Received xx xxxxx xxxx; | Revised xx xxxxx xxxx; | Accepted xx xxxxx xxxx
DOI: xxx/xxxx

PREPRINT - SUBMITTED TO MAGNETIC RESONANCE IN MEDICINE

Strategies to Minimize Out-of-Distribution Effects in Data-Driven MRS Quantification

Julian P. Merkofer[1] | Antonia Kaiser[2] | Anouk Schrantee[3] | Oliver J. Gurney-Champion[3,*] | Ruud J. G. van Sloun[1,*]

[1]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[2]CIBM Center for Biomedical Imaging, École Polytechnique Fédérale de Lausanne, EPFL, Lausanne, Switzerland
[3]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Center, Amsterdam, The Netherlands
[*]These authors share last authorship.

Correspondence
Julian P. Merkofer, Eindhoven University of Technology, PO Box 513 5600 MB Eindhoven, The Netherlands.
Email: j.p.merkofer@tue.nl

Present Address
Eindhoven University of Technology
Electrical Engineering Department
Groene Loper 19, 5612 AP Eindhoven, The Netherlands

Funding Information
This work was in part funded by Spectralligence (EUREKA IA Call, ITEA4 project 20209) and received support from the NVIDIA Academic Hardware Grant Program.

Summary

Purpose: This study systematically compared data-driven and model-based strategies for metabolite quantification in magnetic resonance spectroscopy (MRS), focusing on resilience to out-of-distribution (OoD) effects and the balance between accuracy, robustness, and generalizability.
Methods: A neural network designed for MRS quantification was trained using three distinct strategies: supervised regression, self-supervised learning, and test-time adaptation. These were compared against model-based fitting tools. Experiments combined large-scale simulated data, designed to probe metabolite concentration extrapolation and signal variability, with 1H single-voxel 7T in-vivo human brain spectra.
Results: In simulations, supervised learning achieved high accuracy for spectra similar to those in the training distribution, but showed marked degradation when extrapolated beyond the training distribution. Test-time adaptation proved more resilient to OoD effects, while self-supervised learning achieved intermediate performance. In-vivo experiments showed larger variance across the methods (data-driven and model-based) due to domain shift. Across all strategies, overlapping metabolites and baseline variability remained persistent challenges.
Conclusion: While strong performance can be achieved by data-driven methods for MRS metabolite quantification, their reliability is contingent on careful consideration of the training distribution and potential OoD effects. When such conditions in the target distribution cannot be anticipated, test-time adaptation strategies can ensure consistency between the quantification, the data, and the model, enabling reliable data-driven MRS pipelines.

KEYWORDS:
machine learning, magnetic resonance spectroscopy, metabolite quantification, generalization, test-time adaptation, out-of-distribution effects

1 | INTRODUCTION

Magnetic resonance spectroscopy (MRS) is a non-invasive technique for measuring the metabolic composition of tissues, providing valuable insights into neurological disorders and serving as a tool to characterise tumours for treatment stratification and monitoring.[1,2,3] However, its clinical utility is constrained by several challenges, including an inherently low signal-to-noise ratio (SNR), spectral overlap of metabolites, unparameterized baseline effects, and various experimental artifacts.[4,5] Together, these factors make accurate quantification of metabolite concentrations a challenging task, complicating the analysis and interpretation of MRS data.

Traditionally, metabolite quantification in MRS has relied on model-based fitting approaches such as linear combination modeling (LCM) and peak fitting.[6] These methods construct a theoretical representation of the expected signal and optimize its parameters to match the measured spectrum, typically using nonlinear least-squares algorithms like the Levenberg–Marquardt method[7,8]. Frequency-domain LCM is widely adopted[9,10,11,12,13], fitting a linear combination of basis spectra (the idealized signal contributions of individual metabolites) to the data while accounting for baseline distortions, frequency and phase shifts, and lineshape variations[14]. These purely model-based methods are generally computationally intensive, may require user expertise for proper setup and interpretation, and the inherently ill-posed nature yields non-unique solutions, requiring methods to employ specific regularizers, constraints, or other priors.[6,15,16,17,9] To overcome these challenges, increasing attention has been given to machine learning (ML) methods as their ability to learn from data offer a promising alternative.[18,19]

Early data-driven approaches to MRS quantification explored direct regression from spectra to metabolite concentrations using supervised learning techniques, including random forests[20] and convolutional neural networks (CNNs)[21,22,23,24,25]. These models are trained to predict metabolite amplitudes directly from the input spectrum by minimizing a loss function between predicted and ground truth concentrations. This requires access to reference concentrations during training, which are typically only available for synthetic data. More recently, self-supervised strategies have emerged that integrate a physics-based signal model into the training process[26,27,28]. In these methods, the neural network (NN) estimates all relevant signal parameters, then reconstructs the spectrum using a forward signal model as done in traditional LCM. The distinction lies in the optimization strategy: instead of explicitly solving a least-squares problem, a network learns to reduce the reconstruction error via stochastic gradient descent amortized over the training dataset. This brings several advantages including the use of learned priors as a regularizer that lowers prediction variance, reduces sensitivity to noise, and helps the optimizer avoid suboptimal local minima.

Nevertheless, the effectiveness of ML models depends strongly on the quality and diversity of the training data.[29,30,31] Most of the previous work has relied heavily on simulated data with minimal in-vivo testing.[18] While investigations have explored diverse NN architectures, spectroscopic input types, the use of ensemble learning[32], and methods for uncertainty estimation[24,33,34], a systematic analysis of the critical aspects of robustness and generalization is lacking. In particular, the influence of training paradigms on a model's susceptibility to bias, its ability to maintain performance under challenging conditions, and its capacity to generalize to unseen, potentially out-of-distribution (OoD), in-vivo MRS data has not been thoroughly studied and documented. Furthermore, test-time adaptation (TTA)[35,36,37,38,39,40], where NNs are updated during inference, has received little attention in the context of MRS, despite its potential to mitigate domain shift.

This study contributes to this ongoing effort by systematically comparing different data-driven strategies for MRS metabolite quantification, explicitly focusing on their inherent data biases and their resilience to OoD samples. We evaluate a supervised regression approach, a self-supervised learning method that incorporates a signal model during training, and TTA techniques, as well as compare these to purely model-based fitting. By assessing the performance of these strategies on both carefully controlled synthetic data and 7T in-vivo human brain proton spectra, we aim to provide valuable insights into the trade-offs between accuracy, robustness, and generalizability of different strategies for MRS quantification.

2 | METHODS

This section outlines the simulation framework, in-vivo data acquisition and processing, quantification strategies, evaluation metrics, and performed experiments.

2.1 | Simulated Data

Simulated spectra offer access to ground truth metabolite concentrations and acquisition parameters, providing a controlled setting for both model optimization and analysis.

2.1.1 | Signal Model

To simulate proton MRS spectra, we define a parametric signal model in the frequency domain, denoted by $X(f \mid \boldsymbol{\theta})$, where f is frequency and $\boldsymbol{\theta}$ represents the set of signal model parameters. The modeled spectrum is expressed as:

$$X(f \mid \boldsymbol{\theta}) = e^{i(\phi_0 + f\phi_1)} \sum_{m=1}^{M} a_m \; S_m(f) + B(f), \qquad (1)$$

where ϕ_0 and ϕ_1 are zeroth- and first-order phases, a_m is the amplitude of the m-th metabolite, and $B(f)$ denotes a spectral baseline, modeled as a complex-valued K-order polynomial. Each metabolite basis function is defined as:

$$S_m(f) = \mathcal{F}\{s_m(t) \; e^{-(\gamma + \varsigma^2 t + i\epsilon)t}\}, \qquad (2)$$

where γ and ς denote global (for all metabolites) Lorentzian and Gaussian linewidth broadening parameters, respectively, and ϵ represents a global frequency shift. The operator $\mathcal{F}\{\cdot\}$ denotes the Fourier transform, and $s_m(t)_{m=1}^{M}$ are the time-domain basis functions representing the idealized signal contributions of individual metabolites. Macromolecules (MMs) are included, experiencing the same broadening, phasing, and shifting as the metabolites. To simulate realistic measurement conditions, the observed spectrum $Y(f)$ is defined as:

$$Y(f) = X(f \mid \boldsymbol{\theta}) + N, \qquad (3)$$

where N denotes complex Gaussian noise.

2.1.2 | Parameter Ranges

The simulation ranges for the signal model parameters

$$\boldsymbol{\theta} = \left\{a_1, ..., a_M, \gamma, \varsigma, \epsilon, \phi_0, \phi_1, b_1, ..., b_{2(K+1)}\right\}, \qquad (4)$$

were designed to capture the full variability observed in the in-vivo data. Therefore, metabolite concentration bounds were derived from a combination of literature values reported by De Graaf[41], and the empirical distributions obtained by fitting all in-vivo spectra using both LCModel[9] and FSL-MRS[13] (details are reported in Section 2.2). For each metabolite, the lower and upper bounds were defined as the minimum and maximum observed values across these three sources. This ensured that the entire dynamic range of concentrations present in our in-vivo dataset was represented, while avoiding unrealistically narrow ranges.

For the remaining signal parameters γ, ς, ϵ, ϕ_0, ϕ_1, b_1, ..., $b_{2(K+1)}$, and N, literature guidance was limited, so we used our own in-vivo data as the initial reference. Since this 7T in-vivo data, acquired with high SNR, good shimming, and consistent processing, resulted in relatively narrow parameter distributions, we deliberately selected wider simulation ranges. This ensured the simulated data reflected the greater variability that may be encountered in broader clinical or research settings.

An overview of all simulation parameter distributions is provided in Table 1. The individual metabolite range reported in De Graaf 2019[41] and the obtained ranges of LCModel and FSL-MRS are listed in Tables B9, B10, and B11 in Appendix B.

2.1.3 | Training Data Generation

Synthetic examples were generated *ad-hoc* at training time, with each batch composed of newly sampled signals. This setup allowed for virtually unlimited data during training and reduced the risk of overfitting to a discrete synthetic distribution. The signal parameters $\boldsymbol{\theta}$ were drawn independently from their respective distributions and passed through Equation (3) to generate spectra. Training was done with a batch size of 16 and every 256 batches, a set of 1024 new samples was used for validation (resulting in a 20/80% validation/training split).

2.2 | In-Vivo Data

Data was obtained from 61 healthy volunteers as part of the BrainBeats study. The study adhered to the guidelines of the Institutional Review Board of the University of Amsterdam (the Netherlands). All participants provided written informed consent. Four subjects were excluded following visual inspection due to insufficient data quality, resulting in a final cohort of 57 participants.

2.2.1 | Acquisition

In-vivo single-voxel MRS data were acquired in the anterior cingulate cortex using a semi-LASER sequence with TE/TR = 36/5000 ms on a 7T Philips scanner, as part of an interleaved fMRI/MRS protocol[42]. Acquisition parameters included a volume-of-interest of 25 x 18 x 18 mm^3, 1024 sample points, and a spectral bandwidth of 3000 Hz. Water suppression was performed using VAPOR and shimming was optimized using HOSDLT[43]. For each subject, spectra were obtained across three scan sessions. A total of 64 signal averages were used from the first session, 3×64 from the second, and 2×64 from the third, resulting in 342 spectra across all subjects. Further acquisition details are provided in Appendix C, Table C14, as part of the minimum reporting standards for in-vivo MRS (MRSinMRS)[44].

TABLE 1 Overview of the notations and distribution ranges of the simulation parameters. Metabolite bounds were set using the minimum and maximum values from De Graaf[41] and fits to all in-vivo spectra using LCModel[9] and FSL-MRS[13]. $\mathcal{U}[p_{min}, p_{max}]$ and $\mathcal{CN}(p_{mean}, p_{var})$ denote continuous and complex Gaussian distributions, respectively, and curly braces $\{\cdot\}$ indicate discrete sets of values.

Parameter	Notation	Range	Unit	Parameter	Notation	Range	Unit
Alanine (Ala)	a_1	$\mathcal{U}[0.0, 1.6]$	mM	Number of Metabolites (+MMs)	M	$\{21\}$	–
Ascorbate (Asc)	a_2	$\mathcal{U}[0.0, 4.9]$	mM	Frequency Shifts	ϵ	$\mathcal{U}[-10, 10]$	$\mathrm{rad\,s^{-1}}$
Aspartate (Asp)	a_3	$\mathcal{U}[0.0, 4.8]$	mM	Lorentzian Broadening	γ	$\mathcal{U}[2, 25]$	$\mathrm{s^{-1}}$
Creatine (Cr)	a_4	$\mathcal{U}[3.9, 12.3]$	mM	Gaussian Broadening	ς	$\mathcal{U}[2, 25]$	$\mathrm{s^{-1}}$
Gamma-Aminobutyric Acid (GABA)	a_5	$\mathcal{U}[0.0, 4.0]$	mM	Zeroth-Order Phase	ϕ_0	$\mathcal{U}[-0.5, 0.5]$	rad
Glutamine (Gln)	a_6	$\mathcal{U}[0.0, 6.8]$	mM	First-Order Phase	ϕ_1	$\mathcal{U}[-10^{-5}, 10^{-5}]$	$\mathrm{rad\,Hz^{-1}}$
Glutamate (Glu)	a_7	$\mathcal{U}[6.0, 17.9]$	mM	Baseline Polynomial Order	K	$\{2\}$	–
Glycine (Gly)	a_8	$\mathcal{U}[0.0, 1.0]$	mM	Baseline Coefficients	b_1	$\mathcal{U}[-600, 200]$	a.u.
Glycerophosphocholine (GPC)	a_9	$\mathcal{U}[0.0, 3.6]$	mM		b_2	$\mathcal{U}[-800, 300]$	a.u.
Glutathione (GSH)	a_{10}	$\mathcal{U}[0.0, 3.6]$	mM		b_3	$\mathcal{U}[-1000, 600]$	a.u.
Myo-Inositol (mIns)	a_{11}	$\mathcal{U}[4.0, 12.1]$	mM		b_4	$\mathcal{U}[-600, 1000]$	a.u.
Lactate (Lac)	a_{12}	$\mathcal{U}[0.0, 3.1]$	mM		b_5	$\mathcal{U}[-1600, 200]$	a.u.
N-Acetylaspartylglutamate (NAAG)	a_{13}	$\mathcal{U}[0.0, 2.5]$	mM		b_6	$\mathcal{U}[-400, 1000]$	a.u.
N-Acetylaspartate (NAA)	a_{14}	$\mathcal{U}[7.5, 16.3]$	mM	Random Walk Step Size	–	$\mathcal{U}[0, 10^5]$	a.u.
Phosphocholine (PCh)	a_{15}	$\mathcal{U}[0.0, 2.4]$	mM	Random Walk Smoothing	–	$\mathcal{U}[1, 10^5]$	a.u.
Phosphocreatine (PCr)	a_{16}	$\mathcal{U}[0.0, 5.5]$	mM	Random Walk Min. Bound	–	$\mathcal{U}[-10^6, 0]$	a.u.
Phosphoethanolamine (PE)	a_{17}	$\mathcal{U}[0.0, 5.2]$	mM	Random Walk Max. Bound	–	$\mathcal{U}[0, 10^6]$	a.u.
Scyllo-Inositol (Scyllo)	a_{18}	$\mathcal{U}[0.0, 0.6]$	mM	Complex Gaussian Noise*	N	$\mathcal{CN}(0, \sigma^2)$	a.u.
Serine (Ser)	a_{19}	$\mathcal{U}[0.0, 7.3]$	mM	Noise Variance	σ^2	$\mathcal{U}[10, \sqrt{2} \cdot 5000]$	a.u.
Taurine (Tau)	a_{20}	$\mathcal{U}[1.2, 6.0]$	mM	Chemical Shift Limits	f_{min}	$\{0.5\}$	ppm
Macromolecules (MMs)	a_{21}	$\mathcal{U}[0.0, 400.0]$	mM		f_{max}	$\{4.0\}$	ppm

*SNR ranging from 0 - 40 dB as computed from the ground truth noise and metabolite-only signals over 0.5 to 4.0 ppm.

