SUPPORTING INFORMATION

BAYESIAN INVERSE REGRESSION FOR VASCULAR MAGNETIC RESONANCE FINGERPRINTING

Fabien Boux, Florence Forbes, Julyan Arbel, Benjamin Lemasson and Emmanuel L. Barbier

CONTENTS

S.I. EXAMPLES OF SYNTHETIC SCALABLE SIGNALS

We represent below four scalable signals in the time and frequency domains. In the time domain, the synthetic scalable signals used in this study correspond to the absolute value of a linear combination of non-linear functions (i.e. sine and exponential functions) (cf. Section III-A.1 of the main manuscript). Moreover, the choice of frequencies and decay rates yields overlapping contributions to the synthetic scalable signals in the frequency domain, not compatible with a simple separation of the contributions to the signal in that domain. Note that these signals are real-valued vectors.

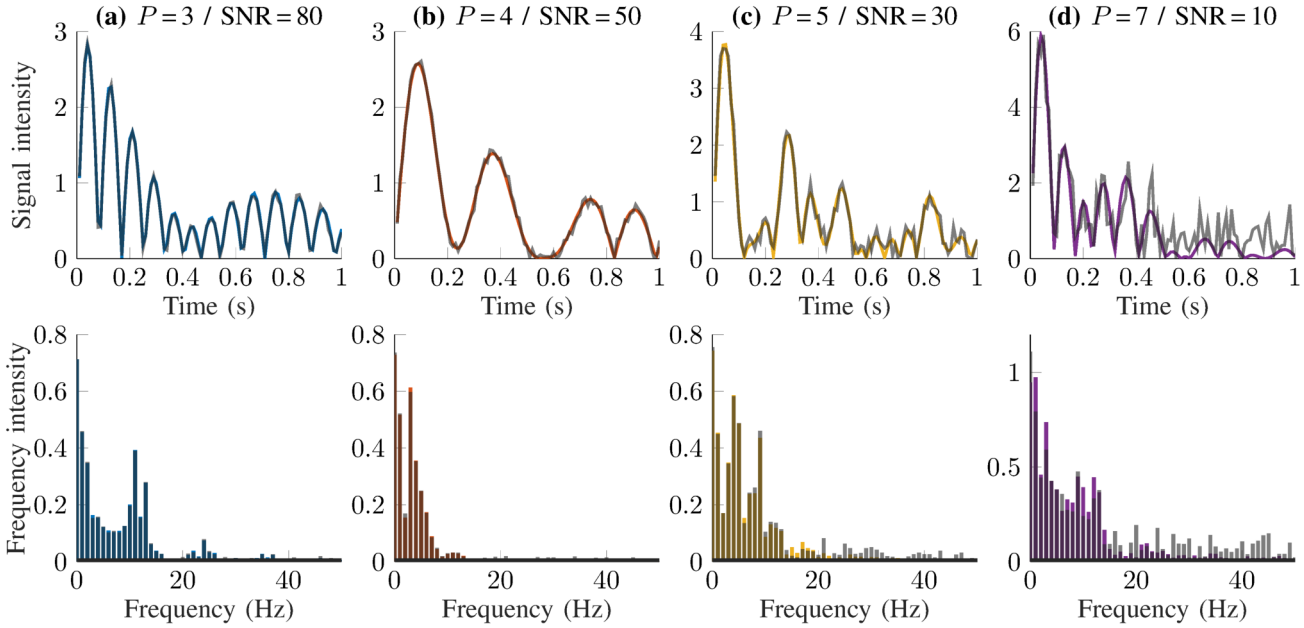


Fig. S1. Synthetic scalable signals for different numbers of parameters and SNR levels, in the time (upper row) and in the frequency domain (lower row). Color curves represent synthetic scalable signals randomly generated with different number of parameters **(a)** $\boldsymbol{P} = 3$, **(b)** $\boldsymbol{P} = 4$, **(c)** $\boldsymbol{P} = 5$, and **(d)** $\boldsymbol{P} = 7$. Black curves represent these signals once noise has been added with **(a)** SNR = 80, **(b)** SNR = 50, **(c)** SNR = 30, and **(d)** SNR = 10. The signal oscillation frequencies are (**(a)** to **(d)** in Hz): (6.1, 6.4, 5.1), (0.6, 3.5, 2.6, 1.1), (3.8, 4.4, 1.5, 5.1, 7.7) and (5.5, 3.6, 3.5, 6.3, 6.5, 6.0, 6.2).

S.II. MRI EXPERIMENTS

A. Animals

The study design was approved by the local institutional animal care and use committee. All animal procedures complied with French government guidelines and were performed under permit 380820 and A3851610008 (for experimental and animal care facilities) from the French Ministry of Agriculture (Articles R214–117 to R214–127 published on 7 February 2013). This study complies with the ARRIVE guidelines (Animal Research: Reporting in Vivo Experiments) [1]. Animals 7 weeks old at the start of the experiments (Charles River, France) were housed in groups of 3-4 in Plexiglas cages under standard laboratory condition (12 h light/dark cycle with lights off at 7:00 p.m. and controlled temperature in $22 \pm 2\,^{\circ}$C). Water and standard laboratory chow were provided *ad libitum*. All procedures were performed under anesthesia by isoflurane (IsoFlo, Abbot Laboratories Ltd, Berkshire, UK). 9LGS cells were implanted in the brain of male Fisher rats. One µl of cell suspension in serum-free RPMI1640 medium containing 10^4 cells were inoculated. MRI was performed 10 days after tumor implantation. C6 cells were implanted in the brain of male Wistar rats. Five µl of cell suspension in serum-free RPMI1640 medium containing 10^5 cells were inoculated. MRI was performed 20 days after tumor implantation. Animals were euthanized by intra-cardiac injection of Pentobarbital 200 mg.kg^{-1} (Dolethal, Vétoquinol Inc, France).

B. Closed-form expression fitting (CEF) method

Parameters BVf and VSI are estimated from the gradient echo sampling of the free induction decay and spin echo according to [2]. The changes in relaxation rates ΔR_2^* and ΔR_2 induced by the injection of the ultrasmall superparamagnetic iron oxide particles (USPIO) contrast agent are computed using gradient echo (GE) signal intensities and spin echo (SE) signal intensities, respectively. The pre-injection and post-injection relaxation times are obtained by fitting the GE signal intensities to an exponential function. It allows to compute ΔR_2^*. ΔR_2 is directly calculated from the two SE signal intensities. BVf and VSI are computed using:

$$\mathrm{BVf} = \frac{3}{4\pi\,\gamma\,B_0\,\Delta\chi_{\mathrm{USPIO}}}\,\Delta R_2^*\,,$$

$$\mathrm{VSI} = 0.425\left(\frac{\mathrm{ADC}}{\gamma\,B_0\,\Delta\chi_{\mathrm{USPIO}}}\right)^{\frac{1}{2}}\left(\frac{\Delta R_2^*}{\Delta R_2}\right)^{\frac{3}{2}}\,,$$

where $\gamma = 2.6752 \times 10^8$ rad.s^{-1}.T^{-1} is the gyromagnetic ratio, $B_0 = 4.7$ T is the magnetic field, $\Delta\chi_{\mathrm{USPIO}} = 3.5$ ppm (SI unit) is the susceptibility difference between blood in the presence and in the absence of USPIO and the apparent diffusion coeficient for water ADC = 800 µm^2.s^{-1}.

S.III. DB-DL AND DB-SL TUNING

A. Choice of the number of epochs used to learn the DB-DL model

The dictionary-based deep learning approach was implemented using the Deep Learning toolbox in the Matlab environment (R2019a; The Mathworks Inc., Natick, Ma, USA). One can observe that for 2 000 epochs, the average RMSE across parameters reaches a plateau, for the number of parameters used in this study (3 to 7).

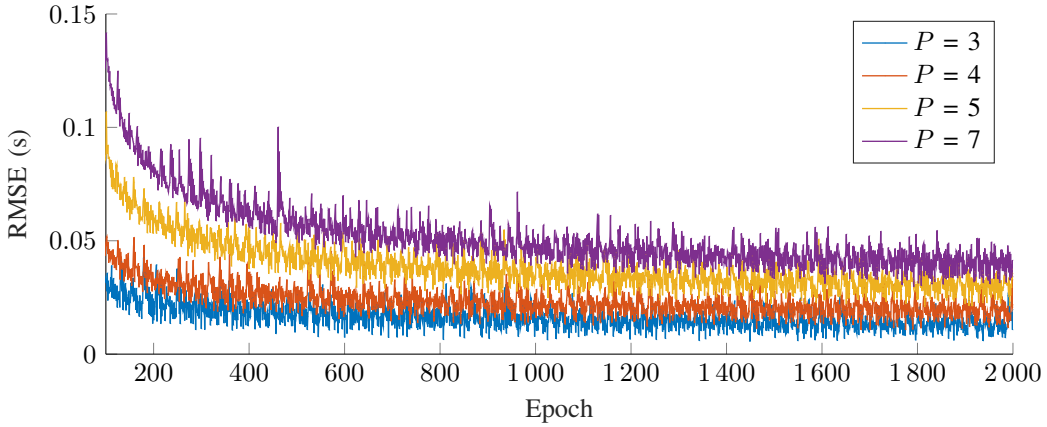


Fig. S2. RMSE on parameter estimates as a function of the number of epochs for the dictionary-based deep-learning (DB-DL) method, using synthetic scalable signals and different numbers of parameters ($\boldsymbol{P}$ = 3, 4, 5 and 7). The network is trained with the ADAM gradient descent algorithm, the learning rate is set to 0.001 and the loss function defined as the mean square error.

B. Choice of the number of Gaussian distributions used to learn the DB-SL model

GLLiM handles the modeling of non-linear relationships with a piecewise linear model [3]. In this approach, there is only one parameter to tune: the number of pieces, i.e. the number K of Gaussian distributions used to represent the function to learn. Here, using synthetic scalable signals, we evaluate the influence of K on the average RMSE across parameter estimates and considering $P = 3$ to $P = 7$ parameters. One can observe that beyond $K = 40$, the average RMSE plateaus. For $P = 7$ and for $K > 100$, the RMSE begins to slowly increase, suggesting that the number of dictionary entries is not sufficient to properly approximate the GLLiM model since the size of the model increases with K (complexity in $\mathcal{O}(KPS)$).

