The Impact of Training Data on MRS Metabolite Quantification with Deep Learning

Z. Ma[1], O. Karakus[1], S. M. Shermer[2], F. C. Langbein[1]

[1]School of Computer Science and Informatics, Cardiff University, UK

[2]Physics, Faculty of Science, Swansea University, UK

Primary category - subcategory: **AI and Machine Learning - Analysis/Processing**
Secondary category - subcategory: **Contrast Mechanisms - Spectroscopy**
General keyword: **Spectroscopy**
Other keywords: **Magnetic Resonance Spectroscopy, Metabolite Quantification, Deep Learning, Simulation**

Synopsis

Motivation: We aim to improve metabolite quantification in MRS with deep learning by investigating the impact of training data quality on model performance.

Goal(s): The goal is to investigate dataset factors, specifically the variability of metabolite models and noise realism, impacting the quantification performance of a deep learning architecture trained on varying datasets.

Approach: We evaluate deep learning models for metabolite quantification on experimental phantom spectra that were trained on varying simulated datasets.

Results: Results show that training datasets significantly impact quantification performance. Specifically, more realistic noise models yield improvements.

Impact: This study underscores the importance of training data quality in deep learning for MRS. By demonstrating the impact of noise model realism, it provides insights for developing more accurate metabolite quantification models, potentially improving clinical diagnosis and monitoring neurological disorders.

Introduction

Deep learning has potential to enhance metabolite quantification in magnetic resonance spectroscopy (MRS)[1]. However, machine learning methods struggle to surpass traditional techniques like LCMODEL[2] in accuracy[3]. While complex deep learning architectures have been explored, the impact of training data quality on quantification performance remains understudied. Limited availability of real data, due to costs and the difficulty in ascertaining ground truth concentrations, often necessitates the use of simulated datasets, which can lack the diversity of real spectra.

We investigate the role of two key factors in simulated training datasets: (1) the variability of chemical shifts and J-couplings in metabolite models and (2) the realism of the noise model. A Y-shaped autoencoder with state-of-the-art performance is employed to quantify GABA, Creatine, Glutamine, Glutamate, and NAA with MEGAPRESS[4]. We train and validate it on varying simulated

datasets and analyze its performance on phantom spectra[5]. Results show an influence of training dataset simulation on model performance.

Method

To generate training datasets, we simulate basis spectra for each metabolite with FID-A[6] using MEGAPRESS yielding edit-off and edit-on edited spectral shapes. Weighted sums of these basis shapes create the simulated spectra. The weights represent relative concentrations, sampled between 0.0 and 1.0 using Sobol low-discrepancy sampling to evenly represent concentrations for 10^5 samples. Afterwards, noise is added to each spectrum.

We explore two basis spectra simulations (single, multi) and two noise models (gg, adc), resulting in four datasets: single-adc, single-gg, multi-adc, and multi-gg. Single has a fixed basis shape per metabolite using the default FID-A metabolite chemical shifts and J-couplings. For multi we select a range of metabolite models from the literature[7-10] to cover a wider range of spectral shapes; for each spectrum, a random basis shape per metabolite is chosen. Adc noise adds Gaussian noise in the time domain; gg noise adds generalized Gaussian noise to the frequency domain instead. Adc noise parameters are estimated using frequency ranges without metabolite signal in phantom spectra, while MCMC estimation[11] is used for the gg noise parameters.

The impact of the training dataset is investigated with a parameterized Y-shaped autoencoder architecture[4] in Figure 1. The parametrization enables the selection of the best-performing architecture parameters, such as the number of neurons per layer, using Gaussian process optimization. To focus on the impact of the dataset, optimal parameters have been selected via the average performance in five-fold cross-validation on two simulated datasets: single-adc and multi-gg, giving architectures A and B respectively. Both architectures are trained on the four datasets and their performance is compared on experimental phantom spectra[5].

