molecules

MDPI

Article

Unsupervised Analysis of Small Molecule Mixtures by Wavelet-Based Super-Resolved NMR

Aritro Sinha Roy [1] and Madhur Srivastava [1,2,*]

[1] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14850, USA
[2] National Biomedical Center for Advanced ESR Technology, Cornell University, Ithaca, NY 14850, USA
* Correspondence: ms2736@cornell.edu

Abstract: Resolving small molecule mixtures by nuclear magnetic resonance (NMR) spectroscopy has been of great interest for a long time for its precision, reproducibility, and efficiency. However, spectral analyses for such mixtures are often highly challenging due to overlapping resonance lines and limited chemical shift windows. The existing experimental and theoretical methods to produce shift NMR spectra in dealing with the problem have limited applicability owing to sensitivity issues, inconsistency, and/or the requirement of prior knowledge. Recently, we resolved the problem by decoupling multiplet structures in NMR spectra by the wavelet packet transform (WPT) technique. In this work, we developed a scheme for deploying the method in generating highly resolved WPT NMR spectra and predicting the composition of the corresponding molecular mixtures from their ^{1}H NMR spectra in an automated fashion. The four-step spectral analysis scheme consists of calculating the WPT spectrum, peak matching with a WPT shift NMR library, followed by two optimization steps in producing the predicted molecular composition of a mixture. The robustness of the method was tested on an augmented dataset of 1000 molecular mixtures, each containing 3 to 7 molecules. The method successfully predicted the constituent molecules with a median true positive rate of 1.0 against the varying compositions, while a median false positive rate of 0.04 was obtained. The approach can be scaled easily for much larger datasets.

Keywords: NMR; shift spectra; wavelet packet transform; automated small molecule mixture analysis

check for updates

Citation: Sinha Roy, A.; Srivastava, M. Unsupervised Analysis of Small Molecule Mixtures by Wavelet-Based Super-Resolved NMR. *Molecules* **2023**, *28*, 792. https://doi.org/10.3390/molecules28020792

Academic Editor: Carmelo Corsaro

Received: 7 December 2022
Revised: 27 December 2022
Accepted: 3 January 2023
Published: 13 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Identification of the components of small molecule mixtures is a crucial and challenging step in the research and development activities in the pharmaceutical drug discovery [1–3], metabolomics [4–6], natural product synthesis [7–9], food quality control [10,11], and environmental sciences [12,13]. Different types of nuclear magnetic resonance (NMR) spectroscopic methods, high-performance liquid chromatography (HPLC), and mass spectrometry (MS) are widely used across the associated industries for this purpose. The main advantages of NMR over the other techniques are that (1) its results are highly reproducible, (2) it requires very little sample preparation effort, and (3) it is a nondestructive method [14–16]. However, its relatively poor resolution and sensitivity often make NMR an essential, but non-exhaustive analytic tool [5,9]. While recent developments for sensitivity improvement have largely been successful [17–19], limited progress has been made towards achieving the desired resolution. This is primarily due to the limited range of chemical shift windows (~10 ppm) and overlapping resonance lines of small molecules. It is possible to enhance the resolution in homonuclear decoupled ^{1}H NMR spectroscopy by producing pure shift spectra [20–24]. The technique has failed to gain wide applicability owing to its experimental complexity and poor sensitivity [20,25]. While multi-dimensional NMR can improve the resolution by revealing some of the overlapped components, it comes at the cost of a high signal acquisition time and experimental noise, which makes it unsuitable for automated high-throughput studies. To overcome the limited chemical shift window of 1D ^{1}H NMR,

pseudo-2D NMR methods are employed using diffusion coefficients, relaxation parameters, and other suitable molecular parameters for spectral simplification and differentiation in NMR [26–29]. However, the accuracy of the extraction of such molecular parameters and, hence, the efficiency of separating spectral components for a mixture depends heavily on the molecular size distribution, the extent of spectral overlapping, and the magnetic properties of the molecules in a mixture [29]. Therefore, these methods are often complementary to each other and cannot be applied without user interference or prior knowledge about the mixture or data pre-processing, which adds further limitations to the applicability of the methods [26,30]. In order to resolve the problem theoretically, the maximum entropy method has been used in converting NMR to pure shift spectra by deconvolution [31,32]. One of its major drawbacks is the requirement of prior knowledge about the scalar coupling patterns and the coupling constants, which is reasonable for ^{13}C NMR, but unsuitable for ^{1}H NMR spectroscopy. Apart from this, a series of spectral analysis tools has been developed, which include peak matching strategies [33–35], spectral editing [36,37], similarity measure [38,39], and deep-learning-based tools [40–42], for identifying small molecule mixture constituents from the corresponding NMR spectra. However, those applications can be seldom generalized, often suffer from low reliability, and/or require extremely large and specifically designed training datasets. As a result, none of the methods are suitable for high-throughput analysis of small molecule mixtures using ^{1}H NMR as the primary tool.

In a recent work, we showed that the wavelet packet transform (WPT) can work as a multi-resolution signal processing tool in transforming an ^{1}H NMR to a pure shift spectrum [43]. Successive decomposition of a spectrum by WPT yields pairs of approximation and detail components at each level, which contain some of the low- and high-frequency spectral features from the chemical shift domain, respectively. The approximation component produced at the final level of decomposition of an ^{1}H NMR spectrum produces only singlet structures, while the multiplet structures are transferred to the various detail components. We illustrate that the former can be used to calculate a simple pure shift spectrum, and the robustness of the WPT-based NMR spectral analysis method against a significant level of noise has been established [43]. An overview of the wavelet transform theory is provided in the Appendix A.

Automating the task of molecular identification in the study of metabolites and other small molecule mixtures without a priori knowledge remains a major challenge, especially using 1D NMR as the primary analytical tool [44–46]. In absence of an automated mixture analysis, the study time increases significantly, while the accuracy of the analysis varies widely based on the user inputs and interpretations, as well as the nature of the molecular mixtures [44,47,48]. In this work, we developed an automated method for predicting molecular compositions from the corresponding 1D ^{1}H NMR spectra without a priori knowledge and demonstrate its applicability across a wide range of molecules. The problem of the automated identification of mixture components from the corresponding 1D ^{1}H NMR spectra can be divided into two parts: (1) predicting the number of molecules in a mixture and (2) predicting their identities. For this purpose, we created an extensive database of 1000 augmented NMR spectra of molecular mixtures, each containing 3 to 7 spectra of the constituent molecules. A library of WPT shift NMR was created from the 500 MHz NMR spectra of 74 molecules. The mixed NMR spectra were analyzed in an automated fashion by implementing a four-step algorithm. The algorithm in its first two steps calculates a WPT shift spectrum from an NMR spectrum and obtains a potential molecular composition by matching the peaks in the shift spectrum with those of the spectral library. Next, the composition is optimized by employing a gradient descent method to minimize the mean-squared error in predicting the WPT shift spectrum of the mixture. The top 15 entries from the potential composition are forwarded to the next step, where another gradient-descent-based minimization in predicting the WPT spectrum of the mixture produces the final list of molecules.

