Politecnico di Torino

1859

POLITECNICO DI TORINO

Polytechnic University of Turin

DEPARTMENT OF MECHANICAL AND AEROSPACE ENGINEERING

Master's Degree in Biomedical Engineering

An automated data processing tool for Hyperpolarized Nuclear Magnetic Resonance: advancing precision medicine research

SUPERVISOR:

Prof.ssa Kristen Mariko Meiburger

AUTHOR:

Chiara Bernardi

CO-SUPERVISOR:

Prof. Filippo Molinari

PROJECT SUPERVISORS:

David Gomez Cabeza

Irene Marco Rius

IBEC

Institute for Bioengineering of Catalonia

MIPMED

MOLECULAR IMAGING for PRECISION MEDICINE

Turin - academic year 2023/2024

List of Figures

Abstract

In biomedical research, a comprehensive approach to detecting and monitoring diseases during different stages and measuring metabolic processes is essential for early diagnosis, managing chronic illnesses, and improving health outcomes. Precision medicine exemplifies this approach by tailoring treatment to individual patients, thereby enabling more accurate diagnoses, better disease prediction, and personalized therapies. Central to the advancement of precision medicine are innovative technologies such as Hyperpolarized Nuclear Magnetic Resonance (HP-NMR) spectroscopy, which significantly enhances the sensitivity of molecular analysis. HP-NMR facilitates real-time, non-invasive observation of molecular processes, providing unprecedented insights into dynamic biological phenomena. However, the complex signals generated by HP-NMR require efficient and accurate pre-processing methods to extract meaningful information. Addressing these challenges is essential for fully realizing HP-NMR's potential in advancing precision medicine.

To address these challenges, this thesis focuses on developing an automated tool for processing HP-NMR data. The primary goal is to overcame the inefficiency and potential inaccuracies associated with manual NMR data processing. Manual methods are time-consuming and prone to human error, making them unreliable for handling the intricate and voluminous data produced by HP-NMR spectroscopy. Hence, my primary research question is: "How can an automated tool improve the accuracy and efficiency of preprocessing HP-NMR spectra?"

To answer this question, the research explores several methodologies, including phase correction and noise reduction techniques. I evaluated three distinct approaches for phase correction: a coarse and fine tuning strategy, entropy minimization, and absolute spectrum comparison. Similarly, I tested three techniques for noise reduction: deep learning Autoencoders, singular value decomposition (SVD), and moving average filters. I tested these methods separately to determine their performance in their respective tasks.

The results indicate that the automated tool provides the accuracy and efficiency of NMR data processing. Each phase correction and noise reduction method shows varying strengths and limitations, but collectively, they contribute to a more reliable and standardized preprocessing workflow. The tool's ability to improve signal clarity and accuracy holds promise for advancing precision medicine by enabling better diagnostic and therapeutic decisions.

Hence, this thesis demonstrates the effectiveness of an automated tool in preprocessing HP-NMR spectra, addressing the challenges of noise and phase distortion. Future research should focus on refining these methodologies and exploring their applications in diverse NMR datasets to further enhance the tool's robustness and applicability in precision medicine.

Contents

Chapter 1

Introduction

Hyperpolarized Nuclear Magnetic Resonance (HP-NMR) spectroscopy stands as a powerful analytical technique that greatly amplifies signal strength, enabling detailed exploration of molecular structures and dynamics. This non-invasive and quantitative method furnishes real-time insights, proving indispensable across diverse domains such as disease detection and treatment at varying stages, metabolic process measurements, protein structure analysis, and other scientific inquiries. HP-NMR spectroscopy is widely applicable across diverse fields such as physics, chemistry, biological structure, and medicine, highlighting its critical role as a versatile and high-throughput analytical tool [15]. It is essential to enhance its effectiveness and productivity to fully capitalize on its potential in these multifaceted applications.

This research narrows its focus within the broader context of precision medicine, which aims to improve diagnostic accuracy and tailor treatments based on individual disease profiles. Precision medicine facilitates targeted approaches that empower healthcare providers to assess disease risks, prevent illnesses, identify suitable treatments, and monitor treatment responses effectively. This personalized approach holds promise for optimizing medicine by ensuring treatments are more efficacious, with minimized side effects and enhanced decision-making capabilities [13]. Cutting-edge technologies such as HP-NMR play a pivotal role in advancing precision medicine through their ability to enhance the sensitivity of molecular analyses. Therefore, optimizing the processing of HP-NMR signals is crucial for achieving substantial advancements in this field.

This research is motivated by the pivotal role of HP-NMR in precision medicine, where the signals it captures necessitate efficient preprocessing methods for extracting crucial information. Current investigations underscore the necessity for advancements in automated data processing tools to exceed limitations inherent in manual interpretation. This thesis introduces an automated NMR data processing tool designed to address challenges including sensitivity enhancement, chemical shift, and baseline correction, with specific emphasis on phase correction and noise reduction.

Central to this research is the development of an automated tool for NMR data processing aimed at enhancing the efficiency and accuracy of HP-NMR data interpretation. This objective is driven by the need to overcome current limitations in manual processing methods, particularly in sensitivity enhancement, chemical shift correction, baseline adjustment, phase correction, and noise reduction. Key research questions guiding this study include: How can automated processing improve the efficiency of HP-NMR spectroscopy data interpretation? What are the primary challenges in automating HP-NMR data processing, and how can they be effectively addressed? To what extent does the developed tool improve the accuracy of HP-NMR data analysis compared to manual methods?

Figure 1.1 shows the general workflow followed in this study . This thesis delves into the fundamental principles of NMR spectroscopy, signal acquisition, and hyperpolarization methods such as Dynamic Nuclear Polarization (DNP), which aim to enhance sensitivity. It explains the acquired signal known as Free Induction Decay (FID) in the time domain and its Fourier transform, providing information in the frequency domain. Chapter 4 shifts the focus to data processing techniques, emphasizing their importance in enhancing sensitivity, managing chemical shifts, correcting baselines, phase adjustment, and denoising. Chapter 5 thoroughly describes the materials and methodologies used in the research. Chapter 6 presents and analyzes the results of phase correction and noise reduction methods, offering quantitative and qualitative assessments that compare their performance against existing methods. The thesis concludes in Chapter 7 by summarizing findings and contributions to NMR data processing, emphasizing implications for precision medicine and suggesting future research directions. Each chapter builds on the previous ones, systematically addressing research questions and showcasing advancements in NMR spectroscopy methodologies.

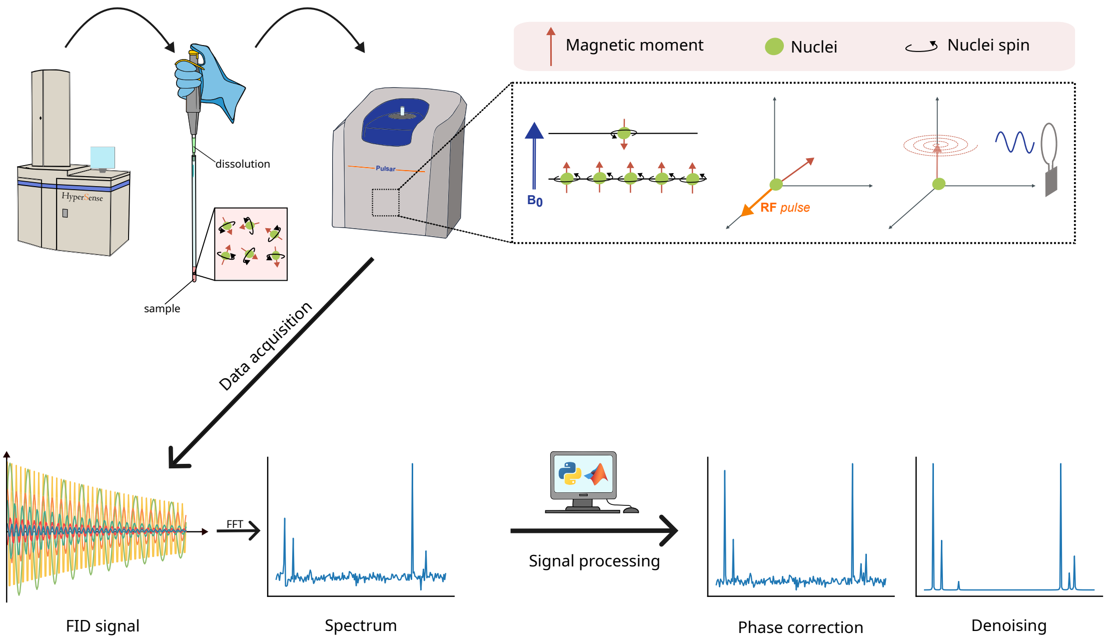


Figure 1.1: General workflow: involves analyzing a sample using hyperpolarized Nuclear Magnetic Resonance (HP-NMR) spectroscopy. The sample is placed in an NMR tube and hyperpolarized to enhance sensitivity. The tube is then placed in a homogeneous magnetic field inside a benchtop NMR spectrometer. The nuclei in the sample align into two energy states. When appropriate frequency radiation is applied, the nuclei absorb energy. As they return to equilibrium, they release this energy, generating a signal that reflects the sample's composition. The Free Induction Decay (FID) signal, initially in the time domain, is converted into a spectrum via Fourier transform. This initiates the signal processing pathway to optimize the spectrum for analysis.

Chapter 2

State of Art

Hyperpolarized Nuclear Magnetic Resonance (HP-NMR) spectroscopy significantly enhances the sensitivity of molecular analysis, facilitating real-time, non-invasive observation of molecular processes [15]. However, the complex signals generated by HP-NMR necessitate efficient and accurate pre-processing methods, such as phase correction and noise reduction, to extract meaningful information. Developing new methods in these areas is essential for advancing the effectiveness of HP-NMR in precision medicine.

Current methods for phase correction and denoising in HP-NMR exhibit significant limitations, highlighting the need for innovative approaches. Traditional phase correction methods, including manual adjustments, entropy minimization [11], and coarse and fine tuning [4], often lack accuracy, efficiency, and ease of use. Manual phase correction, while precise, is highly time-consuming and prone to human error, making it impractical for large datasets. Automated methods, such as entropy minimization, are computationally intensive and may not always converge to the optimal solution due to the complexity of the required mathematical models. The coarse and fine tuning approach, although structured, lacks the precision needed for highly sensitive HP-NMR data and does not effectively handle iterative optimization of phase parameters.

Conventional denoising methods such as moving average filters and Stationary Wavelet Transform (SWT) [2] also face challenges. Moving average filters, while simple to implement, tend to smooth out important spectral features, reducing the overall quality of the data. SWT can better preserve signal characteristics compared to moving averages but still may not achieve the desired noise reduction level, particularly in high noise scenarios.

To address these shortcomings, I propose two novel methods: an absolute spectrum comparison method for phase correction and an autoencoder-based method for denoising.

The absolute spectrum comparison method for phase correction maximizes similarity with the absolute value of the spectrum, offering a novel criterion for phase correction.

This method relies on simpler mathematical operations, making it more accessible and less computationally demanding. It incorporates a two-step iterative process: the first step optimally searches for the value of first-zero order correction, (ph_0), followed by a second step that iteratively finds the optimal value of first-oder phase correction,(ph_1). This structured approach ensures a more precise and accurate phase correction by utilizing more effective optimization algorithms.

For denoising, the application of an autoencoder leverages deep learning to distinguish between noise and true signal components. It demonstrates superior performance even under high noise conditions, ensuring the integrity of the signal is maintained. By taking into account the specific characteristics of the signal, the autoencoder-based method preserves important features of each peak in the spectra, which are critical for accurate signal evaluation.

These innovative methods address the significant shortcomings of existing techniques and provide a more robust framework for processing HP-NMR data. The absolute spectrum comparison method for phase correction and the autoencoder-based approach for denoising enhance the accuracy and efficiency of data processing while ensuring the preservation of critical details within the HP-NMR spectra, thereby improving the overall quality of the analytical results.

In summary, the development of the absolute spectrum comparison method for phase correction and the autoencoder-based denoising method marks a significant step forward in the pre-processing of HP-NMR data. These novel approaches address the limitations of existing methods and provide enhanced tools for accurate and efficient data analysis, supporting the advancement of precision medicine through improved HP-NMR spectroscopy.

Chapter 3

Nuclear Magnetic Resonance (NMR) Spetroscopy and Hyperpolarization

Nuclear Magnetic Resonance (NMR) spectroscopy is a non-invasive technique that leverages the magnetic properties of atomic nuclei to detect their chemical environment. NMR spectroscopy is versatile, applicable in both liquid and solid states, in one-dimensional (1D), two-dimensional (2D) , and multidimensional (nD) experiments. In 2D and nD NMR spectroscopy, data appear in a space defined by two or more time axes (or frequency axes in the frequency domain), providing detailed structural information. This versatility allows NMR to provide detailed information about a sample's structure, composition, purity, molecular weight, dynamics, and diffusion properties at the nanometer scale [24]. Additionally, NMR techniques are invaluable for metabolic studies at the molecular level, enabling the identification of metabolic processes and the tracking of metabolite flux in living systems, both in *vitro* and in *vivo* [42].

When a sample surrounded by a magnetic field and exposed to Radiofrequency (RF) radiation (energy) at the appropriate frequency, the nuclei in the sample can absorb the energy. After the nuclei absorb this energy, the duration and manner which they dissipate that energy provide information about various dynamic processes [44].

Despite its detailed analytical capabilities, NMR spectroscopy suffers from inherently low sensitivity due to the weak interaction between nuclear spins and magnetic fields. This limitation sets the stage for the challenges addressed in subsequent sections, which also present potential solutions to overcome the low sensitivity of conventional NMR, particularly through a hyperpolarized approach.

This chapter begins with an overview of the fundamental principles underlying NMR spectroscopy. It explores nuclear spins and magnetization, which are crucial for understanding how nuclei interact with magnetic fields to generate measurable signals.

3.1 Fundamentals of NMR

This section provides a comprehensive overview of the fundamental principles underlying Nuclear Magnetic Resonance (NMR) spectroscopy. It begins with a discussion on nuclear spins and magnetization, essential concepts for understanding NMR. Following the discussion on spins and magnetization, the phenomenon of resonance and its significance in NMR experiments receives detailed analysis. Finally, we investigate the relaxation processes that allow the nuclear spin system to return to equilibrium after excitation. This foundational knowledge is crucial for interpreting the results and applications of NMR spectroscopy.

3.1.1 Nuclear spins and Magnetization

Understanding the arrangement and presence of atoms in a chemical compounds is central to many scientific inquiries. Figure 3.2.a shows the atomic level in which the nucleus is a dense, positively charged entity described by a set of a quantum properties, one of which is *nuclear spins*, denoted by the quantum number I. The nuclear spin is fundamentally related to the composition of protons and neutrons within the nucleus of an atom or isotope. Indeed, only atomic nuclei with nuclear angular momentum are analyzable using NMR spectroscopy [50] [28]. This category includes nuclei with an odd number of protons, an odd number of neutrons, or both. Conversely, nuclei with an even number of protons and neutrons possess a nuclear spin quantum number of zero, making them unaffected by a magnetic field and, therefore, unsuitable for NMR spectroscopy. Consequently, the most common isotopes of the carbon, nitrogen, and oxygen (^{12}C, ^{14}N and ^{16}O), which lack a nuclear spin and therefore remain undetectable through NMR spectroscopy [44].

This work focuses on ^{13}C, a carbon isotope with a spin number $I = 1/2$. Consequently, its nuclear spin adopts either a positive (spin up) or negative (spin down) orientation. Any charge particle in motion generates a corresponding magnetic field [44]. When the nucleus spins anti-clockwise, it generates a magnetic field represented by an arrow pointing upwards, referred to as the magnetic moment μ, akin to a small magnet with a north pole at the arrow's tip and south pole at the tail. Conversely, a clockwise spin results in a magnetic moment pointing downwards. A nucleus with a spin of $I = \pm 1/2$ can have only these two configurations. Figure 3.2.b presents all of the aforementioned information.

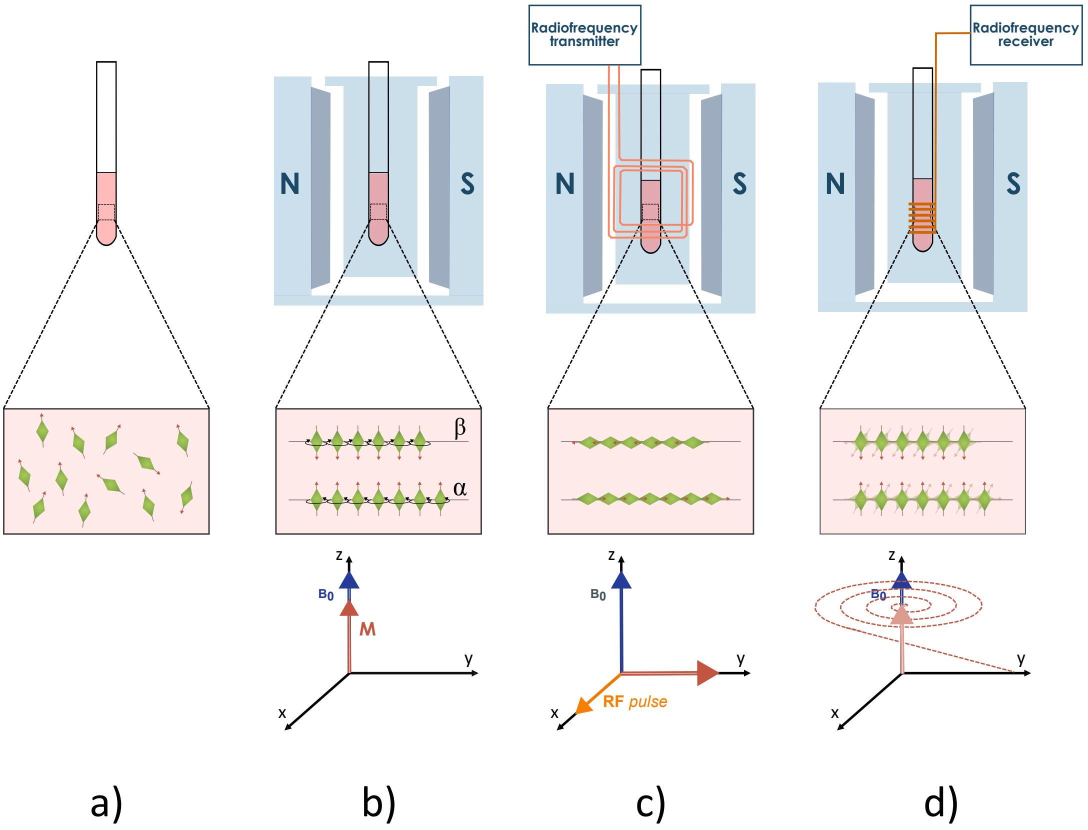


Figure 3.1: Fundamental principles of NMR spectroscopy. **(a)** Represents an unperturbed system where atomic nuclei in the sample are randomly oriented, leading to a net magnetic moment of zero as the individual magnetic moments cancel each other out. **(b)** Shows the sample places in a homogeneous static magnetic field B_0, causing magnetization. In this state, nuclei align with the external magnetic field, with the majority in the lower energy α state (parallel to B_0) and fewer in the higher energy β state (anti parallel to B_0). **(c)** Depicts the application of a short radiofrequency (RF) pulse, causing the nuclei to resonate and flip from the lower energy state to the higher energy state by absorbing energy. **(d)** The RF pulse is removed, and the nuclei return to their initial state through precession, a process known as relaxation.

The relationship between the nuclear spin (I) and the magnetic moment (μ) follows the equation:

$$\mu = \gamma I \tag{3.1}$$

Where γ is the gyromagnetic ratio, a constant dependent on the isotope [9]. NMR spectroscopy detects only atomic nuclei with $I \neq 0$ (NMR-active nuclei, such as ^{1}H , ^{2}H, ^{13}C and, ^{15}N) [9].

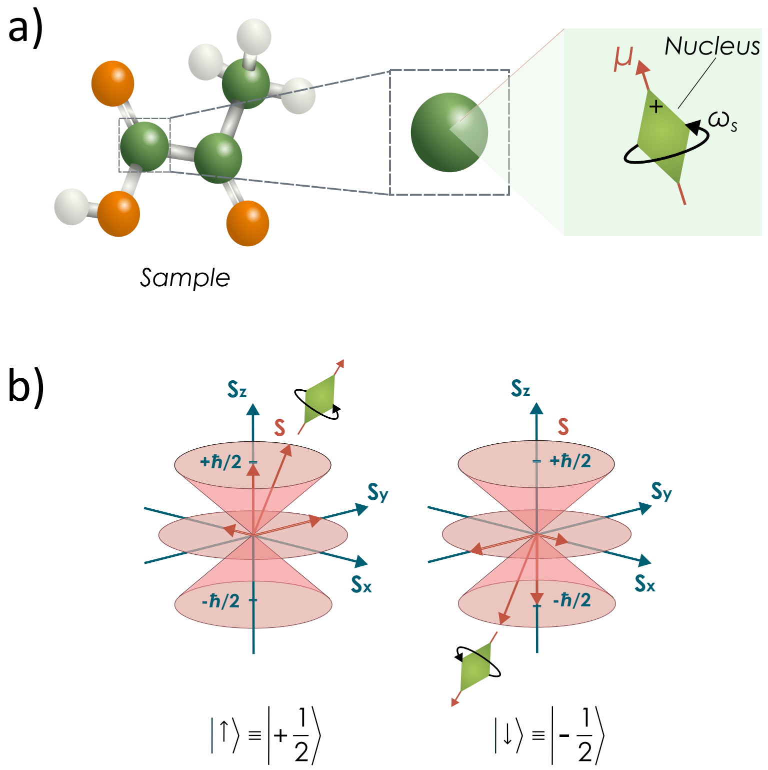

a)
Nucleus
μ
+
ω_s
Sample
b)
Sz
S
+ħ/2
Sy
Sx
-ħ/2
$|\uparrow\rangle \equiv |+\frac{1}{2}\rangle$
$|\downarrow\rangle \equiv |-\frac{1}{2}\rangle$

Figure 3.2: Structure and properties of the sample at the atomic level. **(a)** Illustration of the sample at the atomic level, highlighting the nucleus with its positive charge and quantum spin property (I). The nucleus has both an angular momentum (ω_s) and a magnetic moment (μ). **(b)** The spin positive (spin up) with $I = 1/2$, spinning counterclockwise and the magnetic moment pointing upwards, or negative (spin down) with $I = -1/2$, spinning clockwise with the magnetic moment pointing downwards.

Placing a sample in a strong magnetic field causes the magnetic moments of individual nuclei to align with the external field, a process called magnetization, showed in Figure 3.1.b [9] [50] [44]. The force of this magnetic alignment is defined by the gyromagnetic ratio γ [9].

In the field of a larger magnet, the orientation of the small magnet is no longer random as depicted in Figure3.1.a. Instead, one particular orientation becomes more probable. The most favorable orientation aligns parallel to the external magnetic field and corresponds to the positive nuclear spin $1/2$, representing the low-energy state. Conversely, the less favorable state aligns anti-parallel to the field, associated with the negative nuclear spin $-1/2$, representing the high-energy state. [44] [9]. These two orientations correspond to the two spin states labelled as α and β [30]. It is a quantum mechanical requirement that any nuclear spins with $I = 1/2$ be in one of the two states in a magnetic field.

The energy difference (ΔE) between these levels depends on the magnetic field and gyromagnetic ratio, affecting the sensitivity of the technique [9]. This relationship follows:

$$\Delta E = E_\beta - E_\alpha = \gamma \frac{h}{2\pi} B_0 \tag{3.2}$$

Where h is the Planck's constant (6.63×10^{-27} erg sec) and B_0 is the magnetic field surrounding the nucleus [44]. A stronger external magnetic field results in a larger ΔE [30].

According to Planck's equation ($E = h\nu$), the relationship in Equation 3.2 becomes:

$$h\nu = \gamma \frac{h}{2\pi} B_0 \tag{3.3}$$

By eliminating h from both sides and converting the frequency from Hz to radians/s (multiplying by 2π), the equation becomes:

$$\omega = \gamma B_0 \tag{3.4}$$

This is the Larmor equation, ω is the frequency of the precessional motion of the nucleus into the field, also called *Larmor frequency*, proportional to the magnetic field and the gyromagnetic ratio of the nucleus. Thus, in the presence of an external magnetic field, nuclei with different gyromagnetic ratios distinguish themselves by their precession frequencies. The Larmor equation defines both the precession frequency of the magnetic moment about the direction of the external field and the energy splitting associated with the transitions between quantized nuclear magnetic states [31].

