nature portfolio

Peer Review File

Multi-Parameter Molecular MRI Quantification using Physics-Informed Self-Supervised Learning

Corresponding Author: Dr Or Perlman

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors present a self-supervised approach for quantifying biophysical parameters, in a variety of quantitively MRI applications, using the combination of an MLP and a differential forward simulation. The use of efficient code and GPU acceleration in the multi-pool spin simulation allows the authors to train the quantification network in an impressively short timeframe. The weights of the trained network are then frozen and the inference on unseen brain data can then be performed in seconds. This flexible approach should be applicable to many quantitively MRI sequences, a highly active topic in MRI.

General comments:
- I commend the authors for the rigorous validation of their method in multiple qMRI applications, both in phantom and in vivo. However, I'm missing an analysis of the potential biases introduced by the NN. At various points in the manuscript, I get the impression that the NN introduces bias to ultimately achieve a lower variance of the parameter estimates. This is a common occurrence in Bayesian approaches that can achieve a lower variance compared to maximally efficient unbiased estimators at the cost of introducing bias. In many qMRI applications, biased estimates are a very steep price to pay for reduced computation time. Examples of where I think to perceive bias are listed below in the "specific comments" section.

- Even though the results shown are impressive, I am unsure about the novelty of the presented method. Yes, the here presented approach is novel in the field of qMRI but this kind of training loop has been used in NN applications. It would also strengthen the manuscript if the authors compared it to supervised NN-based fitting, especially if a better quantification can be achieved besides just faster training.

- I think the training data distribution requires further discussion since, in my mind, it is a bonus and a drawback of the presented approach. Dictionary-based NN training requires the estimation of the application-specific parameter distribution, something with is a-priori unknown. The here presented approach circumvents this by training on the in-vivo data itself, thus perfectly matching the parameter distribution. Along with the benefits come however some drawbacks. The quantification of lesions is often of vital importance, but their volume is often significantly less than 1% of the total brain volume. It remains uninvestigated if the presented approach can work well in these important regions which are poorly represented in the total parameter distribution. Dictionary-based approaches on the other hand have the flexibility of over-representing the expected values of lesions (if they are known) or simply using a very broad parameter distribution that covers both the healthy and diseased states.

Specific comments:
- Phantom results (Fig. 2): Stating an excellent agreement between the methods is in my opinion not a nuanced enough analysis of the presented results. In most vials, especially vials A, D, and F dictionary fitting appears closer to the ground truth (GT) value. More specifically, in (f) a large and obvious difference (~10-20%) is shown between the dictionary matching and NBMF. Considering that dictionary matching is a maximum likelihood estimator it can be considered as a proxy for the GT.

- Generally, the Pearson coefficient of correlation and the SSIM don't seem like particularly sensitive or easy-to-interpret metrics. I would challenge the authors to find a more meaningful metric considering biophysical parameters are being compared.

- Single-subject NBMF (Fig. 5): This comment is related to both the bias as well as the training data distribution mentioned

in “general comments”. Again “good agreement” seems like an oversimplified analysis. One can see a clear visual difference in the parameter estimates from the single-subject training versus the reused network. This seems like a major observation and a potential pitfall of the presented technique. The argument for computation speed builds mostly on the assumption that the network generalizes well between volunteers. Unfortunately, the choice of metrics and the plotting ranges in e) seem purposefully chosen to diminish these differences.

- Computation times (Table 1): I don’t understand why the supervised training based on a dictionary would be significantly slower compared to the proposed training. In both cases a likely similar number of forward simulations have to be performed, the only difference is if the parameter distribution follows the in-vivo distribution or some heuristically chosen one. Note, that the "dictionary" doesn’t have to be sampled on a Cartesian grid and can follow any a-prior distribution. The simulation pipeline in ref. 38 is likely much less optimized than the here proposed code thus it seems unfair to cite their computation times and the computation using the author's code should be stated.

- Noise model and loss function: What is the motivation behind using uniform noise? This is a very uncommon noise distribution for MRI. Although uniform noise is used to model undersampling artifacts in MRF, undersampling is not noise, and modeling it as uniform noise is a strong assumption. In line with this question, why was an L1-norm used instead of a L2-norm?

- VBMF: Assuming proper convergence, then non-linear least squares should be very close to a maximally efficient estimator. How can NBMF outperform VBMF even though both fit the exact same biophysical model? In my mind, this can only happen because NBMF is NOT unbiased. The lower R2 in brain parenchyma for VBMF seems to point in this direction as well.
I don’t agree with the given explanation (“This a known attribute of neural networks…”) since the forward model (M(t) simulator) is the same in both cases.

Reviewer #2

(Remarks to the Author)
Thank you to the authors for submitting this very interesting and well written work. This work presents a flexible, deep learning alternative to dictionary based methods (which have to precompute the signal evolution for a given protocol in advance) by embedding the simulation of the physics of CEST imaging into the training in a differentiable way. Furthermore, the proposed framework is self-supervised, allowing for training on a single subject. I believe that the incorporation of the ODEs into the training process is of great interest to the medical imaging community for the enforcement of model/data consistency in inverse problem solving where forward model computation is not straightforward (e.g. as in MR image reconstruction). The validation and statistical analysis seemed comprehensive and appropriate (except for one point which I detail below). Reproducibility of this work is feasible as the authors say they will release source code and sample data after publication.

Major Comments:
I had some questions/concerns about the comparison to dictionary based methods which I wanted to raise.
1. One main difference to dictionary based methods seems to be that the dictionary is generated from a fixed set of varying parameters of interest from a fixed protocol, using a Bloch simulator to generate signals from the corresponding parameter set. In contrast, the proposed approach is more flexible to the set of parameters and the protocol since the Bloch simulator can simply be adapted to different parameter sets/protocols. Furthermore, From Table 1, it seems that the signal generation is faster using the proposed GPU solution of the ODEs. In this case, is it possible to use the proposed Bloch simulator to generate dictionaries for the dictionary approach more quickly, and then train a network on this data? In that case, this version of DL dictionary matching would seem to take a similar amount of time as the proposed approach, with quality of the estimates being the difference.

2. The other main difference is the fact that the dictionary based methods (as described in the text) rely solely on the synthetically generated dictionary for fitting, whereas the proposed method (at least for the first subject) trains in a self-supervised way on the actual data of the acquired subject using the flexible Bloch simulator to enforce consistency of the measurements to the physics of the acquisition. I noted that there are no side by side comparisons (in-vivo) of a deep learning dictionary fitting vs. the proposed approach, which I thought would be a natural/fair comparison, especially if the point above is answered affirmatively. I would be interested to know why the authors chose to omit this/would they expect the fitting to be similar? I can think of an argument for the dictionary based method being better: the underlying parameters can be chosen freely and hence can be chosen to encompass a wide range of possibilities, whereas the proposed approach is constrained to the parameters available in the subject used for training. While the proposed approach seems to generalize well to unseen parameters, this is something worth investigating further. However, I also see the argument that the proposed approach would be better, as mentioned in the text: the proposed approach is trained using the actual acquired data in conjunction with the ODE model and thus could plausibly be more robust on real data (with regard to data consistency) than the dictionary approaches, which only train on synthetic data. In any case, unless the authors have a compelling argument against comparing to a DL dictionary matching approach using the same protocol as tested in the manuscript, I think this comparison is warranted, and would strengthen the paper to show the effect of self-supervised training/inclusion of data consistency in training.

3. I would also somewhat disagree with the characterization in the paper that the proposed approach is necessarily better for routine or clinical use, based on considerations of time. If the model can be reused for different subjects, then the only time

difference to DL dictionary approaches would be the initial training time. In practice, I would think that standardized protocols would be constructed for routine use; hence, the need for flexibility or the concern that generating a dictionary would take too long would be diminished, since it would only need to be done once for each protocol. Since the inference time is also similar for both methods, it would seem both methods would be comparable for routine use. For research, where one wants to test a variety of different protocols, I can definitely appreciate the benefit of the time savings. Furthermore, if the model needed to be trained for each subject, the fitting/training time of around 18min for the proposed approach could also be infeasible for clinical use, though of course this could be reduced with optimization, better hardware, etc. I would be interested to know the author's comments on these points.

Minor Comments:

1. Given the above discussion, do the authors think that combining the self-supervised and dictionary approaches (e.g. training a single network using a mix of both strategies) would make sense?

2. It seems that with a differentiable/efficient forward model and a pretrained parameter estimator, one could, for example, output a map which uses both at inference (e.g. use the pretrained parameter estimator as a prior within a Tikhonov regularization). Do the authors think this would improve the parameter estimates/be worth the additional computation?

3. In Figure 5, can the authors comment on the outliers in the SSIM plot for k_{ssw}? 0.6-0.7 SSIM is quite low compared to the performance for the rest, are there any factors to explain this (e.g. slice position, etc).

