Bayesian inverse regression for vascular magnetic resonance fingerprinting

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

Bayesian inverse regression for vascular magnetic resonance fingerprinting

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

***Abstract*— Standard parameter estimation from vascular magnetic resonance fingerprinting (MRF) data is based on matching the MRF signals to their best counterparts in a grid of coupled simulated signals and parameters, referred to as a dictionary. To reach a good accuracy, the matching requires an informative dictionary whose cost, in terms of design, storage and exploration, is rapidly prohibitive for even moderate numbers of parameters. In this work, we propose an alternative dictionary-based statistical learning (DB-SL) approach made of three steps: 1) a quasi-random sampling strategy to produce efficiently an informative dictionary, 2) an inverse statistical regression model to learn from the dictionary a correspondence between fingerprints and parameters, and 3) the use of this mapping to provide both parameter estimates and their confidence indices. The proposed DB-SL approach is compared to both the standard dictionary-based matching (DBM) method and to a dictionary-based deep learning (DB-DL) method. Performance is illustrated first on synthetic signals including scalable and standard MRF signals with spatial undersampling noise. Then, vascular MRF signals are considered both through simulations and real data acquired in tumor bearing rats. Overall, the two learning methods yield more accurate parameter estimates than matching and to a range not limited to the dictionary boundaries. DB-SL in particular resists to higher noise levels and provides in addition confidence indices on the estimates at no additional cost. DB-SL appears as a promising method to reduce simulation needs and computational requirements, while modeling sources of uncertainty and providing both accurate and interpretable results.**

***Index Terms*— Quantitative imaging, magnetic resonance fingerprinting, multiparametric mapping, dictionary learning, inverse regression, brain vascular structure.**

I. Introduction

Magnetic resonance fingerprinting (MRF) is a novel approach to quantitative magnetic resonance imaging that allows the estimation of multiple tissue properties in a single acquisition [1], [2]. The acquisition, which consists in repeating measurements with varying experimental conditions, generates a signal evolution (or *fingerprint*) that depends on the parameters of the studied tissue. To estimate these parameters, a large database, referred to as a *dictionary* and containing a large number of possible signal evolutions, is simulated from biophysical models. A comparison is performed between an acquired signal and the signals in the dictionary to find the best match according to an objective function. The tissue parameters are then estimated to the values that generated the best signal evolution match. In MRF, parameter estimation accuracy therefore depends on the number of dictionary entries, which increases exponentially with the number of parameters. For applications with many parameters such as vascular MRF [3], the required memory size and simulation time as well as the parameter estimation time (or reconstruction time) quickly become a limit.

To compress the dictionary while limiting the loss of information, several authors have used singular value decomposition to project the dictionary in a well-chosen subspace [4]–[8]. However, this compression does not allow a reduction in simulation time. It has also been proposed to directly find a mapping from the fingerprints to the parameter space using kernel regression [9], maximum likelihood approach [10] or neural network approaches [11]–[18]. The resulting compact representation offers the advantage over the discrete MRF grid of a continuous exploration of parameter values. These approaches significantly reduce the reconstruction time, but not the simulation time due to the need to span a high dimensional fingerprint space. To limit the simulation time, Cohen *et al.* [14] studied a mapping obtained from a sparse set of dictionary entries. The study, carried out with only two parameters, led to a modest reduction of dictionary entries (up to 60-fold). Consider a dictionary of 100×100 vascular entries simulated in 1 hour. If the number of parameters increases from 2 to 4 parameters, and always considering 100 values per parameter, then the dictionary computation time increases from 1 hour to more than 1 year. As an illustration, when applying vascular MRF in stroke and brain tumors models, Lemasson *et al.* report that the largest dictionary used (4 parameters and about 30 values per parameter) was generated on a 30-node cluster in about 24 hours [19]. In such settings, an approach that greatly reduces the need for simulation, continuously represents the parameters without loss of precision, relies on an explainable model and reduces the reconstruction time becomes highly desirable [20].

To reach this goal, we adopt in this work the Bayesian framework, which has already been employed with MRF for multiple tissue components within a single voxel [21] or for

This work was supported by the French National Research Agency - project cerebrovascular dynamics in epilepsy: endothelial-pericyte interface - Epicyte, ANR-16-CE37-0013. This work was performed on the IRMaGe platform, member of France Life Imaging network (grant ANR-11-INBS-0006). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

F. Boux, B. Lemasson and E. L. Barbier are with Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut Neurosciences, GIN, 38000 Grenoble, France (phone: +33-4-56-52-05-88, e-mail: emmanuel.barbier@univ-grenoble-alpes.fr).

F. Boux, F. Forbes and J. Arbel are with Univ. Grenoble Alpes, Inria, CNRS, G-INP, 38000 Grenoble, France.

spatial modeling [22]. More specifically, we use the Gaussian locally linear mapping (GLLiM) model [23], which allows to learn a mapping from fingerprints to parameters and provides a full posterior distribution per fingerprint. This distribution can then be used to compute an estimated value and a confidence index for each parameter.

In this vascular MRF study, the proposed approach referred to as the dictionary-based statistical learning (DB-SL) method is compared to both the standard dictionary-based matching (DBM) method and to a dictionary-based deep learning (DB-DL) method. The comparison is made on various types of signals including synthetic standard and vascular signals and real vascular signals acquired in tumor bearing rats.

II. MRF AS AN INVERSE PROBLEM

In inverse problems, the overall issue is to provide information on some parameters of interest $\mathbf{x}$ given an observed signal $\mathbf{y}$, using a known *direct* or *forward model* that describes how the parameters $\mathbf{x}$ translate into a signal $\mathbf{y}$. Among inverse problems, MRF exhibits the following difficulties: 1) the direct model is (highly) non-linear, as a (complex) series of equations or simulation tools; 2) the $\mathbf{y}$'s are high-dimensional signals and 3) many $\mathbf{y}$'s need to be inverted (one for each voxel in an image); 4) the vector of parameters $\mathbf{x}$ is multidimensional and predicting each component of $\mathbf{x}$ independently is likely to be sub-optimal. The last point may be moderated by the degree to which a joint modeling can be carried out and preferred, depending on the parameters interactions, on the complexity of the model and the amount of data available for its estimation.

Most methods to solve inverse problems can be classified into two main categories, optimization-based and learning-based methods. In the next section, we refer to standard MRF as a matching method. We show that it can be seen as a penalized optimization, which does not require statistical modeling, while the method we propose belongs to statistical learning approaches.

A. Dictionary-based matching (DBM) methods

MRF requires a large database $\mathcal{D}_f$, referred to as a dictionary [1]. It is made of N entries of coupled fingerprint and parameters $(\mathbf{x}, \mathbf{y})$. The S-dimensional fingerprints $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ are generated by running the simulation model f for N different values of the P-dimensional magnetic and physiological parameters $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. In the DBM method, a P-dimensional grid is generated with sampled values in a pre-set interval for each parameter. Then, to invert an observed $\mathbf{y}_{\text{obs}}$, it is compared with the signals in $\mathcal{D}_f$ to find the best match according to an objective function $\mathrm{d}(\cdot,\cdot)$, usually a standard distance or dissimilarity measure (*e.g.* in MRF, the dot product). With $\mathcal{D}_f = \{(\mathbf{x}_n, \mathbf{y}_n = f(\mathbf{x}_n)),\, n = 1{:}N\}$, $\mathbf{x}$ is thus estimated as the argument of the following minimization:

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \,\in\, \mathcal{D}_f}{\arg\min}\, \mathrm{d}(\mathbf{y}_{\text{obs}}, f(\mathbf{x})). \tag{1}$$

Solutions are sought in $\mathcal{D}_f$ only, while in a non-constrained optimization the minimization is over the whole continuous space of parameter values. The performance of the method depends directly on the space discretization *i.e.* the choice of the number of dictionary entries and the number of parameters. The DBM grid is by essence discontinuous and more likely to suffer from a sensitivity issue, meaning that two similar fingerprints may be mapped to not so close parameters values. This is well explained and illustrated in a very recent review paper by J. Asslander [24]. The larger the number N of entries $(\mathbf{x}_n, \mathbf{y}_n)$, the more accurate the estimates but the larger the simulation time and memory requirement. Even for moderate number of parameters, the required number of elements in the dictionary renders grid search intractable on a desktop computer. In addition, each new $\mathbf{y}_{\text{obs}}$, requires the computation and comparison of N matching scores $\mathrm{d}(\mathbf{y}_{\text{obs}}, \mathbf{y}_n)$, which can be costly if N is very large and if many inversions are desired. The regression or learning method that we propose and describe in the next sections is more efficient with respect to these aspects.

