



RESEARCH ARTICLE

Magnetic Resonance in Medicine

Denoising single MR spectra by deep learning: Miracle or mirage?

Martyna Dziadosz^{1,2,3} | Rudy Rizzo^{1,2,3}  | Sreenath P. Kyathanahally⁴  | Roland Kreis^{1,2} 

¹MR Methodology, Department for Diagnostic and Interventional Neuroradiology, University of Bern, Bern, Switzerland

²Translational Imaging Center (TIC), Swiss Institute for Translational and Entrepreneurial Medicine, Bern, Switzerland

³Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland

⁴Department System Analysis, Integrated Assessment and Modelling, Eawag - Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Correspondence

Roland Kreis, MR Methodology, University Bern, Freiburgstr. 3, CH-3010 Bern, Switzerland.

Email: roland.kreis@insel.ch

Funding information

H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 813120; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung; European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement, Grant/Award Number: 813120 (inspire-med); Swiss National Science Foundation, Grant/Award Number: 320030-175984

Purpose: The inherently poor SNR of MRS measurements presents a significant hurdle to its clinical application. Denoising by machine or deep learning (DL) was proposed as a remedy. It is investigated whether such denoising leads to lower estimate uncertainties or whether it essentially reduces noise in signal-free areas only.

Methods: Noise removal based on supervised DL with U-nets was implemented using simulated ¹H MR spectra of human brain in two approaches: (1) via time-frequency domain spectrograms and (2) using 1D spectra as input. Quality of denoising was evaluated in three ways: (1) by an adapted fit quality score, (2) by traditional model fitting, and (3) by quantification via neural networks.

Results: Visually appealing spectra were obtained; hinting that denoising is well-suited for MRS. However, an adapted denoising score showed that noise removal is inhomogeneous and more efficient for signal-free areas. This was confirmed by quantitative analysis of traditional fit results as well as DL quantification following DL denoising. DL denoising, although apparently successful as judged by mean squared errors, led to substantially biased estimates in both implementations.

Conclusion: The implemented DL-based denoising techniques may be useful for display purposes, but do not help quantitative evaluations, confirming expectations based on estimation theory: Cramér Rao lower bounds defined by the original data and the appropriate fitting model cannot be circumvented in an unbiased way for single data sets, unless additional prior knowledge can be incurred in the form of parameter restrictions/relations or applicable substates.

KEYWORDS

deep learning, denoising, machine learning, MR spectroscopy, parameter estimation, quantification

1 | INTRODUCTION

Magnetic resonance spectroscopy (MRS) is a non-invasive technique for identifying and quantifying metabolites in

vivo. Its primary limitation is probably the low SNR^{1–3} achievable in clinically relevant resolution and acquisition times. The low SNR limits the number of quantifiable metabolites and measurement precision, especially

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Magnetic Resonance in Medicine* published by Wiley Periodicals LLC on behalf of International Society for Magnetic Resonance in Medicine.

for low-concentration metabolites, such as lactate or γ -aminobutyric acid (GABA). To ameliorate the situation in general or specific circumstances, numerous denoising approaches have been suggested to enhance SNR in MRS: wavelet feature analysis,^{4–6} apodization,⁷ principal component analysis,^{8–10} signal decomposition,^{6,11–15} or deep-learning (DL).¹⁶

However, it is questionable¹⁷ under which conditions denoising can really improve estimation uncertainty or whether it may mostly just serve as a cosmetic technique to ameliorate appearance by reducing noise in signal-free areas.¹⁸ Denoising techniques based on spatial-spectral (or dynamic-spectral) separability have proven to minimize uncertainty for fitting metabolite content in scenarios with multiple spectra,^{8,9,11–13,19} where common features may be distinct from random noise. However, for single spectra, reduction of noise would directly translate into lower uncertainties in subsequent traditional modeling. This would lead to a contradiction with basic estimation theory,²⁰ which states (1) that all unbiased estimators are limited by the Cramér-Rao-lower-bounds (CRLB); and (2) that, in the case of a spectrum with random uncorrelated noise, ordinary least squares estimation constitutes an efficient unbiased estimator (i.e., one that can attain the CRLB limit).^{17,20}

To test and demonstrate applicability of the above, denoising by DL was implemented in two novel fashions, first using a time-frequency representation of the spectroscopy data, where signal and noise are spread out over two dimensions and a neural network might, therefore, distinguish signal from noise more effectively than in a single domain, and second, using a frequency-domain representation of the data with real and imaginary channel. It was then investigated (1) whether denoising acts uniformly in the spectrum or depends on the amount of local signal, (2) whether denoising introduces bias for subsequent traditional modeling, and (3) whether the benefit of denoising might show better if quantification is performed by a DL algorithm rather than traditional modeling, which is usually valid for white Gaussian noise only.

2 | METHODS

2.1 | Data preparation

Spectra mimicking human brain were synthesized in VESPA²¹ for a semi-LASER sequence ($TE = 35$ ms, $B_0 = 3$ T).^{22,23} Basis datasets were composed of 16 metabolites: aspartate (Asp), GABA, glucose (Glc), glutamine (Gln), glutamate (Glu), glycine (Gly), glutathione (GSH), lactate (Lac), myo-inositol (mI),

NAA, N-acetylaspartylglutamate (NAAG), phosphoethanolamine (PE), scyllo-inositol (sI), taurine (Tau), total choline (tCho) (1:1-mixture of glycerophosphorylcholine + phosphorylcholine), total creatine (tCr) (1:1-mixture of creatine + phosphocreatine). To account for substantial variations in metabolite content in pathological conditions, metabolite concentrations ranged from 0 to double that seen in healthy brain²⁴ (except for tCho up to five-times normal). A reference signal, representing a downsampled non-suppressed water signal supposedly obtained from a separate acquisition, was added at 0.5 ppm to facilitate quantification.²⁵ The following parameters were varied to mimic in vivo conditions: shim with Gaussian line-broadening of 2–5 Hz added to the inherent line broadening of 1.1–3.9 Hz because of assumed metabolite T_2 s^{26,27} (tCr [CH_2]: 111 ms, tCr [CH_3]: 169 ms, NAA [CH_3]: 289 ms, all other protons: 185 ms), overall SNR of the spectrum of 5 to 40 (termed global SNR and defined in time-domain as absolute signal intensity at time 0 versus the standard deviation of the noise), and the intensity of the macro-molecular background signal²⁷ at $\pm 33\%$ of the norm. A metabolite-specific SNR (metabolite-SNR) was defined and used to display and categorize results. It was defined as: global SNR times each metabolite's ground truth (GT) concentration times the number of relevant protons (chemical shift between 0 and 4.1 ppm). Use of these two measures of SNR allows to display the outcome of denoising as function of the generally used overall SNR reflecting total signal power as well as the individual signal of a metabolite per spectrum, needed because of the large concentration range covered for each metabolite.

In a first denoising algorithm (referred to as 2D-UNet) a time-frequency representation (spectrogram), which has proven beneficial for audio signal processing as well as for quantification of MR spectra,²⁵ was selected as raw data form for denoising. In this approach, synthesized time-domain signals were transformed into spectrograms using a short-time Fourier transformation (“stft” in MATLAB) with window size 128 and overlap interval of 97 points converting time-domain signals of 4096 points into 128×128 spectrograms. For the second investigated denoising method, termed 1D-UNet, 1D spectra were used as input after truncation to 1024 complex points covering a range of 8.11 ppm.

