Received: 12 September 2022 | Revised: 14 February 2023 | Accepted: 16 February 2023 | Published on: 13 March 2023

DOI: 10.1002/mrm.29628

RESEARCH ARTICLE

Magnetic Resonance in Medicine

Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies

Misha P. T. Kaandorp[1,2] | **Frank Zijlstra**[1,2] | **Christian Federau**[3,4] | **Peter T. While**[1,2]

[1]Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway

[2]Department of Circulation and Medical Imaging, NTNU – Norwegian University of Science and Technology, Trondheim, Norway

[3]Institute for Biomedical Engineering, University and ETH Zurich, Zurich, Switzerland

[4]AI Medical, Zurich, Switzerland

Correspondence
Misha P. T. Kaandorp, Department of Radiology and Nuclear Medicine, St. Olav's University Hospital, Trondheim, Norway.
Email: mpkaando@stud.ntnu.no

Funding information
Norges Forskningsråd, Grant/Award Number: Grant/Award Number: 302624

Purpose: The development of advanced estimators for intravoxel incoherent motion (IVIM) modeling is often motivated by a desire to produce smoother parameter maps than least squares (LSQ). Deep neural networks show promise to this end, yet performance may be conditional on a myriad of choices regarding the learning strategy. In this work, we have explored potential impacts of key training features in unsupervised and supervised learning for IVIM model fitting.

Methods: Two synthetic data sets and one in-vivo data set from glioma patients were used in training of unsupervised and supervised networks for assessing generalizability. Network stability for different learning rates and network sizes was assessed in terms of loss convergence. Accuracy, precision, and bias were assessed by comparing estimations against ground truth after using different training data (synthetic and in vivo).

Results: A high learning rate, small network size, and early stopping resulted in sub-optimal solutions and correlations in fitted IVIM parameters. Extending training beyond early stopping resolved these correlations and reduced parameter error. However, extensive training resulted in increased noise sensitivity, where unsupervised estimates displayed variability similar to LSQ. In contrast, supervised estimates demonstrated improved precision but were strongly biased toward the mean of the training distribution, resulting in relatively smooth, yet possibly deceptive parameter maps. Extensive training also reduced the impact of individual hyperparameters.

Conclusion: Voxel-wise deep learning for IVIM fitting demands sufficiently extensive training to minimize parameter correlation and bias for unsupervised learning, or demands a close correspondence between the training and test sets for supervised learning.

KEYWORDS
diffusion-weighted magnetic resonance imaging, gliomas, intravoxel incoherent motion, IVIM, supervised deep learning, unsupervised deep learning

1 | INTRODUCTION

The intravoxel incoherent motion (IVIM)[1] model for DWI shows great potential to be an alternative to dynamic susceptibility contrast (DSC) imaging for providing prognostic perfusion-based cancer imaging biomarkers, which can be used to characterize tumors and monitor treatment response without the use of a contrast agent.[2–4] The IVIM model describes the attenuation of the MRI signal as a bi-exponential function of the diffusion weighting (b value). The IVIM model parameters of interest include a diffusion coefficient (D) associated with the slow diffusion of water molecules, a pseudo-diffusion coefficient (D^*) which is linked to the fast diffusion effect caused by capillary microcirculation, and a perfusion fraction (f) representing the signal contribution from the pseudo-diffusion component. The IVIM model has shown promise in several anatomies,[3,5–7] including the brain.[2,8,9] However, fitting IVIM parameters remains challenging in the in-vivo brain due to low SNR and low f.[2,9,10]

Several data-processing algorithms have been proposed to improve the IVIM parameter estimation quality.[11] New IVIM fitting approaches are typically compared to conventional non-linear least squares (LSQ), or a two-step segmented approach, which has been shown in some studies to be less sensitive to noise than LSQ, particularly for the pseudo-diffusion estimates.[10,12] More advanced Bayesian inference algorithms have been proposed, demonstrating improved precision and accuracy.[13,14] However, these Bayesian approaches are typically very time-consuming, and may lead to biased parameter estimates or features disappearing due to over-smoothing.[15]

Recently, deep learning or deep neural networks (DNNs) were introduced as a promising alternative for IVIM fitting. These DNNs can either be trained unsupervised, where the network is optimized on the error between the estimated IVIM signal and the measured signal; or trained supervised, where the network is optimized on the error between the estimated parameters and the ground truth parameters. The approach of unsupervised learning, originally proposed by Barbieri et al.,[16] allows the DNN to be trained directly on the in vivo signals, without the need for labeled data. Kaandorp et al.[17] observed unexpected parameter correlations with this unsupervised network, but resolved these by optimizing various hyperparameters (IVIM-NET$_{optim}$). However, although IVIM-NET$_{optim}$ showed promising results in the pancreas,[17] applying it to brain data showed poor anatomy generalization and high D^* values.[18]

In contrast, the approach of supervised learning for IVIM fitting, originally proposed by Bertleff et al.,[19] carries the assumption that the in vivo test data is well represented by the training data, which could limit the network's performance in terms of generalizability. Gyori et al.[20] performed an in-depth investigation of the supervised approach for simple two-compartment and three-compartment diffusion models based on the spherical mean technique, and showed that these supervised approaches produce smooth parameter maps that are significantly influenced by the training distribution, which could possibly lead to deceptive parameter maps.

Based on these observations, we hypothesize that the performance of both unsupervised and supervised approaches for IVIM fitting is strongly influenced by the choice of both training data and network hyperparameters; hence, we identify the need to test and compare the limitations of both approaches. In this work, we explore the impact of learning rate, network size, training length, and training distribution on the convergence behavior of the unsupervised and supervised DNN loss terms, and on the accuracy of the parameter estimates. We also explore the impact of a suite of other hyperparameters on the final loss terms. We demonstrate the possible pitfalls associated with early stopping, training length, and different training data distributions, using both simulations and in-vivo data from glioma patients.

2 | METHODS

2.1 | In vivo data and preprocessing

We analyzed in vivo data from 28 glioma patients, which had been acquired and used in previous studies.[2,18,21] Those studies were approved by the local ethics committee at the University of Lausanne, and patient consent was waived. Data were acquired using a 3T MRI scanner (Trio, Verio, or Skyra; Siemens, Erlangen, Germany). A Stejskal-Tanner diffusion-weighted spin-echo EPI pulse sequence was used with 16 b values: 0, 10, 20, 40, 80, 110, 140, 170, 200, 300, 400, 500, 600, 700, 800, 900 mm^2/s in three orthogonal directions, and the corresponding trace was calculated (no averaging). A total of 20 patients were included from one study[2] with scan parameters: TR = 4000 ms, TE = 99 ms, FOV = 297 × 297 mm^2, acquisition matrix = 256 × 256, with varying slices (9-21), and receiver bandwidth = 1086 Hz/pixel. A total of eight patients were included from another study[21] with scan parameters: TR = 4000 ms, TE = 93 ms, FOV = 270 × 270 mm^2, acquisition matrix = 225 × 225, with 20 slices, and receiver bandwidth = 1106 Hz/pixel. In addition, both studies used slice thickness = 4 mm, in-plane resolution of 1.2 × 1.2 mm^2, 75% partial Fourier encoding, parallel imaging with an acceleration factor of

2, and fat was suppressed with a spectrally selective saturation routine. The total acquisition time for both studies was 3 min and 7 s.

To pre-process the DWI data, we first removed background voxels by using both a manual threshold, as used in previous studies,[2,21] and a Gaussian filter, which removed other background artifacts, including ghosting artifacts. Data were normalized to the SNR of the white matter (WM) of each patient, defined as the ratio of mean to standard deviation, estimated for a homogeneous region of interest (ROI), which had an SNR of approximately 30.

2.2 | Networks

We initially implemented the original DNN architecture of Barbieri et al.[16] in PyTorch. The network architecture was a multi-layer perceptron with three hidden layers and exponential linear unit (ELU) activation functions.[22] The network input consisted of the measured DWI signal $S(b)$ at each b value, and the network output consisted of the three IVIM parameters plus an extra parameter S_0, the signal at $b = 0$. These parameters were further constrained by absolute value functions to avoid the possibility of exploding gradients due to negative output, and scaled (but not constrained) to appropriate physical ranges: $0 \le S_0 \le 1$, $0 \le D \le 3\times10^{-3}$ mm^2/s, $0 \le f \le 50\%$, and $3\times10^{-3} \le D^* \le 100\times10^{-3}$ mm^2/s (corresponding to network output 0–1). The upper scaling value of D is approximately equal to the self-diffusion coefficient of free water.[23]

The networks were trained either unsupervised or supervised. The unsupervised networks were trained using a loss term equal to the mean-squared error (MSE) between the input signal $S(b)$ and the estimated IVIM signal $S_{net}(b)$, denoted in this manuscript as "signals-MSE":

$$\text{signals-MSE} = \mathcal{L}\left(S(b), S_{net}(b)\right) = \sum_{b\in B} \|S(b) - S_{net}(b)\|^2 \tag{1}$$

where

$$S_{net}(b) = S_0\left(f e^{-bD^*} + (1-f)e^{-bD}\right) \tag{2}$$

and B is the set of b values. The supervised networks were trained using a loss term equal to the MSE between the output parameters $\hat{\theta}_{net}$ of the network and the normalized IVIM parameters $\hat{\theta}$ (scaled between 0 and 1), denoted in this manuscript as "parameters-MSE":

$$\text{parameters-MSE} = \mathcal{L}\left(\hat{\theta}, \hat{\theta}_{net}\right) = \sum_{S_0,D,F,D*} \left\|\hat{\theta} - \hat{\theta}_{net}\right\|^2. \tag{3}$$

