The AGNOSTIC MRS Benchmark Dataset: Deep Learning for Out-of-voxel Artifacts

Aaron T. Gudmundson[a,b], Christopher W. Davies-Jenkins[a,b], İpek Özdemir[a,b], Saipavitra Murali-Manohar[a,b], Helge J. Zöllner[a,b], Yulu Song[a,b], Kathleen E. Hupfeld[a,b], Alfons Schnitzler[c], Georg Oeltzschner[a,b], Craig E. L. Stark[d], and Richard A. E. Edden[a,b]

[a] Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins School of Medicine

[b] F. M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute

[c] Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[d] Department of Neurobiology and Behavior, University of California, Irvine, Irvine, CA.

Word Count: 4466

Number of Figures: 6

Number of Tables: 1

*Corresponding Author: Richard A. E. Edden, Ph.D.

Department of Radiology and Radiological Science, Johns Hopkins School of Medicine

600 N. Wolfe St, Park 338,

Baltimore, MD 21287

E-mail Address: edden@jhu.edu

Abstract:
Purpose:

Neural networks are potentially valuable for many challenges associated with MRS data. The purpose of this manuscript is to describe the AGNOSTIC dataset, which contains 259,200 synthetic MRS examples. To demonstrate the utility, we use AGNOSTIC to train two Convolutional Neural Networks (CNNs) to address out-of-voxel (OOV) echoes. A Detection Network was trained to identify the point-wise presence of OOV echoes, providing proof of concept for real-time detection. A Prediction Network was trained to reconstruct OOV echoes, allowing subtraction during post-processing.

Methods:

AGNOSTIC was created using 270 basis sets that were simulated across 18 field strengths and 15 echo times. The synthetic examples were produced to resemble *in vivo* brain data with combinations of metabolite, macromolecule, and residual water signals, and noise.

Complex OOV signals were mixed into 85% of synthetic examples to train two separate U-net CNNs for the detection and prediction of OOV signals.

Results:

AGNOSTIC is available through Dryad and all Python 3 code is available through GitHub. The Detection network was shown to perform well, identifying 95% of OOV echoes. Traditional modeling of these detected OOV signals was evaluated and may prove to be an effective method during linear-combination modeling. The Prediction Network greatly reduces OOV echoes within FIDs and achieved a median $\log_{10}$ normed-MSE of –1.79, an improvement of almost two orders of magnitude.

Conclusion:

The AGNOSTIC benchmark dataset for MRS is introduced and various dataset features are described. As an exemplar use of AGNOSTIC, two CNNs were developed to detect and predict OOV echoes.

Keywords: Magnetic Resonance Spectroscopy, Synthetic Data, Simulation, Deep Learning, Out-of-voxel Artifacts, Human Brain

1. Introduction:

Proton (^{1}H) magnetic resonance spectroscopy (MRS) non-invasively measures levels of endogenous neurometabolites. MRS-visible metabolites are present at millimolar concentrations in the brain, yielding detectable signals with relatively low signal-to-noise ratio (SNR) which mutually overlap. *In vivo* spectra suffer from several artifacts that complicate modeling and interpretation of the data, including eddy current effects and out-of-voxel (OOV) echoes[1]. While there is some degree of standardization and consensus around pre-processing, modeling, and quantification of MRS data[2–5], this is an evolving field that lacks a single ideal solution due to the complexity of the problem, and which therefore is likely to benefit from recent advances in machine learning.

Deep learning (DL) uses a network consisting of a series of computational layers to process information[6]. Iterative training allows features of the data to be identified and weighted to estimate a final function which predicts a desired output based on a given input[7]. Supervised learning involves training the network based on a pre-defined target, associating ground-truth parameters with each input. An extensive, balanced, and diverse dataset is preferred to increase the generalizability of the DL outcome. High-dimensional data, such as medical images or time series, are demonstrated to be the most beneficial set of data for several computer vision tasks, such as classification, registration, segmentation, reconstruction, and object detection[8,9].

DL has been developed for MRS data as a proof-of-concept in many applications, including metabolite quantification[10–14], phase and frequency correction[15–17], reconstruction of missing data[18], accelerated post-processing[19,20], denoising[21], super-resolution[22], artifact removal[23], and anomaly detection[24]. Despite the potential, these methods have yet to be shown to generalize outside of small datasets with a single fixed acquisition protocol. Before these methods can be adopted more broadly, appropriate tools must be developed to evaluate performance. A key barrier is the lack of a benchmark dataset to play the role that MNIST and ImageNet have played in the field of Computer Vision[25,26]. Such a dataset lowers the barrier to entry for neural network development in MRS and allows performance comparisons between models.

OOV echoes are a substantial issue for *in vivo* MRS, and an under-studied potential DL application. MRS voxel localization is achieved via a combination of RF pulses and magnetic field gradients, with the intended coherence transfer pathway selected both by phase cycling and

dephasing “crusher” gradient scheme[27]. OOV signals arise from gradient echoes – signals from outside the shimmed voxel of interest are refocused by evolution in local field gradients that are either inherent (from air-tissue-bone interfaces) or arising from second-order shim terms[28]. Therefore, brain regions close to air cavities (e.g., medial prefrontal cortex) or which require stronger shim gradients (e.g., thalamus, hippocampus, etc.) most commonly exhibit OOV artifacts[28]. OOV echoes seldom occur at the time of the primary echo, so they manifest in the spectrum as broad peaks with strong first-order phase “ripple” that can obscure metabolite resonances. While acquisition strategies can mitigate OOV echoes to some extent, by careful consideration of crusher schemes or voxel orientation[29,30], post-processing strategies remain valuable where complete elimination is not possible.

This manuscript develops **A**daptable **G**eneralized **N**eural-Network **O**pen-source **S**pectroscopy **T**raining dataset of **I**ndividual **C**omponents (AGNOSTIC), a dataset consisting of 259,200 synthetic MRS examples. AGNOSTIC spans a range of field strengths, echo times, and clinical profiles, representing metabolite signals, macromolecule (MM) background signals, residual water signals, and Gaussian noise as separate components. To date, DL applications to MRS have relied upon narrow in-house-generated training datasets that limit the generalizability of the solutions developed and comparisons between tools; AGNOSTIC is proposed as a benchmark dataset to fill this gap. In order to demonstrate the utility of this resource, we then illustrate a specific augmentation of the AGNOSTIC dataset to train neural networks for the detection and prediction of OOV echoes.

2. Methods:

2.1. AGNOSTIC Synthetic Dataset

The parameter space that AGNOSTIC spans is deliberately broad, comprising: 18 field strengths; 15 echo times; broad distributions of metabolite, MM, and water amplitudes; and densely sampled time-domain to allow down-sampling. Calculations were carried out using an in-house Python 3[31] programming script using NumPy[32]. The dataset is structured as a zipped NumPy[32] archive file (.npz) and can be opened as a Python 3[31] dictionary object. This dictionary contains complex-valued NumPy[32] arrays of time-domain data corresponding to the metabolite, macromolecule, water, and noise components which can be combined in different ways depending on the application. For instance, a denoising model may want to target the combined metabolite, MM, and water signal without noise. Within the file, all the acquisition parameters (field strength, echo time, spectral width, etc.), simulation parameters (signal to noise, full-width half-max, concentrations, T_2 relaxation, etc.), and data augmentation options are specified as detailed below.