2.2 | Denoising methods

Convolutional Autoencoders are commonly used to denoise images.²⁸ However, they have a problem with degradation (increasing the complexity of a network reduces performance on test and training data). A

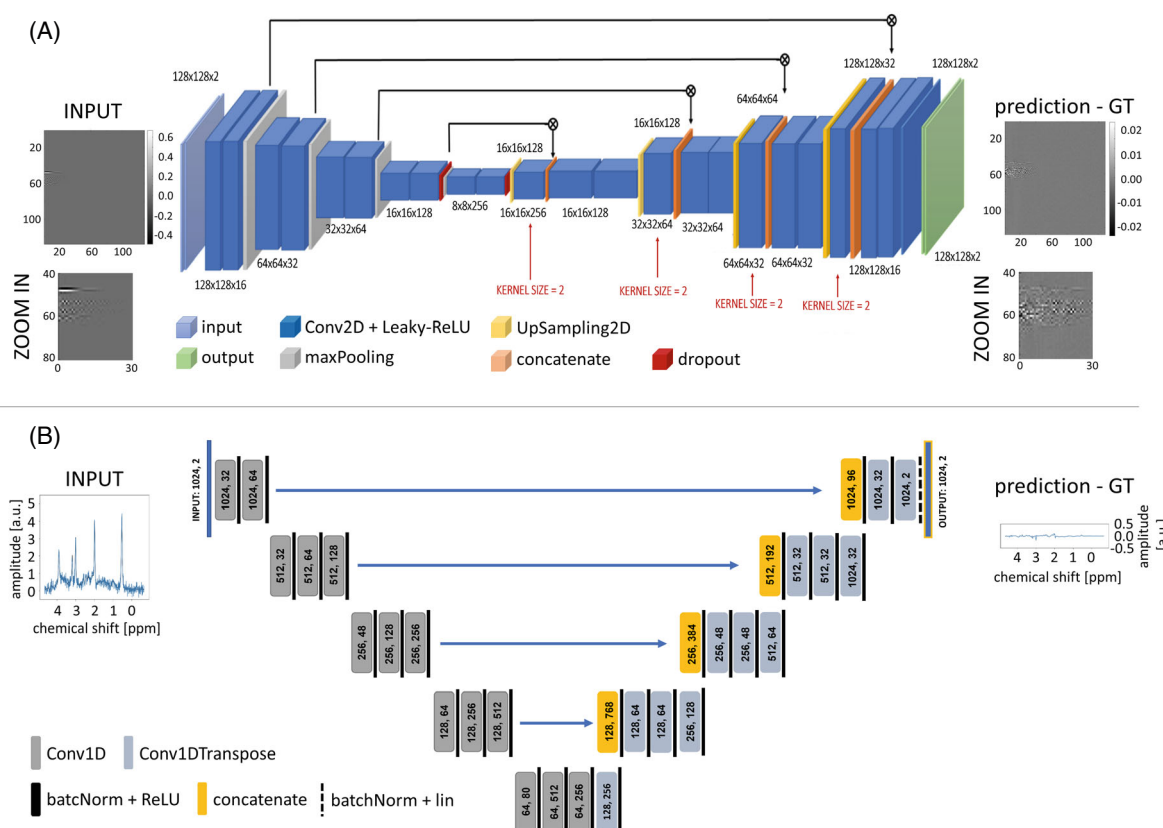


FIGURE 1 Illustration of the DL denoising networks used and illustrative sample data. (A) A sketch of the 2D-UNet with relevant parameters is indicated. The tensor size is indicated as $A \times B \times C$ where $A \times B$ is the layer size and C is the number of features. Sample input data is shown on the left of the network (absorption channel of the 2D time-frequency representation i.e., two-channel spectrograms, of 128×128 matrix size). MRS relevant signal (i.e., 0–4.5 ppm) is evident in the 2D representation on the y-axis (frequency domain) between bins 40 and 80 and on the x-axis (time domain) up to bin 30, where the FID evolution is visible (i.e., ZOOM IN panel). On the right, the corresponding residue spectrogram (prediction–ground truth [GT]) is presented. The effect of denoising yields bigger residual signal in amplitude in correspondent MRS signal areas (evident in ZOOM IN panel). Data were scaled by the maximum of the absolute values of the spectrogram to stay in the range $[-1, 1]$, keeping 0 in place. Size of training/test/validation sets: 18 000/1000/1000. The network implements dropout 0.6 and kernel size = 5. It was trained for 30 epochs with early stopping (patience = 3), batch size = 128 and mean square error as a loss. Adam optimizer with an initial learning rate of 2×10^{-4} was deployed. (B) A sketch of the 1D-UNet with relevant parameters is indicated, supplemented by a sample input spectrum (absorption part only) on the left and a resulting residue spectrum on the right. The tensor size is indicated as A, C where A is the layer size and C is the number of features. Size of training/test/validation sets: 18 000/1000/1000. The network implements kernel size = 2. It was trained for 200 epochs with early stopping (patience = 10), batch size = 50 and mean square error as a loss. Adam optimizer with an initial learning rate of 2×10^{-4} was deployed.

well-known solution is to use skip connections; therefore, Convolutional Autoencoders with symmetric skip connections were used for denoising in this work. Such architectures are also referred to as U-Nets.

For 2D-denoising, a network derived from 2D audio-signal processing,²⁹ where noise was predicted from magnitude and phase channels, was adapted. Our 2D-UNet network³⁰ implementation used real and imaginary channels to map the original spectrograms into apparently noise-free representations. The detailed architecture is depicted in Figure 1A. The network consists of an encoder and a decoder with skip connections: the encoder comprises nine 2D convolutional layers

(with Leaky-rectified linear unit [ReLU], maxpooling, and dropout layers), whereas the decoder symmetrically expands the data with information on skip connections (with Leaky-ReLU, up-sampling, and concatenating layers). Input and output data were scaled into distributions between -1 and $+1$, keeping 0 in place. Leaky-ReLU is the final activation function to yield an output ranging from -1 to $+1$. The model was assembled with the Adam-optimizer using the Huber loss function and taking the mean-square-error as metric. The dataset was composed of 20 000 two-channel spectrograms split into 80%/10%/10% (training/testing/validation). The algorithm was trained iteratively five times, and the

result with the smallest loss was chosen. The resulting spectrograms were transformed back into 1D-spectra in MATLAB.

A second denoising algorithm called 1D-UNet was also tested. It uses plain 1D spectra as input, forecasting noise-free spectra from the original noisy real and imaginary channels. Figure 1B shows the detailed architecture. Like the 2D-UNet, the 1D-UNet consists of an encoder and a decoder with skip connections. Here, the encoder is composed of fourteen 1D convolutional layers (with batch normalization and ReLU activation), whereas the decoder symmetrically expands the stream retrieving high-level features via skip connections on the encoder side. The model was assembled with the Adam-optimizer using the mean-squared-error as loss function and deploying an early-stopping criterion monitoring the minimization of validation loss with patience of 10 iterations. A total of 20 000 two-channel spectra were divided into 80%/10%/10% (training/testing/validation). The output with the smallest loss was selected after the algorithm underwent five repeated training cycles.

The Keras Tuner and Tensorflow framework (Tensorflow 2.3.0; Google) were used to tune network parameters on three graphics processing units (NVIDIA Titan Xp, NVIDIA Titan RTX, NVIDIA GeForce RT 2080 Ti).

2.3 | Denoising efficiency

The effectiveness of denoising was analyzed in two ways: first, by adapted denoising scores inspired by fit the quality score used in model fitting (MF), and second by analyzing variance of estimates in traditional MF, as well as variance in predictions from DL quantification. The denoising effect was evaluated as a denoising score (DS) calculated as mean absolute deviation from GT relative to the mean absolute deviation from zero of pure noise, similar to the definition of fit quality.³¹ DS was determined in three versions: first without weights for the whole spectral range (20 ppm in 2D-UNet; 8.1 ppm in 1D-UNet) and second focusing on the main area-of-interest (0.0–4.5 ppm). In addition, weighting with the GT spectrum automatically yields a score (DS_w) that is sensitive to the most important signal intensities and crucial spectral ranges (although calculated over the full spectral widths):

$$DS = \sqrt{\frac{\sum_i (I_i^{DN} - I_i^{GT})^2}{\sum_i (I_i^N)^2}};$$

$$DS_w = \sqrt{\frac{\sum_i \omega_i (I_i^{DN} - I_i^{GT})^2}{\sum_i \omega_i (I_i^N)^2}}; \quad \omega_i = |I_i^{GT}|$$

with I_i^{GT} , I_i^{DN} , and I_i^N the intensities of GT, the denoised and noise-only spectra as a function of frequency index i , running over the concatenated real and imaginary channels of the spectra, and ω_i the weights for the weighted score (DS_w) as calculated from the absolute intensities of the GT spectrum. The inverse of DS can be regarded as a denoising factor, that is, offering an indicator of how much the noise amplitude is reduced in size after denoising (although without guarantee that the new or remaining “noise” keeps its white Gaussian characteristic).

