NOISE REDUCTION AND ARTEFACT REMOVAL IN MR SPECTROSCOPY USING WAVELET ANALYSIS AND DEEP LEARNING

Master thesis Technical Medicine

Author:

Martin Bastiaan Schilder, BSc.

Supervisory committee:

Medical supervisor:	Dr. M.H. Lequin
Technical supervisor UT:	Dr. J.M. Wolterink
Technical supervisor UMCU:	Dr. A Bhogal
Process supervisor:	Drs. P.A. van Katwijk
Chairman of committee:	Prof. dr. ir. B. ten Haken
External member:	Dr. ir. L. Alic

UNIVERSITY OF TWENTE.

UMC Utrecht

Abstract

Introduction: Magnetic Resonance Spectroscopy is an imaging technique that grants information on tissue metabolism. After performing a Fourier transform, metabolites in the spectrum can be quantified. However, noise and artefacts from both lipids and unsuppressed water impair accurate metabolite quantification. To mitigate the effects of noise and artefacts, deep learning could be a solution. Additionally, the added value of the continuous wavelet transform and resulting scalogram in noise reduction and artefact removal are addressed in this study.

Methods: 8000 spectra were simulated and contaminated with noise. 1000 spectra were contaminated with a residual water artefact. Another 1000 spectra were contaminated with a lipid artefact. After that, the data was split into a training set of 6400 spectra and scalograms and a validation set of 1600 spectra and scalograms. A convolutional autoencoder, a U-Net and a dilated convolutional neural network were trained on the spectra and the scalograms. After training, the standard deviation of the residual and peak height differences for the three peaks was calculated. 10 *in vivo* spectra were also tested.

Results: Parameters were found to transform simulated and *in vivo* spectra using the continuous wavelet transform and to reconstruct spectra using the inverse continuous wavelet transform with minimal loss of information. All networks significantly improved the residual and peak height differences. When presented with *in vivo* spectra, the output either had prominent peaks that were not visible in the input spectrum or missed peaks that were visible in the input spectrum.

Conclusion: Deep learning networks reduced noise and removed artefacts with variable degrees of success. The added value of the wavelet transforms was proven in reducing noise and removing lipid artefacts. Contrarily, it was not proven for removing the residual water artefacts. Results from *in vivo* input to the networks could not be trusted. Future studies should focus on simulating more realistic spectra and quantifying metabolite concentrations in simulated spectra.

Keywords: Magnetic resonance spectroscopy, continuous wavelet transform, deep learning, artefact removal

Abbreviations

Asp: aspartic acid

B_0: main magnetic field

CHESS: chemical shift selective saturation

CNN: convolutional neural network

Cr: creatine

CRLB: Cramér–Rao lower bound

CSI: chemical shift imaging

CWT: continuous wavelet transform

FID-A: FID Appliance

f_L: Larmor frequency

Glx: sum of glutamate and glutamine

GPC: glycerophosphocreatine

GSH: glutathione

ICWT: inverse continuous wavelet transform

IDH1: isocitrate dehydrogenase 1

J1: the number of frequency scales

MRI: magnetic resonance imaging

MRS: magnetic resonance spectroscopy

MRSI: magnetic resonance spectroscopic imaging

NAA: N-Acetyl aspartate

NAAG: N-Acetyl aspartyl glutamate

PC: phosphocholine

PCr: phosphocreatine

ppm: parts per million

PReLU: parametric rectified linear unit

PRESS: point resolved spectroscopy

ReLU: rectified linear unit

RF: radiofrequency

s_0: lowest scaling factor

sd: standard deviation

SE: spin echo

SNR: signal-to-noise ratio

SSIM: structural similarity index measure

tCho: total choline

tCr: total creatine

TE: echo time

Table of Contents

1. Introduction

Brain cancer, in both paediatric and adult patients, is a disease with a poor prognosis. Paediatric brain tumours have a mortality rate of 30% (1,2). For adults, this number is even higher due to the higher incidence of aggressive brain tumours (3). To diagnose a brain tumour, radiologists use magnetic resonance imaging (MRI) because of its superior soft tissue contrast (4). MRI grants the radiologist anatomical and structural information of the imaged organ. However, structural information from MRI is not in all cases sufficient for differential diagnosis or treatment monitoring since it lacks information on tissue metabolism. Proton Magnetic Resonance Spectroscopy (^{1}H MRS, further referred to as MRS) is an advanced imaging technique that can add to traditional MRI by obtaining insight into tissue metabolism (4,5).

The insight into tissue that MRS offers is of great value in brain cancer patients, both in children and adults. Firstly, MRS can aid in the differential diagnosis. The detection of 2-Hydroxyglutarate can predict isocitrate dehydrogenases (IDH) mutant glioma with a specificity of 95% and a sensitivity of 91% (6). This is an important finding, since IDH mutant glioma respond well to temozolomide therapy in combination with radiotherapy, whereas patients with an IDH wildtype glioma hardly benefit from temozolomide therapy (7). Also, MRS can help distinguish between medulloblastoma and ependymoma, since medulloblastoma typically have a high taurine concentration, whereas ependymoma have a very high myo-inositol concentration (8).

Not only does MRS have value at the initial moment of diagnosis, but it also allows for treatment monitoring. Most of adult glioblastoma patients are treated with chemoradiotherapy. After a few month following the start of therapy, the effect of treatment is monitored with MRI. Yet, MRI is on itself not always conclusive, since true disease progression under therapy cannot in all cases be distinguished from therapy effects, which mimic radiological signs of progression. This leaves radiologists and oncologists with the question whether to switch therapy or wait for the next scan to be conclusive, risking disease progression. MRS might be a solution to this question as well, as it was found that it was 84% accurate in distinguishing true disease progression from pseudoprogression (9).

The reason MRS can be helpful is because its ability to obtain signal from metabolites and to quantify these signals in metabolite concentrations. In tissues, different metabolites are found. These molecules shield the protons with a varying degree from the magnetic field. The result is that protons experience a slightly different magnetic field, and will precess at a slightly different frequency. By Fourier transforming the signal acquired in MRS, the contribution from different frequencies can be analysed. These different frequencies in the signal correspond to known metabolites and can thereafter be quantified using a linear combination of model spectra of known metabolite concentrations, either from *in vitro* experiments or from simulations (10). In pathological conditions, the concentrations of these metabolites change (11). To give an example, in tumorous tissue, the concentration of choline is elevated, since choline is a marker of membrane turnover. Contrarily, the concentration of N-Acetyl aspartate (NAA), as a marker of neuronal integrity, is lower in tumorous tissue since the integrity of these neurons is disrupted by tumour growth (12–15). Metabolite concentrations between different pathological conditions also varies (6,8,16). This enables the possibility for classification (17).

However, MRS also has downsides, which have the ability to impair accurate and reliable quantifications, and thereby prevent MRS to have clinical impact. The downside of MRS is threefold. Firstly, the metabolite concentrations are very low compared to water, resulting in a relatively low signal-to-noise ratio (SNR). Solutions to improve SNR could be to use larger voxels or to take more signal averages. However, a larger voxel might cover a volume that not only takes into account the lesion, but also healthy tissue, thereby mixing signals from diseased tissue with healthy tissue. More signal averages are also not always feasible in clinical routine because they cost additional scanning time. One factor contributing to the SNR is fluctuations of the main magnetic field, which results in

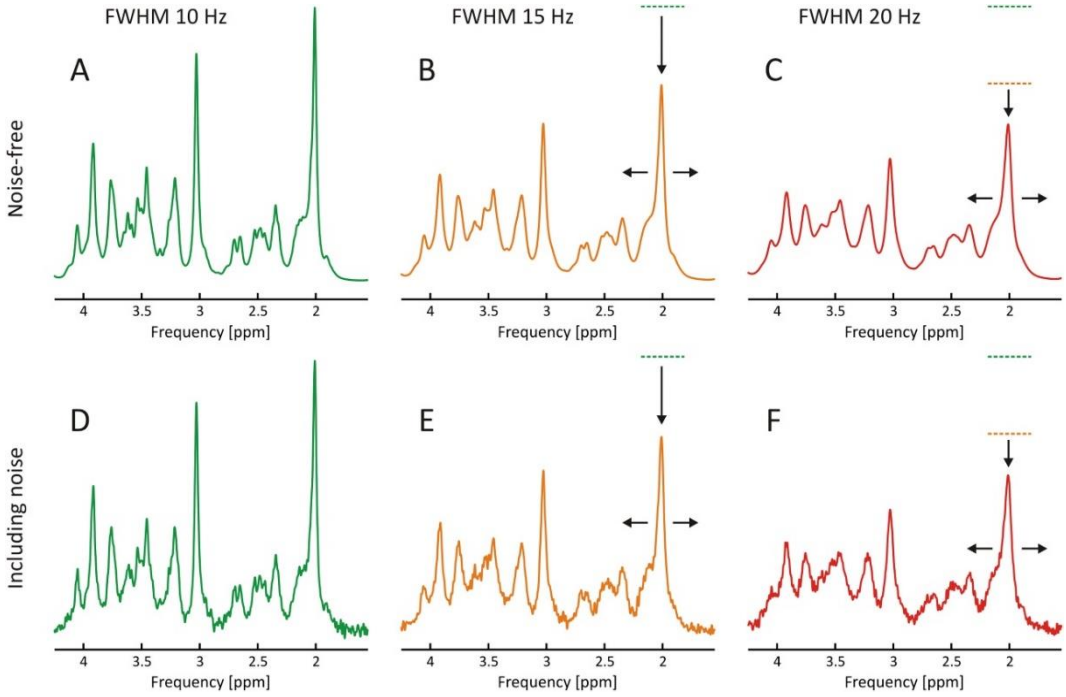


Figure 1.1: Simulation of MR Spectrum of the human brain at 7T. The linewidth is proportionate to the full width at half maximum (FWHM). The linewidth increases from A to C, e.g. due to magnetic field inhomogeneities. With increasing linewidth, the peaks start to overlap. In D, E and F, some noise has also been added. However, MR spectra are usually subject to higher noise levels. From Juchem et al. (20).

more noise and broader peaks. Fluctuations in the magnetic field can happen near bone structures or at tissue-air interfaces (18). Broader peaks in turn lead to more overlapping peaks and less reliable quantification (19). An example of peak broadening is seen in Figure 1.1. Secondly, some metabolites, such as glutamate and glutamine, overlap as they resonate at frequencies that are in close proximity to each other. It can therefore be hard to distinguish which metabolite contributed to the signal. This in turn leads to more uncertainty. Thirdly, artefacts can superimpose on the metabolites (12,20–22). Superimposing artefacts leads to less reliable metabolite quantification because artefacts are not part of the set of model spectra used for fitting, also known as the basis set (23). An example of an artefact is a lipid artefact. These artefacts originate from adjacent fatty tissue, such as fat in subcutaneous tissue, or locations with cellular membrane breakdown or necrosis (18). Due to the oscillating nature of the point spread function, signal from voxels containing the actual lipid, leaks into neighbouring voxels, thereby introducing a lipid artefact (24). Another artefact that superimposes on the spectrum, is the water artefact. This is the result of incomplete water suppression. The right tail of the water peak disturbs the peaks in the range of 2.0 to 4.7 parts per million (ppm) (25). An example of an *in vivo* water artefact can be seen in Figure 1.2. Low SNR, overlapping peaks and artefacts impair the accurate quantification of metabolite concentrations. This leads to high uncertainties in metabolite concentrations, and this uncertainty cannot be used in medical decision-making.

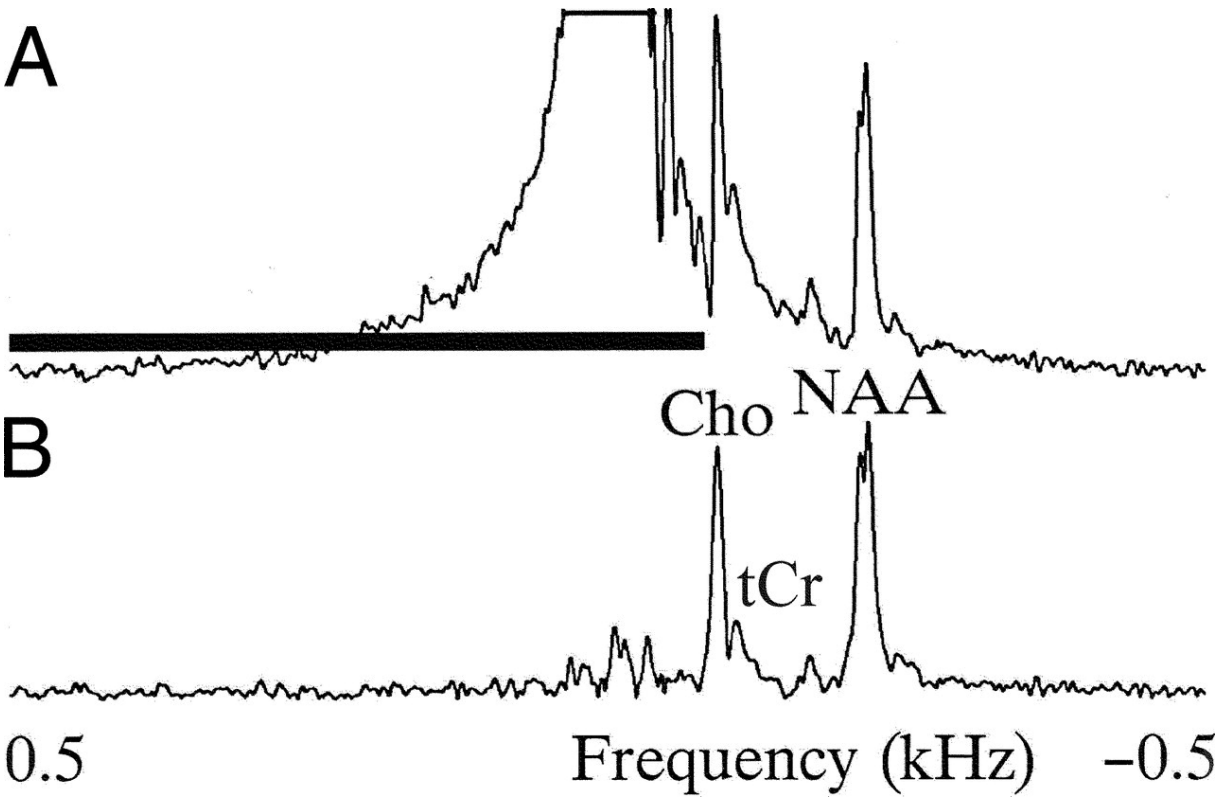


Figure 1.3. A: A big water peak overlaps with choline (Cho) and creatine (tCr). B: The water peak has been removed using a Hankel-Lanczos singular-values decomposition filter. The black horizontal band indicates the region that contains the water peak. From Jansen et al. (26).

To quantify metabolite concentrations more reliably, the effects of noise and artefacts have to be mitigated. Approaches that have been applied are denoising using a statistical selection of principal components (26) and using receive-coil sensitivity profiles to remove spurious echo artefacts (27). In recent years, deep learning solutions have also been explored. Deep learning networks aim to learn features that optimally represent the data of interest, such that peaks from metabolites are separated from spectra that also contain noise and artefacts (28). Multiple studies have already been performed that addressed the issues in MRS using deep learning. Lee *et al*. demonstrated with the use of a convolutional neural network (CNN) that it was possible to perform noise reduction, line narrowing and baseline removal in simulated brain spectra (29). In a study by Kyathanahally *et al.*, ghost artefacts were detected using a CNN and thereafter removed from spectra using stacked what-where autoencoders and spectrograms, which are the 2D representations of spectra obtained using the Fourier transform (23).

Besides using the Fourier transform to represent a spectrum in a higher dimension through a spectrogram, wavelet decomposition can also achieve this. With wavelet decomposition, the spectrum is decoupled in the ppm-frequency domain using the continuous wavelet transform (CWT). Compared to the Fourier transform, the CWT preserves location information. Wavelet coefficients are obtained by continuously varying the scale parameter and position parameter to select the different portions of the spectrum and analyse the different scale variations. The wavelet coefficients are then plotted in the scalogram for each scale of the wavelet (30). An example of a spectrum and its corresponding scalogram can be found in Figure 1.3. Through an inverse continuous wavelet transform (ICWT), the spectrum can again be obtained. In the scalogram, coefficients corresponding to specific metabolite peaks, noise and artefacts are isolated. Due to this isolation in 2D scalograms, deep learning networks potentially pick up and remove unwanted features easier in 2D scalograms than in a 1D spectrum, and with that leading to more reliable metabolite concentrations. Yet, it is not known if this is true in this specific application. The hypothesis is that by decoupling information in the scalogram, a deep learning network can more easily detect the isolated features that represent either metabolites, noise or artefacts and will thus outperform networks trained on spectra.

To train deep learning networks, usually, there is a need for a lot of data, and for the networks that

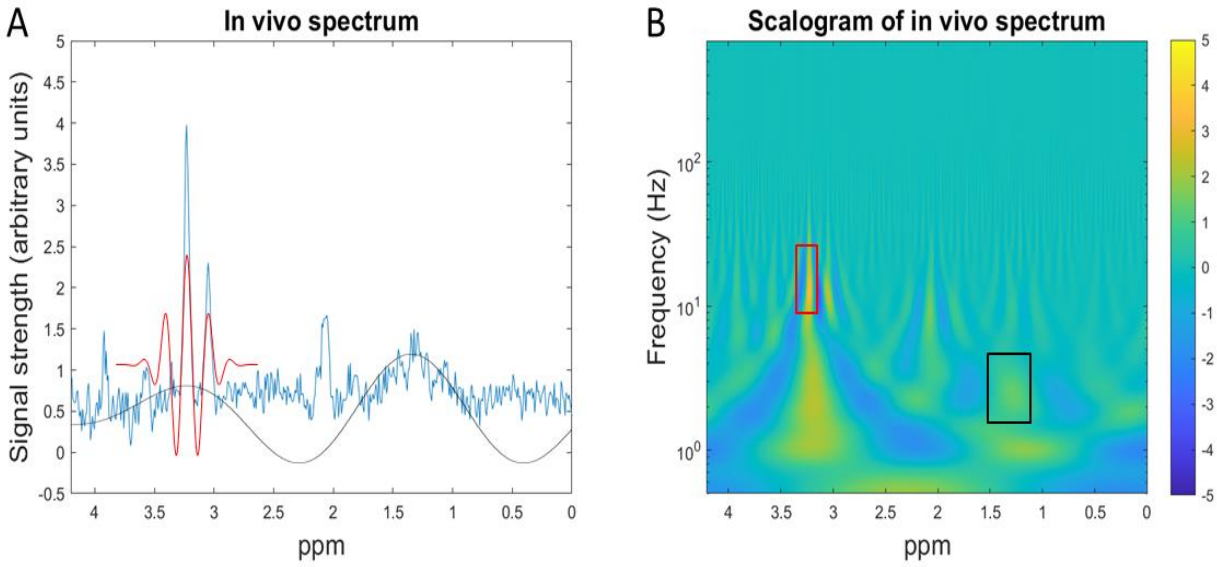


Figure 1.3. A: In vivo spectrum in blue. In red is a wavelet with a high frequency, and in black is a wavelet with a low frequency. B: Corresponding scalogram. The red and blue boxes indicate the location of the wavelet coefficients in the scalogram. Spatial information was preserved.

were trained for this research, also paired data. Paired data means an input and output pair of spectra or scalograms exists, one containing metabolites only, and the other also containing contaminations. These two spectra need to be the exact same spectra with the exact same metabolite concentrations. However, such paired data lacks in *in vivo* MRS. Therefore, the networks were trained on simulated data.

The goal of this thesis is therefore twofold: the first goal is to quantify metabolites in the spectra more accurately and reliably by removing noise and artefacts from the spectra using deep learning networks that were trained on simulated data. The second goal is to determine the added value of wavelet decomposition and use the decoupled information in the scalogram as input for a deep learning network.

2. Background

Magnetic Resonance Spectroscopy

^{1}H atoms, when placed in a constant magnetic field, will align with the magnetic field and precess around their axis. The frequency with which the ^{1}H atoms precess is called the Larmor frequency (F_L), and is calculated with Equation 1 (2.1):

$$f_L = \frac{\gamma}{2\pi} \times B_0 \tag{2.1}$$

In Equation 1, γ is the gyromagnetic ratio (42.6 MHz/T) and B_0 is the magnetic field strength. MRS is based on the fact that hydrogen atoms in a molecule are shielded by neighbouring chemical bonds. Due to this shielding, the hydrogen atoms experience a slightly different magnetic field strength and will therefore resonate at a different Larmor frequency. This is also known as chemical shift. The effective Larmor frequency is calculated by adding the shielding factor σ to Equation 2.2 (4):

$$f_{L,eff} = (1 - \sigma) \times \frac{\gamma}{2\pi} \times B_0 \tag{2.2}$$

The relative difference between f_L and $f_{L,eff}$ is called chemical shift (4,25). The chemical shift can be visualized in a spectrum. As it is known which part of the spectrum corresponds with what metabolite, distinct peaks can be allocated to these metabolites. For instance, choline can be found at 3.22 ppm, and NAA can be found at 2.01 ppm (4,31,32). An example of a spectrum can be found in Figure 2.1.

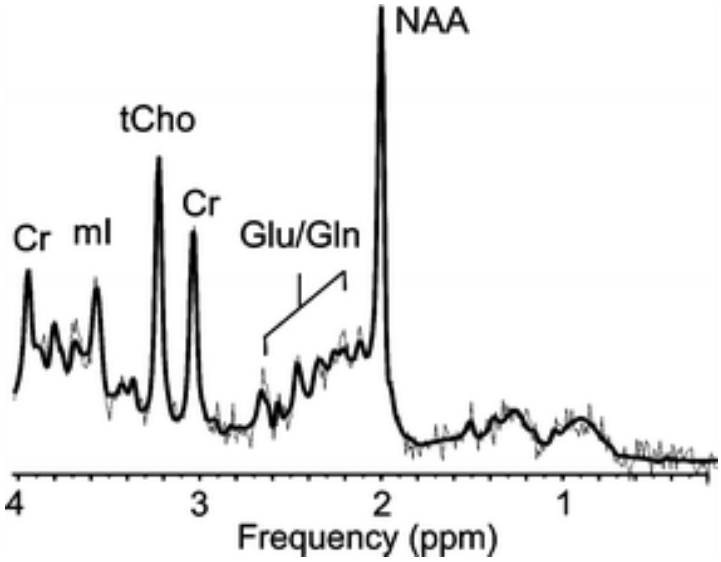


Figure 2.1 MR spectrum of healthy paediatric brain. From Panigrahy et al. (33).