2.1.1. Basis Set Simulation:

Metabolite spectra are based upon density-matrix-simulated basis functions[33–36]. A total of 270 basis sets were created across 18 field strengths (1.4 T – 3.1 T in steps of 0.1 T) and 15 echo times (10 ms – 80 ms in steps of 5 ms). The Point RESolved Spectroscopy (PRESS) pulse sequence[37] was simulated using ideal pulses. The simulated spectral width, centered on 4.7 ppm, was 63.62 ppm for all field strengths (e.g., 8 kHz at 3 T; 4 kHz at 1.5 T). The simulated "acquisition window" was started immediately after the last pulse to generate points before the echo. Each metabolite signal was output as a N x 16684 NumPy[32] array, where N is the number of spins for a given metabolite and 16684 is the fixed length of complex time points (300 points before the echo maximum, with an appropriate padding number of zeros and followed by the simulated pre-echo signal, and 16384 points after the echo).

39 brain metabolite basis functions were simulated: Adenosine Triphosphate (ATP); Acetate (Ace); Alanine (Ala); Ascorbate (Asc); Aspartate (Asp); β-hydroxybutyrate (bHB); β-hydroxyglutarate (2HG); Citrate (Cit); Cysteine (Cys); Ethanolamine (EA); Ethanol (EtOH); Creatine (Cr); y-Amino-Butyric Acid (GABA); Glucose (Glc); Glutamine (Gln); Glutamate (Glu); Glycerophosphocholine (GPC); Glutathione (GSH); Glycerol (Glyce); Glycine (Gly); Water (H_2O); Homocarnosine (HCar); Histamine (Hist); Histidine (His); Lactate (Lac); Myo-

Inositol (mI); N-Acetyl-Aspartate (NAA); N-Acetyl-Aspartate-Glutamate (NAAG); Phenylalanine (Phenyl); Phosphocholine (PCho); Phosphocreatine (PCr); Phosphoethanolamine (PE); Scyllo-Inositol (sI); Serine (Ser); Taurine (Tau); Threonine (Thr); Tryptophan (Tryp); Tyrosine (Tyr); and Valine (Val). GABA was separately simulated using two different spin-system enumerations[38,39]. Both α-glucose alpha and β-glucose were simulated.

2.1.2. Assembly of Metabolite Component:

Individual metabolite basis functions were linearly combined to give a metabolite spectral component, weighted by metabolite concentrations sampled from distributions defined by our recent meta-analysis[40], including both healthy and clinical cohort ranges. From the full basis sets, 22 metabolites were selected which had defined concentration ranges available in our recent meta-analysis[40]. Concentrations were selected with equal probability from a range defined by ±2.5 standard deviations from the meta-analysis mean of each cohort[40].

Transverse decay of time-domain data was simulated with exponential and Gaussian components. Intrinsic T_2 relaxation times were based upon the 1.5 T results from a relaxation meta-regression[40] and modeled as an exponential decay[41]. Additional "T_2*" contributions were modeled to achieve a full-width half-maximum (*FWHM*) linewidth of the NAA singlet between 3 Hz and 18 Hz with a uniform distribution by applying appropriate amounts of Gaussian decay. A small amount of jitter (between 20 s^{-2} and 100 s^{-2}) was added to the gaussian decay rate such that each metabolite would undergo a similar, but not identical, amount of Gaussian decay to better replicate the variability observed for *in vivo* data.

2.1.3. Macromolecular Component:

Fourteen MM signals were modeled at: 0.92 ppm; 1.21 ppm; 1.39 ppm; 1.67 ppm; 2.04 ppm; 2.26 ppm; 2.56 ppm; 2.70 ppm; 2.99 ppm; 3.21 ppm; 3.62 ppm; 3.75 ppm; 3.86 ppm; and 4.03 ppm[42,43]. MM chemical shifts were jittered by ± 0.03 ppm. Each MM signal was simulated as a singlet with exponential decay rate sampled uniformly from a range specified by literature of MM T_2 time constants[44] and additional Gaussian decay to reach published linewidths[43,44]. MM amplitudes were sampled uniformly from within published ranges[43,44].

2.1.4. Noise Component:

Noise was generated from a normal distribution, with independent random real and imaginary points. The noise was scaled such that the signal-to-noise ratio of the NAA singlet

(SNR_{NAA}, defined as *NAA height divided by the standard deviation of the noise*) was between 5 and 80, uniformly sampled. The noise amplitude values are also stored within the archive file.

2.1.5. Residual Water Component:

The residual water basis signal was simulated as a singlet at 4.7 ppm. In order to simulate imperfect water suppression, this residual water signal was used to model up to 5 unique water signals with variable ppm locations, phases, and amplitudes[45]. The ranges for these parameters are listed in Table 1. The final water component was scaled to be between 1× and 20× the maximum value of the frequency-domain metabolite spectrum. The water components used, along with their corresponding parametrizations, are stored within the NumPy[32] archive file.

Table 1. Parametrization of the residual water signal

Component	Location		Phase / deg		Amplitude	
1	4.679	4.711	−10	10		1.00
2	4.599	4.641	15	45	.35	.55
3	4.801	4.759	−60	−30	.35	.55
4	4.449	4.541	45	−70	.10	.25
5	4.859	4.901	105	135	.10	.25

2.1.6. Frequency and Phase Shifts:

Within the NumPy[32] archive file, frequency and phase shifts are specified for each entry in the dataset, but not applied to the time-domain components. Frequency shifts were sampled uniformly from the range −0.313 ppm to +0.313 ppm. Zero-order phase shifts were sampled uniformly from the range −1.57 radians to +1.57 radians. First-order phase shifts were sampled uniformly from the range −0.34 to +0.34 radians per ppm.

2.2. Exemplar Application to AGNOSTIC: Machine Learning for Out-Of-Voxel Artifacts:

The primary motivation for the AGNOSTIC dataset is as a training resource for the development of processing, modeling, and analysis tools for MRS. Synthetic spectra with known ground truths are valuable in a range of applications, from the development and validation of traditional linear combination modeling algorithms to training DL models.

In order to demonstrate the utility of the dataset, an exemplar application is presented, in which the AGNOSTIC dataset is supplemented by simulated artifacts (in this case out-of-voxel OOV echoes) and used to train DL models to detect and predict the artifact signals. The AGNOSTIC dataset was developed as building blocks which can be combined to train a variety of different models. A strength of this dataset is that custom user-defined components can be utilized. We demonstrate this point here by building an OOV dataset to train and evaluate a DL model to identify and suppress OOV artifacts.

2.2.1. Simulation of Out-Of-Voxel Echoes:

OOV artifacts were defined as complex time-domain signals with a time point (τ_{OOV}), width (W_{OOV}), frequency (ω_{OOV}), phase (ϕ_{OOV}) and amplitude (a_{OOV}) as shown in Figure 1. τ_{OOV} describes the timepoint of the top of the OOV echo and was sampled randomly from a uniform distribution between 10 ms and 400 ms. W_{OOV} describes the Gaussian decay rate and was sampled randomly from a uniform distribution between 500 s^{-2} and 8000 s^{-2}, resulting in a FWHM echo duration between 18 ms and 74 ms. ω_{OOV} describes the offset in the frequency domain, and was sampled randomly from a uniform distribution in order to produce OOVs that occur between 1 ppm and 4 ppm.

$$Out\ of\ Voxel\ Echo\ =\ a_{OOV}\left(e^{-W_{OOV}(t\ -\ \tau_{OOV})^2}\right)\left(e^{-i\omega t}\right)\left(e^{-i\Phi}\right) \qquad [1]$$

2.2.2. Integration of OOV Echoes into AGNOSTIC for the Training Dataset:

To build the training dataset, we combined metabolite, water, MM, and noise components from the AGNOSTIC dataset. We then added OOV signals to 85% of the dataset and a complex zeros array in the remaining 15%. Finally, we applied the included frequency and phase shifts specified within the AGNOSTIC dataset. The network input consisted of the combined metabolite, water, MM, noise, and OOV signals as a complex time-domain signal. This input was normalized so that the absolute maximum among the real and imaginary values was 1. Finally, training data were converted to a TensorFlow Dataset[46].