In conclusion, I think this work makes a valuable contribution to the literature for the computationally efficient incorporation of a complex, ODE forward model into inverse problem solving for biomedical parameter estimation (in this case for CEST). However, I think that the paper can be strengthened by adding a direct in-vivo comparison to DL dictionary based approaches (e.g. using the GPU forward model for data generation) as well as a more nuanced discussion of the advantages of each approach.

Reviewer #3

(Remarks to the Author)
This paper describes a physics-based deep learning framework for rapid model fitting of the human brain tissue proton spin properties. In general, the manuscript is well-presented and easy to follow. Please find below some concerns for further improving the manuscript.

- On page 2, "Excellent agreement was obtained between the NBMF-estimated and known L-arginine concentrations (Fig. 2a-b),..." From Fig. 2c, it seems that dictionary matching shows better consistency with the known concentrations. Please also supplement the dictionary matching-concentration results and compare them with the NBMF-concentration results. Also, provide the dictionary matching-pH results and compare them with the NBMF-pH results.
- On page 3, when discussing the "Computational complexity and timing," it is essential to mention the in-plane matrix size and slice number for the whole brain data, as the voxel number affects the total computational timing.
- On page 6, in addition to the readout module, please also include the key parameters for acquiring the WASABI (B0 & B1), saturation recovery (T1), and multi-echo sequences (T2).
- It is suggested to add quantitative maps of other human subjects, possibly as supplementary figures, to demonstrate the robustness of the proposed methods.
- Regarding the use of 3.5 ppm concentration as the amide signal (fs, ks), this should be noted as a limitation in the Discussion section, as there may be other contributing factors.
- In Figure 2, for the tube CDF, the pH was varied from 4.0 to 5.0, which is far from the typical physiological pH of around 7.2. Please include tubes with higher pH values closer to physical conditions to better mimic the in vivo case.
- In Figure 2, in plot f, only half of the "pH" label is shown. Please correct this.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the authors for their extensive analyses and their rigorous investigation of the critiques raised during the previous revision! The presented techniques now seem more transparent, and the benefits and drawbacks are more clearly defined. I only have the following minor critiques:

Minor comments:
- “Even with a faster generation, the transfer of synthetic data-trained NN to experimental data raises concerns about biased estimates, a steep price to pay for a reduced computation time.”: The formulation with “steep price to pay” seems unscientific and should probably be changed in the paper. I apologize for using the same language in the last revision and I should have been less figurative.

- “Supplementary Figure 3. (a, c): A comparison between the voxelwise training/test modeling errors of self-supervised learning (NBMF) (a) and dictionary-based supervised (c) learning.”: Are the labels (a) and (c) swapped in the caption? Otherwise, both the results and the rest of the caption do not make sense to me. Adding titles on the left and right sides of the figure could help to avoid confusion.

Reviewer #2

(Remarks to the Author)
I thank the authors for their comprehensive rebuttal and modification/additions to the text.

The additions to the supplementary material and change of language in the text have greatly improved the paper and have clearly and fairly shown the advantages (and some disadvantages) of the proposed approach of rapid self-supervised learning of parameter maps using a novel GPU accelerated forward model, in comparison to e.g. a supervised dictionary based approach based on simulated signals.

In my opinion, the paper is suitable for publication.

Reviewer #3

(Remarks to the Author)
The authors addressed my comments very well. The manuscript has been improved significantly.

Revised manuscript COMMSPHYS-24-1318.R1

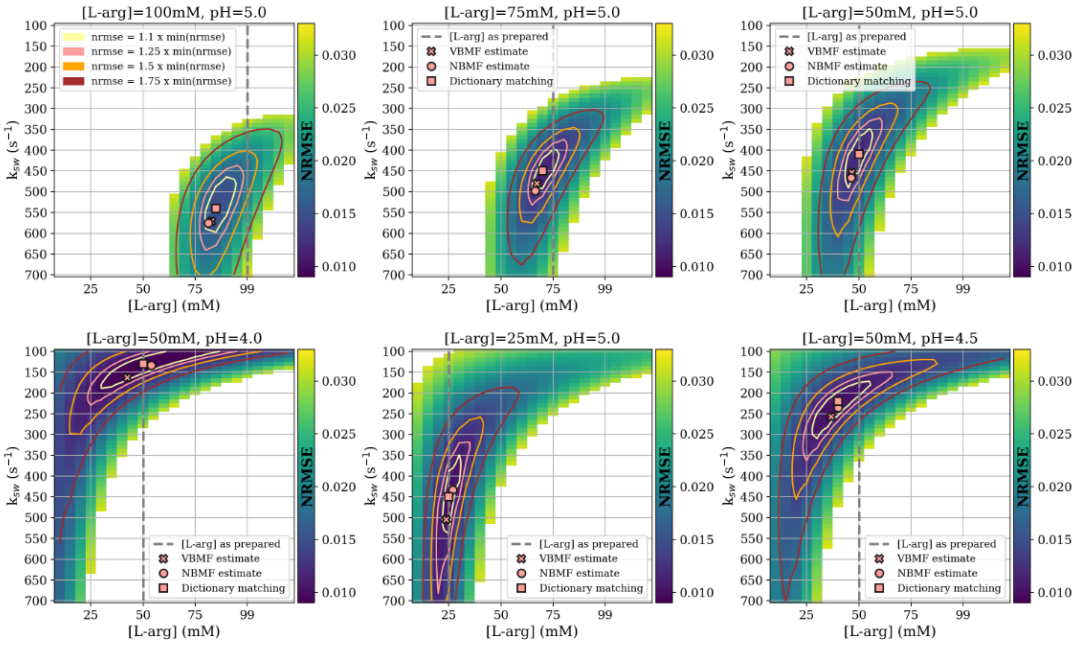
Response to Reviewer 1

We thank the reviewer for the positive response to our work and the thoughtful and constructive comments. In addition to the response below, please find related changes in the revised manuscript highlighted in yellow.

1. Comment: "I commend the authors for the rigorous validation of their method in multiple qMRI applications, both in phantom and in vivo. However, I'm missing an analysis of the potential biases introduced by the NN. At various points in the manuscript, I get the impression that the NN introduces bias to ultimately achieve a lower variance of the parameter estimates. This is a common occurrence in Bayesian approaches that can achieve a lower variance compared to maximally efficient unbiased estimators at the cost of introducing bias. In many qMRI applications, biased estimates are a very steep price to pay for reduced computation time. Examples of where I think to perceive bias are listed below in the "specific comments" section."

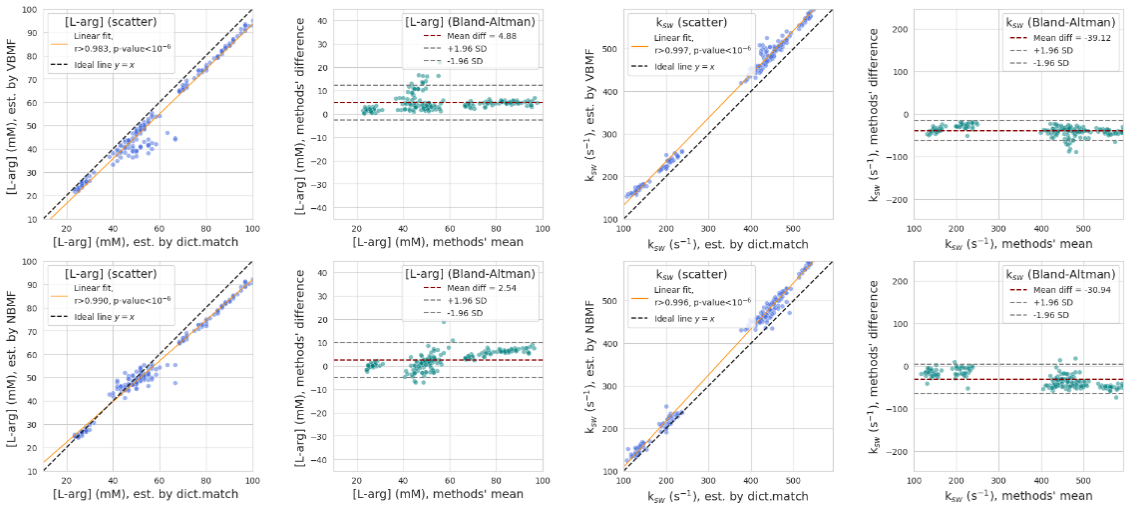
Response: An extensive bias analysis was performed and provided in the new supplementary information sections #1 and #2. The new findings include:

1. An in-vitro analysis (where the ground truth exists from orthogonally measured lab equipment), demonstrating that the self-supervised NN (NBMF) introduces a certain bias, yet it generally falls inside the same narrow confidence interval (<10% from the minimum possible error) as the reference traditional dot-product dictionary matching (Supplementary Fig. 1).