In analyzing the performance of the method, we used the true positive and false positive rates, which represent the number of accurate and false predictions with respect to the actual compositions of a molecular mixture, respectively. After the first optimization step, we ob-

tained an average true positive rate of 1.0, while the average false positive rate was very high (0.3). This elimination step removed the molecules (choices) with zero or very low probability to be present in the composition from the probable list. Among the remaining choices, the top entries by their calculated probability of existence describe the true compositions for all the cases in the augmented dataset. In fact, we observed that, for mixtures with 3 to 4 molecules, a true positive rate of 1.0 could have been obtained considering only the top 6 to 8 entries, respectively. Therefore, selecting the optimal number of entries from the potential list of molecules without a priori information requires a second optimization. In this identification step, the top 15 entries obtained at the end of the elimination step were optimized by another gradient descent method, which resulted in a median true positive rate of 1.0, while reducing the false positive rate to 0.04 for the analysis.

2. Results and Discussion

As a benchmark, we analyzed the dataset of mixed spectra by matching them with the pure NMR spectra of the individual molecules, which is the most commonly used strategy at present [33,34]. From the summary of the analysis shown in Figure 1, it can be seen that the average true positive rate is ∼0.7 for the entire dataset, as well as for the mixtures with different numbers of constituent molecules in them. Both subplots in the figure show large variations in the true positive rate, which demonstrates the high uncertainty involved in the analysis. The false positive rate for all the cases was equal to 0. It should also be noted that this kind of direct matching may not be feasible for much larger libraries than the one with 74 molecules used in this work.

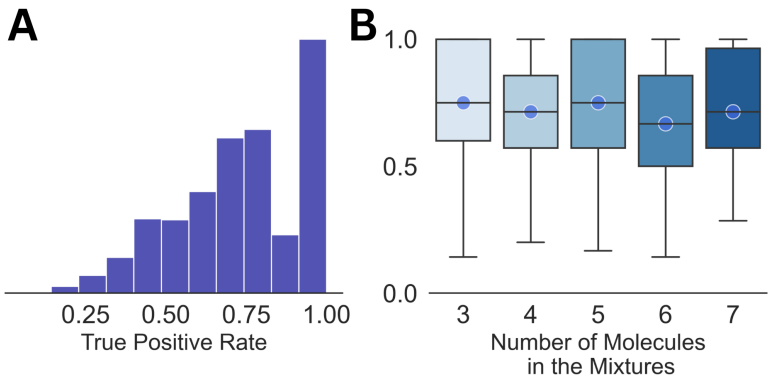

Figure 1. The distribution of the true positive rates for the entire dataset (**A**) and against the size of the mixture (**B**) is shown. The circles in (**B**) emphasize the median for each of the distributions.

The results obtained in Step (3) of our scheme (Figure 2) are summarized in Figure 3. At this stage of our analysis, a median true positive rate of 1.0 was obtained for the entire spectral dataset. This observation demonstrates the robustness of the WPT shift representation of a NMR spectrum and its ability to enhance the resolution [43]. However, the impressive true positive rate was associated with a very high false positive rate across all the cases, with a median value of 0.3. Both the average false positive rate and its variations increased with the size of the mixtures or the increasing complexity of the corresponding spectra.

Looking at the individual analyses and the corresponding mixture compositions, we noticed that, while ∼30 molecules were present in each prediction on average, leaving a few outliers, the top 6 to 15 entries by their probability of existence contained all the components of the mixtures. Hence, in the final step of our spectral analysis scheme, we employed another optimization, which used the top 15 entries from a predicted molecular composition in Step (3). This step resulted in a massive reduction in the false positive rate from 0.3 in Step (3) to 0.04, shown in Figure 4, while the true positive rate remained mostly unaffected, except for the case with molecular mixtures comprising seven molecules. Even for the latter case, we obtained a median true positive rate of 1.0 with a standard deviation of 0.08.

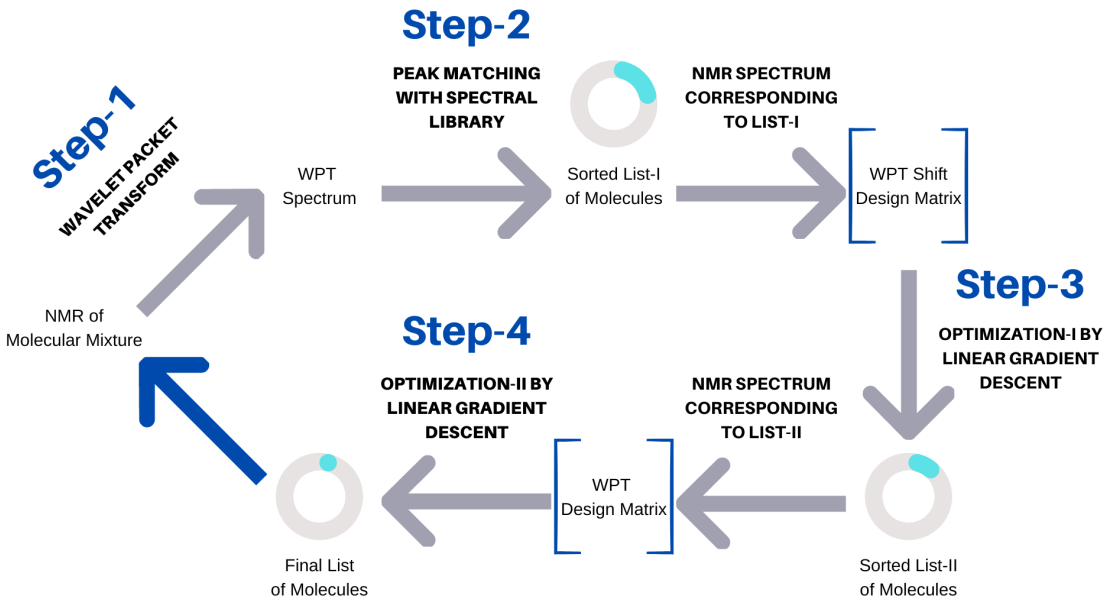

Figure 2. Schematic representation of the spectral analysis algorithm.

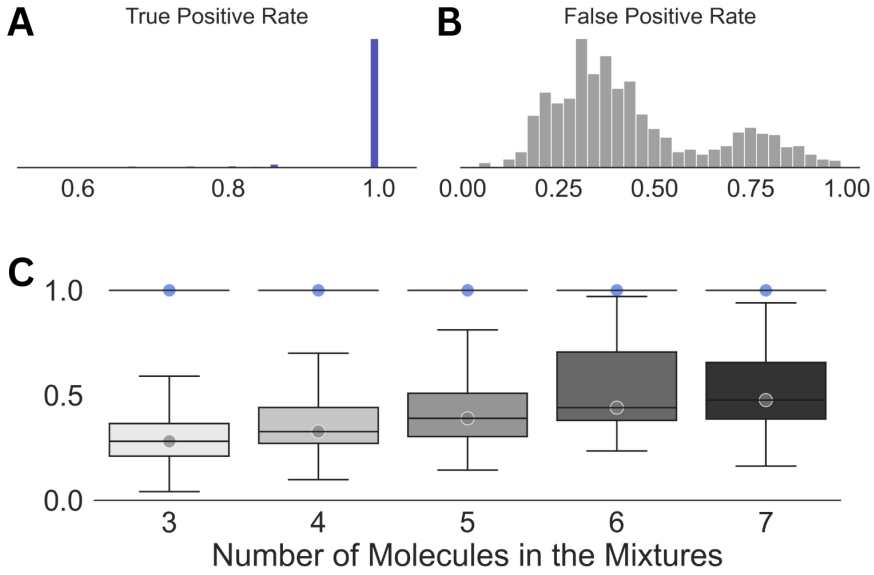

Figure 3. Summary of the results obtained in Step (3) of the analysis. Distributions of the true positive (**A**) and the false positive rates (**B**) for the entire dataset along with those against the size of the mixtures (**C**) are shown. The circles in (**C**) emphasize the median true positive rate (blue) and false positive rate (gray) for each of the distributions. For all the cases, a true positive rate of 1.0 was achieved (standard deviation = 0).