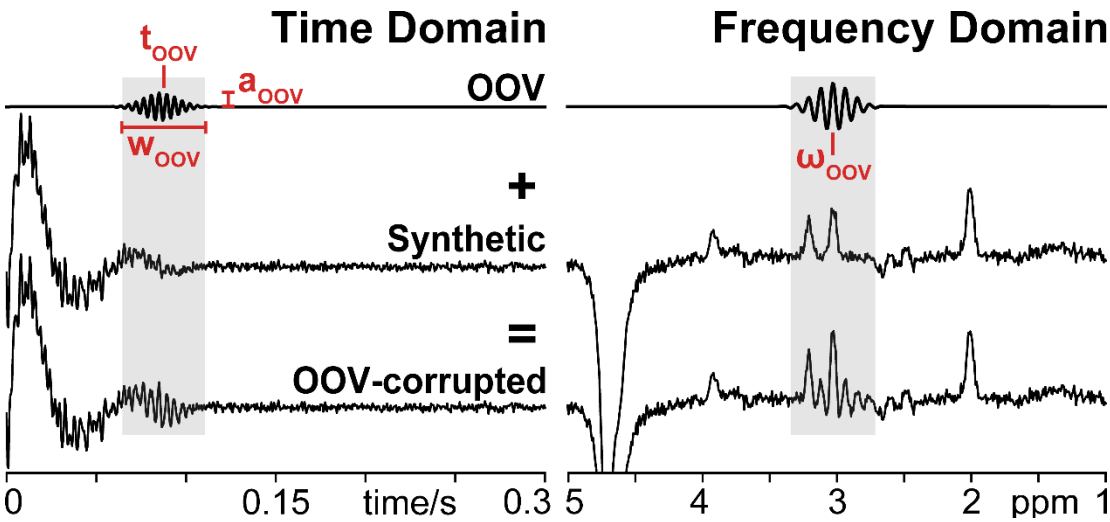

Figure 1. Simulation of OOV echoes and OOV-corrupted synthetic data: OOV echoes were simulated as complex time-domain signals with a center timepoint (τ_{OOV}), width (W_{OOV}), frequency (ω_{OOV}), phase (ϕ_{OOV}), amplitude (a_{OOV}). OOV echoes were added to 85% of synthetic data to create datasets for training and evaluation.

2.2.3. Detection Network:

The first exemplar network is designed to detect OOV echoes within time-domain data by identifying the points in the spectra that have been contaminated by OOV echoes. This Detection Network is a fully Convolutional Neural Network (CNN) designed using TensorFlow2[46] with Keras[47] in a Python 3[31] environment. The network consists of contracting encoding layers and expanding decoding layers with a total of 1.543 million parameters, as shown in Figure 2. Each layer was initialized (kernel_initializer) with "he_normal"[48]. Each convolutional layer (except the output layer which uses a sigmoid activation) includes batch normalization and a leaky rectified linear unit (ReLu) activation function[49]. The network is designed to receive a time-domain input signal and return a binary mask of the same size as the input with ones placed in OOV-detected regions and zeros elsewhere. A ground-truth binary mask was determined as the 5% level of the Gaussian OOV kernel. For training, the input and output of this network is a 60 x 2048 x 2 x 1 tensor, where 60 is the batch number of input examples, 2048 is the number of time points, 2 is the real/imaginary dimension, and 1 is the channel dimension.

The Dice coefficient[50–52] of the overlap between the network output and the correct binary OOV location vector was used as a training loss function, calculated as 2x the intersection divided by the union plus 1; where 1 was used to avoid division by 0. The Adam optimizer was

used with a learning rate of 0.0003. Training was performed using an 8 GB NVIDIA GeForce RTX 3070 GPU. A clustering algorithm was applied to the final network output, which zeroed any group of time points in which the network detected OOV echo that was smaller than 5 consecutive time points, to dampen spurious output.

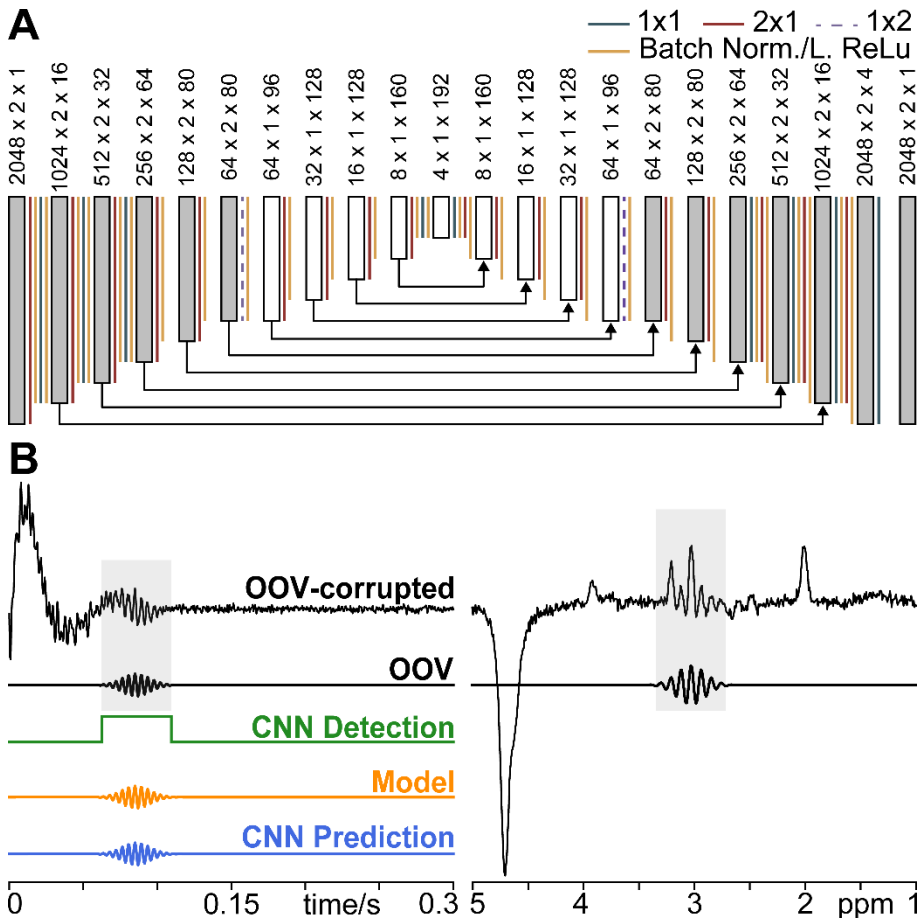

Figure 2. Convolutional Neural Network Architecture, Input, and Output:

A) Fully convolutional neural network architecture used for both the Detection and Prediction Network. Convolutional strides, batch normalization, and Leaky ReLu activation functions are denoted by a colored line. Dark gray blocks represent complex data with the 2nd dimension representing real and imaginary components, while white blocks represent the network abstracted single dimension. Arrows show residual connections. Note, inputs and outputs are all time-domain signals; Frequency-domain is shown for convenient visualization. **B)** OOV-corrupted synthetic example and the isolated OOV. The complex OOV-corrupted data was used as the Detection and Prediction Network input. The target Output is the isolated OOV.