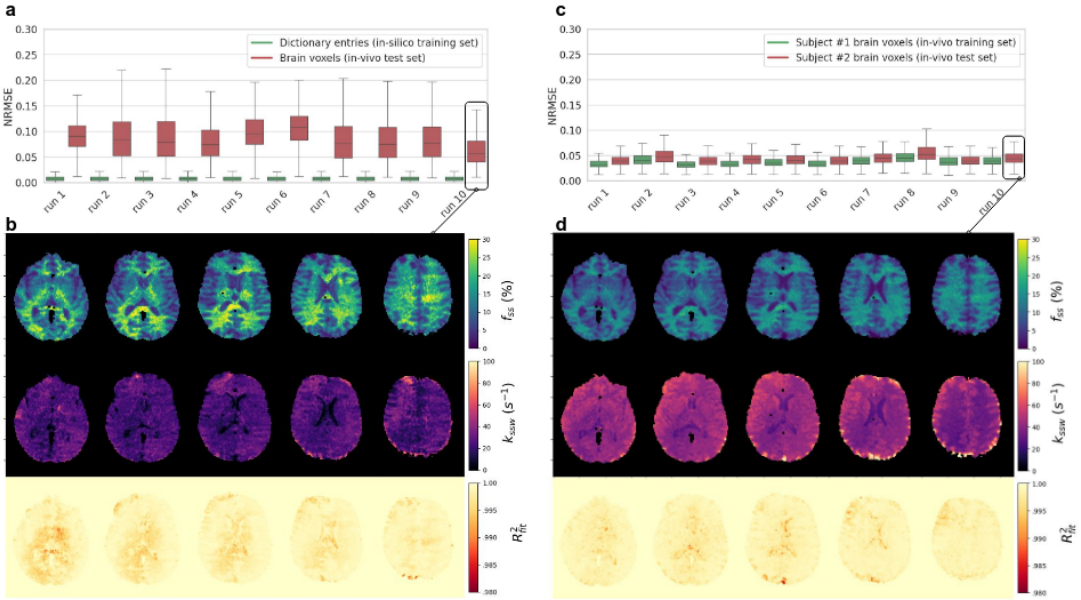
Supplementary Figure 1. Single-voxel uncertainty and bias analysis in-vitro. Each panel describes a single voxel taken from the center of the respective L-arginine vial shown in Fig. 2. Dot-product matching was performed with all dictionary entries in the {[L-arg], k_{sw}} parameter space to create modeling error maps, dictionary-based minima estimates, and NRMSE at c×NRMSE$_{min}$, c=1.1, 1.25, 1.5, 1.75.

In addition, a significant correlation was obtained between self-supervised parameter estimates and traditional dot-product matching (Supplementary Fig. 2, Bottom panel, Pearson's r>0.99, p<10^{-6}):



Supplementary Figure 2. In-vitro CEST experiment analysis: comparing VBMF (**top**) and NBMF (**bottom**) quantitative estimates of concentration (**left**) and exchange rate (**right**) to those obtained using traditional dictionary matching[1].

2. An in-vivo analysis (where a perfect ground truth does not exist, yet the consistency with the raw data can be calculated) highlighted a clear improvement in data consistency for self-supervised learning compared to dictionary-based supervised learning (Supplementary Fig. 3).



Supplementary Figure 3. **(a, c)**: A comparison between the voxelwise training/test modeling errors of self-supervised learning (NBMF) (**a**) and dictionary-based supervised (**c**) learning. The modeling NRMSE capturing the circuit error of processing the normalized measured signal **m** through the neural reconstructor $\mathscr{R}$ and the physical forward model $\mathscr{F}$ (applied on the resulting parameter estimates $\mathscr{R}(\mathbf{m})$) is computed as $||\mathscr{F}(\mathscr{R}(\mathbf{m})) - \mathbf{m}||$, aggregated across training/test set entries and shown for 10 runs of the pipeline (differing in randomness realizations at initialization and training noise injection). **(b, d)**: Quantitative parameter estimates $\{f, k\} = \mathscr{R}(\mathbf{m})$ for five representative slices alongside R^2_{fit}=1-NRMSE^2 maps. For the supervised learning (**b**), the reconstructor with the best post-hoc test results out of $\{\mathscr{R}_j; j = 1..10\}$ is selected, while for NBMF (**d**), a typical run is used.

Please see additional details and discussion, in the response to specific comments #2, #6, and #9.

2. Comment: ”Even though the results shown are impressive, I am unsure about the novelty of the presented method. Yes, the here presented approach is novel in the field of qMRI but this kind of training loop has been used in NN applications. It would also strengthen the

manuscript if the authors compared it to supervised NN-based fitting, especially if a better quantification can be achieved besides just faster training."

Response: A comparison between self-supervised fitting (NBMF) and supervised learning for in-vivo human data was conducted and analyzed in the new Supplementary Section 2. An improved quantification was obtained using NBMF (Supplementary Fig. 3). While there is no perfect ground truth in-vivo, the NRMSE of the modelling error (representing the consistency between the contrast-weighted images reconstructed based on the quantitative parameter maps and the originally acquired contrast-weighted raw images) was evidently lower for NBMF (Supplementary Figure 3c) compared to supervised learning (Supplementary Figure 3a). Moreover, when repeating the quantification procedure ten times using either NBMF or the supervised approach, the repeatability of the reconstructed parameter maps and their agreement with the raw data was improved and more consistent (see the narrower distribution of the box-plot analysis in Supplementary Figure 3c). Additional analysis details and discussion are available in Supplementary Section 2.

The core novelty of the paper resides in the implementation and application of automatic differentiation to make any Bloch-based biophysical model amenable to large scale gradient-based learning, whether of local parameters (model fitting) or the NN weights parametrizing a mapping, in a self-supervised fashion. This unlocks improved consistency with the actual subject data within clinically relevant scan times.

3. Comment: "I think the training data distribution requires further discussion since, in my mind, it is a bonus and a drawback of the presented approach. Dictionary-based NN training requires the estimation of the application-specific parameter distribution, something which is a-priori unknown. The here presented approach circumvents this by training on the in-vivo data itself, thus perfectly matching the parameter distribution. Along with the benefits come however some drawbacks. The quantification of lesions is often of vital importance, but their volume is often significantly less than 1% of the total brain volume. It remains uninvestigated if the presented approach can work well in these important regions which are poorly represented in the total parameter distribution. Dictionary-based approaches on the other hand have the flexibility of over-representing the expected values of lesions (if they are known) or simply using a very broad parameter distribution that covers both the healthy and diseased states."

Response: The following paragraph was added to the Discussion chapter (first section):
"Supervised NN training using a synthetic signal dictionary requires the estimation of the application-specific parameter distribution, which is often unknown in advance. The self-supervised (NBMF) approach circumvents this challenge by training on the in-vivo data itself, offering improved parameter distribution matching. When it comes to re-using the trained network on new unseen subjects, one drawback of this approach lies in its reliance on previously represented proton exchange parameters. Dictionary-based approaches, on the other hand, have the flexibility for representing the expected abnormality values (if they are known) or simply using a very broad parameter distribution that covers both the healthy and diseased states. A future patient cohort investigation is needed to examine the clinical utility of transferring the self-supervised quantification approach, when trained on healthy volunteers, for quantification in unseen pathology (e.g., small lesions)."

4. Comment: "Phantom results (Fig. 2): Stating an excellent agreement between the methods is in my opinion not a nuanced enough analysis of the presented results. In most vials, especially vials A, D, and F dictionary fitting appears closer to the ground truth (GT) value. More specifically, in (f) a large and obvious difference (~10-20%) is shown between the dictionary matching and NBMF. Considering that dictionary matching is a maximum

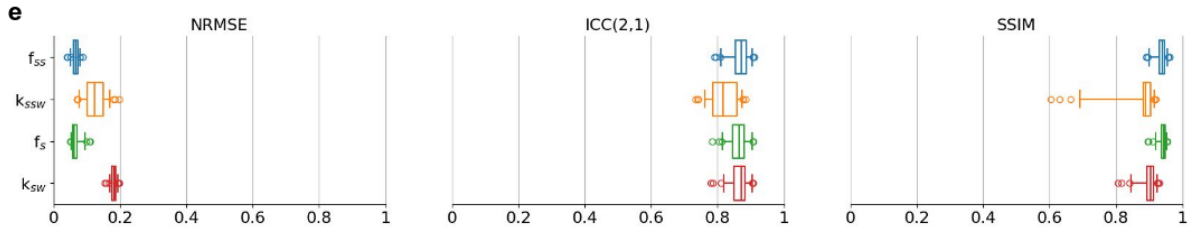
likelihood estimator it can be considered as a proxy for the GT." As requested by reviewer #3, we added

Response: The description of the in-vitro agreement between the methods was toned down (Results chapter, "In-vitro CEST quantification" section). An extended analysis of the in-vitro study is now provided in Supplementary Section 1. While some deviations are observed (as indicated by the reviewer), the overall agreement between the NBMF/VBMF methods and dot-product matching yielded a Pearson's $r > 0.983$, $p<10^{-6}$ ($r > 0.99$ in most cases, Supplementary Fig. 2). The exact per-vial reconstructed parameter values are provided in Supplementary Table 1, and Bland-Altman analysis is provided in Supplementary Fig. 2.

