Science Arts & Métiers (SAM)

is an open access repository that collects the work of Arts et Métiers Institute of Technology researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: https://sam.ensam.eu
Handle ID: .http://hdl.handle.net/10985/23068

This document is available under CC BY license

To cite this version :

Omar CHEHAB, Alexandre DEFOSSEZ, Alexandre GRAMFORT, Jean-Remi KING, Jean-Christophe LOISEAU - Deep Recurrent Encoder: an end-to-end network to model magnetoencephalography at scale - Neurons, Behavior, Data analysis, and Theory - Vol. 1, - 2022

Any correspondence concerning this service should be sent to the repository Administrator : scienceouverte@ensam.eu

Arts et Métiers Institute of Technology

NBDT

ORIGINAL ARTICLE

Journal Section

Deep Recurrent Encoder: an end-to-end network to model magnetoencephalography at scale

Omar Chehab[1,2*] | **Alexandre Défossez**[1*] | **Jean-Christophe Loiseau**[3] | **Alexandre Gramfort**[2] | **Jean-Remi King**[1,4]

[1]Facebook AI Research

[2]Université Paris-Saclay, Inria, CEA, Palaiseau, France

[3]ENSAM, Paris, France

[4]École normale supérieure, PSL University, CNRS, Paris, France

Correspondence
Jean-Remi King, Facebook AI Research
Email: jeanremi@fb.com

Funding information
Alexandre Gramfort, Inria: ANR-20-CHIA-0016, and European Research Council Starting Grant SLAB ERC-StG-676943; Jean-Remi King, École normale supérieure, PSL University: ANR-17-EURE-0017 and the Fyssen Foundation.

Understanding how the brain responds to sensory inputs from non-invasive brain recordings like magnetoencephalography (MEG) can be particularly challenging: (i) the high-dimensional dynamics of mass neuronal activity are notoriously difficult to model, (ii) signals can greatly vary across subjects and trials, and (iii) the relationship between these brain responses and the stimulus features is non-trivial. These challenges have led the community to develop a variety of preprocessing and analytical (almost exclusively linear) methods, each designed to tackle one of these issues. Instead, we propose to address these challenges through a specific end-to-end deep learning architecture, trained to predict the MEG responses of multiple subjects at once. We successfully test this approach on a large cohort of MEG recordings acquired during a one-hour reading task. Our Deep Recurrent Encoder (DRE) reliably predicts MEG responses to words with a three-fold improvement over classic linear methods. We further describe a simple variable

Abbreviations: MEG, magnetoencephalography; EEG, electroencephalography.

[*] Equally contributing authors.

importance analysis to investigate the MEG representations learned by our model and recover the expected evoked responses to word length and word frequency. Lastly, we show that, contrary to linear encoders, our model captures modulations of the brain response in relation to baseline fluctuations in the alpha frequency band. The quantitative improvement of the present deep learning approach paves the way to a better characterization of the complex dynamics of brain activity from large MEG datasets.

KEYWORDS
Magnetoencephalography (MEG), Encoding, Forecasting, Reading, Deep Learning

1 | INTRODUCTION

A major goal of cognitive neuroscience consists of identifying how the brain responds to distinct experimental conditions. While descriptive statistics and statistical tests are classically used to analyze neural data [1], this approach is not suited to predict how the brain should react to new conditions. The resulting models of the brain can thus be particularly challenging to compare. By contrast, predictive encoding models [2, 3] can be directly trained to predict brain responses to various experimental conditions, and compared on their ability to accurately predict novel conditions. For example, encoding models allow the estimation of integration constants in the brain [4, 5], the hierarchical organization of visual [6] and speech processing [7, 8]. Beyond MEG, predictive models have enabled automatic segmentation [9] and dynamical system identification [10, 11]. In functional Magnetic Resonance Imaging, predictive encoding models are starting to emulate complex neural processing [12] and are a step towards discovering new phenomena [13, 14]. Yet, this general objective of developing encoding models faces three major challenges when working with non-invasive and time-resolved signals collected by magneto- and electro-encephalography (M/EEG).

