE45-0017 and RHU4 RECORDS, Programme d'Investissements d'Avenir, ANR-18-RHUS-0004), Région Île de France (SESAME 2016), and Fondation Foch.

Contributors

S.G.D. and D.A. conceived the study. S.G.D., P.M. and C.R. defined parameters for mass spectrometry breath analysis. S.G.D., G.S., S.C., J. F., H.S., E.N., L-J.C., P.D., P.M., D.A. performed the experiments and analyzed and/or interpreted results. S.G.D., P.M., N.H., D.A. collected epidemiological and clinical data. P.M. and N.H. assisted in patient recruitment. C.R. and E.T. developed software and analyzed the data. S.G.D. and C.R. checked the underlying data. S.G.D., C.R. and D.A. drafted the manuscript. E.T., P.M., P.D., L-J.C., H.S, E.N., G.S., S.C., N.H., J.F. revised the manuscript. All authors read and approved the final version of the manuscript.

Declaration of Competing Interests

DA has received a grant from Agence Nationale de la Recherche to conduct the RECORDS program, of which this study is part (ANR-18-RHUS-0004). S.G.D, C.R., E.T. and D.A. are named as inventors on a patent application covering breath analysis in COVID-19. The authors declare no other conflicts of interest.

Data sharing statement

The study protocol and the datasets generated during and/or analysed during the current study, including deidentified participant data will be available with publication from the corresponding author on reasonable request. The *ptairMS* R package used for data analysis is publicly available at https://github.com/camilleroquencourt/ptairMS

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.103154.

References

1 COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html (accessed October 19th, 2020).
2 Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. JAMA 2020;324(8):782–93.
3 Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of COVID-19 - final report. N Engl J Med 2020.
4 Recovery_Collaborative_Group, Horby P, Lim WS, et al. Dexamethasone in hospitalized patients with COVID-19 - preliminary report. N Engl J Med 2020.
5 Lucas C, Wong P, Klein J, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature 2020;584(7821):463–9.
6 Hadjadj J, Yatim N, Barnabei L, et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science 2020;369(6504):718–24.
7 Kuri-Cervantes L, Pampena MB, Meng W, et al. Comprehensive mapping of immune perturbations associated with severe COVID-19. Sci Immunol 2020;5(49).
8 Shen B, Yi X, Sun Y, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. Cell 2020;182(1) 59-72 e15.
9 Kataoka H, Saito K, Kato H, Masuda K. Noninvasive analysis of volatile biomarkers in human emanations for health and early disease diagnosis. Bioanalysis 2013;5(11):1443–59.

10 Rattray NJ, Hamrang Z, Trivedi DK, Goodacre R, Fowler SJ. Taking your breath away: metabolomics breathes life in to personalized medicine. Trends Biotechnol 2014;32(10):538–48.
11 Amann A, de Lacy Costello B, Miekisch W, et al. The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. J Breath Res 2014;8(3):034001.
12 de Lacy Costello B, Amann A, Al-Kateb H, et al. A review of the volatiles from the healthy human body. J Breath Res 2014;8(1):014001.
13 Koo S, Thomas HR, Daniels SD, et al. A breath fungal secondary metabolite signature to diagnose invasive aspergillosis. Clin Infect Dis 2014;59(12):1733–40.
14 Nakhleh MK, Jeries R, Gharra A, et al. Detecting active pulmonary tuberculosis with a breath test using nanomaterial-based sensors. Eur Respir J 2014;43(5):1522–5.
15 Coronel Teixeira R, Rodriguez M, Jimenez de Romero N, et al. The potential of a portable, point-of-care electronic nose to diagnose tuberculosis. J Infect 2017;75(5):441–7.
16 Suarez-Cuartin G, Giner J, Merino JL, et al. Identification of pseudomonas aeruginosa and airway bacterial colonization by an electronic nose in bronchiectasis. Respir Med 2018;136:111–7.
17 Schnabel R, Fijten R, Smolinska A, et al. Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia. Sci Rep 2015;5:17179.
18 Filipiak W, Beer R, Sponring A, et al. Breath analysis for in vivo detection of pathogens related to ventilator-associated pneumonia in intensive care patients: a prospective pilot study. J Breath Res 2015;9(1):016004.
19 Schnabel RM, Boumans ML, Smolinska A, et al. Electronic nose analysis of exhaled breath to diagnose ventilator-associated pneumonia. Respir Med 2015;109(11):1454–9.
20 Bos LD, Schultz MJ, Sterk PJ. Exhaled breath profiling for diagnosing acute respiratory distress syndrome. BMC Pulm Med 2014;14:72.
21 Bos LD, Weda H, Wang Y, et al. Exhaled breath metabolomics as a noninvasive diagnostic tool for acute respiratory distress syndrome. Eur Respir J 2014;44(1):188–97.
22 Schubert JK, Muller WP, Benzing A, Geiger K. Application of a new method for analysis of exhaled gas in critically ill patients. Intensive Care Med 1998;24(5):415–21.
23 van Geffen WH, Bruins M, Kerstjens HA. Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study. J Breath Res 2016;10(3):036001.
24 Traxler S, Bischoff AC, Sass R, et al. VOC breath profile in spontaneously breathing awake swine during Influenza A infection. Sci Rep 2018;8(1):14857.
25 Ackermann M, Verleden SE, Kuehnel M, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in COVID-19. N Engl J Med 2020;383(2):120–8.
26 Force ADT, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin definition. JAMA 2012;307(23):2526–33.
27 Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270(24):2957–63.
28 Vincent JL, Moreno R, Takala J, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. Intensive Care Med 1996;22(7):707–10.
29 Trefz P, Pugliese G, Brock B, Schubert JK, Miekisch W. Effects of elevated oxygen levels on VOC analysis by means of PTR-ToF-MS. J Breath Res 2019;13(4):046004.
30 Hansel A, Jordan A, Holzinger R, Prazeller P, Vogel W, Lindinger W. Proton transfer reaction mass spectrometry: on-line trace gas analysis at the ppb level. Int J Mass Spectrom Ion Process 1995;149-150:609–19.
31 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57(1):289–300.
32 Breiman L. Random forests. Machine Learning 2001;45(1):5–32.
33 Thevenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. J Proteome Res 2015;14(8):3322–35.
34 Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Adv Neural Inform Process Syst 2000;13.
35 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005;67(2):301–20.
36 Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics 2009;10:62.
37 Trefz P, Schmidt M, Oertel P, et al. Continuous real time breath gas monitoring in the clinical environment by proton-transfer-reaction-time-of-flight-mass spectrometry. Anal Chem 2013;85(21):10321–9.
38 Brock B, Kamysek S, Silz J, Trefz P, Schubert JK, Miekisch W. Monitoring of breath VOCs and electrical impedance tomography under pulmonary recruitment in mechanically ventilated patients. J Breath Res 2017;11(1):016005.
39 van de Kant KD, van Berkel JJ, Jobsis Q, et al. Exhaled breath profiling in diagnosing wheezy preschool children. Eur Respir J 2013;41(1):183–8.
40 Corradi M, Pignatti P, Manini P, et al. Comparison between exhaled and sputum oxidative stress biomarkers in chronic airway inflammation. Eur Respir J 2004;24(6):1011–7.
41 Rahman I. Oxidative stress, chromatin remodeling and gene transcription in inflammation and chronic lung diseases. J Biochem Mol Biol 2003;36(1):95–109.

Bibliography

Anton Amann, Ben de Lacy Costello, Wolfram Miekisch, Jochen Schubert, Bogusław Buszewski, Joachim Pleil, Norman Ratcliffe, and Terence Risby. The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *Journal of Breath Research*, 8(3):034001, June 2014. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/8/3/034001. URL http://stacks.iop.org/1752-7163/8/i=3/a=034001?key=crossref.6e7280a35b5e0d4fb0ad16f618f47439.

Martin Andersson. A comparison of nine pls1 algorithms. *Journal of Chemometrics*, 23:518 – 529, 10 2009. doi: 10.1002/cem.1248.

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. ISSN 00029947. URL http://www.jstor.org/stable/1990404.

Manor Askenazi, Hisham Ben Hamidane, and Johannes Graumann. The arc of Mass Spectrometry Exchange Formats is long, but it bends toward HDF5: plain HDF5 as a mass spectrometry exchange format. *Mass Spectrometry Reviews*, 36(5):668–673, September 2017. ISSN 02777037. doi: 10.1002/mas.21522. URL https://onlinelibrary.wiley.com/doi/10.1002/mas.21522.

Sung-June Baek, Aaron Park, Young-Jin Ahn, and Jaebum Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *The Analyst*, 140(1):250–257, 2015. ISSN 0003-2654, 1364-5528. doi: 10.1039/C4AN01061B. URL http://xlink.rsc.org/?DOI=C4AN01061B.

Amel Bajtarevic, Clemens Ager, Martin Pienz, Martin Klieber, Konrad Schwarz, Magdalena Ligor, Tomasz Ligor, Wojciech Filipiak, Hubert Denz, Michael Fiegl, Wolfgang Hilbe, Wolfgang Weiss, Peter Lukas, Herbert Jamnig, Martin Hackl, Alfred Haidenberger, Bogusław Buszewski, Wolfram Miekisch, Jochen Schubert, and Anton Amann. Noninvasive detection of lung cancer by analysis of exhaled breath. *BMC Cancer*, 9(1):348, December 2009. ISSN 1471-2407. doi: 10.1186/1471-2407-9-348. URL http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-9-348.

Eric Barat, Claude Comtat, Thomas Dautremer, Thierry Montagu, and Regine Trebossen. A non-parametric bayesian approach for PET reconstruction. In *2007 IEEE Nuclear Science Symposium Conference Record*, pages 4155–4162, Honolulu, HI, USA, 2007a. IEEE. ISBN 978-1-4244-0922-8. doi: 10.1109/NSSMIC.2007.4437035. URL http://ieeexplore.ieee.org/document/4437035/.

Eric Barat, Thomas Dautremer, and Thierry Montagu. Nonparametric bayesian inference in nuclear spectrometry. In *2007 IEEE Nuclear Science Symposium Conference Record*, pages 880–887, Honolulu, HI, USA, October 2007b. IEEE. ISBN 978-1-4244-0922-8 978-1-4244-0923-5. doi: 10.1109/NSSMIC.2007.4436469. URL http://ieeexplore.ieee.org/document/4436469/. ISSN: 1082-3654.