2.3 | Training and test data

In addition to the in vivo data described in Section 2.1, two synthetic data sets were considered for the training and testing of the networks. For the synthetic data, parameter combinations were drawn from the two distributions described below. IVIM signals were simulated using Eq. 2 for the same set of b values as the in vivo data. Rician noise was added to the signals such that when $S_0 = 1$ the SNR was 200. The following three data sets were used for training and testing:

1. Uniform distribution (synthetic): IVIM signals were simulated uniformly with the following parameter ranges: $0 \le S_0 \le 1$, $0 \le D \le 3\times10^{-3}$ mm^2/s, $0 \le f \le 50\%$, and $3\times10^{-3} \le D^* \le 100\times10^{-3}$ mm^2/s (i.e., identical to the scaling range of the output of the network). The uniform test set consisted of 100 000 randomly generated IVIM signals.
2. Patient distribution (synthetic): To provide a realistic patient distribution, we sampled IVIM parameter combinations obtained from conventional model fitting applied to the in vivo data. We performed a segmented fit using the Levenberg–Marquardt algorithm for each patient, where we first fitted D for $b > 200$ s/mm^2 and afterward fitted f and D^* using all b values while fixing D. S_0 was derived from the normalized $b = 0$ image. The upper bound for the segmented fit of f was 100%, while the upper bounds of S_0, D and D^* were unconstrained. All parameters were constrained to a lower bound of 0. We then simulated IVIM signals using the parameter combinations obtained from the segmented fit. We only included parameter combinations for which $D^* \le 100\times10^{-3}$ mm^2/s. Segmented fitting was chosen over LSQ because it was found to be more robust to noise.
3. In vivo data from glioma patients, as described in Section 2.1.

For the patient-derived data ((2) and (3) above), we used 20 subjects for training and 8 subjects for testing. Figure 1 displays the parameter distributions for the synthetic test sets, which illustrates that the range and weighting of parameters differed considerably between the two synthetic test sets. Training was performed with 500 batches per epoch and batch size 128 using an Adam optimizer,[24] similar to previous work,[17] for 50 000 epochs. Note that alternative hyperparameters were also considered, as described in Section 2.4.2 below. Training and testing were performed on a single core of an Intel Xeon Gold 6226R CPU, with each training of 50 000 taking approximately 16.5 h (1.18 s per epoch).

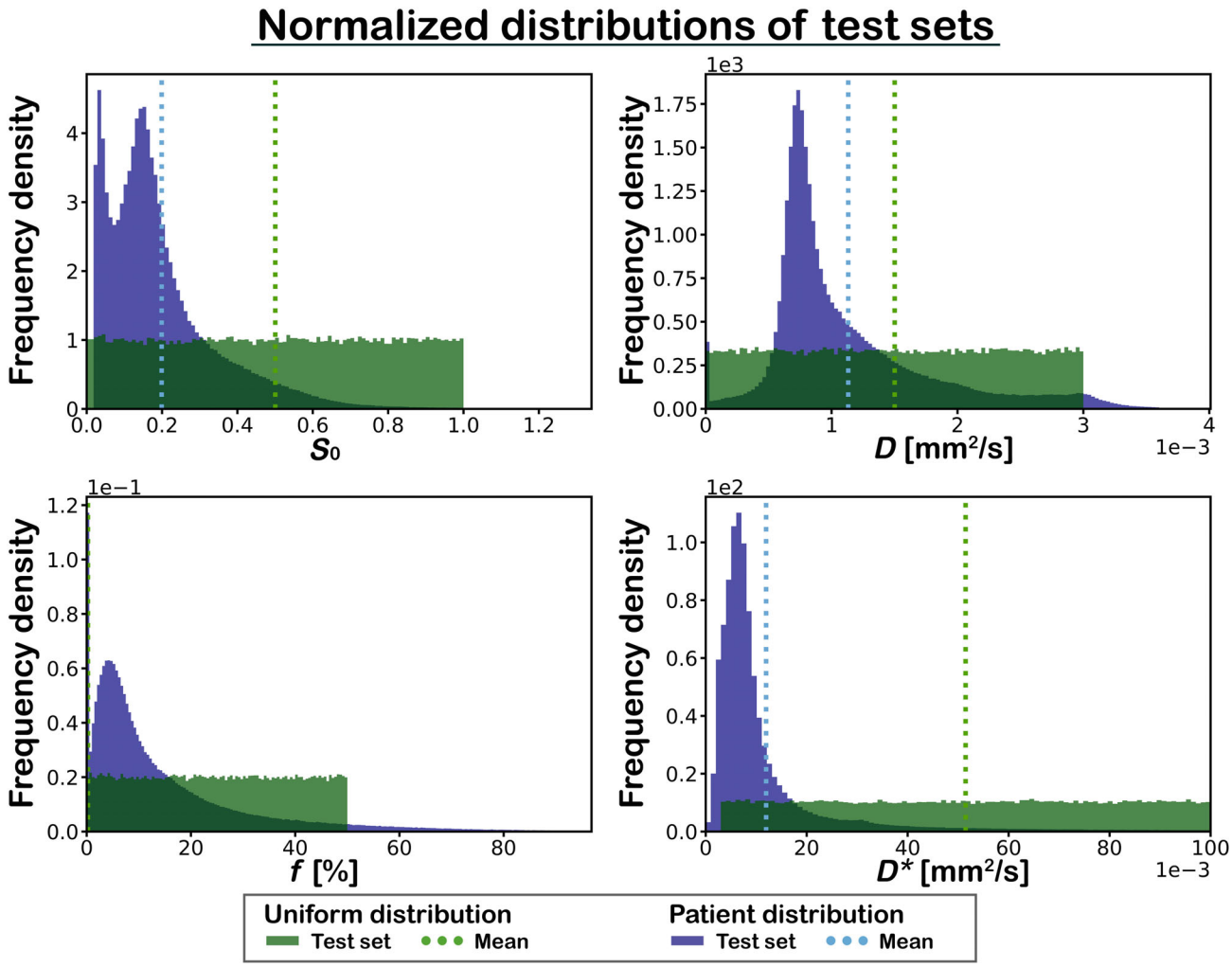

FIGURE 1 Normalized distributions for each IVIM parameter for the uniform distribution synthetic test set and the patient distribution synthetic test set. The histograms were normalized so that the area under the curve is 1 (i.e., density curve). For S_0, D, and f, the upper 0.1% of the patient distribution test set is not displayed for visualization purposes. The dotted lines represent the parameter means for each test set.

2.4 | Evaluation

2.4.1 | Simulations: Learning rate and network size

We tested the impact of the learning rate and network size for the different strategies of unsupervised and supervised learning using the uniform distribution synthetic data set. We compared initially the performance of four networks by considering two different numbers of hidden units (16, 64) and two different learning rates (1×10^{-3}, 1×10^{-4}). For each network, we computed the signals-MSE (unsupervised loss), parameters-MSE (supervised loss) and Spearman's correlation coefficient between the pseudo-diffusion parameters ($\rho(D^*,f)$) at the end of each epoch. Previous work[17] demonstrated that DNNs may predict erroneous correlations between the pseudo-diffusion parameters; hence, it is important to test for correlations in assessing the performance of the networks. The stability of the networks was assessed in terms of the convergence of signals-MSE or parameters-MSE on the test set, and the hyperparameters corresponding to the most stable network were used in all subsequent analyses. For the most stable network, we evaluated the suitability of the early-stopping criterion used in previous approaches[16,17] which is triggered when the test loss does not improve over 10 epochs.

2.4.2 | Simulations: Hyperparameters

We also investigated the influence of several other hyperparameters on the loss values at the early-stopping point and at the last epoch (Epoch-50000). We trained and tested networks that differed by one hyperparameter with respect to the baseline network that had been declared the most

stable (Section 2.4.1; see also Section 3.1). Here we assessed both learning strategies (unsupervised, supervised) using the uniform distribution synthetic data set. The different hyperparameters that were examined included (baseline in square parentheses): batch size ([128], 256, 512), activation function ([ELU], ReLu, sigmoid), number of hidden units (16, [64], 128), batch normalization ([without], with), dropout ([0%], 10%, 20%), number of hidden layers (2, [3], 4, 5, 6), optimizer ([Adam], SGD, RMSProp), and learning rate (1×10^{-3}, $[1\times10^{-4}]$, 1×10^{-5}).

2.4.3 | Simulations: Learning strategy and training distribution

Using the hyperparameters corresponding to the most stable network (Section 2.4.1), four further networks were trained by considering the two different learning strategies (unsupervised, supervised) and the two synthetic data sets (uniform distribution, patient distribution). These four networks were considered in all subsequent analyses. These networks were tested on the patient distribution test set to assess generalizability. For each network, we computed the signals-MSE, parameters-MSE, Spearman's $\rho(D^*,f)$, and normalized MSE per parameter at the end of each epoch.

2.4.4 | Simulations: Training length

The performance of the four networks was further evaluated at different validation points during training by comparing the estimated parameters with the ground truth for individual data points of the uniform distribution test set. To enable further qualitative assessment, parameter maps and root-mean-square error (RMSE) maps were generated for each network from the synthetic patient data. Here, the RMSE was calculated between the expected DWI signal for the estimated parameters and the ground truth DWI signal. Four different validation points were used: (i) early stopping; (ii) when D^*-MSE was at a minimum (Min(D^*-MSE)); (iii) Epoch-5000; (iv) the last epoch (Epoch-50000), representing extensive training. Comparisons were also made to parameter maps estimated by LSQ using the Levenberg–Marquardt algorithm, and a segmented approach as described in Section 2.3.

