Heliyon 10 (2024) e36998

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

NMR-Onion - a transparent multi-model based 1D NMR deconvolution algorithm

Mathies Brinks Sørensen [a], Michael Riis Andersen [b], Mette-Maya Siewertsen [a], Rasmus Bro [c], Mikael Lenz Strube [d], Charlotte Held Gotfredsen [a,*]

[a] Department of Chemistry, Technical University of Denmark, Kgs Lyngby, DK-2800, Denmark
[b] Department of Applied Mathematics and Computer Science, Kgs Lyngby, DK-2800, Denmark
[c] Department of Food Science, University of Copenhagen, Frederiksberg, DK-1958, Denmark
[d] Department of Biotechnology and Biomedicin, Kgs Lyngby, DK-2800, Denmark

ARTICLE INFO

Keywords:
Statistical evidence
Time domain models
Open source
Deconvolution
High sensitivity
Extensive overlaps
Computationally efficiency

ABSTRACT

We introduce NMR-Onion, an open-source, computationally efficient algorithm based on Python and PyTorch, designed to facilitate the automatic deconvolution of 1D NMR spectra. NMR-Onion features two innovative time-domain models capable of handling asymmetric non-Lorentzian line shapes. Its core components for resolution-enhanced peak detection and digital filtering of user-specified key regions ensure precise peak prediction and efficient computation. The NMR-Onion framework includes three built-in statistical models, with automatic selection via the BIC criterion. Additionally, NMR-Onion assesses the repeatability of results by evaluating post-modeling uncertainty. Using the NMR-Onion algorithm helps to minimize excessive peak detection.

1. Introduction

Deconvolution of 1D spectra in NMR spectroscopy is a key step when elucidating complex 1D ^{1}H NMR spectra. The spectra contain extensive structural, quantitative, and dynamic information. This information, extractable even from 1D spectra, is vital in many fields of science studying complex mixtures, such as metabolomics [1] and *in situ* samples. Due to hundreds of overlapping signals from compounds in varying concentrations, traditional manual spectral analysis is inadequate. Automated extraction methodologies, combined with spectral database comparisons, can facilitate this process. The most common data extraction approaches involve deconvolution [2][3] or binning of frequency buckets [4][5]. The latter, especially with intelligent bucketing [6], has been widely applied in metabolomics. This approach has been successfully applied in disease diagnostics [7][8], natural product identification [9][10], foodomics [11], and drug discovery [12]. Unlike binning, deconvolution focuses on resolving single peaks, potentially revealing information from small peaks otherwise lost in the binning process. However, the increased information content also increases complexity in spectral interpretation, mathematical modeling, and computational demands. Over the past 30 years, significant advancements have addressed these complexities. Bretthorst's pioneering work established a probabilistic framework for modeling the shape and number of NMR signals based on free induction decay (FID) data [2]. Building on Bretthorst's mathematical framework, the Craft method [13] introduced probabilistic modeling with digital filtering, reducing computational complexity. These approaches

* Corresponding author.
E-mail address: chg@kemi.dtu.dk (C.H. Gotfredsen).

https://doi.org/10.1016/j.heliyon.2024.e36998
Received 8 December 2023; Received in revised form 23 August 2024; Accepted 26 August 2024

Available online 30 August 2024

focus on time-domain data, where peaks are modeled as sums of exponentially damped sinusoids, corresponding to Lorentzian line shapes in the frequency domain [14]. Recent research combined Gaussian and Lorentzian line shapes into pseudo Voigt line shapes for a frequency domain model [15,16]. This model is implemented in the commercial MNOVA GSD software [17] and the R software package rNMRfit [18]. In rNMRfit, a baseline correction method that produces robust results was further implemented. Additionally, the frequency domain methods have been integrated into algorithms designed to match deconvoluted NMR signals with databases. This has resulted in the automatic detection of compounds, as demonstrated by popular frameworks such as NMRbatman [19], Bayesil [20], and the commercial program Chenomx [21].

Finally, the latest methods applied within 1D deconvolution are based on deep learning methodology. Specifically, the methods employed are based on transfer learning, training neural networks on larger simulated datasets. One transfer learning example is found within the DEEP picker network [22], in which signals are simulated from a classic Voigt FID model. A more complex example is found within the Vogit Fitter1D algorithm [23], which is combined with the Deep Picker1D algorithm [23]. The workflow of the two algorithms is almost identical to the deep picker algorithm, but the Voigt Fitter1D is capable of removing "odd" shapes such as shimming errors, broad peaks, and narrow peaks. Whilst deep learning holds a lot of promise for deconvolution, the weakness comes from training data, as currently most of the training data is based on simulated data rather than real data, potentially leading to large training bias and a loss of generalization.

Following 30 years of research, the field of deconvolution still faces challenges. These challenges include providing statistical evidence of model quality compared to other models (model selection), addressing parameter uncertainty, ensuring generalization, and establishing statistical evidence for the presence or absence of highly overlapping peaks. To address these issues, we propose a novel five-step process called NMR-Onion. This method evaluates model quality by selecting the optimal model for estimating frequencies, coupling constants (within a frequency distance matrix), amplitudes, and parameter uncertainty within a user-specified region of interest (ROI). Furthermore, as part of the evaluation of parameters uncertainty, confidence intervals are generated, enabling a statistically-based assessment of overlapping signals. The NMR-Onion framework is carried out as a hybrid approach between the frequency and time domain, in which the frequency domain is utilized for peak detection and the time domain is utilized for the actual modeling of individual signals (e.g. peaks). The motivation for choosing the hybrid approach is: (1) The frequency domain peak picking approach is very well established, fast, and reliable through the application of Savitzky Golay filtering derivatives [24]. In contrast, the time domain is highly convoluted, and while subspace methods exist for detecting frequencies, they scale poorly with number of points and number of frequencies [25,26]. (2) Time domain models are more robust at handling frequency domain artifacts such as baseline distortions, which affects the full frequency domain, while in the time domain, only the first few points are affected which can be removed or down-weighted [27,28]. Another benefit comes in the form of reduced parameter space, as phase and amplitudes are estimated from the time coupled parameters [29,13], whereas the frequency domain requires explicit estimation through optimization [18].

The five-step process that constitutes the NMR-Onion algorithm is outlined below and visually presented in Fig. 1.

- Step 1: The computational burden of the algorithm during the fitting process is reduced by applying a digital band-pass filter, which generates a user-specified region of interest (ROI).
- Step 2: Peaks within a ROI are identified by applying the first and second-order derivative Savitzky Golay filter [24] in tandem with resolution enhancement [14].
- Step 3: Multiple times domain models are applied to the ROI. To address the multimodality in frequency estimations, a peak detection algorithm is employed to generate initial parameter inputs.
- Step 4: The best model is selected using a likelihood-based information criterion [30]
- Step 5: Parameter uncertainty is evaluated applying the wild bootstrap algorithm [31].

2. Theory

2.1. Model formulation

The NMR-Onion algorithm aims to identify chemical shifts, intensities, distances within a given multiplet (for obtaining J - spin-spin coupling constant information), and the number of underlying signals present in an NMR spectrum. Utilizing the time domain, the spectrum can be expressed mathematically as a sum of damped complex sinusoids:

$$y(t) = \sum_{k=1}^{K} A_k \cdot exp(2j\pi v_k t + j\phi_k) \cdot \Psi_n(\boldsymbol{\rho_k}). \tag{1}$$

In equation (1) the term $\Psi_n(\boldsymbol{\rho_k})$ represents the n'th decay function, where $\boldsymbol{\rho_k}$ denotes a vector which represents all parameters associated with the n-th decay. The decay function, is in its most simple form, expressed as a negative exponential function, $\Psi_1(\boldsymbol{\rho_k}) = exp(-\alpha_k t)$, as described by Keeler [14] and in equation (2).

$$y(t) = \sum_{k=1}^{K} A_k \cdot exp(2j\pi v_k t + j\phi_k) \cdot \Psi_1(\boldsymbol{\rho_k}). \tag{2}$$

The parameters in equation (2) are defined as follows: v denotes the frequency (Hz), A represents the amplitude, ϕ denotes the phase, and α signifies decay rate for the k'th sinusoid. The validity of equation (2) holds under the assumption that no artifacts affect the

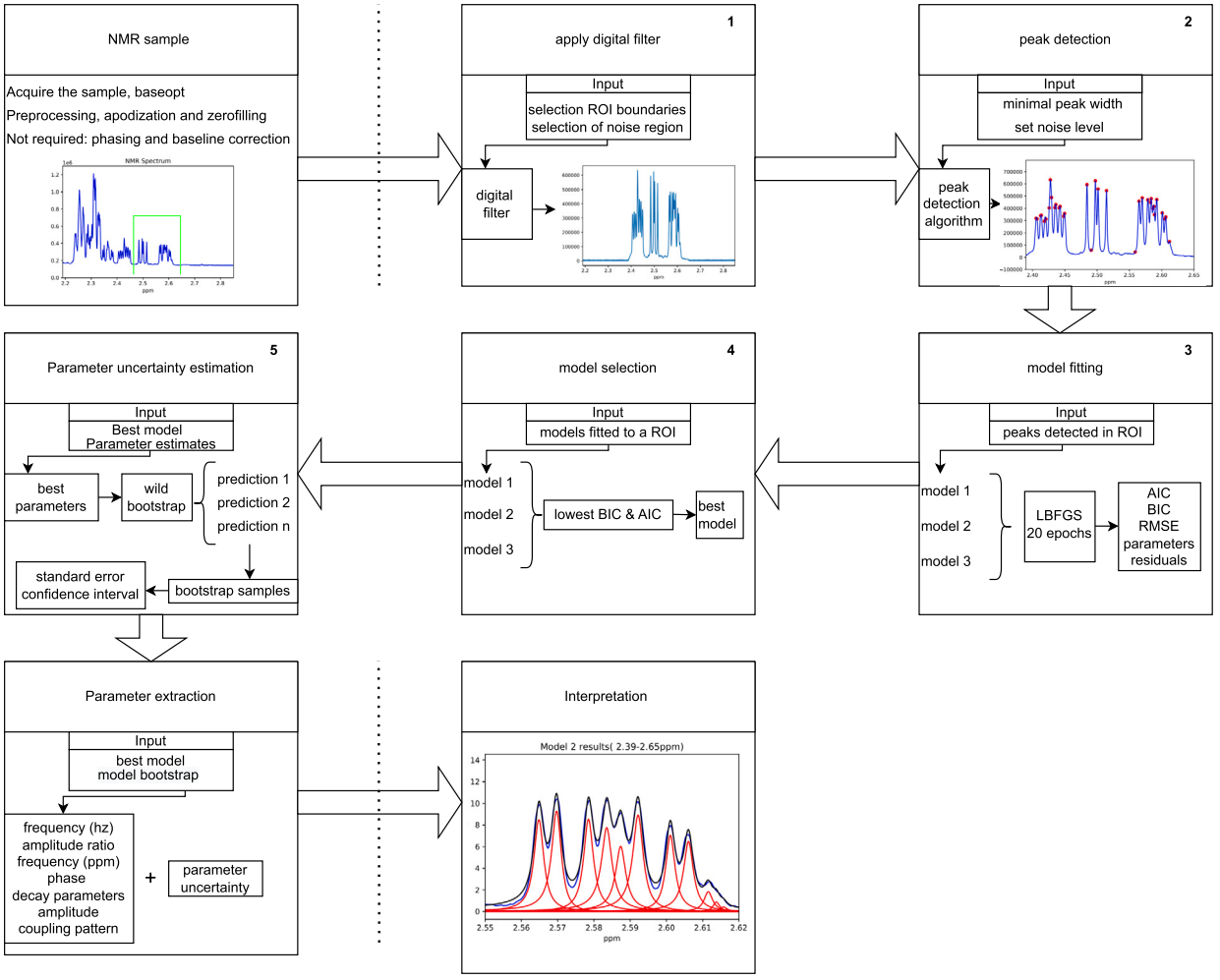

Fig. 1. Visual representation of the NMR-Onion algorithm outlined in the preceding five steps. The steps detailed below are labeled with numbers 1-5 in the figure.