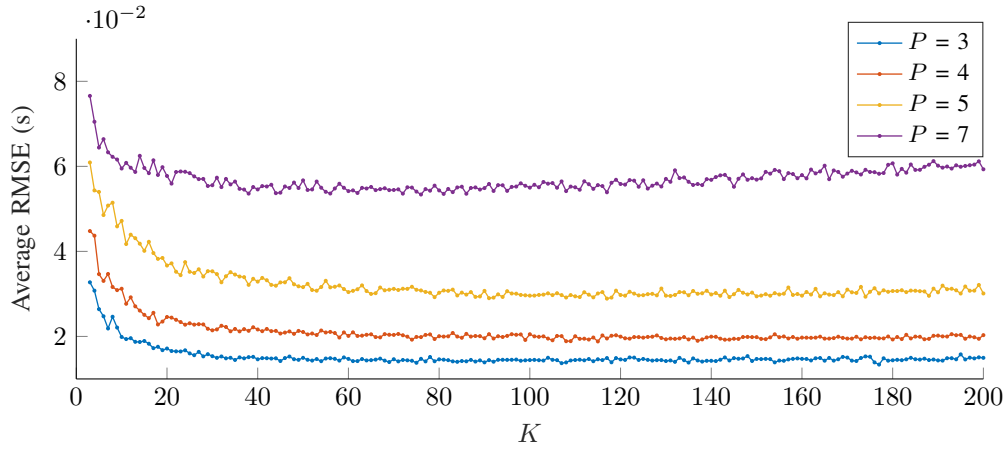


Fig. S3. Impact of the number $\boldsymbol{K}$ of Gaussian distributions on dictionary-based statistical learning (DB-SL) method performance, using synthetic scalable signals. The plot shows the RMSE (over $\boldsymbol{M}$ = 10 000 test signals) with respect to $\boldsymbol{K}$ varying from 3 to 200. Four dictionaries with the same number of entries ($\boldsymbol{N}$ = 10 000) but different number of parameters ($\boldsymbol{P}$ = 3, 4, 5 and 7) are used.

C. Addition of noise to the training signals for DB-DL and DB-SL

Considering a learning set (dictionary) of synthetic scalable signals simulated from a quasi-random sampling of the parameters, we evaluate the impact of adding noise to these signals on the estimation of the DB-DL and DB-SL models. For each parameter number ($P = 5$ and 7), two dictionary sizes (N) are considered. The signal-to-noise ratio (SNR) on the learning signals is denoted SNR_{dico}. The RMSE of the parameter estimates is then evaluated as a function of the SNR on the $M = 1\,000$ test signals. One can observe that the addition of noise to the dictionary signals generally improves the RMSE averaged across parameter estimates, except when the SNR on the test signals is high (here, above 65 to 80, depending on the number of parameters).

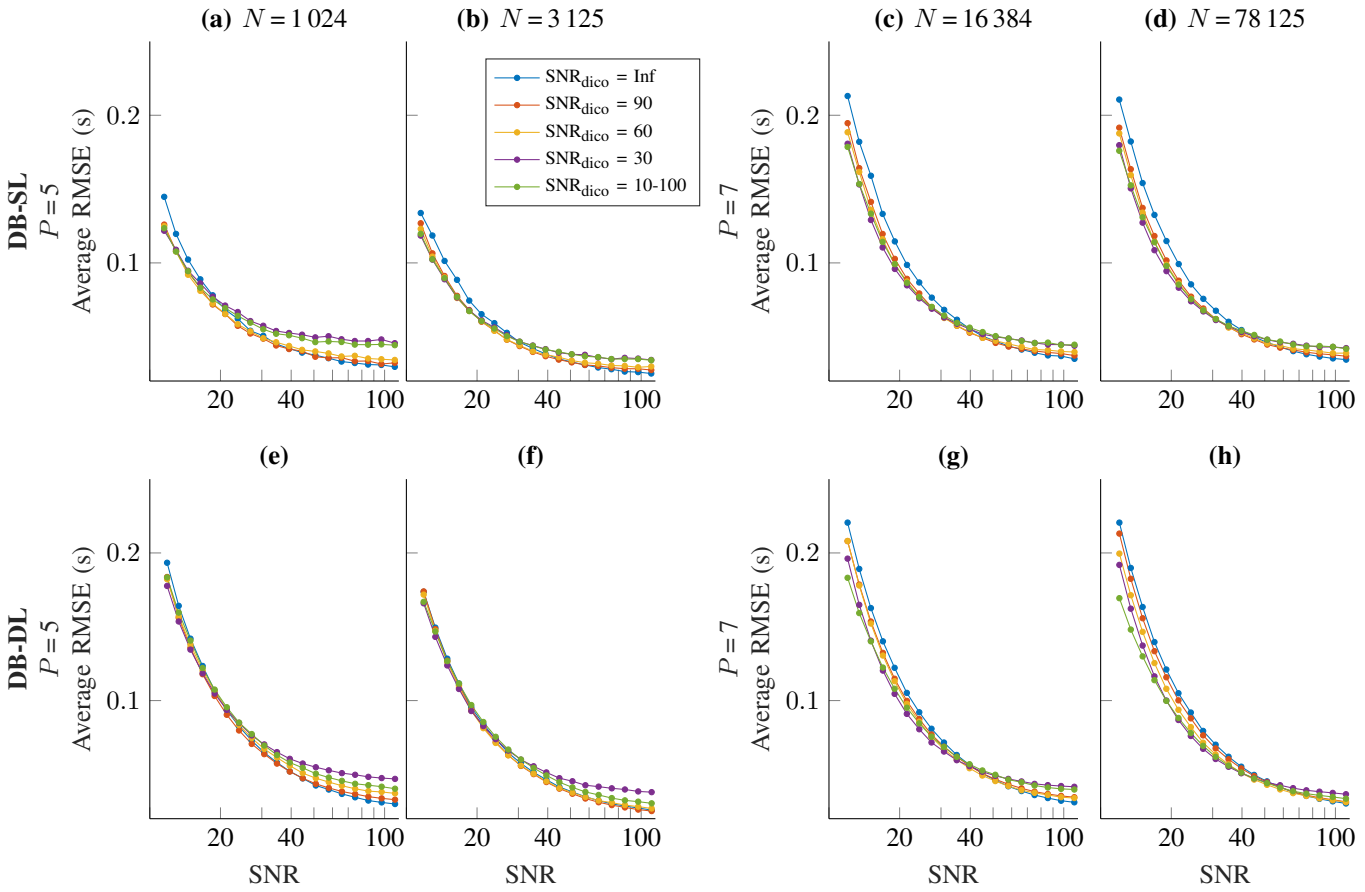


Fig. S4. Impact of the dictionary signal SNR on the estimates provided by the dictionary-based learning methods (DB-DL and DB-SL). Using synthetic scalable signals, average RMSE ($M = 10\,000$) are given as a function of the test signal SNR for different dictionary sizes N (for $P = 5$, $N = 4^5 = 1\,024$ and $N = 5^5 = 3\,125$; for $P = 7$, $N = 4^7 = 16\,384$ and $N = 5^7 = 78\,125$). The first row **(a-d)** shows the average RMSE of the DB-SL method and the second row **(e-h)** shows the average RMSE of the DB-DL method. Figures **(a, b)** and **(e, f)** show results obtained with 5 parameters; Figures **(c, d)** and **(g, h)** results with 7 parameters. The SNR_{dico} refers to the noise level added to the dictionary signals used to learn the model. In the insert, SNR_{dico} = 10-100 corresponds to the addition of noise with an SNR randomly varying between 10 and 100.

S.IV. Additional comparisons of the 3 parameter sampling strategies for different dictionary sizes (DBM and DB-SL)

We evaluate the average RMSE across parameter estimates obtained with DBM and DB-SL for three parameter sampling strategies (regular (*i.e.* a grid), random and a quasi-random), different values of P, and different dictionary sizes. Whatever the simulation condition, the DBM approach yields the smallest RMSE when a regular sampling is used to build the dictionary. Conversely, for the DB-SL model, the RMSE is minimal when a quasi-random sampling is chosen to build the dictionary used to learn the model.

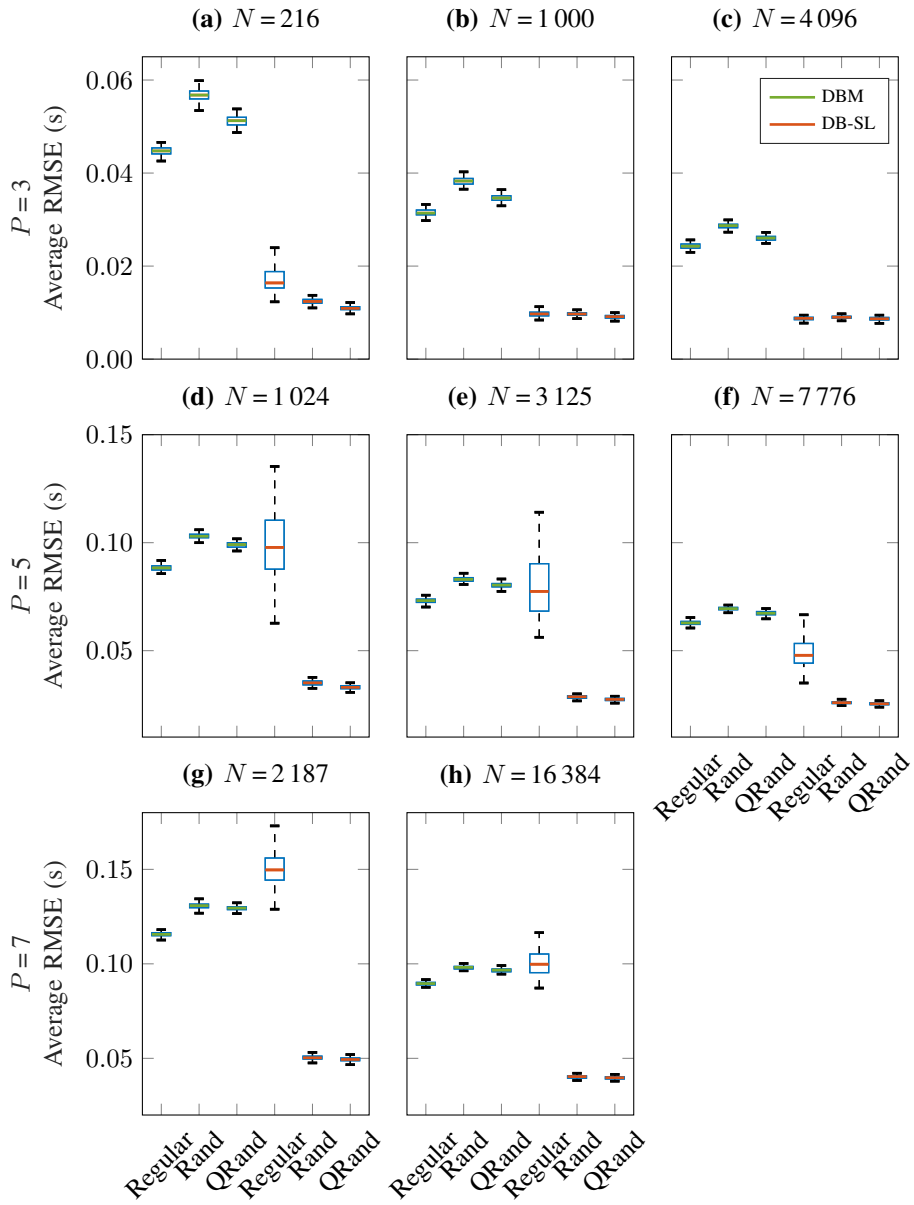


Fig. S5. Effect of parameter sampling strategies on the dictionary-based methods (DBM and DB-SL), using synthetic scalable signals. Distributions of average RMSE ($M = 1\,000$) for regular grid (Regular), random (Rand) and quasi-random (QRand) dictionary samplings. For each box, the central red or green mark indicates the median; the bottom and top edges indicate the 25th and 75th percentiles, respectively. The whiskers extend to the minimum and maximum values (dashed lines). Note that no noise is added on dictionary signals in this initial experiment.