B. Dictionary-based learning (DBL) methods

In contrast to the DBM method, regression and learning methods can adapt to handle massive inversions of high dimensional data. The main principle is to transfer the computational cost, from 2-signal matchings to the learning of an inverse operator $\mathcal{F}^{-1}$. Equivalently, the goal is to learn a mapping from the fingerprint space to the parameter space, for any $\mathbf{y}$, with cost-less evaluation of $\mathcal{F}^{-1}(\mathbf{y})$. The dictionary $\mathcal{D}_f$ can be used to estimate $\mathcal{F}^{-1}$. Learning or regression methods adapted to high dimensions include inverse regression methods, *i.e.* sliced inverse regression [25], partial least squares [26], approaches based on mixtures of regressions with different variants, *e.g.* Gaussian locally linear mapping (GLLiM) [23], mixtures of experts [27], cluster weighted models [28] and kernel methods [9]. Inverse regression methods are flexible in that they reduce the dimension in a way optimal to the subsequent mapping estimation task that can itself be carried out by any kind of standard regression tool. In that sense, the inverse regression methods are said to be non-parametric or semi-parametric. Similarly, in [9], the authors propose a regression with an appropriate kernel function to learn the non-linear mapping. The procedure has the advantage to be semi-parametric but a serious limit is that the components of f are optimized in each dimension separately.

As regards application to MRF, deep learning tools have also been proposed by several groups [11]–[18]. For comparison purpose, a neural network with an architecture inspired from the DRONE network of [14] is thus considered. In the sequel, we use a fully-connected neural network, referred to as dictionary-based deep learning (DB-DL). The first and last layers are the S-node input and P-node output layers which match the sizes of the input signal $\mathbf{y}$ and the output parameters $\mathbf{x}$, respectively. The others are hidden layers. In [14], the DRONE model has 2 hidden layers of 300 nodes. Here, we consider instead 6 hidden layers and reduce the number of nodes to 100 in order to reduce the number of trainable parameters and limit overfitting issues. For example, considering $S = 100$ and $P = 3$, the number of trainable parameters of the neural network is 60 903. In addition, the activation functions

in DRONE, hyperbolic tangents (hidden layers) and sigmoid functions (output layer), are particularly susceptible to the vanishing gradient problem as already discussed in [14]. In our implementation, more robust ReLU activation functions are used instead.

C. Proposed dictionary-based statistical learning (DB-SL) method

In the same vein as [9] and in contrast to deep learning approaches, we propose to use the GLLiM method that exploits Gaussian mixture models [23]. Compared to other regression methods that focus on providing point-wise estimates, GLLiM provides a full probability distribution selected in a family of parametric models, *e.g.* mixture of Gaussian distributions, where the parameters are denoted by $\boldsymbol{\theta}$. The inversion operator is defined as $\mathcal{F}^{-1}(\mathbf{y}) = p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is estimated from the dictionary. More specifically, GLLiM handles the modeling of non-linear relationships with a piecewise linear model. Each $\mathbf{y}$ is seen as the noisy image of $\mathbf{x}$ obtained from a K-component mixture of affine transformations. This is modeled by introducing a latent variable $z \in \{1,\dots,K\}$ such that

$$\mathbf{y} = \sum_{k=1}^{K} \delta_k(z)\,(\boldsymbol{A}_k\mathbf{x} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k), \tag{2}$$

where $\delta_k(z)$ indicates membership in the region k of $\mathbf{x}$, having the value 1 if it belongs to the region and the value 0 otherwise. $\boldsymbol{A}_k$ is a $P \times S$ matrix and $\boldsymbol{b}_k$ a vector in $\mathbb{R}^P$ that characterize an affine transformation. Variable $\boldsymbol{\epsilon}_k$ corresponds to an error term in $\mathbb{R}^P$ which is assumed to be zero-mean and not correlated with $\mathbf{x}$, capturing both the modelling noise and the reconstruction error due to the affine approximations. In GLLiM, $\boldsymbol{\epsilon}_k$ follows a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}_k$, whose density function is denoted by $\mathcal{N}(.;\mathbf{0},\boldsymbol{\Sigma}_k)$. Then, $\mathbf{x}$ follows a mixture of K Gaussian distributions defined by $p(\mathbf{x}|z=k) = \mathcal{N}(\mathbf{x};\boldsymbol{c}_k,\boldsymbol{\Gamma}_k)$ and $p(z=k) = \pi_k$, where $\pi_k \in [0,1]$, $\boldsymbol{c}_k \in \mathbb{R}^L$ and $\boldsymbol{\Gamma}_k \in \mathbb{R}^{L\times L}$ are respectively the weight, the mean vector and covariance matrix of the kth Gaussian distribution. It follows that $\boldsymbol{\theta} = \{\pi_k,\boldsymbol{c}_k,\boldsymbol{\Gamma}_k,\boldsymbol{A}_k,\boldsymbol{b}_k,\boldsymbol{\Sigma}_k\}_{k=1:K}$ is the set of parameters defining the model. The conditional probability distribution of interest can then be derived as

$$p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k^*(\mathbf{y})\,\mathcal{N}(\mathbf{x};\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*,\boldsymbol{\Sigma}_k^*) \tag{3}$$

$$\text{with } w_k^*(\mathbf{y}) = \frac{\pi_k\,\mathcal{N}(\mathbf{y};\boldsymbol{c}_k^*,\boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K}\pi_j^*\,\mathcal{N}(\mathbf{y};\boldsymbol{c}_j^*,\boldsymbol{\Gamma}_j^*)}$$

with a parameterization $\boldsymbol{\theta}^* = \{\pi_k^*,\boldsymbol{c}_k^*,\boldsymbol{\Gamma}_k^*,\boldsymbol{A}_k^*,\boldsymbol{b}_k^*,\boldsymbol{\Sigma}_k^*\}_{k=1:K}$ easily expressed as an analytical function of $\boldsymbol{\theta}$. The mixture setting provides some guaranties that when choosing K large enough it is possible to approximate any reasonable relationship [27]. Automatic model selection criteria can also be used to select K (see [23]).

The $p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})$ distribution provides both estimates of the parameters $\mathbf{x}$ and information about the confidence to be placed in these estimates. In this work, estimates are defined through the expectation and the confidence indices as the square root of the covariance matrix diagonal element vector:

$$\hat{\mathbf{x}} = \mathbb{E}\left[\mathbf{x}|\mathbf{y};\boldsymbol{\theta}\right], \tag{4}$$

$$\text{CI} = \sqrt{\text{diag}\left(\text{Var}\left[\mathbf{x}|\mathbf{y};\boldsymbol{\theta}\right]\right)}, \tag{5}$$

with $\mathbb{E}\left[\mathbf{x}|\mathbf{y};\boldsymbol{\theta}\right] = \sum_{k=1}^{K} w_k^*(\mathbf{y})(\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*)$, and

$$\text{Var}\left[\mathbf{x}|\mathbf{y};\boldsymbol{\theta}\right] = \sum_{k=1}^{K} w_k^*(\mathbf{y})\left[\boldsymbol{\Sigma}_k^* + (\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*)(\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*)^T\right] - \left(\sum_{k=1}^{K} w_k^*(\mathbf{y})(\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*)\right)\left(\sum_{k=1}^{K} w_k^*(\mathbf{y})(\boldsymbol{A}_k^*\mathbf{y}+\boldsymbol{b}_k^*)\right)^T,$$

where $\text{diag}(\cdot)$ denotes the function returning the diagonal elements of a matrix. For the CI, computed from the estimated posterior $p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})$, to be a good indicator of the parameter estimation error, it is required that the inverted $\mathbf{y}$ follows the same model used to computed $\boldsymbol{\theta}$. The use of a unique $\boldsymbol{\theta}$ parameter for all inversions provides a great gain when massive inversions are required but it also assumes that the same model is valid for all fingerprints and that the dictionary $\mathcal{D}_f$ is a good representation of them. In practice, acquired fingerprints may come with different noise levels. An interesting feature of GLLiM is to adapt to this case at a very low cost. When the observed $\mathbf{y}$ comes with some covariance matrix $\boldsymbol{\Sigma}_\eta$ corresponding to a centered Gaussian noise variable $\boldsymbol{\eta}$, the initial dictionary $\mathcal{D}_f$ may not be fully adapted if it has not been generated with this same additional measurement error. Another training set should be simulated and used instead, with a corrected likelihood corresponding to $\mathcal{N}(\mathbf{y};f(\mathbf{x}),\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_{\boldsymbol{\eta}})$. Fortunately, it is straightforward to check that the structure of the Gaussian mixture approximation avoid the re-learning of the GLLiM model. Indeed, it suffices to change the estimated $\boldsymbol{\Sigma}_k$'s into $\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_\eta$ and to report this change when computing $\boldsymbol{\theta}^*$.

Because S is much larger than P in MRF applications, it is important that the model (2) involving $\boldsymbol{\theta}$ is estimated first and then used to derive model (3) that has a similar structure. The number of model parameters $\boldsymbol{\theta}$ can be drastically reduced by choosing constraints on covariance matrices $\boldsymbol{\Sigma}_k$ without inducing oversimplifications on the target model (3). In this work, equal diagonal covariance matrices are used as they yield the best results: for $1 \le k \le K$, $\boldsymbol{\Sigma}_k = \mathbf{D}_S$, where $\mathbf{D}_S \in \mathbb{R}^{S\times S}$ is a diagonal matrix. For example, with $S = 100$, $P = 3$ and $K = 50$, the number of parameters $\boldsymbol{\theta}$ is equal to 20 599 while a direct estimation of $\boldsymbol{\theta}^*$ would involve 272 999 parameters (see [23] for more details). The complexity of this GLLiM model is more generally in $\mathcal{O}(KPS)$.