In practice, a large number of nuclei in a sample placed in a magnetic field will split into two sub-populations with respective energy levels. [44] [50]. The population distribution in different energy states, when the nuclear spin system is unperturbed, follows the Boltzmann equation:

$$\frac{N_{upper}}{N_{lower}} = \frac{N_\beta}{N_\alpha} = e^{-\Delta E/kT} \tag{3.5}$$

Where N_{upper} and N_{lower} represent the population of nuclei in upper and lower energy states, respectively, k is the Boltzmann constant, and T is the absolute temperature (°K) [44].

The small excess of spins precessing in the low-energy state, randomly distributed across the precessional cone's surface, generates a macroscopic magnetization vector *M*, aligned with the magnetic field. The NMR experiment involves manipulation of the orientation of these magnetisation vectors, and therefore, it is convenient to define an axis system where the B_0 field aligs with the z-direction depicted in Figure 3.1.b [31].

3.1.2 Resonance Phenomenon

In an NMR spectrometer, energy is required to excite protons from the lower energy state (α spin state) to the higher energy state (β spin state). This energy comes from electromagnetic radiation B_1 in the radio frequency (RF) region, typically applied as a short pulse. If the RF radiation's energy matches the energy gap (ΔE), or in other words, if its frequency match the resonance frequency, the proton will flip its magnetic moment from the lower energy state to the higher energy state, and the nuclei resonate with the electromagnetic radiation [30] [44].

When the nuclei absorb the B_1 energy, the difference between spins up and down decreases, reducing the macroscopic magnetization in the z-direction. However, magnetization will persist; instead, it flips away from the direction of the static B_0 field by an angle θ because the nuclear spins are no longer randomly distributed but tend to point with the B_1 field, as showed in Figure 3.3. The flip angle achieved by the pulse depends on the nucleus's nature, the strength of the B_1 field, and the pulse duration:

$$\theta = \gamma B_1 t \tag{3.6}$$

Pulses also have phases, often applied in the *x, y, -x* or -y direction [31]. For instance, when the RF is applied on x, it corresponds to 90° pulse.

Common flip angles are 90° ($\pi/2$) and 180° (π), illustrated in Figure 3.3.c and 3.3.d, respectively. Researchers select various smaller (Figure 3.3.a) and larger (Figure 3.3.b) angles for different purposes. A 90° angle provides the largest possible M_{xy} and detectable NMR signal, requiring a known B_1 strength and duration. The displacement angle of the sample's magnetic moment is linearly related to the product of B_1 field strength and time. For fixed B_1 field strength, a 90° displacement takes half the time of 180° displacement, as Equation 3.6 indicates. With flip angles smaller than 90°, less time is needed to displace M_z, and achieving larger transverse magnetization per unit excitation time [43] [31].

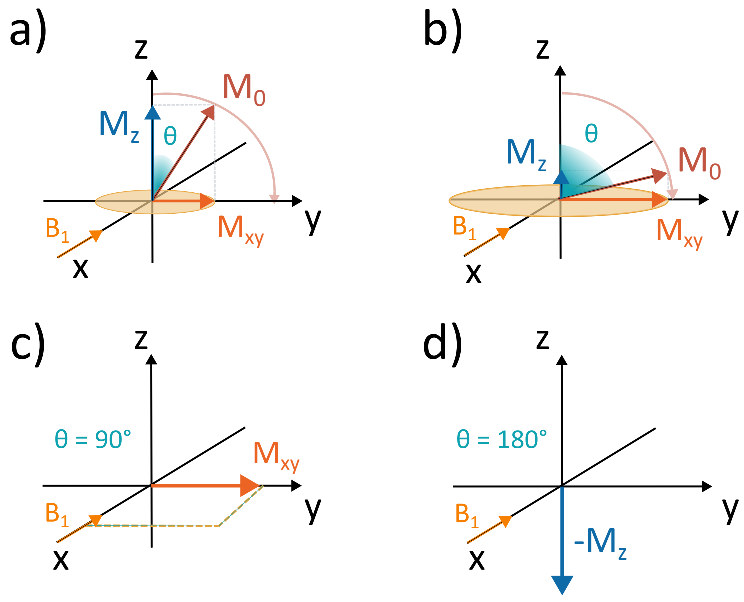


Figure 3.3: Flip angles θ and their effects on longitudinal magnetization M_z. Flip angle θ, which is the angle of displacement of the longitudinal magnetization vector M_z from its equilibrium position when a RF pulse is applied, depends on the duration and amplitude of the pulse. **(a)** Shows a small θ (less than 45°), producing a small M_z. **(b)** Depicts a larger θ (around 75° to 90°), resulting in a larger M_z. **(c)** illustrates θ equal to 90°, which generates the maximum transverse magnetization M_{xy} . **(d)** Represents a θ equal to 180°, which inverts the longitudinal magnetization, aligning it along the negative z-axis.

3.1.3 Relaxation Processes

Relaxation is the process by which a nuclear spin system return to thermal equilibrium after absorbing RF energy (Figure 3.4.b). Relaxation process, which neither emit nor absorb radiation, allow the nuclear spin system to redistribute the population of nuclear spins [44].

After switching off the B_1 field, the spins gradually lose the coherence, and the macroscopic magnetisation returns to the direction of the static B_0 field. These phenomena follow an exponential decay, described by the Bloch equations:

$$M_x(t) = [M_x(0)cos(\omega t) - M_y(0)sin(\omega t)]e^{-t/T_2} \tag{3.7}$$

$$M_y(t) = [M_x(0)sin(\omega t) + M_y(0)cos(\omega t)]e^{-t/T_2} \tag{3.8}$$

$$M_z(t) = M_{eq} + [M_z(0) - M_{eq}]e^{-t/T_2} \tag{3.9}$$

These equations define how the macroscopic magnetisation evolves in the direction of each axis, showing how the precessing macroscopic magnetization returns to the direction of the static field B_0 [31].

The Bloch equations illustrate that the recovery time of the macroscopic magnetization aligners with the static field B_0 depends on the Larmor frequency and two parameters called T_1 and T_2. These parameters, integral to the Bloch equations, characterize the relaxation processes of the macroscopic magnetization.

T_1 called longitudinal relaxation, impacting M_z and determining how quickly the magnetisation returns to alignment with the static field (Figure 3.4.a). Conversely, T_2 relates to transverse magnetization, affecting M_x and M_y and delineates the rate at which coherence is lost following the cessation of B_1 excitation (Figure 3.4.c) [31].

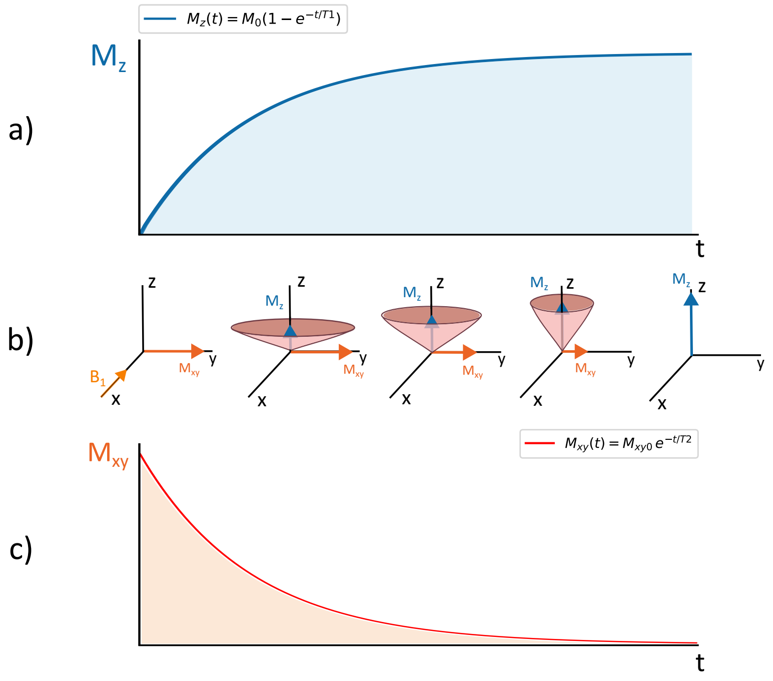


Figure 3.4: Relaxation phenomenon, removing the RF pulse the magnetic moment returns to its equilibrium state, parallel to the static magnetic field B_0. **(a)** Shows the recovery of longitudinal magnetization M_z, determined by T_1, known as longitudinal relaxation. **(b)** Depicts the relaxation process, where M_z returns to equilibrium and transverse magnetization M_{xy} is lost. **(c)** Illustrates the loss of transverse magnetization M_{xy}, determined by T_2, known as transverse relaxation.

3.1.4 NMR experiments

3.1.5 Composition and operation of NMR

An NMR system comprises five essential components: a stable superconducting magnet for generating a homogeneous magnetic field, a radio frequency (RF) transmitter for producing electromagnetic radiation, a sensitive RF receiver for detecting signals from resonating nuclei, a console to control RF pulses and digitize received signals, and software for data interpretation (Figure 3.5) [23].

The steps below detail the process of obtaining an NMR spectrum of a molecule of interest.

Samples, often dissolved in a solvent, are positioned precisely within the probe in the magnetic field. Each NMR-active nucleus in the sample possesses a microscopic magnetic moment. Initially, these magnetic moments align to form a net macroscopic magnetization vector parallel to the static magnetic field B_0.

Excitation begins with a broad-band RF pulse generated by the spectrometer's probe coils, causing the macroscopic magnetization to rotate, typically to the xy plane.

The resulting precession of the magnetization induces weak currents in the probe coils, known as Free Induction Decay (FID), which is recorded over time by the spectrometer. The FID, exhibiting a complex exponential decay pattern, is converted into the frequency domain using Fourier Transform (FT).

Multiple scans are often accumulated to improve the signal-to-noise ratio (SNR) ,factor discussed in Section 3.2.1, necessary for peak identification and structural elucidation [23].

3.1.6 Benchtop NMR Spectroscopy

Benchtop NMR spectroscopy, showed in Figure 3.5, provids a compact and versatile alternative to traditional high-field NMR systems. Unlike superconducting magnet-based systems requiring cryogenic cooling, benchtop NMR utilizes rare earth permanent magnets, operating at lower static fields (typically 0.5 - 2.5 Tesla). This technology democratizes NMR applications, making it accessible in various laboratory and manufacturing settings [35].

Despite lower resolution compared to high-field NMR, benchtop NMR instruments are easier to maintain and operate. They detect a smaller proportion of aligned spins per million nuclei due to the lower total static field, yet remain effective for diverse chemical and material analyses. Exciting aligned spins with RF pulses and recording their relaxation-induced precession forms the basis of benchtop NMR, where accumulated signal improves detection sensitivity and noise reduction [35].

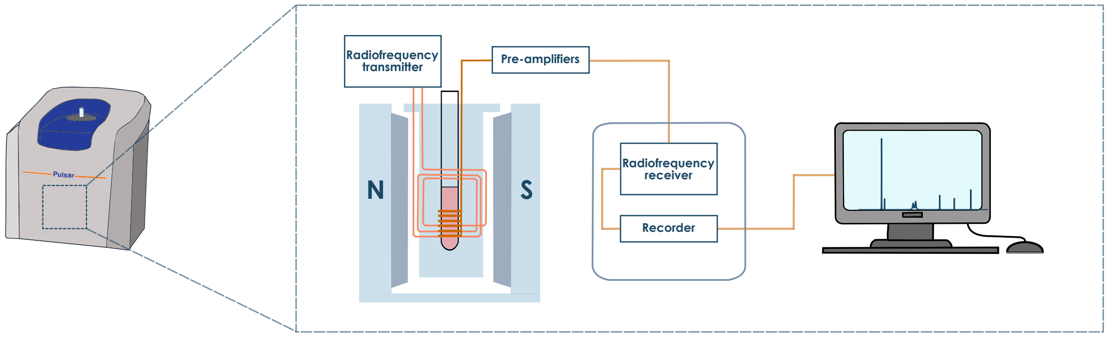


Figure 3.5: Components of a standard NMR system: a rare earth permanent magnet, which generates a homogeneous magnetic field around the centrally placed sample. A radiofrequency (RF) transmitter sends an RF pulse to the sample. Detection of the relaxation phenomena occurs through a pre-amplifier and an RF receiver. The signal then passes through a recorder before being displayed on a visualization tool.

3.2 Hyperpolarization method in NMR spectroscopy

3.2.1 Signal-to-Noise Ratio (SNR)

Noise presents a significant challenge in analytical techniques such as Nuclear Magnetic Resonance (NMR) spectroscopy, impacting data quality and sensitivity. Noise in signals is generally considered additive, modeled a signal $x(t)$ by:

$$x(t) = s(t) + n(t) \tag{3.10}$$

Where $s(t)$ is the true signal, and $n(t)$ represents the noise, typically approximated as the Gaussian noise with a flat power spectrum. In NMR spectra, noise appears along the baseline, making it difficult to distinguish trues signals from background noise, thereby reducing sensitivity.

Signal-to-Noise Ratio (SNR) quantifies the noise in a signal. SNR, expressed in decibels (dB) measures the signal strength relative to the background noise:

$$SNR_{db} = 20\ log_{10}\left(\frac{A_s}{\sigma_n}\right) \tag{3.11}$$

Where A_s is the signal amplitude and σ_n is the noise standard deviation. A higher SNR indicates a stronger signal relative to noise, essential for accurate data interpretation.

Enhancing SNR in NMR spectroscopy often involves increasing the number of scans, as already mentioned in Section 3.1.4. Each scan adds data, and averaging these scans tends to cancel out random noise while clarifying the consistent signal. Achieving an adequate SNR may require thousands of scans, especially for discerning structural peaks from background noise [23].

3.2.2 NMR Sensitivity and Boltzmann Equation

NMR sensitivity is inherently limited by the weak interaction of magnetic nuclei with magnetic fields which leads to low nuclear magnetization and spin polarization at thermal equilibrium [15]. The term spin polarization defines the degree to with the spin is aligned with a given direction [26].

These week interactions lead to small population difference between nuclear energy levels, described by the Boltzmann equation 3.5. For proton in a 18.8 T magnetic field (800 MHz) at room temperature, the population ratio is approximately 0.999872, meaning only 128 more nuclei are in the lower energy state than in the upper state per 1,000,128 nuclei (as showed in Figure 3.6.a and 3.6.b). This small excess generates the NMR signal, with the majority of nuclei canceling each other out [44].

Combining Equations 3.2 and 3.3, the Boltzmann equation 3.5 may be rewritten as:

$$\frac{N_{upper}}{N_{lower}} = \frac{N_\beta}{N_\alpha} = e^{-\nu h/kT} \tag{3.12}$$

Referring to the Equations 3.3 and 3.12, using stronger magnetic fields increases the population ratio and thus the sensitivity. Figure 3.6.a depicts the linear relation between the population ratio and the magnetic field B_0. Another method to improve sensitivity is increasing the number of nuclei in the sample, either by raising the concentration or increasing sample volume [44].

Factors that contributes to lower SNR include non-uniform magnetic field strength, which causes nuclei to achieve the Larmor condition (Equation 3.4) at different frequencies, resulting in a broader signal. The design and geometry of the receiver coil also affect sensitivity. Biological samples often have a high dielectric constant, leading to additional signal loss [44].

3.2.3 Spin Hyperpolarization

Low spin polarization levels lead to weak signal intensities in NMR, resulting in limited sensitivity in experiments. This limitation hinders the application of magnetic-resonance-based techniques across various fields, such as rapid analytical NMR in combinatorial synthesis and screening, chemical and pathogen detection, portable NMR, and magnetic resonance in imaging (MRI) for in-field chemical sensing or emergency medical diagnosis.

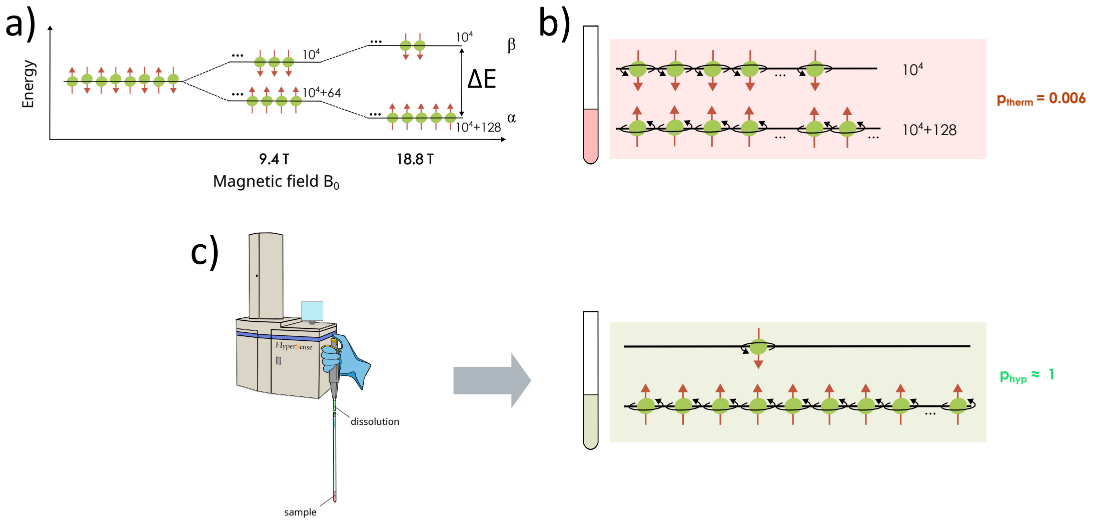


Figure 3.6: Effects of magnetic field strength and hyperpolarization on nuclear spin states. **(a)** Dependence on the magnetic field strength B_0 in separating nuclei into two different energy states for a spin $I = 1/2$, along with the relative population for each energy state assuming approximately 2 million protons in the sample (an unrealistic value as typically higher numbers are present in reality). **(b)** Thermal equilibrium depicting protons in a magnetic field of 18.8 T (corresponding to 800 MHz in ^{1}H NMR frequency) at room temperature. **(c)** Hyperpolarization effect where all spins align in the same direction as B_0, resulting in signal enhancement ($p_{hyp} \sim 1$) compared to their relative thermal equilibrium state (b).

The primary constrains include the limited amount or concentration of the sample, short feasible observation times, and the impracticality of using large supercondution magnets to induce sufficient polarization [15].

Enhancing NMR-based techniques' sensitivity can significantly broaden their applications. Hyperpolarization, which increases nuclear spin polarization by driving the system into a non-equilibrium state, offers a solution. This approach can achieve signal enhancements of four to five orders of magnitude or more, providing stronger signals than those available under thermal equilibrium [15].

To illustrate the underlying concepts in relatively simple terms, consider an ensemble of nuclei with a nonzero spin. The interaction of the spins with an applied static magnetic field generates nuclear magnetization, which remains small due to the weak interaction. Essentially, the magnetic field attempt to orient nuclear spins along one direction but cannot effectively compete with thermal randomization, resulting in weak spin orientation preference [15].

Spin polarization, indicating the population on energy levels, characterizes the degree of spin orientation. For isolated spin 1/2 species with a gyromagnetic ratio γ in a static magnetic fields, polarization p is relatively straightforward:

$$p = sgn(\gamma)\frac{n_\alpha - n_\beta}{n_\alpha + n_\beta} \tag{3.13}$$

where n_i is the number of species in state i, and $sgn(\gamma)$ is the sign (+1 or -1) of gyromagnetic ratio. Signal intensity in an NMR spectrum is directly proportional to nuclear spin polarization [15].

The polarization of a thermally equilibrated spin system, p_{therm}, depends on the magnetic field strength, temperature, and the nuclei's gyromagnetic ratio, γ_n. The factor sgn(γ) ensures p_{therm} is always positive. At ambient temperatures, spin polarization of a sample at thermal equilibrium in modern NMR instruments is on the order of 10^{-4} to 10^{-5} for the ^{1}H nuclei, and even lower for other nuclei with smaller γ_n. Only one in every 10.000 100.000 spins contributes to the observable signal due to opposing contribution from spins in α and β states (Figure 3.6.b) [15].

Increasing the intrinsically low polarization levels at thermal equilibrium involves enhancing nuclear spin polarization by driving the spin system into a non-equilibrium state, a process known as hyperpolarization. Techniques for hyperpolarizaztion can achieve spin polarization $p_{hyp} \sim 1$, corresponding to signal enhancements of 4-5 order of magnitude (Figure 3.6.c) [15].

Here is the signal enhancement defined as:

$$\varepsilon = \frac{I}{I_0} \tag{3.14}$$

where I and I_0 are the intensities of the hyperpolarized and thermally polarized spin system, respectively, under identical experimental conditions [15].

Significan signal enhancements of several orders of magnitude significantly widen the scope of applications of NMR and MRI, even on high-field instruments [15]. Diverse hyperpolarization techniques offering substantial sensitivity improvements include dynamic nuclear polarization (DNP), which replies on electron - nuclear polarization transfer, to utilize the higher polarization of electron spins. DNP can yield nuclear spin states with polarization levels far exceeding those achievable by the highest-field spectrometers under Boltzmann equilibrium [15]. Metabolites have also been hyperpolarized using parahydrogen-induced polarization (PHIP), a technique base on the use of H_2 in its singlet nuclear spin state which is called parahydrogen [15].

Ex situ dynamic nuclear polarization (DNP) produces liquid-phase samples with spin polarizations up to 50 %, providing NMR sensitivity equivalent to averaging about 1.000.000

scans. However, this process necessitates obtaining the comprehensive spectrum within just one or a few transients [36]. These advancements in hyperpolarization significantly enhance NMR sensitivity, enabling a broader range of novel and advanced applications.

3.3 FID signal and Fourier Transform

The previous sections detailed the principles of Nuclear Magnetic Resonance and acquisition of a signal that decays due to relaxation processes. This signal, know as Free Induction Decay (FID), is time-dependent. However, analysis requires a signal in the frequency domain, which is achieved through a mathematical process called Fourier Transform (FT), resulting in a spectrum.

This section begins with a discussion of the mathematical equation of the FID and its characteristics, followed by an exploration of spectrum features such as phase and lineshapes, which are closely associated with the Fourier transform.

3.3.1 The Free Induction Decay (FID)

The FID is the detectable NMR signal resulting from the precession of nuclear spin magnetization out of equilibrium around the magnetic field. When this magnetization vector has a component in the xy plane, it generates an oscillating voltage in both the detection coils (one aligned with the x-axis and one with y-axis) surrounding the sample. The FID, which is a time-domain signal, is directly proportional to the magnetization and is influenced by numerous instrumental factors [20].

Consider this signal as arising from a vector of length S_0 rotating at frequency f_0 (Figure 3.7.a). The x and y components of the vector give $S_x(t)$ and $S_y(t)$ (they are represented in Figure 3.7.b with the blue dotted line). It is convenient to regard $S_x(t)$ and $S_y(t)$ as the real and imaginary part of a complex number $S(t)$:

$$\begin{aligned} S(t) &= S_x(t) + iS_y(t) \\ &= S_0\, cos\ (\ 2\pi f_0\, t\ ) + i\, S_0\, sin\ (\ 2\pi f_0\, t\ ) \\ &= S_0\, e^{\,i\, 2\pi f_0\, t} \end{aligned} \tag{3.15}$$

However, the relaxation time T_2 limits the duration of the Nuclear Magnetic Resonance (NMR) signal, describing the rate at which the transverse magnetization decays over time [25]. Consequently, the FID follows an exponential decay with a time constant T_2 (Figure 3.7.b and Figure 3.7.c with the red line) and the Equation 3.15 becomes as:

$$\begin{aligned} S(t) &= S_0 \, e^{\,i\,2\pi f_0\,t} \cdot e^{-t\,/\,T_2} \\ &= S_0 \, e^{-t\,/\,T_2} \, cos\,(\,2\pi f_0\,t\,) + i\,S_0\,e^{-t\,/\,T_2}\,sin\,(\,2\pi f_0\,t\,) \end{aligned} \tag{3.16}$$

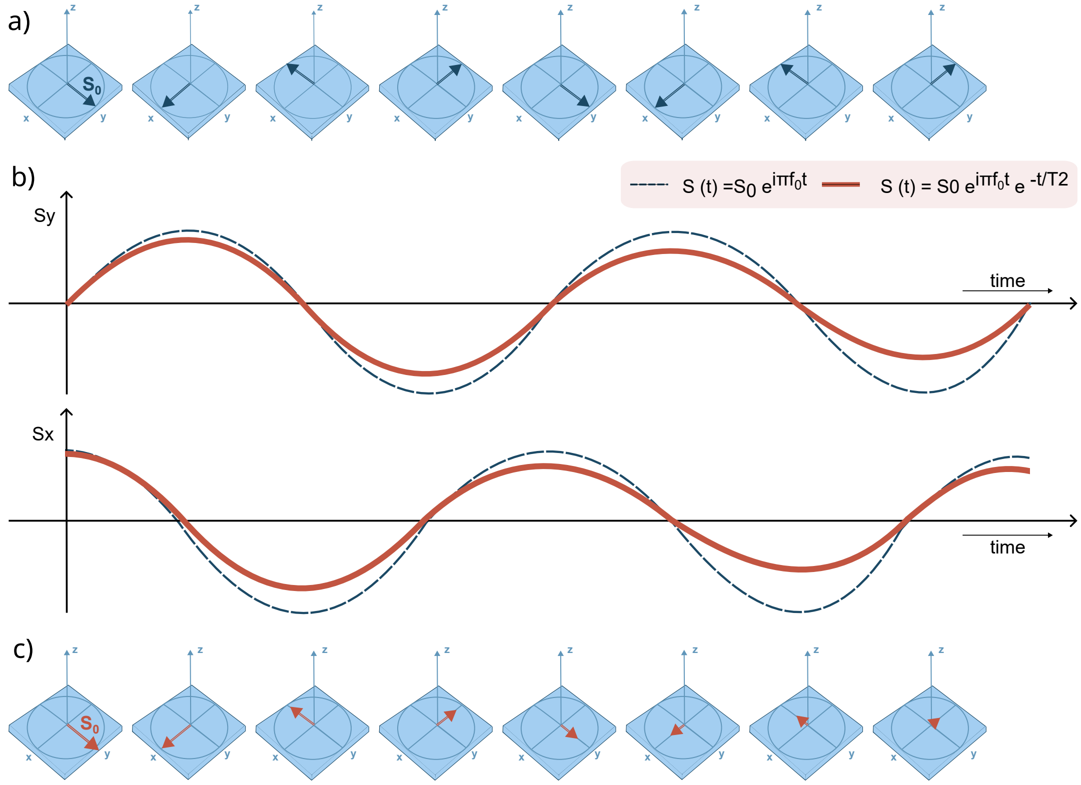


Figure 3.7: Evolution of the signal over the time: **(a)** A vector of amplitude S_0 rotating at frequency f_0 without decay, maintaining a constant amplitude over time; **(b)** The real (S_x) and imaginary (S_y) components of the signal $S(t)$, depicting both non-decaying (blue dotted line) and decaying (red line) amplitudes over time; **(c)** A vector with amplitude S_0 rotating at frequency f_0 exhibiting exponential decay over time.