2.4.5 | In vivo

To demonstrate that our simulations are comparable to real in vivo data, we evaluated the performance of the four networks (i.e., trained on synthetic data) on the in vivo glioma patient data, together with an additional unsupervised network that was directly trained on the in vivo signals. We computed the signals-MSE and $\rho(D^*,f)$ at the end of each epoch. In order to provide a fair assessment of the bulk of the data, we performed a correction for outliers by excluding voxels lying within the top 0.7% of signals-MSE, chosen to mimic classical outlier detection.[25] These voxels contained signals that were poorly represented by a bi-exponential function, and therefore represented signal behavior not seen by the four networks during training.

Performance was also assessed qualitatively for the in vivo glioma patient data by generating parameter maps and RMSE maps at the early-stopping point, Epoch-5000 and Epoch-50000. These maps were compared to parameter maps estimated by LSQ, the segmented approach, the four networks from Section 2.4.3 at Epoch-50000, and IVIM-NET$_{\text{optim}}$.[17] IVIM-NET$_{\text{optim}}$ was applied as in the original article,[17] with the exception that the range of the S_0 parameter was 0–2 instead of 0.7–1.3 to encompass the signal scaling in this study.

An additional important consideration is to assess the impact that the individual fitting approaches may have on the quantitative values of each parameter for certain tissue types of interest. Hence, for each IVIM parameter, we pooled the voxel-wise estimates for in vivo tumor regions of all patients (one patient was excluded as no tumor mask was available) and compared the distribution of parameter estimates for each fitting approach at the early-stopping point and at Epoch-50000.

3 | RESULTS

3.1 | Simulations: Learning rate and network size

Figure 2 shows the loss curves of the first 5000 epochs for both the unsupervised and supervised learning strategies trained and tested on the uniform distribution using different network size and learning rate. For both learning strategies, using fewer hidden units reduced convergence speed, whereas higher learning rate resulted in spiky convergence of the loss term, which could result in sub-optimal solutions, particularly in unsupervised learning. Therefore, the network containing hidden units = 64 and learning rate = 1×10^{-4} was considered the most stable in both strategies, and was used in all subsequent analyses.

Early stopping (patience = 10 epochs) resulted in sub-optimal solutions prior to true convergence, at epoch 237 for unsupervised learning and epoch 118 for supervised learning (Figure 2). Furthermore, the

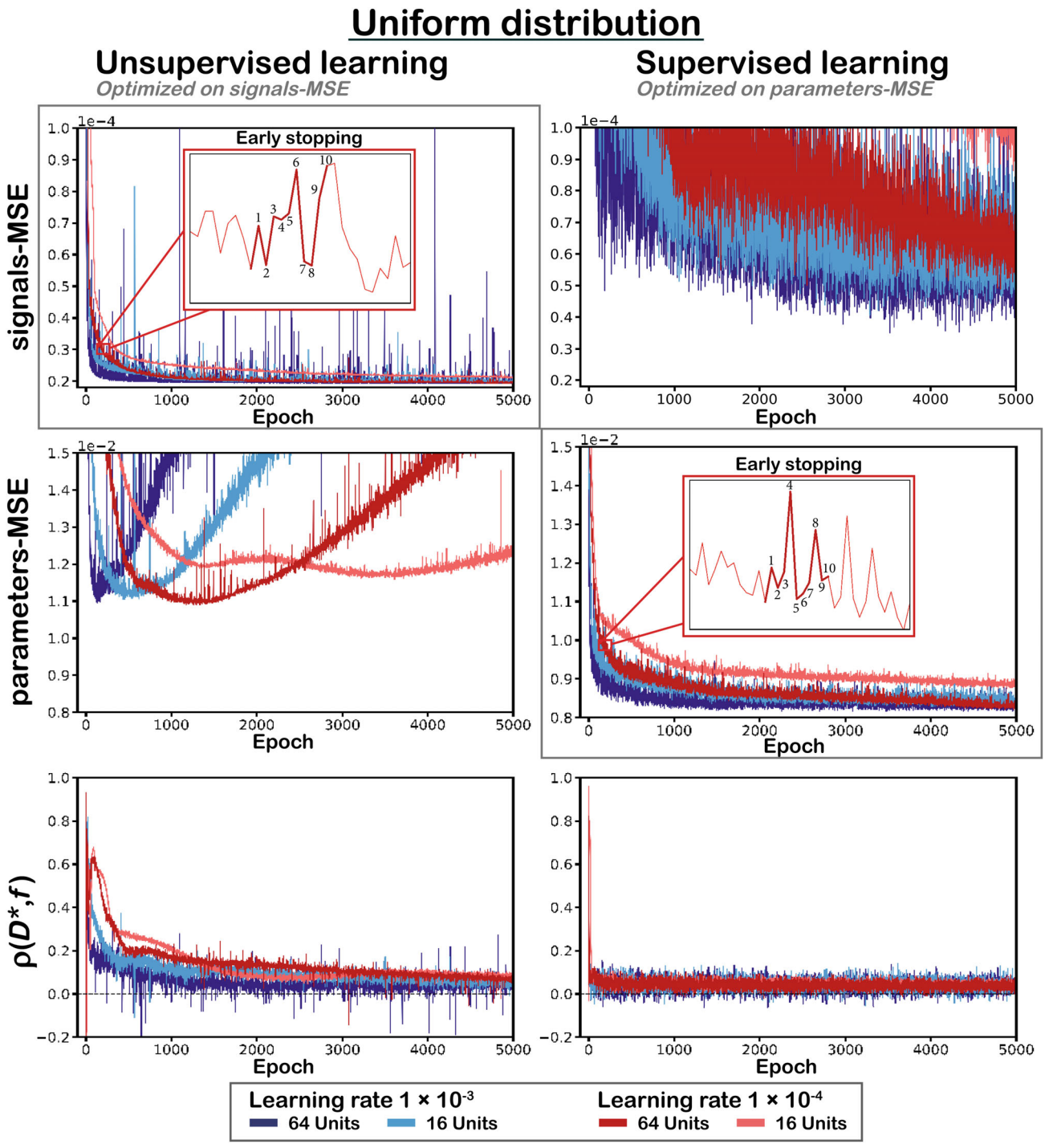

FIGURE 2 Test curves showing the metrics signals-MSE, parameters-MSE, and Spearman's $\rho(D^*,f)$ for unsupervised (left) and supervised (right) networks with different number of hidden units (16, 64) and learning rate (1×10^{-3}, 1×10^{-4}), trained and tested on synthetic data from the uniform distribution, and evaluated over 5000 epochs. For the most stable network, the red square highlights where the test loss did not improve over 10 consecutive epochs and the early-stopping criterion used in prior work[16,17] would have ended the training, which results in sub-optimal convergence, as well as correlated parameters for the unsupervised case at epoch 223 ($\rho(D^*,f) = 0.62$).

pseudo-diffusion parameters were correlated for unsupervised learning ($\rho(D^*,f) = 0.62$). Extending training resolved these correlations and reduced parameters-MSE. However, training substantially longer resulted in increased parameters-MSE. For supervised learning, both parameters-MSE and signals-MSE decreased with longer training length. As expected, unsupervised learning yielded lower signals-MSE (unsupervised loss), whereas supervised learning yielded lower parameters-MSE (supervised loss) in all cases.