As noted by the reviewer, there is no absolute ground truth for proton exchange rate estimation (unlike compound concentration), and dot-product serves as a maximum likelihood estimator. As reqested by reviewer #3, the dictionary-based dot-product results were added to Fig. 2c,d and are shown side by side next to the results obtained using NBMF.

5. Comment: "Generally, the Pearson coefficient of correlation and the SSIM don't seem like particularly sensitive or easy-to-interpret metrics. I would challenge the authors to find a more meaningful metric considering biophysical parameters are being compared."

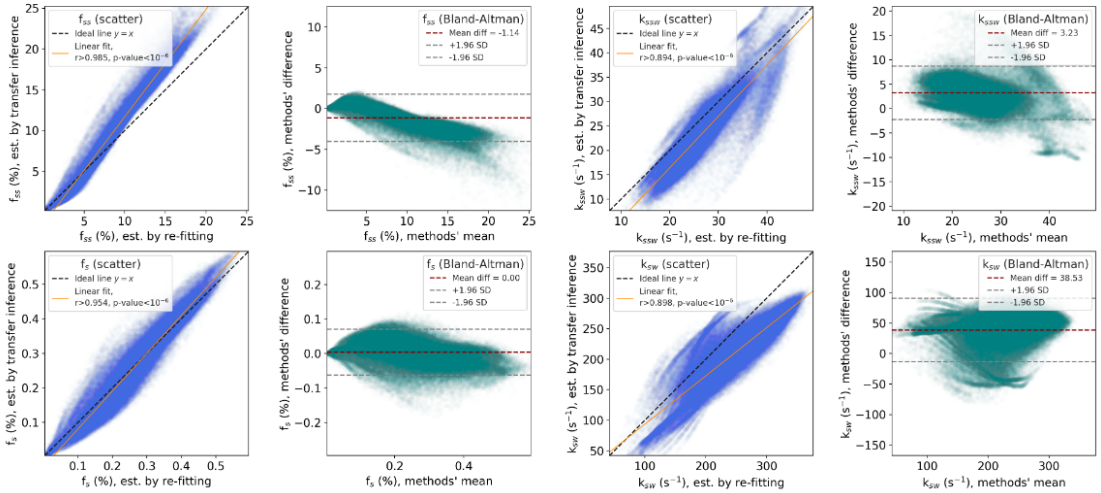
Response: While re-examining relevant literature, we found that various quantitative MRI papers utilized the NRMSE metric alongside SSIM [e.g., assisting ref. 1, 2]. In addition, some papers utilized the ICC metric [e.g., assisting ref. 3, 4], as a better estimate of the adherence obtained by the reconstructed parameter map to reference values. Accordingly, Fig. 5e now shows the NRMSE, ICC, and SSIM values across 50 different slices in the subject's whole brain:



The values are also explicitly mentioned in the revised results section called "Quantifying the semisolid-MT and CEST proton exchange parameters in the human brain":

"The resulting agreement metrics (Fig. 5e) were NRMSE: 7±1%, 12±3%, 7±1%, and 18±1%; Intraclass correlation coefficient ICC(2,1): 0.87±0.03, 0.82±0.04, 0.86±0.03, 0.86±0.03; SSIM: 0.93±0.02, 0.87±0.07, 0.94±0.01, 0.90±0.03, for the f_{ss}, k_{ssw}, f_s, and k_{sw}, respectively. Additional analysis is provided in Supplementary Figure 4."

Bland-Altman and scatter plots were also added in Supplementary Fig. 4.



Supplementary Figure 4. Comparing the two modes of NBMF operation (controlled fitting and neural inference) for in-vivo human brain semisolid-MT (**top**) and CEST-MRF (**bottom**) quantification. Estimated proton volume fraction (**left**) and exchange rate (**right**) values are shown.

Assisting reference 1. Lu, Lan, et al. "Acceleration of Simultaneous Multislice Magnetic Resonance Fingerprinting With Spatiotemporal Convolutional Neural Network." *NMR in Biomedicine* 38.1 (2025): e5302.

Assisting reference 2. Balsiger, Fabian, et al. "Spatially regularized parametric map reconstruction for fast magnetic resonance fingerprinting." *Medical image analysis* 64 (2020): 101741.

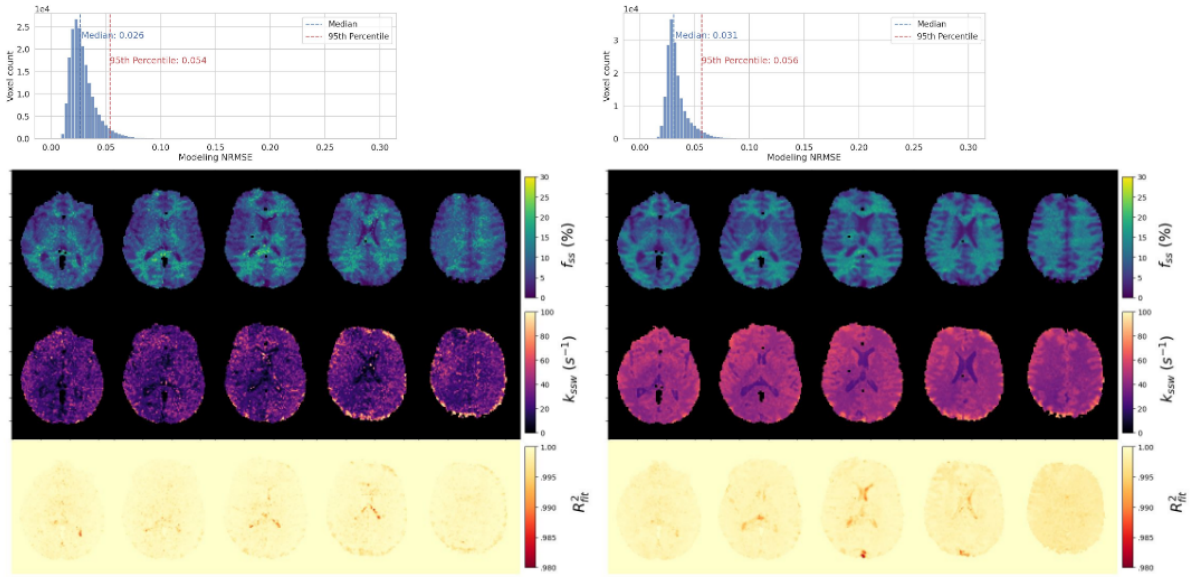
Assisting reference 3. Buonincontri, Guido, et al. "Three dimensional MRF obtains highly repeatable and reproducible multi-parametric estimations in the healthy human brain at 1.5 T and 3T." *Neuroimage* 226 (2021): 117573.

Assisting reference 4. Choi, Moon Hyung, et al. "3D MR fingerprinting (MRF) for simultaneous T1 and T2 quantification of the bone metastasis: initial validation in prostate cancer patients." *European Journal of Radiology* 144 (2021): 109990.

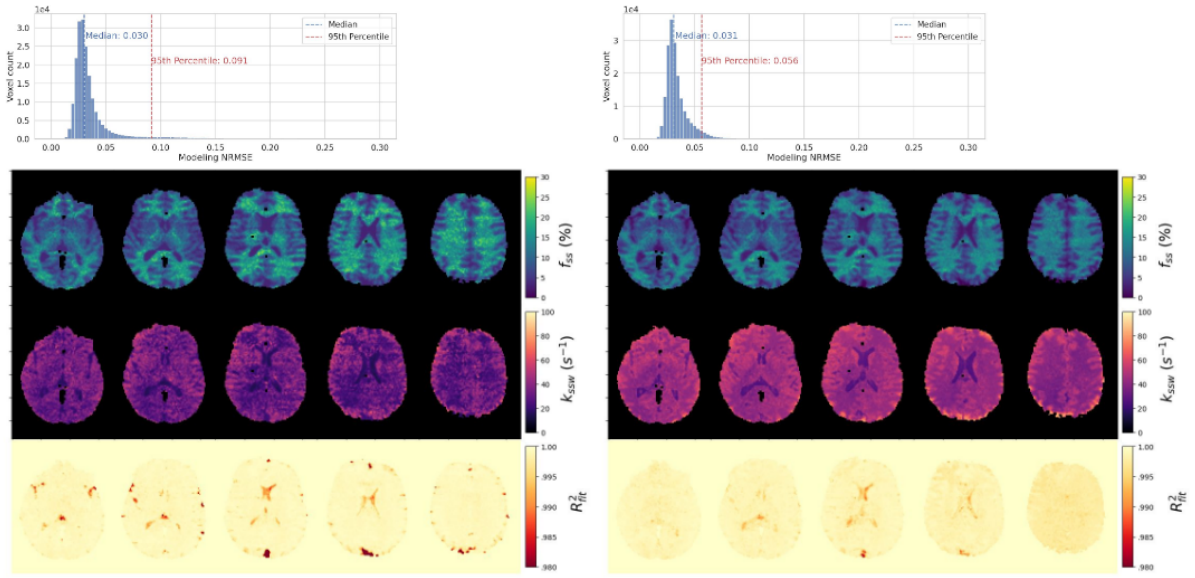
6. Comment: "Single-subject NBMF (Fig. 5): This comment is related to both the bias as well as the training data distribution mentioned in "general comments". Again "good agreement" seems like an oversimplified analysis. One can see a clear visual difference in the parameter estimates from the single-subject training versus the reused network. This seems like a major observation and a potential pitfall of the presented technique. The argument for computation speed builds mostly on the assumption that the network generalizes well between volunteers. Unfortunately, the choice of metrics and the plotting ranges in e) seem purposefully chosen to diminish these differences."