2.2.4. Modeling:

Modeling of the OOV echoes was performed as an optimization problem and solved with SciPy[53] minimization routines. Here, the non-gradient Powell[54,55] optimizer was used to determine the five OOV parameters (τ_{OOV}, W_{OOV}, ω_{OOV}, ϕ_{OOV}, and a_{OOV}), minimizing the mean squared error (MSE) between the model and the data within the time window identified by the Detection Network. Initial values for τ_{OOV}, W_{OOV}, and the a_{OOV} are inferred from the Detection Network's output center timepoint, the detection duration, and the standard deviation of the target signal within the detected region.

Optimization was performed as three sequential optimization steps. The first optimization is used to determine τ_{OOV}, W_{OOV}, and the a_{OOV} by minimizing the MSE between the absolute values of the model and the data (i.e., removing frequency and phase from the model) in the time domain. The second optimization determines ω_{OOV} by absolute-mode minimization in the frequency domain. The third optimization refines the values determined by optimizations 1 and 2 and determines ϕ_{OOV} by complex optimization in the time domain.

2.2.5. Prediction Network:

The second exemplar network is designed to predict the OOV echoes found within time-domain data. This prediction network is also a CNN designed using TensorFlow2[46] with Keras[47] in a Python 3[31] environment, with the same architecture as the detection network (as shown in Figure 2). The network is designed to receive a time-domain input signal containing a combination of the "correct" time-domain signal and the OOV artifact and return a time-domain output signal that only contains the OOV signal, amplified 10x, where the amplification served to globally weight the entire echo. For training, a weighted mean squared error was used as a loss function with an ADAM[56] optimizer and learning rate of 0.0003. Training was again performed on an 8 GB NVIDIA GeForce RTX 3070 GPU.

2.2.6 Evaluating the Performance of Networks and Modeling:

In the final testing set, OOV artifacts were present in 6,137 of the total 7,200 examples (85.2%). The detection network was evaluated using the Dice coefficient[50–52], the overlap

between the ground-truth binary OOV mask and the cluster-thresholded network output. As well as computing global success, the dependence of detection success on various attributes of the OOV echo and the underlying spectrum were also investigated.

Both modeling and the prediction network return a pure OOV signal, and in both cases, the MSE between the prediction/model and the ground-truth OOV echo is used for evaluation. If the ground-truth echo datapoints are E_i and the model or echo prediction is M_i, we calculate the fractional remaining OOV amplitude as:

$$\text{Fractional OOV Remaining} = \frac{\sum |M_i - E_i|^2}{\sum |E_i|^2} \qquad [2]$$

where the bars represent the complex amplitude. The sum is taken over the ground-truth range of the OOV echo. In order to visualize a wide range of success and failure, we take the $\log_{10}$ of this quantity for plotting (i.e., a $\log_{10}$ value of 1 is no change, anything positive is a manipulation that is worse than doing nothing, and a negative value show the order of magnitude of improvement). Note that E_i is the ground-truth echo signal, not the signal from which the echo is being removed which also contains metabolite, macromolecule, and noise components.

The timing of the OOV was found to be a key parameter determining the success of detection and prediction, and as a result, the evaluation metrics were calculated for the following time-bins (based on the known value of τ_{OOV}): 10-20 ms; 20-40 ms; 40-60 ms; 60-80 ms; 80-120 ms; 120-200 ms; 200-300 ms; 300-400 ms.

2.2.7 *In Vivo* Proof-of-Principle

As a proof-of-principle demonstration of this exemplar use of the AGNOSTIC dataset, the network was applied to 256 transients of *in vivo* data, selected because they contain prominent OOV echoes and were excluded during quality assessment in a recent study[57]. These data were collected on a 2.89 T Siemens scanner using the MEGA-PRESS[58,59] pulse sequence with a TE of 68 ms and TR of 1.75 s, and a spectral width of 2.4 kHz. Note that this challenges the generality of the training because the network has never seen data acquired at 2.89 T, nor at 2.4 kHz spectral width, nor at TE 68 ms, nor with MEGA-Editing, nor with actual real RF pulses. Raw data from a 25 x 25 x 25 mm^3 voxel in the cerebellum were loaded and coil

combined in Osprey[60]. Time-domain data were saved as a MATLAB[61] .mat file and loaded as a Python 3[31] object using SciPy[53]. The data were normalized (as above with training data) to be used as input for the neural networks.

One challenge of *in vivo* data (and the reason that this network demonstration focuses substantially on synthetic data) is that no ground truth is available. Therefore, the degree of success in removing OOV echo signals from time-domain data D_i is:

$$\text{Fractional Reduction in standard deviation} = 1 - \frac{\sigma(M_i - D_i)}{\sigma(D_i)} \qquad [3]$$

where σ denotes the standard deviation. Note that, in contrast to the metric used for synthetic data in Equation 1, only D_i is available, not the ground truth E_i, which substantially changes the ceiling of success. It is still expected that substantial signal variance remains after OOV removal, since D_i contains metabolite signals and noise. The range over which this standard deviation is calculated is the 50% level of the normalized histogram of the detection network's output across the 128 averages.

3. Results:

3.1. AGNOSTIC Synthetic Dataset:

The AGNOSTIC dataset contains 259,200 examples, consisting of 960 examples from each of the eighteen field strengths and fifteen echo times (i.e., 960x18x15=259,200). A representative set of ten spectra are shown in Figure 3, illustrating the diversity of field strengths, TEs, SNR, and linewidth within the dataset.

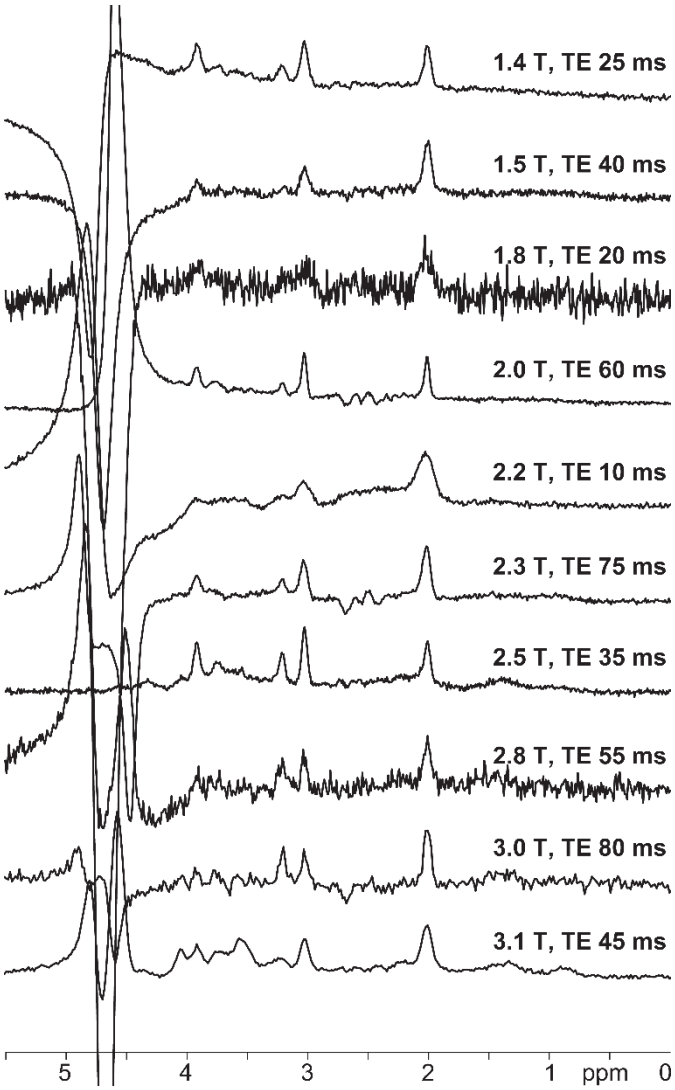

Figure 3. AGNOSTIC synthetic dataset:

10 representative spectra from the AGNOSTIC dataset. The 10 examples show the diversity of field strength, TE, linewidths, and residual water signal present among the data. Note, examples are shown here in the frequency-domain to better illustrate the heterogeneity, but the dataset provides time-domain examples.

