

A sparse EEG-informed fMRI model for hybrid EEG-fMRI neurofeedback prediction

Claire Cury, Pierre Maurel, Rémi Gribonval, Christian Barillot

▶ **To cite this version:**

Claire Cury, Pierre Maurel, Rémi Gribonval, Christian Barillot. A sparse EEG-informed fMRI model for hybrid EEG-fMRI neurofeedback prediction. Frontiers in Neuroscience, 2020, 13, 10.3389/fnins.2019.01451 . inserm-02090676v3

HAL Id: inserm-02090676

https://inserm.hal.science/inserm-02090676v3

Submitted on 1 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A sparse EEG-informed fMRI model for hybrid EEG-fMRI neurofeedback prediction

Claire Cury [1,2,*], **Pierre Maurel** [1], **Rémi Gribonval** [2] **and Christian Barillot** [1]

[1] *Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074, Empenn team ERL U 1228, F-35000 Rennes, France.*
[2] *Univ Rennes, CNRS, Inria, IRISA UMR 6074, PANAMA team, F-35000 Rennes, France.*

Correspondence*:
Claire Cury
claire.cury.pro@gmail.com

ABSTRACT

Measures of brain activity through functional magnetic resonance imaging (fMRI) or Electroencephalography (EEG), two complementary modalities, are ground solutions in the context of neuro-feedback (NF) mechanisms for brain rehabilitation protocols. While NF-EEG (real-time neurofeedback scores computed from EEG signals) have been explored for a very long time, NF-fMRI (real-time neurofeedback scores computed from fMRI signals) appeared more recently and provides more robust results and more specific brain training. Using simultaneously fMRI and EEG for bi-modal neurofeedback sessions (NF-EEG-fMRI, real-time neurofeedback scores computed from fMRI and EEG) is very promising to devise brain rehabilitation protocols. However, fMRI is cumbersome and more exhausting for patients. The original contribution of this paper concerns the prediction of bi-modal NF scores from EEG recordings only, using a training phase where EEG signals as well as the NF-EEG and NF-fMRI scores are available. We propose a sparse regression model able to exploit EEG only to predict NF-fMRI or NF-EEG-fMRI in motor imagery tasks. We compared different NF-predictors steaming from the proposed model. We showed that predicting NF-fMRI scores from EEG signals adds information to NF-EEG scores and significantly improve the correlation with bi-modal NF sessions, compared to classical NF-EEG scores.

Keywords: Optimisation, EEG, sparsity, machine learning, Neurofeedback,EEG-fMRI, sparsity

1 INTRODUCTION

Neurofeedback approaches (NF) provide real-time feedback to a subject about its brain activity and help him or her perform a given task (Hammond, 2011; Sulzer et al., 2013). The estimation of neurofeedback information, is done through online brain functional feature extraction, to provide this real-time feedback to the subject. NF appears to be an interesting approach for clinical purposes, for example in the context of rehabilitation and psychiatric disorders (Sulzer et al., 2013; Birbaumer et al., 2009; Wang et al., 2017). Functional magnetic resonance imaging (fMRI) and electro-encephalography (EEG) are the most used noninvasive functional brain imaging modalities in neurofeedback. EEG measures the electrical activity of the brain through electrodes located on the scalp. EEG has an excellent temporal resolution (milliseconds), but a limited spatial resolution (centimeters) implying a lack of specificity. Furthermore, source localisation in EEG is a well-known ill-posed inverse problem (Grech et al., 2008).

On the other hand, blood oxygenation level dependent (BOLD) fMRI, measures a delayed hemodynamic response to neural activity with a good spatial resolution, and a temporal resolution of 1 or 2 seconds

depending on the sequence used. Therefore fMRI is more specific than EEG, making the fMRI an adequate modality for neurofeedback (NF-fMRI) (Thibault et al., 2018). However the use of the MRI scanner is costly, exhausting for patients since staying perfectly still when suffering is challenging and time consuming, hence NF-fMRI sessions cannot be repeated too many times for the same subject or patient.

During the past few years, the use of simultaneous EEG-fMRI recording has been used to understand the links between EEG and fMRI in different states of the brain activity and received recognition as a promising multi-modal measurement of the brain activity (Perronnet et al., 2018; Abreu et al., 2018). However this bi-modal acquisition is cumbersome for subjects or patients, due to the use of the fMRI scanner. The methodology to extract information from fMRI with EEG have been also intensively investigated (some methods involved in the process are reviewed here (Abreu et al., 2018)). Indeed, both modalities are sensitive to different aspect of brain activity, with different speeds. EEG provides in real time a direct measure of the changes in electrical potential occurring in the brain, while fMRI indirectly estimates brain activity by measuring changes in BOLD signal reflecting neuro-vascular activity, which occurs in general few seconds after a neural event (Friston et al., 1994; Logothetis et al., 2001). Several studies have investigated correlations between EEG signal and BOLD activity, in specific and simple tasks (de Munck et al., 2007; Goncalves et al., 2008; Engell et al., 2012; Magri et al., 2012; Scheeringa et al., 2011), and found different relationships between certain frequency bands on the EEG signal with the BOLD signal. All those studies reveal the existence of a link between EEG and fMRI, but this relationship highly varies with the task, the location in the brain and the considered frequency bands.

In the literature, the term EEG-informed fMRI refers to methods extracting features from EEG signals in order to derive a predictor of the associated BOLD signal in the region of interest under study. A recent review (Abreu et al., 2018) gives a good overview of the principal EEG-informed fMRI methods and their limitations. Different strategies have been investigated, depending on the type of activity under study (epilepsy, resting state, open/closed eyes, relaxation): either by selecting one channel on interest, either by using multiple channels, before extracting features of interest. For example in (Leite et al., 2013; Formaggio et al., 2011), authors used a temporal independent component analysis to select the best channel reflecting the epileptic seizures. In (Schwab et al., 2015), authors used a spatial, spectral and temporal decomposition of the EEG signals to map EEG on BOLD signal changes in the thalamus. From a more symmetrical way, in (Noorzadeh et al., 2017) it has been proposed a method for the estimation of brain source activation, improving its spatio-temporal resolution, compared to EEG or BOLD fMRI only. However, in the context of neurofeedback, using simultaneous recording of EEG-fMRI to estimate neurofeedback scores computed in real time from features of both modalities (NF-EEG-fMRI) is a

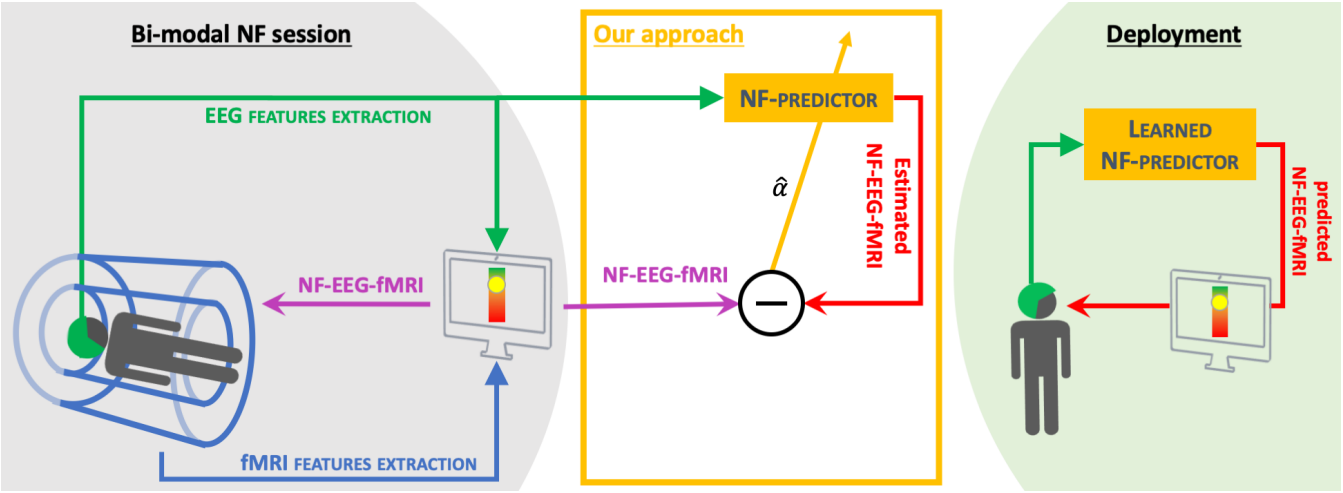


Figure 1. Objective : From bi-modal neurofeedback sessions (NF-EEG-fMRI) (see (Perronnet et al., 2017) or section 3 for more details), we propose a method to learn a NF-predictor. The final goal of this method is to propose NF sessions using EEG only, with the quality of bi-modal NF sessions. Therefore reducing the use of fMRI.

recent application that have been first introduced, and its feasibility demonstrated by (Zotev et al., 2014; Perronnet et al., 2017; Mano et al., 2017). The recent methodology synchronising both signals for real time neurofeedback (Mano et al., 2017), allows the setup of a new kind of data named NF-EEG-fMRI data, such as the dataset presented by (Perronnet et al., 2017), which we used for the present study. Furthermore, it has been shown in (Perronnet et al., 2017), that the quality of neurofeedback session is improved when using simultaneously both modalities, in NF-EEG-fMRI sessions. Thus, being able to reproduce in real time a NF-EEG-fMRI session when using EEG only, would reduce the use of fMRI in neurofeedback, while increasing the quality of NF-EEG sessions. To export fMRI information outside the scanner, most of the methods intend to predict the fMRI BOLD signal activity on a specific region of interest by learning from EEG signal recorded simultaneously, inside the fMRI scanner. Indeed, the method proposed in (Meir-Hasson et al., 2014), uses a ridge regression model with a ℓ_2 regularisation, based on a time/frequency/delay representation of the EEG signal from a single channel. Results show a good estimation of the BOLD signal in the region of interest, but the use of the fMRI neurofeedback in this study is only to serve the paradigm. The method aims at better targeting the amygdala in NF-EEG sessions.

