



**Daniel Cañueto**

**Improvement of sample classification and metabolite  
profiling in  $^1\text{H}$ -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

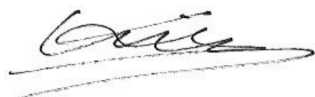
Avda. Països Catalans, 26

43007 Tarragona

I CERTIFY that the present study, entitled "**Improvement of sample classification and metabolite profiling in <sup>1</sup>H-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







































































































































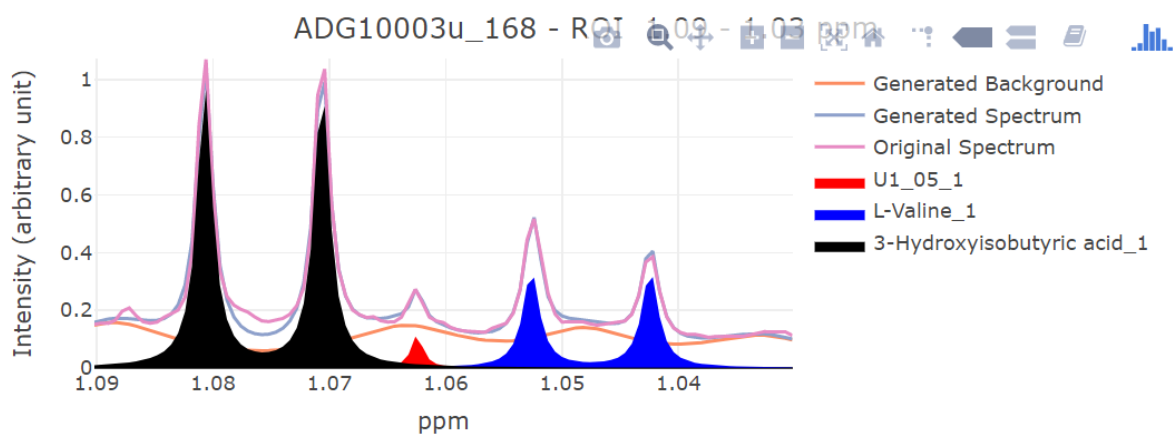












Here you have some indicators of the quantification.

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<sup>-10</sup>).

#### 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 reimplementations 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

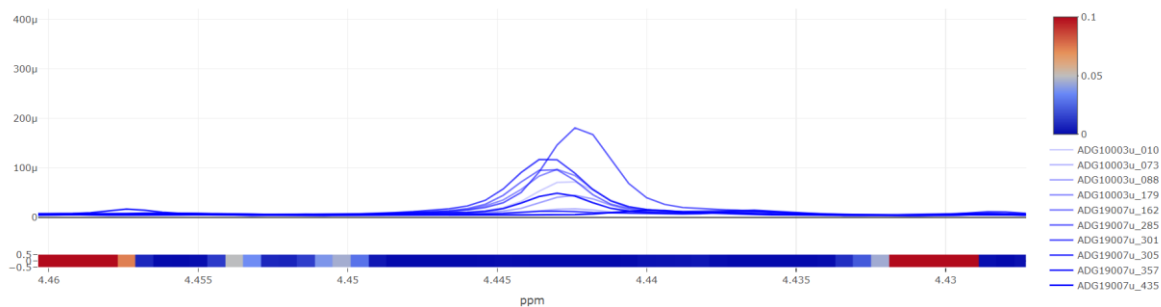
One of the main advantages of Dolphin in comparison with alternative profiling tools was the interactive visualization of spectra, a necessary requisite to perform a robust exploratory analysis of the sources of variability present in the spectra. The generation of interactive figures was achieved thanks 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.

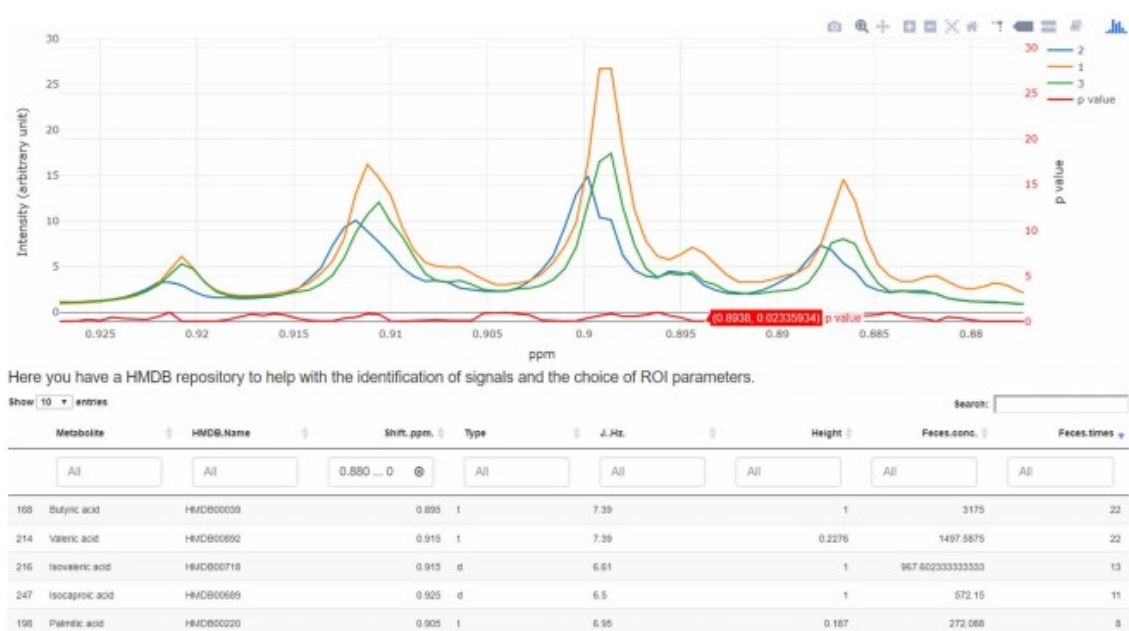
Accordingly, two different kinds of interactive figures of multiple spectra were generated:

- 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.<sup>11</sup> 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).<sup>12</sup> 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.<sup>13,14,15</sup> 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).<sup>16</sup> 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...	<a href="#">details</a>
Blood	Detected and Quantified	60.00-115.0 uM	Adult (>18 years old)	Male	Normal	Vancouver Co...	<a href="#">details</a>
Blood	Detected and Quantified	86.6 +/- 18.8 uM	Adult (>18 years old)	Both	Normal	21359215	<a href="#">details</a>
Breast Milk	Detected and Quantified	39.9 +/- 7.9 uM	Adult (>18 years old)	Female	Normal	24027187	<a href="#">details</a>
Cerebrospinal Fluid (CSF)	Detected and Quantified	43 +/- 12 uM	Adult (>18 years old)	Both	Normal	18502700	<a href="#">details</a>
Cerebrospinal Fluid (CSF)	Detected and Quantified	65.2 (51.8-78.6) uM	Adult (>18 years old)	Both	Normal	7108550	<a href="#">details</a>