One challenge with making this dataset available is its size — 75 GB — but we do make it freely available on Dryad. The basis sets from which these are constructed are more manageable — 9 GB — and can also be accessed through Dryad. Code for generating the AGNOSTIC dataset locally is available at: https://github.com/agudmundson/agnostic.

3.2. Exemplar Application to AGNOSTIC: Machine Learning for Out-Of-Voxel Artifacts:

3.2.1. Detection Network:

Of the 6,137 examples where OOV artifacts were present, the Detection Network correctly identified 5,827 (94.9%) with a median Dice score of 0.974 (0.941–0.985 interquartile range) and missed 310 (5.05%) with a Dice score of 0.00. In the 1063 examples that did not include OOV artifacts, the network correctly ignored 912 (85.8%) and falsely detected OOV echoes in 151 (14.2%). Figure 4 shows the Detection Network's output for a synthetic OOV-corrupted example.

Analysis of the factors that determined success indicated that the time at which OOV signals occur is most critical. Therefore, OOV echoes were further broken down into eight time-bins, and the Dice score plotted in Figure 5. The median Dice scores — 0.165, 0.858, 0.892, 0.934, 0.960, 0.974, 0.978, and 0.978 — are poor in the first bin and improve thereafter. Note that these bins are not spaced equally to emphasize poor performance extremely early. The number of examples in each bin is 161, 289, 282, 329, 622, 1256, 1565, and 1633, respectively.

3.2.2. Modeling

The modeling optimization converged in 5,824 of the 5,827 examples where the detection network detected OOV artifacts and provided initial values. Across this subset of the examples, the modeling achieved a median $\log_{10}$ (fractional OOV remaining) of −2.19 (−2.90 – −1.19 inter-quartile range), i.e., a median reduction of more than two orders of magnitude. Figure 5 shows the resulting model for a synthetic OOV-corrupted example.

These values — broken down into 8 time-bins — are shown in Figure 5. The median $\log_{10}$(fractional OOV remaining) decreases across the time bins: 1.663, 1.324, 0.680, 0.223, −1.586, −2.276, −2.491, and −2.567.

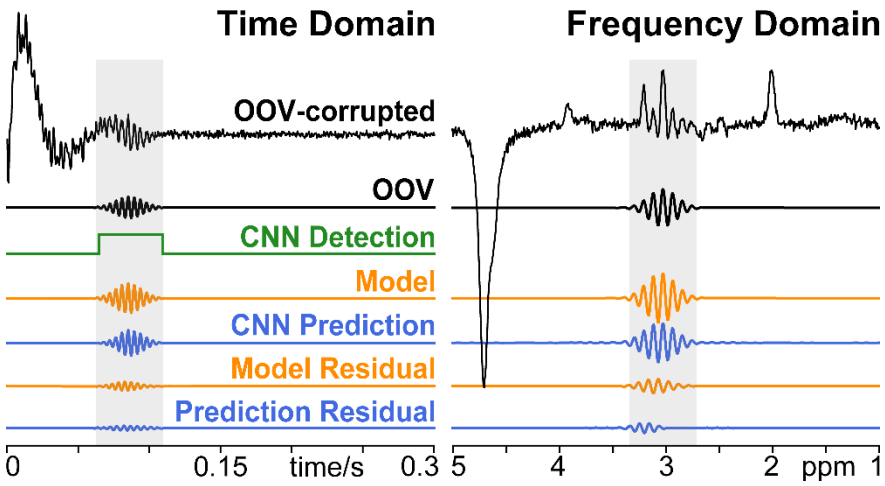

Figure 4. OOV-corrupted example:

OOV-corrupted synthetic example and the isolated OOV. Results from Detection Network (green), Model (orange), and Prediction Network (blue) are shown below the ground truth OOV-corrupted and OOV. OOV residuals are shown for the Model (orange) and Prediction Network (blue) demonstrating remaining signal after subtraction. Note, frequency-domain is shown for convenient visualization, but the Detection Network, Modeling, and Prediction Network all operate on time-domain signals.

3.2.3. Prediction Network:

In the 6,137 examples where OOV artifacts were present, the prediction network achieved a median $\log_{10}$ normed-MSE of −1.79 (−2.21 – −1.11 inter-quartile range). In the 5,824 examples where OOV artifacts were successfully modeled, the prediction network achieved a median $\log_{10}$ normed-MSE of −1.85 (−2.24 – −1.24 inter-quartile range). Figure 5 shows the Prediction Network's output for a synthetic OOV-corrupted example.

OOVs were further broken down into 8 time-bins (Figure 5) early — the number of examples in each bin is 86, 226, 261, 312, 592, 1208, 1538, 1601. The median $\log_{10}$(fractional OOV remaining) decreases across the time bins: −0.207, −0.583, −0.862, −1.250, −1.577, −1.878, −2.005, and −2.052.

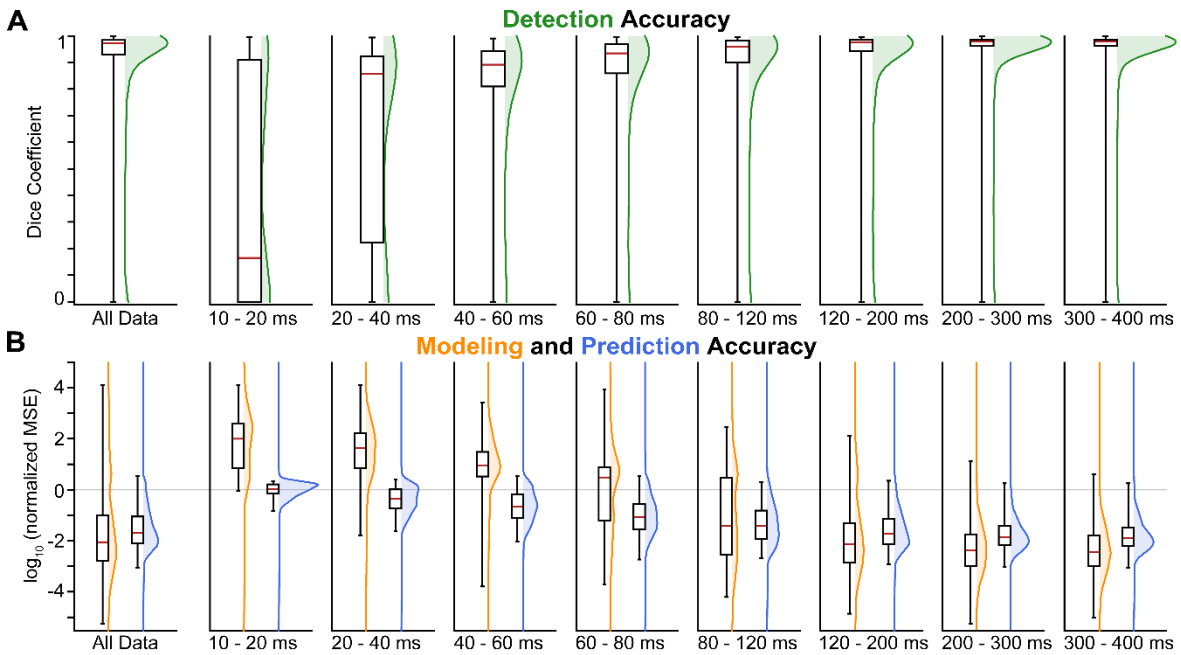

Figure 5. Evaluation of Detection Network, Modeling, and Prediction Network: A testing set with 7200 (2400 examples with 3 different OOV echoes) unseen examples was used to evaluate the **A)** Detection Network and **B)** Modeling and Prediction Network. Performance across the whole test set is shown on the left-hand side. Performance across the binned center timepoint (τOOV) is shown across the right-hand side.