MRS can be divided into two techniques. Firstly, single-voxel MRS. In single-voxel MRS, one region of interest is targeted. These voxels are typically big and cover a whole lesion. As data needs to be acquired from a single-voxel only, acquisition time is relatively short and spectral quality is relatively good due to the relatively big voxel size. The disadvantage of single-voxel MRS is that the data gathered from the used voxel represents that part of the tumour covered by the voxel. As a result, in the case of a heterogeneous tumour, commonly seen in paediatric patients, the obtained spectrum is not by definition representative. It might for instance average the most aggressive part with necrotic tissue in the core of the tumour. This could be resolved by acquiring multiple voxels at the cost of a longer scan time, which is clinically not always feasible.

Secondly, there is magnetic resonance spectroscopic imaging (MRSI), which is also called chemical shift imaging (CSI). The difference with single-voxel MRS is that, in MRSI, a grid of voxels is placed located around the tumour, or even over the whole brain. Usually, these voxels are smaller than voxels in single-voxel MRS. This leads to a relatively low SNR in MRSI compared to single-voxel MRS, because there are fewer protons in a smaller voxel that give rise to peaks. The advantage of MRSI is that it provides spatial information on metabolite concentrations and can cover large tumours with different tumour components (33,34). An example of MRSI in a glioblastoma patient can be found in Figure 2.2. The most prominent changes are an increase in the choline peak and a decrease in the NAA peak, indicative of increased membrane turnover and loss of neuronal integrity (35).

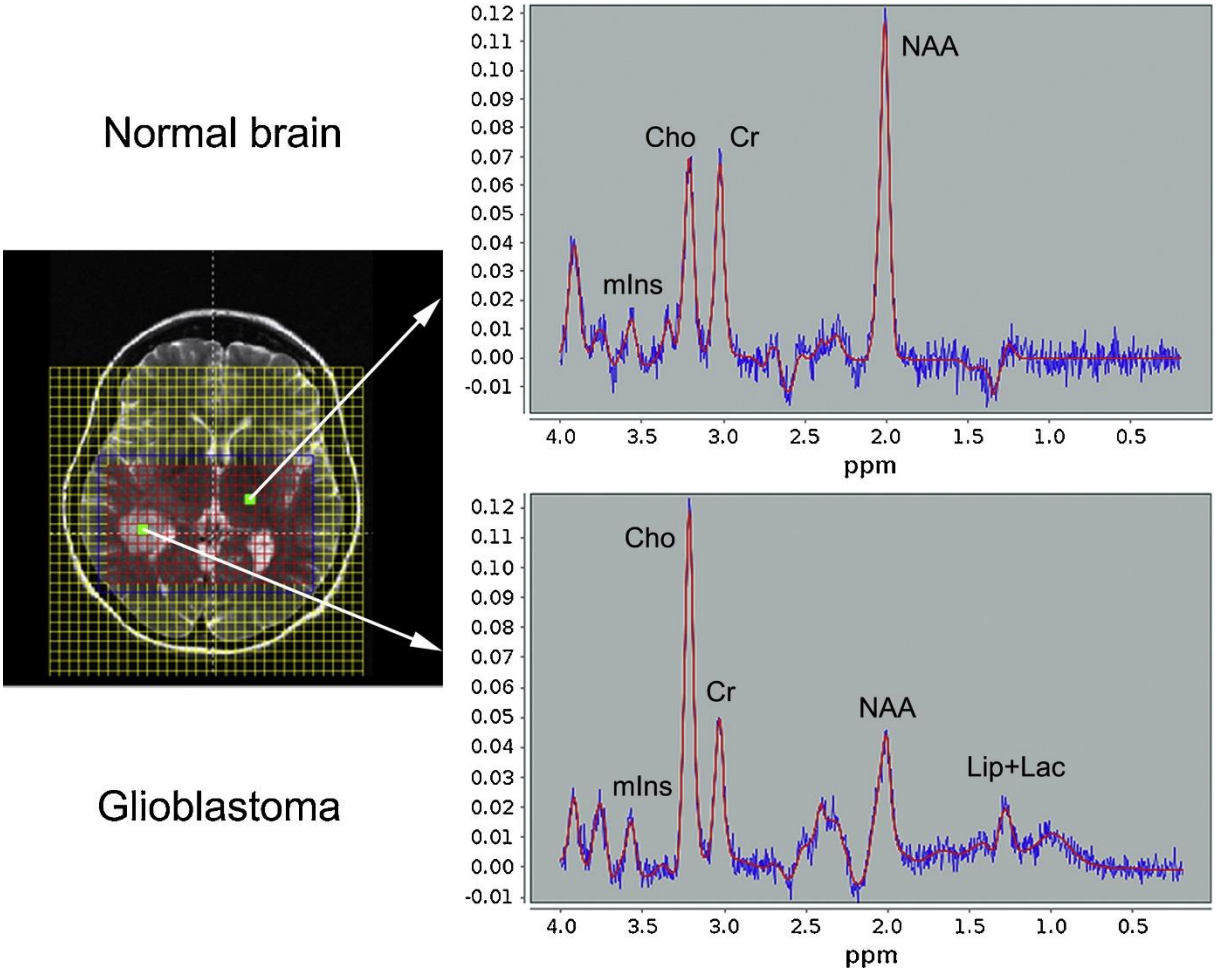


Figure 2.2. MRSI and spectra from two voxels of an adult patient. The top spectrum has been acquired from healthy brain tissue. The bottom spectrum has been acquired from tumorous tissue. Figure from Bulik et al. (36).

Nowadays, the most commonly used sequence in single-voxel MRS and MRSI is Point Resolved Spectroscopy Sequence (PRESS). PRESS is a 90^o-180^o-180^o spin-echo (SE) pulse sequence. Herein, the 90^o radiofrequency (RF) pulse is an excitation pulse, and the 180^o RF pulses are refocusing pulses needed to rephase the excited protons. The RF pulses are applied simultaneously with G_Z, G_X and G_Y field gradients. Thus, only the protons that lie in the intersection of the three crossing planes experience all three RF pulses, forming a cuboid-shaped voxel. After the last 180^o RF-pulse, the SE signal is acquired from this voxel (12,36,37). The PRESS sequence can be found in Figure 2.3.

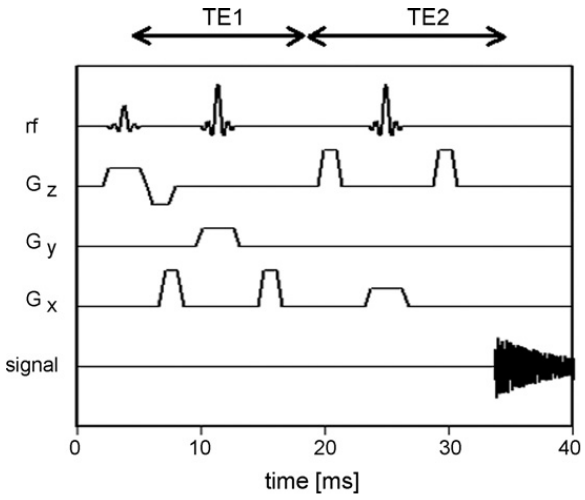


Figure 2.3: PRESS sequence. From Klose et al. (37).

One major factor to consider in MRS is the echo time (TE). TE is an acquisition parameter that can be modulated to better isolate certain metabolites. If, for instance, TE is chosen to be short (<40 ms), the SNR is higher compared to longer TE. This is because most metabolites have a relatively short T2-time (38). Contrary, if TE >135 ms, the SNR is lower. The advantage of a longer TE is that lactate peaks can be distinguished from lipid peaks. This because lactate has a longer T2 time compared to lipids (39). Using a 3T MR system, at a TE of approximately 144 ms, the lactate peak becomes inverted due to J-coupling (40,41).

At the University Medical Center Utrecht, the LCModel software is used to quantify metabolites. LCModel analyses an *in vivo* MR spectrum and fits a linear combination (LC) of *in vitro* spectra from individual metabolite solutions. The area under the curve is then proportionate to the concentration of the metabolites. As the signal also includes noise, peaks of different metabolites can overlap or there is a strong baseline signal from macromolecules, the concentration of these metabolites is uncertain. This uncertainty is expressed in the Cramer-Rao lower bound (CRLB) (34).

Metabolites

In MRS, a wide variety of metabolites can be measured. This section lists the most important metabolites. Their location in the spectrum, function in healthy tissue and clinical significance are listed below.

Alanine

Alanine can be found at two resonance frequencies. These are 1.47 ppm and 3.77 ppm. After leucine, it is the second most occurring amino acid. Alanine concentrations are linked to metabolic pathways since it is synthesised through reductive amination of pyruvate (12). Alanine can be elevated in meningiomas, a brain tumour arising from the membranous layers surrounding the brain (11,31).

Aspartic acid

Aspartic acid (Asp) is an amino acid that acts as an excitatory neurotransmitter. Aspartic acid is the anionic form of aspartate. Aspartate transports electrons produced during glycolysis across the mitochondrial inner membrane. Asp can be found at three resonance frequencies, namely 3.89 ppm and 2.67 ppm and 2.80 ppm (12).

Choline

The total choline (tCho) peak is one of the most prominent peaks in and MR spectrum and can be found at 3.21 ppm. This peak arises mostly from phosphorylcholine (PC) and glycerophosphorylcholine (GPC). Free choline (Cho) also contributes to the signal. Smaller peaks can be observed in between 3.5 and 4.5 ppm. Choline-containing compounds are related to membrane turnover and increased tCho is seen in cancer. Yet, as multiple metabolites contribute to the tCho signal, the exact interpretation is complex (12,31).

Citrate

Citrate is located between the 2.57 and 2.72 ppm range. However, due to the overlap of creatine and glutamine, citrate cannot always be detected accurately. Usually, citrate levels are elevated in brain tumours (12).

Creatine

The total creatine (tCr) is found at 3.03 ppm. Both creatine (Cr) and phosphocreatine (PCr) contribute to this peak. A smaller peak at 3.93 can be observed. As the frequency shift of Cr and PCr is not very big, the peaks are not separable at 3T. Creatine is involved in the energy metabolism of cells. In chronic phases of disease, creatine levels can be decreased. As creatine does not originate from the brain, systemic disease may impact creatine levels (12,31).

Glutamate and glutamine

At lower field strengths, the glutamate and glutamine (Glx) sum is quantified more reliably than separately. The peak is located at 2.0-2.46 ppm and 3.6-3.8 ppm. Glutamate is a precursor of γ-aminobutyric acid (GABA), an inhibitory neurotransmitter. Glutamate also plays a role in the synthesis of other metabolites, peptides and proteins. Glutamine is synthesised from glutamate in astroglia. It has been suggested that glutamine functions as fuel for multiple types of cancer (12,31).

Glutathione

Glutathione (GSH) is a tripeptide that consists of glycine, cysteine and glutamate. GSH is an antioxidant, thereby neutralizing free radicals and reactive oxygen species. In Parkinson's disease and in other neurodegenerative diseases affecting the basal ganglia, altered GSH levels have been reported. GSH can be found in the spectrum at 3.77 ppm, overlapping with glutamate, and at 2.98 ppm, 2.55 ppm and 2.15 ppm (12,42).

Lactate

The lactate has two distinct peaks, that are found at 1.3 ppm and 4.1 ppm. Lactate is a product of anaerobic glycolysis. It is elevated in cases of insufficient blood flow. This can be the case low-grade tumours, or in the necrotic core of a high-grade tumour (8,12,13). In the case of glioblastoma, lactate concentrations are an important factor in outcome prediction. A long TE is required to measure lactate concentrations (14,15).

Lipids

Lipids are mostly found at 0.9 ppm and 1.3 ppm. Lipids also show resonances at different places in the spectrum, though these resonances are not as significant. Lipids in the brain are found in the cell membrane and in myelin sheaths. The presence of lipid peaks in a spectrum might be associated with grade IV tumours and bad prognosis (12,13).

Myo-Inositol

The myo-inositol peaks are found between at 3.27 ppm, 3.55 and 4.05 ppm. Myo-inositol is a cyclic sugar alcohol. Its function is relatively unknown, though it said to play a role as glial marker (12). Myo-inositol is usually elevated in low-grade diffuse gliomas (15).

N-Acetyl aspartate

NAA is located at 2.01 ppm. In healthy brain proton MRS, the NAA peak is the most prominent peak in the spectrum (31). The NAA resonance is used as a marker of neuronal density. A decrease in the height of the NAA peak might be a sign of neuronal disorder (12,15). When looking at the ratio tCho/NAA, a higher ratio is associated with higher tumour grading and poorer prognosis (13–15).

Metabolites that contain N-Acetyl contribute to the NAA peak. One of these metabolites is N-Acetyl Aspartyl Glutamate (NAAG). As NAAG resonates at 2.04 ppm, which is very close to NAA, it usually is impossible to detect NAAG separately from NAA (12).

Taurine

Taurine resonates at two locations in the spectrum: 3.25 ppm and 3.42 ppm. These overlap with both the choline peak and the myo-inositol peak in lower field strengths. The concentration of taurine is strongly age-dependent (12). Taurine levels are elevated in medulloblastoma (8,16).

The continuous wavelet transform

A signal can be decomposed through the wavelet transform into a sum of scaled and translated mother wavelets. Common types of mother wavelets are discussed later in this section. Also, there exist multiple variations of the wavelet transform. These include the CWT, the Discrete Wavelet Transform and the Cross Wavelet Transform (43). As this study uses only the CWT, this is the only variation of the wavelet transform that will be discussed.

Mathematics

The CWT is the sliding dot product of a scaled and translated mother wavelet $\psi^*_{a,b}\left(\frac{t-b}{a}\right)$ with a function $f(t)$. This sliding dot product is also known as a convolution. The CWT is described using Equation 2.3:

$$W_f(a,b) = \int_{-\infty}^{\infty} f(t)\psi^*_{a,b}(t)dt = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} \psi^*_{a,b}\left(\frac{t-b}{a}\right) f(t)dt \tag{2.3}$$

In Equation 3, f(t) is the signal, ψ Is the mother wavelet, a is the scale factor and b is the 'placement', or translation, in the time domain. However, in the case of MR spectra, the time domain is replaced by the ppm domain, which is also a frequency domain. The most commonly used wavelet is the Morlet wavelet (43). Through the ICWT, the wavelet coefficients are again transformed into the spectrum. The ICWT is described using Equation 2.4:

$$f(t) = \frac{1}{C_\psi}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{a}\,\psi\left(\frac{t-b}{a}\right) w(a,b)\,\frac{dadb}{a} \tag{2.4}$$

In Equation 4, C_ψ is the admissibility condition of the wavelet. The admissibility condition must be met in order not to lose information:

$$C_\psi = \int \frac{|\Psi(\omega)|^2}{|\omega|} < +\infty \tag{2.5}$$

In Equation 5, $\Psi(\omega)$ denotes the Fourier transform of a wavelet $\psi(t)$. The admissibility condition implies that the Fourier transform of $\psi(t)$ vanishes at the zero frequency, such that for $\omega = 0$

$$|\Psi(\omega)|^2 = 0 \tag{2.6}$$

Equivalently, the wavelet must be oscillatory, like a wave, and have a zero-integrated area in time-domain, or in this case, ppm-domain:

$$\int \psi(t)\, dt = 0 \tag{2.7}$$

This means that, upon performing an ICWT, the area under the curve of a reconstructed spectrum could result in a shift in baseline (44,45).

Parameters

When performing a wavelet transform, a couple of parameters should be taken into account. Firstly, there is the mother wavelet. Examples include the Morlet, Morse and Mexican hat. Secondly, there are amount of frequency scales *J1* used to describe the signal in the wavelet domain. This can also be seen as the number of different wavelet frequencies. Lastly, there are the distance between the frequency scales dJ and the lowest wavelet scale s_0. These last three are intertwined. The lowest wavelet scale describes the frequency of the base wavelet. If chosen too low in combination with a low amount of frequency scales, the continuous wavelet transform can describe only the high frequencies. If a high value for the lowest wavelet scale is chosen, the continuous wavelet transform can only describe low frequencies. If dJ is chosen too low in combination with a low amount of frequency scales, high frequency waves are not described in the wavelet transform. Therefore, to describe a signal, a sufficient combination of all three is required (45).

Visualisation

The CWT can be visualised in 2D using a scalogram. On the horizontal axis is the ppm range of the original spectrum and on the vertical axis is the frequency range in Hertz of the wavelet. Thus, noise with a high frequency can be found at the top of the scalogram, whereas a low frequency can be found at the bottom of a scalogram. In Figures 2.4 to 2.7, there is a demonstration of the CWT. In the first example, a low-frequent sine wave is used. The amplitudes of the sine wave are seen in the scalogram. The reconstructed sine wave, using the ICWT, and the residual are also plotted. As is visible from Figure 2.4D, after the CWT and ICWT, no significant residual is seen.

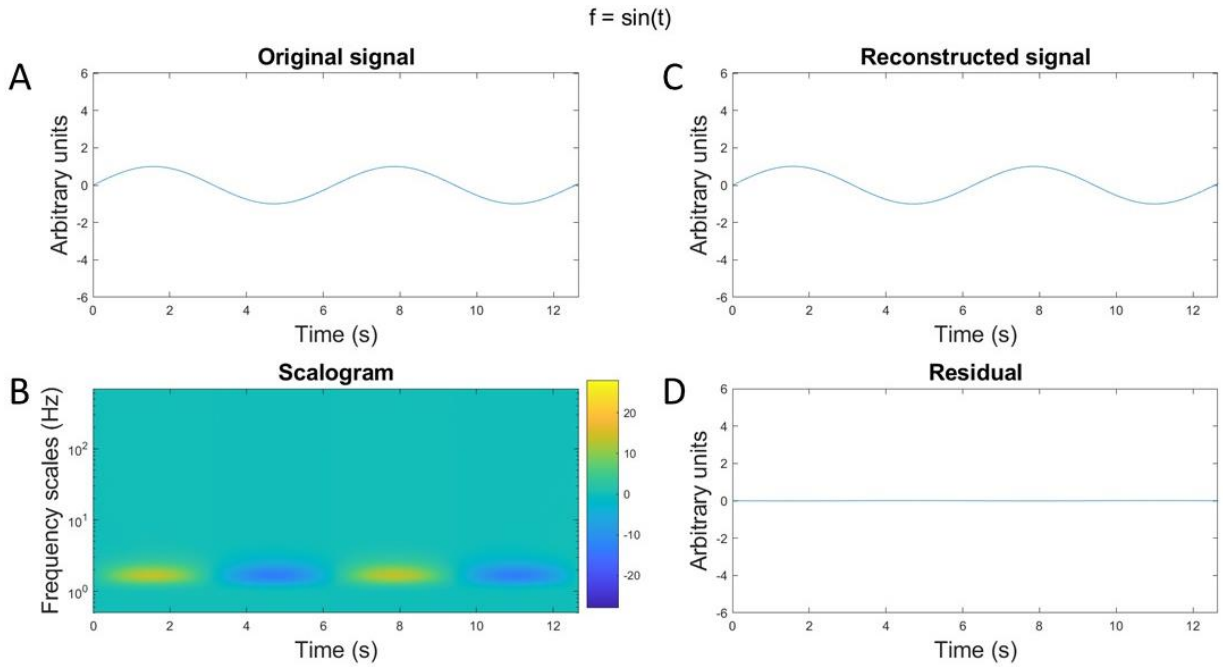


Figure 2.4. A: Original signal, consisting of a low-frequent sine wave. B: Scalogram of this signal. Yellow indicates high wavelet coefficients. Blue indicates low wavelet coefficients. C: Reconstructed signal. D: Residual after reconstructing. No information was added or removed.

A sine wave with a higher frequency and amplitude is now added to the previous example. The result is seen below in Figure 2.5. In Figure 2.5, wavelet coefficients with higher frequencies and stronger coefficients are visualized in the resulting scalogram.

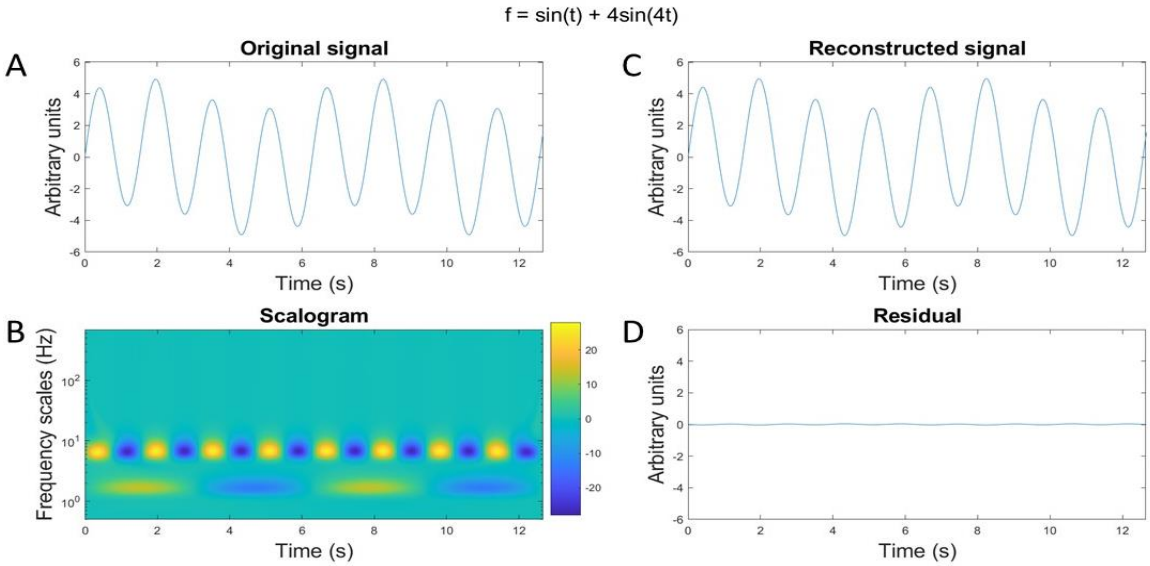


Figure 2.5. A: The original signal, consisting of a low-frequent sine wave and a high-frequent sine wave. B: Scalogram of signal in A. Notice how compared to Figure 2.4B there are now relatively stronger coefficients in a higher frequency. C: Reconstruction of signal in A. D: Residual after reconstruction.