S.V. Impact of scaling-up the dictionary on estimation accuracy (DBM, DB-DL and DB-SL)

The average RMSE across parameter estimates obtained with DBM, DB-DL and DB-SL as a function of the SNR on the test signals is evaluated as the dictionary size scales up, *i.e.* the number of parameters P and the number of entries N increase. Whatever the SNR on the test signals, the DBM approach yields always higher average RMSE than at least one of the two DBL methods and most of the time than the two DBL methods. When the SNR on the test signals is below about 40, DB-SL yields the smallest RMSE. Above a SNR of about 70, the DB-DL yields the most accurate estimates.

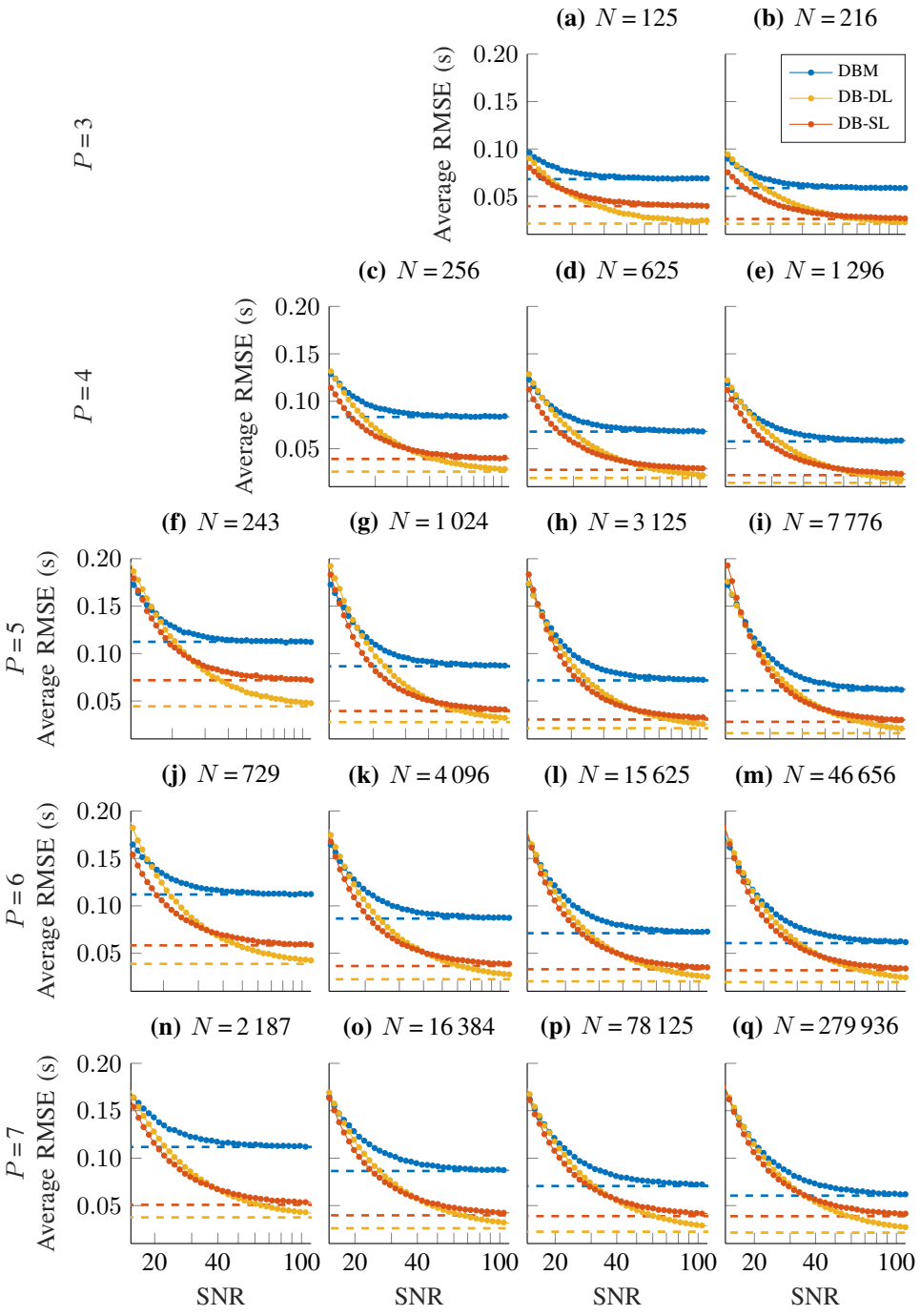


Fig. S6. Impact of dictionary size and test signal SNR on the RMSE obtained with dictionary-based matching (DBM) and dictionary-based learning (DB-DL and DB-SL) methods, using synthetic scalable signals. Average RMSE are given as a function of the SNR for different numbers of parameters (P between 3 and 7) and numbers of entries. N in alphabetical order from **(a)** to **(q)** is equal to $5^3 = 125$, $6^3 = 216$, $4^4 = 256$, $5^4 = 625$, $6^4 = 1\,296$, $3^5 = 243$, $4^5 = 1\,024$, $5^5 = 3\,125$, $6^5 = 7\,776$, $3^6 = 729$, $4^6 = 4\,096$, $5^6 = 15\,625$, $6^6 = 46\,656$, $3^7 = 2\,187$, $4^7 = 16\,384$, $5^7 = 78\,125$ and $6^7 = 279\,936$. The dashed lines represent the average RMSE in the absence of noise on the test signals.

S.VI. Processing time and memory requirements (DBM, DB-DL and DB-SL)

The computer footprints of DBM, DB-DL and DB-SL were estimated on the same two computers (a desktop and a high-performance computer) for different numbers of parameters P from 4 to 7 and increasing values of the dictionary size N. Measures correspond to the processing of $M = 10\,000$ signals. Note that the considered settings here were all tractable on both a desktop and high performance computer, which may not be the case for more complex settings.

Results are reported in the following table. For DBL methods, estimation times and memory requirements are rather stable. Estimations times are short (less than 1.6 s) while the learning times increase with N and P up to 4 343.1 s (about 1 hour 13 minutes). In contrast for DBM, estimation times increase moderately and memory requirements significantly with N and P. In terms of overall performance, DB-DL always shows the smallest memory requirements and often the smallest estimation times. For the learning times, DB-SL is often competitive with smaller times than for DB-DL. However, a fair comparison requires comparable methods implementations and optimization. In our settings, the DB-DL method has been implemented using a commercially available Matlab toolbox (Deep Learning Toolbox) while the DB-SL method uses in-house Matlab code, without any compiled routines.

	Dictionaries		DBM		DB-DL			DB-SL		
	P	N	Estimation (s)	Memory (MB)	Learning (s)	Estimation (s)	Memory (MB)	Learning (s)	Estimation (s)	Memory (MB)
High performance	4	625	0.02 ± 0.00	0.5	34.5	0.08 ± 0.00	0.3	1.0	0.71 ± 0.01	4.2
	4	1 296	0.04 ± 0.00	1.1	37.3	0.09 ± 0.00	0.3	3.0	0.71 ± 0.01	4.2
	5	1 024	0.03 ± 0.00	0.9	34.5	0.08 ± 0.00	0.3	1.8	0.80 ± 0.02	4.3
	5	3 125	0.07 ± 0.00	2.6	49.7	0.07 ± 0.00	0.3	8.8	0.80 ± 0.01	4.3
	5	7 776	0.15 ± 0.00	6.5	72.2	0.07 ± 0.00	0.3	24.8	0.79 ± 0.01	4.3
	6	729	0.03 ± 0.00	0.6	31.8	0.07 ± 0.00	0.3	0.8	0.87 ± 0.01	4.3
	6	4 096	0.09 ± 0.00	3.5	55.3	0.07 ± 0.00	0.3	11.8	0.88 ± 0.01	4.3
	6	15 625	0.28 ± 0.00	13.2	120.3	0.07 ± 0.00	0.3	96.0	0.86 ± 0.01	4.3
	6	46 656	0.78 ± 0.02	39.6	368.0	0.06 ± 0.00	0.3	700.0	0.87 ± 0.03	4.3
	7	2 187	0.06 ± 0.00	1.9	40.3	0.07 ± 0.00	0.3	5.2	0.93 ± 0.01	4.3
	7	16 384	0.30 ± 0.01	14.0	115.3	0.07 ± 0.00	0.3	109.0	0.95 ± 0.01	4.3
	7	78 125	1.28 ± 0.04	66.9	609.6	0.05 ± 0.00	0.3	911.1	0.95 ± 0.02	4.3
	7	279 936	4.22 ± 0.08	239.6	2440.6	0.06 ± 0.00	0.3	3531.9	0.95 ± 0.01	4.3
Desktop	4	1 296	0.06 ± 0.01	1.1	39.1	0.34 ± 0.00	0.3	1.4	1.11 ± 0.01	4.2
	5	7 776	0.26 ± 0.01	6.5	119.9	0.21 ± 0.00	0.3	30.8	1.38 ± 0.01	4.3
	6	46 656	1.46 ± 0.02	39.6	622.8	0.20 ± 0.00	0.3	523.8	1.41 ± 0.01	4.3
	7	279 936	8.60 ± 0.03	239.6	3271.4	0.14 ± 0.00	0.3	4343.1	1.59 ± 0.01	4.3