3.2.4. *In Vivo* Proof-of-Principle:

The Detection Network identified an OOV in 243 of 256 transients (94.9%). In these 243 OOV-detected transients, the modeling achieved a median reduction in standard deviation of 71.0 % (60.2 -75.3% inter-quartile range). The Prediction Network achieved a median reduction in standard deviation of 69.65% (66.33 %/72.7 % inter-quartile range) in this subset. In the full set of 256 transients, the Prediction Network achieved a median 69.4 % (65.3 – 72.6 % inter-quartile range) reduction in standard deviation. The standard deviation of the noise floor was found to account for a median of 10.3% (9.35–11.6 % inter-quartile range) of the standard deviation of signal within the time window for the 256 averages. A representative *in vivo* example is shown in Figure 6.

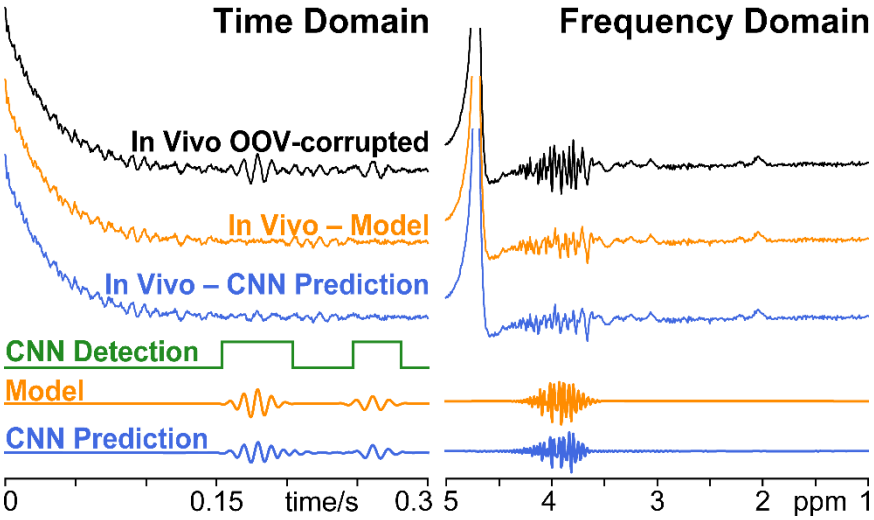

Figure 6. In Vivo MEGA-PRESS Example:

In vivo OOV-corrupted example. Results from Detection Network (green), Model (orange), and Prediction Network (blue) are shown below. Note, frequency-domain is shown for convenient visualization, but the Detection Network, Modeling, and Prediction Network all operate on time-domain signals.

4. Discussion:

AGNOSTIC is a benchmark MRS dataset for training and evaluating performance across various models. In order to make these synthetic data representative of in vivo brain MRS datasets, a total of 22 brain metabolites and 14 MM peaks were simulated within 270 basis sets, spanning field strengths from 1.4 T to 3.1 T and TEs from 10 to 80 ms. Parameterized water residual and noise were included. SNR and linewidths were assigned at random, independent of B_0 or TE. The broad span of the dataset is key in training networks that generalize. While AGNOSTIC is broad in these dimensions, it does only represent simulated data for PRESS[37] acquisitions, and may benefit from expansion to include other pulse sequences, such as STEAM[62], SPECIAL[63,64], LASER[65], and semi-LASER[66,67], and edited versions including MEGA[58,59] and Hadamard-encoded[68–71] schemes. AGNOSTIC simulations used ideal pulses, and thus fail to capture effects associated with spatially heterogeneous coupling evolution. The extent to which these limitations matter will depend on the applications that AGNOSTIC synthetic data are being used for.

The Detection network was highly successful, identifying 94.9% of the testing set where OOV artifacts were present. The precise value of this success metric is obviously impacted by the parameters of the OOVs – a later minimum OOV time would tend to increase performance, and earlier would degrade it. It is noteworthy that, although the training datasets never contained more than one OOV echo, the detection and prediction networks were able to handle more than one OOV echo in vivo data, presumably because CNNs operate locally within the FID. It is also encouraging that the networks generalized well to the *in vivo* data (Figure 6), which was collected with unseen acquisition parameters, i.e., edited MEGA-PRESS[58,59] data acquired at 2.89 T with a TE of 68 ms, and 2.4 kHz spectral width.

In the exemplar OOV application, the success of the networks depended heavily on the timing of the OOV signal. The earliest OOV echoes were most challenging, unsurprisingly since such signals are broad Gaussian resonances that are indistinguishable from within-voxel MM and baseline signals. Indeed, the only feature that differentiates OOV signals from other broad components of the model is timing. It is conceptually helpful to consider this in the Fourier domain, even though all network processing is performed in the time domain. In the frequency domain, a mismatch between the echo-top and the acquisition start is represented as a first-order phase error of the signal associated with that echo. Where insufficient first-order phase exists to

be represented within the linewidth of the signal in question, (which in the time domain corresponds to substantial truncation of the lefthand side of the echo), the network struggles to identify OOV signals.

In the context of this study, modeling and prediction are treated as two alternative approaches to OOV characterization. For early OOV signals, the modeling approach tended to mis-attribute non-artifact signal as OOV signal, a result that the metric scored as worse than no intervention. The median performance of the Prediction network, even for very early OOV signals, was close to zero. Both modeling and prediction performance improve as the OOV moves later in the acquired signal, with modeling improving faster than the network, and performing better than prediction beyond 120 ms. This strong performance of the model at least in part reflects the exact match between the generative model of the synthetic OOV artefacts and the model that is being used to extract them. More moderate performance might be expected for real in vivo examples – but the same may also be true for networks which have been trained with the same synthetic data and may have learned specifically to identify OOV signals that have a Gaussian kernel.

One key difference between most DL applications and applications in MRS, is the strict requirement to preserve amplitude fidelity in network outputs. A common approach to artifacts in DL is to return an artifact-free version of the network input. In contrast, the approach taken here is to return the artifact, which has the following benefits: it avoids networks over-learning the formulaic pattern of typical spectra; it reduces the impact of the lack of sequence diversity within the AGNOSTIC dataset; and it is less likely to impact the amplitudes of metabolite signals.

The ultimate goal of this work is to extract metabolite levels from MRS data that are not impacted by OOV artifacts. This problem can be addressed at several points: either by not acquiring data that contain OOV artifacts; by removing OOV artifacts post-acquisition; and by incorporating appropriate OOV model components into quantification model so that the impact of OOV is minimized. While the work presented here focuses primarily on the second context, it raises important potential applications in the other contexts. One motivator for developing the Detection network is the possibility of real-time deployment during sequence acquisition to trigger sequence changes when OOV artifacts are detected. The modeling applied here was time-

restricted to a given window and ignored other components of the spectrum, but demonstrates potential for future integration within a full linear-combination model.