Challenge 1: rich response dynamics

M/EEG signals are known and promoted for their excellent temporal resolution. While this ability to measure cognitive processes at a millisecond time-scale offers unique opportunities for fine chronometry of neural responses in humans, it also makes such signals notoriously difficult to analyze. For example, brain responses to audio streams overlap in time making their identification difficult. To address this issue in the context of encoding models, it is standard to employ a Temporal Receptive Field (TRF) model [15, 16, 17, 18, 19, 20, 21, 22, 23, 24][1]. TRF models are commonly designed to predict neural responses to exogenous stimulation by fitting a linear regression model with a fixed time-lag window of past sensory stimuli. By doing so, the predictions derived from TRFs are only influenced by stimuli descriptors, enabling them to modulate their response based on previous brain activity. Consequently, unless the basal activity from previous time points is introduced as an exogenous feature, TRF cannot learn to capture neuronal adaptation responses [27], nor can it learn to vary an evoked response as a function of the pre-stimulus alpha power [28, 29].

[1] also referred to as Finite Impulse Response (FIR) analysis in fMRI [25], and Distributed Lag modeling in statistics [26]

Challenge 2: inter-trial and inter-subject variability

Neuronal recordings in general, and M/EEG in particular, can be extremely variable across trials and subjects [30, 3, 31, 29, 32]. To reduce the nuisance factors behind these variations such as eye blinks, head movements, cardiac, and face muscle activity which corrupt MEG recordings, it is common to make use of multiple sessions and subjects within a study. For example, several methods based on spatial filtering [33, 34, 35, 36, 37] or "hyper-alignment" use linear models such as canonical correlation analysis (CCA), partial least square regression (PLS), multi-view ICA and back-to-back regressions (B2B) [38, 39, 40, 41, 42, 43] to isolate the brain responses shared across trials and/or individuals. However, these denoising techniques can also remove relevant signals. For example, VanRullen [29] have repeatedly shown that evoked responses to sensory input can be modulated by pre-stimulus alpha activity in a predictable way. Averaging trials, or filtering out this variability during preprocessing would therefore prevent the identification of such phenomenon.

Challenge 3: identifying the relationship between brain responses and stimulus features

A large part of cognitive neuroscience aims to identify *how* the brain responds to stimulus features. For example, are V1 neurons tuned to respond to luminance, contrast, oriented lines, or faces? When and where does this elicit a response? To tackle this issue, it is common to present many stimuli to the subject, and fit a general linear model (GLM) to predict brain responses given a set of hypothetical features [2]. This approach can be limited, as GLMs only reveal brain responses to features predetermined by the analyst [44, 25, 45, 46] and understanding interactions between features often requires explicitly modeling these interactions (e.g. as cross-terms to form a quadratic polynomial) and feeding them to a linear regression [47].

Here, we propose to simultaneously address these three core challenges with a unique end-to-end “Deep Recurrent Encoding" (DRE) neural network trained to robustly predict brain responses from both (i) past MEG activity and (ii) current experimental conditions. We test DRE on 99 subjects recorded with MEG during a one-hour long reading task, and show that our model (1) better predicts MEG responses than standard models, (2) efficiently captures inter-trial and inter-subject variability, and (3) identifies feature-specific responses as well as interactions between basal activity and these evoked responses.

2 | MATERIALS AND METHOD

This section presents, with consistent and self-contained mathematical notations, a methodological progression from linear to nonlinear encoding models of neural dynamics as observed with MEG. It also discusses the statistical and computational benefits of recurrent models, as well as the novel methodological ideas proposed with the DRE model.