TABLE S1

COMPUTATIONAL TIMES AND MEMORY REQUIREMENTS OF DICTIONARY-BASED MATCHING (DBM), DICTIONARY-BASED DEEP LEARNING (DB-DL) AND DICTIONARY-BASED STATISTICAL LEARNING (DB-SL) METHODS, USING SYNTHETIC SCALABLE SIGNALS. COMPUTATIONAL ESTIMATION TIMES OF $M = 10\,000$ SIGNALS AND LEARNING TIMES (DB-DL AND DB-SL ONLY: TIME TO GENERATE THE MODEL) AND MEMORY REQUIREMENTS TO STORE THE DICTIONARY (DBM) AND TO STORE THE MODEL (DB-DL AND DB-SL). RESULTS ARE GIVEN FOR DIFFERENT COMBINATIONS OF NUMBER OF PARAMETERS (P BETWEEN 4 AND 7) AND NUMBER OF DICTIONARY ENTRIES (N BETWEEN 625 AND 279 936). THE FIRST PART OF THE TABLE (THE FIRST 13 LINES) SHOWS RESULTS FOR THE PROCESSING OF DATA ON A 32-NODE HIGH PERFORMANCE COMPUTER (*Intel Xeon Gold* 6130, 2.1 GHZ, 384 GB SYSTEM MEMORY) AND THE SECOND PART OF THE TABLE (THE LAST 4 LINES) SHOWS RESULTS FOR THE PROCESSING ON A 4-NODE DESKTOP COMPUTER (*Intel Core* I7-4770, 3.4 GHZ, 32 GB SYSTEM MEMORY). THE LEARNING TIME FOR THE NEURAL NETWORK ALSO CORRESPONDS TO THE CPU IMPLEMENTATION. NOTE THAT USING A GPU IMPLEMENTATION (*GPU Nvidia RTX2080 Ti*, 4352 CUDA CORES, 11 GB MEMORY), WE OBSERVE A SPEED GAIN ABOUT 10-FOLD FOR THE DEEP LEARNING APPROACH BUT A SPEED GAIN FOR THE PROPOSED APPROACH WOULD BE PROBABLY ALSO OBSERVED WITH A GPU IMPLEMENTATION.

S.VII. Confidence index (DB-SL)

A. Non-updated CI versus RMSE

The non-updated CI (see section II-C in the manuscript for details) is compared to the RMSE computed for different SNR values (20, 30, 40, 60 and 100). Figure S7 below shows that the CI is always proportional to the RMSE ($R^2 \geq 0.62$) but the proportionality coefficient α differs from one and appears proportional to the SNR. The proposed procedure for updating the CI addresses this issue (cf. Fig. 4 in the main manuscript and Fig. S8 below).

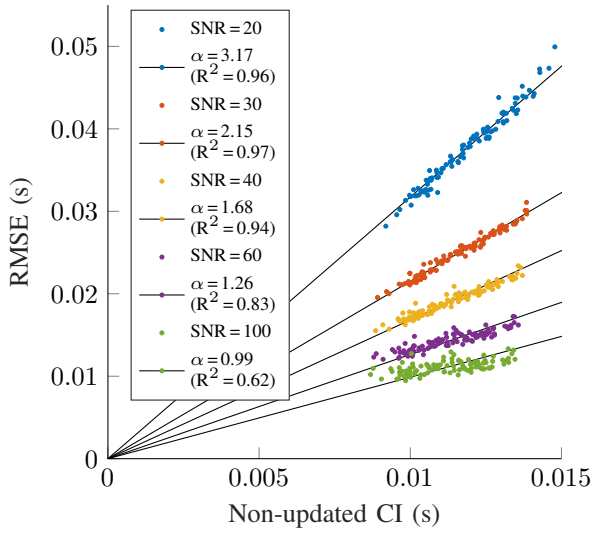


Fig. S7. Correlation between the non-corrected confidence index (CI) and the RMSE, using synthetic scalable signals. RMSE ($M = 10\,000$ test signals) as a function of the CI for SNR = 20 (blue), 30 (orange), 40 (yellow), 60 (purple) and 100 (green). Black lines correspond to the regression coefficients α between the RMSE and the CI for each SNR values. R^2 is the coefficient of determination.

B. CI outside of the parameter space covered by a block dictionary

Using synthetic scalable signals and a block dictionary, the average RMSE and CI are estimated for a parameter range beyond the one covered by the block dictionary. Fig. S8a reproduces Fig. 3c, but with a different colorbar. Within the dictionary blocks, the CI and RMSE are in good agreement (Fig. S8c; $R^2 = 0.964$ and $p < 0.001$). Outside of the blocks, however, the CI remains low and is only partially related to the RMSE ($R^2 = 0.384$; $p < 0.001$). We therefore represent CI as missing values when parameter estimates are obtained outside the learning blocks.

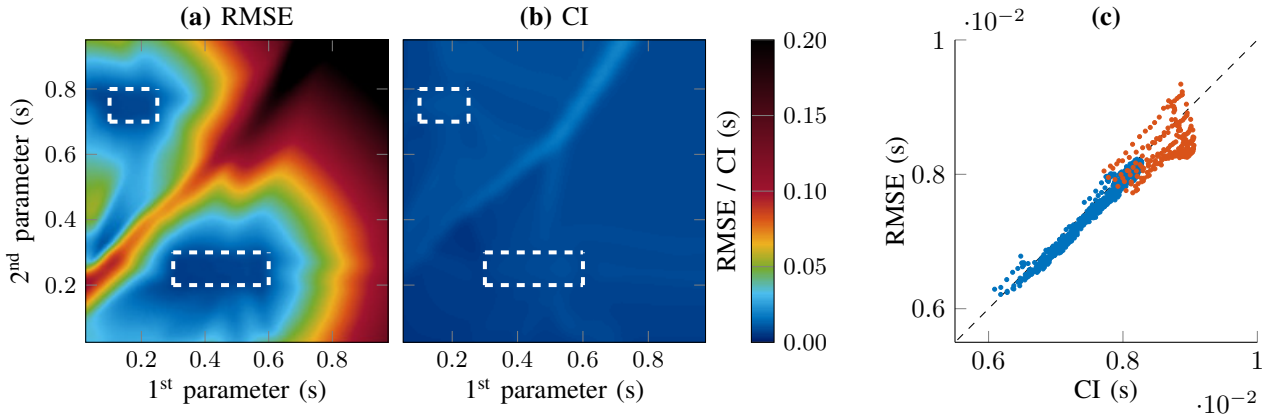


Fig. S8. RMSE and CI outside the parameter space covered by a block dictionary. Figures **(a, b)** show the average RMSE and CI ($M = 2\,000\,000$), respectively. Figure **(c)** shows the correlation between the average RMSE and the CI for the test signals located within the dictionary blocks (blue and orange colors correspond to the two blocks). The white dashed lines delimit the blocks. The average RMSE is computed from signals in a 50×50 ms sliding window, moving in 5 ms step in the parameter space.

S.VIII. Synthetic MRF signals (DBM, DB-DL and DB-SL)

A. Impact of the signal to noise ratio of test signals

1) Synthetic bSSFP signals: To compare DBM, DB-DL and DB-SL on a standard relaxometry application of magnetic resonance fingerprinting (MRF) [4], we use an MRF acquisition based on an inversion-recovery balanced steady state free-precession (bSSFP) sequence to generate synthetic signals. According to [4], after an initial inversion pulse, the flip angle is a series of repeating sinusoidal curves with a period of 250 repetition times (TR) and alternating maximum flip angles (FA). In the odd periods, the FA is calculated as $\mathrm{FA}_t = 10 + 50 \times \sin\left(\frac{2\pi}{500}t\right) + \mathrm{rand}(2)$, where t varies from 1 to 250 and $\mathrm{rand}(2)$ is a function to generate normally distributed random numbers (mean is zero) with a standard deviation of 2. In the even periods, we divide the previous period's FA by two. A 50-TR delay is added between each period. The TR pattern is generated following a procedural noise between 10.5 and 14 ms. We consider $S = 1\,000$ samples. FA and TR patterns as well as example synthetic MRF signals are shown in Fig. S9a-c. The parameters of interest ($P = 3$) are the relaxation times T_1 (between 200 and 3 000 ms) and T_2 (between 20 and 300 ms), and the off-resonance Δf (between -200 and 200 Hz).

2) Complex-valued signals: In contrast to the synthetic scalable and vascular MRF signals used in the main study, standard MRF signals are vectors of complex-valued samples. For the DBM method, the scalar product extends to complex data. To consider these complex samples using the two DBL methods, signals are doubled in size and composed of the real and imaginary parts of the initial signals as it has been proposed by other works [5], [6].

3) Results: To compare dictionary-based matching (DBM) and dictionary-based learning (DBL) methods on synthetic MRF signals, we generate two dictionaries of $N = 4\,096$ and $N = 226\,981$ entries. The average root mean square error (RMSE) is computed over $M = 10\,000$ test signals, for SNR values between 10 and 110 (Fig. S9d-f).

We observe that each method has its advantages with results that depend on the parameter of interest. For the small dictionary, DB-SL outperforms DBM in all conditions and DB-DL at low SNR ($\mathrm{SNR} < 40$) where DB-DL provides higher RMSE than DB-SL. For the large dictionary, DB-DL is always outperformed except for T_2 at very high SNR ($\mathrm{SNR} > 70$). In all other cases, DBM provides the best estimations for T_1, DB-SL provides the best estimations for T_2 ($\mathrm{SNR} < 70$), while for Δf DB-SL is more accurate at low SNR ($\mathrm{SNR} < 30$) and DBM at high SNR. For that parameter, both methods outperform DB-DL, whatever the SNR level.