Our challenge here, is to learn EEG activation patterns (see section 2.2) from hybrid (or bi-modal) NF-EEG-fMRI sessions (Perronnet et al., 2017), and improve the correlation with NF-EEG-fMRI of NF scores using EEG signal only. The motivation of this is multiple: since we are considering a new kind of data, we want to provide a simple method characterising NF-EEG-fMRI in EEG, leading to understandable model to confirm existing relations between EEG and fMRI in neurofeedback scores, or

to discover new relationships. Neurofeedback features in fMRI come from the BOLD activation in one or more region of interest. We propose an original alternative to source reconstruction in the context of neurofeedback by taking advantage of the recent and unique dataset of NF-EEG-fMRI we have access to. Indeed we directly intent to predict NF scores, without dealing with source reconstruction or spatial filtering to estimate BOLD-fMRI signal first on a specific region of interest, as proposed by a previous approach (Noorzadeh et al., 2017). To our knowledge, this problem of prediction of hybrid neurofeedback scores (without source reconstruction) is new, and has not yet been explored in the literature. Also we want the activation pattern to be applicable in real-time when using new EEG data. The main objective of this paper (Figure 1) is to design a method able to exploit EEG only, and predict an NF score of quality comparable to the NF score that could be achieved with a simultaneous NF-EEG-fMRI session. The approach is based on a machine learning mechanism. During a training phase, both EEG and fMRI are simultaneously acquired and used to compute and synchronise, in real time, NF-EEG and NF-fMRI scores, both being combined into an hybrid NF-EEG-fMRI score (Mano et al., 2017). EEG signals and NF scores are used to learn activation patterns. During the testing phase, the learned NF-predictor (also called activation pattern) is applied to unseen EEG data, providing simulated NF-EEG-fMRI scores in real time. Sparse regularisation is exploited to build a model called NF-predictor. The model used for the NF-predictor uses an adapted prior for brain activation patterns, using a mixed norm giving a structured sparsity, to spatially select electrodes and then select the most relevant frequency bands.

In section 2 we present the proposed model and the methods used to solve it. Then we will experiment our learning model on neurofeedback sessions with motor imagery task, which unique data are presented in section 3. Section 4 presents results on a dataset of 17 healthy subjects with 3 NF sessions of motor imagery each, one is used to learn the model, and the two others are used to test the model. Section 5 provides a discussion of the proposed framework.

2 PROBLEM AND METHOD

Considering that, during a learning phase, we have access to reference scores $y(t)$ and a temporal representation (potentially non-linear) of EEG signals $\boldsymbol{X}$ (called a design matrix, presented in section 2.1); the approach consists in choosing a vector of parameters $\boldsymbol{\alpha}$ such that $y(t) \approx q(\boldsymbol{X}(t), \boldsymbol{\alpha})$ for all t, where q is some parametric function. $\boldsymbol{\alpha}$ is a matrix matching the size of $\boldsymbol{X}(t)$, here we consider

$$q(\boldsymbol{X}(t), \boldsymbol{\alpha}) := \langle \boldsymbol{X}(t), \boldsymbol{\alpha} \rangle = \sum_{i=1}^{E} \sum_{j=1}^{F} \boldsymbol{X}_{i,j}(t) \boldsymbol{\alpha}_{i,j}. \tag{1}$$

To determine this parameter $\boldsymbol{\alpha}$, a regularisation is used to select an optimal parameter $\widehat{\boldsymbol{\alpha}}$ that fits the training data (see Figure 2 and section 2.3), while avoiding over-fitting, as detailed in Section 2.2.

Only a few brain regions are expected to be activated by a given cognitive task, therefore the electrodes configuration is said to be spatially sparse. However frequency bands of each electrodes are not necessarily sparse, and might even be smooth depending on the frequency band sampling.

From here, we will use the following notations :

- $y_{\mathrm{e}}(t) \in \mathbb{R}, \forall t \in \{1, ..., T\}$ are the T neurofeedback scores estimated from EEG signals (noted $S_{\mathrm{EEG}} \in \mathbb{R}^{E \times T_{\mathrm{EEG}}}$), measured from E electrodes during T_{EEG} samples of time.
- $y_{\mathrm{f}}(t) \in \mathbb{R}, \forall t \in \{1, ..., T\}$ are the T neurofeedback scores extracted from Blood Oxygen Level Dependent imaging (BOLD) signal of functional-MRI acquisitions $S_{\mathrm{fMRI}} \in \mathbb{R}^{V \times T_{\mathrm{fMRI}}}$, with V the number of voxels and T_{fMRI} the number of acquired volumes.
- $y_{\mathrm{c}}(t) = y_{\mathrm{e}}(t) + y_{\mathrm{f}}(t) \in \mathbb{R}, \forall t \in \{1, ..., T\}$ a combination of both NF scores (more details are provided in section 3).
- $y(t) \in \mathbb{R}, \forall t \in \{1, ..., T\}$ is a set of neurofeedback scores that can be y_{e}, y_{f} or y_{c}.

First, to build our predictor, relevant information from EEG data need to be extracted and organised to form what we call a design matrix.

2.1 Structured design matrices from EEG signal

The design matrix $\boldsymbol{X}_0 \in \mathbb{R}^{T \times E \times B}$, where E is the number of electrodes and B the number of frequency bands, contains relevant information extracted from the EEG signal. Each temporal matrix of $\boldsymbol{X}_0$, $\boldsymbol{X}_0(t) \in \mathbb{R}^{E \times B}\, \forall t \in \{1; \ldots; T\}$ is a frequency decomposition corresponding to the past 2 seconds of S_{EEG}. We used a Hamming time window of 2 seconds, to estimate the average *power* of each frequency band $b \in \{1; \ldots; B\}$ (defined below) on each channel $\in \{1; \ldots; E\}$. Each time window of EEG signal is overlapped by 1.75 seconds (0.25 seconds shift), to match with the 4Hz sample of the $\boldsymbol{y}$ values. The B frequency bands have an overlap of 1 Hz with the next band, and are defined between a minimum frequency b_{min} Hz and a maximum frequency b_{max} Hz (see section 3). We chose to use several relatively narrow frequency bands to let the model select the relevant bands for each electrodes. Furthermore it has been suggested (de Munck et al., 2009; Rosa et al., 2010) to use different frequency bands when working with coupling EEG-fMRI data.

The model also has to be able to predict y_{f} scores, derived from BOLD signal (see section 3). There is no linear relationship between BOLD signal and average power on frequency bands from EEG signal. Therefore, to better match y_{f} scores, we decided to apply a non-linear function to $\boldsymbol{X}_0$, used in fMRI

to model BOLD signals (Pedregosa et al., 2013; Lindquist et al., 2009), the canonical Hemodynamic Response function (HRF). We convolved $\boldsymbol{X}_0$ on its temporal dimension with the HRF, formed by 2 gamma functions, for a given delay of the first gamma function to compensate the response time of BOLD signal, as suggested in (Meir-Hasson et al., 2014; Moosmann et al., 2008). The HRF will temporally smooth and give a BOLD-like shape to the design matrix and increase a potential linear relationship between y_{f} and design matrix. Since HRF is known to vary considerably across brain regions and subjects (Handwerker et al., 2004), it is therefore recommended to consider different delays, but also to chose a range of values corresponding to the task asked. For the type of task addressed in the experimental part, the observed delay is around 4 seconds, therefore we convolved X0 with 3 different HRFs leading to 3 new design matrices $\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$ with respectively peak locations of 3, 4 and 5 seconds. We started from the canonical HRF and changed the peak parameter (respectively using 3 4 and 5) to induce the 3 different HRF used to induce time delay to the initial design matrix $\boldsymbol{X}_0$. By doing so, we keep the total of design matrices of a reasonable size

Those design matrices are concatenated in their 2nd dimension to form the $\boldsymbol{X}_c \in \mathbb{R}^{T\times M\times B}$ matrix, with $M = 4 * E$. Therefore, for each time t, $\boldsymbol{X}_c(t) = [\boldsymbol{X}_0(t); \boldsymbol{X}_3(t); \boldsymbol{X}_4(t); \boldsymbol{X}_5(t)]$. We also denote $\boldsymbol{X}_d(t) = [\boldsymbol{X}_3(t); \boldsymbol{X}_4(t); \boldsymbol{X}_5(t)]$ the design matrix of the different delays.

2.2 Optimisation

EEG data are now represented into a structured design matrix $\boldsymbol{X}_c$, we can search for a weight matrix $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{M\times B}$, such that $\sum_{m,h} \hat{\boldsymbol{\alpha}}(m,h)\boldsymbol{X}_c(t,m,h)$ estimates as well as possible the NF score $y(t), \forall t \in \{1;\ldots;T\}$. Note: the methodology is presented for the design matrix $\boldsymbol{X}_c$, but can be used for $\boldsymbol{X}_0$ or $\boldsymbol{X}_d$.

To identify the $\hat{\boldsymbol{\alpha}}$, called activation pattern, we propose the following strategy, which consists in learning, for a given subject and a NF session, the optimal $\hat{\boldsymbol{\alpha}}$ by solving the following problem:

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{t=1}^{T} \frac{1}{2}\big(y(t) - q(\boldsymbol{X}_c(t), \boldsymbol{\alpha})\big)^2 + \phi_\lambda(\boldsymbol{\alpha}) \tag{2}$$

with ϕ_λ a regularisation term, λ a weighting parameter for the regularisation term. This $\hat{\boldsymbol{\alpha}}$ is then applied to a design matrix $\boldsymbol{X}_c^{test}$ from a new session, to predict its NF scores

$$\tilde{y}_{\widehat{\boldsymbol{\alpha}}}(t) = q(\boldsymbol{X_c}^{test}(t), \widehat{\boldsymbol{\alpha}}) \; \forall t \in \{1;\ldots;T\} \tag{3}$$

Equation (2) is of the form $\operatorname{argmin}(g_1(\boldsymbol{\alpha}) + g_2(\boldsymbol{\alpha}))$ and its resolution can be done using the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) (Beck and Teboulle, 2009), which is a two-step approach of the Forward-Backward algorithm (Combettes and Wajs, 2005) making it faster. FISTA requires the same

conditions as the Forward-Backward algorithm, meaning a convex differentiable with Lipschitz gradient term g_1 and a convex term g_2 that is not necessarily differentiable but smooth enough to make its proximal map computable.