D. Dictionary sampling strategy

The dictionary design depends on the sampling strategy of the parameter space. In MRF, regular grids of P-dimensional parameter values are generally considered. In [16], authors show that in a regression context, the random sampling strategy provides better estimation of the parameters than the use of a regular grid. However, this strategy entails a risk of imperfectly covering the parameter space coverage.

Fig. 1a shows a two-dimensional projection of $N = 1\,000$ points from a uniform grid in the 3D-hypercube ($P = 3$). Each

parameter is described by 10 different values. Note that with 1 000 points in 3D, only 100 distinct combinations appear in the 2D projection plane, each representing 10 different values of the third variable. This sampling scheme is not optimal in terms of information content. A significant improvement over the grid can be achieved by scrambled nets [29], [30]. In this paper, the Sobol sequence is generated [31] and scrambled [32]. We show the projection of $N = 1\,000$ points from the scrambled Sobol sequence (Fig. 1c) referred to as quasi-random in the remainder of the manuscript.

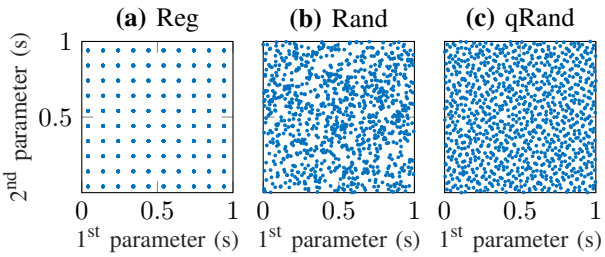

Fig. 1. Illustration of the dictionary sampling strategies. Plots show the 2-dimensional projections of $\boldsymbol{N}$ = 1 000 dictionary entries of the 3-dimensional parameter space ($\boldsymbol{P}$ = 3) obtained from **(a)** a regular grid sampling (Reg), **(b)** a random sampling (Rand) and **(c)** a quasi-random sampling (qRand) obtained from a scrambled Sobol sequence.

III. Analysis framework

A. Signals

1) Synthetic scalable signals: The sensitivity of the standard MRF signals to each parameter is variable. In addition, parameters cannot readily be added to the simulation tool that produces the MRF or the vascular MRF signals. To produce signals that are equally sensitive to each parameter and dependent on a variable number of parameters (*i.e.* P may be set to any value), scalable signals that mimic MRF signals are introduced in equation (6). The parameters of the synthetic scalable signal have physical units to help understand their structure but no physical meaning,

$$y_t = \left| \sum_{i=1}^{P} \sin\left(50\,\phi_i\,t\right) \exp\left(-\frac{t}{x_i}\right) \right|, \tag{6}$$

where x_i are the elements of $\mathbf{x}$, t varies from 10 to 1 000 ms in 10 ms steps ($S = 100$), the ϕ_i values are between 0.1 and 1 and $|\cdot|$ is the absolute value function. The values of parameters $\mathbf{x}$ are in the range of 10 to 1 000 ms. The vector ϕ is defined randomly such that none of the terms are equal. This makes the parameters x_i non-exchangeable: permutations of the $\mathbf{x}$ elements cannot lead to the same signal $\mathbf{y}$. Note that the relationship between $\mathbf{x}$ and $\mathbf{y}$ is non-linear. Examples of synthetic scalable signals are given in Supp. Fig. S1.

To create a noisy signal, a Gaussian zero-mean random variable with standard deviation σ_{noise} is added to the signal $\mathbf{y}$. The absolute value of the noisy signal is then considered. The signal-to-noise ratio is defined as: SNR $= I_{\max}/\sigma_{\text{noise}}$, where $I_{\max}$ is the maximum signal intensity. The same procedure is used to add noise to the following signals.

2) Synthetic standard MRF signals: Although the focus of our work is on vascular MRF, the proposed approach is applicable to other fingerprinting settings. We therefore evaluate the methods on standard MRF signals and in particular assess the impact of aliasing noise that might result from highly undersampled images. Using the aliasing noise model proposed by [33] and a simulated brain phantom with variable T_1, T_2 and off-resonance values, we produce signals similar to those introduced in [1]. Supp. S.VIII details the MRF signals simulated from the relaxation times T_1, T_2 and the off-resonance Δf parameters ($P = 3$).

3) Synthetic vascular MRF signals: Vascular MRF signals are ratio of the gradient echo sampling of the free induction decay and spin echo (GESFIDSE) signals measured pre- and post-injection of ultrasmall superparamagnetic iron oxide particles (USPIO) [3]. Eight sampled time points are obtained after the 90-degree pulse and 24 sampled time points after the 180-degree pulse ($S = 32$). These signals mainly depend on the vascular properties of the tissues, which in our application are specified by three parameters ($P = 3$): blood volume fraction (BVf), vessel size index (VSI) and tissue oxygen saturation (StO_2). The simulation tool [34] takes into account intrinsic relaxations, magnetic field perturbations induced by susceptibility interfaces (vessels), water proton diffusion and compartmentalization of the contrast agent in the vessels. Due to the complexity of the tool, simulations are extremely time-consuming. In our setting, simulation of a single synthetic vascular MRF signal takes about 10 seconds and a dictionary of 100 000 signals is generated on a 32-core high-performance computer (*Intel Xeon Gold 6130*, 2.1 GHz) in about 67 hours (different from [19]).

4) Acquired vascular MRF signals: Experimental data are acquired at 4.7 T (Bruker, Ettlingen, Germany) and have been introduced in [19]. The field of view was $30 \times 30\,\text{mm}^2$ and the voxel size was $234 \times 234 \times 800\,\mu\text{m}^3$. A turbo spin-echo sequence is acquired to identify anatomical structures and tumor tissues. Then, two GESFIDSE sequences ($S = 32$) are acquired, before and after UPSIO injection. See Supp. S.II-A for details on animal preparation and data acquisition.

B. Analysis pipeline

For the DB-SL method, the simulated and acquired data are processed using custom code developed in the Matlab environment (The MathWorks Inc., Natick, Ma, USA). This code and the numerical experiment scripts are available[1]. The DB-DL approach is implemented using the Deep Learning toolbox in the Matlab environment (R2019a; The Mathworks Inc., Natick, Ma, USA). Data from tumor bearing rats are processed using the *Medical software for Processing multi-Parametric images Pipelines*[2] [35].

1) Dictionary design: The dictionary is generated in two steps. First, combinations of parameter values in the parameter space are sampled using one of the sampling strategies in section II-D. Then, for each combination of parameter values, the associated fingerprint is simulated using either equation (6)

[1]https://github.com/nifm-gin/DB-qMRI
[2]https://github.com/nifm-gin/MP3

for synthetic scalable signals, the method described in Supp. S.VIII-A for synthetic MRF signals, or the simulation tool described in section III-A.3 for vascular MRF signals. For DBL methods, a low level, zero-mean Gaussian noise (typically SNR = 60) is added to the dictionary signals to promote robust learning [16] (see section IV-A.1). Note that in this work, we use magnitude signals since both scalable and vascular MRF signals are real-valued vectors. An evaluation using complex-valued signal is proposed for standard MRF signals (Supp. S.VIII).

2) Dictionary-based analysis: The dictionary is fully stored for the DBM method or summarized by a neural network or a parametric model $\boldsymbol{\theta}$ for the DBL methods. To obtain these models, we use a neural network architecture described in section II-B for the DB-DL and the GLLiM regression described in section II-C for the DB-SL. The model learning, a potentially time-consuming step, is performed only once, after the production of the dictionary.

In DBM, given an observed signal $\mathbf{y}_{\text{obs}}$, an estimate $\hat{\mathbf{x}}$ of the true $\mathbf{x}_{\text{obs}}$ is calculated as the minimization argument of equation (1) among the couples $(\mathbf{x}, \mathbf{y})$ in the dictionary. The observed signal and the signals in the dictionary are previously normalized to have unit Euclidean norm. The parameters are normalized to have zero mean and unit variance using scaling and translating factors that are then used to rescale the estimates.

In DB-DL, the trained neural network is used to compute an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}_{\text{obs}}$. 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. To ensure convergence, the maximum number of epochs is set to 2 000 (Supp. Fig. S2). The proposed neural network design provided better results than the initial design of [14] when using synthetic scalable signals (not shown).

In DB-SL, the estimate $\hat{\mathbf{x}}$ of $\mathbf{x}_{\text{obs}}$ is computed using equation (4) and a confidence index (CI) using equation (5). To obtain an accurate CI, an estimation of the signal noise variance is required. This estimate can be derived from the data SNR and then used as explained in section II-C to update $\boldsymbol{\theta}$ adequately. For DB-SL, the model requires only the setting of the calibration value K. In our study, the precise K value is not critical and different K values give similar results as long as they are sufficiently large ($K \geq 50$ in our study), see Supp. Fig. S3 for an illustration.