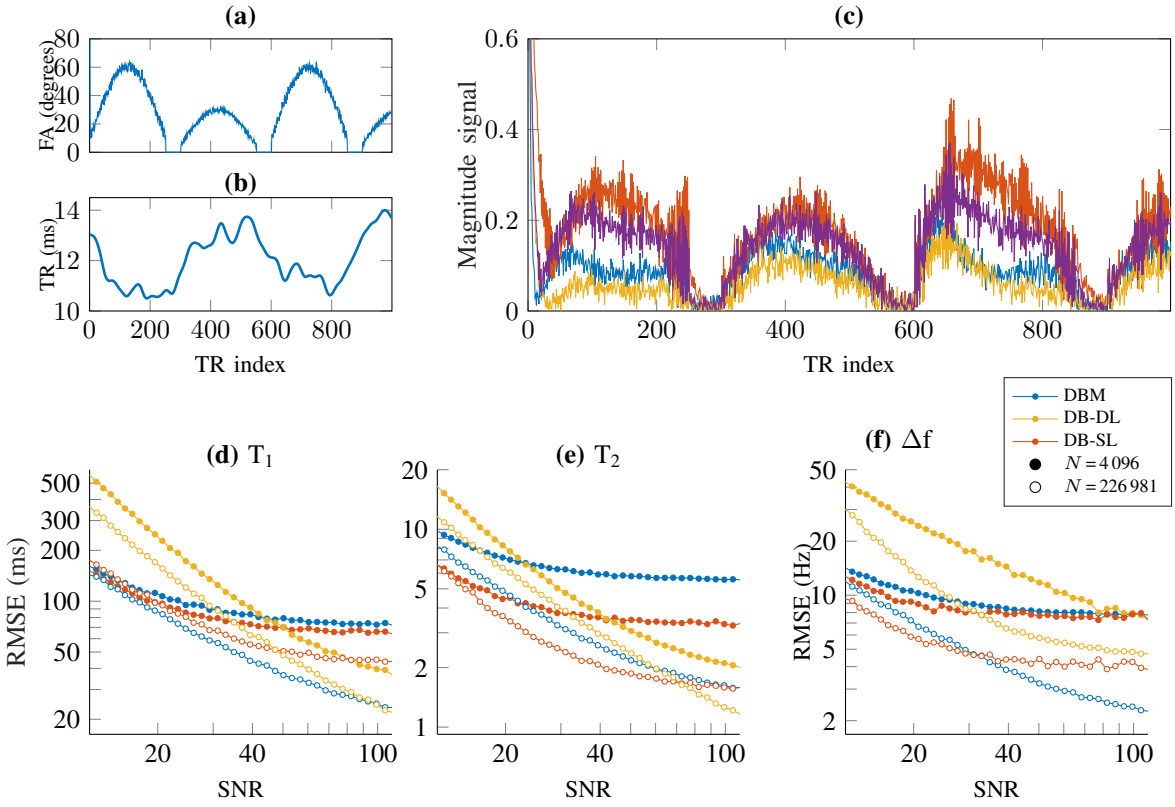


Fig. S9. Comparison of dictionary-based matching (DBM) and dictionary-based learning (DB-DL and DB-SL) methods, using synthetic MRF signals. Figures **(a, b)** show the acquisition sequence settings: flip angles (FA) and repetition times (TR). Figure **(c)** shows four synthetic MRF signal magnitudes ($S = 1\,000$). Figures **(d-f)** show the RMSE ($M = 10\,000$ test signals) on relaxation times ($\mathrm{T_1}$ and $\mathrm{T_2}$) and off-resonance (Δf) parameters ($P = 3$) using the DBM and the two DBL methods with $N = 4\,096$ and $N = 226\,981$ dictionary entries.

B. Bias and variance analysis

1) Bias-variance decomposition: In [7], the authors proposed to investigate the statistical properties of the reconstruction methods using a bias-variance analysis. We propose to conduct a similar analysis here. M vectors of parameters $\{\mathbf{x}_m\}_{m=1:M}$ are considered and for each vector $\mathbf{x}_m$, the corresponding signal is simulated and then perturbed by adding to it successively I ($I = 100$) random Gaussian realizations, according to a fixed signal-to-noise ratio (SNR). It follows that the same parameter vector $\mathbf{x}_m$ is now associated to I signals $\mathbf{y}_{m,i}$. From each of these I signals, a vector of parameters is estimated using one of the three methods DBM, DB-SL or DB-DL, which leads to I values $\hat{\mathbf{x}}_{m,i}$ to approximate the same $\mathbf{x}_m$.

To compare the bias, variance and RMSE, the following empirical quantities are then computed for the m^{th} parameter vector $\mathbf{x}_m$:

$$\text{Bias}_m = \hat{\mathbb{E}}\left[\mathbf{x}_m - \hat{\mathbf{x}}_m\right] , \tag{1}$$

$$\text{Var}_m = \hat{\mathbb{E}}\left[\left(\hat{\mathbf{x}}_m - \hat{\mathbb{E}}\left[\hat{\mathbf{x}}_m\right]\right)^2\right] , \tag{2}$$

where $\hat{\mathbb{E}}\left[\,\cdot\,\right]$ is the empirical mean for the $I = 100$ Monte Carlo simulations: we have $\hat{\mathbb{E}}\left[\hat{\mathbf{x}}_m\right] = I^{-1}\sum_i^I \hat{\mathbf{x}}_{m,i}$. The variance in equation (2) measures the stability of the estimator $\hat{\mathbf{x}}_m$ or how much it moves around its mean performance. The variance depends only on the estimator and is not per se a measure of error as it does not involve the target $\mathbf{x}_m$. So in addition to the biais, another important quantity to measure the performance of an estimator is the Root Mean Square Error (here written in its empirical version):

$$\text{RMSE}_m = \sqrt{\hat{\mathbb{E}}\left[(\mathbf{x}_m - \hat{\mathbf{x}}_m)^2\right]} . \tag{3}$$

An important relationship between these quantities is referred to as the bias-variance trade-off or the bias-variance decomposition which states that:

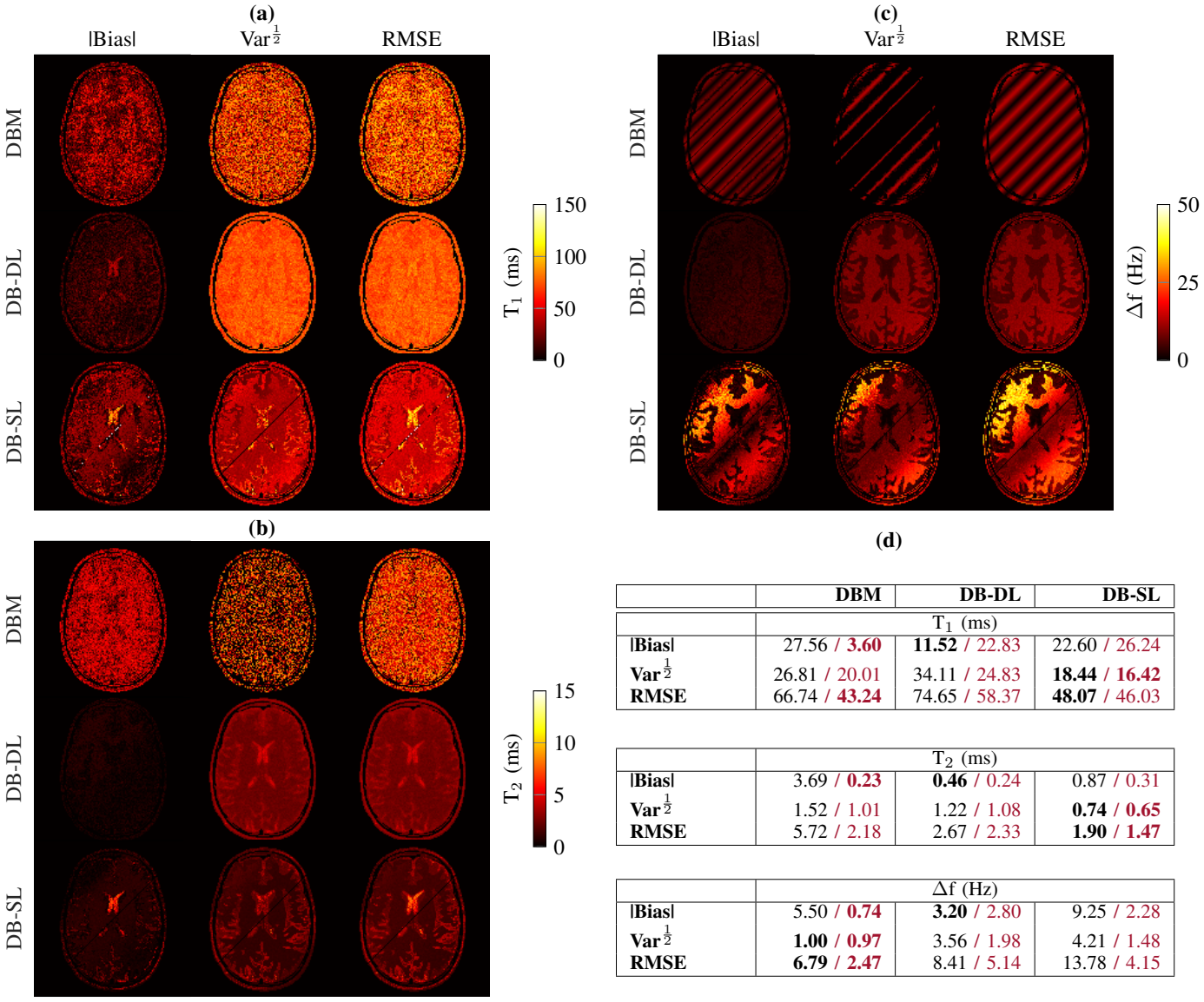
$$\text{RMSE}_m = \sqrt{\text{Bias}_m^2 + \text{Var}_m} . \tag{4}$$

To investigate the statistical properties of the methods, we therefore conduct this bias-variance analysis. A 4-region human brain phantom is generated with a resolution of 128×128. In each region, T_1 and T_2 values are sampled according to a normal distribution around a central value (1 000, 1 400, 1 800 and 2 200 for T_1 and 60, 120, 180 and 240 for T_2). An off-resonance map (Δf) is generated by linear increase following each direction, between -200 and 200 Hz, Fig. S11c. This phantom results in $M = 7\,622$ test signals. 100 random noises (complex thermal noise) are generated according to SNR = 40 for each test signal.

To produce the parameters estimations $\hat{\mathbf{x}}_{m,i}$, the applied instances of the methods are first the ones corresponding to a dictionary of $N = 4\,096$ entries. Bias, variance and RMSE are computed according to equations (1), (2) and (3), respectively. Resulting maps are presented in Fig. S10a-c. Mean values for each map and those obtained with a larger dictionary of $N = 226\,981$ entries, are given in Fig. S10d.

2) Results: The analysis is similar to the one described in [7] up to some obvious typos therein. The bias-variance analysis was carried out for DBM, DB-DL and DB-SL using two dictionary sizes $N = 4\,096$ and $N = 226\,981$ and for each parameter, T_1, T_2 and Δf. In the small dictionary case, for T_1 and T_2, we observed that the DB-DL method was the less biased method with a mean bias about twice smaller than the one of the DB-SL method, but exhibited larger variances. Consequently, DB-DL RMSE were higher for both T_1 and T_2 than DB-SL RMSE, see Fig. S10d. We observed that compared to the DBM method, DBL methods provided brain region specific errors, *i.e.* brain structures were visible in the error maps. The DBM bias, variances and RMSE were always higher than the DB-SL ones for T_1 and T_2. In contrast, DBM improved when a larger dictionary was used and provided smaller bias but not always enough to compensate its higher variances.

For the off-resonance Δf, DBM provided the smallest variances and RMSE. We suspect that a better handling of complex-valued computation could be the reason for this better performance compared to DBL methods. Interestingly, the grid sampling of the dictionary could be visualized as diagonal error lines. Black lines (corresponding to a small error) were located at the dictionary values. DB-SL provided larger error on Δf.