A third addition that can be made, is white noise with an SNR of 3. White noise has an even higher frequency than the second sine wave, so the resulting coefficients are higher in the scalogram than the coefficients of both sine waves. The result can be found in Figure 2.6.

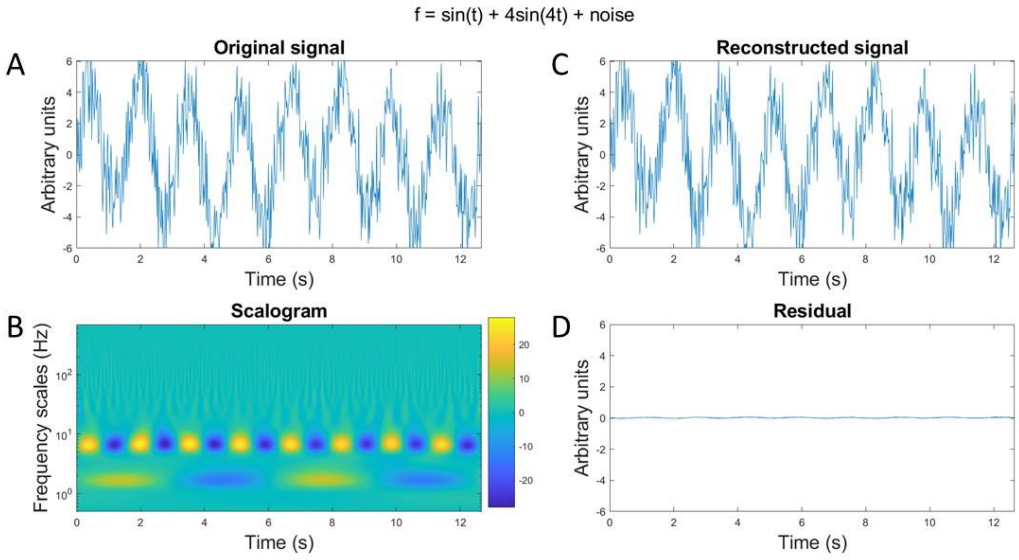


Figure 2.6. A: The original signal, consisting of a low-frequent sine wave, a high-frequent sine wave and noise with an even higher frequency. B: Scalogram of signal in A. See how the high-frequent noise is now above the sine waves. C: Reconstruction of signal in A. D: Residual after reconstruction.

Lastly, a baseline is introduced to the signal. The resulting scalogram in Figure 2.7B is no different from the scalogram in Figure 2.6B. In Figure 2.7, it can be seen in the bottom right tile, a residual is left when comparing the original signal and the reconstructed signal. This is due to the admissibility condition of the wavelet.

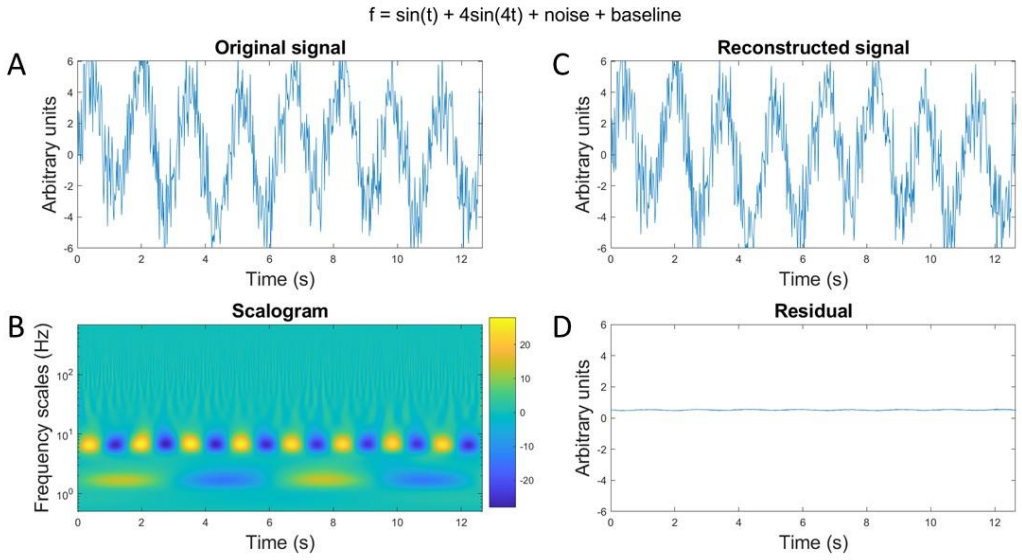


Figure 2.7. A: Signal, consisting of a low-frequent sine wave, a-high frequent sine wave, high-frequent noise and a baseline. B: Scalogram of signal in A. C: Reconstruction of signal using the ICWT. D: Residual after reconstruction. Notice how there still is a baseline left.

Deep learning

Deep learning is a technique that allows artificial neural networks to learn representations of data with multiple levels of abstraction (46). Using these data representations, deep learning models can have a variety of tasks, such as classification (47), segmentation (48) or reconstruction (49). For supervised networks, the networks learn data representation of a certain input by comparing the output of a network to a ground truth. The output of such networks is computed through a series of convolutions, weights and non-linear activation functions. A loss is then computed by comparing the output of the network and ground truth. Using this loss, backpropagation is used to calculate the gradient of this loss with respect to the weights and the weights are updated with the goal that the loss becomes lower. With an optimizing algorithm, such as gradient descent or Adam, the weights are adapted, such that the output is as close to the ground truth as possible, according to the applied loss function (46,50). Besides supervised learning, there is also unsupervised and semi-supervised learning. As this study does not use unsupervised or semi-supervised techniques, they will not be discussed.

Network architectures

In this thesis, three network architectures will be used. These are the convolutional autoencoder, the U-Net and the dilated convolutional neural network. All three networks and the differences between the networks are briefly explained.

Convolutional autoencoder

In essence, a convolutional autoencoder is a series of downsampling convolutions, resulting in a latent space representation of the input, that is then up sampled to the shape of the input through a series

of transposed convolutions Convolutional autoencoders differ from 'standard' autoencoders, since convolutions are used in both the encoding and decoding steps. The advantage here is that a convolutional autoencoder learns image structures better due to the convolutions. (46,51).

U-Net

A U-Net can be regarded as a convolutional autoencoder, with the additional purpose of preserving spatial features. It does so using skip connections. Skip connections combine features from the contracting path with the upsampled output in the expanding path. The result is that image features from multiple resolution levels can be retained. Despite the fact that the first U-Net was designed for segmentation purposes (52), it can also work well for pixel-wise regression purposes (53). In Figure 2.8, the architecture of a U-Net is shown (52).

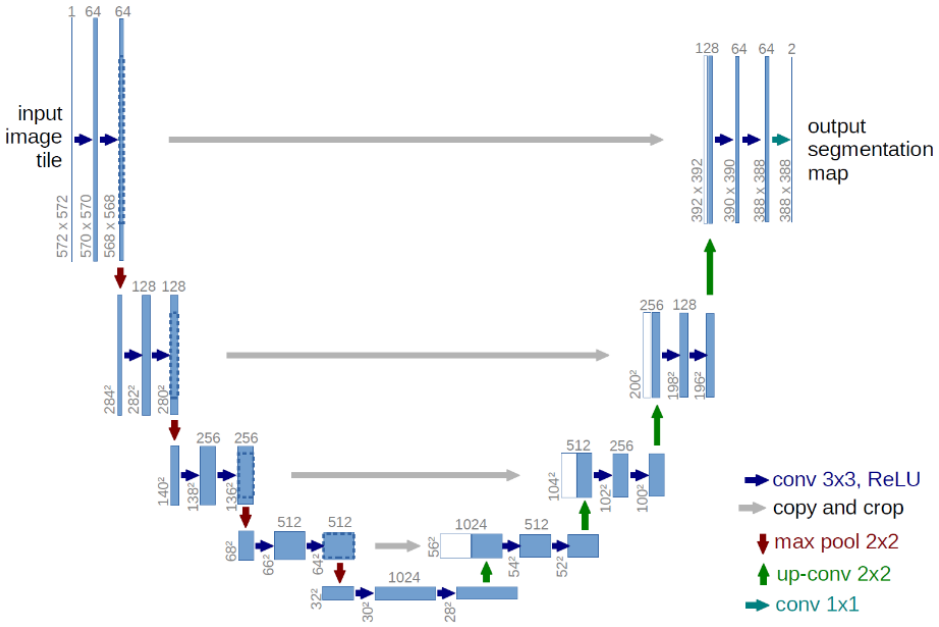
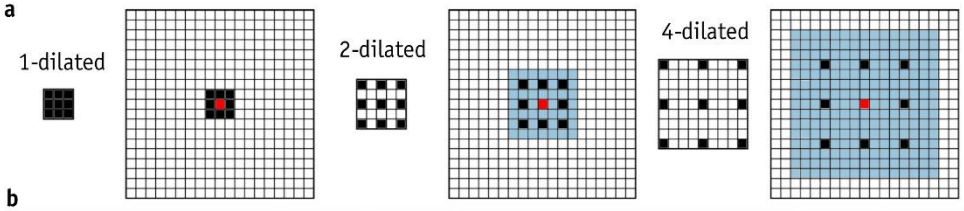


Figure 2.9: Architecture of a U-Net. From Ronneberger et al. (53).

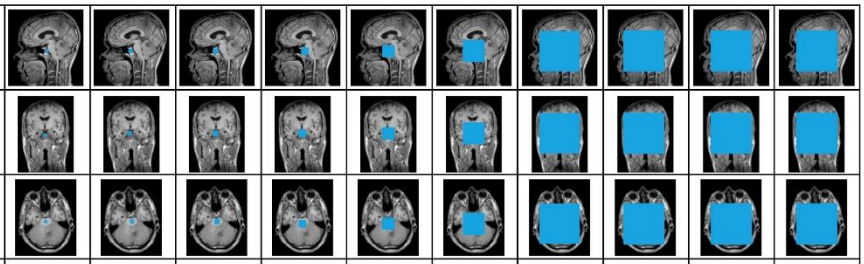
Dilated convolutional neural network

A dilated convolutional neural network stacks convolution layers on top of each other. These convolutions do have an additional parameter in the kernel, namely the dilation parameter. Using dilated convolutions, a bigger context can be learned, compared to non-dilated convolutions. Thus, a bigger part of the image is used to predict the value of a single pixel value. In a study by Wolterink *et al.*, a dilated convolutional neural network was used to synthesize computed tomography (CT) images from MR images. In Figure 2.10A, the dilated convolution is shown. In Figure 2.10b, the network architecture, including the visual representation of the receptive field, is shown (54). The biggest difference between the dilated CNN and the autoencoder and the U-Net is that the dilated CNN only does down convolutions, and there are no upsampling layers. Another difference is that a dilated CNN uses padding before a training or testing sample enters the network, since the output size must match the input size of the spectrum or scalogram to compute the loss.

a



b



Layer	1	2	3	4	5	6	7	8	9	10
Convolution	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	3 x 3	1 x 1	1 x 1
Dilation	1	1	2	4	8	16	32	1	1	1
Field	3 x 3	5 x 5	9 x 9	17 x 17	33 x 33	65 x 65	129 x 129	131 x 131	131 x 131	131 x 131
Channels	32	32	32	32	32	32	32	64	64	1
Parameters	448	9376	9376	9376	9376	9376	9376	18752	2112	65

Figure 2.10. A: Example of a 1-dilated, 2-dilated and 4-dilated convolution. B: Visualisation of receptive field per layer in the network. From Wolterink et al. (55).

3. Methods

The methods developed in this thesis are intended to be used on *in vivo* data. In supervised learning, a pair of spectra is needed, namely an input spectrum (or scalogram) and a target spectrum (or scalogram). The input spectrum contains information of interest and contamination, whereas the target spectrum only contains information of interest. This *in vivo* data will always contain noise, and in some cases an artefact. Clean ground truth data is not available. Thus, the *in vivo* data can not be used for training of neural networks. Consequently, deep learning networks were trained on paired simulated data. Paired data means to have an input, in this case, a contaminated spectrum, and a target, in this case, a clean spectrum, which contains only the metabolites. This data was simulated to match the *in vivo* data as closely as possible.

Acquisition and analysis of *in vivo* spectra

10 patients (8 males, 2 females. Mean age = 6.91 yrs at diagnosis) were retrospectively included from a dataset from the St. Jude Children's Research Hospital. This data was shared with the University Medical Center Utrecht by St Jude Children's Research Hospital per data sharing agreement. All patients were diagnosed with a tumour located in the pons. If histological classification was performed, the World Health Organisation (WHO) grading was a grade III or grade IV tumour.

The MR data were collected using a Siemens 3T MRI system (Siemens, Erlangen, Germany). A voxel grid of 16 × 16 voxels (voxel size was 10mm × 10mm × 15mm (LR × AP × CC) was placed over a transverse slice containing the tumour. Spectra were acquired using a PRESS sequence (TE/TR = 135ms/1700ms, spectral width = 1200 Hz, number of points = 1048).

From the spectra, the SNR and linewidth were calculated using the op_getSNR and op_getLW functions from the FID Appliance (FID-A) software (55) and the acquisition parameters were copied for the simulations.

Simulation of spectra

In MATLAB 2021B (MathWorks, Natick, MA, USA), 8000 spectra were simulated using the FID-A software. The spectra were simulated with a PRESS sequence, TE=135ms, linewidth=6 Hz, spectral width=1200 Hz, B_0=3T, number of datapoint=1024. The simulated metabolites and concentration ranges can be found in Table 3.1. The concentrations were simulated based on concentrations of metabolites in healthy brain tissue (29) using a uniform distribution, except Cr, GPC, NAA and PCh. For Cr, a wider range was used than is typically found in healthy tissue since it varies more in diseased brain tissue, for GPC and PCh, higher concentrations were simulated than found in healthy tissue and for NAA the concentrations are simulated.

Table 3.1: Ranges of metabolite concentrations.

Metabolite	Lower limit (mmol L^{-1})	Upper limit (mmol L^{-1})
Ala	0.1	1.5
Asp	1.0	2.0
Cit	0.2	0.6
Cr	3.5	11.5
GPC	0.5	3.0
GSH	1.5	3.0
Gln	3.0	6.0
Glu	6	12.5
Ins	4	9
Lac	0.2	1.0
NAA	5.5	15
NAAG	0.5	2.5
PCh	0.5	3.0
PCr	3.0	5.5
Tau	2.0	6.0

The simulated spectra were obtained by linear addition of spectra of single metabolites and were stored in three channels, namely the real part of the spectrum, the imaginary part of the spectrum and the magnitude of the spectrum. The spectra in the three channels were resized from shape 468,1 to shape 512,1 for the purpose of being used in the deep learning networks.

Thereafter, all the clean spectra were contaminated with noise and baseline drift. 1000 of these spectra were contaminated with a lipid artefact. Another 1000 were contaminated with a residual water artefact. The noise that was added was white Gaussian noise, such that the signal-to-noise ratio (SNR) of the spectra was between 5 and 7. All spectra received a baseline shift of 0-1.5 arbitrary units to simulate baseline effects. The lipid artefact usually originates from an adjacent voxel with high lipid concentrations and propagates to another voxel by the point spread function. This point spread function is responsible for the spatial mixing of signals (21). To generate a lipid artefact, a spin system was first built for lipids, with resonances and relative peak heights based on a study by Hamilton *et al.* (56). The lipid artefact was generated using a linewidth of 40 Hz and a concentration of 250-500 mmol L^{-1}, such that it resembles a typical outer volume lipid artefact found in chemical shift imaging. The residual water peak is the result of imperfect suppression. The most common water suppression technique is called chemical shift selective saturation (CHESS). CHESS selectively excites and dephases water magnetization. However, CHESS is sensitive to B_1 inhomogeneities since imperfect 90^o excitation leaves partial longitudinal water magnetization, thereby resulting in a residual water artefact (57). To mimic a residual water artefact, a water peak was simulated with broader linewidth due to dephasing (lw = 72 Hz) and a significantly lower than *in vivo* concentration due to longitudinal magnetization suppression (concentration = 200 to 500 mmol L^{-1}). The contaminated spectra were also stored in a real, imaginary and magnitude channel.

To evaluate the performance of the networks on different noise levels, different amplitudes of a residual water artefact and different amplitudes of a lipid artefact, three additional test sets were synthesized. Each additional test set consisted of 100 spectra and corresponding scalograms. These three sets were not used during training, but only after training to specifically test performance on increasingly strong contamination. The first additional test set was used to test the capacity of the networks to reduce noise. The spectra were contaminated with noise with an SNR between 3 and 7.

The second test set was used to test the capacity of the networks to remove residual water signal. This test set contained spectra with water concentrations between 0 and 700 mmol L^{-1}. The third test set contained spectra with lipid contamination, with increasing concentration from 0 to 750 mmol L^{-1}.

In both the test set with increasing residual water artefacts and the test set with the increasing lipid artefact, an SNR of 6 was used. The SNR was not varied in these sets, since it could influence the resulting fitting.

Wavelet analysis

In order to address the potential value of the wavelet analysis, the clean spectrum and corresponding contaminated spectrum were transformed using the continuous wavelet transform. A prerequisite is that the information loss after performing a continuous wavelet transform and inverse continuous wavelet transform is minimalized. As discussed in the Background section, a residual baseline may be there due to the admissibility condition, as long as it is a flat baseline. However, using Matlab's default settings, performing a continuous wavelet transform and inverse continuous wavelet transform, the residual is not a flat baseline, as can be seen in Figure 3.1.

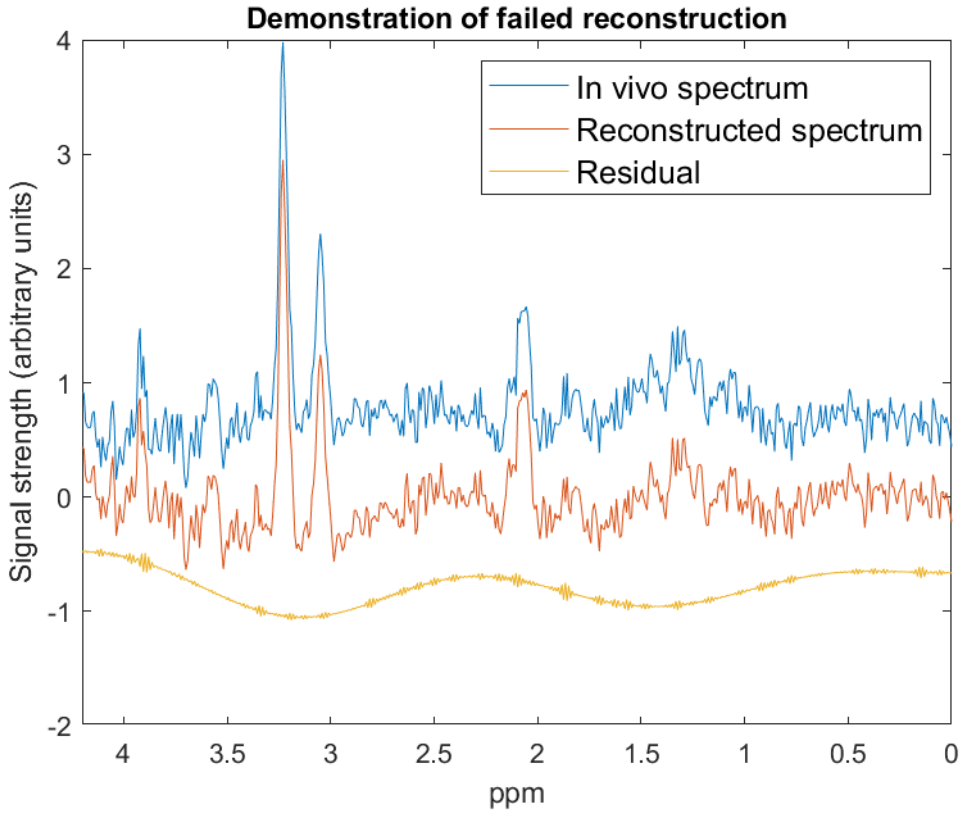


Figure 3.4: Original spectrum (blue) and reconstructed spectrum (red). The residual is not a flat line. Thus, the reconstruction has failed.

The Matlab function in version 2021B lacks the option to input arguments for the inverse wavelet transform. Therefore, the functions by Jon Erickson (Jon Erickson (2022). continuous wavelet transform and inverse (https://www.mathworks.com/matlabcentral/fileexchange/20821-continuous-wavelet-transform-and-inverse), MATLAB Central File Exchange. Retrieved October 26, 2022) were used, as these did grant this option.

Ideally, the standard deviation of the residual between the spectrum and the reconstruction is 0. This would mean that the residual is a straight line, and no information, except a baseline, was lost. In order

to achieve this, a grid search was performed over the following parameters: the number of frequency scales $J1$ (128,256,512,1048,2048), the lowest wavelet scale s_0 (0.25,0.5,0.75,1,1.25), the spacing between the frequency scales dJ (0.5, 0.25, 0.05, 0.025, 0.005) and the wavelet parameter (2, 4, 6, 8). The mother wavelet was not chosen as one of the parameters since the literature advises to use the Morlet wavelet (43). The combination of parameters that lead to the lowest residual standard deviation were considered the optimal parameters. The optimal parameters were used for the final continuous wavelet transform to obtain scalograms.

Deep learning

1D networks

The networks that used the corrupted one-dimensional (1D) spectra as input and predicted clean spectra were trained in Python (version 3.8.13, Python Software Foundation, Wilmington, DE) using PyTorch (version 1.12.1) (58). Networks that were trained were a convolutional autoencoder, a U-Net and a dilated CNN. The convolutional autoencoder and the U-Net were built using code from Medical Open Network for Artificial Intelligence (MONAI, version 0.9.1). The dilated CNN was based on code by Blanken *et al.* (59). All three networks were trained with three channels, namely the real part of the spectrum, the imaginary part of the spectrum and the magnitude part of the spectrum.

1D convolutional autoencoder architecture

A 1D convolutional autoencoder was trained with 5 convolution layers. The filters per layer were 8, 16, 32, 32 and 32. The kernel size was 3. A dilation factor of 2 was used to increase the receptive field of the convolutional autoencoder. A stride of 2 was used. To upsample, transposed convolutions were used with kernel size 3. As activation function, the Parametric Rectified Linear Unit (PReLU) was used. The number of trainable parameters in the 1D convolutional autoencoder was 41465. The output of this network has shape 512,1.