Elettra Barberis, Elia Amede, Shahzaib Khoso, Luigi Castello, Pier Paolo Sainaghi, Mattia Bellan, Piero Emilio Balbo, Giuseppe Patti, Diego Brustia, Mara Giordano, Roberta Rolla, Annalisa Chiocchetti, Giorgia Romani, Marcello Manfredi, and Rosanna Vaschetto. Metabolomics diagnosis of covid-19 from exhaled breath condensate. *Metabolites*, 11(12), 2021. ISSN 2218-1989. doi: 10.3390/metabo11120847. URL https://www.mdpi.com/2218-1989/11/12/847.

Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003. doi: https://doi.org/10.1002/cem.785. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.785.

J Beauchamp. Inhaled today, not gone tomorrow: pharmacokinetics and environmental exposure of volatiles in exhaled breath. 5(3):037103, jun 2011. doi: 10.1088/1752-7155/5/3/037103. URL https://doi.org/10.1088/1752-7155/5/3/037103.

Jens Behrmann, Christian Etmann, Tobias Boskamp, Rita Casadonte, Jörg Kriegsmann, and Peter Maab. Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics*, 34(7):1215–1223, 11 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx724. URL https://doi.org/10.1093/bioinformatics/btx724.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*, 57:289 – 300, 11 1995. doi: 10.2307/2346101.

Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25, 11 2015. doi: 10.1007/s11749-016-0481-7.

Robert S. Blake, Christopher Whyte, Ceri O. Hughes, Andrew M. Ellis, and Paul S. Monks. Demonstration of Proton-Transfer Reaction Time-of-Flight Mass Spectrometry for Real-Time Analysis of Trace Volatile Organic Compounds. *Analytical Chemistry*, 76(13):3841–3845, July 2004. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac0498260. URL https://pubs.acs.org/doi/10.1021/ac0498260.

Kaatje Bollaerts, Paul H. C. Eilers, and Iven Mechelen. Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, 59(2):451–469, November 2006. ISSN 00071102. doi: 10.1348/000711005X84293. URL http://doi.wiley.com/10.1348/000711005X84293.

Agnes W. Boots, Lieuwe D. Bos, Marc P. van der Schee, Frederik-Jan van Schooten, and Peter J. Sterk. Exhaled Molecular Fingerprinting in Diagnosis and Monitoring: Validating Volatile Promises. *Trends in Molecular Medicine*, 21(10):633–644, October 2015. ISSN 14714914. doi: 10.1016/j.molmed.2015.08.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1471491415001574.

Lieuwe Bos, Marcus Schultz, and Peter Sterk. Exhaled breath profiling for diagnosing acute respiratory distress syndrome. *BMC pulmonary medicine*, 14:72, 04 2014a. doi: 10.1186/

1471-2466-14-72.

Lieuwe D.J. Bos, Hans Weda, Yuanyue Wang, Hugo H. Knobel, Tamara M.E. Nijsen, Teunis J. Vink, Aeilko H. Zwinderman, Peter J. Sterk, and Marcus J. Schultz. Exhaled breath metabolomics as a noninvasive diagnostic tool for acute respiratory distress syndrome. *European Respiratory Journal*, 44(1):188–197, 2014b. ISSN 0903-1936. doi: 10.1183/09031936.00005614. URL https://erj.ersjournals.com/content/44/1/188.

Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifier. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 5, 08 1996. doi: 10.1145/130385.130401.

L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

Richard G. Brereton and Gavin R. Lloyd. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, 28(4):213–225, 2014. doi: https://doi.org/10.1002/cem.2609. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.2609.

Andreas Brezger and Stefan Lang. Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50(4):967–991, February 2006. ISSN 01679473. doi: 10.1016/j.csda.2004.10.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947304003214.

Beate Brock, Svend Kamysek, Josephine Silz, Phillip Trefz, Jochen K Schubert, and Wolfram Miekisch. Monitoring of breath VOCs and electrical impedance tomography under pulmonary recruitment in mechanically ventilated patients. *Journal of Breath Research*, 11(1):016005, January 2017. ISSN 1752-7163. doi: 10.1088/1752-7163/aa53b2. URL https://iopscience.iop.org/article/10.1088/1752-7163/aa53b2.

R.S. Brown and N.L. Gilfrich. Design and performance of a matrix-assisted laser desorption time-of-flight mass spectrometer utilizing a pulsed nitrogen laser. *Analytica Chimica Acta*, 248 (2):541–552, August 1991. ISSN 00032670. doi: 10.1016/S0003-2670(00)84673-5. URL https://linkinghub.elsevier.com/retrieve/pii/S0003267000846735.

Tobias Bruderer, Thomas Gaisl, Martin T. Gaugg, Nora Nowak, Bettina Streckenbach, Simona Müller, Alexander Moeller, Malcolm Kohler, and Renato Zenobi. On-Line Analysis of Exhaled Breath: *Focus Review*. *Chemical Reviews*, 119(19):10803–10828, October 2019. ISSN 0009-2665, 1520-6890. doi: 10.1021/acs.chemrev.9b00005. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.9b00005.

Thomas Burger. Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *Journal of Proteome Research*, 17, 10 2017. doi: 10.1021/acs.jproteome.7b00170.

Joris Cadow, Matteo Manica, Roland Mathis, Tiannan Guo, Ruedi Aebersold, and María Rodríguez Martínez. On the feasibility of deep learning applications using raw mass spectrometry data. *Bioinformatics*, 37(Supplement):i245–i253, 07 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab311. URL https://doi.org/10.1093/bioinformatics/btab311.

Kim-Anh Cao, Simon Boitard, and Philippe Besse. Sparse pls discriminant analysis: biologically

relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*, 12:253, 06 2011. doi: 10.1186/1471-2105-12-253.

Luca Cappellin, Franco Biasioli, Alessandra Fabris, Erna Schuhfried, Christos Soukoulis, Tilmann D. Märk, and Flavia Gasperi. Improved mass accuracy in PTR-TOF-MS: Another step towards better compound identification in PTR-MS. *International Journal of Mass Spectrometry*, 290(1):60–63, February 2010. ISSN 13873806. doi: 10.1016/j.ijms.2009.11.007. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380609003571.

Luca Cappellin, Franco Biasioli, Pablo M. Granitto, Erna Schuhfried, Christos Soukoulis, Fabrizio Costa, Tilmann D. Märk, and Flavia Gasperi. On data analysis in PTR-TOF-MS: From raw spectra to data mining. *Sensors and Actuators B: Chemical*, 155(1):183–190, July 2011a. ISSN 09254005. doi: 10.1016/j.snb.2010.11.044. URL https://linkinghub.elsevier.com/retrieve/pii/S0925400510009135.

Luca Cappellin, Michael Probst, Limtrakul Jumras, Franco Biasioli, Schuhfried Erna, Christos Soukoulis, Tilmann Märk, and Flavia Gasperi. Proton transfer reaction rate coefficients between h30+ and some sulphur compounds. *International Journal of Mass Spectrometry*, 295: 43–48, 07 2011b. doi: 10.1016/j.ijms.2010.06.023.

Luca Cappellin, Eugenio Aprea, Pablo Granitto, Ron Wehrens, Christos Soukoulis, Roberto Viola, Tilmann D. Märk, Flavia Gasperi, and Franco Biasioli. Linking GC-MS and PTR-TOF-MS fingerprints of food samples. *Chemometrics and Intelligent Laboratory Systems*, 118:301–307, August 2012a. ISSN 01697439. doi: 10.1016/j.chemolab.2012.05.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743912001219.

Luca Cappellin, Thomas Karl, Michael Probst, Oksana Ismailova, Paul M. Winkler, Christos Soukoulis, Eugenio Aprea, Tilmann D. Märk, Flavia Gasperi, and Franco Biasioli. On Quantitative Determination of Volatile Organic Compound Concentrations Using Proton Transfer Reaction Time-of-Flight Mass Spectrometry. *Environmental Science & Technology*, 46(4): 2283–2290, February 2012b. ISSN 0013-936X, 1520-5851. doi: 10.1021/es203985t. URL https://pubs.acs.org/doi/10.1021/es203985t.

Rosamaria Capuano, Iuliia Khomenko, Felicia Grasso, Valeria Messina, Anna Olivieri, Luca Cappellin, Roberto Paolesse, Alexandro Catini, Marta Ponzi, Franco Biasioli, and Corrado Natale. Simultaneous proton transfer reaction-mass spectrometry and electronic nose study of the volatile compounds released by plasmodium falciparum infected red blood cells in vitro. *Scientific Reports*, 9, 08 2019. doi: 10.1038/s41598-019-48732-x.

Tian-Lu Chen, Yu Cao, Yinan Zhang, Jiajian Liu, Yuqian Bao, Congrong Wang, Wei Jia, and Aihua Zhao. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-based complementary and alternative medicine : eCAM*, 2013:298183, 02 2013. doi: 10.1155/2013/298183.

Xing Chen, Keda Zhang, Zhihong Yin, Mingliang Fang, Weidan Pu, Zhening Liu, Lei Li, Pablo Sinues, Robert Dallmann, Zhen Zhou, and Xue Li. Online real-time monitoring of exhaled breath particles reveals unnoticed transport of nonvolatile drugs from blood to breath. *Analytical Chemistry*, 93(12):5005–5008, 2021. doi: 10.1021/acs.analchem.1c00509. URL https://doi.org/10.1021/acs.analchem.1c00509. PMID: 33724781.

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do,

Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141): 20170387, 2018. doi: 10.1098/rsif.2017.0387. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0387.

Marie Chion, Christine Carapito, and Frédéric Bertrand. Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics. 2021. URL https://arxiv.org/abs/2108.07086.

J. N. Coles and M. Guilhaus. Resolution limitations from detector pulse width and jitter in a linear orthogonal-acceleration time-of-flight mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 5(8):772–778, August 1994. ISSN 1044-0305, 1879-1123. doi: 10.1016/1044-0305(94)80010-3. URL http://link.springer.com/10.1016/1044-0305(94)80010-3.

Kevin R Coombes, Jr Fritsche, Herbert A, Charlotte Clarke, Jeng-neng Chen, Keith A Baggerly, Jeffrey S Morris, Lian-chun Xiao, Mien-Chie Hung, and Henry M Kuerer. Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization. *Clinical Chemistry*, 49(10):1615–1623, 10 2003. ISSN 0009-9147. doi: 10.1373/49.10.1615. URL https://doi.org/10.1373/49.10.1615.