2.2.2 | Processing

Processing of the in-vivo MRS data involved several steps. Coil combination was performed using a custom method that estimated phase correction and amplitude weighting parameters for each coil from an unsuppressed water reference scan acquired at the beginning of the time series. All subsequent processing was performed using FSL-MRS. Individual transients were frequency- and phase-aligned within the 0.2–4.2 ppm range and, subsequently, averaged. Eddy current correction was then applied using the unsuppressed water reference, followed by removal of nuisance peaks using Hankel singular value decomposition (HSVD). Finally, the spectra were frequency- and phase-aligned to creatine (Cr) at 3.027 ppm.

2.2.3 | Analysis

Metabolite quantification of the processed in-vivo spectra was performed using both LCModel and FSL-MRS, using a basis set matched to the acquisition parameters (7T field strength, semi-LASER sequence, TE = 34 ms, 1024 points, 3000 Hz bandwidth), see Appendix C, Table C14 for details on fit settings. The same basis set was employed for both synthetic spectrum generation and quantification to maintain consistency across training and evaluation. It consisted of 20 metabolites together with a single macromolecular baseline; detailed specifications are provided in Table 1.

2.3 | Quantification Strategies

Quantification in MRS aims to estimate metabolite concentrations ${a_m}_{m=1}^{M}$ from an observed spectrum. Standard tools such as LCModel fit a parametric signal model using the Levenberg–Marquardt algorithm, which combines gradient descent with Gauss-Newton updates[45]. We explicitly linked data-driven learning with traditional fitting by comparing strategies that all relied on gradient-based updates. We included a purely model-based baseline that replaced Levenberg-Marquardt with direct gradient descent, alongside NN-based approaches. The strategies, summarized in Figure 1, provided a common framework to investigate how data-driven quantification could mitigate OoD effects and explore the bias–variance trade-offs between learned priors and adaptive fitting.

Let $\mathbf{y}, \mathbf{x}(\boldsymbol{\theta}) \in \mathbb{C}^L$ denote the observed and modeled spectra sampled at L frequency points in the range $f_{\min} \leq f \leq f_{\max}$, corresponding to 0.5-4.0 ppm throughout this work. Unless stated otherwise, all methods were optimized using Adam[46] with a learning rate of 1×10^{-4}.

FIGURE 1 The schematic provides an overview of the relevant parameter estimation methods. From top to bottom: purely model-based fitting using LCM, supervised regression trained to predict metabolite amplitudes directly, self-supervised training that utilizes the underlying signal model to map estimated parameters to their corresponding spectrum, and TTA to refine predictions via gradient-based optimization at inference. The NN employed for the latter three is a simple multi-layer perceptron (MLP) with three fully connected layers, which takes the normalized real and imaginary parts of the frequency domain spectra as input and outputs the estimated parameters of the chosen signal model.

To ensure consistency across signal intensities, all methods applied normalization and scaling:

$$\mathbf{y} \leftarrow \mathbf{y} \;/\; \|\mathbf{y}\|_2. \tag{5}$$

The norm was then propagated forward to scale the predicted metabolite amplitudes and baseline parameters:

$$\hat{a}_m \leftarrow \hat{a}_m \cdot \|\mathbf{y}\|, \quad \hat{b}_k \leftarrow \hat{b}_k \cdot \|\mathbf{y}\|. \tag{6}$$

This procedure allowed the models to operate on normalized inputs while preserving the effective signal amplitude in the outputs.

2.3.1 | Purely Model-Based Fitting

Our purely model-based approach directly optimized the signal model parameters $\boldsymbol{\theta}$ using gradient descent (Adam with a learning rate of 1×10^{-1}, for 1000 epochs). No NN was involved; instead, the parameters were treated as learnable tensors and refined iteratively to minimize the residual between the modeled ($\mathbf{x}(\boldsymbol{\theta})$) and observed spectra ($\mathbf{y}$):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{x}(\boldsymbol{\theta})\|_2^2. \tag{7}$$

To enforce constraints, such as positivity for metabolite amplitudes and lineshape parameters, we applied the same activation functions used in the learning-based strategies (see Section 2.3.5 for details).

2.3.2 | Supervised Regression

In the supervised setting, a NN $g_{\boldsymbol{\psi}}$ with trainable weights $\boldsymbol{\psi}$ maps each input spectrum $\mathbf{y}$ to signal model parameters $\hat{\boldsymbol{\theta}} = g_{\boldsymbol{\psi}}(\mathbf{y})$. The training objective minimizes the mean absolute error (MAE) between predicted and reference parameters, amortized over the training dataset $\mathcal{D}_{train}$:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \frac{1}{|\mathcal{D}_{train}|} \sum_{(\mathbf{y},\boldsymbol{\theta}) \in \mathcal{D}_{train}} \mathcal{L}_{\mathrm{MAE}^*}(\boldsymbol{\theta}, g_{\boldsymbol{\psi}}(\mathbf{y})). \tag{8}$$

The scaled MAE loss is defined as

$$\mathcal{L}_{\mathrm{MAE}^*}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \left| \frac{\boldsymbol{\theta} - p_{\min}}{p_{\max} - p_{\min}} - \frac{\hat{\boldsymbol{\theta}} - p_{\min}}{p_{\max} - p_{\min}} \right|, \tag{9}$$

where $p_{\min}$ and $p_{\max}$ denote the lower and upper bounds of each parameter, derived from the simulation priors. This scaling ensures balanced optimization across all

components of $\boldsymbol{\theta}$. Although strictly only the metabolite concentrations need to be estimated, we chose to predict all signal parameters to allow direct comparison with alternative methods and to ensure that the resulting fits are structurally analogous. After optimization, the trained network $g_{\boldsymbol{\psi}^*}$ is fixed and used for rapid prediction (inference) on unseen test data.

2.3.3 | Self-Supervised Regression

The self-supervised strategy integrates the physics-based signal model directly into the training loop. Unlike supervised learning, this method trains the network $g_{\boldsymbol{\psi}}$ without requiring reference parameters $\boldsymbol{\theta}$. Instead, the optimization objective minimizes the reconstruction error between the modeled spectrum $\mathbf{x}(\hat{\boldsymbol{\theta}}) = \mathbf{x}(g_{\boldsymbol{\psi}}(\mathbf{y}))$ and the observed spectrum $\mathbf{y}$:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \frac{1}{|\mathcal{D}_{train}|} \sum_{\mathbf{y}\in\mathcal{D}_{train}} \|\mathbf{y} - \mathbf{x}(g_{\boldsymbol{\psi}}(\mathbf{y}))\|_2^2. \quad (10)$$

The key distinction from purely model-based least-squares fitting is the optimization target: minimizing the loss by updating the network weights $\boldsymbol{\psi}$ enables the network to learn priors from the dataset $\mathcal{D}_{train}$. Furthermore, the trained network $g_{\boldsymbol{\psi}^*}$ can instantly output parameter predictions during inference, rather than requiring iterative parameter fitting for each new spectrum.

2.3.4 | Test-Time Adaptation

TTA refers to refining model predictions during inference, allowing the network to adjust dynamically to previously unseen data $\mathcal{D}_{test}$. This is done by fine-tuning the pretrained network $g_{\boldsymbol{\psi}^*}$ using least-squares. Unless stated otherwise, all TTA procedures in this work initialize the network from the supervised pretrained model, with alternative initializations reported in Appendix A.

Test-Time Instance Adaptation: Adapts the pretrained network $g_{\boldsymbol{\psi}^*}$ to a single spectrum $\mathbf{y} \in \mathcal{D}_{test}$. The approach is particularly relevant for clinical deployment, where predictions must be robust to single-subject variability. It fine-tunes the network weights $\boldsymbol{\psi}$ for $j \in \{1, ..., J = 50\}$ steps using least-squares:

$$\boldsymbol{\psi}^{(j+1)} = \arg\min_{\boldsymbol{\psi}^{(j)}} \|\mathbf{y} - \mathbf{x}(g_{\boldsymbol{\psi}^{(j)}}(\mathbf{y}))\|_2^2. \quad (11)$$

After adaptation, the network $g_{\boldsymbol{\psi}^{(J)}}$ produces the parameter prediction $\hat{\boldsymbol{\theta}} = g_{\boldsymbol{\psi}^{(J)}}(\mathbf{y})$. This procedure is repeated independently for each spectrum, allowing the network to refine predictions in response to distribution shifts while retaining priors learned from the training dataset.

Test-Time Online Adaptation: Updates the network $g_{\boldsymbol{\psi}^*}$ continuously as new batches $\mathcal{B}_i \subset \mathcal{D}_{test}$ arrive. This strategy allows the model to adjust continually to evolving data characteristics, making it suitable for streaming or high-throughput acquisition settings. For each batch, the network weights $\boldsymbol{\psi}$ are adapted,

$$\boldsymbol{\psi}_{i+1} = \arg\min_{\boldsymbol{\psi}_i} \frac{1}{|\mathcal{B}_i|} \sum_{\mathbf{y}\in\mathcal{B}_i} \|\mathbf{y} - \mathbf{x}(g_{\boldsymbol{\psi}_i}(\mathbf{y}))\|_2^2, \quad (12)$$

so that the updated model $g_{\boldsymbol{\psi}_{i+1}}$ and the corresponding predictions $\hat{\boldsymbol{\theta}}_i = g_{\boldsymbol{\psi}}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{B}_i$ are obtained as part of the same update process. The adapted weights $\boldsymbol{\psi}_{i+1}$ then initialize the model for the next incoming batch $\mathcal{B}_{i+1}$ (default batch size $|\mathcal{B}_i| = 16$).

Test-Time Domain Adaptation: Refines the pretrained network $g_{\boldsymbol{\psi}^*}$ using the entire test dataset $\mathcal{D}_{test}$ to account for distribution shifts. Domain adaptation is particularly beneficial in research settings where the full test dataset is available before inference, but could also be used to recalibrate a network to a new institute/scanner. The network weights are adapted by minimizing:

$$\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} \frac{1}{|\mathcal{D}_{test}|} \sum_{\mathbf{y}\in\mathcal{D}_{test}} \|\mathbf{y} - \mathbf{x}(g_{\boldsymbol{\psi}_i}(\mathbf{y}))\|_2^2. \quad (13)$$

The adapted network $g_{\boldsymbol{\psi}^*}$ is then used to produce predictions $\hat{\boldsymbol{\theta}} = g_{\boldsymbol{\psi}^*}(\mathbf{y})$ for all spectra in $\mathcal{D}_{test}$. A mini batch size of 16 is used and optimization is run for 1000 epochs.

2.3.5 | Neural Network Architecture

All data-driven strategies used the same NN architecture, illustrated in Figure 1: a three-layer MLP with ELU[47] activations. Outputs were constrained via activation functions to enforce physical plausibility: metabolite amplitudes are passed through a softplus function to ensure non-negativity, linewidths were softplus-transformed and shifted by 1, frequency offsets and other linear parameters remained unconstrained, and the first-order phase was restricted to a narrow range using a scaled hyperbolic tangent activation.

The choice of an MLP over other architectures such as CNNs was intentional: our analysis focused on the impact of training strategy rather than architectural design. By using a general-purpose function approximator, we maximized the NN's flexibility while minimizing architecture-specific biases. The MLP was optimized using sweep runs for hyperparameter tuning (depth, width, activation, etc.). Exact architecture implementation details are provided in Appendix B, Table B12 along with results obtained with a CNN in Appendix A.

2.4 | Evaluation

The prediction accuracy was assessed using both absolute and relative error metrics. The MAE was computed directly between the predicted amplitudes $\hat{a}_m$ and the true concentrations a_m:

$$\mathrm{MAE} = \frac{1}{M-1} \sum_{m=1}^{M-1} |\hat{a}_m - a_m| \,, \tag{14}$$

where $M-1$ is the number of metabolites (excluding MMs).

To enable fair comparison of relative estimates, we calculated an optimal scaling factor w_{opt} that minimized the absolute error between the scaled relative metabolite concentration estimates $\hat{a}_m$ and absolute ground truth values a_m:

$$w_{\mathrm{opt}} = \arg\min_{w} \sum_{m=1}^{M-1} |w\ \hat{a}_m - a_m| \,. \tag{15}$$

Using this scaling, the mean optimally scaled absolute error (MOSAE) is defined as

$$\mathrm{MOSAE} = \frac{1}{M} \sum_{m=1}^{M-1} |w_{\mathrm{opt}}\ \hat{a}_m - a_m| \,. \tag{16}$$

This metric allowed a fair comparison of concentration estimates independent of a water reference or other metabolite references such as total creatine (tCr).

We further assessed the agreement between predicted and true concentrations via linear regression

$$a_m = \alpha\ \hat{a}_m + \beta, \tag{17}$$

yielding four interpretable metrics: slope α (proportional bias), intercept β (constant bias), coefficient of determination R^2 (explained variance), and root mean squared error (RMSE)

$$\sigma = \sqrt{\frac{1}{|\mathcal{D}_{test}|} \sum_{u=1}^{|\mathcal{D}_{test}|} \left((\hat{a}_m)_u - (a_m)_u\right)^2}. \tag{18}$$

2.5 | Experiments

The primary purpose of the experiments is to investigate the inherent data biases and resilience of these methods to OoD effects and domain shift, ultimately assessing trade-offs between accuracy, robustness, and generalizability.