Cerebrospinal Fluid (CSF)	Detected and Quantified	64.95 (37.5-92.4) uM	Adult (>18 years old)	Both	Normal	Geigy Scientific ...	<a href="#">details</a>
Feces	Detected but not Quantified		Children (6 - 18 years old)	Both	Normal	27609529	<a href="#">details</a>
Feces	Detected but not Quantified		Children (6 - 18 years old)	Not Specified	Normal	27609529	<a href="#">details</a>
Saliva	Detected but not Quantified		Adult (>18 years old)	Male	Normal	22308371	<a href="#">details</a>

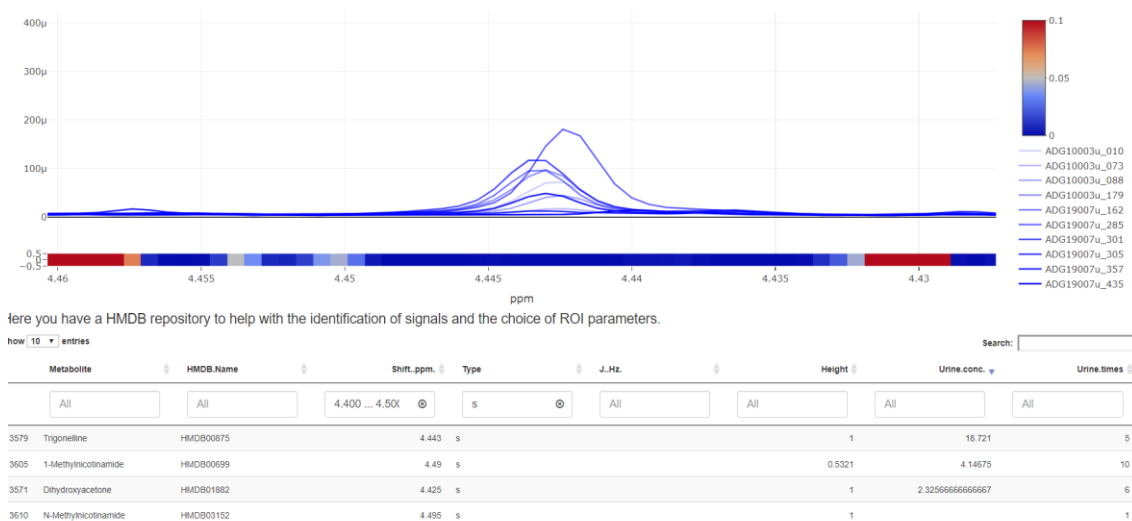
  

```

</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 D,
      <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.

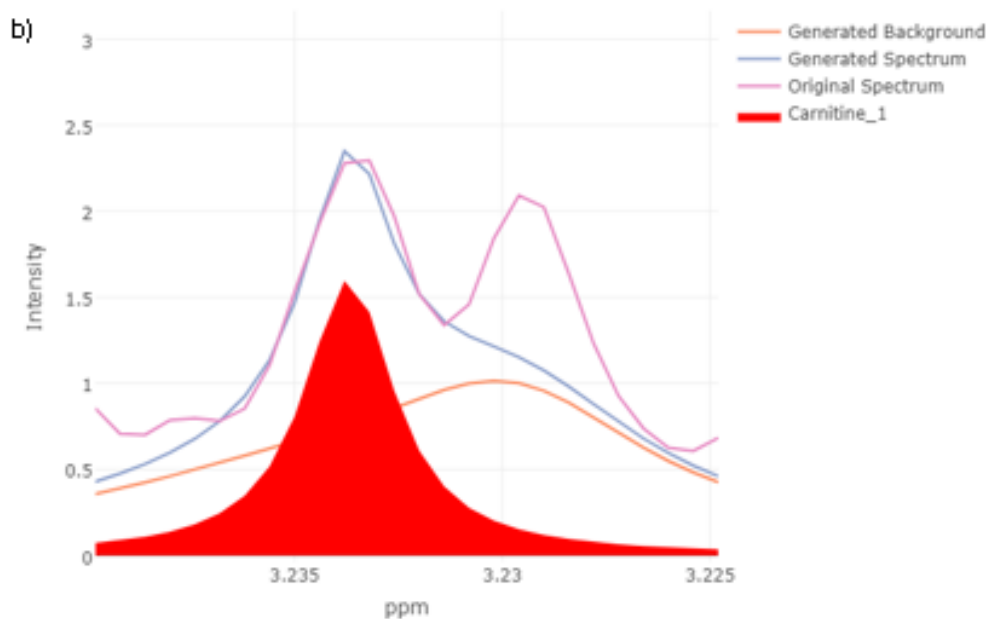
a)

Show 10 entries Search

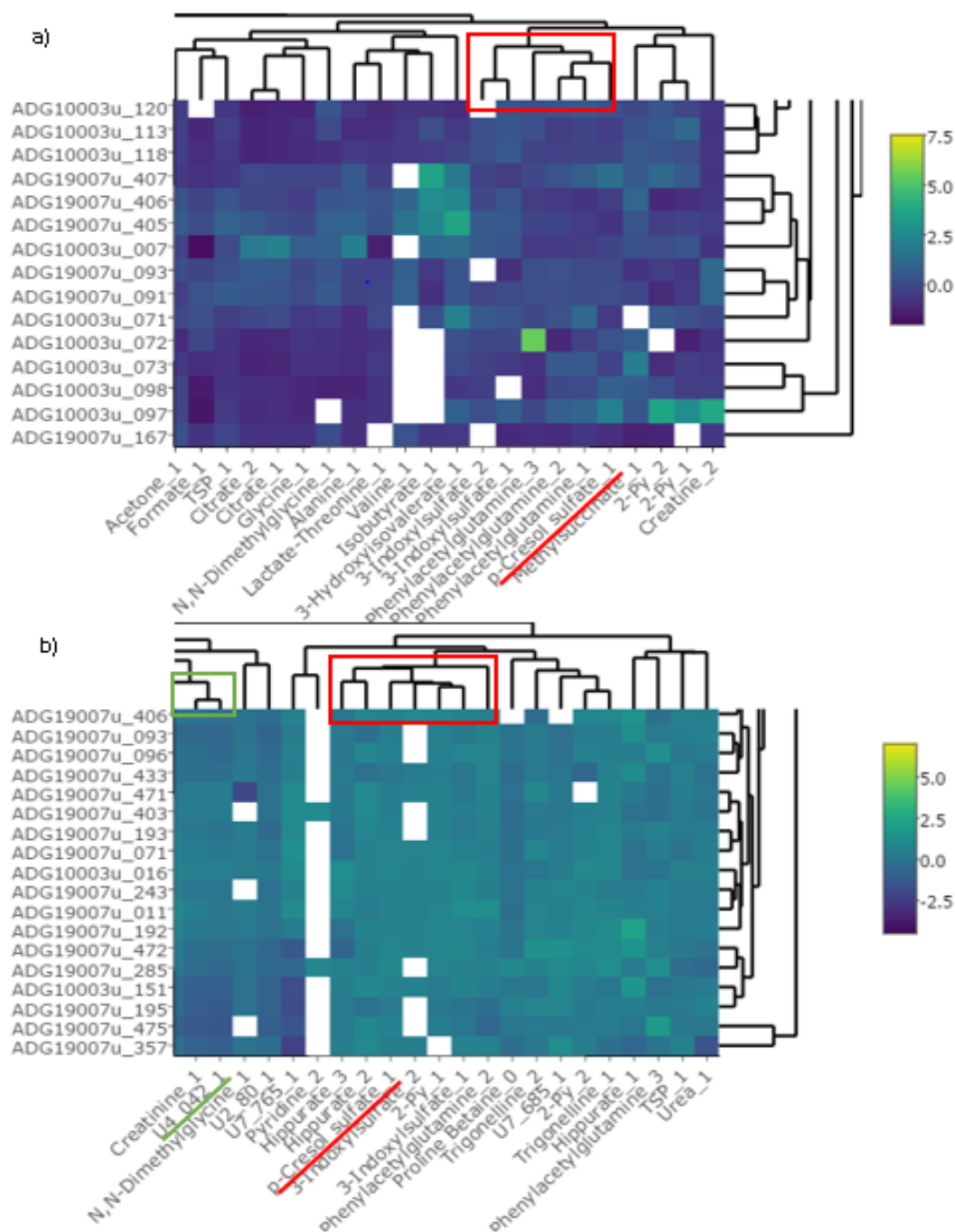
creatine_1	cis-Aconitate_1	Prolise Betaine_0	Carnitine_1	Betaine_1	Trimethylamine N-oxide_1	T
	0.0002		-0.0055		-0.0004	
	-0.0001		-0.0042		0.0002	
	0.0001		-0.0039		0	
	0.0002		-0.0032		0.0004	
	-0.0001		-0.003		-0.0014	
	0.0007		-0.002		0.0002	
	0.0003		-0.002		-0.0004	
	-0.0009		-0.0011		0.0002	
	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.<sup>17,18</sup> 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.<sup>19</sup>
- MTBLS237: This study contains 114 spectra of human faecal extract.<sup>20</sup>

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.<sup>21,22</sup>

## 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 <sup>1</sup>H-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 $^1\text{H}$ -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.<sup>1,2</sup> <sup>1</sup>H-NMR spectroscopy is a high-throughput technique that quantifies metabolite concentrations in a reliable and reproducible manner.<sup>3</sup> <sup>1</sup>H-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.<sup>4</sup>

A metabolite can be characterized in an <sup>1</sup>H-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.<sup>5</sup> 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 Ca<sup>2+</sup> or Mg<sup>2+</sup>

concentration) of the sample.<sup>6</sup> 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.<sup>7</sup> A recent article showed that the pH and ionic strength of human urine samples can be extrapolated from chemical shift information.<sup>8</sup> A wide range of diseases (e.g., tumours<sup>9</sup>) are characterized by metabolic alkalosis/acidosis<sup>10</sup> or ionic imbalance<sup>8</sup>: 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.<sup>11</sup> 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 DFTMP<sup>12</sup> or Chenomx-based pH calibration.