M. Corradi, P. Pignatti, P. Manini, R. Andreoli, M. Goldoni, M. Poppa, G. Moscato, B. Balbi, and A. Mutti. Comparison between exhaled and sputum oxidative stress biomarkers in chronic airway inflammation. *European Respiratory Journal*, 24(6):1011–1017, 2004. ISSN 0903-1936. doi: 10.1183/09031936.04.00002404. URL https://erj.ersjournals.com/content/24/6/1011.

S M Cristescu, H A Gietema, L Blanchet, C L J J Kruitwagen, P Munnik, R J van Klaveren, J W J Lammers, L Buydens, F J M Harren, and P Zanen. Screening for emphysema via exhaled volatile organic compounds. *Journal of Breath Research*, 5(4):046009, December 2011. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/5/4/046009. URL http://stacks.iop.org/1752-7163/5/i=4/a=046009?key=crossref.c32325fed4ac4e42a95c31b898674471.

I D Currie and M Durban. Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 2(4):333–349, December 2002. ISSN 1471-082X, 1477-0342. doi: 10.1191/1471082x02st039ob. URL http://journals.sagepub.com/doi/10.1191/1471082x02st039ob.

Pijush Das, Anirban Roychowdhury, Subhadeep Das, Susanta Roychoudhury, and Sucheta Tripathy. sigfeature: Novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. *Frontiers in Genetics*, 11:247, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00247. URL https://www.frontiersin.org/article/10.3389/fgene.2020.00247.

Carl de Boor. *A practical guide to splines*. 1978.

Nicolien C. de Clercq, Tom van den Ende, Andrei Prodan, Mark I. van Berge Henegouwen, Suzanne S. Gisbertz, Sybren L. Meijer, Sandor Schokker, Jacques J.G.H.M. Bergman, Nadia Haj Mohammad, Jelle P. Ruurda, Richard van Hillegersberg, Stella Mook, Nicole C.T.

van Grieken, Tanja D. de Gruijl, Mark Davids, Maarten F. Bijlsma, Maarten C.C.M. Hulshof, Hanneke W.M. Van Laarhoven, and Max Nieuwdorp. Intestinal and tumor microbiome analysis combined with metabolomics of the anti-pd-l1 phase ii perfect trial for resectable esophageal adenocarcinoma. *Journal of Clinical Oncology*, 38(15_suppl):4556–4556, 2020. doi: 10.1200/JCO.2020.38.15_suppl.4556. URL https://doi.org/10.1200/JCO.2020.38.15_suppl.4556.

E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications*. Wiley, 2007. ISBN 9780470512135. URL https://books.google.fr/books?id=6D_Zz2cvgvUC.

B de Lacy Costello, A Amann, H Al-Kateb, C Flynn, W Filipiak, T Khalid, D Osborne, and N M Ratcliffe. A review of the volatiles from the healthy human body. *Journal of Breath Research*, 8(1):014001, January 2014. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/8/1/014001. URL http://stacks.iop.org/1752-7163/8/i=1/a=014001?key=crossref.b6b6a6911efb2c74c533c2e3a6bae189.

Johan J. de Rooi and Paul H.C. Eilers. Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, 117:56–60, August 2012. ISSN 01697439. doi: 10.1016/j.chemolab.2011.11.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743911002292.

Alexis Delabrière, Ulli M Hohenester, Benoit Colsch, Christophe Junot, François Fenaille, and Etienne A Thévenot. proFIA: a data preprocessing workflow for flow injection analysis coupled to high-resolution mass spectrometry. *Bioinformatics*, 33(23):3767–3775, December 2017. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btx458. URL https://academic.oup.com/bioinformatics/article/33/23/3767/3965327.

Eugene Demidenko. Mixed models: Theory and applications. *Mixed Models: Theory and Applications*, 01 2004. doi: 10.1002/0471728438.

Philippe Devillier, Helene Salvator, Emmanuel Naline, Louis-Jean Couderc, and Stanislas Grassin-Delyle. Metabolomics in the Diagnosis and Pharmacotherapy of Lung Diseases. *Current Pharmaceutical Design*, 23(14), May 2017. ISSN 13816128. doi: 10.2174/1381612823666170130155627. URL http://www.eurekaselect.com/149636/article.

Corrado Di Natale, Roberto Paolesse, Eugenio Martinelli, and Rosamaria Capuano. Solid-state gas sensors for breath analysis: A review. *Analytica chimica acta*, 824:1–17, 2014.

Paul Dierckx. Curve and Surface Fitting with Splines. page 5, 1995.

Natalia Drabińska, Cheryl Flynn, Norman Ratcliffe, Ilaria Belluomo, Antonis Myridakis, Oliver Gould, Matteo Fois, Amy Smart, Terry Devine, and Benjamin De Lacy Costello. A literature survey of volatiles from the healthy human breath and bodily fluids: the human volatilome. *Journal of Breath Research*, 15, 03 2021. doi: 10.1088/1752-7163/abf1d0.

P. Du, W. A. Kibbe, and S. M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, September 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl355. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl355.

Maria Durban, Iain Currie, and Paul Eilers. Using P-splines to smooth two-dimensional Poisson

data. page 8, 2002.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Rob Tibshirani. Least angle regression" (with discussions). *The Annals of Statistics*, 32, 01 2004.

P. Eilers and B Marx. *Practical Smoothing: The Joys of P-splines.* Cambridge, cambridge university press. edition, 2021. URL doi:10.1017/9781108610247.

Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121, May 1996. ISSN 0883-4237. doi: 10.1214/ss/1038425655. URL http://projecteuclid.org/euclid.ss/1038425655.

Paul H C Eilers, Brian D Marx, and Maria Durban. Twenty years of P-splines. page 38, 2015.

Reef Einoch Amor, Morad K. Nakhleh, Orna Barash, and Hossam Haick. Breath analysis of cancer in the present and the future. *European Respiratory Review*, 28(152):190002, June 2019. ISSN 0905-9180, 1600-0617. doi: 10.1183/16000617.0002-2019. URL http://err.ersjournals.com/lookup/doi/10.1183/16000617.0002-2019.

Andrew M. Ellis and Christopher A. Mayhew. *Background*, chapter 1, pages 1–23. John Wiley and Sons, Ltd, 2014. ISBN 9781118682883. doi: https://doi.org/10.1002/9781118682883.ch1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118682883.ch1.

Mariana Valente Farraia, João Cavaleiro Rufo, Inês Paciência, Francisca Mendes, Luís Delgado, and André Moreira. The electronic nose technology in clinical diagnosis: A systematic review. *Porto biomedical journal*, 4(4), 2019.

Charlotte Feil, Frank Staib, Martin R. Berger, Thorsten Stein, Irene Schmidtmann, Andrea Forster, and Carl Christoph Schimanski. Sniffer dogs can identify lung cancer patients from breath and urine samples. *BMC Cancer*, 21, 2021.

R. Fernández del Río, M.E. O'Hara, A. Holt, P. Pemberton, T. Shah, T. Whitehouse, and C.A. Mayhew. Volatile Biomarkers in Breath Associated With Liver Cirrhosis — Comparisons of Pre- and Post-liver Transplant Breath Samples. *EBioMedicine*, 2(9):1243–1250, September 2015. ISSN 23523964. doi: 10.1016/j.ebiom.2015.07.027. URL https://linkinghub.elsevier.com/retrieve/pii/S2352396415300797.

W Filipiak, V Ruzsanyi, P Mochalski, A Filipiak, A Bajtarevic, C Ager, H Denz, W Hilbe, H Jamnig, M Hackl, A Dzien, and A Amann. Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. 6(3):036008, aug 2012. doi: 10.1088/1752-7155/6/3/036008. URL https://doi.org/10.1088/1752-7155/6/3/036008.

The ARDS Definition Task Force. Acute Respiratory Distress Syndrome: The Berlin Definition. *JAMA*, 307(23):2526–2533, 06 2012.

Mario Fordellone, Andrea Bellincontro, and Fabio Mencarelli. Partial least squares discriminant analysis: A dimensionality reduction method to classify hyperspectral data. 06 2018.

Maria Franco-Villoria, Massimo Ventrucci, and Håvard Rue. A unified view on bayesian varying coefficient models. *Electronic Journal of Statistics*, 13(2):5334–5359, 2019.

Thomas Frenzel, Andreas Miller, and Karl-Heinz Engel. A methodology for automated com-

parative analysis of metabolite profiling data. *European Food Research and Technology*, 216: 335–342, 04 2003. doi: 10.1007/s00217-002-0659-y.

Andrzej Galecki and Tomasz Burzykowski. *Linear Mixed Effects Models Using R. : A Step-by-Step Approach*. 03 2013. ISBN 978-1-4614-3899-1. doi: 10.1007/978-1-4614-3900-4.

Feng Gan, Guihua Ruan, and Jinyuan Mo. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems*, 82 (1-2):59–65, May 2006. ISSN 01697439. doi: 10.1016/j.chemolab.2005.08.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743905001589.

Julian W Gardner and Philip N Bartlett. A brief history of electronic noses. *Sensors and Actuators B: Chemical*, 18(1-3):210–211, 1994.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, page 16, 2004.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2010.03.014. URL https://www.sciencedirect.com/science/article/pii/S0167865510000954.

Quentin Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Coute, and Thomas Burger. Technical brief calibration plot for proteomics (cp4p): A graphical tool to visually check the assumptions underlying fdr control in quantitative experiments. *Proteomics*, 16, 11 2015. doi: 10.1002/pmic.201500189.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. doi: 10.1080/00401706.1979.10489751. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1979.10489751.

Pablo M. Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, September 2006. ISSN 01697439. doi: 10.1016/j.chemolab.2006.01.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743906000232.

Stanislas Grassin-Delyle, Camille Roquencourt, Pierre Moine, Gabriel Saffroy, Stanislas Carn, Nicholas Heming, Jérôme Fleuriet, Hélène Salvator, Emmanuel Naline, Louis-Jean Couderc, Philippe Devillier, Etienne A. Thévenot, and Djillali Annane. Metabolomics of exhaled breath in critically ill COVID-19 patients: A pilot study. *EBioMedicine*, 63:103154, January 2021. ISSN 23523964. doi: 10.1016/j.ebiom.2020.103154. URL https://linkinghub.elsevier.com/retrieve/pii/S2352396420305302.