2.5.1 | Controlled Simulation Experiments

For each test scenario, 10,000 synthetic spectra were generated, focusing on two main categories of perturbations.

Metabolite Concentration Effects:

- ID (Mid-Range): Training and testing on the central 50% of the concentration range (baseline).
- OoD (Full-Range): Training on mid-range, testing across the full range (extrapolation stress test).
- ID (Full-Trained): Training and testing on the full range (reference).

Signal Parameter Perturbations: Test scenarios with wider ranges of SNR, linewidth, frequency/phase shifts, and baseline variations than seen in training.

2.5.2 | In-Vivo Data Testing

From the 342 original 7T acquisitions with 64 transients each, we produced 1,710 spectra by forming subsets of 4, 8, 16, and 32 transients to control SNR. Since true in-vivo concentrations are unknown, FSL-MRS fits from the full 64 averages served as pseudo ground truth for signal parameter estimates and error estimation. LCModel and mixed references are described in Appendix A. The in-vivo spectra were then filtered to match the simulation conditions: ID (Mid Range), OoD (Full Range), and ID (Full Trained). This setup enabled a systematic study of two factors: the domain shift from synthetic to real data, and the impact of training on narrow versus broad synthetic ranges when applied to in-vivo spectra.

3 | RESULTS

3.1 | Results on Simulated Data

3.1.1 | Overall Quantification Performance

The performance, measured by MOSAE across three metabolite concentration scenarios, diverged significantly between methods reliant on learned priors and adaptive strategies (Table 2). In the ideal, restricted-range in-distribution (ID) setting, supervised regression achieved the lowest errors (0.2764 ± 0.0032). However, when restricted-range models were evaluated on the full concentration range (OoD extrapolation), supervised regression showed the strongest decrease in accuracy, with errors nearly doubling (0.5537 ± 0.0058). Domain adaptation also improved over baseline regression. Training on the full concentration range removed the extrapolation gap, but all data-driven methods exhibited increased errors in this broader setting (e.g. supervised: 0.3896 ± 0.0047).

Classical fitting approaches (purely model-based gradient descent, FSL-MRS, LCModel) yielded similar error

TABLE 2 Comparison of quantification methods across test scenarios of 10,000 spectra. **ID (Mid-Range)**: Models trained and tested on the central 50% metabolite concentration range. **OoD (Full-Range)**: Models trained on the central 50% concentration range, but tested across the entire range of concentrations to assess extrapolation. **ID (Full-Trained)**: Models trained and tested on the full concentration range. Lowest error values are highlighted in **bold**.

Method	MOSAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	**0.2764 (± 0.0032)**	0.5537 (± 0.0058)	**0.3896 (± 0.0047)**	**0.1565**
Self-Supervised	0.3734 (± 0.0044)	0.5401 (± 0.0059)	0.4548 (± 0.0059)	0.1574
Test-Time Instance Adaptive	0.3767 (± 0.0049)	**0.4420 (± 0.0052)**	0.4378 (± 0.0060)	30.4637
Test-Time Online Adaptive	0.2788 (± 0.0032)	0.5520 (± 0.0058)	0.3923 (± 0.0048)	0.2454
Test-Time Domain Adaptive	0.3621 (± 0.0044)	0.4717 (± 0.0053)	0.4478 (± 0.0055)	104.5304
Purely Model-Based	0.5299 (± 0.0070)	0.5238 (± 0.0071)	0.5229 (± 0.0071)	424.9928
FSL-MRS	0.5369 (± 0.0073)	0.5210 (± 0.0073)	0.5201 (± 0.0074)	962.7825
LCModel	0.5704 (± 0.0067)	0.5843 (± 0.0072)	0.5832 (± 0.0072)	78.0310

magnitudes across all three scenarios, but with runtimes orders of magnitude slower than the data-driven models. Extended results, including equivalent MAE performance, CNN-based baselines, self-supervised initialization of TTA, and alternative iteration counts for instance adaptation, are provided in Tables A1 and A2 in Appendix A. A comparison of MAE and MOSAE showed similar relative performance trends across quantification methods. Furthermore, similar observations regarding performance degradation from ID to OoD were found for the CNN-based baselines.

3.1.2 | Metabolite Distributions

Analysis of metabolite distributions revealed that for Glu, the predicted values closely matched the uniform ground truth, with only minor deviations (Figure 2). However, for GABA the predicted distributions were noticeably narrower. The predictions of the supervised method exhibited a confinement to its training distribution, with few estimates observed outside its trained range, particularly evident for lower SNR spectra. Self-supervised training similarly showed a bias towards its training distribution, with no significant trend observed in relation to SNR variations. In contrast, test-time instance adaptation demonstrated better coverage of the full range of concentrations. In the full-range (ID) scenario, its RMSE was slightly higher. The purely model-based approach showed constant performance and reasonable agreement with the ground truth distribution, though it consistently exhibits a high RMSE. Overall, higher RMSEs were observed for the OoD cases compared to the ID cases for both supervised and self-supervised methods, whereas test-time instance adaptive and purely model-based approaches maintained their performance across these scenarios.

The corresponding plots for the remaining quantification methods and the respective figures illustrating optimally scaled concentrations for relative metabolite quantification comparison are in Appendix A, Figures A1, A2, A3. Alternative metabolites showed similar effects and are provided in Appendix A, Figure A4 and A5 for OoD and Figure A6 and A7 for ID.

3.1.3 | Signal Parameter Perturbations

Test-time instance adaptive had the least effect on MOSAE in OoD data as compared to the other NN methods (Figure 3).

For ID spectra, all methods maintain uniform performance and good visual fits. However, under OoD conditions, frequency shifts caused minor degradation, while OoD phase shifts led to more pronounced errors for data-driven methods. Other signal parameter variations (SNR, linewidth, MMs, baseline, random nuisance effects) showed only minor OoD degradation, as can be seen in Appendix A, Figures A8, A9, A10, and A11.

3.1.4 | Performance Across Parameter Ranges

Figure 4 complements the single-spectrum scatter plots of Figure 3, highlighting systematic trends across methods and scenarios. Data-driven approaches generally show increased errors under extreme OoD conditions, while TTA and model-based methods maintain more stable performance across all parameters and metabolites. The observed trends hold for additional metabolites and signal parameters, as can be seen in Appendix A, Figures A12, A13, A14, and A15.

FIGURE 2 Scatter plots with marginal histograms comparing predicted versus true concentrations of glutamate (Glu) and gamma-aminobutyric acid (GABA) across 10,000 simulated spectra. Models were evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE 3 Scatter plots of quantification accuracy (MOSAE) across 10,000 simulated spectra as a function of ground truth frequency and zeroth-order phase shifts. Data-driven methods (supervised, self-supervised, and test-time instance adaptive) were compared against purely model-based fitting. Each point represents a single spectrum, illustrating method-specific sensitivity to core signal parameter variations. Example spectra from ID and OoD regions are shown above the scatter plots, with lines connecting each spectrum to its corresponding point, including fitted signals and residuals for all methods.

3.2 | Results on In-Vivo Data

3.2.1 | Overall Quantification Performance

The MOSAEs reported for in-vivo data (Table 3) were generally higher across most data-driven methods, indicating a domain shift between synthetic and in-vivo spectra. Supervised and self-supervised models exhibited the strongest increase in error. TTA methods retained lower errors, with domain adaptation achieving the best overall performance among adaptive approaches (0.3425 ± 0.0102 MOSAE in the OoD scenario). The purely model-based approach performed comparably well to the best adaptive methods (0.3431 ± 0.0106 MOSAE in the OoD scenario).

Additional results utilizing alternative pseudo ground truths (including the mean of FSL-MRS and LCModel, and LCModel alone) confirm these general trends across the methods (Tables A5, A6, A7, and A8 in Appendix A). Furthermore, extended initialization experiments reveal that test-time instance adaptation initialized from scratch achieved the lowest deviations from FSL-MRS pseudo ground truths, outperforming models initialized from the pretrained network (Tables A3 and A4 in Appendix A).

3.2.2 | Metabolite Distributions

Analyzing the metabolite distributions, supervised and self-supervised models exhibit narrow distributions with reduced slopes, consistent with regression toward the mean (Figures 5 and 6). This effect was less pronounced when considering relative concentrations (see Appendix, Figures A16 and A17). Test-time online adaptation remained similar to the supervised baseline, while instance and domain adaptation produced broader distributions with slopes closer to unity. The purely model-based approach was closely aligned with FSL-MRS, whereas LCModel showed the largest deviation.

FIGURE 4 Mean quantification error (MOSAE) across 10,000 simulated spectra for all methods under ID and OoD conditions. Results are shown for key signal parameters (SNR, linewidth, baseline, random walk) and metabolites (Aspartate (Asp), Cr, N-acetylaspartylglutamate (NAAG), N-acetylaspartate (NAA)). Each curve represents the mean error across spectra binned by the corresponding parameter value or metabolite concentration, summarizing method performance trends and sensitivity to challenging conditions.

TABLE 3 Comparison of quantification methods on 1,710 in-vivo spectra using pseudo ground truth: **FSL-MRS**. The spectra were filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)** Lowest error values are highlighted in **bold**. Comparisons where FSL-MRS is evaluated against its own estimates (using different numbers of averages: 4, 8, 16, 32, 64 vs. 64) are shown in *italics*.

Method	MOSAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	0.5164 (± 0.0170)	0.5533 (± 0.0136)	0.5562 (± 0.0159)	**0.1145**
Self-Supervised	0.5773 (± 0.0217)	0.6249 (± 0.0171)	0.5037 (± 0.0138)	0.1347
Test-Time Instance Adaptive	0.4498 (± 0.0144)	0.4818 (± 0.0115)	0.4489 (± 0.0118)	212.2983
Test-Time Online Adaptive	0.4793 (± 0.0158)	0.5156 (± 0.0127)	0.5101 (± 0.0143)	0.3400
Test-Time Domain Adaptive	**0.3052 (± 0.0120)**	**0.3425 (± 0.0102)**	0.3490 (± 0.0104)	312.6671
Purely Model-Based	0.3168 (± 0.0130)	0.3431 (± 0.0106)	**0.3431 (± 0.0106)**	3154.4558
FSL-MRS	*0.2104 (± 0.0119)*	*0.2210 (± 0.0097)*	*0.2210 (± 0.0097)*	475.4806
LCModel	0.7027 (± 0.0262)	0.7611 (± 0.0207)	0.7611 (± 0.0207)	274.6317

As in the simulations, the transition from ID to OoD testing was reflected in the distributions, with adaptive models extrapolating more effectively. Additional metabolites showed similar effects and are shown in Appendix A, Figures A18, A19, A20, and A21.

3.2.3 | Signal Parameter Perturbations

Most data-driven models maintain stable deviations from the pseudo ground truth across different SNR and linewidth conditions (Figure 7). Fitting-based methods show stronger variability: LCModel produces larger errors at high SNR compared to FSL-MRS, and all fitting approaches show increasing errors with broader linewidths. Supervised and self-supervised regressions display some inaccuracies even under ID conditions, while TTA methods remain relatively stable.

Extended sensitivity analyses, including frequency and phase shifts, baseline, and MM variations, showed the same stable deviations from the pseudo ground truth, as seen in Appendix A, Figures A22, A23, A24, and A25. As with the simulations, overall performance trends aggregated in binned means across the full parameter ranges are provided in Appendix A, Figures A26, A27, A28, and A29.

4 | DISCUSSION

Our core findings revealed a coupled bias–variance and computational tradeoff. Supervised regression achieved the lowest errors under ideal ID simulated conditions (0.2764 ± 0.0032 MOSAE in the restricted range case). This reflects a low variance model that relies heavily on its learned prior. Once forced to extrapolate beyond its trained concentration range, that same prior introduced systematic bias, nearly doubling the error (0.5537 ± 0.0058 MOSAE) and constraining metabolite estimates such as GABA to the training interval. TTA reduced this bias by relaxing the prior per sample, but at the cost of higher variance and additional computation. Instance adaptation proved substantially more resilient to concentration extrapolation in simulation (0.4420 ± 0.0052 MOSAE), and domain adaptation achieved the best overall adaptive performance when tested on unseen in-vivo spectra (0.3425 ± 0.0102 MOSAE OoD against the FSL-MRS reference).

4.1 | Performance of Data-Driven Methods

The simulations, although simplified, incorporated central challenges of MRS such as low SNR, peak overlap, broad linewidths, and baseline variability. Importantly, these factors make quantification inherently difficult even in the idealized case where the forward signal model used for fitting is identical to the generated spectra. The challenged performance of purely model-based methods in this setting suggests that the simulations reproduce meaningful aspects of the quantification problem and therefore provide a valuable evaluation for different strategies.

Across the simulated test scenarios, data-driven methods achieved low errors ID (Table 2) and maintained stable performance under perturbations in SNR, linewidth, etc. (Figure 4 and Appendix A). This indicates that learning-based approaches capture relevant spectral structure and yield consistent metabolite estimates when confronted with moderate variations in signal characteristics. However, the priors learned by these methods

FIGURE 5 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations of Cr and glutathione (GSH) across 1,710 in-vivo spectra. Models were evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Data-driven methods include supervised, self-supervised, and test-time instance adaptive approaches, compared with purely model-based fitting.

FIGURE 6 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations of Cr and GSH across 1,710 in-vivo spectra. Models were evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Methods included test-time online adaptive approaches, test-time domain adaptive approaches compared with FSL-MRS (Newton) and LCModel.

FIGURE 7 Scatter plots showing quantification accuracy (MOSAE) across 1,710 in-vivo spectra as a function of estimated SNR, linewidth. All methods are compared, with FSL-MRS re-estimating for 4, 18, 16 and 32 number of signal averages (NSA) then comparing against the 64 NSA pseudo ground truth. Each point represents one spectrum, illustrating method-specific sensitivity to core signal parameter variations.

reflect the source distribution, and shifts in the target metabolite concentrations can introduce mismatches that degrade performance.

Interestingly, training networks from scratch on individual spectra outperformed direct model-based optimization using the same signal model (Appendix A, Tables A1, A2, A3, A4). Here, the improvement did not come from priors across multiple examples but from the architectural structure of the network and the properties of gradient-based training, which together act as a form of implicit regularization.