In conclusion, we have presented the AGNOSTIC dataset for deep learning in MRS and demonstrated an exemplar use case to develop CNNs to detect and predict out-of-voxel artifacts.

Funding Information:

This work has been supported by The Henry L. Guenther Foundation, Sonderforschungsbereich (SFB) 974 (TP B07) of the German Research foundation, and the National Institute of Health, grants T32 AG00096, R00 AG062230, R21 EB033516, R01 EB016089, R01 EB023963, K00AG068440, P30 AG066519, R21 AG053040, R01 AG076942, P30 AG066519 and P41 EB031771.

References:

1. Kreis R. Issues of spectral quality in clinical1H-magnetic resonance spectroscopy and a gallery of artifacts. *NMR Biomed*. 2004;17(6):361-381. doi:10.1002/nbm.891

2. Near J, Harris AD, Juchem C, et al. Preprocessing, analysis and quantification in single voxel magnetic resonance spectroscopy: experts' consensus recommendations. *NMR Biomed*. 2020;(December 2019):1-23. doi:10.1002/nbm.4257

3. Wilson M, Andronesi O, Barker PB, et al. Methodological consensus on clinical proton MRS of the brain: Review and recommendations. *Magn Reson Med*. 2019;82(2):527-550. doi:10.1002/mrm.27742

4. Öz G, Deelchand DK, Wijnen JP, et al. Advanced single voxel 1H magnetic resonance spectroscopy techniques in humans: Experts' consensus recommendations. *NMR Biomed*. 2021;34(5):1-18. doi:10.1002/nbm.4236

5. Maudsley AA, Andronesi OC, Barker PB, et al. Advanced magnetic resonance spectroscopic neuroimaging: Experts' consensus recommendations. *NMR Biomed*. 2021;34(5):1-22. doi:10.1002/nbm.4309

6. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444. doi:10.1038/nature14539

7. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. The MIT Press; 2016. http://www.deeplearningbook.org/.

8. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. 2019;29(2):102-127. doi:10.1016/j.zemedi.2018.11.002

9. Gassenmaier S, Küstner T, Nickel D, et al. Deep Learning Applications in Magnetic Resonance Imaging: Has the Future Become Present? *Diagnostics*. 2021;11(12):2181. doi:10.3390/diagnostics11122181

10. 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. doi:10.1002/mrm.27727

11. 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):1689-1706. doi:10.1002/mrm.28234

12. Hatami N, Sdika M, Ratiney H. Magnetic resonance spectroscopy quantification using deep learning. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2018;11070 LNCS:467-475. doi:10.1007/978-3-030-00928-1_53

13. Chandler M, Jenkins C, Shermer SM, Langbein FC. MRSNet: Metabolite Quantification from Edited Magnetic Resonance Spectra With Convolutional Neural Networks. 2019:1-12. http://arxiv.org/abs/1909.03836.

14. Rizzo R, Dziadosz M, Kyathanahally SP, Shamaei A, Kreis R. Quantification of MR spectra by deep learning in an idealized setting: Investigation of forms of input, network architectures, optimization by ensembles of networks, and training bias. *Magn Reson Med*. 2023;89(5):1707-1727. doi:10.1002/mrm.29561

15. Tapper S, Mikkelsen M, Dewey BE, et al. Frequency and phase correction of J-difference edited MR spectra using deep learning. *Magn Reson Med*. 2021;85(4):1755-1765. doi:10.1002/mrm.28525

16. Shamaei A, Starcukova J, Pavlova I, Starcuk Z. Model□informed unsupervised deep learning approaches to frequency and phase correction of MRS signals. *Magn Reson Med*. 2023;89(3):1221-1236. doi:10.1002/mrm.29498

17. Ma DJ, Le HAM, Ye Y, et al. MR spectroscopy frequency and phase correction using convolutional neural networks. *Magn Reson Med*. 2022;87(4):1700-1710. doi:10.1002/mrm.29103

18. Lee H, Lee HH, Kim H. Reconstruction of spectra from truncated free induction decays by deep learning in proton magnetic resonance spectroscopy. *Magn Reson Med*. 2020;84(2):559-568. doi:10.1002/mrm.28164

19. Gurbani SS, Sheriff S, Maudsley AA, Shim H, Cooper LAD. Incorporation of a spectral model in a convolutional neural network for accelerated spectral fitting. *Magn Reson Med*. 2019;81(5):3346-3357. doi:10.1002/mrm.27641

20. Iqbal Z, Nguyen D, Thomas MA, Jiang S. Deep learning can accelerate and quantify

simulated localized correlated spectroscopy. *Sci Rep*. 2021;11(1):8727. doi:10.1038/s41598-021-88158-y

21. Chen D, Hu W, Liu H, et al. Magnetic Resonance Spectroscopy Deep Learning Denoising Using Few In Vivo Data. *IEEE Trans Comput Imaging*. 2023:1-12. doi:10.1109/TCI.2023.3267623

22. Gassenmaier S, Afat S, Nickel D, et al. Application of a Novel Iterative Denoising and Image Enhancement Technique in T1-Weighted Precontrast and Postcontrast Gradient Echo Imaging of the Abdomen. *Invest Radiol*. 2021;56(5):328-334. doi:10.1097/RLI.0000000000000746

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):851-863. doi:10.1002/mrm.27096

24. Jang J, Lee HH, Park JA, Kim H. Unsupervised anomaly detection using generative adversarial networks in 1H-MRS of the brain. *J Magn Reson*. 2021;325:106936. doi:10.1016/j.jmr.2021.106936

25. Fei-Fei L, Deng J, Li K. ImageNet: Constructing a large-scale image database. *J Vis*. 2010;9(8):1037-1037. doi:10.1167/9.8.1037

26. Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Process Mag*. 2012;29(6):141-142. doi:10.1109/MSP.2012.2211477

27. Bodenhausen G. Reflections of pathways: A short perspective on ‘Selection of coherence transfer pathways in NMR pulse experiments.’ *J Magn Reson*. 2011;213(2):295-297. doi:10.1016/j.jmr.2011.08.004

28. Starck G, Carlsson A, Ljungberg M, Forssell-Aronsson E. k-space analysis of point-resolved spectroscopy (PRESS) with regard to spurious echoes in in vivo (1)H MRS. *NMR Biomed*. 2009;22(2):137-147. doi:10.1002/nbm.1289

29. Song Y, Zöllner HJ, Hui SCN, Hupfeld KE, Oeltzschner G, Edden RAE. Impact of gradient scheme and non□linear shimming on out□of□voxel echo artifacts in edited

MRS. *NMR Biomed*. 2023;36(2). doi:10.1002/nbm.4839

30. Ernst T, Chang L. Elimination of artifacts in short echo time1H MR spectroscopy of the frontal lobe. *Magn Reson Med*. 1996;36(3):462-468. doi:10.1002/mrm.1910360320

31. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.

32. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

33. Fano U. Description of States in Quantum Mechanics by Density Matrix and Operator Techniques. *Rev Mod Phys*. 1957;29(1):74-93. doi:10.1103/RevModPhys.29.74

34. Blum K. *Density Matrix Theory and Applications*. 1st ed. Boston, MA: Springer US; 1981. doi:10.1007/978-1-4615-6808-7

35. Farrar T. Density matrices in NMR spectroscopy: Part I. *Concepts Magn Reson*. 1990;2:1-12.

36. Sørensen OW, Eich GW, Levitt MH, Bodenhausen G, Ernst RR. Product operator formalism for the description of NMR pulse experiments. *Prog Nucl Magn Reson Spectrosc*. 1984;16:163-192. doi:10.1016/0079-6565(84)80005-9

37. Bottonmley PA. SELECTIVE VOLUME METHOD FOR PERFORMING LOCALIZED NMR SPECTROSCOPY. 1984. https://mriquestions.com/uploads/3/4/5/7/34572113/bottomley_press_patent_1984.pdf.

38. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR Biomed*. 2000;13(3):129-153. doi:10.1002/1099-1492(200005)13:3<129::AID-NBM619>3.0.CO;2-V

39. Near J, Leung I, Claridge T, Cowen P, Jezzard P. Chemical shifts and coupling constants of the GABA spin system. *Proc Intl Soc Mag Reson Med*. 2012;20(1993):4386. http://cds.ismrm.org/protected/12MProceedings/files/4386.pdf.

40. Gudmundson AT, Koo A, Virovka A, et al. Meta-analysis and Open-source Database for In Vivo Brain Magnetic Resonance Spectroscopy Studies of Health and Disease. 2023.

doi:10.1101/2023.02.10.528046

41. Juchem C, de Graaf RA. B0 magnetic field homogeneity and shimming for in vivo magnetic resonance spectroscopy. *Anal Biochem.* 2017;529:17-29. doi:10.1016/j.ab.2016.06.003

42. Cudalbu C, Behar KL, Bhattacharyya PK, et al. Contribution of macromolecules to brain 1H MR spectra: Experts' consensus recommendations. *NMR Biomed*. 2021;34(5):1-24. doi:10.1002/nbm.4393

43. Giapitzakis IA, Avdievich N, Henning A. Characterization of macromolecular baseline of human brain using metabolite cycled semi-LASER at 9.4T. *Magn Reson Med*. 2018;80(2):462-473. doi:10.1002/mrm.27070

44. Murali-Manohar S, Borbath T, Wright AM, Soher B, Mekle R, Henning A. T2 relaxation times of macromolecules and metabolites in the human brain at 9.4 T. *Magn Reson Med*. 2020;84(2):542-558. doi:10.1002/mrm.28174

45. Lin L, Považan M, Berrington A, Chen Z, Barker PB. Water removal in MR spectroscopic imaging with L2 regularization. *Magn Reson Med*. 2019;82(4):1278-1287. doi:10.1002/mrm.27824

46. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale Machine Learning on Heterogeneous Distributed Systems. 2015. doi:10.5281/zenodo.4724125

47. Chollet F, others. Keras. 2015.

48. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Vol 2015 Inter. IEEE; 2015:1026-1034. doi:10.1109/ICCV.2015.123

49. Agarap AF. Deep learning using rectified linear units (relu). *arXiv Prepr arXiv180308375*. 2018.

50. Dice LR. Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by□: Ecological Society of America Stable URL□:

http://www.jstor.org/stable/1932409. *Ecology*. 1945;26(3):297-302.

51. Carass A, Roy S, Gherman A, et al. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Sci Rep*. 2020;10(1):1-19. doi:10.1038/s41598-020-64803-w

52. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *K Danske Vidensk Selsk*. 1948;5(4):1-34.

53. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

54. Powell MJD. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput J*. 1964;7(2):155-162. doi:10.1093/comjnl/7.2.155

55. Powell MJD. A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation. In: Gomez S, Hennart JP, eds. *Advances in Optimization and Numerical Analysis*. Dordrecht: Springer Netherlands; 1994:51-67. doi:10.1007/978-94-015-8330-5_4

56. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. 2015:1-15.

57. Zöllner HJ, Thiel TA, Füllenbach ND, et al. J-difference GABA-edited MRS reveals altered cerebello-thalamo-cortical metabolism in patients with hepatic encephalopathy. *Metab Brain Dis*. 2023;38(4):1221-1238. doi:10.1007/s11011-023-01174-x

58. Mescher M, Tannus A, O'Neil Johnson M, Garwood M. Solvent suppression using selective echo dephasing. *J Magn Reson - Ser A*. 1996;123(2):226-229. doi:10.1006/jmra.1996.0242

59. Mescher M, Merkle H, Kirsch J, Garwood M, Gruetter R. Simultaneous in vivo spectral editing and water suppression. *NMR Biomed*. 1998;11(6):266-272. doi:10.1002/(SICI)1099-1492(199810)11:6<266::AID-NBM530>3.0.CO;2-J

60. Oeltzschner G, Zöllner HJ, Hui SCN, et al. Osprey: Open-source processing, reconstruction & estimation of magnetic resonance spectroscopy data. *J Neurosci Methods*. 2020;343(June):108827. doi:10.1016/j.jneumeth.2020.108827

61. The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). 2022. https://www.mathworks.com.

62. Frahm J, Merboldt KD, Hänicke W. Localized proton spectroscopy using stimulated echoes. *J Magn Reson*. 1987;72(3):502-508. doi:10.1016/0022-2364(87)90154-5

63. Mekle R, Mlynárik V, Gambarota G, Hergt M, Krueger G, Gruetter R. MR spectroscopy of the human brain with enhanced signal intensity at ultrashort echo times on a clinical platform at 3T and 7T. *Magn Reson Med*. 2009;61(6):1279-1285. doi:10.1002/mrm.21961

64. Mlynarik V, Gambarota G, Frenkel H, Gruetter R. Localized short-echo-time proton MR spectroscopy with full signal-intensity acquisition. *Magn Reson Med*. 2006;56(5):965-970. doi:10.1002/mrm.21043

65. Garwood M, DelaBarre L. The return of the frequency sweep: Designing adiabatic pulses for contemporary NMR. *J Magn Reson*. 2001. doi:10.1006/jmre.2001.2340

66. Scheenen TWJ, Heerschap A, Klomp DWJ. Towards 1H-MRSI of the human brain at 7T with slice-selective adiabatic refocusing pulses. *Magn Reson Mater Physics, Biol Med*. 2008;21(1-2):95-101. doi:10.1007/s10334-007-0094-y

67. Scheenen TWJ, Klomp DWJ, Wijnen JP, Heerschap A. Short echo time1H-MRSI of the human brain at 3T with minimal chemical shift displacement errors using adiabatic refocusing pulses. *Magn Reson Med*. 2008;59(1):1-6. doi:10.1002/mrm.21302

68. Oeltzschner G, Saleh MG, Rimbault D, et al. Advanced Hadamard-encoded editing of seven low-concentration brain metabolites: Principles of HERCULES. *Neuroimage*. 2019;185(September 2018):181-190. doi:10.1016/j.neuroimage.2018.10.002

69. Chan KL, Puts NAJ, Schär M, Barker PB, Edden RAE. HERMES: Hadamard encoding and reconstruction of MEGA-edited spectroscopy. *Magn Reson Med*. 2016;76(1):11-19. doi:10.1002/mrm.26233

70. Chan KL, Edden RAE, Barker PB. Simultaneous editing of GABA and GSH with Hadamard encoded MR spectroscopic imaging. 2019;(July 2018):21-32. doi:10.1002/mrm.27702

71. Saleh MG, Oeltzschner G, Chan KL, et al. Simultaneous edited MRS of GABA and glutathione. *Neuroimage*. 2016;142:576-582. doi:10.1016/j.neuroimage.2016.07.056