1D U-Net architecture

A 1D U-Net was trained with 6 convolution layers. Each layer in the network had 32 filters. In the downsampling part of the network, the kernel size was 3 and the stride was 2. In the upsampling part of the U-Net, transposed convolutions with kernel size = 3 were used. As activation function, the PReLU was used. The number of trainable parameters in the 1D U-Net was 41773. The output of this network has shape 512,1.

1D dilated CNN architecture

A 1D dilated CNN was trained with 9 convolution layers and 64 filters per layer. Prior to training, reflection padding was used so the shape of the output of the network would match the target shape. Reflection padding pads the input tensor using the reflection of the input boundary. An example of reflection padding is shown below in Figure 3.2. The kernel size in this network was 3 and the Rectified Linear Unit (ReLU) was used as activation function. The number of trainable parameters in the 1D dilated CNN was 100803. The output of this network has shape 512,1.

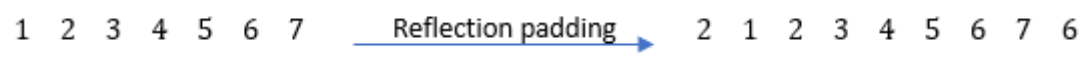


Figure 3.5: Example of reflection padding.

Training of 1D networks

The dataset consisting of 8000 spectra was split into a training and a testing set using an 80/20 split. Thus, the networks were trained on 6400 samples and tested on 1600 samples. The split was stratified such that the training set contained 800 spectra containing a lipid artefact and 800 spectra containing

a residual water artefact. Batch normalization was used in all three networks. The networks were trained using an Adam optimizer, using an L1 loss, a learning rate of 0.0001 and a batch size of 4. The convolutional autoencoder was trained for 50 epochs, the U-Net was trained for 200 epochs and the dilated CNN was trained for 200 epochs. During training, both the loss on the training set and the loss on the validation set were computed. The convolutional autoencoder was trained with fewer epochs to prevent overfitting the training data.

Testing of 1D networks

Next to testing om the 1600 spectra from the dataset, the 1D networks were also tested on the three additional test sets. These test sets contained spectra with increasing noise levels, an increasingly strong residual water artefact and an increasingly strong lipid artefact. Lastly, the networks were also tested on 10 *in vivo* spectra.

2D networks

The networks that used the scalograms of the corrupted spectra as input and predicted scalograms of clean spectra were trained in Python (version 3.9.2) using PyTorch (version 1.12.1) (58). Networks that were trained were a convolutional autoencoder, a U-Net and a dilated CNN. The convolutional autoencoder and the U-Net were build using code from MONAI and the dilated CNN was based on code by Blanken *et al.* (59).

2D convolutional autoencoder architecture

A 2D convolutional autoencoder was trained with 5 convolution layers. The number of filters per layer were 32, 32, 32, 64 and 64. The kernel size was 3×3. A dilation factor of 2 and a stride of 2 were used in each layer. To upsample, transposed convolutions were used with kernel size 3×3. The PReLU was used as activation function. The number of trainable parameters in the 2D convolutional autoencoder was 149961. The output of this network has shape 512,2048.

2D U-Net architecture

A 2D U-Net was trained. It had 5 convolution layers, with 32 filters per layer. In the downsampling part of the network, the kernel size was 3×3. The stride was 2 in each layer. A kernel size of 3×3 was used in the upsampling part of the network. The PReLU was used as activation function. The number of trainable parameters in the 2D U-Net was 123373. The output of this network has shape 512,2048.

2D dilated CNN architecture

A 2D dilated CNN was trained with 5 convolution layers and 32 filters per layer. Prior to training, reflection padding was used so the shape of the output of the network would match the target scalogram. The kernel size in this network was 3×3 and the Rectified Linear Unit (ReLU) was used as activation function. The number of trainable parameters in the 2D dilated CNN was 38307. The output of this network has shape 512,2048.

Training of 2D networks

The dataset consisting of 8000 scalograms was split in a training and a testing set again using an 80/20 split. Thus, the networks were trained on 6400 scalograms and tested on scalogram samples. The split was stratified such that the training set contained 800 scalograms containing a lipid artefact and 800 scalograms containing a residual water artefact. Batch normalization was used in all three networks. The networks were trained using an Adam optimizer, a learning rate of 0.001 and a batch size of 4. Initially, these networks were trained on an L1 loss. However, this loss was not able to fully capture some of the structural details in the scalogram, resulting in reconstructed spectra that did not contain the details of the target spectra. Therefore, the structural similarity index measure (SSIM) loss was also used. However, this resulted in scalograms with correct structures, yet inaccurate values for the

wavelet coefficients. This resulted in spectra with too low peaks. Consequently, in the end, a sum of the L1 loss and the SSIM loss was used to compute the loss. The convolutional autoencoder, U-Net and dilated CNN were all trained for 80 epochs. During training, both the loss on the training set and the loss on the validation set were computed.

Testing of 2D networks

The 2D networks were tested on the 1600 scalograms from the validation set, and on the three additional test sets. These additional test sets contain the scalograms of spectra with increasing noise, residual water artefact and lipid artefact. Scalograms from the *in vivo* spectra were also tested.

Reconstruction of spectra

Spectra were reconstructed from the outputted scalograms from the 2D networks using the ICWT. The parameters that came out of the optimization grid search that was performed earlier were used in the reconstruction.

Residual analysis

To compare performance between the networks, A residual analysis was performed. This residual analysis was performed on both the 1600 spectra that were used for validation during training, as well as on the 3 additional test sets. First, the standard deviation of the difference between prediction and ground truth was calculated as a gross estimation of prediction accuracy. The formula for standard deviation can be found in Equation 3.1.

$$SD = \sqrt{\frac{\sum|x - \bar{x}|^2}{n - 1}} \tag{3.1}$$

In Equation 3.1, n is the number of points in x, x are the values in the residual and $\bar{x}$ is the mean value of the residual.

Next to the standard deviation of the residual, peak height differences were calculated for NAA, tCr and tCho. To find the peak heights, the maximum value in the ppm range of 1.97-2.08 ppm was computed for NAA, the maximum value in the range of 2.98-3.08 ppm was computed for tCr and the maximum value in the range of 3.15-3.28 ppm was computed for tCho. For each of the spectra, the absolute percentage error was calculated per metabolite, using Equation 3.2.

$$Absolute\ percentage\ error = 100\% \times \frac{\left|peak\ height_{predicted} - peak\ height_{Ground\ truth}\right|}{peak\ height_{Ground\ truth}} \tag{3.2}$$

To check if the differences between standard deviation and peak height differences were significantly different, a Student t-test was performed. A P value of < 0.05 was considered statistically significant. Bonferroni correction was used to decrease risk of a type I error, which can occur when performing multiple statistical tests (60). All statistical analysis was performed in Matlab.

In vivo spectra

The outputted *in vivo* spectra were first inspected visually to check if the output was trustworthy. The height of the creatine, choline and NAA peaks were used as reference points. Also, it was inspected if the spectra showed unexpected large peaks. If the reference points visibly matched and there were no unexpected large peaks, the *in vivo* spectra were deemed trustworthy.

4. Results

Simulated spectra and wavelet analysis

8000 clean and contaminated spectra were successfully simulated to train and validate the network. The grid search for the optimal parameters for the CWT and ICWT yielded a standard deviation of 0.0023 arbitrary units. The resulting parameters were the number of frequency scales $J1$ = 2048, the lowest wavelet scale s_0 = 0.5 Hz, spacing between the frequency scales dJ = 0.005 Hz and wavelet parameter = 4. In Figure 4.1, an *in vivo* spectrum has been transformed with the CWT using the mentioned parameters and has thereafter been inversely transformed with the ICWT, resulting in the reconstructed spectrum. As can be seen from the residual, only minimal information was added to the spectrum.

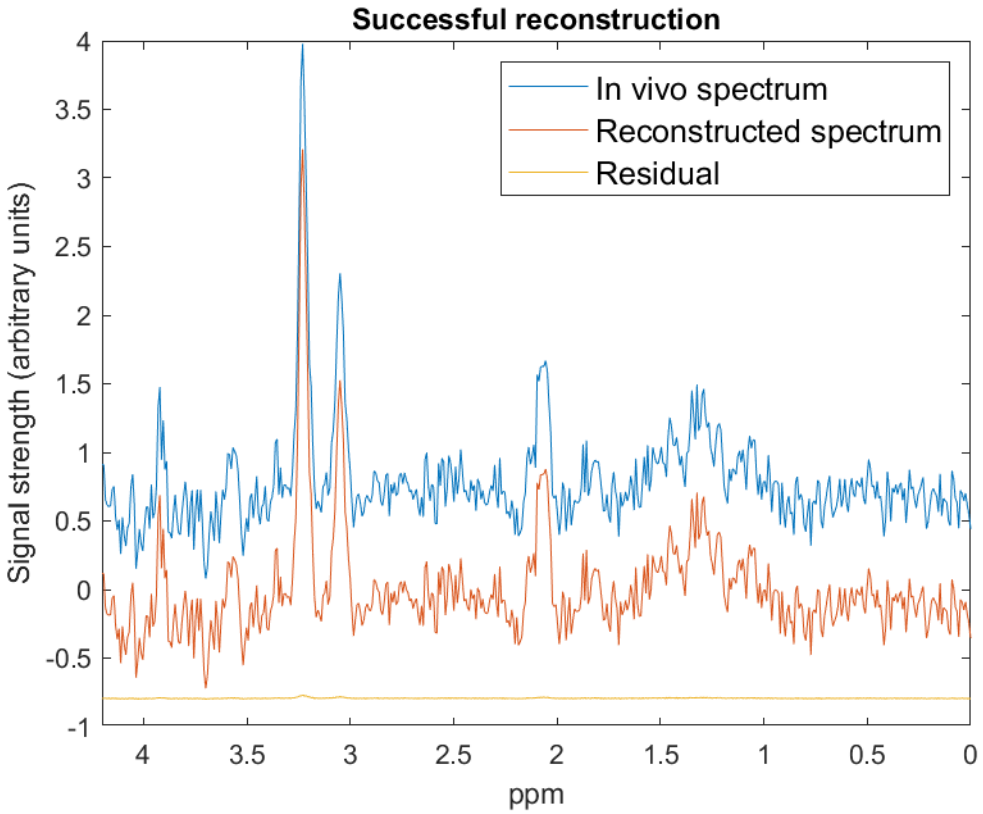


Figure 4.1: In blue is an in vivo MR spectrum. In orange is the reconstructed spectrum. In yellow is the difference between the in vivo spectrum and the reconstructed spectrum.

The three additional test sets, containing increasing noise, lipid artefact and residual water artefact contamination were also simulated successfully. Of all spectra, scalograms were also acquired. Figures 4.2, 4.3 and 4.4 show an example of a clean spectrum, a contaminated spectrum and the respective scalograms of these spectra. The examples are of a noise-contaminated, residual water-contaminated and lipid-contaminated spectrum, respectively.

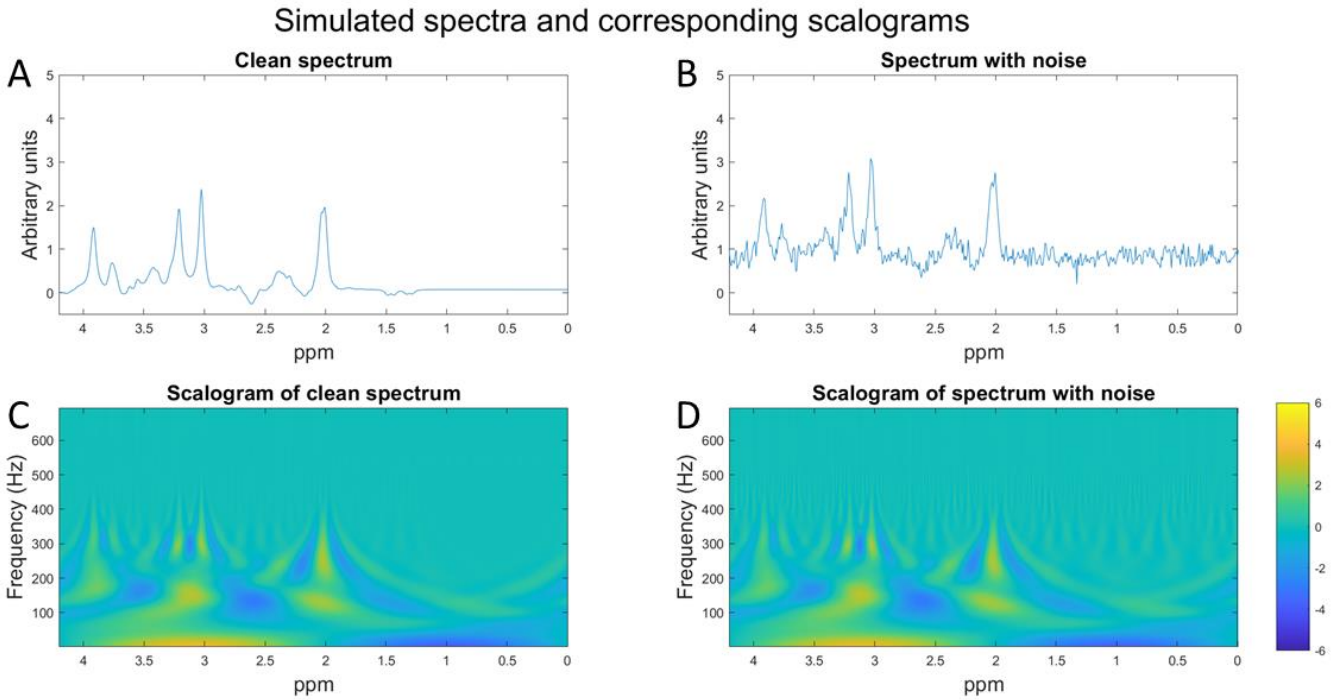


Figure 4.2. A: Simulated MR spectrum containing only metabolites. B: Simulated MR spectrum with noise and baseline shift. C: Scalogram of spectrum in A. D: Scalogram of spectrum in C. Notice how there are now stronger wavelet coefficients, most notably between 300 and 400 Hz.

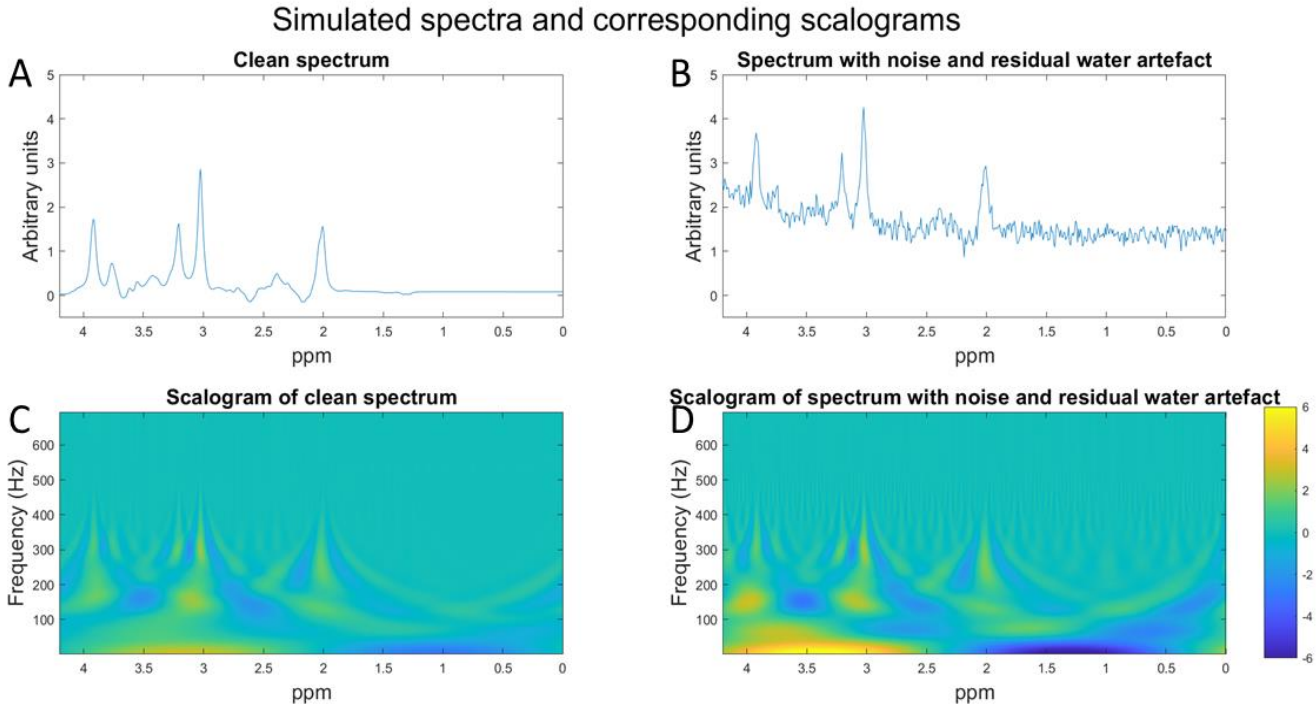


Figure 4.3. A: Simulated MR spectrum containing only metabolites. B: Simulated MR spectrum with noise, baseline drift and a residual water artefact. The baseline shift and the right tail of the water artefact result in a higher baseline. C: Scalogram of spectrum in A. D: Scalogram of spectrum in B. Notice that there are now strong coefficients at the low frequencies, which describe the low-frequent residual water artefact.

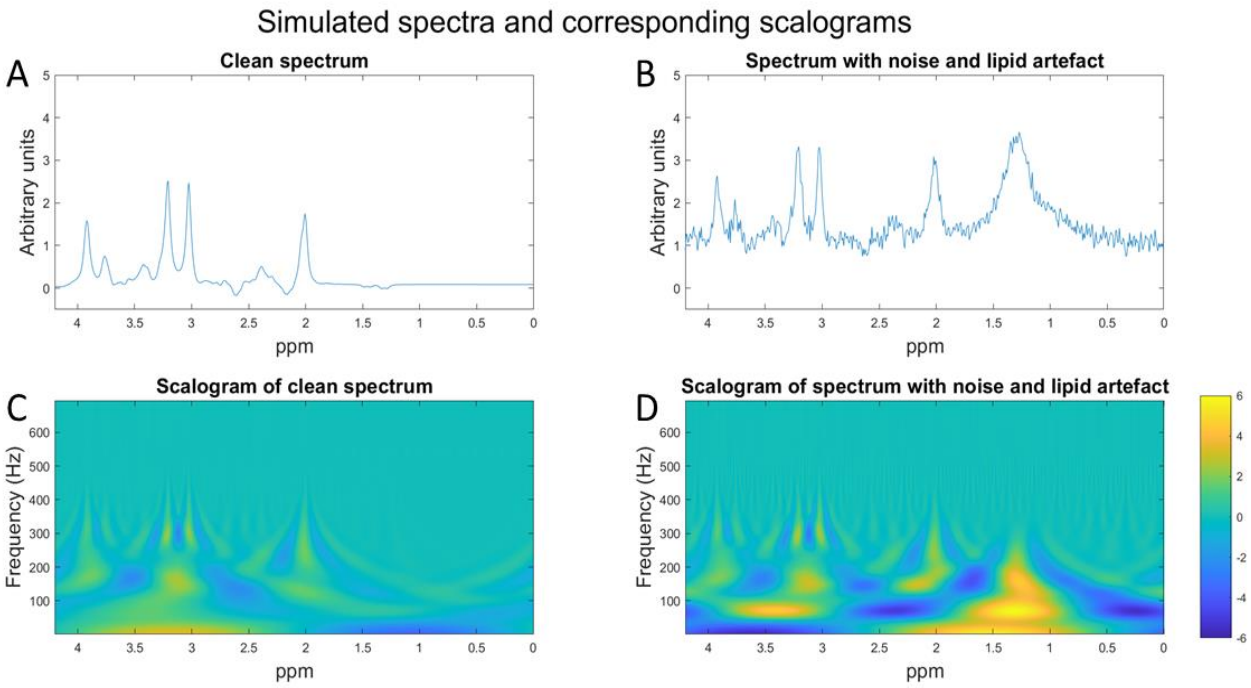


Figure 4.4. A: Simulated MR spectrum containing only metabolites. B: Simulated MR spectrum with noise, a baseline shift and a lipid artefact. C: Scalogram of spectrum in A. D: Scalogram of spectrum in C. Scalogram of spectrum in B. The lipid artefact results in strong coefficients in the 1-1.5 ppm range.

Deep Learning

Six networks were trained and validated after each epoch with the goal to perform both noise reduction and artefact removal. The loss on both the training set and the validation set can be found in Table 4.1. From the networks that were optimized on the spectra, the dilated CNN had the lowest loss on the validation set. From the networks that were optimized on the scalograms, the convolutional autoencoder had the lowest loss on the validation set.

Table 4.1: Training losses of the networks. The losses of the 1D networks have been calculated on the spectra. The losses of the 2D networks have been calculated on the scalograms. Losses cannot be compared one-on-one, since the 2D networks use a summation of two loss functions.

Network	Type of loss function	Training loss	Validation loss
1D convolutional autoencoder	MAE	0.019	0.041
1D U-Net	MAE	0.016	0.040
1D dilated CNN	MAE	0.017	0.023
2D convolutional autoencoder	MAE + SSIM	0.043	0.097
2D U-Net	MAE + SSIM	0.141	0.146
2D dilated CNN	MAE + SSIM	0.124	0.126

Simulated data

After training, the spectra returned by the 1D networks were inspected qualitatively prior to quantification. Figure 4.5 shows an example of how the 1D dilated CNN outputs a spectrum contaminated with a residual water artefact. In Figure 4.6, a network output can be found when the input contained a spectrum that was contaminated with noise and lipid artefact. This output was given

by the U-Net. When comparing the residual in Figures 4.5C and 4.6C, the prediction of the network on the spectrum with a residual water artefact is more accurate.