2.1 | Problem formalization

In the case of MEG, the measured magnetic fields x reflect a tiny subset of brain dynamics h — specifically a partial and macroscopic summation of the synaptic input to cortical pyramidal cells. Given the physics of electromagnetic fields propagation, it is standard to assume that these neuronal magnetic fields have a linear, stationary, and instantaneous relationship with the magnetic fields measured via MEG sensors [48]. We refer to this “readout operator” as C, a matrix which is subject-specific because it depends on the location of pyramidal neurons in the cortex and thus on the anatomy of each subject. Furthermore, the brain dynamics governed by a function f evolve according to their

past and to external stimuli u [49]. In sum, we can formulate the problem as follows:

$$\begin{cases} x_{\text{current}} &= Ch_{\text{current}} \\ h_{\text{current}} &= f(h_{\text{past}}, u_{\text{current}}) \end{cases}$$

2.2 | Operational Objective

Here, we aim to parameterize f with θ, and subsequently learn θ and C to obtain a statistical (as opposed to biologically constrained as in [50]) generative model of observable brain activity that accurately predicts MEG activity $\hat{x} \in \mathbb{R}^{d_x}$ given an initial state and a series of past stimuli.

Notations

We denote by $u_t \in \mathbb{R}^{d_u}$ the stimulus with d_u encoded features at time t, $x_t \in \mathbb{R}^{d_x}$ the MEG recording with d_x sensors, $\hat{x}_t$ its estimation, and by $h_t \in \mathbb{R}^{d_h}$ the underlying brain activity. Because the true underlying brain dynamics are never known, h will always refer to a model estimate. To facilitate the parametrization of f, it is common in the modeling of dynamical systems to explicitly create a "memory buffer" by concatenating successive lags. We adopt the bold notation $\boldsymbol{h}_{t-1:t-\tau_h} := \left[h_{t-1}, \dots, h_{t-\tau_h}\right] \in \mathbb{R}^{d_h \tau_h}$ for flattened concatenation of $\tau_h \in \mathbb{N}$ time-lagged vectors. With these notations, the dynamical models considered in this paper are described as:

$$\begin{cases} x_t &= Ch_t \\ h_t &= f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \boldsymbol{u}_{t:t-\tau_u}) \end{cases} \tag{1}$$

where

- $f : \mathbb{R}^{d_h \tau_h + d_u(\tau_u+1)} \to \mathbb{R}^{d_h}$ governs brain dynamics given the preceding brain states and external stimuli
- $C \in \mathbb{R}^{d_h \times d_x}$ is a linear, stationary, instantaneous, and subject-specific observability operator that makes a subset of the underlying brain dynamics observable to the MEG sensors.

2.3 | Models

Temporal Receptive Field (TRF)

Temporal receptive fields (TRF) [15] are arguably the most common model for predicting neural time series in response to exogeneous stimulation. The TRF equation is that of control-driven linear dynamics:

$$h_t = f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \boldsymbol{u}_{t:t-\tau_u}) = B\boldsymbol{u}_{t:t-\tau_u} \quad , \tag{2}$$

where $B \in \mathbb{R}^{d_h \times d_u \cdot (\tau_u+1)}$ is the convolution kernel that maps the stimuli to the brain response and $\theta = \{B\}$. By definition, the TRF kernel encodes the input-output properties of the system, namely, its characteristic time scale, its memory, and thus its ability to sustain an input over time. A computational drawback is that the TRF kernel size scales linearly with the duration of the neural response to the stimulus. For example, a dampened oscillation evoked by the stimulus could last one hundred time samples ($\tau_u = 99$) and would require $B \in \mathbb{R}^{d_h \times 100 d_u}$ to reach 100 steps in the past, even though oscillatory dynamics can be compactly written as a second-order differential equation expressing

h_t in terms of only two of its own past states (h_{t-1}, h_{t-2})[2]. Emulating this, we will introduce a recurrent component to the TRF model to tackle the issue of dimensionality.

