



UNIVERSITAT ROVIRA I VIRGILI

IMPROVEMENT OF SAMPLE CLASSIFICATION AND METABOLITE PROFILING IN 1H-NMR BY A MACHINE LEARNING-BASED MODELLING OF SIGNAL PARAMETERS

Daniel Cañueto Rodríguez

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

Daniel Cañueto

**Improvement of sample classification and metabolite
profiling in ^1H -NMR by a machine learning-based
modelling of signal parameters**

DOCTORAL THESIS

Supervised by Dr. Nicolau Cañellas Alberich

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica (DEEEA)



UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2018



UNIVERSITAT ROVIRA I VIRGILI

Escola Tècnica Superior d'Enginyeria

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica

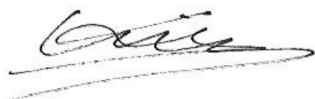
Avda. Països Catalans, 26

43007 Tarragona

I CERTIFY that the present study, entitled "**Improvement of sample classification and metabolite profiling in ¹H-NMR by a machine learning-based modelling of signal parameters**", presented by **Daniel Cañueto Rodríguez** for the award of the degree of Doctor, has been carried out under my supervision at the Department of Electronic, Electrical and Automatic Control Engineering of this university and meets the requirements to qualify for International Mention.

Tarragona, August 2018

Doctoral Thesis Supervisor



Dr. Nicolau Cañellas Alberich

Acknowledgements

The development of this PhD thesis could not have been possible without the help, collaboration and understanding of many people. The following lines are dedicated to them.

First of all, I would like to thank my supervisor Nico for all the effort he put into helping me. I have read and heard many stories of supervisors who spend little time on their PhD students. Instead, you have dedicated hours and hours, if necessary in your free time, to train me, guide me and stop me when you thought it necessary. Sometimes I followed your advice to the letter, sometimes I decided to go ahead; however, your recommendations always helped me to make better decisions. I would also like to thank the other leaders of the SIPOMICS/MILAB group for teaching me with their experience and patience. I hope that the following years will bring you many successes and scientific articles. I also do not want to forget the people from the MetaboLights group during my PhD stay in Cambridge for making me feel part of the group. Reza, once again, thank you for believing in me. Lastly, I want to thank the firm I currently work for; their understanding has greatly helped during these last months finishing this PhD thesis.

It's time to acknowledge my rear-guard buddies at the bunker now. First of all, Pep, who was the trailblazer in the research on automatic profiling and was my guide during my first months. I am so Capricorn and you are so Cancer... Nonetheless, you were always very generous in easing my acclimatization to the group and always helped me as much as possible. I hope our future careers let us work (and relax from it) regularly again. I also wanna thank the other already retired metabolomics war heroes: Xavi, Rubén & Pere. Having such PhD mates my first 2 years was a stroke of luck that I can only realise now. You're a great group of talented, helpful and cheerful researchers. If I had not had you (and Pep) as mates and models, my knowledge of programming, statistics or engineering wouldn't be the current one. It's time for Sonia, Alex and Carla: the current embattled soldiers at the bunker. I hope I did my best to help you during the months we were together, and I wish you good luck and lots of First Author in the time you have left. And, of course, I want to thank many other people from the Tarragona and Reus bunkers for the good times we had together.

Now it is time for my family. If I am not mistaken, I am the first doctor in my (both maternal and paternal) family. I am probably even the first scientist in the family. If I am the first of all, it is because all my ancestors did not have the facilities and possibilities that I have been able to enjoy. If I am able to write this section it is only thanks to you serving 12-hour 6 days a week serving tables or getting up early to clean front desks every day. Thank you, parents and sister, for believing in me. For having sacrificed so much so I was able to write this. For putting up with me these last months. We can be very different people, but I know how much I owe you and how much I

have to learn from you. I also want, of course, acknowledge my friends, who have always been there to listen and support me during my ramblings about a world which is so different from their own. I know I have been keeping you guys kind of abandoned these last few months. I promise to make it up to you.

I would also like to thank the entire scientific community who devote their efforts and free time to sharing ideas and helping others. It is incalculable how much I have learned and learned to do (and not to do) thanks to the information shared by someone thousands of miles away only because they believe in science and they want to do their bit to improve it.

And finally, thanks to you, reader. If you are reading this, there are three possibilities: you are a reviewer, you are someone checking to see if you have been included in the acknowledgements or you are a doctoral student trying to get ideas. If it is the third option, please do not hesitate to contact me; I will be happy to share and discuss everything I have learned. And if you believe in an idea, persevere. Now it is the time to take risks.

Science advances one funeral at a time.

Max Planck

Abstract

NMR is an important analytical technique to characterize the metabolites levels present in a biological mixture. Each metabolite signal in a ^1H NMR spectrum can be constructed with three parameters: the chemical shift (i.e., the location in the spectrum), the half bandwidth (i.e., the half-width at half-height) and the intensity. The area below the signal can be estimated to calculate concentrations in a process called metabolite profiling. Automatic profiling tools are based on the optimization of the combination of metabolite signals which best fits to the spectrum lineshape and the later calculation of the area below each signal fitted.

The chemical shift and the half bandwidth of a signal are determined by the chemical environment of the nucleus mediating the signal. This chemical environment is affected by the high variance in the sample and matrix properties or in the sample storage and preparation during the study of complex matrices. As a result, the chemical shift and half bandwidth (and in some cases even the relative intensity) might show relevant variability. This variability obliges to widen greatly the search space necessary to consider during the optimization of lineshape fitting. Current strategies of automatic profiling tools can reduce the search space but at the expense of restrictions in the protocol or the matrix to analyse or of inability to profile unidentified metabolites. Current improvements in NMR capabilities and protocols require a solution to model the variability in the signal parameters which can be as sample-, protocol- and matrix-independent as possible.

To achieve this objective, the strategy consisted of inferring the specific properties of each sample from the values of the same signal parameters collected during a first profiling iteration. The multicollinearity in the signal parameter information can be exploited with ML-based workflows to generate narrow and accurate predictions of the expected parameter values in a signal according to the values present in collinear signals. These predictions can be used to reduce greatly the search space during the optimization of fitting in a second profiling iteration and improve the quality of metabolite profiling. Likewise, the difference between the expected and the obtained parameter values can be analyzed to parameterize the quality of quantifications more effectively than standard methods (e.g., fitting error).

The modelling of the signal parameters was also used to enhance the discrimination of metabolomics samples. Chemical shifts encode information about sample properties (e.g., pH, ionic strength) so they can be combined with metabolite concentrations to identify more differences between kinds of samples. Lastly, metabolite identification tools based on this work were built and added to a new open-source profiling tool especially designed to handle complex matrices.

List of Abbreviations

1D	1-Dimensional
2D	2-Dimensional
API	Application Programming Interface
CPMG	Carr–Purcell–Meiboom–Gill
CSI	Chemical Shape Indicator
DL	Deep Learning
FID	Free Induction Decay
GC	Gas Chromatography
GUI	Graphical User Interface
HMDB	Human Metabolome DataBase
LC	Liquid Chromatography
ML	Machine Learning
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PI	Prediction Interval
PLS-DA	Partial Least Squares – Discriminant Analysis
ppm	parts per million
PQN	Probabilistic Quotient Normalization
RF	Random Forest
(AU)ROC	(Area Under the) Receiver Operating Characteristic
ROI	Region Of Interest
TSP	TrimethylSilylPropanoic acid

List of Publications

Cañueto, D., Gómez, J., Salek, R. M., Correig, X. & Cañellas, N. rDolphin: a GUI R package for proficient automatic profiling of 1D ¹H-NMR spectra of study datasets. *Metabolomics* **14**, 24 (2018). DOI: 10.1007/s11306-018-1319-y

Cañueto, D., Salek, R. M., Correig, X. & Cañellas, N. Improving sample classification by harnessing the potential of ¹H-NMR signal chemical shifts. *Sci. Rep.* **8**, 11886 (2018). DOI: 10.1038/s41598-018-30351-7

Cañueto, D., Navarro, M., Bulló, M., Correig, X. & Cañellas, N. Maximizing the quality of NMR automatic metabolite profiling by a machine learning-based prediction of signal parameters. *Submitted to Analytical Chemistry* (2018).

Hernández-Alonso, P., **Cañueto, D.**, Giardina, S., Salas-Salvadó, J., Cañellas, N., Correig, X. & Bulló, M. Effect of pistachio consumption on the modulation of urinary gut microbiota-related metabolites in prediabetic subjects. *J. Nutr. Biochem.* **45**, 48–53 (2017). DOI: 10.1016/j.jnutbio.2017.04.002

Spicer, R., Salek, R. M., Moreno, P., **Cañueto, D.** & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **13**, 106 (2017). DOI: 10.1007/s11306-017-1242-7

Hernández-Alonso, P., Giardina, S., **Cañueto, D.**, Salas-Salvadó, J., Cañellas, N. & Bulló, M. Changes in Plasma Metabolite Concentrations after a Low-Glycemic Index Diet Intervention. *Mol. Nutr. Food Res.* 1700975 (2018). DOI: 10.1002/mnfr.201700975

Huaman, J.W., Mego, M., Manichanh, C., Cañellas, N., **Cañueto, D.**, Seguro, H., Jansana, M., Malagelada, C., Accarino, A., Vulevic, J., Tzortzis, G., Gibson, G., Saperas, E., Guarner, F. & Azpiroz, F. Effects of Prebiotics vs a Diet Low in Fodmaps in Patients with Functional Gut Disorder. *Gastroenterology* 0, (2018) DOI: 10.1053/j.gastro.2018.06.045

Miranda, J., Simões, R., Paules, C., **Cañueto, D.**, Pardo Cea, M.A, García-Martín, M.L., Crovetto, F., Fuertes-Martin, R., Domenech, M., Gómez-Roig, M.D., Eixarch, E., Estruch, R., Hansson, S.R., Amigó, N., Cañellas, N., Crispi, F. & Gratacós, E. "Metabolic profiling and targeted lipidomics reveals a disturbed lipid profile in mothers and fetuses with intrauterine growth restriction". Accepted in Scientific Reports. (2018).

List of Congresses

Cañueto, D., Navarro, M., Yanes, O., Correig, X. & Cañellas, N. Maximizing $^1\text{H-NMR}$ Metabolic Profiling Quality Through Post-Profiling Prediction of Expected Metabolite Signals Parameters. MetaboMeeting Conference 2017, Birmingham (UK), 11-13 December 2017.

Gómez, J., **Cañueto, D.**, Brezmes, J., Salek, R.M., Correig, X. & Cañellas, N. Dolphin: Improving automatic targeted metabolite profiling using $^1\text{H-NMR}$. ISPROF 2015 - International Symposium on Profiling, Lisbon (Portugal), 21-24 September 2015.

Contents

Acknowledgements	i
Abstract	v
List of Abbreviations	vii
List of Publications	viii
List of Congresses	ix
1 Introduction	1
2 Goals	7
3 Background Information	11
3.1 Introduction to NMR information and its application to metabolomics studies	13
3.1.1 Introduction to metabolomics.....	13
3.1.2 The basic metabolomics study workflow	13
3.1.3 Analytical platforms in metabolomics studies	14
3.1.4 Parameters of metabolite signals in ¹ H-NMR spectra	15
3.1.5 Types of ¹ H-NMR quantification data to use in metabolomics studies	18
3.1.6 Additional information to metabolite concentration in NMR spectra.....	19
3.1.7 Reactivity of chemical shift to pH and ionic strength variability.....	20
3.2 ¹ H-NMR metabolite profiling	23
3.2.1 Introduction	23
3.2.2 Challenges in ¹ H NMR metabolic profiling	24
3.2.3 Introduction to automatic ¹ H-NMR metabolite profiling	29
3.2.4 Current state-of-the-art automatic profiling tools for complex matrices.....	32
3.2.5 Consequences of limitations of current state-of-the-art metabolic profiling tools	34
3.2.6 Improvement of reproducibility of profiling as a means to solve current challenges in metabolite profiling	36
3.3 Introduction to Machine Learning.....	37
3.3.1 Supervised learning	37
3.3.2 Unsupervised learning.....	38
3.3.3 Basics of the training of prediction models during supervised learning	41
3.3.4 Applications of Machine Learning in metabolomics	44
4 rDolphin: a GUI R package for proficient automatic profiling of 1D ¹H-NMR spectra of study datasets	53
4.1 Introduction	55
4.2 Methods.....	56

4.2.1	Improvement of input and output structures	57
4.2.2	Implementation in open-source code	58
4.2.3	Tools for the effective interactive visualization of spectra	58
4.2.4	Generation of matrix-specific information of suggested signals to annotate	61
4.2.5	Flexibility to correct suboptimal quantifications.....	63
4.2.6	Identification of unknown or wrongly identified signals	67
4.3	Results and Discussion.....	67
4.4	Limitations	69
4.5	Achievements	69
5	Improving sample classification by harnessing the potential of ¹H-NMR signal chemical shifts.....	73
5.1	Introduction.....	75
5.2	Materials and Methods	76
5.2.1	Datasets	76
5.2.2	Spectra pre-processing and profiling.....	77
5.2.3	Multivariate analysis	78
5.2.4	Reproducibility of study workflow	78
5.3	Results	79
5.3.1	Exploratory visualization of PCA information	79
5.3.2	Classification results	79
5.4	Discussion	82
5.4.1	Relationship between chemical shift and metabolic alkalosis/acidosis.....	82
5.4.2	Effect of class-dependent signal misalignment on fingerprinting approaches	83
5.4.3	Future directions and challenges	85
5.5	Achievements	86
5.6	Apendix	90
5.6.1	Filtering of unreliable metabolite relative concentrations and chemical shifts ...	90
5.6.2	Filtering of non-informative signal chemical shifts	90
5.6.3	Univariate tests in non-aligned and aligned fingerprint data.....	90
5.6.4	Supplementary Tables	91
5.6.5	Supplementary Figures.....	93
6	Maximizing the quality of NMR-based automatic metabolite profiling by predicting the expected metabolite signal parameters	95
6.1	Introduction.....	97
6.2	Materials and Methods	99
6.2.1	Datasets	99

6.2.2	¹ H-NMR metabolite profiling workflow	100
6.2.3	Prediction pipeline of expected signal parameter values	100
6.2.4	Evaluation of improvement in profiling data quality	103
6.3	Results	103
6.3.1	Accurate predicted values with narrow PIs which can be used to maximize profiling performance.....	103
6.3.2	High accuracy of the calculated anomaly score for detecting improvable quantifications	107
6.4	Discussion	108
6.4.1	Future directions.....	110
6.5	Achievements	110
6.6	Apendix	114
6.6.1	Workflows of MS profiling data	114
6.6.2	Values of algorithm parameters used during lineshape fitting.....	114
6.6.3	Signal-specific lineshape fitting error calculation	114
6.6.4	Results in urine dataset.....	114
6.6.5	Signal parameter prediction pipeline.....	115
7	Conclusions and Future Directions.....	119
7.1	Conclusions	121
7.2	Future directions.....	122
	List of Figures	127
	List of Tables	131

1 Introduction

Metabolomics consists of the study of small molecules called metabolites which are present in cells, tissues, organs, and biofluids from organisms.¹ The characterization of the metabolites present in a mixture enables the study of the biological pathways where these metabolites participate in, hence contributing to the characterization of the factors behind the shown phenotype.² The current standard analytical techniques to characterize the metabolites levels present in a mixture are mass spectrometry (MS) and nuclear magnetic resonance (NMR).³ NMR consists of the alignment through a radio frequency pulse of the excitable atom nuclei present in a mixture. This alignment results in a spectrum which contains a sum of Lorentzian signals distributed along the spectrum which correspond to the metabolites present in the mixture. The area below a metabolite signal is proportional to the metabolite concentration. This quantitative property renders NMR with high-throughput potential for the quantification of metabolite concentrations in the samples of metabolomics studies. From the possible excitable nuclei, ¹H, because of its abundance in nature and the reliability, remains the default choice when acquiring NMR spectra.²

The recommended choice to perform the quantification of metabolites in NMR spectra remains the profiling of metabolites through the deconvolution of the metabolite signals.⁴ Any metabolite signal has three parameters from which the signal can be constructed: the chemical shift (i.e., the location in the spectrum), the half bandwidth (i.e., the half-width at half-height) and the intensity.⁵ Theoretically, a metabolite signal should always have the same chemical shift, have a half bandwidth always perfectly collinear to the half bandwidths of the other signals and have an intensity perfectly collinear to the one of the other signals from this metabolite. These constraints ease greatly the search by optimization algorithms of the combination of signals in a spectrum which is best able to fit the spectrum lineshape. As a result, several ¹H-NMR automatic profiling tools have appeared in order to improve the quality and reduce the duration of the metabolite profiling process when compared to options such as manual profiling or fingerprinting.⁶⁻⁸

However, the chemical shift and the half bandwidth of a signal are determined by the chemical environment of the nucleus mediating the signal. This chemical environment is affected by the high variance in the sample and matrix properties (e.g., pH, ionic strength, presence of macromolecules) and in the sample storage (freezing) and preparation (e.g., thawing, buffering, dilution) prevalent in the study of complex matrices.^{9,10} Also, there are differences in the spectrum acquisition output depending on the spectrometer used. As a result, the chemical shift and half bandwidth might show high variability in complex matrices even after buffering.¹¹ These sources of variability also worsen the perfect collinearity expected in the half bandwidths of all metabolite signals. In addition, they can also distort the expected relative intensity ratios between the signals of a same metabolite.

All these variations from the expected behaviour of the signal parameters break the assumptions necessary to perform a reliable optimization of lineshape fitting and challenge the performance of NMR automatic profiling tools. Default strategies to circumvent limitations are based on the reduction of the search space during optimization by the enforcement of restrictions during sample preparation and spectrum acquisition (e.g., use of a specific type of spectrometer, adoption of a specific sample protocol), the restriction of the tool to specific matrices or the implementation of bioinformatic solutions based on empirical observation. Although these strategies have eased the implementation of automatic approaches in metabolomics studies, they do not suppose a full solution to the original problem: how to model the variability present in the signal parameters if we cannot fully parameterize the sources of variability present in the sample to be analysed or the spectrometer to be used. The development of a modelling of the signal parameters without the need of prior information is a necessary step to implement an automatic profiling which can be as sample-, protocol- and matrix-independent as possible. These needs are even more necessary when considering current sensitivity improvements in NMR as they will mean a higher number of metabolites to profile and to monitor during automatic profiling.¹²

The current lack of effective modelling of the variability in the signal parameters also hinders the harnessing of the information encoded in them. The signal parameters encode, in their patterns of variability, specific properties of the sample analysed (pH, ionic strength, temperature, the chemical environment of the proton mediating the signal). Therefore, the information present in these variability patterns might be exploited to maximize the quality and quantity of information extracted from the samples. For example:

- Although occasional uses of the information encoded in the chemical shift have already been developed,¹³ metabolomics studies do not exploit the information present in the chemical shift or the half bandwidth to try improve the discrimination between different kinds of samples related to differences in pH or ionic strength.
- The modelling of the signal parameters should also enable the analysis of differences between the expected and the obtained signal parameters values during metabolite profiling in order to detect wrong annotations and suboptimal quantifications.
- Lastly, the protons with similar chemical environment should present similar reactivity to fluctuations in the pH or the ionic strength. Therefore, the signals whose protons have similar chemical environments should show collinear variability patterns in their chemical shift and half bandwidth. This collinearity might help in the identification of unknown metabolites by the analysis of how the patterns in the parameters of their signals correlate with the ones of known metabolites.

Possible reasons to the not yet developed implementation of these strategies to exploit the benefits of the modelling of the signal parameters are:

- The need to collect all the signal parameters of several datasets from different matrices in order to develop strategies which can model the signal parameters ensuring the best balance between accuracy and generalizability.
- The lack of open source implementations of state-of-the-art algorithms which are able to handle the non-linearities and the noise present in the signal parameter datasets.

To overcome the first challenge, it is necessary to first develop an open-source tool which is able to collect and export the signal parameters values for their analysis. The development of this tool should be ideally based on the redesign of an already existent one in order to take advantage of the already developed profiling workflow. This redesign should ensure a reproducible collection of signal parameters which is flexible to any matrix in order to achieve high-quality datasets which can maximize the quality of the development of the modelling workflow. To overcome the second challenge, the current emergence of the machine learning (ML) field ensures the availability of new options and of open-source implementations of established and emerging options. These options promise a better handling of possible challenges in the modelling of the signal parameters (e.g., removal of noise, feature selection, overfitting of nonlinearities).^{14,15} As a result, it should be now possible the development of solutions to model the signal parameters in a manner which is effective and approachable for any profiling tool.

References

1. Fiehn, O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171 (2002).
2. Bharti, S. K. & Roy, R. Quantitative 1H NMR spectroscopy. *TrAC Trends Anal. Chem.* 35, 5–26 (2012).
3. Zhang, A., Sun, H., Wang, P., Han, Y. & Wang, X. Modern analytical techniques in metabolomics analysis. *Analyst* 137, 293–300 (2012).
4. Aalim M. Weljie, †,‡, Jack Newton, †, Pascal Mercier, †, Erin Carlson, † and Carolyn M. Slupsky*†, §. Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. (2006). doi:10.1021/AC060209G
5. Dona, A. C. et al. A guide to the identification of metabolites in NMR-based metabolomics/metabolomics experiments. *Comput. Struct. Biotechnol. J.* 1–19 (2016). doi:10.1016/j.csbj.2016.02.005
6. Gómez, J. et al. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Anal. Bioanal. Chem.* 406, 7967–7976 (2014).
7. Ravanbakhsh, S. et al. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* 10, e0124219 (2015).
8. Hao, J., Astle, W., De iorio, M. & Ebbels, T. M. D. Batman-an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. *Bioinformatics* 28, 2088–2090 (2012).
9. Emwas, A.-H. et al. Recommendations and Standardization of Biomarker Quantification Using NMR-based Metabolomics with Particular Focus on Urinary Analysis. *J. Proteome Res.* (2016). doi:10.1021/acs.jproteome.5b00885
10. Emwas, A.-H. et al. Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics* (2014). doi:10.1007/s11306-014-0746-7
11. Xiao, C., Hao, F., Qin, X., Wang, Y. & Tang, H. An optimized buffer system for NMR-based urinary metabolomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst* 134, 916–925 (2009).
12. Ardenkjaer-Larsen, J. H. On the present and future of dissolution-DNP. *J. Magn. Reson.* 264, 3–12 (2016).
13. Spraul, M. et al. Mixture analysis by NMR as applied to fruit juice quality control. *Magn. Reson. Chem.* 47, S130–S137 (2009).
14. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* [Hardcover]. (2013). doi:10.1007/978-1-4614-6849-3
15. Gron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, And Tensorflow: Concepts, Tools, And... Techniques to Build Intelligent Systems.* (O'Reilly Media, 2018).

2 Goals

The aim of this PhD thesis was the exploration of the use of trending ML-based techniques and of robust ML-based workflows to model and then exploit the information present in the different parameters collected for each signal during the metabolite profiling of ¹H-NMR datasets. In particular, the applications considered were the enhanced classification of samples in metabolomics studies and the enhancement of the quality of automatic profiling in ¹H-NMR datasets.

The steps to perform these goals were the next ones:

- First, the high throughput collection of accurate signal parameter information during the metabolite profiling of complex matrices. For this purpose, it was projected a redesign of the Dolphin profiling workflow. This redesign needed to ease the adaptation of the tool to the matrix properties and to identify potentially suboptimal quantifications (and to provide tools to improve them). To solve these challenges, most developed solutions should be based on ML methods. At the same time, the current emphasis in the need of reproducibility in scientific research should be enabled during the redesign and be promoted with the use of public metabolomics study datasets. The work related to this step is included in Chapter 4.
- Second, the implementation of robust ML workflows to use reliably the collected signal parameter information in the previous step to help enhance the performance of classification models during the multivariate analysis of metabolomics studies. This enhancement should be based on the addition of information about the sample properties not encoded by the metabolite concentrations (and by the signal intensities). Due to the high variability of the half bandwidth information estimated during lineshape fitting, only chemical shift information would be added to metabolite concentration information to try help increase the classification performance. The work related to this step is included in Chapter 5.
- Last, the enhancement of automatic metabolite profiling workflows thanks to the prediction of the expected signal parameter values in a dataset. This aim should be achieved thanks to the use of information previously collected from the same dataset to exploit the multicollinearity between the parameters of all signals in order to extract the information about the sources of the variability in the parameters. This prediction should be open source and implementable in any profiling tool. In addition, it should be useful for any dataset of any matrix and avoid the need of prior information about the metabolites to profile. As a result, the approach should be able to solve challenges such as the profiling of not known metabolites or matrices or the limitations caused by the lack of standardisation in the metabolomics study workflows. The work related to this step is included in Chapter 6.

3 Background Information

3.1 Introduction to NMR information and its application to metabolomics studies

The understanding of the challenges necessary to overcome during the thesis requires the enumeration of the concepts behind these challenges (e.g., the different parameters of the NMR signals, the translation of NMR signals into metabolite quantifications, the reactivity of the signal parameters to certain sources of variability). The 3.1 subsection introduces NMR and these concepts.

3.1.1 Introduction to metabolomics

Metabolomics consists of the study of small molecules (molecular weight of 50-2000 Da) called metabolites which are present in cells, tissues, organs, and biofluids from organisms.¹ The characterization of the metabolites present in these different matrices enables the study of the biological pathways where these metabolites participate in, hence contributing to the characterization of the factors behind the shown phenotype.²

In contrast to other -omics fields such as genomics or transcriptomics, the material which is characterized in metabolomics studies is much more linked to processes external to the organism that can interfere with the shown phenotype. For example, although there exist genetic biomarkers which may be associated with the onset of cardiovascular diseases, the information present in metabolomic datasets of blood samples can provide better biomarkers for the multiple environmental and behavioural factors behind the onset of these diseases.³ This tighter link with the phenotype renders metabolomics great potential as a vital contributor to emerging trends in research like precision medicine (the need for creating personalized treatments which adapt to the individual phenotype of every patient).⁴ Since its emergence in 1999, metabolomics has progressively gained importance in biomedical research, increasing its scholarly output to 3130 yearly results (with a total above 18000 results).⁵

3.1.2 The basic metabolomics study workflow

There exist multiple options to study metabolomics samples depending on the objectives of the study, on the biological matrix to study and on the resources available. Extensive literature regarding the currently ongoing study of best practices to perform metabolomics studies depending

on these multiple factors is available elsewhere.^{6,7,8,9,10,11} To guide the reader during the next sections, it is appropriate to first enumerate the standard steps of the workflow of a metabolomics study:

1. Collection of samples and storage in freezing conditions.
2. Sample preparation:
 - 2.1. Thawing of samples.
 - 2.2. Optional extraction of the desired part of the sample to analyse (for example, protein removal or selection of the lipidic/aqueous part of the sample).
 - 2.3. Mixture of the sample with compounds which enhance the quality of the acquired information of the sample. Some examples are bacterial growth inhibitors, buffer phosphate (to reduce the pH/ionic strength variability between samples) or spectrum references.
3. Acquisition of a spectrum through an analytical platform.
4. Pre-processing of the spectrum.
5. Exploratory analysis of the spectra dataset.
6. Quantification of the spectra dataset (fingerprinting, metabolite profiling).
7. Pre-processing of the quantification dataset prior to later statistical analysis (removal of outliers, imputation of missing values, scaling to give equal importance to all variables, transformation to correct typical right-tailed distributions).
8. Statistical analysis (e.g., univariate, multivariate, network analysis).
9. Biological interpretation of the results of the analysis.

3.1.3 Analytical platforms in metabolomics studies

The current standard analytical techniques to characterize the metabolites levels present in a mixture are mass spectrometry (MS) and nuclear magnetic resonance (NMR).¹² MS consists of the ionization of the metabolites and the sorting of the resulting ions based on their mass-to-charge ratio.¹³ MS is usually coupled to gas (GC) or liquid (LC) chromatography, a separation technique which enhances the separation between the metabolites present in the mixture.¹³

On the other hand, NMR consists of the alignment through a radio frequency pulse of the excitable atom nuclei present in a mixture.² The nucleus of an atom consists of neutrons and protons. Each one of these particles shows an intrinsic property called spin. If the sum of protons and neutrons is odd (e.g., ¹H, ¹³C, ¹⁹F, ³¹P, ¹⁵N), nuclei have non-zero spin and are excitable by a magnetic field.

This excitation aligns the nuclei with or against a constant magnetic field in a spectrometer. However, the weak oscillating magnetic fields (called radio frequency pulses) irradiated by the spectrometer can disrupt this alignment. Following one of these pulses, the nuclei return to equilibrium. A time domain emission signal (called a free induction decay -FID-) is recorded by the spectrometer as the nuclei relax back to equilibrium. Then, a frequency-domain spectrum from this FID is obtained through the Fourier transform.¹⁴ Each metabolite present in the mixture shows a series of characteristic signals in the acquired spectrum. The properties of these signals are mediated by the properties of their excitable nuclei and by the metabolite concentration (hence rendering NMR its quantitative potential in metabolomics studies).¹⁵

After FID acquisition, the FID needs to be pre-processed into an NMR spectrum which can be analysed. The conversion of the FID into an NMR spectrum is performed through Fast Fourier transform with specific apodization parameters. Later, the spectrum is phase corrected and the baseline is removed to enhance its quality therefore its quantitative potential.^{16,17} Lastly, all the spectra of a dataset are aligned according to a common reference. This reference can be added during sample preparation (usually the singlet of trimethylsilylpropanoic acid -TSP-) or internal (the doublet of the β -anomer of D-glucose or the singlet of formic acid in blood).¹⁵

In addition, special pulses can be performed to create a second dimension in the acquired spectrum. The second dimension can add much higher resolution or additional information to the resulting spectrum.^{18,19} This information can be provided through the combination of information of different nuclei (heteronuclear spectroscopy; e.g., HMBC, HSQC) or through the exploitation of different chemical properties of a nucleus (homonuclear spectroscopy; e.g., COSY, TOCSY, jRES).^{18,19} However, these special pulses also reduce the relationship metabolite signal and concentration. In addition, most nuclei have too low abundance in metabolite structures to provide enough sensitivity for quantitative purposes.²⁰ As a result, the quantitative potential of NMR as analytical technique is mostly reduced to 1D spectra with ¹H as nucleus (and ¹³C in lower proportion).

3.1.4 Parameters of metabolite signals in ¹H-NMR spectra

In ¹H-NMR spectra, each metabolite signal follows generally a Lorentzian function lineshape (although it sometimes has a certain Gaussian component therefore it shows a Voigt lineshape^{21,22}). This lineshape can be parameterized through its intensity, its location in a spectrum

(‘chemical shift’ in the NMR terminology) and its half width at half height (also called ‘half bandwidth’) (Figure 3-1).

- Peak height – intensity of the peak relative to the baseline (average noise)
- Peak width – width (in hertz) at half the intensity of the peak
- Line-shape – NMR peaks generally resemble a Lorentzian function
 - A – amplitude or peak height
 - $(LW_{1/2})$ – peak width at half height (Hz)
 - X_0 – peak position (Hz)

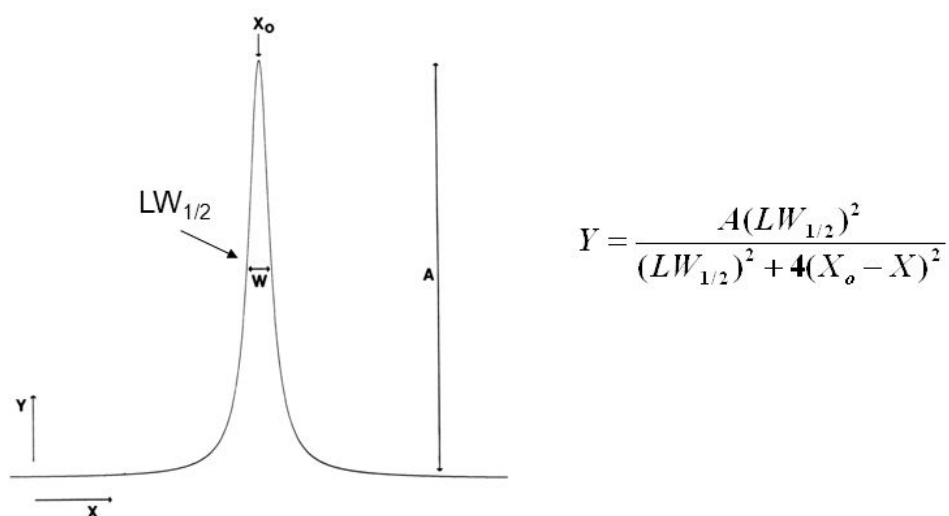


Figure 3-1 Parameters of metabolite signals in $^1\text{H-NMR}$ spectra. Extracted from ²³.

With the value of these three parameters, the Lorentzian function can be fitted according to the formula shown in Figure 3-1. The area under this signal is proportional to the concentration of the metabolite mediating this signal.

However, it is necessary to consider that the nucleus mediating the signal can be coupled to other equivalent nuclei of the metabolite chemical structure. In that case, the signal is split into a number of Lorentzian functions equal to the sum of coupled nuclei plus one (Figure 3-2). The number of peaks of these multiplets is called ‘multiplicity’ and is constant. The separation between the peaks is called ‘j-coupling’ and is generally also constant. The intensity ratio of the peaks tends to follow the Pascal’s triangle (Figure 3-3).