For visualization purposes, we plot the predicted NMR spectra from the component analysis for a set of four representative cases and compared those with the corresponding mixed NMR spectra, shown in Figure 5. The descriptions for the representative mixtures are given in Table 1. In Figure 5A, a simple visual inspection could remove the false entries: astaxanthin, indolelactive acid, and L fucose. The probable cause for their inclusion in the final prediction was partial overlap between the molecular and mixed NMR spectra. In contrast, the top four molecules of the prediction in Figure 5B corresponded to the composition of the molecular mixture 23. Two of the three false positives, nicotinuric acid and linalyl acetate, could be discarded by visual inspection, and the probability of the third one, sulcatone, is less than half of that of 1,8-cineole. Figure 5C illustrates a similar analysis for a mixture containing five molecules, predicted by the top five molecules in

the analysis. An easy elimination of the false positives, nicotine and catechin, by visual inspection is achievable in this case as well. In the last example, Figure 5D, the top seven molecules in the prediction capture all six molecules in the corresponding mixture. The false positive, shikimic acid, shows up in this list because of its high degree of overlap with the mixed spectrum. However, the missing peak in the mixed spectrum between 7 and 8 ppm could be used to remove it from the predicted composition. The rest of the false positives, nicotine and sucrose, can be eliminated by visual comparison of the actual and the predicted spectra. Our method's performance summary is given in Table 2.

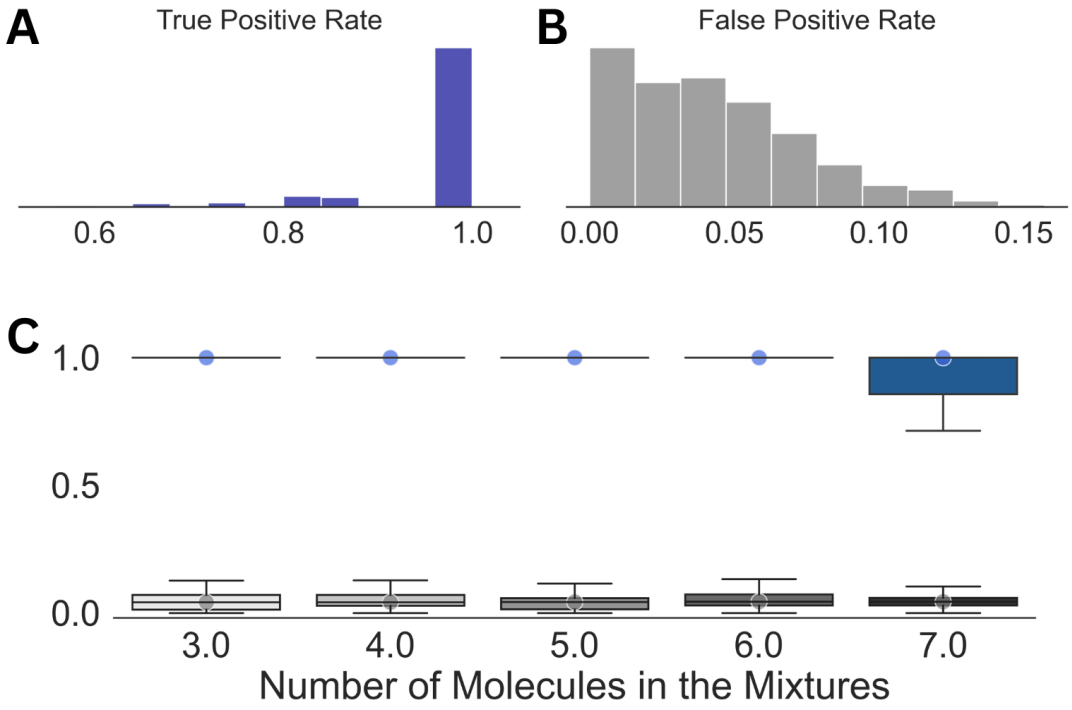

Figure 4. Summary of the results obtained after Step (4) in the analysis. Distributions of the true positive (**A**) and false positive rates (**B**) for the entire dataset along with those against the size of the mixtures (**C**) are shown. The circles in (**C**) emphasize the median true positive rate (blue) and false positive rate (gray) for each of the distributions.

Table 1. Representative set of molecular mixtures and the corresponding prediction summary.

Mixture No.	Number of Molecules	Molecules (Proportions %)	True Positive Rate	False Positive Rate
5	3	Caffeine (39), ribitol (33), cis-jasmone (28)	1.0	0.04
23	4	Nerolidol (35), 1,8-cineole (22), leaf alcohol (22), furfuryl alcohol (21)	1.0	0.04
35	5	Sorbitol (28), eugenol (26), ribitol (18), ascorbic acid (15), salicylic acid (13)	1.0	0.03

Table 1. *Cont.*

Mixture No.	Number of Molecules	Molecules (Proportions %)	True Positive Rate	False Positive Rate
20	6	Ribitol (20), eugenol (19), cis-jasmone (18), 5-methylfurfural (17), ascorbic acid (15), 1,8-cineole (12)	1.0	0.04