2.4 | Spectral modeling

FitAID³² was used for traditional MF. All metabolite and macromolecule spectra used for data synthesis plus a simple line for the reference peak were used as base spectra and the model was constrained with equal phase, frequency shift and Lorentzian broadening for all components. CRLB were only calculated and interpreted for the fits of the original noisy spectra. CRLB cannot be determined readily after denoising because the noise will not necessarily be white Gaussian and the model may not be correct either.

In case of the 2D-UNet, a complementary quantification was performed using a neural network based on an optimized shallow convolutional neural network (CNN) architecture and zoomed spectrograms as inputs.²⁵ While the initial network parameters had been optimized for similar noisy spectra, for the current context the network was retrained both, for the original and the denoised spectrograms.

Because of the reduced spectral range after 1D-UNet denoising the calculation of equivalent spectrograms needed for the input to the above DL quantification network was not possible. Therefore, a different DL quantification network was used. To complement 1D denoising, a 1D quantification network was chosen. A modified InceptionNet-1D from Rizzo et al.²⁵ was selected for straight metabolite quantification because it performed best from all networks based on 1D spectral input tested in Rizzo et al.²⁵ The modifications accounted only for input–output dimensionality, matching the number of data points here considered and where real and imaginary spectral components are now supplied to the network in two separate channels. In both cases, quantification relied on the ratio of metabolite-to-reference peak area estimates as described before.²⁵

3 | RESULTS

Both suggested denoising methods provide—at least visually—excellent results, as demonstrated for a few cases

in Figure 2A. The denoised spectra closely match GT in appearance, and the extent of denoising is reflected in the denoising score DS (Figure 2B), suggesting a dramatic decrease in the noise floor when considering the whole spectrum (DS_{full} reduced to 10% or with a denoising factor of 10 at low SNR). At higher SNR, denoising seems more successful for the 1D U-net with a threefold reduction in noise throughout. If restricting the determination of denoising quality to the spectral range of interest ($DS_{45\text{ ppm}}$), the DS factor becomes larger (i.e., apparently less efficient denoising)—in particular for the 2D U-net, because the original spectral input range is four times larger compared to the 1D-UNet case. The denoising-score weighted with GT intensities (DS_w) is somewhat higher, but similar to the score calculated over the actual spectral range of interest. DS_w tends toward one on average for the 2D U-net at higher SNR suggesting that denoising is ineffective where there is signal and the apparent excellent denoising performance in the original denoising score stems from areas that lack MR signals. For the 1D U-net, DS_w remains substantially below 1 also at high SNR, promising substantial (approximately 2- to 3-fold) removal of noise also in areas with strong MR signals.

Figure 3 shows metabolite-specific results of quantification for both original and denoised data using traditional least-squares MF for both denoising methods and three metabolites of characteristically low, medium, and high metabolite-SNR: sI, NAAG and NAA (results for further metabolites are provided in Figure S1). Each subplot represents the relationship between GT and the estimated metabolite content for one metabolite. To aid visual detection of systematic patterns, metabolites have been arranged in an approximately ascending order of metabolite-SNR. As complementary information, the marginal distributions for the estimated (horizontal) and GT (vertical) metabolite contents were plotted outside the correlation plot. The metabolite-SNR value was generally used to group the metabolites into three groups: low, medium, and high SNR metabolites. Hence, for each metabolite, a comparison between the correlation plot of estimation results for original noisy and denoised spectra visualizes the gross effects of denoising.

(1) For all metabolites, particularly at low SNR, there are much fewer outliers with drastically overestimated metabolite content if denoised input data is used for MF quantification.

(2) While the slope of the linear correlation line between GT and estimated values is close to one for the noisy data, it deviates strongly for the denoised spectra hinting at a systematic bias at low and high concentrations.

(3) In accordance with the latter finding, the marginal distributions for the denoised cases show a higher density of cases at the center of the tested concentration range, which is not seen for traditional modeling and is reminiscent of quantitation of low SNR spectra by CNNs.^{25,33} For MF, a deviation from a uniform distribution is apparent at very low metabolite content, which is the effect of bounding the concentrations to positive values.³⁴

(4) Quantitative aspects of estimation success can be drawn from the numbers in the upper right corner representing the slope and offset of the correlation line (a , q), the square of the Pearson correlation constant (R^2), and the root mean square error (σ) over the whole test set. Although σ may remain in size with and without denoising (or even become smaller after denoising for single metabolites), a smaller slope a of the correlation line for almost all cases indicates that bias is introduced with denoising and that the small σ values are often achieved through bias toward the mean training concentration for low SNR data.

Figure 4 displays the results of denoising followed by quantification by DL in an equivalent plot, again for both denoising schemes and the same three metabolites as quantified by the respective 1D or 2D DL quantification schemes (results for further metabolites presented in Figure S2.) Essentially, the same effects for denoising are found with DL as for MF, except that DL quantification of the original noisy spectra already shows the main artifacts of denoising found above with traditional modeling for the denoised cases: bias and bell-shaped distributions of estimated results are already present without denoising. Slopes and mean square errors from denoised input indicate that quantification by a DL network may be somewhat better at low SNR than when using traditional MF, but the outcome is clearly inferior to MF of the original noisy spectra. The relative performance of the two DL quantification methods can best be compared by the plots of DL results of the (identical) noisy data. They show the above effect to a different extent and this is also somewhat metabolite-specific.

Figure 5 contains plots of deviations of estimates from GT as a function of overall SNR for the 2D scheme. It shows the scattering of estimates in relation to the CRLB from traditional modeling (yellow, without denoising) and it visualizes bias inflicted by denoising with subsequent quantification by traditional MF. An equivalent plot as function of metabolite SNR is presented in Figure S3. In addition, it becomes apparent that the noise-dependence of fitting precision expected for MF (decreasing CRLB with increasing SNR) is almost absent if spectra are denoised before quantification (most evident for low and intermediate SNR metabolites).