Results

Figure 2 shows the mean absolute error (MAE) distributions for the phantom spectra for architectures A and B trained on four datasets. Models trained on gg generally perform better than models trained on adc. For A, there is little difference between multi and single, but the single-gg model performs best. For B, we see an improvement using gg for single which also shows the best performance, but no improvement independent of noise for multi.

Figure 3 compares the MAE distributions per metabolite, showing similar but more subtle behavior compared to Figure 2, with greater uncertainty in the Glutamine and Glutamate concentrations due to the similarity of their signal.

To carefully analyze the difference in the MAE distributions, we present statistical tests to compare the overall MAE distributions from Figure 2 in Figures 4 and 5.

Discussion

Architecture A has been optimized for single-adc, i.e. single basis spectra shapes. B has been optimized for the more complex multi-gg data, i.e. multiple basis spectra shapes. In both cases, we see that gg noise improves the results, except for B in the multi case which is slightly worse. Likely this is because architecture B has been optimized for the multi case. B trained on single-gg performs best, with A on single-gg being the second best.

Overall, models trained on single data outperform those trained on multi data. Notably, gg noise yields superior results compared to adc noise. Training with multiple basis shapes does not improve accuracy. Further improvements may be achieved by refining single basis shape models to match experimental data better.

Conclusion

Our results demonstrate that the training dataset, particularly the choice of noise model, impacts quantification performance, highlighting the importance of realistic simulations. Our results are limited to a few cases but clearly indicate potential for investigating the training datasets. Focusing on improving the realism of simulations or obtaining large real datasets may yield substantial improvements in quantifying metabolites in MRS spectra.

References

1. van de Sande DMJ, Merkofer JP, Amirrajab S, et al. A review of machine learning applications for the proton MR spectroscopy workflow. Magnetic Resonance in Medicine 2023; 90(4):1253–1270.
2. Provencher, SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magnetic Resonance in Medicine 1993; 30(6):672–679.
3. Rizzo R, Dziadosz M, Kyathanahally SP, et al. 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. Magnetic Resonance in Medicine 2024; 89(5):1707–1727.
4. Ma Z, Karakus O, Schermer SM, Langbein FC. Metabolite quantification from edited magnetic resonance spectra with deep learning. Preprint 2024. https://mrs.qyber.dev/paper-mrsnet-autoencoder.
5. Shermer SM, Jenkins C, Chandler M, Langbein FC. Magnetic resonance spectroscopy data for GABA quantification using MEGAPRESS pulse sequence. IEEE Data Port 2019; DOI: 10.21227/ak1d-3s20.
6. Near J, Simpson R, Jezzard P, et al. FID-A - Advanced MR spectroscopy processing and simulation. Github, Version 1.2 Sep 25 2018. https://github.com/CIC-methods/FID-A/tree/V1.2.
7. Michaelis T, Merboldt KD, Hänicke W, Gyngell ML, Bruhn H, Frahm J. Identification of cerebral metabolites in localized 1H NMR spectra of human brain in vivo. NMR in Biomedicine 1991; 4(1):90–98.
8. Kaiser L, Young K, Meyerhoff D, Mueller S, Matson G. Analysis of localized J-difference GABA editing: Theoretical and experimental study at 4 T. NMR in Biomedicine 2008; 21(1):22–32.
9. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. NMR in Biomedicine 2000; 13(1):129–153.
10. Near J, Leung I, Claridge T, Cowen P, Jezzard P. Chemical shifts and coupling constants of the GABA spin system. 20th International Society for Magnetic Resonance in Medicine (ISMRM) Annual Meeting & Exhibition, Melbourne, Australia 2012; 4386.
11. Karakus O, Kuruoğlu EE, Altınkaya MA. Beyond trans-dimensional RJMCMC with a case study in impulsive data modeling. Signal Processing 2018; 153:396–410.