experimental data. However, in practice, artifacts such as shimming, eddy currents, temperature fluctuations, receiver gain settings, sample conditions, etc., may impact the acquired spectrum in an unpredictable fashion [32]. To the best of our knowledge, no existing approach has successfully accounted for all distortions. Therefore, we aim to extend the damping term of $\Psi(\boldsymbol{\rho_k})$ to accommodate some of the aforementioned distortions by introducing a flexible decay model instead of a purely exponential decay. To achieve this flexibility, we propose two novel time domain models. The first model is a weighted sum of Gaussian and exponential decays, analogous to a pseudo-Voigt shape in the frequency domain, whereas the second model consists of an exponential power law model. The pseudo-Voigt (equation (3)) will result in a redefined decay term of equation (1):

$$\Psi_2(\boldsymbol{\rho_k}) = (1-\eta_k)\cdot\exp(-\alpha_k t) + \eta_i \cdot \exp(-\alpha_k t^2) \tag{3}$$

resulting in the full model of the weighed sums being

$$y(t) = \sum_{k=1}^{K} f(A_k, \nu_k, \phi_k)\cdot\Psi_2(\boldsymbol{\rho_k}) \tag{4}$$

With f (see equation (5)) being equal to the harmonic term of equation (1):

$$f(A_k, \nu_k, \phi_k) = A_k \cdot \exp(2j\pi\nu_k t + j\phi_k). \tag{5}$$

The addition of the η term introduces a weighting between an exponential and Gaussian decay type. When $\eta = 0$, a pure exponential decay is obtained, reducing equation (4) to equation (2). If $\eta = 1$ a pure Gaussian decay is achieved since the $\exp(-\alpha_k t)$ term becomes zero. A further generalization of (2) is achieved in the second model with an introduction of a power term $\Psi_3(\boldsymbol{\rho_k}) = \exp(-\alpha_k t^{\beta_k})$ as the decay function, resulting in:

$$y(t) = \sum_{k=1}^{K} f(A_k, \nu_k, \phi_k) \cdot \Psi_3(\boldsymbol{\rho_k}) \tag{6}$$

The addition of the β term introduces a stretched exponential decay when $\beta > 1$ and a compressed exponential when $0 < \beta < 1$. Finally, equation (6) reduces to the classic exponential decay (equation (2)) when $\beta = 1$.

The advantage of the power law decay model of equation (6) and exponential mixture model of equation (4) lies in their flexibility to describe non-Lorentzian peak shapes. To further enhance the flexibility of the models presented in equations (2), (4) and (6), asymmetric line shapes are incorporated inspired by the approach outlined in the works of Matviychuk [32], introducing a complex skewing term of $exp(j\gamma_k)$ for each signal. Though NMR signals are in principle symmetric, this is not always the case, as slight shimming errors or eddy currents may affect the spectrum in a non-predictable fashion, introducing asymmetric lineshapes [32,33]. This results in equations (2), (4), and (6) being formulated with added terms to accommodate asymmetric line shapes as:

$$y(t) = \sum_{k=1}^{K} f(A_k, \nu_k, \phi_k) \cdot \Psi_1(\boldsymbol{\rho_k}) \cdot \exp(\exp(j\gamma_k)t) \tag{7}$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \nu_k, \phi_k) \cdot \Psi_2(\boldsymbol{\rho_k}) \cdot \exp(\exp(j\gamma_k)t) \tag{8}$$

$$y(t) = \sum_{k=1}^{K} f(A_k, \nu_k, \phi_k) \cdot \Psi_3(\boldsymbol{\rho_k}) \cdot \exp(\exp(j\gamma_k)t) \tag{9}$$

respectively. The $exp(j\gamma_k) \cdot t$ causes the peak to skew leftward if $\gamma > 0$ and rightward if $\gamma < 0$, with γ constrained to $[-\frac{\pi}{2} : \frac{\pi}{2}]$. With the formulated models, a routine for parameter estimation can be developed by transforming equations (7), (8), and (9) into a non-linear least squares optimization (NLS-opt) problem. The NLS-opt problem involves setting up a matrix formulation of all models, incorporating a residual term E as the following:

$$Y = ZA^T + E. \tag{10}$$

Here, A is a $1 \times K$ where each element represents a complex amplitude $a_k = A_k \cdot \exp(2j\pi\phi_k)$ and Z is a $N \times K$ matrix consisting of the time-dependent terms from equations (7), (8) and (9). Each column of the Z-matrix represents a single sinusoid with its own subset of parameters, while the rows correspond to the signal values at the n-th time point. Y is a $1 \times N$ vector where each element represents a measured time point in the FID. Finally, E is a $1 \times N$ residual vector assumed to be identically and independently distributed according to a Gaussian distribution $E \overset{\text{i.i.d}}{\sim} N(0, \sigma)$. The model of equation (10) is simplified by integrating out the none-time dependent terms, following the same method originally suggested by Bretthorst [2], the A matrix can be expressed as a function of the Z matrix (see equation (11)):

$$A = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{YZ} \tag{11}$$

This enables for the model to be turned into an NLS-opt problem, by reformulating equation (10) as minimization of the sum of squared errors (SSE) loss function depending soley on the Z-matrix, as presented in equations (12) and (13).

$$E = Y - AZ = Y - ((\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{YZ})Z \tag{12}$$

where

$$\text{SSE} = E^H E \tag{13}$$

and

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta}(\text{SSE}) \tag{14}$$

Here, H denotes the complex conjugate transpose of the matrix, also referred to as the Hermitian transposed matrix, and $\hat{\boldsymbol{\theta}}$ represents a vector of estimated parameters. Ideally, the SSE formulation of equation (14) poses a standardized loss function for minimizing. However, owing to the model's nature as a superposition of sinusoids, the likelihood of spurious signals arising is very high. Therefore, to minimize the variance of phases a penalty term is added into equation (14)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} SSE + \frac{1}{K}\sum_{k=1}^{K}(\phi_k - \bar{\phi})^2 \tag{15}$$

The penalty term of equation (15), where $\bar{\phi}$ denotes the mean phase vector, ensures that the phases do not deviate excessively, thereby preventing large reversed-phase peaks from occurring. One should note that the magnitude of a phase is very small, ranging from $[-\pi : \pi]$, compared to the magnitude of the SSE in an NMR spectrum. Therefore, to ensure the penalty criterion has an effect, all FID data is normalized using the Frobenius norm [34]. Optimal estimation of the parameters from the loss function of equation (15) may appear as a straightforward optimization problem at first glance. However, achieving an estimation of a global minimum

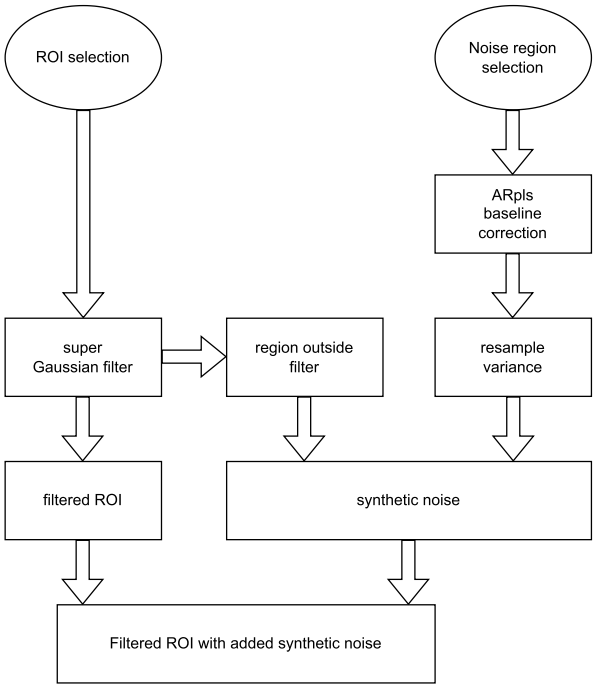

Fig. 2. Visual representation of the digital filter workflow. The ellipses indicate user input, specifying the targeted ROI and noise region (in ppm). The output is a digitally filtered ROI containing only the signal within the target range, while the rest of the spectrum is filled with synthetic noise.

to obtain optimal parameters for (7), (8), and (9) is notoriously challenging. This is due to several factors: the model's unknown number of components (K), its non-linearity, multi-modality with respect to the frequencies (ν_k), and the computational expense of the optimization algorithms involved. Therefore, the following subsections provide insights into reducing computational burdens, estimating model order, and the handling of parameter multimodality. The numerical constraint details for each parameter can be found in the constraining parameters supplementary (section A).

2.2. Computational bottlenecks

When applying the NMR-Onion algorithm to analyze a metabolomic 1D ^{1}H NMR spectrum, it is common to have more than 1000 peaks. This poses a significant computational challenge due to the size of the Z-matrix described by equation (10), which becomes large with $K > 1000$ and $32768 < N < 131072$ time points, depending on the sample acquisition protocol. Fortunately, 1D ^{1}H NMR data exhibits sparsity, with non-overlapping regions being independent. These spectra often contain large regions of redundant noise, and typically only specific regions of the spectrum are of interest. Hence, by focusing on smaller regions of interest (ROIs), the dimensions of the Z-matrix can be substantially reduced, resulting in fewer columns (K). This reduction is achieved through the application of a digital band-pass filter as outlined in step 2 of Fig. 1.

Various methods exist for implementing digital filters. For example, the CRAFT algorithm utilized a finite impulse response (FIR) filter in conjunction with a Blackman window function [13], while Djermoune employed wavelet packet-like filter banks in an adaptive subband filtering scheme [35]. In NMR-Onion, we have adapted the super-Gaussian band-pass digital filter proposed by Hulse and Foroozandeh [36], making modifications to accommodate baseline artifacts and improve noise estimation in the data. This approach was chosen because integrating prior knowledge about noise levels is advantageous for NMR deconvolution, facilitating easier separation of signals from noise.