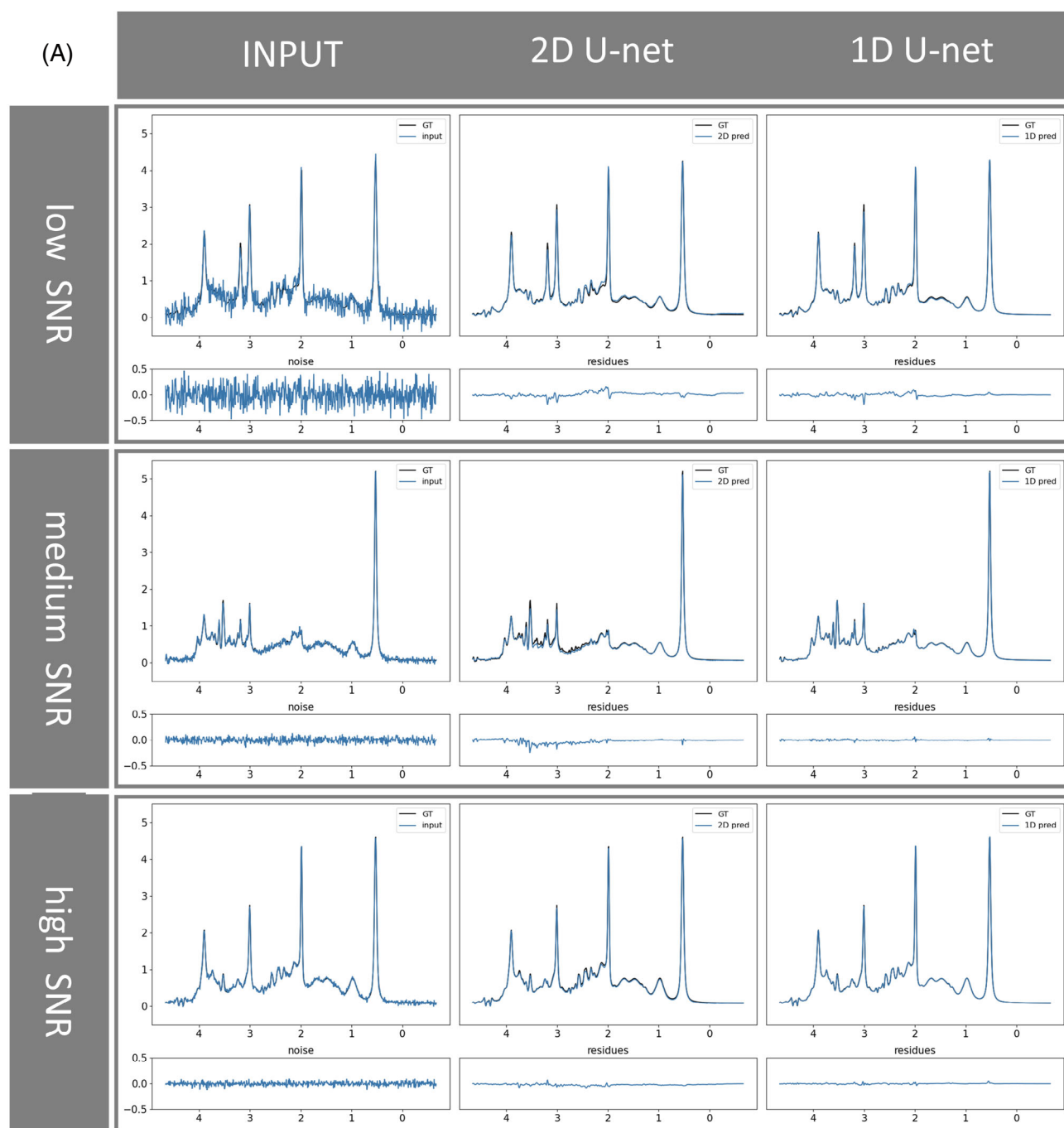


FIGURE 2 Illustration of sample spectra and the effects of denoising. (A) Three illustrative simulated cases of low (upper row), moderate (middle row), and high (lower row) SNR plotted as input on the left and the denoised output from the 2D U-Net in the middle and from the 1D U-Net on the right, zoomed to the relevant spectral range. The input and prediction spectra are overlaid with the ground truth spectrum. The corresponding noise (below input) and respective residues (prediction–ground truth [GT]) are plotted below the spectra for all cases. Residues highlight that noise suppression is most successful in the areas without signal (below 0.5 ppm). (B) Illustration of the denoising efficiency as reflected by the denoising scores. The score for all test spectra is shown without weighting for the full spectral range (denoising score [DS], left), when confined to the metabolite area (0–4.5 ppm) ($DS_{4.5\text{ppm}}$, middle), and when signal-weighting was applied (DS_w , right). DS factors below unity prove that the noise level has successfully been decreased. The scores are lowest when calculated over the full spectral range indicating that noise suppression is most successful in areas without signal. Moreover, the score is lowest for low SNR (largest denoising effect), but remains substantially below one for most (2D network) or all (1D network) cases promising better accuracy for subsequent quantification—although the latter turned out not to be the case.

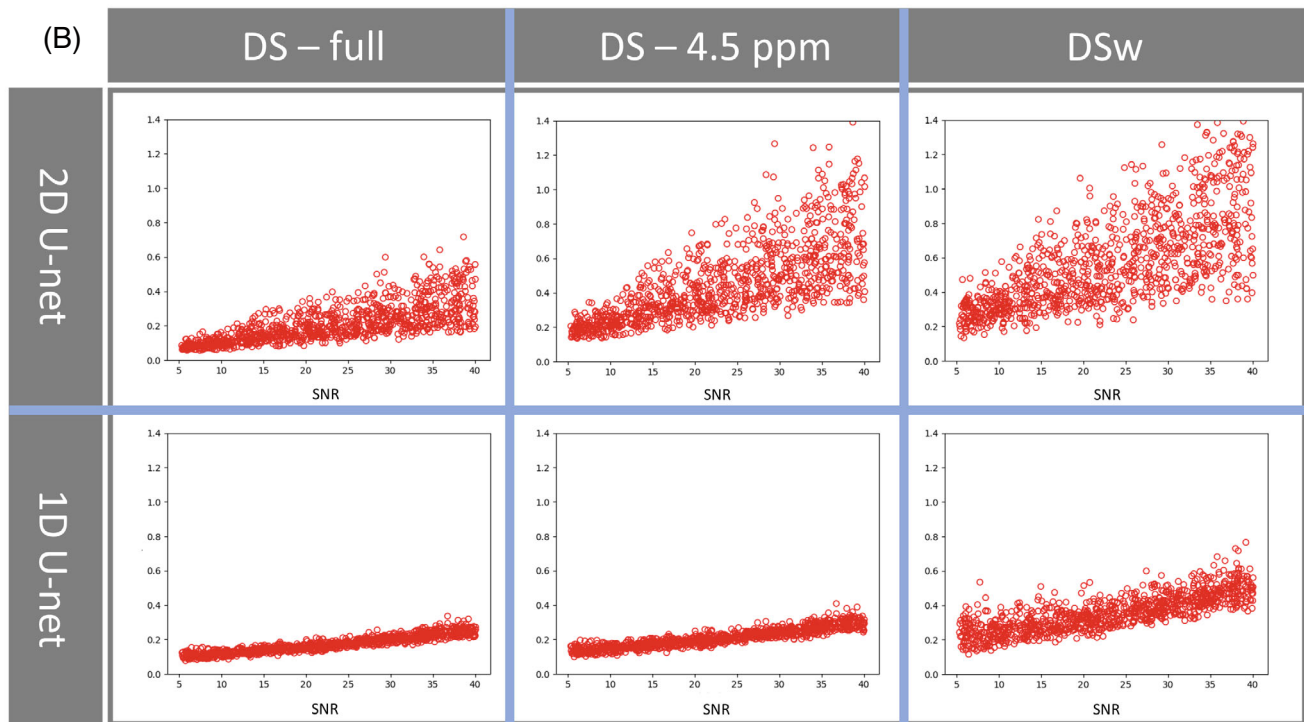


FIGURE 2 (Continued)

4 | DISCUSSION

Machine learning and in particular DL applications with astounding performances in MRI are expanding at a tremendous pace—and also in MRS, such algorithms show promise and are used in more and more contexts. They range from quality assessment^{35–39} to artifact detection,^{40,41} MRSI reconstruction,^{42–45} quantification,^{2,25,46–48} water removal,^{49,50} or denoising.¹⁶ Here, we present DL methods for denoising of single spectra, where one network is a simple U-Net with 1D spectra as input, while the other is a U-Net with input of a time-frequency representation of the MRS data that can potentially denoise simultaneously in time and frequency domain. Both methods seem exceptionally successful on visual inspection and using a global metric. When focusing on the relevant frequency range, or when using a newly introduced weighted denoising score (DS_w) this denoising success is somewhat lower but still apparently substantial, particularly for low SNR data. However, on inspection of variance and bias from subsequent quantification by MF or DL networks, it becomes obvious that the currently evaluated denoising schemes are mainly useful for display purposes, but not in a quantitative setting.

This result does not surprise given that general limits of estimation theory^{20,51} provide lower boundaries for estimated parameters that are valid for any unbiased estimation method. If denoising followed by optimal

(traditional or novel) estimation methods would indeed yield narrower confidence limits this would conflict with these lower boundaries. Comparison of full versus weighted denoising scores suggest that the denoising algorithms easily denoise the MR signals in the signal-free areas of the spectrum, but less so in the signal-containing areas (where this effect is stronger for the 2D than the 1D network). Given that (1) the weighted denoising score is substantially below unity for the majority of cases (in particular with the 1D denoising method), but that (2) the subsequent quantification results in similar mean square quantification error and substantial bias toward the mean training concentrations, it has to be concluded that denoising is apparent only and leads to a biased representation of the ground truth data. It should be stressed that the use of the word “noise” in the denoised spectrum is not appropriate if one associates white Gaussian characteristics with noise in MRS.³¹ As seen in Figure 2, this remaining “noise” is frequency-dependent and could just as well be considered residuals plus noise if the denoising process is viewed as an estimation process searching for a best representation of the ground truth spectrum.