Here $g_1(\boldsymbol{\alpha}) = \sum_t^T \frac{1}{2}(y(t) - q(\boldsymbol{X}_c(t), \boldsymbol{\alpha}))^2$ is a sum of convex and differentiable functions with

$$\bigtriangledown g_1(\boldsymbol{\alpha}) = \sum_t -\boldsymbol{X}_c(t)(y(t) - q(\boldsymbol{X}_c(t), \boldsymbol{\alpha})) \tag{4}$$

since $\forall i \in \{1; \ldots; M\}, j \in \{1; \ldots; B\}, [\frac{\partial q(\boldsymbol{X}_c(t), \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}(i,j)}]_{i,j} = \boldsymbol{X}_c(t, i, j)$. By representing $\boldsymbol{X}_c(t)$ and $\boldsymbol{\alpha}$ as vectors of size $M * B$, we can easily note that $\frac{\partial g_1}{\partial \boldsymbol{\alpha}}$ is a sum of Lipschitz functions. Therefore, the Lipschitz constant of $\frac{\partial g_1}{\partial \boldsymbol{\alpha}}$ is $L = \|X_V^\intercal X_V\|$ with $X_V \in \mathbb{R}^{T*M*B}$ the vectorised version of $\boldsymbol{X}_c$.

The NF-predictor uses structured design matrix to have a better control on the interpretation of results and to better optimise the weights $\hat{\boldsymbol{\alpha}}$. Therefore we have to adopt an optimisation strategy coherent with this structure. The activation pattern of the NF-predictor:

1. has to be spatially sparse to regulate the model as EEG signals are noisy and to select the most relevant electrodes on each frequency bands,
2. has to be smooth across different overlapped frequency bands,
3. has to allow non-relevant frequency bands to be null.

The term g_2 is the prior term. Here, for $g_2(\boldsymbol{\alpha}) = \phi_\lambda(\boldsymbol{\alpha})$, we chose to use a ℓ_{21} mixed norm (Ou et al., 2009) followed by a ℓ_1-norm (noted ℓ_{21+1}-norm in (Gramfort et al., 2011)) to fit all structure conditions mentioned above.

$$\phi_\lambda(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_{21} + \rho\|\boldsymbol{\alpha}\|_1 \tag{5}$$

with $\rho \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}^+$. We chose not to estimate the parameter ρ, to keep computation time reasonable. Indeed, ρ weights the induced spatial sparsity over EEG channels, and we chose to fix this parameter for all subjects, as we hypothesis that there is no reason, for the number of electrodes involved in the activation pattern, to significantly change between subjects. However the estimation of λ parameter is needed (since we do not have hypothesis on its behaviour) and presented in the next section. The ℓ_{21} mixed norm that writes $\|\boldsymbol{\alpha}\|_{21} = \sum_m \sqrt{\sum_b \alpha_{m,b}^2}$ satisfies conditions 1) and 2). The ℓ_1 norm defined as $\|\boldsymbol{\alpha}\|_1 = \sum_{m,b} |\alpha_{m,b}|$ satisfies condition 3) since ℓ_p norms with $p \leq 1$ are known to promote sparsity. The last key point of FISTA algorithm is the proximal map associated to the ℓ_{21+1} norm $\text{Prox}_{\ell_{21+1}} : \mathbb{R}^{M \times B} \rightarrow$

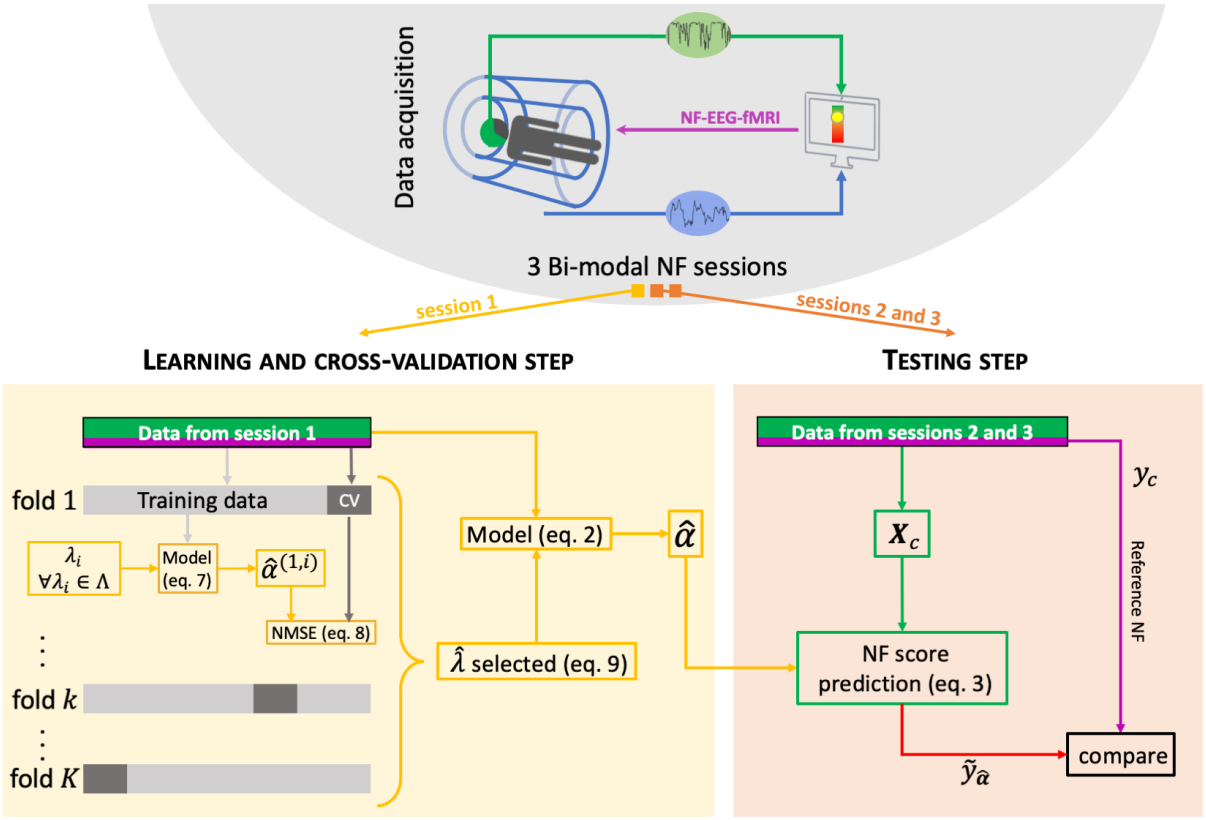


Figure 2. Machine learning scheme. For each subject, a bimodal neurofeedback session (NF-EEG-fMRI session 1 here) is used for the learning step, then the learned activation pattern $\hat{\alpha}$ is apply to the other sessions (2 and 3) for the testing step. The learning data are split K times into a training set (90% of the learning set) and a cross-validation (CV) set (10% of the learning set). The optimal $\hat{\lambda}$ parameter is the one minimising the variance and the bias in the learning step.

$\mathbb{R}^{M\times B}, \boldsymbol{\beta} \mapsto \text{argmin}_{\boldsymbol{\alpha}}(\phi_\lambda(\boldsymbol{\alpha}) + 1/2\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|^2)$, defined as

$$(\text{Prox}_{\ell_{21+1}}(Y))_{m,b} = \frac{Y_{m,b}}{|Y_{m,b}|}(|Y_{m,b}| - \rho)^+(1 - \frac{\lambda}{\sqrt{\sum_b(|Y_{m,b}| - \rho)^{+2}}})^+ \tag{6}$$

with operator $(.)^+ = max(., 0)$. One can note that by cancelling either the λ parameter or the ρ parameter, we retrieve the proximal map associated to the ℓ_{21} (when $\rho = 0$) and to the ℓ_1 norm (when $\lambda = 0$), which demonstrations can be found in the appendix of (Gramfort et al., 2012). For the stopping criterion of FISTA, a large enough number of iteration has been used, allowing the model to converge before reaching the last iteration. All elements and conditions are gathered to run the FISTA algorithm.

2.3 λ parameter selection

The parameter λ is important in the optimisation problem and we decided to estimate it automatically. The following process chooses the best λ among a list of $\Lambda = \{\lambda_1; ...; \lambda_l\}$ sorted in increasing order. First of all, the data must be split into 2 sets. In our case subjects have 3 NF-EEG-fMRI sessions : one session

is used as the learning set, and another NF-EEG-fMRI session is used as the testing set (see Figure 2). For each value λ_i of Λ, the learning set, formed by T neurofeedback scores with their associated design matrices, is divided $K = 50$ times into a training set of indices R_k, representing 90% of the T data, and a cross-validation set CV_k composed by composed by the remaining 10% of the learning set. A model $\hat{\boldsymbol{\alpha}}^{(k,i)}$ is estimated on the training dataset k composed by R_k neurofeedback scores $y(j)$ and the associated design matrices $\boldsymbol{X}_c(j)$ with λ_i, i.e.:

$$\hat{\boldsymbol{\alpha}}^{(k,i)} = \arg\min \sum_j \| y(j) - q(\boldsymbol{X}_c(j), \hat{\boldsymbol{\alpha}}^{(k,i)})\|^2 + \lambda_i \|\boldsymbol{\alpha}^{(k,i)}\|_{21} + \rho \|\boldsymbol{\alpha}^{(k,i)}\|_1 \tag{7}$$

with $j \in R_k$ and $\boldsymbol{X}_c(j) \in \mathbb{R}^{M \times B}$. For the current λ_i evaluation, we stop the process when $\sum_k |\hat{\boldsymbol{\alpha}}^{(k,i)}|_0 / K < 2$. There is no need to investigate the next λ_i, the current one is sparse enough, and the next one might lead to null models.