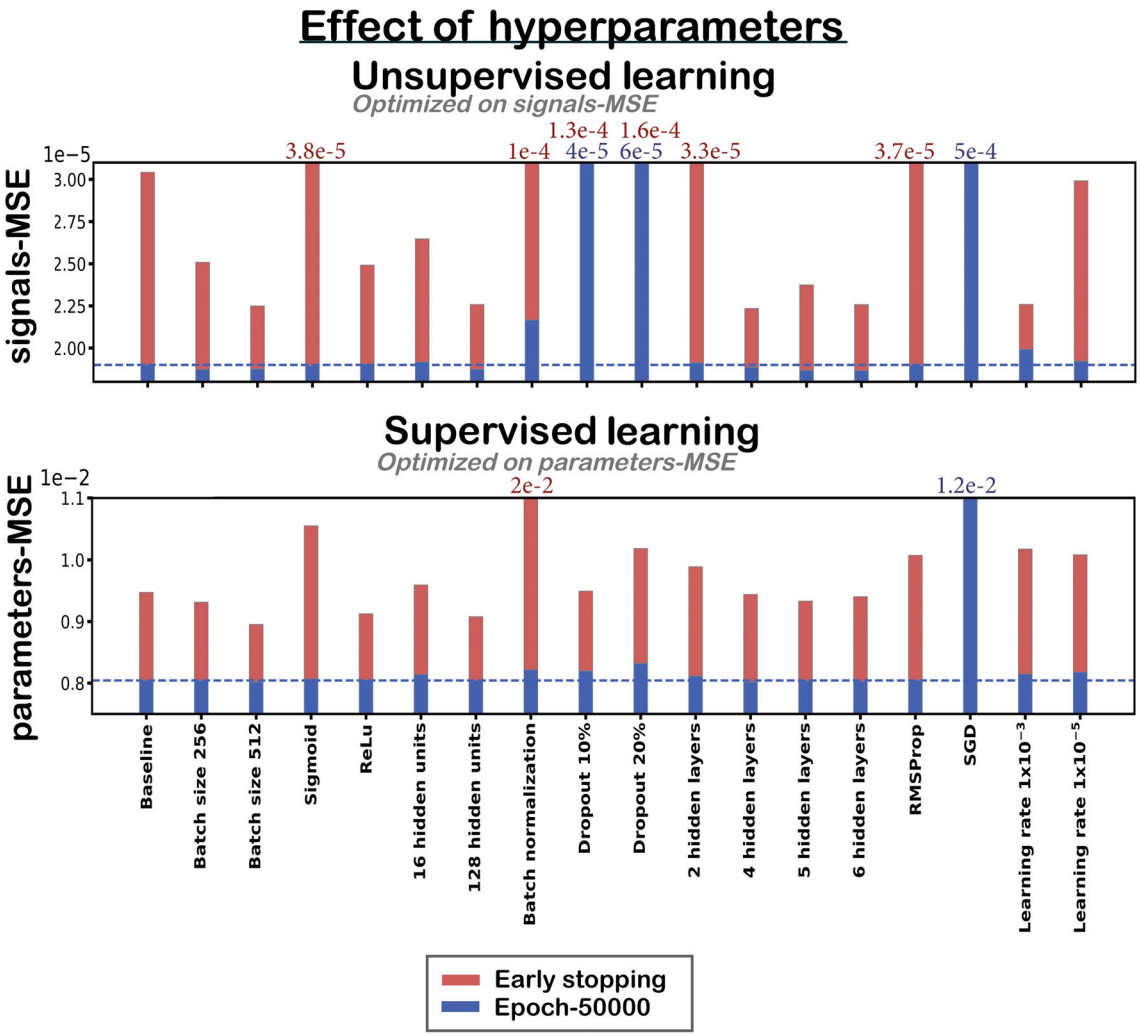

FIGURE 3 Performance of the networks that differed by one hyperparameter with respect to the baseline network that had been declared the most stable in the previous analysis in Figure 2 (see Sections 2.4.1 and 3.1). Results are shown for both learning strategies (unsupervised and supervised) when trained on the uniform distribution and evaluated at the early-stopping point and after training for 50 000 epochs (Epoch-50000). Details about the hyperparameters considered and the baseline set can be found in Section 2.4.2. The blue dashed line represents the loss for the baseline network at Epoch-50000. The values printed above the plots are the loss values at the early-stopping point (red) and at Epoch-50000 (blue) that could not be displayed within the chosen range of the plot. The SGD optimizer had no early-stopping point.

3.2 | Simulations: Hyperparameters

Figure 3 displays loss values at two different validation points during training for the networks that differed by one hyperparameter from the baseline network determined in Section 3.1. Every network performed substantially better when training was extended beyond early stopping. The networks with greater learning capacity (i.e., more hidden layers or more hidden units) showed a marginal improvement in the final loss value. Adding batch normalization resulted in a marginally higher loss. Adding dropout resulted in a substantially higher loss for unsupervised learning, whereas for supervised learning the loss was only marginally higher. Increasing batch size resulted in a marginally lower loss. The choice of activation function had little impact on the loss after extensive training, whereas the choice of optimizer may be an important consideration given the high loss for SGD. In general, the variability in loss values between networks was far greater at the early-stopping point than after extensive training.

3.3 | Simulations: Learning strategy and training distribution

In Figure 4, we show the loss curves of the four networks, as described in Section 2.4.3, tested on the patient distribution test set. Figure 4A shows that for both

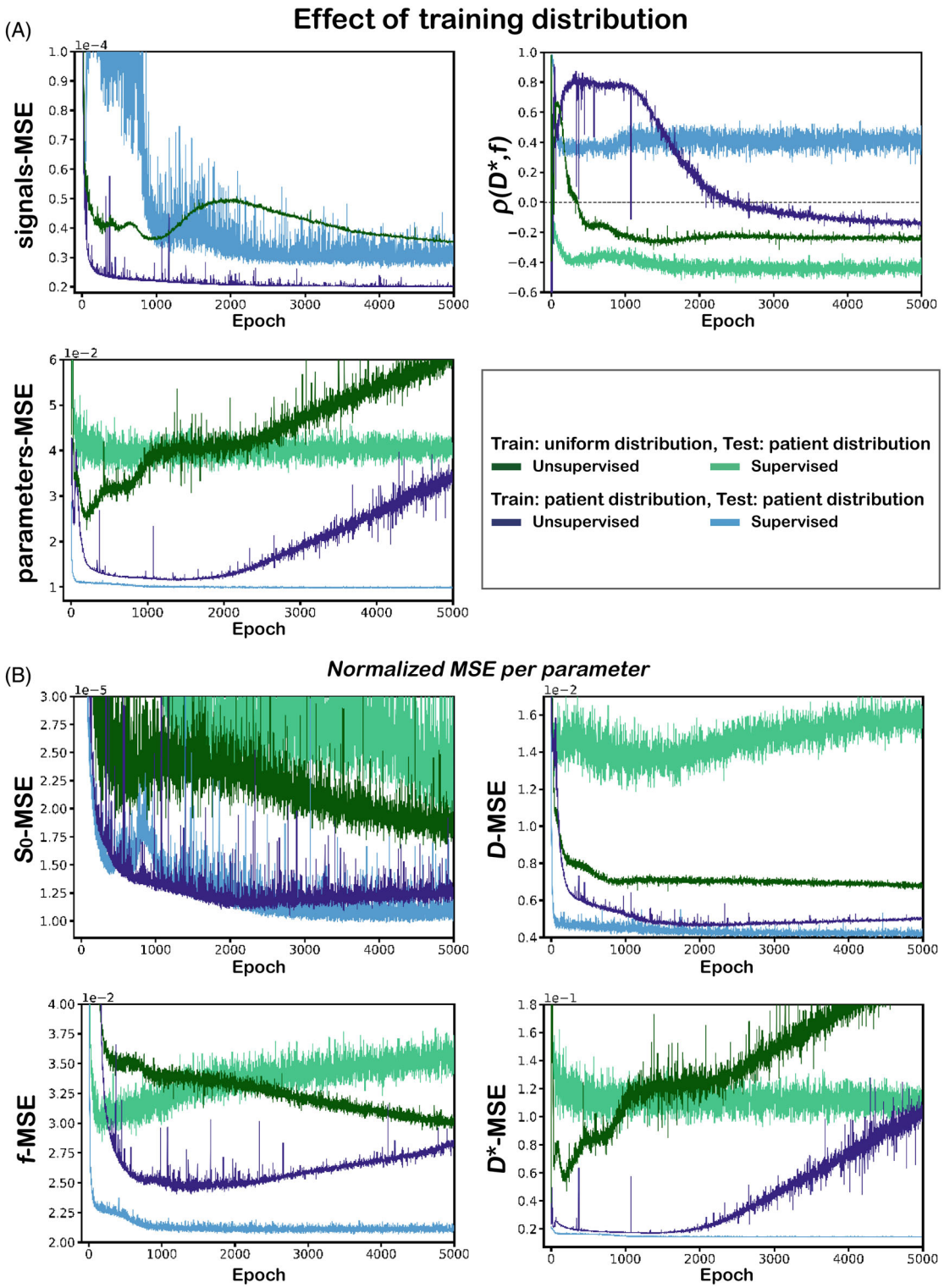

FIGURE 4 (A) Test curves showing the metrics signals-MSE, parameters-MSE, and Spearman's $\rho(D^*,f)$ for the four networks described in Section 2.4.3, that trained either unsupervised or supervised (hidden units = 64 and learning rate = 1×10^{-4}) on synthetic data from either the uniform distribution or patient distribution, and tested on the patient distribution test set over 5000 epochs. (B) Test curves showing the normalized MSE per parameter for each network.

unsupervised and supervised learning, training and testing on the same distribution yields lower loss and faster convergence compared to training and testing on different distributions. For the unsupervised networks, training on the uniform distribution (i.e., different to test set) shows an increase in test loss (signals-MSE) as training is extended (e.g., between epoch 1000 and 2000). Moreover, for unsupervised learning, the undesirable parameter correlation persists longer when training on the patient distribution compared to training on the uniform distribution. Here, early stopping occurred at epoch 181 with $\rho(D^*,f) = 0.73$ when trained on the patient distribution. The increase in parameters-MSE observed in unsupervised learning is mainly due to an increase in errors in D^*, as evidenced by the similarity in plots between D^*-MSE (Figure 4B) and parameters-MSE (Figure 4A), although increases in errors for S_0, D, and f also occur. Note that Min(D^*-MSE) occurs at epoch 208 when training on the uniform distribution and epoch 1552 for the patient distribution. Conversely, in supervised learning, the signals-MSE, parameters-MSE, and $\rho(D^*,f)$ gradually decrease with training length, but parameters-MSE is substantially greater when the testing and training distributions are different.

3.4 | Simulations: Training length

Figure 5 displays scatter plots comparing the estimations with the known ground truth values for the four networks at different validation points during training, when applied to the uniform distribution test set. This figure clearly shows the impact of early stopping, especially when training on the patient distribution, which results in poor accuracy. Early in training, both strategies estimate parameters that are apparently biased toward the mean of the training distributions (see Figure 1), particularly for D^*. As training progresses in unsupervised learning, the estimates corresponding to low S_0 signals (i.e., low SNR) exhibit higher variability and display a distribution tending toward that of LSQ. Conversely, as training progresses in supervised learning, the bias toward the mean of the distributions persists for a greater proportion of data points. These data correspond to either low S_0 (i.e., low SNR), where a bias occurs for every IVIM parameter; low f, which results in uncertainty in D^*; or low D^*, which results in uncertainty in f.

In addition, Figure 5 shows that early in training the parameter values that are overrepresented in the training distribution (see Figure 1) are estimated more accurately than underrepresented parameter values. Furthermore, there is poor-fitting behavior for low D values with high S_0 when training on the patient distribution. This is particularly evident for supervised learning, and is due to a lack of these parameter combinations occurring in the patient data.

Figure 6 displays parameter maps and RMSE maps at different validation points during training of the four networks in a synthetic patient slice. At the early-stopping point, the parameter maps are relatively inaccurate, particularly for D^* when trained on the uniform distribution, and appear smooth and homogeneous. As found in Figure 5, early in training, the parameter estimates in voxels corresponding to low SNR signals (e.g., WM voxels) are biased toward the mean of the synthetic training distribution. As training progresses in unsupervised learning, the parameter estimates in WM voxels exhibit a distribution tending toward that of LSQ, which is also visualized by comparable extremities in D^* and RMSE maps, although f maps are generally more comparable to the segmented fit. There are increases in the RMSE at Epoch-50000 when training on the uniform distribution (red arrows), which are not present when training on the patient distribution.