Landheer et al.¹⁷ have recently shown that CRLBs are practically valid lower limits for estimation variance when the model is known and correct. They also concluded that DL methods may only offer benefits in dealing with data with artifacts or if the true model is not fully known in a parametric form. Here, we show that this conclusion is

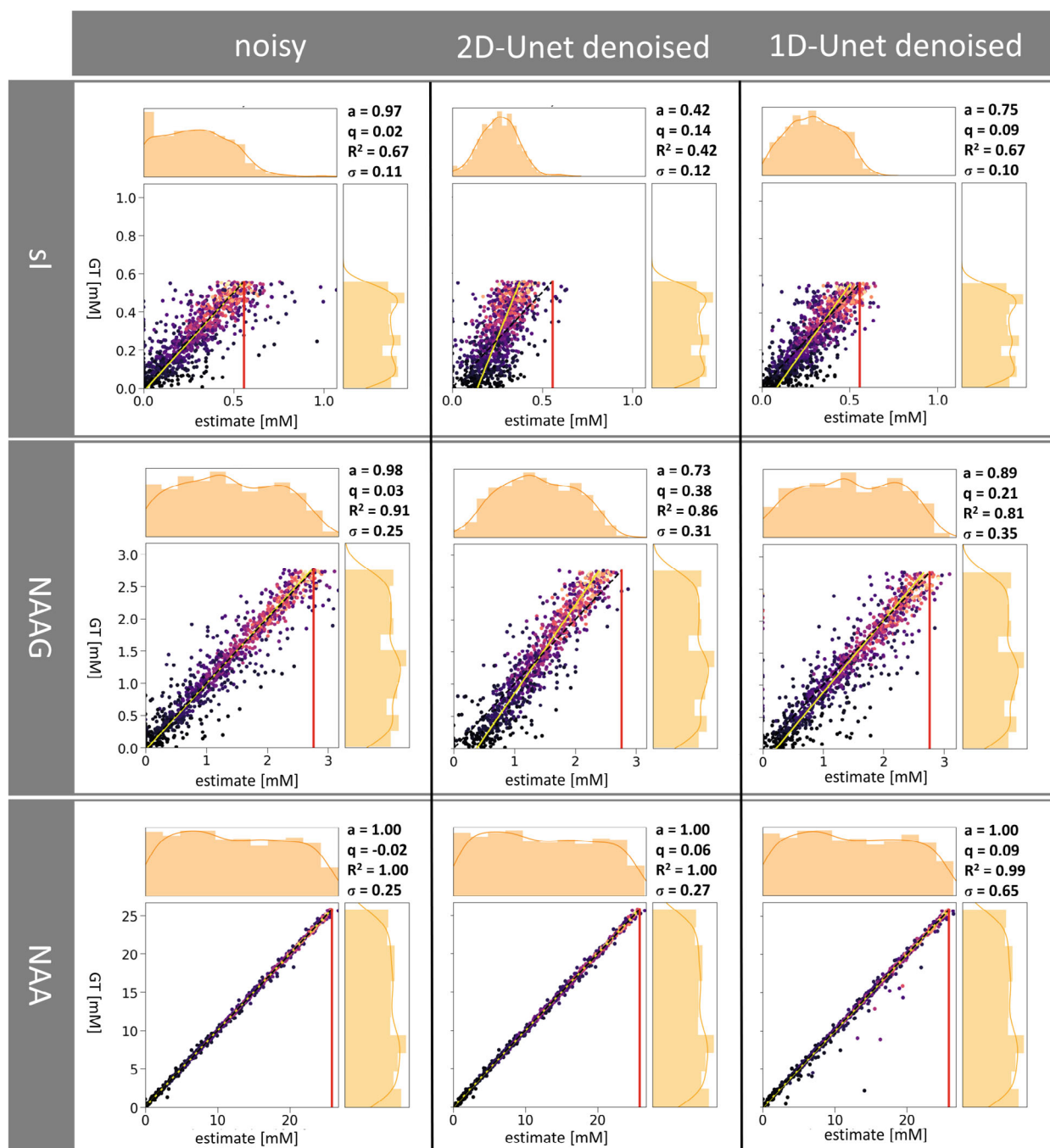


FIGURE 3 Illustration of quantification results from model fitting with FiTAID for original and denoised spectra from both denoising networks. Correlations between estimates and ground truth (GT) are depicted in each subplot. This is presented along with the marginal concentration distributions for the estimates (horizontal) and GT (vertical). One exemplary metabolite from each group of low (sI, top), medium (NAAG, middle) and high metabolite-SNR (NAA, bottom) are depicted. Quantification results from original spectra are presented on the left, those from denoised spectra in the middle (2D U-Net) and on the right (1D U-Net). Detailed numeric characterization values are given in the upper right corners: a , regression coefficient (slope); q , regression intercept; R^2 , squared Pearson regression score; and σ , the root-mean-square error over the whole dataset. The metabolite-SNR of the spectra is color-coded with lighter colors for higher SNR, with the color scale metabolite-specific to maximize the contrast and numerically interpretable from Figure S3. The dotted black regression line maps identity between GT and estimates. The estimated regression line is yellow. The vertical red line indicates the maximum GT concentration used in the test set. Results for 10 further metabolites are illustrated in Figure S1.