However, certain exceptions exist to this Pascal’s triangle structure because of complex phenomena such as second order effects. One of these relevant second-order effects is called ‘roofing’. Roofing consists of the distortion of the intensity ratio of the peaks because of the presence of

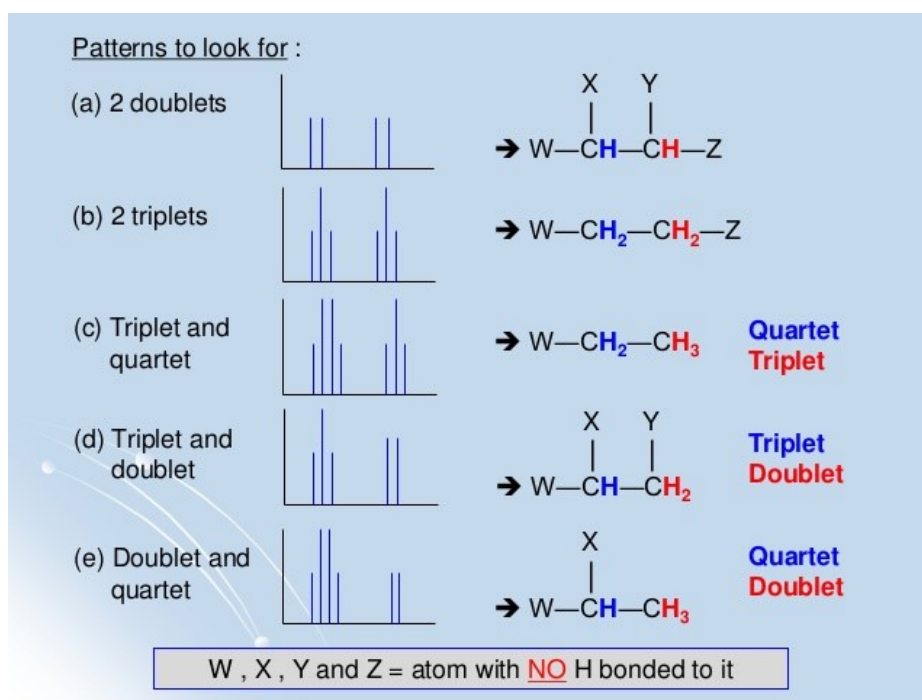


Figure 3-2 Kinds of multiplets and relationship with chemical structure. Extracted from ²⁴.

n = 0										1					singlet	
n = 1										1	1					doublet
n = 2										1	2	1				triplet
n = 3										1	3	3	1			quartet
n = 4										1	4	6	4	1		quintet
n = 5										1	5	10	10	5	1	sextet

Figure 3-3 Pascal triangle structure of the peak intensities in multiplets. Figure extracted from ²⁵.

nuclei from the same metabolite with relatively similar chemical shift.²⁶ For example, in the case of the two citric acid signals, they tend to have a distance between them of approx. 0.15 parts per million (ppm). 0.15 ppm is equivalent to 90 Hz in 600 MHz spectra. The j-coupling of both signals is approx. 16 Hz. According to Dona et al,²⁶ when there is a ratio *distance between signals / j-coupling* lower than 10, it is possible to find influences of the metabolite signals with the expected intensity ratio which create roofing in the citric acid signals (Figure 3-4):

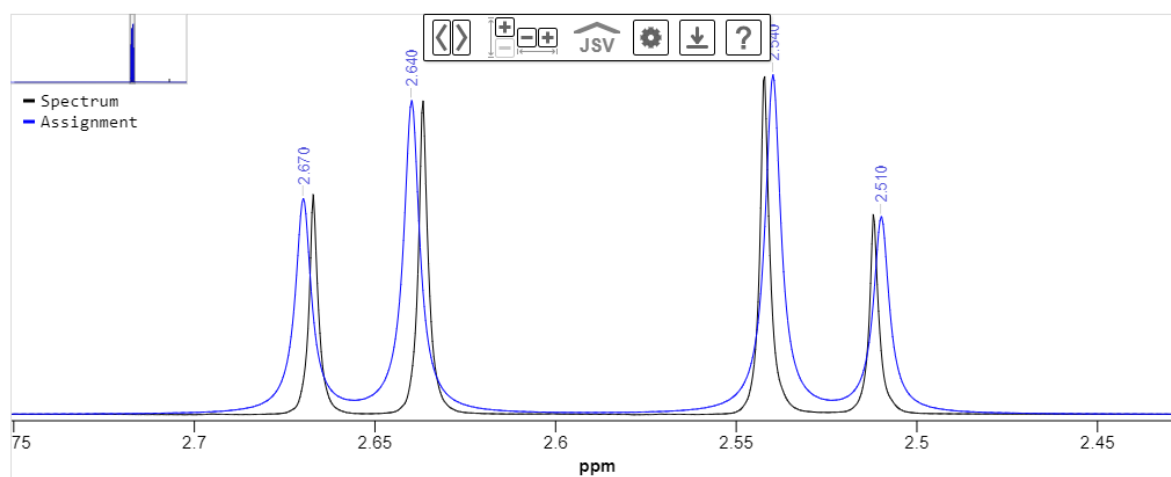


Figure 3-4 Roofing of the citric acid doublets. Extracted from ²⁷.

This roofing can be parameterized to allow the construction of the metabolite signal. However, a consequence of the relationship between the roofing and the chemical shift (in Hz) of the metabolite signals is that there is an inverse relationship between the frequency of the spectrometer and the roof effect (i.e., the lower the frequency, the lower the separation in Hz between the signals and the higher the roofing interactions).

3.1.5 Types of ¹H-NMR quantification data to use in metabolomics studies

The acquisition of pre-processed spectra in metabolomics studies enables the evaluation of the quantified metabolic information present in the samples analysed. Therefore, the similarities and differences in the metabolic state between different biological samples can be analysed. These are the two principal methods to compare the metabolic information of different samples:

- **Fingerprinting:** it consists of the analysis of samples by the intensity of bins of their NMR spectrum. These bins are obtained through the split of the spectrum into regions of constant width (e.g., 0.01 ppm) or into regions of variable width in order to minimize the split of the signal of a metabolite into different bins.²⁸ After the data analysis, the bins with relevant results can be associated to the metabolite signals typical from the spectral region and matrix analysed.
- **Profiling:** it consists of the analysis of samples by the estimated metabolite concentrations in their spectrum. In contrast to fingerprinting, annotation of metabolite signals is not posterior but prior to the quantification workflow.

The quantification of areas below the bin or the signal provides relative metabolite concentrations i.e., comparisons between the metabolite concentrations of different samples. If this relative concentration is needed to be translated into an absolute concentration, relative concentrations need to be compared to the one of a compound with known concentration which has been injected to the sample. Usually, the compound used is the same used for reference alignment. The formula to translate relative concentrations into optimal absolute concentrations (called Serkova formula²⁹) is:

$$[\textit{Metabolite}] = [\textit{Reference}] * (A_S/A_R) * (H_S/H_R)$$

where [Metabolite] indicates the metabolite absolute concentration, [Reference] indicates the reference absolute concentration, A_S indicates the area of the metabolite signal quantified, A_R indicates the area of the reference signal quantified, H_S indicates the number of protons in the metabolite signal quantified, and H_R indicates the number of protons in the reference signal quantified.

3.1.6 Additional information to metabolite concentration in NMR spectra

Metabolomic studies focus on the information about metabolite concentrations present in ¹H-NMR datasets. However, there is much more information present in the ¹H-NMR about the samples analysed. This information is provided in the previously mentioned signal parameters (chemical shift, half bandwidth, multiplicity, roof effect) and in other parameters such as heteronuclear coupling constants or the stability of lattice relaxation times (T_{1s}) and spin-spin relaxation times (T_{2s}).²⁶

Most of this information is not related to the sample properties but to the sample preparation protocol; therefore, it does not provide sample information and its use is reduced to help during metabolite identification. On the other hand, the chemical shifts, j-coupling constants (as their value is mediated by the chemical shift of each signal peak) and half bandwidths of signals are reactive to changes in the sample properties. These changes alter the metabolite chemical structure therefore they modify their behaviour towards pulses. pH, ionic strength and temperature in a sample influence the chemical shift of metabolite signals. Regarding half bandwidth, it is influenced by factors such as the paramagnetic species (e.g., dissolved oxygen gas) in the sample (however, any information present in temperature or paramagnetic species is lost during sample preparation).

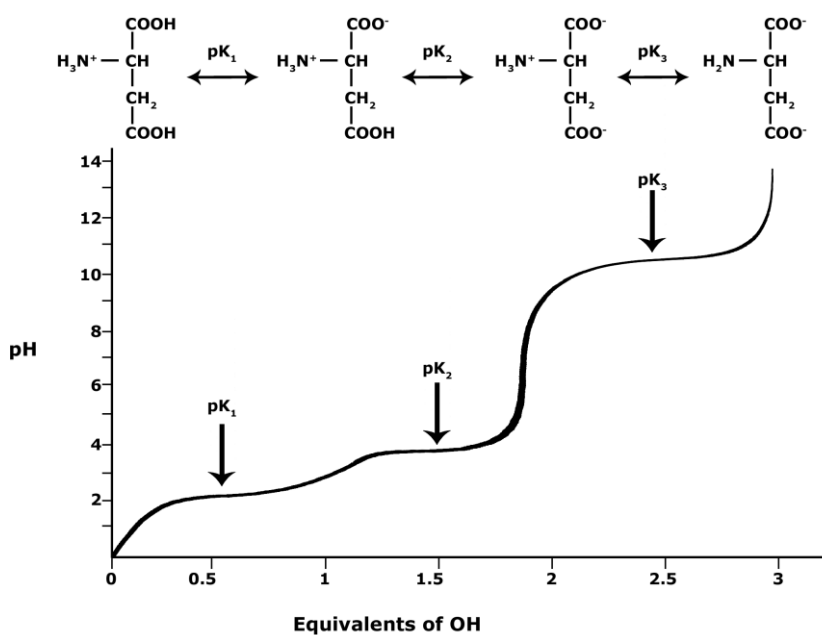


Figure 3-5 Relationship between pH decrease and deprotonation of the nuclei of functional groups. Extracted from

36

The information about pH and ionic strength given by the chemical shifts has already been proved to be beneficial for the quality control of fruit juice.³⁰ A wide range of diseases (e.g., tumours³¹) are characterized by metabolic alkalosis/acidosis³² or ionic imbalance³³; these diseases could be better identified in the NMR data with the help of chemical shift information. In addition, theoretical proof of the potential of chemical shift information to separate samples is already available.³⁴ Even so, chemical shift information is still not used to characterize these sample properties and possible differences between classes. Possible causes are the masking of pH and ionic strength by phosphate buffering and the high variability dilution in certain matrices.³⁵ As a result, the noise in the chemical shift information is increased and the use and interpretability of this information is reduced.

3.1.7 Reactivity of chemical shift to pH and ionic strength variability

The reactivity of chemical shift to pH changes can be predicted through the metabolite pKas. pKa indicates the pH where half the number of molecules of a metabolite have deprotonated (i.e., lost an H⁺) a functional group. The higher the pH the lower the percentage of molecules with a protonated functional group. In Figure 3-5, it can be observed a metabolite with three functional groups that can be deprotonated: an amino group (-NH₃⁺) and two carboxylic acids (-COOH). Each

functional group has a different pKa. As the pH increases and surpasses each pKa, each of the groups starts deprotonating until total deprotonation of the functional groups:

The deprotonation in a functional group causes changes in the chemical structure of the metabolite therefore the signal chemical shifts are altered. More concretely, the chemical shift of the signals becomes lower (Figure 3-6):³⁵ In Figure 3-6, it can also be observed how the intensity and rate of chemical shift lowering is different for every metabolite pKa. This is dependent on the metabolite structure (for example, the signals mediated by protons located in aromatic rings are more robust to pH variations) and on the distance between the proton mediating the signal and the deprotonated functional group (the higher the number of atoms between them, the lesser the chemical environment of the proton mediating the signal will be affected by the deprotonation).

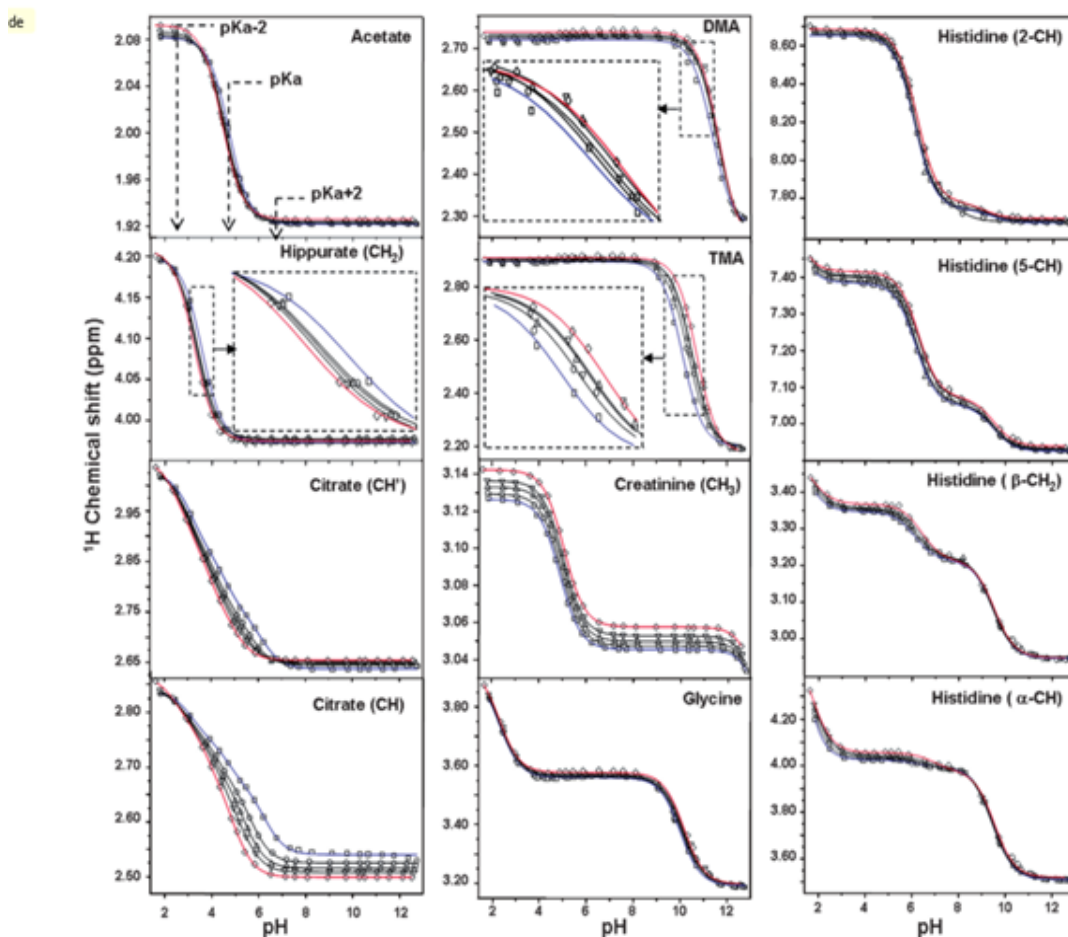


Fig. 2 The pH dependence of the ¹H NMR chemical shifts for some urinary metabolites in the model solutions containing NaCl of 0 M (□), 0.1 M (○), 0.2 M (△), 0.5 M (▽) and 1.0 M (◇); salt concentration increased from blue to red lines; the inserts showed regional expansions; open symbols were the measured data points and the solid curves were calculated data from eqn (1).

Figure 3-6 Relationship between pH decrease and chemical shift decrease. The intensity of the chemical shift decreases and of the pH range where this decrease happens is specific from very metabolite signal. Extracted from ³⁵.

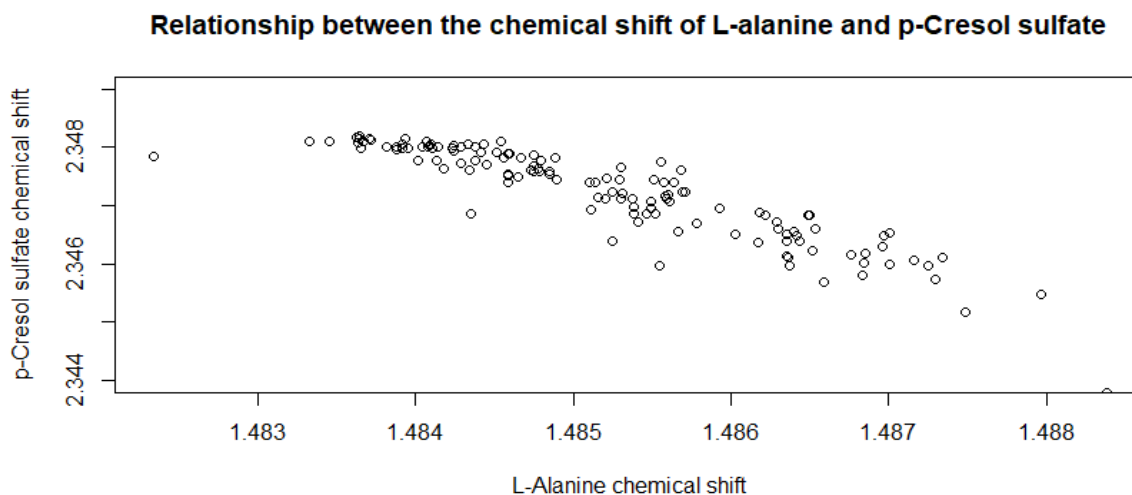


Figure 3-7 Inverse relationship between the chemical shifts of signals caused by the choice of a reference non-resistant to pH changes.

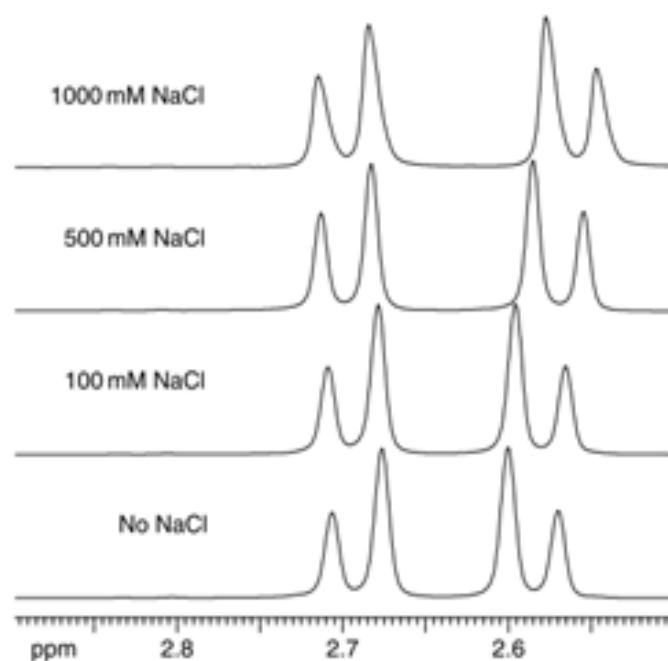


Figure 3-8 Chemical shift and lineshape changes mediated by the variability in sodium concentration present in human urine matrix. Extracted from ¹⁵.

In addition, the choice of reference alignment is a further factor in the chemical shift reactivity. The chemical shift of a signal represents the distance between the signal and the reference signal. The pKa of TSP is approximately 5, which makes its signal chemical shift sensitive to pH variation and causes signals with lower sensitivity than the TSP signal (like the ones in the phenolic

region³⁷) to seem to move in the opposite direction to other signals. For example, the inverse relationship in the human urine matrix between the chemical shift of a p-Cresol sulphate signal located at 2.345 ppm and an alanine signal located at 1.485 ppm is shown in Figure 3-7.

Regarding ionic strength, the electronegativity of ions can create changes in the chemical environment of the proton mediating the signal. The higher the concentration the greater the alteration in the chemical shift. The pattern of the change is specific of every ion. Figure 3-8 shows the different changes in the citric acid signals in the human matrix mediated by sodium.

There are several ions with the electronegativity and concentration requirements to influence the chemical shift and every ion can influence it in different ways. Consequently, the study of ionic strength needs to be much more complex and much lesser literature on the monitoring of this influence on chemical shift is available. A recent article shares the modelling of the chemical shift found for dozens of metabolites present in human urine with different pHs and ion concentrations.³⁷ However, the amount of metabolites studied through NMR can be of hundreds and there is no standardization in the choice of reference signals. For example, although most current studies use TSP as reference, most available data online (e.g., the Human Metabolome Database - HMDB-²⁷ or the Biological Magnetic Resonance Bank -BMRB-³⁸) is still based on DSS, which gives signals a 0.015 ppm higher than the one found when using TSP. These limitations hinder the evaluation and monitoring of the chemical shift to expand the information found on the samples.

3.2 ¹H-NMR metabolite profiling

The main goal of this thesis was the maximization of the quality of automatic metabolite profiling in ¹H-NMR datasets by the modelling of the parameters of the signals present in each spectrum. This section aims to help the reader understand better why the proficient automatic profiling of NMR spectra has not been yet fully accomplished (especially in complex matrices) and the current state-of-the-art of the different automatic profiling options. In addition, some context is provided about how current limitations might suppose even a bigger bottleneck to the application of future platform improvements (e.g., sensitivity, resolution) and research requirements (i.e. reproducibility).

3.2.1 Introduction

As explained in the *Types of ¹H-NMR quantification data to use in metabolomics studies* section, fingerprint approaches are based on the quantification of spectral bins. On the other hand, profiling approaches quantify metabolite concentrations. Their different workflows imply variations in the identification of metabolites and on the amount and quality of the data which is obtained.³⁹ Profiling is deemed to provide more resistance against factors such as the overlap of signals of different metabolites or the appearance of a macromolecule baseline.⁴⁰ In addition, the achievement of metabolite concentrations enables the comparison of concentrations with bibliography of the study of metabolic networks. Accordingly, although fingerprinting approaches still present benefits for the metabolomics field (ease to perform, evaluation of metabolite signals very difficult to profile), the trending approach of quantification in metabolomics studies is the metabolite profiling of spectra datasets.

There are two main approaches to perform the quantification of the area below the metabolite signals in a spectrum:

- Integration of the area below the metabolite signal.
- Deconvolution of the metabolite signal lineshape which best represents the spectrum lineshape where the signal is located.

In comparison with integration, the deconvolution of signals achieves a much better isolation of the area below the signal of interest from the areas of other signals or from the baseline (Figure 3-9). Therefore, deconvolution is the method chosen to perform the quantification of metabolite concentrations in most current profiling tools. Nevertheless, recent emergent methods such as blind source separation promise possible further improvements in signal deconvolution.⁴¹

3.2.2 Challenges in ¹H NMR metabolic profiling

3.2.2.1 Matrix-independent NMR challenges

NMR is a highly versatile, reproducible analytical technique and sample preparation does not require the destruction of the sample as in MS implementation. In addition, sample preparation and acquisition are quick to perform and with automatable potential. These properties confer the technique high-throughput potential. Nonetheless, NMR has been reduced into a secondary position compared to MS in metabolomics studies when trying to characterize the metabolome of a matrix of interest.⁴²

The reasons behind this loss of relevancy are the current low sensitivity and resolution inherent to the technique. NMR is several grades of magnitude less sensitive than MS,⁴² limiting its

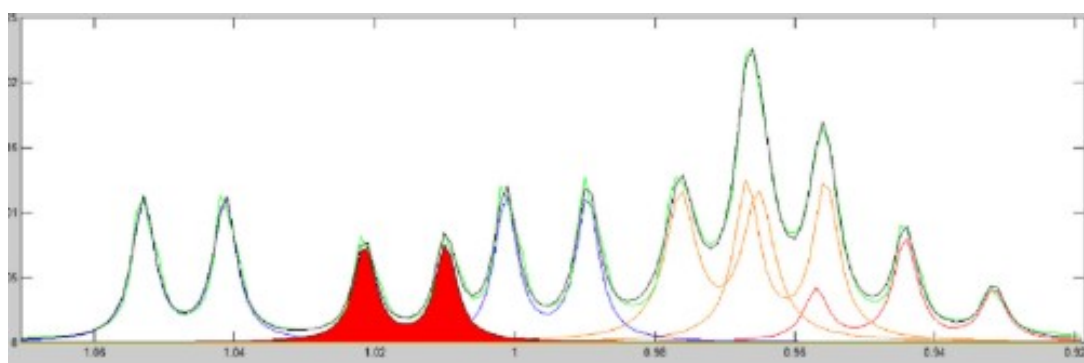


Figure 3-9 The deconvolution of signals permits the isolation of the signal of interest from the other signals. As a result, the quantification of the area below the signal is improved.

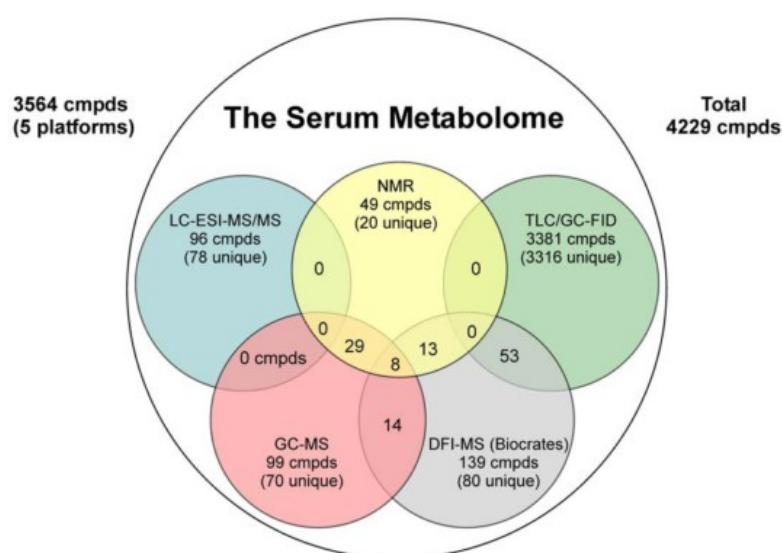


Figure 3-10 Venn diagram of the different metabolites which can be characterized with every combination of platforms. NMR provides a much lower number of metabolites than other compounds, reducing its potential to characterize the metabolome. Figure extracted from ⁴³.

capacity to characterize the metabolome of the matrix to study. For example, in human serum, each kind of MS is able to detect at least almost a hundred of compounds, a vast majority of them unique. In contrast, NMR is able to detect 49 compounds, only 20 of them unique (Figure 3-10).⁴³

Regarding resolution limitations, although a typical ¹H-NMR spectrum is able to contain hundreds of Lorentzian functions, most studies with ¹H-NMR approaches only identify and quantify dozens of metabolites. Metabolites can create different multiplets occupying different regions of the spectrum. Metabolites tend to share functional groups (methyl, amine, carboxy, phenyl) hence the metabolites tend to overcrowd regions associated with these functional groups and leave other

regions empty. As a result, low-intensity signals cannot stand out enough to be reliably annotated and/or quantified.

Promising approaches are being currently developed to increase sensitivity up to five orders of magnitude such as dynamic nuclear polarization.⁴⁴ Nonetheless, the quantitative potential of these approaches still remains a work in progress with no discernible generalization to metabolomic studies in the short or medium term.²⁰ Furthermore, the improvement in sensitivity cannot help quantify a higher amount of metabolites if their metabolite signals appear in the regions already crowded by metabolite signals with much higher concentration because of low resolution.

Regarding resolution, the analysis of other nuclei in 1D or 2D spectra can help isolate the signals according to the added metabolites-specific information. However, their analysis is limited by current sensitivity limitation to provide enough signal intensity for non-standard nuclei.²⁰ Furthermore, the need of more complex pulses creates additional costs and noise, as well as a higher complexity when quantifying volumes below the resulting signals.²⁰ These challenges are also shared by homonuclear 2D spectra. Multiple examples of quantitation of these kinds of spectra and nuclei are available in NMR literature.¹⁹ However, the quantitative potential shown in these studies has not been translated into its standardization in metabolomic studies.²⁰

An emerging promising approach called pure shift promises to enhance resolution through the conversion of multiplets into single peaks.^{45,46} This conversion does not only allow additional space for low-intensity signals to stand out but it also increases the intensity of the resulting peak. However, the technique is right now a work in progress with no clear quantitative potential for application in metabolomics studies.⁴⁵ In addition, the technique would harden the identification of signals as the characteristic shape of multiplets would be lost.

To sum up, multiple promising techniques are being developed to solve the two current most important NMR limitations in its potential to help characterize the metabolome. Nonetheless, these techniques are right now plagued with challenges which do not allow seeing its standard implementation in metabolomic studies in the short or medium term. In addition, the enhancements achieved to solve limitations will be hampered by other limitations. For example, sensitivity improvements will not translate into huge advancements in metabolite quantification if the signals of the additional metabolites cannot stand out because they overlap with much higher intensity signals.

As a result, only a holistic improvement of NMR capabilities from different perspectives will greatly improve the potential of NMR to characterize the metabolome. Meanwhile, it is assumable

that the use of 1D ^1H -NMR spectra will remain during next years as the *de facto* NMR implementation in metabolomic studies. Henceforth, it is recommendable that approaches for the quantification of metabolite concentrations in 1D ^1H -NMR spectra keep advancing. If possible, these approaches should be engineered to be translatable to the properties of future improvements in NMR.

3.2.2.2 Matrix-specific NMR limitations

Although low sensitivity and resolution are NMR limitations generalizable to all matrices, other limitations are more specific to the matrix studied. These additional limitations are generally found in matrices of complex mixtures such as human blood (serum/plasma) or urine. In the samples of these matrices there are dozens of metabolites added to different ions (Ca^{2+} , Mg^{2+} , Na^+ are the most relevant ones) and macromolecular compounds (such as protein). These matrices are tightly linked to the phenotype so they may show a high variability proportional to the inherent phenotypic variability. The chemical environment of the proton mediating the signal is altered by the sample properties; therefore, the variability in sample properties results in variability of the signal parameters in NMR spectra. As a result, a metabolite signal can have a different chemical shift or half bandwidth on each spectrum. In addition, the proton mediating every signal has its own chemical environment so it has its own reactivity patterns to this variability. This signal parameter variability challenges the accurate signal deconvolution and quantification of areas. Depending on the matrix to study, the sources of variability are different hence the approaches to solve the generated complexity in the spectra dataset need to be different. The matrix-specific added complexity represents another bottleneck in the development of the high-throughput potential of NMR.

Some relevant kinds of matrix-specific complexity are these ones:

- Baseline created by macromolecules or of broad signals mediated by lipids (Figure 3-11). Typical of serum/plasma matrices. The quantification of the area below the signal requires the accurate deconvolution of the signal from the baseline or lipid signal. This problem is usually reduced with the application of a Carr–Purcell–Meiboom–Gill (CPMG) sequence (Figure 3-11),⁴⁷ although this sequence can reduce the intensity of the signals of metabolites bound to protein.⁴⁸ Current extraction protocols are being developed in order to minimize protein presence in samples.¹¹

Characteristic ^1H NMR spectra for a T1DM patient

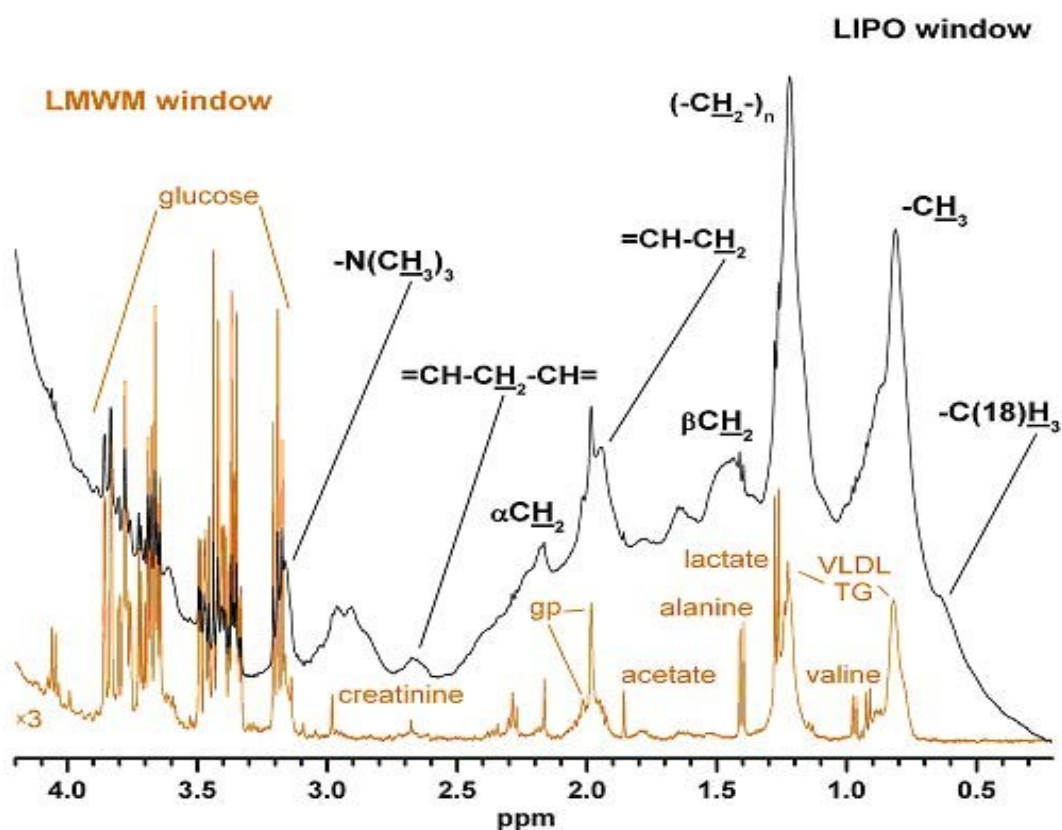


Figure 3-11 Baseline of lipids and macromolecules present in the human blood matrix. After applying CPMG sequence during spectrum acquisition, the original spectrum lineshape (black line) most baseline and broad signals are removed from the spectrum (brown line).