For supervised learning on the uniform distribution, the estimates for D^* in the WM voxels are biased toward the mean of the training distribution, even after extensive training, which is consistent with Figure 5 for data with low S_0. In addition, the voxels that represent the tumor in this synthetic patient slice (purple arrow in $b = 0$ image) possess very low f values, such that the simulated signals contain little information for estimating accurate D^*. Hence, the supervised network estimates inaccurate D^* close to the mean of this distribution, resulting in high residuals. In contrast, the supervised network trained on the patient distribution shows lower residuals, which are more comparable to the results of the unsupervised networks, because the bias toward the mean of the patient distribution yields values that are closer to the actual ground truth.

3.5 | In vivo

Figure 7 displays loss curves and parameter maps for the four networks applied to a representative slice of the in vivo glioma patient data, as well as corresponding results for the additional unsupervised network trained directly on the in vivo data. These results are in broad agreement with those displayed in Figures 4 and 6 for the simulations. Similar to Figure 4, training and testing on the same dataset yielded the lowest loss term and fastest convergence (Figure 7A). Furthermore, early stopping resulted in sub-optimal solutions at epoch 312 prior to true convergence, where parameters were strongly correlated ($\rho(D^*,f) = 0.94$). Similar to Figure 6, early stopping resulted in seemingly smooth and homogeneous parameter maps. As training was extended, parameter maps tended toward

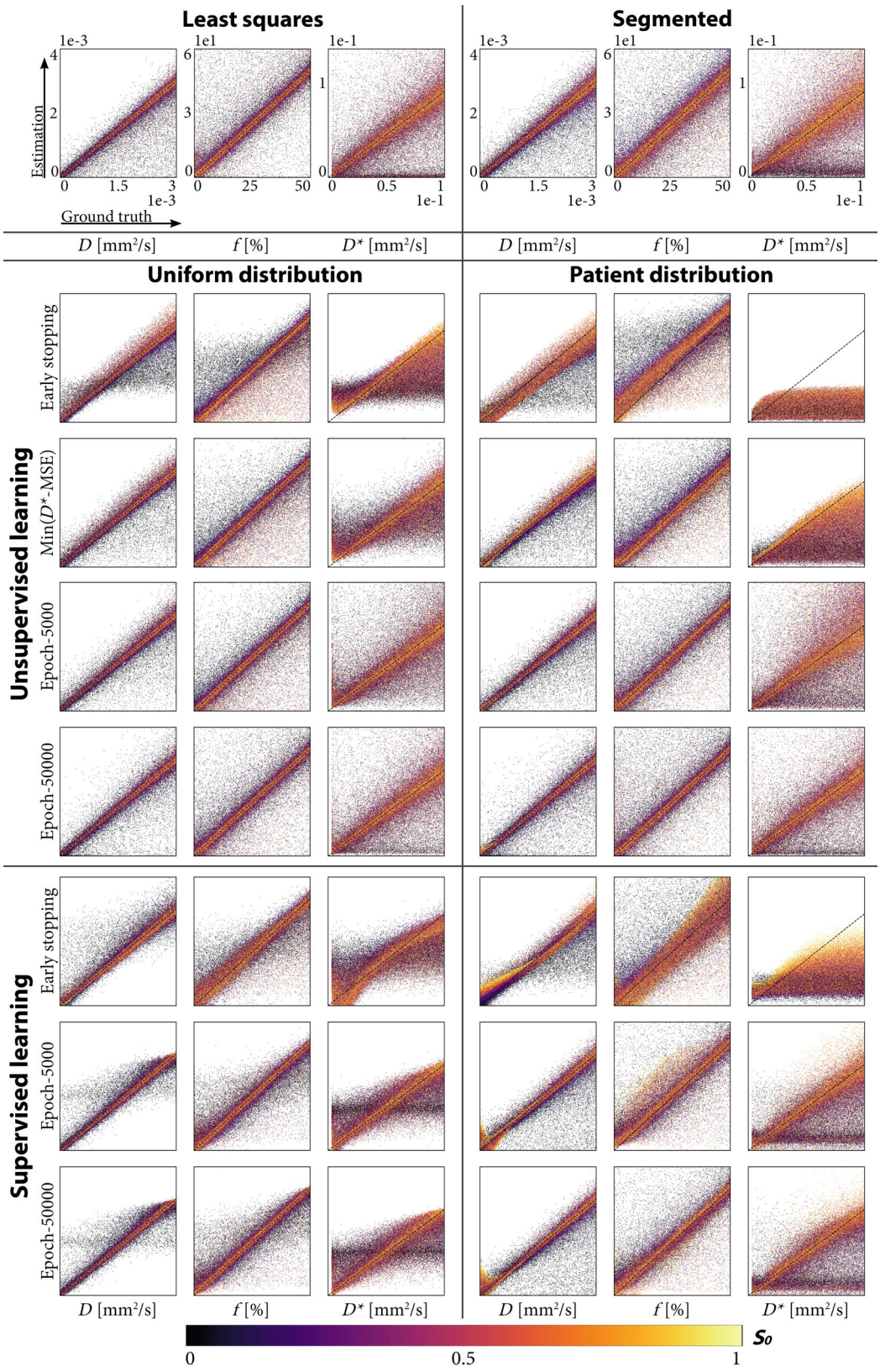

FIGURE 5 Scatter plots of estimated parameter values against ground truth at each of the four validation points (Early stopping, Min(D*-MSE), Epoch-5000, and Epoch-50000) for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution), and tested on the uniform distribution test set. Corresponding plots for least squares and the segmented approach are also shown. All data points are colored by their S_0-value, where $S_0 = 0$ (black) corresponds to SNR = 0 and $S_0 = 1$ (bright yellow) corresponds to SNR = 200.

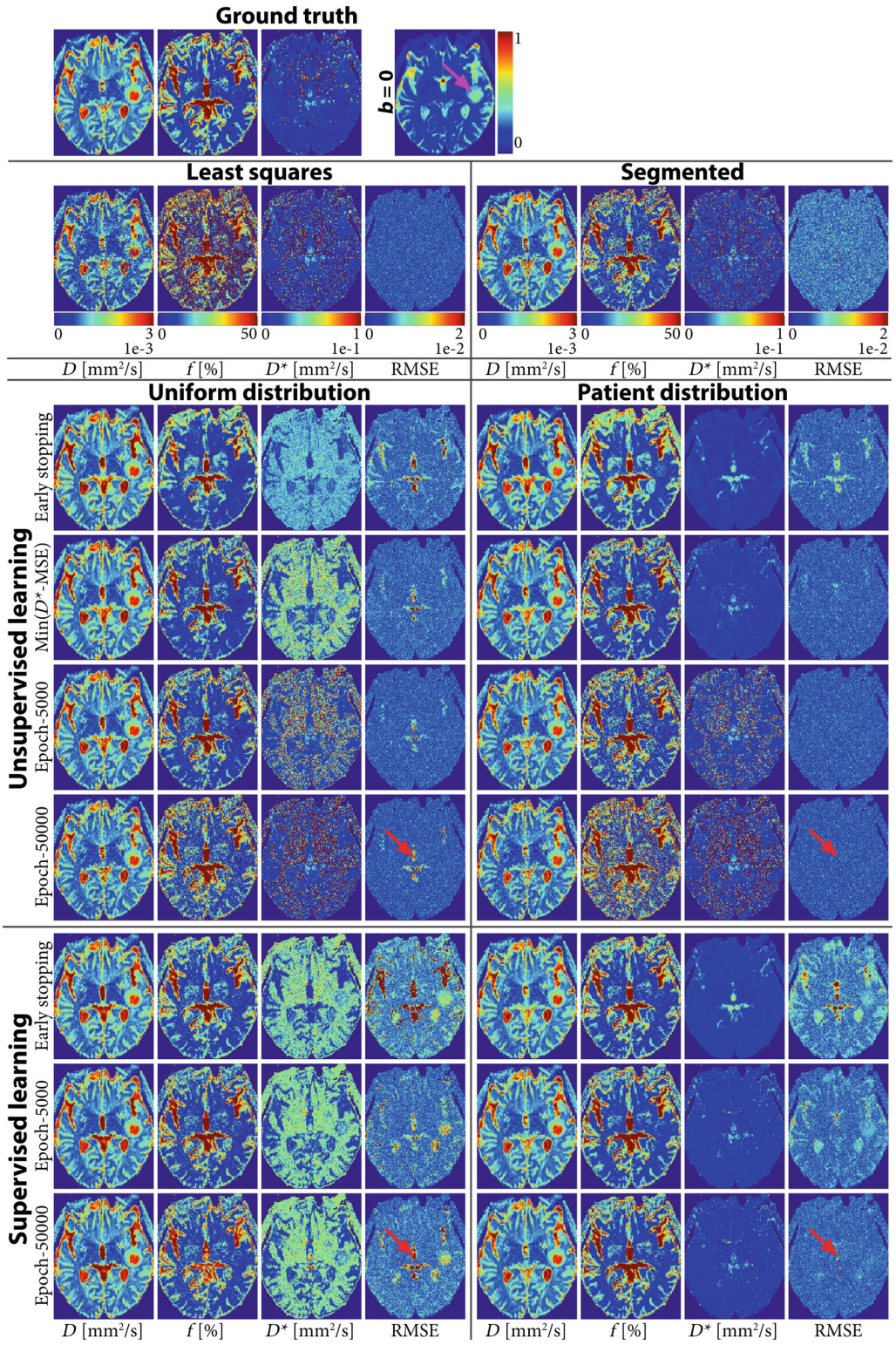

FIGURE 6 IVIM parameter maps and RMSE maps generated for a representative slice of a synthetic glioma patient. Maps are displayed at each of the four stopping points (Early stopping, Min(D*-MSE), Epoch-5000, and Epoch-50000) for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution). The ground truth parameter maps are generated by applying the segmented approach on the real patient data. The red arrows in the Epoch-50000 RMSE maps indicate out-of-distribution data for the uniform distribution, where $D > 3 \times 10^{-3}$ mm^2/s or $f > 50\%$, which lie inside of the domain of the patient distribution, but outside of the domain of the uniform distribution. Corresponding maps for least squares and the segmented approach fitted on the synthetic glioma patient are also shown. The purple arrow in the ground truth $b = 0$ image indicates the location of a tumor.