(d)

	DBM	DB-DL	DB-SL
	T$_1$ (ms)		
|Bias|	27.56 / **3.60**	**11.52** / 22.83	22.60 / 26.24
Var$^{\frac{1}{2}}$	26.81 / 20.01	34.11 / 24.83	**18.44** / **16.42**
RMSE	66.74 / **43.24**	74.65 / 58.37	**48.07** / 46.03
	T$_2$ (ms)		
|Bias|	3.69 / **0.23**	**0.46** / 0.24	0.87 / 0.31
Var$^{\frac{1}{2}}$	1.52 / 1.01	1.22 / 1.08	**0.74** / **0.65**
RMSE	5.72 / 2.18	2.67 / 2.33	**1.90** / **1.47**
	Δf (Hz)		
|Bias|	5.50 / **0.74**	**3.20** / 2.80	9.25 / 2.28
Var$^{\frac{1}{2}}$	**1.00** / **0.97**	3.56 / 1.98	4.21 / 1.48
RMSE	**6.79** / **2.47**	8.41 / 5.14	13.78 / 4.15

Fig. S10. Monte Carlo approach for bias-variance analysis, using synthetic MRF signals. Figures **(a-c)** represent the absolute bias (|Bias|), the root variance (Var$^{\frac{1}{2}}$) and the root mean square error (RMSE) maps computed at each voxel ($M = 7\,622$) using 100 different Gaussian noise draws according to SNR = 40. The maps are produced using a dictionary with $N = 4\,096$ entries and for T_1, T_2 and Δf parameters separately (specified in the color bars). In each block of maps, the first row corresponds to DBM method, the second row to DB-DL method and third row to DB-SL method. Table **(d)** contains global mean bias, variance and RMSE (*i.e.* mean values across voxels) for 2 dictionary sizes, $N = 4\,096$ (in black) and for $N = 226\,981$ (in red). On each line, and for each dictionary size, the best value appears in bold.

C. Impact of aliasing noise

To investigate the estimation accuracy of dictionary-based methods in presence of aliasing noise, we use the two previous dictionaries ($N = 4\,096$ and $N = 226\,981$) and the previous brain phantom ($M = 7\,622$). Aliasing noise is added to test signals as described below.

1) Aliasing noise as modulated Gaussian noise: In [8], authors introduced a tool to rapidly assess the efficiency of an MRF sequence in the presence of both thermal and aliasing noise. In this work, we use this tool to assess the robustness to aliasing noise of the DBM, DB-DL, and DB-SL methods. In MRI, the acquired signal contains both Gaussian noise $\boldsymbol{\eta}_{\text{thermal}}$ as the result of thermal contributions (as considered in the main manuscript), and correlated aliasing noise $\boldsymbol{\eta}_{\text{aliasing}}$ due to spatial undersampling artifacts. According to this work [8], the aliasing noise can be modeled as proportional to the signal at each sample. The aliasing noise is a zero-mean Gaussian noise with standard deviation σ_{aliasing}. It results that a noisy signal $\mathbf{y}_{\text{noisy}}$ is computed from an original signal $\mathbf{y}$ as:

$$\mathbf{y}_{\text{noisy}} = \mathbf{y} + \boldsymbol{\eta}_{\text{thermal}} + |\mathbf{y}|\,\boldsymbol{\eta}_{\text{aliasing}}\,. \tag{5}$$

In our experiments on aliasing artifacts, we focus on the case where aliasing noise is dominant so that $\boldsymbol{\eta}_{\text{thermal}} = 0$. This procedure provides typical undersampled MRF signals, see examples in Fig. S11d-f below.

2) Results: For the two previous dictionaries ($N = 4\,096$ and $N = 226\,981$) and the previous brain phantom ($M = 7\,622$), aliasing noise is therefore added to test signals according to equation (5), with σ_{aliasing} between 0 and 1.5. Examples of resulting signals for 3 different values of σ_{aliasing} (0.22, 0.46 and 1.42) are shown in Fig. S11d-f.

We observe for all methods (matching and learning) that the RMSE increases significantly with σ_{aliasing}. For T_1, RMSE remains below 300 ms (about 10 % of the mean) until aliasing noise levels of 0.3 for DB-DL and 0.6 for DB-SL. For T_2, RMSE remains below 30 ms (about 10 %) until aliasing noise levels of 1.1 for DB-DL and 1.2 for DB-SL. Note that the impact of the number of dictionary entries is barely visible when the aliasing noise level is above 0.4. For T_1 and T_2, DBM is at best equivalent to the best learning method and often worse. The DB-SL method stands between DBM and DB-DL for low aliasing noise levels and becomes the most accurate for high aliasing noise levels. For Δf, DBM shows better RMSE for a large dictionary, but is below DB-DL and DB-SL in the case of small dictionary and low aliasing noise levels. The good performance of DBM in estimating Δf might be due to a better handling of complex-valued numbers.

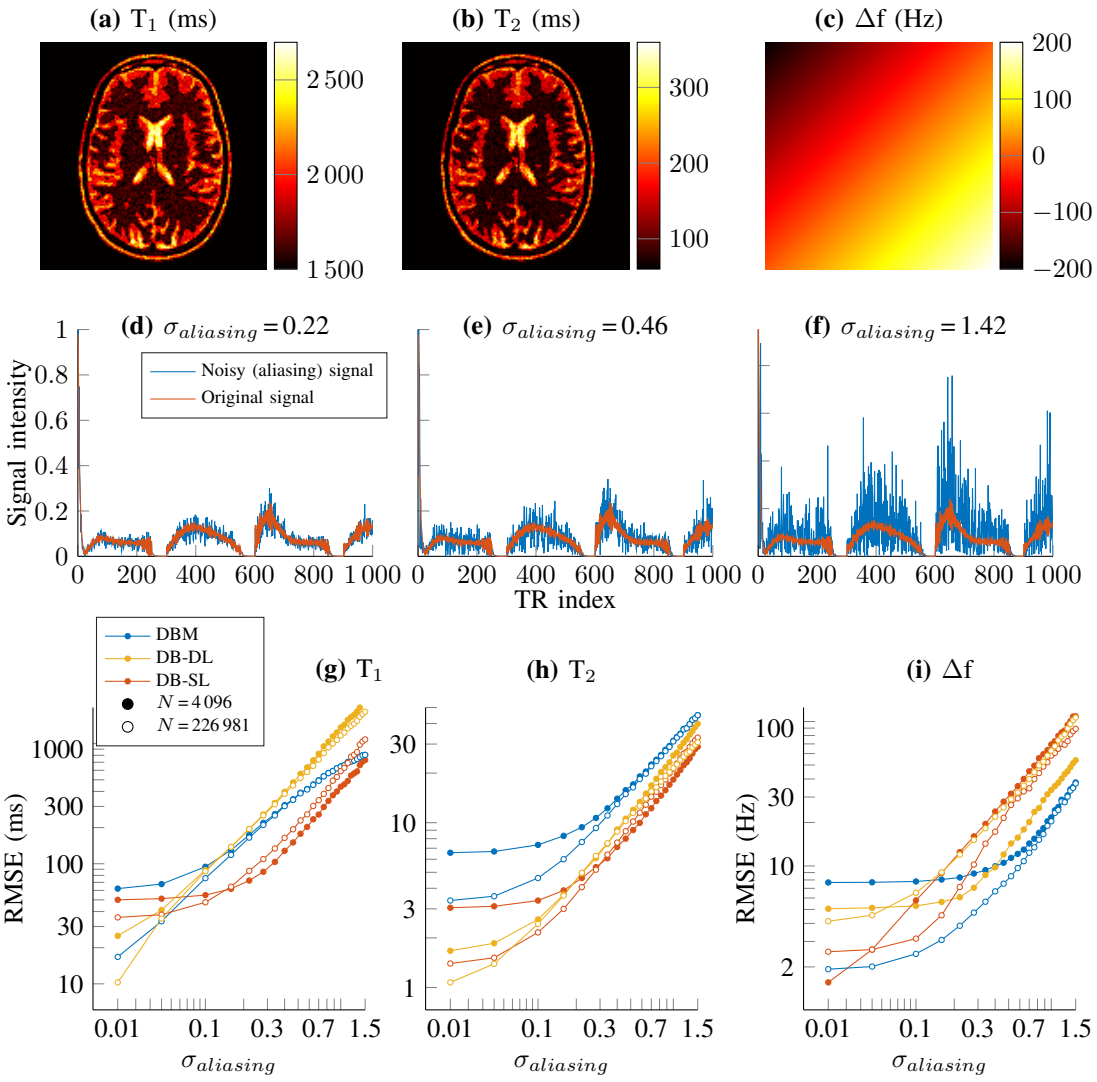


Fig. S11. Impact of aliasing noise in k-space on estimates, using synthetic MRF signals. Figures **(a-c)** show parameter maps ($P = 3$: T_1, T_2 and Δf) used for simulating the $M = 7\,622$ test signals. Figures **(d-f)** show an example of a signal altered by aliasing noise levels of 0.22, 0.46, and 1.42. Figures **(g-i)** show RMSE on estimated parameters for the three dictionary-based methods using $N = 4\,096$ and $N = 226\,981$ dictionary entries. RMSE are given as a function of the aliasing noise level. Note that no thermal noise is added to the test signals in this simulation.

S.IX. Evaluation of a block-dictionary using synthetic Vascular MRF signals (DBM, DB-DL and DB-SL)

We investigate the behavior of DBM and DBL methods outside the limits of the parameter space used to produce the dictionary, using synthetic vascular MRF signals. Two parameters are considered VSI and BVf with values ranging respectively from 1 to 30 μm and 0 to 30%. $M = 100\,000$ test signals are then generated from these values. The three methods are then used to recover parameters VSI and Bvf but using a dictionary that does not cover the full parameter ranges.

In a first step, we design a block dictionary, *i.e.* two regions of the parameter space (two blocks) that are used to produce the dictionary (first two rows of Fig. S12). In a second step, we add 10 dictionary entries outside of these two blocks (last two rows of Fig. S12). The RMSE obtained for the two dictionaries are computed and represented using color bars. For each dictionary, the BVf (1st and 3rd rows) and VSI (2nd and 4th rows) RMSE are shown separately. The different methods DBM, DB-DL and DB-SL correspond to columns.

With DBM, one can see that the RMSE rapidly increases outside of the blocks. For the DBL methods, the RMSE remains low further away from the block boundaries. The addition of extra dictionary entries has a greater impact on the RMSE obtained with the DBL methods than on that obtained with DBM. Moreover, the additional entries reduces the RMSE for low VSI values whereas their impact on large VSI values remains limited (compare Fig. S12 (d-f) to (j-l)).