When the sample contains multiple detectable nuclei, the FID reflects the combined contributions of each nucleus [25]. Therefore, Equation 3.16 becomes:

$$S(t) = \sum_{i}^{n} S_{0,n}\,e^{-t\,/\,T_{2,n}}\,cos\,(\,2\pi f_{0,n}\,t\,) + i\,S_{0,n}\,e^{-t\,/\,T_{2,n}}\,sin\,(\,2\pi f_{0,n}\,t\,) \tag{3.17}$$

Where n denotes the number of detected nuclei.

3.3.2 Fourier Transformation

The Fourier transform converts a time-domain signal into the frequency domain, which is essential for interpreting NMR parameters more effectively. Due to the linear nature of the Fourier transform, the FID contains superimposed frequencies corresponding to different chemical environments of the nuclei in the sample, resulting in peaks with specific widths and frequencies, both positive and negative, in the transformed signal [25].

Since the time-domain signal comprises both real and imaginary components (illustrated in Figure 3.8.a), the frequency-domain signal also includes these components, the real part, known as the *absorption mode* (showed in Figure 3.8.b), and the *imaginary part* (Figure 3.8.c), called the *dispersion mode.*

The term 'absorption' relates to the energy absorbed by nuclei transitioning between energy levels in a magnetic field, it is a even function typically appearing as peaks centered around the resonance frequency of the nuclei [33] [4].

'Dispersion' refers to the energy dispersed by the nuclei during these transitions, it is a odd function manifesting antisymmetric which contributes to a long positive tail on one side of the peak and a long negative tail on the other side [33] [4].

The absorption lineshape typically displays a positive value, but there are situations where this may not hold true (as explained in Section 4.4). Whereas the dispersion lineshape exhibits both positive and negative parts.

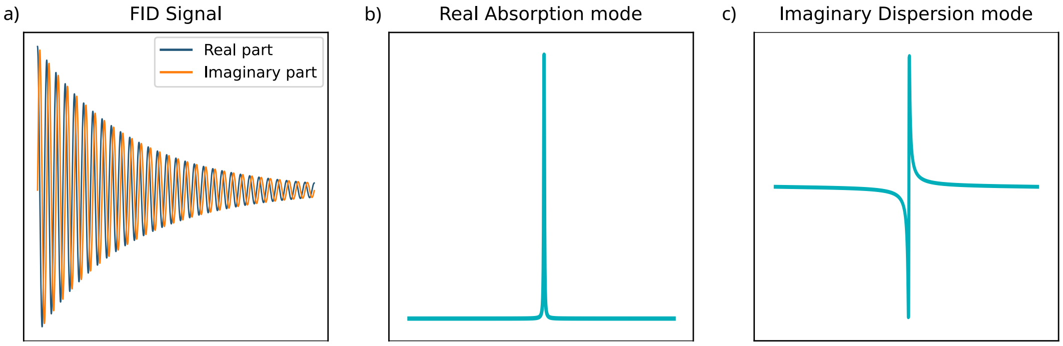


Figure 3.8: Illustration of the Fourier transformation of the time-domain signal FID. **(a)** The time-domain free induction decay (FID) signal, comprising both real and imaginary components. **(b)** Absorption mode the real part of the frequency-domain signal **(c)** Dispersione mode the imaginary part of the frequency-domain signal.

Introducing key features of the signal reveals the close link between spectrum parameters and the FID signal. A non-decaying signal (as shown in Figure 3.9) transforms into a single peak through the Fourier transform. However, the FID, which decays over time as nuclei return to a stable state after excitation removal, transforms into a peak with a broader linewidth. The area under the line (the peak intensity) correlates with the number of contributing nuclei, aiding in quantifying their relative amounts in the sample [22] [25]. This changing from a single peak to a broader peak indicates that the frequency-domain lineshape is directly related to the FID's behavior.

The equation correlating the time-domain and frequency-domain signals is:

$$W_{1/2} = \frac{1}{\pi T_2} \tag{3.18}$$

Here, $W_{1/2}$ represents the Full Width at Half Maximum, linked to the frequency domain, and T_2 denotes the relaxation time, related to the time domain [22]. This inverse relationship means that as T_2 decreases, $W_{1/2}$ increases, resulting in broader lines (Figure 3.9). Despite the area under the line remaining constant, the peak magnitude diminishes.

Thus, the peak height (the peak magnitude) in the frequency domain is directly proportional to the signal amplitude $S(t)$; an increase in S_0 leads to a corresponding rise in peak height [25].

A potential issue arises if the signal is not recorded until complete decay, leading to a 'truncated' signal. This truncation affects the frequency domain, reducing peak height and introducing ripples at the entire baseline, an undesirable artifact. Section 5.2 address this issue.

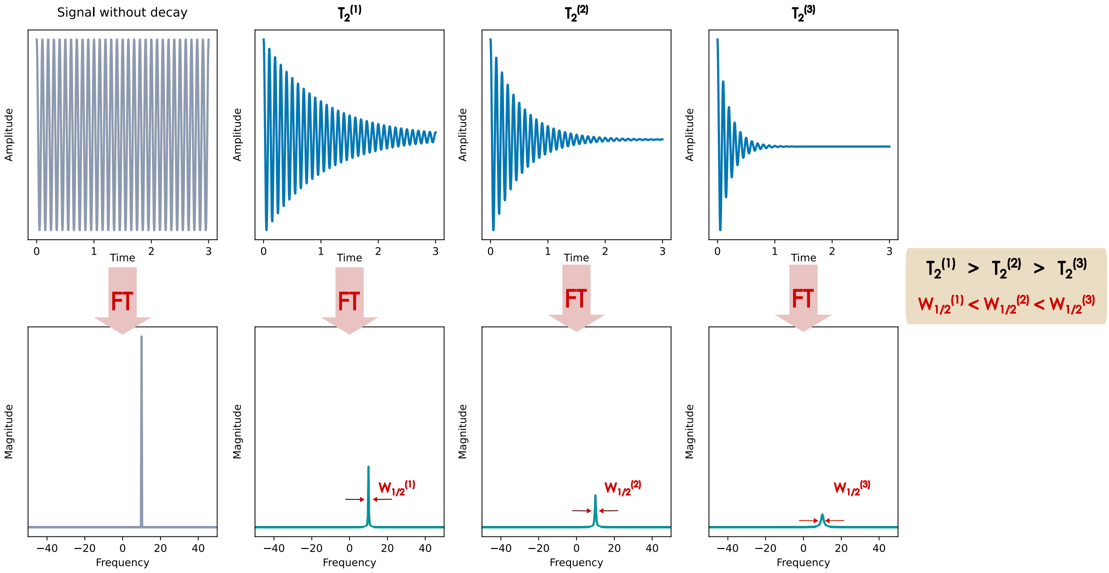


Figure 3.9: Fourier Transform of Non-Decaying and Decaying Signals: The figure illustrates a non-decaying signal that transforms into a single peak through the Fourier transform. Subsequent figure demonstrate the inverse proportionality between the relaxation time T_2 and the Full Width at Half Maximum $W_{1/2}$ increases. This leads to a decrease in the height of each peak.

Chapter 4

Data Processing

The analysis of raw NMR spectra necessitates initial processing steps such as phase correction, baseline correction, and noise reduction. These corrections are crucial for both improving the visual quality of spectra, thereby facilitating analysis, and ensuring the accuracy of quantitative results [41]. This chapter delves into these fundamental data processing techniques, emphasizing their importance in enhancing sensitivity, managing chemical shifts, correcting baselines, phase adjustment, and denoising. Each section provides a detailed exploration of methods employed to optimize the quality and reliability of NMR data, underscoring their significance in obtaining precise and interpretable results.

To evaluate the effectiveness of these processing methods, various metrics are utilized. These metrics assess the accuracy and quality of the processed spectra, ensuring that the enhancements contribute to the overall reliability of the data. By systematically applying these techniques and evaluating their impact through specific metrics, we aim to achieve the highest standards in NMR data analysis.

4.1 Sensitivity enhancement

When recording a FID signal, background noise is also captured. This interference predominantly originates from the intrinsic thermal noise generated by the instrument's detection coil.

Hence, because NMR has limited sensitivity, it is essential to improve the Signal-to-Noise ratio (SNR) of the resultant spectrum. Optimisation of the SNR signal is achieved by processing FID signal [25] . In data acquisition, it is customary to extend the recording duration beyond the decay of the FID to capture all the precessing phenomenon accurately, thereby mitigating the risk of information loss and artifact introduction in spectral analysis.

However, prolonged recording periods leads to the predominance of noise over signal as the FID weakens. Hence, employing techniques to shorten the acquisition time becomes essential. Reducing acquisition time can enhance the SNR since the most substantial signal components are typically present in the initial segment of the FID. Nevertheless, exercising caution to avoid excessively truncating the acquisition duration is essential, as doing so may result in overlooking critical portions of the FID, thereby diminishing the SNR and introducing ripple artifacts in the spectrum [25].

One effective method is to deliberately multiply the FID by a weighting, called *apodization function*, that starts at 1 and gradually decreases to zero. This approach emphasizes the early part of the FID, where the signal is strongest, and attenuates the later part, where the signal is weakest. By doing so, the essential segment of the signal remains preserved, while minimizing the noise contribution.

A typical formulation for this function is an negative exponential, as shown in the following equation:

$$W(t) = e^{-R_{LB}t} \tag{4.1}$$

In this equation, R_{LB} represents a rate constant modifiable to regulate the decay rate of the weighting function, while t denotes the acquisition time [25].

From Figure 4.1, it is possible to see the effect of the weighting function and different rate constants. In Figure 4.1.g and 4.1.h, the improvement in SNR compared to the spectrum in Figure 4.1.b is evident. However, Figure 4.1.h, which uses the more rapidly decaying weighting function shown in Figure 4.1.d, demonstrates a further reduction in the noise level.

It is important to note that a more rapidly decaying signal leads to broader line widths and a decrease in peak intensity, as shown in figures 4.1.g and 4.1.h. Thus, although the application of this function can reduce noise interference in the signal tail, it also causes greater line broadening, potentially resulting in the loss of small peaks and further reducing SNR. Therefore, two conflicting effects are at play: a more rapidly decaying function attenuates interference and noise but also broadens the lines, thereby reducing the SNR. To address this issue, normalizing the data is crucial, as shown in Figure 4.1.i and 4.1.j.

4.2 Chemical Shift

The nuclei of different elements, each with distinct gyromagnetic ratios, generate signals at different frequencies when subjected to a specific magnetic field. However, nuclei of the

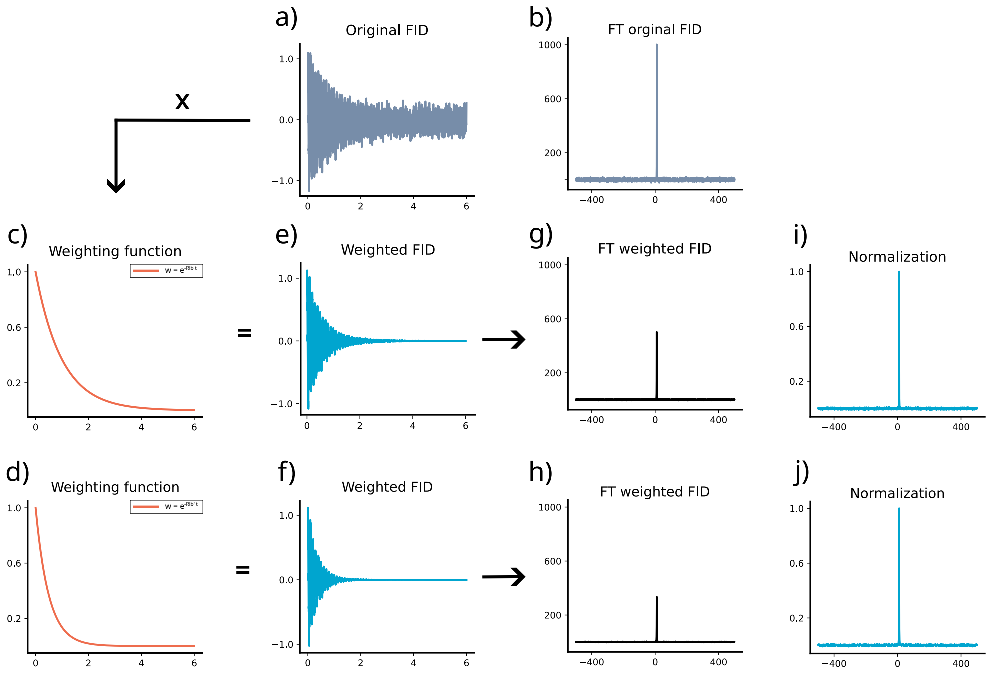


Figure 4.1: Improving Signal-to-Noise Ratio (SNR) using weighting functions: the original FID signal and its spectrum are shown in **(a)** and **(b)**. Multiplying the FID by the weighting function in **(c)** results in the signal in **(e)** and its Fourier transform in **(g)**. The more rapidly decaying weighting function in **(d)** produces the signal in **(f)** and its spectrum in **(h)**. Normalized signals in **(i)** and **(j)** highlight the SNR improvement. Weighting functions enhance SNR by emphasizing the stronger early signal components and minimizing later noise, despite causing broader linewidths and reduced peak intensity.

same type can resonate at different frequencies when the local magnetic field affecting the nucleus deviates slightly from that of another similar nucleus [22].

The variation in the local magnetic field is illustrated in Figure 4.2. When a molecule containing the nucleus of interest (nucleus B in the Figure Figure 4.2) is placed in a magnetic field (B_0), it induces electron currents within the molecule, perpendicular to the applied magnetic field. These induced currents create a small magnetic field opposite to B_0, effectively shielding the nucleus. Consequently, the magnetic field perceived by a second nucleus (A in the Figure 4.2) will be very slightly altered from the applied field B_0 due to the both contribution of B_0 and the induced magnetic field, affecting the frequency at which the nucleus resonates [22]. The shielding and the resulting resonance frequency

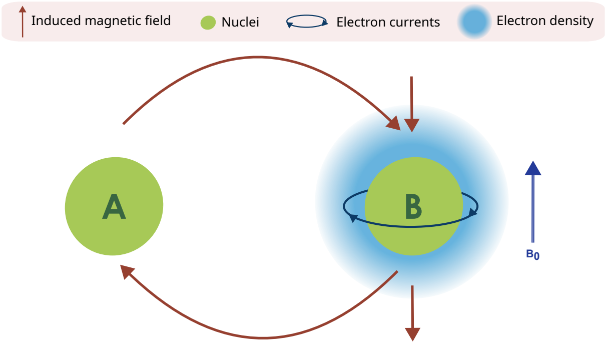


Figure 4.2: Magnetic Shielding Effect. Nucleus B placed in an external magnetic field B_0 induces electron currents that create a small magnetic field opposing B_0, shielding nearby nucleus A

are determined by the specific characteristics of the electronic environment surrounding the nucleus [22].

To eliminate this variability in NMR, the frequencies are usually measured relative to a frequency standard proportional to the magnetic field. It is convenient to define the chemical shift (expressed in term of parts per million ppm) as:

$$chemical\ shift\ (\delta) = \frac{\upsilon_{sample} - \upsilon_{ref}}{\upsilon_{ref}} * 10^6 \quad (ppm) \tag{4.2}$$

Where υ is resonance frequency. In order to establish a chemical shift scale, it is necessary to choose a reference substance, which is defined to have a chemical shift of 0.0 ppm. The most used reference substance is TMS (tetramethylsilane) [34].

4.3 Baseline correction

The baseline of a spectrum is the flat, horizontal line that connects points not related to the signal (i.e, noise). Ideally, the baseline should be perfectly flat and smooth to facilitate accurate identification and quantification of the peaks corresponding to different nuclear spins in the sample [14].

Deviations frequently arise from distortions in the initial points of the FID, often due to transmitter breakthrough. This effect occurs because the detector requires a recovery period from the pulse effect, despite being switched off during the pulse application [14]. Another potential cause is if the signal recording is terminated before the FID has fully decayed. This premature truncation results in oscillations around the base of the peaks [25].

Baseline distortion is also influenced by the first-order phase correction, which will be examined in detail in Section 5.3.

Baseline distortions can significantly impact the accuracy and repeatability of manual and automatic phase correction results, especially for methods that rely on maximizing the integral of absorption spectrum or maximizing the number of baseline points [4]. Therefore, a baseline correction is necessary. However, a robust baseline recognition method, immune to phase and baseline distortion, must be implemented to determinate the position of the left and right tails of the peaks [4]. The baseline recognition used in this study and implemented by Qingjia Bao, et.al. [4], is based on absolute value of the derivative spectra versus frequency ω, as shown in the following equation:

$$ADSpec = \left| \frac{\partial S}{\partial \omega} \right| \tag{4.3}$$

The absolute derivative spectra have the advantage of eliminating low-frequency baseline distortions through derivative operation, while all signals appear as absorption peaks [4].

Working with the absolute derivative spectra makes it immune to phase distortion, which is an important consideration [4]. In a spectrum with Lorentzian peaks, an unphased signal is represented as follows:

$$S(\omega) = A\,[\,a(\omega) + i\,d(\omega)\,]\;[\,cos(\phi) + i\,sin(\phi)\,] \tag{4.4}$$

where ϕ denotes the phase distortion of peaks. The absorption $a(\omega)$ and the dispersion $d(\omega)$ modes are defined as follows:

$$a(\omega) = \frac{A/T_2}{(1/T_2)^2 + (\omega - \omega_0)^2} \tag{4.5}$$

$$d(\omega) = \frac{A(\omega - \omega_0)}{(1/T_2)^2 + (\omega - \omega_0)^2} \tag{4.6}$$

Where A is the amplitude, T_2 is the transverse relaxation time, and ω_0 is the center frequency of the peak .

Assuming that ϕ is approximately constant, considering Equation (4.4), the absolute derivative spectra is described as follows:

$$ADSpec = \left|\frac{\partial S}{\partial \omega}\right| = \frac{A}{(1/T_2)^2 + (\omega - \omega_0)^2} = T_2 \cdot a(\omega) \tag{4.7}$$

This shows that it only depends on $a(\omega)$, so the peaks describe a pure absorption mode and are independent of phase distortion [4].

To perform the derivative, the standard numeric derivative algorithm has as major drawback of increasing the noise level. For this reason, in their work Bao et al. [4] used the Continues Wavelet Transform (CWT). In particular, the Haar wavelet has been employed because it helps in detecting changes or discontinuities, highlighting areas of abrupt change.

After baseline recognition, the baseline correction is applied. This process involves two steps. First, a baseline model is constructed using the Whittaker smoother, with all baseline points identified through the baseline recognition procedure. Subsequently, the spectrum is corrected by subtracting the baseline model [4].

4.3.1 Whittaker smoother

The Whittaker smoother, based on penalized least squares, is highly efficient, providing continuous control over smoothness and automatic interpolation [14].

Given a series y with distorted baseline of length m, where observations are regularly sampled at uniform intervals (a common scenario for most applications), the goal is to fit a smooth series z to y. Achieving this necessitates balancing two conflicting objectives: fidelity to raw data and smoothness z. A smoother z will deviate more from y [14].

Express the roughness of z in terms of d^{th} differences, typically with d typically being 1 or 2. For instance, first differences are given by:

$$\Delta z_i = z_i - z_{i-1} \tag{4.8}$$

Squaring and summing these differences provides an effective measure of the roughness of z:

$$R = \sum_{i=1}^{m} (\Delta z_i)^2 \tag{4.9}$$

Measuring the lack of fit to the data (fidelity) using the conventional sum of squares of differences:

$$F = \sum_{i=1}^{m} (y_i - z_i)^2 \tag{4.10}$$

A balanced combination of these two objectives is given by the following sum:

$$Q = F + \lambda R \tag{4.11}$$

where λ is a user-chosen parameter that trades off the smoothness of z against its fit to the data y [14]. The scalar smoothing parameter λ significantly influences the output z, and its optimal value varies based on the application. When λ tends to zero, the penalization on the estimate is minimal, resulting in a non-smoothed curve closely resembling the input data. Conversely, large values of λ lead to an oversmoothed curve with a poor fit. The optimal λ value yields a smooth cure that accurately reflects the underlying data, eliminating roughness and randomness [12] [14].

The goal is to penalize the least squares finding the series z that minimizes Q. The larger the value of λ is, the greater the influence of R on Q, resulting in a smoother z.

To simplify the algebra, it is advantageous to introduce matrices and vectors:

$$Q = |y - z|^2 + \lambda |Dz|^2 \tag{4.12}$$

Where D is a matrix such that $Dz = \Delta z$ [14].

Using results from matrix calculus, the vector of partial derivatives is found as:

$$\frac{\partial Q}{\partial z} = -2(y - z) + 2\lambda D' D z \tag{4.13}$$

Equating this to 0 leads to the linear system of equations:

$$(I + \lambda D' D) z = y \tag{4.14}$$

where I is the identity matrix [14].

Data may often contains missing values do to several reasons. To address this, modify the smoother by assigning an arbitrary value, such as 0, to the missing elements of y, and introduce a weight vector w. Set $w_i = 0$ for missing observations and $w_i = 1$ otherwise [14]. The measure of fit in Equation 4.10 is then changed to:

$$F = \sum_{i=1}^{m} w_i(y_i - z_i)^2 = (y - z)'W(y - z) \tag{4.15}$$

4.4 Phase correction

NMR spectra are mostly presented in absorption mode due to its advantages over magnitude or power mode, such as higher resolution and more accurate quantitative information regarding spin concentrations. However, post-Fourier Transform, the spectra often appear in dispersion modes other than absorption. There are many reasons attributed to this inconsistency, including the misalignment of the reference phase relative to the receiver phase detector, amplifier dead time, and phase shift introduced by the digital filter used for noise reduction. Therefore, phase correction is an essential procedure in NMR data processing [4].

Examining the FID signal, one can readily identify the factor causing a spectrum to deviate from absorption mode. Decomposing the FID signal into its x and y components could show that these components may differ at time zero. For instance, $S_x(t = 0)$ may assume a distinct nonzero value, whereas $S_y(t = 0)$ equals zero.

However, this is not necessarily always the case; the situation could be reversed or fall anywhere in between. This general scenario indicated that the signal is phase-shifted or has a phase error [25]. The phenomenon is illustrated in Figure 4.3.