3) Closed-form expression fitting (CEF) analysis: Vascular MRF signals can also be analyzed by fitting of a non linear biophysical model [3], [36]. The closed-form expression fitting (CEF) analysis method refers to this multiple-operation procedure. First, relaxation rates are extracted by fitting the intensities of MRI signals (synthetic or acquired). Then, these relaxation rates are used to compute the BVf, VSI and StO_2 parameters using two equations, described in Supp. S.II-B.

4) Performance evaluation: To compare the methods performance in terms of parameter estimation, a set of M test signals is generated in the same way as for the dictionaries. The parameters values are randomly sampled in the parameter space and then the associated signals are computed. For each parameter, we compute the root mean square error (RMSE) as the square root of the quadratic mean of the differences between the estimated and the true parameter values.

IV. Results

A. Synthetic scalable signals

1) Effect of sampling strategy on parameter accuracy: To promote robust learning, we first evaluate the addition of noise to the training signals, as proposed in [16]. Results are reported in Supp. Fig. S4a-d for DB-SL and Fig. S4e-h for DB-DL. The use of training signals with a SNR of 60 appears optimal for the two DBL methods. We then investigate the impact of three parameter sampling strategies, regular, random and quasi-random, using synthetic scalable signals and the DBM, DB-DL and DB-SL methods. We consider successively $P = 3$, 5 and 7 and the corresponding numbers of entries in the dictionary $N = 216$, 1 024 and 2 187, respectively. For each value of P, $M = 1\,000$ test signals are generated from parameters randomly sampled in the parameter space. The RMSE between the estimated and the true parameter values is then computed (see section III-B.4) and divided by the number of parameters to obtain the average RMSE. To characterize the distribution of the average RMSE, the whole procedure is repeated 100 times (Table I).

Method	P	RMSE (ms)		
		Regular	Random	qRandom
DBM	3	**44.8** $\pm$ 1.0	56.9 $\pm$ 1.3	51.2 $\pm$ 1.1
	5	**88.4** $\pm$ 1.3	103.0 $\pm$ 1.2	99.0 $\pm$ 1.4
	7	**115.6** $\pm$ 1.3	120.6 $\pm$ 1.0	129.5 $\pm$ 1.3
Mean		**82.9**	93.5	93.2
DB-DL	3	12.0 $\pm$ 1.5	10.0 $\pm$ 1.8	**9.1** $\pm$ 1.8
	5	39.3 $\pm$ 3.0	22.1 $\pm$ 1.7	**20.4** $\pm$ 1.6
	7	63.3 $\pm$ 3.1	32.7 $\pm$ 1.2	**32.2** $\pm$ 1.6
Mean		38.2	21.6	**20.6**
DB-SL	3	17.2 $\pm$ 3.0	12.4 $\pm$ 0.7	**11.0** $\pm$ 0.5
	5	100.0 $\pm$ 17.8	35.1 $\pm$ 1.1	**33.0** $\pm$ 1.0
	7	150.4 $\pm$ 10.7	50.4 $\pm$ 1.3	**49.4** $\pm$ 1.2
Mean		89.2	32.6	**31.1**

TABLE I
IMPACT OF THE PARAMETER SAMPLING STRATEGY ON ACCURACY. AVERAGE RMSE ($M = 1\,000$ TEST SYNTHETIC SCALABLE SIGNALS) ACROSS THE PARAMETER ESTIMATES FOR THE THREE SAMPLING STRATEGIES AND $P = 3$, 5 AND 7 PARAMETERS (MEAN $\pm$ STANDARD DEVIATION; BEST VALUE PER LINE IS IN BOLD).

Regardless of the sampling strategy, the average RMSE increases with P, the number of parameters. For DBM, the non-regular sampling strategies yield an increased RMSE (12.8 % for random and 12.4 % for quasi-random), compared to the regular sampling and across all conditions. As reported previously, random sampling gives a 53.5 % lower average RMSE than regular sampling, for the two DBL methods and across all conditions [16]. The quasi-random sampling further reduces the average RMSE by 4.6 % for the two DBL methods and across all conditions. These observations are also valid for other conditions presented in Supp. Fig. S5 for DBM and DB-SL. For DB-DL, there is a reduction of 18.8 %, 44.0 % and 47.3 % in average RMSE between regular and quasi-random

sampling for 3, 5 and 7 parameters, respectively. A similar improvement is observed for the DB-SL method (reduction in RMSE of 14.4 %, 60.3 % and 63.8 %). Therefore, in the following, a regular sampling is used for DBM and a quasi-random sampling is used for the DBL methods.

2) Impact of the dictionary size and SNR on parameter accuracy: To study this impact for DBM and DBL methods, we generate four scalable signals dictionaries for $P = 5$ and 7 parameters (a total of 8 conditions). The number of dictionary entries N is chosen so as to keep similar densities, *i.e.* a constant number of values per parameter (for $P = 5$: $N = 3^5$, 4^5, 5^5 and 6^5 and for $P = 7$: $N = 3^7$, 4^7, 5^7 and 6^7). For each condition, we evaluate the average RMSE, using $M = 10\,000$ signals. To characterize the impact of SNR, the procedure is repeated for test signals with SNR between 10 and 110 (Fig. 2). This experiment allows to assess the impact of the dictionary size and SNR for each method separately (Fig. 2). Comparison between methods can also be made but is better illustrated in Supp. Fig. S6.

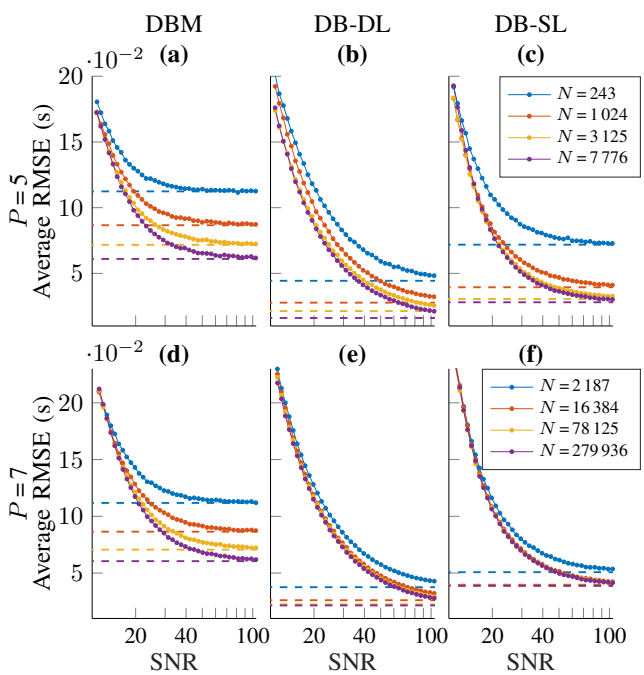

Fig. 2. Impact of dictionary size and SNR on DBM and DBL methods, using synthetic scalable signals. Average RMSE are given with respect to the SNR for different numbers of parameters and dictionary entries. Average RMSE ($M = 10\,000$ test signals) for the **(a, d)** DBM, **(b, e)** DB-DL and **(b, d)** DB-SL methods. The upper row **(a-c)** shows the results for $P = 5$ parameters and the lower row **(d-f)** for $P = 7$ parameters. The dashed lines represent the average RMSE in the absence of noise on the test signals.

As expected for the DBM method, the average RMSE decreases as the number of entries N increases. The average RMSE decreases as the SNR increases to about SNR = 60 and then plateaus near the value obtained in absence of noise. For the DBL methods, the average RMSE also decreases as the SNR increases. Again, the highest SNR yields an average RMSE close to that obtained in the absence of noise. For the DBM method, the average RMSE is comparable between 5 and 7 parameters. N has a lower impact for DBL methods than for DBM. For $P = 7$ and across SNR values, between the smallest and the largest dictionary size, the average RMSE decreases by 29.1 ± 14.9 % for the DBM method, while it decreases by only 18.3 ± 9.5 % for the DB-DL method and by 13.3 ± 7.0 % for the DB-SL method. Moreover, the two highest N yield similar average RMSE for the DBL methods (differences are only 2.6 ± 0.9 % for DB-DL and 0.1 ± 0.7 % for DB-SL), suggesting that an increase in the number of entries would not further improve the average RMSE. In terms of methods comparison, the average RMSE obtained with DBL methods are always lower than with DBM in all configurations of N and P. In particular, compared to DBM, DB-SL reduces the average RMSE, for 5 (resp. 7 parameters) by 19.9 ± 11.8 % (resp. 5.8 ± 6.5 %), while reducing the number of entries by a factor of 8 (resp. 128). Similar conclusions hold for additional values of P and N as reported in Supp. Fig. S6, which includes the settings of Fig. 2, also shows that for SNR ≤ 30 (across all simulation conditions), the DB-SL RMSE are always smaller than the DB-DL RMSE (7.7 ± 7.0 % smaller); while for SNR ≥ 60, the DB-DL RMSE are always smaller than the DB-SL RMSE (22.8 ± 6.7 % smaller).