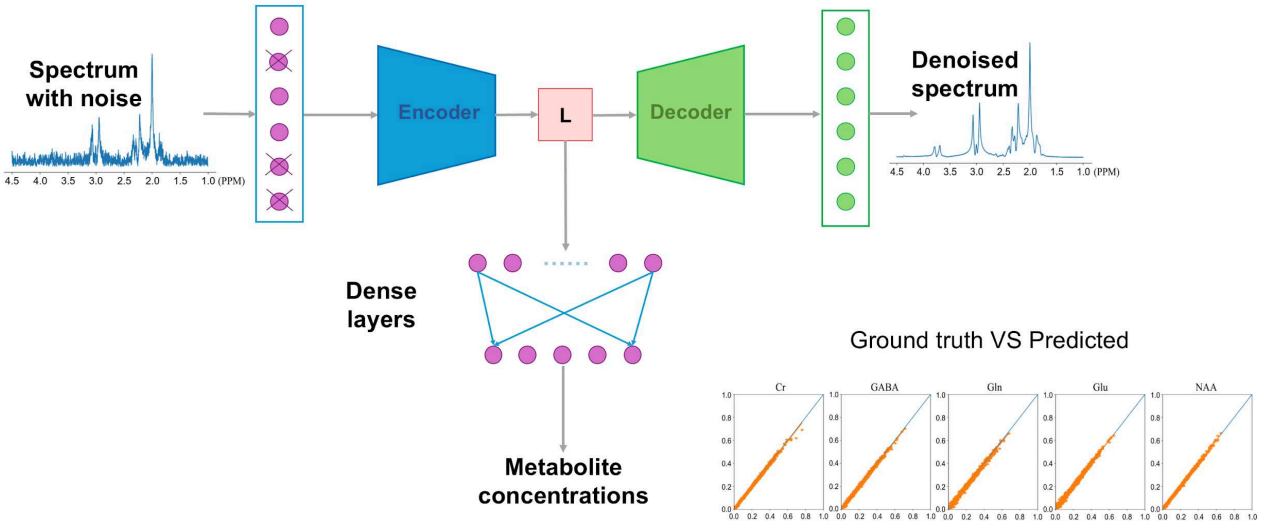

Figure 1: Y-shaped full-connected autoencoder structure to quantify and denoise MR spectra. An encoder embeds the input spectra into a latent space ***L*** with a decoder constructing a denoised spectrum and a dense network predicting metabolite concentrations.

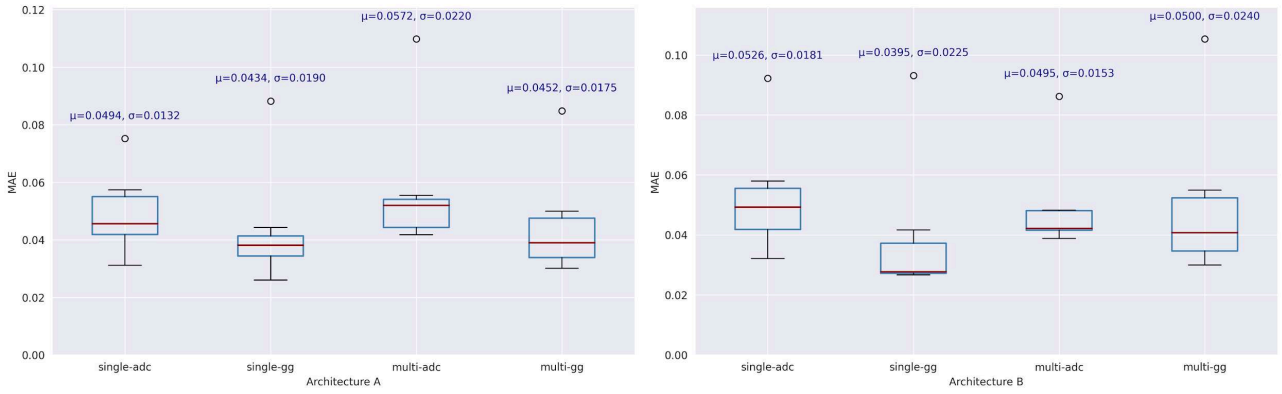

Figure 2: Overall accuracy performance on experimental phantom spectra of two architecture variants A and B of the Y-shaped fully-connected autoencoder trained on multiple simulated datasets (single-adc, single-gg, multi-adc, multi-gg). Variant A has been optimised for the single-adc and variant B for the multi-gg simulated dataset. We show the MAE distributions with mean ***μ*** and standard deviation ***σ***.