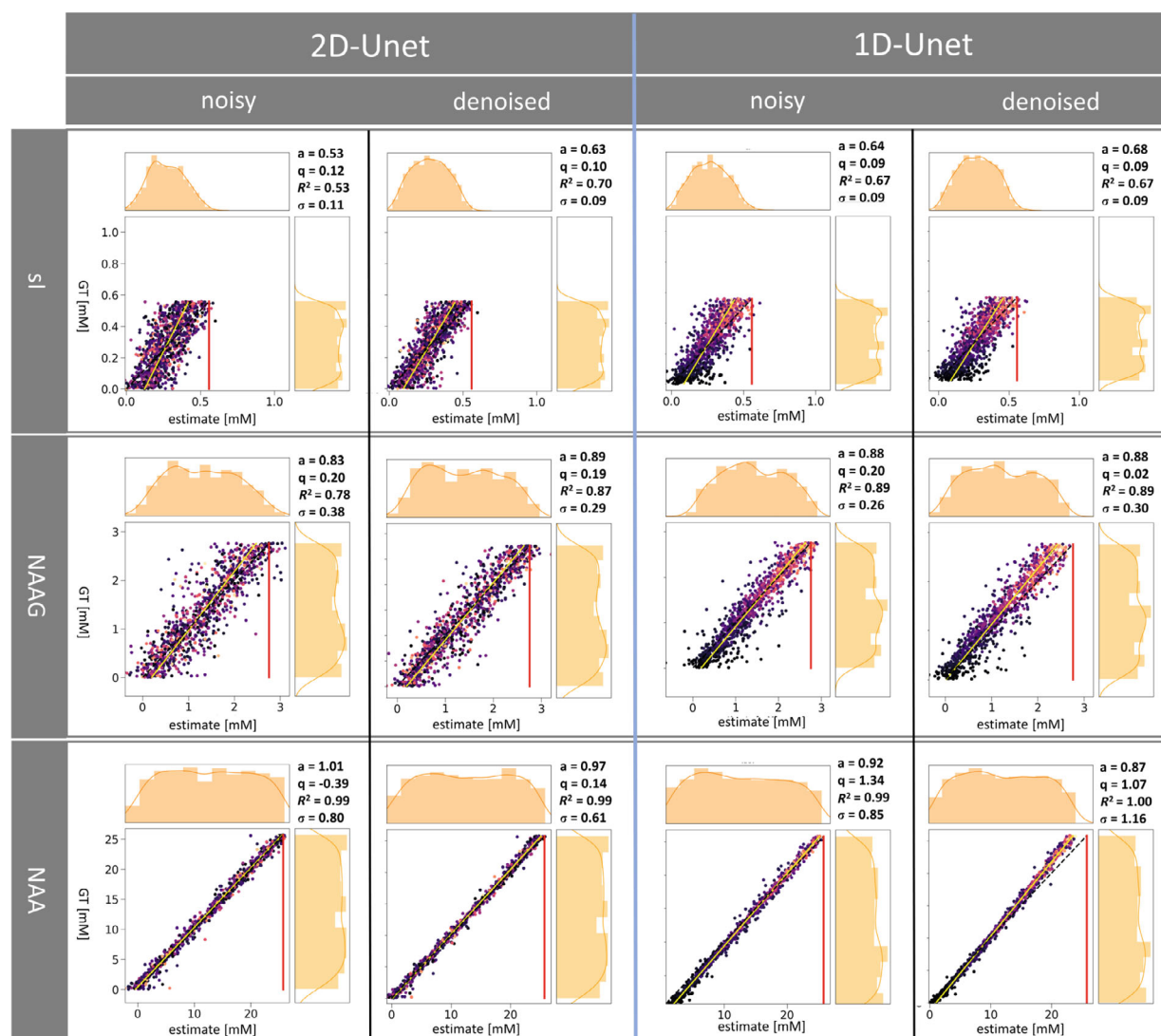


FIGURE 4 Illustration of quantification results from deep learning (DL) quantification networks for both denoising schemes comparing outcome for original (noisy) and denoised spectra. Outcome for DL quantification with a 2D network for original (noisy) and 2D-denoised spectra is shown on the left, resulting data for 1D-denoising with 1D DL quantification is given on the right. The correlation between estimates and ground truth (GT) is depicted in each subplot. This is presented along with the marginal concentration distributions for the estimates (horizontal) and GT (vertical). The plots depict one representative metabolites for each of the three metabolite-SNR groups: low (sI, top), medium (NAAG, middle) and high SNR (NAA, bottom). Detailed numeric characterization values are given in the upper right corners: a , regression coefficient (slope); q , regression intercept; R^2 , squared Pearson regression score; and σ , the root-mean-square error over the whole dataset. The metabolite-SNR of the spectra is color-coded with lighter colors for higher SNR (for metabolite-specific color code see Figure S3) The vertical red line indicates the maximum GT concentration used in the test set. Results for 10 further metabolites are illustrated in Figure S2.

valid for denoising as preprocessing step, and we show how this can introduce bias in outcome or restrictions for available parameter space.

The main goal of this study was, therefore, to document the properties of denoising through DL as a preprocessing step in metabolite quantification using synthetic data in an idealized setup. Some detailed findings from denoising followed by MF or DL estimation as categorized by representation strength of the metabolite signals are summarized below:

(1) Metabolites with low and medium SNR (e.g., sI, Lac, NAAG): denoising appears to be effective because it prevents outliers and may, therefore, lead to lower mean deviations in the estimates (σ), but it strongly biases the fit results to a Gaussian distribution around the center of the training concentration range, similar to what was reported in a different context.^{25,52} This trend is also reflected in the correlation coefficients that may increase with denoising, but the slope of the correlation line is decreasing—reflecting estimation bias. It should be noted

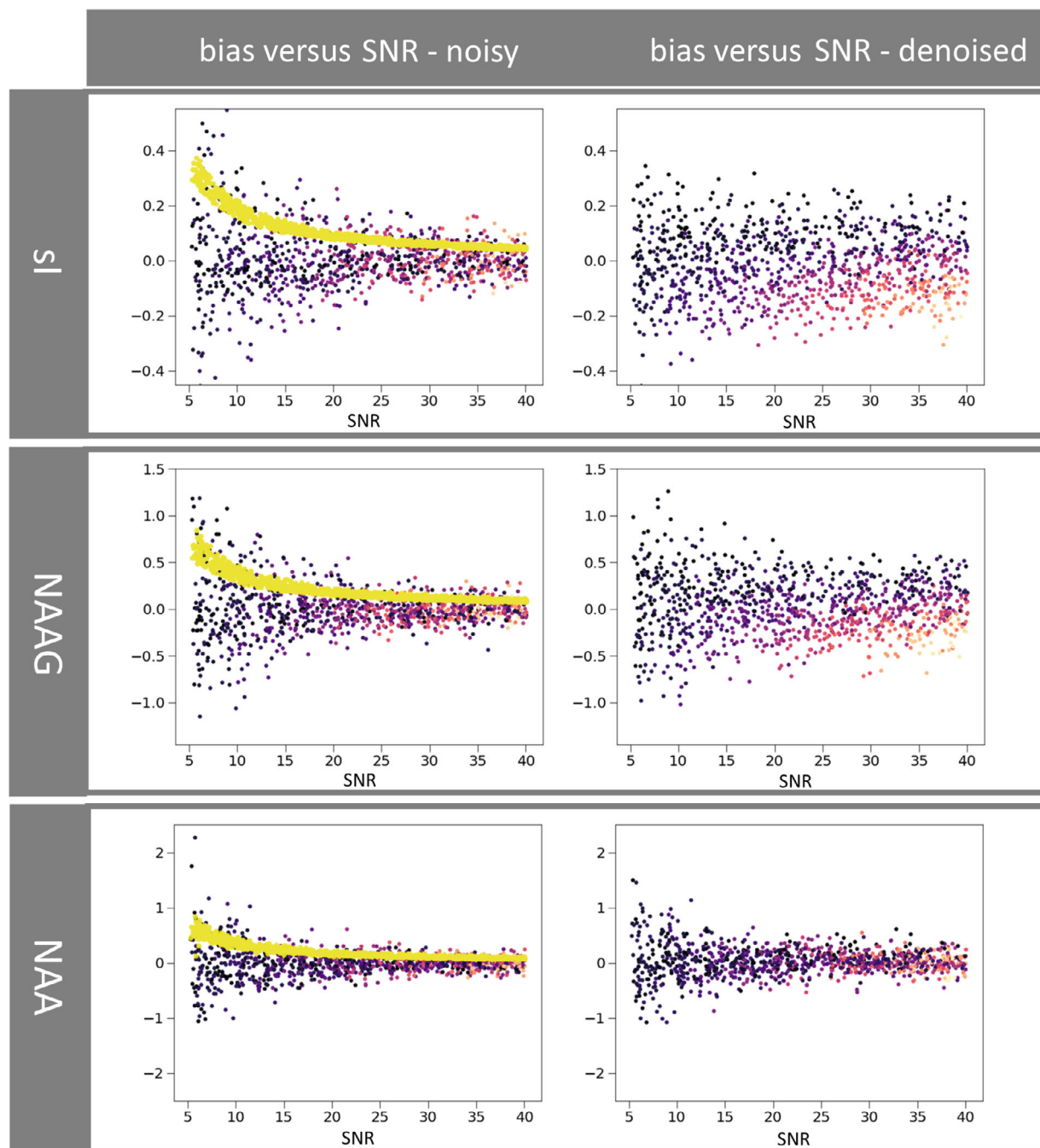


FIGURE 5 Comparison of estimation outcome from MF (presented as deviation from ground truth in mM) as function of overall SNR (for definition, see Methods) for exemplary metabolites with low (sI), moderate (NAAG) and high SNR (NAA). The left column contains the results for original (noisy) input data, the right column those for denoised input from the 2D U-Net network. The coloring of data points reflects metabolite-SNR (warmer colors for higher SNR, as defined in the color bar in Figure 3) and the yellow dots represent CRLB obtained with traditional modeling of the original noisy data in FiTAID. Distributions as function of metabolite-SNR are presented in Figure S3.

that this bias is much stronger than a potential small bias encountered from additional terms in the cost function of traditional modeling that guide the minimization progress (e.g., Levenberg Marquardt).