To date, several tools have been developed to automatically quantify metabolite concentrations in 1D <sup>1</sup>H-NMR spectra datasets,<sup>13–15</sup> 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 <sup>1</sup>H-NMR datasets.<sup>16</sup> 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 MetaboLights<sup>17</sup> (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.<sup>18</sup> 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).<sup>19</sup> 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.<sup>20</sup> The original study analysed <sup>1</sup>H-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.<sup>18-20</sup>

### 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 <sup>1</sup>H-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.<sup>21</sup> 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.<sup>22</sup> 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.<sup>23,24</sup> 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,<sup>25</sup> upsampling to maximize the robustness of the models against the class imbalance problem in datasets,<sup>26</sup> 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](https://github.com/danielcanueto/chemical_shift_classification).

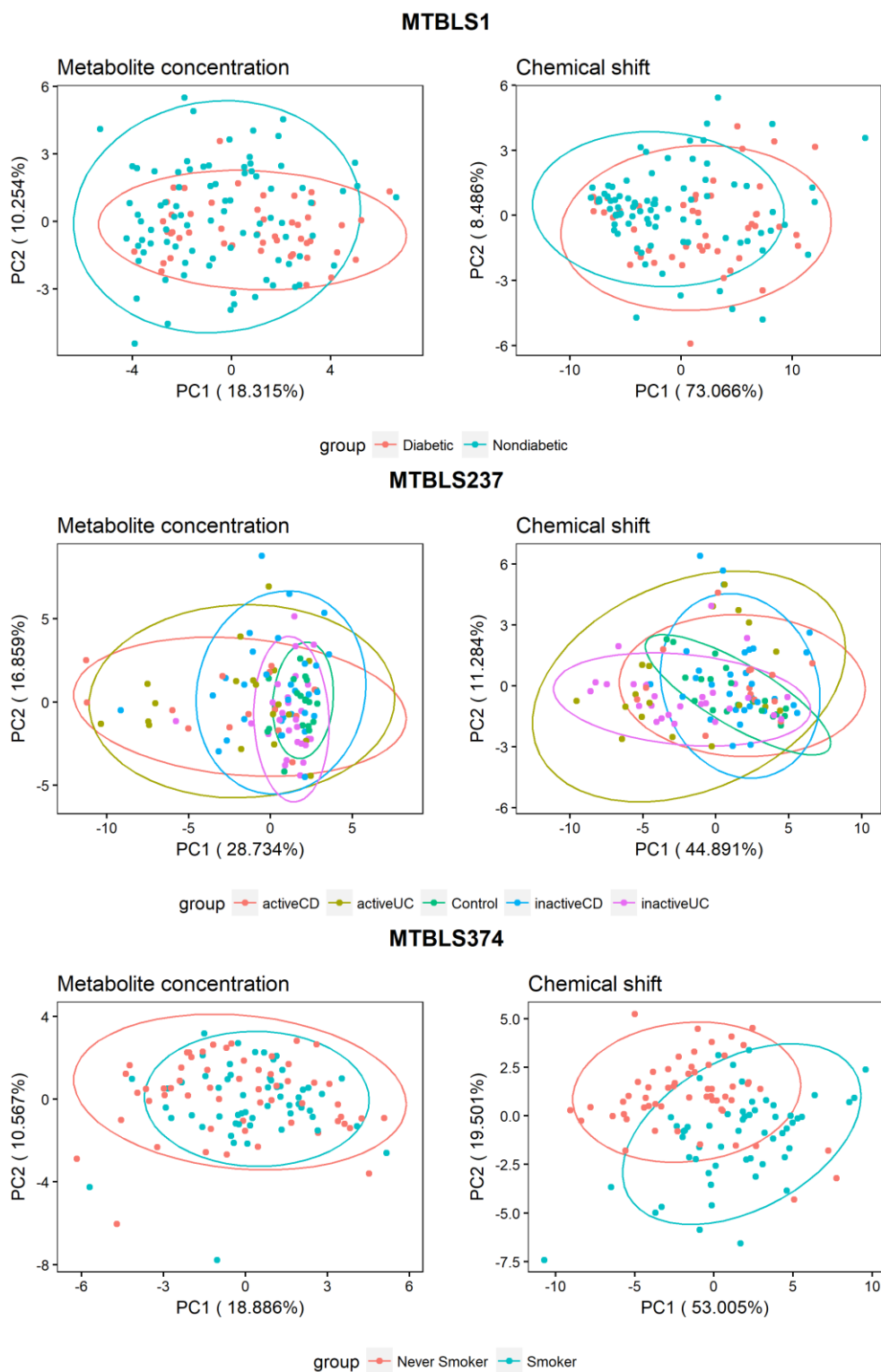
## 5.3 Results

### 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
<b>Accuracy</b>	0.929	0.933	0.795
<b>kappa</b>	0.840	0.849	0.559
<b>AUROC</b>	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
<b>Active UC vs Inactive UC</b>			
<b>Accuracy</b>	0.863	0.826	0.876
<b>kappa</b>	0.635	0.555	0.698
<b>AUROC</b>	0.870	0.811	0.917
<b>Active CD vs Inactive CD</b>			
<b>Accuracy</b>	0.801	0.808	0.721
<b>kappa</b>	0.505	0.526	0.331
<b>AUROC</b>	0.768	0.777	0.661
<b>Active UC vs Active CD</b>			
<b>Accuracy</b>	0.730	0.717	0.668
<b>kappa</b>	0.462	0.438	0.339
<b>AUROC</b>	0.768	0.743	0.682
<b>Inactive UC vs Inactive CD</b>			
<b>Accuracy</b>	0.808	0.771	0.797
<b>kappa</b>	0.617	0.545	0.594
<b>AUROC</b>	0.870	0.810	0.841

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

	<b>Both sets of information</b>	<b>Concentration information</b>	<b>Chemical shift information</b>
