Machine Learning Approach for Metabolite Profiling in Complex Mixtures via NMR Data

Rahul Madhav
School of Biological Sciences
National Institute of Science Education and Research, Bhubaneswar
rahulmadhav.m@niser.ac.in

Rabmit Das
School of Biological Sciences
National Institute of Science Education and Research, Bhubaneswar
rabmit.das@niser.ac.in

Abstract

In our project, we try to develop a machine learning algorithm for metabolite profiling and quantification in complex mixtures via NMR data. We chose urine as our complex mixture to develop the model on. We synthetically generated the dataset using RDolphin and added noise by varying positions in a small range and changing peak heights. We plan to use a simple regression model in the beginning and go on to use CRNN which has shown promising results in the past.

1 Introduction

We encounter numerous complex mixtures in various fields such as food science, biomedical research, metabolomics, health science, etc. Typically, we utilize analytical chemistry tools such as chromatography coupled with mass spectrometry (GC-MS, LC-MS) for analysing these mixtures. While these techniques offer high sensitivity and resolution, they often entail extensive sample preparation and have limited coverage of chemical classes. Additionally, they may require the use of multiple separation and detection methods to comprehensively characterize complex mixtures, resulting in increased complexity and time investment.

Nowadays, NMR spectroscopy has become increasingly prevalent in various industries and research fields. This technique is being widely adopted for the analysis of complex mixtures due to its non-destructive and non-invasive nature. It allows researchers and professionals to simultaneously detect and identify multiple compounds within mixtures without the need for extensive sample preparation or separation steps. NMR is being utilized in diverse areas such as biomedical research, environmental science, food analysis, and industrial manufacturing, highlighting its versatility and effectiveness in modern analytical practices.

Analysing NMR spectra presents several challenges due to their high complexity, comprising thousands of data points from numerous metabolites with varying sensitivities. Extracting informative signals requires significant computational power and dedicated screening pipelines. Additionally, clusters from different metabolites may heavily overlap, posing difficulties in accurate signal decompositions, particularly for low-concentration metabolites neighbouring dominant ones. Furthermore, NMR spectra are affected highly by the environment and instrumental variations, leading to alterations in signal patterns and positional shifts compared to reference libraries.

Despite the above obstacles, the most common method for metabolite analysis in NMR spectra presently depends on expert annotation. NMR professionals use software like Chenomx to manually

match probable metabolite candidates from a reference library to the mixture, resulting in a final fitting report. However, this procedure takes a lot of time, relies on professional judgement, and is prone to subjectivity, which may jeopardise repeatability. The complexities and variations involved limit the widespread application and hinder the efficient interpretation of NMR spectra, particularly in clinical settings. Hence, exploring the integration of machine learning (ML) holds promise. ML algorithms could offer automated and objective metabolite identification and quantification, overcoming the limitations of manual annotation and enhancing reproducibility and efficiency in NMR-based metabolomics analyses.

2 Related Works

There have been several attempts in developing a ML model to identify and quantify metabolites and most of them outperforms the available automated alternatives.

For instance, BQuant adopts a probabilistic strategy for this purpose. This method still relies on database-based classification and measurement of metabolites within neighbouring features of the 1H NMR plot. It involves representing the sample as a combination of reference profiles from a database and inferring the classification and quantification of metabolites through Bayesian model selection.

DeepMID (deep-learning-based mixture identification method), uses a pseudo-Siamese convolutional neural network (pSCNN) and a spatial pyramid pooling (SPP) layer to use to identify plant flavours (mixtures)

SMART-Miner uses a CNN based algorithm for metabolite identification from 1H-13C HSQC spectra Also, lot of targeted profiling have been done using PCA-based pattern recognition.

3 Methodology

3.1 Synthetic Data Generation

Manual annotation of metabolites and spectral features is typically necessary for training ML models, which can be labour-intensive and time consuming, especially in cases of overlapping peaks and ambiguous signals. Furthermore, the variability in experimental conditions, sample matrices, and instrument parameters across different datasets adds to the complexity of data preprocessing and standardization. Consequently, given the constraints of time and resources, opting for synthetic datasets may offer a more feasible approach for us.

We followed the instructions provided on https://github.com/danielcanueto/rDolphin. The packages utilized are from the same library they used for blood sample analysis. The parameter "Chemical shift tolerance (ppm)" in the Region of Interest (ROI) was adjusted from the default value of 0.002 to 0.02 to get better mapping results.