4.2 | Role of Physics-Informed Models

Incorporating the signal model into training enabled self-supervised learning without ground truth metabolite concentrations, which is especially valuable for TTA methods. However, residual minimization can be ambiguous, as different parameter sets may yield similar fits. We observed that while the residual continues to decrease during training, the quantification error increases, indicating overfitting to the residual and limiting further improvements.

By incorporating the physics model into the optimization, the problem is no longer a black-box mapping from spectra to concentrations. Instead, the optimization landscape becomes physics-informed: certain parameters collapse to constrained, meaningful subspaces. For example, phase, which was previously an arbitrary number, becomes cyclic and physically interpretable. This constraining improves plausibility, but the model has no incentive to produce estimates outside the training distribution, as doing so increases errors during training. Consequently, while extrapolations are slightly smoother, overall OoD performance for metabolite concentrations remains limited (Figure 4).

4.3 | Regularization via Learned Priors

For a fixed likelihood, the Cramér-Rao bound (CRB) sets a lower bound on the variance of an unbiased estimator.[15] Reducing variance beyond this bound requires either additional information or the introduction of bias. Data-driven methods learn priors on metabolite parameters from the training data. These priors act as a form of regularization, lowering variance in the predictions, but they can also introduce systematic bias, particularly when the test/target distribution differs from the training/source distribution. TTA addresses this by adapting the network parameters to incoming test spectra, retaining some of the variance reduction offered by the learned priors while mitigating bias caused by domain shift.

4.4 | Adaptation & Initialization Effects

Supervised regression provides an effective initialization for TTA methods, placing estimates near the global optimum rather than in local minima defined only by residual minimization. Residual-based updates then refine these estimates on test spectra, improving alignment with the new distribution while remaining anchored to accurate starting points.

While our main results focused on overall performance, supplementary iteration experiments (Appendix A, Tables A1 and A2) indicated that too many fine-tuning steps can lead to overfitting on individual spectra, whereas a moderate number of updates provided more stable improvements. Overall, TTA offered a practical means of adapting to distribution shifts, but its computational cost and sensitivity to the number of updates remain important considerations.

4.5 | From Simulations to In-Vivo: Domain Shift

A key limitation was the lack of ground truth metabolite concentrations in-vivo, which prevented direct evaluation of absolute errors. We therefore assessed generalization only relative to alternative model-based quantification approaches. Consistent with prior work[48,17], even established tools can disagree substantially. To maintain a controlled comparison, we adopted a single signal model as the reference pseudo ground truth, with results from alternative models provided in Appendix A (Tables A5, A6, A7, and A8).

Methods that shared the same signal model, including FSL-MRS, deviated less from each other than from models using a different parametrization such as LCModel. This indicates that differences in the signal model contributed more to variability than generalization or domain shift itself. Across metabolites, deviations were primarily systematic offsets rather than random errors. Supervised and self-supervised methods generalize most poorly relative to the FSL-MRS reference, but they exhibit similar trends for ID and OoD spectra as observed in simulations, with reasonable performance for ID cases. TTA, however, performed well under both conditions, with domain adaptation in particular showing strong performance by converging reliably. Instance adaptation remained more sensitive to the number of iterations, mirroring the iteration-dependent effects seen in simulations.

4.6 | Training on Target Data

Supervised approaches are constrained to simulated training due to their reliance on ground-truth metabolite concentrations, and are consequently more affected by domain shift. In contrast, physics-informed methods, including self-supervised and TTA, can be trained directly on unlabeled in-vivo data. Data leakage is not an issue because these methods train solely on the measured spectrum and the forward model, never accessing ground truths.

When network initialization is ignored, training a self-supervised model on target data is equivalent to test-time domain adaptation, which refines the network using the entire test dataset. Domain adaptation effectively bridges systematic domain shifts, achieving strong overall adaptive performance (Table 3).

For instance adaptation, the impact of domain shift from supervised initialization is evident: more iterations progressively reduce deviation from all tested pseudo ground truths (Appendix A, Tables A3, A4, A5, A6, A7, and A8). When initialized with self-supervised priors, instance adaptation starts from a better baseline, achieving the best overall performance (Appendix A, Tables A5, A6, A7, and A8). Training from scratch, which removes any domain shift entirely, produces the lowest overall deviation from the pseudo ground truth FSL-MRS (Appendix A, Tables A3 and A4).

Online adaptation provides a practical trade-off between performance and speed, consistently outperforming the supervised baseline on in-vivo spectra (Table 3).

4.7 | Limitations

One of the main limitations of this study is the simplified signal model used for both the simulations and LCM. The model included only global linewidths and frequency shifts, a single MM component, and a second-order polynomial baseline. It did not account for local variations, metabolite-specific lineshapes, or distortions such as eddy currents. Incorporating additional parameters or using more flexible models could better capture the complexity of real spectra, but this comes at the cost of reduced interpretability, as it becomes harder to disentangle the influence of individual effects.

Another limitation concerns the training strategy. The networks were optimized for gradual and steady convergence on simulated data with known ground truths, but these settings may not translate optimally to in-vivo spectra, where variability is higher and ground truth is unavailable.

A further challenge observed across all data-driven methods is phase handling. These approaches are particularly sensitive to OoD phase variations. Model architecture plays a significant role: some architectures have demonstrated more robust phase handling in ID settings[49,50,51], but it remains unclear whether such robustness extends to OoD cases. Alternative input representations may also help address this limitation.[32]

Finally, other limitations must be considered, including potential biases from preprocessing, sensitivity to specific artifacts, and generalization across acquisition protocols. Addressing these factors is essential for developing robust MRS quantification methods that perform reliably across different datasets and experimental conditions.

4.8 | Broader Implications

Despite these limitations, the results offer several insights for the development of data-driven methods in MRS. Learned priors and adaptive strategies can stabilize parameter estimation and handle moderate variability, suggesting that networks trained on realistic simulations can provide robust quantification even in challenging conditions. Furthermore, training with in-vivo data with representative metabolite and signal perturbation parameters can lead to well performing methods ID and with proper adaptation strategies in place also OoD. These findings extend beyond metabolite quantification, emphasizing the value of incorporating prior knowledge, physics-informed modeling, and adaptive updates when designing data-driven pipelines in MRS.

5 | CONCLUSION

Data-driven strategies for MRS metabolite quantification, including supervised, self-supervised, and TTA, were systematically compared with purely model-based approaches. While data-driven methods showed strong performance in simulated ID scenarios, inherent biases and sensitivity to OoD conditions were observed. TTA techniques were found to significantly enhance robustness and generalizability in OoD settings for both simulated and in-vivo data, mitigating the impact of domain shift. The study highlighted the importance of integrating prior knowledge, physics-informed modeling, and adaptive updates for developing robust data-driven MRS pipelines.

ACKNOWLEDGMENTS

This work was in part funded by Spectralligence (EUREKA IA Call, ITEA4 project 20209). We further acknowledge support from the NVIDIA Academic Hardware Grant Program, and acknowledge the CIBM Center for Biomedical Imaging for providing expertise and resources to conduct this study.

DATA AVAILABILITY STATEMENT

The source code used in our experiments for the methods, data simulation, and analysis can be found at https://github.com/julianmer/OoD-Robust-MRS-Quantification. The in-vivo data are available from the corresponding author upon reasonable request.

REFERENCES

1. Faghihi Reza, Zeinali-Rafsanjani Banafsheh, Mosleh-Shirazi Mohammad-Amin, et al. Magnetic Resonance Spectroscopy and Its Clinical Applications: A Review. *Journal of Medical Imaging and Radiation Sciences.* 2017;48(3):233–253.
2. Maudsley Andrew A., Andronesi Ovidiu C., Barker Peter B., et al. Advanced magnetic resonance spectroscopic neuroimaging: Experts' consensus recommendations. *NMR in Biomedicine.* 2020;34.
3. Horská Alena, Berrington Adam, Barker Peter B., Tkáč Ivan. Magnetic Resonance Spectroscopy: Clinical Applications:241–292. Cham: Springer International Publishing 2023.
4. Kreis Roland. Issues of spectral quality in clinical 1H-magnetic resonance spectroscopy and a gallery of artifacts. *NMR in Biomedicine.* 2004;17.
5. Hurd Ralph E.. Artifacts and pitfalls in MR spectroscopy:30–43. Cambridge University Press 2009.
6. Near Jamie, Harris Ashley D., Juchem Christoph, et al. Preprocessing, Analysis and Quantification in Single-Voxel Magnetic Resonance Spectroscopy: Experts' Consensus Recommendations. *NMR in Biomedicine.* 2021;34(5):e4257.
7. Levenberg Kenneth. A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics.* 1944;2:164-168.
8. Marquardt Donald W.. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of The Society for Industrial and Applied Mathematics.* 1963;11:431-441.
9. Provencher Stephen W.. Estimation of Metabolite Concentrations from Localizedin Vivo Proton NMR Spectra. *Magnetic Resonance in Medicine.* 1993;30(6):672–679.
10. Soher Brian J., Semanchuk Philip, Todd David, et al. Vespa: Integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Magnetic Resonance in Medicine.* 2023;90(3):823-838.
11. Gajdošík Martin, Landheer Karl, Swanberg Kelley M., Juchem Christoph. INSPECTOR: free software for magnetic resonance spectroscopy data inspection, processing, simulation and analysis. *Scientific Reports.* 2021;11.
12. Oeltzschner Georg, Zoellner H, Hui Steve C. N., et al. Osprey: Open-source processing, reconstruction & estimation of magnetic resonance spectroscopy data. *Journal of Neuroscience Methods.* 2020;343.
13. Clarke William T., Stagg Charlotte J., Jbabdi Saad. FSL-MRS: An End-to-end Spectroscopy Analysis Package. *Magnetic Resonance in Medicine.* 2021;85(6):2950–2964.
14. Poullet Jean-Baptiste, Sima Diana M., Van Huffel Sabine. MRS Signal Quantitation: A Review of Time- and Frequency-Domain Methods. *Journal of Magnetic Resonance.* 2008;195(2):134–144.
15. Landheer Karl, Juchem Christoph. Are Cramér-Rao lower bounds an accurate estimate for standard deviations in in vivo magnetic resonance spectroscopy?. *NMR in Biomedicine.* 2021;34(7):e4521.
16. Marjańska Małgorzata, Terpstra Melissa. Influence of fitting approaches in LCModel on MRS quantification focusing on age-specific macromolecules and the spline baseline. *NMR in Biomedicine.* 2021;34(5):e4197. e4197 NBM-19-0058.R2.
17. Zöllner Helge J., Považan Michal, Hui Steve C.N., Tapper Sofie, Edden Richard A.E., Oeltzschner Georg. Comparison of different linear-combination modeling algorithms for short-TE proton spectra. *NMR in Biomedicine.* 2021;34(4):e4482. e4482 NBM-20-0312.
18. Sande Dennis M. J., Merkofer Julian P., Amirrajab Sina, et al. A review of machine learning applications for the proton MR spectroscopy workflow. *Magnetic Resonance in Medicine.* 2023;90(4):1253-1270.
19. Luo Yao, Zheng Xiaoxu, Qiu Mengjie, et al. Deep learning and its applications in nuclear magnetic resonance spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy.* 2025;146-147:101556.
20. Das Dhritiman, Coello Eduardo, Schulte Rolf F., Menze Bjoern H.. *Quantification of Metabolites in Magnetic Resonance Spectroscopic Imaging Using Machine Learning.* 2017.
21. Hatami Nima, Sdika Michaël, Ratiney Hélène. *Magnetic Resonance Spectroscopy Quantification Using Deep Learning.* 2018.
22. Chandler M., Jenkins C., Shermer S. M., Langbein F. C.. MRSNet: Metabolite Quantification from Edited Magnetic Resonance Spectra With Convolutional Neural Networks. *arXiv preprint arXiv:1909.03836.* 2019;.
23. Shamaei Amirmohammad, Starčuková Jana, Starčuk Jr. Zenon. *A Wavelet Scattering Convolutional Network for Magnetic Resonance Spectroscopy Signal Quantitation:.* 2021.
24. Lee Hyeong Hun, Kim Hyeonjin. Deep Learning-Based Target Metabolite Isolation and Big Data-Driven Measurement Uncertainty Estimation in Proton Magnetic Resonance Spectroscopy of the Brain. *Magnetic Resonance in Medicine.* 2020;84(4):1689–1706.