Response: Bias analysis: An extensive bias, confidence, and uncertainty analysis is now provided in Supplementary Section 1. As shown in Supplementary Fig. 1, the quantitative estimates of compound concentration and proton exchange rate by either NBMF or VBMF generally fall within a similar parameter estimation confidence region that represents a data consistency error that deviates by less than 10% from the minimal possible NRMSE. Please see additional details and discussion in Supplementary Section 1.

A related analysis of the bias in human subjects is provided in the new Supplementary Section 3, where Both NBMF and VBMF demonstrate a similar consistency with raw imaged data and dot-product estimates. This is demonstrated by the visually similar R^2_{fit} error maps (bottom panel in Supplementary Fig. 5 and 6) and is quantified by the consistency in NRMSE distribution (histograms at the top panel in Supplementary Figs. 5 and 6).



Supplementary Figure 5. Bottom: Semisolid MT-MRF reconstruction using Dictionary Matching (**left**) and NBMF (**right**), representative slices. **Top**: Corresponding distributions of the voxelwise modeling error.



Supplementary Figure 6. Bottom: Semisolid MT-MRF reconstruction using VBMF (**left**) and NBMF (**right**), representative slices. **Top**: Corresponding distributions of the voxelwise modeling error.

Single-subject training versus the reused network: We agree with the reviewer that some differences between the reconstruction are visible, and the agreement description was toned down in the manuscript (results section). To further visualize and quantify the differences between the two reconstruction modes we added a box-plot analysis using the Normalized Root-Mean-Square (NRMSE) and the Intraclass Correlation Coefficient (ICC) metrics in Fig. 5e. In addition, scatter and Bland-Altman plots were calculated and added as Supplementary Fig. 4.

Importantly, we acknowledge the limitations of the quantification transfer-mode performance in the Discussion, and offer future routes for improvement:

"Notably, while this work presented a proof of concept for rapid inference by a network trained on a single subject, robustness and consistency of the transfer leaves a clear room for improvement (Fig. 5 and Supplementary Fig. 4). Future work could study different NN architectures with spatial awareness (via convolutional or attention layers), as well as larger datasets composed of multiple subjects and a combination of both dictionary-based synthetic data and real-world scans. Subsequent efforts could also improve the modeling accuracy by taking under consideration the contributions of additional proton pools (such as amine and guanidinium) to the 3.5 ppm signal."

We respectfully disagree that our argument for the novelty and/or speed-up achievement builds mostly on the assumption of generalization between subjects. The fundamental speed-up (via autodiff of ODE solutions) is that of the simple fitting ("VBMF") of a generic BM-governed evolution, augmented with neural network learning the mapping ("NBMF") to more quickly converge to smoother and data-consistent results within a timeframe which is practical in itself (<30 min, compared to several hours using conventional model fitting [ref. 48]). The transferability to new subjects based on the learnt self-supervised reconstruction network serves as an additional proof-of-concept option to benefit from the developed scheme.

7. Comment: "Computation times (Table 1): I don't understand why the supervised training based on a dictionary would be significantly slower compared to the proposed training. In both cases a likely similar number of forward simulations have to be performed, the only difference is if the parameter distribution follows the in-vivo distribution or some heuristically chosen one. Note, that the "dictionary" doesn't have to be sampled on a Cartesian grid and can follow any a-prior distribution. The simulation pipeline in ref. 38 is likely much less optimized than the here proposed code thus it seems unfair to cite their computation times and the computation using the author's code should be stated."

Response: Table 1 was expanded to facilitate two separate analyses:
(a) A comparison with previously published reports, highlighting the motivation to explore acceleration strategies (especially in the context of traditional, non-neural-network-based Bloch fitting).
(b) A unified benchmarking using our own implementation of dictionary generation, matching, and supervised learning, which is further described in detail in Supplementary Section 2. All variants were implemented using the accelerated formulation developed as part of this work (GPU-based JAX formulation of the Bloch McConnell numerical solution).

(a) NBMF computational complexity compared to previously reported studies of semisolid-MT and CEST quantification

Method / Stage	Numerical BM model direct fitting	Dictionary-based pattern matching	NN-based mapping (supervised learning)	**NBMF** (MT+CEST)
Preparatory steps using synthetic data + 1^{st} subject mapping	None	Hours to days; e.g., 21.3h (x56 nodes) for 79M-entries dictionary generation[12]	Hours; e.g., ~4.5h for 60K-entries dictionary generation and network training[38]	<30min (fit&train)
N^{th} subject tissue parameters quantification (N>1)	Hours to days; e.g., 6h recon of a 256 × 256 pixel slice[48]	Hours to days; e.g., 2.31h for reconstructing a 128 × 128 pixel slice[12]	~1 sec[12, 38] (NN inference)	1sec (inference) **or**: <30min (re-fitting)

(b) Unified benchmarking of all methods using the accelerated approach developed as part of this work*

Method / Stage	Autodiff-based (AD) model fitting **(VBMF)**	Dictionary-based pattern matching	NN-based mapping (supervised learning)	Self-supervised NN + AD **(NBMF)**
Preparation + 1^{st} subject mapping	3 min	25 sec& (generation) + 93 sec (matching)	25 sec (generation) + 58 sec (training)	3 min (fit&train)
N^{th} subject mapping (N>1)	3 min	93 sec (matching)	1 sec (NN inference)	1 sec (inference)
Consistency with raw per-subject data	✓ (up to convergence)	✓ (up to grid resolution)	× (empirically, in our test)	✓ (up to prior)
Implicit NN-based smoothing prior	×	×	✓	✓

*A GPU-based JAX formulation of the Bloch McConnell ODEs numerical/analytical solutions. The comparison was based on a semisolid MT MRF protocol using a dictionary consisting of 400K entries and a whole-brain quantification task (194K voxels), see additional details in Supplementary Section 2.

& Dictionary generation for the 79M entries CEST and semisolid MT cartesian grid dictionary used in ref[12] took 80 min, exhibiting a linear scaling of the computational complexity.

Notably, the accelerated formulation developed as part of this paper also facilitates a rapid use of supervised NN training, with comparable timing to self-supervision, given that a non-cartesian sampling grid is used for dictionary generation. The benefit of using the self-supervised approach compared to supervised training is highlighted in Supplementary Figure 3, in the context of consistency with the raw acquired data and per-subject discrepancy minimization.

This information was added to the Results chapter, under the section “Computational complexity, timing, and comparison with alternative approaches”.

8. Comment: “Noise model and loss function: What is the motivation behind using uniform noise? This is a very uncommon noise distribution for MRI. Although uniform noise is used to model undersampling artifacts in MRF, undersampling is not noise, and modelling it as uniform noise is a strong assumption. In line with this question, why was an L1-norm used instead of a L2-norm?”

Response: We apologize for the mistaken description of the method detail, which is now corrected in the revised version (Methods section). A Gaussian noise and the L2 loss were used.

9. Comment: “VBMF: Assuming proper convergence, then non-linear least squares should be very close to a maximally efficient estimator. How can NBMF outperform VBMF even though both fit the exact same biophysical model? In my mind, this can only happen because NBMF is not unbiased. The lower R2 in brain parenchyma for VBMF seems to point in this direction as well. I don’t agree with the given explanation (“This a known attribute of neural networks…”) since the forward model (M(t) simulator) is the same in both cases”.