In Figure 4.3.a, the signal begins out along the x-axis and precessing towards y-axis. The real part of the FID (corresponding to S_x) shows up as a cosine wave, while the imaginary part (corresponding to S_y) appears as a damped sine wave. Fourier transform yields a spectrum where the real part contains the absorption mode lineshape, and the imaginary part shows the dispersion mode. Fig. 4.3.b illustrates the effect of a 45° phase shift, ϕ. Here, S_y starts out at finite value rather than at zero, resulting in both the real and imaginary parts of the spectrum displaying a mixture of absorption and dispersion modes, rather than a pure abosorption mode lineshape. A similar phenomenon occurs with a 90° phase shift, as illustrated in Figure. 4.3.c. Finally, in Figure. 4.3.d, a 180° phase shift produces a negative absorption mode signal in the real part of the spectrum.

It can be concluded that the appearance of the spectrum depends on the position of

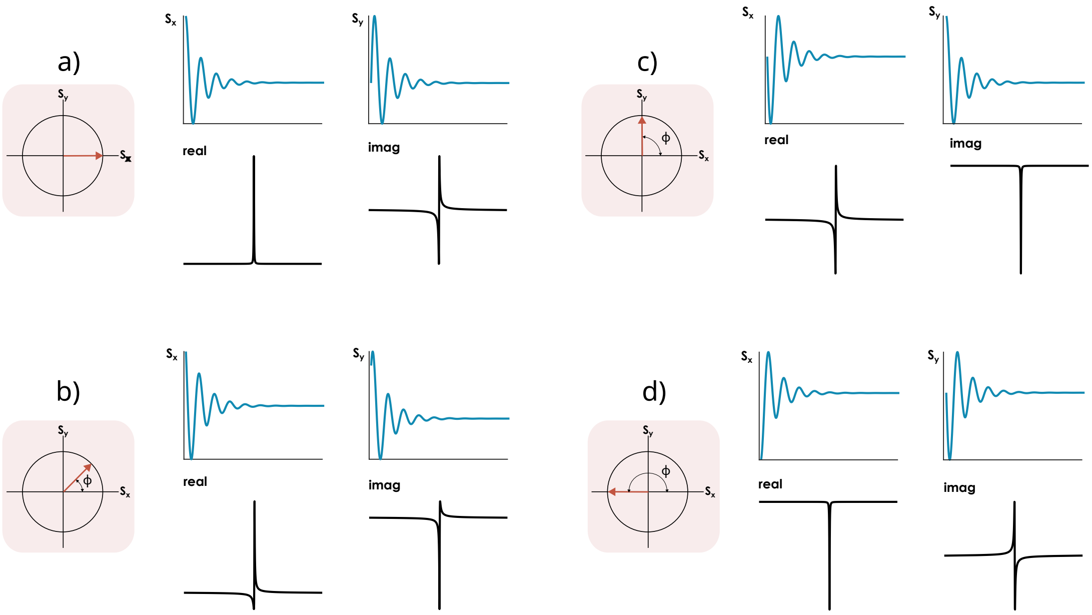


Figure 4.3: Effect of Phase Shift on NMR spectra: **(a)** Ideal absorption mode whit S_x starts at maximum and S_y at zero. **(b)** 45° phase shift causing mixed absorption and dispersion modes. **(c)** 90° phase shift showing further deviations. **(d)** 180° phase shift resulting in a negative absorption mode. In each diagram, the vector indicates the signal's position at time zero.

the signal at time zero, specifically on the phase of the signal at this initial point [25]. Mathematically, this phase distortion can be incorporated into the complex FID signal defined by the Equation 3.16 as:

$$S(t) = S_0 e^{i\phi}\, e^{\,i\,2\pi f_0\,t} \cdot e^{-t\,/\,T_2} \tag{4.16}$$

Phase correction is a process of mixing the real and imaginary parts directly obtained after Fourier transformation of the FID signal $S(t)$ [4]. The corrected spectrum $S'(\omega)$ is determined by directly multiplying the original complex spectrum, which includes an initial phase error, $S(\omega)$, by a phase shift term $\Delta\phi$. This shift term encompasses two main parameters: the zero-order phase correction ph_0 and the first-order phase correction ph_1.

$$\Delta\phi \;= \phi_0 + \phi_1 \cdot \omega \tag{4.17}$$

The phase correction follow the equation:

$$S'(\omega) = S(\omega) \cdot e^{i\Delta\phi} \tag{4.18}$$

Including Equation 4.17, the Equation 4.18 becomes:

$$\begin{aligned} S'(\omega) &= S(\omega) \cdot e^{i(\phi_0+\phi_1\omega)} \\ &= S(\omega) \cdot (cos(\phi_0 + \phi_1\omega) + i\ sin(\phi_0 + \phi_1\omega)) \end{aligned} \tag{4.19}$$

Where ϕ_0 is measured in radians, and ϕ_1 in radians per unit frequency.
By carefully tuning ph_0 and ph_1, a spectrum can be achieved where the real part corresponds to the absorption mode and the imaginary part to the dispersion mode.

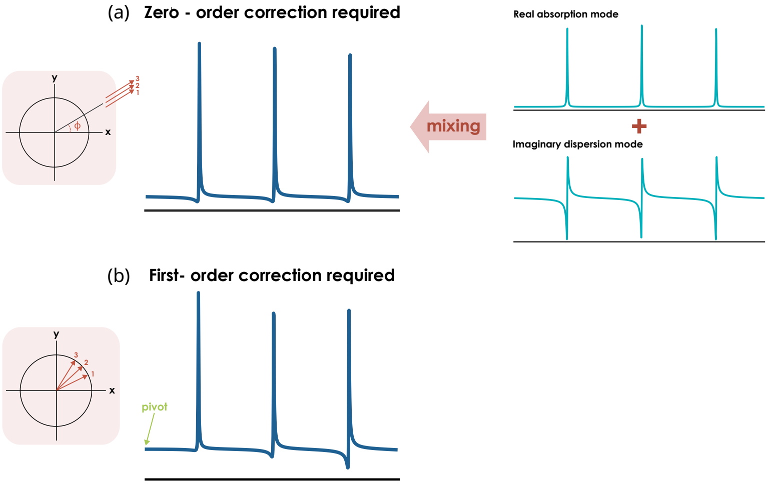

(a) Zero - order correction required
y
x
φ
3
2
1
Real absorption mode
mixing
+
Imaginary dispersion mode
(b) First- order correction required
y
x
3
2
1
pivot

Figure 4.4: Effects of zero-order and first-order phase shifts on spectra. **(a)** an example of a spectrum with a zero-order phase shift by an angle ϕ resulting in a mixture of the real and imaginary parts. All signals exhibit the same phase distortion. **(b)** An example of a spectrum with a first-order phase shift, where the phase error depends on frequency; the greater the offset from the pivot point, the greater the phase error.

The zero-order phase shift arises from a discrepancy in the relative phase between the transmitter pulse and receiver. This disparity causes a fusion of the desired real part of the spectrum with a portion of the corresponding imaginary part, resulting in a downward deviation of one side of the base of each peak below the baseline (Figure 4.4.a). Correcting this zero-order phase undoes this mixing. Notably, this correction uniformly affects all frequencies in the same way, and is deemed frequency-independent [38]. On the other hand, the first-order phase shift induces a frequency-dependent phase distortion (Figure 4.4.b). This distortion originates from delays occurring the pulse sequence and detection processes, leading to a phase error proportional to the chemical shift. When these delays are small compared to the frequency offset, corrective measures can rectify the phase error [38]. Otherwise, in scenarios characterized by significant delays, attempting such correction may result in the introduction of baseline distortion, which was introduced in the Section 4.3.

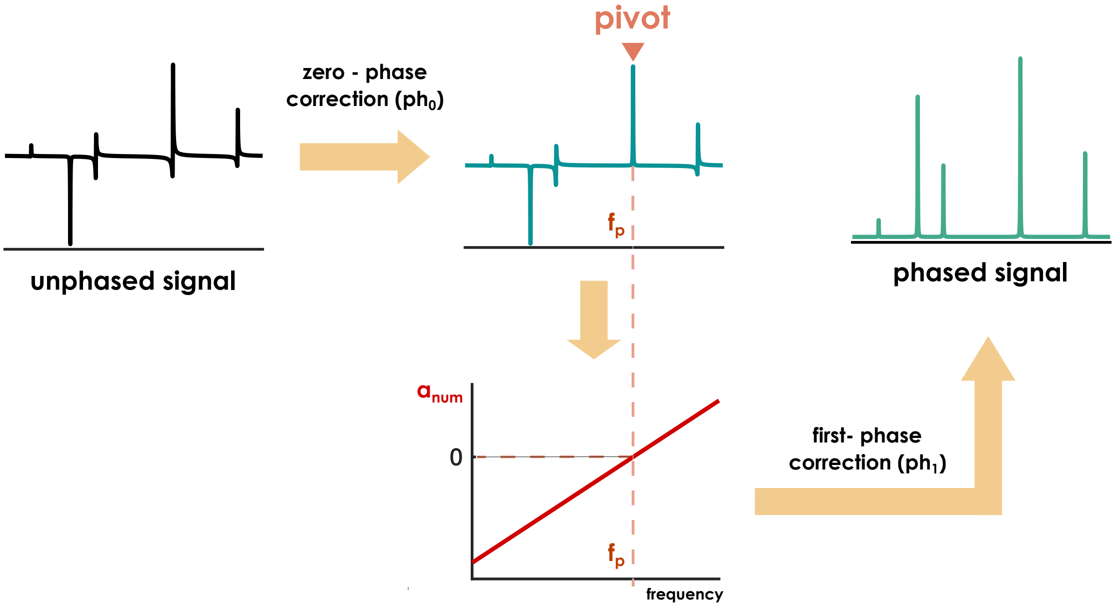

pivot
zero - phase correction (ph_0)
unphased signal
f_p
a_{num}
0
f_p
frequency
first- phase correction (ph_1)
phased signal

Figure 4.5: Phase Correction process: phase correction begins by selecting a strong peak in the spectrum as a reference point (*pivot*). First, adjust the zero-order phase to ensure the pivot exhibits a pure absorption mode. Subsequently, calculate the vector a_{num}, which is zero at the pivot frequency, f_p, and use it for frequency-dependent first-order phase correction.

To correct the phase, it is crucial to select a strong peak in the spectrum as the a reference point called pivot. Initially, the zero-order phase is adjusted to ensure that the pivot exhibits a pure-absorption mode. Subsequently, the first-order correction is fine-tuned until the signal at the opposite end of the spectrum also attains a pure-absorption mode. All the process is showed in Figure 4.5.

While manual phase correction offers satisfactory results through a carefully tuning of the ph_0 and ph_1 parameters, this approach is laborious and dependent on user expertise. Consequently, I propose automatic phase correction methods to streamline the process. We will elucidate this aspect further in Section 5.3.

4.5 Noise reduction

Noise reduction is essential in signal processing, particularly in the context of NMR spectroscopy, where noise from electronic interference, thermal fluctuations, and sample impurities can obscure spectral features and complicate data analysis. Enhancing data quality by attenuating this unwanted noise is therefore crucial. This section introduces three distinct noise reduction techniques: deep learning Autoencoders, singular value decomposition (SVD), and moving average filters. Each method employs different principles and computational approaches to effectively reduce noise and improve signal clarity.

4.5.1 Deep Learning - Autoencoder

Deep Learning (DL) operates within the domain of machine learning, specifically leveraging Artificial Neural Networks (ANNs) designed to emulate human brain data processing [46]. Unlike conventional machine learning, which relies on predefined features, deep learning automatically extracts hierarchical features from raw data. This capability is particularly effective for tasks such as noise reduction, scaling efficiently with data volume. Despite requiring extensive training time due to numerous parameters, DL models execute rapidly during testing compared to traditional algorithms [46].

This study focuses on applying noise reduction techniques to signals. Various mathematical methods exist for filtering noise, such as subtracting the mean signal when noise is consistent across multiple signals [51]. However, these methods assume prior knowledge about noise characteristics, which may not always be available. In such cases, learning noise patterns from example data becomes crucial, and DL techniques like Autoencoders are advantageous [51].

Autoencoders compress input data into a latent space representation and then reconstruct the original data. Comprising an encoder, latent space, and decoder, this architecture captures essential features in a lower-dimensional format and restores the input from this representation (Figure 4.6). In the context of signal noise, an Autoencoder processes noisy input data to learn and subsequently remove noise, reconstructing the clean signal from the noisy version.

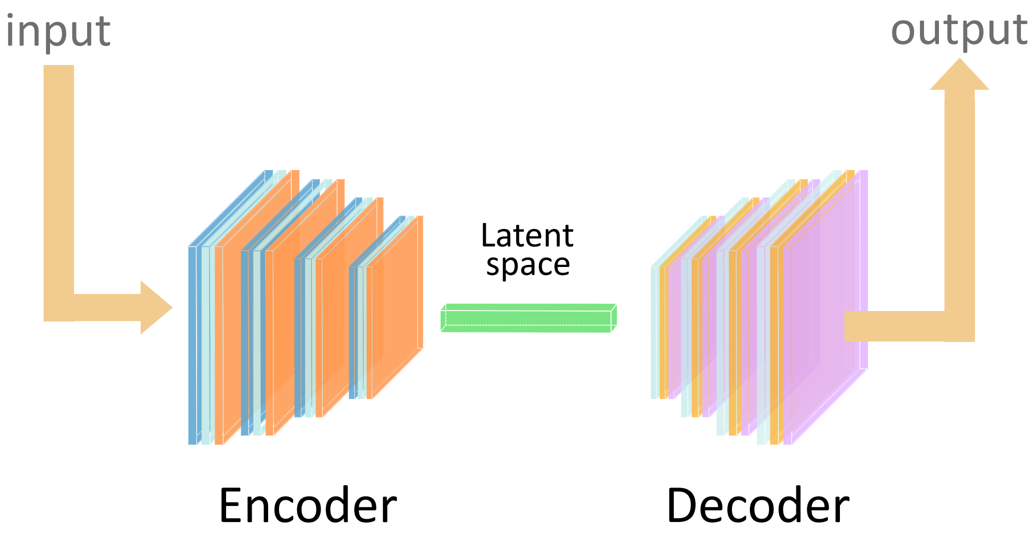


Figure 4.6: General Autoencoder architecture: Autoencoder compress input data into a latent space representation and then reconstruct the original data. Comprising an encoder, latent space, and decoder, this architecture captures essential features in a lower-dimensional format.

While effective for specific tasks like signal denoising, Autoencoders may perform poorly in other domains without retraining, highlighting their domain-specific adaptability [51].

An auntoencoder uses a mathematical operation called convolution to extract features from input data by applying a filter (or kernel) across the data. In a one-dimensional convolution (Conv1D) the kernel slides over the input data and computes a dot product between the kernel weights and the input values at each position.

Mathematically, for an input signal x and a kernel w, the convolution operation y at position t follows:

$$y(t) = \sum_{i=0}^{k-1} x(t+i) \cdot w(i) \tag{4.20}$$

where k denotes the size of the kernel. The size of the filter determines the length of the segment of the input data over which the convolution is computed. A larger kernel size captures broader features, while a smaller kernel size focuses on finer details. However, they are generally important both to control the initialization of each kernel's weights and to add constraints during the optimization process to prevent the weights from growing too large.

Other important parameters include stride and padding. Stride is the number of positions the kernel moves in each step, e.g. a stride of 1 means that the kernel only moves one position at a time. A higher stride value results in fewer convolutions and a smaller output size, but this may lead to coarse data extraction. Padding involves adding extra values (typically zeros) to the input data to control the spatial dimension of the output. There are three types of padding: same padding, which keeps the output size the same as the input size adding an equal number of elements to each side; valid padding, which denotes that there is no padding resulting in a reduction of the output size and casual padding which adds elements only to the left side of the input, also ensuring that each output only depends on the current and past inputs, not the future ones [39] [1].

Transposed convolutional layers, also known as deconvolutional layers, are integral to the decoder in Autoencoders. They upsample the compressed data back to its original dimensions. Unlike traditional convolutional layers that reduce spatial dimensions, transposed convolutional layers increase them. They achieve this by applying a reverse operation to the convolution: instead of computing dot products, they spread input values across the output space according to the kernel weights. This process helps in reconstructing the input data from its compressed representation.

To ensure the network learns complex patterns, activation functions introduce nonlinearity into the network. Common activation functions include sigmoid, hyperbolic tangent (tanh), rectified linear unit (ReLU), and leaky rectified linear unit (Leaky ReLU).

This work introduces the final two activation function previously mentioned. ReLU applies a threshold, activating neurons by outputting the input directly if positive and zero otherwise (Figure 4.7.a) . Leaky ReLU, showed in Figure 4.7.b, allows a small nonzero gradient (a) when the input is negative.

One common problem in training neural networks is overfitting. Overfitting occurs when a model achieves a good fit on the training data but does not generalize well to new, unseen data. In other words, the model learns patterns specific to the training data that are irrelevant to other data [8]. Several regularization techniques help mitigate this, including early stopping, dropout, weight initialization techniques, and batch normalization [47].

Batch normalization, in particular, makes neural networks faster and more stable by adding extra layers that perform standardizing and normalizing operations on the input

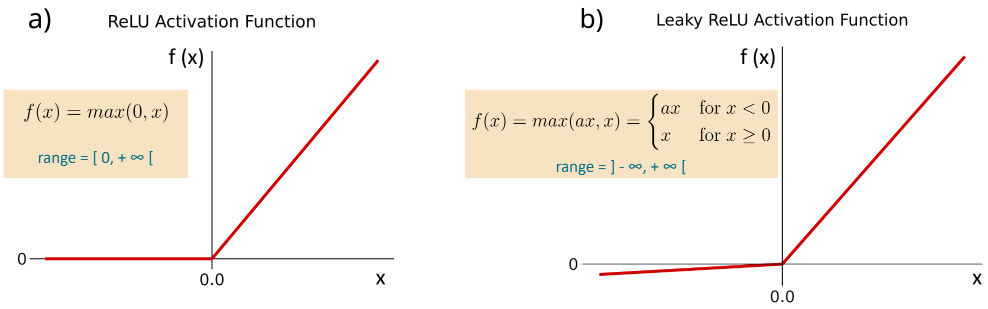


Figure 4.7: Activation Functions: **(a)** The ReLU (Rectified Linear Unit) function outputs the input directly if it is positive; otherwise, it outputs zero. **(b)** The Leaky ReLU function permits a small, non-zero slope (a) for negative input values, allowing a minor gradient when the input is negative.

of a layer coming from a previous layer. This process happens in batches, not as a single input. Batch normalization is a two-step process: first, the input is normalized, and then it is rescaled and offset [47].

Deep neural networks suffer from the degradation problem, where performance declines as the network depth increases. Autoencoders, with multiple convolutional and deconvolutional layers, also experience performance issues during image or signal reconstruction due to information loss. Residual networks with skip connections address this problem. Adding skip connections from the encoder to the decoder in Autoencoders helps improve performance. These connections directly send feature maps from an earlier encoder layer to a later decoder layer, helping the decoder form clearer decompressions of the input signal or image [49].

4.5.2 Stationary Wavelet Transform technique

The technique proposed by Adam R. Altenhof et.al [2] for denoising frequency-domain NMR data utilizes the Wavelet transform (WT) method to represent signals using orthonormal basis functions known as wavelets. Unlike the traditional Discrete Wavelet transform (DWT), which decimates components through downsampling, the Stationary Wavelet transform (SWT) retains undecimated values, preserving the original signal length at each decomposition level. This approach benefits from analyzing highly localized frequencies within a signal, aiding in the identification and removal of noise through thresholding [2].

The denoising procedure begins with SWT decomposing the real component of the frequency-domain data into approximation A_k and detail D_k components across k levels. Next, signal windowing isolates baseline noise in each A_k component. A crucial part of this method is the application of a thresholding routine, allowing selection among hard, soft, and modified thresholding. In general, the threshold constant λ is derived from the windowed noise per decomposition level, where $\lambda = \sigma_{noise}\sqrt{2\log(n)}$, with n the representing number of data points.

Hard thresholding sets any spectral intensities below a certain threshold λ to zero. Soft thresholding not only sets the spectral intensities below the threshold to zero but also shrinks the remaining coefficients towards zero by subtracting the threshold value, ensuring a more continuous signal by avoiding abrupt changes. The modified thresholding technique, as described by Wang and Dai [53], introduces an additional parameter, alpha α, to adjust the thresholding process, which ranges between 0 and 1. For each decomposition level, the modified thesholding applies the formula:

$$d_i = \begin{cases} d_i - \alpha\frac{\lambda_i^4}{d_i^3} & |d_i| \geq \lambda_i \\ (1-\alpha)\frac{d_i^5}{\lambda_i^4} & |d_i| < \lambda_i \end{cases} \tag{4.21}$$

This formula allows for nuanced adjustments to the data, reducing noise while preserving important signal features.

Finally, the inverse SWT (ISWT) then reconstructs the denoised NMR spectrum.

4.5.3 Rolling Window Technique (Moving average)

The moving average filter employs a rolling window technique. This method groups observations into sets of size n and shifts the window one observation at a time across the dataset. As the window moves, it aggregates data using a summary statistic, the average in the case of a moving average filter. For each data point, the filter replaces its value with the average of its neighboring points. Most observations are part of n-1 groups, except those near the beginning or end, which are included in fewer groups. This process dampens rapid fluctuations in the signal while preserving slower fluctuations in the smoothed signal. Section 5.6 provides a more detailed discussion of the implementation of the median filter.

4.6 Evaluation Metrics for Performance and Accuracy

In data analysis and model evaluation, various metrics assess performance and accuracy, providing quantitative measures to compare predicted outcomes against actual values.

These metrics help identify the strengths and weakness of the methods used. This section discusses the evaluation metrics employed in this work. Each metric evaluates signal A and B, where signal B is the reference signal and A represents the predicted signal from the method under evaluation for its similarity to the reference signal.

4.6.1 Correlation Coefficient

The correlation coefficient is a statistical measure describing the degree to which two variables move in relation to each other. It ranges from -1 to 1 where, 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship.

The correlation coefficient matrix, r, for A and B, consists of correlation coefficients for each pairwise variable combination:

$$r = \begin{pmatrix} \rho\left(A, A\right) & \rho\left(A, B\right) \\ \rho\left(B, A\right) & \rho\left(B, B\right) \end{pmatrix} \tag{4.22}$$

Where ρ represents the Pearson correlation coefficient, which, for variables with n scalar observations, is defined as:

$$\rho(A, B) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A_i - \mu_A}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right) = \frac{cov(A, B)}{\sigma_A \sigma_B} \tag{4.23}$$

Since A and B are always directly correlated with themselves, the diagonal entries of the matrix r are 1 [17].

4.6.2 Euclidean Norm

The Euclidean norm, also known as the L2 norm or Euclidean distance. This measure represents the length of the straight line connecting the points in Euclidean space. The result is always non-negative, with a value of 0 indicating that the two points coincide, while larger values signify greater distances between them [55]. For two vectors $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_n)$, the Euclidean distance is defined as:

$$||A - B||_2 = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{4.24}$$

4.6.3 Root Mean Square Error

The Root Mean Square Error (RMSE) quantifies the average magnitude of the errors between estimated and actual values. RMSE, derived from the Mean Square Error (MSE), provides a more interpretable measured as it retains the same units as the original data. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2} \tag{4.25}$$

Lower RMSE values signify better model performance indicating smaller discrepancies between predicted and actual values [10].

4.6.4 Least - Squares Method

The least-squares method approximates solutions by minimizing the sum of the squares of the residuals, which are the differences between observed and calculated values. To find the least-squares solution x for a system of linear equations represented in matrix form as $Ax = B$, solve:

$$\underset{x}{min} \ || \ Ax - B \ ||_2^2 \tag{4.26}$$

When x is close to 1, A and B are highly similar, indicating a good fit between observed and calculated values. As x deviates from 1, the similarity between A and B decreases, reflecting a poorer fit [5].

4.6.5 Structural Similarity index (SSIM)

The structural Similarity Index (SSIM) measures the uniformity of a spectrum or pattern by comparing predicted values with the actual values. SSIM is particularly useful for comparing signals because it accounts for structural information, offering a more comprehensive assessment of similarity than metrics focusing solely on magnitude differences.

The SSIM equation is:

$$SSIM(A, B) = \frac{(2\mu_A\mu_B)(2\sigma_{AB})}{(\mu_A^2 + \mu_B^2)(\sigma_A^2 + \sigma_B^2)} + c \tag{4.27}$$

Here, μ_A and μ_B represent the means of A and B, σ_A^2 and σ_B^2 denote variances of A and B, σ_{AB} is the covariance between A and B. The constant c is a small constant that ensures the SSIM is bound over a range of $[-1, +1]$, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates perfect dissimilarity [2].

Chapter 5

Methods

This chapter focuses on developing automated methods for processing hyperpolarized NMR spectroscopy spectra, crucial for enhancing spectral analysis quality and accuracy to facilitate more reliable data interpretation.