We then apply the model $\hat{\boldsymbol{\alpha}}^{(k,i)}$ to the corresponding cross-validation set of CV_k NF scores, to obtain estimated values of $y(s)$, $\tilde{y}(s) = q(\boldsymbol{X}_c(s), \hat{\boldsymbol{\alpha}}^{(k,i)})$, with $s \in CV_k$. For each one of the 50 partitioning into training and cross-validation sets, we computed the normalised mean squared error NMSE for a given set of data $\{y, \boldsymbol{X}_c\}$, for the training sets and the cross-validation sets.

$$\text{NMSE}(\{y, \boldsymbol{X}_c\}, \hat{\boldsymbol{\alpha}}^{(k,i)}) = \frac{\sum_s (y(s) - q(\boldsymbol{X}_c(s), \hat{\boldsymbol{\alpha}}^{(k,i)}))^2}{\sum_s (y(s) - \bar{y})^2} \tag{8}$$

with $\bar{y} = 1/n \sum_s^n y(s)$. The optimal $\hat{\lambda}$ parameter is defined as the one minimising the error during training and cross-validation. Considering only the errors from the training set $\text{NMSE}(\{y(R_k), \boldsymbol{X}_c(R_k)\}, \hat{\boldsymbol{\alpha}}^{(k,i)})$ would introduce bias, and considering only the error of the cross-validation set $\text{NMSE}(\{y(CV_k), \boldsymbol{X}_c(CV_k)\}, \hat{\boldsymbol{\alpha}}^{(k,i)})$, would introduce variance. Then the optimal $\hat{\lambda}$ is the λ_i that minimises :

$$\sum_{k=1}^{K} [\text{NMSE}(\{y(R_k), \boldsymbol{X}_c(R_k)\}, \hat{\boldsymbol{\alpha}}^{(k,i)}) + \text{NMSE}(\{y(CV_k), \boldsymbol{X}_c(CV_k)\}, \hat{\boldsymbol{\alpha}}^{(k,i)})] \tag{9}$$

$\hat{\lambda}$ parameter is the optimal parameter used for the model estimation. If there are several candidates, to favor sparsity, the larger of these candidate is chosen.

3 DATA ACQUISITION AND PRE-PROCESSING

We used an existing dataset presented in (Perronnet et al., 2018), composed of 17 healthy subjects that were scanned using the hybrid Neurofeedback platform from Neurinfo (Rennes, France) coupling EEG

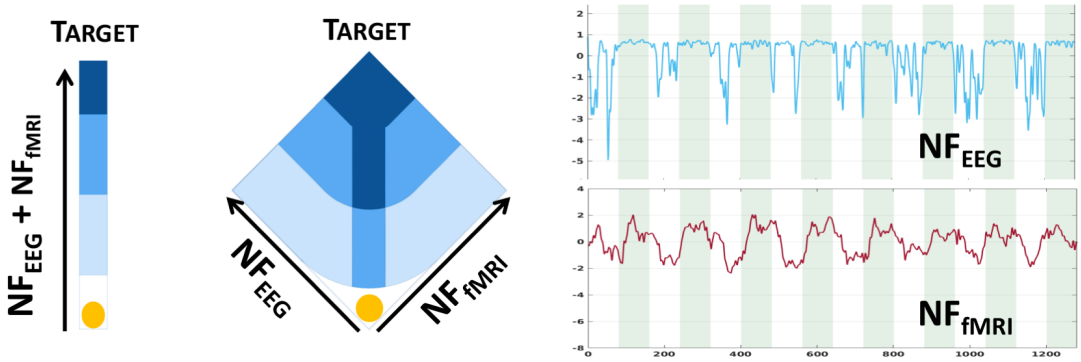


Figure 3. Bi-modal neurofeedback metaphors (1D on the left, 2D on the middle), displayed during sessions (Perronnet et al., 2018). 1D: the ball's position represents the sum of y_e and y_f. 2D: the left axis represents the y_e and the right axis represents the y_f scores. The 2 plots on the right show NF scores from EEG and from fMRI, green areas are task, white areas are rest. The goal is to bring the ball in the dark blue area.

and fMRI signal (Mano et al., 2017). Data are now available online in BIDs format on openneuro : `https://openneuro.org/datasets/ds002338` (Lioi et al., 2019). A 64-channel MR-compatible EEG solution from Brain Products (Brain Products GmbH, Gilching, Germany) has been used, the signal was sampled at 5kHz, FCz is the reference electrode and AFz the ground electrode. For the fMRI scanner, we used a 3T Verio Siemens scanner with a 12 channels head coil (repetition time (TR) / echo time (TE) = 2000/23ms, FOV = 210×210mm^2, voxel size = $2 \times 2 \times 4$mm^3, matrix size =105×105 with 16 slices, flip angle = 90°). All subjects are healthy volunteers, right-handed and had never done any neurofeedback experiment before. They all gave written informed consent in accordance with the Declaration of Helsinki, as specified in the study presenting the data (Perronnet et al., 2017). They all had 3 NF motor imagery sessions of 320 seconds each, after a session dedicated to the calibration. One session consists in 8 blocks alternating between 20 seconds of rest, eyes open, and 20 seconds of motor imagery of their right hand. The neurofeedback display was uni-dimensional (1D) for 9 subjects (Figure 3 left), and bi-dimensional (2D) for 8 subjects (Figure 3 middle). For both, the goal was to bring the ball into the dark blue area (Perronnet et al., 2018).

NF scores y_e and y_f being from different modalities, were standardised before summing to form y_c. In this study NF scores refer to standardised scores, except when they are predicted. For this study, y_e have been computed from the commonly used, in neurofeedback, the Laplacian operator, centred around the region of interest, channel C3 here. For each time interval I_t the spatial filtering is noted Lap(C3, I_t). The temporal segments I_t are spaced by 250ms, and a length of 2 seconds (therefore an overlapping of 1,75 seconds), as for the design matrix construction. The power of the frequency band [8Hz - 30Hz] is then extracted via the Power Spectral Density function PSD:

$$y_e(t) = -PSD_{[8-30]}(\text{Lap(C3},I_t))$$

One may note the presence of the minus operator used here, for the sake of coherence with y_{f} (Figure 3 right).

The neurofeedback scores y_{f} have been computed from the maximal intensity of BOLD signal covering the right-hand motor area and the supplementary motor area, one score is computed per volume acquired (i.e 1 per second). Then scores y_{f} are re-sampled and smoothed (using a Savitzky-Golay filter, known to avoid signal distortion) to fit the 4Hz y_{e} scores ($T = 1280$).

An active set have been selected on design matrices to avoid potentially correlated noise, due to head movement during resting blocks, obstructing signal from channels of interest. Indeed in coupling EEG-fMRI acquisitions, subjects are lying into the MRI scanner, therefore outer electrodes can be in contact with the bed or holds. We excluded outer electrodes and kept 28 electrodes, the 3 central lines have 7 electrodes (FCz is the reference), 3 frontal electrodes and 3 posterior electrodes.

Potential outliers in the design matrices (i.e. observations $>$ mean$\pm$3std) were thresholded in the NF-EEG-fMRI session used as learning set, and bad observations from annotations on the EEG signal were removed as their corresponding NF scores. For frequency dimension of the design matrix $\boldsymbol{X}_0$ construction (cf section 2), and therefore for the other design matrices, we chose $b_{min} = 8$, $b_{max} = 30$ to cover alpha and beta frequency bands involved in motor tasks. We considered $B = 10$ frequency bands, leading to bands of 3 Hz wide with an overlap of 1Hz.

As mentioned in the previous section, for the regularisation of the NF-predictor, we used 15 values of λ_i from 100 to 3000 and fixed the parameter $\rho = 1500$.

4 EXPERIMENTS AND RESULTS

As said in the previous section, we have 3 neurofeedback sessions per subjects. For each subjects, we will consider 1 session as learning set, and the 2 others as testing sets. Leading to 3 different learning sets, and 6 different testing sets per subjects.

4.1 Experiments and validation

We tested different NF-predictors for the prediction of different NF scores, $\forall t \in \{1; \ldots; T\}$:

1. $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{c}}}(t) = q(\boldsymbol{X}_c^{test}(t), \hat{\boldsymbol{\alpha}}_c)$ with $\hat{\boldsymbol{\alpha}}_c$ (eq. 2), learned from $\boldsymbol{X}_c$ and y_{c}
2. $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{f}}}(t) = q(\boldsymbol{X}_d^{test}(t), \hat{\boldsymbol{\alpha}}_f)$ with $\boldsymbol{X}_d = [\boldsymbol{X}_3; \boldsymbol{X}_4; \boldsymbol{X}_5]$, and $\hat{\boldsymbol{\alpha}}_f$ (eq. 2), learned from $\boldsymbol{X}_d$ and y_{f}
3. $y_{\mathrm{e}}(t) + \tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{f}}}(t)$ using NF-predictor 2

The advantage of the last NF-predictor is the possibility of using the 2D score visualisation (Figure 3) to display NF scores. Also in this case, only NF-fMRI scores has to be learned. We run the following experiments to test our different NF-predictors:

- Model validation: we used the learning set to assess if the different NF-predictors could model accurately the NF scores. For each subject and each NF-predictor, we estimated correlations with the reference NF score to quantify the quality of prediction. As a reference, we compared to the correlation of y_e to y_c, which is part of the y_c score, to assess if, in the validation process, the model can better predict y_c scores than y_e.
- Model prediction: we apply the learned activation patterns to a new NF-EEG-fMRI recorded session (testing set, Figure 2). For each subject and each pair of session (6 learning/testing pair per subject), we compared correlations between NF-predictors $\tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{c}}$ and $y_\mathrm{e} + \tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{f}}$ to reference score y_c, and between the prediction of $\tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{f}}$ to y_f.
- To observe the captured structure in the activation pattern $\hat{\boldsymbol{\alpha}}_c$ and $\hat{\boldsymbol{\alpha}}_f$ respectively learned to predict y_c and y_f, we re-shaped the averaged activation patterns of a subject over sessions, into the matrices corresponding to the design matrices defining $\boldsymbol{X}_c$ respectively $\boldsymbol{X}_d$ (see section 2.1), and displayed results of the first dimension (electrodes) and of the second dimension (frequency bands).