<b>Accuracy</b>	0.899	0.806	0.883
<b>kappa</b>	0.797	0.614	0.766
<b>AUROC</b>	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 <sup>1</sup>H-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.<sup>27</sup> A lower pH increases the chemical shift of signals (i.e., the signal moves to the left in a spectrum).<sup>28</sup> 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 region<sup>29</sup>) 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.<sup>30</sup> 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.<sup>31</sup> 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 smokers<sup>32</sup> and in cigarette smoke that contains oxidants with acidic properties.<sup>33</sup> 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

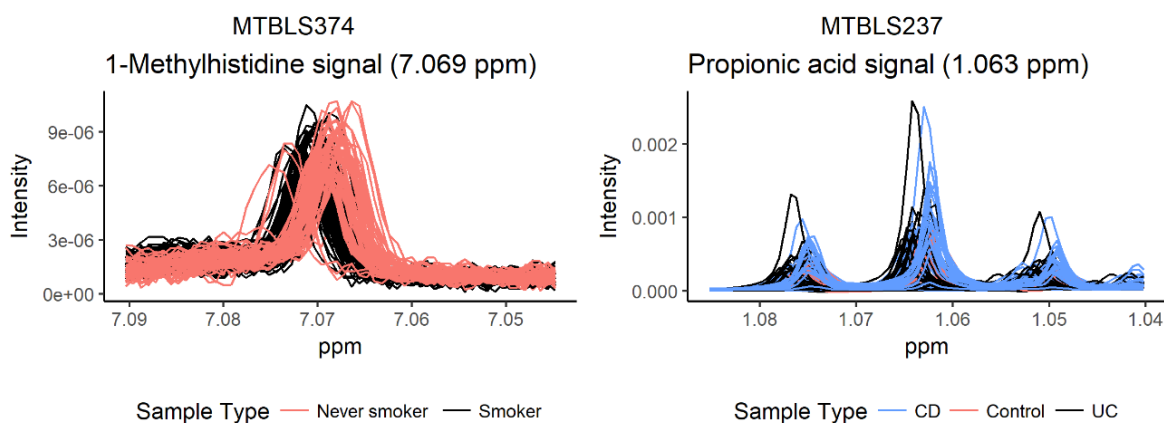
All the datasets evaluated were processed using fingerprinting approaches in the original studies, in contrast to the profiling approach used here. Fingerprinting approaches perform the classification by looking for significant spectral differences between groups and identifying the metabolites involved 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.<sup>34</sup>

Profiling is deemed to provide more resistance against signal overlap or baseline appearance through the deconvolution of signals in the spectrum lineshape.<sup>35</sup> 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 algorithms<sup>36</sup>) 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 algorithm<sup>37</sup> (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.<sup>38</sup> 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.<sup>39</sup> 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.<sup>8</sup> 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 <sup>1</sup>H-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](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
<b>Sensitivity</b>	0.838	0.855	0.713
<b>Specificity</b>	0.986	0.982	0.841
<b>Pos Pred Value</b>	0.968	0.963	0.723
<b>Neg Pred Value</b>	0.915	0.923	0.835

*Table 5.5 Additional classification indicators in the MTBLS1 dataset.*

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

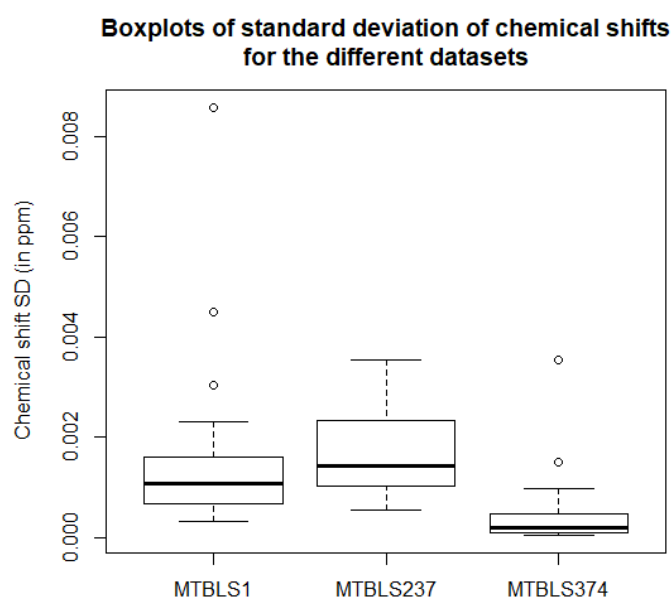
	<b>Both sets of information</b>	<b>Concentration information</b>	<b>Chemical shift information</b>
<b>Active UC vs Inactive UC</b>			
<b>Sensitivity</b>	0.866	0.839	0.882