It begins with foundational processing steps, including data extraction from FID signals, parameter selection from instrument settings, and essential spectral adjustments. These steps ensure data readiness for subsequent analysis by providing accurate preprocessing.

Next, the chapter explores phase and baseline correction techniques, detailing both automated algorithms and manual methods. Automated approaches include coarse and fine tuning, entropy minimization, and maximizing similarity between phased and absolute vale of the unphased signals. Baseline correction methods additionally address distortions encountered during phase correction, ensuring a clear baseline for accurate peak identification and quantification.

Additionally, the chapter covers the generation of synthetic datasets simulating real-world NMR spectroscopy data. These datasets incorporate realistic noise patterns and signal characteristics, essential for training and evaluating noise reduction techniques.

Moreover, the chapter investigates advanced techniques such as autoencoder neural networks for spectral noise reduction. Autoencoders train to reconstruct clean spectra from noisy inputs, utilizing deep learning to enhance signal clarity and quality. Alternative methods such as Stationary Wavelet Transform and Rolling Window techniques are also under exploration for their potential to complement or substitute neural network-based approaches.

In summary, this chapter aims to establish a robust framework for automated spectral processing in hyperpolarized NMR spectroscopy. By integrating diverse methodologies from initial data preprocessing to advanced noise reduction techniques it aims to enhance the efficiency and reliability of spectral analysis, contributing to advancements in both research and practical applications of NMR spectroscopy.

5.1 Materials

I evaluated the performance of the methods using signals collected by the Pulsar NMR benchtop from Oxford Instruments [40]. I previously discussed the operational principles of this instrument in Chapter 3. In total, I analyzed 31 experiments on ^{13}C, a carbon isotope, each comprising a variable number of spectra. This variability arises because some spectra contain multiple signals from different time instants, as the instrument permits sequential acquisitions during chemical reactions to monitor changes in specific functional groups. Overall, I analyzed 2021 spectra.

For each experiment, the Pulsar instrument provides one file containing the FID signal and another containing the list of parameters and tool settings. Each experiment has a parameter set-up routine, resulting in variability in the spectra's parameters, such as the number of points associated with each spectrum. I used these spectra to assess both phase correction and denoising methods. For phase correction methodologies, I evaluated the metrics across the entire experiment, computing an average of the metrics for experiments containing multiple spectra.

For noise reduction methodologies, I generated synthetic data to closely represent the real data, as Deep Learning techniques require a large training dataset. In evaluating the effectiveness of the denoising techniques on both synthetic and real data, I assessed each spectrum individually rather than averaging metrics for experiments with multiple spectra. This approach ensured a comprehensive assessment of the denoising methods.

5.2 Basics processing

After extracting the data, it is important to select key parameters from the settings of the Pulsar benchtop instrument [40]. These parameters include the Receiver Points (RP) expressed in Hz, which indicate the number of data points collected in each scan. The Channel Frequency Offset (O) given in HZ is the deviation from the base frequency for the nucleus channel (such as Hydrogen, Carbon, or another nucleus). The Base Frequency (SF) measured in MHz represents the frequency of the reference substance. Finally, the *Filter* (*bound*) refers to the bandwidth, which determines the range of frequencies allowed during data acquisition.

An inspection of the data revealed a delay at the beginning of the FID signal (Figure

5.1.a), starting at zero values and lacking significant information. Applying the FT to this FID signal produces a spectrum with a ripple baseline (Figure 5.1.b), a phenomenon similar to the effects mentioned in Section 3.3.2 when the signal undergoes truncation. The ripples effects in the resulting baseline distortion often cause automated algorithms to misidentify them as peaks, or they may obscure nearby weaker peaks leading to misinterpretation of the signal. To remove this delay, it was quantified using two parameters: the RP provided by the instrument and the Actual Points (AP), which is the number of points in the signal considering the starting offset. The number of samples associated with the delay was calculated using the following equation:

$$delay = AP - RP - 1 \tag{5.1}$$

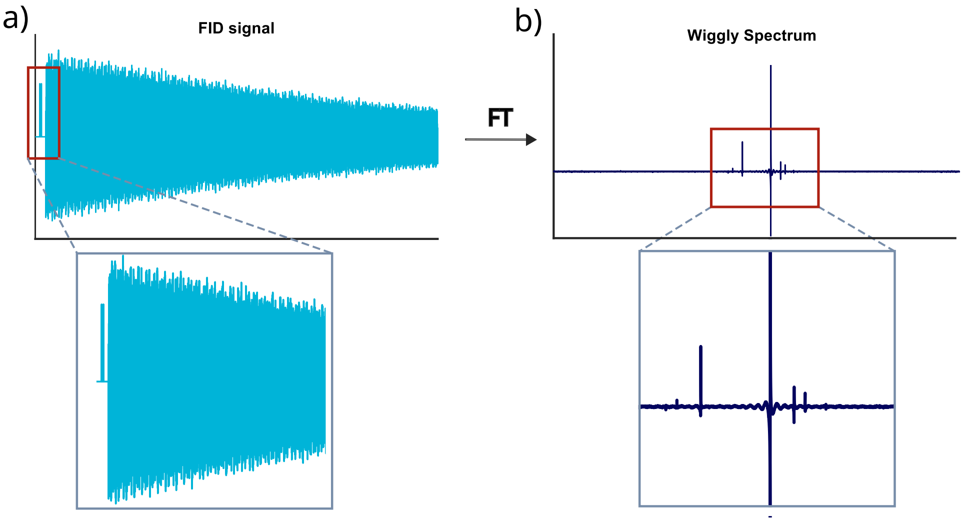


Figure 5.1: Impact of an initial delay in the FID signal on the resulting spectrum. **(a)** Initial segment of the FID signal showing a delay with no significant information. **(b)** Spectrum obtained by applying the Fourier Transform to the FID signal, displaying a ripple baseline caused by the initial delay.

Chapter 4 on Data Processing explains the sensitivity enhancement method, which the subsequent procedure implemented and showed in Figure 5.2. Considering Equation 4.1 the only modifiable parameter is R_{LB}. To apply weighting function that matches the linewidth of the highest peak in the spectrum select the appropriate value of R_{LB} is necessary. For this reason, R_{LB} is defined as the width at half prominence of the highest peak R_2, expressed in frequency.

The calculation of the weighting function utilized Equation 4.1 and normalized it to start at 1.

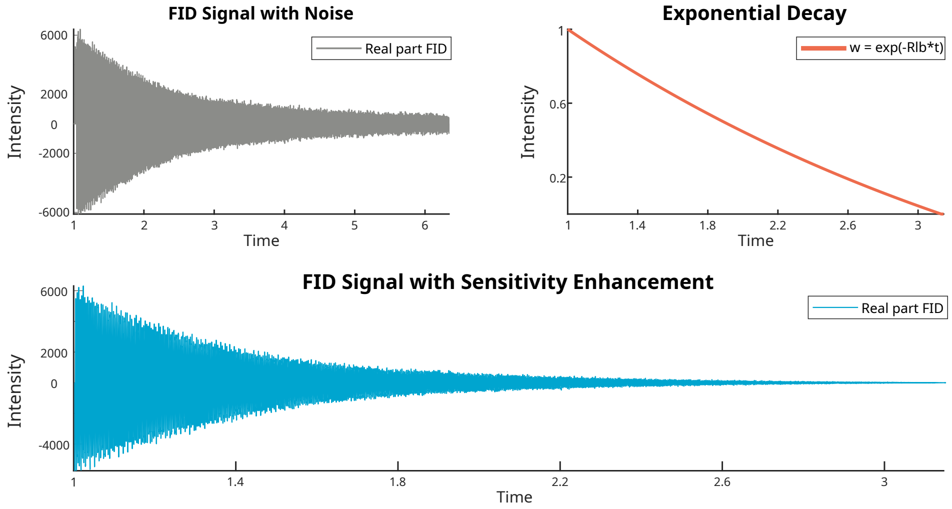


Figure 5.2: Sensitivity enhancement applied to a real signal. The noisy FID signal is shown in gray, while the red line represents the weighting function, an exponentially decaying function. The light blue line depicts the result of multiplying the noisy FID signal by the weighting function, which decays more rapidly than the noisy signal.

The subsequent step involved expressing all spectra in terms of chemical shift to ensure their independence from the magnetic field strength of the NMR spectrometer. This standardization facilitates the comparison of chemical shift across different experiments and spectrometers. I adjusted the spectra using a constant O, which requires careful consideration. Essentially, the observed and reference frequencies are centered around this offset. Given that the reference frequency provided by the instrument is expressed in MHz, the chemical shift in ppm was calculated using the following equation:

$$ppm = \frac{(f + O) - (SF + O)}{SF + O} * 10^6 \tag{5.2}$$

With some simplification, the Equation (5.2) becomes:

$$ppm = \frac{(f - SF)}{SF + O} * 10^6 \tag{5.3}$$

The highest peak in each spectrum, corresponding to the $[1 -^{13} C]pyruvate$ signal, should appears at approximately 171 ppm [7]. However, the alignment of these peaks was inaccurate due to the procedure conducted, as showed in Figure 5.3.b, requiring further adjustments. Primarily, one must calculate the disparity in ppm terms between the observed highest peak position in the spectrum and its actual value. Once this shift value is determined, it is subtracted from the Equation (5.3) obtaining the spectrum in Figure 5.3.c.

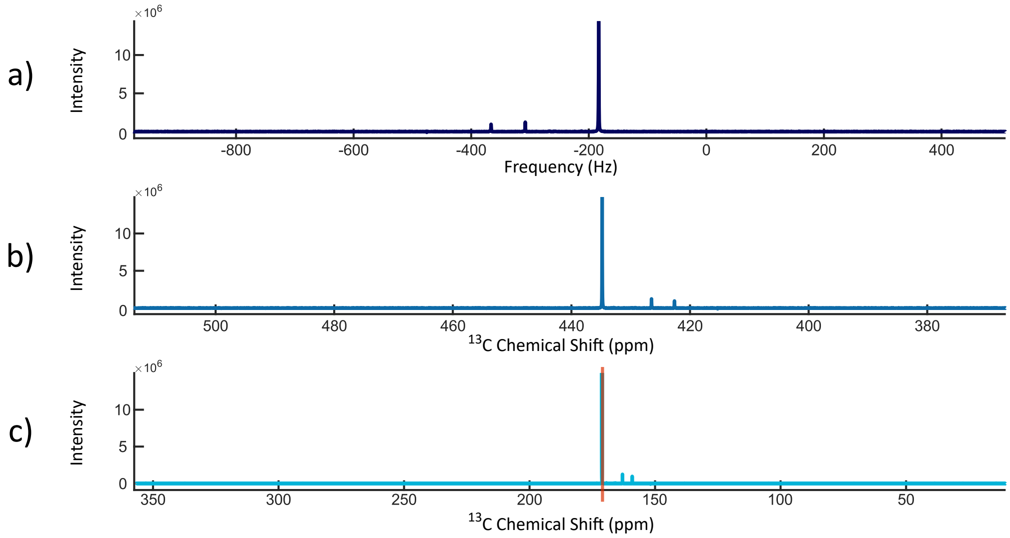


Figure 5.3: Process of chemical shift correction in NMR spectra. **(a)** Illustrates a spectrum with frequency (Hz) on x-axis and signal intensity on the y-axis. **(b)** shows the spectrum transformed into chemical shift representation, where the highest peak associated with $[1 -^{13} C]pyruvate$ is not centered at 171 ppm. **(c)** Shows the the corrected spectrum in chemical shift following adjustment, resulting in the peak precisely centered at 171 ppm, highlighted by the vertical orange line.

5.3 Phase and Baseline correction

Manual correction, although regarded as the gold standard, demands a profound understanding of zero and first-order components. Additionally, manual correction may require significant time due to high volume of spectra to correct. Also, manual correction may entail errors that vary depending on the user's experience level.

To address the challenges inherent in manual correction, automated algorithm have been developed.

In this chapter section, I introduce three automated phase correction algorithms utilized in this project. The first approach employs a strategy combining "coarse tuning" and "fine tuning", while the second focuses on minimizing the signal's entropy. Lastly, the third approach involves comparing the unphased signal with its absolute magnitude.

Additionally, in each method, I applied baseline correction after phase correction to rectify any baseline distortions that may have arisen from first-order corrections.

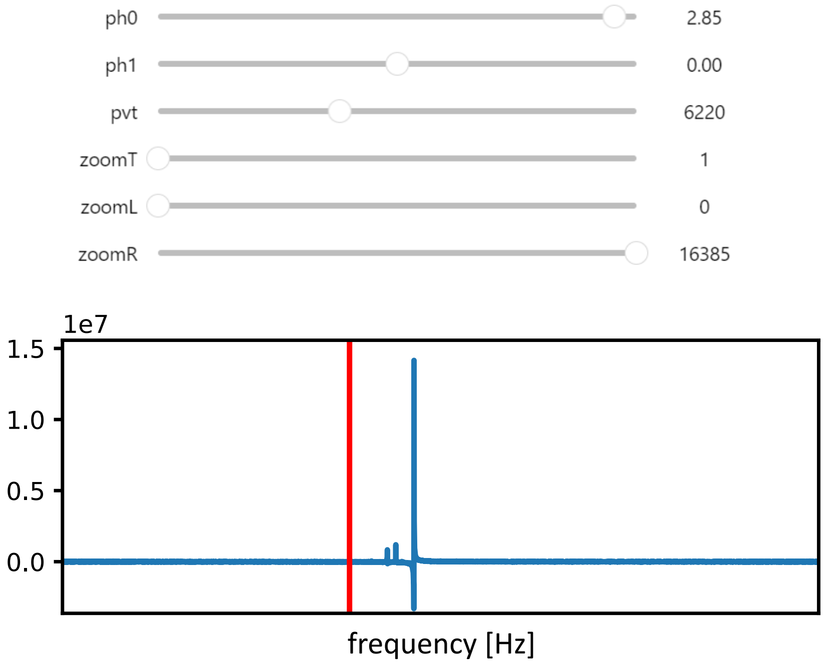
5.3.1 Manual Phase Correction

A developed algorithm enables manual phase correction of the provided spectrum. This algorithm incorporates an interactive function that automatically generates user interface controls for data exploration and interaction, as showed in Figure 5.4. Adjusting the cursors associated with *ph0* and ph_1 values will modify these values according to Equation 4.19. Additionally, adjusting a slider linked to the pivot parameter causes a red vertical line to shift.

In this study, the pivot was chosen as the highest peak, and the vector ω in Equation 4.19 was determined by:

$$\omega = \frac{range(-pivot, pivot + n)}{n} \tag{5.4}$$

Where *pivot* represents a frequency and n denotes the spectrum size.



ph0 2.85
ph1 0.00
pvt 6220
zoomT 1
zoomL 0
zoomR 16385

Figure 5.4: Interactive interface for manual phase correction, showing adjustable parameters (*ph0*, ph_1, *pvt*) and their effects on spectrum adjustment. Zoom functionalities (*zoomT*, *zoomR*, *zoomL*) allow users to examine specific spectral regions in detail.

In the initial phase correction step, the adjustment of the *ph0* involves moving a slider to modify the phase of the chosen pivot, typically the highest signal in our scenario. Should a ph_1 correction be required, one can adjust the slider relative to the pivot until the vertical line aligns with the selected pivot. Utilizing a visualization tool, we can zoom in on both the x-axis and y-axis to scrutinize the correction's impact meticulously. Utilizing a visualization tool, users can zoom in to observe the detailed effects of the correction on the data.

Following this phase correction, the baseline may become distorted because of ph_1 correction, demanding baseline correction.

5.3.2 Coarse and Fine Tuning method

The first automated phase correction method is based on the technique proposed by Qingjia Bao, et.al. [4], modified to better fit the available data. This method exploits two key differences between the absorption spectrum and the incorrectly phased signal: the tail peaks are shorter and more symmetrical in a correctly phased spectrum, and the spectrum does not contain negative peaks.

According to these criterion, the automatic phase correction implements two steps: coarse tuning and fine tuning.

Coarse tuning initially employs a baseline recognition method, as discussed further in Section 5.3.5, to locate the left and right tail ends of peaks, followed by a function to quantify the height differences between these ends. Subsequently, the spectra undergoes phasing by minimizing this function using a simple method.

After identifying signal-free baseline regions, the next step involves determining the signal area by excluding these baseline regions and marking the left and right ends as the start and end points of identified peaks.

Following this, the algorithm utilizes the height differences between the start and end points of each peak to construct an Objective Function (OF). This function is used to minimize deviations in order to find optimal values of ph_0:

$$\mathrm{OF} = \sum_{k=1}^{PR} \left| \Re\left(S_{\mathrm{b}}(s_k) \cdot e^{i\frac{\pi}{180}(\phi_0)}\right) - \Re\left(S_{\mathrm{b}}(e_k) \cdot e^{i\frac{\pi}{180}(\phi_0)}\right) \right| \tag{5.5}$$

Here, PR represents the number of recognized peaks, s_k denotes the start index of the k-th peak, and e_k signifies the end index. $S_{\mathrm{b}}(s_k)$ and $S_{\mathrm{b}}(e_k)$ are the data values at the start and end indices of the k-th peak, respectively. The length of the spectrum is denoted by L, and $\Re$ denotes the real part of the complex number.

This optimization process iteratively calculates a few points for each recognized peak, ensuring computational efficiency.

To adjust the zero-order phase shift using the preliminary tuning results, the phase data correction is computed as:

$$S_{CT} = S_b \, e^{i \cdot \frac{\pi}{180}(x)} \tag{5.6}$$

where S_b is the signal before the correction and x is the best solution founded through the minimization of the Equation 5.5, expressed in degrees. Subsequently, if the real part of S_{CT} at the maximum peak index is negative, performing the inversion of the entire array ensuring correct peak orientation.

After the initial 'coarse tuning' negative points might appear in the spectra, requiring further refinement through 'fine tuning'. This process devices a new Penalty Function (PF) based on the absence of negative point in the absorption mode spectra, except for those within negative or distorted peaks. This refined penalty functions aids in achieving more accurate phase corrections.

Before the next steps, to mitigate the impact of baseline distortion, the baseline recognition is necessary.

To establish this custom PF, it is important the categorization of the peaks into three classes: positive, negative and distorted. Qingjia Bao, et.al. [4] introduce a simple and effective method to categorize the peaks in the spectra after 'coarse tuning'. After recognition and categorization of the peaks, the spectrum undergoes further phasing by minimizing the custom PF:

$$PF = -\sum_{j=1}^{N} (\, S_{temp}(j) - |\, S_{temp}(j)\, |\,) \tag{5.7}$$

Where S_{temp} is the temporary spectrum calculated as:

$$S_{temp} = S_{CT} \cdot e^{i\phi_1 \cdot a_{num}} \tag{5.8}$$

It only contains first-order phase correction since 'coarse tuning' is effective for pivot phase correction.

The significant of formula 5.7 is straightforward: if a negative point appears in temporary spectrum, the square of that point is added to the penalty value.

A Genetic Algorithm (GA) determinates the optimum ph_1 value for minimizing the PF equation, as the optimisation algorithm used in the previous step encountered a local minimum and failed to provide the correct ph_1 value.

To correct the spectra using the fine tuning result, the best solution found through minimizing 5.7, denoted as Y, is applied:

$$S_{FT} = S_{CT} \cdot e^{i\,y\,\cdot a_{num}} \tag{5.9}$$

5.3.3 Entropy minimization method

The second automated phase correction approach relies on the technique proposed by Li Chen, et.al. [11], with slight adjustments made to adapt it more effectively to the available data. Claude Shannon introduced the concept of *entropy* as a quantitative measure of uncertainty [48].

In a correct Fourier transform NMR spectrum, the real part contains only non-negative spectral bands, in contrast, the imaginary part possesses both positive and negative values. Consequently, only the entropy of real parts of the phased spectra is considered in the objective function. Shannon introduced the following equation to measure the information uncertainty, called entropy S, of the probability distribution h:

$$S = -\sum_j h_j \ln h_j \tag{5.10}$$

Therefore, this equation is applied to the spectrum, defining the probability distribution h as:

$$h_i = \frac{|R_i|}{\sum_i |R_i|} \tag{5.11}$$

Where R_i is the spectrum's real part.

Entropy, closely linked to the region above the signal in a spectrum, yielding identical values for both negative and positive spectra. However, since the goal is to achieve a spectrum in absorption mode, entropy minimization should lead to the former scenario, where the final spectrum comprises solely positive values. Therefore, the OF, acting as a Shannon-type information entropy measure for phase correction, incorporates a PF. This function serves to prevent the occurrence of the latter scenario, where the spectrum is predominantly negative, by penalizing such instances.

The OF is given by the following equation:

$$min\left(-\sum_i h_i \ln h_i\right) \tag{5.12}$$

The PF operates on the principle that an increase in entropy value occurs if the maximum value of R_i differs from the maximum value of its absolute value. This adjustment ensures that the zero-order and the first-order phase correction factors are optimized effectively, mitigating the risk of the optimization algorithm producing inaccurate results.

The present contribution implements phase correction optimization using two GAs. The first genetic algorithm was employed to determinate x, the optimal ph_0 value, that minimized the objective function 5.12. Essentially, it seeks the ph_0 value that generates a new signal with minimized entropy when applied to the original signal. The optimal number found is used to correct the signal:

$$R_{ga_1} = S_b * e^{-ix} \tag{5.13}$$

As previously mentioned, entropy is associated with the region above the signal, emphasizing the importance of implementing baseline correction after each phase adjustment.

Subsequently, the second GA is employed with slightly different features. In this case, the aim is to find y, representing the optimal ph_1 value. Given the frequency-dependent nature of this correction, it is crucial to define a vector (a_{num}) closely tied to the pivot value, corresponding to the Equation 5.4. The final phase correction is then executed according to the following equation:

$$R_{ga_2} = R_{ga_1} * e^{-iy \cdot a_{num}} \tag{5.14}$$

5.3.4 Absolute spectrum method

The third method for automatic phase correction is based on the principle that a spectrum in absorption mode contains only positive values. A similar signal exhibiting these characteristics is the absolute value of the spectrum. Thus, the algorithm determines the value of ph_0 and ph_1 by comparing the phase-corrected signal with the absolute value of the phase-incorrected signal.

To evaluate the correction, the objective function employs a similarity metric, namely the Mean Square Error (MSE) defined as:

$$mse = \frac{1}{N} \sum_{j=1}^{N} (R_{abs_j} - R_j)^2 \tag{5.15}$$

Here, N is the sample number, R_j represents the j-th element of the real spectrum, and R_{abs_j} denotes the j-th element of the real part of the absolute value of the spectrum, calculated as:

$$R_{abs} = real(\sqrt{R_b^2 + I_b^2}) \tag{5.16}$$

Where R_b and I_b are respectively the real and imaginary parts of the signal before phase correction. As in the previous method, penalizing signals with negative points is important, accomplished by employing the follow penalty function implemented by Li Chen, et.al. [11]:

$$P(R_i) = \gamma \left[\sum_i F(R_i) R_i^2 \right] \tag{5.17}$$

Here, γ, set to *0.1*, is a penalty factor that balances the contributions of entropy and penalty parts. The function F is defined as:

$$F(y) = \begin{cases} 0 & y \geq 0 \\ 1 & y > 0 \end{cases} \tag{5.18}$$

The objective function is expressed as:

$$min \left(\frac{1}{N} \sum_i (R_{abs_j} - R_j)^2 + P(R_i) \right) \tag{5.19}$$

Optimization of phase correction is achieved using two genetic algorithms (GAs). The first GA determines the optimal ph_0 value, denoted as x, that minimizes the objective function 5.19. This value generates a new signal with minimized MSE between the spectrum with zero-order phase correction and the absolute value of the signal before correction calculated as in Equation 5.16. The obtained optimal number is then used to correct the signal:

$$R_{ga_1} = S_b * e^{ix} \tag{5.20}$$

Subsequently, the ph_1 correction is applied following a similar principle. However, since baseline distortion may occur after ph_1 correction, the objective function is applied considering the signal with the baseline corrected. This choice is made because first-order phase correction may result in a distortion of the baseline, as showed in Figure 5.5. Without baseline correction, the smallest MSE value might correspond to an incorrect ph_1 value. This happens because the resulting spectrum's baseline may be closer to the absolute value of the signal, even if the phase correction is wrong. Since non-peak point outnumber peak points, the baseline values disproportionately affect the MSE value. Consequently, the algorithm may converge to a suboptimal solution. After baseline correction, the optimal value of ph_1, denoted as y, is determined, and the final phase correction is executed using the following equation:

$$R_{ga_2} = R_{ga_1} * e^{iy \cdot a_{num}} \tag{5.21}$$

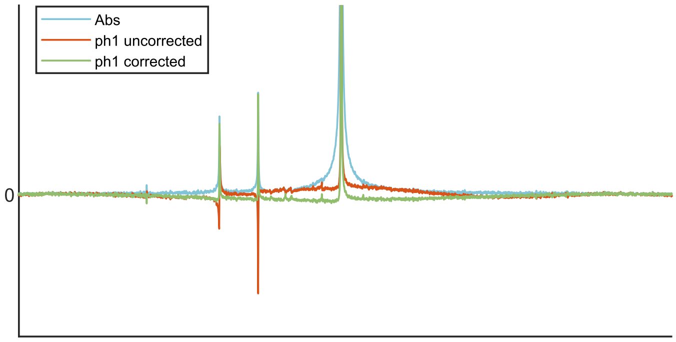


Figure 5.5: Baseline distortion across varying values of the first-order correction ph_1. The absolute signal intensity is shown in blue, while the red spectrum illustrates an uncorrected ph_1 value, evident from the presence of negative peaks. The green spectrum represents a corrected ph_1 value, albeit with a significantly distorted baseline compared to the red and blue spectra.