(2) Metabolites with high SNR (e.g., mI, NAA): the above effects are less conspicuous, but close inspection of mean errors and slopes shows that also these metabolites

do not profit from denoising. In addition, it is seen that the outcome from traditional modeling is better than from DL quantification also for these metabolites—without and also with denoising.

It should be noted that these negative conclusions about denoising do not necessarily apply to situations where non-random inherent variation between multiple

spectra is present (and can be extracted in a decomposition approach in, e.g., MRSI, series of diffusion-weighted or functional spectra). In addition, denoising may offer a practical benefit of alleviating the finding of the global χ^2 minimum with reduced chances of getting stuck in local noise-related minima, especially allowing fitting strategies that are particularly susceptible to noise.⁶

4.1 | Limitations

A limited synthetic data set was chosen to offer a broad enough space to probe realistic performance but not to cover the whole range of in vivo spectra with all potential sources of variance (range of shim settings, missing variation in phase, frequency, and lineshape). It was, therefore, not tried to apply the algorithm on experimental data.

Furthermore, the investigation was limited to spectra with known GT model, where traditional MF can converge on GT parameter values—only (and truly) limited by CRLB. If artifacts like spurious echoes or unknown background signals had been included, denoising would no longer be the correct word,³¹ but the task would change to improvement of spectral quality,⁴⁰ from which no message about denoising per-se could be derived. Although, undoubtedly, DL processing may be very useful for such tasks.

In terms of denoising followed by DL-based quantification, combined approaches are conceivable to perform better. It would be of interest to optimize quantification network parameters for input obtained by DL denoising rather than to just use denoised spectra in training. Implementing an interleaved scheme, aiming at simultaneous minimization of losses of both networks might be promising.

As only two single implementations of supervised DL-based denoising have been used, even if optimized in multiple ways, it is obviously not possible to draw conclusions for DL denoising in general. However, the conclusions are identical for both very different methods and are in-line with expectation based on theory. We, therefore, expect that the estimation bias inflicted by the denoising schemes may be taken as representative of the general consequences of denoising of single spectra by DL. It is important to mention that denoising based on extraction of common features or subspace representations of multiple spectra with different acquisition history (like spatial¹⁹ or temporal¹³ dependence or diffusion-weighting)^{9,10} is a distinct approach from what was investigated here. In these cases, the benefit of denoising may well be real in comparison to sequential analysis of the data and the outcome equivalent to simultaneous modeling of the overall dataset. However, the latter is only possible if a model

and appropriate software for a simultaneous approach is available. If not, denoising based on principle component analysis, subspace or tensor representations will be beneficial because they are model-free or provide representations that can be used to define the appropriate model for an overall fit.

5 | CONCLUSIONS

This work first aimed at investigating whether denoising via DL can effectively remove noise in spectral areas with metabolite signals and therefore, differentiate between signal and noise. The second target was to quantify estimation uncertainties when using denoising as a preprocessing step before traditional or DL-based spectral modeling and quantification. We demonstrated that the two implemented DL denoising schemes are suitable for creating visually appealing spectra. However, the weighted denoising score proved that denoising is more effective in signal-free areas. Furthermore, it was demonstrated that DL denoising effectively establishes soft constraints for the allowed parameter space, resulting in fit outcomes confined to the training range of metabolite concentrations and consequently substantial estimation bias. The extent of bias may of course be smaller for other implementations of denoising, but care will always have to be used to watch for resulting estimation bias as function of SNR and as function of the training range.

According to the current assessment and in line with expectation,¹⁸ denoising as preprocessing step in preparation of parameter estimation does not provide any benefits if the model is known and denoising does not extend to removal of non-random features.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 813120 (inspire-med) and the Swiss National Science Foundation (320030-175984). Open access funding provided by University of Bern.

DATA AVAILABILITY STATEMENT

The main part of the code will be available on Github (<https://github.com/bellarude>). For questions, please contact the authors.

ORCID

Rudy Rizzo  <https://orcid.org/0000-0003-4572-5120>

Sreenath P. Kyathanahally  <https://orcid.org/0000-0002-7399-8487>

Roland Kreis  <https://orcid.org/0000-0002-8618-6875>

REFERENCES

- Zhang Y, Shen J. Effects of noise and linewidth on in vivo analysis of glutamate at 3 T. *J Magn Reson*. 2020;314:106732.
- Macri MA, Garreffa G, Giove F, et al. In vivo quantitative 1H MRS of cerebellum and evaluation of quantitation reproducibility by simulation of different levels of noise and spectral resolution. *Magn Reson Imaging*. 2004;22:1385-1393.
- Bartha R. Effect of signal-to-noise ratio and spectral linewidth on metabolite quantification at 4 T. *NMR Biomed*. 2007;20:512-521.
- Ahmed OA. New denoising scheme for magnetic resonance spectroscopy signals. *IEEE Trans Med Imaging*. 2005;24:809-816.
- Cancino-De-Greiff HF, Ramos-Garcia R, Lorenzo-Ginori JV. Signal de-noising in magnetic resonance spectroscopy using wavelet transforms. *Concepts Magn Reson*. 2002;14:388-401.
- Papakostas GA, Karras DA, Mertzios BG, van Ormondt D, Graveron-Demilly D. Chapter 30. Two-stage evolutionary quantification of in vivo MRS metabolites. In: Tran QN, Arabnia H, eds. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*. Morgan Kaufmann; 2015:537-560.
- Goryawala M, Sullivan M, Maudsley AA. Effects of apodization smoothing and denoising on spectral fitting. *Magn Reson Imaging*. 2020;70:108-114.
- Froeling M, Prompers JJ, Klomp DWJ, van der Velden TA. PCA denoising and Wiener deconvolution of ^{31}P 3D CSI data to enhance effective SNR and improve point spread function. *Magn Reson Med*. 2021;85:2992-3009.
- Jelescu I, Veraart J, Cudalbu C. MP-PCA denoising dramatically improves SNR in large-sized MRS data: an illustration in diffusion-weighted MRS. *Proceedings of the Virtual 28th Annual Meeting of ISMRM*, Toronto, CA; 2020:395.
- Mosso J, Simicic D, Şimşek K, Kreis R, Cudalbu C, Jelescu IO. MP-PCA denoising for diffusion MRS data: promises and pitfalls. *Neuroimage*. 2022;119634:119634.
- Nguyen HM, Peng X, Do MN, Liang ZP. Denoising MR spectroscopic imaging data with low-rank approximations. *IEEE Trans Biomed Eng*. 2013;60:78-89.
- Liu Y, Ma C, Clifford BA, Lam F, Johnson CL, Liang ZP. Improved low-rank filtering of magnetic resonance spectroscopic imaging data corrupted by noise and B0 field inhomogeneity. *IEEE Trans Biomed Eng*. 2016;63:841-849.
- Brender JR, Kishimoto S, Merkle H, et al. Dynamic imaging of glucose and lactate metabolism by ^{13}C -MRS without hyperpolarization. *Sci Rep*. 2019;9:3410.
- Belkić D, Belkić K. Automatic self-correcting in signal processing for magnetic resonance spectroscopy: noise reduction, resolution improvement and splitting overlapped peaks. *J Math Chem*. 2019;57:2082-2109.
- Laleg-Kirati T-M, Zhang J, Achten E, Serrai H. Spectral data de-noising using semi-classical signal analysis: application to localized MRS. *NMR Biomed*. 2016;29:1477-1485.
- Lei Y, Ji B, Liu T, Curran W, Mao H, Yang X. Deep learning-based denoising for magnetic resonance spectroscopy signals. *Proceedings of SPIE 11600, Medical Imaging*. Vol 1160006; 2021.
- Landheer K, Juchem C. Are Cramer-Rao lower bounds an accurate estimate for standard deviations in in vivo magnetic resonance spectroscopy? *NMR Biomed*. 2021;34:e4521.
- van Ormondt D, van der Veen JW, Graveron-Demilly D. Signal denoising does not improve the precision of metabolite quantitation. *Proceedings of the BENELUX Chapter Meeting of ISMRM*, Arnhem, NL; 2020.
- Clarke WT, Chiew M. Uncertainty in denoising of MRSI using low-rank methods. *Magn Reson Med*. 2022;87:574-588.
- van den Bos A. *Parameter Estimation for Scientists and Engineers*. Wiley; 2007.
- Soher BJ, Semanchuk P, Todd D, Steinberg J, Young K. VeSPA: integrated applications for RF pulse design, spectral simulation and MRS data analysis. *Proceedings of the 19th Annual Meeting of ISMRM*. CANA, Montreal; 2011:1410.
- Oz G, Tkac I. Short-echo, single-shot, full-intensity proton magnetic resonance spectroscopy for neurochemical profiling at 4 T: validation in the cerebellum and brainstem. *Magn Reson Med*. 2011;65:901-910.
- Scheenen TW, Klomp DW, Wijnen JP, Heerschap A. Short echo time 1H-MRSI of the human brain at 3T with minimal chemical shift displacement errors using adiabatic refocusing pulses. *Magn Reson Med*. 2008;59:1-6.
- Marjanska M, McCarten JR, Hodges J, et al. Region-specific aging of the human brain as evidenced by neurochemical profiles measured noninvasively in the posterior cingulate cortex and the occipital lobe using (1)H magnetic resonance spectroscopy at 7 T. *Neuroscience*. 2017;354:168-177.
- Rizzo R, Dziadosz M, Kyathanahally SP, Shamaei A, Kreis R. Quantification of MR spectra by deep learning in an idealized setting: investigation of forms of input, network architectures, optimization by ensembles of networks, and training bias. *Magn Reson Med*. 2023;89:1707-1727.
- Traber F, Block W, Lamerichs R, Gieseke J, Schild HH. 1H metabolite relaxation times at 3.0 tesla: measurements of T1 and T2 values in normal brain and determination of regional differences in transverse relaxation. *J Magn Reson Imaging*. 2004;19:537-545.
- Hoefemann M, Adalid V, Kreis R. Optimizing acquisition and fitting conditions for (1) H MR spectroscopy investigations in global brain pathology. *NMR Biomed*. 2019;32:e4161.
- Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res*. 2010;11:3371-3408.
- Dong L-F, Gan Y-Z, Mao X-J, Yang Y, Shen C. Learning deep representations using convolutional auto-encoders with symmetric skip connections. *2018 IEEE international conference on acoustics, Speech and Signal Processing (ICASSP)*; 2018: 3006-3010.
- Belz V. Speech enhancement. <https://github.com/vbelz/Speech-enhancement>.
- Kreis R, Boer V, Choi IY, et al. Terminology and concepts for the characterization of in vivo MR spectroscopy methods and MR spectra: background and experts' consensus recommendations. *NMR Biomed*. 2020;34:e4347.
- Chong DG, Kreis R, Bolliger CS, Boesch C, Slotboom J. Two-dimensional linear-combination model fitting of magnetic resonance spectra to define the macromolecule baseline using FiTAID, a fitting tool for arrays of interrelated datasets. *Magma*. 2011;24:147-164.
- Rizzo R, Dziadosz M, Kyathanahally SP, Reyes M, Kreis R. Reliability of quantification estimates in MR spectroscopy: CNNs vs.