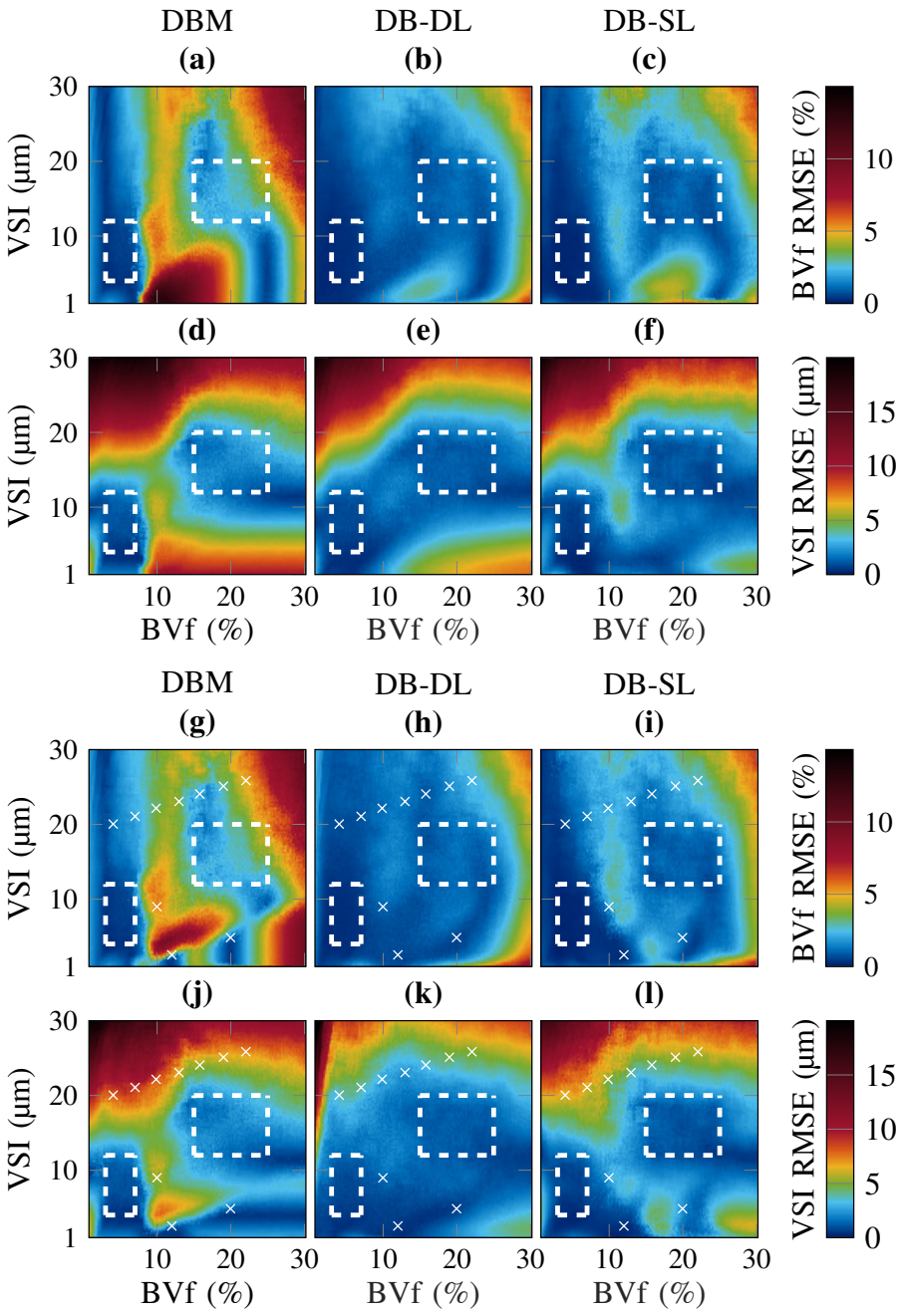


Fig. S12. Estimation accuracy outside of the limits of the parameter space covered by the dictionary. The accuracy is evaluated on two parameters VSI and BVf using synthetic vascular MRF signals. $M = 100\,000$ test signals are used. The white dashed lines delimit the subspace covered by the dictionary (first two rows) with 10 additional entries (last two rows). RMSE are computed for BVf (1st and 3rd rows) and VSI (2nd and 4th rows) separately with the three columns corresponding respectively to DBM, DB-DL and DB-SL. Average RMSE are computed from signals in a 5 µm × 3 % sliding window, moving in 0.5 µm and 0.2 % steps in the parameter space.

S.X. Dictionary design impact on parameter maps for acquired vascular MRF data (DB-DL and DB-SL)

When using real vascular MRF data, RMSE cannot be computed due to the absence of ground truth parameters. The impact of the dictionary design is then assessed qualitatively by looking at changes in parameter estimation when going from a large reference dictionary to a more restricted one. In this section the maps obtained with a large dictionary ($N = 167\,216$) are compared with two reduced dictionaries (both of size $N = 4\,373$), a block dictionary with 10 additional entries corresponding to the one used in the previous section and Fig. S12, and a sub-sampled dictionary obtained by reducing the density of the large dictionary but not its range.

A. Block dictionary

The Figure below shows results for the DBL methods and both the reference dictionary and the block one. The parameter maps obtained with both dictionaries appear comparable, with some notable differences. The maps reconstructed with the block dictionary exhibit smaller values (*e.g.* bottom of the BVf maps) and appear smoother than that obtained with the large dictionary. With the block dictionary, the range of parameter values appear narrower with DB-DL than with DB-SL. This suggests that DB-SL is less sensitive to a restriction of the dictionary and more able to recover the large dictionary results. Note that many values are estimated outside of the blocks, as visible on the CI maps where missing CI values correspond to such estimates.

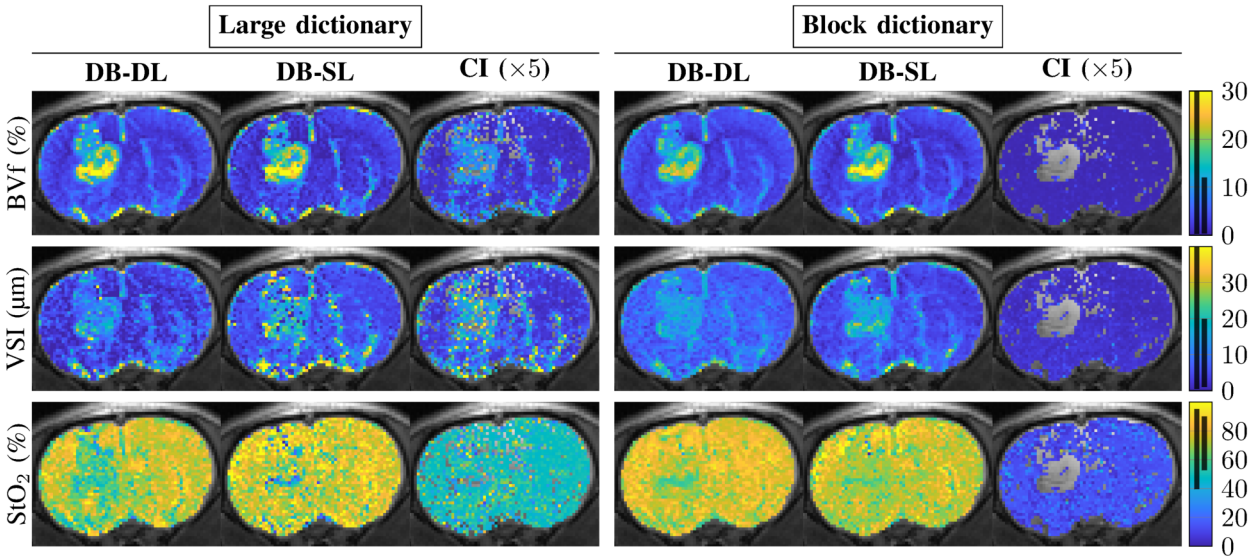


Fig. S13. Vascular parameter estimate maps using a block dictionary composed of two parameter sub-spaces and 10 additional entries (see Fig. S12). The estimated maps for BVf, VSI and StO_2 are shown, using DB-DL (first and fourth columns) and DB-SL (second and fifth columns). The third and sixth columns show the DB-SL confidence index (CI) maps. In the color bars, the black lines represent the parameter ranges covered by the two dictionaries: the short (resp. long) line for the small block (resp. large) dictionary. The number of large and block dictionary entries is respectively $N = 167\,216$ and $N = 4\,373$.

B. Sub-sampled dictionary

The next Figure S13 shows the maps obtained with the large dictionary and with a sub-sampling of this dictionary. This sub-sampled dictionary has the same number of entries as the block dictionary $N = 4\,373$ but the range of parameter values is that of the large dictionary. The parameter maps obtained are more similar in particular with DB-SL confirming that the method is rather robust to a reduction of the number of entries and all the more so as the entries left cover appropriately the parameter values that can occur in practice. Another conclusion from the comparison is that the parameter values range has an impact on the CI values. For instance, when focusing on the two large vessels that perfuse the right hemisphere and are visible in both the BVf and VSI maps, we observe that the CI values are less underestimated in the sub-sampled case than in the block case. However, there are more missing values in the CI maps than for the large dictionary. In the CI map, a value is considered missing for all parameters as soon as one parameter value is found outside the simulated parameter range. The BVf and VSI missing values mainly correspond to high StO_2 and low VSI values, outside the simulated range.

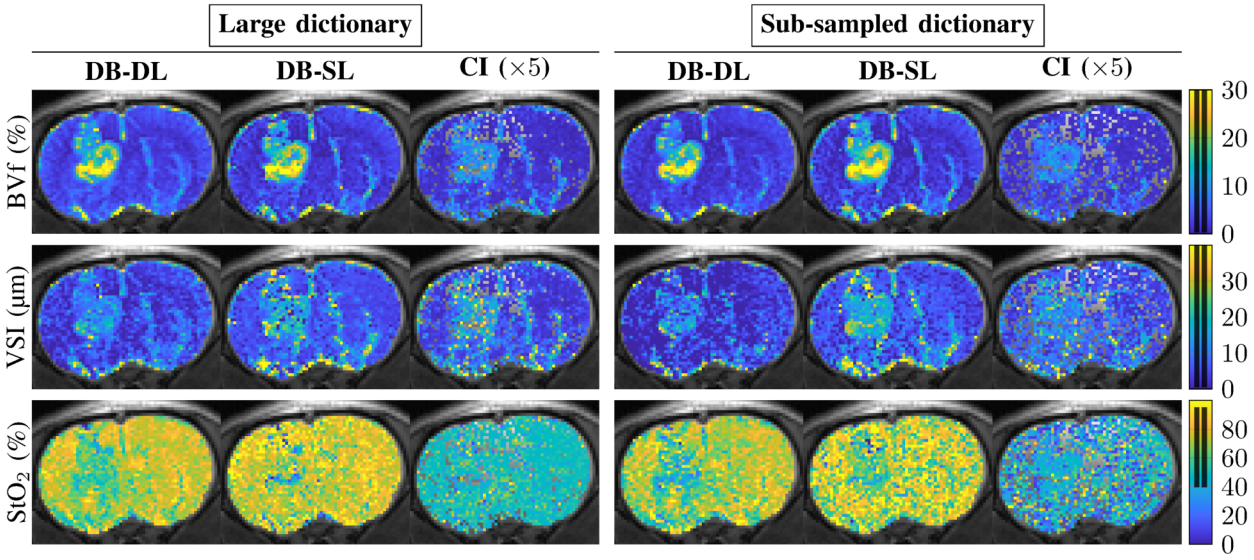


Fig. S14. Vascular parameter estimate maps using a dictionary obtained by sub-sampling the large one. The estimated maps for BVf, VSI and StO_2 are shown, using DB-DL (first and fourth columns) and DB-SL (second and fifth columns). The third and sixth columns show the DB-SL confidence index (CI) maps. In the color bars, the black lines represent the parameter ranges covered by the two dictionaries. Compared to the previous figure, the two dictionaries covered the same parameter ranges. The number of sub-sampled dictionary entries is $N = 4\,373$ extracted from the large dictionary.