25. Iqbal Zohaib, Nguyen Dan, Thomas Michael Albert, Jiang Steve. Deep Learning Can Accelerate and Quantify Simulated Localized Correlated Spectroscopy. *Scientific Reports.* 2021;11(1):8727.
26. Gurbani Saumya S., Sheriff Sulaiman, Maudsley Andrew A., Shim Hyunsuk, Cooper Lee A.D.. Incorporation of a Spectral Model in a Convolutional Neural Network for Accelerated Spectral Fitting. *Magnetic Resonance in Medicine.* 2019;81(5):3346–3357.
27. Shamaei Amirmohammad, Starcukova Jana, Starcuk Zenon. Physics-Informed Deep Learning Approach to Quantification of Human Brain Metabolites from Magnetic Resonance Spectroscopy Data. *Computers in Biology and Medicine.* 2023;158:106837.
28. Chen Dicheng, Lin Meijin, Liu Huiting, et al. Magnetic Resonance Spectroscopy Quantification Aided by Deep Estimations of Imperfection Factors and Macromolecular Signal. *IEEE Transactions on Biomedical Engineering.* 2024;71(6):1841-1852.
29. Bishop Christopher M.. *Pattern Recognition and Machine Learning.* Information Science and StatisticsNew York: Springer; 2006.
30. Gudmundson Aaron T., Davies-Jenkins Christopher W., Özdemir İpek, et al. Application of a 1H brain MRS benchmark dataset to deep learning for out-of-voxel artifacts. *Imaging Neuroscience.* 2023;1:1-15.
31. Mohammed Sedir, Budach Lukas, Feuerpfeil Moritz, et al. The effects of data quality on machine learning performance on tabular data. *Information Systems.* 2025;132:102549.
32. Rizzo Rudy, Dziadosz Martyna, Kyathanahally Sreenath P., Shamaei Amirmohammad, Kreis Roland. 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.* 2023;89(5):1707–1727.
33. Lee Hyeong Hun, Kim Hyeonjin. Bayesian Deep Learning–Based 1H-MRS of the Brain: Metabolite Quantification with Uncertainty Estimation Using Monte Carlo Dropout. *Magnetic Resonance in Medicine.* 2022;88(1):38–52.
34. Rizzo Rudy, Dziadosz Martyna, Kyathanahally Sreenath P., Reyes Mauricio, Kreis Roland. *Reliability of Quantification Estimates in MR Spectroscopy: CNNs vs Traditional Model Fitting.* 2022.
35. Sun Yu, Wang Xiaolong, Liu Zhuang, Miller John, Efros Alexei, Hardt Moritz. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In: III Hal Daumé, Singh Aarti, eds. *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 119: :9229–9248PMLR; 2020.
36. Wilson Garrett, Cook Diane J.. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology.* 2020;11(5):1–46. Epub 2020 Jul 5.
37. Kouw Wouter M., Loog Marco. A Review of Domain Adaptation without Target Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2021;43(3):766-785.
38. Fang Yuqi, Yap Pew-Thian, Lin Weili, Zhu Hongtu, Liu Mingxia. Source-free unsupervised domain adaptation: A survey. *Neural Networks.* 2024;174:106230.
39. Li Jingjing, Yu Zhiqi, Du Zhekai, Zhu Lei, Shen Heng Tao. A Comprehensive Survey on Source-Free Domain Adaptation. *IEEE Transactions on Pattern Analysis & Machine Intelligence.* 2024;46(08):5743-5762.
40. Liang Jian, He Ran, Tan Tieniu. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision.* 2025;133(1):31–64.
41. De Graaf Robin A.. *In Vivo NMR Spectroscopy: Principles and Techniques.* Hoboken, NJ: John Wiley & Sons, Inc; 3rd ed ed.2019.
42. Schrantee Anouk, Najac Chloe, Jungerius Chris, et al. A 7T interleaved fMRS and fMRI study on visual contrast dependency in the human brain. *Imaging Neuroscience.* 2023;1:1-15.
43. Boer Vincent O., Andersen Mads, Lind Anna, Lee Nam Gyun, Marsman Anouk, Petersen Esben T.. MR spectroscopy using static higher order shimming with dynamic linear terms (HOS-DLT) for improved water suppression, interleaved MRS-fMRI, and navigator-based motion correction at 7T. *Magnetic Resonance in Medicine.* 2020;84(3):1101-1112.
44. Lin Alexander, Andronesi Ovidiu, Bogner Wolfgang, et al. Minimum Reporting Standards for in Vivo Magnetic Resonance Spectroscopy (MRSinMRS): Experts' Consensus Recommendations. *NMR in Biomedicine.* 2021;34(5).
45. Gavin Henri P.. *The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems.* Accessed: 2025-07-18; 2011.
46. Kingma Diederik P., Ba Jimmy. Adam: A Method for Stochastic Optimization. In: Bengio Yoshua, LeCun Yann, eds. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ; 2015.
47. Clevert Djork-Arné, Unterthiner Thomas, Hochreiter Sepp. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In: Bengio Yoshua, LeCun Yann, eds. *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, ; 2016.
48. Bhogal Alex A., Schür Remmelt R., Houtepen Lotte C., et al. 1H–MRS processing parameters affect metabolite quantification: The urgent need for uniform and transparent standardization. *NMR in Biomedicine.* 2017;30(11):e3804. e3804 NBM-17-0072.R2.
49. Tapper Sofie, Mikkelsen Mark, Dewey Blake E., et al. Frequency and Phase Correction of J-difference Edited MR Spectra Using Deep Learning. *Magnetic Resonance in Medicine.* 2021;85(4):1755–1765.
50. Ma David J., Le Hortense A-M., Ye Yuming, et al. MR Spectroscopy Frequency and Phase Correction Using Convolutional Neural Networks. *Magnetic Resonance in Medicine.* 2022;87(4):1700–1710.
51. Shamaei Amirmohammad, Starcukova Jana, Pavlova Iveta, Starcuk Jr. Zenon. Model-Informed Unsupervised

Deep Learning Approaches to Frequency and Phase Correction of MRS Signals. *Magnetic Resonance in Medicine.* 2023;89(3):1221–1236.
52. Rayleigh . The Problem of the Random Walk. *Nature.* 1905;72:318-318.
53. Susnjar Antonia, Kaiser Antonia, Simicic Dunja, et al. Reproducibility Made Easy: A Tool for Methodological Transparency and Efficient Standardized Reporting Based on the Proposed MRSinMRS Consensus. *NMR in Biomedicine.* 2025;38(6):e70039. e70039 NBM-24-0224.R2.
54. Cudalbu Cristina, Behar Kevin L., Bhattacharyya Pallab K., et al. Contribution of Macromolecules to Brain 1H MR Spectra: Experts' Consensus Recommendations. *NMR in Biomedicine.* 2021;34(5):e4393.

APPENDIX

A ADDITIONAL MATERIALS

This section provides supplementary analyses to support the findings and discussions of this work.

A.1 Additional Results on Simulated Data

This part offers extended results from the simulated MRS data, including:

- **Tables A1 and A2**: These tables provide comprehensive comparisons of quantification methods across 10,000 simulated spectra, detailing both MAE and MOSAE, respectively. Three primary scenarios for metabolite concentration ranges are covered:
  - **ID (Mid-Range)**: Models trained and tested on the central 50% metabolite concentration range.
  - **OoD (Full-Range)**: Models trained on the central 50% concentration range but tested across the entire range of concentrations to assess extrapolation.
  - **ID (Full-Trained)**: Models trained and tested on the full concentration range.

  These tables also include performance metrics for CNN-based baselines, self-supervised initialization of TTA models, and various iteration counts for instance adaptation.

- **Figures A1, A2, A3**: These scatter plots with marginal histograms compare predicted versus true concentrations for Glu and GABA across 10,000 simulated spectra. They are evaluated under OoD and ID scenarios, including optimally scaled concentrations for relative metabolite quantification comparison.

- **Figures A4, A5, A6, A7**: These figures show scatter plots with marginal histograms comparing predicted versus true concentrations for other metabolites such as NAA, Cr, Glu, GSH, and GABA, under both OoD (mid-range trained, full-range tested) and ID (full-range trained) conditions.

- **Figures A8, A9, A10, A11**: These scatter plots illustrate quantification accuracy (MOSAE) for various signal parameter SNR, linewidth, zeroth-order phase shift, frequency offset, MM, baseline, and random signal corruptions.
  - The random signal corruption is generated by $R(f)$, a complex-valued random walk that introduces arbitrary spectral distortions. This process is defined as a bounded, smoothed stochastic process with independent real and imaginary components[52]. This random walk component $R(f)$ is exclusively added during evaluation to assess robustness to random spectral artifacts and is excluded during the training phase.

- **Figures A12, A13, A14, A15**: These figures summarize quantification performance across various metabolites and other signal parameters, displaying the mean MOSAE within parameter bins.

A.2 Additional Results on In-Vivo Data

This section presents further results from the in-vivo data evaluation, including:

- **Tables A3, A4, A5, A6, A7, A8**: These tables extend the in-vivo quantification performance analysis using different pseudo ground truths (FSL-MRS, mean of FSL-MRS and LCModel, and LCModel), providing both MAE and MOSAE across the ID and OoD scenarios. They also include results for CNN-based models and various TTA initialization and iteration settings.

- **Figures A16, A17**: These scatter plots with marginal histograms compare optimally scaled predicted versus pseudo-true concentrations for Glu

and GABA across 1,710 in-vivo spectra, under ID and OoD conditions.

- **Figures A18, A19, A20, A21**: These figures show predicted versus pseudo-true concentrations for NAA, Cr, Glu, GSH, and GABA in-vivo, under both OoD and ID full-range scenarios.
- **Figures A22, A23, A24, A25**: These scatter plots illustrate MOSAE across in-vivo spectra as a function of estimated SNR, linewidth, zeroth-order phase shift, frequency offset, MM, and baseline variation.
- **Figures A26, A27, A28, A29**: These figures provide a summary of quantification performance for various metabolites and signal parameters in in-vivo spectra, showing mean MOSAE within parameter bins.

B IMPLEMENTATION DETAILS

This section offers additional information about the specific implementation choices made throughout this work.

B.1 Metabolite Concentration Ranges

The origins of the metabolite concentration ranges used for simulation are detailed in the following tables. Bounds were derived from a combination of literature values reported by De Graaf 2019[41] and empirical distributions obtained by fitting all in-vivo spectra using LCModel[9] and FSL-MRS[13] (Tables B9, B10, and B11).

B.2 Model Architecture & Setup Configuration

Layer-wise architectures and configuration parameters used for training and testing the models are provided in Tables B12 and B13.

C MRS IN MRS

This section presents tables that detail the acquisitions, experimental setup, processing, and data analysis methods employed in the study, adhering to the MRSinMRS guidelines[44], made easier by REMY[53].

D HARDWARE & SOFTWARE ENVIRONMENT

D.1 Simulation Environment

All simulated experiments and runtime benchmarks were executed on a workstation with the following specifications:

- **CPU**: AMD Ryzen 9 7950X
- **GPU**: NVIDIA RTX 6000 Ada
- **Memory**: 128 GB DDR5-6000
- **OS**: Ubuntu 24.04.2 LTS
- **Software**: Python 3.10.16, PyTorch 2.6.0, CUDA 12.4

Runtimes reported in tables 2, A1 and A2 reflect per-sample inference times measured on simulated test data, using the most efficient configuration for each method (e.g., GPU where applicable, multiprocessing, etc.).

D.2 In-Vivo Environment

All in-vivo experiments were executed on a high-performance computing cluster managed by SLURM. Jobs were scheduled on dual-socket AMD EPYC 7662 nodes (128 cores, 256 threads) with the following resource allocation:

- **CPU**: 8 CPUs per task
- **GPU**: NVIDIA A100 with 10–40 GB memory (depending on job configuration)
- **Memory**: 128 GB system memory
- **OS**: Red Hat Enterprise Linux 8.10 (Ootpa)
- **Software**: Python 3.11.13, PyTorch 2.7.1, CUDA 12.8

Runtimes reported for in-vivo experiments in tables 3, A3, A4, A5, A6, A7, and A8 correspond to these allocated resources, and were measured within SLURM-managed jobs on dedicated compute nodes.

TABLE A1 Comparison of quantification methods across test scenarios of 10,000 spectra. **ID (Mid-Range)**: Models trained and tested on the central 50% metabolite concentration range. **OoD (Full-Range)**: Models trained on the central 50% concentration range, but tested across the entire range of concentrations to assess extrapolation. **ID (Full-Trained)**: Models trained and tested on the full concentration range.

Method	MAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	**0.2962 (± 0.0033)**	0.5970 (± 0.0059)	**0.4121 (± 0.0047)**	**0.1565**
Self-Supervised	0.4037 (± 0.0046)	0.5846 (± 0.0060)	0.4863 (± 0.0060)	0.1574
TT Instance Adaptive	0.4071 (± 0.0051)	**0.4767 (± 0.0054)**	0.4699 (± 0.0061)	30.4637
TT Online Adaptive	0.3025 (± 0.0033)	0.5917 (± 0.0059)	0.4182 (± 0.0048)	0.2454
TT Domain Adaptive	0.3977 (± 0.0046)	0.5166 (± 0.0054)	0.4888 (± 0.0056)	104.5304
Purely Model-Based	0.5881 (± 0.0071)	0.5819 (± 0.0072)	0.5822 (± 0.0072)	424.9928
Supervised (CNN)	0.3212 (± 0.0034)	0.6483 (± 0.0063)	0.4326 (± 0.0049)	0.1642
Self-Supervised (CNN)	0.4075 (± 0.0046)	0.5780 (± 0.0060)	0.4866 (± 0.0059)	0.1643
TT Instance Adaptive (CNN)	0.4649 (± 0.0051)	0.6364 (± 0.0066)	0.4479 (± 0.0056)	40.7961
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.4676 (± 0.0058)	0.5412 (± 0.0063)	0.5247 (± 0.0072)	40.9133
TT Instance Adaptive (Self-Sup. Init.)	0.4375 (± 0.0053)	0.5362 (± 0.0060)	0.5152 (± 0.0070)	30.5271
TT Instance Adaptive (From Scratch Init.)	0.5140 (± 0.0059)	0.6119 (± 0.0069)	0.6038 (± 0.0069)	4567.4828
TT Online Adaptive (Self-Sup. Init.)	0.4107 (± 0.0046)	0.5876 (± 0.0060)	0.4967 (± 0.0061)	0.2362
TT Domain Adaptive (Self-Sup. Init.)	0.4677 (± 0.0050)	0.6299 (± 0.0064)	0.5632 (± 0.0065)	105.9836
TT Instance Adaptive (10 Iter.)	0.3170 (± 0.0036)	0.5285 (± 0.0053)	0.4245 (± 0.0050)	12.7126
TT Instance Adaptive (50 Iter.)	0.4071 (± 0.0051)	0.4767 (± 0.0054)	0.4378 (± 0.0060)	30.5217
TT Instance Adaptive (100 Iter.)	0.4660 (± 0.0061)	0.4873 (± 0.0060)	0.5071 (± 0.0069)	59.3287
TT Instance Adaptive (500 Iter.)	0.5312 (± 0.0070)	0.5138 (± 0.0070)	0.5363 (± 0.0074)	285.3065

TABLE A2 Comparison of quantification methods across test scenarios of 10,000 spectra. **ID (Mid-Range)**: Models trained and tested on the central 50% metabolite concentration range. **OoD (Full-Range)**: Models trained on the central 50% concentration range, but tested across the entire range of concentrations to assess extrapolation. **ID (Full-Trained)**: Models trained and tested on the full concentration range.