- Signal misalignment across a spectrum dataset. Signal misalignment is most profound in urine, where high dilution variability mediates high pH and ionic strength variability (as dilution lowers acidity and ionic strength). Several buffers or chelators have been proposed to mini-mize this variability hence reducing signal misalignment.⁴⁸ Nonetheless, with current sample preparation protocols, the typical range of the chemical shift of signals still tends to be of 0.01 ppm. As a result, the reliable annotation of signals can be compromised (especially in urine, a matrix with a high number of detectable metabolites of high concentration variability -as they come from different metabolic systems in different levels-).
- Half bandwidth variability. Related to factors such as the presence of paramagnetic species. An optimal solution to reduce this hindrance is the calculation of the half bandwidth from a chemical shape indicator (CSI; generally, the TSP signal) which can serve as a reference to estimate the expected half bandwidth of other signals.

- Broadening and area loss of the CSI. Typical in serum/plasma. It is mediated by the binding of this CSI to albumin or to the NMR tube wall. It hampers the potential of TSP as reference to estimate the absolute concentration of metabolites and to be used as CSI. The development of protein extraction protocols¹¹ or the addition of an ERETIC signal⁴⁹ are two different means to overcome this hindrance.

3.2.3 Introduction to automatic ¹H-NMR metabolite profiling

During last years, several automatic profiling tools of ¹H-NMR spectra have appeared in the metabolomics literature with the potential to be generally used in metabolomics studies.^{50,51,52,53} These tools promise to solve the time requirements and the uncertainty in metabolite identification and concentration quantification involved in manual profiling tools (which also report automatic capabilities) such as Chenomx or AMIX.⁵⁴ The mostly constant properties of the parameters of metabolite signals in a ¹H-NMR spectrum (theoretical constant chemical shift, half bandwidth proportional to a CSI, lower platform-based coefficient of variation) facilitate the possibility of the automatic deconvolution of a spectrum region as a sum of metabolite signals. Automatic profiling tools are usually based in optimization solvers which, starting from a set of starting estimates of the parameter values, find in the search space a minimum which should represent the fitting of the spectrum lineshape with lowest fitting error.

However, the mentioned sources of variability in *Challenges in NMR metabolic profiling* section force the widening of value range necessary to consider in each parameter during lineshape fitting and, therefore, the presence of a wide range of local minima where the optimization algorithm meets the completion criteria. With such a wide range of local minima, it is possible that, in a percentage of cases, the minimum at which an optimization solver stops is not the global minimum of the search space. In addition, with such complexity to monitor in the spectrum lineshape during optimization, the global minimum of the search space may not find the actual signal parameter values but the ones which can help best replicate the complex lineshape. As a result, automatic profiling may provide sometimes wrong metabolite identifications (an important bottleneck in metabolomics) and suboptimal quantifications.

Solutions developed by automatic profiling tools to handle this variability consist of the minimization of the search space during the optimization of lineshape fitting. There are two main pathways to achieve this minimization. The first one consists of the development of bioinformatics solutions based on empirical observations. The solutions are mostly based on rules that permit the

reduction of the search space necessary to consider during the optimization of lineshape fitting. Examples of these solutions are the calculation of the half bandwidth of a signal from the half bandwidth of a CSI or the simultaneous lineshape fitting of all the signals of a same metabolite based on the expected ratios between its signal intensities. Limitations of this approach are:

- It requires previous knowledge of the metabolites to profile (such as the relative intensities of their signals or the ration of their half bandwidths with the ones of the CSI).
- These solutions tend to be matrix-specific as any matrix has its own intricacies. For example, the deconvolution of signals in human blood matrices needs to minimize the incorporation of area from the macromolecule baseline or of broad lipid signals. Therefore, the accurate estimation of the intensity and the half bandwidth is a priority. When dealing with the human urine matrix, in contrast, these problems are not prevalent. On the other side, sample pH and ionic strength variability are prevalent. As a result, chemical shift may have large variability and its monitoring constitutes the priority during the automatic profiling of this matrix. This high range of bioinformatics solutions to develop depending on every matrix reduces the potential of the solutions developed.
- The assumptions which are based on can be optimal but are not fully accomplished. For example, internal exploratory analyses of two human urine datasets with different lab protocols show that general assumptions are not present even in the same matrix. More concretely, the relative intensities of the signals of a same metabolite can be different (as shown in the hippurate signals in Figure 3-12) and the relationships between the half bandwidths of signals and the ones of a CSI such as TSP (as shown in the ratio between a creatinine signal and the TSP one in Figure 3-13) show differences. It is known that signal lineshapes sometimes do not follow a strict Lorentzian lineshape: Voigt lineshapes (with a % percentage of Gaussian lineshape) might be necessary to fit when shimming variations and other possible kinds of effects appear mediated by sample properties or preparation.²² The breaking of assumptions can be not observed when doing controlled experiments based on spike-ins but appear when dealing with actual samples from complex matrices. As a result, bioinformatic solutions like the simultaneous lineshape fitting of all the signals of a same metabolite might be not robust to the breaking of these assumptions as the simultaneous lineshape fitting requires a prior strict estimation of chemical shifts, half bandwidths and relative intensities which cannot be ensured.

The other general approach is the restriction of the tool to specific matrices, sample preparation or spectrum acquisition. With this strategy, the variability in the dataset is reduced enough until the tool can handle the remaining variability. However, this approach reduces the scope of potential applications of the tool and forces researchers to implement a specific protocol. The changing

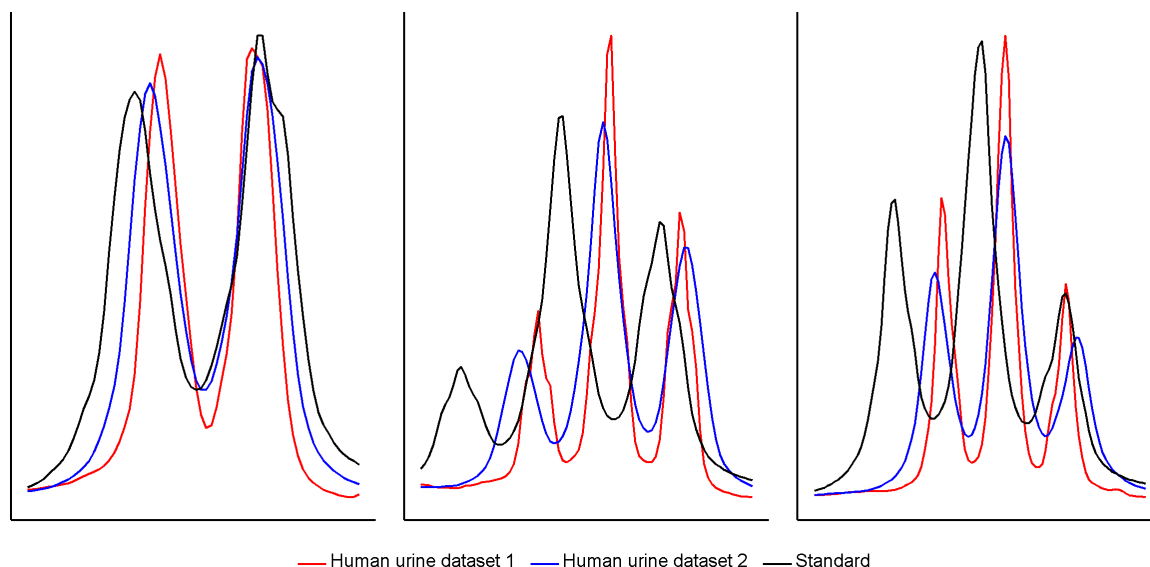


Figure 3-12 The relative intensities of signals of a same metabolite can be not constant. The three hippurate signals at the 7.85-7.5 ppm region are shown for two datasets of human urine and for the BMRB standard. After normalizing the spectra by the left signal, the other two signals show clear differences in relative intensity even when coming from the same matrix. This variability is mediated by shimming differences and possible other effects related to differences in samples properties or preparation. As a result, the simultaneous lineshape fitting of all metabolites can be compromised as the assumption of constant relative intensity is not accomplished.

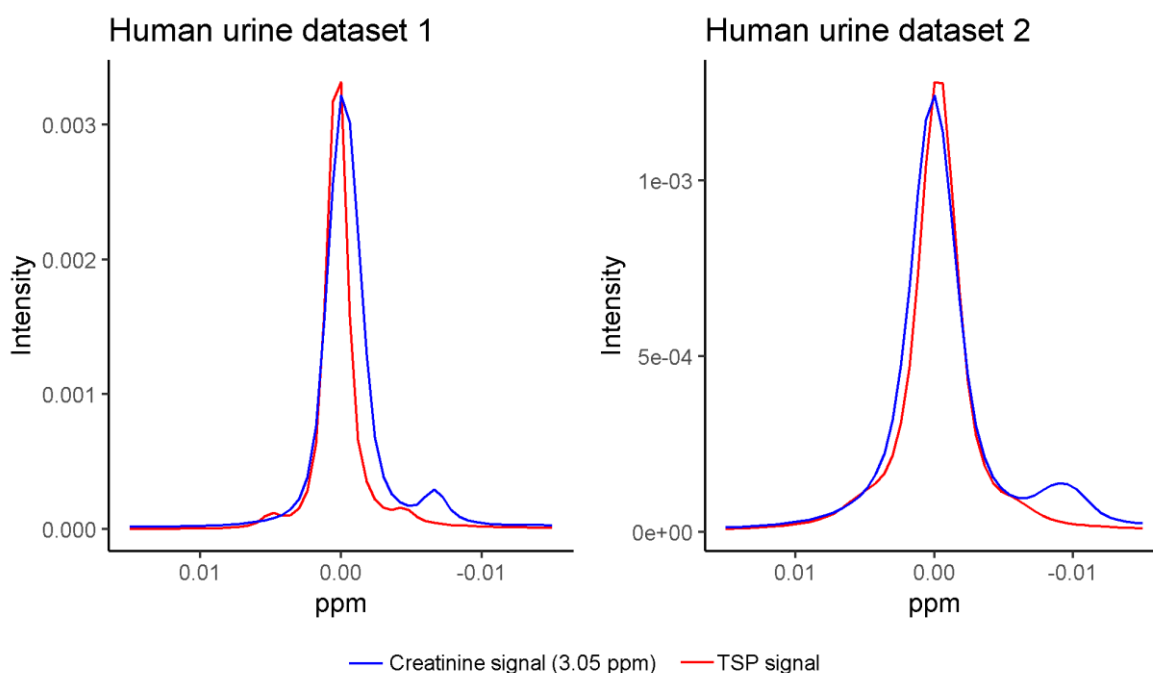


Figure 3-13 The ratio between half bandwidths of signals can be not constant. The TSP signal is used as CSI to estimate the expected half bandwidth of the rest of signals in a spectrum. However, in datasets of the same matrix (human urine), differences between the ratio of the half bandwidth of a signal such as a creatinine one and the one of the CSI signal can be observed. More concretely, on the dataset 1, the ratio creatinine/TSP is much higher than on the dataset 2. As a result, the assumption of constant ratio between half bandwidths is not accomplished and the estimation of accurate half bandwidths is compromised.

of established protocols can be hard to implement as it is necessary to perform previous investigation and to convince other researchers of the need to change already known practices. In addition, the need of specific spectrometers can be not feasible by some research groups. Lastly, the enforcement of protocols is not robust to the constant progress in the development of protocols which can maximize the quality of the datasets.^{9,11}

3.2.4 Current state-of-the-art automatic profiling tools for complex matrices

Next sections provide an enumeration of some current important tools and of some of the solutions they have developed to ensure the best balance between quality and scope during profiling.

3.2.4.1 Focus

This tool has a heavily automatic approach to detect and quantify metabolite peaks (not signals). Focus does not require the annotation of signals previous to its quantification: it automatically detects peaks which are constant across a spectra dataset and then quantifies them through area integration. In addition, it is robust to peak misalignment and to half bandwidth variability.⁵³

However, his heavily automatic approach might be prone to wrong annotations caused by high metabolite concentration variability or by signal misalignment. Likewise, integration can be more fragile than deconvolution in the case of highly overlapped signals. In addition, it cannot relate between the peak intensities of a same signal, then it cannot use this information to help perform more accurate quantifications. Lastly, its implementation in MATLAB code requires the payment of a MATLAB license.

3.2.4.2 Dolphin

The Dolphin approach consists of the lineshape fitting of signals in regions of interest (ROIs). A ROI is a spectrum region with matrix-specific signals that consistently stand out from the spectrum lineshape.⁵⁵ The split of spectra in ROIs with specific annotated signals enhances lineshape fitting quality and computing time: the lineshape to fit is much shorter and the number of parameters needed to optimize to perform effective lineshape fitting is much smaller.

The Dolphin original implementation required the use of *j*-resolved spectra in order to control the variability of certain signal parameters (e.g. chemical shift, *j*-coupling, half bandwidth).⁵² In a later implementation called Dolphin 1D, the need for this kind of 2D spectrum was avoided through the use of a GUI which enables the supervision of optimal parameter starting estimates and ranges.⁵⁶ Then, the Levenberg-Marquardt algorithm (with lower and upper bounds) calculates the combination of signal parameter values (and of baseline shape) which minimizes the lineshape fitting error of the ROI.

Dolphin, like FOCUS, presents computational and time-consuming requirements within reach of standard computers. In addition, it allows the user a much higher range of possibilities when performing quantification. The flexibility of its workflow enables its potential use even in the most complicated matrices (like human urine). Nonetheless, it also shares with FOCUS the inability to relate signals from the same metabolite and the use of MATLAB compiled (therefore, not open-source) code. In addition, it does not provide bioinformatic solutions to predict signal half bandwidths or chemical shifts for each signal. Consequently, broad parameter value ranges might be necessary during ROI lineshape fitting and optimization might easily fall into local minima and generate suboptimal resolutions.

3.2.4.3 Batman

Batman is an open-source R tool which shows the ability to relate signals from the same metabolite.⁵⁰ Therefore, it can simultaneously lineshape fit the signals of a same metabolite and take advantage of this restriction to improve the quality of lineshape fitting. In addition, it presents bioinformatic solutions to have flexibility to changes in the chemical shift of metabolite signals and can adapt to the half bandwidth of signals.

Nonetheless, its use is not recommended in the human urine matrix, requires extensive parameter tuning that cannot be performed through GUI (therefore its correct use might be limited to people experienced with programming languages) and is computationally intensive.

3.2.4.4 Bayesil

Bayesil is a web-based tool which promises the automatic profiling of spectra from human blood and CSF taking advantage of state-of-the-art ML techniques such as Bayesian methods and simulated annealing.⁵¹ Bayesil is implemented through a browser-based GUI so it avoids the need to

learn a programming language. Bayesil shares with Batman the ability to relate signals from the same metabolite and to adapt to the spectrum-specific signal half bandwidth of signals.

To guarantee an optimal performance of the tool, Bayesil has a different approach to the previous tools. It does not require the intense parameter tuning of the tool but that the research group prepares the samples and acquires the spectra in the desired way within a limited range of spectrometers in order to guarantee an optimal performance. This strategy might be at odds with the deep-ingrained protocols and practices of research labs and is not robust to the current improvements in sample preparation in the matrices analysed (only blood and cerebrospinal fluid, currently). Lastly, the tool works in a black-box manner and outputs limited information (concentration and a quality indicator).

3.2.4.5 B.I.QUANT-UR

This Bruker-based tool reports the robust automatic profiling of NMR spectra of samples of the human urine matrix, a matrix of high interest for the metabolomics community because of the ease of its sample extraction and its high volume of metabolite information.⁵⁷ In addition, it provides the interesting option to choose between several levels of profiling reliability.

To solve challenges such as signal overlap and misalignment and inter and intra-concentration variability, it uses complex ML based models.³³ However, in order to ensure these models can be used in the spectrum analysed, it requires that the research group prepares the samples and acquires the spectra in a standard way (and through Bruker spectrometers). Also, similarly to Bayesil, the necessary restrictions might be difficult to implement in labs and be not robust to future improvement in study workflows. In addition, this tool is license-based and works in a black box manner.

3.2.5 Consequences of limitations of current state-of-the-art metabolic profiling tools

Each of the described tools presents a series of strengths and weaknesses. These specific strengths and weaknesses are derived from their strategies to solve the matrix-independent and matrix-specific limitations and the changes they provoke to the signal parameters. No tool can simultaneously present flexible strategies to the properties of any matrix and show high enough

robustness. In addition, some of the current challenges will only be exacerbated when the progressing improvements in sensitivity and resolution are achieved.

The range of tools with different strengths and weaknesses enables a flexibility to choose the tool that has the best balance between the time-consuming limitations and the profiling quality requirements for each spectra dataset. Nonetheless, this flexibility benefits expert users but not newcomers to the metabolomics field. For newcomers, lack of familiarization with the nuances in NMR spectra and metabolite profiling (derived from the tight link between metabolome and phenotype) create a steep learning slope during the familiarization with the parameters to tune in these tools. Ideally, every research group may have a researcher familiarized with the different profiling tools so he would be able to implement the tool which best suits the traits of the studied dataset. Unfortunately, this option can be out of reach for most research groups. A most common solution is the use of manual profiling tools, which give a sense of security about the performed quantification, or of fingerprinting approaches.

Nonetheless, the quality of fingerprinting is compromised by the spectral complexity explained on *Matrix-specific NMR limitations* section: features can be misaligned or may have a baseline component which distorts statistical results. Furthermore, fingerprinting is more focused on finding differences between sample groups than in the actual characterization of the metabolites present in the matrix. Likewise, manual profiling is a much more time-consuming process and the sense of security can be false if the user has no expertise.⁵⁴ For example, newcomers may not accurately represent the baseline influence when fitting signals or may wrongly annotate signals to metabolites not typical from the matrix studied.

The range of strengths and weaknesses specific to each tool also explains the current lack of standardization of the metabolite profiling part of metabolomic study workflows.⁵⁸ Current lack of workflows robust to any relevant matrix greatly hardens the implementation of metabolomics approaches into the study of e.g. diseases and its integration with other -omics workflows.

To sum up, there is a current need of improvement in metabolic profiling in NMR into a standardisable approach which is robust but flexible to the different properties of every matrix and the changes created in the signal parameters. The finding of this optimal approach will contribute to the consolidation of metabolomics and to the achievements of more reproducible results.

3.2.6 Improvement of reproducibility of profiling as a means to solve current challenges in metabolite profiling

Reproducibility of study findings is one of the biggest challenges in science research. A 2016 survey by Nature shows that 90% of researchers are concerned about reproducibility in science.⁵⁹ According to this survey, more than 70% of scientists have failed to reproduce the results of previous experiments and more than half have failed to reproduce even the ones of their own experiments. Generation of non-reproducible information involves the wastage of an enormous amount of public funds ineffectively invested in hypotheses by this non-reproducible information.⁶⁰ Multiple approaches have been developed to enhance the reproducibility of studies.^{61,62} However, “publish or perish” incentives make researchers prone to selective implementation of these approaches.

In the context of metabolomics studies, the high influence of the phenotype in metabolism produces high inter- and intra-variability in the estimated metabolite concentrations. This variability needs to be added to the complex challenges during the profiling of study datasets. In order to monitor all this complexity, various techniques need to be applied to samples (e.g., buffering of samples), to spectra (e.g., apodization, binning) or to metabolite profiling data (e.g., removal of inaccurate values and imputation, data normalization and transformation, removal of samples with not enough data quality). Multiple options exist to perform each one of these techniques, and the emergence of this -omics field implies current lack of standardization in the study workflows.

Current initiatives to improve the quality of metabolomics research are the standardisation of metabolomic study workflows^{9,58,63} or the reproducibility of metabolomics studies through the access to the study data in public repositories.^{64,65} In the latter case, nonetheless, the process performed in intermediate steps of the workflow (e.g., metabolite identification, metabolite profiling, exploratory analyses to choose data transformation techniques) cannot be evaluated so there is flexibility to generate an outcome which best aligns with the researcher’s incentives. The generation of an open data standard for the description, storage and language of NMR data seems a promising step in this direction.⁶⁶ However, this language would be only able to reproduce the inputs and outputs of metabolite profiling but not an interactive evaluation (and optimization, if necessary) of this workflow. The generation of solutions to review the process of metabolite identification and choice of metabolite quantification performed in a study might greatly help reduce the current challenges regarding correct metabolite identification and variability in quantification.^{54,67}

3.3 Introduction to Machine Learning

Machine learning (ML) is an application of artificial intelligence which builds algorithms that can perform statistical analysis to automatically learn from previous experience without explicit programming. Multiple applications of ML can be seen in metabolomics literature, from its use in multivariate analysis to complement statistical modelling techniques to its implementation to improve metabolite profiling or metabolic networks.⁶⁸

The implementation of ML solutions to achieve the goals of this thesis required the knowledge and familiarization with the two main approaches associated to ML (supervised learning and unsupervised learning), the nuances of the different choices of algorithms and the requirements to apply during the implementation of ML-based prediction models. Next subsections introduce these different themes in order to understand better the choices made during the implementation of ML-based workflows in next chapters.

3.3.1 Supervised learning

In supervised learning, from a series of inputs (called features), the algorithm tries to create a model able to predict an output (called target). The training of the model is based on the evaluation of the rules and combinations of features which are best able to reproduce the correctly labelled target of each observation. This model will be then used to correctly predict unlabelled target observations. Variations of this approach are semi-supervised learning, where only a percentage of the target to use during the training of the model is labelled, and reinforcement learning, in which the features work as feedback to orient the algorithm in a dynamic environment in the shape of rewards and punishments.

The most important applications of supervised learning are the prediction of the class of a discrete target (classification) or of the value of a continuous target (regression). Depending on the aim (classification or regression), the size of the training dataset of features, the computing restrictions and the interpretability of results, the choice of algorithm will need to be different:

- If size and computing assets do not suppose any limitation, trending deep learning (DL) approaches tend to show currently the best performance (Figure 3-14). DL models are based on a set of multiple layers in which every layer performs the nonlinear processing of information coming from the previous layer to progressively extract and transform the

information given by the inputs. In addition, DL approaches have supposed the efficient use of new kinds of inputs such as images or texts.

- If the size of the training dataset or the computing requirements are limited (as it usually happens in the metabolomics field) or it is required the interpretability of the most important factors explaining the results, DL approaches may not provide improvements in performance compared to more traditional forms of ML (linear models, tree-based algorithms). Improvements in dataset size limitations (e.g., data augmentation⁶⁹), computing limitations (e.g., transfer learning⁷⁰) and interpretability⁷¹ are currently being developed in the flourishing DL field. However, even if they can be required to perform ML in images or texts, they are not necessary to apply in tabular numeric and categorical data such as the one typical in metabolomics studies.
- When choosing between conventional learning methods, linear models are much less computationally intensive.⁷⁴ In contrast, tree based models can handle much better the presence of non-linearities and do not require pre-processing (centring, scaling of features) so they generally provide better performance than linear models.⁷⁵ On the other hand, they are more computationally intensive (there is an exponential relationship between the size/dimensionality of the dataset and the computing requirements). A recent survey compared the performance of several conventional learning algorithms in 165 datasets.⁷³ This survey showed that the current best conventional machine learning algorithms are tree-based ones (gradient tree boosting, random forest -RF-) (Figure 3-15).

In the context of metabolomics datasets, where it is common to find low sample sizes, high number of features (with a high percentage of correlated or non-informative features) and non-linearities typical from biological data), the application of tree-based algorithms seems possibly the preferable standard to choose. In the *Basics of the training of prediction models during supervised learning* section, a detailed description of the factors to control during supervised learning is enumerated.

3.3.2 Unsupervised learning

In unsupervised learning, there is no target from the one to try to learn rules or associations. In contrast, learning is based on the analysis of the structure of the features to try to uncover patterns which can provide further information. The two most common approaches to unsupervised learning in metabolomics are:

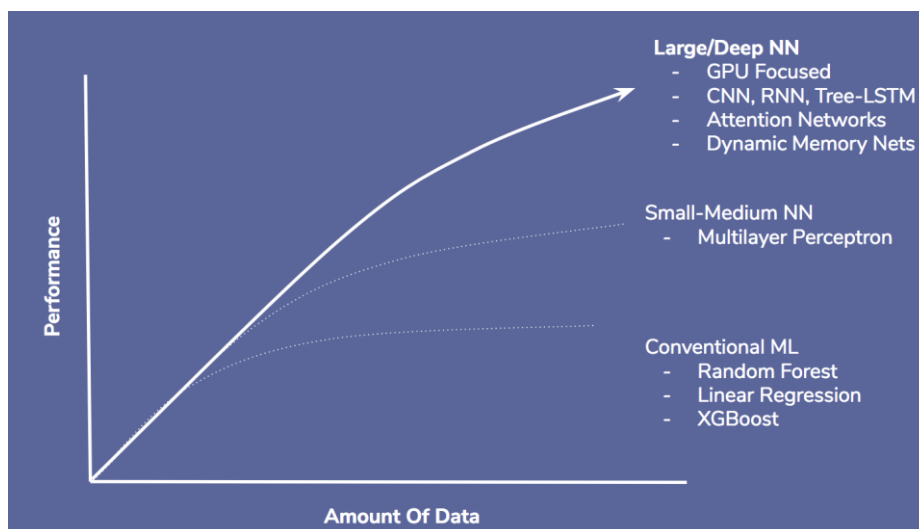


Figure 3-14 As the amount of data increases, the performance of DL approaches trumps the one of traditional ML techniques. Figure extracted from ⁷².

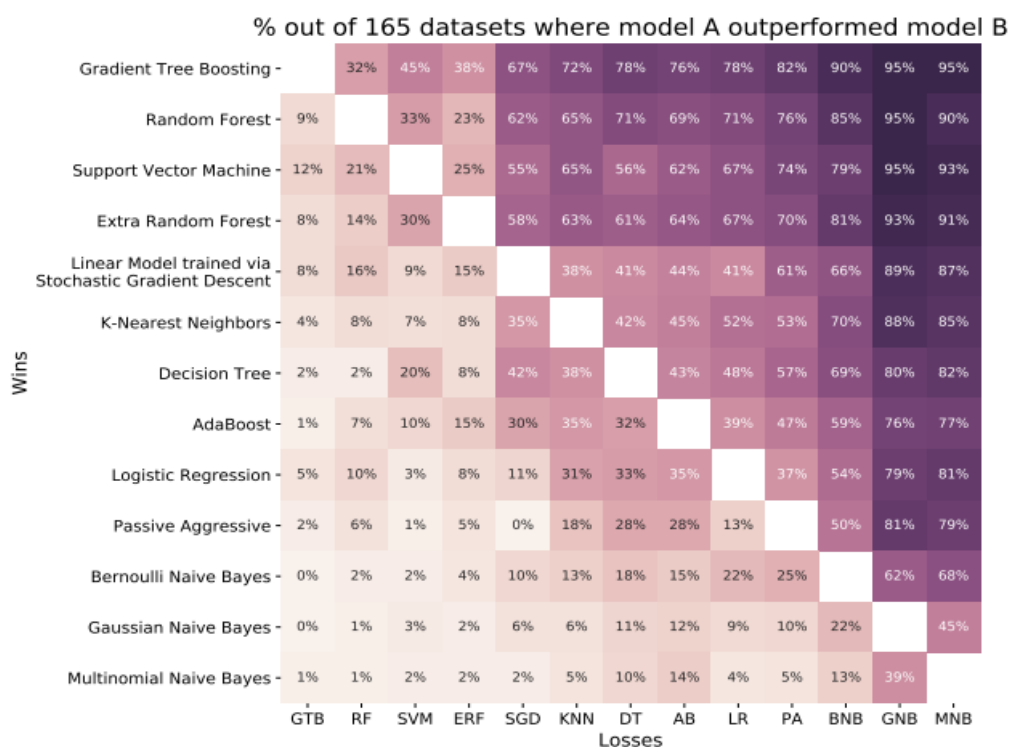
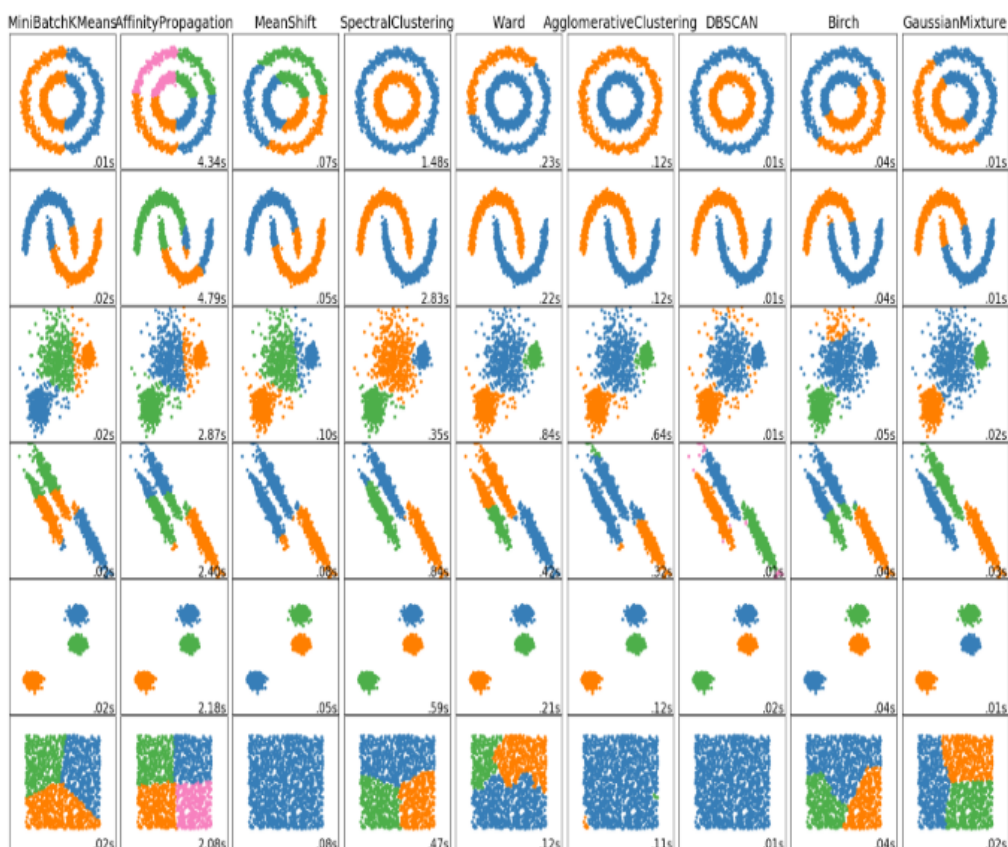


Fig. 2. Heat map showing the percentage out of 165 datasets a given algorithm outperforms another algorithm in terms of best accuracy on a problem. The algorithms are ordered from top to bottom based on their overall performance on all problems. Two algorithms are considered to have the same performance on a problem if they achieved an accuracy within 1% of each other.

Figure 3-15 Tree-based algorithms showed the best performance in the evaluation of different traditional ML algorithms in 165 different datasets. Figure extracted from ⁷³.



A comparison of the clustering algorithms in scikit-learn

Figure 3-16 Comparison (in accuracy and speed) of different clustering algorithms when dealing with different data patterns. Figure extracted from ⁷⁶.

- The clustering of observations into different groups where observations are similar within the groups and different from observations of other groups. Generally, this clustering is based on the analysis of differences between observations and the optimization of the grouping of observations until finding one that creates the most homogeneous clusters. Multiple algorithms to perform clustering exist, each one with its advantages and disadvantages when dealing with the patterns present in the data. Probably, the most important factor is the pre-specification (e.g., k-means) or not (e.g., affinity propagation) of the number of clusters to identify. Figure 3-16 shows the evaluation of the performance (in accuracy and speed) of different clustering algorithms in datasets with different data patterns. The ‘scikit-learn’ Python module documentation provides extensive information about the right choice of clustering algorithm to implement depending on the data properties.⁷⁶
- Blind source separation. Most features in metabolomics datasets share information (therefore presenting correlations between features, for example) or contain non-relevant information. In blind source separation, the algorithm tries to reorganize features to keep the

relevant information in a much smaller set of features. Benefits of this kind of dimensionality reduction in metabolomics can be the generation of uncorrelated and/or less noisy features which can enhance the performance and interpretability of supervised learning approaches, the reduction of computing requirements during supervised learning (and, therefore, the possible choice of more computing-intensive algorithms) and the easing of the exploratory analysis of the dataset (e.g., through the visualization of 2D or 3D representations of the original datasets).

The most common algorithm is principal components analysis (PCA) and its derivations (e.g., robust PCA, sparse PCA, kernel PCA). In PCA, the set of possibly correlated features are converted into uncorrelated features called principal components (PCs).⁷⁴ This conversion is performed in a way that the first PC has the largest possible variance of the dataset, the second PC has the largest possible remaining variance and so on. Related approaches are independent components analysis (ICA) or the trending t-Distributed Stochastic Neighbour Embedding (t-SNE).⁷⁷ t-SNE provides higher performance than traditional methods; however, its high flexibility makes it prone to provide the overfit results to the desires of the user.⁷⁸

3.3.3 Basics of the training of prediction models during supervised learning

To perform the prediction of a numeric or categorical target through a supervised learning approach, it needs to be evaluated multiple factors which can interfere in the quality/cost (in computing time) balance during model training. Some examples of these factors are the target and feature types (numeric, categorical), the size of the inputs, and the presence of nonlinearities, non-normalities or missing values in the dataset. Depending on these factors, the approach to performing the training of prediction models will need to be modified in order to best adapt to the particularities of the dataset.