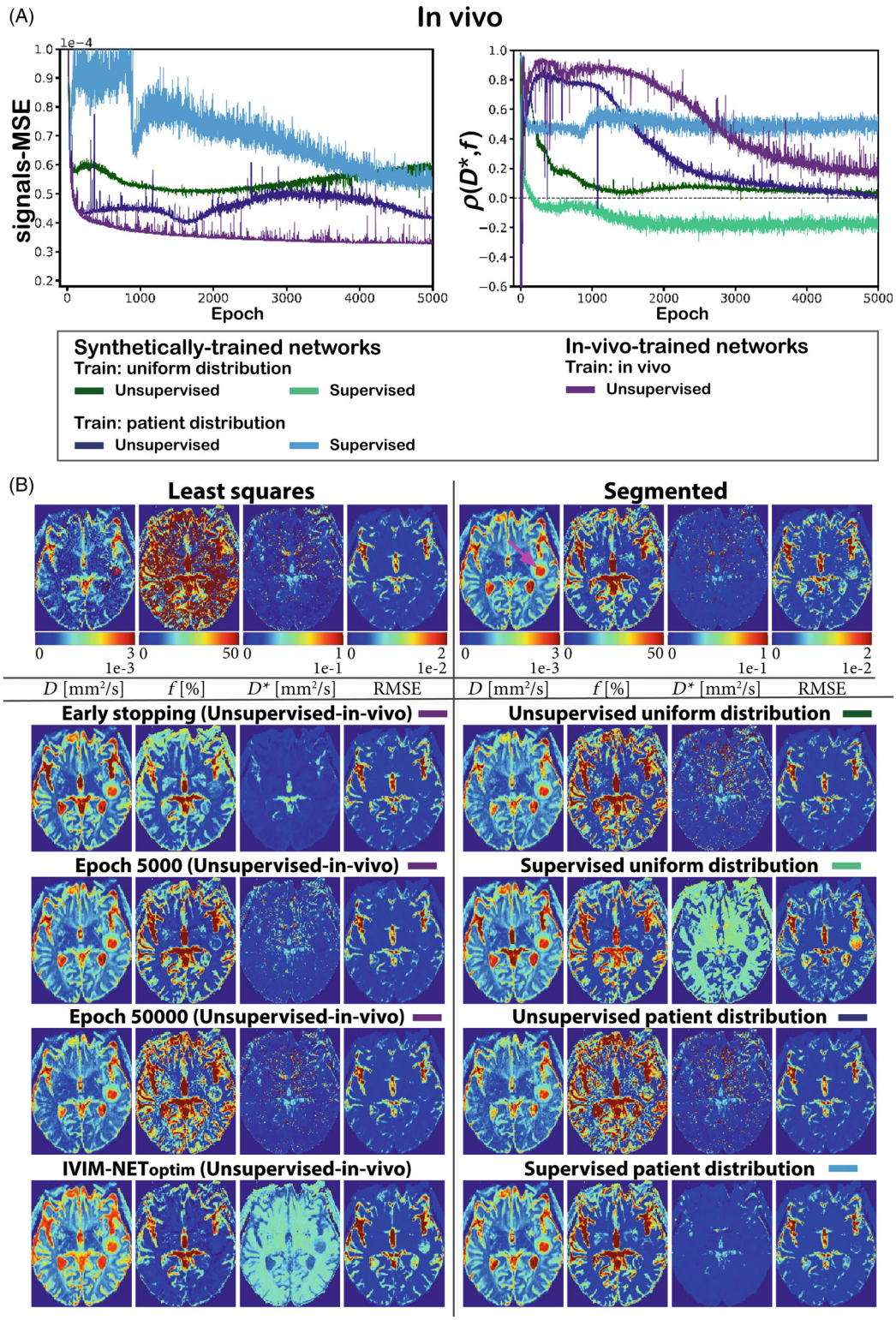

FIGURE 7 (A) Test curves showing the metrics signals-MSE and Spearman's $\rho(D^*,f)$ for the four networks described in Section 2.4.3 (i.e., trained either unsupervised or supervised on synthetic data from either the uniform distribution or patient distribution), and tested on the in vivo data over 5000 epochs. In addition, test curves for an unsupervised network trained directly on the in vivo signals are displayed. (B) IVIM parameter maps and RMSE maps for a representative slice from the in vivo glioma patient data for each of the networks described in (A). For the unsupervised network trained directly on the in vivo data (left), maps corresponding to three stopping points are displayed (Early stopping, Epoch-5000, and Epoch-50000), whereas the maps shown for the four synthetically-trained networks (right) are after training for 50 000 epochs. Corresponding maps are also shown for least squares (top left), the segmented approach (top right) and IVIM-NET$_{optim}$ (bottom left). The purple arrow in the D map for the segmented approach indicates the location of a tumor.

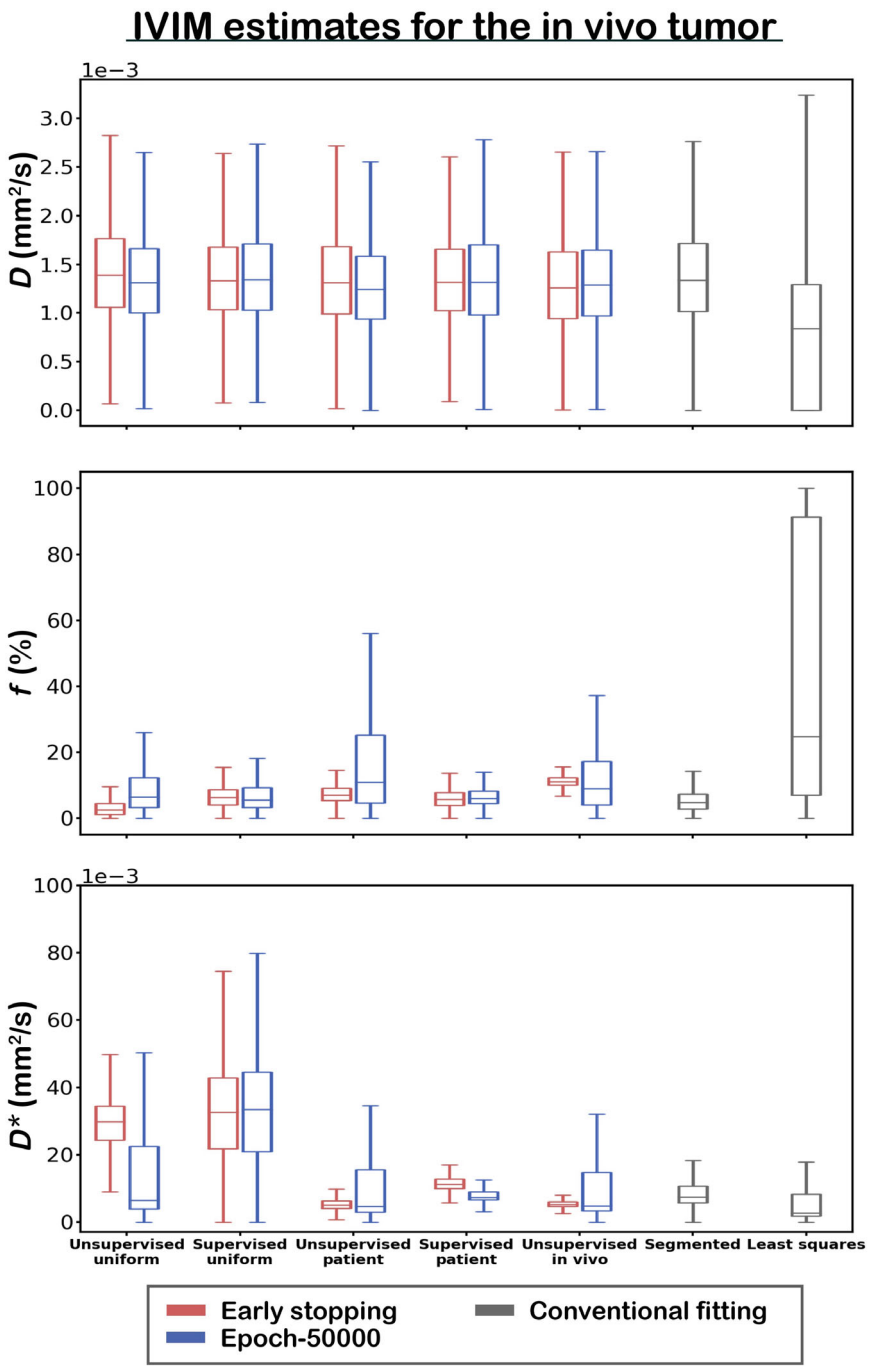

FIGURE 8 Boxplots of the voxel-wise IVIM parameter estimates for in vivo tumor regions of all patients for each of the networks described in Figure 7A, at the early-stopping point and after training for 50 000 epochs (Epoch-50000). Corresponding boxplots are also shown for least squares and the segmented approach. The outliers of the boxplots (i.e., those data exceeding 1.5 times the interquartile range above the upper quartile and below the lower quartile) are not displayed for visualization purposes.

those of LSQ, particularly for D^*. Generally, the residual maps of Figure 7B displayed higher RMSE than those of Figure 6. The parameter maps and RMSE maps of the unsupervised networks that were trained on either the uniform or patient distributions were similar to those for the unsupervised network trained directly on the in vivo data. In contrast, IVIM-NET$_{optim}$ displayed inferior RMSE and high D^* (cf.[18]), and parameter estimates were also considerably different for voxels in the tumor region.