4.2 Results

4.2.1 Model and prediction

The model validation of the NF-predictors (i.e. the learned activation patterns $\hat{\boldsymbol{\alpha}}$ are applied to the learning set) results are shown at Figure 4 left side. Pearson's correlation coefficients between the prediction and the ground truth are computed for each of the 3 learning sessions and for all subjects. Correlations are very high (> 0.8) for the 2 NF-predictors of y_c scores, letting think that the model is adapted to the problem of NF prediction. Indeed, the model could fit NF-fMRI scores using only EEG signals information with a median correlation of r = 0.81 with y_f.

For the evaluation of the model prediction, the learned activation patterns are applied to the testing sets i.e. the 2 other unseen NF sessions, for each one of the 3 NF sessions. The significance of the differences between the correlations to the reference NF scores is assessed by applying a Fisher transformation to the correlation coefficients. The p-values obtained with Student's t-test, after this transformation, indicate the significance of the differences between the correlation coefficients distributions. Table 1, shows that $y_\mathrm{e} + \tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{f}}$, whose median correlation is r = 0.74, better predicts y_c than y_e only, despite the fact that y_e is part of y_c. The paired t-test, gives a p-value $p = 6.6\text{e-}4$ ($t = 3.52$), meaning that the prediction of NF-fMRI scores by $\tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{f}}$ significantly adds information to y_e. Predicting y_c scores with $y_\mathrm{e} + \tilde{y}_{\hat{\boldsymbol{\alpha}}_\mathrm{f}}$ is also

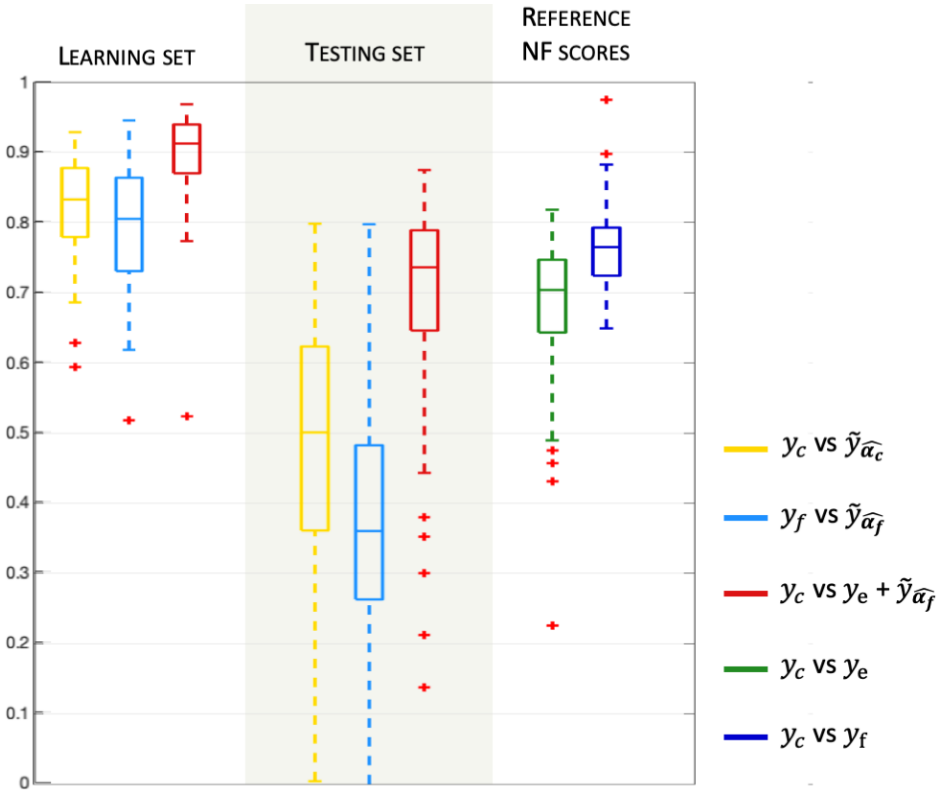


Figure 4. Model validation and prediction. Boxplots (median and quartiles) of Pearson's correlation coefficients over all subjects and sessions, between NF-predictors and y_c or y_f. The right part of the plot indicates the correlation of the reference NF scores y_e and y_f versus the corresponding reference bi-modal NF scores $y_c (= y_e + y_f)$.

Table 1. Model prediction : Pearson's correlation coefficients over all subjects and sessions, between NF-predictors and y_c. The table shows p-values of t-test, the hypothesis is : models in columns correlates better with y_c than models in rows.

	paired t-test		
Correlation (y_c,.)	$\tilde{y}_{\widehat{\alpha}_c}$	$y_e + \tilde{y}_{\widehat{\alpha}_f}$	y_e
$\tilde{y}_{\widehat{\alpha}_c}$	N.A.	$t = -23.17$ $p = 1$	$t = -12.76$ $p = 1$
$y_e + \tilde{y}_{\widehat{\alpha}_f}$	$t = 23.17$ $p =$ 2.e-42	N.A.	$t = 3.52$ $p =$ 6.6e-4
y_e	$t = 12.76$ $p =$ 9e-23	$t = -3.51$ $p = 1$	N.A.

better than $\tilde{y}_{\widehat{\alpha}_c}$ ($p =$ 2e-42, $t = 23.17$), which is expected since the models predicts NF-fMRI scores and directly use NF-EEG scores (with all the potential noise coming from the EEG measures) which are part of the reference score, as y_e is in $y_e + \tilde{y}_{\widehat{\alpha}_f}$, and therefore correlates well with y_c. Furthermore the proposed model is able to predict y_f scores with a fair correlation of 0.36 in median and 0.35 in average. Thus predicting NF-fMRI scores, instead of predicting bi-modal NF scores, seem to be the best way of predicting those bi-modal NF scores. Figure 5 gives examples of predictions of y_c with $y_e + \tilde{y}_{\widehat{\alpha}_f}$ and $\tilde{y}_{\widehat{\alpha}_c}$.

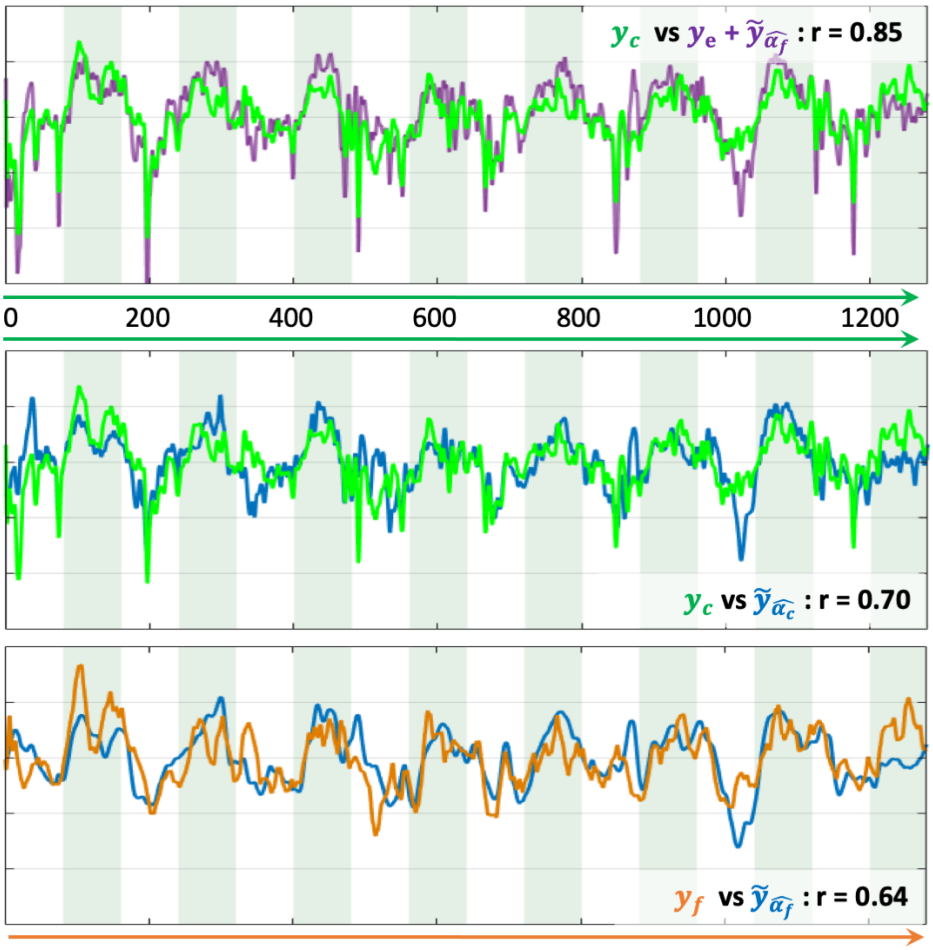


Figure 5. Examples of prediction of NF scores. The x-axis is the temporal axis in milliseconds. Vertical bands indicate the rest and the task blocks. The correlation coefficient r indicates the correlation between each pair of time-series NF scores. Top: prediction of y_c with $y_e + \tilde{y}_{\widehat{\alpha}_f}$. Middle : prediction of y_c with $\tilde{y}_{\widehat{\alpha}_c}$. Bottom : prediction of y_f with $\tilde{y}_{\widehat{\alpha}_f}$.

It also illustrates that, even if the correlations of $\tilde{y}_{\widehat{\alpha}_c}$ are lower than $y_e + \tilde{y}_{\widehat{\alpha}_f}$ (Table 1), $\tilde{y}_{\widehat{\alpha}_c}$ can predict correctly the reference score y_c.

During the learning session (i.e. the NF session used to learn the predictor), subjects might unevenly focus on their fMRI or EEG feedbacks, and when subjects focus more on NF-fMRI than on NF-EEG, the EEG-signals might lose coherence with respect to the NF-fMRI scores. Even though, the EEG signals could predict NF-fMRI scores with a correlation of 0.36 in median and mean 0.35 (cf Figure 4, which is a fair correlation between such different modalities. Examples of NF prediction are given at the bottom of Figure 5, the plot shows the prediction $\tilde{y}_{\widehat{\alpha}_f}$ of NF-fMRI scores on a testing set.