By eliminating the costly dictionary matching operation, DBL methods can greatly reduce computation time when N increases. For 7 parameters and $N = 78\,125$, inverting 10 000 test signals takes 1.3 s with DBM, 5 ms for DB-DL, and 0.9 s with DB-SL. When N increases to 279 936, the estimation time increases to 4.2 s for DBM while it remains stable at 6 ms for DB-DL and 0.9 s for DB-SL. In terms of memory, these dictionaries require 66.9 MB ($N = 78\,125$) and 239.6 MB ($N = 279\,936$) whereas they required only 0.3 MB (DB-DL) and 4.3 MB (DB-SL) once summarized by a model. Note that the differences in estimation time/memory requirement between DB-DL and DB-SL are not coming from differences in model complexities but more likely from the methods implementation (*e.g.* for $S = 100$ and $P = 3$, 20 599 DB-SL parameters correspond to about 4.2 MB and 60 903 DB-DL parameters to about 0.3 MB). Additional comparisons between methods in terms of speed and memory are given in Supp. Table S1. We observe that the number of dictionary entries has little effect on the estimation time or on the memory size once the model is learned.

3) Boundary behavior: The DBL methods estimate parameter values using a continuous function that is not limited to the parameter space covered by the dictionary entries. To investigate the behavior of DBM and DBL methods outside of the limits of this parameter space, we define a dictionary ($N = 10\,000$) composed of two disjoint blocks in the parameter space, generate $M = 2\,000\,000$ test signals and evaluate the average RMSE for each parameter value. The experiment is then repeated after the addition of three new dictionary entries, outside of the two initial blocks.

The three methods yield similar estimation accuracy in the blocks covered by the dictionary entries (Fig. 3a-c). Outside of these blocks, the average RMSE obtained with the DBM method increases with the distance to the blocks. For the DB-DL and DB-SL methods, the average RMSE remains below 100 ms, well beyond the limits of the dictionary blocks. In particular, the error is reduced in between the two blocks. As

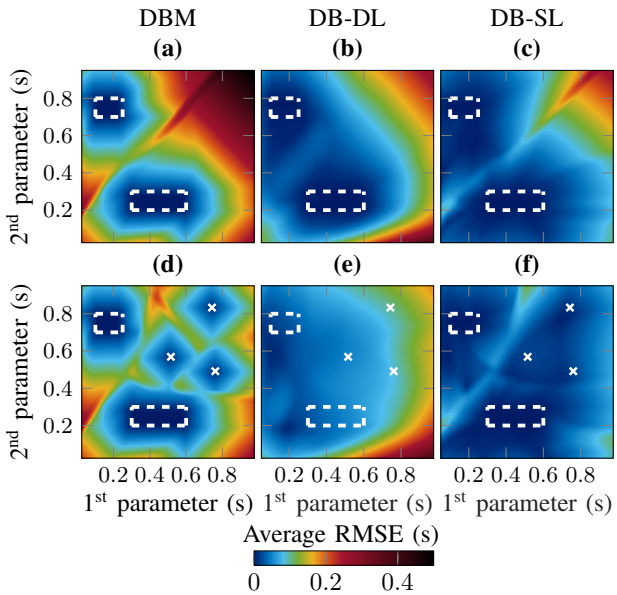

Fig. 3. Estimation accuracy outside the limits of the parameter space covered by the dictionary, using synthetic scalable signals. Average RMSE ($M = 2\,000\,000$ test signals) in the parameter space ($P = 2$) obtained **(a, d)** with the DBM, **(b, e)** with the DB-DL and **(c, f)** with the DB-SL method. The white dashed lines delimit the blocks covered by the dictionaryand the 3 white marks in **(d-f)** are 3 additional dictionary entries. The average RMSE is computed from signals in a 50×50 ms sliding window, moving in 5 ms steps in the parameter space.

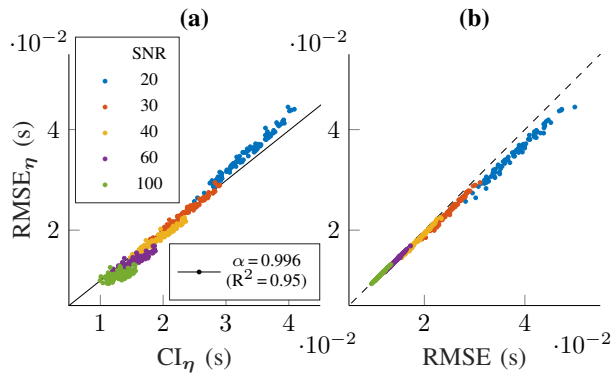

Fig. 4. $RMSE_\eta$ vs confidence index (CI_η) and RMSE (non-updated model), using synthetic scalable signals. $SNR_{test} = 20$ (blue), 30 (orange), 40 (yellow), 60 (purple) and 100 (green). $M = 10\,000$ test signals. **(a)** The black line represents the proportional regression coefficient α between $RMSE_\eta$ and CI_η for all SNR values. R^2 is the coefficient of determination. **(b)** The dashed black line is the identity function.

expected, for the DBM method, extra dictionary entries yield improved estimates near each additional entry. For DBL methods, these extra entries decrease the RMSE in a much larger space than the neighborhood of the entries (Fig. 3e-f). This is particularly true for the DB-SL method (RMSE < 100 ms). However surprisingly, for the DB-DL method, the RMSE inside of the blocks is altered by the addition of the three extra entries, while this is not the case for DB-SL. A similar experiment, using synthetic vascular signals, is reported in section IV-B.3.

4) Confidence index: We investigate the relationship between the CI, available with the DB-SL method, and the RMSE. We generate $N = 10\,000$ dictionary entries and $M = 10\,000$ test signals. We then add different noise levels to the test signals to obtain a SNR = 20, 30, 40, 60 and 100. A single initial regression model is computed. For each SNR, this model is then updated based on the noise level (denoted by η) which corresponds to the SNR values of the test signals (see section II-C). We compute the RMSE and CI for the initial model (*i.e.* without accounting for the noise level) and $RMSE_\eta$ and CI_η using the updated model. For each SNR value, the experiment is repeated 100 times. Supp. Fig. S7, shows that the non-updated CI is proportional to but not equal to the RMSE in the SNR value range. Supp. Fig. S7 also shows that the scaling factor between RMSE and non-updated CI depends on the added noise level.

As expected, $RMSE_\eta$ and CI_η increase as the SNR decreases (Fig. 4a). $RMSE_\eta$ and CI_η are proportional and comparable in the simulated SNR range (slope: 0.99, $R^2 = 0.95$). Note that CI_η may slightly under or over-estimate the $RMSE_\eta$ (mean difference: 7.8 %). Overall, CI_η appears to be a good indicator of the $RMSE_\eta$. Interestingly, the inclusion of noise in the model slightly improves the estimation accuracy. On average, the $RMSE_\eta$ is 4.11 % lower than the RMSE (Fig. 4b). In the following, for DB-SL, RMSE and CI refer to $RMSE_\eta$ and CI_η (updated model). Outside of the dictionary parameter space, the CI is underestimated and no longer reliable (Supp. Fig. S8). This issue is further investigated in section IV-B.3

B. MRF signals

1) Synthetic standard MRF signals: We compare the DBM and the DBL methods using the standard MRF signals proposed by Ma *et al.* [1]. These signals are complex-valued vectors ($S = 1000$). DBM extends straightforwardly to complex values and DBL methods are performed by concatenating the real and imaginary parts of the initial signals, as proposed in other studies [11], [13]. 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). The comparison is carried out for two dictionary sizes, $N = 4\,096$ (small) and $N = 226\,981$ (large). The average RMSE is computed on $M = 10\,000$ test signals, for SNR values between 10 and 110. More information and detailed results are provided in Supp. S.VIII.

We observe that each method has its advantages with results that depend on the parameter of interest (Supp. Fig. S9). For the small dictionary, DB-SL outperforms DBM in all conditions and DB-DL at low SNR (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 high SNR (SNR > 70). In all other cases, DBM provides the best estimations for T_1, DB-SL provides the best estimations for T_2 (SNR < 70), while for Δf, DB-SL is more accurate at low SNR (SNR < 30) and DBM at high SNR. For that parameter, Both methods outperform DB-DL whatever the SNR level.

To gain further insights into the RMSE differences between the three methods, we carry out an additional bias-variance analysis keeping the same two dictionary sizes for learning.

Following a similar analysis to [10], we use a human brain phantom resulting in $M = 7\,622$ test signals with added noise for a SNR = 40 (see details in Supp. S.VIII-B). In terms of RMSE comparison the same conclusions hold but the bias-variance decomposition exhibits that lower RMSE are essentially coming from lower variances. In the small dictionary case, for T_1 and T_2, DB-DL yields mean bias about twice smaller than DB-SL but with larger variances leading to higher RMSE for DB-DL (see table in Supp. Fig. S10d). The brain phantom allows the representation of errors as brain maps from which it is visible that compared to DBM, DBL methods yield tissue-dependent errors with brain structures visible in the maps (Supp. Fig. S10). The DBM bias, variances and RMSE are always higher than the DB-SL ones for T_1 and T_2. In contrast, DBM improves for the larger dictionary and provides smaller bias but not always enough to compensate its higher variances. For Δf, DBM provides the smallest variances and RMSE. We suspect that a better handling of complex-valued computation could be the reason for this superior performance.