5.3.5 Baseline recognition and Correction

A baseline correction is applied to the output of the phase correction (shown in Figure 5.6.a). For this purpose, a specialized function was designed by Qingjia Bao, et.al. [4] for baseline recognition and peak detection in spectral data, employing a sliding window method and threshold application.

The function requires three inputs: real spectral data, a noise level scaling factor set to 6, and a Continuous Wavelet Transform (CWT) filter factor set to 0.001. Its output is a structure array containing peak information [4].

The process begins with the initializing and preparing the spectral data, where the real part is extracted and stored, and its size is determined. Subsequently, the function performs a Continuous Wavelet Transform on both the real and imaginary parts of the spectrum using the *Haar wavelet.* The results have been combined and the absolute value applied obtaining the absolute derivative spectra as showed in Figure 5.6.b.

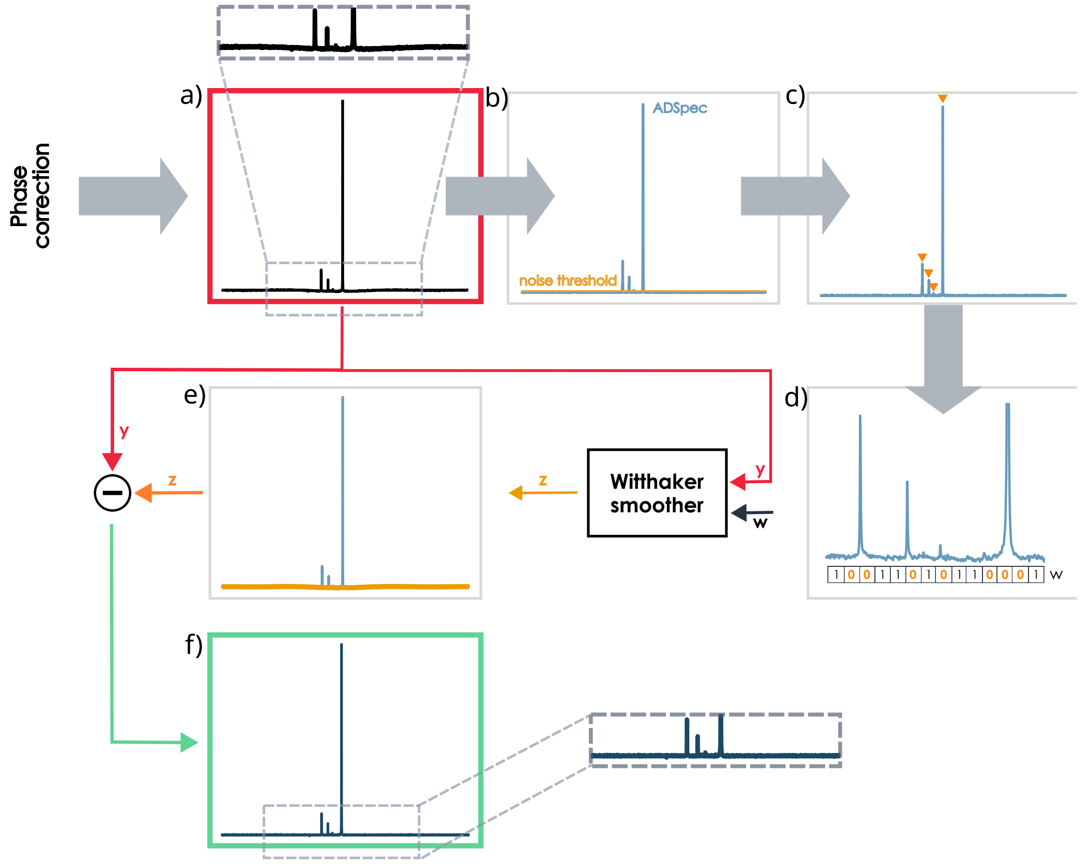


Figure 5.6: Algorithm scheme for baseline recognition and baseline correction

The noise level is then calculated using one of two methods. In the first method, noise is computed based on the first segment of the data, while the second method divides the data into sixteen segments, calculates the standard deviation for each segment, and selects the minimum value as the noise level. The chosen noise value is then scaled by the noise factor parameter.

Thereafter, a sliding window traverses the spectrum. Its specified width is typically one-thousandth of the spectrum's total width. The height is calculated as the difference between the maximum and the minimum values of the signal within the specified window width. For each window, comparison to the scaled noise level is conducted. Points within the window are classified as either baseline or signal based on this comparison. If the window height exceeds the noise level, a potential peak is detected, and the peak start point is recorded. Conversely, if the window height is below the noise level, the end of a peak is recorded.

Detected peaks (Figure 5.6.c) are stored in a structure array, which includes start, end, and maximum slope information. These peaks undergo filtering based on their length, retaining only those that meet a minimum length criterion (set to 10 samples, this number reflects empirical determination.).

Consequently to the baseline recognition, the baseline correction function is designed to correct baseline variations in spectral data by identifying and excluding peak regions. To achieve this, a binary weights vector, with a length matching that of the spectrum, is generated based on the peaks identified through the previous procedure. Samples corresponding to the peaks are assigned a value of 0, while those corresponding to the baseline are designated a value of 1 (Figure 5.6.d).

Employing the Wittaker smoother (discussed in more detail in Section 4.3.1), the function constructs a sparse matrix to incorporate the exclusion zones identified by the peaks. In Matlab, the implementation of baseline correction is facilitated by the availability of built-in functions such as *diff()* for computing the first derivative matrix *D*. For shorter data series, typically comprising less than 1000 points, direct computation of the baseline using dense matrices is viable. However, beyond this scale, computational time and storage requirements escalate significantly. To mitigate this issue, sparse matrices are utilized to reduce memory consumption and computation overhead [14]. With sparse matrices, only the nonzero elements of the identify matrix are stored.

The matrix undergoes *Cholesky decomposition* to compute the baseline (Figure 5.6.e). The selection of Cholesky decomposition over direct solution method is predicated on its suitability for spare systems and its avoidance of unnecessary bandwidth optimization, which can be time-intensive for extensive dataset [14]. In instances where peak detection fails or encounters errors, the baseline is set to zero. Subsequently, the baseline-corrected data is derived by subtracting the estimated baseline from the original data (Figure 5.6.f).

This methodology guarantees efficient and precise baseline correction, even for sizable spectral datasets, without compromising computational efficiency or memory usage [14].

5.4 Simulated spectroscopic data Generation

To train the proposed network and compare it with other noise reduction methods, I developed a simulated dataset using Python. This dataset simulates NMR spectroscopy data from the Pulsar benchtop from Oxford instrument [40], considering signals already in the correct phase. Initially, the process generated noise-free signals, then introduced random noise to these pure signals (Figure 5.8).

Each generated signal has a duration of 10 seconds and contains 5120 samples, making it easily divisible into segments of 512 samples, which is important for network training as already mentioned in Section 5.5.1. The starting point is Equation 3.16, in which the adjustable parameters include the number of peaks n, magnitude S_0, relaxation time T_2, and Larmor frequency f_0. These parameters closely reflect real data characteristics, featuring very high peaks next to very small peaks and narrower linewidths, reflecting varying amplitude values S_0 and high relaxation times T_2.

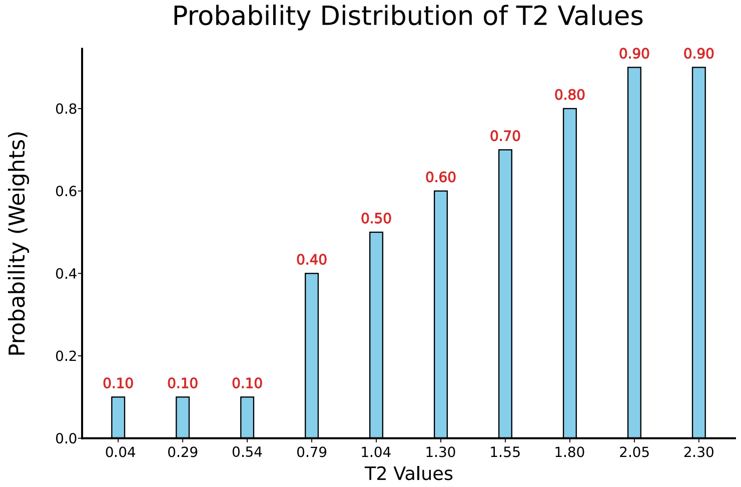


Figure 5.7: Chosen T2 values used for generating synthetic data to closely emulate real signals, each value being assigned an increasing probability.

Real signals predominantly feature low peaks. In the simulation, for a given FID signal and number n of nuclei, 80% of the nuclei received an amplitude value between 5 and 50 to replicate the prevalence of small peaks, while 20% received a value between 300 and 500.

To ensure the synthetic signals closely represent real data, the selection of the T_2 parameter required careful consideration. Since real signals typically feature small amplitude peaks, the simulation favored higher relaxation times. Section 3.3.2 discusses the relationship between amplitude and relaxation time in detail. The method assigns probabilities to each T_2 value (x-axis in Figure 5.7) using a weight vector. This approach increases the selection likelihood for larger T_2 values by providing higher probabilities (y-axis).

Consequently, the number of nuclei in the sample is a random value between 2 and 10 to produce spectra with varying numbers of peaks. The Larmor frequency has a value randomly selected from an array of frequencies ranging from -400 to $+400Hz$.

The calculation of the Fourier transform for each FID signal produced and stored the spectra (Figure 5.8.a).

Afterward, I add noise with a normal (Gaussian) distribution N to the pure signals, following the additive noise model as described in Equation 5.22. This process is characterized by the following equation:

$$S_{noisy}(t) = S_{pure}(t) + noise\ factor \cdot N(\mu_n, \sigma_n) \tag{5.22}$$

For this signal, selecting three parameters is necessary: the mean value of the distribution (μ_n, the standard deviation ((σ_n, which must be non-negative), and the dimension of the noise.

Initially, the mean of the distribution is set to 0, with the standard deviation randomly chosen between 50 and 500, and the length matching that of the spectrum

Additionally, another crucial factor is the *noise factor*, which indicates the amount of noise added to the original signal. This factor plays a significant role in simulating real-world conditions where signals are often corrupted by various sources of noise. Therefore, the noise factor can has a random value in a range between 3 and 8.

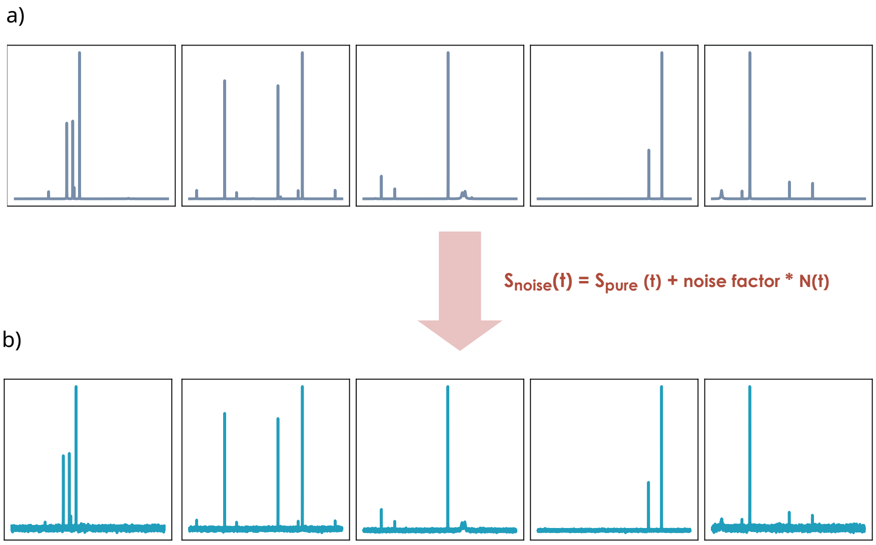


Figure 5.8: Simulated spectroscopic data: **(a)** noise-free synthetic spectra, and **(b)** spectra affected by Gaussian noise with a zero mean (μ_m) and standard deviation σ chosen from the range between 50 and 500.

5.5 Autoencoder for Spectral Noise Reduction

This section explores the application of Autoencoders for noise reduction in spectral data. It details the methodology, including data pre-processing, model implementation, and training processes. The following subsections discuss the essential steps involved in preparing the data, constructing the Autoencoder architecture, and the specifics of the training process to achieve optimal noise reduction performance.

5.5.1 Data Pre-processing

Before training the Autoencoder for denoising spectra, several pre-processing steps are essential to prepare the data for input into the neural network. These steps include normalization, segmenting the data, and reshaping it to fit the requirements of TensorFlow/Keras (Python), each step playing a crucial role in optimizing the neural network's performance and efficiency.

Normalization is the first step, scaling the data to fall within a specific range [0,1]. This ensures consistency in the input scale, allowing the network to learn more effectively and converge faster. For both signal, noisy and pure, normalization is achieved by dividing the signal by its maximum value, optimizing the loss function used, the MSE. This step is essential because input data can vary significantly in range, rendering the incomparable without normalization. For instance, one feature x_1 might range from 10 and 50, while another feature x_2 ranges from 1000 to 5000. These disparities lead to different value scales, causing varied weight updates and optimization steps, which can distort the shape of the loss function. Consequently, a lower learning rate would be necessary to prevent overshooting, resulting in a slower learning process. Thus, normalization stabilizes and enhances the optimization process [3].

The dataset initially comprised dimensions of (10.000, 5120). To process the data more effectively, each signal is segmented into smaller segments of 512 samples, resulting in a new dataset with dimension of (100.000, 512). This segmentation is advantageous for several reasons. Firstly, smaller segments reduce the computational load on the network, making the process more efficient. Additionally, neural networks often perform better when the input size is a power of 2, aligning well with the architecture of the underlying hardware (e.g., GPUs), leading to optimized memory usage and faster computations.

After segmenting, the data undergoes reshaping by adding an extra dimension to represent the number of channels, transforming the dataset size to (100.000, 512, 1). This additional dimension indicates a single channel for one-dimensional signals, similar to working with grayscale images [51].

The next step involves dividing the reshaped dataset into training and validation sets, with 70% allocated for training and 30% for validation. Another dataset of the same initial size serves as the test set, following the same pre-processing procedures as the training set.

5.5.2 Autoencoder Architecture

Figure 5.9 illustrates the Autoencoder architecture implemented in this study. The encoder component of the model compresses the input signal into a lower-dimensional representation through successive Conv1D layers. The input layer accepts the input signal with a shape of (512, 1). Subsequent Conv1D layers increase the number of filters (32, 64, 128, 256, 512, and 1024), use a kernel size of 3, stride of 2, and 'same' padding to preserve the input size. Each Conv1D layer employs the ReLU activation function (mentioned in Section 4.5.1), and includes both kernel constraint and a kernel intializer.

Kernel constraints keep kernel weights from becoming excessively large during optimization, which can lead overfitting and numerical instability [6]. The Max Norm which constraints used in this study restricts the maximum norm of the weights as follows:

$$|| \, W \, || \leq c \tag{5.23}$$

where c is set to 2.0.

The neural network requires an initial set of weights that are iteratively updated. Kernel initialization sets these initial values using a statistical distribution or function. He Normal initialization is particularly effective with ReLU activation function, addressing issues such as the inactivation of ReLU neuron. He Normal initializes the weights with values draw from a normal distribution with a mean of 0 and a standard deviation of $\sqrt{2/n}$, where n is the number of neurons in the previous layer. This method helps prevent vanishing or exploding gradients, improving stability and convergence during training [32].

Each layer's output passes through a LeakyReLU activation with a small negative slope (0.003) to allow a gradient when the units inactive.

Some layers incorporate skip connections, where output from an earlier layer is added to a later layer's output.

The bottleneck layer, representing the compressed latent space, is the flatted output of the last convolution layer and applying a dense layer with 500 neurons and ReLU activation, capturing the input signal's essential features compactly.

The decoder mirrors the encoder, using transposed convolutional layers to upsample the data back to its original dimensions. Each transposed layer corresponds to an encoder layer, reversing the encoding process. The first transposed layer applies 1024 filter, with the final one applying a single filter to match the original input dimension. Batch normalization layers follow the Leaky ReLU activations in the decoder to standardize inputs and accelerate training.

The decoder also employs skip connections, combining high-level features from the encoder with upsampled features to aid in reconstruction.

The output layer uses the sigmoid activation function, suitable for reconstructing normalized input data.

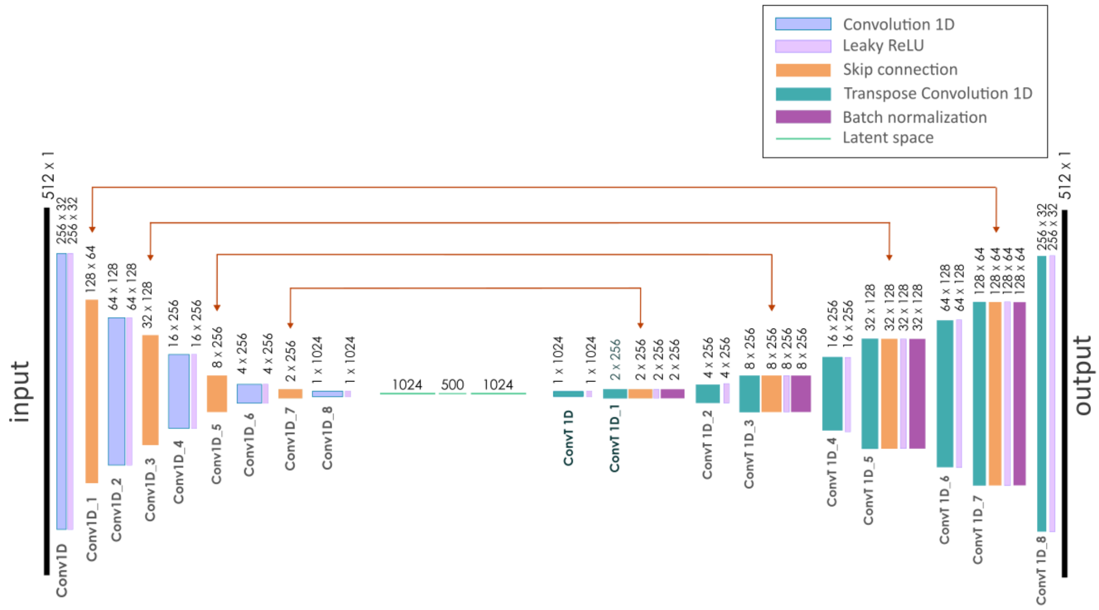


Figure 5.9: Autoencoder Architecture implemented in this study. The encoder compresses the input signal into a lower-dimensional representation using Conv1D layers with increasing filters (32 to 1024), ReLU activation, kernel constraints, and He Normal initialization. LeakyReLU activations (slope 0.003) maintain gradient flow. Skip connections and a bottleneck layer with a dense layer capture essential features. The decoder mirrors the encoder with transposed convolutions, batch normalization, and skip connections for reconstruction. Output uses sigmoid activation for normalized data.

5.5.3 Training the Autoencoder

The training process involved the compilation and fitting of the model using the Adam optimizer, set with a learning rate of 1e-7, and the mean squared error (MSE) loss function.

During this phase, the model underwent training with the noisy and pure signals input for a total of 20 epochs. Preliminary experiments show that model performance plateaus after this number of epochs, with no significant improvements in validation loss beyond this point, leading to the selection of this specific number of epochs. A batch size of 32 facilitates the training process, while reserving a portion of the dataset for validation monitors the model's performance and generalization capabilities throughout the training period. This approach ensured an optimal balance between training efficiency and model accuracy, preventing overfitting while maximizing the model's ability to learn from the data.

5.6 Alternative Denoising Methods

5.6.1 SWT Denoising Technique

In implementing the method discussed in Section 4.5.2, the process starts by calculating the noise threshold for each decomposition level using a region-specific binary spectrum that distinguishes peak regions from noise. Following this calculation, the wavelet denoising function applies the SWT up to k decomposition level, typically ranging from 5 to 7 for most NMR spectra using the modified threshold method with $\alpha = 0$. In this study, both levels underwent evaluation to determine their respective performances. Finally, the function reconstructs the denoised spectrum using the ISWT with the 'bior2.4' wavelet [2].

5.6.2 Rolling Window Technique

The implementation of a Moving average, discussed in Section 4.5.3, is straightforward. This technique uses a window of a specific size to perform a mathematical operation on the signal. The window size, a key parameter, determines the number of observations used for the average calculation. In this study, I used a simple function in Pandas to apply this filter and denoise the signal.

Two critical parameters for this function are the window size and the window type. The window size determines the extent of noise reduction; a larger window results in greater noise reduction but also increases the loss of information. Additionally, selecting the appropriate window type is essential. The simplest type is a uniform window, were all points have equal weight. Other types include the Gaussian, Barlett, Hamming, and Parzen windows, each of which may require additional parameter settings. For example, the Gaussian window necessitates specifying a standard deviation.

In the Rolling Window technique, I evaluated two types of moving average windows: a simple rolling window and a Gaussian window. Both windows had a length of 7, with the

Gaussian window having a standard deviation of 3. This evaluation aimed to determine the effectiveness of each window type in denoising the signal.

Chapter 6

Results and discussion

In this chapter, I present the outcomes of applying various signal processing techniques to NMR spectra, highlighting the advantages and limitations of each method. The chapter is divided into sections focusing on phase correction methods and denoising techniques.

Overall, this chapter provides a comprehensive comparison of different signal processing techniques for NMR spectra, highlighting the advantages and limitations of each method.

The initial section evaluates and benchmarks three automatic phase correction methods: combined coarse and fine tuning (CFT_{ph}), inspired by Qingjia Bao et al. [4]; entropy minimization (H_{ph}), based on the work of Li Chen et al. [11]; and absolute similarity maximization (Abs_{ph}), developed and implemented specifically for this study. Each method underwent refinement to better align with the objectives.

I assess these methods using metrics such as correlation coefficients, Euclidean distances, and root mean square errors (RMSE) to determine their accuracy compared to manually corrected references. I provide a detailed analysis of each method's performance, considering their strengths and weaknesses in handling different types of NMR signals. The second section explores various denoising techniques applied to synthetic and real NMR data. I assess the performance of the Autoencoder ($Auto$), stationary wavelet transform (SWD), and rolling window smoother (RWD) methods. The evaluation is based on metrics such as Structural Similarity Index (SSIM), correlation coefficient, and Euclidean distance. I also investigate how these methods perform under different noise levels and compare their effectiveness in reducing noise in real NMR spectra. Through visual representations and detailed analysis, I identify the Autoencoder as the most effective method, with SWD and RWD demonstrating varying degrees of effectiveness.

6.1 Performance Assessment of Phase Correction Methods for NMR Spectra

In this section, I evaluate and benchmark three automatic phase correction methods for NMR signals: combined coarse and fine tuning (CFT_{ph}) detailed in Section 5.3.2; entropy minimization (H_{ph}) explained in Section 5.3.3; and absolute value maximization (Abs_{ph}) discussed in Section 5.3.4. Figure 6.1 presents an example of the output from these three phase correction methods, comparing them to the signal before phase correction and the signal with manual phase correction.

The chosen metrics provide insights into the methods' performance across different NMR signals types. Non-hyperpolarized samples, with inherently varying signal-to-noise ratios (SNR), can lead to differences in metric values, potentially skewing comparisons. Normalizing signal amplitudes ensures comparability without altering underlying noise characteristics. To mitigate the issue of noise remaining constant post-normalization, evaluations focused on signal regions containing prominent peaks, where the true signal information is concentrated, rather than across the entire spectra.