<b>Specificity</b>	0.915	0.885	0.891
<b>Pos Pred Value</b>	0.912	0.882	0.895
<b>Neg Pred Value</b>	0.881	0.858	0.888
<b>Active CD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

*Table 5.6 Additional classification indicators in the MTBLS237 dataset.*

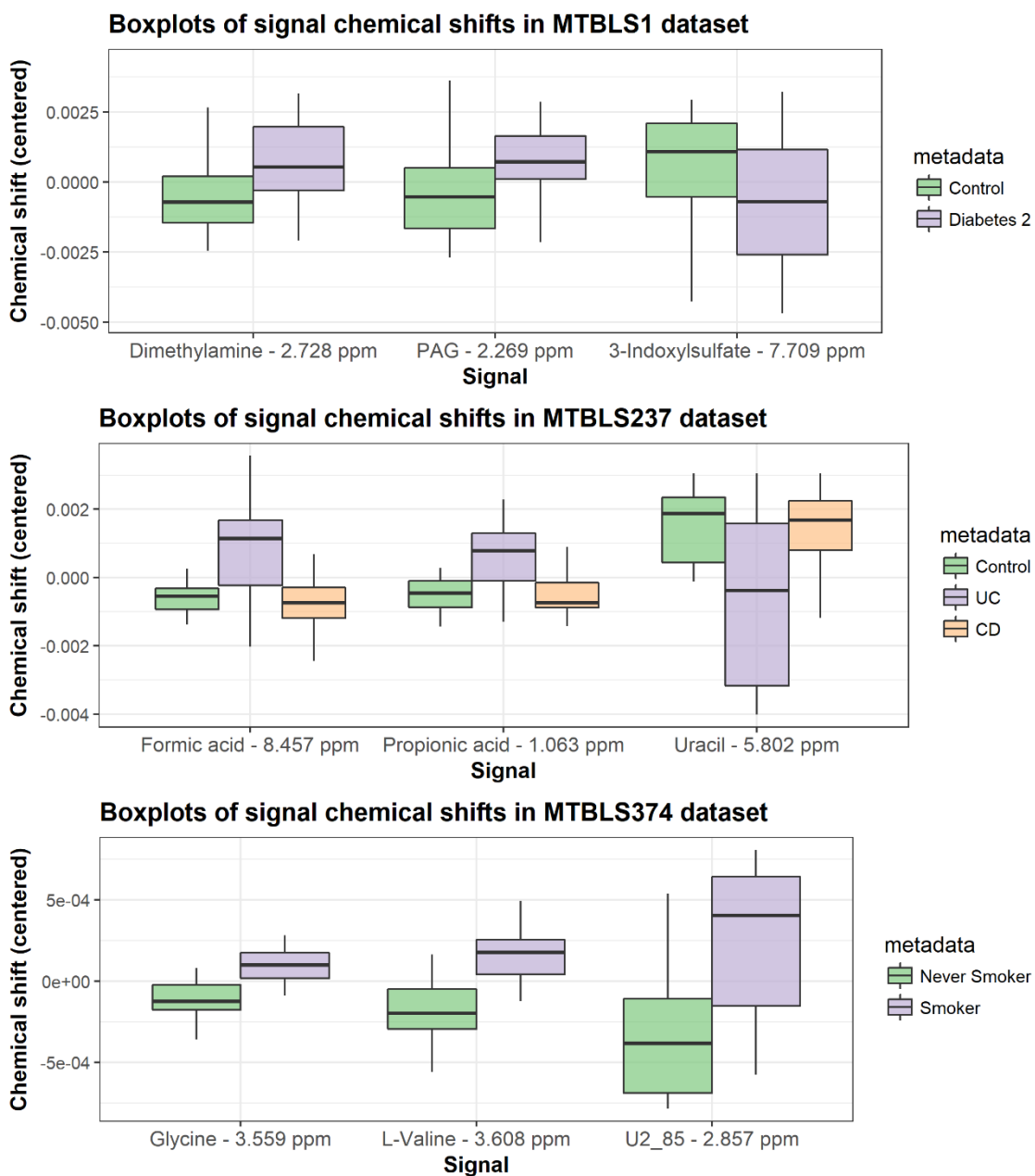
	Both sets of information	Concentration information	Chemical shift information
<b>Sensitivity</b>	0.893	0.810	0.870
<b>Specificity</b>	0.907	0.805	0.887
<b>Pos Pred Value</b>	0.906	0.809	0.886
<b>Neg Pred Value</b>	0.895	0.810	0.874

*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.<sup>1,2</sup> 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.<sup>3,4</sup> This area can be quantified by area integration or signal deconvolution. In the case of 1D <sup>1</sup>H-NMR spectra, three signal parameters need to be estimated to deconvolute a signal: intensity, chemical shift and half bandwidth.<sup>3</sup> 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.<sup>5-7</sup> 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.<sup>8-9</sup>

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 overlap<sup>10</sup>) (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)).<sup>11</sup> 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 metabolomics<sup>12</sup>) 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 others<sup>13</sup>). 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.<sup>14,15</sup>

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  $^1\text{H}$ -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 PQN<sup>16</sup> were performed through rDolphin.<sup>17</sup>

In the serum dataset, sample collection details are available in Hernández-Alonso, P. *et al.*<sup>18</sup> For each sample, 300  $\mu\text{l}$  aliquots were mixed with 300  $\mu\text{l}$  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.<sup>17</sup>

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 Appendix.

## 6.2.2 <sup>1</sup>H-NMR metabolite profiling workflow

Automatic metabolic profiling was performed using the rDolphin R package<sup>17</sup>, 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.<sup>19</sup> The values of the algorithm parameters used during lineshape fitting are available in Appendix. 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 Appendix) 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.