The choice of the optimal approach to performing the prediction of the target is typically based on a process of exploratory data analysis. During this analysis, the different factors to analyse are evaluated through the summary of valuable information or the visualization of the interactions between features. Examples of findings that can be achieved during the exploratory analysis to maximize the quality of ML implementations are:

- A RF-based imputation of the missing values in a dataset may be of higher quality than a median-based one, but it may be too computing intensive when the dataset has high

dimensionality. The analysis of the dimensionality of the dataset guides the algorithms chosen during prediction or in the previous steps of the process.

- It may be necessary to perform data transformation in datasets of biological origin, in which ones it is common to find right-tailed distributions. If not performed, some algorithms (e.g., linear models) may be distorted by the extreme values of the distribution. The changes applied to the dataset to enhance prediction performance are called feature engineering. Some typical examples of this kind of feature engineering are: normalization, transformation, and scaling.
- Feature engineering can consist also of the addition of new features of higher quality from the combination of the information of several features. For example, a PCA can concentrate the information the noisy information of multiple correlated features in a first principal component with low noise as the random variance has been relegated to later principal components.
- There can be a high number of features in the feature dataset without any relationship with the target. These features will add noise during the prediction of the target thus lowering the performance. Feature selection before prediction minimizes the noise added during prediction.

The appearance of open-source tools to apply complex algorithms (e.g., RF, neural networks) in trending programming languages (e.g., R, Python) enables the generation of more accurate target predictions through the complex combination of the information of many features. The complexity of the models generated can be tweaked to further maximize the quality of the predictions generated. This step is iteratively performed through a process called hyperparameter tuning. In this process, the parameters of algorithms are tuned to adapt them to the particular patterns of the data studied and therefore increase the performance of prediction. For example, in the case of RF, some hyperparameters that can be optimized are the number of trees in the forest, the maximum number of features considered for splitting a node or the minimum number of data points placed in a node before the node is split.

However, the trained complex prediction models, in order to be effective in new data, have to be generalizable, i.e., the models should only learn information about the target and the features studied and should not learn information which is particular to the dataset where the model was trained. Learned non-generalizable information, when used in new data, will only add noise to the predictions thus lowering prediction performance. The explained process is called overfitting. These are several examples of techniques implemented to avoid overfitting: ^{74,75}

- **Regularization:** it consists of applying a cost to the heightening of the complexity in the model created until finding the right balance between quality and complexity of the model. Standard methods of regularization in linear models are LASSO and RIDGE. In the case of tree-based models, regularization is performed through the application of limits in the splitting of nodes or in depth of trees. In the case of DL approaches, L1 and L2 regulation, dropout of nodes, early stopping and data augmentation are common choices.
- **Training-validation split:** consists of the split of the dataset used during model training in a training subset and a validation subset, where improvements in the model in the training subset are validated in the validation subset. This split enables checking how the addition of non-generalizable complex information during the training of the model can worsen the prediction in another subset. To enhance the capabilities of this strategy against overfitting, validation can be performed in several iterations of different random splits (k-fold cross-validation). K-fold cross-validation also enables the generation of prediction intervals (PIs) which inform more accurately about the uncertainty present in the prediction.
- **Bootstrap resampling:** it is a similar strategy to the one applied in cross-validation. However, in each iteration of bootstrap resampling, the dataset is not split into a training and a validation subset, but a subset of observations is removed or replaced by other observations of the same dataset.
- **Representativeness of the training dataset:** another approach to achieve generalizable models consists of ensuring that the dataset used during model training is as similar as possible to the different datasets where the model will be implemented. To ensure this similarity:
 - The quality of the training dataset can be enriched to avoid the overfitting of spurious information. An example of this enrichment is the removal of noise in the dataset through the analysis of dataset values which behave as outliers.
 - The use of prediction models can be restricted to datasets which reproduce the particularities of the dataset where models were trained. For example, several machine-learning-based approaches in NMR only ensure their effectiveness when restrictions in sample preparation and in spectrum acquisition and pre-processing are applied in the dataset in which their models will be applied to (e.g., Bayesil, BI-Quant-Ur).

To evaluate the quality of the trained model, the quality of its predictions is evaluated in a different subset called test set. In the case of regression, evaluation is generally performed through indicators which parameterize the residuals between the predicted target values and the real target values. In the case of classification, the performance of classification is evaluated through indicators which analyse the percentage of true positives and negatives and of false positives and negatives.

Some of these indicators are classification accuracy, Cohen's kappa (a more robust indicator against chance classification and class imbalance) or the receiver operating characteristic (ROC) curve.⁷⁹

The high number of factors to control during the training of machine learning models implies that multiple different metric values can be achieved, and there can be small differences between these values. Robust comparisons between the different models can be performed thanks to the split of the dataset into different subsets performed during k-fold cross-validation or bootstrap resampling. The metric values achieved in each subset can be compared between different models through hypothesis testing. The most recommended to minimize possible limitations is the K-fold cross-validated paired t-test procedure.⁸⁰

3.3.4 Applications of Machine Learning in metabolomics

3.3.4.1 Data analysis in metabolomics studies

After quantifying metabolite concentrations, the collected concentrations can be analysed to identify differences between sample groups which provide valuable inferences about the condition studied in the metabolomics study. However, metabolites are not isolated but are components of complex metabolic networks and they have limited ability to factor the different important confounders (e.g. age, gender, BMI, diseases, time of sample extraction, sample protocol etc) in metabolomics studies.^{81,82}

In order to disentangle the effect of multiple factors and analyse their combined effect into the observed phenotype, some statistical modelling methods can be applied in metabolomics studies (e.g., the analysis of ratios of metabolite concentrations,⁸³ multivariate methods of hypothesis testing). However, the challenges in hypothesis testing to adequately control for several assumptions (normality, homogeneity of variance, control for interactions) make ML models an interesting alternative to apply in metabolomics studies. In comparison to statistical modelling techniques, ML techniques do not try to demonstrate or refute a hypothesis (e.g., this metabolite is related or not to the phenomenon to study) but try to provide a model which provides an optimal performance of the prediction of an outcome (e.g., the classification between case and condition samples).¹

¹ However, it should be considered that the first aim of ML methods is not the inference of information from the data analysis but the performance in the prediction of an outcome. I.e., the analysis of important features of a ML model

The most common method of multivariate analysis in metabolomics analysis is PLS-DA.⁸⁴ However, the limitations of this algorithm (e.g., overfitting and overoptimistic scores plots) have been extensively documented and recommendations to use other algorithms for metabolomics studies such as support vector machine or RF have been proposed.^{85,86} Recently, there have appeared various attempts to perform DL-based methods in the data analysis of metabolomics studies despite the theoretical challenges in metabolomics datasets (e.g., low dataset size) to take advantage of its potential in comparison to conventional learning methods.^{87,88}

Lastly, it is necessary to remember that previously enumerated unsupervised ML unsupervised methods (e.g. PCA, feature selection algorithms) are commonly used in metabolomics studies to help enhance the visualization of information about the dataset or to discard non-informative features.

3.3.4.2 Improvement of data pre-processing

ML methods are also applied to help during the pre-processing of metabolomics datasets prior to data analysis to enhance the quality of the metabolite concentrations collected.⁶⁸ As shown during the *Unsupervised learning* section, the use of ML algorithms to help pre-process the datasets is a constant when it is necessary to handle the nuances of metabolomics datasets (e.g., tabular data with more features than spectra and with lots of correlated or non-relevant features).

Other examples of ML applications in MS and NMR are the next ones:

- To ensure a reliable peak annotation for the accurate identification and quantification of metabolites.⁸⁹
- To enable the data integration of metabolomics datasets with the datasets from different -omics.⁹⁰
- To foster the analysis of metabolite pathways (and of their dynamics).⁹¹

In the case of automatic profiling in NMR spectra, some examples of ML approaches already implemented to help during automatic profiling are the Bayesian methods used by BAYESIL and

should be a means to provide interpretability about the model and not an end. ML methods have several limitations in sample size, signal/noise ratio or objective which might recommend the use of statistical modelling techniques in order to reduce the introduction of non-reproducible knowledge into the metabolomics literature.⁹²

BATMAN^{50,51} to guide the deconvolution of signals or the complex regression models implemented to predict chemical shifts by BI QUANT UR.³³

References

1. Fiehn, O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
2. Bharti, S. K. & Roy, R. Quantitative 1H NMR spectroscopy. *TrAC - Trends Anal. Chem.* **35**, 5–26 (2012).
3. Rhee, E. P. & Gerszten, R. E. Metabolomics and Cardiovascular Biomarker Discovery. *Clin. Chem.* **58**, 139–147 (2012).
4. Beger, R. D. *et al.* Metabolomics enables precision medicine: “A White Paper, Community Perspective”. *Metabolomics* **12**, 149 (2016).
5. Kell, D. B. & Oliver, S. G. The metabolome 18 years on: a concept comes of age. *Metabolomics* **12**, 148 (2016).
6. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Front. Bioeng. Biotechnol.* **3**, 23 (2015).
7. Bjerrum, J. T. *Metabonomics*. **1277**, (Springer New York, 2015).
8. Lindon, J. C., Nicholson, J. K. & Holmes, E. *The handbook of metabonomics and metabolomics*. (Elsevier, 2007).
9. Emwas, A.-H. *et al.* Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics* (2014). doi:10.1007/s11306-014-0746-7
10. Yen, S. *et al.* Metabolomic analysis of human fecal microbiota: A comparison of feces-derived communities and defined mixed communities. *J. Proteome Res.* **14**, 1472–1482 (2015).
11. Nagana Gowda, G. A., Gowda, Y. N. & Raftery, D. Expanding the Limits of Human Blood Metabolite Quantitation Using NMR Spectroscopy. *Anal. Chem.* **87**, 706–715 (2015).
12. Zhang, A., Sun, H., Wang, P., Han, Y. & Wang, X. Modern analytical techniques in metabolomics analysis. *Analyst* **137**, 293–300 (2012).
13. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* **26**, 51–78 (2007).
14. Keeler, J. *Understanding NMR spectroscopy*. (John Wiley and Sons, 2010).
15. Lindon, J. C., Nicholson, J. K. & Holmes, E. *The Handbook of Metabonomics and Metabolomics*. *The Handbook of Metabonomics and Metabolomics* (2007). doi:10.1016/B978-0-444-52841-4.X5000-0
16. Smolinska, A., Blanchet, L., Buydens, L. M. C. & Wijmenga, S. S. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Anal. Chim. Acta* **750**, 82–97 (2012).

17. Vettukattil, R. Preprocessing of Raw Metabonomic Data. in *Methods in molecular biology (Clifton, N.J.)* **1277**, 123–136 (2015).
18. Bingol, K., High, N. & Field, M. Multidimensional approaches to NMR-based metabolomics *Kerem*. **86**, 47–57 (2015).
19. Giraudeau, P. Quantitative 2D liquid-state NMR. *Magn. Reson. Chem.* (2014). doi:10.1002/mrc.4068
20. Giraudeau, P. Challenges and perspectives in quantitative NMR. *Magn. Reson. Chem.* (2016). doi:10.1002/mrc.4475
21. Markley, J. L. *et al.* The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **43**, 34–40 (2017).
22. Marshall, I., Higinbotham, J., Bruce, S. & Freise, A. Use of Voigt lineshape for quantification of in vivo 1H spectra. *Magn. Reson. Med.* **37**, 651–7 (1997).
23. Watts, M. Obtaining an NMR Spectra. Available at: <http://slideplayer.com/slide/4898972/16/images/95/NMR+Peak+Description+LW1/2.jpg>. (Accessed: 26th August 2018)
24. Harlington Community School. Quick Guide High Resolution 1H Nuclear Magnetic Resonance Spectroscopy. Available at: <https://pt.slideshare.net/varmar71/quick-review-a2-high-resolution-nmr>. (Accessed: 26th August 2018)
25. University of Calgary. Chapter 13: Spectroscopy. Available at: <http://www.chem.ucalgary.ca/courses/350/Carey5th/Ch13/ch13-nmr-5.html>. (Accessed: 26th August 2018)
26. Dona, A. C. *et al.* A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput. Struct. Biotechnol. J.* 1–19 (2016). doi:10.1016/j.csbj.2016.02.005
27. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
28. De Meyer, T. *et al.* NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm. *Anal. Chem.* **80**, 3783–3790 (2008).
29. Serkova, N., Florian Fuller, T., Klawitter, J., Freise, C. E. & Niemann, C. U. 1H-NMR-based metabolic signatures of mild and severe ischemia/reperfusion injury in rat kidney transplants. *Kidney Int.* **67**, 1142–1151 (2005).
30. Spraul, M. *et al.* Mixture analysis by NMR as applied to fruit juice quality control. *Magn. Reson. Chem.* **47**, S130–S137 (2009).
31. Corbet, C. & Feron, O. Tumour acidosis: from the passenger to the driver's seat. *Nat. Rev. Cancer* **17**, 577–593 (2017).
32. Galla, J. H. Metabolic alkalosis. *J. Am. Soc. Nephrol.* **11**, 369–75 (2000).

33. Takis, P. G., Schäfer, H., Spraul, M. & Luchinat, C. Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nat. Commun.* **8**, 1662 (2017).
34. Cloarec, O. *et al.* Evaluation of the Orthogonal Projection on Latent Structure Model Limitations Caused by Chemical Shift Variability and Improved Visualization of Biomarker Changes in ¹H NMR Spectroscopic Metabonomic Studies. *Anal. Chem.* **77**, 517–526 (2005).
35. Xiao, C., Hao, F., Qin, X., Wang, Y. & Tang, H. An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst* **134**, 916–925 (2009).
36. No Title. Available at:
https://files.mtstatic.com/site_4463/9669/0?Expires=1535279012&Signature=LOnoATu9tzrkOneRceb4rXBNqKW-UE3YOLdQE55GtZDwnWWjBVc-5yPsIsJbVhz3py0jz-fr8duycR79Jr0wh56SnWb8SxsggqvTgsEwnTweh2U3BJm1yQOiEVnWWcmz5nLRqX7bThZ2oG173qtkeziFV~74rN7EJ-OhB9oq-t0_&Key-Pair-. (Accessed: 26th August 2018)
37. Tredwell, G. D., Bundy, J. G., De Iorio, M. & Ebbels, T. M. D. Modelling the acid/base 1H NMR chemical shift limits of metabolites in human urine. *Metabolomics* **12**, 1–10 (2016).
38. BMRB - Biological Magnetic Resonance Bank. Available at:
<http://www.bmrwisc.edu/>. (Accessed: 26th August 2018)
39. Viant, M. R., Ludwig, C. & Günther, U. L. Chapter 2. 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling. in *Metabolomics, Metabonomics and Metabolite Profiling* 44–70 (Royal Society of Chemistry, 2007).
doi:10.1039/9781847558107-00044
40. Aalim M. Weljie, †, ‡, Jack Newton, †, Pascal Mercier, †, Erin Carlson, † and Carolyn M. Slupsky*†, §. Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. (2006). doi:10.1021/AC060209G
41. Toumi, I., Caldarelli, S. & Torrèsani, B. A review of blind source separation in NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **81**, 37–64 (2014).
42. Emwas, A.-H. M. The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research. in *Methods in molecular biology (Clifton, N.J.)* **1277**, 161–193 (2015).
43. Psychogios, N. *et al.* The human serum metabolome. *PLoS One* **6**, e16957 (2011).
44. Liu, G. *et al.* One-thousand-fold enhancement of high field liquid nuclear magnetic resonance signals at room temperature. *Nat. Chem.* **9**, 676–680 (2017).
45. Zangger, K. Pure shift NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **86–87**, 1–20 (2015).

46. Aguilar, J. A., Faulkner, S., Nilsson, M. & Morris, G. A. Pure Shift 1H NMR: A Resolution of the Resolution Problem? *Angew. Chemie Int. Ed.* **49**, 3901–3903 (2010).
47. J. Patrick Loria, †, Mark Rance, *, ‡ and & Arthur G. Palmer III*, †. A Relaxation-Compensated Carr–Purcell–Meiboom–Gill Sequence for Characterizing Chemical Exchange by NMR Spectroscopy. (1999). doi:10.1021/JA983961A
48. Chatham, J. C. & Forder, J. R. Lactic acid and protein interactions: implications for the NMR visibility of lactate in biological systems. *Biochim. Biophys. Acta* **1426**, 177–84 (1999).
49. Serge Akoka, *, †, Laurent Barantin, ‡ and & Trierweiler†, M. Concentration Measurement by Proton NMR Using the ERETIC Method. (1999). doi:10.1021/AC981422I
50. Hao, J. *et al.* Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* **9**, 1416–27 (2014).
51. Ravanbakhsh, S. *et al.* Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* **10**, e0124219 (2015).
52. Gómez, J. *et al.* Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Anal. Bioanal. Chem.* **406**, 7967–7976 (2014).
53. Alonso, A. *et al.* Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis. (2014).
54. Sokolenko, S. *et al.* Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1H-NMR spectra in synthetic mixtures and urine with additional insights on choice of pulse sequences and robotic sampling. *Metabolomics* **9**, 887–903 (2013).
55. Lewis, I. A., Schommer, S. C. & Markley, J. L. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn. Reson. Chem.* **47**, S123–S126 (2009).
56. Overbeek, R. A. *9th International Conference on Practical Applications of Computational Biology and Bioinformatics.*
57. Bruker. B.I.QUANT-UR Reproducible Metabolite Quantification in Urine | Bruker. Available at: <https://www.bruker.com/products/mr/nmr-preclinical-screening/biquant-ur.html>. (Accessed: 26th August 2018)
58. Rocca-Serra, P. *et al.* Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **12**, 14 (2016).
59. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–4 (2016).
60. Macleod, M. R. *et al.* Biomedical research: increasing value, reducing waste. *Lancet* **383**, 101–104 (2014).
61. Centre for Open Science. Preregistration of Studies. Available at: <https://cos.io/prereg/>.

- (Accessed: 26th August 2018)
62. Benjamin, D. J. *et al.* Redefine statistical significance. (2017).
doi:10.31234/OSF.IO/MKY9J
 63. Emwas, A.-H. *et al.* Recommendations and Standardization of Biomarker Quantification Using NMR-based Metabolomics with Particular Focus on Urinary Analysis. *J. Proteome Res.* (2016). doi:10.1021/acs.jproteome.5b00885
 64. Kale, N. S. *et al.* MetaboLights: An Open-Access Database Repository for Metabolomics Data. in *Current Protocols in Bioinformatics* 14.13.1-14.13.18 (John Wiley & Sons, Inc., 2016). doi:10.1002/0471250953.bi1413s53
 65. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
 66. Schober, D. *et al.* nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. *Anal. Chem.* **90**, 649–656 (2018).
 67. van der Hooft, J. J. J. & Rankin, N. Metabolite Identification in Complex Mixtures Using Nuclear Magnetic Resonance Spectroscopy. in *Modern Magnetic Resonance* 1–33 (Springer International Publishing, 2017). doi:10.1007/978-3-319-28275-6_6-2
 68. Cuperlovic-Culf, M. Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. *Metabolites* **8**, 4 (2018).
 69. Wang, J. & Perez, L. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning.*
 70. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).
 71. Olah, C. *et al.* The Building Blocks of Interpretability. *Distill* **3**, e10 (2018).
 72. Moore, J. Deep Misconceptions About Deep Learning – Towards Data Science. Available at: <https://towardsdatascience.com/deep-misconceptions-about-deep-learning-f26c41faceec>. (Accessed: 26th August 2018)
 73. Olson, R. S., Cava, W. La, Mustahsan, Z., Varik, A. & Moore, J. H. Data-driven advice for applying machine learning to bioinformatics problems. in *Biocomputing 2018* 192–203 (WORLD SCIENTIFIC, 2018). doi:10.1142/9789813235533_0018
 74. Kuhn, M. & Johnson, K. *Applied Predictive Modeling.* (Springer New York, 2013). doi:10.1007/978-1-4614-6849-3
 75. Efron, B. & Hastie, T. *Computer Age Statistical Inference.* (2016). doi:10.1017/CBO9781316576533
 76. Scikit-learn. 2.3. Clustering. Available at: <http://scikit-learn.org/stable/modules/clustering.html>. (Accessed: 26th August 2018)
 77. Maaten, L. van der, Learning, G. H.-J. of M. & 2008, undefined. Visualizing Data using

- t-SNE. *seas.harvard.edu*
78. Wattenberg, M., Viégas, F. & Johnson, I. How to Use t-SNE Effectively. *Distill* **1**, e2 (2016).
 79. Davis, J. & Goadrich, M. *The Relationship Between Precision-Recall and ROC Curves*.
 80. Dietterich, T. G. & G., T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
 81. Emwas, A. H. M., Salek, R. M., Griffin, J. L. & Merzaban, J. NMR-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* **9**, 1048–1072 (2013).
 82. Slupsky, C. M. *et al.* Investigations of the Effects of Gender, Diurnal Variation, and Age in Human Urinary Metabolomic Profiles. *Anal. Chem.* **79**, 6995–7004 (2007).
 83. Petersen, A.-K. *et al.* On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics* **13**, 120 (2012).
 84. Considine, E. C., Thomas, G., Boulesteix, A. L., Khashan, A. S. & Kenny, L. C. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics* **14**, 7 (2018).
 85. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* (2015). doi:10.1016/j.aca.2015.02.012
 86. Trainor, P., DeFilippis, A. & Rai, S. Evaluation of Classifier Performance for Multiclass Phenotype Discrimination in Untargeted Metabolomics. *Metabolites* **7**, 30 (2017).
 87. Date, Y. & Kikuchi, J. Application of a Deep Neural Network to Metabolomics Studies and Its Performance in Determining Important Variables. *Anal. Chem.* **90**, 1805–1810 (2018).
 88. Alakwaa, F. M., Chaudhary, K. & Garmire, L. X. Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. *J. Proteome Res.* **17**, 337–347 (2018).
 89. Shen, H., Zamboni, N., Heinonen, M. & Rousu, J. Metabolite Identification through Machine Learning— Tackling CASMI Challenge Using FingerID. *Metabolites* **3**, 484–505 (2013).
 90. Acharjee, A., Ament, Z., West, J. A., Stanley, E. & Griffin, J. L. Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC Bioinformatics* **17**, 440 (2016).
 91. Costello, Z. & Martin, H. G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Syst. Biol. Appl.* **4**, 19 (2018).
 92. Harrell, F. Road Map for Choosing Between Statistical Modeling and Machine Learning. Available at: <http://www.fharrell.com/post/stat-ml/>. (Accessed: 26th August 2018)

4 rDolphin: a GUI R package for proficient automatic profiling of 1D ^1H -NMR spectra of study datasets

Abstract

The adoption of automatic profiling tools for ¹H-NMR-based metabolomic studies still lags behind other approaches in the absence of the flexibility and interactivity necessary to adapt to the properties of study data sets of complex matrices and in the impossibility to export all the information collected during profiling. rDolphin is an open-source redesign of the Dolphin profiling workflow that fully integrates these needs to provide the best balance between accuracy, reproducibility and ease of use. rDolphin incorporates novel techniques to optimize exploratory analysis, metabolite identification, and validation of the profiling output quality. To demonstrate these benefits, the information and quality achieved in two public datasets of complex matrices are maximized.

4.1 Introduction

¹H-NMR is a high throughput analytical technique that allows the quantification of metabolites in biofluids, tissues or cell culture extracts in a reliable and reproducible manner. However, variability in the sample properties and preparation and during the spectra acquisition and pre-processing incorporates complexity to the generated spectra.¹ Examples of this complexity are baseline artefacts due to macromolecules, broad signals originated from lipids, signal overlap and misalignment, and variability of signal shapes. Profiling approaches can provide more resilience to these sources of variability than fingerprinting approaches.²

In comparison to manual profiling (e.g., through Chenomx), with high variability depending on the user,³ automatic profiling promises more robustness as well as lower time-consuming demands. Several tools that perform automatic 1D ¹H-NMR profiling have been published, such as Dolphin,⁴ BATMAN,⁵ or BAYESIL⁶ (additional options are in Spicer et al, 2017).⁷ Nonetheless, they may require extensive knowledge of the dataset properties or expertise in programming languages to be fully utilized. Other possible drawbacks are the reliance on strict parameter settings for sample preparation and data acquisition, or the inability to handle unknown signals. Most automatic tools may lose performance when applied to a large number of complex biological samples: the complexity to be monitored may become too demanding to be efficiently controlled in a single profiling iteration, therefore several iterations with different parameters are needed to avoid wrong annotations and suboptimal quantifications. These issues hamper the reproducibility of the profiling-based approaches compared to fingerprint-based ones. The challenges become exacerbated if the user does not have previous expertise with the studied matrix.

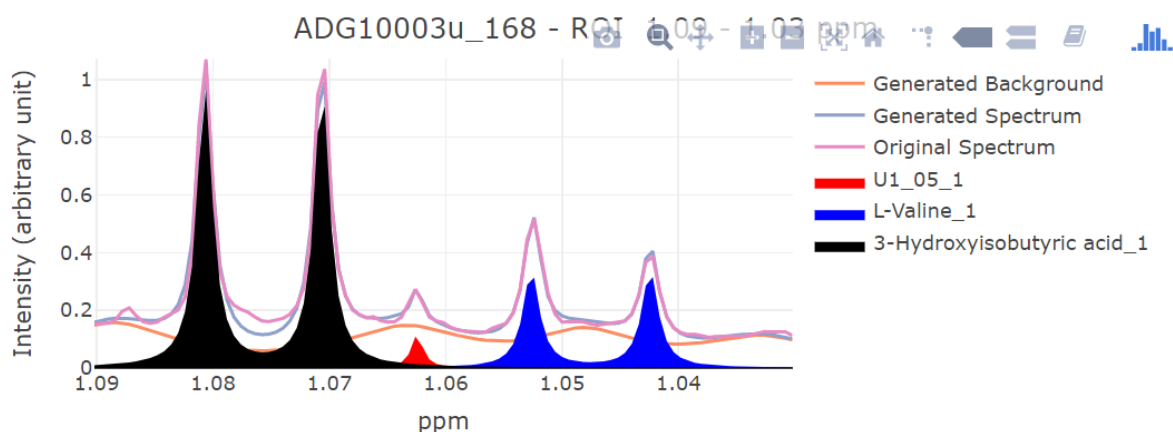
Dolphin is a project that uses the region of interest (ROI)⁸ concept to allow the flexible and time-affordable automatic profiling of 1D ¹H-NMR spectra. Dolphin implements an approach based on the estimation of the baseline and the signal parameters (chemical shift, intensity, half bandwidth, j-coupling) values which maximize the quality in the lineshape fitting of the signals in the analysed ROI (Figure 4-1). The supervision of adequate starting estimates for this calculation is performed through graphical user interface (GUI). The challenges observed during the profiling of study datasets of complex matrices emphasized the need for a new framework with novel solutions to optimize exploratory analysis and a matrix-specific metabolite annotation. This framework should also enable the interactive validation and optimization of the performed annotations and quantifications without the need to generate new profiling iterations. It should as well allow the loading and analysis of metabolite profiling sessions on any computer: the enhanced reproducibility may help heighten the standardization and quality of profiling approaches.⁹ Hereafter, it is presented a redesign implementation of Dolphin workflow using open-source R software, called rDolphin. rDolphin has been specifically redesigned in order to satisfy these needs during the profiling of study datasets of complex matrices.

4.2 Methods

The Dolphin workflow has been extensively tested with around 4050 different spectra across 25 different metabolomics studies with varying biological complexity, ranging from urine, blood to aqueous and lipidic cell extract. The workflow requires as input pre-processed spectra which can be then normalized or aligned in different ways through the tool. Then, the profiles of the metabolite signals to fit in a ROI are adapted to the dataset and metabolite relative concentrations are quantified and outputted (and they can be converted to absolute ones through TSP or ERETIC concentration).

The redesign of the original MATLAB-based Dolphin workflow to elaborate strategies was necessary to maximize the quality and reproducibility of the profiling process in complex matrices. Some capabilities of the redesigned workflow should be the next ones:

- The profiling process should be able to be exported and analysed by other researchers. In order to accomplish this requirement, the profiling tool used should be written in an open-source language (e.g., R, Python) so it can be installed and used by other researchers. In addition, the computing requirements of the tool should be within reach for most research groups.



Here you have some indicators of the quantification.

Show entries

Search:

	Quantification (arbitrary unit)	Fitting Error	Signal/total area ratio
U1_05_1	0.0004	0.02	13.2672
L-Valine_1	0.0019	0.0184	44.3372
3-Hydroxyisobutyric acid_1	0.0051	0.0299	74.4341

Figure 4-1 Example of lineshape fitting in the 1.09–1.03 ppm region of the human urine MTBLS1 dataset. Signal area quantifications and fitting quality indicators are shown below the interactive Plotly figure.

- All necessary information to load the profiling process in the researcher’s server should be saved in a format compatible with this server. This information includes the parameters used during profiling, the performed signal annotations and the profiling output (with associated quality indicators and figures).
- The loaded profiling session should be possible to be improved if necessary. This would include the generation of new profiling iterations with new parameters and the individual correction of quantifications according to the review of quality indicators.

To satisfy these requirements, several changes and additions (based on novel strategies) in the Dolphin workflow were implemented. The next sections provide a description of each one of them.

4.2.1 Improvement of input and output structures

The use of inputs and outputs in XLS format from the Dolphin workflow hardened the integration of these inputs and outputs with the ones of the other steps of metabolomics study workflows. Accordingly, these inputs and outputs were changed to CSV format, a text-based format much more flexible to be correctly interpreted by different tools and programming languages. The use of the CSV format also avoids the limitations of the XLS format to handle in an efficient way the dimensionality of spectra datasets. As a result, the input of spectra datasets not acquired by Bruker technology was enabled.

In addition, total flexibility to tune the parameters of the profiling workflow was enabled through the generation of an additional input CSV file where to specify any parameter modifications. Lastly, trending pre-processing methods were added as optional pre-processing options during the redesign (e.g., probabilistic quotient normalization -PQN⁻¹⁰).

4.2.2 Implementation in open-source code

Open-source capabilities guaranteed that the profiling workflow of the tool can be integrated into any study workflow of any research group. In addition, the use of open-source code allowed the use of state-of-the-art statistical techniques that can help improve metabolic profiling. R, as the most prevalent programming language in the scientific community, ensures the maximum scope to the developed reimplementation and the easiest learning slope for researchers still not proficient with programming. In addition, the code of the tool was shared in a GitHub repository, allowing other researchers and developers correct bugs and suggest improvements.

In order to support the interactive evaluation and optimization of the profiling process, a Shiny GUI was incorporated. However, as R is not a general-purpose language but a statistical purpose language, its capabilities to prepare interfaces are not as developed as the ones of other languages, and Shiny has limitations in computing efficiency which render it more suited to the generation of light-weight standalone web apps. Consequently, in contrast to Dolphin, the console-based use of the tool by R functions was also enabled in rDolphin. In addition, this console-based use of the tool facilitated the smooth integration of the tool workflow with other steps of a metabolomics study within a terminal-based pipeline.

4.2.3 Tools for the effective interactive visualization of spectra

One of the main advantages of Dolphin in comparison with alternative profiling tools was the interactive visualization of spectra, a necessary requisite to perform a robust exploratory analysis of the sources of variability present in the spectra. The generation of interactive figures was achieved thanks to the Application Programming Interface (API) 'Plotly' (Figure 4-2). The generated figures can be zoomed in and out and they provide additional information about the data point where the cursor is located.

The biggest drawback of this API is the high computing demands of the figures generated. As a result, a dataset of dozens of spectra cannot be optimally visualized in common servers. To overcome this drawback, it was considered that the dimensionality of the spectra dataset should be reduced to one that could, notwithstanding, maintain as much relevant information as possible about the total dataset so exploratory analysis was of high quality. In addition, the reduction of spectra to observe would help researchers focus on the most important phenomena to control during profiling.

Accordingly, two different kinds of interactive figures of multiple spectra were generated:

- A figure with a spectra subset which is representative of the variance present in the dataset. To achieve this purpose, it was necessary to perform a row-wise dimensionality reduction of the dataset by means of the clustering of the spectra dataset and the selection of an exemplar for each cluster. This clustering and selection of exemplars was achieved thanks to the affinity propagation algorithm provided by the 'apcluster' R package.¹¹ This algorithm performs the clustering of data without the need of specifying the number of clusters to identify. The automatic selection of exemplars able to represent the total variance of the spectra dataset supposes a novel contribution to the evaluation of the NMR spectra datasets.

To ensure a high-quality representation of the total spectra dataset through the selection of exemplars, it was necessary to maximize the performance of the algorithm. Accordingly, first it is necessary to filter non-informative features (i.e., spectrum regions without relevant presence of metabolite signals). Then, it is necessary to scale the data to give equal importance to each feature during spectra clustering. Later, it is necessary to remove outlier spectra as they will create individual clusters which do not optimally represent the most important variability patterns in the dataset. Only after these processes, the algorithm can be efficiently applied to generate a number of approximately 10 exemplars to be visualized through Plotly figures (Figure 4-2).

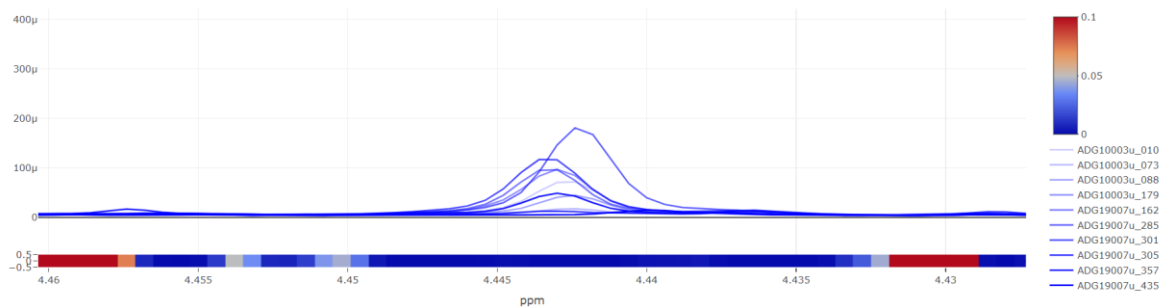


Figure 4-2 Reduction of a 132 spectra dataset into 10 representative exemplars (whose sample names are specified below right). This interactive figure is created by the Plotly API.