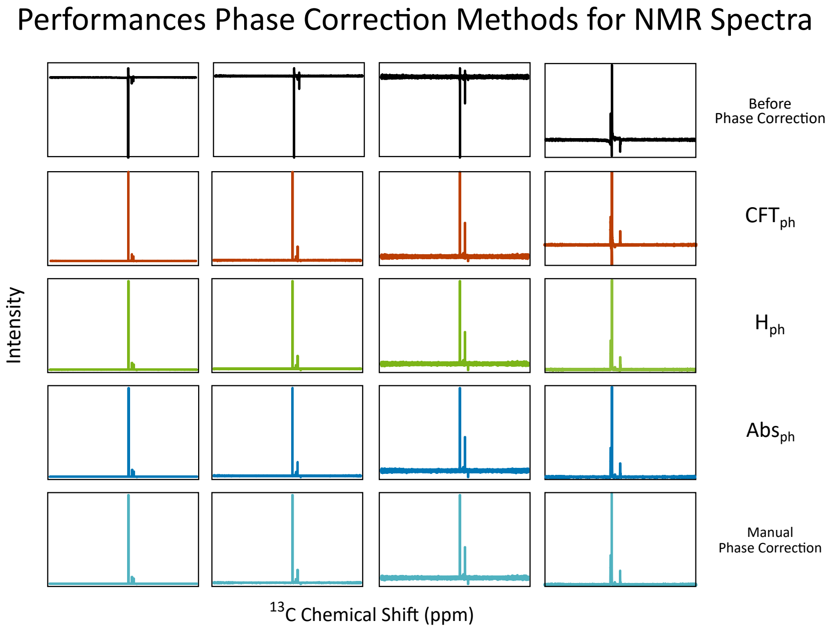


Figure 6.1: Performance of three phase correction methods for NMR spectra: combined coarse and fine tuning (CFT_{ph}), entropy minimization (H_{ph}), and absolute value maximization (Abs_{ph}). These methods are evaluated against signals prior to phase correction and signals corrected manually.

6.1.1 Evaluation of Phase Correction Method Performance

The correlation coefficient, analyzed via box plots in Figure 6.2, indicates the linear relationship between automatic and manual corrected signals. The x-axis represents the three different phase correction methods, while the y-axis shows the correlation coefficient values, indicating the degree of linear relationship between the phase-corrected signal and the ground truth obtained by the manual correction. Each box plot includes the median (central line), interquartile range (IQR, the box) and potential outliers (points outside the whiskers).

A correlation coefficient value close to 1 suggests strong linear relationships. The three methods in question demonstrate a linear correlation for most tested signals, as evidenced by a median value close to 1 for all of them. However, the CFT_{ph} method shows greater variability in the results, denoted by the larger IQR. The H_{ph} and Abs_{ph} methods exhibit comparable performance, as evidenced by the similarity between their box plots when evaluating the correlation coefficient. Both methods have a high median value and relatively narrow IQR, suggesting consistency and low variability in performance.

All three methods produce values that deviate more from the mean correlation coefficient, with outliers primarily due to signals that are challenging to correct. However, the H_{ph} and Abs_{ph} methods, perform better on a wider range of spectra, as evidence by fewer outliers and closer to the median value.

Upon evaluating the Euclidean distance through heatmap visualization (Figure 6.3), the findings indicate varying degrees of proximity between phase spectra and reference spectra across different experiments. A heatmap visualizes data using colors to represent values, providing a graphical representation of information [21]. In this context, each cell within the heatmap denotes the Euclidean distance calculated between the reference signal (manually corrected) and the signals phased by the three methods across various experiments. The color gradient ranges from dark blue to dark red, indicating ascending Euclidean distance values.

The CFT_{ph} technique exhibits variability in Euclidean distances across experiments, indicating moderate performance variance and occasional deviation from the reference spectrum, as observed from a mixture of blue and dark red shades. This variability suggests that in certain experiments, CFT_{ph} yields higher Euclidean distances compared to the H_{ph} and Abs_{ph} methods, which consistently demonstrate lower Euclidean distances indicated by uniform lighter shades. These methods consistently perform well across experiments, characterized by lower Euclidean distances, reflecting minimal variance and closer alignment with the reference spectra.

In Figure 6.4 the RMSE values depict the dissimilarity between signals, specifically between the output signals of automatic phase correction and the reference spectrum. This visualization enables as assessment of how these metrics vary across different correction method. The correlation matrix highlights both the self-correlation of each method and its correlation with others, represented on the x- and y- axes by their respective RMSE values.

Upon examining these metrics, it become evident that the CFT_{ph} method exhibits poorer performance compared to the other two, as indicated by higher RMSE values, reaching a maximum of 0.601, as shown in the first plot on the diagonal. This suggest a greater discrepancy between automatically corrected signal and the manually corrected reference. Conversely, the H_{ph} and Abs_{ph} methods consistently demonstrate better performance, with RMSE values peaking around 0.16. Notably, the main diagonal of the correlation matrix reveals lower RMSE values for both, underscoring their stability and reliability across experiments. Furthermore, the correlation coefficient r between H_{ph} and Abs_{ph} is notably higher at 0.62, indicating a stronger similarity in RMSE trend compared to other method pairs. This implies that the performance consistency between H_{ph} and Abs_{ph} is more pronounced relative to the first method.

Figure 6.5 illustrates the solution x that minimizes the Equation 4.26. This equation achieves its minimum then the result is 0, indicating equality between A and b with x equal to 1. The degree of similarity between A and b directly influences how closely x approaches 1. In our context, A represents the automatically corrected signal and b the manually corrected reference. As observe from previous metrics, the CFT_{ph} method exhibits slightly lower performance compare to the other two methods, suggesting that A and b may diverge more across different experiments, leading x to deviate from 1. Indeed, in the Figure 6.5, the CFT_{ph} method generally performs similarly to H_{ph} and Abs_{ph} , but typically yields a lower x values the other methods. In contrast, considering this metric, the H_{ph} and Abs_{ph} methods exhibits slightly different trends, particularly in challenging experimental scenarios. The results reveal that CFT_{ph} often shows more variable and occasionally less accurate corrections compared to H_{ph} and Abs_{ph}.

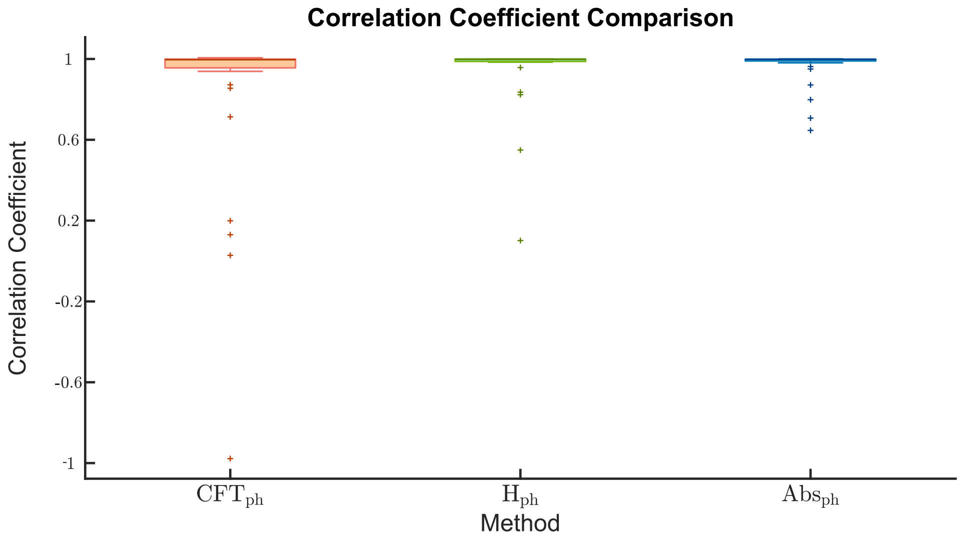


Figure 6.2: Correlation coefficient values computed between spectra corrected using three methods (CFT_{ph}, H_{ph},Abs_{ph}) and manually phase corrected signals.

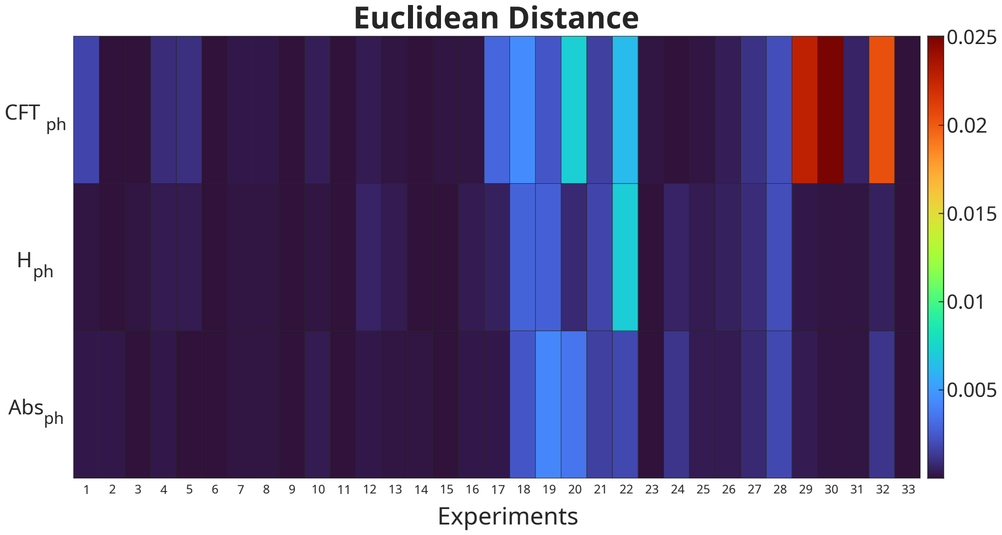


Figure 6.3: Heatmap visualization of Euclidean distances (L2 norm) computed between phase corrected spectra using three methods(CFT_{ph}, H_{ph},Abs_{ph}) and a reference spectrum (manually corrected).

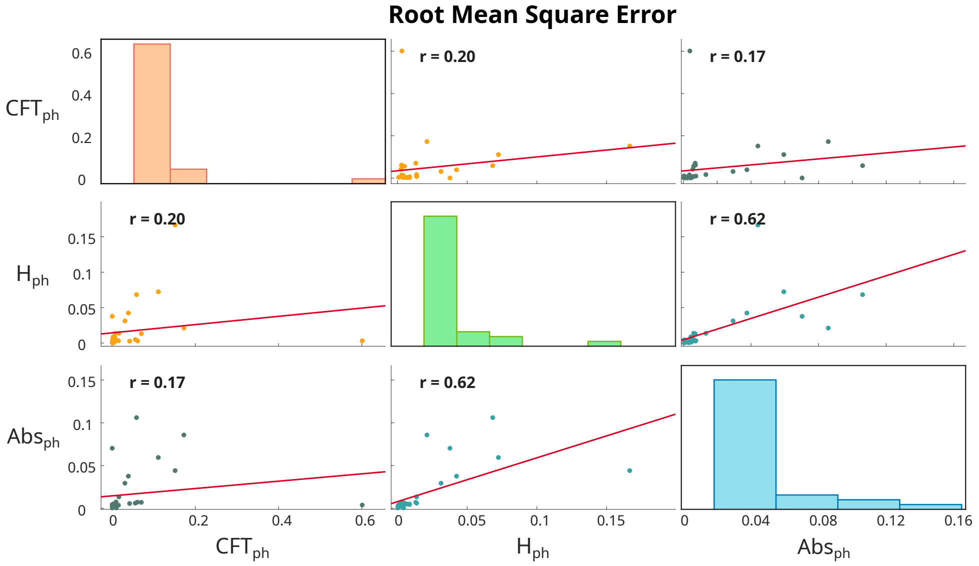


Figure 6.4: RMSE (Root Mean Square Error) values that quantify dissimilarities between automatic phase-corrected signals and a reference spectrum. This visualization allows assessment of these metrics across different correction methods. The correlation matrix depicts both the self-correlation of each method and its correlation with others, shown on the x- and y-axes by corresponding RMSE values.

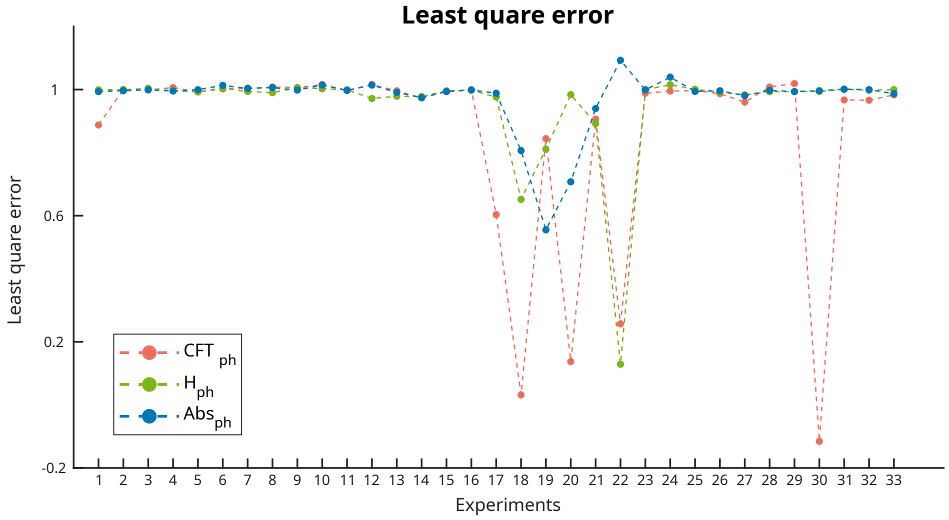


Figure 6.5: Solution x minimizing Equation 4.26. The degree of similarity between automatically corrected signal, A, and manually corrected reference, b. Solution for the methods: CFT_{ph}, H_{ph},Abs_{ph}

6.1.2 Discussions

Based on the results of using different phase correction methods on NMR spectra, it is evident that these strategies yield varying performance outcomes due to their inherent methodologies.

The coarse and fine tuning (CFT_{ph}) approach involves identifying and classifying peaks as positive, negative, or distorted during the fine-tuning before applying the penalty function in Equation 5.7. This function targets positive or negative peaks by adding the square of each negative peak's value, while distorted peaks have their values set to zero, thereby avoiding penalties [4]. Avoid distorted peaks can lead to incorrect ph_1 phase choices during the optimization of the objective function. In contrast, entropy minimization (H_{ph}) and similarity maximization with the absolute spectra (Abs_{ph}) apply a penalty function that indiscriminately penalizes all negative points in the spectrum.

Despite the CFT_{ph} method's lower performance compared to H_{ph} and Abs_{ph}, it operates significantly faster. This speed advantage primarily arises from CFT_{ph} using recognized peaks rather than all data points when performing the objective function in Equation 5.5. Additionally, the efficiency comes from employing optimization algorithms to find the optimal values ofph_0 and ph_1, incorporating both global and local optimizers.

Local optimization algorithms start with a randomly generated hypothesis and subsequently optimize it using a greedy algorithm [27]. These algorithms make locally optimal choices at each step, aiming to find a global optimum solution. However, decisions are based solely on current information without considering future implications [18]. The local optimization process generates an initial hypothesis randomly and repeats this process multiple times, each time with a different randomly generated hypothesis [27]. Consequently, local optimizers often find the nearest minimum, which is typically a local minimum unless the starting point is exceptionally well-chosen.

Local optimization techniques struggle with global optimization problems. They frequently become trapped in local minima and cannot generate or utilize the global information needed to find the global minimum for functions with multiple local minima. In contrast, the genetic algorithm (GA) addresses optimization problems by mimicking biological evolution principles. It repeatedly modifies a population of individual points using rules modeled on gene combinations in biological reproduction. Due to its stochastic nature, the genetic algorithm increases the chances of finding a global solution by cleverly sampling the parameter space to approach the optimum [37].

Although both algorithms may spend considerable time around a reached minimum point [19], genetic algorithms tend to be slower due to their structural complexity and the intricate genetic operators involved, which contribute to slow computational speeds [16]. However a local optimizer, although faster, may provide suboptimal solutions as it stops at a local extremum. This speed does not guarantee accuracy and may yield incorrect values by not finding the global minimum. Hence, a global optimizer is preferred for achieving the global minimum (or maximum). The CFT_{ph} method employs a local optimizer for *ph0* and a global optimizer for ph_1, while H_{ph} and Abs_{ph} use a global optimizer (GA) for both phase correction values, resulting in increased computational time.

The quality of the results and the convergence speed of genetic algorithms can be significantly influenced by algorithm parameters, particularly the initial population of solutions [52]. Genetic algorithms typically consist of two main processes: selecting individuals for the production of the next generation and manipulating these selected individuals through crossover and mutation techniques to form the new generation [45]. A large difference between the initial population and the next generation can lead to prolonged times for the genetic algorithm to reach the optimal solution [45]. Optimizing the crossover and mutation processes, along with adaptive parameter adjustments, can help minimize the differences between consecutive populations. However, it is essential to ensure that the difference is not too small to avoid local optima [45] [29]. Optimizing crossover and mutation processes can prevent local optima and ensure feasible mutated paths, ultimately improving search efficiency and convergence speed [29], thereby increasing the overall efficiency of automatic phase correction methods in NMR spectroscopy.
Another factor impacting computational time is baseline correction. Baseline correction involves recognizing where there is only noise in the spectrum and identifying and labeling peaks. For long spectra, and with baseline recognition based on a sliding window, this process can significantly affect computational time, especially if performed repeatedly. In this study, the automatic methods for phase correction involve two steps: the first step searches for the optimal value of ph_0, and the second searches for the optimal value of ph_1. Baseline correction primarily occurs at three points: after applying the optimal ph_0 value, during the iterative process to find the optimal ph_1 value (applying baseline correction for each tested phase value), and finally, after applying the optimal ph_1 value to obtain the correctly phase final spectrum. To minimize computational time, the iterative search for the optimal ph_0 value excludes baseline correction. As discussed in Section 5.3.5, only the first-order phase correction affects the baseline, potentially distorting it. If not corrected at each iteration, this distortion could lead to selecting a suboptimal ph_1 value. SInce ph_0 correction does not impact the baseline, applying baseline correction at each iteration for ph_0 is unnecessary and time-consuming.

6.2 Performance Assessment of Denoising Methods for NMR Spectra

In this study, I applied various denoising methodologies, namely Autoencoder (*Auto*), Stationary Wavelet Transform (*SWD*), and Rolling Window Smoother (*RWD*), as discussed in Section 4.5, to synthetic noisy NMR signals and real NMR spectra. Figure 6.6 illustrates an example of the output signals of these three methods, compared with the pure signal and noisy signals at varying noise levels.

Before comparison, I normalized the output signals from the rolling window and SWD methods, as well as the ground truths, to ensure comparability. The Autoencoder is already configured to provide normalised signals.

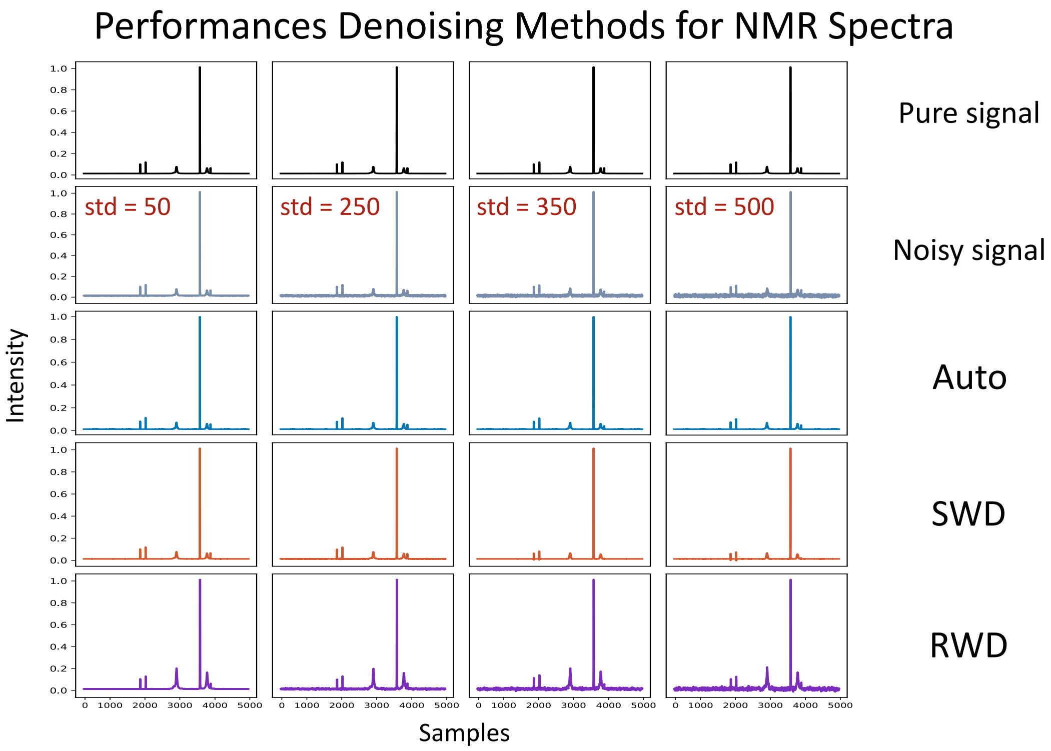


Figure 6.6: Performance of three denoising methods for NMR spectra: Autoencoder (*Auto*), Stationary Wavelet Transform (*SWD*), and Rolling Window Smoother (*RWD*). These methods are evaluated against pure signals and noisy signal with varying noise levels.

6.2.1 Evaluation of Noise Reduction Method Performance

The evaluation metrics used are Structural Similarity Index (SSIM), Correlation Coefficient, and Euclidean distance, calculated across 10,000 synthetic spectra.

Firstly, focusing on the SSIM scores (Figure 6.7), the Autoencoder method demonstrates the highest median SSIM value, indicating superior performance in preserving structural similarity. This is supported by its narrow interquartile range (IQR) and few outliers, suggesting consistent and reliable denoising results. The SWD method also shows a high median SSIM value but with a wider IQR and more outliers, implying some variability in its performance. In contrast, the RWS exhibits the lowest median SSIM value and the widest IQR, indicating less consistent performance and more frequent deviations from the ground truth.

Moving to the correlation coefficient results (Figure 6.8), the Autoencoder method maintains a high median value, indicating a strong linear relationship between the denoised signals and the ground truth. The SWD method performs similarly but with slightly less consistency, as evidenced by a marginally lower median correlation coefficient. The RWS again shows the lowest median correlation coefficient among the methods, suggesting weaker linear alignment with the ground truth.

In terms of Euclidean distances (Figure 6.9), which reflect how close the denoised signals are to the ground truth in a spatial sense, the Autoencoder exhibits the lowest median Euclidean distance, indicating closer proximity to the noise-free signals. This is complemented by a narrow IQR, indicating consistent performance across different data points. Conversely, the SWD method shows a higher median Euclidean distance and a wider IQR, suggesting less accurate denoising and greater variability. The rolling window smoother displays the highest median Euclidean distance and the widest IQR, highlighting the least accurate denoising performance with considerable variability and numerous outliers.

From the box plot analysis, it is evident that the Autoencoder neural network generally performs the best in terms of maintaining structural similarity to the ground truth signals, with high median SSIM, strong linear correlation, and minimal Euclidean distance. The SWD also performs well but with greater variability. In contrast, the rolling window smoother demonstrates the lowest median SSIM and highest variability. This suggests that the Autoencoder is the most reliable method for denoising NMR spectroscopy signals among the those tested.