Peter J. Grenn. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 12 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711. URL https://doi.org/10.1093/biomet/82.4.711.

P.J. Grenn and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models*. School of Mathematics University of Bristol UK, 1994.

Oswaldo Gressani and Philippe Lambert. Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis*, 154:107088, February 2021. ISSN 01679473. doi: 10.1016/j.csda.2020.107088. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947320301791.

J.H. Gross. Springer Berlin Heidelberg, 2011. doi: https://10.1007/978-3-642-10711-5.

M. Guilhaus, D. Selby, and V. Mlynski. Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrometry Reviews*, 19(2):65–107, 2000. doi: https://doi.org/10.1002/(SICI)1098-2787(2000)19:2<65::AID-MAS1>3.0.CO;2-E. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2787%282000%2919%3A2%3C65%3A%3AAID-MAS1%3E3.0.CO%3B2-E.

A. Guirao, L. Molins, I. Ramón, G. Sunyer, N. Viñolas, R. Marrades, D. Sánchez, J.J. Fibla, M. Boada, J. Hernández, R. Guzmán, A. Libreros, A. Gómez-Caro, C. Guerrero, and A. Agustí. Trained dogs can identify malignant solitary pulmonary nodules in exhaled gas. *Lung Cancer*, 135:230–233, 2019. ISSN 0169-5002. doi: https://doi.org/10.1016/j.lungcan.2019.06.008. URL https://www.sciencedirect.com/science/article/pii/S016950021930501X.

Alexander Gundlach-Graham, Lyndsey Hendriks, Kamyar Mehrabi, and Detlef Gu. Monte Carlo Simulation of Low-Count Signals in Time-of-Flight Mass Spectrometry and Its Application to Single-Particle Detection. *Anal. Chem.*, page 9, 2018.

Yu Guo, Armin Graber, Robert Mcburney, and Raji Balasubramanian. Sample size and statistical power considerations in high-dimensionality data settings: A comparative study of classification algorithms. *BMC bioinformatics*, 11:447, 09 2010. doi: 10.1186/1471-2105-11-447.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 01 2002. doi: 10.1023/A:1012487302797.

Eugen Hartungen, Armin Wisthaler, Tomas Mikoviny, Dagmar Jaksch, Elena Boscaini, Patrick Dunphy, and Tilmann Märk. Proton-transfer-reaction mass spectrometry (ptr-ms) of carboxylic acids. *International Journal of Mass Spectrometry*, 239:243–248, 12 2004. doi: 10.1016/j.ijms.2004.09.009.

David A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977. doi: 10.1080/01621459.1977.10480998. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480998.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. doi: https://doi.org/10.1111/j.2517-6161.1993.tb01939.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/

j.2517-6161.1993.tb01939.x.

Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays. *Technical report, Stanford Statistics Department*, 1, 12 2001.

Joshua Heinemann, Aurélien Mazurie, Monika Tokmina-Lukaszewska, Greg Beilman, and Brian Bothner. Application of support vector machines to metabolomics experiments with limited replicates. *Metabolomics*, 10, 12 2014. doi: 10.1007/s11306-014-0651-0.

Ben Henderson, Dorota Ruszkiewicz, Maxim Wilkinson, Jonathan Beauchamp, S.M. Cristescu, Stephen Fowler, Dahlia Salman, Fabio Di Francesco, Gudrun Koppen, Jens Langejuergen, Olaf Holz, Andria Hadjithekli, Sergi Moreno, Michele Pedrotti, Pablo Sinues, Gitte Slingers, Michael Wilde, Tommaso Lomonaco, Delphine Zanella, and Charles Thomas. A benchmarking protocol for breath analysis: The peppermint experiment. *Journal of Breath Research*, 14, 06 2020. doi: 10.1088/1752-7163/aba130.

Jens Herbig and Jonathan Beauchamp. Towards standardization in the analysis of breath gas volatiles. *Journal of Breath Research*, 8(3):037101, September 2014. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/8/3/037101. URL http://stacks.iop.org/1752-7163/8/i=3/a=037101?key=crossref.4e3bdba7f2701356e225d9050ae0ae5e.

Jens Herbig, Thorsten Titzmann, Jonathan Beauchamp, Ingrid Kohl, and Armin Hansel. Buffered end-tidal (BET) sampling—a novel method for real-time breath-gas analysis. *Journal of Breath Research*, 2(3):037008, September 2008. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/2/3/037008. URL http://stacks.iop.org/1752-7163/2/i=3/a=037008?key=crossref.6c1463e66d0714dce9de2a6cfe05bf66.

Jens Herbig, Markus Müller, Simon Schallhart, Thorsten Titzmann, Martin Graus, and Armin Hansel. On-line breath analysis with PTR-TOF. *Journal of Breath Research*, 3(2):027004, June 2009. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/3/2/027004. URL http://stacks.iop.org/1752-7163/3/i=2/a=027004?key=crossref.cd366040f247a8c9557af6c2233e90f2.

Benjamin Heuclin, Frédéric Mortier, Catherine Trottier, and Marie Denis. Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):24–50, 2021. doi: https://doi.org/10.1111/rssc.12447. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12447.

A. Hoerl and R. Kennard. Ridge regression. *In Encyclopedia of Statistical Sciences*, 8:129–136, 1988.

R. Holzinger. PTRwid: A new widget tool for processing PTR-TOF-MS data. *Atmospheric Measurement Techniques*, 8(9):3903–3922, September 2015. ISSN 1867-8548. doi: 10.5194/amt-8-3903-2015. URL https://www.atmos-meas-tech.net/8/3903/2015/.

DONALD R. Hoover, JOHN A. Rice, COLIN O. WU, and LI-PING YANG. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 12 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.4.809. URL https://doi.org/10.1093/biomet/85.4.809.

I. Horvath, Z. Lazar, N. Gyulai, M. Kollai, and G. Losonczy. Exhaled biomarkers in lung cancer.

European Respiratory Journal, 34(1):261–275, July 2009. ISSN 0903-1936, 1399-3003. doi: 10.1183/09031936.00142508. URL http://erj.ersjournals.com/cgi/doi/10.1183/09031936.00142508.

Wadah Ibrahim, Michael Wilde, Rebecca Cordell, Dahlia Salman, Dorota Ruszkiewicz, Luke Bryant, Matthew Richardson, Robert C Free, Bo Zhao, Ahmed Yousuf, Christobelle White, Richard Russell, Sheila Jones, Bharti Patel, Asia Awal, Rachael Phillips, Graham Fowkes, Teresa McNally, Clare Foxon, Hetan Bhatt, Rosa Peltrini, Amisha Singapuri, Beverley Hargadon, Toru Suzuki, Leong L Ng, Erol Gaillard, Caroline Beardsmore, Kimuli Ryanna, Hitesh Pandya, Tim Coates, Paul S Monks, Neil Greening, Christopher E Brightling, Paul Thomas, and Salman Siddiqui. Assessment of breath volatile organic compounds in acute cardiorespiratory breathlessness: a protocol describing a prospective real-world observational study. *BMJ Open*, 9(3): e025486, March 2019. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2018-025486. URL https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2018-025486.

Wadah Ibrahim, Rebecca L. Cordell, Michael J. Wilde, Matthew Richardson, Liesl Carr, Ananga Sundari Devi Dasi, Beverley Hargadon, Robert C. Free, Paul S. Monks, Christopher E. Brightling, Neil J. Greening, and Salman Siddiqui. Diagnosis of covid-19 by exhaled breath analysis using gas chromatography–mass spectrometry. *ERJ Open Research*, 7(3), 2021. doi: 10.1183/23120541.00139-2021. URL https://openres.ersjournals.com/content/7/3/00139-2021.

Ronald Iman. Use of a t-statistic as ah approximation to the exact distribution of the wilcoxon signed ranks test statistic. *Communications in Statistics - Theory and Methods*, 3:795–806, 01 1974. doi: 10.1080/03610927408827178.

Alyssa Imbert, Magali Rompais, Mohammed Selloum, Florence Castelli, Emmanuelle Mouton-Barbosa, Marion Brandolini, Emeline Chu-Van, Charlotte Joly, Aurélie Hirschler, Pierrick Roger, Thomas Burger, Sophie Leblanc, Tania Sorg, Sadia Ouzia, Yves Vandenbrouck, Claudine Médigue, Christophe Junot, Myriam Ferro, Estelle Pujos-Guillot, and Etienne Thévenot. Prometis, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Scientific Data*, 8, 12 2021. doi: 10.1038/s41597-021-01095-3.

Neal Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics (Oxford, England)*, 21:3066–73, 08 2005. doi: 10.1093/bioinformatics/bti482.

Arlene John, Jishnu Sadasivan, and Chandra Sekhar Seelamantula. Adaptive savitzky-golay filtering in non-gaussian noise. *IEEE Transactions on Signal Processing*, 69:5021–5036, 2021. doi: 10.1109/TSP.2021.3106450.

Caroline Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17, 03 2016. doi: 10.1038/nrm.2016.25.

A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Märk, H. Seehauser, R. Schottkowsky, P. Sulzer, and T.D. Märk. A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *International Journal of Mass Spectrometry*, 286(2-3):122–128, September 2009. ISSN 13873806. doi: 10.1016/j.ijms.2009.07.005. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380609002371.

A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications.

pages 1200–1205, 2015. doi: 10.1109/MIPRO.2015.7160458.

J. Kaiser. Nonrecursive digital filter design using the i-sinh window function. *Proceedings of the IEEE*, 1977.

Edward D. Kantz, Saumya Tiwari, Jeramie D. Watrous, Susan Cheng, and Mohit Jain. Deep neural networks for classification of lc-ms spectral peaks. *Analytical Chemistry*, 91(19):12407–12413, 2019. doi: 10.1021/acs.analchem.9b02983. URL https://doi.org/10.1021/acs.analchem.9b02983. PMID: 31483992.

Mikko Katajamaa and Matej Orešič. Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158(1-2):318–328, July 2007. ISSN 00219673. doi: 10.1016/j.chroma.2007.04.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0021967307006966.

Michael G. Kenward and James H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997, 1997. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533558.

Young-Ju Kim and Chong Gu. Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356, 2004. doi: https://doi.org/10.1046/j.1369-7412.2003.05316.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1046/j.1369-7412.2003.05316.x.