Recurrent Temporal Receptive Field (RTRF)

A Recurrent Temporal Receptive Field (RTRF) is a linear auto-regressive model. The RTRF with exogenous input can model time-series from its own past (e.g., past brain activity) *and* from exogenous stimuli. Unrolling the recurrence reveals that current brain activity can be expressed in terms of past activity. This corresponds to recurrent dynamics with control:

$$h_t = f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \boldsymbol{u}_{t:t-\tau_u}) = A\boldsymbol{h}_{t-1:t-\tau_h} + B\boldsymbol{u}_{t:t-\tau_u} \quad , \tag{3}$$

where the matrix $A \in \mathbb{R}^{d_h \times (d_h \cdot \tau_h)}$ encodes the recurrent dynamics of the system and $\theta = \{A, B\}$.

The dependency of h_t on h_{t-1} in (3) means we need to unroll the expression of h_{t-1} in order to compute h_t. However, it has been shown that linear models perform poorly in this case, as terms of the form A^t (A to the power t) will appear, with either exponentially exploding or vanishing eigenvalues. This rules out optimization with first order methods due to the poor conditioning of the problem [51], or using a closed-form solution. To circumvent unrolling the expression of h_{t-1}, we need to obtain it from what is measured at time $t-1$. This however assumes the existence of an inverse relationship from x_{t-1} to h_{t-1}, which we assume here to be linear by using the pseudo inverse of C: $h_{t-1} = C^\dagger x_{t-1}$. As a result, h_t and x_t are identifiable to one another, and (3) can be solved in closed form as a regular linear system [52]. Initializing the RTRF dynamics with the pre-stimulus data can be written as:

$$h_t = C^\dagger x_t \qquad \forall t \in \{0, \ldots, \tau_h - 1\} \quad , \tag{4}$$

where τ_h is chosen to match the pre-stimulus duration τ.

Though the recurrent component of the RTRF is able to reduce the receptive field τ_u of TRF, it is nevertheless constrained to maintain a 'sufficiently big' receptive field τ_h to initialize over τ_h steps. The following model, DRE, will avoid this issue, and will also not require that h_t and x_t are identifiable via linear inversion.

Deep Recurrent Encoder (DRE)

DRE is an architecture based on the Long-Short-Term-Memory (LSTM) computational block [53]. It is useful to think of the LSTM as a "black-box nonlinear dynamical model", which composes the RTRF building block with nonlinearities and a memory module which reduces the need for receptive fields, so that $\tau_h = 1$ and $\tau_u = 0$. It is employed here to capture nonlinear dynamics evoked by a stimulus. A single LSTM layer can be formulated as [53]:

$$\begin{cases} h_t &= f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \boldsymbol{u}_{t:t-\tau_u}) = o_t \odot \tanh(m_t) \\ m_t &= d_t \odot m_{t-1} + i_t \odot \tilde{m}_t \\ \tilde{m}_t &= \tanh(Ah_{t-1} + Bu_t) \end{cases} \quad , \tag{5}$$

where the $\tanh$ nonlinearity is applied element-wise, $\odot$ is the Hadamard (element-wise) product, and $(d_t, i_t, o_t) \in (\mathbb{R}^{d_m})^3$ are data-dependent vectors with values between 0 and 1 modeled as forget (or drop) input and output gates, respectively. The memory module $m_t \in \mathbb{R}^{d_m}$ thus interpolates between a "past term" $m_{t-1} \in \mathbb{R}^{d_m}$ and a "prediction

[2] A sine wave can be produced by a simple linear auto-regressive (AR) model of order 2

term" $\tilde{m}_t \in \mathbb{R}^{d_m}$, taking h_t as input. The "prediction term" (See (5) last equation) resembles that of the previous RTRF model except that it is here composed with a tanh nonlinearity which conveniently normalizes the signals.

Again, the dependency of h_t on h_{t-1} in (3) meant that we needed to unroll the expression of h_{t-1} to compute h_t. While this is numerically unstable for the RTRF, the LSTM is designed such that h_t and its gradient are stable even for large values of t. As a result, h_t and x_t do not need to be identifiable to one another. In other words, contrary to RTRF, the LSTM allows h_t to represent a hidden state containing potentially more information than its corresponding observation x_t.