Keeping the previous setting, the robustness of the methods to aliasing noise is then assessed using the tool introduced in [33] (details in Supp. S.VIII-C). Typical undersampled MRF signals obtained this way are shown in Supp. Fig. S11d-f while Fig. S11g-i shows the RMSE for T_1, T_2 and Δf for increasing levels of aliasing noise. DB-DL appears less robust than DB-SL. Overall, DB-SL provides good estimate accuracy even in presence of aliasing noise and is the best performing method when estimating T_1 and T_2 for the strongest aliasing noise levels used in this study and this for the two dictionary sizes. However, for Δf, the DBM method remains the method of choice, except for a low aliasing noise level in the small dictionary case. These results are consistent with the one obtained using the scalable signals.

2) Synthetic vascular MRF signals: We compare the three dictionary-based methods and the CEF method. The dictionaries (grid and quasi-random sampling) are simulated with a BVf between 0.25 and 30 %, a VSI between 0.5 and 50 µm and a StO_2 between 30 and 95 %. Among the 170 100 combinations, some signals cannot be produced, due to simulation constraints (*e.g.* a very large BVf cannot be produced with distant, small, vessels or small BVf with large vessels). The obtained N values reduce then to $N = 164\,524$ for the grid and to $N = 167\,216$ for quasi-random sampling.

For each method, $M = 100\,000$ test signals (SNR = 100) are generated. To analyze the BVf RMSE, test signals are divided into three parts: small, medium and large vessel sizes; for VSI RMSE: low, medium and large blood volumes.

For all vessel diameters, the BVf RMSE increases with BVf (Fig. 5a-c). The DBM and CEF methods yield similar RMSE for BVf values below 10 %. The DB-SL method always yields the lowest error with an RMSE of 3.10 % for CEF, 3.61 % for DBM, 2.14 % for DB-DL and 1.87 % for DB-SL.

For VSI values smaller than 15 µm, the behavior of the RMSE is similar in all methods, except for the large BVf range. Above 15 µm, the CEF method yields larger errors than the three dictionary-based approaches and the RMSE obtained with CEF is linearly correlated with the VSI value

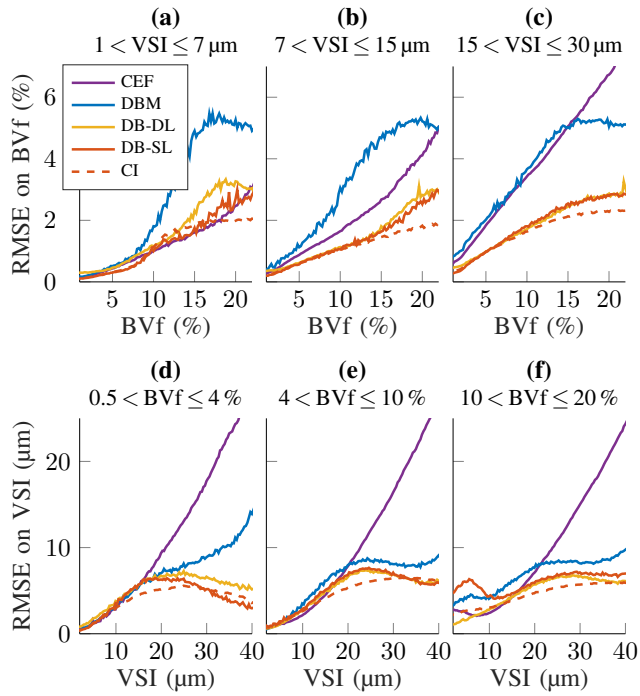

Fig. 5. Comparison of the RMSE on BVf and VSI obtained with the three dictionary-based methods (DBM, DB-DL, and DB-SL) and with the closed-form expression fitting (CEF) method, using synthetic vascular MRF signals. **(a, b, c)** show RMSE ($M = 100\,000$ test signals) on BVf for three ranges of VSI and **(d, e, f)** show the RMSE on VSI for three ranges of BVf. The dashed lines represent the average confidence indices (CI) on BVf (first row) and VSI (second row) obtained with DB-SL. The data are shown after 1-dimensional sliding window filtering (3 % for BVf and 5 µm for VSI). The dictionary dimensions are $\boldsymbol{P} = 3$, $\boldsymbol{S} = 32$ and $\boldsymbol{N} = 164\,524$ for DBM, and $\boldsymbol{N} = 167\,216$ for the 2 DBL methods.

($R^2 \geq 0.99$). This linear behavior has already been reported in [37]. Considering VSI, DB-SL yields a 1.77 µm smaller RMSE than DBM, on average with an RMSE of 12.70 µm for CEF, 7.55 µm for DBM, 45.76 µm for DB-DL, and 5.78 µm for DB-SL. The CI appears again to be a good indicator of the RMSE, with maximum differences between CI and RMSE of 1.77 % for BVf and 4.60 µm for VSI and average differences of 0.43 % and 1.24 µm.

3) Acquired vascular MRF signals: The DBL methods are then applied to vascular MRF signals collected from rats bearing 9L and C6 tumors. We quantify BVf, VSI and StO_2 with both DB-DL and DB-SL, using two dictionary sizes. The large dictionary ($N = 170\,100$) is the one used previously in section IV-B.2. The small dictionary ($N = 4\,320$) is simulated with BVf between 0.33 and 12 %, VSI between 1 and 20 µm and StO_2 between 40 and 90 %.

All methods yield consistent estimates (Fig. 6), in which the tumor and the large vessels can be easily depicted on BVf and VSI maps. StO_2 appears constant in healthy tissues. However, for DBM, there are many isolated high values that correspond to estimates at the dictionary boundaries, suggesting non accurate estimates. In [19], the authors proposed to apply a spatial Gaussian filtering to increase the SNR. Using the DBM method, this additional step allowed them to produce more spatially homogeneous maps. Interestingly, with the DBL methods, this step is no longer required. DBL is therefore

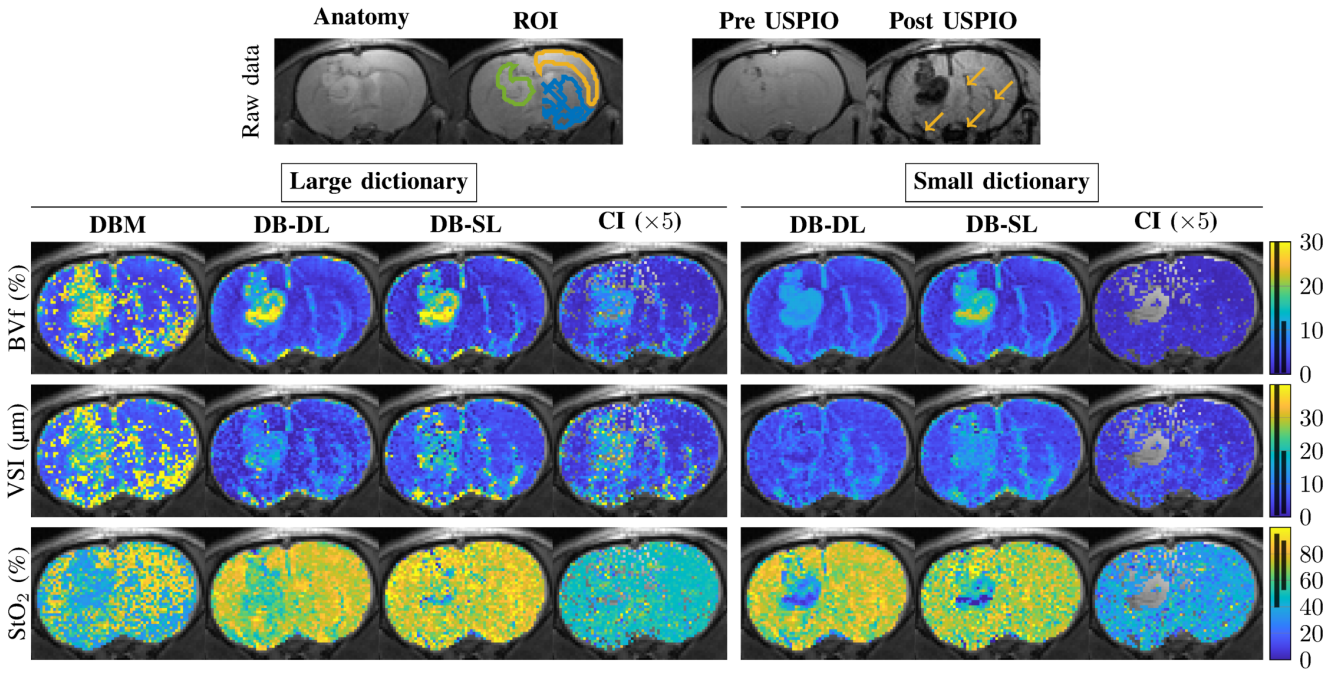

Fig. 6. Maps of vascular parameter estimates of a 9L rat tumor model. The first row shows the anatomical image and regions of interest (left) and the MGEFIDSE pre and post USPIO injection (right) for the second echo time (6.3 ms). The tumor, cortex and striatum are respectively delineated with green, 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 DBM (first column), DB-DL (second and fifth) and DB-SL (third and sixth columns). The fourth and seventh 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 dictionary: $N = 164\,524$ for DBM; 70 values for BVf between 0.25 and 30 %, 90 values for VSI between 0.5 and 50 µm and 27 values for StO_2 between 30 and 95 % and $N = 167\,216$ for DBL. Small dictionary: $N = 4\,218$ for DBM; 36 values for BVf between 0.33 and 12 %, 20 values for VSI between 1 and 20 µm and 6 values for StO_2 between 40 and 90 % and $N = 4\,119$ for DBL. Missing values in CI maps correspond to estimates outside the parameter space covered by the dictionary, where CI is no longer reliable.

more likely to preserve small structure information, which may be otherwise removed by spatial filtering.