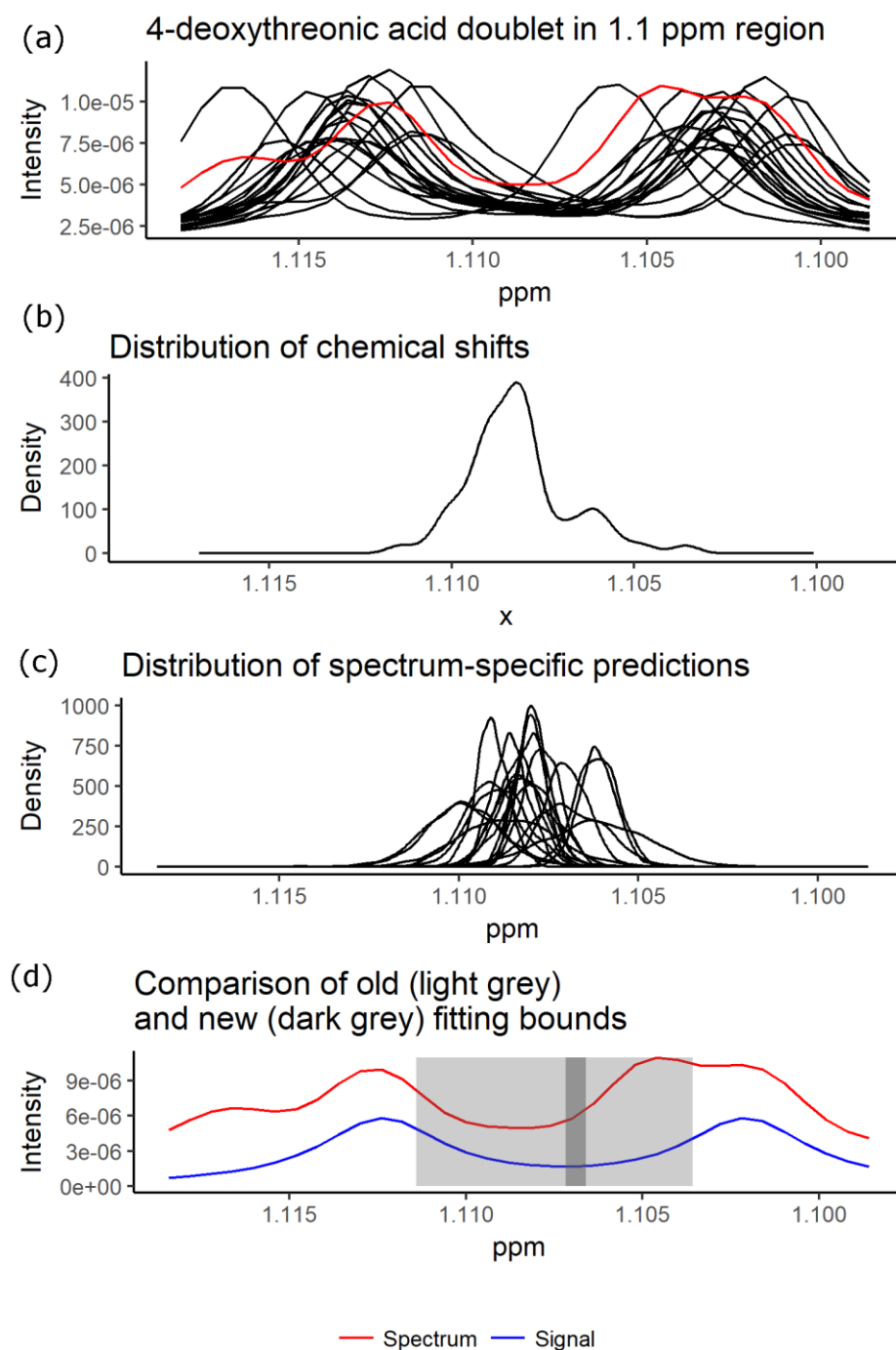
## 6.2.3 Prediction pipeline of expected signal parameter values

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.<sup>20</sup> 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.<sup>21,22</sup> 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.<sup>23</sup> 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

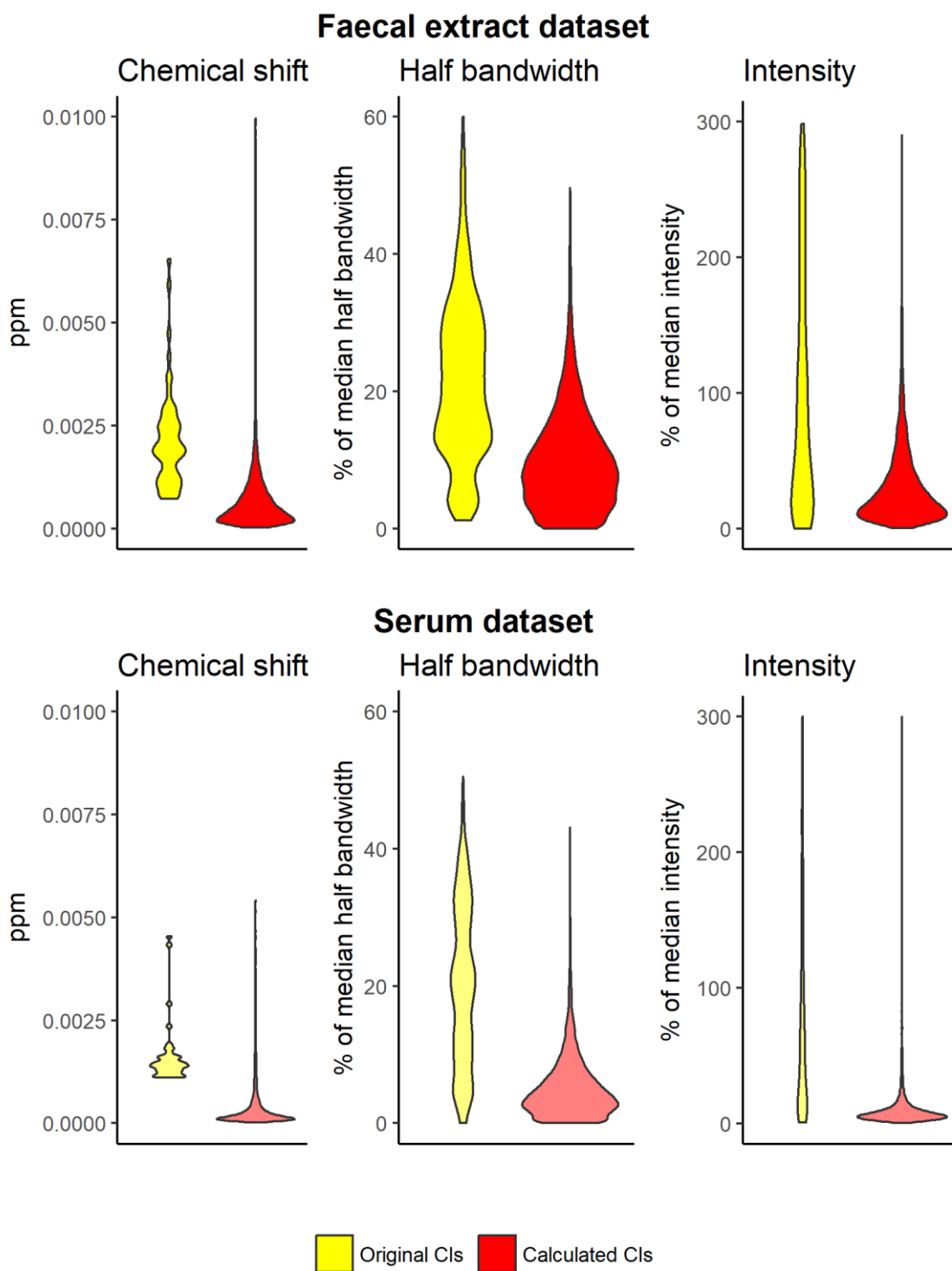
### 6.3.1 Accurate predicted values with narrow PIs which can be used to maximize profiling performance

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.