- traditional model fitting. *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer; 2022:715-724.
34. Rizzo R, Kreis R. Accounting for bias in estimated metabolite concentrations from cohort studies as caused by limiting the fitting parameter space. *Proceedings of the Virtual 29th Annual Meeting of ISMRM*; 2021:2011.
 35. Wright AJ, Arús C, Wijnen JP, et al. Automated quality control protocol for MR spectra of brain tumors. *Magn Reson Med*. 2008;59:1274-1281.
 36. Kyathanahally SP, Mocioiu V, Pedrosa de Barros N, et al. Quality of clinical brain tumor MR spectra judged by humans and machine learning tools. *Magn Reson Med*. 2018;79:2500-2510.
 37. Pedrosa de Barros N, McKinley R, Wiest R, Slotboom J. Improving labeling efficiency in automatic quality control of MRSI data. *Magn Reson Med*. 2017;78:2399-2405.
 38. Jang J, Lee HH, Park J-A, Kim H. Unsupervised anomaly detection using generative adversarial networks in 1H-MRS of the brain. *J Magn Reson*. 2021;325:106936.
 39. Lee HH, Kim H. Deep learning-based target metabolite isolation and big data-driven measurement uncertainty estimation in proton magnetic resonance spectroscopy of the brain. *Magn Reson Med*. 2020;84:1689-1706.
 40. Kyathanahally SP, Döring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magn Reson Med*. 2018;80:851-863.
 41. Gurbani SS, Schreibmann E, Maudsley AA, et al. A convolutional neural network to filter artifacts in spectroscopic MRI. *Magn Reson Med*. 2018;80:1765-1775.
 42. Luo J, Zeng Q, Wu K, Lin Y. Fast reconstruction of non-uniform sampling multidimensional NMR spectroscopy via a deep neural network. *J Magn Reson*. 2020;317:106772.
 43. Lam F, Li Y, Peng X. Constrained magnetic resonance spectroscopic imaging by learning nonlinear low-dimensional models. *IEEE Trans Med Imaging*. 2020;39:545-555.
 44. Motyka S, Hingerl L, Strasser B, et al. k-Space-based coil combination via geometric deep learning for reconstruction of non-cartesian MRSI data. *Magn Reson Med*. 2021;86:2353-2367.
 45. Iqbal Z, Nguyen D, Hangel G, Motyka S, Bogner W, Jiang S. Super-resolution 1H magnetic resonance spectroscopic imaging utilizing deep learning. *Front Oncol*. 2019;9:1010.
 46. Gurbani SS, Sheriff S, Maudsley AA, Shim H, Cooper LAD. Incorporation of a spectral model in a convolutional neural network for accelerated spectral fitting. *Magn Reson Med*. 2019;81:3346-3357.
 47. Lee HH, Kim H. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. *Magn Reson Med*. 2019;82:33-48.
 48. Shamaei A, Starcukova J, Starcuk Z Jr. Physics-informed deep learning approach to quantification of human brain metabolites from magnetic resonance spectroscopy data. *Comput Biol Med*. 2023;158:106837.
 49. Louis MS, Coello E, Liao H, Joshi A, Lin A. Quantification of non-water-suppressed proton spectroscopy using deep neural networks. *Proceedings of the Virtual 28th Annual Meeting of ISMRM*, Toronto; 2021:1298.
 50. Nagaraja BH, Debals O, Sima DM, Himmelreich U, De Lathauwer LD, Van Huffel S. Tensor-based method for residual water suppression in 1H magnetic resonance spectroscopic imaging. *IEEE Trans Biomed Eng*. 2019;66:584-594.
 51. de Beer R, van Ormondt D. Analysis of NMR data using time domain fitting procedures. In: Rudin M, ed. *In-Vivo Magnetic Resonance Spectroscopy I: Probeheads and Radiofrequency Pulses Spectrum Analysis NMR (Basic Principles and Progress)*. Vol 26. Springer; 1992.
 52. Györi NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn Reson Med*. 2022;87:932-947.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Figure S1. Illustration of quantification results from model fitting with FiTAID for original and denoised spectra from both denoising networks for 10 further metabolites—complementing Figure 3.

Figure S2. (A, B) Illustration of quantification results from deep learning (DL) quantification networks for both denoising schemes comparing outcome for original (noisy) and denoised spectra for 10 further metabolites—complementing Figure 4.

Figure S3. Comparison of estimation outcome from MF (presented as deviation from ground truth in mM) as function of metabolite-SNR (for definition, see Methods) for exemplary metabolites with low (sI), moderate (NAAG) and high (NAA) SNR—complementing Figure 5.

How to cite this article: Dziadosz M, Rizzo R, Kyathanahally SP, Kreis R. Denoising single MR spectra by deep learning: Miracle or mirage?. *Magn Reson Med*. 2023;90:1749-1761. doi: 10.1002/mrm.29762