For the large dictionary, within the tumor, the mean BVf and VSI obtained with the DBM and DBL methods are similar, but with the DBL methods, we can distinguish sub-regions within the lesion. Mean values and standard deviations in the tumor are 19.85 ± 4.99 %, 14.83 ± 6.20 µm for DBM, 17.15 ± 6.12 %, 12.52 ± 5.16 µm for DB-DL, and 18.02 ± 6.31 %, 16.28 ± 6.21 µm for DB-SL. For StO_2, the DBL methods provide significantly larger values than DBM, closer to the expected values for healthy tissue [38]. Values for striatum are 55.03 ± 16.59 % for DBM, 63.11 ± 10.79 for DB-DL, and 72.4 ± 10.4 % for DB-SL.

For the small dictionary, the contrasts are similar. The estimates obtained by the DBM method are limited to the space spanned by the dictionary, while the DBL methods yield estimates outside this range and closer to the parameter values obtained with the large dictionary. For example with DB-SL, the mean tumor BVf is 14.30 % (18.2 % for the large) while the maximum dictionary value is 12 %. CI values related to estimates outside of the parameter space covered by the dictionary are removed from the CI maps, as we suspect that they are not reliable (Supp. Fig. S8).

The DBL method with a small dictionary produces estimates similar to those obtained with a large dictionary, except for the largest values that are underestimated. This results in a slight reduction in mean ROI values. Another way to reduce the dictionary size is to use a dictionary made of blocks with a few additional entries to reduce the RMSE, as presented in section IV-A.3 or to reduce the number of entries by subsampling the dictionary without affecting its range (Supp. Fig. S12 and S13). In these two cases, the DBL methods yield similar parameter maps as the ones obtained with the large dictionary (Supp. Fig. S13 and S14), with the maps resulting from the sub-sampled dictionary being closer. Moreover, the CI maps obtained with the sub-sampled dictionary exhibit less missing values in the tumor than that obtained with the block dictionary and allows to recover more reliable CI values. Note that DB-SL is the first method to provide error maps for BVf, VSI and StO_2.

A second example of a C6 rat tumor model is given in Supp. Fig. S15. The results are similar to those observed for the 9L tumor. DBL and DBM produce comparable parameter maps. With DBL, however, maps are more spatially homogeneous and, using DB-SL, CI maps can be produced.

At last, averaging parameter values across 8 animals (4 for each tumor model), Supp. Fig. S16 shows, by regions of interest (ROI), mean vascular estimates obtained with the large dictionary. Overall, values obtained with all methods are comparable, even if some differences may readily be observed. Further analyses would require a ground truth, which is not available. The mean CIs in the tumor are 3.14 % for BVf, 5.91 µm for VSI and 16.19 % for StO_2, while in the cortex the mean CI are 1.09 % for BVf, 2.29 µm for VSI and 16.54 % for

StO_2. These results are in agreement with previous results that pointed better estimates for BVf than for VSI and low signal sensitivity to StO_2 [19]. The CI is therefore on average 2 to 3 times lower in the tumor than in the cortex for BVf and VSI but is similar for StO_2.

V. Discussion

This work introduces a statistical method for estimating vascular MRF parameters based on dictionary learning. This method is compared to the standard matching (DBM) method and to a deep learning (DB-DL) method. Overall, the two learning methods yield more accurate estimates than DBM, whatever the type of synthetic signal and for most noise levels (thermal and aliasing types). Learning methods also produce accurate estimates further away from the dictionary boundaries than DBM. Finally, learning methods compute estimates faster than DBM and with smaller memory requirements. Among the learning methods, the DB-SL method appears to perform best for signals with low SNR (typically below 40) and DB-DL for signals with high SNR (above 70). The additional confidence index (CI), provided by DB-SL, appears as a reliable estimate of the RMSE within the parameter range covered by the dictionary. When considering acquired vascular MRF signals, we observe that the number of dictionary entries can be divided by about 40 using the learning methods and still lead to accurate maps. The maps produced with the DBL methods are spatially more homogeneous than those obtained with the DBM method while preserving structures not observed with DBM, notably in the lesions. The additional tissue contrast provided by the DBL methods could therefore contribute to improved tumor characterization [39].

Regarding the design of the dictionary, we first observe that a quasi-random sampling of the parameter space gives more accurate estimates for the two learning methods, in agreement with [16], and that a regular sampling is more suited for DBM. We then evaluate two strategies to further reduce the simulation cost while maintaining estimates accuracy: a block dictionary, possibly with a few additional entries at distance from these blocks, and a sub-sampled dictionary. When DBM is carried out with a block dictionary, the RMSE quickly increases with the distance to the blocks (Fig. 3, Supp. Fig. S12). Dictionary undersampling, *i.e.* reducing the dictionary density, also increases the RMSE on parameter estimates (Fig. 2, Supp. Fig. S6). DBM is therefore not the method of choice to reduce the simulation cost. In contrast, learning methods maintain low RMSE at larger distances from the blocks and better exploit additional entries in the dictionary (Fig. 3, Supp. Fig. S12). Regarding DB-DL, [14] reported increased deviations from the true values at the boundaries of the training space, likely due to the vanishing gradient of the activation function in these regions. In this work, using a different activation function (*i.e.* ReLU), we observe that the DB-DL performance remains good although not as stable as for DB-SL (Fig. 3, Supp. Fig. S12). Note however that the gain or loss due to additional entries may depend on the nature and sensitivity of the relationship between signals and parameters. For instance, additional entries outside of the blocks degrade the DB-DL RMSE in the blocks when using synthetic scalable signals (Fig. 3b,e) but this phenomenon is not observed for the synthetic vascular signals (Supp. Fig. S12). For DB-SL, a remaining not satisfying feature in the block setting is the accuracy loss on CI outside of the blocks. This behavior seems to be related to the parameter range covered by the dictionary and can be overcome with the use of a sub-sampled dictionary (Supp. Fig. S14). With a reduced simulation cost, similar to that of a block dictionary, CI estimates remain comparable to that obtained with the reference approach (large dictionary). This suggests that optimal results could be obtained with DB-SL by considering at the learning phase (i) parameter intervals that correspond to the expected ones and (ii) a variable parameter density, with higher densities where higher accuracy is desired.

Regarding the inversion, the regression approach usually involves a calibration. In GLLiM, the number K of Gaussian distributions is the only calibration value to be adjusted. This can be done automatically using a standard information criterion such as AIC or BIC [40], [41], as illustrated in [23], but at the cost of additional learning time. In contrast, neural networks [11]–[18] used to solve similar inverse problems are very sensitive to their many complex calibration settings: architecture, batch sizes, learning rates, among others. Moreover, other networks such as recurrent neural network [13], [42] have been proposed for MRF and could also be evaluated. The single parameter calibration of DB-SL appears more interpretable, more reproducible and less critical (Supp. Fig. S3). Despite these differences in algorithm architecture, the learning times of DB-DL and DB-SL are in the same range (Supp. Tab. S1). A finer comparison of their actual computational costs would require to put their implementations on an equal footing, which is out of the scope of this study.

Beyond vascular MRF, the proposed approach could also benefit standard MRF data. With a small dictionary composed of synthetic standard MRF signals, we observe that DB-SL outperforms DB-DL and DBM for high termal noise levels (SNR below 40) in estimating T_1 and T_2. Learning methods also outperform DBM in case of strong aliasing noise and small dictionary. The results obtained on Δf (Supp. Fig. S9, S10 and S11) suggest that DB-DL and DB-SL estimates could be more accurate if complex-valued signals were properly handled, instead of the concatenation approach used in this study. Further work includes the generalization of GLLiM to complex Gaussian distributions while complex-valued neural networks could also be investigated.

In conclusion, this first evaluation of the DB-SL method appears promising. It reduces the simulation time and the memory required for dictionary storage, improves parameter accuracy, reduces the estimation time and provides a first confidence index on parameter estimates. The DB-SL method has been tested on different types of MRF signals. It has the potential to scale efficiently when the number of parameters increases and to offer the possibility to account for more physiological and experimental contributions to the signal.