<b>Metabolites in faecal extract</b>	<b>Original Spearman's rho correlation</b>	<b>Spearman's rho correlation after new profiling iteration</b>
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

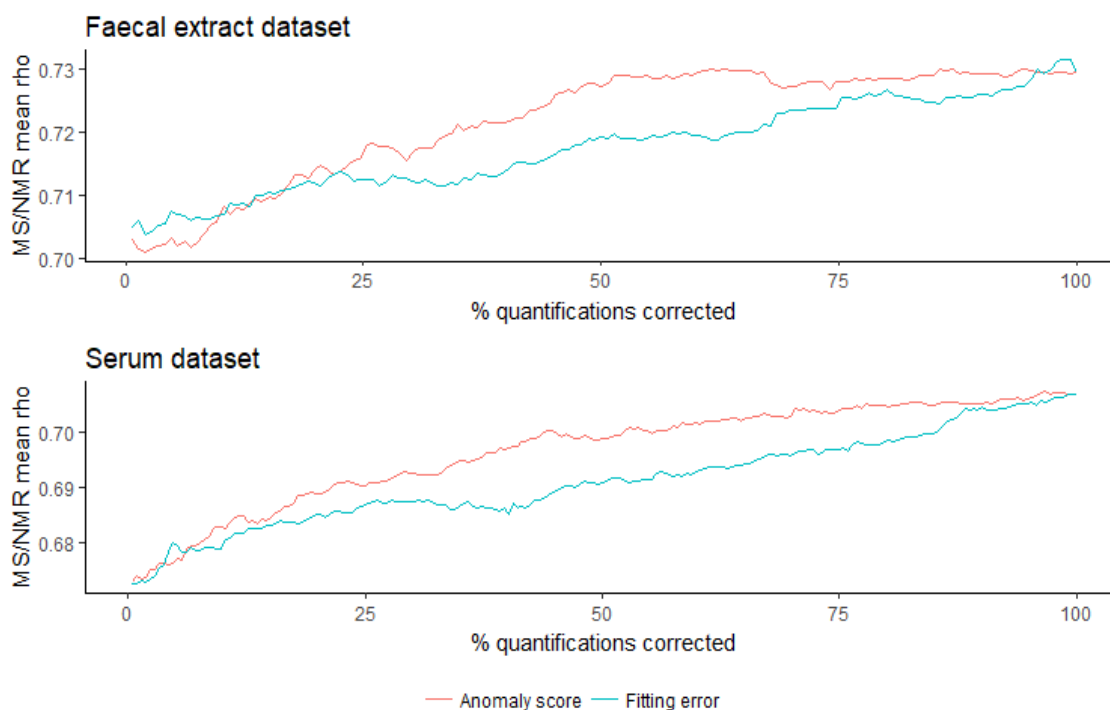
<b>Metabolites in serum</b>	<b>Original Spearman's rho correlation</b>	<b>Spearman's rho correlation after new profiling iteration</b>
Glycine	0.547	0.658
Hexanoic acid	0.556	0.677
5-aminovaleric acid	0.608	0.681
L-Isoleucine	0.629	0.632
L-Alanine	0.656	0.662
L-Valine	0.664	0.655
L-Leucine	0.697	0.715
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.<sup>12,14,15</sup> 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.<sup>10</sup>
- 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.<sup>24,25</sup>

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.<sup>13</sup> 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.<sup>26</sup> 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.<sup>27</sup> 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.<sup>28</sup> 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

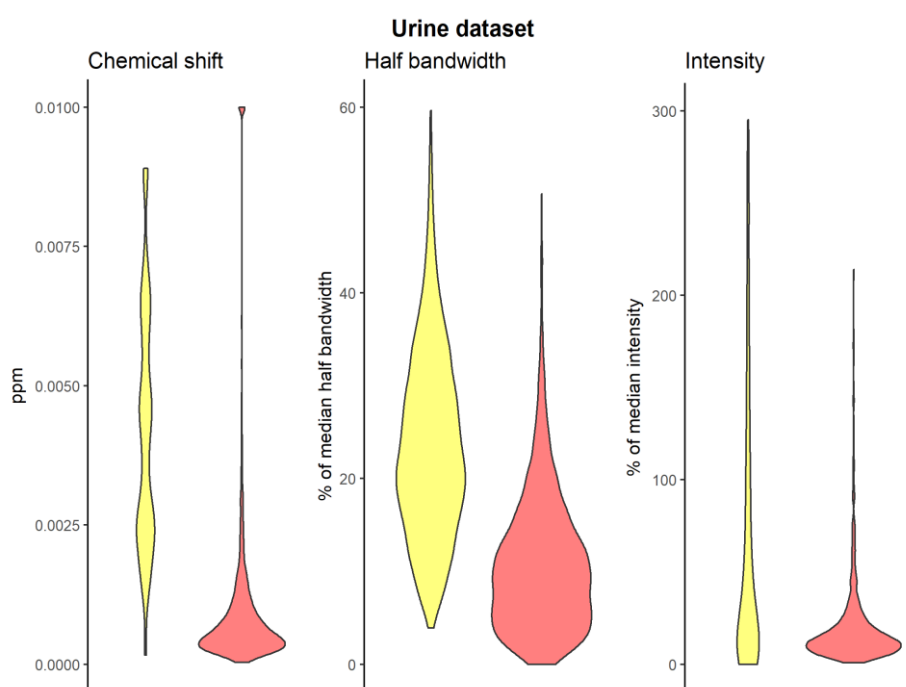
For chemical shift, the median range in the spectrum-specific 95% PIs calculated was 5.69e-04 ppm. This value is lower than the bucket width (6e-04 ppm) and is a reduction of 87.16% in the median range in the spectrum-unspecific 95% PIs (4.43e-03 ppm) (Figure 6-4; left).

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%.