Method	MOSAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	**0.2764 (± 0.0032)**	0.5537 (± 0.0058)	**0.3896 (± 0.0047)**	**0.1565**
Self-Supervised	0.3734 (± 0.0044)	0.5401 (± 0.0059)	0.4548 (± 0.0059)	0.1574
TT Instance Adaptive	0.3767 (± 0.0049)	**0.4420 (± 0.0052)**	0.4378 (± 0.0060)	30.4637
TT Online Adaptive	0.2788 (± 0.0032)	0.5520 (± 0.0058)	0.3923 (± 0.0048)	0.2454
TT Domain Adaptive	0.3621 (± 0.0044)	0.4717 (± 0.0053)	0.4478 (± 0.0055)	104.5304
Purely Model-Based	0.5299 (± 0.0070)	0.5238 (± 0.0071)	0.5229 (± 0.0071)	424.9928
FSL-MRS	0.5369 (± 0.0073)	0.5210 (± 0.0073)	0.5201 (± 0.0074)	962.7825
LCModel	0.5704 (± 0.0067)	0.5843 (± 0.0072)	0.5832 (± 0.0072)	78.0310
Supervised (CNN)	0.3017 (± 0.0033)	0.6073 (± 0.0062)	0.4084 (± 0.0049)	0.1642
Self-Supervised (CNN)	0.3767 (± 0.0045)	0.5335 (± 0.0059)	0.4553 (± 0.0058)	0.1643
TT Instance Adaptive (CNN)	0.4229 (± 0.0048)	0.5911 (± 0.0063)	0.4169 (± 0.0055)	40.7961
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.4316 (± 0.0056)	0.5027 (± 0.0061)	0.4843 (± 0.0068)	40.9133
TT Instance Adaptive (Self-Sup. Init.)	0.4004 (± 0.0051)	0.4985 (± 0.0059)	0.4768 (± 0.0067)	30.5271
TT Instance Adaptive (From Scratch Init.)	0.4492 (± 0.0055)	0.5497 (± 0.0066)	0.5395 (± 0.0066)	4567.4828
TT Online Adaptive (Self-Sup. Init.)	0.4004 (± 0.0051)	0.4985 (± 0.0059)	0.4594 (± 0.0060)	0.2362
TT Domain Adaptive (Self-Sup. Init.)	0.4007 (± 0.0045)	0.5571 (± 0.0061)	0.5081 (± 0.0063)	105.9836
TT Instance Adaptive (10 Iter.)	0.2932 (± 0.0035)	0.4979 (± 0.0053)	0.3961 (± 0.0049)	12.7126
TT Instance Adaptive (50 Iter.)	0.3767 (± 0.0049)	0.4420 (± 0.0052)	0.4378 (± 0.0060)	30.5217
TT Instance Adaptive (100 Iter.)	0.4283 (± 0.0058)	0.4496 (± 0.0058)	0.4684 (± 0.0066)	59.3287
TT Instance Adaptive (500 Iter.)	0.4859 (± 0.0066)	0.4708 (± 0.0066)	0.4913 (± 0.0070)	285.3065

FIGURE A1 Scatter plots with marginal histograms comparing predicted versus true concentrations of Glu and GABA across 10,000 simulated spectra. Models are evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Methods include test-time online adaptive approaches and test-time domain adaptive approaches. Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE A2 Scatter plots with marginal histograms comparing optimally scaled predicted versus true concentrations of Glu and GABA across 10,000 simulated spectra. Models are evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Data-driven methods include supervised, self-supervised, and test-time instance adaptive approaches, compared with purely model-based fitting. Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE A3 Scatter plots with marginal histograms comparing optimally scaled predicted versus true concentrations of Glu and GABA across 10,000 simulated spectra. Models are evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Methods include test-time online adaptive approaches, test-time domain adaptive approaches compared with FSL-MRS (Newton) and LCModel. Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE A4 Scatter plots with marginal histograms comparing predicted versus true concentrations for NAA, Cr, Glu, GSH, and GABA across 10,000 simulated spectra. Models trained on mid-range concentrations are evaluated across the full concentration range (OoD). Data-driven methods include supervised, self-supervised, and test-time instance adaptive compared against purely model-based fitting. Points are colored by SNR, and regression lines with corresponding statistics (R^2, slope, intercept, RMSE) are included.

FIGURE A5 Scatter plots with marginal histograms comparing predicted versus true concentrations for NAA, Cr, Glu, GSH, and GABA across 10,000 simulated spectra. Models trained on mid-range concentrations are evaluated across the full concentration range (OoD). Adaptive methods include test-time online adaptive and test-time domain adaptive. Points are colored by SNR, and regression lines with corresponding statistics are included.

FIGURE A6 Scatter plots with marginal histograms comparing predicted versus true concentrations for NAA, Cr, Glu, GSH, and GABA under the ID full-range scenario, where models are trained and tested on the full concentration range. This figure shows data-driven methods: supervised, self-supervised, and test-time instance adaptive against purely model-based. Points are colored by SNR, and regression lines with corresponding statistics are included.

FIGURE A7 Scatter plots with marginal histograms comparing predicted versus true concentrations for NAA, Cr, Glu, GSH, and GABA under the ID full-range scenario. This figure shows adaptive methods: test-time online adaptive and test-time domain adaptive. Points are colored by SNR, and regression lines with corresponding statistics are included.

FIGURE A8 Scatter plots showing quantification accuracy (MOSAE) across 10,000 simulated spectra as a function of ground truth SNR, linewidth, zeroth-order phase shift, and frequency offset. Data-driven methods include supervised, self-supervised, and test-time instance adaptive compared against purely model-based fitting. Each point represents one spectrum, illustrating method-specific sensitivity to core signal parameter variations.

FIGURE A9 Scatter plots showing quantification accuracy (MOSAE) across 10,000 simulated spectra under MM, baseline, and random signal corruptions. Data-driven methods include supervised, self-supervised, and test-time instance adaptive compared against purely model-based fitting. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations.

FIGURE A10 Scatter plots showing quantification accuracy (MOSAE) across 10,000 simulated spectra as a function of ground truth SNR, linewidth, zeroth-order phase shift, and frequency offset. Adaptive and classical methods include test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations.

FIGURE A11 Scatter plots showing quantification accuracy (MOSAE) across 10,000 simulated spectra under MM, baseline, and random signal corruptions. Adaptive and classical methods include test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations.

FIGURE A12 Summary of quantification performance across 10,000 simulated spectra for eight metabolites (Ala, Asc, Asp, Cr, GABA, Gln, Glu, Gly). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under OoD conditions.

FIGURE A13 Summary of quantification performance across 10,000 simulated spectra for eight metabolites (GPC, GSH, mIns, Lac, NAAG, NAA, PCh, PCr). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under OoD conditions.

FIGURE A14 Summary of quantification performance across 10,000 simulated spectra for four metabolites (PE, Scyllo, Ser, Tau). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under OoD conditions.

FIGURE A15 Summary of quantification performance under OoD signal perturbations. For each signal parameter, SNR, linewidth, frequency offset, phase shift, MM baseline, and polynomial baseline, the MOSAE is averaged within parameter bins. All methods are overlaid in each subplot, including model fitting (purely model-based gradient decent, FSL-MRS (Newton), and LCModel), supervised, self-supervised, along with the TTA strategies.

TABLE A3 Comparison of quantification methods on in-vivo data using pseudo ground truth: **FSL-MRS**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	**MAE ↓ (± SE)**			**Time ↓ (ms/sample)**
	ID (Mid-Range)	**OoD (Full-Range)**	**ID (Full-Trained)**	
Supervised	0.7953 (± 0.0265)	0.8633 (± 0.0211)	0.7623 (± 0.0233)	**0.1145**
Self-Supervised	0.8173 (± 0.0267)	0.8942 (± 0.0218)	0.7073 (± 0.0215)	0.1347
TT Instance Adaptive	0.4674 (± 0.0140)	0.5063 (± 0.0116)	0.4770 (± 0.0120)	212.2983
TT Online Adaptive	0.6751 (± 0.0220)	0.7440 (± 0.0179)	0.7055 (± 0.0212)	0.3400
TT Domain Adaptive	0.3403 (± 0.0122)	0.3922 (± 0.0111)	0.3950 (± 0.0111)	312.6671
Purely Model-Based	0.3500 (± 0.0134)	0.3802 (± 0.0112)	0.3802 (± 0.0112)	3154.4558
FSL-MRS	*0.2390 (± 0.0130)*	*0.2560 (± 0.0109)*	*0.2560 (± 0.0109)*	475.4806
LCModel	0.7288 (± 0.0259)	0.7935 (± 0.0205)	0.7935 (± 0.0205)	274.6317
Supervised (CNN)	0.7484 (± 0.0221)	0.8245 (± 0.0180)	0.7757 (± 0.0190)	0.7374
Self-Supervised (CNN)	0.8474 (± 0.0327)	0.8963 (± 0.0248)	0.5071 (± 0.0127)	0.1551
TT Instance Adaptive (CNN)	0.6163 (± 0.0179)	0.6662 (± 0.0143)	0.5326 (± 0.0124)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.5948 (± 0.0197)	0.6474 (± 0.0156)	0.4592 (± 0.0121)	270.3693
TT Instance Adaptive (Self-Sup. Init.)	0.5419 (± 0.0178)	0.6040 (± 0.0148)	0.4559 (± 0.0119)	205.7727
TT Instance Adaptive (From Scratch Init.)	**0.3272 (± 0.0121)**	**0.3739 (± 0.0109)**	**0.3761 (± 0.0109)**	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.6353 (± 0.0217)	0.7001 (± 0.0176)	0.5357 (± 0.0150)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.3752 (± 0.0140)	0.4260 (± 0.0122)	0.4219 (± 0.0123)	312.5201
TT Instance Adaptive (10 Iter.)	0.6069 (± 0.0189)	0.6744 (± 0.0156)	0.6571 (± 0.0192)	44.8087
TT Instance Adaptive (50 Iter.)	0.4674 (± 0.0140)	0.5063 (± 0.0116)	0.4770 (± 0.0120)	212.2983
TT Instance Adaptive (100 Iter.)	0.4220 (± 0.0134)	0.4657 (± 0.0114)	0.4404 (± 0.0118)	414.0273
TT Instance Adaptive (500 Iter.)	0.4171 (± 0.0141)	0.4565 (± 0.0121)	0.4574 (± 0.0127)	2061.0019

TABLE A4 Comparison of quantification methods on in-vivo data using pseudo ground truth: **FSL-MRS**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	**MOSAE ↓ (± SE)**			**Time ↓ (ms/sample)**
	ID (Mid-Range)	**OoD (Full-Range)**	**ID (Full-Trained)**	
Supervised	0.5164 (± 0.0170)	0.5533 (± 0.0136)	0.5562 (± 0.0159)	**0.1145**
Self-Supervised	0.5773 (± 0.0217)	0.6249 (± 0.0171)	0.5037 (± 0.0138)	0.1347
TT Instance Adaptive	0.4498 (± 0.0144)	0.4818 (± 0.0115)	0.4489 (± 0.0118)	212.2983
TT Online Adaptive	0.4793 (± 0.0158)	0.5156 (± 0.0127)	0.5101 (± 0.0143)	0.3400
TT Domain Adaptive	0.3052 (± 0.0120)	0.3425 (± 0.0102)	0.3490 (± 0.0104)	312.6671
Purely Model-Based	0.3168 (± 0.0130)	0.3431 (± 0.0106)	0.3431 (± 0.0106)	3154.4558
FSL-MRS	*0.2104 (± 0.0119)*	*0.2210 (± 0.0097)*	*0.2210 (± 0.0097)*	475.4806
LCModel	0.7027 (± 0.0262)	0.7611 (± 0.0207)	0.7611 (± 0.0207)	274.6317
Supervised (CNN)	0.5543 (± 0.0184)	0.5923 (± 0.0141)	0.5302 (± 0.0131)	0.7374
Self-Supervised (CNN)	0.7274 (± 0.0278)	0.7538 (± 0.0211)	0.4423 (± 0.0125)	0.1551
TT Instance Adaptive (CNN)	0.5834 (± 0.0184)	0.6199 (± 0.0141)	0.4991 (± 0.0123)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.5509 (± 0.0206)	0.5902 (± 0.0161)	0.4254 (± 0.0118)	270.3693
TT Instance Adaptive (Self-Sup. Init.)	0.5029 (± 0.0185)	0.5518 (± 0.0150)	0.4178 (± 0.0116)	205.7727
TT Instance Adaptive (From Scratch Init.)	**0.2982 (± 0.0120)**	**0.3307 (± 0.0102)**	**0.3328 (± 0.0102)**	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.5339 (± 0.0202)	0.5842 (± 0.0160)	0.4080 (± 0.0110)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.3169 (± 0.0123)	0.3528 (± 0.0104)	0.3419 (± 0.0107)	312.5201
TT Instance Adaptive (10 Iter.)	0.4654 (± 0.0150)	0.5018 (± 0.0122)	0.5179 (± 0.0147)	44.8087
TT Instance Adaptive (50 Iter.)	0.4498 (± 0.0144)	0.4818 (± 0.0115)	0.4489 (± 0.0118)	212.2983
TT Instance Adaptive (100 Iter.)	0.4028 (± 0.0135)	0.4361 (± 0.0110)	0.4100 (± 0.0114)	414.0273
TT Instance Adaptive (500 Iter.)	0.3932 (± 0.0143)	0.4221 (± 0.0117)	0.4253 (± 0.0124)	2061.0019

TABLE A5 Comparison of quantification methods on in-vivo data using pseudo ground truth: **Mean of FSL-MRS and LCModel**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	**MAE ↓ (± SE)**			**Time ↓ (ms/sample)**
	ID (Mid-Range)	**OoD (Full-Range)**	**ID (Full-Trained)**	
Supervised	0.7251 (± 0.0386)	0.7498 (± 0.0219)	0.8030 (± 0.0245)	**0.1145**
Self-Supervised	0.6487 (± 0.0353)	0.6902 (± 0.0206)	0.7501 (± 0.0220)	0.1347
TT Instance Adaptive	0.5806 (± 0.0250)	0.5671 (± 0.0133)	0.6576 (± 0.0154)	212.2983
TT Online Adaptive	0.6354 (± 0.0319)	0.6617 (± 0.0185)	0.7498 (± 0.0228)	0.3400
TT Domain Adaptive	0.4586 (± 0.0217)	0.4723 (± 0.0122)	0.4832 (± 0.0125)	312.6671
Purely Model-Based	0.4646 (± 0.0233)	0.4868 (± 0.0130)	0.4868 (± 0.0130)	3154.4558
FSL-MRS	*0.5018 (± 0.0242)*	*0.5256 (± 0.0138)*	*0.5256 (± 0.0138)*	475.4806
LCModel	*0.4296 (± 0.0200)*	*0.4534 (± 0.0117)*	*0.4534 (± 0.0117)*	274.6317
Supervised (CNN)	0.6885 (± 0.0322)	0.7062 (± 0.0182)	0.8248 (± 0.0217)	0.7374
Self-Supervised (CNN)	0.8069 (± 0.0436)	0.8180 (± 0.0242)	0.5382 (± 0.0137)	0.1551
TT Instance Adaptive (CNN)	0.5977 (± 0.0233)	0.5977 (± 0.0130)	0.7218 (± 0.0168)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.4798 (± 0.0219)	0.5139 (± 0.0125)	0.4984 (± 0.0129)	270.3693
TT Instance Adaptive (Self-Sup. Init.)	**0.4250 (± 0.0183)**	**0.4605 (± 0.0109)**	0.5286 (± 0.0132)	205.7727
TT Instance Adaptive (From Scratch Init.)	0.4541 (± 0.0218)	0.4790 (± 0.0124)	**0.4768 (± 0.0124)**	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.5103 (± 0.0238)	0.5410 (± 0.0142)	0.6028 (± 0.0160)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.5091 (± 0.0244)	0.5301 (± 0.0138)	0.5094 (± 0.0135)	312.5201
TT Instance Adaptive (10 Iter.)	0.6057 (± 0.0285)	0.6258 (± 0.0162)	0.7280 (± 0.0212)	44.8087
TT Instance Adaptive (50 Iter.)	0.5806 (± 0.0250)	0.5671 (± 0.0133)	0.6576 (± 0.0154)	212.2983
TT Instance Adaptive (100 Iter.)	0.5328 (± 0.0251)	0.5315 (± 0.0135)	0.6031 (± 0.0153)	414.0273
TT Instance Adaptive (500 Iter.)	0.4819 (± 0.0238)	0.5045 (± 0.0133)	0.5380 (± 0.0141)	2061.0019