The initial modification made to the filter involved integrating a baseline correction step prior to estimating the noise level within a noise region. For baseline correction, the algorithm of asymmetrically re-weighted penalized least squares smoothing (ARpls) [37] was applied. In addition, the noise level was determined using a resampling approach designed to enhance the robustness of noise estimation. Assuming the noise follows an average Gaussian white noise pattern, 1000 samples were randomly drawn from a Gaussian distribution with a mean of 0 and variance obtained from the baseline-corrected noise region. The mean of these resampled values was then used to establish an artificial noise floor for the filter. The adjusted filtering process is illustrated in Fig. 2.

Another bottleneck, in addition to model dimensions, arises from the optimization technique, programming language, and the specification of loss function derivatives. To address these challenges, we employ the modern PyTorch framework to implement the loss function defined in equation (15). We employ the quasi-Newton optimizer based on the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm [38], which demonstrated considerably faster convergence for the NMR-Onion algorithm compared to the Scipy implementation [39]. Moreover, PyTorch's autograd module provides automatic differentiation (AD), further enhancing the robustness and efficiency of the optimization process [40]. AD furnishes both gradient and Hessian information,

crucial for swiftly and effectively estimating model parameters without the manual effort of identifying and implementing first-order (gradient) and second-order (Hessian) derivatives. Finally, we have included the option of applying a GPU (currently only applicable for Linux) during the parameter estimation step which leads to an approximately 5-fold increase in speed compared to running on a CPU (see results section for more information).

2.3. Peak detection

In addition to computational challenges, managing the multi-modality of frequencies and accurately estimating the number of signals for modeling purposes presents considerable difficulty. Initially addressed by Bertthorst [2], this challenge involved employing a search pattern algorithm [41] in conjunction with maximum power spectral density to determine initial frequency values. Sequential fitting of peaks continued until the Bayesian generalized likelihood criterion identified the correct number of components. Rubtsov and Griffin proposed an alternative method using reversible Monte Carlo Markov Chain (MCMC) jumps for model order selection and parameter estimation [29]. Frequency-based peak detection methods the first and second-order derivative Savitzky-Golay (SG) filters [24] are popular choices, especially with high-field NMR data that provide high-resolution spectra [42]. The NMR-Onion algorithm enhances peak detection using third-order SG filter derivatives combined with resolution enhancements [14] and employs a clustering algorithm that merges potentially overlapping peaks. This approach draws from the rNMRfind algorithm [43], implemented in Python as the third building block detailed in Fig. 1. The detection algorithm requires user input to define the minimal peak width parameter. Generally, as demonstrated in the original rNMRfind work [43], increasing this parameter reduces the number of detected peaks, offering a more conservative detection approach. The default noise threshold is set at 2 times the standard deviation determined by the interquartile range (IQR) of the first principal component from SG-filtered first and second-order derivative spectra of the real and imaginary spectrum.

2.4. Model selection

Following model fitting, the model selection process described in the fourth building block of Fig. 1 is executed. Various approaches can be employed for selection criteria, but for the NMR-Onion algorithm, we opt for the user to perform the selection utilizing either the Akaike Information Criterion (AIC) [44] or the Bayesian information criterion (BIC) [45] defined as

$$AIC = -2 \cdot \mathcal{L}(\theta) + 2 \cdot p \cdot K \tag{16}$$

$$BIC = -2 \cdot \ \mathcal{L}(\theta) + 2 \cdot p \cdot K \log(N) \tag{17}$$

Here p represents the number of parameters per sinusoid, K denotes the total number of sinusoids, and $\mathcal{L}(\theta)$ denotes the log-likelihood of equation (4) which according to the works of Nadler [46] may be formulated as

$$\mathcal{L}(\theta) = \frac{-N}{2}(1 + \log(2\pi)) - \frac{N}{2}\log[E^H E]. \tag{18}$$

Equation (18) corresponds to a Gaussian log-likelihood computed by utilizing the maximum likelihood estimator (root mean squared error) for the variance. The final model is determined based on its ability to provide the most accurate description of the data utilizing the minimal number of components, as indicated by achieving the lowest value from either equation (16) (AIC) or (17) (BIC). The primary distinction between using (16) and (17) lies in the stringency of the penalty imposed based on the number of model parameters. BIC imposes a stricter penalty on models with more parameters compared to AIC. Consequently, BIC tends to favor simpler models over AIC. The NMR-Onion algorithm considers both approaches as valid and allows for testing of both under different scenarios. Other criteria, such as the kmap criterion [47], have also been developed but are primarily applicable to non-decaying sinusoids and do not explicitly address the weighting of decay rate and flexibility constants.

2.5. Parameter uncertainties

An often overlooked aspect in deconvolution algorithms is the assessment of uncertainties, particularly regarding the repeatability of peaks and the procedure for estimating parameter uncertainties. Various approaches exist for uncertainty estimation, with a common method involving the application of the second order derivative of the model's negative log-likelihood function to derive the Fisher information used in Wald approximation confidence intervals [48]. However, as demonstrated by Wilson [3], the profile likelihood of parameters, especially frequencies, are far from representing a second-order curvature. Therefore, to accurately quantify uncertainties, alternative methods are necessary. A robust approach involves employing various Monte Carlo Markov Chain (MCMC) schemes, as illustrated by Jie Hao [19], although this approach intensifies computational cost.

In the NMR-Onion algorithm, we choose to implement a frequency approach rather than a Bayesian model, utilizing the ad hoc method of wild bootstrapping [31] to estimate parameter uncertainties. This scheme is illustrated in Fig. 1 step five and detailed in Algorithm 1 above. The purpose of the bootstrap method shown in Algorithm 1 is to estimate the confidence interval (CI) for each parameter. This is especially important for frequencies, as overlapping CIs would suggest that highly overlapping peaks might not be consistently detected in replicates, as within a replicate, the resolved peak might merge into a single peak. We classify peaks with overlapping CIs as potential resolved peaks (PRPs). To assess the repeatability of the PRPs, independent experimental replicates should be produced to determine whether the overlapping peaks are consistent or a result of random sample variations.

Algorithm 1 Wild bootstrap algorithm.

Compute residuals based on best model fit
$\varepsilon_i = y_i - \hat{y}_i$
draw bootstrap samples
for $b = 1, 2, .., B$ **do**
  for $i = 1, 2..., N$ **do**
    $\bar{\varepsilon}_i^b \sim Z_i^b \cdot \varepsilon_i$, $Z_i^b \sim N(0, 1)$
    $\hat{y}_i^b = \hat{y}_i + \bar{\varepsilon}_i^b$
    $\boldsymbol{\theta_b} = \arg\min_{\theta_b} SSE(\hat{y}_i^b)$
  end for
end for

Generate α level confidence interval for the k'th parameter
$\boldsymbol{\theta}_{k_C I} = [\boldsymbol{\theta}_{k_{b_{\alpha/2}}}, \boldsymbol{\theta}_{k_{b_{1-\alpha/2}}}]$

Here $\boldsymbol{\theta}_{k_b}$ is the k'th parameter CI of b'th estimation at the upper and lower CI value of α

Generate sample variance for the k'th parameter
$\bar{\boldsymbol{\theta}_k} = \frac{1}{b} \sum_{b=1}^{B} \theta_{k_b}$
$\boldsymbol{\theta}_{k_{var}} = \frac{1}{b} \sum_{b=1}^{B} (\boldsymbol{\theta}_{k_b} - \bar{\boldsymbol{\theta}_k})^2$

return $\boldsymbol{\theta_b}, \boldsymbol{\theta}_{k_C I}, \boldsymbol{\theta}_{k_{std}}$

Table 1
Experiment 1: Phenol:isopropanol sample mixture compositions.

Sample No.	Phenol (mM)	Isopropanol (mM)
1	1	1
3	2	1
3	1	2
4	0.5	0.5
5	0.1	0.1

3. Data acquisition

For this study, spectral data were obtained using two different experimental setups. The first experiment consisted of sample mixtures between phenol and isopropanol dissolved in 90:10 H_2O : D_2O and D_2O. The mixture ratios of phenol: isopropanol were set according to Table 1, producing a total of 5 data sets. The second experimental setup involved a sample of the complex molecule phytosteroid Diosgenin dissolved in chloroform. Two samples with an identical final concentration (4 mM) were made. All spectral data were acquired on a Bruker AVANCE III HD 800 MHz spectrometer equipped with a 5 mm TCI cryoprobe. The ^{1}H pulse programs depended on the solvent used, where spectra acquired with only 10% D_2O in water a zgespg pulse sequence was used whereas for samples dissolved in $CDCl_3$ a zg and zg30 pulse sequence scheme was used, for all spectra the baseopt rectangular filter setting was used to minimize baseline and first-order phase distortions. All spectra were acquired at 25 °C, with a relaxation delay of 2 s, 128 scans, and 64 K data points. For the data analysis, the NMR-Onion program was run on a virtual-box Ubuntu (64-bit) Linux operating system with a Processor Intel(R) Core(TM) i9-9880H CPU @ 2.30 GHz, 2304 MHz, 8 Core(s), 16 Logical Processor(s). Furthermore, all experimental data were normalized applying the Frobenius norm before initiating any model fitting within the NMR-Onion algorithm. The modeling process in NMR-Onion utilized a learning rate (lr) of 0.1 over 20 epochs (iteration cycles), employing an exponential learning rate schedule that reduced the learning rate by 30% per epoch.

Post data acquisition, pre-processing was done in Bruker TopSpin version 4.0.7 [49]. This included apodization applying an exponential line broadening of 0.3 Hz from which a subsequent automatic phase and baseline correction was carried out. Finally, the transfer of preprocessed data from Topspin to a Python environment was carried out applying the NMR-Glue [50] package, enabling the importing of Bruker data along with experimental setup parameters (acquisition and pre-processing).

All acquired spectral experimental data (raw and processed) for this paper is available for download on our GitHub: https://www.github.com/Mabso1/NMR-onion.

4. Results

4.1. Simulation study

In order to test the fitting capabilities of the NMR-Onion algorithm, a simulation study was conducted. Specifically, detection and spectral reconstruction were evaluated for 500 datasets simulated at four signal-to-noise (SNR) levels being 30, 20, 10, and 1 dB (1000,100,10 and 1.3 in the definition of S/N) simulated as average Gaussian noise. Each dataset was generated such that the highest multiplicity was that of an octet, whilst mixing of closely spaced signals ensured that multiplets were also present in the simulated data. The setting of each parameter for the simulations can be found within the supplementary (section B). In addition, the noise threshold and peak width filtering were held constant throughout all experiments (see supplementary section B for details).