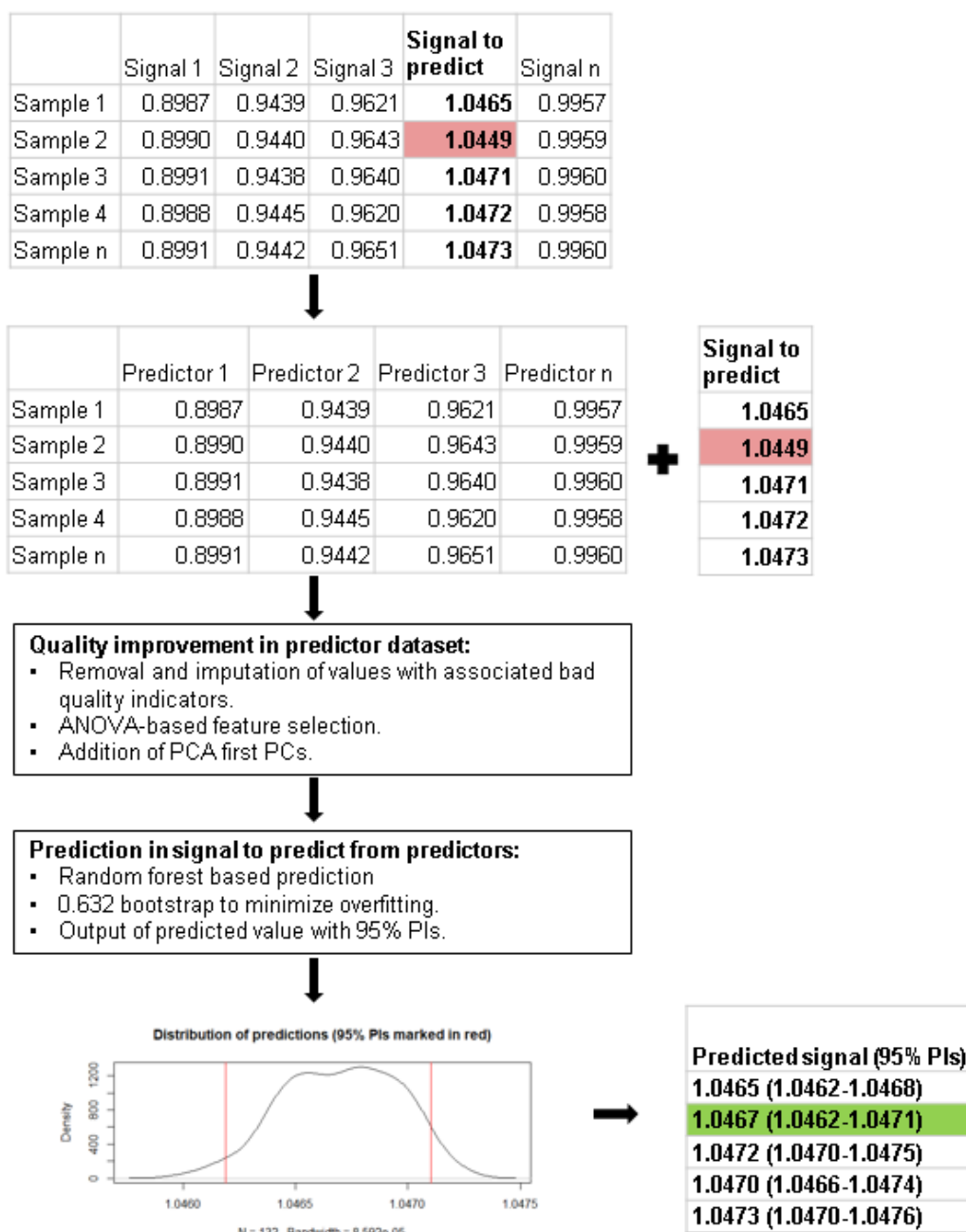
**Figure 6-4** The spectrum-specific 95% PIs of the parameter values PIs are much narrower than the spectrum-unspecific 95% PIs. Chemical shift PIs are generally lower than the bucketing applied ( $6e-4$  ppm). The narrow PIs enhance the performance of error minimization algorithms to end in the right 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.<sup>13</sup> 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.<sup>16</sup> 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 <sup>1</sup>H-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 <sup>1</sup>H-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.<sup>1</sup>
- 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.<sup>2</sup> 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).<sup>3,4,5</sup> 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

<b>Figure 3-1 Parameters of metabolite signals in <sup>1</sup>H-NMR spectra.....</b>	<b>16</b>
<b>Figure 3-2 Kinds of multiplets and relationship with chemical structure..</b>	<b>17</b>
<b>Figure 3-3 Pascal triangle structure of the peak intensities in multiplets.....</b>	<b>17</b>
<b>Figure 3-4 Roofing of the citric acid doublets.....</b>	<b>18</b>
<b>Figure 3-5 Relationship between pH decrease and deprotonation of the nuclei of functional groups.....</b>	<b>20</b>
<b>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. ....</b>	<b>21</b>
<b>Figure 3-7 Inverse relationship between the chemical shifts of signals caused by the choice of a reference non-resistant to pH changes.....</b>	<b>22</b>
<b>Figure 3-8 Chemical shift and lineshape changes mediated by the variability in sodium concentration present in human urine matrix.....</b>	<b>22</b>
<b>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. ....</b>	<b>25</b>
<b>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.....</b>	<b>25</b>
<b>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). ....</b>	<b>28</b>
<b>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.....</b>	<b>31</b>
<b>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.</b>	

As a result, the assumption of constant ratio between half bandwidths is not accomplished and the estimation of accurate half bandwidths is compromised.....	31
<b>Figure 3-14 As the amount of data increases, the performance of DL approaches trumps the one of traditional ML techniques.....</b>	<b>39</b>
<b>Figure 3-15 Tree-based algorithms showed the best performance in the evaluation of different traditional ML algorithms in 165 different datasets. ....</b>	<b>39</b>
<b>Figure 3-16 Comparison (in accuracy and speed) of different clustering algorithms when dealing with different data patterns. ....</b>	<b>40</b>
<b>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. ....</b>	<b>57</b>
<b>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. ....</b>	<b>60</b>
<b>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. ....</b>	<b>60</b>
<b>Figure 4-4 Example of available information of reported concentrations in the HMDB website (top) and the equivalent information present in XML format (down). ....</b>	<b>62</b>
<b>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. ....</b>	<b>63</b>
<b>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. ....</b>	<b>65</b>
<b>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</b>	

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

<b>Table 5.1 Chemical shift information shows discriminative potential in the MTBLS1 dataset.</b> However, it cannot enhance the excellent results given by concentration information during RF classification. ....	81
<b>Table 5.2 Adding chemical shift information to concentration information improved the classification between the five different kinds of sample in the MTBLS237 dataset.</b> Several quality indicators of the models generated are shown. ....	81
<b>Table 5.3 Adding chemical shift information to concentration information provides the best classification of samples in the MTBLS374 dataset.</b> 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
<b>Table 5.4 Ranked predictors in RF classification of samples with both con-centration and chemical shift information in the MTBLS374 dataset.</b> There are few predictors because of the recursive feature ex-traction of non- discriminative features.....	91
<b>Table 5.5 Additional classification indicators in the MTBLS1 dataset.</b> .....	91
<b>Table 5.6 Additional classification indicators in the MTBLS237 dataset.</b> .....	92
<b>Table 5.7 Additional classification indicators in the MTBLS374 dataset.</b> .....	93
<b>Table 6.1 The predicted signal parameter information increases Spearman’s rho correlation between metabolite concentrations in MS and NMR data in both datasets.</b> 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