Response: We generally agree with the reviewer’s analysis and accept that the previously given concise explanation needed to be replaced and expanded. This previous description (Supplementary section 5) was replaced by the following lines:

“The deviation of NBMF from pure voxelwise discrepancy-minimization can be interpreted as a "biased" estimation, that is, with respect to a Maximum-Likelihood under an assumption of a physical model with additive Gaussian noise. The enforcement of a neural mapping reflects an implicit regularizing prior leading to Maximum A-Posteriori (MAP) estimation[8,9]. Such a data-driven prior encapsulates information from other voxels, including aggregated insight on the distribution and the model gap.”

Supporting Ref. 8:

8. Xu, A. & Heagy, L. J. A Test-Time Learning Approach to Reparameterize the Geophysical Inverse Problem with a Convolutional Neural Network. IEEE Trans. Geosci. Remote. Sens. 62, 1–12 (2024).

Supporting Ref. 9:

9. Haltmeier, M. & Nguyen, L. V. Regularization of Inverse Problems by Neural Networks (2023). Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision. Cham: Springer International Publishing, 2023. 1065-1093.

Response to Reviewer 2

We thank the reviewer for the positive response to our work and the thoughtful and constructive comments. In addition to the response below, please find related changes in the revised manuscript highlighted in green.

Major Comments:

1. Comment: “One main difference to dictionary based methods seems to be that the dictionary is generated from a fixed set of varying parameters of interest from a fixed protocol, using a Bloch simulator to generate signals from the corresponding parameter set. In contrast, the proposed approach is more flexible to the set of parameters and the protocol since the Bloch simulator can simply be adapted to different parameter sets/protocols. Furthermore, From Table 1, it seems that the signal generation is faster using the proposed GPU solution of the ODEs. In this case, is it possible to use the proposed Bloch simulator to generate dictionaries for the dictionary approach more quickly, and then train a network on this data? In that case, this version of DL dictionary matching would seem to take a similar amount of time as the proposed approach, with quality of the estimates being the difference.”

Response: We thank the reviewer for this useful insight. We have followed this suggestion and added a unified benchmarking to Table 1b using our own implementation of dictionary generation, matching, and supervised learning (additional details are available in Supplementary Section 2). All variants were implemented using the accelerated formulation developed as part of this work (GPU-based JAX formulation of the Bloch McConnell numerical solution).

(a) NBMF computational complexity compared to previously reported studies of semisolid-MT and CEST quantification

Stage \ Method	Numerical BM model direct fitting	Dictionary-based pattern matching	NN-based mapping (supervised learning)	**NBMF** (MT+CEST)
Preparatory steps using synthetic data + 1^{st} subject mapping	None	Hours to days; e.g., 21.3h (x56 nodes) for 79M-entries dictionary generation[12]	Hours; e.g., ~4.5h for 60K-entries dictionary generation and network training[38]	<30min (fit&train)
N^{th} subject tissue parameters quantification (N>1)	Hours to days; e.g., 6h recon of a 256 × 256 pixel slice[48]	Hours to days; e.g., 2.31h for reconstructing a 128 × 128 pixel slice[12]	~1 sec[12, 38] (NN inference)	1sec (inference) **or**: <30min (re-fitting)

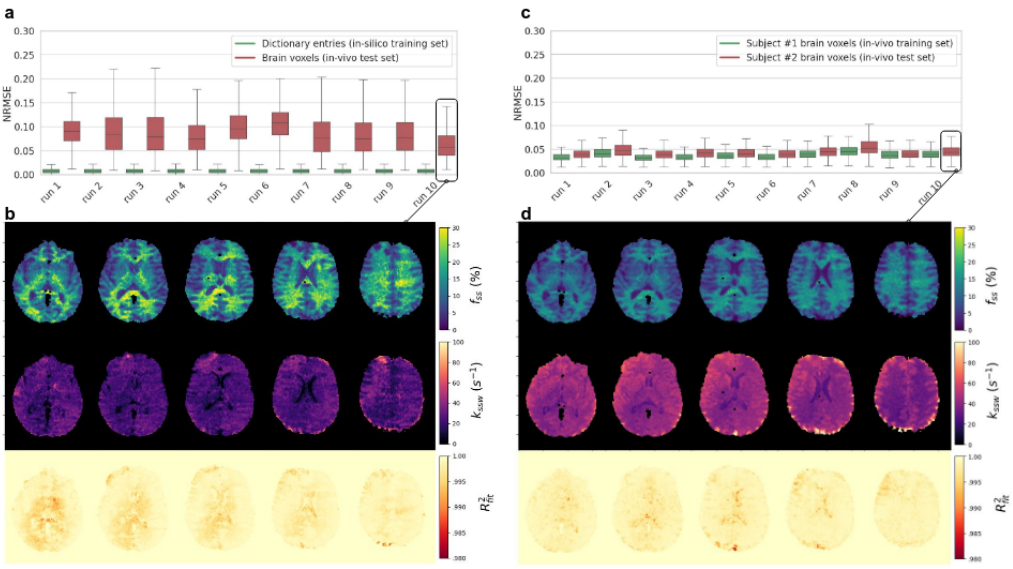
(b) Unified benchmarking of all methods using the accelerated approach developed as part of this work*

Stage \ Method	Autodiff-based (AD) model fitting **(VBMF)**	Dictionary-based pattern matching	NN-based mapping (supervised learning)	Self-supervised NN + AD **(NBMF)**
Preparation + 1^{st} subject mapping	3 min	25 sec& (generation) + 93 sec (matching)	25 sec (generation) + 58 sec (training)	3 min (fit&train)
N^{th} subject mapping (N>1)	3 min	93 sec (matching)	1 sec (NN inference)	1 sec (inference)
Consistency with raw per-subject data	✓ (up to convergence)	✓ (up to grid resolution)	× (empirically, in our test)	✓ (up to prior)
Implicit NN-based smoothing prior	×	×	✓	✓

*A GPU-based JAX formulation of the Bloch McConnell ODEs numerical/analytical solutions. The comparison was based on a semisolid MT MRF protocol using a dictionary consisting of 400K entries and a whole-brain quantification task (194K voxels), see additional details in Supplementary Section 2.

& Dictionary generation for the 79M entries CEST and semisolid MT cartesian grid dictionary used in ref[12] took 80 min, exhibiting a linear scaling of the computational complexity.

The accelerated formulation developed as part of this paper also facilitates a rapid use of supervised NN training, with comparable timing to self-supervision. Notably, increasing the dictionary size (e.g., to 79M entries, as in ref.[12] increases the dictionary generation time to 80 min, which is longer than self-supervised learning, yet is much shorter than previous implementations (>60 hours in[12])). The benefit of using the self-supervised approach, under these implementation conditions compared to supervised training is highlighted in Supplementary Figure 3, in the context of the quality of estimates, consistency with the raw acquired data, and per-subject discrepancy minimization.



Supplementary Figure 3. (**a, c**): A comparison between the voxelwise training/test modeling errors of self-supervised learning (NBMF) (**a**) and dictionary-based supervised (**c**) learning. The modeling NRMSE capturing the circuit error of processing the normalized measured signal **m** through the neural reconstructor $\mathscr{R}$ and the physical forward model $\mathscr{F}$ (applied on the resulting parameter estimates $\mathscr{R}(\mathbf{m})$) is computed as $||\mathscr{F}(\mathscr{R}(\mathbf{m})) - \mathbf{m}||$, aggregated across training/test set entries and shown for 10 runs of the pipeline (differing in randomness realizations at initialization and training noise injection). (**b, d**): Quantitative parameter estimates $\{f, k\} = \mathscr{R}(\mathbf{m})$ for five representative slices alongside R^2_{fit}=1-NRMSE^2 maps. For the supervised learning (**b**), the reconstructor with the best post-hoc test results out of $\{\mathscr{R}_j; j = 1..10\}$ is selected, while for NBMF (**d**), a typical run is used.

2. Comment: “The other main difference is the fact that the dictionary based methods (as described in the text) rely solely on the synthetically generated dictionary for fitting, whereas the proposed method (at least for the first subject) trains in a self-supervised way on the actual data of the acquired subject using the flexible Bloch simulator to enforce consistency of the measurements to the physics of the acquisition. I noted that there are no side by side comparisons (in-vivo) of a deep learning dictionary fitting vs. the proposed approach, which I thought would be a natural/fair comparison, especially if the point above is answered affirmatively. I would be interested to know why the authors chose to omit this/would they expect the fitting to be similar? I can think of an argument for the dictionary based method being better: the underlying parameters can be chosen freely and hence can be chosen to encompass a wide range of possibilities, whereas the proposed approach is constrained to the parameters available in the subject used for training. While the proposed approach seems to generalize well to unseen parameters, this is something worth investigating further. However, I also see the argument that the proposed approach would be better, as mentioned in the text: the proposed approach is trained using the actual acquired data in conjunction with the ODE model and thus could plausibly be more robust on real data (with regard to data consistency) than the dictionary approaches, which only train on synthetic data. In any case, unless the authors have a compelling argument against comparing to a DL dictionary matching approach using the same protocol as tested in the manuscript, I think this comparison is warranted, and would strengthen the paper to show the effect of self-supervised training/inclusion of data consistency in training.”