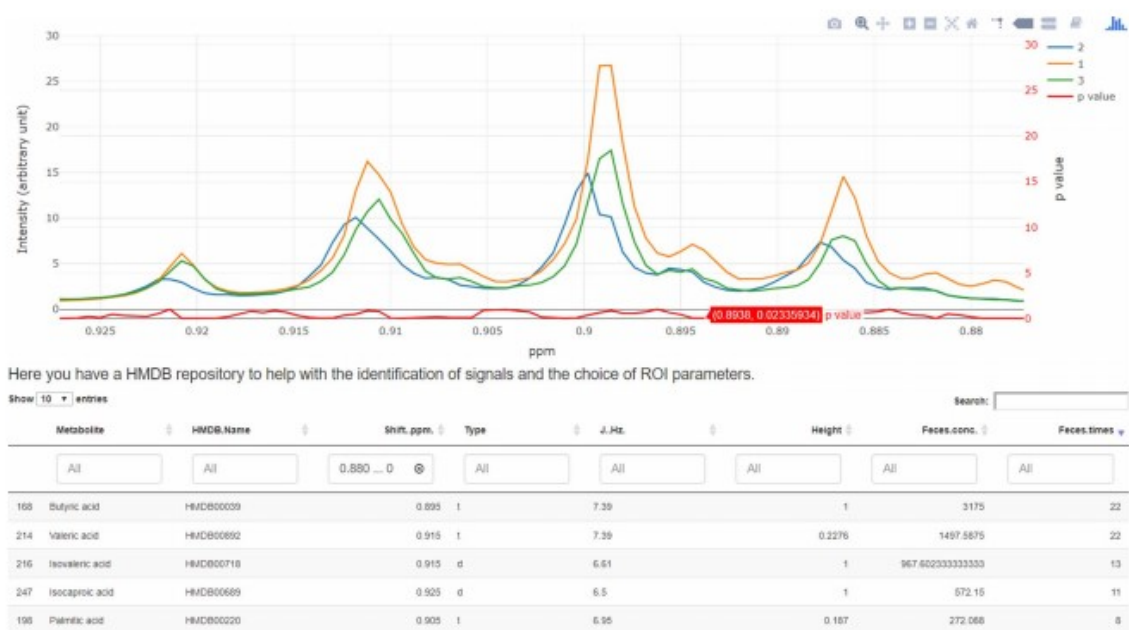


Figure 4-3 Exploratory analysis of human faecal extract MTBLS237 dataset with rDolphin. Differences between the median spectrum of three kinds of sample in the 0.92–0.88 ppm region are shown on an interactive figure. Fingerprint analysis information is also provided by the red trace below the median spectra

- A figure with the median spectrum of each group of samples analysed (Figure 4-3). This kind of figure simultaneously accomplishes two objectives: the generation of a spectra dataset whose visualization is much less computationally intensive and the generation of information that helps the researcher focus on the most important regions to spectrum to analyse during profiling. Metabolite profiling tends to be a combination of targeted + untargeted approach: a standard number of metabolites consistently observed in the analysed matrix is profiled but this number can be increased to add metabolites which might be harder and less reliable to profile but might contain relevant insights about the metabolome in the analysed dataset. The visualization of the median spectrum of different sample groups enables the selective addition to profiling of metabolites which will be

relevant in the spectra dataset analysed. In addition, to increase the quality of the exploratory visualization of differences between sample groups, data of fingerprint analysis was added to the tool (Figure 4-3).

This fingerprint information consists of a basic automatic hypothesis testing workflow performed for each bucket through the ‘p_values’ function. In two-sample tests, non-normality in every group of samples is checked using Shapiro-Wilk tests. If a group of samples shows no normality, a Mann-Whitney test is performed; if all groups show normality a Welch t-test is performed (an explanation of why Welch’s t-tests are better than Student’s t-tests is available in Lakens, 2015).¹² The p-values estimated in every study were then Benjamini-Hochberg adjusted.

4.2.4 Generation of matrix-specific information of suggested signals to annotate

No current tool provided a platform to suggest a ranking of signals to be annotated in a determined spectrum location according to the matrix studied. Chenomx or AMIX inform of possible signals according to a general database of metabolite signals, but this strategy is liable to wrong annotations of signals of metabolites not present in the studied matrix. In addition, the lack of filtering by matrix greatly expands the range of possible options of signal annotation, making this process more complex. On the other hand, Bayesil or BI-QUANT UR automatically annotate signals according to the studied matrix. However, this approach is not flexible to the appearance of metabolites because of the study design or of future sensitivity improvements.

Some challenges might explain this lack of current workflows to help the user annotate reliably signals on matrices. There exists a high range of possible different matrices to study and each matrix can have different sample preparation and spectrum acquisition protocols which vary the number of metabolites that can be profiled.^{13,14,15} In addition, relatively recent emergence of metabolomics implies a constantly improving process where identifications of typical metabolites in a matrix are still in progress. Consequently, the metabolomics literature is still flawed with wrong annotations which harden the creation of reliable rankings. Any approach to creating a ranking of signals by matrix will need to adapt to the constant evolution in the study workflows and metabolite identifications currently happening in the metabolomics field.

To enable a dynamic ranking which can adapt to this progress, rDolphin incorporated a novel metabolite signal repository based on public information that can be found in the Human

Metabolome Database (HMDB).¹⁶ The HMDB provides extensive information about 114,064 metabolite entries which can be helpful for researchers interested in the study of the metabolome. Apart from other information, the database contains information of signals mediated by each metabolite in several kinds of NMR spectra (this information is the foundation of the Bayesil tool developed by the same research group) as well as information about the presence of the metabolite in previous metabolomics studies (with information about the concentration value and the matrix studied) (Figure 4-4; top). For each metabolite, the information about the times it appeared in previous metabolomics studies and the concentration reported in each appearance can be extracted from an XML file available to the metabolomics community (Figure 4-4; bottom).

Matrix	Status	Concentration	Age Group	Sex	Condition	Reference	Details
Blood	Detected and Quantified	40.0-95.0 uM	Adult (>18 years old)	Female	Normal	Vancouver Co...	details
Blood	Detected and Quantified	60.00-115.0 uM	Adult (>18 years old)	Male	Normal	Vancouver Co...	details
Blood	Detected and Quantified	86.6 +/- 18.8 uM	Adult (>18 years old)	Both	Normal	21359215	details
Breast Milk	Detected and Quantified	39.9 +/- 7.9 uM	Adult (>18 years old)	Female	Normal	24027187	details
Cerebrospinal Fluid (CSF)	Detected and Quantified	43 +/- 12 uM	Adult (>18 years old)	Both	Normal	18502700	details
Cerebrospinal Fluid (CSF)	Detected and Quantified	65.2 (51.8-78.6) uM	Adult (>18 years old)	Both	Normal	7108550	details
Cerebrospinal Fluid (CSF)	Detected and Quantified	64.95 (37.5-92.4) uM	Adult (>18 years old)	Both	Normal	Geigy Scientific ...	details
Feces	Detected but not Quantified		Children (6 - 18 years old)	Both	Normal	27609529	details
Feces	Detected but not Quantified		Children (6 - 18 years old)	Not Specified	Normal	27609529	details
Saliva	Detected but not Quantified		Adult (>18 years old)	Male	Normal	22308371	details


```

</concentration>
<concentration>
  <biofluid>Blood</biofluid>
  <concentration_value>86.6 +/- 18.8</concentration_value>
  <concentration_units>uM</concentration_units>
  <subject_age>Adult (sgt:18 years old)</subject_age>
  <subject_sex>Both</subject_sex>
  <subject_condition>Normal</subject_condition>
  <references>
    <reference>
      <reference_text>Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B,
      <pubmed_id>21359215</pubmed_id>
    </reference>
  </references>
</concentration>
<concentration>
  <biofluid>Breast Milk</biofluid>
  <concentration_value>39.9 +/- 7.9</concentration_value>
  <concentration_units>uM</concentration_units>
  <subject_age>Adult (sgt:18 years old)</subject_age>
  <subject_sex>Female</subject_sex>
  <subject_condition>Normal</subject_condition>
  <comment>Samples collected in the morning on day 90 postpartum. Metabolite was measured by using 1H NMR spectroscopy.</comment>
  <references>
    <reference>
      <reference_text>Smilowitz JT, O'Sullivan A, Barile D, German JB, Lonnerdal B, Slupsky CM: The human milk metabolome reveals
      <pubmed_id>24027187</pubmed_id>
    </reference>
  </references>
</concentration>
<concentration>
  <biofluid>Cerebrospinal Fluid (CSF)</biofluid>
  <concentration_value>43 +/- 12</concentration_value>
  <concentration_units>uM</concentration_units>
  <subject_age>Adult (sgt:18 years old)</subject_age>
  <subject_sex>Both</subject_sex>
  <subject_condition>Normal</subject_condition>
  <references>
    <reference>
      <reference_text>
      <pubmed_id>
    </reference>
  </references>
</concentration>

```

Figure 4-4 Example of available information of reported concentrations in the HMDB website (top) and the equivalent information present in XML format (down).

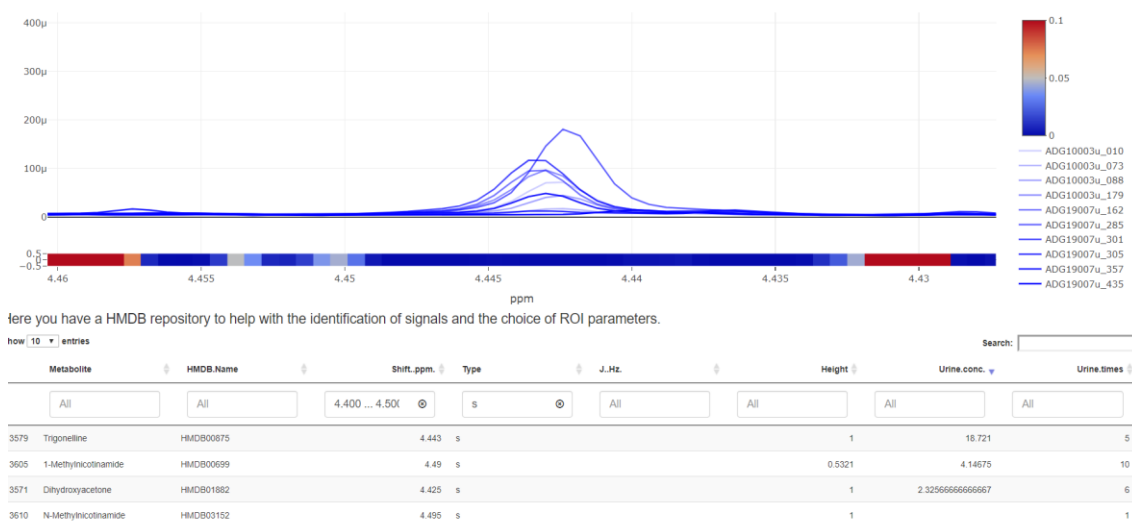


Figure 4-5 The use of HMDB information facilitates the accurate matrix-specific information of metabolite signals. The rDolphin repository of metabolite signals can be filtered by the matrix and the spectrum region. Then, signals can be sorted according to the presence in previous bibliography and of its typical concentration in the matrix analysed. In addition, the repository provides information about the kind of multiplet, the J-coupling and the relative intensity of the signal.

To connect the matrix-specific metabolite concentration information with the information about the signals of each metabolite, it was necessary to find the information about the metabolite signals in NMR 1D spectra and to merge it with its concentration information. The extraction of the information of the metabolite signals required extensive work to curate and merge the information present in thousands of public TXT files with non-consistent data structures. After that, a final general repository of signals of thousands of metabolites was created. This general repository provides information such as the chemical shift, the multiplicity (singlet, doublet, triplet...), the j-coupling and the relative intensity of the signal. The chemical shift information can then be used to limit the database to signals typically located in the spectrum region where the signal to annotate is located. This information can be complemented by the times this signal has appeared in previous metabolomics studies in the same matrix and by the calculated concentration in these studies (Figure 4-5). The combination of these three sources of information provides a robust solution to the occasional inaccuracies present in these evolving databases.

4.2.5 Flexibility to correct suboptimal quantifications

rDolphin provides the first approach to interactively revise and correct individual suboptimal quantifications inside a profiling tool. This possibility to perform an interactive review and

correction of quantifications may help reducing the presence of false positives and negatives in the literature caused by the presence of wrong annotations and suboptimal quantifications.

The first requisite to enable these capabilities was the storage of the profiling session information into a file that could be later loaded on the profiling tool. Ideally, this file would be able to be loaded on any computer then avoiding the need to perform all iterations on the same computer. In addition, the storage of all profiling information into a file that can be loaded on any computer would increase the reproducibility potential of the profiling performed in order to be reviewed by other research groups. These desired objectives were reached through the storage of the profiling information into .RData format of all the information regarding the dataset, the signal parameters and the quantification performed for each signal in each dataset.

After the creation of the necessary structure to be able to load performed profiling sessions, the tool incorporated interactive data tables of different indicators of quality in order to evaluate the annotation and performance in each quantification. Common quality indicators of the quantification are based on the performance of deconvolution in comparison to the spectrum lineshape. In contrast, rDolphin incorporated the difference between the predicted and the expected signal parameter values according to the information collected during profiling (the progress in the analysis of the expected values of the signal parameters became the foundations of Chapter 6; more concrete details about the approach to estimate the expected parameter values is available there).

The values of the five quality indicators calculated (fitting error, signal to total area ratio, difference between expected and predicted chemical shift, difference between expected and predicted half bandwidth, difference between expected and predicted intensity) can be used to generate visual input about the quality of each quantification in the interactive data table provided. The last three indicators are novel information sources enabled by ML-based prediction of these signal parameters according to information extracted from signals correlated to the one of interest. The cells of the data table are red coloured with a shade proportional to the difference between the perfect value and the calculated value (Figure 4-6; a). In addition, the interactivity of the data table enables the ordering of the quantifications according to the value of the quality indicator; this allows the researcher focusing only on the most suspicious quantifications (Figure 4-6).

The saving of the performed quantifications and the interactivity of the data table enabled the loading through a click of the identified suspicious quantification into the GUI in order to be evaluated and, if necessary, updated. The update of the quantification can be performed through the edition of the initial ROI parameters in order to adapt the better to the characteristics of the lineshape fitted ROI. Nonetheless, this edition might not be enough in especially complicated

cases. For these cases, rDolphin incorporated the manual edition of the baseline and signal parameters to be visually evaluated until achieving a combination which satisfies the user.

a)

Show 10 entries Search

creatine_1	cis-Aconitate_1	Prolise Betaine_0	Carnitine_1 *	Betaine_1	Trimethylamine N-oxide_1	T
	0.0002		-0.0055		-0.0004	
	-0.0001		-0.0042		0.0002	
	0.0001		-0.0039		0	
	0.0002		-0.0032		0.0004	
	-0.0001		-0.003		-0.0014	
	0.0007		-0.002		0.0002	
	0.0003		-0.002		-0.0004	
	-0.0009		-0.0011		0.0002	
	0		-0.0008		0.0001	
	0		-0.0004		0.0002	

Showing 1 to 10 of 132 entries

Previous 1 2 3 4 5 ... 14 Next

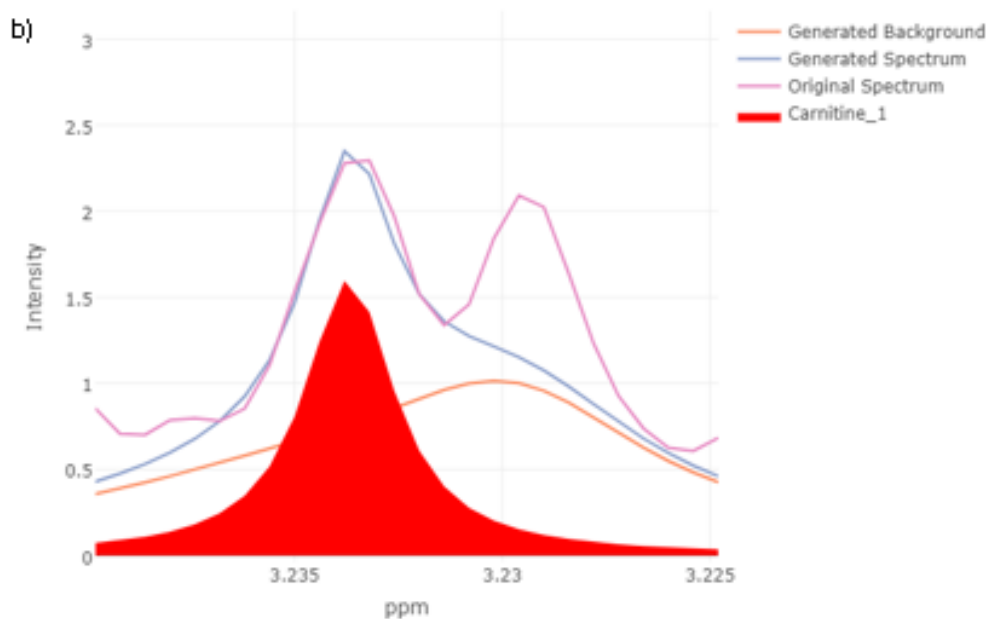


Figure 4-6 rDolphin enables the finding of wrong annotations and suboptimal quantifications through several indicators of quality. In a), possible suboptimal quantifications of carnitine have been ordered by difference between the chemical shift (in ppm) of the performed quantification and the predicted chemical shift. The shade suggests the grade of outlier behaviour. In b), the predicted chemical shift of carnitine is located 0.0042 ppm below than the one of the fitted signal, exactly where the neighbouring signal to its right is located.

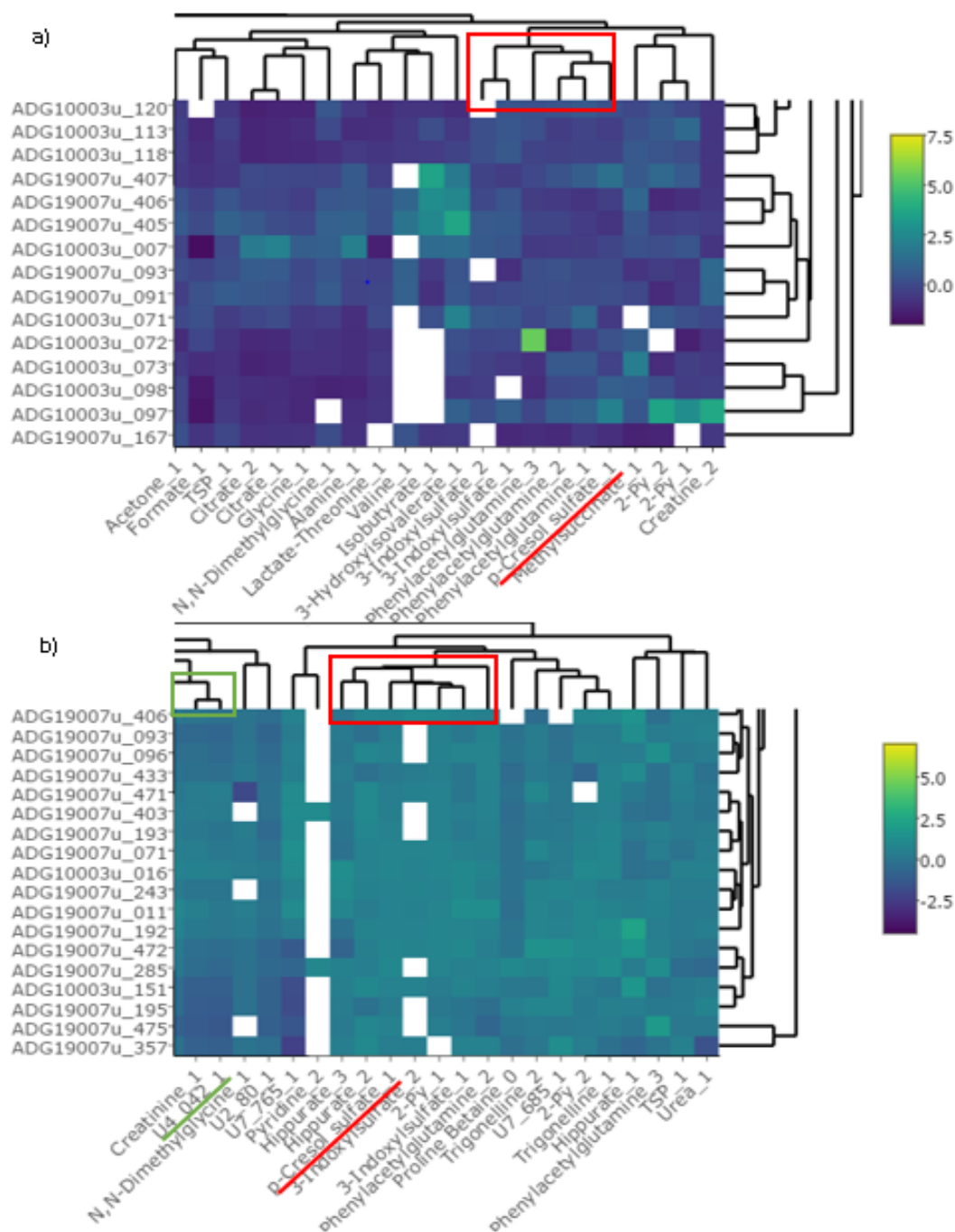


Figure 4-7 The dendrogram heatmaps of rDolphin show the signals with similar quantification (a) and chemical shift (b) patterns. The figures show the dendrograms observed in the MTBLS1 dataset. The singlet at 2.35 ppm (annotated as p-Cresol sulphate in the dendrogram) shows similar quantification patterns to related metabolites such as indoxyl sulphate or phe-nylacetylglutamine. This signal also shows chemical shift patterns similar to the ones of metabolites with similar functional groups such as indoxyl sulphate or hippurate. In b), the strong interrelation between the triplet at 4.042 ppm (annotated as U4_042 in the dendrogram) and a creatinine signal can also be observed.

4.2.6 Identification of unknown or wrongly identified signals

Previous implementation of the Dolphin workflow already provided the option of metabolite identification through STOCSY.^{17,18} However, these tools might be sometimes limited by factors such as signal misalignment, baseline, or correlated metabolite concentrations. In order to maximize metabolite identification capabilities, rDolphin incorporated dendrogram heatmaps of quantification and chemical shift of the profiled signals in order to help in their identification (Figure 4-7). The multicollinearity in the chemical shift and concentration information can be exploited to explore the clustering of signals of not identified metabolites with signals from identified ones. The clustering of the unidentified signal with known metabolite signals provides valuable chemical and biological information to help identify this signal. The clustering of chemical shift information to help identify metabolites is a novelty within profiling tools.

4.3 Results and Discussion

Two public metabolomics datasets from the MetaboLights repository were profiled in order to observe the possible benefits of the redesign of the profiling workflow with the incorporation of novel ML based approaches:

- MTBLS1: This study contains 132 spectra of human urine samples.¹⁹
- MTBLS237: This study contains 114 spectra of human faecal extract.²⁰

The next benefits were observed:

- Improvement and time reduction of exploratory analysis: Thanks to the novel interactive assistance options (Figure 4-2; Figure 4-3; Figure 4-5), the preparation of the necessary information to profile the MTBLS237 dataset (a dataset from a not previously studied matrix) only lasted 3 h on a standard computer. In the case of the MTBLS1 dataset, most efforts during exploratory analysis were focused on the modification of chemical shift information (in order to control for the buffer influence) and on the identification of some metabolite signals. The process lasted <2 h. In the MTBLS1 dataset, 40 metabolites were profiled. In the case of the MTBLS237 dataset, 34 metabolites were profiled. Profiling results show that common challenges found during exploratory analysis were efficiently monitored thanks to the novel options provided.

- Avoidance of possible suboptimal quantifications: Wrong annotations (e.g., in the L-carnitine quantification in the MTBLS1 dataset caused by the combination of chemical shift variability and signal overlap) were found thanks to the information of the predicted chemical shifts (Figure 4-6).

Limited resolution, metabolite concentration variability and signal misalignment create wrong annotations of overlapping signals and limit the effectiveness of automatic approaches to accurately annotate and quantify the signals of interest in all spectra. The information and possibility of maximization of the fitting quality enabled by rDolphin GUI provide the necessary framework to maximize the robustness and quality of NMR profiling strategies. This maximization will be even more important when promising improvements in NMR sensitivity and resolution enable the increase in the number of profiled metabolites in study datasets. These improvements will increase signal overlap and, in the case of pure shift NMR, remove the multiplicities which ease identification: optimal approaches for reliable identification and deconvolution of signals will become even more necessary.

- Two inaccurate metabolite identifications in the original study of the MTBLS1 dataset were demonstrated. These ones are a singlet at 2.35 ppm annotated as oxaloacetate/pyruvate in the MTBLS1 dataset and as p-Cresol sulphate in our database, and a triplet at 4.042 ppm annotated as uridine in the MTBLS1 dataset and as U4_042 in our database. These inconsistencies could be evaluated thanks to the use of dendrogram heatmaps of quantification and chemical shift patterns. The quantification dendrogram heatmap of the singlet at 2.35 ppm showed that the signals with most similar quantification patterns to the one of the singlet were indoxyl sulphate and phenylacetylglutamine signals (Figure 4-7; top). Indoxyl sulphate and phenylacetylglutamine are also uremic solutes like p-Cresol sulphate and they have reported a relationship with p-Cresol sulphate. Likewise, the chemical shift dendrogram heatmap showed that the singlet had similar chemical shift patterns to the ones of other metabolites with phenolic groups such as hippurate and indoxyl sulphate (Figure 4-7; down). In the case of U4_042, it was observed that this broad triplet at 4.04 ppm present in some human urine datasets is a signal closely related to creatinine observable in spectra of internal standards of creatinine. To the knowledge of the author, this identification represents a novelty in the profiling of human urine datasets which stresses the need to enhance the reproducibility of studies.

In addition, the information outputted from these datasets (and from two other public datasets: MTBLS242 and MTBLS374) has been made public in the package GitHub website. The presence of public reproducible profiling datasets, with the profiling workflow performed in every spectrum of the dataset from a complex mixture, is a novelty on the field of NMR metabolite profiling.

4.4 Limitations

- The ROI approach implemented in rDolphin did not relate signals from the same metabolite placed in different ROIs. Nonetheless, Chapter 6 describes the approach designed to relate these signals and, at the same time, avoid the limitations of the simultaneous line-shape fitting of all signals (e.g., fragility to uncertainty in the chemical shift of any of the signals).
- rDolphin cannot deconvolute signals that are more complex than quadruplets (although these complex signals can be decomposed in substructures so to be profiled).
- The information of metabolite identification and concentration for each matrix is limited to the human biofluids present in the HMDB database. In addition, concentration and identification information become less accurate the less studied is the biofluid.
- rDolphin is dependent on multiple novel packages. In addition, these packages are dependent on others and so on. Therefore, rDolphin is directly or indirectly dependent on dozens of packages. As a result, there is fragility to bugs or changes in language, package or operative system which can affect to any one of the packages involved. In addition, the lack of top-down organization in the maintenance of open-source applications causes that the maintenance of the packages developed during scientific research is usually performed voluntarily by individual researchers. These researchers might not be incentivized to perform this maintenance work when progressing in their career and, therefore, the package might become non-usable despite its potential.

In order to minimize these limitations, it has emerged the idea of the containerization of tools into open-source standalone executables that ensure the correct performance of the tool. Containerization of the Dolphin workflow into a container (e.g. a Docker file) available online on a curated wrapper of metabolomics packages remains a promising area being currently explored.^{21,22}

4.5 Achievements

- The building of an open-source automatic profiling tool which provides solutions to handle the challenges typical from complex matrices with the best balance between accuracy, reproducibility and ease to use.

- The collection of the datasets of signal parameters necessary to perform the studies and achievements in Chapter 5 and Chapter 6.
- The generation of indicators of possible wrong annotations and improvable quantifications to help correct individual quantifications. These indicators are based on the ML-based study of the difference between the expected signal parameter value and the obtained one.
- The generation of a ML-based tool able to help during the annotation of metabolite signals thanks to the analysis of clusters of chemical shifts which behave similarly to the signal analysed (and, therefore, should come from metabolite with similar structures).
- The creation of the first public reproducible ¹H-NMR metabolite profiling workflows of metabolomics studies based on already public study datasets in order to enhance the reproducibility of metabolomics study workflows.
- The generation of a metabolite identification tool adapted to minimize wrong annotations of e.g. metabolites not typical from the matrix analysed. This enhanced version of metabolite annotation tool is based on the data mining of open-source HMDB information about the reported concentration and presence information of each metabolite for each matrix and about the parameters of each metabolite signal.
- The row-wise dimensionality reduction of a spectra dataset thanks to the selection of exemplars of spectra clusters able to efficiently represent the variance present in a spectra dataset.

References

1. Sokolenko, S. et al. Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1H-NMR spectra in synthetic mixtures and urine with additional insights on choice of pulse sequences and robotic sampling. *Metabolomics* 9, 887–903 (2013).
2. Aalim M. Weljie, †,‡, Jack Newton, †, Pascal Mercier, †, Erin Carlson, † and Carolyn M. Slupsky*†, §. Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. (2006). doi:10.1021/AC060209G
3. Tredwell, G. D., Bundy, J. G., De Iorio, M. & Ebbels, T. M. D. Modelling the acid/base 1H NMR chemical shift limits of metabolites in human urine. *Metabolomics* 12, 1–10 (2016).
4. Gómez, J. et al. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Anal. Bioanal. Chem.* 406, 7967–7976 (2014).
5. Hao, J. et al. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* 9, 1416–27 (2014).
6. Ravanbakhsh, S. et al. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* 10, e0124219 (2015).
7. Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* 13, 106 (2017).
8. Lewis, I. A., Schommer, S. C. & Markley, J. L. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn. Reson. Chem.* 47, S123–S126 (2009).
9. Rocca-Serra, P. et al. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* 12, 14 (2016).
10. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1 H NMR Metabonomics. *Anal. Chem.* 78, 4281–4290 (2006).
11. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27, 2463–2464 (2011).
12. Lakens, D. The 20% Statistician: Always use Welch’s t-test instead of Student’s t-test. Available at: <http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html>. (Accessed: 18th August 2018)
13. Bouatra, S. et al. The Human Urine Metabolome. *PLoS One* 8, (2013).
14. Psychogios, N. et al. The human serum metabolome. *PLoS One* 6, e16957 (2011).

15. Johnson, S. R. & Lange, B. M. Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front. Bioeng. Biotechnol.* 3, 22 (2015).
16. Forsythe, I. J. & Wishart, D. S. Exploring Human Metabolites Using the Human Metabolome Database. in *Current Protocols in Bioinformatics* 25, 14.8.1-14.8.45 (John Wiley & Sons, Inc., 2009).
17. Olivier Cloarec, † et al. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. (2005). doi:10.1021/AC048630X
18. Wei, S. et al. Ratio Analysis Nuclear Magnetic Resonance Spectroscopy for Selective Metabolite Identification in Complex Samples. *Anal. Chem.* 83, 7616–7623 (2011).
19. Salek, R. M. et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiol. Genomics* 29, 99–108 (2007).
20. Bjerrum, J. T. et al. Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. *Metabolomics* 11, 122–133 (2015).
21. PhenoMeNal – Large-scale Computing for Medical Metabolomics. Available at: <http://phenomenal-h2020.eu/home/>. (Accessed: 18th August 2018)
22. van Rijswijk, M. et al. The future of metabolomics in ELIXIR. *F1000Research* 6, (2017).

5 Improving sample classification by harnessing the potential of ^1H -NMR signal chemical shifts

Abstract

NMR spectroscopy is a technology that is widely used in metabolomic studies. The information that these studies most commonly use from NMR spectra is the metabolite concentration. However, as well as concentration, pH and ionic strength information are also made available by the chemical shift of metabolite signals. This information is typically not used even though it can enhance sample discrimination, since many conditions show pH or ionic imbalance. It is demonstrated how chemical shift information can be used to improve the quality of the discrimination between case and control samples in three public datasets of different human matrices. In two of these datasets, chemical shift information helped to provide an AUROC value higher than 0.9 during sample classification. In the other dataset, the chemical shift also showed discriminant potential (AUROC 0.831). These results are consistent with the pH imbalance characteristic of the condition studied in the datasets. In addition, it is shown that this signal misalignment dependent on sample class can alter the results of fingerprinting approaches in the three datasets. Our results show that it is possible to use chemical shift information to enhance the diagnostic and predictive properties of NMR.