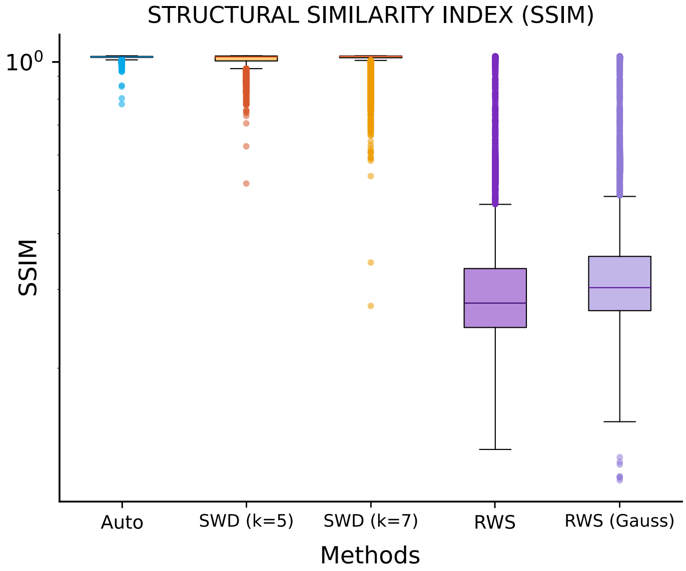


Figure 6.7: Structural Similarity Index (SSIM) values calculated between pure synthetic data and noise-reduced synthetic data processed by three different methods (*Auto*, *SWD*, *RWD*)

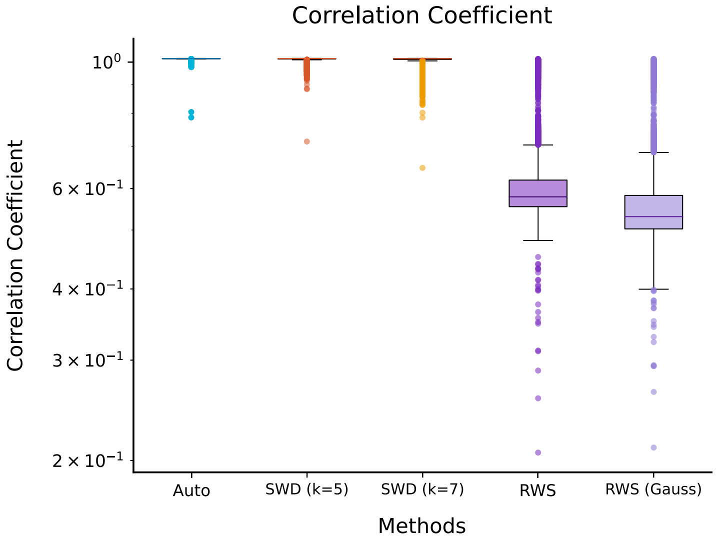


Figure 6.8: Box plots illustrate the correlation coefficient values calculated between pure synthetic data and noise-reduced synthetic data processed by three different methods (*Auto*, *SWD*, *RWD*)

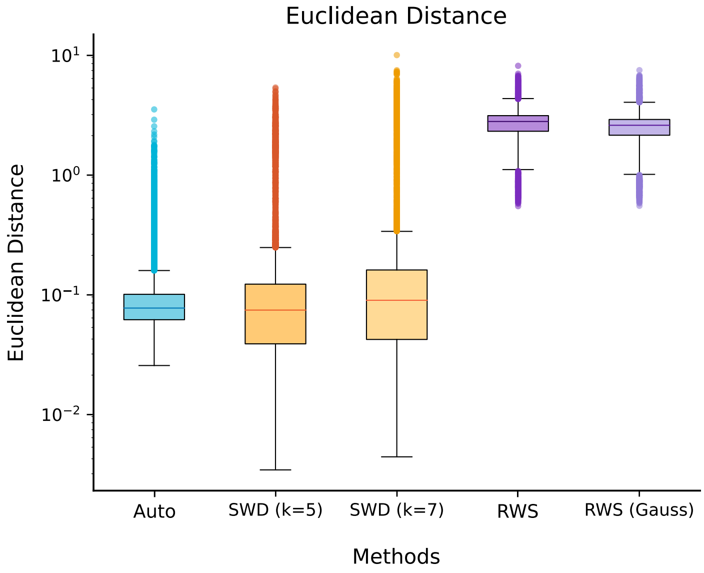


Figure 6.9: Box plots illustrate the Euclidean distance (L2 norm) values calculated between pure synthetic data and noise-reduced synthetic data processed by three different methods ($Auto$, SWD, RWD)

Afterwards, I carried out a further test to evaluate how these methods perform under different noise levels. I divided the data into four groups, each corresponding to an increasing level of noise added to the signals, with a normal distribution, zero mean, and different standard deviations (50, 250, 350, and 500). For each group, the box plots in Figure 6.10 represents how SSIM metric evaluates the aforementioned denoising techniques.

With a standard deviation of noise equal to 50 (Figure 6.10.a), the Autoencoder shows a high median SSIM with relatively low variance, indicating good denoising capability. The SWD method presents a slightly lower median SSIM compared to the Autoencoder but still with robust performance and moderate variance. The RWS has the lowest median SSIM among the three, with higher variance, indicating less effective performance.

As the noise level increases, the Autoencoder maintains relatively high median SSIM values with manageable variance, though performance slightly decreases. The SWD method's median SSIM decreases with increasing noise, showing reduced effectiveness. The RWS continues to exhibit the lowest performance, with decreasing median SSIM and increasing variance.

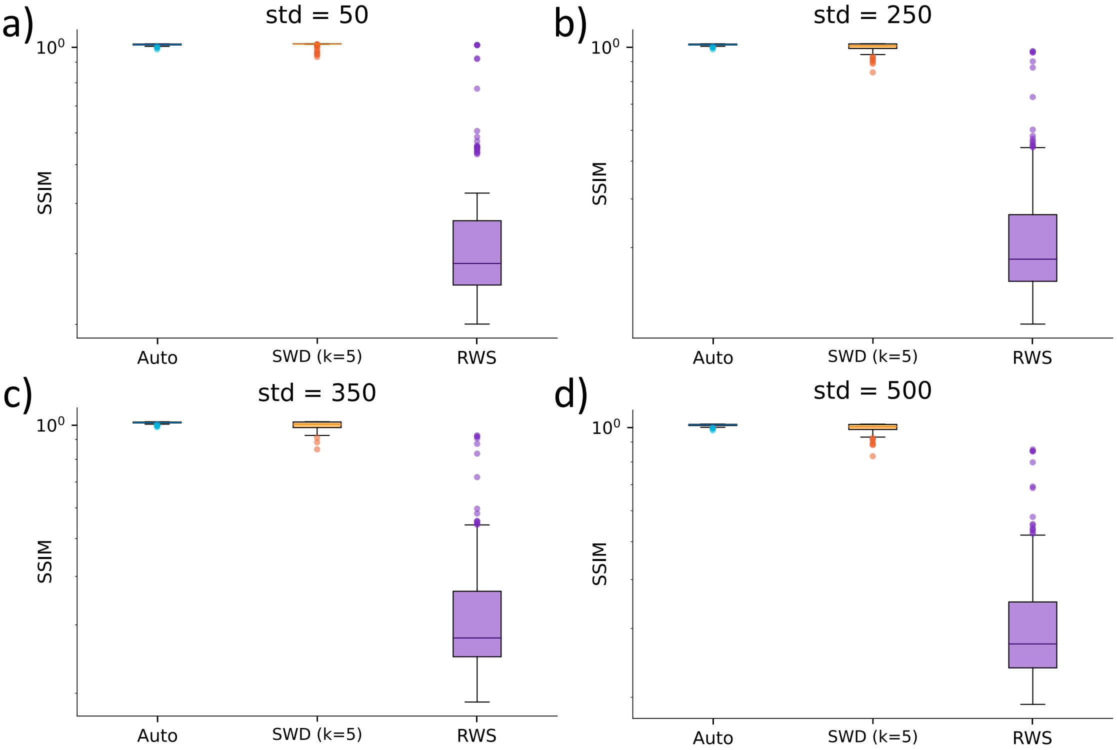


Figure 6.10: Box plots illustrating the Structural Similarity Index (SSIM) values between pure synthetic data and noise-affected synthetic data processed by three methods (*Auto*, *SWD*, *RWD*), varying noise levels. Panels **(a)** to **(d)** display SSIM values calculated between the pure signal and signals affected by noise with standard deviations (std) of 50, 250, 350, and 500, respectively

In this project, I evaluated the performance of three denoising methods on real data data from the Pulsar benchtop from Oxford instrument [40]. I assessed their effectiveness by calculating the noise deviation from 10% of each signal, presumed to contain only noise, and comparing these deviations with the original noisy signals. For these comparisons, I used violin plot (Figure 6.12) and bar plot (Figure 6.11). A violin plot, similar to a box-and-whisker plot, shows the distribution of data points after grouping by one or more variables. Unlike a box plot, each violin is drawn using a kernel density estimate of the underlying distribution [54]. A bar plot presents data with rectangular bars proportional i height to the values they represents, and includes error bars to indicate uncertainty around each estimate [54]. In the performance analysis, *Autoencoder* method achieved the most significant reduction in nose standard deviation, outperforming the other methods. Its performance is evidenced by the lowest bar height in the bar plot and a marked shift toward lower standard deviations in the violin plot, indicating superior denoising performance. *SWD* provided a noticeable reduction in noise but was less effective that the

Autoencoder. The bar plot shows a moderate height for *SWD*, reflecting its intermediate performance in noise reduction, while the violin plot displays a reduction compared to the original noise but with a broader spread, indicating some variability in performance. *RWD* demonstrated the least noise reduction, with standard deviations only slightly lower than the original noisy signals. The bar plot for *RWD* is closest in height to the original noise bar, indicating minimal denoising effectiveness, and the violin plot has a distribution similar to the original noise plot, suggesting that this method is less effective in denoising NMR specta.

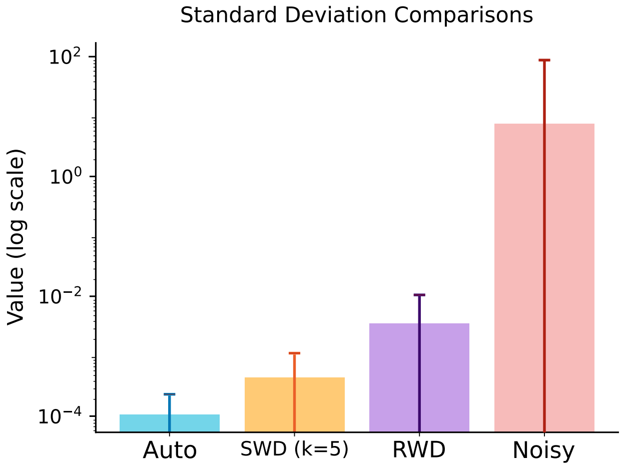


Figure 6.11: Bar plot depicting the distribution of standard deviations calculated from 10 % of the signal after applying three denoising methods (*Auto*, *SWD*, *RWD*). The red violin represents the distribution of standard deviations before noise reduction. The error bars indicates uncertainty around each estimate

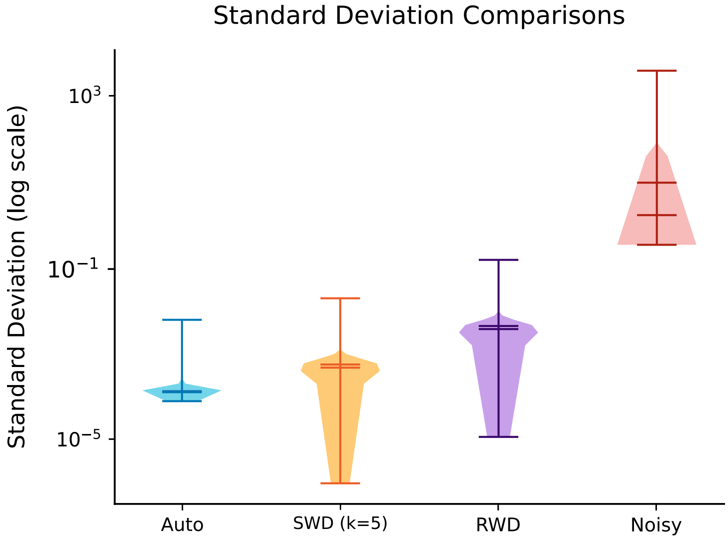


Figure 6.12: Violin plot depicting the distribution of standard deviations calculated from 10 % of the signal after applying three denoising methods (*Auto*, *SWD*, *RWD*). The red violin represents the distribution of standard deviations before noise reduction.

6.2.2 Discussion

Based on the results of applying various denoising methods to NMR spectra, the Autoencoder method demonstrates superior performance compared to stationary wavelet denoising (SWD) and rolling window denoising (RWD). This trend holds across synthetic data with variable noise levels, synthetic data with progressively increasing noise, and real-world data.

The Autoencoder method consistently outperforms the other techniques, maintaining higher Structural Similarity Index (SSIM) values and demonstrating greater robustness in increasing noise levels. It shows a strong linear correlation with the noise-free signal and minimal Euclidean distance. Violin and bar plots further confirm the Autoencoder's significant improvement in denoising NMR spectroscopy signals under varying noise conditions.

In contrast, the SWD method shows moderate performance, which diminishes as noise levels increase. While it provides a more refined approach compared to RWD, SWD is not fully automatic due to the need to set specific variables, such as the threshold for defining peak levels, gaps between recognized peaks, and the number of points to add around peaks. These settings might not be suitable for all spectra types, limiting the generalizability of the SWD method.

RWD demonstrates the lowest performance, particularly at higher noise levels, with lower SSIM values and higher variance. This method is less effective at reducing noise in NMR spectra and requires manual selection of window size, which significantly impacts the results. A large window size averages more points, losing fine details, while a small window size reduces noise insufficiently. Therefore, the choice of window size must be based on the signal's noise level, rendering RWD less automatic and less effective for varied NMR spectra.

Chapter 7

Conclusion

This thesis has addressed the critical question: "How can an automated tool enhance the accuracy and efficiency of preprocessing HP-NMR spectra?" Through the development and evaluation of an automated preprocessing tool, substantial advancements in both accuracy and efficiency demonstrate improvements compared to traditional manual methods.

The focus of this research was on automating baseline correction, phase correction (implemented in Matlab), as well as noise reduction (implemented in Python) for HP-NMR spectra, aiming to minimize human error and enhance reproducibility. The automated phase correction method, particularly the one based on maximizing similarity between corrected and uncorrected absolute signals, proved to be robust and accurate. This approach outperformed manual techniques could significantly reducing user-dependent errors and improving processing speed, particularly with large datasets. Furthermore, the application of an Autoencoder for noise reduction showed superior performance in preserving important signal details compared to other methods, albeit with notable computational resource requirements for training.

Future research directions should prioritize refining phase correction methods with additional constraints and broader spectral considerations. Improvements in denoising techniques tailored to diverse noise conditions and spectrum types would further enhance the efficacy of the automated preprocessing tool. Practical recommendations include extensive deployment in chemical and biological research settings to validate its effectiveness and integration with complementary analytical software for a comprehensive preprocessing suite.

This study aims to significantly advance the field of NMR spectroscopy by introducing an automated preprocessing tool that enhances both accuracy and efficiency. Additionally, it addresses the critical gap in standardizing and automating preprocessing workflows. By advocating for the adoption of automated techniques in routine NMR analysis, this research challenges the conventional reliance on manual methods. Importantly, these advancements contribute directly to precision medicine by improving the reliability of NMR data. This enhancement accelerates the identification of disease biomarkers and metabolic patterns, thereby facilitating more robust and reproducible diagnostic and therapeutic strategies aligned with the goals of precision medicine to deliver personalized healthcare.

In conclusion, integrating automated preprocessing into HP-NMR spectroscopy workflows marks a significant advancement across various fields, illustrating its potential to enhance outcomes in precision medicine and beyond.

Bibliography

[1] The IoT Academy. What is padding in cnn and its types – explained in deep, 4 2024.

[2] Adam R. Altenhof, Harris Mason, and Robert W. Schurko. Desperate: A python library for processing and denoising nmr spectra. *Journal of Magnetic Resonance*, 346:107320, 2023.

[3] Maciej Balawejder. Overview of normalization techniques in deep learning. Published in *Nerd For Tech*, April 2022.

[4] Qingjia Bao, Jiwen Feng, Li Chen, Fang Chen, Zao Liu, Bin Jiang, and Chaoyang Liu. A robust automatic phase correction method for signal dense spectra. *Journal of Magnetic Resonance*, 234:82–89, 2013.

[5] Å. Björck. Algorithms for linear least squares problems. In Emilio Spedicato, editor, *Computer Algorithms for Solving Linear Algebraic Equations*, pages 57–92, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[6] Jason Brownlee. How to reduce overfitting using weight constraints in keras, August 2020. In Deep Learning Performance.

[7] E. Can, J.A.M. Bastiaansen, D.L. Couturier, and et al. [13d]-bicarbonate labelled from hyperpolarized [1-13c]pyruvate is an in vivo marker of hepatic gluconeogenesis in fasted state. *Commun Biol*, 5:10, 2022. Received 31 July 2021, Accepted 07 December 2021, Published 10 January 2022.

[8] Bert Carremans. Handling overfitting in deep learning models. Published in *Towards Data Science*, August 2018.

[9] Hector Zamora Carreras. Nmr spectroscopy: Principles, interpreting an nmr spectrum and common problems, 11 2021.

[10] Tianfeng Chai and R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? *Geosci. Model Dev.*, 7, 01 2014.

[11] Li Chen, Zhiqiang Weng, LaiYoong Goh, and Marc Garland. An efficient algorithm for automatic phase correction of nmr spectra based on entropy minimization. *Journal of Magnetic Resonance*, 158(1):164–168, 2002.

[12] Carlos Cobas. Applications of the whittaker smoother in nmr spectroscopy. *Magnetic*

Resonance in Chemistry, 56, 05 2018.

[13] Carlos Cobas, Felipe Seoane, Santiago DomÃnguez, Stan Sykora, and Antony N. Davies. A new approach to improving automated analysis of proton nmr spectra through global spectral deconvolution (gsd). *Tony Davies Column, Spectroscopy Europe*, February 2011. Mestrelab Research SL, Extra Byte, Castano Primo, Italy, Professor of Analytical Science, SERC, University of Glamorgan, UK, and Director, Analytical Laboratory Informatics Solutions.

[14] Paul H. C. Eilers. A perfect smoother. *Analytical chemistry*, 75 14:3631–6, 2003.

[15] James Eills, Dmitry Budker, Silvia Cavagnero, Eduard Y. Chekmenev, Stuart J. Elliott, Sami Jannin, Anne Lesage, Jorg Matysik, Thomas Meersmann, Thomas Prisner, Jeffrey A. Reimer, Hanming Yang, and Igor V. Koptyug. Spin hyperpolarization in modern magnetic resonance. *Chemical Reviews*, 123(4):1417–1551, 2023. PMID: 36701528.

[16] Wei Gao. An improved fast-convergent genetic algorithm. In *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003*, volume 2, pages 1197–1202 vol.2, 2003.

[17] Agustin Garcia Asuero, Ana Sayago, and Gustavo GonzÃ¡lez. The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry - CRIT REV ANAL CHEM*, 36:41–59, 01 2006.

[18] GeeksforGeeks. Greedy algorithms, May 2024. Last updated: 02 May, 2024.

[19] Eligius M. T. Hendrix and Ana Maria A. C. Rocha. On local convergence of stochastic global optimization algorithms. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Chiara Garau, Ivan Blečić, David Taniar, Bernady O. Apduhan, Ana Maria A. C. Rocha, Eufemia Tarantino, and Carmelo Maria Torre, editors, *Computational Science and Its Applications – ICCSA 2021*, pages 456–472, Cham, 2021. Springer International Publishing.

[20] Joseph P. Hornak. *The Basics of MRI.* Rochester Institute of Technology, 2020. Copyright 1996-2020 J.P. Hornak. All Rights Reserved.

[21] Hotjar. The complete guide to heatmap, January 2024. Last updated: 12 January 2024.

[22] Thomas L James. Fundamentals of nmr. *Online Textbook: Department of Pharmaceutical Chemistry, University of California, San Francisco*, pages 1–31, 1998.

[23] JEOL USA. Nmr basics for the absolute novice, n.d. Accessed: 2024-07-03.

[24] Maria Kaliva and Maria Vamvakaki. Chapter 17 - nanomaterials characterization. In Ravin Narain, editor, *Polymer Science and Nanotechnology*, pages 401–433. Elsevier, 2020.

[25] James Keeler. *Chapter 4 - Fourier transformation and data processing.* Wiley, 2002.

[26] Joachim Kessler. *Description of Polarized Electrons*, pages 7–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[27] Igor Kononenko and MatjaÅ¾ Kukar. Chapter 5 - learning as search. In Igor Kononenko and MatjaÅ¾ Kukar, editors, *Machine Learning and Data Mining*, pages 131–151. Woodhead Publishing, 2007.

[28] LibreTexts. Spin quantum number, n.d. This page is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by LibreTexts.

[29] Guang Rui Liu, Xin Tian, Wenbo Zhou, and Kefu Guo. An improved genetic algorithm in path planning for mobile robot. pages 991–996, 2015.

[30] Xin Liu. *Chapter 6: Structural Identification of Organic Compounds: IR and NMR Spectroscopy*, chapter 6.5, page NMR Theory and Experiment. KPU Pressbooks, 2021. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

[31] Ricardo O. Louro. Chapter 4 - introduction to biomolecular nmr and metals. In Robert R. Crichton and Ricardo O. Louro, editors, *Practical Approaches to Biological Inorganic Chemistry*, pages 77–107. Elsevier, Oxford, 2013.

[32] Mukesh Manral. Guide to commonly used deep learning kernel initializers in real-world projects, 3 2023.

[33] Alan G. Marshall. Dispersion vs. absorption (dispa): A magic circle for spectroscopic line shape analysis. *Chemometrics and Intelligent Laboratory Systems*, 3(4):261–275, 1988.

[34] Mestrelab Research. *MestReNova Manual*, 12 edition, September 2017. 2017 MESTRELAB RESEARCH. Last Revision: 13-Sept-2017.

[35] Miramar Communications Ltd. What is benchtop nmr?, 2024. Oxford Instruments 2024. Website by Miramar Communications Ltd.

[36] Mor Mishkovsky and Lucio Frydman. Progress in hyperpolarized ultrafast 2d nmr spectroscopy. *ChemPhysChem*, 9(16):2340–2348, 10 2008. First published: 31 October 2008.

[37] Sushruta Mishra, Soumya Sahoo, and Mamata Das. Genetic algorithm: An efficient tool for global optimization. *Advances in Computational Sciences and Technology*, 10(8):2201–2211, 2017.

[38] nanalysis. Nmr data processing: Phase correction, 2022.

[39] Georgios Nanos. Deep neural networks padding, 3 2024. Reviewed by: Michal Aibin.

[40] Oxford Instruments plc. Pulsar nmr spectrometer brochure, 2019. Oxford Instruments plc, 2019. All rights reserved. Ref: P-03-19.

[41] Federico M. Paruzzo, Simon Bruderer, Youssef Janjar, Bjoern Heitmann, and Christine Bolliger. Automatic signal region detection in 1 h nmr spectra using deep learning. 2020.

[42] Y. Peng, Z. Zhang, L. He, and et al. Nmr spectroscopy for metabolomics in the living system: recent progress and future challenges. *Anal Bioanal Chem*, 416:2319–2334, April 2024. Received 24 September 2023, Revised 08 December 2023, Accepted 10 January 2024, Published 19 January 2024, Issue Date April 2024.

[43] Radiology Key. Magnetic resonance basics: Magnetic fields, nuclear magnetic characteristics, tissue contrast, image acquisition. Fastest Radiology Insight Engine.

[44] P. C. Riedi and James S. Lord. Fundamentals of nmr. *ChemInform*, 26, 1995.

[45] Rijois Iboy Erwin Saragih and Darsono Nababan. Increase performance genetic algorithm in matching system by setting ga parameter. *Journal of Physics: Conference Series*, 1175(1):012100, mar 2019.

[46] I.H. Sarker. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(5):420, 2021.

[47] Shipra Saxena. Introduction to batch normalization, June 2024.

[48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[49] Metika Sikka. Using skip connections to enhance denoising autoencoder algorithms, 6 2020. Accessed: 2024-06-14.

[50] Serge L. Smirnov and James McCarty. 5.1: Nuclear spin and magnetic field, n.d. This page is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Serge L. Smirnov and James McCarty.

[51] Christian Versloot. Creating a signal noise removal autoencoder with keras, 12 2019. Accessed: 2024-06-14.

[52] Ivan Vlasic, Marko Durasevic, and Domagoj Jakobovic. Improving genetic algorithm performance by population initialisation with dispatching rules. *Computers & Industrial Engineering*, 137:106030, 2019.

[53] Xiaoli Wang and Yongfeng Dai. An improved denoising method based on stationary wavelet transform. In *Proceedings of the 2018 International Symposium on Communication Engineering & Computer Science (CECS 2018)*, pages 481–485, 2018/07.

[54] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[55] Eric W. Weisstein. L$\hat{2}$-norm. MathWorld–A Wolfram Web Resource, n.d. Accessed: 2024-07-03.

Acronyms

ANNs Artificial Neural Networks. 42

AP Actual Points. 51

DL Deep Learning. 42

DWT Discrete Wavelet transform. 45

FID Free Induction Decay. 26–28, 31, 32, 35, 38, 50

FT Fourier Transform. 26, 51

GA Genetic Algorithm. 56, 58, 78

HP-NMR Hyperpolarized Nuclear Magnetic Resonance. 1

ISWT inverse SWT. 46, 68

MSE Mean Square Error. 58, 59, 65

NMR Nuclear Magnetic Resonance. 26, 28, 31, 34, 38, 52, 57

O Channel Frequency Offset. 50, 52

OF Objective Function. 55, 57

PF Penalty Function. 56, 57

RF Radiofrequency. 13

RP Receiver Points. 50, 51

SF Base Frequency. 50

SNR Signal-to-Noise ratio. 31, 32

SWT Stationary Wavelet transform. 45, 68

WT Wavelet transform. 45