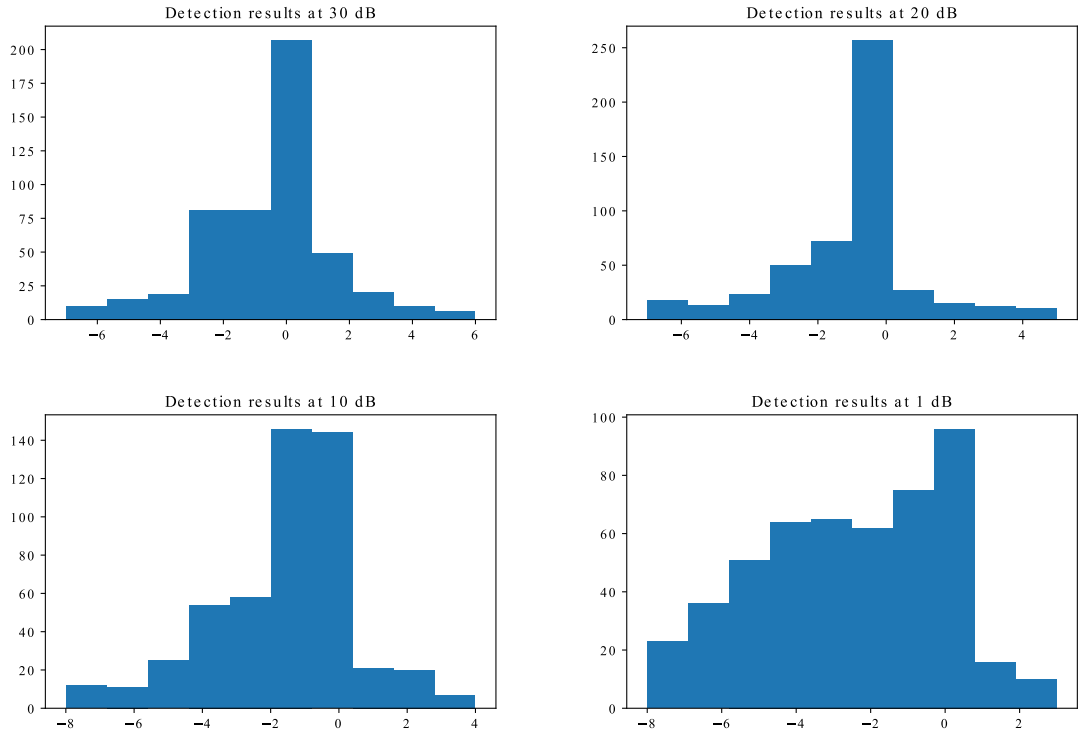

Fig. 3. Visual representation of the detection algorithm performance at different SNR levels. A 0 on the missed peaks axis indicates perfect detection, whilst a negative number indicates how many peaks have been missed within a dataset. Likewise, a positive number indicates how many falsely detected peaks have been added within a dataset. To get the exact numbers, the reader is referred to Table 2 (A) showcases the highest SNR setting of 30 dB. (B) showcases the second-highest setting of 20 dB. (C) showcases the second lowest setting of 10 db. (D) showcases the lowest setting of 1 dB.

Table 2
Results of detection algorithm at different SNR levels. The table shows under-detection (under), over-detection (over) and perfect detection (perfect).

SNR level (dB)	Total	Over	Under	Perfect
30	500	85	206	207
20	500	65	174	259
10	500	48	306	144
1	500	26	376	96

To quantify the detection capabilities of the algorithm, the results were divided into three categories being over-detection, perfect detection, and under-detection. We defined over-detection to occur when the total number of detected signals within a simulated dataset is higher than the total number of actual signals. Likewise, we define under-detection to occur when the number of total detected signals within a simulated dataset is lower than total actual signals. Finally, perfect detection occurs when total number of detected signals within a dataset is equal to the actual number of signals.

The resulting signal detection at each SNR setting is summarized in Table 2 and Fig. 3A-D. The implications of the findings are analyzed within the discussion section, whilst examples of under, over, and perfect detection are shown within the supplementary (section B).

In addition to detection, spectral reconstruction was also evaluated for each SNR setting. As the noise is simulated from average Gaussian noise, a perfect fit would be identified by containing only Gaussian noise within the residuals. Hence we employ the Shapiro Wilk normality test at a 5% significance level, investigating each of the 500 simulations at different SNR levels. The results are summarized in Table 3. In addition, normality plots showcasing some of the residuals are found within the supplementary (section B).

The implication of the findings within Table 3 is further analyzed within the discussion section.

4.2. GPU vs CPU

The testing of GPU speed vs CPU speed was done on a Linux Debian laptop with a 1060 6 GB NVIDIA GTX GPU and a 7th generation i7 core processor. The speed was evaluated by generating multiple runs at one SNR setting of 40 dB, the level was chosen

Table 3
Spectral reconstruction evaluation based on residual normality analysis utilizing the Shapiro Wilk normality test (significance level 0.05). The table shows the number of p-values above 0.05 (fail to reject normality) and below 0.05 (rejects normality).

SNR level (dB)	Total	$p < 0.05$	$p > 0.05$
30	500	477	33
20	500	459	41
10	500	409	91
1	500	384	116

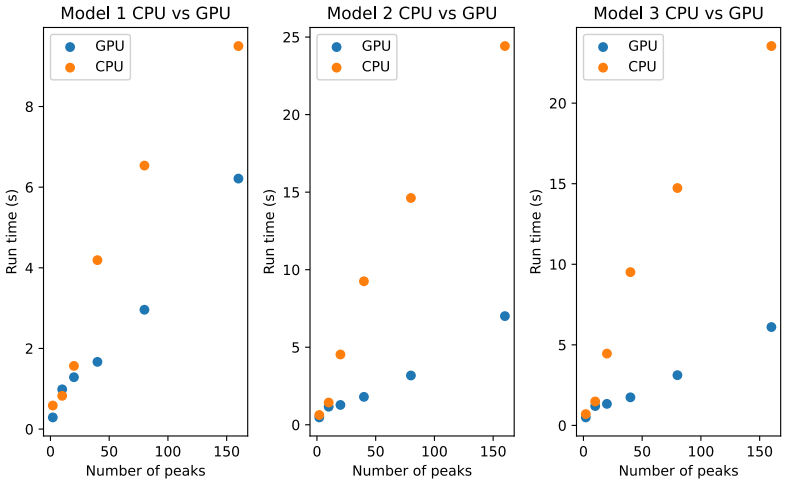

Fig. 4. Visualization of the average GPU vs CPU run times results per epoch. (A) model 1 GPU vs CPU epoch run time. (B) model 2 GPU vs CPU epoch run time. (C) model 2 GPU vs CPU epoch run time.

Table 4
Experiment 1: Depiction of experimental coupling patterns and theoretical peak positions.

Compound	δ (ppm)	Coupling pattern	J (Hz)
Phenol	6.84	broad doublet	~8.9
Phenol	6.91	triplet of triplets	8.0, 0.95
Phenol	7.24	doublet of doublets	7.7, 8.6
Isopropanol	1.16	doublet	6.4
Isopropanol	4.01	septet	6.4

to ensure every peak was not consumed by the noise floor. It should be noted that the purpose was to evaluate the speed of detection, hence every peak position was given to the algorithm passing the detection step. For the series of experiments, the number of model parameters was set at 10, 50, 100, 200, 400, 800 for model 1 and 12, 60, 120, 240, 480, and 960 parameters for model 2 and model 3 (2, 10, 20, 40, 80 and 160 peaks). In total, 100 samples were generated at each parameter setting for each model. The settings of the parameters can be found within the supplementary (section C). The resulting average run time per epoch for each model is outlined within Fig. 4.

From Fig. 4 A-C it is evident that the GPU settings are much faster than solely relying on the CPU when the number of components increases (at a low number of components the results are almost identical). The usage of a GPU led to an approximately 2-fold speed increase for model 1, while models 2 and 3 saw an increase of approximately 5-fold when the number of peaks exceeded 40.

4.3. Case study 1

In the initial experimental setup with real data, the objective is to validate the NMR-Onion algorithm using readily identifiable peak frequencies and coupling patterns covering a large area of the proton spectrum. The primary peaks of the phenol:isopropanol composition are summarized in Table 4

In addition to confirming peak validity, the dilution series (see Table 1) was designed to establish the lower limit of detection. Consequently, regions of interest (ROIs) were defined based on theoretical peak positions applying the filtering process detailed in Section 2.2, resulting in a total of four ROIs (see Table 5).

Table 5
Experiment 1: Region of interest and noise region.

Region No.	lower cutoff (ppm)	higher cutoff (ppm)
1	0.9	1.2
2	3.8	4.1
3	6.6	7.0
4	7.1	7.5
Noise region	-0.1	-0.2

Table 6
Summery of experiment 1 SNR values (both in dB and traditional NMR definition of S/N) for each of the targeted ROIs.

Region No.	Sample No.	SNR (dB)	SNR
1	1	46.9	48978.0
2	1	35.0	3162.3
3	1	38.2	6606.9
4	1	38.2	6606.9
1	2	47.2	52481.0
2	2	34.7	2951.2
3	2	34.3	2691.2
4	2	34.0	2511.9
1	3	34.4	2754.2
2	3	30.7	1174.9
3	3	37.0	5011.9
4	3	36.9	4897.8
1	4	39.7	9332.5
2	4	27.7	588.8
3	4	30.4	1096.5
4	4	30.3	1071.5
1	5	34.2	2630.3
2	5	22.2	165.9
3	5	23.9	245.5
4	5	23.5	223.9

To provide a comprehensive assessment of detection capabilities, the SNR of each ROI was calculated across all concentration levels, as $SNR = 10\log_{10}\frac{S}{N}$ and used to compare the performance of the algorithm (see Table 6). In addition to the dB measurements, we have also included the more traditional SNR definition of NMR being S/N.

Graphical results for the highest and lowest SNR samples are depicted in Figs. 5 and 6. The optimal model for each ROI was automatically selected utilizing BIC of equation (17), (AIC provided similar results).

From Fig. 5 A and Fig. 6 A, the targeted doublet is clearly identified, likewise the targeted septet is found within Fig. 5 B and Fig. 6 B. For figure C, for both the lowest and the highest SNR sample doublets of doublets were identified. The third ROI comprises two sub-ROIs that display second-order effects. Therefore, caution is advised when employing a first-order multiplicity analysis. The first sub-region (Fig. 5 D and 6 D), is observed to be a triplet of triplets as expected if applying a 1st order multiplicity analysis and disregarding J_{para}, whereas for sample 5 broader signals than for sample 1 are observed and the small J coupling constant is not resolved when visually expected but only after deconvolution. For the second sub-ROI of the third region, sample one (Fig. 5 E) is a doublet of multiples, whereas sample 5 (Fig. 6 E) shows different splits making a 1st order multiplicity analysis non-applicable. Regarding the residuals of each model, none satisfy the assumption of being white noise. This issue will be elaborated upon in sections 4.5 and 5.

In addition to conducting individual deconvolution analyses, the stability of models across five identical regions in different samples was investigated. The best models selected across all four regions from 6 are summarized in 7.

From Table 7, it is evident that Model 3, the power law model (equation (9)) is consistently selected as the superior model, indicating a higher level of generalization. Model 2, the mixture model (equation (8)) is less general but still frequently chosen. Interestingly, model 1, the traditional model of exponential decay (equation (7)) is selected only once. It should be noted that each model tested on the data includes a skew term, as outlined in equations (see equation (7), (8) and (9)). To get an overview of the skewness, the distribution of values for γ is visualized in Fig. 7, depicting skewness across each ROI for every sample.

From Fig. 7 it seems that ROI 1 has the least skewness variance compared to the other ROIs. Whereas the other ROIs emit higher variance, e.g. peaks are skewed in the left and right direction. This makes sense, as the patterns are more complicated the other ROIs and splits are occurring.