4.2.2 Activation patterns

We now focus on the learned models to evaluate the sparsity, over sessions and subjects and observe the dispersion of the learned patterns over sessions and frequency bands of a subject. For each subjects and

Table 2. Sparsity of the learned models $\hat{\boldsymbol{\alpha}}_c$ and $\hat{\boldsymbol{\alpha}}_f$. Proportion of zero coefficients in the learned models.

	$\hat{\boldsymbol{\alpha}}_0$	$\hat{\boldsymbol{\alpha}}_3$	$\hat{\boldsymbol{\alpha}}_4$	$\hat{\boldsymbol{\alpha}}_5$	all
$\hat{\boldsymbol{\alpha}}_f$	x	0.92 ± 0.03	0.96 ± 0.03	0.91 ± 0.03	0.93 ± 0.04
$\hat{\boldsymbol{\alpha}}_c$	0.87 ± 0.05	0.92 ± 0.03	0.96 ± 0.02	0.93 ± 0.03	0.92 ± 0.05

each learning session, we computed the proportion of zero coefficients of the activation patterns (Table 2). By construction of the design matrix $\boldsymbol{X}_c$, $\hat{\boldsymbol{\alpha}}_c$ and $\hat{\boldsymbol{\alpha}}_f$ can be split into different activation patterns, as shown in columns of Table 2. The models could select relevant coefficients of the design matrices to predict reference NF scores. In average, there is 87 non-zeros coefficient on $\hat{\boldsymbol{\alpha}}_c$ to predict y_{c}, and 57 non-zeros coefficients on $\hat{\boldsymbol{\alpha}}_f$ to predict y_{f}.

To display the activation patterns of a subject over its 3 sessions, we denote $\hat{\boldsymbol{\zeta}}_c = \sum_{s=1}^{3} |\hat{\boldsymbol{\alpha}}_c^{(s)}| \in \mathbb{R}^{M \times B}$ the absolute activation pattern of $\hat{\boldsymbol{\alpha}}_c$ (respectively $\hat{\boldsymbol{\zeta}}_d^f = \sum_{s=1}^{3} |\hat{\boldsymbol{\alpha}}_d^{(s)}| \in \mathbb{R}^{M \times B}$ for $\hat{\boldsymbol{\alpha}}_f$), and $\hat{\boldsymbol{\gamma}}_c = \sum_{s=1}^{3} \hat{\boldsymbol{\alpha}}_c^{(s)} \in \mathbb{R}^{M \times B}$ the average activation pattern (respectively $\hat{\boldsymbol{\gamma}}_d^f = \sum_{s=1}^{3} \hat{\boldsymbol{\alpha}}_d^{(s)} \in \mathbb{R}^{M \times B}$). We can display heat maps for each of the 4 absolute activation patterns at each electrodes $\sum_{b \in B} \widehat{\boldsymbol{\zeta}}_0, \sum_b \widehat{\boldsymbol{\zeta}}_3, \sum_b \widehat{\boldsymbol{\zeta}}_4$ and $\sum_b \widehat{\boldsymbol{\zeta}}_5$ (Figure 6, top line of each panels); and colour maps of the 4 average patterns $\sum_b \widehat{\boldsymbol{\gamma}}_0, \sum_b \widehat{\boldsymbol{\gamma}}_3$, $\sum_b \widehat{\boldsymbol{\gamma}}_4$ and $\sum_b \widehat{\boldsymbol{\gamma}}_5$ (Figure 6, bottom line of each panels) showing the sign of the strongest and most stable coefficients across all subjects and sessions. Activation patterns displayed at Figure 6, represent the dispersion of the learned parameters for one example subject over its 3 bi-modal NF sessions for which he received a bi-dimensional display. The subject used had a median correlation with the corresponding reference score of 0.44 for $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{f}}}$, of 0.76 for $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{c}}}$ and of 0.81 for $y_{\mathrm{e}} + \tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{f}}}$. All maps present different distributions of the non-zero coefficients. As expected when learning y_{c} scores, the most intense heat map $\sum_b \widehat{\boldsymbol{\zeta}}_0$ with a maximum value of 45, concentrates its non-zero values on C3 channel (above the right-hand motor area) and corresponds to the design matrix $\boldsymbol{X}_0$, directly extracted from EEG signal without non-linear temporal delay. The next heat map in intensity order is $\sum_b \widehat{\boldsymbol{\zeta}}_5$ with two peaks with a parietal positioning above Pz and P4 channels, corresponding to the Brodmann area 7, involved in the visuo-motor coordination. As presented in an interesting study (Sitaram et al., 2017), this brain area is active during generalised neurofeedback when feedback is presented visually. In (Sitaram et al., 2017) it is also mentioned that this part of the cortex is part of the executive control network connected to the thalamus. The activation of the executive control network during motor imagery task is relevant, and indicates that the subject is trying to do the task. It is also interesting to observe that the main activation peaks of $\sum_b \widehat{\boldsymbol{\zeta}}_5$ (Pz and P4) have opposite signs with the activation peak of $\sum_b \widehat{\boldsymbol{\zeta}}_0$ (above C3), suggesting a negative correlation between neural activation measured at C3 and the neural activation of posterior parietal Pz and P4 channels , part of the control network in neurofeedback. Panel B of Figure 6 presents comparable learned activation patterns to predict $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_{\mathrm{f}}}$, however more spread over sessions than for the model on panel A, even if the model as around 93% of null coefficients (Table 2). This suggests that, even for a same

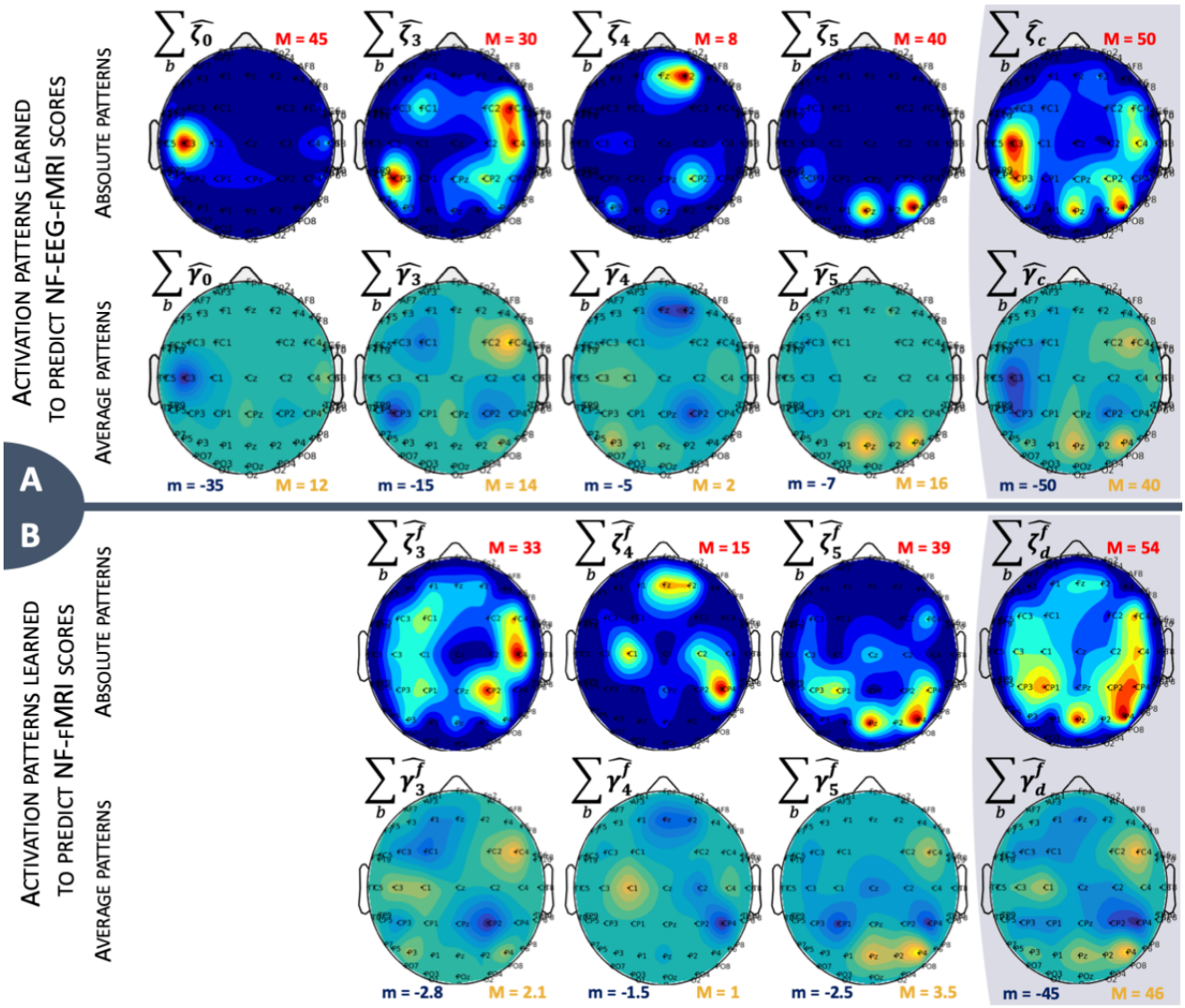


Figure 6. Activation patterns over sessions and frequency bands b for one example subject. There is two lines for each model, panel A represents activation patterns for $\hat{\alpha}_c$ and panel B represents activation patterns for $\hat{\alpha}_f$. The last column is the sum of the other ones. Top line: Heat map representing the distribution of non-zero coefficients. Maximum value is indicated for each map with the red letter M, dark blue areas represent zero values. Bottom line: Average activation patterns, representing the sign of the main non-zero values across subject and sessions. Minimum and maximum values are indicated for each map with the letters m and M, green areas represent zero values.

subject, the relation between EEG and fMRI changes over sessions, probably depending on the strategy used by the subject.