REFERENCES

[1] 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.

[2] B. Bipin Mehta, S. Coppo, D. Frances McGivney, J. Ian Hamilton, Y. Chen, Y. Jiang, D. Ma, N. Seiberlich, V. Gulani, and M. Alan Griswold, "Magnetic resonance fingerprinting: a technical review," *Magnetic Resonance in Medicine*, vol. 81, no. 1, pp. 25–46, 2019.

[3] T. Christen, N. A. Pannetier, W. W. Ni, D. Qiu, M. E. Moseley, N. Schuff, and G. Zaharchuk, "MR vascular fingerprinting: A new approach to compute cerebral blood volume, mean vessel radius, and oxygenation maps in the human brain," *NeuroImage*, vol. 89, pp. 262–270, 2014.

[4] D. F. McGivney, E. Pierre, D. Ma, Y. Jiang, H. Saybasili, V. Gulani, and M. A. Griswold, "SVD compression for magnetic resonance fingerprinting in the time domain," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2311–2322, 2014.

[5] S. F. Cauley, K. Setsompop, D. Ma, Y. Jiang, H. Ye, E. Adalsteinsson, M. A. Griswold, and L. L. Wald, "Fast group matching for MR fingerprinting reconstruction," *Magnetic Resonance in Medicine*, vol. 74, no. 2, pp. 523–528, 2015.

[6] B. Zhao, K. Setsompop, E. Adalsteinsson, B. Gagoski, H. Ye, D. Ma, Y. Jiang, P. Ellen Grant, M. A. Griswold, and L. L. Wald, "Improved magnetic resonance fingerprinting reconstruction with low-rank and subspace modeling," *Magnetic Resonance in Medicine*, vol. 79, no. 2, pp. 933–942, 2018.

[7] M. Yang, D. Ma, Y. Jiang, J. Hamilton, N. Seiberlich, M. A. Griswold, and D. McGivney, "Low rank approximation methods for MR fingerprinting with large scale dictionaries," *Magnetic Resonance in Medicine*, vol. 79, no. 4, pp. 2392–2400, 2018.

[8] J. Assländer, M. A. Cloos, F. Knoll, D. K. Sodickson, J. Hennig, and R. Lattanzi, "Low rank alternating direction method of multipliers reconstruction for MR fingerprinting," *Magnetic Resonance in Medicine*, vol. 79, no. 1, pp. 83–96, 2018.

[9] G. Nataraj, J. F. Nielsen, C. Scott, and J. A. Fessler, "Dictionary-free MRI PERK: Parameter estimation via regression with kernels," *arXiv*, vol. 37, no. 9, pp. 2103–2114, 2017.

[10] 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.

[11] 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.

[12] E. Hoppe, G. Körzdörfer, T. Würfl, J. Wetzl, F. Lugauer, J. Pfeuffer, and A. Maier, "Deep learning for magnetic resonance fingerprinting: A new approach for predicting quantitative parameter values from time series," *Studies in Health Technology and Informatics*, vol. 243, pp. 202–206, 2017.

[13] 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.

[14] O. Cohen, B. Zhu, and M. S. Rosen, "MR fingerprinting deep RecOnstruction NEtwork (DRONE)," *arXiv*, vol. 80, no. 3, pp. 885–894, 2017.

[15] F. Balsiger, A. S. Konar, S. Chikop, V. Chandran, O. Scheidegger, S. Geethanath, and M. Reyes, "Magnetic Resonance Fingerprinting Reconstruction via Spatiotemporal Convolutional Neural Networks," in *arXiv*. Springer, 2018, pp. 39–46.

[16] M. Barbieri, L. Brizi, E. Giampieri, F. Solera, G. Castellani, C. Testa, and D. Remondini, "Circumventing the Curse of Dimensionality in Magnetic Resonance Fingerprinting through a Deep Learning Approach," *arXiv*, 2018.

[17] P. Song, Y. C. Eldar, G. Mazor, and M. R. Rodrigues, "HYDRA: Hybrid Deep Magnetic Resonance Fingerprinting," *arXiv*, vol. 46, no. 11, pp. 4951–4969, 2019.

[18] M. Golbabaee, D. Chen, P. A. Gómez, M. I. Menzel, and M. E. Davies, "Geometry of deep learning for magnetic resonance fingerprinting," in *arXiv*. IEEE, 2018, pp. 7825–7829.

[19] B. Lemasson, N. Pannetier, N. Coquery, L. S. Boisserand, N. Collomb, N. Schuff, M. Moseley, G. Zaharchuk, E. L. Barbier, and T. Christen, "MR vascular fingerprinting in stroke and brain tumors models," *Scientific Reports*, vol. 6, p. 37071, 2016.

[20] F. Boux, F. Forbes, J. Arbel, and E. L. Barbier, "Dictionary-Free Mr Fingerprinting Parameter Estimation Via Inverse Regression," in *Proceedings of the 26th Annual Meeting, ISMRM, Paris*, 2018, pp. 5–6.

[21] D. McGivney, A. Deshmane, Y. Jiang, D. Ma, C. Badve, A. Sloan, V. Gulani, and M. Griswold, "Bayesian estimation of multicomponent relaxation parameters in magnetic resonance fingerprinting," *Magnetic Resonance in Medicine*, vol. 80, no. 1, pp. 159–170, 2018.

[22] S. Metzner, G. Wübbeler, and C. Elster, "Approximate large-scale Bayesian spatial modeling with application to quantitative magnetic resonance imaging," *AStA Advances in Statistical Analysis*, vol. 103, no. 3, pp. 333–355, 2019.

[23] 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.

[24] J. Assländer, "A Perspective on MR Fingerprinting," *Journal of Magnetic Resonance Imaging*, pp. 1–10, 2020.

[25] K.-C. Li, "Sliced Inverse Regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, 1991.

[26] R. D. Cook and L. Forzani, "Partial least squares prediction in high-dimensional regression," *Annals of Statistics*, vol. 47, no. 2, pp. 884–908, 2019.

[27] H. D. Nguyen, F. Chamroukhi, and F. Forbes, "Approximation results regarding the multiple-output mixture of linear experts model," *arXiv*, 2017.

[28] S. Ingrassia, S. C. Minotti, and G. Vittadini, "Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions," *Journal of Classification*, vol. 29, no. 3, pp. 363–401, 2012.

[29] B. J. Collings and H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods.* Siam, 1993, vol. 88, no. 422.

[30] A. B. Owen, "Randomly Permuted (t,m,s)-Nets and (t, s)-Sequences," in *Monte Carlo and quasi-Monte Carlo methods in scientific computing*. Springer, 1995, pp. 299–317.

[31] P. Bratley and B. L. Fox, "Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator," *ACM Transactions on Mathematical Software (TOMS)*, vol. 14, no. 1, pp. 88–100, 1988.

[32] J. Matoušek, "On the {L_2}-discrepancy for Anchored Boxes," *Journal of Complexity*, vol. 14, no. 4, pp. 527–556, 1998.

[33] 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.

[34] N. A. Pannetier, C. S. Debacker, F. Mauconduit, T. Christen, and E. L. Barbier, "A Simulation Tool for Dynamic Contrast Enhanced MRI," *PLoS ONE*, vol. 8, no. 3, p. e57636, 2013.

[35] C. Brossard, O. Montigon, F. Boux, A. Delphin, T. Christen, E. L. Barbier, and B. Lemasson, "MP3: Medical Software for Processing Multi-Parametric Images Pipelines," *Frontiers in Neuroinformatics*, vol. 14, p. 53, 2020.

[36] 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.

[37] I. Troprès, N. Pannetier, S. Grand, B. Lemasson, A. Moisan, M. Péoc'h, C. Rémy, and E. L. Barbier, "Imaging the microvessel caliber and density: Principles and applications of microvascular MRI," *Magnetic Resonance in Medicine*, vol. 73, no. 1, pp. 325–341, 2015.

[38] B. Lemasson, T. Christen, R. Serduc, C. Maisin, A. Bouchet, G. Le Duc, C. Rémy, and E. L. Barbier, "Evaluation of the relationship between MR estimates of blood oxygen saturation and hypoxia: Effect of an antiangiogenic treatment on a gliosarcoma model," *Radiology*, vol. 265, no. 3, pp. 743–752, 2012.

[39] A. Arnaud, F. Forbes, N. Coquery, N. Collomb, B. Lemasson, and E. L. Barbier, "Fully Automatic Lesion Localization and Characterization: Application to Brain Tumors Using Multiparametric Quantitative MRI Data," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1678–1689, 2018.

[40] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[41] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[42] E. Hoppe, F. Thamm, G. Körzdörfer, C. Syben, F. Schirrmacher, M. Nittka, J. Pfeuffer, H. Meyer, and A. Maier, "Magnetic resonance fingerprinting reconstruction using recurrent neural networks," in *arXiv*. IEEE, 2019, pp. 1537–1540.