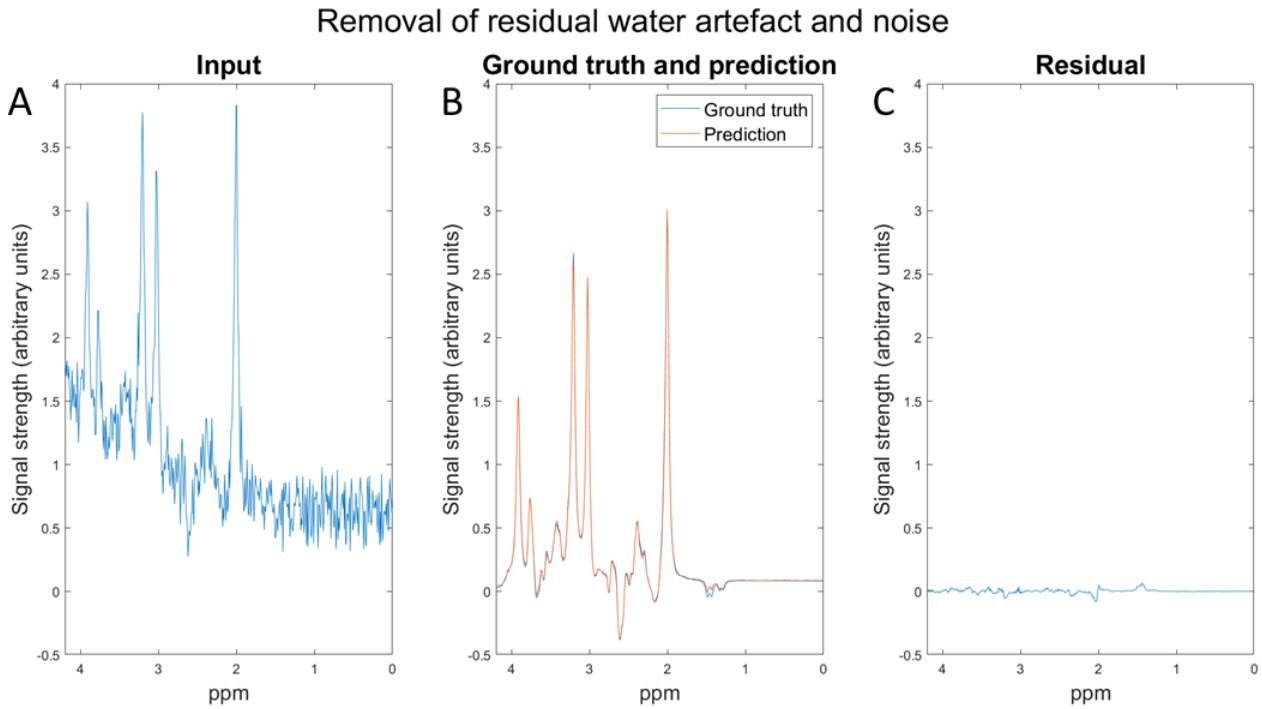


Figure 4.5. A: Spectrum contaminated with noise and residual water artefact. B: Ground truth (b) and prediction of 1D dilated CNN (r). C: Difference between ground truth and prediction.

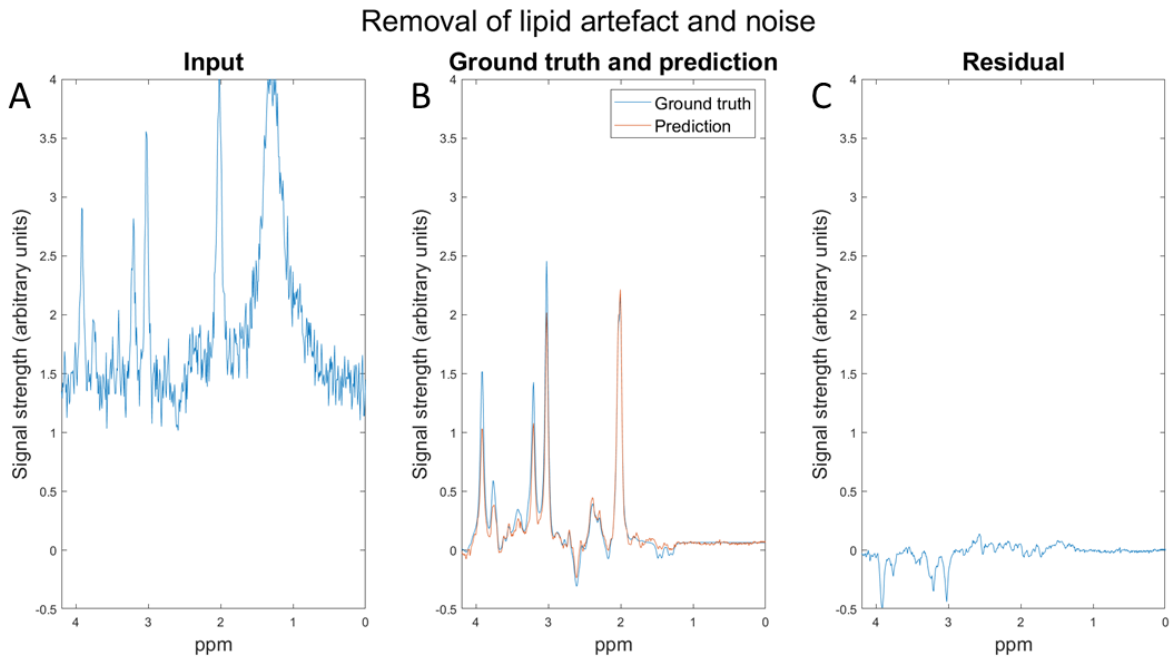


Figure 4.6. A: Spectrum contaminated with noise and lipid artefact. B: Ground truth (b) and prediction of 1D U-Net (r). C: Difference between ground truth and prediction.

For the 2D networks, the output scalograms were compared prior to performing the ICWT. Before training on the summation of the MAE and SSIM loss, they were used individually. In Figure 4.7, an example is visualized after optimizing a U-Net on only the MAE. In Figure 4.8, an example is visualized after optimizing on only the SSIM loss. All results hereafter were obtained on a summation of the MAE and the SSIM loss.

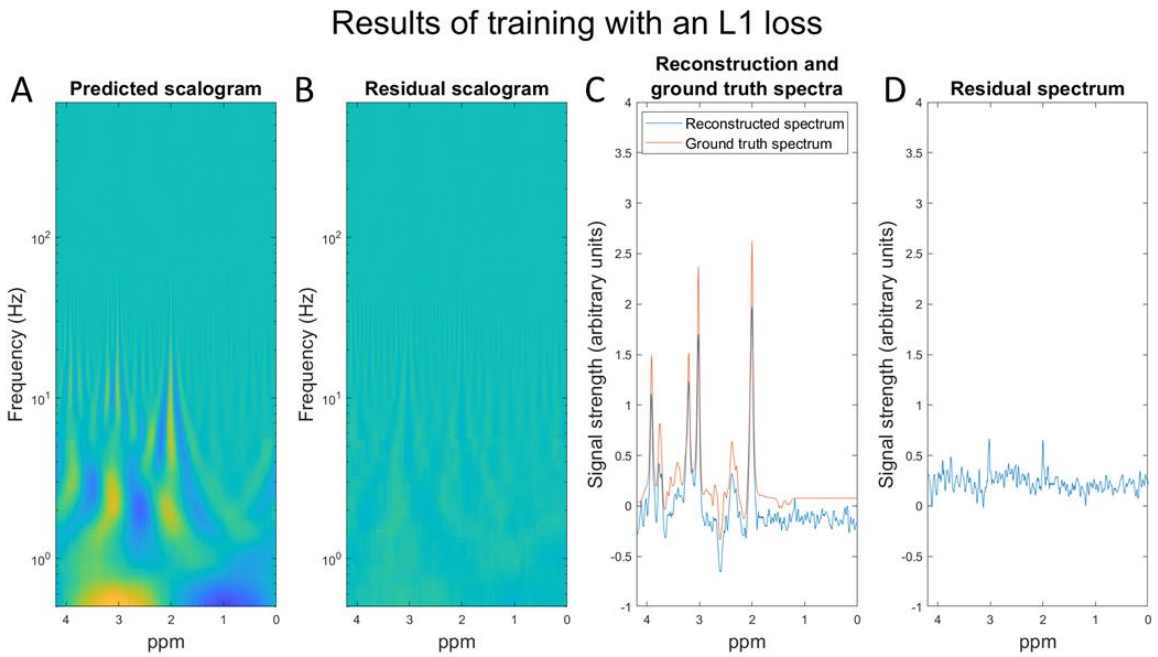


Figure 4.7. A: Predicted scalogram for a U-Net trained on an L1 loss. B: Difference between prediction and ground truth. C: Ground truth spectrum (r) and reconstructed spectrum (b). D: Residual spectrum. Not how there still are low-frequent rolls and noise left.

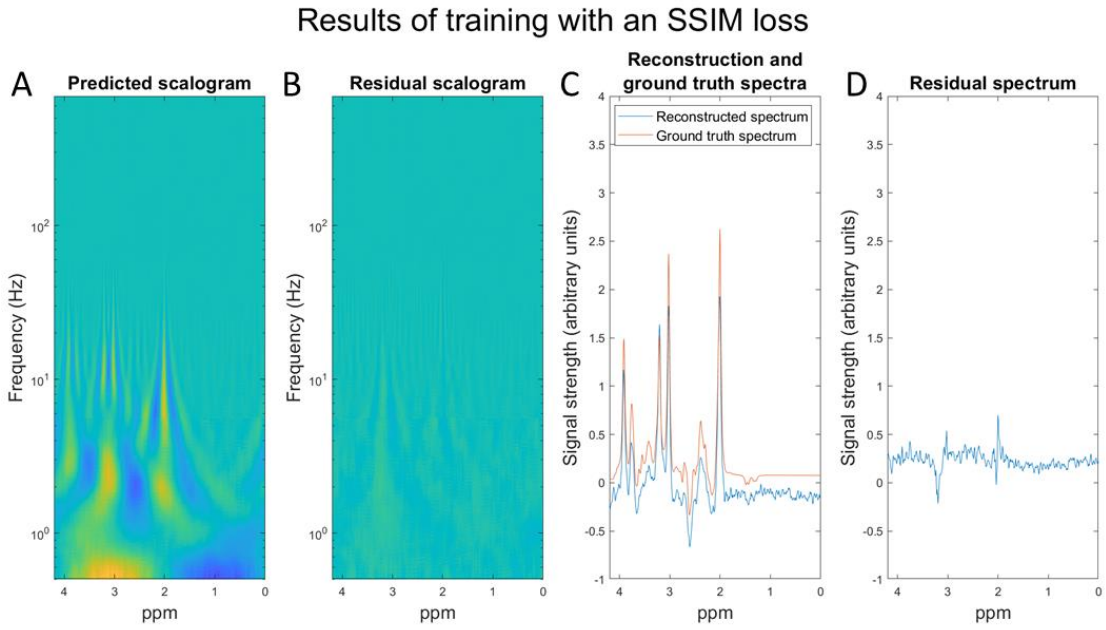


Figure 4.8. A: Predicted scalogram when U-Net was trained on an SSIM loss. B: Difference between predicted scalogram and ground truth scalogram. C: Reconstructed spectrum (b) and ground truth spectrum (r). D. Residual spectrum. Compared to the L1 loss, there is less noise. However, the differences in peak height are bigger due to the wavelet coefficients not being strong enough.

Figure 4.9 shows an example of how the 2D convolutional autoencoder outputs a scalogram of a clean spectrum when the input was a scalogram of a spectrum containing a residual water artefact. In Figure 4.10, an example can be found of how the 2D dilated CNN outputs a scalogram when presented with a scalogram that contains a lipid artefact.

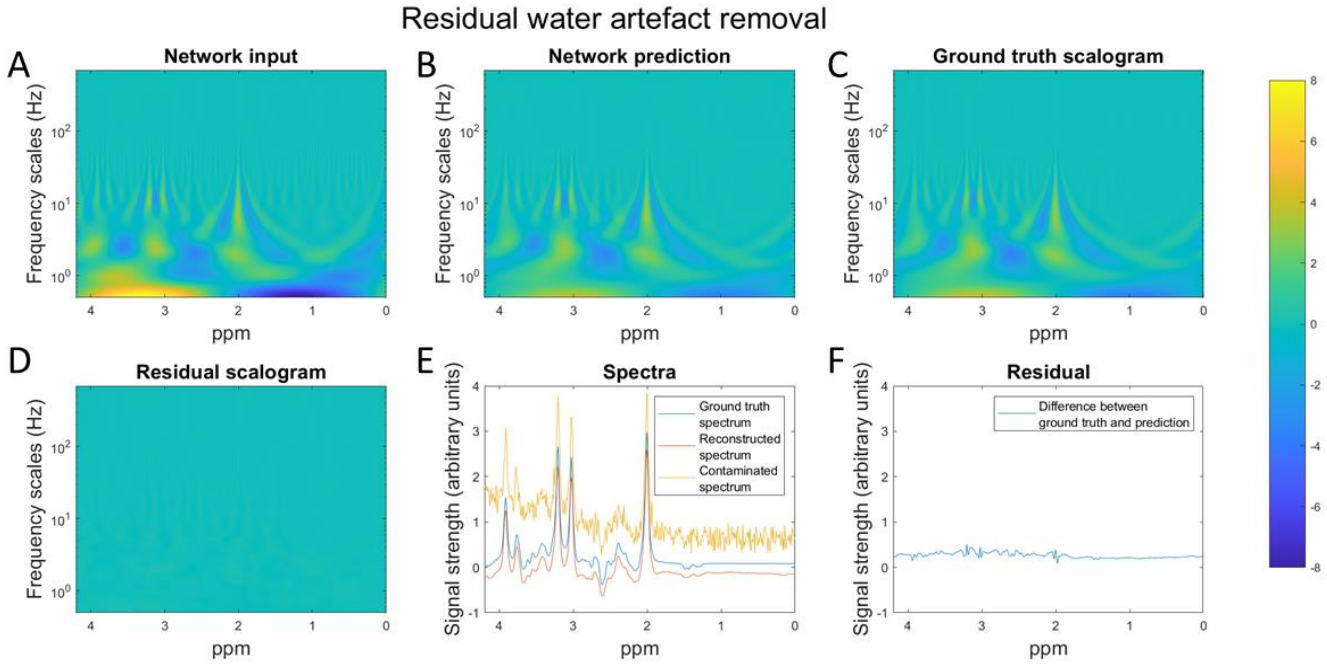


Figure 4.9. A: Scalogram of a spectrum that contains noise and a residual water artefact. B: Prediction of convolutional autoencoder. C: Ground truth scalogram. D: Difference between network prediction and ground truth scalogram. E: Spectrum that contains noise and a residual water artefact (y), ground truth spectrum (b), the reconstructed spectrum from predicted scalogram (r). D: Difference between ground truth spectrum and reconstructed spectrum. Due to the network not accurately predicting the scalogram, there are peaks visible in the residual.

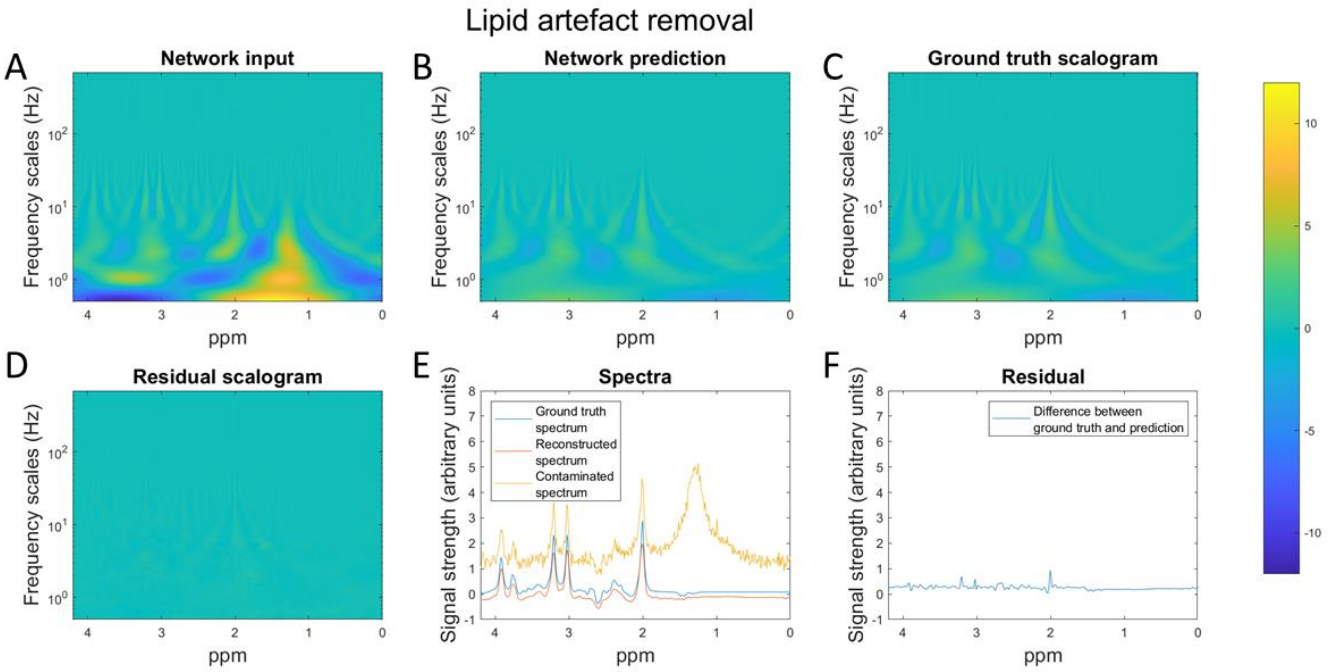


Figure 4.10. A: Scalogram of a spectrum that contains noise and a lipid artefact. B: Prediction of the dilated CNN. C: Ground truth scalogram. D: Difference between prediction and ground truth. E: Ground truth spectrum (b), reconstructed spectrum (r) and lipid contaminated spectrum (y). F: Difference between ground truth spectrum and reconstructed spectrum. Due to the network not accurately predicting the scalogram, there are peaks visible in the residual.

Residual analysis of simulated spectra

During training, 1600 spectra and scalograms were used to validate the network to prevent overfitting. After training, the 1600 spectra from the 1D networks and the reconstructed spectra from the 2D networks were compared based on the standard deviation of the residual, mean peak height difference for the NAA peak, mean peak height difference for the tCr peak and mean peak height difference for the tCho peak. Table 4.2 shows the performance of the networks for these metrics.

Table 4.2. Summary of SD of residual and mean peak height difference of NAA, tCr and tCho for all six networks. The mean peak height difference is in percentage.

	Average SD of the residual (sd)	Mean peak height difference NAA in percentage (sd)	Mean peak height difference tCr in percentage (sd)	Mean peak height difference tCho in percentage (sd)
Noisy spectrum	0.2353 (0.2133)	41.46 (23.91)	40.53 (23.92)	45.49 (28.11)
1D autoencoder	0.0650 (0.0192)	2.83 (2.11)	11.41 (5.92)	14.53 (6.63)
1D U-Net	0.0659 (0.0188)	2.39 (1.88)	12.55 (4.56)	13.11 (5.82)
1D dilated CNN	0.0557 (0.0218)	4.49 (5.82)	13.57 (6.53)	12.48 (6.52)
2D autoencoder	0.0565 (0.0261)	5.22 (4.40)	9.54 (5.74)	4.61 (3.39)
2D U-Net	0.0842 (0.0156)	12.51 (4.96)	12.67 (7.65)	15.45 (10.66)
2D dilated CNN	0.0741 (0.0182)	12.94 (7.47)	15.01 (9.09)	18.53 (12.44)

To test significance, paired t-tests were performed between the SD of the residual and peak height differences of NAA, tCr and tCho of the noisy spectrum and all six networks. Bonferroni correction was used. Therefore, differences were significant for $p < 0.0024$. The mean SD of the residual, mean peak height difference for NAA, mean peak height difference for tCr and mean peak height difference for tCho all differed significantly between the networks, with exception of the SD of the residual between the 1D dilated CNN and the 2D convolutional autoencoder, and the peak height difference of tCr between the 1D U-Net and the 2D U-Net.

Noise reduction capacity

After testing the network's performance on the validation set, the three test sets, containing noise-contaminated, residual water artefact-contaminated, and lipid artefact-contaminated spectra, were individually tested. The first set that was tested contained spectra with increasing SNR. The results based on the SD of the residual on the noise-contaminated spectra can be found in Figure 4.11. The results based on the mean peak height difference for the NAA peak, the tCr peak and the tCho peak can be found in Figure 4.13.

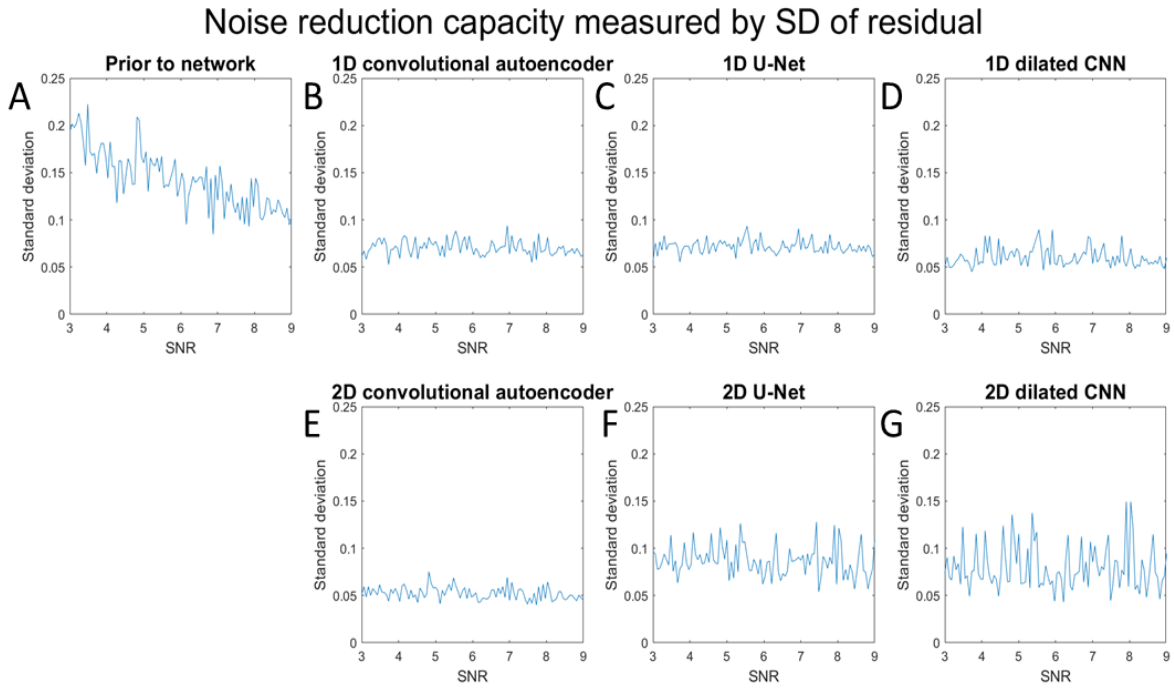


Figure 4.11. A: Standard deviation of residual before denoising. B: Standard deviation of residual after denoising by 1D convolutional autoencoder. C: Standard deviation of residual after denoising by 1D U-Net. D: Standard deviation of residual after denoising by 1D dilated CNN. E: Standard deviation of residual after denoising by 2D convolutional autoencoder. F: Standard deviation of residual after denoising by 2D U-Net. G: Standard deviation of residual after denoising by 2D dilated CNN.