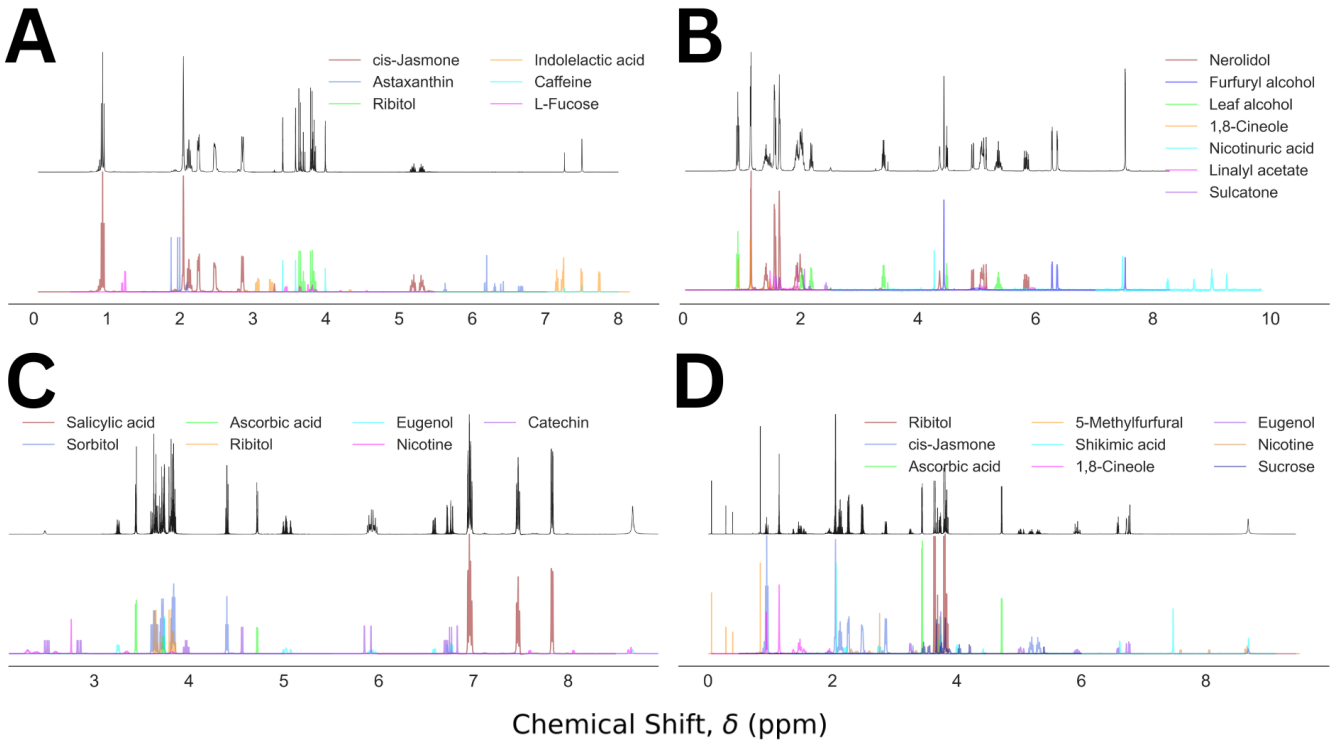

Figure 5. Mixed NMR spectra (black) and the predicted components (color coded) for Mixture Numbers 5 (**A**), 23 (**B**), 35 (**C**), and 20 (**D**), containing 3, 4, 5, and 6 molecules, respectively.

Table 2. Summary of the automated molecular mixture analyzer's performance for the augmented NMR dataset.

Parameters	True Positive Rate	False Positive Rate
Mean	0.97	0.05
Median	1.0	0.04
Standard Deviation	0.09	0.03

3. Materials and Methods

3.1. NMR to WPT Spectral Conversion

Recently, we utilized the properties of wavelet transforms for two different types of magnetic resonance spectroscopies, extracting hidden features from continuous wave electron spin resonance (cw-ESR) spectra [49] and producing highly resolved shift spectra from standard ^{1}H NMR spectra [43]. In the latter case, the input NMR spectrum is decomposed by WPT, yielding a pair of approximation and detail components, effectively separating the low- and high-frequency components in the chemical shift domain. The term frequency is used in a generic sense here, and for a particular multiplet structure, the decomposition

is continued until the derived approximation component produces a broad singlet encompassing the spectral domain. Subsequently, the multiplet in the original NMR spectrum is replaced by the peak position and height of the approximation component in obtaining the shift spectrum. This process is continued for an entire spectrum to obtain the corresponding WPT shift spectrum, while the approximation component itself is called the WPT spectrum. An example of such a spectral conversion for glutathione is shown in Figure 6.

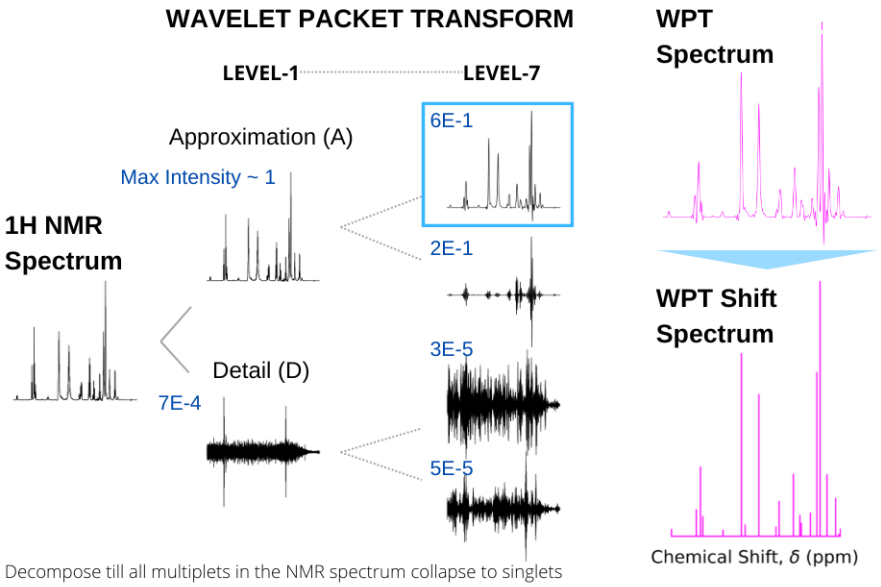

Figure 6. Conversion of the 500 MHz ^{1}H NMR spectrum of glutathione (left) to WPT and WPT shift spectra (right). In calculating the WPT shift from the WPT spectrum, only the peaks above a threshold were taken into consideration. The wavelet decomposition at Level 1 and Level 7 by the Daubechies-9 wavelet (Db9) is shown, and the maximum amplitudes of each of the components are given in blue. A decomposition at Level 7 was selected, where all the multiplets in the original NMR spectrum were reduced to singlets.

A detail description of the method and wavelet transforms can be found in [43,50].

3.2. Spectral Library and Augmented Dataset Creation

We built a spectral library with the NMR spectra of 74 small molecules. Both experimental and predicted NMR spectra were used in the library based on data availability from a peer-reviewed publication [40] and the Human Metabolome Database [51]. The corresponding WPT spectral library for the molecules was computed using the Daubechies-9 (Db9) wavelet, and a full reduction of all multiplets to singlets in a spectrum was used as the criterion to select the optimal decomposition level. We mixed the NMR spectra of 20 molecules from the library to create an augmented dataset of 1000 spectra, shown in Figure 7. Only 20 molecules were chosen in creating the augmented spectra for two reasons, (1) setting a known list of true negatives and (2) making the analysis realistic, because the mixtures in most applications usually contain structurally related molecules. Each mixed spectrum was calculated by mixing 3 to 7 randomly selected molecules' NMR spectra in varying proportions from 0.15 to 0.4. In creating a mixed spectrum, the component spectra were added in such a way that the number of data points in the mixed spectrum equaled the mean length of the individual spectra [43].

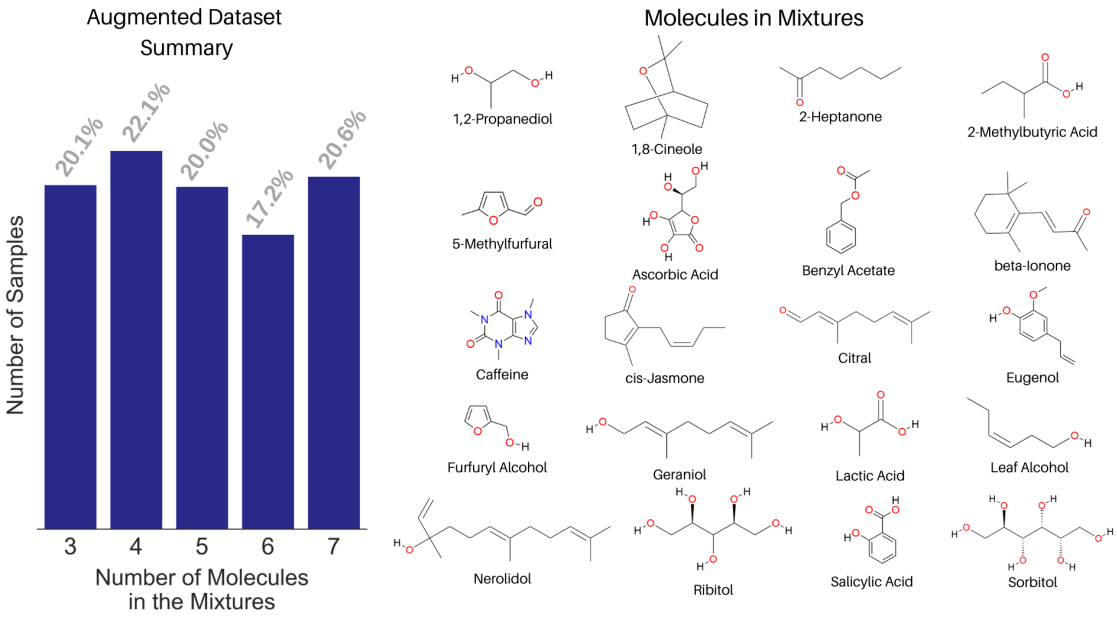

Figure 7. Summary of the augmented NMR spectral dataset with the fraction of samples against the number of constituent molecules in the mixtures (left) and the structure of the 20 molecules used in creating the augmented dataset (right).