At last, it is also possible to display the frequency profile of each average activation patterns. The 4 frequency profiles are $\sum_{e\in\{1,\dots,E\}} \widehat{\gamma}_0$, $\sum_e \widehat{\gamma}_3$, $\sum_e \widehat{\gamma}_4$ and $\sum_e \widehat{\gamma}_5$, $e \in \{1, ..., E\}$, summing weights over electrodes. Figure 7 shows that the most used frequency bands over all sessions and electrodes are alpha band ([8-12] Hz) and lower beta band ([13-17] Hz), the last 4 frequency bands are not displayed as they only have null coefficients. We can observe the effect of ℓ_2 regularisation which allows continuity in the

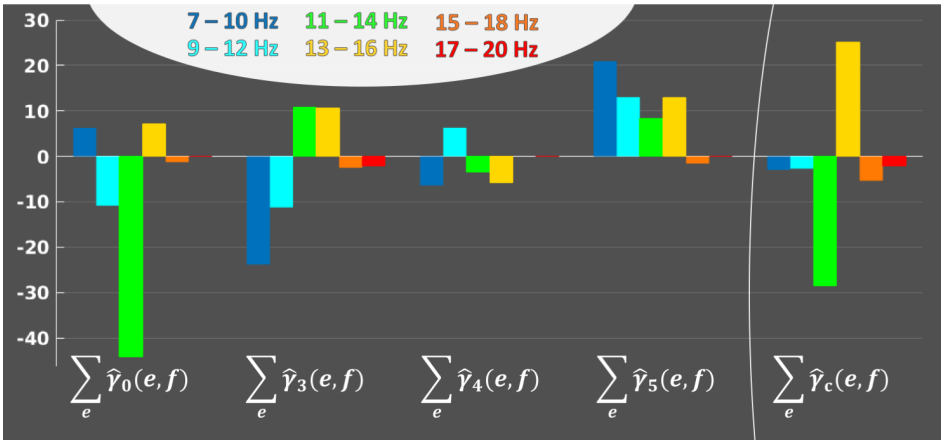


Figure 7. Frequency profiles, across subjects and sessions, of each average activation patterns, to represent the implication of each frequency bands in the activation patterns. $e \in \{1, ..., E\}$. For each average activation pattern, the y-axis indicates the sum of weights over electrodes, for each frequency bands $f \in \{1, ..., B\}$ on the horizontal axis.

frequency bands, and the effect of the subsequent ℓ_1 regularisation which removed the smaller coefficients located in the high frequencies. When considering all activation patterns (Figure 7 right side), there is a change of sign between alpha and lower beta. Each activation pattern shows a different frequency profile, which, together with Figure 6, tends to demonstrate that patterns have complementary information.

4.3 Ablation study

We run an ablation study to understand the impact of the non-linearity part of the model, obtained by the convolution of different HRFs inducing different delays for the prediction of NF-fMRI scores, by analysing results of the model without HRFs convolutions. Therefore, in this ablation study we used the exact same processes as described in section 4 (with a learning step and a testing step), except that we used the design matrix $\boldsymbol{X}_0$ alone to learn the model parameters to predict y_c scores and y_f scores.

Results of the ablation study are displayed at Figure 8 and are to be compared with results of the proposed model presented at Figure 4. Figure 8 shows that during learning step, the prediction of $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_\text{f}}$, using X_0 only, correlates with y_f with a correlation of only 0.48 in mean and median. This prediction is significantly improved by the use of the different HRFs functions (paired t-test, $t = -14,64$, $p = 1e$-26), as in the proposed model $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_\text{f}}$ correlates with y_f with a correlation of 0.81 in median and mean. Therefore, it is natural to observe that the prediction of the bi-modal NF score by $y_\text{e} + \tilde{y}_{\widehat{\boldsymbol{\alpha}}_\text{f}}$ is also significantly improved by the use of non-linearity (paired t-test, $t = -11,87$, $p = 4e$-21). Figure 8 shows that for the testing step, the prediction of $\tilde{y}_{\widehat{\boldsymbol{\alpha}}_\text{f}}$, using X_0 only, correlates with y_f with a correlation of 0.14 in mean and 0.15 in median, which is significantly lower than the proposed model (paired t-test, $t = -9.56$, $p = 2e$-18). The prediction of the bi-modal NF scores by $y_\text{e} + \tilde{y}_{\widehat{\boldsymbol{\alpha}}_\text{f}}$ is also significantly lower than the proposed model

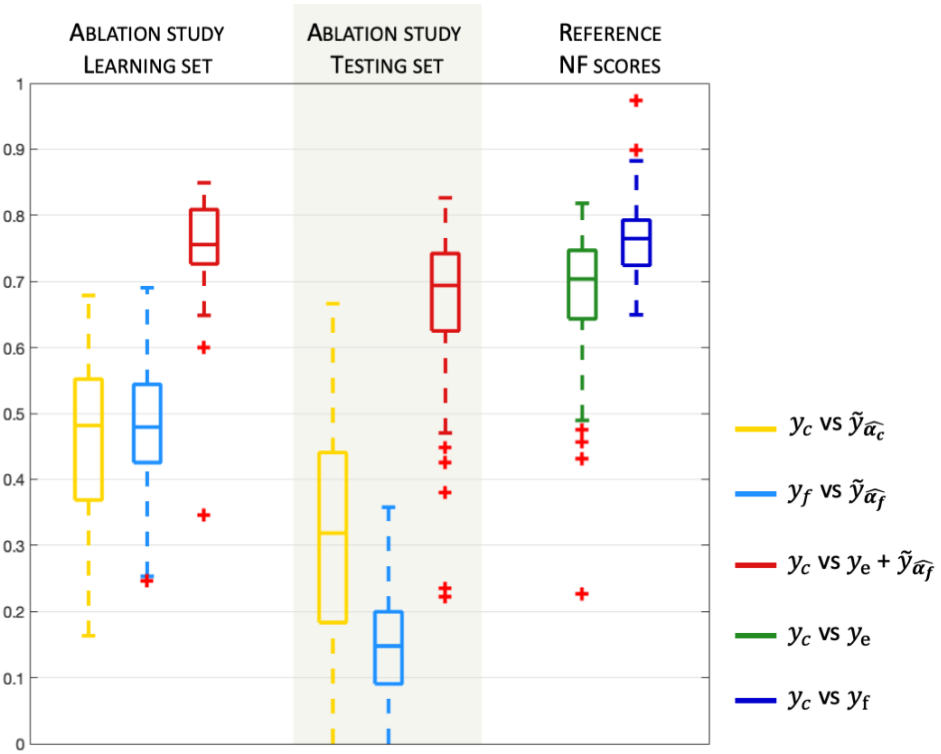


Figure 8. Ablation study. Removing non-linearity, keeping X_0 only in the model to understand the importance of the non-linearity induced by the use of different HRF functions.

(paired t-test, $t = -2.74$, $p = 3e\text{-}3$). Also, when using X_0 only, the prediction of the bi-modal NF scores using $y_e + \tilde{y}_{\widehat{\alpha}_f}$ do not significantly improve over y_e alone (paired t-test, $t = -0.69$, $p = 0.5$).

5 DISCUSSION AND CONCLUSION

The model validation supports that the optimisation strategy we chose for our problem is adapted to the model, as is the choice of the different design matrices. The ablation study supports the use of different non-linear delays to improve the prediction of NF-fMRI scores using EEG signal. The evaluation of the model prediction strongly suggests that predicting only NF-fMRI scores from EEG signal while applying a Laplacian on EEG signal appears to be the best solution. Indeed, for $\tilde{y}_{\widehat{\alpha}_c}$, when predicting the bi-modal NF scores, the variability between EEG signals induces decreasing correlation with the reference NF score y_c. Also NF-EEG scores can always be computed from available EEG signals, which leave only NF-fMRI scores to estimate from EEG signals. However, one might want to improve the selection of features for the computation of NF-EEG scores, but this raises different questions about the validation and the reference score. Here we assumed that the given NF scores are relevant to the task and good enough to be considered as reference scores. As expected, for the prediction of the bi-modal NF scores by $\tilde{y}_{\widehat{\alpha}_c}$, when decomposing the activation pattern $\hat{\boldsymbol{\alpha}}_{\boldsymbol{c}}$ (from $\tilde{y}_{\widehat{\alpha}_c}$), into the 4 matrices corresponding to the different design matrices, the weights corresponding to $\boldsymbol{X}_0$ (0 second of delay) are mainly located above C3, which is the centre of

the Laplacian used for the computation of the NF-EEG scores. This seem to support the fact that with a delay of 0 seconds, only the part of the NF coming from the EEG brings information.

A possibility to increase the prediction of NF-fMRI scores, would be to use more NF sessions as learning sessions, since as observed, EEG signals bring variability in the prediction. Each new bi-modal neurofeedback session could be added to the subject-specific model, to better adapt the model to the subject or patient. This will be investigated in a next study. Given the improved correlation of the proposed NF predictor with bimodal NF scores, it would be interesting in future work to validate its improved performance in actual NF sessions, compared to classical NF-EEG scores. In particular to assess the response of subjects to the predicted bi-modal NF scores and in particular the predicted NF-fMRI scores learned by the proposed model over a standard NF-EEG neurofeedback session, a new and large enough study is needed, as subjects can learn at different pace to regulate their own brain activity.

Presently, the proposed model learns an individual or specific model for each subject that allows a personalised model for adapted neurofeedback sessions. Also a change in strategy for the task (here no specific strategy to imagine moving their right hand was given to the subjects) might impact the learned model, as the relation between EEG and fMRI signals can change. However, in a future work, we are investigating an adaptation of the methodology for the extraction of a common model, taking into account the differences between sessions and subjects, allowing the prediction of NF-EEG-fMRI scores on new subjects who did not participated to the model construction. The model might be less specific, but this would give access to neurofeedback sessions of bi-modal quality using EEG only, for subjects with MRI contraindications, and/or drive a subject-specific model estimation, respecting the strategy used by the subject to progress in its neurofeedback task. The learned NF predictor is of course specific to the particular task considered during the learning session. One can expect to generalise the approach to other tasks where spatial sparsity is relevant. Its extension beyond such tasks is more challenging.