Using a paired t-test, the mean standard deviation of the residual between all networks was significantly different. Bonferroni correction was used. Thus, all significant differences were found with $p < 0.0024$. This is illustrated in Figure 4.12.

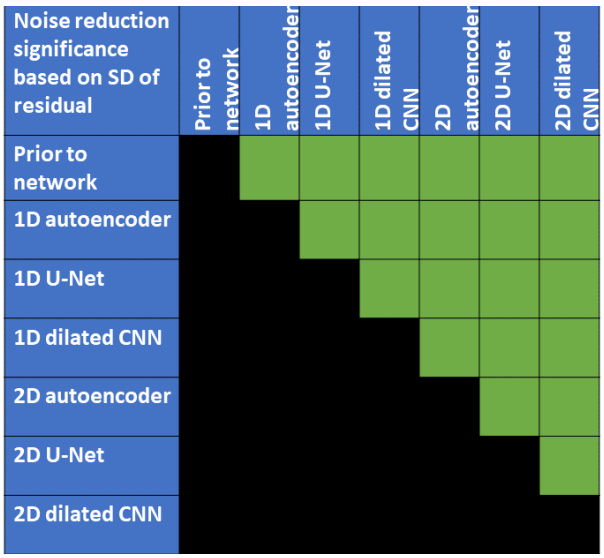


Figure 4.12: The average standard deviation of the residual was significantly different for each of the networks. Green indicates a significant difference. Red indicates there was no significant difference.

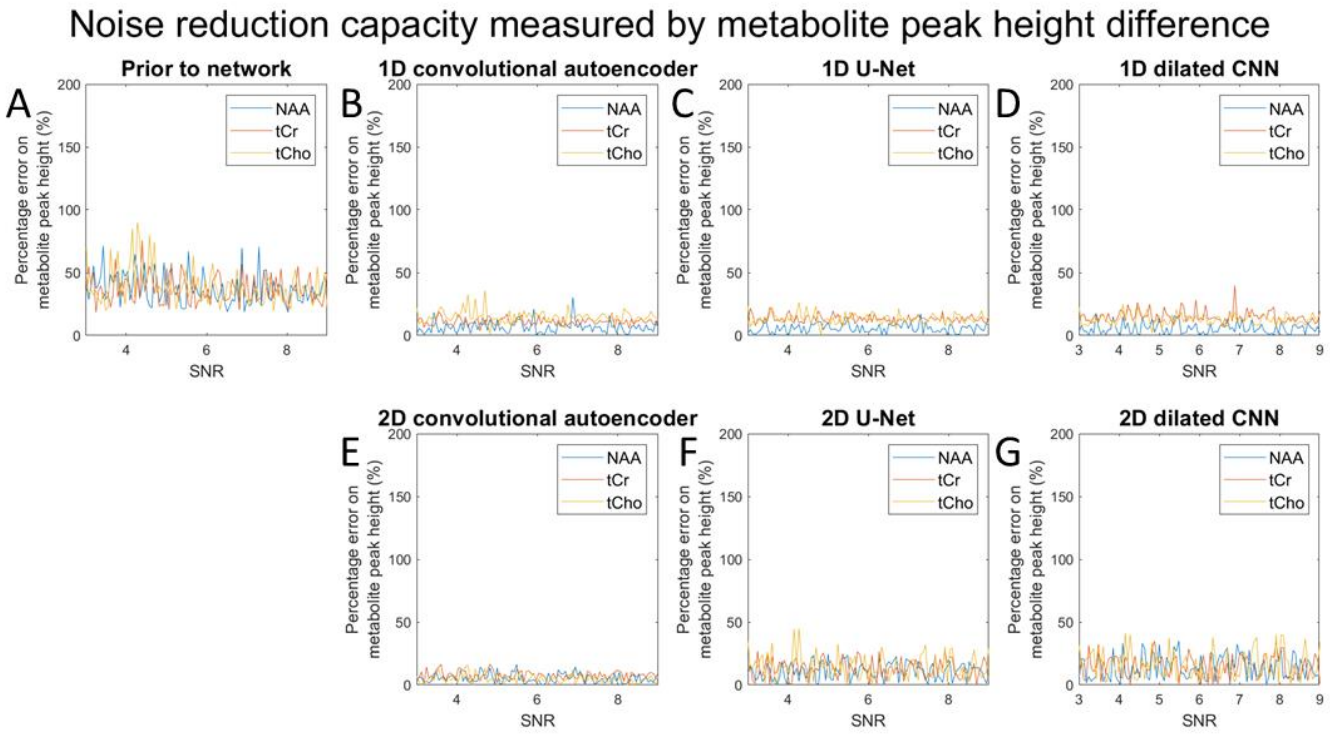


Figure 4.13. Percentage error of metabolites with increasing SNR. A: Before denoising by any network. B: After denoising using the 1D autoencoder. C: After denoising with the 1D U-Net. D: After denoising with the dilated CNN. E: After denoising with the 2D convolutional autoencoder. F: After denoising with the 2D U-Net. G: After denoising with the 2D dilated CNN.

Using a paired t-test, it was tested if the mean peak height differences were significantly different for each of the networks. Bonferroni correction was used. The differences were significant if $p < 0.0024$. The result is visualised in Figure 4.14.

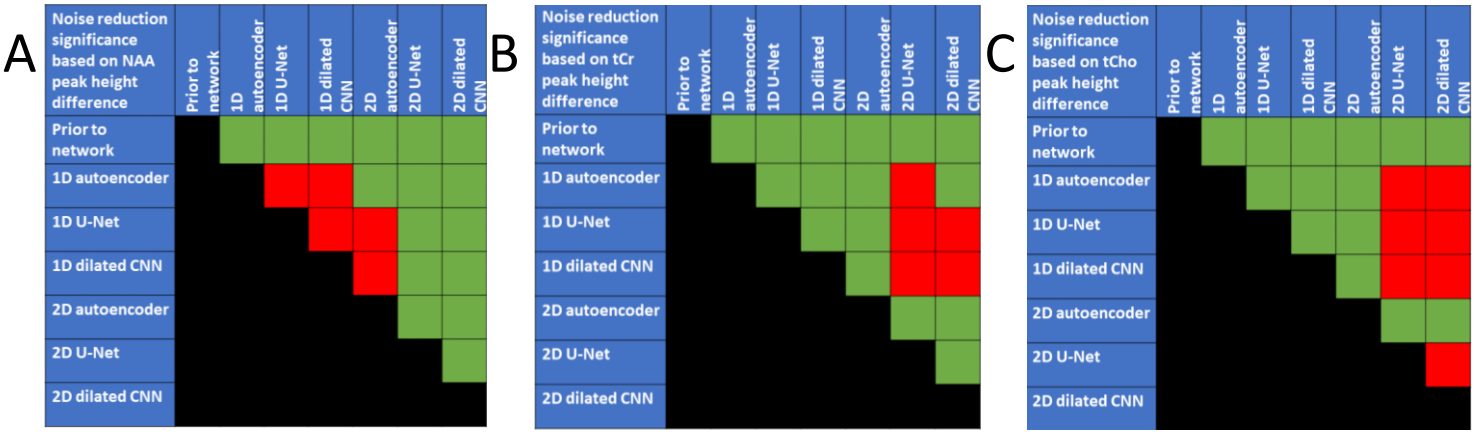


Figure 4.14. Significant differences in mean peak height difference for NAA (A), tCr (B) and tCho (C). Green indicates a significant difference. Red indicates no significant difference.

Residual water artefact

The second individual set that was tested contained spectra with an increasingly strong residual water artefact. These spectra were given to the networks as input. The results based on the SD of the residual on the noise-contaminated spectra can be found in Figure 4.15. Since the values prior to denoising and artefact removal did not come from a nornal distribution, no t-test was performed between the network input and the network outputs. The results based on the mean peak height difference for the NAA peak, the tCr peak and the tCho peak can be found in Figure 4.17.

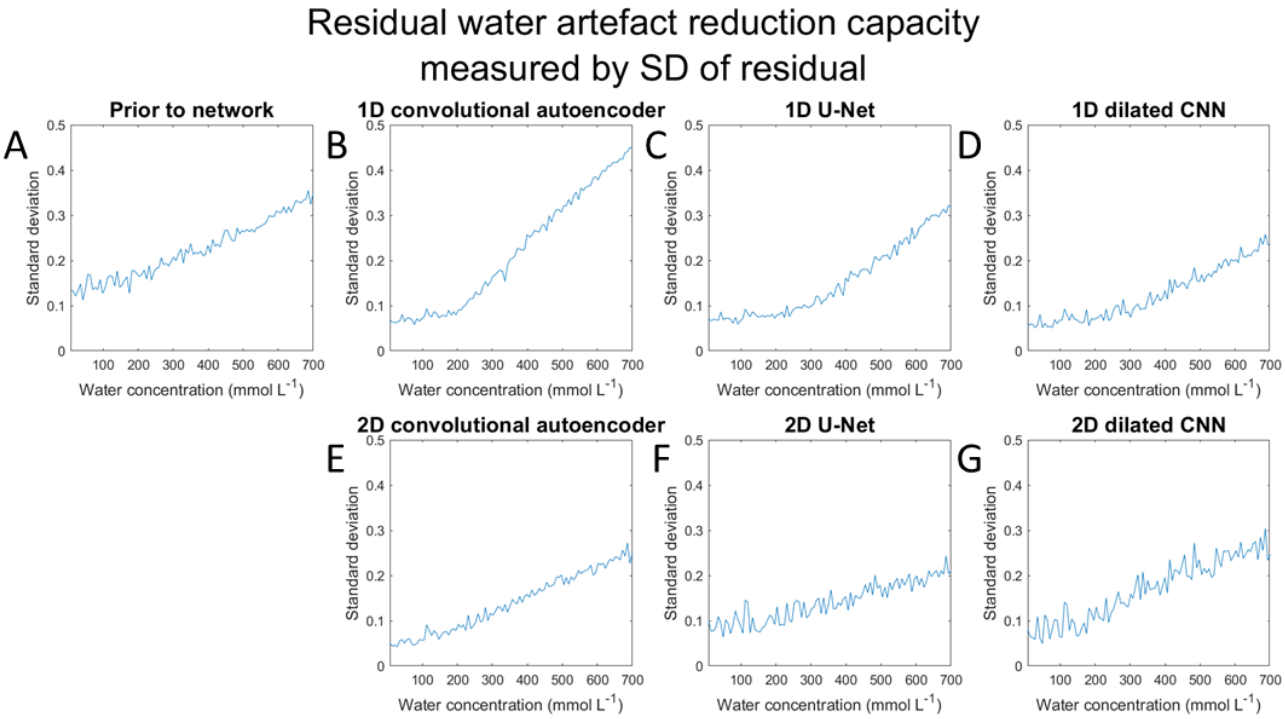


Figure 4.15. A: Standard deviation of residual before artefact removal. B: Standard deviation of residual after artefact removal by 1D convolutional autoencoder. C: Standard deviation of residual after artefact removal by 1D U-Net. D: Standard deviation of residual after artefact removal by 1D dilated CNN. E: Standard deviation of residual after artefact removal by 2D convolutional autoencoder. F: Standard deviation of residual after artefact removal by 2D U-Net. G: Standard deviation of residual after artefact removal by 2D dilated CNN.

Using a paired t-test, we tested if the mean peak height differences were significantly different for each of the networks. The differences were significant if $p < 0.0033$. The result is visualised in Figure 4.16.

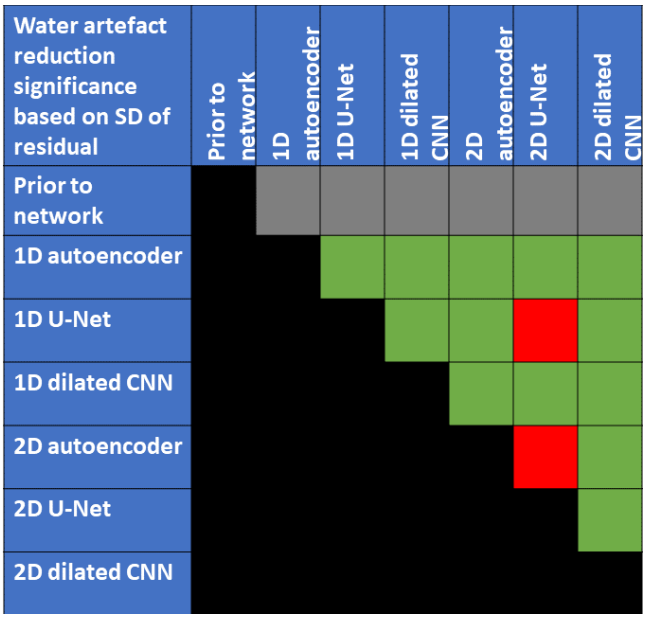


Figure 4.16: The average standard deviation of the residual was significantly different for most of the networks. Green indicates a significant difference. Red indicates there was no significant difference. Grey indicates a t-test was not possible because there was no normal distribution.

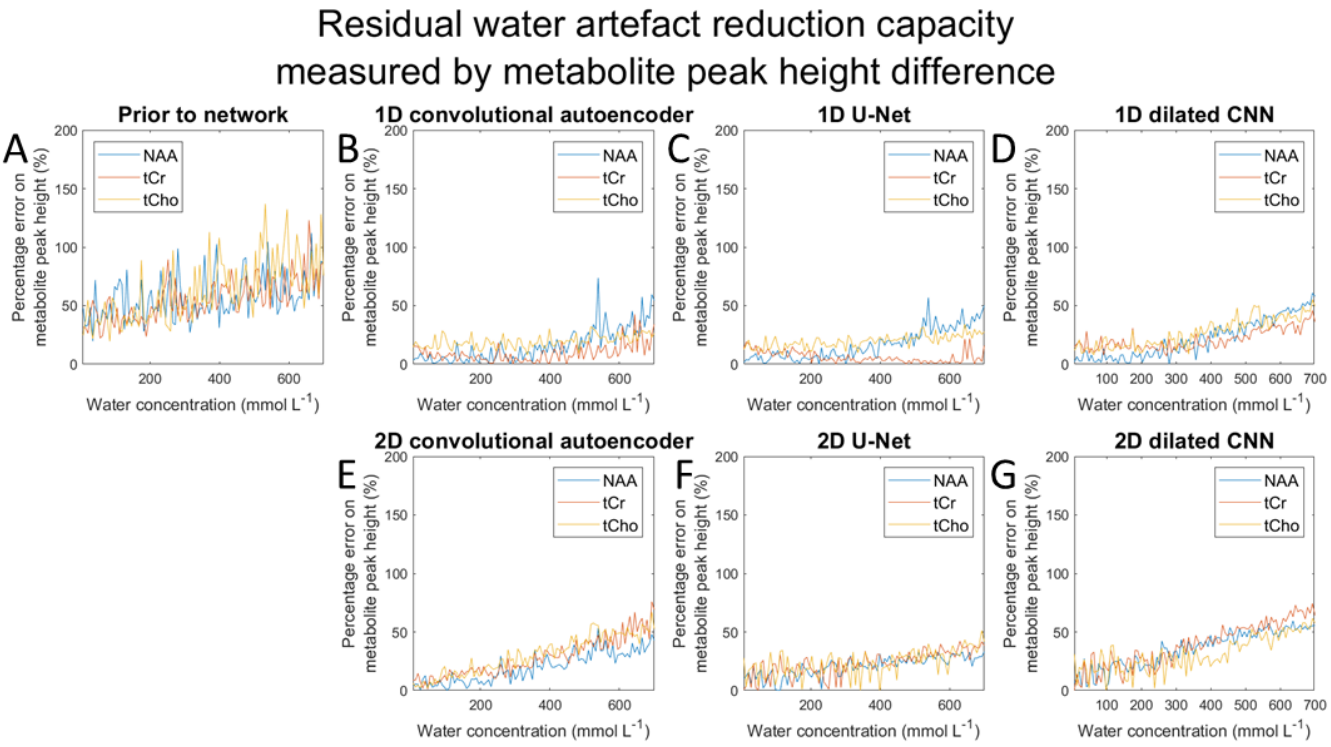


Figure 4.17 Percentage error of metabolites with increasingly strong residual water artefact. A: Before artefact removal by any network. B: After artefact removal using the 1D autoencoder. C: After artefact removal with the 1D U-Net. D: After artefact removal with the dilated CNN. E: After artefact removal with the 2D convolutional autoencoder. F: After artefact removal with the 2D U-Net. G: After artefact removal with the 2D dilated CNN.

Using a paired t-test, it was tested if the mean peak height differences were significantly different for each of the networks after removing the residual water artefact. Bonferroni correction was used. The

differences were significant if $p < 0.0024$. The result is visualised in Figure 4.18.

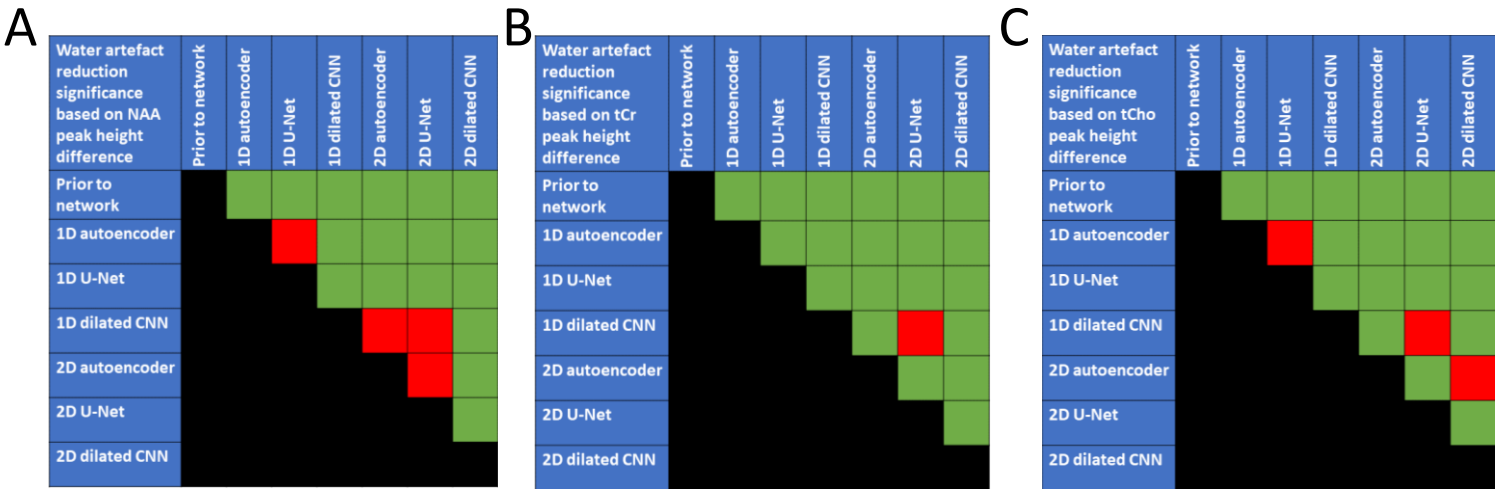


Figure 4.18. Significant differences in mean peak height difference for NAA (A), tCr (B) and tCho (C). Green indicates a significant difference. Red indicates no significant difference.

Lipid artefact

The third individual set that was tested contained spectra with an increasingly strong lipid artefact. These spectra were given to the networks as input. The results based on the SD of the residual on the noise-contaminated spectra can be found in Figure 4.19. Since the values prior to denoising and artefact removal did not come from a nornal distribution, no t-test was performed between the network input and the network outputs. The results based on the mean peak height difference for the NAA peak, the tCr peak and the tCho peak can be found in Figure 4.21.

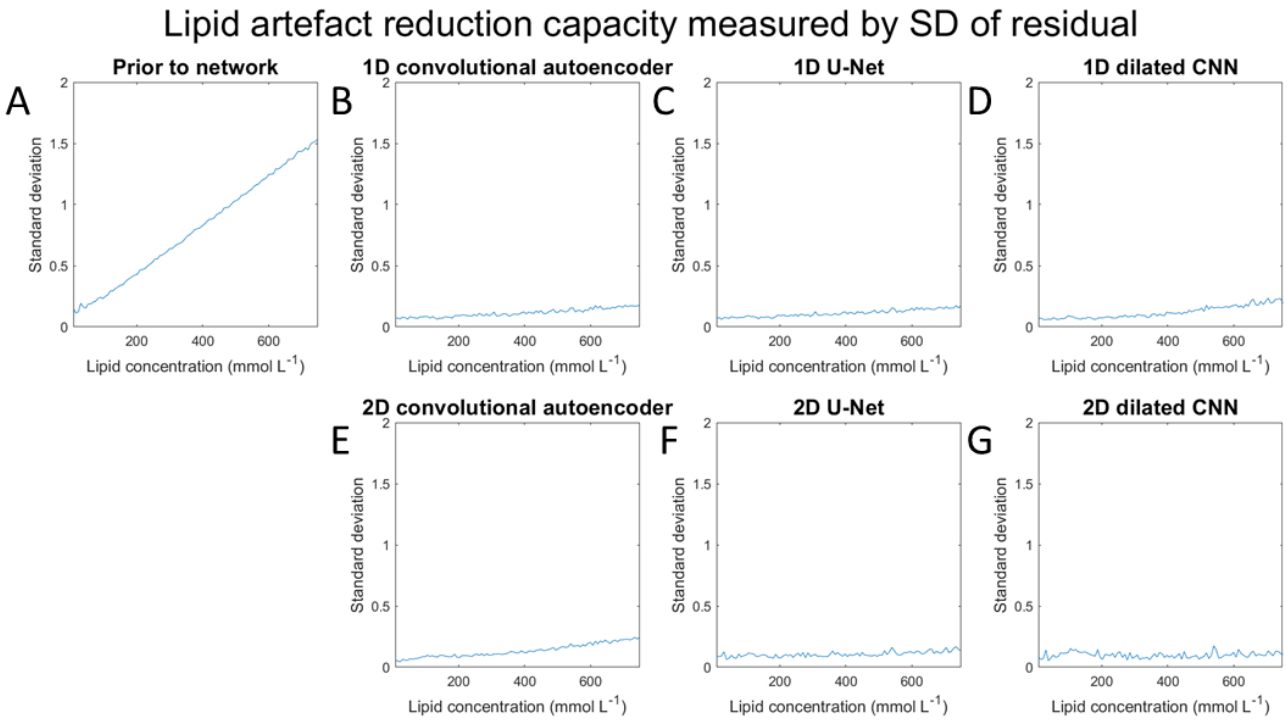


Figure 4.19. A: Standard deviation of residual before lipid artefact removal. B: Standard deviation of residual after lipid artefact removal by 1D convolutional autoencoder. C: Standard deviation of residual after lipid artefact removal by 1D U-Net. D: Standard deviation of residual after lipid artefact removal by 1D dilated CNN. E: Standard deviation of residual after lipid artefact removal by 2D convolutional autoencoder. F: Standard deviation of residual after lipid artefact removal by 2D U-Net. G: Standard deviation of residual after lipid artefact removal by 2D dilated CNN.