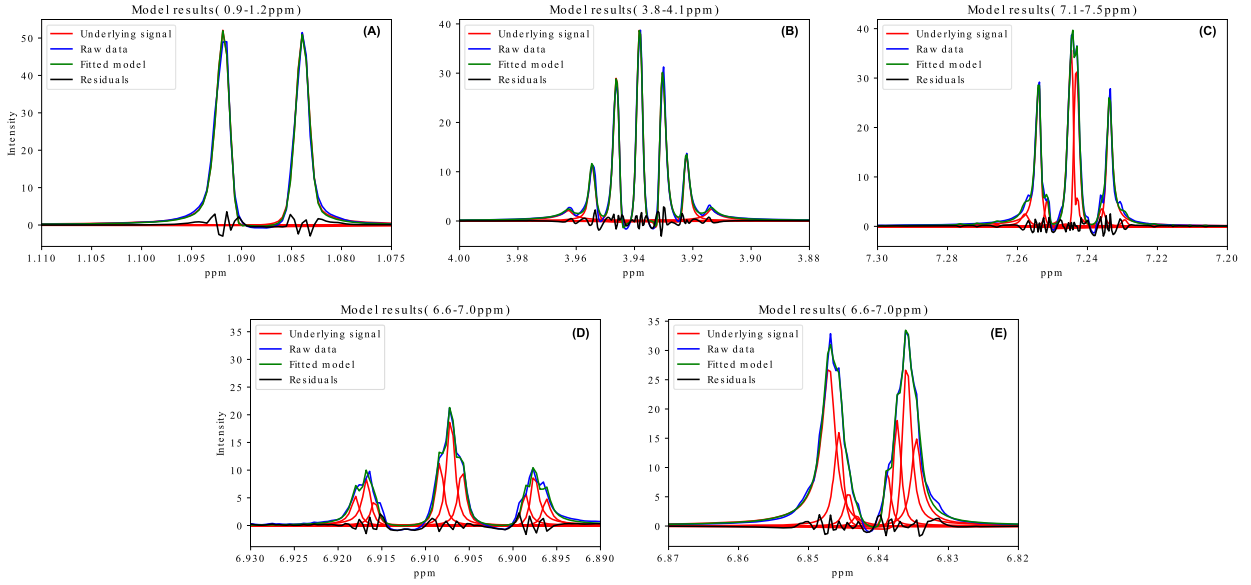

Fig. 5. Visual model deconvolution of sample No. 1 (highest SNR sample - see Table 6 for SNR values). (A) Region 1: Targeted doublet zoom. (B) Region 2: Targeted septet zoom. (C) Region 4: Targeted doublet of doublet zoom. (D) Region 3: Targeted triplet of triplets subpart zoom. (E) Region 3: Targeted board doublet subpart zoom.

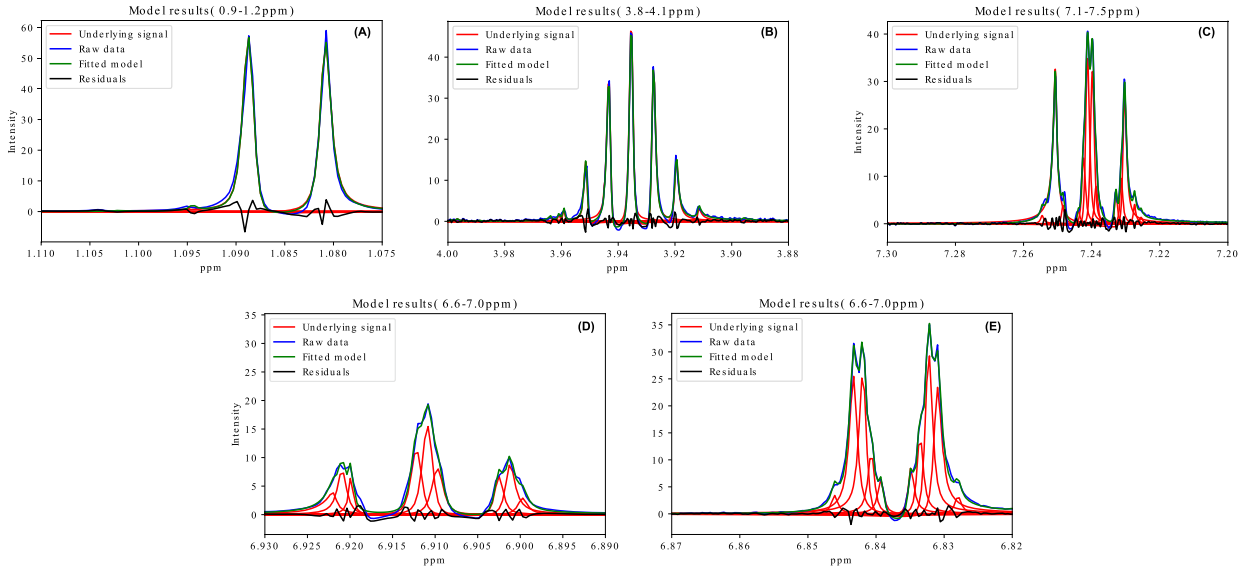

Fig. 6. Visual model deconvolution of sample No. 5 (lowest SNR sample- see Table 6 for SNR values). (A) Region 1: Targeted doublet zoom. (B) Region 2: Targeted septet zoom. (C) Region 4: Targeted doublet of doublet zoom. (D) Region 3: Targeted triplet of triplets subpart zoom. (E) Region 3: Targeted board doublet subpart zoom.

Table 7
Experiment 1 summary of best-selected models across all regions of interests (ROI) for 5 different samples based on BIC.

ROI	model 1	model 2	model 3
1	0	2	3
2	0	1	4
3	1	1	3
4	0	0	5
Total	1	4	15

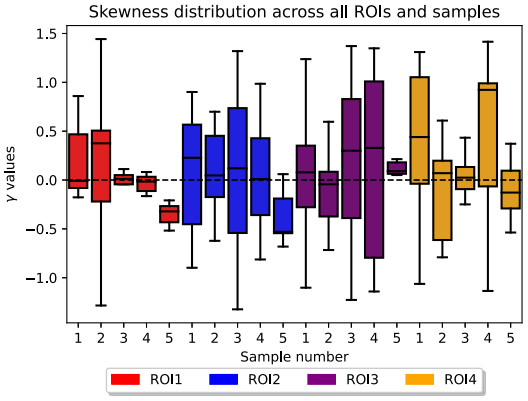

Fig. 7. The figure depicts the distribution of skewness values (γ) across the four regions of interest (ROI) for the 5 samples of case study 1. Values set at 0 (dashed line), would indicate perfect symmetry, while negative and positive values indicate a left and right skewness respectively. The maximum skewness is set at $\pi/2$ and $-\pi/2$.

Table 8
Experiment 1, summary of potential false peaks (PRPs) found across all datasets.

ROI	Total PRPs	Targeted peaks
1	2	0
2	3	0
3	130	66
4	39	9

Table 9
Experiment 2 region of interest and noise region.

Region No.	lower cutoff (ppm)	higher cutoff (ppm)
1	3.35	3.60
Noise region	-0.1	-0.2

A key feature of NMR-Onion lies in the ability to detect PRPs in highly overlapping signals. This feature was utilized to investigate how many of the total peaks detected in each region are less likely to appear in replicates, as they originate from highly overlapping peaks. The results are, for all 5 experiments, summarized in Table 8 including both targeted peaks (see Table 4) and peaks from ^{13}C satellites:

From Table 8, it is observed that the majority of PRPs are found in the third and fourth regions, while the first and second regions contain very few. This observation aligns well with the visual results outlined in Figs. 5 and 6, as peaks are highly overlapping and exhibit second-order effects and the presence of small unresolved J coupling constants. However, it should be noted that many of the PRPs do not originate from the targeted peaks listed in Table 4, but rather from the smaller ^{13}C satellites and some impurities which had CI overlaps in samples 1-3 where detection was possible. This was particularly evident in ROI 3, where second-order effects caused different multiplicity patterns within the signals. In ROI 4, it was revealed that the targeted peaks, where PRPs were identified, occurred only in sample 5 and sample 2 (see more in the discussion section).

Finally, it was noted that across each sample, the consistently detected peaks appeared within the CIs of the first sample (or any other sample's CI), indicating the model's consistent and accurate prediction of peak locations across varying concentrations.

The analysis of the PRPs in this case may not significantly enhance the study, as it would require replicates of the same concentrations to accurately identify specific PRPs arising from sample-to-sample variations. Furthermore, Experiment 1 exhibits highly distinguishable regions, which diminishes the impact of PRP detection. Therefore, to effectively demonstrate the value of the PRP feature, a second case study was designed utilizing a more complex molecule.

4.4. Case study 2

The second case study focuses on analyzing a sample containing the complex phytosteroid Diosgenin molecule. The objective of this experiment is to showcase how NMR-Onion accurately identifies peaks and detects PRPs across two in principle identical samples. We focused on a specific ROI and noise region outlined in Table 9.

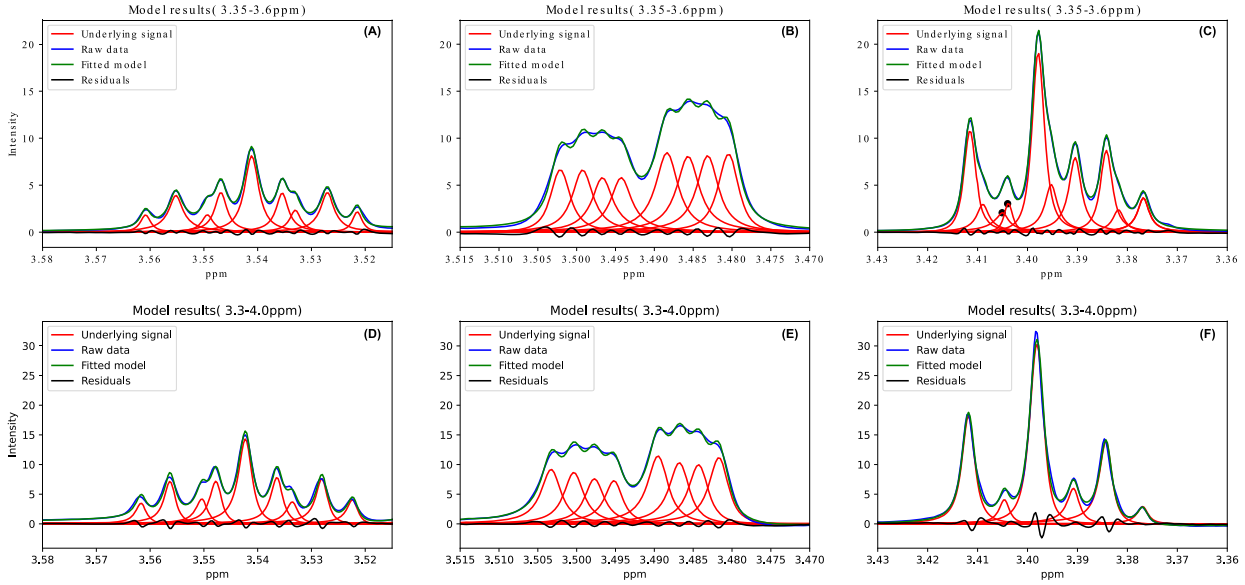

Fig. 8. Visual model deconvolution of two samples within case study 2, where black dots indicate potentially resolved peaks. (A) Sample 1: Sub-region 1 zoom. (B) Sample 1: Sub-region 2 Zoom. (C) Sample 1: Sub-region 3 zoom. (D) Sample 2: Sub-region 1 Zoom. (E) Sample 2: Sub-region 2 zoom. (F) Sample 2: Sub-region 3 zoom.