TABLE A6 Comparison of quantification methods on in-vivo data using pseudo ground truth: **Mean of FSL-MRS and LCModel**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	**MOSAE ↓ (± SE)**			**Time ↓ (ms/sample)**
	ID (Mid-Range)	**OoD (Full-Range)**	**ID (Full-Trained)**	
Supervised	0.5115 (± 0.0242)	0.5372 (± 0.0134)	0.6572 (± 0.0193)	**0.1145**
Self-Supervised	0.4936 (± 0.0210)	0.5190 (± 0.0124)	0.6098 (± 0.0163)	0.1347
TT Instance Adaptive	0.5641 (± 0.0258)	0.5506 (± 0.0137)	0.6402 (± 0.0160)	212.2983
TT Online Adaptive	0.4975 (± 0.0235)	0.5156 (± 0.0129)	0.6351 (± 0.0179)	0.3400
TT Domain Adaptive	0.4360 (± 0.0220)	0.4441 (± 0.0122)	0.4566 (± 0.0126)	312.6671
Purely Model-Based	0.4390 (± 0.0230)	0.4641 (± 0.0129)	0.4641 (± 0.0129)	3154.4558
FSL-MRS	*0.4658 (± 0.0235)*	*0.4845 (± 0.0133)*	*0.4845 (± 0.0133)*	475.4806
LCModel	*0.4111 (± 0.0201)*	*0.4336 (± 0.0117)*	*0.4336 (± 0.0117)*	274.6317
Supervised (CNN)	0.5713 (± 0.0240)	0.5647 (± 0.0130)	0.6846 (± 0.0173)	0.7374
Self-Supervised (CNN)	0.7343 (± 0.0360)	0.7452 (± 0.0199)	0.5085 (± 0.0136)	0.1551
TT Instance Adaptive (CNN)	0.5843 (± 0.0241)	0.5798 (± 0.0130)	0.6969 (± 0.0175)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.4553 (± 0.0225)	0.4831 (± 0.0127)	0.4697 (± 0.0132)	270.3693
TT Instance Adaptive (Self-Sup. Init.)	**0.4034 (± 0.0186)**	**0.4335 (± 0.0109)**	0.4945 (± 0.0136)	205.7727
TT Instance Adaptive (From Scratch Init.)	0.4357 (± 0.0222)	0.4565 (± 0.0125)	**0.4542 (± 0.0125)**	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.4495 (± 0.0194)	0.4723 (± 0.0114)	0.5258 (± 0.0135)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.4652 (± 0.0236)	0.4797 (± 0.0132)	0.4601 (± 0.0129)	312.5201
TT Instance Adaptive (10 Iter.)	0.5017 (± 0.0239)	0.5137 (± 0.0129)	0.6444 (± 0.0183)	44.8087
TT Instance Adaptive (50 Iter.)	0.5641 (± 0.0258)	0.5506 (± 0.0137)	0.6402 (± 0.0160)	212.2983
TT Instance Adaptive (100 Iter.)	0.5186 (± 0.0258)	0.5134 (± 0.0139)	0.5831 (± 0.0157)	414.0273
TT Instance Adaptive (500 Iter.)	0.4583 (± 0.0245)	0.4766 (± 0.0136)	0.5091 (± 0.0144)	2061.0019

TABLE A7 Comparison of quantification methods on in-vivo data using pseudo ground truth: **LCModel**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	MAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	0.9662 (± 0.0470)	0.9896 (± 0.0257)	1.1211 (± 0.0290)	**0.1145**
Self-Supervised	0.8086 (± 0.0409)	0.8429 (± 0.0233)	1.0475 (± 0.0270)	0.1347
TT Instance Adaptive	0.8734 (± 0.0383)	0.9130 (± 0.0214)	1.0334 (± 0.0240)	212.2983
TT Online Adaptive	0.9033 (± 0.0417)	0.9356 (± 0.0232)	1.0788 (± 0.0278)	0.3400
TT Domain Adaptive	0.7955 (± 0.0356)	0.8500 (± 0.0205)	0.8558 (± 0.0210)	312.6671
Purely Model-Based	0.7648 (± 0.0360)	0.8562 (± 0.0218)	0.8562 (± 0.0218)	3154.4558
FSL-MRS	0.8368 (± 0.0383)	0.9293 (± 0.0233)	0.9293 (± 0.0233)	475.4806
LCModel	*0.3248 (± 0.0211)*	*0.3405 (± 0.0125)*	*0.3405 (± 0.0125)*	274.6317
Supervised (CNN)	0.8757 (± 0.0414)	0.9077 (± 0.0232)	1.1076 (± 0.0284)	0.7374
Self-Supervised (CNN)	1.0287 (± 0.0492)	1.0251 (± 0.0274)	0.8611 (± 0.0215)	0.1551
TT Instance Adaptive (CNN)	0.8242 (± 0.0339)	0.8470 (± 0.0192)	1.0844 (± 0.0258)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.6809 (± 0.0300)	0.7303 (± 0.0177)	0.8428 (± 0.0209)	270.3693
TT Instance Adaptive (Self-Sup. Init.)	**0.6556 (± 0.0279)**	**0.6977 (± 0.0163)**	0.8760 (± 0.0215)	205.7727
TT Instance Adaptive (From Scratch Init.)	0.7909 (± 0.0358)	0.8580 (± 0.0210)	**0.8553 (± 0.0209)**	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.7032 (± 0.0307)	0.7430 (± 0.0180)	0.9387 (± 0.0229)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.8379 (± 0.0381)	0.8977 (± 0.0220)	0.8702 (± 0.0217)	312.5201
TT Instance Adaptive (10 Iter.)	0.8840 (± 0.0400)	0.9137 (± 0.0220)	1.0636 (± 0.0270)	44.8087
TT Instance Adaptive (50 Iter.)	0.8734 (± 0.0383)	0.9130 (± 0.0214)	1.0334 (± 0.0240)	212.2983
TT Instance Adaptive (100 Iter.)	0.8432 (± 0.0383)	0.8931 (± 0.0217)	0.9775 (± 0.0240)	414.0273
TT Instance Adaptive (500 Iter.)	0.7846 (± 0.0360)	0.8544 (± 0.0213)	0.8819 (± 0.0222)	2061.0019

TABLE A8 Comparison of quantification methods on in-vivo data using pseudo ground truth: **LCModel**. The spectra are filtered to create equivalent scenarios to the simulated test scenarios: **ID (Mid-Range)**, **OoD (Full-Range)**, and **ID (Full-Trained)**.

Method	MOSAE ↓ (± SE)			Time ↓ (ms/sample)
	ID (Mid-Range)	OoD (Full-Range)	ID (Full-Trained)	
Supervised	0.7992 (± 0.0364)	0.8315 (± 0.0202)	1.0168 (± 0.0262)	**0.1145**
Self-Supervised	0.6666 (± 0.0297)	0.7009 (± 0.0172)	0.9477 (± 0.0241)	0.1347
TT Instance Adaptive	0.8577 (± 0.0393)	0.8972 (± 0.0221)	1.0184 (± 0.0249)	212.2983
TT Online Adaptive	0.7959 (± 0.0364)	0.8297 (± 0.0201)	1.0029 (± 0.0252)	0.3400
TT Domain Adaptive	0.7749 (± 0.0360)	0.8302 (± 0.0209)	0.8349 (± 0.0214)	312.6671
Purely Model-Based	0.7426 (± 0.0361)	0.8292 (± 0.0219)	0.8292 (± 0.0219)	3154.4558
FSL-MRS	0.7905 (± 0.0382)	0.8743 (± 0.0230)	0.8743 (± 0.0230)	475.4806
LCModel	*0.3129 (± 0.0211)*	*0.3288 (± 0.0124)*	*0.3288 (± 0.0124)*	274.6317
Supervised (CNN)	0.7817 (± 0.0365)	0.7990 (± 0.0202)	1.0111 (± 0.0261)	0.7374
Self-Supervised (CNN)	0.9606 (± 0.0437)	0.9652 (± 0.0246)	0.8377 (± 0.0219)	0.1551
TT Instance Adaptive (CNN)	0.8104 (± 0.0349)	0.8309 (± 0.0198)	1.0695 (± 0.0266)	256.5153
TT Instance Adaptive (CNN, Self-Sup. Init.)	0.6585 (± 0.0306)	0.7057 (± 0.0182)	**0.8200 (± 0.0214)**	270.3693
TT Instance Adaptive (Self-Sup. Init.)	**0.6330 (± 0.0283)**	**0.6751 (± 0.0167)**	0.8496 (± 0.0221)	205.7727
TT Instance Adaptive (From Scratch Init.)	0.7748 (± 0.0365)	0.8427 (± 0.0214)	0.8398 (± 0.0214)	29970.4098
TT Online Adaptive (Self-Sup. Init.)	0.6548 (± 0.0281)	0.6901 (± 0.0163)	0.8871 (± 0.0220)	0.3331
TT Domain Adaptive (Self-Sup. Init.)	0.8035 (± 0.0382)	0.8621 (± 0.0220)	0.8327 (± 0.0217)	312.5201
TT Instance Adaptive (10 Iter.)	0.8000 (± 0.0369)	0.8357 (± 0.0205)	1.0080 (± 0.0257)	44.8087
TT Instance Adaptive (50 Iter.)	0.8577 (± 0.0393)	0.8972 (± 0.0221)	1.0184 (± 0.0249)	212.2983
TT Instance Adaptive (100 Iter.)	0.8272 (± 0.0393)	0.8764 (± 0.0223)	0.9604 (± 0.0247)	414.0273
TT Instance Adaptive (500 Iter.)	0.7606 (± 0.0370)	0.8307 (± 0.0219)	0.8567 (± 0.0228)	2061.0019

FIGURE A16 Scatter plots with marginal histograms comparing optimally scaled predicted versus pseudo-true (FSL-MRS estimates) concentrations of Glu and GABA across 1,710 in-vivo spectra. Models are evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Data-driven methods include supervised, self-supervised, and test-time instance adaptive approaches, compared with purely model-based fitting. Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE A17 Scatter plots with marginal histograms comparing optimally scaled predicted versus pseudo-true (FSL-MRS estimates) concentrations of Glu and GABA across 1,710 in-vivo spectra. Models are evaluated under two scenarios for the full concentration range: trained on mid-range concentrations (OoD) or trained on the full range (ID). Methods include test-time online adaptive approaches, test-time domain adaptive approaches compared with FSL-MRS (Newton) and LCModel. Points are colored by SNR, and regression lines with corresponding statistics (slope α, intercept β, R^2, and RMSE σ) are shown.

FIGURE A18 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations for NAA, Cr, Glu, GSH, and GABA across 1,710 in-vivo spectra. Models were trained on mid-range concentrations and evaluated across the full concentration range to assess extrapolation performance. This figure shows data-driven methods: supervised, self-supervised, and test-time instance adaptive against purely model-based. Points are colored by SNR, and regression lines with corresponding statistics (R^2, slope, intercept, RMSE) are included.

FIGURE A19 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations for NAA, Cr, Glu, GSH, and GABA across 1,710 in-vivo spectra. Models were trained on mid-range concentrations and evaluated across the full concentration range. This figure shows adaptive and classical methods: test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Points are colored by SNR, and regression lines with corresponding statistics (R^2, slope, intercept, RMSE) are included.

FIGURE A20 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations for NAA, Cr, Glu, GSH, and GABA across 1,710 in-vivo spectra. Models were trained and tested across the full concentration range. This figure shows data-driven methods: supervised, self-supervised, and test-time instance adaptive against purely model-based. Points are colored by SNR, and regression lines with corresponding statistics (R^2, slope, intercept, RMSE) are included.

FIGURE A21 Scatter plots with marginal histograms comparing predicted versus pseudo-true (FSL-MRS estimates) concentrations for NAA, Cr, Glu, GSH, and GABA across 1,710 in-vivo spectra. Models were trained and tested across the full concentration range. This figure shows adaptive and classical methods: test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Points are colored by SNR, and regression lines with corresponding statistics (R^2, slope, intercept, RMSE) are included.

FIGURE A22 Scatter plots showing quantification accuracy (MOSAE) across 1,710 in-vivo spectra as a function of estimated SNR, linewidth, zeroth-order phase shift, and frequency offset. Data-driven methods include supervised, self-supervised, and test-time instance adaptive compared against purely model-based fitting. Each point represents one spectrum, illustrating method-specific sensitivity to core signal parameter variations.

FIGURE A23 Scatter plots showing quantification accuracy (MOSAE) across 1,710 in-vivo spectra under estimated macromolecular baseline (MM), baseline variation, and random signal corruptions. Data-driven methods include supervised, self-supervised, and test-time instance adaptive compared against purely model-based fitting. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations.

FIGURE A24 Scatter plots showing quantification accuracy (MOSAE) across 1,710 in-vivo spectra as a function of estimated SNR, linewidth, zeroth-order phase shift, and frequency offset. Adaptive and classical methods include test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Each point represents one spectrum, illustrating method-specific sensitivity to core signal parameter variations.

FIGURE A25 Scatter plots showing quantification accuracy (MOSAE) across 1,710 in-vivo spectra under macromolecular baseline (MM), baseline variation, and random signal corruptions. Adaptive and classical methods include test-time online adaptive, test-time domain adaptive, FSL-MRS (Newton), and LCModel. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations.