For the supervised networks, a bias toward the mean of the training distribution occurred, which resulted in high RMSE for the network trained on the uniform distribution. In contrast, the parameter maps and RMSE maps of the supervised network trained on the patient distribution are comparable to those for the unsupervised networks. Despite the differences in parameter estimates for the different unsupervised and supervised approaches, RMSE maps were generally similar for low SNR regions (e.g., WM regions).

In Figure 8, we show boxplots of all voxel-wise IVIM parameter estimates for in vivo tumor regions of all patients for each fitting approach. For D, all of the

approaches yield a similar distribution of estimates, with the exception of LSQ, which tends to estimate lower values. Greater variability between approaches is observed for f, and especially for D *, where at the early-stopping point, the networks trained on synthetic data (unsupervised and supervised) estimate the bulk of the parameter values toward the mean of the respective synthetic training distribution (see Figure 1). After extensive training (Epoch-50000), the unsupervised networks display broader distributions of parameter estimates, but apparently with less bias, given the observed similarity between the distributions for each network. Conversely, the bias observed for the supervised networks appears to persist even after extensive training, which is consistent with the results from the simulations. Note also in Figure 8 that LSQ yields a substantially broader distribution of f values in the tumor, which is also reflected in Figure 7 with considerably higher estimates for f, whereas the segmented approach produces a much narrower distribution of lower f values.

4 | DISCUSSION

This work highlights the impact of learning rate, network size, training length and training distribution on unsupervised and supervised DNN IVIM fitting to DWI data. In simulations, we showed that high learning rate and early stopping may lead to correlated parameter estimates and sub-optimal model fitting. We showed that a network with more hidden units increases convergence speed, a lower learning rate results in stable convergence, and extending training beyond early stopping leads to reduced parameter correlations and parameter error. However, extensive training resulted in an increased sensitivity to noise. For unsupervised learning, parameter estimates in regions with low SNR (or low f), that is, with underlying uncertainty, displayed variability similar to LSQ estimates. For supervised learning, parameter estimates with underlying uncertainty were affected strongly by the distribution of the training data, showing biases toward the mean of the training distribution, which resulted in smooth, yet possibly deceptive parameter maps. Moreover, we showed that early in training the networks were more accurate at estimating parameter combinations that were well represented in the training set, compared with underrepresented combinations, but the accuracy of the latter improved with extensive training. The in vivo results were in broad agreement with the simulations, and fitting residuals were almost identical between approaches, particularly for the unsupervised networks. However, the pseudo-diffusion maps varied considerably, demonstrating the difficulty of fitting D* in these regions.

Previous work demonstrated that the use of DNNs for IVIM fitting can lead to erroneous correlation between estimates for the pseudo-diffusion coefficient (D*) and the perfusion fraction (f).[17] Whereas that work resolved those correlations by exhaustive hyperparameter optimization, the present work has demonstrated that extending training beyond early stopping is a simpler alternative, which provides consistent results across many hyperparameters (Figure 3). Of course, this begs the question of how much training is sufficient or optimal. In the approach of Barbieri et al.,[16] the early stopping criterion resulted in sub-optimal convergence at approximately 50 epochs, whereas we trained for 50 000 epochs (i.e., 1000 times longer) to ensure adequate convergence. In the simulations, extensive training in unsupervised learning resulted in an increase in MSE per parameter. Therefore, it could be argued that the point where D*-MSE is lowest might be a good choice of stopping point in unsupervised learning (Figure 4). However, the minimum D*-MSE observed in the simulations is really only an artifact of the persisting bias in the parameter estimates at that point in training (Figures 5, 6), coupled with the chosen metric for gauging accuracy. For unsupervised learning, extensive training was found to ameliorate this bias and also further reduce the fitting residuals, and may therefore be favored in spite of the higher parameter variability. Indeed, there were still improvements after training for 5000 epochs (Figure 5).

In contrast, for supervised learning, bias toward the mean of the training distribution was found to persist even after extensive training, for data with low SNR (or low f). While this approach produced smooth parameter maps, little confidence can be placed in the estimated maps for D* (Figures 6, 7). Similar bias was demonstrated by Gyori et al. for the spherical mean technique diffusion model.[20] Since parameter accuracy for supervised learning is highly dependent on the choice of training data (Figure 4), it may be of benefit to derive ground truth parameter combinations from conventional model fitting, as done for the synthetic patient distribution described in this study, such that the degree of bias is reduced. Nevertheless, it is important to remain aware that this inherent bias, although reduced, will ultimately still be present in regions of low SNR, in spite of the possibly alluring smoothness of the parameter maps.

We showed that the choice of learning rate strongly impacts the convergence behavior of the loss term, particularly for unsupervised learning. A higher learning rate was associated with spiky convergence, which has the potential for poor parameter estimates if the termination of training coincides with such a spike in the loss term. In contrast, a lower learning rate mitigates this spiky convergence at the cost of more training epochs. In this work, we have used a fixed learning rate. An alternative approach would

be to vary the learning rate over the training process, using a so-called learning rate schedule, which could speed up convergence while avoiding spiky behavior at the end of training.

We showed that a network with a higher number of hidden units in each layer resulted in increased convergence speed. Previous approaches[16,17] set the hidden units equal to the number of measured b values (e.g., 16 in this study). Consequently, if a sparse protocol were used for acquiring DWI data, then the hidden units would be very low and could limit the learning capacity of the DNN. Therefore, we recommend using a fixed number of hidden units that is considerably higher than the number of b values (e.g., 64 in this study) in order to avoid a low learning capacity. Adding even more network parameters may make the network more powerful, but this would also extend computation time, and ultimately may not be beneficial if the network is already sufficiently powerful for the application.

In general, we found that altering hyperparameters of the network did not produce a considerable improvement in performance, provided that training was sufficiently long to ensure convergence (Figure 3). The marginally lower loss that resulted from a higher batch size is expected because the gradients calculated at each training step are more reliably defined. Furthermore, we found that using batch normalization, dropout, or the SGD optimizer could result in inferior performance, in particular for unsupervised learning. Therefore, use of these methods must be carefully evaluated. Performance of batch normalization may be improved by increasing batch size.

For the in vivo data (Figure 7), the parameter maps and RMSE maps for the unsupervised networks that were trained on synthetic data were very similar to those for the additional unsupervised network that was trained directly on the in vivo data (Epoch-50000). This suggests that a single unsupervised network may possibly be trained using synthetic data generated by a broad range of parameter combinations, and utilized for every anatomy, provided that the amount of training were sufficiently extensive. It is important to note that these parameter maps were also very similar to those obtained by conventional model fitting. As such, the practical benefit of these unsupervised networks may simply be a possible reduction in computation time. Correspondence between the results of the unsupervised approach and LSQ fitting is expected since both minimize the L2-norm on the signals (signals-MSE, Eq. 1). This is most evident for D^*, whereas for D and f there is perhaps greater similarity with the maps of the segmented approach. One supposition is that, similar to the segmented approach, an unsupervised network may first estimate adequate D, wherefrom it can learn adequate f and subsequently adequate D^*. This may explain why f and D^* are correlated if training is terminated early.

Every network performed poorly on data that lay outside of the respective training distribution, so-called out-of-distribution (OOD) data. In Figure 6, the networks trained on the synthetic data from the patient distribution showed low RMSE for all voxels in the representative slice, especially after extensive training. Conversely, the networks trained on synthetic data from the uniform distribution showed high RMSE for the voxels associated with parameters that lay outside of the training distribution (i.e., $D > 3\times10^{-3}$ mm^2/s and $f > 50\%$; red arrows). The impact of these OOD data in the test set was also observed in the plot of signals-MSE in Figure 4A, where the unsupervised network trained on the uniform distribution displayed an increase in signals-MSE as training progressed. However, if the patient distribution test set was instead truncated using the same bounds as the uniform distribution, the difference in signals-MSE between the two unsupervised networks was found to be minimal (data not shown). One further example of the consequences of OOD data is shown in Figure 5, where the networks trained on the synthetic patient distribution show inaccurate estimation of low D, despite high S_0. This is due to the lack of certain parameter combinations occurring in the patient distribution, while being present in the uniform distribution test set, which were therefore not seen during training.

Note that the poorly-fitted OOD data in Figure 6 correspond to parameter values that are not physically realistic (i.e., $D > 3\times10^{-3}$ mm^2/s and $f > 50\%$). This demonstrates one limitation of this study, which uses conventional model fitting to generate and provide synthetic patient data for training and testing. Specifically, we used the two-step segmented approach, since this was found to be more robust to noise than LSQ for the in vivo data considered in this study (Figures 7 and 8). These conventional methods are applied under the assumption that the in vivo data are well approximated by a bi-exponential function. However, the IVIM model does not account for other possible biophysical processes or imaging artifacts (e.g., subject motion) that may influence the signal. These factors make it difficult to generate synthetic data and corresponding ground truth parameters that provide an accurate representation of real data and the underlying biophysics. Future work should consider how to provide more appropriate ground truth data for training.