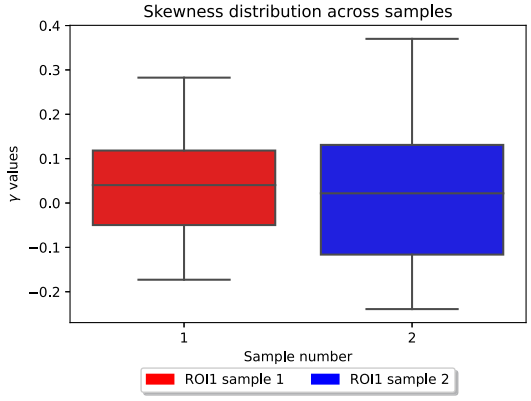

Fig. 9. The figure depicts the distribution of skewness values (γ) across the one region of interest (ROI) for the 2 samples of case study 2. Values set at 0 (dashed line), would indicate perfect symmetry, while negative and positive values indicate a left and right skewness respectively. The maximum skewness is set at $\pi/2$ and $-\pi/2$.

The underlying model of the ROI described in Table 9 was chosen using the same approach outlined in Section 4.3, leading to both datasets being modeled by the exponential power law model (equation (9)). The graphical representation based on the exponential power law in each replicate is depicted in Fig. 8, highlighting three distinct sub-regions.

Two of the sub-regions shown Fig. 8 A, B and D, E, exhibited no obvious difference as the same peaks were detected in both samples. However, the third sub-ROI revealed a possible PRP in the first sample (Fig. 8 C), where the same peaks were absent in the corresponding region of the second sample (Fig. 8 F). This detected PRP suggests that replicates may have a low probability of resolving the same peaks consistently. It is noteworthy that the algorithm does not detect as many peaks in the second sample as in the first sample; however, residual analysis reveals that signals absent in the second sample correspond to signals found in the first sample (See more in the discussion section). Alongside assessing signal uncertainties, the skewness is visualized in Figure In addition to evaluating signal uncertainties, the skewness is also visualized in Fig. 9 for the two datasets of case study 2.

From Fig. 9, it is observed that the two cases have more or less identical skewness, and overall the median skewness is much smaller compared to that of case study 1.

4.5. Comparison with other software

To comprehensively evaluate the results of the NMR-Onion algorithm, we choose to compare it with MNOVA GSD, one of the most popular and widely utilized algorithms in the field. The experimental data from Case Study 1 and Case Study 2 were analyzed using the MNOVA GSD algorithm, and the results are illustrated in Fig. 10 A-E and Fig. 11 A-C. Unfortunately, a direct comparison of metrics such as root mean squared error or BIC/AIC between NMR-Onion and MNOVA GSD are not possible as the internal data

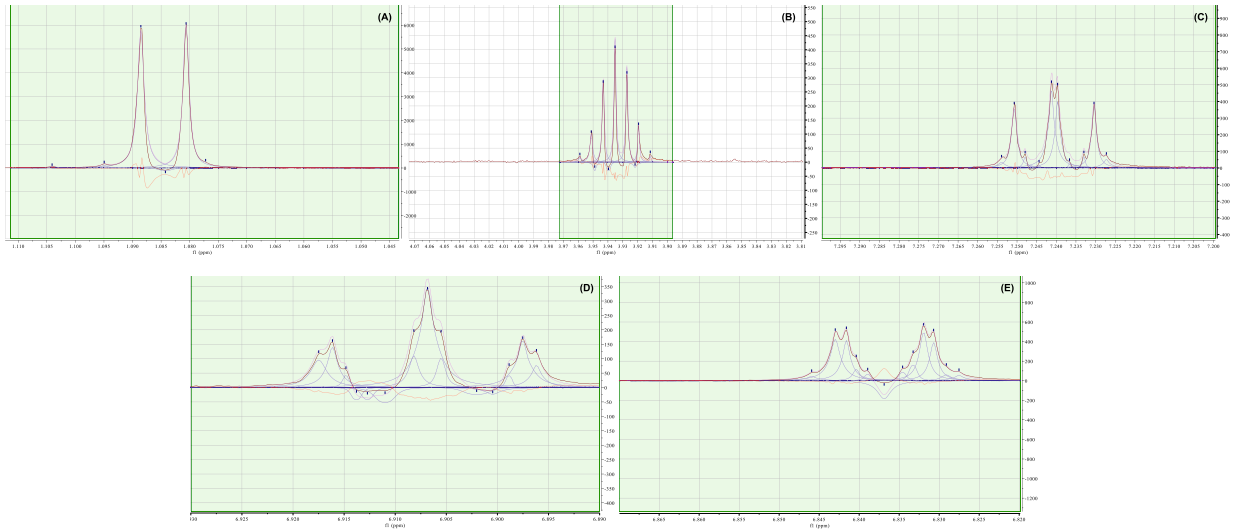

Fig. 10. Visual model deconvolution of sample 5 as detailed in Tables 1 and 6. The dark red lines represent the original spectrum, the purple lines depict the fitted spectrum, the blue line denotes the underlying signals, the orange lines indicate the residuals, and the black ticks mark the detected peaks. (A) Region 1: targeted doublet of zoom. (B) Region 2: Targeted septet zoom. (C) Region 4: Targeted doublet of doublet zoom. (D) Region 3: triplet of triplets zoom. (E) Region 3: Targeted doublet zoom.

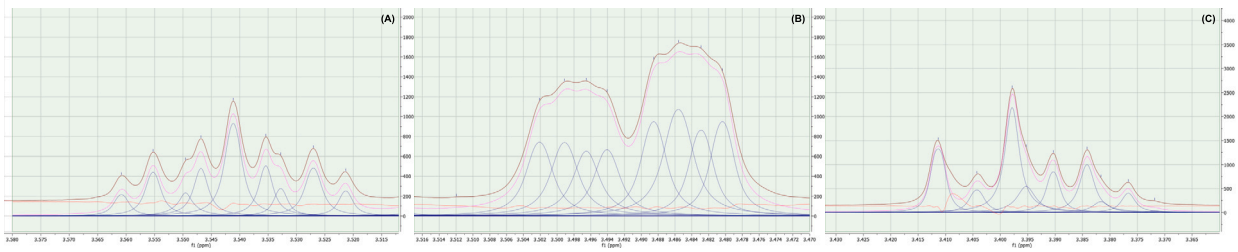

Fig. 11. Visual model deconvolution of sample 1 as detailed in Table 9 and plotted in Fig. 8 A, B and C. The dark red lines represent the original spectrum, the purple lines depict the fitted spectrum, the blue line denotes the underlying signals, the orange lines indicate the residuals, and the black ticks mark the detected peaks. (A) Sub-region 1 zoom. (B) Sub-region 2 zoom. (C) Sub-region 3 zoom.

Table 10
Experiment 1, a comparison of numbers peaks detected by MNOVA and NMR-onion within sample 5.

Region No.	MNOVA	NMR-Onion
1(A)	5	4
2(B)	9	9
3(D)	14	9
3(E)	12	10
4(C)	10	14
Total:	50	46

normalization and loss function formulation of MNOVA GSD cannot be extracted. Therefore, only visual evaluations of residuals are considered here for comparison.

When comparing the MNOVA output shown in Fig. 10 with the results from NMR-Onion depicted in Fig. 6, it is apparent that the residuals produced by both programs do not exhibit a pattern indicative of normally distributed white noise (see more in the discussion section). The summarized results, including the count of detected peaks in each ROI, are presented in Table 10, note that letters A-C corresponding to the sub-plot numbering of Fig. 6 and 10), has been added to each region number.

Upon comparing the number of detected peaks (see Table 10), it is observed that NMR-Onion and MNOVA generally detect a similar number of peaks, though it seems MNOVA are detecting peaks with negative amplitudes as well, as exemplified from Fig. 10 A. Despite many similarities, there are notable differences between NMR-Onion and MNOVA. For instance, in the region around

7.24 ppm, NMR-Onion detects more resolved peaks compared to MNOVA. It is worth noting that these peaks are identified as PRP, suggesting that their detection may be affected by low repeatability due to sample-to-sample variations.

In the second experiment, the results from MNOVA are depicted in Fig. 11. Upon visual inspection, MNOVA and NMR-Onion generally exhibit consistency in their findings for sub-regions 1 and 2. However, a notable difference arises in the third sub-region (Fig. 8 C and 11 C) where NMR-Onion identifies a highly overlapping peak shoulder around 3.4 ppm that MNOVA does not detect. Moreover, the models produced by both MNOVA and NMR-Onion appear to align more closely with the assumptions of white noise compared to the results of the first experiment.

5. Discussion

5.1. Simulation study

From the simulation study, it was evident that the detection algorithm exhibited a minimal over-detection in relation to instances of under-detection and perfect detection across all SNR levels, as outlined in Table 2. Notably, as the SNR lowers, an increase in instances of under-detection occurs. This occurrence may be due to peaks becoming hidden within the noise floor, rendering them obscured from detection. An additional explanation may be that maintaining a constant noise threshold throughout all experimental iterations may lead to the oversight of peaks falling below the noise threshold. The manifestation of over-detection may be attributed to the algorithm operating under fixed noise threshold and peak width filtering, resulting in the erroneous inclusion of random noise spikes or baseline errors as detected peaks. It is imperative to underline that the algorithm is not designed for wholly hands-off operation. Rather, post-detection intervention by a human operator is needed, testing for spurious peaks which may be removed by increasing or decreasing peak width filtering and/or the noise threshold.

As for the spectral reconstruction, lower SNR seemed to correlate with the decrease in the proportion of normally distributed residuals. This pattern may have occurred as the signals are getting harder to distinguish from noise. This aligns with the results of the detection algorithm, which suggested an increase in the proportion of under-fittings as SNR decreased. A possible method for mitigating the need for tuning the parameters would be to incorporate a different type of detection algorithm possibly based on deep learning. Here one could use transfer learning of trained 1D convolutional neural networks for peak detection. However, in order to reduce bias and increase model generality, neural networks trained on real NMR data are required, as simulated data cannot capture every scenario encountered when working with real data.