5.1 Introduction

Metabolomics (or metabonomics) is the study of the metabolome in biofluids, cells or tissues extracted from animals and plants by characterizing the metabolic fingerprint or phenotype (or their underlying mechanisms) in a biological system.^{1,2} ¹H-NMR spectroscopy is a high-throughput technique that quantifies metabolite concentrations in a reliable and reproducible manner.³ ¹H-NMR data can be used to classify samples, so it is a powerful means for capturing diagnostic and predictive properties and has promising potential for personalized medicine.⁴

A metabolite can be characterized in an ¹H-NMR spectrum by its characteristic pattern of signals. The metabolite concentration can be measured by estimating the area below any one of these signals. Likewise, each signal has a specific location determined by its chemical shift (the resonant frequency of its nucleus in a magnetic field). For example, lactate concentration can be quantified from a signal with a chemical shift located at 1.33 ppm or from another signal with a chemical shift located at 4.11 ppm.⁵ The chemical shift (that is to say, the location in a spectrum) of signals is influenced by the pH and the ionic strength (mostly mediated by Ca²⁺ or Mg²⁺

concentration) of the sample.⁶ The information about pH and ionic strength given by the chemical shifts has already been proved to be beneficial for the quality control of fruit juice.⁷ A recent article showed that the pH and ionic strength of human urine samples can be extrapolated from chemical shift information.⁸ A wide range of diseases (e.g., tumours⁹) are characterized by metabolic alkalosis/acidosis¹⁰ or ionic imbalance⁸: these diseases could be better identified in the NMR data with the help of chemical shift information. In addition, theoretical proof of the potential of chemical shift information to separate samples is already available.¹¹ Even so, chemical shift information is still not used to characterize these sample properties and possible differences between classes because the pH and ionic strength can be masked by phosphate buffering and the dilution of matrices varies considerably. These factors hinder the interpretability of the pH information provided by DFTMP¹² or Chenomx-based pH calibration.

To date, several tools have been developed to automatically quantify metabolite concentrations in 1D ¹H-NMR spectra datasets,^{13–15} making it easier to collect additional information, including signal chemical shifts. For example, a recent redesign of the Dolphin NMR tool rDolphin using open-source R language provided more flexible and reproducible automatic metabolite profiling in 1D ¹H-NMR datasets.¹⁶ One additional feature of rDolphin is its ability to capture and output additional information (such as the signal parameter values –including chemical shift– from every quantified signal) for further evaluation. The collection of multiple chemical shifts and the open-source availability of complex algorithms able to combine their information make it possible to use chemical shift information to discriminate samples despite the drawbacks of pH masking and dilution mentioned above. In this study, it is reported an approach to combine the binomial of metabolite concentration and signal chemical shift information in NMR data from metabolomic studies to maximize NMR discriminant potential. To do so, the metabolite concentrations and signal chemical shifts of three public NMR metabolomic study datasets are quantified. It is shown that chemical shift information can be used to separate samples more effectively than just metabolite concentration information.

5.2 Materials and Methods

5.2.1 Datasets

Three NMR datasets from different human matrices from MetaboLights¹⁷ (a public repository of metabolomic studies) were analysed and profiled:

- MTBLS1 MetaboLights dataset: fingerprint NMR data (with adaptive binning) was used to analyse metabolomic changes mediated by type 2 diabetes in mouse, rat, and human urine.¹⁸ The Metabolights dataset provides human urine data of 84 samples from nondiabetics and 48 samples from diabetics.
- MTBLS237 Metabolights dataset: in human faecal extract samples, fingerprint NMR data was used to determine the metabolic profiling of control subjects and patients with active or inactive ulcerative colitis (UC) and Crohn's disease (CD).¹⁹ The spectra dataset analysed consisted of: 20 control samples, 14 active CD samples, 31 inactive CD samples, 19 active UC samples and 28 inactive UC samples.
- MTBLS374 Metabolights dataset: the metabolic serum profiles of smokers and non-smokers were compared in order to study functional alterations caused by smoking through fingerprint data.²⁰ The original study analysed ¹H-NMR fingerprint data, with the help of 2D spectrum information, to identify metabolites. According to the information available on the repository, the spectra dataset analysed in our study consisted of 56 samples from smokers and 57 samples from non-smokers.

Details about sample preparation, spectrum acquisition and main results are available in the original manuscripts. Information about the buffer and dietary restrictions in the original studies is available in the *Datasets* section of Chapter 4. Information about chemical shift variability in metabolite signals after sample preparation is available in Figure 5-3. The ethical issues regarding the studies associated with the used datasets are described in detail in their original articles.¹⁸⁻²⁰

5.2.2 Spectra pre-processing and profiling.

The spectrum pre-processing parameters available in the manuscripts of the studies associated with the datasets used were evaluated to generate ¹H-NMR spectra similar to the ones of the original studies. All datasets were normalised using PQN as it is the recommended normalisation method in recent reviews.²¹ This method analyses the distribution of quotients of the amplitudes of each spectrum with those of a reference spectrum, and then normalises the spectrum by the median of the distribution of quotients.²² Then, data binning (0.0006 ppm) was applied to the spectra before they were profiled by rDolphin. Unreliable relative metabolite concentrations and signal chemical shifts were filtered using a variety of quality indicators (additional information is available in Appendix). Then, univariate outliers for each feature (controlling for sample class) were set as missing values and imputed.

For metabolite concentration information, the final dataset consisted of: MTBLS1, 39 features; MTBLS237, 35 features, MTBLS374, 30 features. For chemical shift information, the features were highly correlated. Consequently, in each dataset, dimensionality was reduced by principal components analysis (PCA) and the dozens of correlated chemical shifts were grouped into 5 independent principal components (enabling the factors influencing signal chemical shifts to be accurately evaluated).

5.2.3 Multivariate analysis

First, an exploratory visualization was performed in both metabolite concentration and chemical shift information datasets to compare their discriminant potential. The visualization was based on the results of a PCA performed to each set of information. During this exploratory visualization, it was also checked that no batch effects exerted an effect on the observed differences.

Next, sample classification was performed using the RF algorithm, a decision tree-based algorithm which combines predictions and uses bootstrapping to maximize the optimization of bias and variance.^{23,24} The modelling workflow provided by the 'caret' R package was used to perform sample classification. The models were trained with an average number of 500 trees, automatic hyperparameter tuning to best adapt to data properties, 500-iteration 0.632 bootstrap resampling to avoid overfitting,²⁵ upsampling to maximize the robustness of the models against the class imbalance problem in datasets,²⁶ and recursive feature elimination to minimize the influence of non-informative features. Classification was performed in three different variable subsets: 1- Only relative metabolite concentrations, 2- Only signal chemical shifts and 3- Using both relative metabolite concentrations and signal chemical shifts. Results were evaluated using classification accuracy, Cohen's kappa (a more robust indicator against chance classification and class imbalance) and the area under the ROC (AUROC). In addition, to further evaluate the trained models, the sensitivity, specificity, positive predicted value and negative predicted value are available in Appendix. Lastly, the variable importance in the models generated with both sets of variables was measured.

5.2.4 Reproducibility of study workflow

To validate and reproduce the results, the profiling output, the data analysis workflow and the links for downloading the datasets analysed are available on github.com/danielcanueto/chemical_shift_classification.

5.3 Results

5.3.1 Exploratory visualization of PCA information

Visualization of the first two principal components (PCs) of the PCAs of metabolite concentrations and signal chemical shifts suggested higher discriminant power in chemical shift information (**Error! Reference source not found.**). In chemical shift figures, less ellipse overlap (or at least more separated centres) was observed. Although more discriminative power in concentration information might be present in later PCs, the noise-related variance might be able to mask this power more intensely. Also, no batch effects were visible on any dataset.

5.3.2 Classification results

- MTBLS1 dataset: Chemical shift information showed potential for discriminating between diabetic and non-diabetic samples during RF classification (AUROC 0.831) (Table 5.1). However, adding chemical shift information did not improve the excellent results obtained with only metabolite concentrations (AUROC 0.979).
- MTBLS237 dataset: Chemical shift information, alone or combined with metabolite concentration information, significantly improved sample discrimination in 6 of the 8 subgroup comparisons: Active UC vs Inactive UC (0.917 vs 0.811 in AUROC), Active UC vs Active CD (0.768 vs 0.743 in AUROC), Inactive UC vs Inactive CD (0.870 vs 0.810 in AUROC), Control vs Active UC (0.948 vs 0.914 in AUROC), Control vs Inactive UC (0.943 vs 0.823 in AUROC) and Control vs Inactive CD (0.854 vs 0.825 in AUROC) (Table 5.2).
- MTBLS374 dataset: RF classification on smoker and non-smoker samples showed much higher AUROC values with chemical shift information than with metabolite concentration (0.937 vs 0.856 in AUROC) (Table 5.3). The combination of both sources of information gave slightly better values than when only chemical shift information was used (AUROC 0.950; Table 5.3, left).

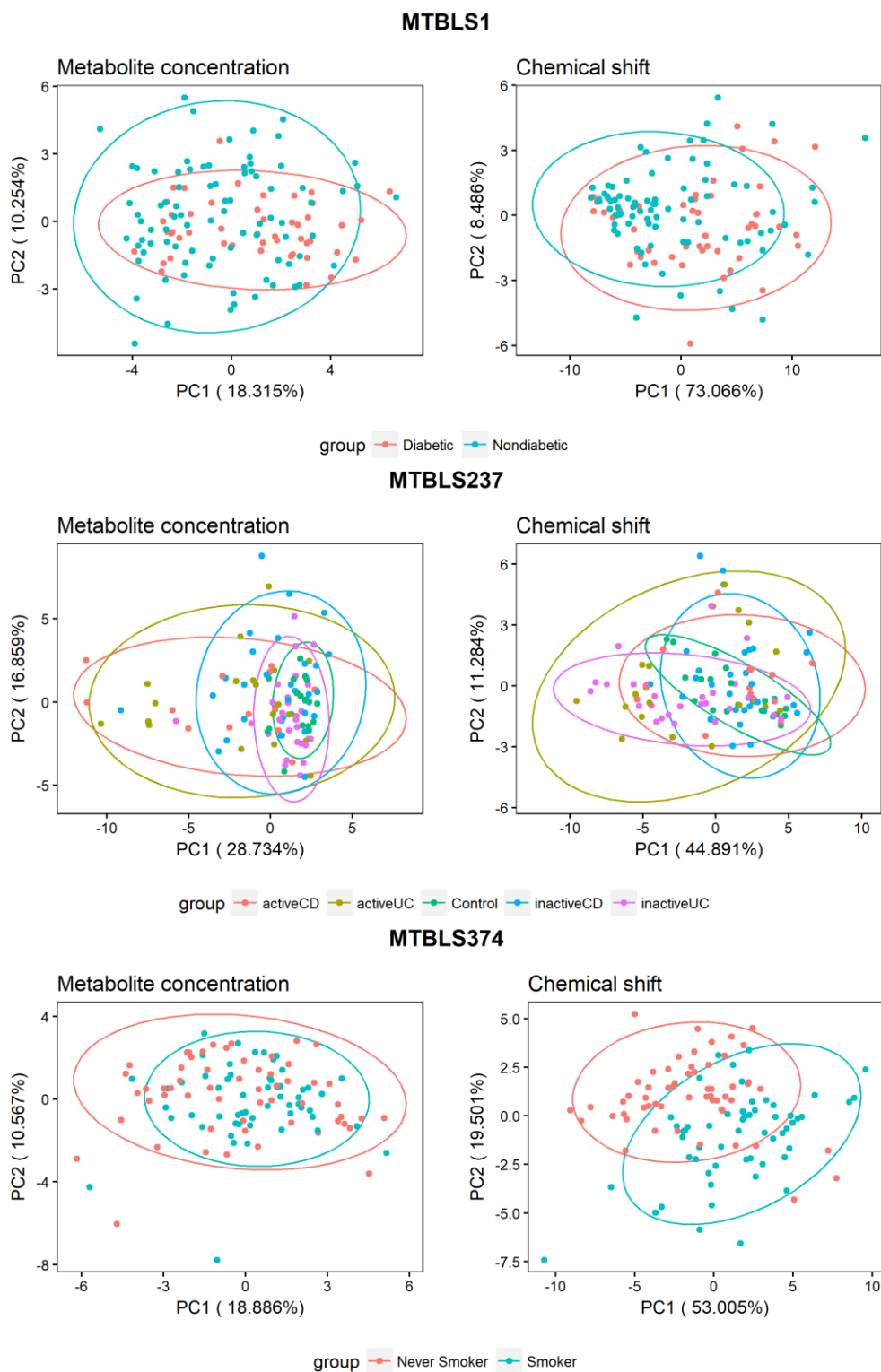


Figure 5-1 Exploratory PCA analysis shows the potential of the chemical shift data in the classification models. The first PCs of the PCA using chemical shifts (right) show better separation than the ones using concentrations (left). Plots also suggest no batch effects necessary to monitor.

	Both sets of information	Concentration information	Chemical shift information
Accuracy	0.929	0.933	0.795
kappa	0.840	0.849	0.559
AUROC	0.980	0.979	0.831

Table 5.1 Chemical shift information shows discriminative potential in the MTBLS1 dataset. However, it cannot enhance the excellent results given by concentration information during RF classification.

	Both sets of information	Concentration information	Chemical shift information
Active UC vs Inactive UC			
Accuracy	0.863	0.826	0.876
kappa	0.635	0.555	0.698
AUROC	0.870	0.811	0.917
Active CD vs Inactive CD			
Accuracy	0.801	0.808	0.721
kappa	0.505	0.526	0.331
AUROC	0.768	0.777	0.661
Active UC vs Active CD			
Accuracy	0.730	0.717	0.668
kappa	0.462	0.438	0.339
AUROC	0.768	0.743	0.682
Inactive UC vs Inactive CD			
Accuracy	0.808	0.771	0.797
kappa	0.617	0.545	0.594
AUROC	0.870	0.810	0.841

	Both sets of information	Concentration information	Chemical shift information
Control vs Active UC			
Accuracy	0.890	0.860	0.882
kappa	0.773	0.714	0.762
AUROC	0.948	0.914	0.926
Control vs Active CD			
Accuracy	0.867	0.861	0.790
kappa	0.719	0.707	0.556
AUROC	0.921	0.916	0.839
Control vs Inactive UC			
Accuracy	0.882	0.804	0.892
kappa	0.753	0.596	0.775
AUROC	0.926	0.823	0.943
Control vs Inactive CD			
Accuracy	0.806	0.787	0.782
kappa	0.589	0.550	0.551
AUROC	0.854	0.825	0.81

Table 5.2 Adding chemical shift information to concentration information improved the classification between the five different kinds of sample in the MTBLS237 dataset. Several quality indicators of the models generated are shown.

	Both sets of information	Concentration information	Chemical shift information
Accuracy	0.899	0.806	0.883
kappa	0.797	0.614	0.766
AUROC	0.950	0.856	0.937

Table 5.3 Adding chemical shift information to concentration information provides the best classification of samples in the MTBLS374 dataset. Several quality indicators of the models generated only with concentration information, only with chemical shift information and with both sources of information are shown.

5.4 Discussion

The results of our studies showed that 1D ¹H-NMR spectra chemical shift information can give greater insight into sample properties and improve sample classification. In the three datasets analysed, chemical shift information led to good sample classification. In addition, in two of them, chemical shift information helped gave AUROC values higher than 0.9 and improved the classification with only metabolite concentration information.

5.4.1 Relationship between chemical shift and metabolic alkalosis/acidosis

The high classification performance observed in the three study datasets seems to be consistent with what has been previously reported about the alkalosis or acidosis characteristics of the conditions in the associated studies.

The MTBLS1 dataset is associated with the study of the changes in human urine caused by type 2 diabetes. Type 2 diabetes mediates lower pH in urine as a result of greater net acid excretion and fewer ammonia buffers.²⁷ A lower pH increases the chemical shift of signals (i.e., the signal moves to the left in a spectrum).²⁸ Accordingly, most signals show a higher chemical shift in the diabetes samples than in the control samples (Figure 5-4; top). Several signal chemical shifts (such as one of indoxyl sulphate in Figure 5-4) show an inverse trend to the other signals. This inverse trend may be mediated by the influence of ionic strength. However, it may also be an artefact of the TSP signal used to reference spectra. The pKa of TSP is approximately 5, which makes its signal chemical shift sensitive to pH variation and causes signals with lower sensitivity (like the ones in the phenolic region²⁹) to seem to move in the opposite direction to other signals.

In the case of the MTBLS237 dataset, alkalosis/acidosis in inflammatory bowel disease (the subtypes of which are UC and CD) has been reported elsewhere in the literature.³⁰ The relationship between faecal pH and the disease could be influenced by the location of lesions and/or the complex acid-base balances. The pH disturbance could have manifested as acidic pH in the UC samples represented by a higher chemical shift (Figure 5-2, right; Figure 5-4, middle), and has been reported in the literature.³¹ As in the MTBLS1 dataset, several signal chemical shifts show an inverse trend that may be mediated by the use of the TSP signal to reference spectra (Figure 5-4; middle).

As for the MTBLS374 dataset, respiratory acidosis is typically seen in lung disease developed by smokers³² and in cigarette smoke that contains oxidants with acidic properties.³³ Signals in the spectra from the smokers group showed a higher chemical shift than the equivalent signals in the non-smokers (Figure 5-2, left; Figure 5-4, bottom). This effect might be mediated by a more acidic pH in smokers' samples as a consequence of smoking, which would be mostly captured by the second principal component of the PCA of signal chemical shifts (Table 5.4). Unlike the other two datasets, this dataset does not contain any signal chemical shift with an inverse trend. This is consistent with the reference signal being glucose, a metabolite with a pKa (approx. 12) that is quite different from the pH of biological samples and thus much more resilient to pH variability.

5.4.2 Effect of class-dependent signal misalignment on fingerprinting approaches

All the datasets evaluated were processed using fingerprinting approaches in the original studies, in contrast to the profiling approach used here. Fingerprinting approaches perform the classification by looking for significant spectral differences between groups and identifying the metabolites involved in the second stage. On the other hand, profiling approaches start by characterizing the metabolites in the samples and then performing statistical analysis in the second stage. Their different workflows imply variations in how metabolites are identified and how their concentrations are quantified.³⁴

Profiling is deemed to provide more resistance against signal overlap or baseline appearance through the deconvolution of signals in the spectrum lineshape.³⁵ However, one factor not evaluated in the differences between fingerprinting and profiling approaches is class-dependent signal misalignment (i.e., the differences in signal chemical shifts between spectra from different sample

classes). Fingerprinting reliability is based on the premise that signals are reasonably well-aligned throughout the spectra dataset and, consequently, the differences are caused by differences in metabolite concentrations. It has been theoretically demonstrated that classification in fingerprint data can be influenced by class-dependent signal misalignment (i.e., that the differences found between classes are actually caused by having the metabolite signals located in different bins). However, approaches to minimize this problem (like the use of signal alignment algorithms³⁶) are still not prevalent in the metabolomics field and were not applied in any of the datasets analysed.

In the three datasets analysed, the results of the univariate analysis in fingerprint data were compared before and after signal alignment using the CluPA algorithm³⁷ (the analysis workflow is available in Appendix). Signal alignment decreased the number of significant bins in all datasets (MTBLS374, -42%; MTBLS1, -7%; MTBLS237, -5%). This decrease means an improvement in the quality of classification models, as it can be ensured that the differences between classes are caused by potential biomarkers and not by signal misalignment.

Results confirmed the effect that class-dependent signal misalignment can exert on the results of fingerprinting data. Therefore, they further recommend the adoption of profiling approaches enabled by recent open-source profiling tools to minimize the generation of non-reproducible results. If the fingerprinting approach is still preferred, the implementation of signal alignment algorithms can minimise non-reproducible results; nonetheless, this alignment will involve losing the information given by chemical shift information.

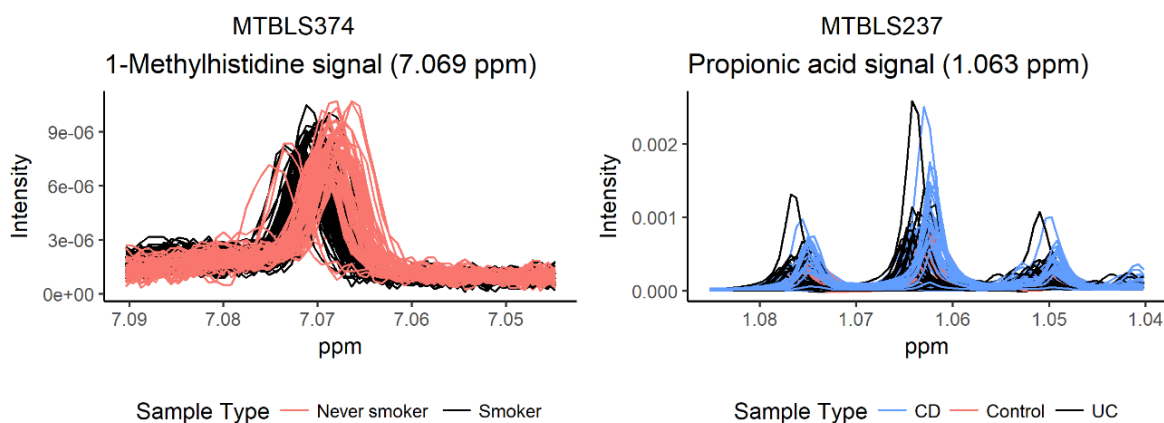


Figure 5-2 Signals can be misaligned in some sample classes. Low pH mediated by the condition studied increases the chemical shift of the signals. The resulting class-dependent signal misalignment can distort the results of the analysis of fingerprint data: features can show significant differences caused by differences in chemical shift (mediated by pH or ionic strength) rather than by differences in metabolite concentration.

5.4.3 Future directions and challenges

Our study workflow uses publicly available datasets and performs data pre-processing, profiling and statistical analysis with open-source tools following community recommendations.³⁸ By sharing this workflow, the hope is to make the use of chemical shift information in NMR studies more straightforward and more widespread. In addition, the resulting reproducibility might help assess some aspects that need to be considered to take maximum advantage of chemical shift information:

- Some matrices present considerable variations in dilution, which can greatly influence their pH and ionic strength (and, therefore, chemical shift). In addition, chemical shift variability is reduced by adding phosphate buffers (sometimes with added chelators such as EDTA) to the sample.³⁹ Both dilution variability and the use of buffers may mask the effects on the chemical shift produced by the condition studied. Consequently, the fact that the discriminative potential observed in MTBLS1 and MTBLS237 datasets was lower than the potential of the MTBLS374 dataset may be due to the higher dilution variability in the matrices studied (human urine and faecal extracts). The use of buffers or chelators should be minimized and sample dilution variability should be reduced if maximum advantage is to be taken of the properties of chemical shift information.
- It has been suggested that chemical shift information could also be translated to sample pHs and ionic concentrations, hence maximizing the information extracted from a dataset.⁸ Nonetheless, the limitations mentioned above raise concerns about the correct use of this information in several commonly studied matrices. In addition, the fact that these matrices commonly use a signal to reference spectra that is not resilient to pH (such as the TSP signal) may further distort the translation of chemical shifts to pH and ionic concentration values. There are several affordable techniques (e.g., pH meter or potentiometer) for directly measuring pH and ion concentrations that make this challenging translation unnecessary.
- Studies aiming to take advantage of chemical shift information should ensure consistent sample preparation and spectra acquisition in all samples to prevent the discrimination between sample classes being mediated by differences in the preparation or acquisition protocol.
- Further improvements in the quality of the classification models generated may be made by extracting more chemical shifts from NMR datasets and filtering noise in the chemical shift information (caused by low resolution with the consequent signal overlap in ¹H-NMR) prior to model training. High-resolution spectra (e.g., 2D NMR) could help isolate

more signals (with their associated chemical shifts) from different nuclei and prevent noise.

5.5 Achievements

- The reliable and optimized exploitation of the potential of chemical shift information to maximize the performance of the classification of samples during the multivariate analysis of metabolomics studies.
- The demonstration of the influence of the chemical shift variability in the results of fingerprint-based analyses of the difference between sample cases (and, therefore, of the further need to promote the development profiling approaches instead of fingerprint-based ones).

References

1. Lindon, J. C., Nicholson, J. K., Holmes, E. & Everett, J. R. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Reson.* 12, 289–320 (2000).
2. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171 (2002).
3. Bharti, S. K. & Roy, R. Quantitative 1H NMR spectroscopy. *Trends Analyt. Chem.* 35, 5–26 (2012).
4. Beger, R. D. et al. Metabolomics enables precision medicine: 'A White Paper, Community Perspective'. *Metabolomics* 12, (2016).
5. 1H NMR Spectrum (HMDB0000190). Human Metabolome Database: 1H NMR Spectrum (HMDB0000190) Available at: http://www.hmdb.ca/spectra/nmr_one_d/1162. (Accessed: 17th February 2018)
6. Dona, A. C. et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput. Struct. Biotechnol. J.* 14, 135–153 (2016).
7. Spraul, M. et al. Mixture analysis by NMR as applied to fruit juice quality control. *Magn. Reson. Chem.* 47 Suppl 1, S130–7 (2009).
8. Takis, P. G., Schäfer, H., Spraul, M. & Luchinat, C. Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nat. Commun.* 8, 1662 (2017).
9. Corbet, C. & Feron, O. Tumour acidosis: from the passenger to the driver's seat. *Nat. Rev. Cancer* 17, 577–593 (2017).
10. Galla, J. H. Metabolic alkalosis. *J. Am. Soc. Nephrol.* 11, 369–375 (2000).
11. Cloarec, O. et al. Evaluation of the Orthogonal Projection on Latent Structure Model Limitations Caused by Chemical Shift Variability and Improved Visualization of Biomarker Changes in 1H NMR Spectroscopic Metabonomic Studies. *Anal. Chem.* 77, 517–526 (2005)
12. Reily, M. D. et al. DFTMP, an NMR reagent for assessing the near-neutral pH of biological samples. *J. Am. Chem. Soc.* 128, 12360–12361 (2006).
13. Hao, J., Aistle, W., De Iorio, M. & Ebbels, T. M. D. BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 28, 2088–2090 (2012).
14. Gómez, J. et al. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. *Anal. Bioanal. Chem.* 406, 7967–7976 (2014).

15. Ravanbakhsh, S. et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One* 10, e0124219 (2015).
16. Cañueto, D., Gómez, J., Salek, R. M., Correig, X. & Cañellas, N. rDolphin: a GUI R package for proficient automatic profiling of 1D 1H-NMR spectra of study datasets. *Metabolomics* 14, (2018).
17. Haug, K. et al. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, D781–6 (2013).
18. Salek, R. M. et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiol. Genomics* 29, 99–108 (2007)
19. Bjerrum, J. T. et al. Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. *Metabolomics* 11, 122–133 (2014).
20. Kaluarachchi, M. R., Boulangé, C. L., Garcia-Perez, I., Lindon, J. C. & Minet, E. F. Multiplatform serum metabolic phenotyping combined with pathway mapping to identify biochemical differences in smokers. *Bioanalysis* 8, 2023–2043 (2016).
21. Emwas, A.-H. et al. Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis. *J. Proteome Res.* 15, 360–373 (2016).
22. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* 78, 4281–4290 (2006).
23. Efron, B. & Hastie, T. *Computer Age Statistical Inference.* (2016).
24. Gromski, P. S. et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* 879, 10–23 (2015).
25. Efron, B. & Tibshirani, R. Improvements on Cross-Validation: The .632 Bootstrap Method. *J. Am. Stat. Assoc.* 92, 548 (1997).
26. Kuhn, M. & Johnson, K. *Applied Predictive Modeling.* (2013).
27. Maalouf, N. M., Cameron, M. A., Moe, O. W. & Sakhaee, K. Metabolic basis for low urine pH in type 2 diabetes. *Clin. J. Am. Soc. Nephrol.* 5, 1277–1281 (2010).
28. Xiao, C., Hao, F., Qin, X., Wang, Y. & Tang, H. An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst* 134, 916–925 (2009).
29. Tredwell, G. D., Bundy, J. G., De Iorio, M. & Ebbels, T. M. D. Modelling the acid/baseH NMR chemical shift limits of metabolites in human urine. *Metabolomics* 12, 152 (2016).
30. Barkas, F., Liberopoulos, E., Kei, A. & Elisaf, M. Electrolyte and acid-base disorders in inflammatory bowel disease. *Ann. Gastroenterol. Hepatol.* 26, 23–28 (2013).

31. Vernia, P. et al. Fecal Lactate and Ulcerative Colitis. *Gastroenterology* 95, 1564–1568 (1988).
32. Broaddus, V. C. et al. *Murray & Nadel's Textbook of Respiratory Medicine*. (Elsevier Health Sciences, 2015).
33. Pryor, W. A. & Stone, K. Oxidants in cigarette smoke. Radicals, hydrogen peroxide, peroxyxynitrate, and peroxyxynitrite. *Ann. N. Y. Acad. Sci.* 686, 12–27; discussion 27–8 (1993).
34. Viant, M. R., Ludwig, C. & Günther, U. L. Chapter 2. 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling. *Metabolomics, Metabonomics and Metabolite Profiling* 44–70.
35. Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* 78, 4430–4442 (2006).
36. Vu, T. N. & Laukens, K. Getting your peaks in line: a review of alignment methods for NMR spectral data. *Metabolites* 3, 259–276 (2013)
37. Vu, T. N. et al. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* 12, 405 (2011).
38. Rocca-Serra, P. et al. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* 12, 14 (2016).
39. Li, N., Song, Y. P., Tang, H. & Wang, Y. Recent developments in sample preparation and data pre-treatment in metabonomics research. *Arch. Biochem. Biophys.* 589, 4–9 (2016).

5.6 Apendix

5.6.1 Filtering of unreliable metabolite relative concentrations and chemical shifts

- First, relative concentrations and chemical shifts of signal quantifications which did not pass a specific threshold in two quality indicators outputted by rDolphin (fitting error, signal area / total spectrum area ratio) were removed.
- Then, outliers for each signal area quantification (controlling by sample type) were removed.
- Next, relative concentrations and chemical shifts with 40% missing values or more were removed from the analysis.
- When more than one signal was quantified for a metabolite, the signal with the lowest number of missing values was considered to be the one able to provide the most accurate relative concentration and the quantification of the other signals was removed.
- Finally, missing values from chemical shifts and relative concentrations were imputed by RF methods.

5.6.2 Filtering of non-informative signal chemical shifts

Inaccurate chemical shift quantification in multiplet integration or lack of meaningful chemical shift variability mediated the presence of chemical shifts with noisy (i.e., non-informative) information in the dataset. Chemical shifts are correlated so an internal consistency in the chemical shift dataset is expected and this consistency can be measured. This internal consistency was analysed with the ‘psych’ R package. The chemical shifts which worsened the internal consistency of the chemical shift dataset were removed.

5.6.3 Univariate tests in non-aligned and aligned fingerprint data

The ‘p_values’ function of the [‘rDolphin’](#) R package contains the workflow that generates univariate tests for every bin. In two-sample tests, non-normality in every group of samples is

checked using Shapiro-Wilk tests. If a group of samples shows no normality, a Mann-Whitney test is performed; if all groups show normality a Welch t-test is performed (to understand why Welch and not Student's t-tests are performed, see this [link](#)). The p-values estimated in every study were then Benjamini-Hochberg adjusted.

5.6.4 Supplementary Tables

Predictors	Importance
PC2 - Chemical shift	100
PC1 - Chemical shift	56.4
Citric acid - quantification	32.181
U2_85 - quantification	24.541
PC3 - Chemical shift	24.39

Table 5.4 Ranked predictors in RF classification of samples with both con-centration and chemical shift information in the MTBLS374 dataset. There are few predictors because of the recursive feature ex-traction of non- discriminative features.

	Both sets of information	Concentration information	Chemical shift information
Sensitivity	0.838	0.855	0.713
Specificity	0.986	0.982	0.841
Pos Pred Value	0.968	0.963	0.723
Neg Pred Value	0.915	0.923	0.835

Table 5.5 Additional classification indicators in the MTBLS1 dataset.

	Both sets of information	Concentration information	Chemical shift information
Active UC vs Inactive UC			
Sensitivity	0.76	0.751	0.6
Specificity	0.842	0.882	0.806
Pos Pred Value	0.776	0.815	0.658
Neg Pred Value	0.837	0.842	0.755
Active CD vs Inactive CD			
Sensitivity	0.595	0.616	0.483
Specificity	0.909	0.909	0.839
Pos Pred Value	0.725	0.745	0.516
Neg Pred Value	0.838	0.845	0.789
Active UC vs Active CD			
Sensitivity	0.674	0.662	0.617
Specificity	0.798	0.783	0.721
Pos Pred Value	0.702	0.684	0.615
Neg Pred Value	0.781	0.769	0.725
Inactive UC vs Inactive CD			
Sensitivity	0.871	0.837	0.852
Specificity	0.847	0.816	0.851
Pos Pred Value	0.864	0.835	0.865
Neg Pred Value	0.861	0.824	0.846

	Both sets of information	Concentration information	Chemical shift information
Active UC vs Inactive UC			
Sensitivity	0.866	0.839	0.882
Specificity	0.915	0.885	0.891
Pos Pred Value	0.912	0.882	0.895
Neg Pred Value	0.881	0.858	0.888
Active CD vs Inactive CD			
Sensitivity	0.828	0.821	0.671
Specificity	0.901	0.897	0.885
Pos Pred Value	0.86	0.854	0.794
Neg Pred Value	0.883	0.878	0.802
Active UC vs Active CD			
Sensitivity	0.842	0.739	0.889
Specificity	0.917	0.863	0.897
Pos Pred Value	0.876	0.793	0.868
Neg Pred Value	0.894	0.829	0.921
Inactive UC vs Inactive CD			
Sensitivity	0.744	0.708	0.741
Specificity	0.852	0.848	0.82
Pos Pred Value	0.768	0.751	0.735
Neg Pred Value	0.841	0.824	0.832

Table 5.6 Additional classification indicators in the MTBLS237 dataset.

	Both sets of information	Concentration information	Chemical shift information
Sensitivity	0.893	0.810	0.870
Specificity	0.907	0.805	0.887
Pos Pred Value	0.906	0.809	0.886
Neg Pred Value	0.895	0.810	0.874

Table 5.7 Additional classification indicators in the MTBLS374 dataset.