3.3. Automated Spectral Analysis Algorithm

The algorithm used can be seen as a four-step process, (1) the conversion of an NMR spectrum to its WPT and WPT shift versions, (2) matching peaks with the WPT shift NMR library and producing a sorted list (L I) of potential components, (3) optimizing L I to L II by applying a linear gradient descent algorithm, and (4) optimizing the top 15 entries of L II to produce the final prediction of the molecular composition of a mixture. The scheme is summarized in Figure 2. Both optimization steps used linear gradient descent algorithms, but the targets (Y) were taken to be WPT shift spectra for Step (3) and the WPT spectra for Step (4). WPT shift spectra are much simplified versions of the corresponding WPT spectra, where only the peak positions and peak heights from the latter are used [43]. The design matrices (X), whose columns correspond to the potential molecules in L I or L II, were constructed from the intersections of the chemical shift values from Y and the WPT shift/WPT spectral intensities for the individual molecules. The cost function, J [52], and the gradient descent minimization are given by

$$\begin{aligned} J(\Theta) &= \sum (Y - X \cdot \Theta)^2 / m \\ \Theta_{i+1} &= \Theta_i - \alpha \, \nabla_\Theta \, J(\Theta_i) \end{aligned} \tag{1}$$

where m is the dimension of Y, Θ contains the probabilities for a set of molecules to be present in a mixture, $\nabla_\Theta J(\Theta)$ represents the gradient of the cost function, and α is the learning rate. For our method, the target Y and the design matrix X are described in Table 3. The chemical shift and intensity values from the WPT shift spectrum (in Step 3) and the WPT spectrum (in Step 4) of an experimental NMR spectrum of a small molecule mixture were used to define Y_1 and Y_2. Each of the columns in X corresponds to the molecules in our database, $(\text{Molecule}_1, \ldots, \text{Molecule}_n)$. The matrix elements, x_{ij}, were calculated by matching the WPT shift (Step 3) and WPT (Step 4) spectrum intensity of Molecule_j to the chemical shift value of δ_i, assigning $x_{ij} = 0$ if δ_i fell outside of the spectral domain of Molecule_j.

Table 3. Calculation of a target Y and the corresponding design matrix X.

Chemical Shift	Target, Y	Design Matrix, X			
		Molecule$_1$	**Molecule$_2$**	**...**	**Molecule$_n$**
δ_1	y_1	x_{11}	x_{12}	...	x_{1n}
δ_2	y_2	x_{21}	x_{22}	...	x_{2n}
⋮	⋮	⋮	⋮	⋯	⋮
δ_m	y_m	x_{m1}	x_{m2}	...	x_{mn}

The value of α in Equation (1) was chosen to be 0.1 in Step (3), and for each iteration in Step (4), α was selected randomly from a uniform distribution in the range of 0.01 and 0.1. The steps in the algorithm are summarized as follows:

1. Calculate WPT and WPT shift spectrum from an NMR spectrum;
2. Match the WPT shift spectrum with the WPT shift spectral library:
   (a) p = count the number of matches for each molecule in the library;
   (b) The probability for a molecule to be in the mixture = p/the number of peaks in the WPT shift spectrum of the molecule;
   (c) Continue for all the molecules in the library, and short-list the ones with non-zero probabilities into the list, L I.
3. Optimize the short-listed molecules by a gradient descent method:
   (a) Define the WPT shift NMR spectrum of a molecular mixture as the target variable, Y_1;
   (b) Create a design matrix, X_1, from the intersection of the chemical shift values from Y_1 and the intensities of the spectra for the molecules in L I;
   (c) Minimize $\sum (Y_1 - X_1 \cdot \Theta)^2 / n_1$, where n_1 is the dimension of Y_1 and Θ is the probabilities associated with the molecules in L I, using a gradient descent method with a learning rate, α = 0.1;
   (d) An optimized list of molecules, L II, associated with non-zero probabilities is obtained.
4. The top 15 entries from L II are used as the input to another optimization step:
   (a) Define the WPT NMR spectrum of a molecular mixture as the target variable, Y_2;
   (b) Create a design matrix, X_2, from the intersection of the chemical shift values from Y_2 and the intensities of the spectra for the molecules in L II;
   (c) Minimize $\sum (Y_2 - X_2 \cdot \Theta)^2 / n_2$ using a gradient descent method with the learning rate chosen randomly from a uniform distribution between 0.01 and 0.1;
   (d) An optimized list of molecules associated with probabilities greater than 0.1 is obtained.

We used thew true positive and false positive rates as the metrics in evaluating the performance of our spectral analysis method, defined as follows:

$$\text{True positive rate} = \frac{\text{True assignments}}{\text{Actual composition}}$$

$$\text{False positive rate} = \frac{\text{False assignments}}{\text{Spectral library} - \text{True assignments}}$$

3.4. An Example of How the Scheme Works

For an illustration, we selected the molecular mixture 82, which contains 7 molecules: lactic acid, caffeine, citral, geraniol, 2-heptanone, furfuryl alcohol, and benzyl acetate. The

NMR spectrum of the mixture is shown in Figure 8, followed by the analysis. The analysis started with the calculation of the WPT spectrum (Y_2) such that all multiplets in the original spectrum were collapsed to singlets, and subsequently, the algorithm identified the peak positions and heights from Y_2 in producing the WPT shift spectrum (Y_1). An automated sorting of molecules followed, where the peaks in Y_1 were matched with the library of the WPT shift spectra of pure molecules, which picked 64 molecules for what we call List I. In the next step, a design matrix, X_1, was created as per the description in Table 3, and the minimization of the quantity, $\sum (Y_1 - X_1 \cdot \theta)^2$, by a gradient descent was performed, where θ denotes the probability of the molecules in List I to be present in the mixture. The minimization was initiated by using a null vector of length 64 as θ. In this particular example, the minimization reduced the potential list of molecules to 62 (List II). In the next step, the top 15 molecules from List II were used to create another design matrix, X_2, and another gradient descent minimization of the quantity, $\sum (Y_2 - X_2 \cdot \theta)^2$, yielded an optimum θ. The final list of molecules corresponded to non-zero θ values, which in this case resulted in 8 molecules. The molecular composition matched the first 7 molecules in the prediction (true positives), while the last entry in the prediction was a false positive.