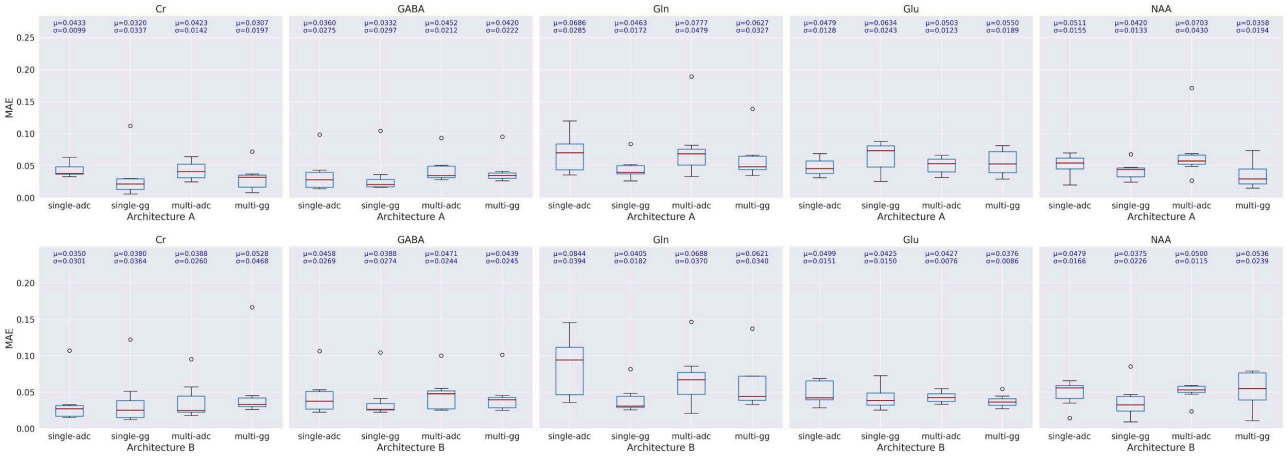

Figure 3: Per metabolite accuracy performance of two architecture variants A and B of the Y-shaped fully-connected autoencoder trained on multiple simulated datasets (single-adc, single-gg, multi-adc, multi-gg). Variant A has been optimised for the single-adc and variant B for the multi-gg simulated dataset. We show the MAE distributions with mean μ and standard deviation σ for each metabolite (Cr, GABA, Gln, Glu, NAA).