5.2. Case study 1

From the results in Section 4.5, regarding the first experiment, it was generally observed that MNOVA and NMR-Onion detected nearly identical numbers of peaks across the range of SNRs set up in Table 6. However, in some cases, NMR-Onion detected more peaks than MNOVA, as exemplified in Fig. 6 C. Furthermore, Table 7, depicts that our novel models, given by equations equation (8) and (9) generally outperformed the traditional pure exponential decay model, making them more suitable for fitting non-Lorentzian line shapes. Our decision not to present results in the time domain in this paper is primarily based on two reasons. Firstly, NMR spectroscopy is conventionally analyzed in the frequency domain, aligning our approach with standard practice in the field. Secondly, our reliance on visual comparisons between software outputs posed challenges in the time domain due to the more convoluted nature of fits. Still, detailed time domain outputs can be generated if required, and are provided in the NMR-Onion tutorial available on the GitHub site (see supporting information). As for the lineshape skewness, the study revealed that peaks are skewed both in the right and left direction. We saw that commonly, the skewness is highest in more convoluted regions compared to regions containing fewer signals.

Another noteworthy finding is presented in Table 8 where ROI 3 exhibited numerous instances of confidence interval overlaps suggesting that the peaks found around the targeted resonances may be potentially resolved peaks and therefore should be further investigated for consistency within independent replicates. However, as demonstrated in the results, while the targeted peaks were consistently present across all samples, they exhibited varying underlying multiplet structures. We believe the variability is attributed to the occurrence of second-order effects, particularly notable in ROI 3. Furthermore, ROI 4 also exhibited a notable abundance of PRPs, akin to the observation in ROI 3. Detailed identification and characterization of these PRPs were not extensively pursued in this study. However, this could be addressed through the addition of more replicates at identical concentrations. Such an approach would enable differentiation between peaks that consistently appear but are highly overlapping, and those that arise due to sample-to-sample variations.

Finally, it was noted that in the first case study the model residuals from both MNOVA and NMR-Onion are heavily deviating form the assumptions of exhibiting white Gaussian noise. We believe this is to be attributed to the imperfections in the data regarding model formulations, which manifest in both MNOVA and NMR-Onion due to non-flat baselines, alongside minimal preprocessing. The rationale behind these imperfections was to assess how our approach could manage complex data with only minimal corrections. Extensive corrections are heavily reliant on manual operator correction rather than automation, potentially introducing a bias in the analysis. Interestingly, the same minimal preprocessing scheme was applied in the second experiment, yet the residuals here align much more closely with the model assumptions, likely owing to a higher SNR.

5.3. Case study 2

In the second experiment, as detailed in sections 4.5 and 4.4, a potential PRP was identified within the first sample (Fig. 8 C), whereas the overlap was absent in the replicate (Fig. 8 F). This suggests that this particular peak may have arisen due to sample variations. As for the skewness distribution, we noted that case study 2 had much less asymmetry compared to that of case study 1. However, we did note that the median skewness was not 0 and had cases that were far from perfect symmetry. Additionally, while the results from MNOVA and NMR-Onion were largely similar, there was a notable difference in the detection of peaks within the last sub-region (3.36-3.43 ppm). Specifically, NMR-Onion detected a peak shoulder around 3.4 ppm that MNOVA did not detect (see Fig. 11 C vs Fig. 8 C). However, we cannot state if NMR-Onion archives more accurate detection in general, as the comparisons are only based on a few spectra. Finally, it should be noticed that Fig. 8 F shows fewer detected peaks compared to Fig. 8 C, as evident from the residuals which indicate missing peaks in the former. Adjusting the peak width cutoff does enable the detection of the missing peaks, but we chose not to change this parameter to maintain consistency across experiments when comparing outputs.

The residuals in case study 2 are notably closer to meeting the model assumptions compared to case study 1. However, it's important to note that neither MNOVA nor NMR-Onion achieves perfect white noise in this study. This observation is rational given the nature of real-world data, where it is impractical to construct a flawless model that accommodates every form of distortion without risking significant overfitting. However, it seems that the model produces some small systematic error in the residuals. The reason for imperfections might be that the loss function is non-convex and a better optimum might be identified should the algorithm run longer (see our GitHub and supplementary, where we ran the model for 40 epochs, reducing the residual error). It should be noted that the error is very small which is evident from the time domain residuals found in the supplementary. Nevertheless, we have developed a model and framework capable of accurately representing a spectrum, achieving minimal residual signals and avoiding significant instances of missed peaks. A potential improvement could involve incorporating a random effect into the model, thereby creating a non-linear mixed-effects model that could potentially capture random distortions. To the best of our knowledge, this approach has not been previously explored and could effectively account for stochastic variations among samples.

5.4. The NMR-Onion algorithm

An essential aspect of the NMR-Onion algorithm lies in its capability to identify potentially resolved peaks through overlapping confidence intervals, which are determined using the wild bootstrap method (see Algorithm 1). The drawback of this approach stems from the significant computational time required, given that the model needs to be refitted 1000 times (default value) or more. This challenge was addressed by decimating the time series signal [51], leveraging the fact that initial parameter values from the fit of the non-decimated ROI were estimated prior to executing the bootstrap. Alternative methods to the wild bootstrap algorithm (see Algorithm 1) do exist, such as those based on Bayesian approaches. However, as highlighted by Wilson [3] the sampling schemes in pure MCMC approaches are often considerably slower than almost any other optimization method. Therefore, one might consider a variational Bayesian (VB) inference sampling scheme [52] as a possible alternative to the wild bootstrap and model fitting, Still, a challenge with Bayesian formulations is the necessity to specify appropriate priors for the parameters to ensure effective sampling. The Zellner prior has demonstrated effectiveness in fitting sinusoids, as observed in the works of Rubtsov and Griffith [29]. Utilizing variational Bayesian (VB) inference with the Zellner prior could potentially enhance computational efficiency of a future version of NMR-Onion. Another novel aspect of the NMR-Onion algorithm involves implementing models and optimization routines using the modern framework of PyTorch. One of the primary advantages of using PyTorch is its automatic differentiation (AD) capabilities. When defining a loss function such as equation (15), PyTorch can automatically compute the gradient and Hessian, thereby optimizing the optimization process, making it faster and more robust. Therefore, automatic differentiation (AD) enables the development of robust models much more easily, as it eliminates the need for manual implementation of derivatives. We believe that in conjunction with the peak detection and digital filter modules, other models can be readily implemented and tested using the PyTorch core optimization framework of NMR-Onion. This makes the development of both time and frequency domain models more accessible for all developers. We experimented with other non-Quasi-Newton optimization approaches such as ADAM [53] and RSM-prop [54] algorithms, but these proved less effective (results not included) compared to LBFGS implementations in both PyTorch [38] and Scipy [39].

In making the optimization routine using LBFGS, computationally feasible, other methods were also explored as alternatives to the digital band-pass filter. We attempted mini-batch stochastic optimization, which is a technique in deep learning capable of handling much larger data sets. However, these attempts did not yield promising results on either real or simulated data. We believe that the outcome can be attributed largely to the LBFGS algorithm in PyTorch not being capable of handling a mini-batch approach rather than the efficacy of the alternative method itself. Regarding the simulated data, the non-quasi-Newton approaches (Adam and RSM-prop) performed well with mini-batches; however, they failed when applied to real data. Therefore, an attractive improvement to our algorithm is to implement mini batches when the LBFGS algorithm of Pytorch is further developed to include stochastic optimization, as this would render manual ROI selection obsolete, instead fitting the full spectrum all at once. We did improve the speed of the algorithm by utilizing the GPU framework made possible by the Pytorch backbone (see section 4.2), this can in part remove the need for the digital filter as suggested by the speed comparisons of CPU vs GPU runs. It is also possible to combine the digital filter and GPU for further speed increase should only a small area of signals be of interest. The downside of relying solely on the GPU is that this will create a barrier to entry for some users due to the cost of GPU's. In the current state of NMR-Onion, multiplets must be manually assigned based on estimated amplitude ratios and coupling constants. This manual process can be challenging, particularly in untargeted studies. Therefore, for future developments of NMR-Onion, implementing automatic assignment of multiplets based

on amplitude ratios and coupling constants is a desired feature. This enhancement would make NMR-Onion suitable for expedited analysis in both targeted and untargeted studies.

6. Conclusion

From the results of this study, we conclude that the NMR-Onion framework offers a robust approach for analyzing 1D ^{1}H NMR spectra. This framework effectively targets specific regions of interest (ROIs) within a spectrum, enabling targeted analysis across a wide range of signal-to-noise ratio (SNR) values. Additionally, we conclude that our novel time domain models are capable of fitting highly overlapping signals, outperforming the traditional exponential decay model formulation. We believe that, with the NMR-Onion framework being open-source, model improvements and further development can be rapidly integrated due to the AD library. The core modules of digital filtering ensure computational feasibility, and the peak detection algorithm effectively handles the multi-modality of the frequencies. Furthermore, we predict that the addition of detecting potentially resolved peaks generated from the NMR-Onion framework, in combination with replicates, will significantly reduce the risk of false conclusions. This would be particularly relevant for large metabolomics samples, where numerous signal overlaps are present, and sample-to-sample variations could potentially play a significant role. With the NMR-Onion algorithm, users are made aware of potential artifacts, minimizing the likelihood of drawing false conclusions based on peaks marked as potentially resolved peaks. With NMR-Onion, we have developed an algorithm capable of statistically evaluating uncertainties in the results, ensuring that users are alerted to potentially resolved peaks. When combined with replicates, this feature helps confirm that highly overlapping peaks are consistent across samples and not a result of sample-to-sample variations.

CRediT authorship contribution statement

Mathies Brinks Sørensen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael Riis Andersen:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Mette-Maya Siewertsen:** Investigation. **Rasmus Bro:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Mikael Lenz Strube:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Charlotte Held Gotfredsen:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data utilized for the article can be found on our Github: https://www.github.com/Mabso1/NMR-onion.

Acknowledgements

We would like to thank the Danish National Research Foundation for the Center for Microbial Secondary Metabolites Grant Number: DNRF137 for funding this project and also thank the NMR Center • DTU and the Villum Foundation are acknowledged for access to the 800 MHz spectrometers.

Appendix A. Supplementary material

Github: https://www.github.com/Mabso1/NMR-onion. Additional information can be found in the online version of the article. Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36998.

References

[1] A.-H. Emwas, E. Saccenti, X. Gao, R.T. McKay, V.A.P.M. dos Santos, R. Roy, D.S. Wishart, Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine, Metabolomics 14 (3) (2018) 31, https://doi.org/10.1007/S11306-018-1321-4.

[2] G.L. Bretthorst, C.-C. Hung, D.A. D'Avignon, J.J.H. Ackerman, Bayesian analysis of time-domain magnetic resonance signals, J. Magn. Reson. (1969) 79 (2) (1988) 369–376, https://doi.org/10.1016/0022-2364(88)90233-8.

[3] A.G. Wilson, Y. Wu, D.J. Holland, S. Nowozin, M.D. Mantle, L.F. Gladden, A. Blake, Bayesian Inference for NMR Spectroscopy with Applications to Chemical Quantification, arXiv:1402.3580, Applications, feb 2014, https://doi.org/10.48550/arXiv.1402.3580.

[4] R.A. Davis, A.J. Charlton, J. Godward, S.A. Jones, M. Harrison, J.C. Wilson, Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform, Chemom. Intell. Lab. Syst. 85 (1) (2007) 144–154, https://doi.org/10.1016/j.chemolab.2006.08.014.