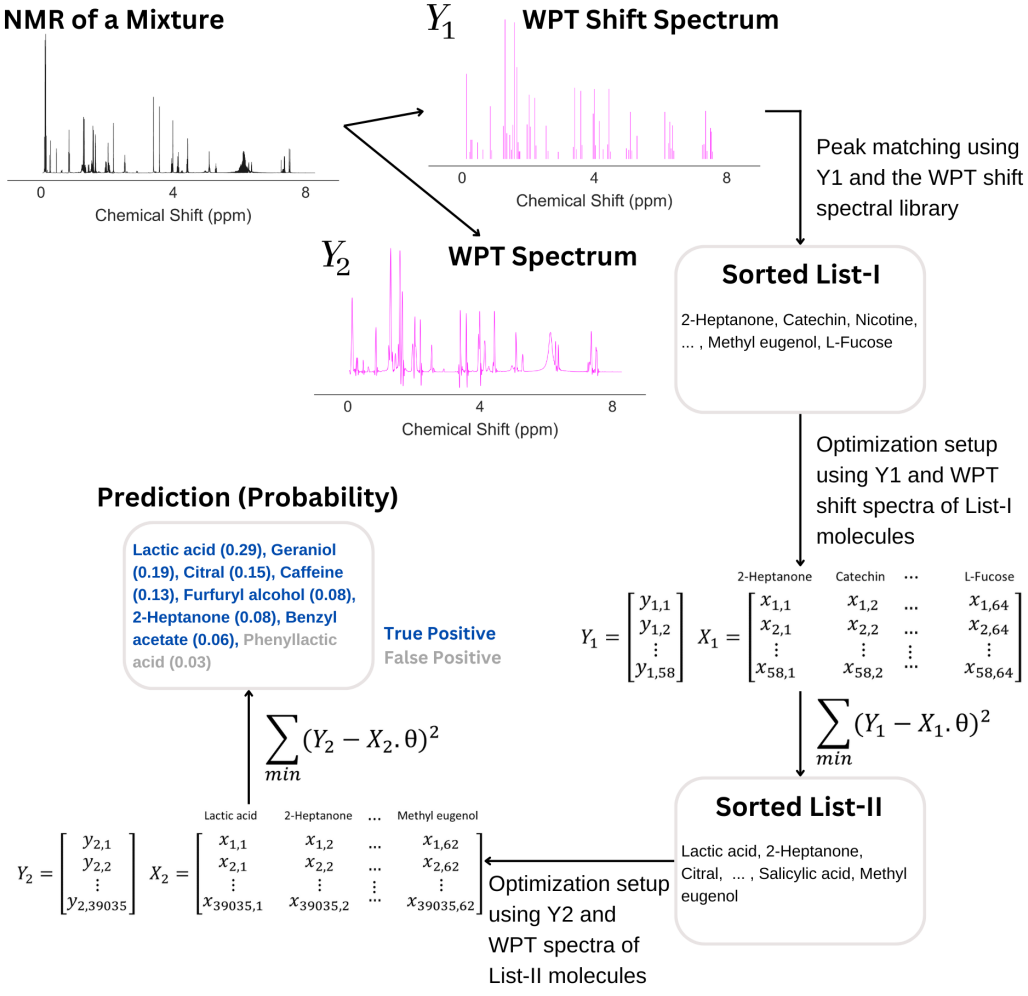

Figure 8. An illustration of how an NMR spectrum is analyzed in predicting the corresponding mixture composition. After calculating the WPT shift (Y_1) and WPT (Y_2) spectra from the NMR spectrum, an automated sorting selected 64 molecules (List I) from the library of 74 molecules by matching the WPT shift spectral peaks of Y_1 and that of the pure molecules from the library. An optimization of List I followed, yielding List II with 62 molecules. Another optimization of the top 15 entries from List II produced the final prediction, containing 8 molecules, with 7 of those corresponding to the true molecular composition of the mixture.

4. Conclusions

Composition analysis of small molecule mixtures is essential across a wide range of biological and organic research activities. While ^{1}H NMR spectroscopy is a very powerful and effective technique in identifying small molecules, the NMR spectra of molecular mixtures are often poorly resolved due to spectral overlapping and the presence of multiplet structures. In this work, we presented an automated spectral analysis algorithm, which enhances spectral resolution by the application of the wavelet packet transform and predicts the associated molecular composition in a probabilistic manner. An augmented dataset of 1000 NMR spectra, corresponding to molecular mixtures containing 3 to 7 molecules, was used to test the efficiency of our method. We obtained a median true positive rate of 1.0 for all the mixtures with zero variation for the mixtures containing up to six molecules; the true positive rate for mixtures with seven molecules had a median and standard deviation of 1.0 and 0.08, respectively. A reasonably low false positive rate of 0.04 was achieved for the dataset. In addition, we demonstrated that the precision of the analysis could be further improved by visual inspection of the actual and predicted NMR spectrum of a molecular mixture, which can be automated as well. We believe that this method can enable high-throughput analysis of small molecule mixture compositions using ^{1}H NMR as the primary or only spectroscopic tool.

Author Contributions: Conceptualization, A.S.R.; methodology, A.S.R.; software, A.S.R.; formal analysis, A.S.R.; resources, M.S.; data curation, A.S.R.; writing—original draft preparation, A.S.R.; writing—review and editing, A.S.R. and M.S.; visualization, A.S.R.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSF Grant Number 2044599.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this paper can be accessed via Signal Science Lab's repository (7 December 2022) https://github.com/Signal-Science-Lab/Unsupervised_Molecular_Mixture_Analysis.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

WPT	Wavelet packet transform
DWT	Discrete wavelet transform
NMR	Nuclear magnetic resonance

Appendix A. Overview of Wavelet Transform

A continuous wavelet transform can be defined as [53]

$$F(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{+\infty} f(\delta) \psi^* \left(\frac{\delta - \tau}{s} \right) dt \tag{A1}$$

where s is the inverse frequency (or frequency range) parameter, τ is the signal localization parameter, δ represents the chemical shift, $f(\delta)$ is the spectrum, $F(\tau, s)$ is the wavelet-transformed signal at a given signal localization and frequency, and $\psi^* \left(\frac{\delta - \tau}{s} \right)$ is the signal probing function called "wavelet". Different wavelets are used to vary the selectivity or sensitivity of adjacent frequencies with respect to signal localization. They are not dependent on a priori information of the signal or its characteristics.

Discrete wavelet transform (DWT) is expressed by two sets of wavelet components (detail and approximation) in the following way [53]:

$$D_j[n] = \sum_{m=0}^{p-1} f[\delta_m] 2^{\frac{j}{2}} \psi[2^j \delta_m - n] \tag{A2}$$

$$A_j[n] = \sum_{m=0}^{p-1} f[\delta_m] 2^{\frac{j}{2}} \phi[2^j \delta_m - n] \tag{A3}$$

where $f[\delta_m]$ is the discrete input spectrum, p is the length of the input signal $f[\delta_m]$, $D_j[n]$ and $A_j[n]$ are the detail and approximation components, respectively, at the jth decomposition level, and $\psi[2^j \delta_m - n]$ and $\phi[2^j \delta_m - n]$ are wavelet and scaling functions, respectively. The maximum number of decomposition levels that can be obtained is N, where $N = \log_2 p$ and $1 \leq j \leq N$. The scaling and wavelet functions, at a decomposition level, are orthogonal to each other, as they represent non-overlapping frequency information. Similarly, wavelet functions at different decomposition levels are orthogonal to each other.