Ł. Komsta. Comparison of Several Methods of Chromatographic Baseline Removal with a New Approach Based on Quantile Regression. *Chromatographia*, 73(7-8):721–731, April 2011. ISSN 0009-5893, 1612-1112. doi: 10.1007/s10337-011-1962-1. URL http://link.springer.com/10.1007/s10337-011-1962-1.

Sophia Koo, Horatio Thomas, S Daniels, Robert Lynch, Sean Fortier, Margaret Shea, Preshious Rearden, James Comolli, Lindsey Baden, and Francisco Marty. A breath fungal secondary metabolite signature to diagnose invasive aspergillosis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 59, 10 2014. doi: 10.1093/cid/ciu725.

Quincey Koziol. HDF5. In David Padua, editor, *Encyclopedia of Parallel Computing*, pages 827–833. Springer US, Boston, MA, 2011. ISBN 978-0-387-09766-4. doi: 10.1007/978-0-387-09766-4_44. URL https://doi.org/10.1007/978-0-387-09766-4_44.

Carsten Kuhl, Ralf Tautenhahn, and Steffen Neumann. LC-MS Peak Annotation and Identification with CAMERA. page 15, 2009.

Tien-Chueh Kuo, Cheng-En Tan, San-Yuan Wang, Olivia A Lin, Bo-Han Su, Ming-Tsung Hsu, Jessica Lin, Yu-Yen Cheng, Ciao-Sin Chen, Yu-Chieh Yang, Kuo-Hsing Chen, Shu-Wen Lin, Chao-Chi Ho, Ching-Hua Kuo, and Yufeng Jane Tseng. Human Breathomics Database. 2020:8, 2020.

Olav Kvalheim, Reidar Arneberg, Bjørn Grung, and Tarja Rajalahti Kvalheim. Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by monte carlo resampling: Determination of optimum number of components in pls regression. *Journal of Chemometrics*, 32:e2993, 01 2018. doi: 10.1002/cem.2993.

Nan M. Laird and James H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, 38

(4):963, December 1982. ISSN 0006341X. doi: 10.2307/2529876. URL https://www.jstor.org/stable/2529876?origin=crossref.

Eva Lange, Clemens Gröpl, Reinert Knut, Oliver Kohlbacher, and Andreas Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 11:243–54, 02 2006. doi: 10.1142/9789812701626_0023.

Eva Lange, Clemens Gröpl, Ole Schulz-Trieglaff, Andreas Leinenbach, Christian Huber, and Knut Reinert. A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics*, 23(13):i273–i281, July 2007. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btm209. URL https://academic.oup.com/bioinformatics/article/23/13/i273/233877.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *JAMA*, 270(24):2957–2963, 12 1993.

Dae-Jin Lee, María Durbán, and Paul Eilers. Efficient two-dimensional smoothing with p-spline anova mixed models and nested bases. *Computational Statistics and Data Analysis*, 61:22–37, 2013. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2012.11.013. URL https://www.sciencedirect.com/science/article/pii/S016794731200415X.

Kyriacos Leptos, David Sarracino, Jacob Jaffe, Bryan Krastins, and George Church. Mapquant: Open-source software for large-scale protein quantification. *Proteomics*, 6:1770–82, 03 2006. doi: 10.1002/pmic.200500201.

Jianbo Li and Riquan Zhang. Penalized spline varying-coefficient single-index model. *Communications in Statistics—Simulation and Computation®*, 39(2):221–239, 2010.

Jingyi Jessica Li and Xin Tong. Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines. *Patterns*, 1(7):100115, 2020.

Aikaterini Liangou, Antonios Tasoglou, Heinz J. Huber, Christopher Wistrom, Kevin Brody, Prahlad G Menon, Thomas Bebekoski, Kevin Menschel, Marlise Davidson-Fiedler, Karl DeMarco, Harshad Salphale, Jonathan Wistrom, Skyler Wistrom, and Richard J. Lee. A method for the identification of covid-19 biomarkers in human breath using proton transfer reaction time-of-flight mass spectrometry. *eClinicalMedicine*, 42:101207, 2021. ISSN 2589-5370. doi: https://doi.org/10.1016/j.eclinm.2021.101207. URL https://www.sciencedirect.com/science/article/pii/S2589537021004880.

Zaiyou Liu and John B Phillips. Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface. *Journal of Chromatographic Science*, 29(6):227–231, 1991.

Celia Isabel López-Lorente, Mo Awchi, Pablo Sinues, and Diego García-Gómez. Real-time pharmacokinetics via online analysis of exhaled breath. *Journal of Pharmaceutical and Biomedical Analysis*, 205:114311, 2021. ISSN 0731-7085. doi: https://doi.org/10.1016/j.jpba.2021.114311. URL https://www.sciencedirect.com/science/article/pii/S0731708521004222.

Benjamin Löser, Alina Grabenschröer, Giovanni Pugliese, Pritam Sukul, Phillip Trefz, Jochen K

Schubert, and Wolfram Miekisch. Changes of Exhaled Volatile Organic Compounds in Post-operative Patients Undergoing Analgesic Treatment: A Prospective Observational Study. *Metabolites*, 10(8):321, August 2020. ISSN 2218-1989. doi: 10.3390/metabo10080321. URL https://www.mdpi.com/2218-1989/10/8/321.

Sankar Mahadevan, Sirish L. Shah, Thomas J. Marrie, and Carolyn M. Slupsky. Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80(19):7562–7570, 2008. doi: 10.1021/ac800954c. URL https://doi.org/10.1021/ac800954c. PMID: 18767870.

S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, 1992. doi: 10.1109/34.142909.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Elena Marchiori, Connie R. Jimenez, Mikkel West-Nielsen, and Niels H. H. Heegaard. Robust svm-based biomarker selection with noisy mass spectrometric proteomic data. pages 79–90, 2006.

Pablo Martinez-Lozano Sinues, Malcolm Kohler, and Renato Zenobi. Human breath analysis may support the existence of individual metabolic phenotypes. *PLOS ONE*, 8(4):1–5, 04 2013. doi: 10.1371/journal.pone.0059909. URL https://doi.org/10.1371/journal.pone.0059909.

Brian D Marx. P-spline varying coefficient models for complex data. In *Statistical modelling and regression structures*, pages 19–43. Springer, 2010.

Brian D Marx and Paul H.C Eilers. Multidimensional Penalized Signal Regression. *Technometrics*, 47(1):13–22, February 2005. ISSN 0040-1706, 1537-2723. doi: 10.1198/004017004000000626. URL http://www.tandfonline.com/doi/abs/10.1198/004017004000000626.

R Daniel Mauldin, William D Sudderth, and Stanley C Williams. Polya trees and random distributions. *The Annals of Statistics*, pages 1203–1221, 1992.

Wolfram Miekisch, Sabine Kischkel, Annika Sawacki, Tina Liebau, Maren Mieth, and Jochen K Schubert. Impact of sampling procedures on the results of breath analysis. *Journal of Breath Research*, 2(2):026007, June 2008. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/2/2/026007. URL http://stacks.iop.org/1752-7163/2/i=2/a=026007?key=crossref.8604809836ce670fcb15d956965e8a96.

Wolfram Miekisch, Jens Herbig, and Jochen K Schubert. Data interpretation in breath biomarker research: pitfalls and directions. *Journal of Breath Research*, 6(3):036007, September 2012. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/6/3/036007. URL http://stacks.iop.org/1752-7163/6/i=3/a=036007?key=crossref.05eaf83b4b4b20424c143d1a0c6701b3.

Miroslav Morháč and Vladislav Matoušek. Peak clipping algorithms for background estimation in spectroscopic data. *Applied Spectroscopy*, 62(1):91–106, 2008. doi: 10.1366/000370208783412762. URL https://doi.org/10.1366/000370208783412762. PMID: 18230214.

M. Müller, M. Graus, T. M. Ruuskanen, R. Schnitzhofer, I. Bamberger, L. Kaser, T. Titzmann, L. Hörtnagl, G. Wohlfahrt, T. Karl, and A. Hansel. First eddy covariance flux measurements by ptr-tof. *Atmospheric Measurement Techniques*, 3(2):387–395, 2010. doi: 10.5194/

amt-3-387-2010. URL https://amt.copernicus.org/articles/3/387/2010/.

M. Müller, T. Mikoviny, and A. Wisthaler. Detector aging induced mass discrimination and non-linearity effects in PTR-ToF-MS. *International Journal of Mass Spectrometry*, 365-366:93–97, May 2014. ISSN 13873806. doi: 10.1016/j.ijms.2013.12.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380613004338.

Markus Müller, Christian George, and Barbara D'Anna. Enhanced spectral analysis of C-TOF Aerosol Mass Spectrometer data: Iterative residual analysis and cumulative peak fitting. *International Journal of Mass Spectrometry*, 306(1):1–8, September 2011. ISSN 13873806. doi: 10.1016/j.ijms.2011.04.007. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380611001576.

Markus Müller, Tomáš Mikoviny, Werner Jud, Barbara D'Anna, and Armin Wisthaler. A new software tool for the analysis of high resolution PTR-TOF mass spectra. *Chemometrics and Intelligent Laboratory Systems*, 127:158–165, August 2013. ISSN 01697439. doi: 10.1016/j.chemolab.2013.06.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0169743913001275.

Morad Nakhleh, Raneen Jeries Zaher, A'laa Gharra, Anke Binder, Yoav Broza, Mellissa Pascoe, Keertan Dheda, and Hossam Haick. Detecting active pulmonary tuberculosis with a breath test using nanomaterial-based sensors. *The European respiratory journal*, 43:1522–1525, 05 2014. doi: 10.1183/09031936.00019114.

Inbar Nardi-Agmon, Manal Abud-Hawa, Ori Liran, Naomi Gai-Mor, Maya Ilouze, Amir Onn, Jair Bar, Dekel Shlomi, Hossam Haick, and Nir Peled. Exhaled Breath Analysis for Monitoring Response to Treatment in Advanced Lung Cancer. *Journal of Thoracic Oncology*, 11(6):827–837, June 2016. ISSN 15560864. doi: 10.1016/j.jtho.2016.02.017. URL https://linkinghub.elsevier.com/retrieve/pii/S1556086416004408.