5.6.5 Supplementary Figures

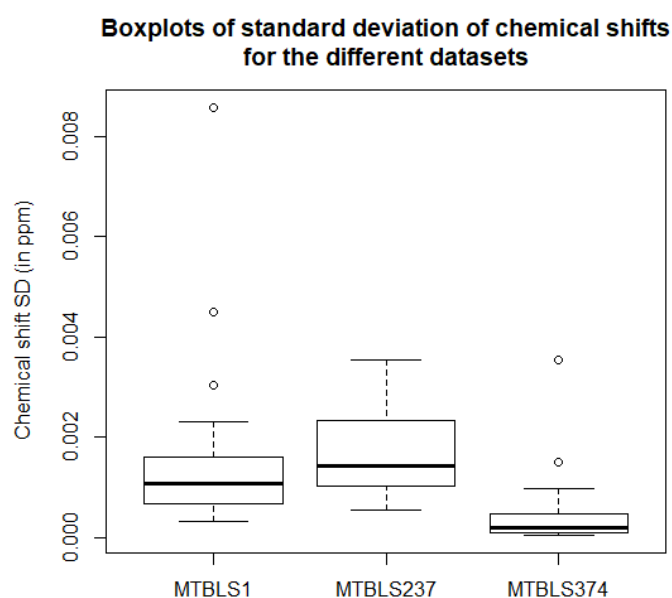


Figure 5-3 Variability (measured by standard deviation) of the chemical shifts analysed in the three datasets. As expected, the dataset of human matrices with higher dilution variability (urine and fecal extracts) show higher chemical shift variability. In all three datasets, the use of buffers does not impede the appearance of chemical shift variability that can be analysed.

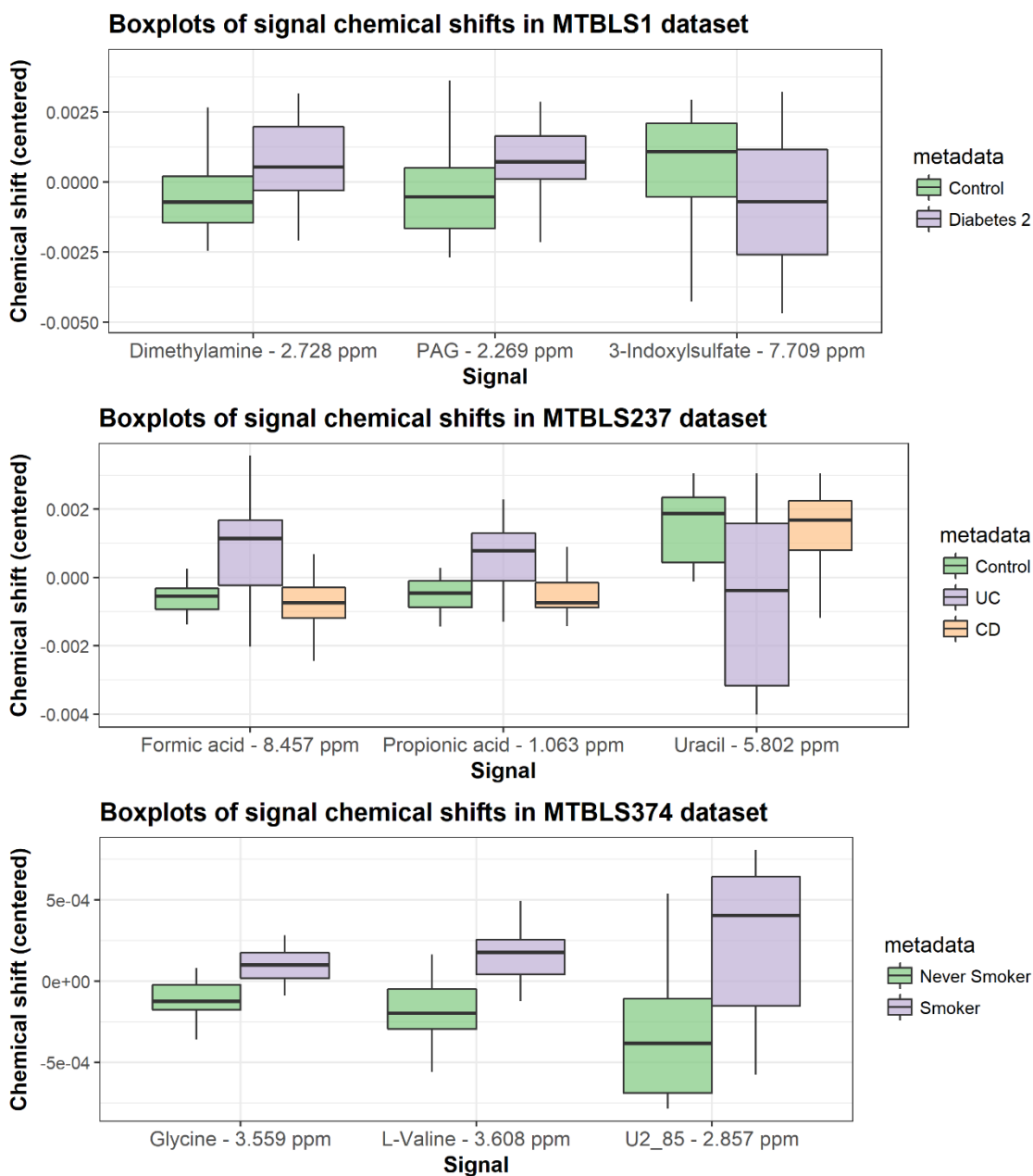


Figure 5-4 Distribution of centred chemical shift of three good chemical shift predictors in the MTBLS1 (top), MTBLS237 (middle) and MTBLS374 (bottom) datasets. Chemical shift patterns in the MTBLS1 and the MTBLS237 datasets showed higher complexity (with some signals with inverse trends) in the chemical shift mediated by the use of TSP as reference.

6 Maximizing the quality of NMR-based automatic metabolite profiling by predicting the expected metabolite signal parameters

Abstract

The quality of automatic metabolite profiling in NMR datasets can be compromised by the multiple sources of variability present in the samples of complex matrixes. These sources cause variability in the value of the metabolite signal parameters (e.g., half bandwidth, chemical shift). To monitor this variability efficiently and avoid suboptimal quantifications or wrong annotations, these tools may need to restrict their use to specific matrixes and strict sample protocols. However, the specific properties of each sample can be inferred from the signal parameters collected during a first profiling iteration as there is a multicollinearity in the signal parameter information which can be exploited to generate narrow and accurate predictions of the expected parameter values. In this study, it is demonstrated that these predictions can help generate better indicators of improvable quantifications than traditional indicators. In addition, the prediction information generated is used to maximize the performance of automatic profiling in a second iteration. Thanks to the ability of our profiling workflow to learn the sample properties, the prediction of signal parameters does not require prior information about the matrix or the protocol, therefore enabling an automatic profiling much more flexible to new matrixes and protocols and to the appearance of unexpected metabolites.

6.1 Introduction

Metabolomic studies characterize the low-molecular-weight components (<1 kDa) called metabolites in samples of biofluids or cell/tissue extracts.^{1,2} The quantification of the metabolite levels in nuclear magnetic resonance (NMR) spectra requires that the area below the metabolite signals to be quantified: this process is called metabolite profiling.^{3,4} This area can be quantified by area integration or signal deconvolution. In the case of 1D ¹H-NMR spectra, three signal parameters need to be estimated to deconvolute a signal: intensity, chemical shift and half bandwidth.³ Once the combination of parameter values that fits the spectrum lineshape with lowest error has been estimated, the signal can be built and the area below the signal can be quantified. Several tools have recently appeared which can automatically estimate signal parameter values.⁵⁻⁷ These tools are usually based on optimization solvers (e.g., the Levenberg-Marquardt algorithm) which evaluate the search space shaped by the range of possible values of each parameter to find a minimum that represents the replication of the spectrum lineshape with the lowest fitting error.⁸⁻⁹

However, automatic approaches are compromised by the multiple sources of variability which can be observed in complex matrices (e.g., macromolecule-based baseline, chemical shift and half bandwidth variability –caused by pH, ionic strength or temperature fluctuations or signal overlap¹⁰) (Figure 6-1 (a)). These sources of variability oblige the ranges of the possible parameter values to be wider during lineshape fitting and, therefore, the presence of a wide range of local minima where the optimization algorithm can meet the completion criteria and, therefore, end into suboptimal resolutions (Figure 6-1 (b)).¹¹ In addition, the possible presence of low-intensity signals adjacent to the ones of interest adds complexity to the spectrum lineshape. Consequently, optimization algorithms may not find the actual parameter values of the signals of interest but the ones which can best help replicate the complex lineshape. As a result of these challenges, automatic profiling tools sometimes provide wrong metabolite identifications (an important bottleneck in metabolomics¹²) and suboptimal quantifications. To reduce the generation of suboptimal fitting resolutions, several bioinformatic solutions can reduce the search space during optimization (e.g., the use of a CSI, the simultaneous lineshape fitting of all the signals from the same metabolite or the modelling of chemical shifts using multiple sources of information, among others¹³). However, these strategies are dependent on prior information. Therefore, they cannot handle unidentified metabolites and might be not robust to small variations in the expected lineshape (e.g., simultaneous lineshape fitting is prone to errors in case of chemical shift variability). Consequently, to ensure optimal performance, some tools can only be used in specific matrices or require restrictive procedures in sample preparation and/or spectrum acquisition. These matrix- and protocol-based restrictions hinder the high-throughput potential of NMR or might mean the incorporation of false positives and negatives into the metabolomics literature when these restrictions are not strictly followed.^{14,15}

To maximize the quality of lineshape fitting during NMR automatic profiling, the ranges of possible parameter values selected during fitting must be as narrow as possible. Likewise, the estimation of these narrow ranges must be robust to the variable and complex properties of metabolomics study datasets in complex matrices. This combination of narrowness and accuracy can be achieved if the information about the sample properties necessary to narrow the ranges is collected from the same dataset during an initial profiling iteration. NMR signals mediated by atoms with similar chemical environments show similar reactivity to the fluctuations in the sample conditions. As a result, there is extensive multicollinearity in their half bandwidth and chemical shift values. This multicollinearity can be exploited to identify signals whose parameters do not behave as expected by this multicollinearity. In addition, accurate spectrum-specific predictions with prediction intervals (PIs) for each signal parameter can be estimated according to the information from the collinear signals. These PIs may be used to create very narrow and accurate value ranges to be used during lineshape fitting in a new profiling iteration. Likewise, the intensities of the

signals from the same metabolite are perfectly collinear. Therefore, the expected intensity of each metabolite signal can also be predicted from the estimated intensities of the other metabolite signals. Consequently, the simultaneous lineshape fitting of all metabolite signals can be avoided. In contrast with other approaches, this prediction workflow is not dependent on prior matrix, protocol or metabolite information: this information is already encoded in the signal parameter values collected during the first profiling iteration. Therefore, it should be able to handle atypical or unidentified metabolites and be more robust to the sample-, matrix- or protocol-based complexities in the spectrum. In addition, the distance between the predicted parameter values and the collected parameter values can be quantified. The quantified distance may be a better profiling quality indicator than some of those in current use (e.g., fitting error) and help further minimise wrong annotations and suboptimal quantifications. To our knowledge, there has been no attempt to provide an open-source flexible automatic signal parameter prediction that maximizes the quality of the information provided by NMR profiling tools. In this study, it is shown how the proposed workflow helps maximize the quality of metabolite profiling in 1D ^1H -NMR datasets.

6.2 Materials and Methods

6.2.1 Datasets

For this study, two datasets were analysed: a faecal extract dataset of 146 samples from a medical treatment study and a serum dataset of 212 samples from a nutritional intervention study.

In the faecal extract dataset, sample collection and preparation and spectrum acquisition and pre-processing. Bucketing (6e-04 ppm as bucket width), referencing to TSP at 0 ppm and PQN¹⁶ were performed through rDolphin.¹⁷

In the serum dataset, sample collection details are available in Hernández-Alonso, P. *et al.*¹⁸ For each sample, 300 μl aliquots were mixed with 300 μl of sodium phosphate buffer. CPMG spectra, at 37°C and with presaturation to suppress the residual water peak, were acquired on a Bruker 600 MHz Spectrometer (Bruker Biospin, Rheinstetten, Germany) equipped with an Avance III console and a TCI CryoProbe Prodigy. CPMG data were pre-processed on the NMR console (TopSpin 3.2, Bruker Biospin, Rheinstetten, Germany) for overfilling, exponential line broadening (0.5 Hz) and phase correction. 0.0006 ppm binning and referencing to the anomer of glucose at 5.233 ppm were performed through rDolphin.¹⁷

In addition, mass spectrometry (MS) profiling data was collected from both datasets. Complete details regarding the MS profiling workflow used in both datasets are available in Apendix.

6.2.2 ¹H-NMR metabolite profiling workflow

Automatic metabolic profiling was performed using the rDolphin R package¹⁷, an open source tool which as well collects the values of the signal parameters and exports them for analysis. rDolphin performs a lineshape fitting based profiling which adjusts spectral regions to a sum of Lorentzian signals, each one of which is characterized by three parameters: intensity, chemical shift and half bandwidth. The fitting process is performed using the Levenberg-Marquardt Non-linear Least-Squares algorithm with lower and upper bounds provided by the 'minpack.lm' R package.¹⁹ The values of the algorithm parameters used during lineshape fitting are available in Apendix. To avoid falling into local minima, the fitting optimization is iterated a number of times proportional to the spectrum lineshape complexity, with signal parameter starting estimates that are randomly initialized for each iteration. After these iterations, the resolution with the least lineshape fitting error is chosen. After lineshape fitting, the areas below the signals are quantified, a specific fitting error for each signal is estimated (procedure explained in Apendix) and the signal parameter values are collected.

A graphical user interface (GUI) is used to select the metabolites to be profiled and the profiling method (area integration, signal deconvolution) for each of the signals. The GUI is also used to supervise the optimal value ranges for each chemical shift and half bandwidth to be used during lineshape fitting. In the case of chemical shift, the median range in both datasets was 0.006 ppm. In the case of half bandwidth, the median range was 50% of the median value. In the case of intensity, the tool automatically calculates the optimal value ranges by analysing the spectrum lineshape.

In the faecal extract dataset, 80 signals (66 through deconvolution and 14 through integration) from 52 different metabolites were profiled. In the serum dataset, 48 signals (43 fitted through deconvolution and 5 through integration) from 33 different metabolites were profiled. In addition, the signal parameter values and fitting errors were collected in both dataset profiling iterations.

6.2.3 Prediction pipeline of expected signal parameter values

After profiling, the collected and outputted signal parameters were used to predict, in each signal, the expected spectrum-specific values (with their PIs) according to the information present in the other signals. These predictions can be used in future steps to evaluate the results and improve the fitting.

To make the spectrum-specific prediction of a signal parameter, the values of the parameter in the other signals are collected to create a dataset of predictors. Then, to enrich the quality of this dataset of predictors, three common steps in machine-learning processes are applied successively: data cleaning to minimize the influence of inaccurate values, feature selection, and feature engineering.²⁰ After these enrichment steps, the signal parameter is predicted using the enriched dataset of predictors during the training of a random forest (RF) based prediction model. The RF algorithm is an ensemble learning method based on the bootstrap aggregation (also called *bagging*) of decision trees.^{21,22} The RF algorithm solves the main drawback of bagging trees (the tendency to create similar decision trees with highly correlated predictions) by adding randomness to the tree construction process. RF models the possible nonlinear factors and showed higher performance during exploratory data analysis and lower variance during prediction. In addition, 0.632 bootstrap resampling is applied to minimize overfitting.²³ Then, for each spectrum, the distribution which best represents the predictions generated during the bootstrap (see (Figure 6-1 (c))) is estimated. From this distribution, the median value (with 95% PIs) is outputted as the spectrum-specific predicted value in the signal parameter analysed (Figure 6-1 (d)). The complete details of the prediction pipeline as well as the specifications regarding intensity prediction are available in Appendix.

It was considered that, if the predictions of parameters were not spectrum-specific, the best possible prediction of this parameter would consist of the median value found for this parameter in all spectra, having as 95% PIs the 95% central distribution of values. Accordingly, to evaluate the narrowness achieved in the spectrum-specific predictions generated, for each signal and parameter, the ranges of the 95% PIs of the spectrum-specific and the spectrum-unspecific predictions were compared.

In addition, a quality indicator based on the difference between the predicted signal parameters and the parameters obtained during profiling was calculated. For each one of the signal parameters with available information, the absolute difference was normalized to 0-1. Subsequently, the values obtained for each signal of each spectrum were averaged. As a result, a 0-1 'anomaly score' was generated, which parameterizes how anomalous the signal parameter values obtained during profiling are.

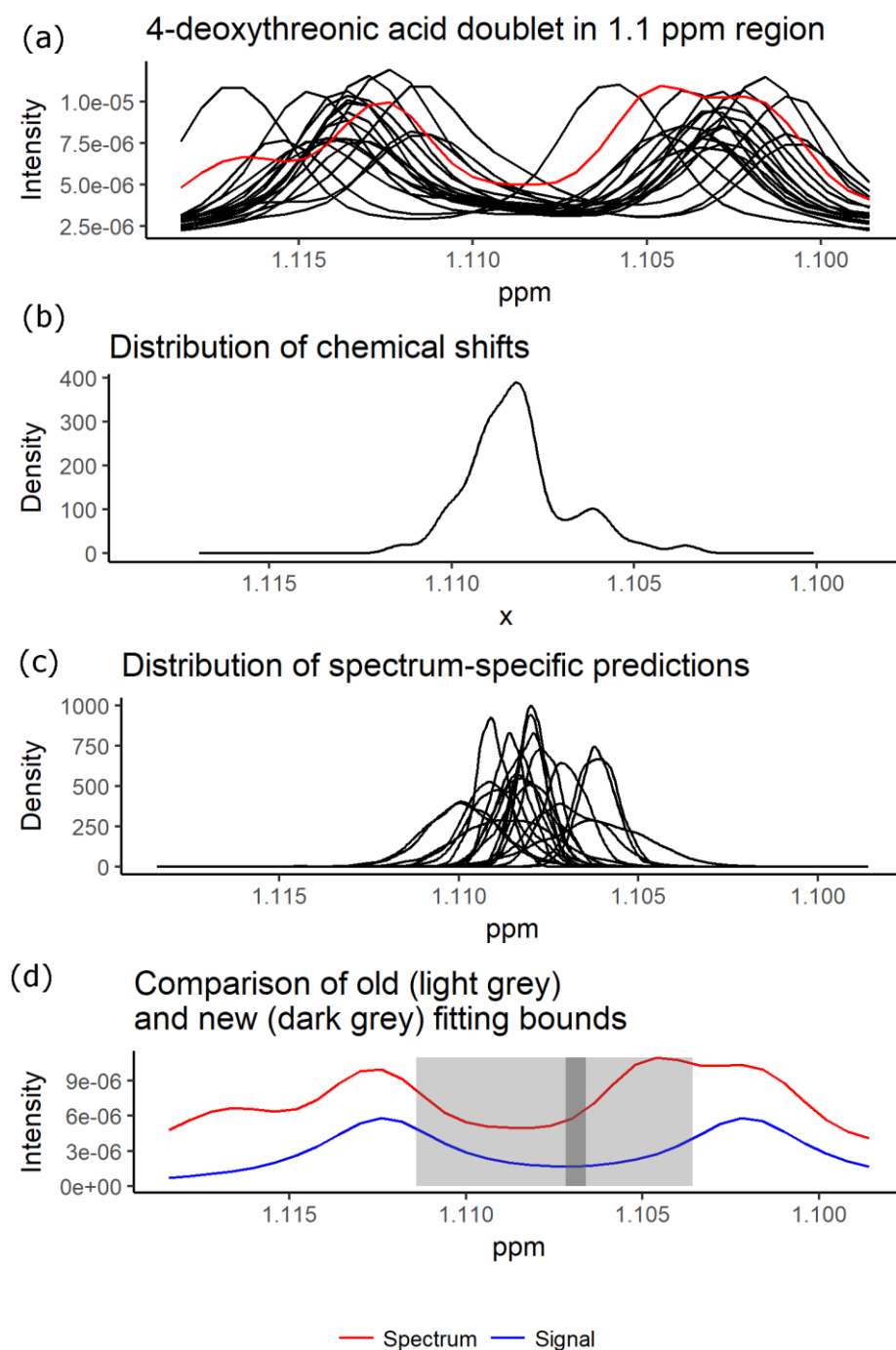


Figure 6-1 The signal parameter prediction pipeline enables narrow and accurate spectrum-specific ranges to be estimated and used during lineshape fitting. The figure shows a difficult signal fitting found with the 4-deoxythreonic acid signal in the urine dataset analysed in Appendix. The chemical shift variability present in this signal (a) forces lineshape fitting algorithms to consider a wide range of possible chemical shift values during the fitting (b). Excessive width can compromise the right assignment of the doublet center when other signals appear adjacent to the signal to be fitted (d). The chemical shift prediction generates spectrum-specific chemical shift distributions of predictions (c). These distributions are very narrow and can help generate much narrower chemical shift ranges (d).

6.2.4 Evaluation of improvement in profiling data quality

The presence of both MS and NMR data made it possible to parameterize the improvement in the quality of profiling data. In both platforms, the concentration of 15 metabolites in the faecal extract dataset and 11 metabolites in the serum dataset was determined. Improvements in profiling quality in NMR data should be associated with an increase in Spearman's rank correlation between the metabolite concentrations collected in NMR data and the ones collected in MS data.

This indicator of profiling data quality was used to evaluate the profiling improvement after a new profiling iteration had been performed using the data of the predicted signal parameters. If the narrow and accurate PIs of signal parameters are used as parameter value ranges during fitting, more accurate resolutions during lineshape fitting should be expected.

In addition, the fitting error and the anomaly score were compared as quality indicators. To make this comparison, the worst quantification from the first profiling iteration according to the quality indicator was identified and corrected by its equivalent in the new profiling iteration. Then, the mean Spearman's rank correlation between MS and NMR data was recalculated and the next worst quantification was identified. This process was iterated until all quantifications from the first profiling iteration had been corrected. It was expected that the better the quality indicator was the more able it would be to identify the quantifications to be corrected, so fewer corrections would be required to meaningfully increase the MS/NMR correlation.

6.3 Results

6.3.1 Accurate predicted values with narrow PIs which can be used to maximize profiling performance

The predictions generated (like the one in Figure 6-1 (d)) showed narrow spectrum-specific PIs for all the signal parameters analysed. For chemical shift, the median range in the spectrum-specific 95% PIs calculated in the faecal extract dataset was $4.7e-04$ ppm. This value is lower than the bucket width ($6e-04$ ppm) and is a reduction of 75.8% in the median range in the spectrum-unspecific 95% PIs ($1.9e-03$ ppm) (Figure 6-2; top left). In the serum dataset, the median range in the spectrum-specific 95% PIs calculated was $1.9e-04$ ppm, a reduction of 87.1% in the median range in the spectrum-unspecific 95% PIs ($1.4e-03$ ppm) (Figure 6-2; down left).

For half bandwidth, the median range in the spectrum-specific 95% PIs calculated in the faecal extract dataset was 8.6% of the predicted half bandwidth. This value is a reduction of 58.4% in the median range in the spectrum-unspecific 95% PIs (20.6% of the predicted half bandwidth) (Figure 6-2; top middle). In the serum dataset, the median range in the spectrum-specific 95% PIs calculated was 4.0% of the predicted half bandwidth, a reduction of 80.3% in the median range in the spectrum-unspecific 95% PIs (20.1% of the predicted half bandwidth) (Figure 6-2; down middle).

For intensity, the median range in the spectrum-specific 95% PIs calculated in the faecal extract dataset was 22.2% of the predicted intensity. This value is a reduction of 92.8% in the median range in the spectrum-unspecific 95% PIs (309.9% of the predicted intensity) (Figure 6-2; top right). In the serum dataset, the median range in the spectrum-specific 95% PIs calculated was 6.9% of the predicted intensity, a reduction of 93.3% in the median range in the spectrum-unspecific 95% PIs (102.9% of the predicted intensity) (Figure 6-2; down right).

Apart from showing narrow PIs, the predictions also helped maximize profiling performance when they were used in a new profiling iteration. When all quantifications were corrected with the predicted information to improve the quality of the lineshape fitting, mean Spearman's rho between MS and NMR metabolite concentrations increased 0.024 points (from 0.706 to 0.730) in the faecal extract dataset (Table 6.1; top) and 0.035 points (from 0.672 to 0.707) in the serum dataset (Table 6.1; down). Of the 25 correlations analysed, 21 of them increased their rho values, the maximum increase being 0.136 points and the maximum decrease 0.038 points. Rho improvements were especially important in the metabolites with the lowest correlation between the quantifications of both platforms: in the faecal extract dataset, the lowest rho value increased from 0.547 to 0.632; in the serum dataset, the lowest rho value increased from 0.515 to 0.589.

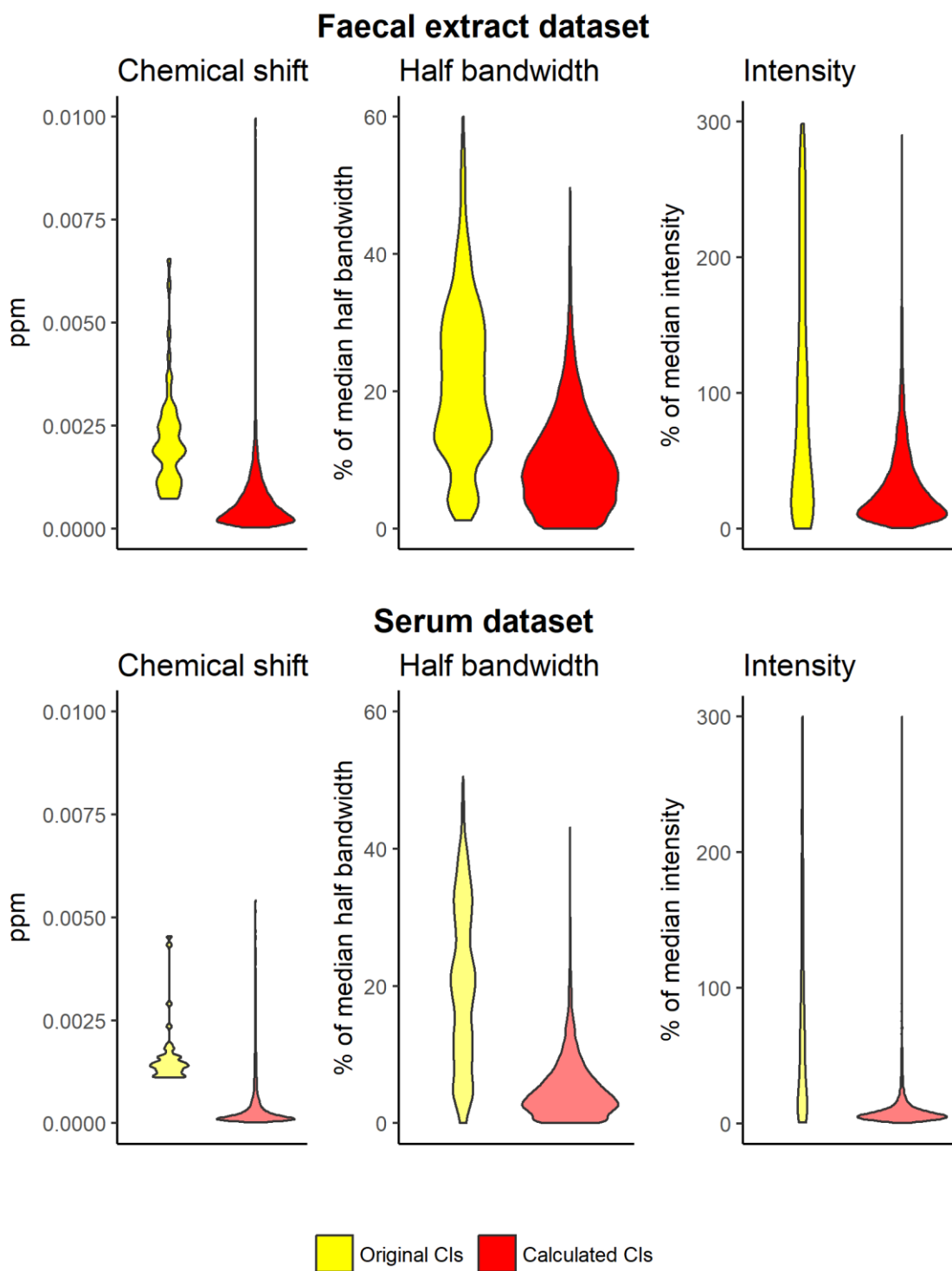


Figure 6-2 The spectrum-specific 95% PIs of the parameter values PIs are much narrower than the spectrum-unspecific 95% PIs. Chemical shift PIs are generally lower than the bucketing applied ($6e-4$ ppm). The narrow PIs enhance the performance of error minimization algorithms to end in the right local minimum.

Metabolites in faecal extract	Original Spearman's rho correlation	Spearman's rho correlation after new profiling iteration
L-Isoleucine	0.513	0.671
D-Glucose	0.556	0.589
Glycine	0.576	0.605
L-Phenylalanine	0.585	0.593
L-Leucine	0.659	0.690
L-Valine	0.672	0.699
Citric acid	0.690	0.695
L-Tyrosine	0.698	0.722
L-Alanine	0.726	0.737
3-Hydroxybutyric acid	0.846	0.872
Lactic acid	0.871	0.901

Metabolites in serum	Original Spearman's rho correlation	Spearman's rho correlation after new profiling iteration
Glycine	0.547	0.658
Hexanoic acid	0.556	0.677
5-aminovaleric acid	0.608	0.681
L-Isoleucine	0.629	0.632
L-Alanine	0.656	0.662
L-Valine	0.664	0.655
L-Leucine	0.697	0.715
2,4-Dihydroxypyrimidine	0.718	0.744
Succinic acid	0.735	0.730
Nicotinic acid	0.765	0.784
Phenylacetic acid	0.794	0.799
Glycerol	0.809	0.771
3-Phenylpropionic acid	0.844	0.873
Lactic acid	0.862	0.832

Table 6.1 *The predicted signal parameter information increases Spearman's rho correlation between metabolite concentrations in MS and NMR data in both datasets. There is a consistent increase in this profiling quality indicator when a new profiling iteration is performed using the PIs as new value ranges during lineshape fitting. The increase is most significant in the metabolites whose profiling was most complicated in the original profiling iteration.*

6.3.2 High accuracy of the calculated anomaly score for detecting improvable quantifications

To parameterize the performance of the fitting error and the calculated anomaly score as quality indicators of quantification, the quality of quantifications was ranked using these quality indicators. This ranking was then used to gradually replace the worst ranked quantification in each metabolite by the equivalent one obtained in the new higher-quality profiling iteration. It was expected that this gradual replacement of quantifications would improve Spearman's rank correlation between MS and NMR data with a logarithmic-like trend, as the first improved quantifications would provide the highest increases in the MS/NMR correlation.

In both datasets, the anomaly score as a quality indicator showed a logarithmic shape (Figure 6-3). Increase in the MS/NMR data correlation stopped improving after correcting approximately the 50% of the quantifications with worst anomaly score. Therefore, the anomaly score showed effectiveness at ranking the quantifications which might be further optimized. In comparison with the anomaly score, the fitting error showed a general lower effectiveness to detect improvable quantifications (as shown by the less logarithmic trend -**Error! Reference source not found.**-).



Figure 6-3 The calculated anomaly score helped identify quantifications which might be further optimized. In both datasets, the anomaly score showed higher performance than the fitting error ranking the quantifications which, if further improved, might further enhance the MS/NMR correlation.

The only subset of quantifications in which the fitting error performed better than the anomaly score was the worst quantifications in the faecal extract dataset (**Error! Reference source not found.**; top). This seems consistent with the high importance of the intensity information in the fitting error compared to half bandwidth or chemical shift information. In the faecal extract dataset, high coefficient of variation and possible fitting of adjacent signals are challenges. Accordingly, occasional high distortions of estimated intensity can be found which are better parameterized by the fitting error. However, after detecting these extreme suboptimal quantifications, the fitting error would be less able than the anomaly score to find quantifications where the characterization of the signal does not behave as expected.

6.4 Discussion

The results of the study showed that predicting signal parameter values with the information collected during a first profiling iteration helps maximize profiling performance. The improvement shown in this study has been demonstrated in biologically complex matrices and not in spike-in samples which cannot fully reproduce the usual complexity of metabolomics studies. Our study also presents a new quality indicator based on the information generated by our machine-learning-based pipeline. This new quality indicator, called the anomaly score, may provide higher-quality information to improve the detection of suboptimal quantifications and enable the detection of wrong annotations, two current bottlenecks in metabolomic studies which contribute to the introduction of false positives and negatives into the metabolomics literature.^{12,14,15} In addition, our machine-learning-based pipeline (contained in the ‘signparpred’ function in the ‘rDolphin’ R package) can be exported to any other profiling tool in any other programming language.

The great benefits of our approach are mediated by the generation of predictions specific to each signal and each spectrum with accurate and narrow PIs. These high-quality predictions ensure that the algorithmic minimization of the signal fitting error prevents the pervasive problem of falling into wrong local minima when numerous parameter values are optimized (dozens of parameters in the case of complex lineshape fittings). Other approaches try to minimize this problem by creating narrow value ranges prior to profiling. However, when dealing with complex matrices, they may have limitations such as:

- Strict sample preparation or spectrum acquisition: difficulty of changing established protocols in labs, less flexibility to adapt the spectrum acquisition process to the properties of samples.
- Half bandwidth and chemical shift prediction: broadening of TSP signal mediated by protein, nonlinear patterns in certain signals in complex matrices, inability to handle unidentified metabolites.¹⁰
- Simultaneous lineshape fitting of all the signals of a same metabolite: variability in the relative intensity of signals depending on the matrix, challenges when signal chemical shift is not predicted exactly, inability to handle unidentified metabolites.
- Algorithm-based signal alignment: signal distortion, wrong annotations.^{24,25}

In contrast, our approach is not dependent on restrictions or extensive previous information about signal properties: it only needs a flexible first profiling iteration that collects information for accurately characterizing the properties of the metabolite signals profiled and of the sample analysed. So, our approach provides a solution to the limitations listed above. Besides, the information obtained about the signal parameters of unidentified metabolites can be studied to find annotated signals with similar patterns (and consequently create valuable inferences about their structure and properties).