To test the feasibility of the proposed workflow, we initially selected 14 representative metabolites. Isoleucine, valine, 3-isobutyric acid, 3-hydroxy-isovalyric acid, alanine, acetic acid, phenyl acetyl glutamine, citric acid, creatine, creatinine, urea, hippuric acid, taurine, hydroxy phenyl acetic acid were the ones which was selected. These metabolites were chosen to represent various pattern complexity and chemical shift values. The number of metabolites and metabolite candidates was selected at random in simulated samples. Because capturing background noise and possible contamination in genuine NMR spectra is challenging using synthetic data alone, we manually incorporated these noises by creating both synthetic and real clean spectra for each sample.

Furthermore, for each simulated combination, we randomly picked one metabolite and combined its genuine NMR spectrum with the remaining synthetic spectra from other samples. This approach incorporates genuine differences inside synthetic samples while preventing them from becoming unduly complex by heavy background noise.

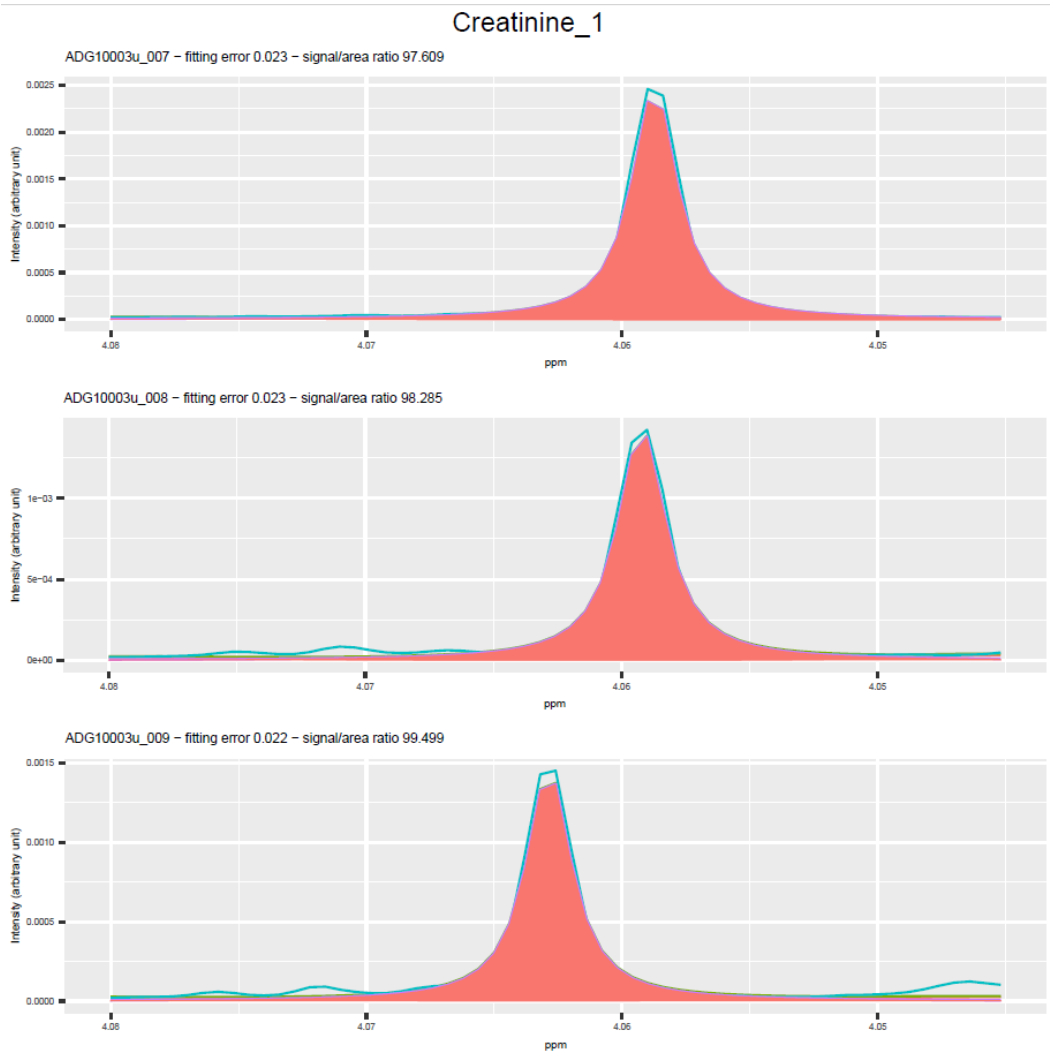

Figure 1: Possibilities of a single peak

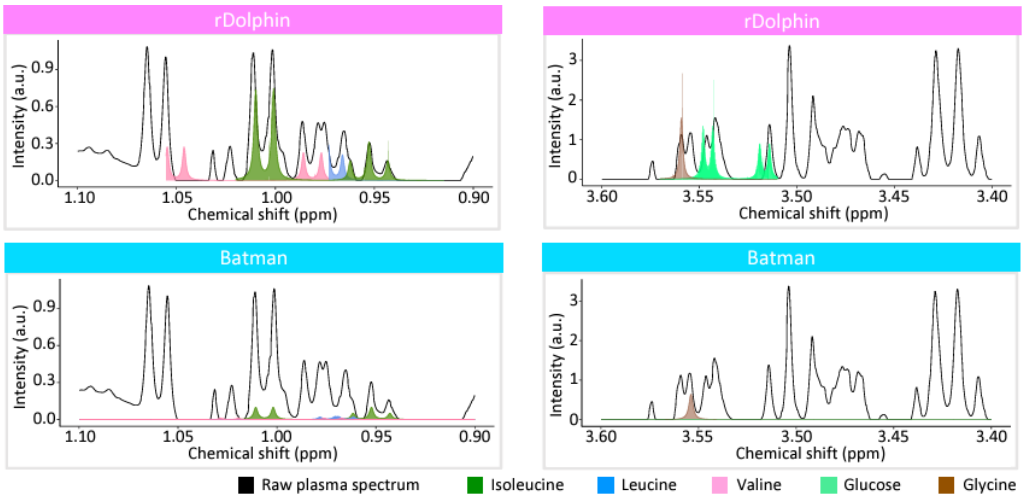

Figure 2: Simulated Data

3.2 NMR Data Preprocessing

In every experimental unprocessed 1H NMR data, potential disturbances from the machine or environment can introduce noise in the spectra, leading to bias in identifying or quantifying metabolites. To address this issue, Martin et al. developed a preprocessing pipeline for NMR data implemented in PepsNMR. This pipeline includes baseline correction, which estimates and removes the baseline from the sample to achieve ideal data. In general, the informative metabolite region in an NMR spectrum ranges from -1 to 11 ppm on the axis. For our study involving fourteen metabolites, the entire region was retained. However, in urine analysis, most metabolite signals are concentrated between 0.7 and 4.5 ppm, thus only bands in this range were kept for computational