[5] S.A.A. Sousa, A. Magalhães, M.M.C. Ferreira, Optimized bucketing for nmr spectra: three case studies, Chemom. Intell. Lab. Syst. 122 (2013) 93–102, https://doi.org/10.1016/j.chemolab.2013.01.006.

[6] T. De Meyer, D. Sinnaeve, B. Van Gasse, E. Tsiporkova, E.R. Rietzschel, M.L. De Buyzere, T.C. Gillebert, S. Bekaert, J.C. Martins, W. Van Criekinge, NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm, Anal. Chem. 80 (10) (2008) 3783–3790, https://doi.org/10.1021/AC7025964.

[7] C. Piras, M. Pibiri, V.P. Leoni, F. Cabras, A. Restivo, J.L. Griffin, V. Fanos, M. Mussap, L. Zorcolo, L. Atzori, Urinary 1H-NMR metabolic signature in subjects undergoing colonoscopy for colon cancer diagnosis, Appl. Sci. (Switzerland) 10 (16) (2020) 5401, https://doi.org/10.3390/APP10165401.
[8] F. Probert, V. Ruiz-Rodado, D. Te Vruchte, E.R. Nicoli, T.D.W. Claridge, C.A. Wassif, N. Farhat, F.D. Porter, F.M. Platt, M. Grootveld, NMR analysis reveals significant differences in the plasma metabolic profiles of Niemann pick C1 patients, heterozygous carriers, and healthy controls, Sci. Rep. 7 (1) (2017) 6320, https://doi.org/10.1038/s41598-017-06264-2.
[9] F.M.M. Ocampos, L.R.A. Menezes, L.M. Dutra, M.F.C. Santos, S. Ali, A. Barison, NMR in chemical ecology: an overview highlighting the main NMR approaches, eMagRes 6 (2) (2017) 325–342, https://doi.org/10.1002/9780470034590.EMRSTM1536.
[10] R.X. Poulin, G. Pohnert, Simplifying the complex: metabolomics approaches in chemical ecology, Anal. Bioanal. Chem. 411 (1) (2019) 13–19, https://doi.org/10.1007/S00216-018-1470-3.
[11] U.K. Sundekilde, N. Eggers, H.C. Bertram, Nmr-based metabolomics of food, in: NMR-Based Metabolomics: Methods and Protocols, 2019, pp. 335–344.
[12] M. Cuperlovic-Culf, A.S. Culf, Applied metabolomics in drug discovery, Expert Opin. Drug Discov. 11 (8) (2016) 759–770, https://doi.org/10.1080/17460441.2016.1195365.
[13] K. Krishnamurthy, CRAFT (complete reduction to amplitude frequency table) – robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR, Magn. Reson. Chem. 51 (12) (2013) 821–829, https://doi.org/10.1002/MRC.4022.
[14] J. Keeler, Understanding NMR Spectroscopy, John Wiley & Sons, Ltd, 2010, p. 526.
[15] I. Marshall, J. Higinbotham, S. Bruce, A. Freise, Use of Voigt lineshape for quantification of *in vivo* 1H spectra, Magn. Reson. Med. 37 (5) (1997) 651–657, https://doi.org/10.1002/MRM.1910370504.
[16] M. Niklasson, R. Otten, A. Ahlner, C. Andresen, J. Schlagnitweit, K. Petzold, P. Lundström, Comprehensive analysis of NMR data using advanced line shape fitting, J. Biomol. NMR 69 (2) (2017) 93–99, https://doi.org/10.1007/s10858-017-0141-6.
[17] Mestrelab, Global Spectral Deconvolution (GSD) - mestrelab resources, https://resources.mestrelab.com/gsd/, 2017.
[18] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka, P. Giraudeau, Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain, J. Magn. Reson. 298 (2019) 91–100, https://doi.org/10.1016/J.JMR.2018.11.004.
[19] J. Hao, M. Liebeke, W. Astle, M. De Iorio, J. Bundy, T. M, D. Ebbels, Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, Nat. Protoc. 9 (6) (2014) 1416–1427, https://doi.org/10.1038/nprot.2014.090.
[20] S. Ravanbakhsh, P. Liu, T.C. Bjorndahl, R. Mandal, J.R. Grant, M. Wilson, R. Eisner, I. Sinelnikov, X. Hu, C. Luchinat, R. Greiner, D.S. Wishart, Accurate, fully-automated NMR spectral profiling for metabolomics, PLoS ONE 10 (5) (2015) e0124219, https://doi.org/10.1371/journal.pone.0124219.
[21] P. Mercier, M.J. Lewis, D. Chang, D. Baker, D.S. Wishart, Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra, J. Biomol. NMR 49 (3–4) (2011) 307–323, https://doi.org/10.1007/S10858-011-9480-X.
[22] D.-W. Li, A. Hansen, C. Yuan, L. Bruschweiler-Li, R. Brüschweiler, Deep picker is a deep neural network for accurate deconvolution of complex two-dimensional nmr spectra, Nat. Commun. 12 (2021) 5229, https://doi.org/10.1038/s41467-021-25496-5.
[23] D.-W. Li, L. Bruschweiler-Li, A.L. Hansen, R. Brüschweiler, DEEP Picker1D and Voigt Fitter1D: a versatile tool set for the automated quantitative spectral deconvolution of complex 1D-NMR spectra, Magn. Reson. 4 (1) (2023) 19–26, https://doi.org/10.5194/mr-4-19-2023.
[24] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.
[25] Y. Hua, T.K. Sarkar, Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise, IEEE Trans. Acoust. Speech Signal Process. 38 (5) (1990) 814–824.
[26] G.R. de Prony, Essai experimental et analytique: sur les lois de la dilatabilite des fluides elastique et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, a differentes temperatures, J. Éc. Polytech. (1795).
[27] Y. Wu, O. Sanati, M. Uchimiya, K. Krishnamurthy, J. Wedell, J.C. Hoch, A.S. Edison, F. Delaglio, SAND: automated time-domain modeling of NMR spectra applied to metabolite quantification, Anal. Chem. 96 (5) (2024) 1843–1851, https://doi.org/10.1021/acs.analchem.3c03078.
[28] Y. Matviychuk, E. Steimers, E. von Harbou, D.J. Holland, Improving the accuracy of model-based quantitative nuclear magnetic resonance, Magn. Reson. 1 (2) (2020) 141–153, https://doi.org/10.5194/mr-1-141-2020.
[29] D.V. Rubtsov, J.L. Griffin, Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy, J. Magn. Reson. 188 (2) (2007) 367–379, https://doi.org/10.1016/J.JMR.2007.08.008.
[30] N. Narisetty, Bayesian model selection for high-dimensional data, in: A.S.S. Rao, C. Rao (Eds.), Handbook of Statistics, vol. 43, Elsevier, 2020, pp. 207–248, https://doi.org/10.1016/bs.host.2019.08.001.
[31] E. Mammen, Bootstrap and wild bootstrap for high dimensional linear models, Ann. Stat. 21 (1) (1993) 255–285, https://doi.org/10.1214/aos/1176349025.
[32] Y. Matviychuk, E. von Harbou, D.J. Holland, An experimental validation of a Bayesian model for quantification in NMR spectroscopy, J. Magn. Reson. 285 (2017) 86–100, https://doi.org/10.1016/J.JMR.2017.10.009.
[33] J.-B. Poullet, D.M. Sima, S. Van Huffel, Mrs signal quantitation: a review of time- and frequency-domain methods, J. Magn. Res. 195 (2) (2008) 134–144, https://doi.org/10.1016/j.jmr.2008.09.005.
[34] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 1985, 662 pp., https://doi.org/10.1017/CBO9780511810817.
[35] E.-H. Djermoune, M. Tomczak, D. Brie, NMR data analysis: a time-domain parametric approach using adaptive subband decomposition, Oil Gas Sci. Technol. - Revue d'IFP Energies Nouvelles 69 (2) (2014) 229–244, https://doi.org/10.2516/OGST/2012092.
[36] S.G. Hulse, M. Foroozandeh, Newton meets Ockham: parameter estimation and model selection of NMR data with NMR-EsPy, J. Magn. Reson. 338 (2022) 107173, https://doi.org/10.1016/J.JMR.2022.107173.
[37] S.-J. Baek, A. Park, Y.J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (1) (2014) 250–257, https://doi.org/10.1039/C4AN01061B.
[38] A. Paszke, LBFGS — PyTorch 1.12 documentation, https://pytorch.org/docs/stable/generated/torch.optim.LBFGS, 2022.
[39] D.M. Cooke, minimize(method = 'L-BFGS-B') — SciPy v1.9.0 Manual, 2004.
[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: NIPS 2017 Workshop on Autodiff, 2017.
[41] R. Hooke, T.A. Jeeves, "Direct search" solution of numerical and statistical problems, J. ACM 8 (2) (1961) 212–229, https://doi.org/10.1145/321062.321069.
[42] M.U.A. Bromba, H. Ziegler, Application hints for Savitzky-Golay digital smoothing filters, Anal. Chem. 53 (11) (1981) 1583–1586, https://doi.org/10.1021/ac00234a011.
[43] R. MacDonald, S. Sokolenko, Detection of highly overlapping peaks via adaptive apodization, J. Magn. Reson. (Calif. 1997) 333 (2021) 107104, https://doi.org/10.1016/J.JMR.2021.107104.
[44] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723, https://doi.org/10.1109/TAC.1974.1100705.
[45] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464, https://doi.org/10.1214/AOS/1176344136.
[46] B. Nadler, L.A. Kontorovich, Model selection for sinusoids in noise: statistical analysis and a new penalty term, IEEE Trans. Signal Process. 59 (4) (2011) 1333–1345, https://doi.org/10.1109/TSP.2011.2105482.
[47] P.M. Djurić, A model selection rule for sinusoids in white Gaussian noise, IEEE Trans. Signal Process. 44 (7) (1996) 1744–1751, https://doi.org/10.1109/78.510621.
[48] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Am. Math. Soc. 54 (3) (1943) 426, https://doi.org/10.2307/1990256.

[49] TopSpin | NMR Data Analysis | Bruker, https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html, 2023.
[50] J.J. Helmus, C.P. Jaroniec, Nmrglue: an open source python package for the analysis of multidimensional NMR data, J. Biomol. NMR 55 (4) (2013) 355–367, https://doi.org/10.1007/S10858-013-9718-X.
[51] R. Crochiere, L. Rabiner, Interpolation and decimation of digital signals - a tutorial review, Proc. IEEE 69 (3) (1981) 300–331, https://doi.org/10.1109/PROC.1981.11969.
[52] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, J. Am. Stat. Assoc. 112 (518) (2016) 859–877, https://doi.org/10.1080/01621459.2017.1285773.
[53] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, 3rd international conference on learning representations, in: ICLR 2015 - Conference Track Proceedings, dec 2014, https://doi.org/10.48550/arxiv.1412.6980.
[54] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude, COURSERA: Neural Netw. Mach. Learn. 4 (2) (2012) 26–31.