Titin Agustin Nengsih, Frédéric Bertrand, Myriam Maumy-Bertrand, and Nicolas Meyer. Determining the number of components in pls regression on incomplete data set. *Statistical Applications in Genetics and Molecular Biology*, 18(6):20180059, 2019. doi: doi:10.1515/sagmb-2018-0059. URL https://doi.org/10.1515/sagmb-2018-0059.

Juliane Obermeier, Phillip Trefz, Josephine Happ, Jochen K. Schubert, Hagen Staude, Dagmar-Christiane Fischer, and Wolfram Miekisch. Exhaled volatile substances mirror clinical conditions in pediatric chronic kidney disease. *PLOS ONE*, 12(6):e0178745, June 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0178745. URL https://dx.plos.org/10.1371/journal.pone.0178745.

Stephen G. Oliver, Michael K. Winson, Douglas B. Kell, and Frank Baganz. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16(9):373–378, 1998. ISSN 0167-7799. doi: https://doi.org/10.1016/S0167-7799(98)01214-1. URL https://www.sciencedirect.com/science/article/pii/S0167779998012141.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 12 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.3.545. URL https://doi.org/10.1093/biomet/58.3.545.

Jorge Pereira, Priscilla Porto-Figueira, Carina Cavaco, Khushman Taunk, Srikanth Rapole, Rahul Dhakne, Hampapathalu Nagarajaram, and José Câmara. Breath Analysis as a Potential and

Non-Invasive Frontier in Disease Diagnosis: An Overview. *Metabolites*, 5(1):3–55, January 2015. ISSN 2218-1989. doi: 10.3390/metabo5010003. URL http://www.mdpi.com/2218-1989/5/1/3.

Michael Phillips. Method for the collection and assay of volatile organic compounds in breath. *Analytical Biochemistry*, 247(2):272–278, 1997. ISSN 0003-2697. doi: https://doi.org/10.1006/abio.1997.2069. URL https://www.sciencedirect.com/science/article/pii/S0003269797920698.

Michael Phillips, Joel Greenberg, and Marilu Sabas. Alveolar gradient of pentane in normal human breath. *Free Radical Research*, 20(5):333–337, 1994. doi: 10.3109/10715769409145633. URL https://doi.org/10.3109/10715769409145633. PMID: 8069391.

Michael Phillips, Nasser Altorki, John H.M. Austin, Robert B. Cameron, Renee N. Cataneo, Joel Greenberg, Robert Kloss, Roger A. Maxfield, Muhammad I. Munawar, Harvey I. Pass, Asif Rashid, William N. Rom, and Peter Schmitt. Prediction of lung cancer using volatile biomarkers in breath1. *Cancer Biomarkers*, 3(2):95–109, April 2007. ISSN 18758592, 15740153. doi: 10.3233/CBM-2007-3204. URL http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/CBM-2007-3204.

Michael Phillips, Renee N. Cataneo, Anirudh Chaturvedi, Peter D. Kaplan, Mark Libardoni, Mayur Mundada, Urvish Patel, and Xiang Zhang. Detection of an extended human volatome with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. *PLOS ONE*, 8(9):null, 09 2013. doi: 10.1371/journal.pone.0075274. URL https://doi.org/10.1371/journal.pone.0075274.

Vincent Picaud, Jean-Francois Giovannelli, Caroline Truntzer, Jean-Philippe Charrier, Audrey Giremus, Pierre Grangeat, and Catherine Mercier. Linear MALDI-ToF simultaneous spectrum deconvolution and baseline removal. *BMC Bioinformatics*, 19(1), December 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2116-3. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2116-3.

Vasyl Pihur, Susmita Datta, and Somnath Datta. Rankaggreg, an r package for weighted rank aggregation. *BMC bioinformatics*, 10:62, 03 2009. doi: 10.1186/1471-2105-10-62.

Pinheiro and Bates. *Mixed-Effects Models in S and S-PLUS*. 2000.

Joachim Pleil, Matthew Stiegel, and Terence Risby. Clinical breath analysis: Discriminating between human endogenous compounds and exogenous (environmental) chemical confounders. *Journal of breath research*, 7:017107, 03 2013. doi: 10.1088/1752-7155/7/1/017107.

Joachim D Pleil, A Hansel, and Jonathan D Beauchamp. Advances in proton transfer reaction mass spectrometry (PTR-MS): applications in exhaled breath analysis, food science, and atmospheric chemistry. *Journal of Breath Research*, May 2019. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7163/ab21a7. URL http://iopscience.iop.org/article/10.1088/1752-7163/ab21a7.

Yotsawat Pomyen, Kwanjeera Wanichthanarak, Patcha Poungsombat, Johannes Fahrmann, Dmitry Grapov, and Sakda Khoomrung. Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal*, 18:2818–2825, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2020.09.033. URL https://www.sciencedirect.com/science/article/pii/S2001037020304177.

Cristian Quiroz-Moreno, Mayra Furlan, Joao Belinato, Fabio Augusto, Guilherme Alexandrino, and Noroska Mogollón. Rgcxgc toolbox: An r-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry. *Microchemical Journal*, 156:104830, 03 2020. doi: 10.1016/j.microc.2020.104830.

Irfan Rahman. Oxidative stress, chromatin remodeling and gene transcription in inflammation and chronic lung diseases. *Journal of biochemistry and molecular biology*, 36:95–109, 02 2003. doi: 10.5483/BMBRep.2003.36.1.095.

Ryne C Ramaker, Emily R Gordon, and Sara J Cooper. R2DGC: threshold-free peak alignment and identification for 2D gas chromatography-mass spectrometry in R. *Bioinformatics*, 34(10): 1789–1791, 12 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx825. URL https://doi.org/10.1093/bioinformatics/btx825.

James Ramsay, Nancy Heckman, and Bernard Silverman. Spline smoothing with model-based penalties. *Behavior Research Methods, Instruments, Computers*, 29:99–106, 01 1996. doi: 10.3758/BF03200573.

Nicholas J.W. Rattray, Zahra Hamrang, Drupad K. Trivedi, Royston Goodacre, and Stephen J. Fowler. Taking your breath away: metabolomics breathes life in to personalized medicine. *Trends in Biotechnology*, 32(10):538–548, October 2014. ISSN 01677799. doi: 10.1016/j.tibtech.2014.08.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0167779914001632.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, 2017. doi: 10.1162/neco_a_00990.

Nader Rifai, Michael Gillette, and Steven Carr. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nature biotechnology*, 24:971–83, 09 2006. doi: 10.1038/nbt1235.

Philippe Rinaudo, Samia Boudah, Christophe Junot, and Etienne A. Thévenot. biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences*, 3, June 2016. ISSN 2296-889X. doi: 10.3389/fmolb.2016.00026. URL http://journal.frontiersin.org/Article/10.3389/fmolb.2016.00026/abstract.

Lee D. Roberts, Amanda L. Souza, Robert E. Gerszten, and Clary B. Clish. Targeted metabolomics. *Current Protocols in Molecular Biology*, 98(1):30.2.1–30.2.24, 2012. doi: https://doi.org/10.1002/0471142727.mb3002s98. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb3002s98.

E Rohée, R Coulon, F Carrel, T Dautremer, E Barat, T Montagu, S Normand, and C Jammes. Benchmark of the non-parametric bayesian deconvolution method implemented in the sinbad code for x/γ rays spectra processing. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 836:91–97, 2016.

E. Rohée, R. Coulon, F. Carrel, T. Dautremer, E. Barat, T. Montagu, S. Normand, and C. Jammes. Qualitative and quantitative validation of the sinbad code on complex hpge gamma-ray spectra. In *2015 4th International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA)*, pages 1–6, 2015. doi: 10.1109/ANIMMA.2015.7465517.

Camille Roquencourt, Stanislas Grassin Delyle, and Etienne Thévenot. ptairMS: real-time processing and analysis of PTR-TOF-MS data for biomarker discovery in exhaled breath. *Bioinformatics*, January 2022. doi: 10.1093/bioinformatics/btac031. URL https://doi.org/10.1093/bioinformatics/btac031.

Andreas F. Ruckstuhl, Matthew P. Jacobson, Robert W. Field, and James A. Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, January 2001. ISSN 00224073. doi: 10.1016/S0022-4073(00)00021-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0022407300000212.

David Ruppert, M.P. Wand, and Raymond J. Carroll. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, 3(0):1193–1256, 2009. ISSN 1935-7524. doi: 10.1214/09-EJS525. URL https://projecteuclid.org/euclid.ejs/1259944245.

Dorota Ruszkiewicz, Daniel Sanders, Rachel o'Brien, Frederik Hempel, Matthew Reed, Ansgar Riepe, J. Baillie, Emma Brodrick, Kareen Darnley, Richard Ellerkmann, Oliver Mueller, Angelika Skarysz, Michael Truss, Thomas Wortelmann, Simeon Yordanov, Charles Thomas, Bernhard Schaaf, and Michael Eddleston. Diagnosis of covid-19 by analysis of breath with gas chromatography-ion mobility spectrometry: A feasibility study. *SSRN Electronic Journal*, 01 2020. doi: 10.2139/ssrn.3675407.

C G Ryan, E Clayton, W L Griffin, and S H Sie. SNIP, a statistics sensitive background treatment for the quantitative analysis of the pixe spectra in geoscience application. page 7, 1988.

Edoardo Saccenti, Huub Hoefsloot, Age Smilde, Johan Westerhuis, and Margriet Hendriks. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10, 12 2013. doi: 10.1007/s11306-013-0598-6.

Shahnorbanun Sahran, Dheeb Albashish, Azizi Abdullah, Nordashima Abd Shukor, and Suria Pauzi. Absolute cosine-based svm-rfe feature selection method for prostate histopathological grading. *Artificial Intelligence in Medicine*, 87, 04 2018. doi: 10.1016/j.artmed.2018.04.002.

Hector Sanz, Clarissa Valim, Esteban Vegas, Josep Oller, and Ferran Reverter. Svm-rfe: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, 19, 11 2018. doi: 10.1186/s12859-018-2451-4.

F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946. ISSN 00994987. URL http://www.jstor.org/stable/3002019.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. pages 515–521, 01 1998.

Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac60214a047. URL http://pubs.acs.org/doi/abs/10.1021/ac60214a047.

Fabian Scheipl, Sonja Greven, and Helmut Küchenhoff. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52(7):3283–3299, March 2008. ISSN 01679473. doi: 10.1016/j.csda.2007.10.022. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947307004306.