Other ways to improve the method proposed here would be to investigate the use of dynamic functional connectivity, a relatively recent field in BOLD fMRI which needs further investigations to be used along with EEG data (Tagliazucchi and Laufs, 2015). Dynamic functional connectivity would allow to take into account the different network (reward, control and learning) involved during neurofeedback sessions (Sitaram et al., 2017). Dynamic functional connectivity study the temporal fluctuations of the BOLD signal across the brain, and appears to be a promising approach in the EEG-fMRI research field (Allen et al., 2018). However, one should be careful with the potential unknown remaining noise coming from the MRI during EEG-fMRI simultaneous recording, that might unsettle the EEG signal coherence between electrodes.

The long-term objective of our project is to learn from EEG-fMRI NF sessions to provide, outside the MRI scanner, enhanced NF-EEG sessions (Figure 1). A future work will investigate the portability of the learned model (on EEG-fMRI neurofeedback data), outside the MRI scanner, bringing new challenges as dealing with the remaining noises in the MRI after artefact correction, and the absence of ground truth once the EEG is measured outside the MRI scanner.

To conclude, the proposed model here is able to provide a good enough prediction of the NF-fMRI scores, to overcome the absence of NF-fMRI and allows to significantly improve the estimation of NF-EEG-fMRI scores when using EEG only.

6 AUTHOR CONTRIBUTIONS

CC guarantor of integrity of entire study. Study concepts and design, data analysis and interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors.

7 CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 ACKNOWLEDGEMENTS

This manuscript has been released as a Pre-Print at (Cury et al., 2019).

Data acquisition was supported by the Neurinfo MRI research facility from the University of Rennes I. Neurinfo is granted by the the European Union (FEDER), the French State, the Brittany Council, Rennes Metropole, Inria, Inserm and the University Hospital of Rennes. This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the “Investissements d’Avenir” program under reference ANR-10-LABX-07-01. It was also funded by Brittany region under HEMISFER project.

REFERENCES

Abreu, R., Leal, A., and Figueiredo, P. (2018). EEG-Informed fMRI: A Review of Data Analysis Methods. *Frontiers in Human Neuroscience* 12

Allen, E. A., Damaraju, E., Eichele, T., Wu, L., and Calhoun, V. D. (2018). EEG Signatures of Dynamic Functional Network Connectivity States. *Brain Topography* 31, 101–116

Beck, A. and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences* 2, 183–202

Birbaumer, N., Ramos Murguialday, A., Weber, C., and Montoya, P. (2009). Chapter 8 Neurofeedback and Brain–Computer Interface. In *International Review of Neurobiology* (Elsevier), vol. 86. 107–117

Combettes, P. L. and Wajs, V. R. (2005). Signal Recovery by Proximal Forward-Backward Splitting. *Multiscale Modeling & Simulation* 4, 1168–1200

Cury, C., Maurel, P., Gribonval, R., and Barillot, C. (2019). A sparse EEG-informed fMRI model for hybrid EEG-fMRI neurofeedback prediction. *bioRxiv*

de Munck, J., Gonçalves, S., Huijboom, L., Kuijer, J., Pouwels, P., Heethaar, R., et al. (2007). The hemodynamic response of the alpha rhythm: An EEG/fMRI study. *NeuroImage* 35, 1142–1151

de Munck, J., Gonçalves, S., Mammoliti, R., Heethaar, R., and Lopes da Silva, F. (2009). Interactions between different EEG frequency bands and their effect on alpha–fMRI correlations. *NeuroImage* 47, 69–76

Engell, A. D., Huettel, S., and McCarthy, G. (2012). The fMRI BOLD signal tracks electrophysiological spectral perturbations, not event-related potentials. *NeuroImage* 59, 2600–2606. doi:10.1016/j.neuroimage.2011.08.079

Formaggio, E., Storti, S. F., Bertoldo, A., Manganotti, P., Fiaschi, A., and Toffolo, G. M. (2011). Integrating EEG and fMRI in epilepsy. *NeuroImage* 54, 2719–2731

Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* 1, 153–171

Goncalves, S. I., Bijma, F., Pouwels, P. W. J., Jonker, M., Kuijer, J. P. A., Heethaar, R. M., et al. (2008). A Data and Model-Driven Approach to Explore Inter-Subject Variability of Resting-State Brain Activity Using EEG-fMRI. *IEEE Journal of Selected Topics in Signal Processing* 2, 944–953

Gramfort, A., Kowalski, M., and Hämäläinen, M. (2012). Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in Medicine and Biology* 57, 1937–1961

Gramfort, A., Strohmeier, D., Haueisen, J., Hamalainen, M., and Kowalski, M. (2011). Functional Brain Imaging with M/EEG Using Structured Sparsity in Time-Frequency Dictionaries. In *IPMI*, eds. G. Székely and H. K. Hahn (Berlin, Heidelberg: Springer Berlin Heidelberg), vol. 6801. 600–611

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., et al. (2008). Review on solving the inverse problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation* 5, 25. doi:10.1186/1743-0003-5-25

Hammond, D. C. (2011). What is Neurofeedback: An Update. *Journal of Neurotherapy* 15, 305–336

Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage* 21, 1639–1651

Leite, M., Leal, A., and Figueiredo, P. (2013). Transfer Function between EEG and BOLD Signals of Epileptic Activity. *Frontiers in Neurology* 4

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* 45, S187–S198

Lioi, G., Cury, C., Perronnet, L., Mano, M., Bannier, E., Lécuyer, A., et al. (2019). Simultaneous MRI-EEG during a motor imagery neurofeedback task: an open access brain imaging dataset for multi-modal data integration. *bioRxiv*

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157. doi:10.1038/35084005

Magri, C., Schridde, U., Murayama, Y., Panzeri, S., and Logothetis, N. K. (2012). The Amplitude and Timing of the BOLD Signal Reflects the Relationship between Local Field Potential Power at Different Frequencies. *Journal of Neuroscience* 32, 1395–1407. doi:10.1523/JNEUROSCI.3985-11.2012

Mano, M., Lécuyer, A., Bannier, E., Perronnet, L., Noorzadeh, S., and Barillot, C. (2017). How to Build a Hybrid Neurofeedback Platform Combining EEG and fMRI. *Frontiers in Neuroscience* 11

Meir-Hasson, Y., Kinreich, S., Podlipsky, I., Hendler, T., and Intrator, N. (2014). An EEG Finger-Print of fMRI deep regional activation. *NeuroImage* 102, 128–141

Moosmann, M., Eichele, T., Nordby, H., Hugdahl, K., and Calhoun, V. D. (2008). Joint independent component analysis for simultaneous EEG–fMRI: Principle and simulation. *International Journal of Psychophysiology* 67, 212–221. doi:10.1016/j.ijpsycho.2007.05.016

Noorzadeh, S., Maurel, P., Oberlin, T., Gribonval, R., and Barillot, C. (2017). Multi-modal EEG and fMRI Source Estimation Using Sparse Constraints. In *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, eds. M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne (Springer International Publishing), vol. 10433, 442–450

Ou, W., Hämäläinen, M. S., and Golland, P. (2009). A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage* 44, 932–946

Pedregosa, F., Eickenberg, M., Thirion, B., and Gramfort, A. (2013). HRF Estimation Improves Sensitivity of fMRI Encoding and Decoding Models. In *2013 International Workshop on Pattern Recognition in Neuroimaging* (Philadelphia, PA, USA: IEEE), 165–169

Perronnet, L., Lécuyer, A., Mano, M., Bannier, E., Lotte, F., Clerc, M., et al. (2017). Unimodal Versus Bimodal EEG-fMRI Neurofeedback of a Motor Imagery Task. *Frontiers in Human Neuroscience* 11

Perronnet, L., Lécuyer, A., Mano, M., Clerc, M., Lotte, F., and Barillot, C. (2018). Learning 2-in-1: towards integrated EEG-fMRI-neurofeedback. *bioRxiv*

Rosa, M. J., Kilner, J., Blankenburg, F., Josephs, O., and Penny, W. (2010). Estimating the transfer function from neuronal activity to BOLD using simultaneous EEG-fMRI. *NeuroImage* 49, 1496–1509

Scheeringa, R., Fries, P., Petersson, K.-M., Oostenveld, R., Grothe, I., Norris, D. G., et al. (2011). Neuronal Dynamics Underlying High- and Low-Frequency EEG Oscillations Contribute Independently to the Human BOLD Signal. *Neuron* 69, 572–583. doi:10.1016/j.neuron.2010.11.044

Schwab, S., Koenig, T., Morishima, Y., Dierks, T., Federspiel, A., and Jann, K. (2015). Discovering frequency sensitive thalamic nuclei from EEG microstate informed resting state fMRI. *NeuroImage* 118, 368–375

Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., et al. (2017). Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience* 18, 86–100. doi:10.1038/nrn.2016.164

Sulzer, J., Haller, S., Scharnowski, F., Weiskopf, N., Birbaumer, N., Blefari, M., et al. (2013). Real time fMRI neurofeedback: Progress and challenges. *NeuroImage* 76, 386–399. doi:10.1016/j.neuroimage.2013.03.033

Tagliazucchi, E. and Laufs, H. (2015). Multimodal Imaging of Dynamic Functional Connectivity. *Frontiers in Neurology* 6

Thibault, R. T., MacPherson, A., Lifshitz, M., Roth, R. R., and Raz, A. (2018). Neurofeedback with fMRI: A critical systematic review. *NeuroImage* 172, 786–807

Wang, T., Mantini, D., and Gillebert, C. R. (2017). The potential of real-time fMRI neurofeedback for stroke rehabilitation: A systematic review. *Cortex* doi:10.1016/j.cortex.2017.09.006

Zotev, V., Phillips, R., Yuan, H., Misaki, M., and Bodurka, J. (2014). Self-regulation of human brain activity using simultaneous real-time fMRI and EEG neurofeedback. *NeuroImage* 85, 985–995. doi:10.1016/j.neuroimage.2013.04.126