S.XI. Additional results on acquired Vascular MRF data (DBM, DB-DL and DB-SL)

A. Data from a rat bearing a C6 tumor

Data acquired on a rat bearing a C6 tumor and the corresponding parameter maps, obtained with DBM, DB-DL and DB-SL, are shown in the figure below. As for the data obtained on a rat bearing a 9L tumor, the maps obtained with the two DBL methods appear more spatially homogeneous than that obtained with the DBM method. Similar conclusions hold.

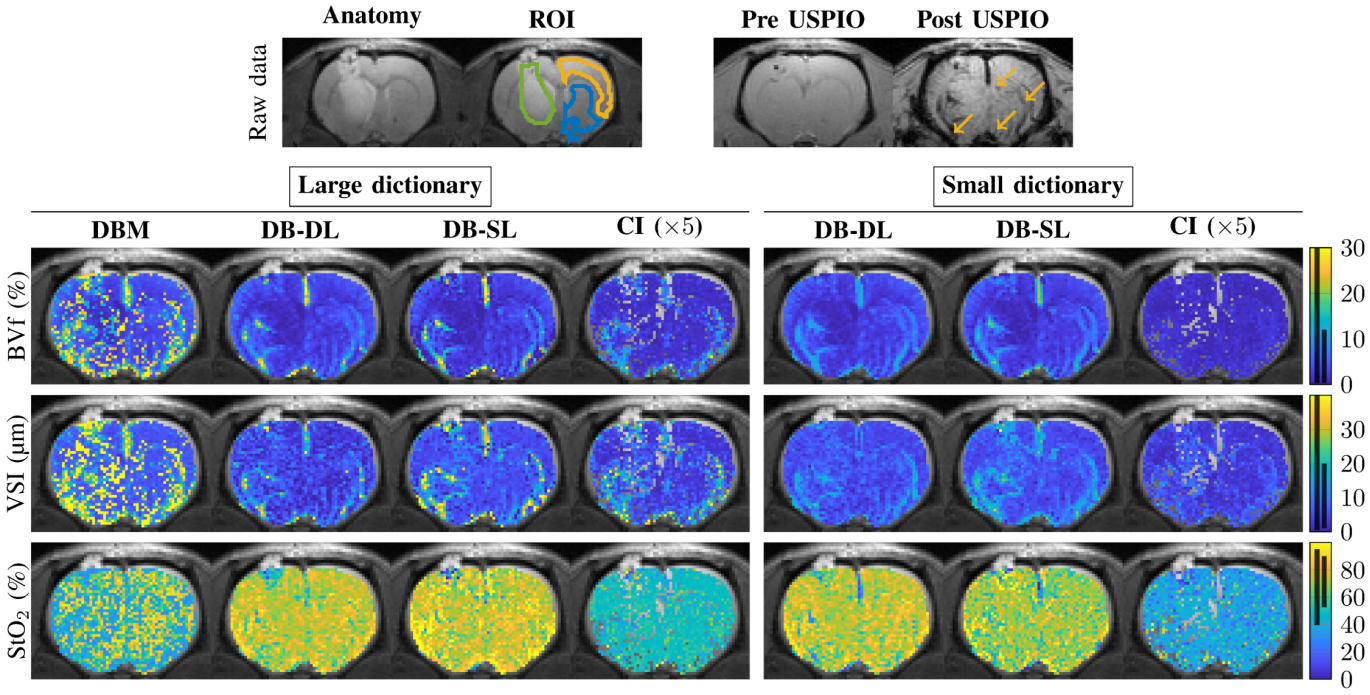


Fig. S15. Vascular parameter estimate maps of a C6 rat tumor model. The first row shows the anatomical image and regions of interest (left) and the MGEFIDSE pre and post USPIO injection (right) for the second echo time (6.3 ms). The tumor, cortex and striatum are respectively delineated with green, yellow and blue lines. The arrows on the post USPIO injection image indicate large vessels. The estimated maps for BVf, VSI and StO_2 are shown below, using DB-DL (first and fourth columns) and DB-SL (second and fifth columns). The third and sixth columns show the DBL confidence index (CI) maps. In the color bars, the black lines represent the parameter ranges covered by the two dictionaries: the short (resp. long) line for the small (resp. large) dictionary. Large and small dictionary entries are $N = 167\,216$ and $N = 4\,119$, respectively.

B. Quantification of parameter estimates in regions of interest

The following figure shows average parameter values for 3 regions of interest and across 8 animals, 4 for each tumor models. There is an overall agreement between the parameter estimates obtained with the four methods (the three dictionary-based and the CEF method). For BVf, the two DBL methods are in good agreement and produce values below that of CEF and DBM. For the VSI parameter, CEF, DBM, and DB-SL are in good agreement, while values produced by DB-DL appear systematically lower. Finally, for StO_2, each method produces a different estimate. CEF generally yields the highest estimates while DBM yields the lowest ones. The two DBL methods produces StO_2 estimates in between the two other methods, with DB-DL estimates below that of DB-SL.

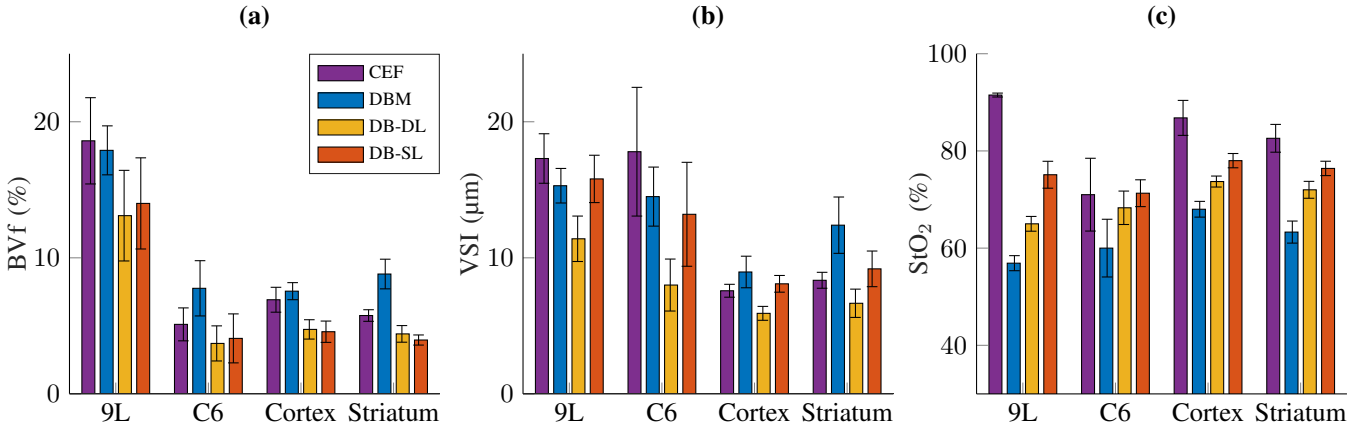


Fig. S16. Mean vascular parameter estimates in 3 regions of interest, the tumor (C6 or 9L), striatum and cortex. Four different methods are compared: the closed-form expression fitting (CEF), the dictionary-based matching (DBM), the dictionary-based deep learning (DB-DL) and the dictionary-based statistical learning (DB-SL) method. Averaged **(a)** BVf, **(b)** VSI and **(c)** StO_2 values across animals (9L and C6 lesions: 4 rats, cortex and striatum: 8 rats) are computed (mean $\pm$ standard deviation).

REFERENCES

[1] G. B. Drummond, D. J. Paterson, and J. C. McGrath, "Animal research: Reporting in vivo experiments: The ARRIVE guidelines," *Experimental Physiology*, vol. 95, no. 8, pp. 842–844, 2010.

[2] I. Troprès, S. Grimault, A. Vaeth, E. Grillon, C. Julien, J. F. Payen, L. Lamalle, and M. Décorps, "Vessel size imaging," *Magnetic Resonance in Medicine*, vol. 45, no. 3, pp. 397–408, 2001.

[3] A. Deleforge, F. Forbes, and R. Horaud, "High-dimensional regression with gaussian mixtures and partially-latent response variables," *Statistics and Computing*, vol. 25, no. 5, pp. 893–911, 2015.

[4] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J. L. Sunshine, J. L. Duerk, and M. A. Griswold, "Magnetic resonance fingerprinting," *Nature*, vol. 495, no. 7440, pp. 187–192, 2013.

[5] E. Hoppe, F. Thamm, G. Körzdörfer, C. Syben, F. Schirrmacher, M. Nittka, J. Pfeuffer, H. Meyer, and A. Maier, "Abstract: Rinq fingerprinting: recurrence-informed quantile networks for magnetic resonance fingerprinting," in *Informatik aktuell*. Springer, 2020, p. 184.

[6] P. Virtue, S. X. Yu, and M. Lustig, "Better than real: Complex-valued neural nets for MRI fingerprinting," in *arXiv*. IEEE, 2017, pp. 3953–3957.

[7] B. Zhao, K. Setsompop, H. Ye, S. F. Cauley, and L. L. Wald, "Maximum Likelihood Reconstruction for Magnetic Resonance Fingerprinting," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1812–1823, 2016.

[8] D. Kara, M. Fan, J. Hamilton, M. Griswold, N. Seiberlich, and R. Brown, "Parameter map error due to normal noise and aliasing artifacts in MR fingerprinting," *Magnetic Resonance in Medicine*, vol. 81, no. 5, pp. 3108–3123, 2019.