The detail component $D_j[n]$ is the discrete form of Equation (A1), where j and n are associated with s and τ, respectively. The approximation component $A_j[n]$ represent the remaining frequency bands not covered by the detail components until the jth level. The signal $f[\delta_m]$ can be reconstructed using the inverse discrete wavelet transform as follows:

$$f[\delta_m] = \sum_{k=0}^{p-1} A_{j_0}[k] \phi_{j_0,k}[\delta_m] + \sum_{j=1}^{j_0} \sum_{k=0}^{p-1} D_j[k] \psi_{j,k}[\delta_m] \tag{A4}$$

where j_0 is the maximum decomposition level from which the input signal needs to be reconstructed. Compared to that, both the approximation and detail components at each level are further decomposed into a set of approximation and detail components. A schematic diagram of DWT and WPT decomposition against increasing levels is shown for comparison in Figure A1 [43].

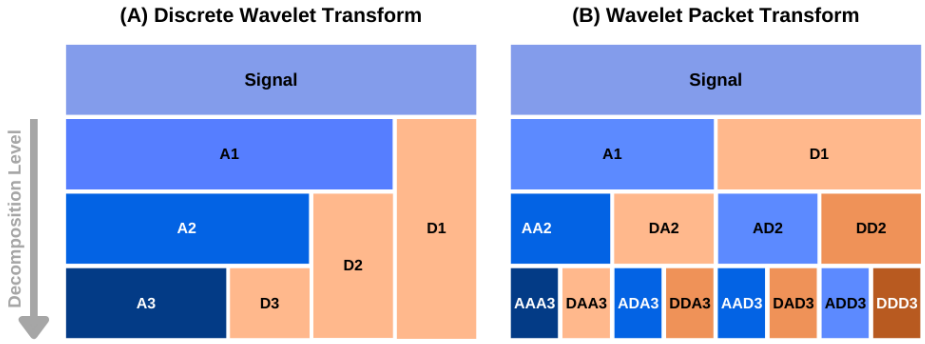

Figure A1. A schematic diagram of data decomposition in discrete (**A**) and packet wavelet transform (**B**) methods. The approximation and detail components at level k are denoted as A_k and D_k in (**A**). In the case of the wavelet packet transform, the approximation and detail components at a decomposition level are denoted by the component name of the previous level followed by A_k or D_k, respectively [43].

References

1. Pellecchia, M.; Sem, D.S.; Wüthrich, K. NMR in drug discovery. *Nat. Rev. Drug Discov.* **2002**, *1*, 211–219. [CrossRef] [PubMed]
2. Shi, L.; Zhang, N. Applications of solution NMR in drug discovery. *Molecules* **2021**, *26*, 576. [CrossRef] [PubMed]
3. Softley, C.A.; Bostock, M.J.; Popowicz, G.M.; Sattler, M. Paramagnetic NMR in drug discovery. *J. Biomol. NMR* **2020**, *74*, 287–309. [CrossRef]
4. Emwas, A.H.; Roy, R.; McKay, R.T.; Tenori, L.; Saccenti, E.; Gowda, G.N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; et al. NMR spectroscopy for metabolomics research. *Metabolites* **2019**, *9*, 123. [CrossRef] [PubMed]

5. Markley, J.L.; Brüschweiler, R.; Edison, A.S.; Eghbalnia, H.R.; Powers, R.; Raftery, D.; Wishart, D.S. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40. [CrossRef]
6. Wishart, D.S. NMR metabolomics: A look ahead. *J. Magn. Reson.* **2019**, *306*, 155–161. [CrossRef] [PubMed]
7. Pauli, G.F.; Jaki, B.U.; Lankin, D.C. Quantitative 1H NMR: Development and potential of a method for natural products analysis. *J. Nat. Prod.* **2005**, *68*, 133–149. [CrossRef] [PubMed]
8. Breton, R.C.; Reynolds, W.F. Using NMR to identify and characterize natural products. *Nat. Prod. Rep.* **2013**, *30*, 501–524. [CrossRef]
9. Robinette, S.L.; Brüschweiler, R.; Schroeder, F.C.; Edison, A.S. NMR in metabolomics and natural products research: Two sides of the same coin. *Acc. Chem. Res.* **2012**, *45*, 288–297. [CrossRef]
10. Capitani, D.; Sobolev, A.P.; Di Tullio, V.; Mannina, L.; Proietti, N. Portable NMR in food analysis. *Chem. Biol. Technol. Agric.* **2017**, *4*, 1–14. [CrossRef]
11. Martínez-Yusta, A.; Goicoechea, E.; Guillén, M.D. A review of thermo-oxidative degradation of food lipids studied by 1H NMR spectroscopy: Influence of degradative conditions and food lipid nature. *Compr. Rev. Food Sci. Food Saf.* **2014**, *13*, 838–859. [CrossRef]
12. Whitfield Åslund, M.L.; McShane, H.; Simpson, M.J.; Simpson, A.J.; Whalen, J.K.; Hendershot, W.H.; Sunahara, G.I. Earthworm sublethal responses to titanium dioxide nanomaterial in soil detected by 1H NMR metabolomics. *Environ. Sci. Technol.* **2012**, *46*, 1111–1118. [CrossRef] [PubMed]
13. Cardoza, L.; Korir, A.; Otto, W.; Wurrey, C.; Larive, C. Applications of NMR spectroscopy in environmental science. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *45*, 209–238. [CrossRef]
14. Pauli, G.F.; Godecke, T.; Jaki, B.U.; Lankin, D.C. Quantitative 1H NMR. Development and potential of an analytical method: An update. *J. Nat. Prod.* **2012**, *75*, 834–851. [CrossRef] [PubMed]
15. Caligiani, A.; Acquotti, D.; Palla, G.; Bocchi, V. Identification and quantification of the main organic components of vinegars by high resolution 1H NMR spectroscopy. *Anal. Chim. Acta* **2007**, *585*, 110–119. [CrossRef] [PubMed]
16. Barison, A.; Pereira da Silva, C.W.; Campos, F.R.; Simonelli, F.; Lenz, C.A.; Ferreira, A.G. A simple methodology for the determination of fatty acid composition in edible oils through 1H NMR spectroscopy. *Magn. Reson. Chem.* **2010**, *48*, 642–650. [CrossRef] [PubMed]
17. Lee, J.H.; Okuno, Y.; Cavagnero, S. Sensitivity enhancement in solution NMR: Emerging ideas and new frontiers. *J. Magn. Reson.* **2014**, *241*, 18–31. [CrossRef] [PubMed]
18. Mompeán, M.; Sánchez-Donoso, R.M.; De La Hoz, A.; Saggiomo, V.; Velders, A.H.; Gomez, M. Pushing nuclear magnetic resonance sensitivity limits with microfluidics and photo-chemically induced dynamic nuclear polarization. *Nat. Commun.* **2018**, *9*, 1–8. [CrossRef] [PubMed]
19. Kovacs, H.; Moskau, D.; Spraul, M. Cryogenically cooled probes—A leap in NMR technology. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 131–155. [CrossRef]
20. Zangger, K. Pure shift NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *86*, 1–20. [CrossRef]
21. Foroozandeh, M.; Morris, G.A.; Nilsson, M. PSYCHE pure shift NMR spectroscopy. *Chem.-Eur. J.* **2018**, *24*, 13988–14000. [CrossRef]
22. Aguilar, J.A.; Nilsson, M.; Morris, G.A. Simple proton spectra from complex spin systems: Pure shift NMR spectroscopy using BIRD. *Angew. Chem.* **2011**, *123*, 9890–9891. [CrossRef]
23. Lupulescu, A.; Olsen, G.L.; Frydman, L. Toward single-shot pure shift solution 1H NMR by trains of BIRD-based homonuclear decoupling. *J. Magn. Reson.* **2012**, *218*, 141–146. [CrossRef] [PubMed]
24. Casta nar, L.; Parella, T. Broadband 1H homodecoupled NMR experiments: Recent developments, methods and applications. *Magn. Reson. Chem.* **2015**, *53*, 399–426. [CrossRef]
25. Giraudeau, P. Challenges and perspectives in quantitative NMR. *Magn. Reson. Chem.* **2017**, *55*, 61–69. [CrossRef] [PubMed]
26. Yuan, B.; Zhou, Z.; Jiang, B.; Kamal, G.M.; Zhang, X.; Li, C.; Zhou, X.; Liu, M. NMR for mixture analysis: Concentration-ordered spectroscopy. *Anal. Chem.* **2021**, *93*, 9697–9703. [CrossRef] [PubMed]
27. Rogerson, A.K.; Aguilar, J.A.; Nilsson, M.; Morris, G.A. Simultaneous enhancement of chemical shift dispersion and diffusion resolution in mixture analysis by diffusion-ordered NMR spectroscopy. *Chem. Commun.* **2011**, *47*, 7063–7064. [CrossRef] [PubMed]
28. Dal Poggetto, G.; Casta nar, L.; Adams, R.W.; Morris, G.A.; Nilsson, M. Relaxation-encoded NMR experiments for mixture analysis: REST and beer. *Chem. Commun.* **2017**, *53*, 7461–7464. [CrossRef]
29. Novoa-Carballal, R.; Fernandez-Megia, E.; Jimenez, C.; Riguera, R. NMR methods for unravelling the spectra of complex mixtures. *Nat. Prod. Rep.* **2011**, *28*, 78–98. [CrossRef] [PubMed]
30. Bernstein, M.A.; Sýkora, S.; Peng, C.; Barba, A.; Cobas, C. Optimization and automation of quantitative NMR data extraction. *Anal. Chem.* **2013**, *85*, 5778–5786. [CrossRef] [PubMed]
31. Delsuc, M.A.; Levy, G.C. The application of maximum entropy processing to the deconvolution of coupling patterns in NMR. *J. Magn. Reson. (1969)* **1988**, *76*, 306–315. [CrossRef]
32. Shimba, N.; Stern, A.S.; Craik, C.S.; Hoch, J.C.; Dötsch, V. Elimination of 13Cα splitting in protein NMR spectra by deconvolution with maximum entropy reconstruction. *J. Am. Chem. Soc.* **2003**, *125*, 2382–2383. [CrossRef] [PubMed]
33. Cui, Q.; Lewis, I.A.; Hegeman, A.D.; Anderson, M.E.; Li, J.; Schulte, C.F.; Westler, W.M.; Eghbalnia, H.R.; Sussman, M.R.; Markley, J.L. Metabolite identification via the madison metabolomics consortium database. *Nat. Biotechnol.* **2008**, *26*, 162–164. [CrossRef] [PubMed]

34. Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB–Constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733–1739. [CrossRef]
35. Tulpan, D.; Léger, S.; Belliveau, L.; Culf, A.; Čuperlović-Culf, M. MetaboHunter: An automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinform.* **2011**, *12*, 1–22. [CrossRef] [PubMed]
36. Vu, T.N.; Laukens, K. Getting your peaks in line: A review of alignment methods for NMR spectral data. *Metabolites* **2013**, *3*, 259–276. [CrossRef]
37. Lepre, C.A. Library design for NMR-based screening. *Drug Discov. Today* **2001**, *6*, 133–140. [CrossRef] [PubMed]
38. dos Santos Ribeiro, H.S.; Dagnino, D.; Schripsema, J. Rapid and accurate verification of drug identity, purity and quality by 1H-NMR using similarity calculations and differential NMR. *J. Pharm. Biomed. Anal.* **2021**, *199*, 114040. [CrossRef] [PubMed]
39. Wei, S.; Zhang, J.; Liu, L.; Ye, T.; Gowda, G.N.; Tayyari, F.; Raftery, D. Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples. *Anal. Chem.* **2011**, *83*, 7616–7623. [CrossRef]
40. Wei, W.; Liao, Y.; Wang, Y.; Wang, S.; Du, W.; Lu, H.; Kong, B.; Yang, H.; Zhang, Z. Deep Learning-Based Method for Compound Identification in NMR Spectra of Mixtures. *Molecules* **2022**, *27*, 3653. [CrossRef] [PubMed]
41. Pomyen, Y.; Wanichthanarak, K.; Poungsombat, P.; Fahrmann, J.; Grapov, D.; Khoomrung, S. Deep metabolome: Applications of deep learning in metabolomics. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2818–2825. [CrossRef]
42. Corsaro, C.; Vasi, S.; Neri, F.; Mezzasalma, A.M.; Neri, G.; Fazio, E. NMR in Metabolomics: From Conventional Statistics to Machine Learning and Neural Network Approaches. *Appl. Sci.* **2022**, *12*, 2824. [CrossRef]
43. Sinha Roy, A.; Srivastava, M. Analysis of Small-Molecule Mixtures by Super-Resolved 1H NMR Spectroscopy. *J. Phys. Chem. A* **2022**, *126*, 9108–9113. [CrossRef] [PubMed]
44. Judge, M.T.; Ebbels, T. Problems, principles and progress in computational annotation of NMR metabolomics data. *Metabolomics* **2022**, *18*, 1–15. [CrossRef]
45. Monge, M.E.; Dodds, J.N.; Baker, E.S.; Edison, A.S.; Fernández, F.M. Challenges in identifying the dark molecules of life. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* **2019**, *12*, 177. [CrossRef] [PubMed]
46. Beniddir, M.A.; Kang, K.B.; Genta-Jouve, G.; Huber, F.; Rogers, S.; Van Der Hooft, J.J. Advances in decomposing complex metabolite mixtures using substructure and network-based computational metabolomics approaches. *Nat. Prod. Rep.* **2021**, 38, 1967–1993. [CrossRef] [PubMed]
47. Weljie, A.M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C.M. Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* **2006**, *78*, 4430–4442. [CrossRef]
48. Ravanbakhsh, S.; Liu, P.; Bjordahl, T.C.; Mandal, R.; Grant, J.R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; et al. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE* **2015**, *10*, e0124219. [CrossRef] [PubMed]
49. Roy, A.S.; Srivastava, M. Hyperfine decoupling of ESR spectra using wavelet transform. *Magnetochemistry* **2022**, *8*, 32. [CrossRef]
50. Srivastava, M. Improving Signal Resolution and Reducing Experiment Time in Electron Spin Resonance Spectroscopy via Data Processing Methods. Ph.D. Thesis, Cornell University, Ithaca, NY, USA, 2018.
51. Wishart, D.S.; Knox, C.; Guo, A.C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D.D.; Psychogios, N.; Dong, E.; Bouatra, S.; et al. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res.* **2009**, *37*, D603–D610. [CrossRef] [PubMed]
52. Ray, S. A quick review of machine learning algorithms. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 35–39.
53. Addison, P. *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*, 2nd ed.; CRC Press: London, UK, 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.