Response: A side-by-side comparison between deep-learning based supervised learning and NBMF is available in Supplementary Figure 3, and is analyzed and discussed in Supplementary Section 2. As hypothesized by the reviewer, the new comparison indeed shows that as NBMF is trained using the actual acquired data, in conjunction with the ODE model, it is more robust on real data (with regard to data consistency) compared to the supervised learning dictionary approach, which only trains on synthetic data. Please also see the response to comment #1.

3. Comment: “I would also somewhat disagree with the characterization in the paper that the proposed approach is necessarily better for routine or clinical use, based on considerations of time. If the model can be reused for different subjects, then the only time difference to DL dictionary approaches would be the initial training time. In practice, I would think that standardized protocols would be constructed for routine use; hence, the need for flexibility or the concern that generating a dictionary would take too long would be diminished, since it would only need to be done once for each protocol. Since the inference time is also similar for both methods, it would seem both methods would be comparable for routine use. For research, where one wants to test a variety of different protocols, I can definitely appreciate the benefit of the time savings. Furthermore, if the model needed to be trained for each subject, the fitting/training time of around 18min for the proposed approach could also be infeasible for clinical use, though of course this could be reduced with optimization, better hardware, etc. I would be interested to know the author's comments on these points.”

Response: We thank the reviewer for raising this issue, which is now addressed and discussed in the first section of the Discussion:

As shown in Table 1, the most time-consuming steps for supervised dictionary-based learning are the dictionary generation step followed by neural network training. However, if the imaging scenario is a priori known (e.g., brain cancer treatment monitoring) and the acquisition protocol parameters are fixed and optimized, these steps can be done once without affecting the rapid inference time for each new subject. Self-supervised data-based learning (NBMF), on the other hand, offers the flexibility of accommodating various imaging

scenarios and is more easily adapted for new acquisition protocols and research directions. That being said, the biophysical modeling developed as part of this work (GPU-based JAX implementation of the Bloch McConnell numerical solution) can also accelerate both dictionary generation and supervised learning (Table 1b), allowing the user to utilize and compare all different approaches.

In addition, we replaced the context of clinical utility with CEST MRF *research opportunities* in the introduction:

"To further push the boundaries of CEST MRF research and expedite its progress, the long dictionary generation time associated with each new application needs to be replaced by a rapid and flexible approach that adequately models multiple proton pools under saturation pulse trains."

Minor Comments:

4. Comment: "Given the above discussion, do the authors think that combining the self-supervised and dictionary approaches (e.g. training a single network using a mix of both strategies) would make sense?"

Response: The following lines were added to the Discussion section (first subsection):
"Future work could study different NN architectures with spatial awareness (via convolutional or attention layers), as well as larger datasets composed of multiple subjects and a combination of both dictionary-based synthetic data with real-world subject information."

5. Comment: "It seems that with a differentiable/efficient forward model and a pretrained parameter estimator, one could, for example, output a map which uses both at inference (e.g. use the pretrained parameter estimator as a prior within a Tikhonov regularization). Do the authors think this would improve the parameter estimates/be worth the additional computation?"

Response: We thank the reviewer for the thought-provoking comment. This kind of more complex data and training flows is definitely something worth considering. Another future avenue would be to employ non-trivial inference pipeline beyond simple neural network application, namely to alternate between classical error-reducing and neural steps ("Plug-and-Play" and "Unrolled network" approaches). Due the increased complexity of these directions, and the volume of the extended analysis provided in the new Supplementary Sections, we will explore these interesting opportunities in future work.

6. Comment: "In Figure 5, can the authors comment on the outliers in the SSIM plot for k_{ssw}? 0.6-0.7 SSIM is quite low compared to the performance for the rest, are there any factors to explain this (e.g. slice position, etc)?".

Response: Indeed, several slices near the bottom or top of the brain showed less structure and less brain pixels, amplifying the measured differences. This effect was more pronounced for the k_{ssw}, as this parameter was previously shown to be more challenging to quantify and disentangle from other signal contributes[28].

7. Comment: "In conclusion, I think this work makes a valuable contribution to the literature for the computationally efficient incorporation of a complex, ODE forward model into inverse problem solving for biomedical parameter estimation (in this case for CEST). However, I think that the paper can be strengthened by adding a direct in-vivo comparison to

DL dictionary based approaches (e.g. using the GPU forward model for data generation) as well as a more nuanced discussion of the advantages of each approach."

Response: We thank the Reviewer for the positive feedback and the constructive suggestions. Please see the comparison with DL dictionary based approaches using the GPU forward model for data generation provided in Supplementary Sections 2, Supplementary Fig. 3, and the summary presented in Table 1b. Additional discussion regarding the advantage of each approach was provided in the Discussion chapter.

Response to Reviewer 3

We thank the reviewer for the positive response to our work and the thoughtful and constructive comments. In addition to the response below, please find related changes in the revised manuscript highlighted in cyan.

1. Comment: "On page 2, "Excellent agreement was obtained between the NBMF-estimated and known L-arginine concentrations (Fig. 2a-b),..." From Fig. 2c, it seems that dictionary matching shows better consistency with the known concentrations. Please also supplement the dictionary matching-concentration results and compare them with the NBMF-concentration results. Also, provide the dictionary matching-pH results and compare them with the NBMF-pH results."

Response: A side by side comparison between the dictionary-matching-based concentration and proton exchange rate parameter maps is available in the updated Fig. 2.

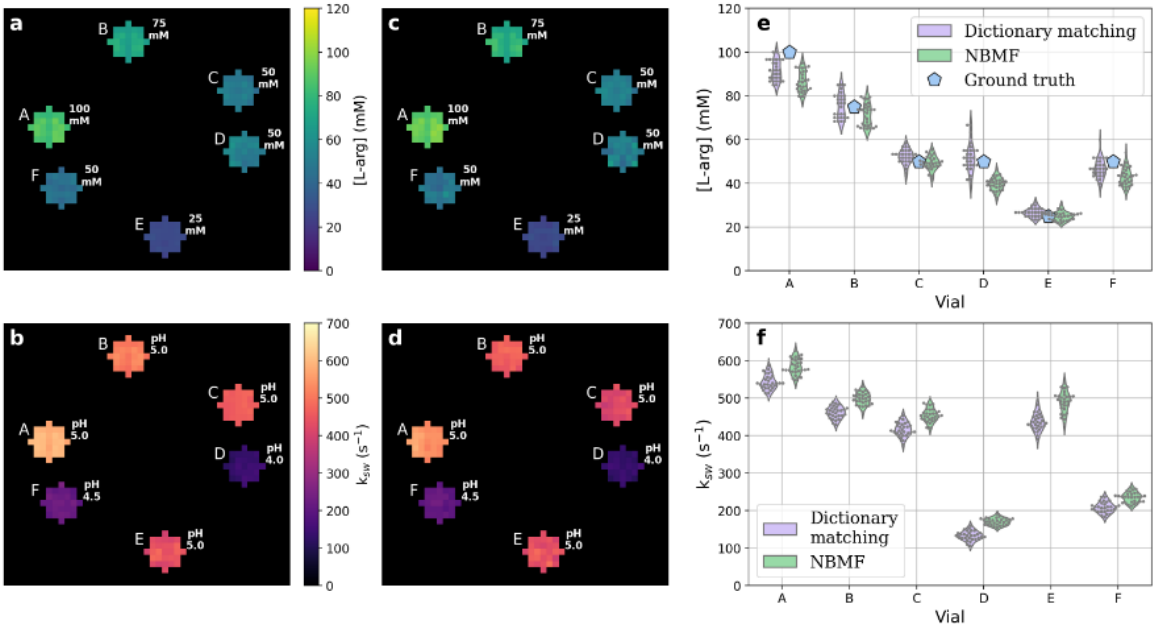
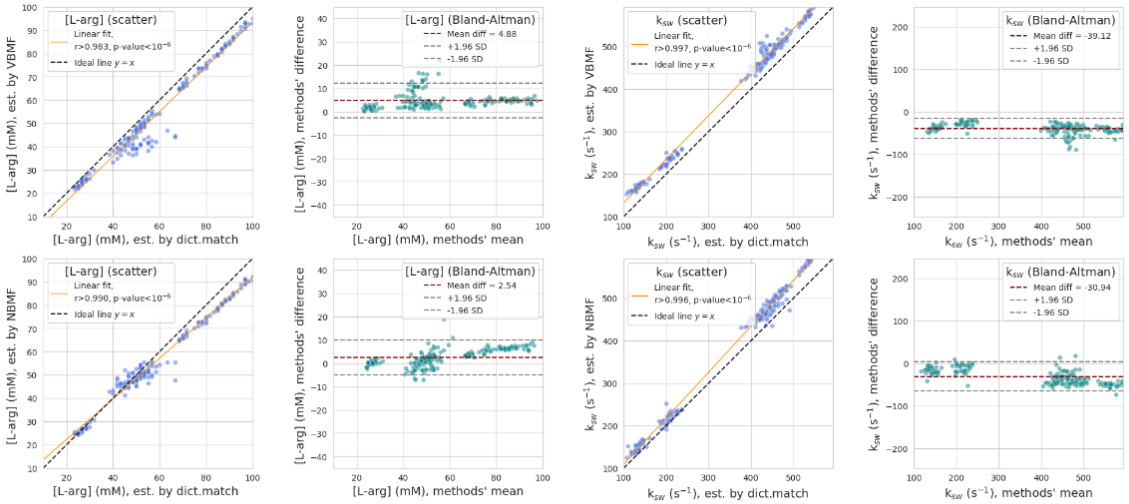