The maximization of the profiling quality shown in the results was not associated with a correlated decrease in the signal fitting error (the standard quality indicator outputted by NMR profiling tools). The mean fitting error of quantifications increased 0.26% in the faecal extract dataset and decreased 0.02% in the serum dataset. This suggests a ceiling in the performance of lineshape fitting approaches when matrices are complex. For example, they may give little importance to the lower intensity signals in the region analysed or not fully monitor the high-intensity baseline present e.g. in serum. Fitting information parameterizes not metabolite properties but spectrum properties. In contrast, the information generated with our prediction pipeline parameterizes metabolite properties. As a result, the new information generated by this workflow leads to next-generation quality indicators which are able to e.g. monitor wrong annotations because the associated chemical shift signal is not consistent with the information present in the whole dataset. This kind of quality control has the potential to filter out suboptimal quantifications more effectively. Consequently, it may be possible to profile many more metabolites without decreasing the profiling data quality.

The variability of chemical shift is one of the biggest challenges to progress in the automatic profiling of NMR datasets, and the PIs achieved during prediction tend to be even lower than the bucket width chosen. Thanks to this accurate chemical shift prediction, signals can be correctly

assigned and the lineshape fitting performance maximized. The fact that chemical shift can be accurately predicted in faecal extract, a matrix with considerable variability in chemical shift and signal overlap, suggests that accurately predicting chemical shift in human urine is achievable. This matrix is of great interest to metabolomics. However, its complexity makes robust automatic profiling a real challenge, and it is recommended that some tools are not used in this matrix. A promising technique for maximizing the quality of NMR profiling in human urine through chemical shift prediction has recently been published.¹³ Nonetheless, this technique cannot be exported to NMR profiling tools because of licensing restrictions and it requires strict sample preparation and spectrum acquisition criteria. The ML pipeline proposed, when tuned to the special conditions of human urine and validated by comparison with MS data, may be a generalizable solution to the signal misalignment problem in human urine. In Appendix, it is shown the current results in a human urine dataset (not validated through MS data).

6.4.1 Future directions

The benefits of our approach should also be observed in 2D NMR spectra and it may help solve some of their current limitations. Current use of 2D for quantitative purposes can be hindered by the lower proportionality between signal volumes and metabolite concentrations.²⁶ This lower proportionality is mediated by the much higher complexity of the pulses used during spectrum acquisition and by the requirement of long experiment times which may lead to greater noise in the acquired data.²⁷ Prediction of the signal properties may help increase this proportionality and expand its quantitative potential. It is plausible the workflow performed could also be helpful to solve the challenges observed in the profiling of datasets of other platforms such as MS. In MS, there are certain biological and technical factors that can interfere with the signal parameter values.²⁸ At present there is considerable interest in solving the challenges present in these datasets. The collection of signal parameter values and the use of our approach may help to this purpose.

6.5 Achievements

- The narrow and accurate prediction of the expected signal parameters in a dataset thanks to information previously collected from this dataset. This achievement liberates profiling tools of the need of requiring prior information about the matrix to study, the metabolites to profile or the sample acquisition or study protocol performed during the study. In

addition, the need for restrictions during sample preparation, spectrum acquisition or matrix to analyse is overcome.

- The improvement of automatic metabolite profiling thanks to the narrow and accurate estimation of ranges of possible parameter values to consider during the lineshape fitting of signals.
- The generation of indicators of possible wrong annotations and improvable quantifications of metabolite concentration of higher quality than the standard ones in lineshape fitting (i.e., fitting error). These indicators are based on the study of the difference between the expected signal parameter value and the obtained one.

References

1. Holmes, E., Wilson, I. D. & Nicholson, J. K. Metabolic Phenotyping in Health and Disease. *Cell* 134, 714–717 (2008).
2. Nicholson, J. K. Global systems biology, personalized medicine and molecular epidemiology. *Mol. Syst. Biol.* 2, 52 (2006).
3. van Duynhoven, J., van Velzen, E. & Jacobs, D. M. Quantification of Complex Mixtures by NMR. in *Annual Reports on NMR Spectroscopy* 181–236 (2013).
4. Fiehn, O. Metabolomics — the link between genotypes and phenotypes. in *Functional Genomics* 155–171 (2002).
5. Hao, J. et al. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* 9, 1416–1427 (2014).
6. Gómez, J. et al. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. *Anal. Bioanal. Chem.* 406, 7967–7976 (2014).
7. Ravanbakhsh, S. et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS One* 10, e0124219 (2015).
8. Roweis, S. Levenberg-Marquardt Optimization, <http://www.cs.nyu.edu/roweis/notes/lm.pdf>
9. Kanzow, Christian, Nobuo Yamashita, and Masao Fukushima. 2004. “Levenberg–Marquardt Methods with Strong Local Convergence Properties for Solving Nonlinear Equations with Convex Constraints.” *Journal of Computational and Applied Mathematics* 172 (2): 375–97.
10. Dona, A. C. et al. A guide to the identification of metabolites in NMR-based metabolomics/metabolomics experiments. *Comput. Struct. Biotechnol. J.* 14, 135–153 (2016).
11. Pardalos, P. M. & Edwin Romeijn, H. *Handbook of Global Optimization*. (Springer Science & Business Media, 2013).
12. van der Hoof, J. J. J. & Rankin, N. Metabolite Identification in Complex Mixtures Using Nuclear Magnetic Resonance Spectroscopy. in *Modern Magnetic Resonance* 1–32 (2016).
13. Takis, P. G., Schäfer, H., Spraul, M. & Luchinat, C. Deconvoluting interrelationships between concentrations and chemical shifts in urine provides a powerful analysis tool. *Nat. Commun.* 8, 1662 (2017).
14. Baran, R. Untargeted Metabolomics Suffers from Incomplete Data Analysis. (2017). doi:10.1101/143818
15. Sokolenko, S. et al. Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1H-NMR spectra in synthetic mixtures and urine

- with additional insights on choice of pulse sequences and robotic sampling. *Metabolomics* 9, 887–903 (2013).
16. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* 78, 4281–4290 (2006).
 17. Cañueto, D., Gómez, J., Salek, R. M., Correig, X. & Cañellas, N. rDolphin: a GUI R package for proficient automatic profiling of 1D 1H-NMR spectra of study datasets. *Metabolomics* 14, (2018).
 18. Hernández-Alonso, P. et al. Changes in Plasma Metabolite Concentrations after a Low-Glycemic Index Diet Intervention. *Mol. Nutr. Food Res.* e1700975 (2018).
 19. Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess and Ben Bolker (2016). minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. R package version 1.2-1. <https://CRAN.R-project.org/package=minpack.lm>
 20. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (2013).
 21. Efron, B. & Hastie, T. *Computer Age Statistical Inference*. (2016).
 22. Gromski, P. S. et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* 879, 10–23 (2015).
 23. Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* 78, 316–331 (1983).
 24. Savorani, F., Tomasi, G. & Engelsen, S. B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.* 202, 190–202 (2010).
 25. Vu, T. N. et al. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* 12, 405 (2011).
 26. Giraudeau, P. Challenges and perspectives in quantitative NMR. *Magn. Reson. Chem.* 55, 61–69 (2017).
 27. Giraudeau, P. Quantitative 2D liquid-state NMR. *Magn. Reson. Chem.* 52, 259–272 (2014).
 28. Vinaixa, M. et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *Trends Analyt. Chem.* 78, 23–35 (2016).

6.6 Apendix

6.6.1 Workflows of MS profiling data

6.6.2 Values of algorithm parameters used during lineshape fitting

Standard algorithm parameters used during lineshape fitting are available at [this link](#). The following parameters were tweaked to maximize quality/speed performance:

- maxiter=500
- ftol=1e-6
- ptol=1e-6
- factor=0.01

6.6.3 Signal-specific lineshape fitting error calculation

1. The spectrum region with the 90% central area below the quantified signal is identified.
2. The root mean squared error from the linear model between the spectrum region lineshape and the fitted lineshape is estimated.
3. The root mean squared error is normalized by the maximum of the spectrum region lineshape.

6.6.4 Results in urine dataset

6.6.4.1 Creation of narrow spectrum-specific PIs

For chemical shift, the median range in the spectrum-specific 95% PIs calculated was 5.69e-04 ppm. This value is lower than the bucket width (6e-04 ppm) and is a reduction of 87.16% in the median range in the spectrum-unspecific 95% PIs (4.43e-03 ppm) (Figure 6-4; left).

For half bandwidth, the median range in the spectrum-specific 95% PIs calculated was 9.66% of the predicted half bandwidth. This value is a reduction of 57.32% in the median range in the spectrum-unspecific 95% PIs (22.62% of the predicted half bandwidth) (Figure 6-4; middle).

For intensity, the median range in the spectrum-specific 95% PIs calculated was 13.42% of the predicted intensity. This value is a reduction of 92.79% in the median range in the spectrum-unspecific 95% PIs (186.03% of the predicted intensity) (Figure 6-4; right).

6.6.4.2 Analysis of coefficient of variation after profiling improvement

The coefficient of variation is a quality indicator of profiling quality (the lower the noise added during profiling, the lower the coefficient of variation). The mean lowering in the coefficient of variation after profiling improvement based on prediction information was 7.8%. In certain metabolite signals, the coefficient of variation decreased more than 25%.

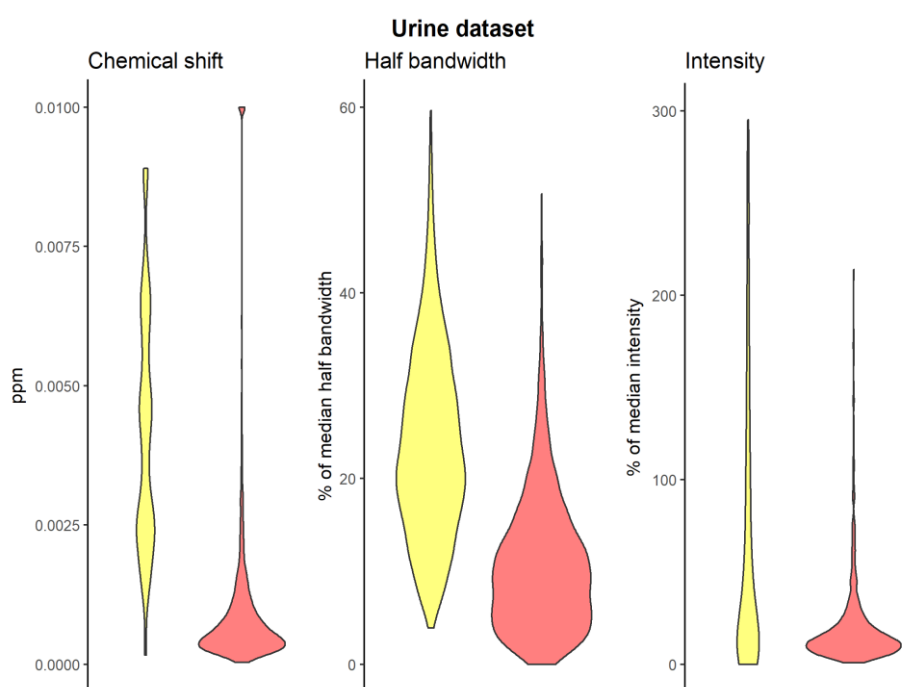


Figure 6-4 The spectrum-specific 95% PIs of the parameter values PIs are much narrower than the spectrum-unspecific 95% PIs. Chemical shift PIs are generally lower than the bucketing applied ($6e-4$ ppm). The narrow PIs enhance the performance of error minimization algorithms to end in the right local minimum.

6.6.5 Signal parameter prediction pipeline

Figure 6-5 shows the signal parameter prediction pipeline when predicting chemical shift and half bandwidth information. Each signal parameter is predicted using the values in the other signal parameters. To maximize the quality of prediction, the steps listed below are followed:

1. A training dataset is built which contains as features all the signal parameters except the ones to be predicted.
2. Data cleaning is applied to the predictor dataset to minimize the influence of inaccurate values in the features (because of wrong annotation or suboptimal quantification) during prediction. Signal parameter values associated with suboptimal fitting error (>0.05) are removed and values are imputed through RF methods.
3. The predictor dataset is enriched by feature selection and feature engineering, two common steps in ML-based pipelines.¹³ Feature engineering is applied by adding the first five PCs of the signal parameter dataset to the predictor dataset. A PCA concentrates the informative features in the first PCs and relegates noise-related variance to later PCs. Consequently, if there is high noise-related variance in the dataset, the addition of these first PCs can enhance prediction performance. Feature selection using the ‘Boruta’ R package is applied to filter non-relevant features to reduce the noise in the dataset.
4. For each signal parameter, a model is trained to perform the signal parameter prediction using the enriched dataset. The 0.632 bootstrap resampling provided by the ‘caret’ R package is applied to minimize overfitting.¹⁶ Then, the distribution which best represents the generated predictions for each signal and each parameter is estimated. From this distribution, the median value (with 95% PIs) is outputted as the predicted value.

When intensity information is predicted, the training dataset consists only of the information provided by other signals from the same metabolite. To adapt to the lower number of predictors, the quality of the predictor dataset is enriched by weighting according to the fitting error and by adding the first PC of the PCA of all the signals of the metabolite analysed.

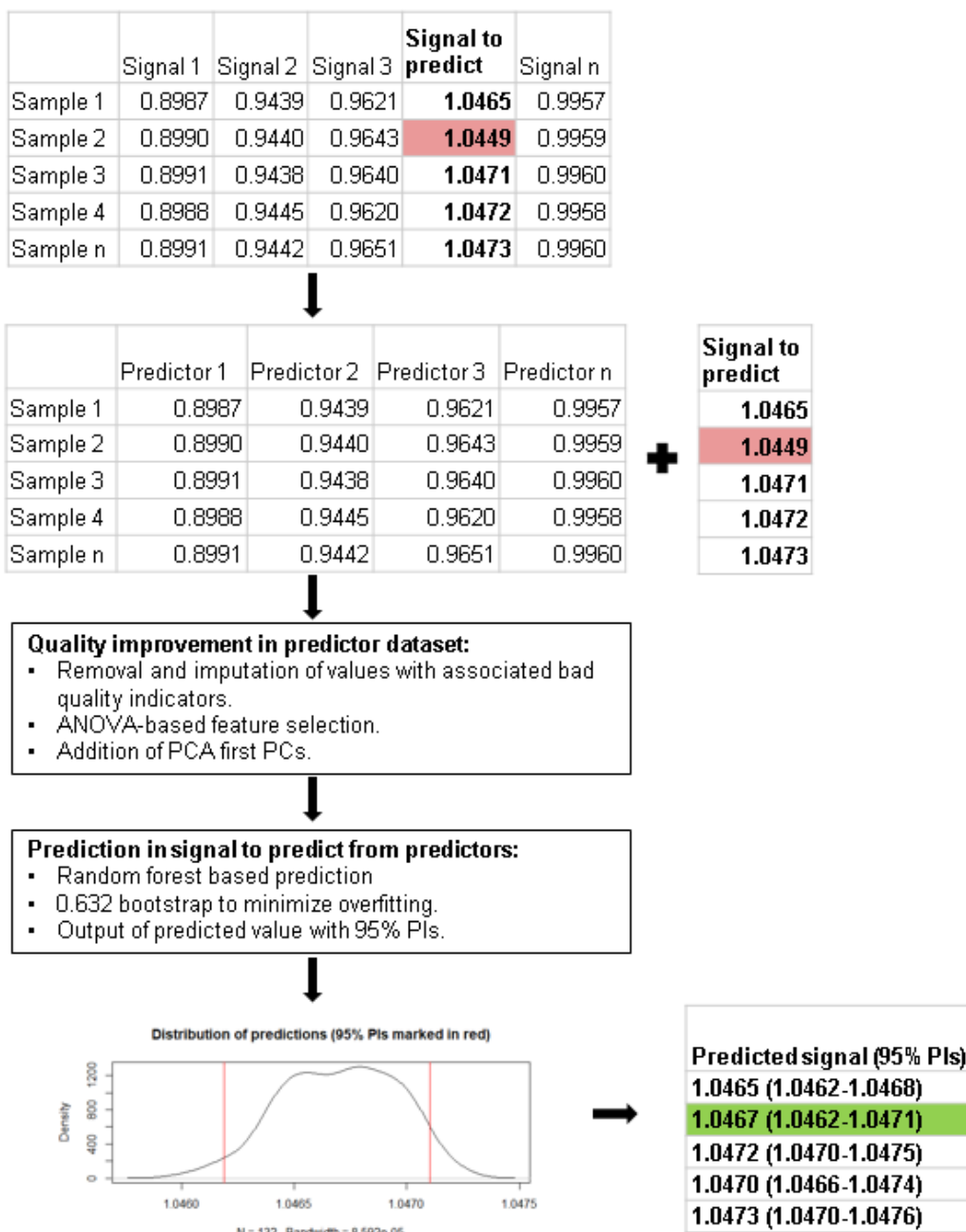


Figure 6-5 Signal parameter prediction pipeline applied to chemical shift information. An inaccurate chemical shift of a signal is shaded in red. For the chemical shifts of the signal, a training dataset is built with the other chemical shifts. The dataset is then cleaned, filtered and enriched to maximize its quality. Next, it is used to train a prediction model for the chemical shifts of the signal analysed. During training, bootstrap resampling avoids overfitting inaccurate values. Then, for each predicted chemical shift, the distribution of the predictions made during the bootstrap iterations is built and the median value and 95% PIs of this distribution are outputted. The predicted value and PIs are shaded in green. The inaccurate chemical shift shaded in red is clearly outside the 95% PIs shaded in green. This process is repeated for each signal.

7 Conclusions and Future Directions

7.1 Conclusions

During this thesis, the evaluation of machine learning-based approaches to model the signal parameters in ¹H-NMR datasets and exploit the possible advantages derived from this modelling accomplished the next achievements:

- The narrow and accurate prediction of the expected signal parameters in a dataset thanks to information previously collected from this dataset. This achievement liberates profiling tools of the need of requiring prior information about the matrix to study, the metabolites to profile or the sample acquisition or study protocol performed during the study. In addition, the need for restrictions during sample preparation, spectrum acquisition or matrix to analyse is overcome.
- The improvement of automatic metabolite profiling thanks to the narrow and accurate estimation of ranges of possible parameter values to consider during the lineshape fitting of signals.
- The generation of indicators of possible wrong annotations and improvable quantifications of metabolite concentration of higher quality than the standard ones in lineshape fitting (i.e., fitting error). These indicators are based on the study of the difference between the expected signal parameter value and the obtained one.
- The reliable and optimized exploitation of the potential of chemical shift information to maximize the performance of the classification of samples during the multivariate analysis of metabolomics studies.
- The generation of a ML-based tool able to help during the identification of metabolite signals. This tool finds clusters of chemical shifts which behave similarly to the signal analysed (and, therefore, should come from metabolite with similar structures).

In addition, during the PhD thesis, additional achievements not directly related to the original objectives were achieved:

- The building of an open-source automatic profiling tool which enhances the flexibility and reproducibility of profiling in order to handle the challenges typical from complex matrices with the best balance between accuracy, reproducibility and ease to use.
- The creation of the first public reproducible ¹H-NMR metabolite profiling workflows of metabolomics studies based on already public study datasets in order to enhance the reproducibility of metabolomics study workflows.
- The generation of a metabolite identification tool adapted to minimize wrong annotations of e.g. metabolites not typical from the matrix analysed. This enhanced version of metabolite annotation tool is based on the data mining of open-source HMDB information about

the reported concentration and presence information of each metabolite for each matrix and about the parameters of each metabolite signal.

- The novel row-wise dimensionality reduction of a spectra dataset thanks to the selection of exemplars of spectra clusters able to efficiently represent the variance present in a spectra dataset.
- The demonstration of the influence of the chemical shift variability in the results of fingerprint-based analyses of the difference between sample cases (and, therefore, of the further need to promote the development profiling approaches instead of fingerprint-based ones).

7.2 Future directions

The achievements during this PhD thesis open the path to possible future achievements within the context of the metabolomics field. In addition, some bottlenecks were discovered during the thesis which might be dealt with in order to maximize the potential of the achievements accomplished during the thesis:

- rDolphin tool should ideally be deployed as a containerized tool, replicating the tendencies in the deployment of data science products to ensure robustness and reproducibility. In this context, the incorporation of this tool into the Phenomenal-H2020 project might help accomplish these objectives.¹
- The finding of the best ML-based solutions was rather based on ad-hoc experience than on a robust analysis through hypothesis testing of the performance metrics collected. In order to maximize the generalizability of the solutions proposed, hypothesis testing should have been performed in order to validate the achievements accomplished. For example, the improvements in sample discrimination achieved in Chapter 5 should have been validated by the K-fold cross-validated paired t-test procedure.² In this context, as far as the author is concerned, there is no current available research on the use of hypothesis testing during the ML-based multivariate analysis in metabolomics studies to validate the insights achieved during this analysis. This study might help uncover possible reproducibility limitations in the insights achieved and help maximize the reproducibility of metabolomics research. Low sample sizes, high phenotypic variability and lack of standardization workflows still prevalent in metabolomics research. These limitations suggest the promising potential of approaches such as the hypothesis testing of metrics or the

bootstrap of statistical tests to investigate and incorporate in standardized metabolomics study workflows.

- Regarding the prediction of the expected signal parameters, further implementations of the analysis of the consistency in the signal parameters should be explored in different kinds of spectra (2D NMR, spectra with recent improvements in sensitivity and resolution such as dynamic nuclear polarization -DNP- or pure shift).^{3,4,5} The implementation of the developed approach in 2D datasets might help monitor the high variability in the data because of the managing of more complex pulses or of nuclei with lower abundance. As a result, the current low proportionality between signal volumes and metabolite concentrations might be enhanced and automatic profiling tools for these other kinds of spectra might be developed. In the case of DNP, the improvements in sensitivity will mean even higher limitations derived from resolution-based constraints in NMR spectra. Consequently, the need of reducing the search space during lineshape fitting through the prediction of signal parameters will become even more necessary. Lastly, regarding pure-shift, the loss of the multiplet shape (from doublets, triplets, multiplets, etc. to singlets) requires the development of strategies to handle the signals from different metabolites with similar chemical shifts.
- In human urine, because of its special interest as matrix for metabolomics studies, further improvements in the prediction of the signal parameters might be explored and with a validation of these improvements with MS data (but, preferably, without losing the generalizability of the developed approach to any matrix).
- The workflow to predict the expected signal parameters might be used to also explore the prediction of the expected metabolite concentrations. Multicollinearity is prevalent in metabolite concentration datasets (e.g., in human serum, it is not uncommon to find Pearson correlations higher than 0.8 between branched-chain amino acids). Therefore, it should be also possible to predict with certain reliability the concentration of a metabolite by the modelling of its concentration with the ones of collinear metabolites. Consequently, it might be possible to find concentrations to correct as the concentration found is not consistent with the expected one. This approach might help to correct the effect of contaminants in the sample, the binding of several metabolites in protein or the lack of stability of the analytical platform.

References

1. PhenoMeNal – Large-scale Computing for Medical Metabolomics. Available at: <http://phenomenal-h2020.eu/home/>. (Accessed: 18th August 2018)
2. Dietterich, T. G. & G., T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
3. Giraudeau, P. Quantitative 2D liquid-state NMR. *Magn. Reson. Chem.* (2014). doi:10.1002/mrc.4068
4. Ardenkjaer-Larsen, J. H. On the present and future of dissolution-DNP. *J. Magn. Reson.* **264**, 3–12 (2016).
5. Zangger, K. Pure shift NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **86–87**, 1–20 (2015).

List of Figures

Figure 3-1 Parameters of metabolite signals in ¹H-NMR spectra.....	16
Figure 3-2 Kinds of multiplets and relationship with chemical structure..	17
Figure 3-3 Pascal triangle structure of the peak intensities in multiplets.....	17
Figure 3-4 Roofing of the citric acid doublets.....	18
Figure 3-5 Relationship between pH decrease and deprotonation of the nuclei of functional groups.....	20
Figure 3-6 Relationship between pH decrease and chemical shift decrease. The intensity of the chemical shift decreases and of the pH range where this decrease happens is specific from very metabolite signal.	21
Figure 3-7 Inverse relationship between the chemical shifts of signals caused by the choice of a reference non-resistant to pH changes.....	22
Figure 3-8 Chemical shift and lineshape changes mediated by the variability in sodium concentration present in human urine matrix.....	22
Figure 3-9 The deconvolution of signals permits the isolation of the signal of interest from the other signals. As a result, the quantification of the area below the signal is improved.	25
Figure 3-10 Venn diagram of the different metabolites which can be characterized with every combination of platforms. NMR provides a much lower number of metabolites than other compounds, reducing its potential to characterize the metabolome.....	25
Figure 3-11 Baseline of lipids and macromolecules present in the human blood matrix. After applying CPMG sequence during spectrum acquisition, the original spectrum lineshape (black line) most baseline and broad signals are removed from the spectrum (brown line).	28
Figure 3 12 The relative intensities of signals of a same metabolite can be not constant. The three hippurate signals at the 7.85-7.5 ppm region are shown for two datasets of human urine and for the BMRB standard. After normalizing the spectra by the left signal, the other two signals show clear differences in relative intensity even when coming from the same matrix. This variability is mediated by shimming differences and possible other effects related to differences in samples properties or preparation. As a result, the simultaneous lineshape fitting of all metabolites can be compromised as the assumption of constant relative intensity is not accomplished.....	31
Figure 3-13 The ratio between half bandwidths of signals can be not constant. The TSP signal is used as CSI to estimate the expected half bandwidth of the rest of signals in a spectrum. However, in datasets of the same matrix (human urine), differences between the ratio of the half bandwidth of a signal such as a creatinine one and the one of the CSI signal can be observed. More concretely, on the dataset 1, the ratio creatinine/TSP is much higher than on the dataset 2.	

As a result, the assumption of constant ratio between half bandwidths is not accomplished and the estimation of accurate half bandwidths is compromised.....	31
Figure 3-14 As the amount of data increases, the performance of DL approaches trumps the one of traditional ML techniques.....	39
Figure 3-15 Tree-based algorithms showed the best performance in the evaluation of different traditional ML algorithms in 165 different datasets.	39
Figure 3-16 Comparison (in accuracy and speed) of different clustering algorithms when dealing with different data patterns.	40
Figure 4-1 Example of lineshape fitting in the 1.09–1.03 ppm region of the human urine MTBLS1 dataset. Signal area quantifications and fitting quality indicators are shown below the interactive Plotly figure.	57
Figure 4-2 Reduction of a 132 spectra dataset into 10 representative exemplars (whose sample names are specified below right). This interactive figure is created by the Plotly API.	60
Figure 4-3 Exploratory analysis of human faecal extract MTBLS237 dataset with rDolphin. Differences between the median spectrum of three kinds of sample in the 0.92–0.88 ppm region are shown on an interactive figure. Fingerprint analysis information is also provided by the red trace below the median spectra.	60
Figure 4-4 Example of available information of reported concentrations in the HMDB website (top) and the equivalent information present in XML format (down).	62
Figure 4-5 The use of HMDB information facilitates the accurate matrix-specific information of metabolite signals. The rDolphin repository of metabolite signals can be filtered by the matrix and the spectrum region. Then, signals can be sorted according to the presence in previous bibliography and of its typical concentration in the matrix analysed. In addition, the repository provides information about the kind of multiplet, the J-coupling and the relative intensity of the signal.	63
Figure 4-6 rDolphin enables the finding of wrong annotations and suboptimal quantifications through several indicators of quality. In a), possible suboptimal quantifications of carnitine have been ordered by difference between the chemical shift (in ppm) of the performed quantification and the predicted chemical shift. The shade suggests the grade of outlier behaviour. In b), the predicted chemical shift of carnitine is located 0.0042 ppm below than the one of the fitted signal, exactly where the neighbouring signal to its right is located.	65
Figure 4-7 The dendrogram heatmaps of rDolphin show the signals with similar quantification (a) and chemical shift (b) patterns. The figures show the dendrograms observed in the MTBLS1 dataset. The singlet at 2.35 ppm (annotated as p-Cresol sulphate in the dendrogram) shows similar quantification patterns to related metabolites such as indoxyl sulphate or phe-nylacetylglutamine. This signal also shows chemical shift patterns similar to the ones of	

metabolites with similar functional groups such as indoxyl sulphate or hippurate. In b), the strong interrelation between the triplet at 4.042 ppm (annotated as U4_042 in the dendrogram) and a creatinine signal can also be observed. 66

Figure 5-1 Exploratory PCA analysis shows the potential of the chemical shift data in the classification models. The first PCs of the PCA using chemical shifts (right) show better separation than the ones using concentrations (left). Plots also suggest no batch effects necessary to monitor. 80

Figure 5-2 Signals can be misaligned in some sample classes. Low pH mediated by the condition studied increases the chemical shift of the signals. The resulting class-dependent signal misalignment can distort the results of the analysis of fingerprint data: features can show significant differences caused by differences in chemical shift (mediated by pH or ionic strength) rather than by differences in metabolite concentration. 84

Figure 5-3 Variability (measured by standard deviation) of the chemical shifts analysed in the three datasets. As expected, the dataset of human matrices with higher dilution variability (urine and fecal extracts) show higher chemical shift variability. In all three datasets, the use of buffers does not impede the appearance of chemical shift variability that can be analysed. 93

Figure 5-4 Distribution of centred chemical shift of three good chemical shift predictors in the MTBLS1 (top), MTBLS237 (middle) and MTBLS374 (bottom) datasets. Chemical shift patterns in the MTBLS1 and the MTBLS237 datasets showed higher complexity (with some signals with inverse trends) in the chemical shift mediated by the use of TSP as reference. 94

Figure 6-1 The signal parameter prediction pipeline enables narrow and accurate spectrum-specific ranges to be estimated and used during lineshape fitting. The figure shows a difficult signal fitting found with the 4-deoxythreonic acid signal in the urine dataset analysed in Appendix. The chemical shift variability present in this signal (a) forces lineshape fitting algorithms to consider a wide range of possible chemical shift values during the fitting (b). Excessive width can compromise the right assignment of the doublet center when other signals appear adjacent to the signal to be fitted (d). The chemical shift prediction generates spectrum-specific chemical shift distributions of predictions (c). These distributions are very narrow and can help generate much narrower chemical shift ranges (d). 102

Figure 6-2 The spectrum-specific 95% PIs of the parameter values PIs are much narrower than the spectrum-unspecific 95% PIs. Chemical shift PIs are generally lower than the bucketing applied (6e-4 ppm). The narrow PIs enhance the performance of error minimization algorithms to end in the right local minimum. 105

Figure 6-3 The calculated anomaly score helped identify quantifications which might be further optimized. In both datasets, the anomaly score showed higher performance than the fitting error ranking the quantifications which, if further improved, might further enhance the MS/NMR correlation. 107

Figure 6-4 The spectrum-specific 95% PIs of the parameter values PIs are much narrower than the spectrum-unspecific 95% PIs. Chemical shift PIs are generally lower than the bucketing applied ($6e-4$ ppm). The narrow PIs enhance the performance of error minimization algorithms to end in the right local minimum. 115

Figure 6-5 Signal parameter prediction pipeline applied to chemical shift information. An inaccurate chemical shift of a signal is shaded in red. For the chemical shifts of the signal, a training dataset is built with the other chemical shifts. The dataset is then cleaned, filtered and enriched to maximize its quality. Next, it is used to train a prediction model for the chemical shifts of the signal analysed. During training, bootstrap resampling avoids overfitting inaccurate values. Then, for each predicted chemical shift, the distribution of the predictions made during the bootstrap iterations is built and the median value and 95% PIs of this distribution are outputted. The predicted value and PIs are shaded in green. The inaccurate chemical shift shaded in red is clearly outside the 95% PIs shaded in green. This process is repeated for each signal. 117

List of Tables

Table 5.1 Chemical shift information shows discriminative potential in the MTBLS1 dataset. However, it cannot enhance the excellent results given by concentration information during RF classification.	81
Table 5.2 Adding chemical shift information to concentration information improved the classification between the five different kinds of sample in the MTBLS237 dataset. Several quality indicators of the models generated are shown.	81
Table 5.3 Adding chemical shift information to concentration information provides the best classification of samples in the MTBLS374 dataset. Several quality indicators of the models generated only with concentration information, only with chemical shift information and with both sources of information are shown.	82
Table 5.4 Ranked predictors in RF classification of samples with both con-centration and chemical shift information in the MTBLS374 dataset. There are few predictors because of the recursive feature ex-traction of non- discriminative features.....	91
Table 5.5 Additional classification indicators in the MTBLS1 dataset.	91
Table 5.6 Additional classification indicators in the MTBLS237 dataset.	92
Table 5.7 Additional classification indicators in the MTBLS374 dataset.	93
Table 6.1 The predicted signal parameter information increases Spearman’s rho correlation between metabolite concentrations in MS and NMR data in both datasets. There is a consistent increase in this profiling quality indicator when a new profiling iteration is performed using the PIs as new value ranges during lineshape fitting. The increase is most significant in the metabolites whose profiling was most complicated in the original profiling iteration.	106