Using a paired t-test, it was tested if the mean peak height differences were significantly different for each of the networks. Bonferroni correction was used. The differences were significant if $p < 0.0033$. The result is visualised in Figure 4.20.

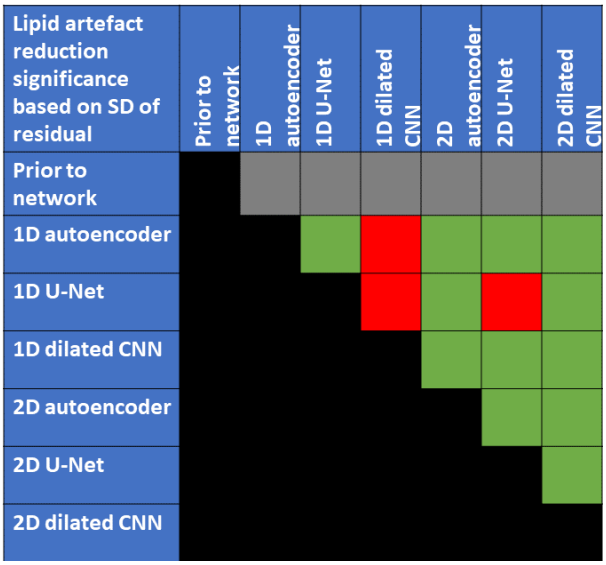


Figure 4.20: The average standard deviation of the residual was significantly different for most of the networks. Green indicates a significant difference. Red indicates there was no significant difference. Grey indicates a t-test was not possible because there was no normal distribution.

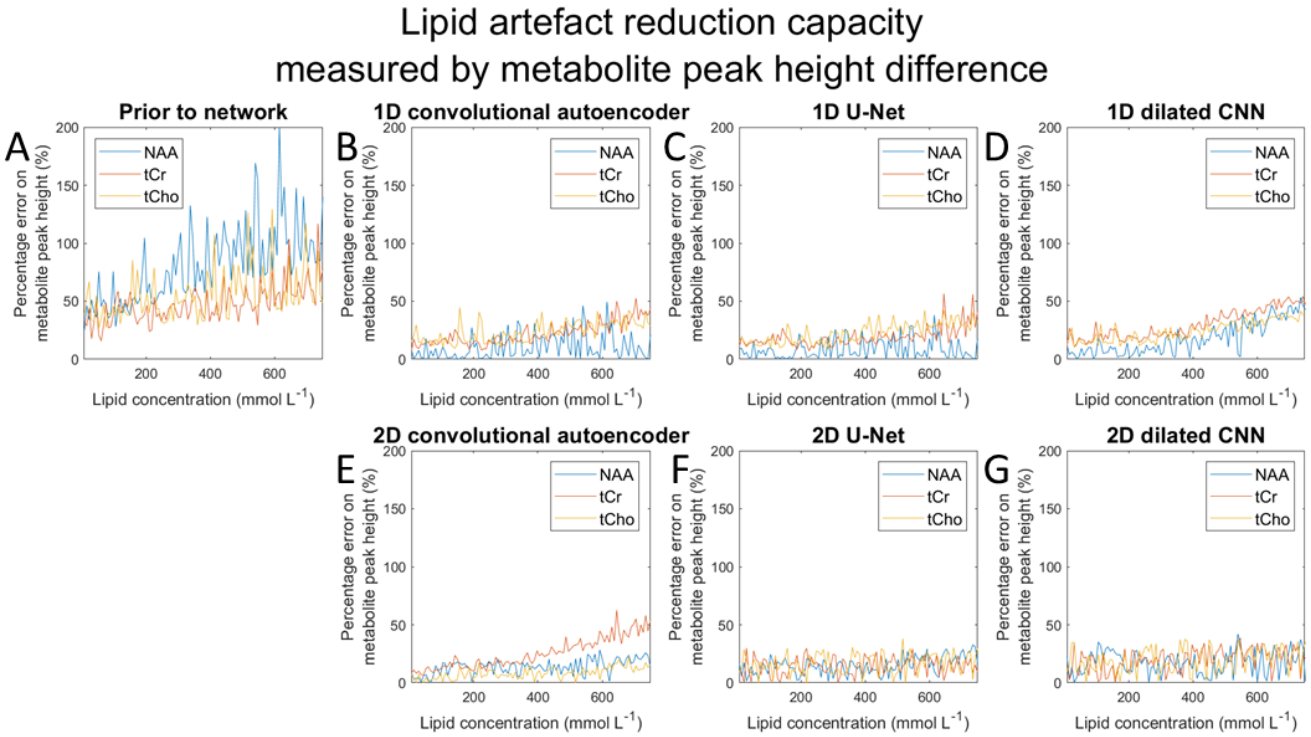


Figure 4.21: Percentage error of metabolites with increasingly strong lipid artefact. A: Before artefact removal by any network. B: After artefact removal using the 1D autoencoder. C: After artefact removal with the 1D U-Net. D: After artefact removal with the dilated CNN. E: After artefact removal with the 2D convolutional autoencoder. F: After artefact removal with the 2D U-Net. G: After artefact removal with the 2D dilated CNN.

A paired t-test tested whether the mean peak height differences were significantly different for each of the networks after removing the lipid artefact. Bonferroni correction was used The differences were significant if $p < 0.0024$. The result is visualised in Figure 4.20.

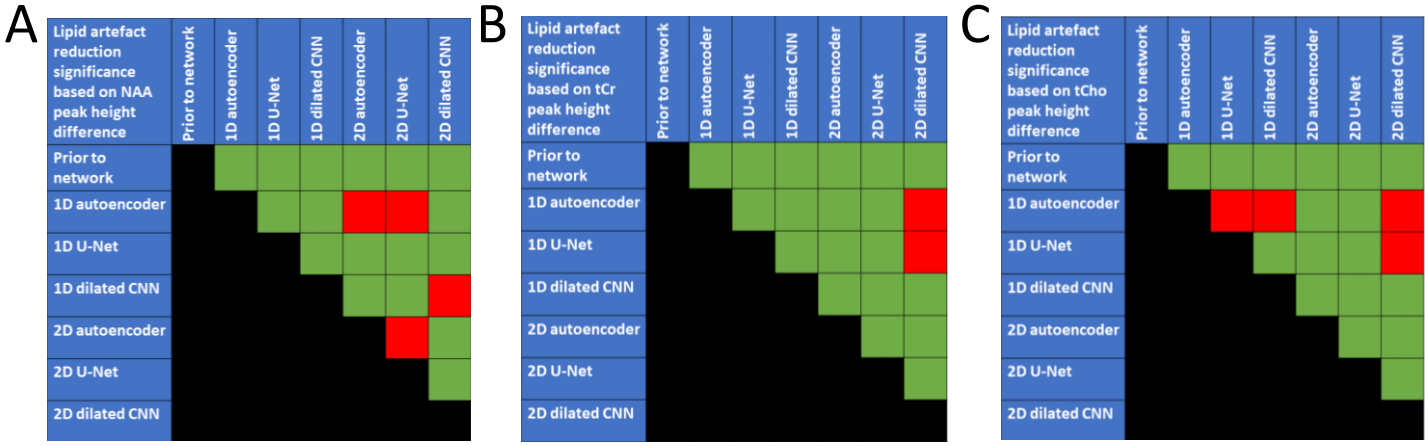


Figure 4.22. Significant differences in mean peak height difference for NAA (A), tCr (B) and tCho (C). Green indicates a significant difference. Red indicates no significant difference.

In vivo spectra

The six networks were tested using the 10 *in vivo* spectra. An example can be found in Figures 4.23 and 4.24. In Figure 4.23, the network output and the difference between input and output are visualised per 1D network. In Figure 4.23D, the spectrum as outputted by the autoencoder can be found. The output does not show a distinct NAA peak. Also, the tCho and tCr peak are not as prominent as in the input spectrum. In Figure 4.23E, the spectrum as outputted by the U-Net is plotted. Like the output by

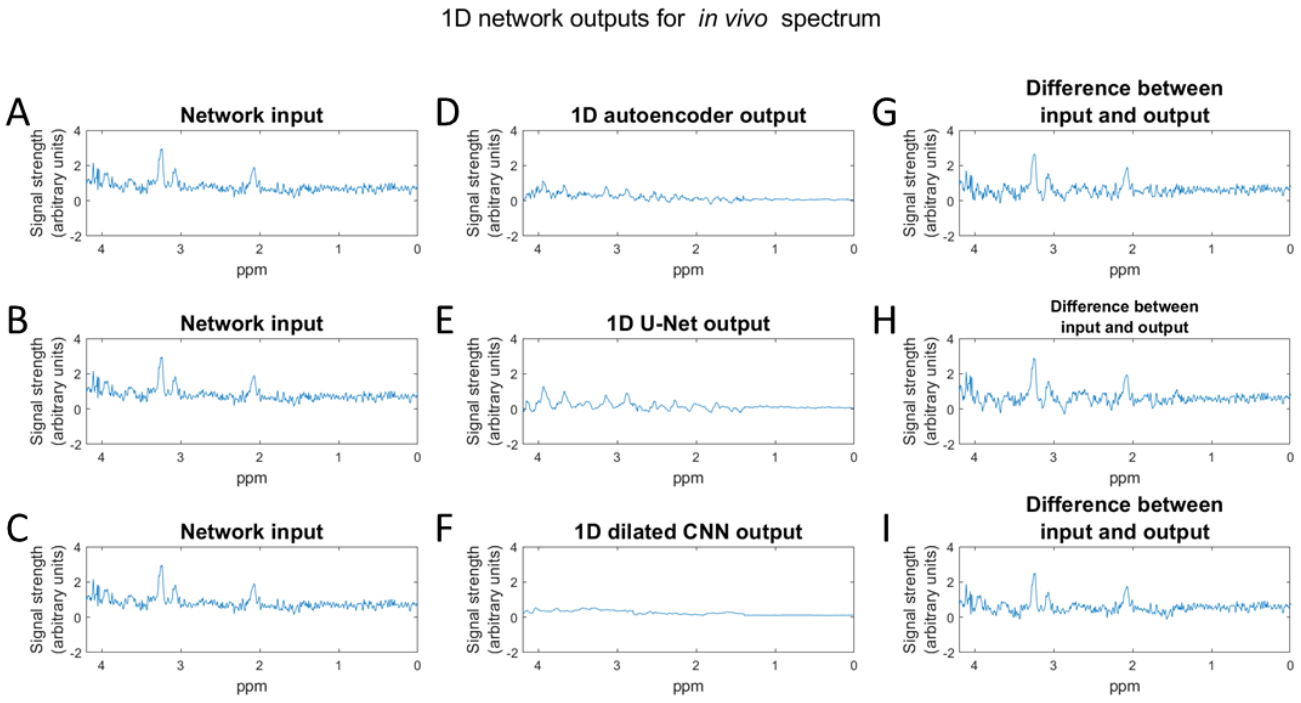


Figure 4.23. A-C: Network input. D-F: Output from the autoencoder, U-Net and dilated CNN, respectively. G-I: Difference between predicted spectrum and network input.

the autoencoder, the NAA peak, tCho peak and tCr peak can not be identified. The output by the dilated CNN is plotted in Figure 4.23F. In this spectrum no peak can be identified at all.

In Figure 4.24, the network output and the difference between input and output are visualized. The inputs for these networks were the scalograms. The outputs of the network are scalograms as well and the reconstructed spectrum from the outputted scalogram is visualized. The difference between the input spectrum before the CWT and the reconstructed spectrum after the ICWT is also shown. In Figure 4.24D, the 2D autoencoder output shows a prominent lipid peak around 1.3 ppm that is not visible in the input. The NAA peak, tCr peak and tCho peak appear to have been preserved. In Figure 4.24E, the 2D U-Net output is shown. The peak heights of NAA, tCr and tCho in the output spectrum are lower compared to the baseline drift than in the input. No clear noise is visible. Figure 4.24F visualizes the output by the dilated CNN. Barely any peak can be identified in this spectrum.

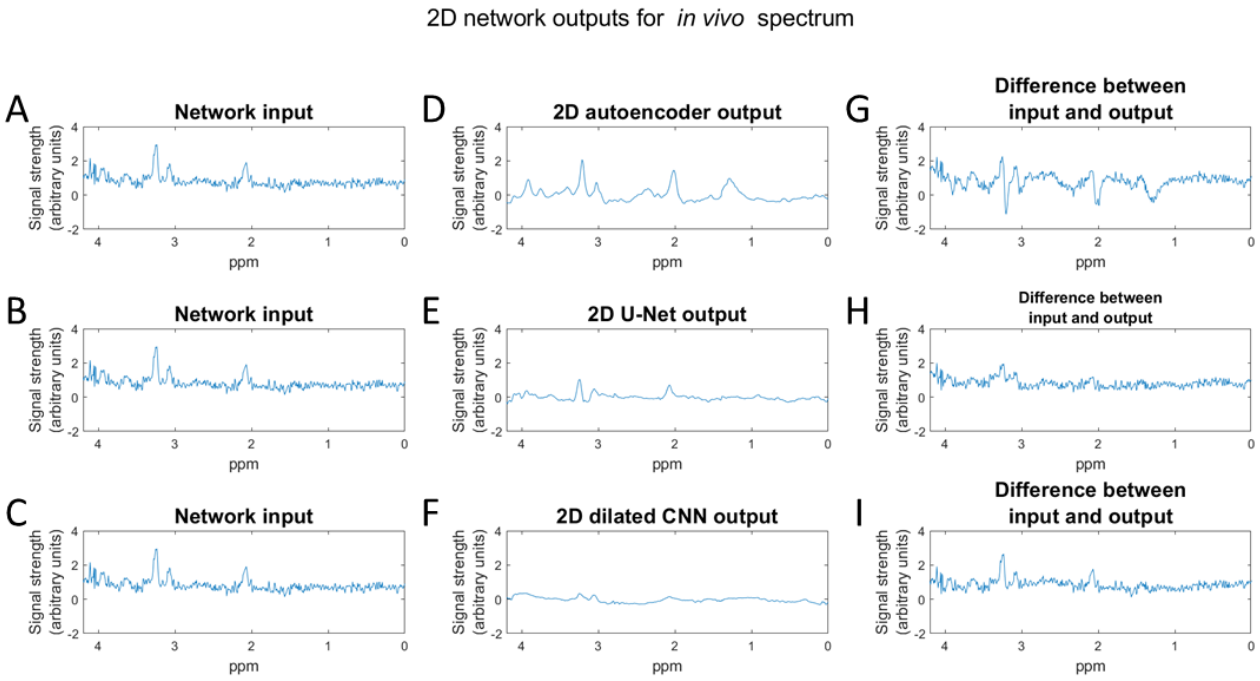


Figure 4.24. A-C: These spectra were transformed using the CWT prior to inputting them to the networks. D-F: Reconstructed spectra from the outputted scalograms from the autoencoder, U-Net and dilated CNN, respectively. G-I: Difference between reconstructed spectrum and network input.

5. Discussion

In this study, we evaluated the ability of deep neural networks to reduce noise and remove artefacts from MR spectra. The deep learning networks that were trained reduced noise and removed artefacts with various degrees of success with respect to the standard deviation of the residual and peak height differences of NAA, tCr and tCho. Overall, each of the networks improved the quality of the synthetic input spectra, as there was a significant difference between the standard deviation of the residual and peak height differences of NAA, tCr and tCho between the input spectra and the output spectra.

Six networks were trained for the purpose of reducing noise and removing artefacts. The best-performing network on the noise-contaminated dataset was the 2D convolutional autoencoder because it either outperformed the other networks or there was no significant difference. The best-performing network on the residual water artefact dataset was the 1D dilated CNN. The best-performing network on the lipid artefact dataset was the 2D dilated CNN. As the networks that were optimized on scalograms had significantly better results on the noise-contaminated test set and the lipid-contaminated test set, the use of the scalograms to reduce noise and remove the lipid artefacts was of added value. However, the added value of the wavelet transform is not proven for residual water artefact removal.

The network outputs from the *in vivo* spectra did not resemble the inputted spectra and were therefore deemed untrustworthy. Still, the results that were achieved on the simulated spectra do show the impact noise reduction and artefact removal in MRS using deep learning could have on more reliable metabolite concentrations.

Performance on simulated data

The performance of the networks on the noise test set was rather good, considering the low values for the standard deviation of the residual and low absolute percentage errors on NAA, tCr and tCho peak height. However, the performance on the residual water artefact test set and the lipid artefact test set was worse. One of the reasons for the difference in performance could be that there were more examples with noise in the training set than examples with a residual water artefact or a lipid artefact.

Performance on *in vivo* data

The network output on the *in vivo* data was deemed untrustworthy. This has to do with reference peaks, such as NAA, tCr and tCho either missing or based on qualitative inspection having a too-low peak when compared to the noisy input. As a consequence of these results, quantifying metabolite concentrations more reliably was not possible. Further applications of MRS, such as classification of the tumors in the pons, could with these results not be improved.

Likely, the *in vivo* spectra were out-of-distribution compared to the training data. This means that the feature distribution in the *in vivo* data differed from the feature distribution data of the training data (61,62). Examples of features could be the SNR or location of the peaks in the spectrum. So if all training data had an SNR ranging between 5 and 7, the result from a test example with an SNR of 25 might not be satisfying.

Upon visualizing the results from the *in vivo* data, it was seen that all spectra did not have the NAA peak at 2.008 ppm, nor at 2.04, where NAAG is located. The NAA peak was in six out of the ten *in vivo* spectra located between 2.05 and 2.07 ppm. This shift of the NAA peak had been reported earlier by Tamrazi *et al.* (63). However, the patients studied in that population were diagnosed with pilocytic astrocytoma, whereas the patients in the *in vivo* dataset have been diagnosed with diffuse intrinsic pontine glioma. It could be that, just like in pilocytic astrocytoma, the tumour biology in diffuse intrinsic pontine glioma caused the peak around 2 ppm to be shifted. Due to the fact that NAA and NAAG were

simulated, yet potentially a different molecule gave rise to the peak around 2 ppm, the deep learning networks might not have been able to accurately predict the outcome of values in this ppm region.

Comparison to literature

Earlier publications to remove artefacts include the work by Lin *et al*. Lin *et al.* proposed to remove residual water artefacts and lipid signals from spectra using an L2-regularization-based post-processing method. Results from the L2-regularization-based method outperform the results from this study due to the fact that peaks were better preserved. Therefore, the quantification of metabolite concentrations was more reliable in the study by Lin *et al.* than in this study. Also, their proposed method worked well on *in vivo* data, which was not the case in this study.

Denoising has been performed by Abdoli *et al.* using a statistical selection of principal components. After denoising simulated spectra, the error in metabolite ratios was reduced by 12-94% (26). In this study, metabolite peak height errors were reduced by 58.5%-86.6% in the test set that contained noise examples only, depending on the noise level, a metabolite of interest, and the network that was used. Yet, the error prior to denoising from the study by Abdoli *et al.* is unknown. Also, in this study, peak heights were used instead of quantified metabolite concentrations. Thus, a one-on-one comparison cannot be made. One strength of their work was that it was well applicable to *in vivo* data. However, the removal of residual water artefacts or lipid artefacts was not considered.

Deep learning approaches to artefact removal have also been attempted. Kyathanahally *et al.*, removed ghosting artefacts from time-domain signals using stacked what-where autoencoders. They found that the ghosting artefacts had been removed successfully based on visual appearance. Their methods succeeded on both simulated data and *in vivo* data (23). Based on the visual appearance of spectra in this study, artefacts in this study were removed with varying degrees of success. However, peak heights were not always preserved in this study, which was the case in the study by Kyathanahally *et al.*

Lee *et al.* used a CNN to perform both denoising and line narrowing on simulated MR spectra of healthy brain tissue. A mean absolute percentage error of 4.64%, 8.65% and 7.87% for NAA, Cr and tCho was reported on the test set (29). These results are similar to the results of the test set for the 2D convolutional autoencoder, which was 5.22%, 9.54% and 4.61% for NAA, tCr and tCho, respectively. Yet, in the analysis in this study, peak height differences were compared, instead of quantified metabolite concentration. The study by Lee *et al.* outperformed this study in terms of the applicability of the trained network on *in vivo* data.

Limitations

The biggest limitation of this study is that quantification of metabolite concentrations was not performed. Instead, the standard deviation of the residual and peak height differences in known ppm regions of the spectra were used as substitutes. Accordingly, all other metabolites have not been quantified. Secondly, there was an unbalance in the training set. Out of the 6400 spectra in the training set, only 800 contained a residual water artefact and another 800 contained a lipid artefact. The other 4800 were noise-contaminated. Not only is this unbalanced, but 800 training examples also might not have been sufficient to learn feature distribution for this contamination. These two are likely the reasons why the networks performed better on the noise set than on the residual water artefact set and lipid artefact set. Thirdly, a thorough hyperparameter tuning was not performed due to time constraints. Examples of hyperparameters that could be tuned are the learning rate, the activation function, network depth and the optimization algorithm. With a more thorough hyperparameter tuning, network performance is likely to improve. A fourth limitation is that loss over the 2D networks was calculated over the scalograms. Yet, the spectra are of interest. Incorporating the CWT at the

beginning of the network and incorporating the ICWT at the end of these networks would still enable the networks to learn features from the scalograms. Additionally, the loss would be calculated over the reconstructed spectra and the ground truth spectra. Potentially, this leads to improved results in terms of the standard deviation of the residual or peak height differences. As a final limitation, the test set that contained only noise-contaminated spectra had only an SNR range of 3-9. To better understand the performance of the network, a wider range of SNR could have been simulated. For a follow-up study, it is advised to simulate an SNR in such a range that the limits of the network's performance are found.