Figure 2. In-vitro study. L-arginine samples were imaged using a pulsed CEST-MRF protocol in a 3T clinical scanner. The NBMF-based L-arginine concentrations (**a**) and proton exchange-rates (**b**) were in good agreement with those obtained by dictionary-based pattern matching (**c**, and **d**, respectively). The ground truth L-arginine concentrations and pH values are mentioned in white text next to each vial. The pixelwise distributions are further compared in (**e**,**f**). Each point in the swarm plot reflects a single 1.8 mm x 1.8 mm x 5.4 mm voxel.

Kindly note that the pH maps presented in the previous manuscript version were drawn from the ground-truth lab pH-meter values as reference, and not from NBMF. To avoid this confusion we now mention the ground truth as white text next to each vial in the revised Fig. 2a-d.

The Pearson's correlation between concentration and proton exchange rate values obtained by both methods was $r>0.983$, $p<10^{-6}$ (new Supplementary Fig. 2).



Supplementary Figure 2. In-vitro CEST experiment analysis: comparing VBMF (**top**) and NBMF (**bottom**) quantitative estimates of concentration (**left**) and exchange rate (**right**) to those obtained using traditional dictionary matching[1].

The wording in the main text that refers to the agreement was nonetheless toned town to "good agreement". A new Supplementary Table 1 was added, which compares NBMF-based compound concentration and proton exchange rate with the respective values obtained using dictionary matching (displayed as mean±std).

Parameter→	[L-arg] (mM)		k_{sw} (s^{-1})	
Vial↓ Estimator→	Dictionary	NBMF	Dictionary	NBMF
A (100mM, pH=5.0)	91.7 ± 4.6	86.6 ± 4.7	543.2 ± 19.5	584.1 ± 21.0
B (75mM, pH=5.0)	75.5 ± 5.7	71.2 ± 5.0	462.0 ± 16.0	499.8 ± 16.0
C (50mM, pH=5.0)	52.1 ± 3.4	49.4 ± 3.2	417.9 ± 18.7	455.0 ± 17.1
D (50mM, pH=4.0)	52.0 ± 6.6	40.1 ± 3.1	132.9 ± 12.2	170.2 ± 9.1
E (25mM, pH=5.0)	26.7 ± 2.0	25.0 ± 1.9	437.6 ± 23.5	491.3 ± 27.6
F (50mM, pH=4.5)	47.0 ± 4.5	43.2 ± 4.1	208.2 ± 14.3	235.3 ± 13.5

Supplementary Table 1. Comparing in-vitro parameter estimates by NBMF to those obtained using dictionary-matching.

The dictionary matching estimates are also available in the right panel of Figure 2e,f (violin+swarm plots) displaying distribution as well as individual voxel values for each vial.

2. Comment: "On page 3, when discussing the "Computational complexity and timing," it is essential to mention the in-plane matrix size and slice number for the whole brain data, as the voxel number affects the total computational timing."

Response: The number of voxels for a subject's whole brain varied between 169K to 194K. This information was added in the section "Computational complexity, timing, and comparison with alternative approaches" (Results chapter). The matrix size and number of slices are now provided in the methodology section: "The in-plane axial matrix size was 116 × 88, with about 50 slices (169K - 194K brain voxels) used per subject."

3. Comment: "On page 6, in addition to the readout module, please also include the key parameters for acquiring the WASABI (B0 & B1), saturation recovery (T1), and multi-echo sequences (T2)"

Response: The key parameters are now available in the “MRI Acquisition” section, of the Methods chapter:
“The WASABI sequence used a preparation scheme realized by a rectangular pulse of 5 ms and nominal B_1=3.7 µT. Twenty-four frequency offsets were equally spaced between -1.8 ppm and 1.8 ppm with a recovery time of 4.5 s. An M_0 image was taken at -300 ppm with a recovery time of 12 s. The saturation recovery T_1 mapping protocol used the following TR (s) values: 10, 6, 4, 3, 2, 1, 0.8, 0.5, 0.4, 0.3, 0.2, 1. The T_2 mapping multi-echo sequence used the following echo times (s): 0, 0.01, 0.025, 0.03, 0.04 0.05, 0.1, 0.2, 0.3, 0.5, 1.0 with a TR = 10 s.”

4. Comment: “It is suggested to add quantitative maps of other human subjects, possibly as supplementary figures, to demonstrate the robustness of the proposed methods.”

Response: Quantitative parameter maps from all four subject used in our study are available in Supplementary Fig. 8. This is now better mentioned in the discussion section.

5. Comment: “Regarding the use of 3.5 ppm concentration as the amide signal (fs, ks), this should be noted as a limitation in the Discussion section, as there may be other contributing factors.”

Response: The following lines were added to the Discussion chapter:
“Subsequent efforts could also improve the modeling accuracy by taking under consideration the contributions of additional proton pools (such as amine and guanidinium) to the 3.5 ppm signal.”

6. Comment: “In Figure 2, for the tube CDF, the pH was varied from 4.0 to 5.0, which is far from the typical physiological pH of around 7.2. Please include tubes with higher pH values closer to physical conditions to better mimic the in vivo case.”

Response: The in-vitro study used L-arginine, which includes Nα-amine exchangeable protons. At physiological pH~7.2 conditions, the proton exchange rate of these proton is >5500 s^{-1}, which is not suitable for 3T clinical CEST imaging. The pH levels were therefore chosen to yield a physiological-relevant intermediate proton exchange rate of 100-600 s^{-1}, that is characteristic for various CEST targets. Notably, similar L-arginine phantom composition and pH range were used in previous quantitive CEST studies[12,14,35,44].

7. Comment: “In Figure 2, in plot f, only half of the "pH" label is shown. Please correct this.”

Response: Thank you for spotting this error. We corrected it in the revised manuscript (the pH values are now shown next to each vial with the wording ‘pH’).

Revised manuscript COMMSPHYS-24-1318B

Response to Reviewer 1

We thank the reviewer for the positive response to our work and the thoughtful and constructive comments.

1. Comment: "Even with a faster generation, the transfer of synthetic data-trained NN to experimental data raises concerns about biased estimates, a steep price to pay for a reduced computation time.": The formulation with "steep price to pay" seems unscientific and should probably be changed in the paper. I apologize for using the same language in the last revision and I should have been less figurative".

Response: The sentence was shortened with the "steep price to pay" term removed.

2. Comment: "Supplementary Figure 3. (a, c): A comparison between the voxelwise training/test modeling errors of self-supervised learning (NBMF) (a) and dictionary-based supervised (c) learning.": Are the labels (a) and (c) swapped in the caption? Otherwise, both the results and the rest of the caption do not make sense to me. Adding titles on the left and right sides of the figure could help to avoid confusion. "

Response: Indeed, the description of the labels in the caption was swapped. Thank you for catching that. It is now corrected in the revised caption of Supplementary Figure 3:

Supplementary Figure 3. **(a, c)**: A comparison between the voxelwise training/test modeling errors of dictionary-based supervised learning (**a**) and self-supervised learning (NBMF) (**c**). The modeling NRMSE capturing the circuit error of processing the normalized measured signal **m** through the neural reconstructor $\mathscr{R}$ and the physical forward model $\mathscr{F}$ (applied on the resulting parameter estimates $\mathscr{R}(\mathbf{m})$) is computed as $||\mathscr{F}(\mathscr{R}(\mathbf{m})) - \mathbf{m}||$, aggregated across training/test set entries and shown for 10 runs of the pipeline (differing in randomness realizations at initialization and training noise injection). **(b, d)**: Quantitative parameter estimates $\{f, k\} = \mathscr{R}(\mathbf{m})$ for five representative slices alongside R^2_{fit}=1-NRMSE^2 maps. For the supervised learning (**b**), the reconstructor with the best post-hoc test results out of $\{\mathscr{R}_j; j = 1..10\}$ is selected, while for NBMF (**d**), a typical run is used.