effectiveness. To suit the CRNN model's input, the NMR spectra were scaled using an interpolation step. Then, normalisation methods were applied to guarantee that all combinations were detected and measured on the same scale. The classification model utilized min-max normalization to adjust the data from 0 to 1, while the quantification model normalized the areas under the curve of the samples to a constant value of 1.

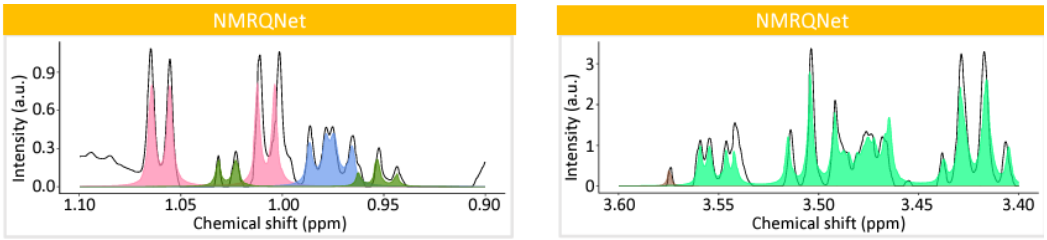

Figure 3: Sample data after adding noise

3.3 Baseline Model

We had planned to try out a linear regression model on the dataset within this time frame which has not been possible due to time constraints. We have decided to go ahead with the previous plan and use a regression model first and then a CRNN model since it has shown promising results in the past.

3.4 Upcoming Plans

By the end of this project, we plan to:

[1] Run a linear regression model for quantification and a random forest model for profiling.

[2] Use two models - First profiling the components of the mixtures (CNN) and second quantifying the classified components (RNN).

[3] Combine the two models (CRNN) and compare the individual accuracy.

References

[1] Wang, Y., Wei, W., Du, W., Cai, J., Liao, Y., Lü, H., Kong, B., & Zhang, Z. (2023) Deep-Learning-Based mixture identification for nuclear magnetic resonance spectroscopy applied to plant flavors. *Molecules, 28(21),*, pp. 7380. https://doi.org/10.3390/molecules28217380

[2] Zheng, C., Zhang, S., Ragg, S., Raftery, D., & Vitek, O. (2011) Identification and quantification of metabolites in 1H NMR spectra by Bayesian model selection. *Bioinformatics, 27(12)*, pp. 1637–1644. https://doi.org/10.1093/bioinformatics/btr118

[3] Weljie, A. M., Newton, J., Mercier, P., Carlson, E. E., & Slupsky, C. M. (2006) Targeted Profiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry, 78(13)*, pp. 4430–4442. https://doi.org/10.1021/ac060209g

[4] https://github.com/danielcanueto/rDolphin

[5] https://github.com/LiuzLab/NMRQNet