Future directions

The metabolite concentrations that were used to simulate spectra were roughly based on metabolite concentrations in healthy brain tissue. Only in NAA, creatine, GPC and PCH, alterations were made. Another change that could improve the results is to simulate a wider range of linewidths for the input spectra. Learning a wider distribution of features might be beneficial for the network's performance. Therefore, in the next study, the argument could be made to simulate a wider range of metabolite concentrations and linewidths. Ultimately, changes in metabolite concentration ranges and linewidths might lead to a more robust performance on both the simulated spectra and scalograms, as well as on the *in vivo* spectra and corresponding scalograms.

The networks that used a scalogram as input have been trained on the real part of the scalogram since this is the only necessary part for an inverse transform (44). However, the scalogram's imaginary, magnitude and phase parts have been disregarded. Potentially, the magnitude and phase part of the scalogram contain information that deep learning networks can pick up easier than from the real part only. This, in turn, might improve results from the scalogram networks. This question is left unanswered.

Another question that has not been answered in this study, is whether the Morlet wavelet is the best mother wavelet to use for this application. Using the Morlet wavelet, the spectrum could be transformed using the CWT. Thereafter, the spectrum could be reconstructed from the scalogram using the ICWT with minimal loss of information. However, it could be that different mother wavelets lead to a different-looking scalogram, from which features can be learned more effectively, resulting in a lower standard deviation in the residual spectrum and lower peak height differences. This, in turn, could lead to more reliable quantification of metabolite concentrations.

The scalograms had a size of (2048, 512) each. This made generating, saving and loading these scalograms a time-costly task. Thereby, network training was very time-consuming, sometimes leading to multiple days of training and validating. In the Methods section, it was described that the optimal parameters for the wavelet transform were chosen to minimize the standard deviation of the residual spectrum. Yet, other options, including the ones with fewer wavelet scales, were not considered. It was therefore not tested if the use of fewer wavelet scales actually leads to significant differences in peak heights. If peak heights do indeed not change significantly, this could be an argument for scalograms with a smaller size. This could save valuable time in generating the scalograms and training networks that use scalograms as training data.

In the deep learning part of this study, alterations could also be made that potentially benefit the results. Firstly, the networks that were trained had to predict either the spectrum or scalogram. However, in some cases, the difference between the clean spectrum or scalogram and noise- and artefact-contaminated spectrum or scalogram is only subtle. It is possible to add the noisy spectrum or scalogram to the prediction of the network at the end. In doing so, the task of the network becomes to predict the difference, instead of predicting the full spectrum or scalogram, which might easier.

Secondly, to improve performance on *in vivo* data, it is an option to blend simulated data with *in vivo* data. However, a ground truth of the *in vivo* data lacks. An option is to use the fitted spectrum as ground truth. That way, the networks can also learn feature distribution from *in vivo* spectra. This approach has already been successful outside the world of medicine. One example is a study by Poucin *et al*., who mixed real-world data and synthetic data in image segmentation tasks. Significant improvements were seen in using a blended dataset (64). Similarly, using a training set that comprises of both synthetic data and *in vivo* data might be beneficial for the performance on both simulated data and *in vivo* data.

Conclusion

Deep learning networks reduced noise and removed artefacts from both synthetic spectra and scalograms. The added value of the wavelet transforms was proven in reducing noise and removing lipid artefacts, but not for removing the residual water artefact. Due to peaks missing or unexpected peaks appearing, the outputted *in vivo* spectra were not trustworthy. Nevertheless, results on the synthetic dataset show the positive impact deep learning could have on reducing noise and removing artefact, leading to more reliable metabolite concentrations. Future studies should focus on more realistic simulations and quantifying metabolite concentrations in simulated spectra.

References

1. Udaka YT, Packer RJ. Pediatric Brain Tumors. Neurol Clin [Internet]. 2018;36(3):533–56. Available from: https://doi.org/10.1016/j.ncl.2018.04.009

2. Jagt-van Kampen CT, Van de Wetering MD, Schouten-van Meeteren AYN. The timing, duration, and management of symptoms of children with an incurable brain tumor: A retrospective study of the palliative phase. Neuro-Oncology Pract. 2015;2(2):70–7.

3. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. Lancet [Internet]. 2018;392(10145):432–46. Available from: http://dx.doi.org/10.1016/S0140-6736(18)30990-5

4. Ulmer S, Backens M, Ahlhelm FJ. Basic principles and clinical applications of magnetic resonance spectroscopy in neuroradiology. J Comput Assist Tomogr. 2016;40(1):1–13.

5. Tognarelli JM, Dawood M, Shariff MIF, Grover VPB, Crossey MME, Cox IJ, et al. Magnetic Resonance Spectroscopy: Principles and Techniques: Lessons for Clinicians. J Clin Exp Hepatol [Internet]. 2015;5(4):320–8. Available from: http://dx.doi.org/10.1016/j.jceh.2015.10.006

6. Suh CH, Kim HS, Jung SC, Choi CG, Kim SJ. 2-Hydroxyglutarate MR spectroscopy for prediction of isocitrate dehydrogenase mutant glioma: A systemic review and meta-analysis using individual patient data. Neuro Oncol. 2018;20(12):1573–83.

7. Kayabolen A, Yilmaz E, Bagci-Onder T. IDH mutations in glioma: Double-edged sword in clinical applications? Biomedicines. 2021;9(7).

8. Lequin M, Hendrikse J. Advanced MR Imaging in Pediatric Brain Tumors, Clinical Applications. Neuroimaging Clin N Am [Internet]. 2017;27(1):167–90. Available from: http://dx.doi.org/10.1016/j.nic.2016.08.007

9. El-Abtah ME, Talati P, Fu M, Chun B, Clark P, Peters A, et al. Magnetic Resonance Spectroscopy Outperforms Perfusion in Distinguishing Between Pseudoprogression and Disease Progression in Patients with Glioblastoma. Neuro-Oncology Adv. 2022;4(August):1–10.

10. Provencher S. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magn Reson Med. 1993;30(6):672–9.

11. Öz G, Alger JR, Barker PB, Bartha R, Bizzi A, Boesch C, et al. Clinical proton MR spectroscopy in central nervous system disorders. Radiology [Internet]. 2014 Mar 5 [cited 2022 Mar 7];270(3):658–79. Available from: https://pubs.rsna.org/doi/abs/10.1148/radiol.13130531

12. de Graaf R. In Vivo NMR Spectroscopy: Principles and Techniques. In: In Vivo NMR Spectroscopy: Principles and Techniques. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2019. p. 49–128.

13. Nakamura H, Doi M, Suzuki T, Yoshida Y, Hoshikawa M, Uchida M, et al. The significance of lactate and lipid peaks for predicting primary neuroepithelial tumor grade with proton MR spectroscopy. Magn Reson Med Sci. 2018;17(3):238–43.

14. Li Y, Lafontaine M, Chang S, Nelson SJ. Comparison between Short and Long Echo Time Magnetic Resonance Spectroscopic Imaging at 3T and 7T for Evaluating Brain Metabolites in Patients with Glioma. ACS Chem Neurosci. 2018;9(1):130–7.

15. Gonçalves FG, Viaene AN, Vossough A. Advanced Magnetic Resonance Imaging in Pediatric Glioblastomas. Vol. 12, Frontiers in Neurology. 2021.

16. Blüml S, Margol AS, Sposto R, Kennedy RJ, Robison NJ, Vali M, et al. Molecular subgroups of

medulloblastoma identification using noninvasive magnetic resonance spectroscopy. Neuro Oncol. 2016;18(1):126–31.

17. Zarinabad N, Abernethy LJ, Avula S, Davies NP, Rodriguez Gutierrez D, Jaspan T, et al. Application of pattern recognition techniques for classification of pediatric brain tumors by in vivo 3T 1H-MR spectroscopy—A multi-center study. Magn Reson Med. 2018;79(4):2359–66.

18. Bertholdo D, Watcharakorn A, Castillo M. Brain Proton Magnetic Resonance Spectroscopy: Introduction and Overview. Neuroimaging Clin N Am [Internet]. 2013;23(3):359–80. Available from: http://dx.doi.org/10.1016/j.nic.2012.10.002

19. Juchem C, de Graaf RA. B0 magnetic field homogeneity and shimming for in vivo magnetic resonance spectroscopy. Anal Biochem. 2017;529:17–29.

20. Chen D, Wang Z, Guo D, Orekhov V, Qu X. Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy. Chem - A Eur J. 2020;26(46):10391–401.

21. Kreis R. Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts. NMR Biomed. 2004;17(6):361–81.

22. Lee HH, Kim H. Deep learning-based target metabolite isolation and big data-driven measurement uncertainty estimation in proton magnetic resonance spectroscopy of the brain. Magn Reson Med. 2020;84(4):1689–706.

23. Kyathanahally SP, Döring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. Magn Reson Med. 2018;80(3):851–63.

24. Skoch A, Jiru F, Bunke J. Spectroscopic imaging: Basic principles. Eur J Radiol. 2008;67(2):230–9.

25. Jansen JFA, Backes WH, Nicolay K, Kooi ME. 1H MR spectroscopy of the brain: Absolute quantification of metabolites. Radiology. 2006;240(2):318–32.

26. Abdoli A, Stoyanova R, Maudsley AA. Denoising of MR spectroscopic imaging data using statistical selection of principal components. Magn Reson Mater Physics, Biol Med. 2016;29(6):811–22.

27. Berrington A, Považan M, Barker PB. Estimation and removal of spurious echo artifacts in single-voxel MRS using sensitivity encoding. Magn Reson Med. 2021;86(5):2339–52.

28. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42(December 2012):60–88.

29. Lee HH, Kim H. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. Magn Reson Med. 2019;82(1):33–48.

30. Rhif M, Abbes A Ben, Farah IR, Martínez B, Sang Y. Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review. 2019 [cited 2022 Feb 17]; Available from: www.mdpi.com/journal/applsci

31. Soares DP, Law M. Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. Clin Radiol [Internet]. 2009;64(1):12–21. Available from: http://dx.doi.org/10.1016/j.crad.2008.07.002

32. Panigrahy A, Nelson MD, Blüml S. Magnetic resonance spectroscopy in pediatric neuroradiology: clinical and research applications. Pediatr Radiol. 2010;

33. Zhang Y, Taub E, Salibi N, Uswatte G, Maudsley AA, Sheriff S, et al. Comparison of reproducibility of single voxel spectroscopy and whole-brain magnetic resonance

spectroscopy imaging at 3T. Vol. 31, NMR in Biomedicine. 2018.

34. Mandal PK. In vivo proton magnetic resonance spectroscopic signal processing for the absolute quantitation of brain metabolites. Eur J Radiol [Internet]. 2012;81(4):e653–64. Available from: http://dx.doi.org/10.1016/j.ejrad.2011.03.076

35. Bulik M, Jancalek R, Vanicek J, Skoch A, Mechl M. Potential of MR spectroscopy for assessment of glioma grading. Clin Neurol Neurosurg [Internet]. 2013;115(2):146–53. Available from: http://dx.doi.org/10.1016/j.clineuro.2012.11.002

36. Klose U. Measurement sequences for single voxel proton MR spectroscopy. Eur J Radiol. 2008;67(2):194–201.

37. Zhu H, Barker PB. MR Spectroscopy and Spectroscopic Imaging of the Brain. Methods Mol Biol [Internet]. 2011 [cited 2022 Mar 10];711:203–26. Available from: https://link.springer.com/protocol/10.1007/978-1-61737-992-5_9

38. Inglese M, Spindler M, Babb JS, Sunenshine P, Law M, Gonen O. Field, coil, and echo-time influence on sensitivity and reproducibility of brain proton MR spectroscopy. Am J Neuroradiol. 2006;27(3):684–8.

39. Madan A, Ganji SK, An Z, Choe KS, Pinho MC, Bachoo RM, et al. Proton T2 measurement and quantification of lactate in brain tumors by MRS at 3 Tesla in vivo. Magn Reson Med. 2015;73(6):2094–9.

40. Lange T, Dydak U, Roberts TPL, Rowley HA, Bjeljac M, Boesiger P. Pitfalls in Lactate Measurements at 3T. Am J Neuroradiol. 2006;27(4).

41. Liu H-S, Chung H-W, Juan C-J, Tsai S-Y, Wang C-Y, Chan C-C, et al. Anomalous J-Modulation Effects on Amino Acids in Clinical 3T MR Spectroscopy. Am J Neuroradiol [Internet]. 2008 [cited 2022 Mar 10];29:1644–8. Available from: www.ajnr.org

42. Sian J, Dexter DT, Lees AJ, Daniel S, Agid Y, Javoy-Agid F, et al. Alterations in glutathione levels in Parkinson's disease and other neurodegenerative disorders affecting basal ganglia. Ann Neurol. 1994;36(3):348–55.

43. Rhif M, Abbes A Ben, Farah IR, Martínez B, Sang Y. Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review. 2019 [cited 2022 Feb 7]; Available from: www.mdpi.com/journal/applsci

44. Poularikas AD. Transforms and applications handbook. Third. Transforms and Applications Handbook, Third Edition. 2010. 1–900 p.

45. de Moortel I, Munday SA, Hood AW. Wavelet analysis: The effect of varying basic wavelet parameters. Sol Phys. 2004;222(2):203–28.

46. Schmidhuber J. Deep Learning in neural networks: An overview. Neural Networks [Internet]. 2015;61:85–117. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003

47. Zhen SH, Cheng M, Tao YB, Wang YF, Juengpanich S, Jiang ZY, et al. Deep Learning for Accurate Diagnosis of Liver Tumor Based on Magnetic Resonance Imaging and Clinical Data. Front Oncol. 2020;10(May):1–14.

48. Zeineldin RA, Karar ME, Coburger J, Wirtz CR, Burgert O. DeepSeg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images. Int J Comput Assist Radiol Surg [Internet]. 2020;15(6):909–20. Available from: https://doi.org/10.1007/s11548-020-02186-z

49. Koonjoo N, Zhu B, Bagnall GC, Bhutto D, Rosen MS. Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. Sci Rep [Internet]. 2021;11(1):1–16. Available from: https://doi.org/10.1038/s41598-021-87482-7

50. Goodfellow I, Bengio Y, Courville A. Deep Learning [Internet]. MIT Press; 2016. Available from: http://www.deeplearningbook.org

51. Gondara L. Medical Image Denoising Using Convolutional Denoising Autoencoders. IEEE Int Conf Data Min Work ICDMW. 2016;0:241–6.

52. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. :1–8.

53. Yao W, Zeng Z, Lian C, Tang H. Pixel-wise regression using U-Net and its application on pansharpening. Neurocomputing [Internet]. 2018;312:364–71. Available from: https://doi.org/10.1016/j.neucom.2018.05.103

54. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT Synthesis Using Unpaired Data. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2017 [cited 2022 Feb 21];10557 LNCS:14–23. Available from: https://link.springer.com/chapter/10.1007/978-3-319-68127-6_2

55. Simpson R, Devenyi GA, Jezzard P, Hennessy TJ, Near J. Advanced Processing and Simulation of MRS Data Using the FID Appliance (FID-A)— An Open Source , MATLAB-Based Toolkit. Magn Reson Med Sci. 2017;33(October 2015):23–33.

56. Hamilton G, Yokoo T, Bydder M, Cruite I, Schroeder ME, Sirlin CB, et al. In vivo characterization of the liver fat 1H MR spectrum. NMR Biomed. 2011;24(7):784–90.

57. Ma J, Wismans C, Cao Z, Klomp DWJ, Wijnen JP, Grissom WA. Tailored spiral in-out spectral-spatial water suppression pulses for magnetic resonance spectroscopic imaging. Magn Reson Med. 2018;79(1):31–40.

58. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32(NeurIPS).

59. Blanken N, Wolterink JM, Brune C, Versluis M. Super-Resolved Microbubble Localization in Single-Channel Ultrasound RF Signals Using Deep Learning. IEEE Trans Med Imaging. 2022;1–1.

60. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34(5):502–8.

61. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 5th Int Conf Learn Represent ICLR 2017 - Conf Track Proc. 2017;1–12.

62. Bitterwolf J, Meinke A, Augustin M, Hein M. Breaking Down Out-of-Distribution Detection: Many Methods Based on OOD Training Data Estimate a Combination of the Same Core Quantities. 2022; Available from: http://arxiv.org/abs/2206.09880

63. Tamrazi B, Nelson MD, Blüml S. MRS of pilocytic astrocytoma: The peak at 2 ppm may not be NAA. Magn Reson Med. 2017;78(2):452–6.

64. Poucin F, Kraus A, Simon M. Boosting Instance Segmentation with Synthetic Data: A study to overcome the limits of real world data sets. Proc IEEE Int Conf Comput Vis. 2021;2021-Octob:945–53.

Appendix I

In this appendix, a description is given of the scripts that are necessary to walk through the pipeline that was built, from simulation to evaluation.

Script 1: simulate_spectra_v2.m

Requirements: Matlab, FID-A toolbox, contwt, invcwt (for the last two, see the reference mentioned on page)

The goal of this script is to generate clean spectra and their contaminated counterparts. Firstly, the spin system is loaded. You can get the spin system from the FID-A toolbox. Then, the spin system for lipids is created. Then, you can go forward to simulation parameters and simulating the PRESS signal and spectrum of the metabolites. I included a table, so you can save the quantifications of the metabolites.

The simulation of different concentrations is done by multiplying the spectrum with a scalar. The scalar is randomized. Please look at the table in the Methods section for a rationale behind these values, and feel free to alter them. The spectra are at the end summed to create a spectrum containing multiple metabolites.

Next, the noise, baseline, water artefacts and lipid artefacts are added to the spectra and the spectra are saved in a real, imaginary and magnitude channel. I think it is wise to remove the magnitude channel, since, in retrospect, it is not needed for fitting and can be computed from the real and imaginary channels anyways. This saves a lot of space as well.

The spectra are then transformed into scalograms using the CWT. Again, the third channel can be removed. Only the real part of the scalogram was used since the imaginary part is neglected in the inverse transform. It was saved as a single to spare space on the SSD and to make loading go faster.

Script 2: Preprocess_StJ.m

Requirements: Matlab, FID-A toolbox, contwt, invcwt

Script meant to make sure the testing data contains the same information as the simulated data. Prior to testing networks on the St. Jude data, you need to manually select the correct voxel, select only the 0-4.2 ppm and perform a wavelet transform like the simulation.

Script 3: Deep learning networks

Requirements: Python (I used the Anaconda and Spyder environment), access to a CUDA processor, see list of libraries in the script

These scripts all follow the same order. First, the libraries are loaded. Then, there is a visualization of spectra or scalograms. After that, a data loader is created. This data loader requires you to store your training and testing data in separate directories. After that, the network is built and the trainer (which iterates through the epochs and updates the weights) is defined). In the next section, you can define your training settings, such as optimizer, learning rate, loss function, etc. The next section starts the training. One error you might run into is a memory error. This can be resolved by setting your batch size lower or by using a network with fewer layers or filters per layer. Then, there are three more sections, meant to loop your test set, your additional test set (e.g. with increasing noise) and your in vivo data.

Script 4: Analyse_scalograms_generated.m

Requirements: Matlab, FID-A toolbox, contwt, invcwt

Using this script, you can load a scalogram (that you have previously saved using one of the deep learning networks). Then, using the inverse wavelet transform, the spectrum can be reconstructed, and compared to the original spectrum. The rest of the code is meant for plotting.

Script 5: sim2sdat.m

Requirements; Matlab, Test01.SPAR file, FID-A toolbox, contwt, invcwt

To fit the data, a mat file will not work in LCModel. You need a .SDAT and corresponding .SPAR file. This script allows you to transform .mat files (containing a spectrum or scalogram) to .SPAR and .SDAT files. The .SPAR file will then contain header data and the .SDAT file will contain spectral data.

Important: make sure that your .SPAR and .SDAT file have the same name! This is a hard necessity for LCModel.

The name of this script can be confusing, since you can use it for spectra, scalograms and in vivo data.

Script 6: create_raw_files.m

Requirements: Matlab, FID-A toolbox.

This script makes .RAW files which are used to fuse into a basis set in LCModel. Please read the comments for the parameters.

.in file

The .in file was used to make the basis set. In short, it tells LCModel which parameters it should use to fuse the .RAW files into a basis set. Make sure there is 1 space at the beginning of each line. You can copy the lines below into a .in text file to make your own makebasis.in file. Seq is the sequence used, echot is the echo time, fwhmba is the full width half maximum of the basis set peaks, hzpppm is the amount of Hertz per ppm (basically the Larmorfrequency at 3T divided by 1.000.000), deltat is the dwell time, which is 0.00083, nunfil is the length of your vector, filbas refers to where the basis set should be stored, filps refers to where the .ps file should be stored, autosc and autoph are your first and second order phase correction (in this case false, since the spectra were simulated without phase errors), idbasi is the name of the basis set, filraw refers to the location of the .RAW file, degzer and degppm are 0 since this is also phase correction, conc is 1 since the .RAW files were simulated with 1mM concentration, ppmapp is the location of the reference peak.

```
$seqpar
seq='PRESS'
echot=135.
fwhmba=.04
$end

$nmall
hzpppm=127.73
deltat=4.0000e-04
nunfil=2048
filbas='.lcmodel/test/output/PRESS3T135ms.basis'
filps='.lcmodel/test/output/PRESS3T135ms.ps'
autosc=.false.
autoph=.false.
idbasi='St Jude basis set (Nov-2022)'
$end
```

```
$nmeach
filraw='.lcmodel/test/rawTE135ms/Ala.RAW'
metabo='Ala'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Asp.raw'
metabo='Asp'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Cit.raw'
metabo='Cit'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Cr.raw'
metabo='Cr'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/GPC.raw'
metabo='GPC'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/GSH.raw'
metabo='GSH'
degzer=0.
```

```
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Gln.raw'
metabo='Gln'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Glu.raw'
metabo='Glu'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/mI.raw'
metabo='Ins'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Lac.raw'
metabo='Lac'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/NAA.raw'
metabo='NAA'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end
```

```
$nmeach
filraw='.lcmodel/test/rawTE135ms/NAAG.raw'
metabo='NAAG'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/PCh.raw'
metabo='PCh'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/PCr.raw'
metabo='PCr'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end

$nmeach
filraw='.lcmodel/test/rawTE135ms/Tau.raw'
metabo='Tau'
degzer=0.
degppm=0.
conc=1.
ppmapp=0., -.4
$end
```