FIGURE A26 Summary of quantification performance across 1,710 in-vivo spectra for eight metabolites (Ala, Asc, Asp, Cr, GABA, Gln, Glu, Gly). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under in-vivo conditions.

FIGURE A27 Summary of quantification performance across 1,710 in-vivo spectra for eight metabolites (GPC, GSH, mIns, Lac, NAAG, NAA, PCh, PCr). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under in-vivo conditions.

FIGURE A28 Summary of quantification performance across 1,710 in-vivo spectra for four metabolites (PE, Scyllo, Ser, Tau). Each subplot corresponds to one metabolite, showing the MOSAE for all methods: purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. This visualization allows comparison of method performance across metabolites under in-vivo conditions.

FIGURE A29 Summary of quantification performance across 1,710 in-vivo spectra. For each signal parameter, SNR, linewidth, frequency offset, phase shift, MM baseline, and polynomial baseline, the MOSAE is averaged within parameter bins. All methods are overlaid in each subplot, including purely model-based gradient descent, FSL-MRS (Newton), LCModel, supervised, self-supervised, and TTA strategies. Each point represents one spectrum, illustrating method-specific robustness to unmodeled spectral deviations in real in-vivo data.

TABLE B9 Overview of the metabolite concentration ranges [mM] taken from De Graaf 2019[41].

Parameter	Notation	Range	Parameter	Notation	Range
Alanine (Ala)	a_1	$\mathcal{U}[0.1, 1.6]$	Lactate (Lac)	a_{12}	$\mathcal{U}[0.2, 1.0]$
Ascorbate (Asc)	a_2	$\mathcal{U}[0.4, 1.7]$	N-Acetylaspartylglutamate (NAAG)	a_{13}	$\mathcal{U}[0.5, 2.5]$
Aspartate (Asp)	a_3	$\mathcal{U}[1.0, 2.0]$	N-Acetylaspartate (NAA)	a_{14}	$\mathcal{U}[7.5, 12.0]$
Creatine (Cr)	a_4	$\mathcal{U}[4.5, 10.5]$	Phosphocholine (PCh)	a_{15}	$\mathcal{U}[0.2, 1.0]$
Gamma-Aminobutyric Acid (GABA)	a_5	$\mathcal{U}[1.0, 2.0]$	Phosphocreatine (PCr)	a_{16}	$\mathcal{U}[3.0, 5.5]$
Glutamine (Gln)	a_6	$\mathcal{U}[3.0, 6.0]$	Phosphoethanolamine (PE)	a_{17}	$\mathcal{U}[1.0, 2.0]$
Glutamate (Glu)	a_7	$\mathcal{U}[6.0, 12.5]$	Scyllo-Inositol (Scyllo)	a_{18}	$\mathcal{U}[0.2, 0.5]$
Glycine (Gly)	a_8	$\mathcal{U}[0.2, 1.0]$	Serine (Ser)	a_{18}	$\mathcal{U}[0.2, 2.0]$
Glycerophosphocholine (GPC)	a_9	$\mathcal{U}[0.4, 1.7]$	Taurine (Tau)	a_{20}	$\mathcal{U}[3.0, 6.0]$
Glutathione (GSH)	a_{10}	$\mathcal{U}[1.7, 3.0]$	Macromolecules (MMs)*	a_{21}	$\mathcal{U}[0.0, 400.0]$
Myo-Inositol (mIns)	a_{11}	$\mathcal{U}[4.0, 9.0]$			

*Not taken from De Graaf 2019.

TABLE B10 The metabolite concentration ranges [mM] obtained by fitting all in-vivo spectra of Section 2.2 using LCModel[9].

Parameter	Notation	Range	Parameter	Notation	Range
Alanine (Ala)	a_1	$\mathcal{U}[0.0, 1.0]$	Lactate (Lac)	a_{12}	$\mathcal{U}[0.0, 2.8]$
Ascorbate (Asc)	a_2	$\mathcal{U}[0.0, 3.7]$	N-Acetylaspartylglutamate (NAAG)	a_{13}	$\mathcal{U}[0.2, 2.0]$
Aspartate (Asp)	a_3	$\mathcal{U}[0.8, 4.0]$	N-Acetylaspartate (NAA)	a_{14}	$\mathcal{U}[9.8, 14.0]$
Creatine (Cr)	a_4	$\mathcal{U}[3.9, 8.0]$	Phosphocholine (PCh)	a_{15}	$\mathcal{U}[0.0, 2.4]$
Gamma-Aminobutyric Acid (GABA)	a_5	$\mathcal{U}[0.0, 4.0]$	Phosphocreatine (PCr)	a_{16}	$\mathcal{U}[0.9, 5.2]$
Glutamine (Gln)	a_6	$\mathcal{U}[1.9, 6.8]$	Phosphoethanolamine (PE)	a_{17}	$\mathcal{U}[0.0, 5.2]$
Glutamate (Glu)	a_7	$\mathcal{U}[11.7, 16.9]$	Scyllo-Inositol (Scyllo)	a_{18}	$\mathcal{U}[0.0, 0.4]$
Glycine (Gly)	a_8	$\mathcal{U}[0.0, 0.5]$	Serine (Ser)	a_{18}	$\mathcal{U}[0.0, 3.6]$
Glycerophosphocholine (GPC)	a_9	$\mathcal{U}[0.0, 3.2]$	Taurine (Tau)	a_{20}	$\mathcal{U}[1.2, 3.1]$
Glutathione (GSH)	a_{10}	$\mathcal{U}[0.0, 1.7]$	Macromolecules (MMs)	a_{21}	$\mathcal{U}[175.7, 318.9]$
Myo-Inositol (mIns)	a_{11}	$\mathcal{U}[5.9, 10.0]$			

TABLE B11 The metabolite concentration ranges [mM] obtained by fitting all in-vivo spectra of Section 2.2 using FSL-MRS[13].

Parameter	Notation	Range	Parameter	Notation	Range
Alanine (Ala)	a_1	$\mathcal{U}[0.0, 0.5]$	Lactate (Lac)	a_{12}	$\mathcal{U}[0.1, 3.1]$
Ascorbate (Asc)	a_2	$\mathcal{U}[0.9, 4.9]$	N-Acetylaspartylglutamate (NAAG)	a_{13}	$\mathcal{U}[0.0, 2.3]$
Aspartate (Asp)	a_3	$\mathcal{U}[0.0, 4.8]$	N-Acetylaspartate (NAA)	a_{14}	$\mathcal{U}[9.7, 16.3]$
Creatine (Cr)	a_4	$\mathcal{U}[5.0, 12.3]$	Phosphocholine (PCh)	a_{15}	$\mathcal{U}[0.0, 2.2]$
Gamma-Aminobutyric Acid (GABA)	a_5	$\mathcal{U}[0.0, 2.2]$	Phosphocreatine (PCr)	a_{16}	$\mathcal{U}[0.0, 5.1]$
Glutamine (Gln)	a_6	$\mathcal{U}[0.0, 3.6]$	Phosphoethanolamine (PE)	a_{17}	$\mathcal{U}[0.0, 3.5]$
Glutamate (Glu)	a_7	$\mathcal{U}[11.1, 17.9]$	Scyllo-Inositol (Scyllo)	a_{18}	$\mathcal{U}[0.0, 0.6]$
Glycine (Gly)	a_8	$\mathcal{U}[0.0, 0.2]$	Serine (Ser)	a_{18}	$\mathcal{U}[0.0, 7.3]$
Glycerophosphocholine (GPC)	a_9	$\mathcal{U}[0.2, 3.6]$	Taurine (Tau)	a_{20}	$\mathcal{U}[1.5, 3.4]$
Glutathione (GSH)	a_{10}	$\mathcal{U}[1.0, 3.6]$	Macromolecules (MMs)	a_{21}	$\mathcal{U}[133.8, 242.1]$
Myo-Inositol (mIns)	a_{11}	$\mathcal{U}[6.5, 12.1]$			

TABLE B12 Layer-wise architecture of the MLP and CNN models, implemented in PyTorch. Convolutional parameters shown as (kernel, stride, padding). ELU activations follow each hidden layer. Parameter counts: MLP = 532K, CNN = 3.0M.

Arch	Layer	Operation	Params (k,s,p)	Input Dim	Output Dim
MLP	Input	BatchNorm1d	-	355	355
	Flatten	Flatten	-	2×355	-
	FC0	Linear + ELU	-	710	512
	FC1	Linear + ELU	-	512	256
	FC2	Linear + ELU	-	256	128
	Output*	Linear	-	128	32
CNN	Input	BatchNorm1d	-	355	355
	Conv0	Conv1d + ELU	(3, 1, 0)	4	8
	Conv1	Conv1d + ELU	(3, 1, 0)	8	16
	Conv2	Conv1d + ELU	(3, 1, 0)	16	32
	Flatten	Flatten	-	-	-
	FC0	Linear + ELU	-	flattened	512
	FC1	Linear + ELU	-	512	256
	FC2	Linear + ELU	-	256	128
	Output*	Linear	-	128	32

***Output vector uses component-wise activations: `softplus` for metabolite amplitudes and linewidths, with a `+1` offset for Gaussian and Lorentzian broadening to ensure values > 1. First-order phase uses a scaled `tanh` activation ($\tanh(x) \times 10^{-4}$) to keep values in the stable range $\mathcal{U}[-10^{-5}, 10^{-5}]$. Remaining parameters (zeroth-order phase, frequency shift, and baseline) are linear.

TABLE B13 Overview of the configuration and training parameters for the MLP and CNN models.

Parameter	Value	Description
Data Settings		
dataType	aumc2_ms	Dataset used for training and evaluation.
basisFmt	7tslaser	Format of the MRS basis set.
path2basis	.../7T_sLASER_OIT_TE34.basis	Path to the basis set.
specType	auto	Automatically selects ppm region.
ppmlim	(0.5, 4.0)	ppm limits of the spectra.
test_size	10000	Number of test samples.
Architecture Settings		
arch	mlp / cnn	Architecture type: MLP or CNN.
activation	elu	Nonlinearity used in hidden layers.
dropout	0.0	Dropout probability.
width	512	Width of first fully connected layer.
depth	3	Number of FC layers (after input / conv layers).
conv_depth	3	Number of Conv1D layers (CNN only).
kernel_size	3	Kernel size for Conv1D layers.
stride	1	Stride for Conv1D layers.
Optimization		
loss	mse_specs / mae_all_scale	Loss function for training.
optimizer	Adam[46]	Optimizer used for training.
batch	16	The batch size.
trueBatch	16	Accumulates the gradients over trueBatch/batch.
check_val_every_n_epoch	None	None, if trained with generator, otherwise the number of epochs between validations.
learning	0.0001	Learning rate.
max_epochs	-1	Maximum number of epochs.
max_steps	-1	Maximum number of steps/iterations.
val_check_interval	256	The number of iterations per between validations.
val_size	1,024	Validation size (in samples).
Adaptation / Inner Loop		
adaptMode	per_spec_adapt	Mode of inner-loop adaptation (instance/online/domain adaptation).
innerEpochs	50	Number of adaptation epochs.
innerBatch	1	Batch size for inner loop.
innerLr	1e-4	Learning rate for inner loop.
innerLoss	mse_specs	Loss function for inner loop.
bnState	train	BatchNorm mode in inner loop.

TABLE C14 MRSinMRS for the data of Section 3.2.

Site (name or number)	Amsterdam UMC
1. Hardware	
a. Field strength [T]	7 T (298030131 MHz)
b. Manufacturer	Philips
c. Model (software version if available)	5.1.7; .1.7;
d. RF coils: nuclei (transmit/receive), number of channels, type, body part	1H, 32 channel, head coil
e. Additional hardware	-
2. Acquisition	
a. Pulse sequence	Semi-LASER
b. Volume of interest (VOI) locations	Anterior cingulate cortex
c. Nominal VOI size [cm^3, mm^3]	$25 \times 18 \times 18$ mm^3
d. Echo time (TE) / repetition time (TR) [ms, s]	36 ms / 5000 ms
e. Total number of excitations or acquisitions per spectrum	64 averages
f. Additional sequence parameters	3000 Hz bandwidth, 1024 sample points,
g. Water suppression method	VAPOR
h. Shimming method, reference peak, and thresholds for "acceptance of shim" chosen	HOS-DLT[43]
i. Triggering or motion correction method (respiratory, peripheral, cardiac triggering)	-
3. Data Analysis Methods and Outputs	
a. Analysis software	In-house Python scripts, FSL-MRS[13] (version 2.1.20), LCModel[9] (version 6.3-1L)
b. Processing steps (deviating from quoted reference or product)	NIfTI-MRS Header (ProcessingApplied): Method: "Custom coil combination (adaptive)", Details: *own_nifti_coil_combination_adaptive, data, reference* *fsl_mrs_preproc* *–data {save_path}/{item}/block{i + 1}/metab.nii.gz* *–reference {save_path}/{item}/block{i + 1}/wref.nii.gz* *–output {save_path}/{item}/block{i + 1}/{sub_folder}* *–hlsvd –conjugate –overwrite –report*
c. Output measure (e.g. absolute concentration, institutional units, ratio)	Absolute concentrations [mM]
d. Quantification references and assumptions, fitting model assumptions	7T Semi-LASER OIT basis set with TE 34ms (metabolite list seen in Table 1, MM[54]) LCModel control: *$LCMODL, nunfil=1024, deltat=3.333e-04,* *hzpppm=hzpppm=2.9803e+02, ppmst=4.0,* *ppmend=0.5, dows=T, doecc=F, neach=50,* *filbas='example.basis', filraw='example.raw',* *filh2o='example.h2o', filps='example.ps',* *filcoo='example.coord', filtab='example.table',* *ltable=7, lcoord=9, lps=8, nsimul=0, echot=36,* *dkntmn=0.5, nuse1=3, chcomb(1)='Glu+Gln',* *hcomb(2)='Cr+PCr', chcomb(3)='NAA+NAAG',* *chcomb(4)='GPC+PCh'atth2o=0.7, wconc=59297,* *$END* FSL-MRS: *from fsl_mrs.utils import fitting* *fitting.fit_FSLModel, method='Newton' ppmlim=(0.5, 4.0),* *baseline_order=2*
4. Data Quality	
a. Reported variables (SNR, linewidth)	S/N = 20.0 - 54.0, FWHM = 0.029 - 0.079 ppm (LCModel estimates)
b. Data exclusion criteria	4 participants excluded based on visual inspection
c. Quality measures of postprocessing model fitting	-
d. Sample spectra (and mean)	