DNNs are promising for enhancing medical imaging technologies, yet for IVIM fitting, performance is strongly dependent on certain design choices, which can make results difficult to interpret, particularly in clinical practice. Clinical assessment is therefore necessary

to determine whether DNNs provide a clinical advantage over other advanced estimators. Comparison to histological characterization could be invaluable, where, for instance, gliomas have been shown to produce a pathologic microvascular network in the periphery through neoangiogenesis, which results in hyperperfusion to provide nutrients for tumor growth, whereas the center of the tumor consists of dead tissue that is hypoperfused.[2,26] A ring of high f values at the location of the tumor would be consistent with this histological finding, which in Figure 7 is most clearly shown for the unsupervised approaches at Epoch-50000. Tumor classification could be another important consideration of clinical assessment,[27] where the choice of fitting approach may be a key factor in the pursuit of quantitative biomarkers (Figure 8).

This work focused on voxel-wise fully-connected network implementations for IVIM fitting, but a detailed investigation of the impact of other network architectures has not been considered. For instance, Vasylechko et al.[28] proposed an unsupervised convolutional neural network based on the U-NET architecture.[29] It is expected that the incorporation of spatial information into the training of a neural network for IVIM fitting may improve performance, but further investigation of this is beyond the scope of the present article and has been left for future work. Lastly, this study mainly focused on the brain, which is a particularly challenging organ for IVIM fitting given the typically low f values. A comprehensive comparison study for other anatomies should be pursued in future work, although similar findings regarding extensive training and biases are expected.

5 | CONCLUSIONS

Using both simulations and in vivo data from glioma patients, we have explored the potential impact of key training features in unsupervised and supervised learning for IVIM model fitting. The main motivation of this work was not to develop a network that could improve IVIM parameter estimation. Rather, our aim was to investigate the possible limitations of voxel-wise deep learning IVIM fitting, by surveying a broad range of reasonable design choices. For both learning strategies, we demonstrated that a high learning rate, small network size, and early stopping can result in sub-optimal solutions and correlated parameters. We demonstrated that extending training beyond early stopping can resolve these correlations and reduce parameter error, providing an alternative to exhaustive hyperparameter optimization. However, extensive training resulted in increased sensitivity to noise, where unsupervised parameter estimates displayed variability similar to estimates from conventional LSQ fitting. In contrast, supervised parameter estimates demonstrated improved precision, but showed strong biases towards the mean of the training distribution. Hence, visual assessment alone is not sufficient for evaluating the quality of the estimates. While the apparent sensitivity to noise in unsupervised learning may be undesirable, it could be argued that the corresponding variability observed in the parameter maps is indicative of the underlying uncertainty, which is indeed useful information. This uncertainty is exemplified by the contrasting D^* maps between approaches, despite the similar residual maps, and illustrates the difficulty in estimating D^* in the brain. Therefore, while deep-learning-based model fitting is a promising tool for IVIM parameter estimation, the impact of design choices on fitting performance and biases must be carefully evaluated.

ACKNOWLEDGMENTS

Misha P. T. Kaandorp, Frank Zijlstra, and Peter T. While gratefully acknowledge support from the Research Council of Norway under FRIPRO Researcher Project 302624.

DATA AVAILABILITY STATEMENT

To stimulate wider clinical implementation of IVIM, we have made the code for our unsupervised and supervised networkstrained on the uniform distribution synthetic data set available on GitHub: https://github.com/Mishakaandorp/Explore_Deep_Learning_IVIM.

ORCID

Misha P. T. Kaandorp https://orcid.org/0000-0002-7340-8256
Frank Zijlstra https://orcid.org/0000-0002-9184-7666
Peter T. While https://orcid.org/0000-0003-2602-0758

REFERENCES

1. Le Bihan D, Breton E, Lallemand D, Aubin M, Vignaud JLM. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*. 1988;168:497-505.
2. Federau C, Meuli R, O'Brien K, Maeder P, Hagmann P. Perfusion measurement in brain gliomas with intravoxel incoherent motion MRI. *Am J Neuroradiol*. 2014;35:256-262. doi:10.3174/ajnr.A3686
3. Klaassen R, Steins A, Gurney-Champion OJ, et al. Pathological validation and prognostic potential of quantitative MRI in the characterization of pancreas cancer: preliminary experience. *Mol Oncol*. 2020;14:2176-2189. doi:10.1002/1878-0261.12688
4. Zhu L, Wang H, Zhu L, et al. Predictive and prognostic value of intravoxel incoherent motion (IVIM) MR imaging in patients with advanced cervical cancers undergoing concurrent chemo-radiotherapy. *Sci Rep*. 2017;7:1-9. doi:10.1038/s41598-017-11988-2

5. Lemke A, Laun FB, Klau M, et al. Differentiation of pancreas carcinoma from healthy pancreatic tissue using multiple *b*-values. *Invest Radiol*. 2009;44:769-775. doi:10.1097/rli.0b013e3181b62271
6. Notohamiprodjo M, Chandarana H, Mikheev A, et al. Combined intravoxel incoherent motion and diffusion tensor imaging of renal diffusion and flow anisotropy. *Magn Reson Med*. 2015;73:1526-1532. doi:10.1002/mrm.25245
7. Luciani A, Vignaud A, Cavet M, et al. Liver cirrhosis: Intravoxel incoherent motion MR imaging-pilot study. *Radiology*. 2008;249:891-899. doi:10.1148/radiol.2493080080
8. Federau C. Intravoxel incoherent motion MRI as a means to measure in vivo perfusion: a review of the evidence. *NMR Biomed*. 2017;30:1-15. doi:10.1002/nbm.3780
9. Federau C, Sumer S, Becce F, et al. Intravoxel incoherent motion perfusion imaging in acute stroke: initial clinical experience. *Neuroradiology*. 2014;56:629-635. doi:10.1007/s00234-014-1370-y
10. Pekar J, Moonen CTW, van Zijl PCM. On the precision of diffusion/perfusion imaging by gradient sensitization. *Magn Reson Med*. 1992;23:122-129. doi:10.1002/mrm.1910230113
11. While PT. Advanced methods for IVIM parameter estimation. In: Bihan D, Iima M, Federau C, Sigmund EE, eds. *Intravoxel incoherent motion (IVIM) MRI: Principles and applications*. Pan Stanford Publishing; 2018:449-484.
12. Meeus EM, Novak J, Withey SB, Zarinabad N, Dehghani H, Peet AC. Evaluation of intravoxel incoherent motion fitting methods in low-perfused tissue. *J Magn Reson Imaging*. 2017;45:1325-1334. doi:10.1002/jmri.25411
13. Barbieri S, Donati OF, Froehlich JM, Thoeny HC. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. *Magn Reson Med*. 2016;75:2175-2184. doi:10.1002/mrm.25765
14. Orton MR, Collins DJ, Koh DM, Leach MO. Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling. *Magn Reson Med*. 2014;71:411-420. doi:10.1002/mrm.24649
15. While PT. A comparative simulation study of bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI. *Magn Reson Med*. 2017;78:2373-2387. doi:10.1002/mrm.26598
16. Barbieri S, Gurney-Champion OJ, Klaassen R, Thoeny HC. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magn Reson Med*. 2020;83:312-321. doi:10.1002/mrm.27910
17. Kaandorp MPT, Barbieri S, Klaassen R, et al. Improved unsupervised physics-informed deep learning for intravoxel incoherent motion modeling and evaluation in pancreatic cancer patients. *Magn Reson Med*. 2021;86:2250-2265. doi:10.1002/mrm.28852
18. Spinner GR, Federau C, Kozerke S. Bayesian inference using hierarchical and spatial priors for intravoxel incoherent motion MR imaging in the brain: analysis of cancer and acute stroke. *Med Image Anal*. 2021;73:102144. doi:10.1016/j.media.2021.102144
19. Bertleff M, Domsch S, Weingärtner S, et al. Diffusion parameter mapping with the combined intravoxel incoherent motion and kurtosis model using artificial neural networks at 3 T. *NMR Biomed*. 2017;30:1-11. doi:10.1002/nbm.3833
20. Gyori 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*. 2021;87:932-947. doi:10.1002/mrm.29014
21. Federau C, O'Brien K. Increased brain perfusion contrast with T2-prepared intravoxel incoherent motion (T2prep IVIM) MRI. *NMR Biomed*. 2015;28:9-16. doi:10.1002/nbm.3223
22. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016: 1–14.
23. Le Bihan D, Iima M. Diffusion magnetic resonance imaging: what water tells us about biological tissues. *PLoS Biol*. 2015;13:1-13. doi:10.1371/journal.pbio.1002203
24. Kingma DP, Ba JL. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015: 1–15.
25. Tukey JW. Exploratory data analysis. *Biometrics*. 1977;33:768.
26. Kumar V, Abbas AK, Fausto N, Aster JC. *Robbins and Cotran pathologic basis of disease*. Elsevier Health Sciences; 2014.
27. Vidić I, Jerome NP, Bathen TF, Goa PE, While PT. Accuracy of breast cancer lesion classification using intravoxel incoherent motion diffusion-weighted imaging is improved by the inclusion of global or local prior knowledge with bayesian methods. *J Magn Reson Imaging*. 2019;50:1478-1488. doi:10.1002/jmri.26772
28. Vasylechko SD, Warfield SK, Afacan O, Kurugol S. Self-supervised IVIM DWI parameter estimation with a physics based forward model. *Magn Reson Med*. 2022;87:904-914. doi:10.1002/mrm.28989
29. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Computer-assisted intervention. In: Navab N, Hornegger J, Wells WM, Frangi AF, Hutchison D, eds. *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI*. Dimensions; 2015:234-241.

How to cite this article: Kaandorp MPT, Zijlstra F, Federau C, While PT. Deep learning intravoxel incoherent motion modeling: Exploring the impact of training features and learning strategies. *Magn Reson Med*. 2023;90:312-328. doi: 10.1002/mrm.29628