R. Schnabel, Rianne Fijten, Agnieszka Smolinska, Jan Dallinga, Marie-Louise Boumans, Ellen Stobberingh, Agnes Boots, Paul Roekaerts, Dennis Bergmans, and Frederik Van Schooten. Analysis of volatile organic compounds in exhaled breath to diagnose ventilator-associated pneumonia. *Scientific reports*, 5:17179, 11 2015. doi: 10.1038/srep17179.

Henny Schwoebel, Roland Schubert, Martin Sklorz, Sabine Kischkel, Ralf Zimmermann, Jochen K. Schubert, and Wolfram Miekisch. Phase-resolved real-time breath analysis during exercise by means of smart processing of PTR-MS data. *Analytical and Bioanalytical Chemistry*, 401(7): 2079–2091, October 2011. ISSN 1618-2642, 1618-2650. doi: 10.1007/s00216-011-5173-2. URL http://link.springer.com/10.1007/s00216-011-5173-2.

Kanako Sekimoto, Shao-Meng Li, Bin Yuan, Abigail Koss, Matthew Coggon, Carsten Warneke, and Joost de Gouw. Calculation of the sensitivity of proton-transfer-reaction mass spectrometry (ptr-ms) for organic trace gases using molecular properties. *International Journal of Mass Spectrometry*, 421:71–94, 2017. ISSN 1387-3806. doi: https://doi.org/10.1016/j.ijms.2017.04.006. URL https://www.sciencedirect.com/science/article/pii/S1387380616302494.

Jose A Seoane, Ian N M Day, Colin Campbell, Juan P Casas, and Tom R Gaunt. Using a Random Forest proximity measure for variable importance stratification in genotypic data. page 12, 2014.

Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, July 2017. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2017.2690524. URL http://ieeexplore.ieee.org/document/7891546/.

Wan-Chi Siu and Kwok-Wai Hung. Review of image interpolation and super-resolution. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–10, 2012.

Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78(3):779–787, February 2006. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac051437y. URL https://pubs.acs.org/doi/10.1021/ac051437y.

David Smith, Patrik Španěl, Jens Herbig, and Jonathan Beauchamp. Mass spectrometry for real-time quantitative breath analysis. *Journal of Breath Research*, 8(2):027101, March 2014. ISSN 1752-7155, 1752-7163. doi: 10.1088/1752-7155/8/2/027101. URL http://stacks.iop.org/1752-7163/8/i=2/a=027101?key=crossref.947798b7ff37f376a62a261671408ec5.

Patrik Španěl, Kseniya Dryahina, and David Smith. A quantitative study of the influence of inhaled compounds on their concentrations in exhaled breath. *Journal of breath research*, 7(1): 017106, 2013.

Daniel Stamate, Min Kim, Petroula Proitsi, Sarah Westwood, Alison Baird, Alejo Nevado-Holgado, Abdul Hye, Isabelle Bos, Stephanie Vos, Rik Vandenberghe, Charlotte Teunissen, Mara Kate, Philip Scheltens, Silvy Gabel, Karen Meersmans, Olivier Blin, Jill Richardson, Ellen Deroeck, Sebastiaan Engelborghs, and Cristina Legido-Quigley. A metabolite-based machine learning approach to diagnose alzheimer-type dementia in blood: Results from the european medical information framework for alzheimer disease biomarker discovery cohort.

Alzheimer's and Dementia: Translational Research and Clinical Interventions, 5:933–938, 12 2019. doi: 10.1016/j.trci.2019.11.001.

Aaron L. Stancik and Eric B. Brauns. A simple asymmetric lineshape for fitting infrared absorption spectra. *Vibrational Spectroscopy*, 47(1):66–69, May 2008. ISSN 09242031. doi: 10.1016/j.vibspec.2008.02.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0924203108000453.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. ISSN 00063444. URL http://www.jstor.org/stable/2331554.

Guillermo Suarez-Cuartin, Jordi Giner, José Merino, Ana Rodrigo-Troyano, Anna Feliu, Lidia Perea, Ferran Sanchez-Reus, Diego Castillo, Vicente Plaza, James Chalmers, and Oriol Sibila. Identification of pseudomonas aeruginosa and airway bacterial colonization by an electronic nose in bronchiectasis. *Respiratory Medicine*, 136, 02 2018. doi: 10.1016/j.rmed.2018.02.008.

Pritam Sukul, Phillip Trefz, Jochen Schubert, and Wolfram Miekisch. Immediate effects of breath holding maneuvers onto composition of exhaled breath. *Journal of Breath Research*, 8:037102, 09 2014. doi: 10.1088/1752-7155/8/3/037102.

Pritam Sukul, Phillip Trefz, Svend Kamysek, Jochen K Schubert, and Wolfram Miekisch. Instant effects of changing body positions on compositions of exhaled breath. *Journal of Breath Research*, 9(4):047105, November 2015. ISSN 1752-7163. doi: 10.1088/1752-7155/9/4/047105. URL http://stacks.iop.org/1752-7163/9/i=4/a=047105?key=crossref.b3a3e383ff839d9cb9e5669d99853c04.

Pritam Sukul, Jochen K. Schubert, Peter Oertel, Svend Kamysek, Khushman Taunk, Phillip Trefz, and Wolfram Miekisch. FEV manoeuvre induced changes in breath VOC compositions: an unconventional view on lung function tests. *Scientific Reports*, 6(1):28029, June 2016. ISSN 2045-2322. doi: 10.1038/srep28029. URL http://www.nature.com/articles/srep28029.

Pritam Sukul, Jochen K. Schubert, Svend Kamysek, Phillip Trefz, and Wolfram Miekisch. Applied upper-airway resistance instantly affects breath components: a unique insight into pulmonary medicine. *Journal of breath research*, 11 4:047108, 2017.

Pritam Sukul, Jochen Schubert, Phillip Trefz, and Wolfram Miekisch. Natural menstrual rhythm and oral contraception diversely affect exhaled breath compositions. *Scientific Reports*, 8:10838, 07 2018. doi: 10.1038/s41598-018-29221-z.

Pritam Sukul, Anna Richter, Jochen K. Schubert, and Wolfram Miekisch. Deficiency and absence of endogenous isoprene in adults, disqualified its putative origin. *Heliyon*, 7(1):e05922, 2021. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2021.e05922. URL https://www.sciencedirect.com/science/article/pii/S240584402100027X.

Lloyd W Sumner, Alexander Amberg, Dave Barrett, Michael H Beale, Richard Beger, Clare A Daykin, Teresa W-M Fan, Oliver Fiehn, Royston Goodacre, Julian L Griffin, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3):211–221, 2007.

Ewa Szymańska, Edoardo Saccenti, Age Smilde, and Johan Westerhuis. Double-check: Validation of diagnostic statistics for pls-da models in metabolomics studies. *Metabolomics : Official journal of the Metabolomic Society*, 8:3–16, 06 2012. doi: 10.1007/s11306-011-0330-3.

Dirk Taeger and Sonja Kuhnt. Statistical hypothesis testing with sas and r. *Statistical Hypothesis Testing with SAS and R*, 02 2014. doi: 10.1002/9781118762585.

Nele Alexandra ten Hagen, Friederike Twele, Sebastian Meller, Paula Jendrny, Claudia Schulz, Maren von Köckritz-Blickwede, Ab Osterhaus, Hans Ebbers, Isabell Pink, Tobias Welte, Michael Peter Manns, Thomas Illig, Anahita Fathi, Marylyn Martina Addo, Andreas Nitsche, Andreas Puyskens, Janine Michel, Eva Krause, Rosina Ehmann, Albrecht von Brunn, Christiane Ernst, Katrin Zwirglmaier, Roman Wölfel, Alexandra Nau, Eva Philipp, Michael Engels, Esther Schalke, and Holger Andreas Volk. Discrimination of sars-cov-2 infections from other viral respiratory infections by scent detection dogs. *Frontiers in Medicine*, 8:2245, 2021. ISSN 2296-858X. doi: 10.3389/fmed.2021.749588. URL https://www.frontiersin.org/article/10.3389/fmed.2021.749588.

Matthew Thiese, Brenden Ronna, and Ulrike Ott. P value interpretations and considerations. *Journal of Thoracic Disease*, 8:E928–E931, 09 2016. doi: 10.21037/jtd.2016.08.16.

Etienne A. Thévenot, Aurélie Roux, Ying Xu, Eric Ezan, and Christophe Junot. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and opls statistical analyses. *Journal of Proteome Research*, 14(8):3322–3335, 2015. doi: 10.1021/acs.jproteome.5b00354. URL https://doi.org/10.1021/acs.jproteome.5b00354. PMID: 26088811.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Thorsten Titzmann, Martin Graus, Markus Müller, Armin Hansel, and Alexander Ostermann. Improved peak analysis of signals based on counting systems: Illustrated for proton-transfer-reaction time-of-flight mass spectrometry. *International Journal of Mass Spectrometry*, 295(1-2): 72–77, July 2010. ISSN 13873806. doi: 10.1016/j.ijms.2010.07.009. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380610002459.

Wouter Touw, Jumamurat Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha van Hijum. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, 14, 07 2012. doi: 10.1093/bib/bbs034.

Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Chuyi Liu, Xianglilan Zhang, Baozhen Shan, Ali Ghodsi, and Ming Li. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature methods*, 16(1):63–66, 2019.

Selina Traxler, Ann-Christin Klemenz, Radost Saß, Phillip Trefz, Peter Gierschner, Beate Brock, Theresa Schwaiger, Claudia Karte, Ulrike Blohm, Charlotte Schröder, Wolfram Miekisch, and Jochen Schubert. Voc breath profile in spontaneously breathing awake swine during influenza a infection. *Scientific Reports*, 8:14857, 10 2018. doi: 10.1038/s41598-018-33061-2.

Phillip Trefz, Markus Schmidt, Peter Oertel, Juliane Obermeier, Beate Brock, Svend Kamysek, Jürgen Dunkl, Ralf Zimmermann, Jochen K. Schubert, and Wolfram Miekisch. Continuous Real

Time Breath Gas Monitoring in the Clinical Environment by Proton-Transfer-Reaction-Time-of-Flight-Mass Spectrometry. *Analytical Chemistry*, 85(21):10321–10329, November 2013. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac402298v. URL https://pubs.acs.org/doi/10.1021/ac402298v.