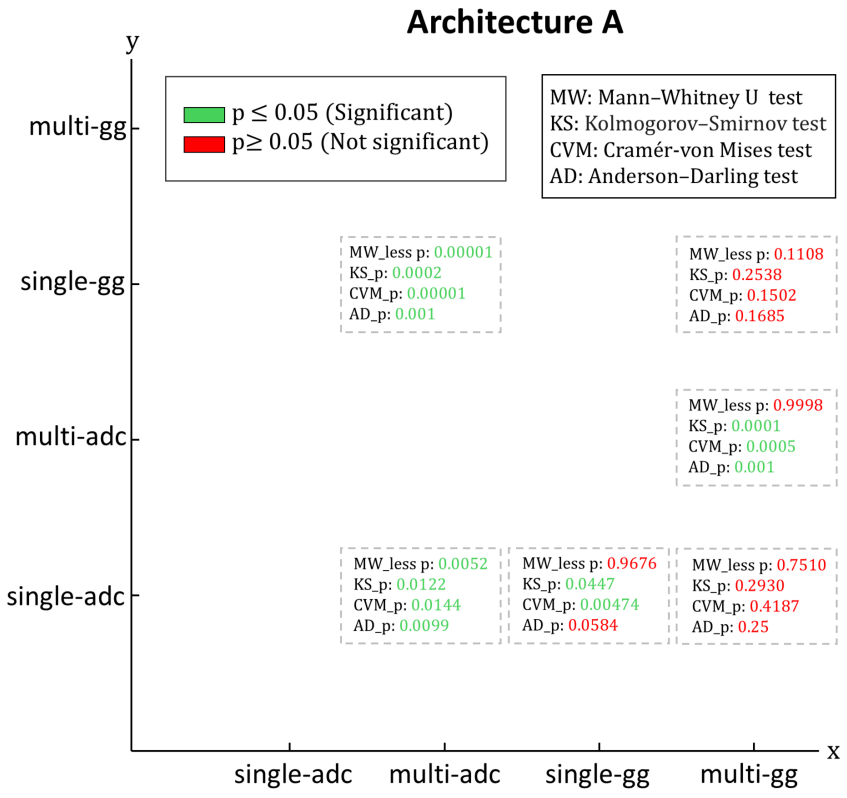

Figure 4: Non-parametric statistical significance tests (y-axis smaller/different from x-axis distribution) for the difference in the MAE distributions for architecture variant A trained with different data. MW (Mann-Whitney U) assesses if one group tends to have lower values than another, KS (Kolmogorov Smirnov) detects overall distributional differences, CVM (Cramér-von Mises) evaluates cumulative distribution discrepancies, AD (Anderson-Darling) focuses on tail differences between distributions.

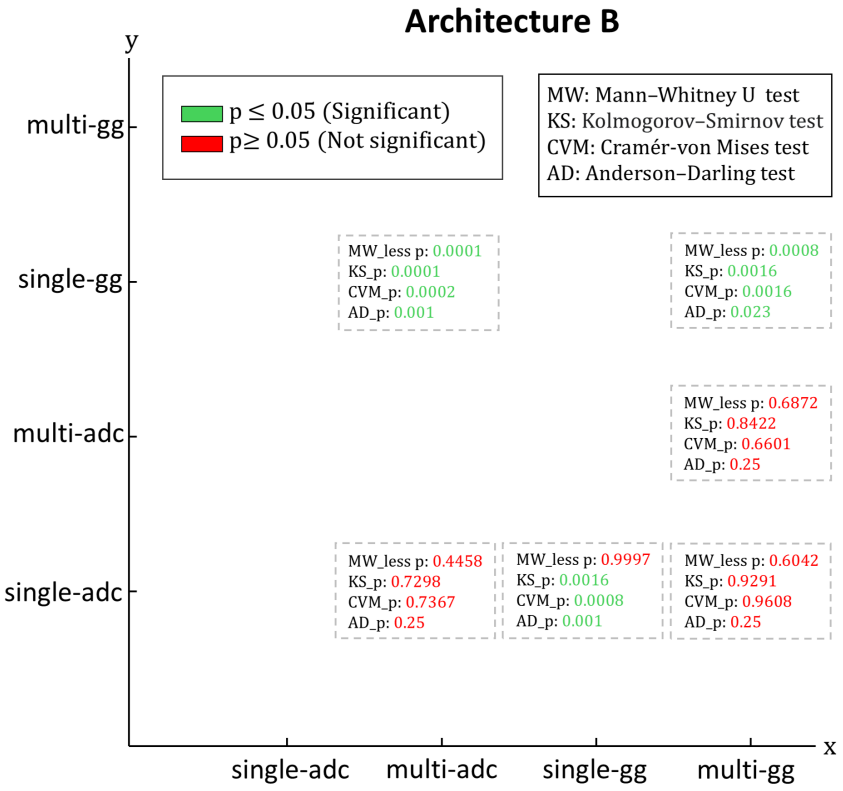

Figure 5: Non-parametric statistical significance tests (y-axis smaller/different from x-axis distribution) for the difference in the MAE distributions for architecture variant B trained with different data. MW (Mann-Whitney U) assesses if one group tends to have lower values than another, KS (Kolmogorov Smirnov) detects overall distributional differences, CVM (Cramér-von Mises) evaluates cumulative distribution discrepancies, AD (Anderson-Darling) focuses on tail differences between distributions.