Phillip Trefz, Jochen K Schubert, and Wolfram Miekisch. Effects of humidity, CO $_2$ and O $_2$ on real-time quantitation of breath biomarkers by means of PTR-ToF-MS. *Journal of Breath Research*, 12(2):026016, March 2018. ISSN 1752-7163. doi: 10.1088/1752-7163/aa9eea. URL https://iopscience.iop.org/article/10.1088/1752-7163/aa9eea.

Phillip Trefz, Giovanni Pugliese, Beate Brock, Jochen K Schubert, and Wolfram Miekisch. Effects of elevated oxygen levels on VOC analysis by means of PTR-ToF-MS. *Journal of Breath Research*, 13(4):046004, July 2019a. ISSN 1752-7163. doi: 10.1088/1752-7163/ab28ec. URL https://iopscience.iop.org/article/10.1088/1752-7163/ab28ec.

Phillip Trefz, Sibylle C. Schmidt, Pritam Sukul, Jochen K. Schubert, Wolfram Miekisch, and Dagmar-Christiane Fischer. Non-Invasive Assessment of Metabolic Adaptation in Paediatric Patients Suffering from Type 1 Diabetes Mellitus. *Journal of Clinical Medicine*, 8(11):1797, October 2019b. ISSN 2077-0383. doi: 10.3390/jcm8111797. URL https://www.mdpi.com/2077-0383/8/11/1797.

Hendrik Treutler and Steffen Neumann. Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data. *Metabolites*, 6(4):37, October 2016. ISSN 2218-1989. doi: 10.3390/metabo6040037. URL http://www.mdpi.com/2218-1989/6/4/37.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 06 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520. URL https://doi.org/10.1093/bioinformatics/17.6.520.

Johan Trygg and Svante Wold. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16(3):119–128, 2002. doi: https://doi.org/10.1002/cem.695. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.695.

Johan Trygg, Elaine Holmes, and Torbjörn Lundstedt. Chemometrics in metabonomics. *Journal of proteome research*, 6 2:469–79, 2007a.

Johan Trygg, Elaine Holmes, and Torbjörn Lundstedt. Chemometrics in metabonomics. *Journal of Proteome Research*, 6(2):469–479, 2007b. doi: 10.1021/pr060594q. URL https://doi.org/10.1021/pr060594q. PMID: 17269704.

M. Unser, A. Aldroubi, and M. Eden. B-spline signal processing. I. Theory. *IEEE Transactions on Signal Processing*, 41(2):821–833, February 1993. ISSN 1053587X. doi: 10.1109/78.193220. URL http://ieeexplore.ieee.org/document/193220/.

Kim D.G. van de Kant, Joep J.B.N. van Berkel, Quirijn Jöbsis, Valéria Lima Passos, Ester M.M. Klaassen, Linda van der Sande, Onno C.P. van Schayck, Johan C. de Jongste, Frederik Jan van Schooten, Eduard Derks, Edward Dompeling, and Jan W. Dallinga. Exhaled breath profiling in diagnosing wheezy preschool children. *European Respiratory Journal*, 41(1):183–188, 2013. ISSN 0903-1936. doi: 10.1183/09031936.00122411. URL https://erj.ersjournals.com/content/41/1/183.

Sandra van den Velde, Marc Quirynen, Paul van Hee, and Daniel van Steenberghe. Differences between Alveolar Air and Mouth Air. *Analytical Chemistry*, 79(9):3425–3429, May 2007. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac062009a. URL https://pubs.acs.org/doi/10.1021/ac062009a.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Omar Vesga, Maria Agudelo, Andres Valencia-Jaramillo, Alejandro Mira-Montoya, Ivan Ossa-Ospina, Esteban Ocampo, Karl Ciuoderis, Laura Perez, Andres Cardona, Yudy Aguilar, Yuli Agudelo, Juan Hernández-Ortiz, and Jorge Osorio. Highly sensitive scent-detection of covid-19 patients in vivo by trained dogs, 06 2021.

Federico Vita, Cosimo Taiti, Antonio Pompeiano, Nadia Bazihizina, Valentina Lucarotti, Stefano Mancuso, and Amedeo Alpi. Volatile organic compounds in truffle (Tuber magnatum Pico): comparison of samples from different regions of Italy and from different seasons. *Scientific Reports*, 5(1):12629, October 2015. ISSN 2045-2322. doi: 10.1038/srep12629. URL http://www.nature.com/articles/srep12629.

Gabriel Vivo Truyols and Peter J. Schoenmakers. Automatic selection of optimal savitzky golay smoothing. *Analytical Chemistry*, 78(13):4598–4608, July 2006. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac0600196. URL https://pubs.acs.org/doi/10.1021/ac0600196.

A. Vlasenko, A. M. Macdonald, S. J. Sjostedt, and J. P. D. Abbatt. Formaldehyde measurements by proton transfer reaction – mass spectrometry (ptr-ms): correction for humidity effects. *Atmospheric Measurement Techniques*, 3(4):1055–1062, 2010. doi: 10.5194/amt-3-1055-2010. URL https://amt.copernicus.org/articles/3/1055/2010/.

Carsten warneke, C. Veen, Stefan Luxembourg, Joost de Gouw, and A. Kok. Measurements of benzene and toluene in ambient air using proton-transfer-reaction mass spectrometry: Calibration, humidity dependence, and field intercomparison. *International Journal of Mass Spectrometry*, 207:167–182, 05 2001. doi: 10.1016/S1387-3806(01)00366-9.

Andreas Wehinger, Alex Schmid, Sergei Mechtcheriakov, Maximilian Ledochowski, Christoph Grabmer, Guenther A. Gastl, and Anton Amann. Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas. *International Journal of Mass Spectrometry*, 265(1):49–59, August 2007. ISSN 13873806. doi: 10.1016/j.ijms.2007.05.012. URL https://linkinghub.elsevier.com/retrieve/pii/S1387380607002382.

Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tian-Lu Chen, and Yan Ni. Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific Reports*, 8, 01 2018. doi: 10.1038/s41598-017-19120-0.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for svms. *Advances in Neural Information Processing Systems*, 13:668–674, 01 2000.

Chalini D Wijetunge, Isaam Saeed, Berin A Boughton, Ute Roessner, and Saman K Halgamuge. A new peak detection algorithm for MALDI mass spectrometry data based on a modified Asymmetric Pseudo-Voigt model. *BMC Genomics*, 16(S12):S12, December 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S12-S12. URL https://bmcgenomics.biomedcentral.com/articles/

10.1186/1471-2164-16-S12-S12.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Michael J. Wilde, Rebecca L. Cordell, Dahlia Salman, Bo Zhao, Wadah Ibrahim, Luke Bryant, Dorota Ruszkiewicz, Amisha Singapuri, Robert C. Free, Erol A. Gaillard, Caroline Beardsmore, C.L. Paul Thomas, Chris E. Brightling, Salman Siddiqui, and Paul S. Monks. Breath analysis by two-dimensional gas chromatography with dual flame ionisation and mass spectrometric detection – Method optimisation and integration within a large-scale clinical study. *Journal of Chromatography A*, 1594:160–172, June 2019. ISSN 00219673. doi: 10.1016/j.chroma.2019.02.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0021967319301311.

Michael J. Wilde, Bo Zhao, Rebecca L. Cordell, Wadah Ibrahim, Amisha Singapuri, Neil J. Greening, Chris E. Brightling, Salman Siddiqui, Paul S. Monks, and Robert C. Free. Automating and extending comprehensive two-dimensional gas chromatography data processing by interfacing open-source and commercial software. *Analytical Chemistry*, 92(20):13953–13960, 2020. doi: 10.1021/acs.analchem.0c02844. URL https://doi.org/10.1021/acs.analchem.0c02844. PMID: 32985172.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, December 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL http://www.nature.com/articles/sdata201618.

David Wishart. Metabolomics for investigating physiological and pathophysiological processes. *Physiological reviews*, 99:1819–1875, 10 2019. doi: 10.1152/physrev.00035.2018.

S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5(3): 735–743, sep 1984. ISSN 0196-5204. doi: 10.1137/0905052. URL https://doi.org/10.1137/0905052.

Svante Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978. doi: 10.1080/00401706.1978.10489693. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1978.10489693.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, October 2001. ISSN 01697439. doi: 10.1016/S0169-7439(01)00155-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0169743901001551.

Simon N Wood. Generalized Additive Models: an introduction with R. page 397, 2006.

Ching Wu, William F. Siems, and Herbert H. Hill. Secondary electrospray ionization ion mobility spectrometry/mass spectrometry of illicit drugs. *Analytical Chemistry*, 72(2):396–403, 2000. doi: 10.1021/ac9907235. URL https://doi.org/10.1021/ac9907235. PMID: 10658336.

Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics*, 10(1), December 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-4.

Tianwei Yu and Hesen Peng. Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection. *BMC Bioinformatics*, 11(1):559, December 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-559. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-559.

Dabao Zhang, Xiaodong Huang, Fred E. Regnier, and Min Zhang. Two-dimensional correlation optimized warping algorithm for aligning gcxgc-ms data. *Analytical Chemistry*, 80(8):2664–2671, 2008. doi: 10.1021/ac7024317. URL https://doi.org/10.1021/ac7024317. PMID: 18351753.

Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *The Analyst*, 135(5):1138, 2010. ISSN 0003-2654, 1364-5528. doi: 10.1039/b922045c. URL http://xlink.rsc.org/?DOI=b922045c.

Wenzhao Zhou, Chaoqun Huang, Xue Zou, Yan Lu, Chengyin Shen, Xiping Ding, Hongzhi Wang, Haihe Jiang, and Yannan Chu. Exhaled breath online measurement for cervical cancer patients and healthy subjects by proton transfer reaction mass spectrometry. *Analytical and Bioanalytical Chemistry*, 409(23):5603–5612, September 2017. ISSN 1618-2642, 1618-2650. doi: 10.1007/s00216-017-0498-0. URL http://link.springer.com/10.1007/s00216-017-0498-0.

Tony Zitek. The appropriate use of testing for covid-19. *Western Journal of Emergency Medicine*, 21, 04 2020. doi: 10.5811/westjem.2020.4.47370.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3647580.

Patrik Španěl and David Smith. Progress in sift-ms: Breath analysis and other applications. *Mass Spectrometry Reviews*, 30(2):236–267, 2011. doi: https://doi.org/10.1002/mas.20303. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mas.20303.