We now motivate three modifications made to the standard LSTM.

First, we help it recognize when (not) to sustain a signal, by augmenting the control u_t with a mask embedding $p_t \in \{0, 1\}$ indicating whether the provided MEG signal generates the current brain response (i.e. 1 before word onset and 0 thereafter). Second, we automatically learn to align subjects with a dedicated subject embedding layer. Indeed, a shortcoming of standard brain encoding analyses is that they are commonly performed on each subject separately. However, this implies that one cannot exploit potential similarities across subjects. Here, we adapt the LSTM in the spirit of Défossez et al. [54] so that a *single* model is able to leverage information across multiple subjects. We do this by augmenting the control u_t with a "subject embedding" $s \in \mathbb{R}^{d_s}$, that is learned for each subject. Note that this amounts to learning a matrix in $\mathbb{R}^{d_s \times n_s}$ that is applied to the one-hot-encoding of the subject number. In order words, each subject has a vectorized representation that is one column of the embedding matrix. Setting $d_s < n_s$ allows us to use the same LSTM block to model subject-wise variability, and to train across subjects simultaneously while leveraging similarities across subjects.

Third, for comparability purposes, RTRF and LSTM should access the same pre-stimulus MEG information $\boldsymbol{x}_{\tau:1}$. Incorporating the initial MEG, before word onset, is done by augmenting the control with $p_t \odot x_t$. The extended control reads: $\tilde{u}_t = \left[u_t, s, p_t, p_t \odot x_t\right]$, and the LSTM with augmented control $\tilde{u}_t$ finally reads:

$$h_t = f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \tilde{\boldsymbol{u}}_{t:t-\tau_u}) = \mathrm{LSTM}_\theta(h_{t-1}, \tilde{u}_t) = \mathrm{LSTM}_\theta(\mathrm{LSTM}_\theta(h_{t-2}, \tilde{u}_{t-1}), \tilde{u}_t) \quad . \tag{6}$$

In practice, to maximize expressivity, two modified LSTM blocks are stacked on top of one another (Figure 1).

Having introduced a nonlinear dynamical system for the brain response h_t, we can also extend the model (1) by challenging the linear instantaneous mixing from the brain response h_t to the measurements x_t. Introducing two new nonlinear functions d and e, respectively parametrized by θ_2 and θ_3, a more general model formally reads:

$$\begin{cases} \boldsymbol{x}_{t:t-\tau_x+1} &= d_{\theta_2}(h_t) \\ h_t &= f_{\theta_1}(\boldsymbol{h}_{t-1:t-\tau_h}, e_{\theta_3}(\tilde{\boldsymbol{u}}_{t:t-\tau_u})) \end{cases} , \tag{7}$$

where τ_x allows us to capture a small temporal window of data around x_t, and τ_u is taken to be much larger than τ_x. Indeed (7) corresponds to (1) if one sets $\tau_x = 1$ and $d_{\theta_2}(h_t) = Ch_t$, as well as $e_{\theta_3}(\tilde{\boldsymbol{u}}_{t:t-\tau_u}) = \boldsymbol{u}_{t:t-\tau_u}$. In more intuitive terms, the DRE model generalizes the linear instantaneous measurement of the previous models with a "convolutional autoencoder" [55]. The e (encoder) function is formed by convolutions and the d (decoder) function uses transposed convolutions, where both functions are two layers deep (Figure 1) [3].

In practice, we use a kernel size $K = 4$ for the convolutions. This impacts the receptive field of the network and the parameter τ_x. Equation (7) implies that the number of time samples in h and x are the same. However, a strong

[3] While "encoding" typically means outputting the MEG with respect to the neuroscience literature, we use "encoder" and "decoder" in the context of deep learning auto-encoders [56] in this paragraph.

benefit of the convolutional auto-encoder is to perform a reduction of the number of time steps by using a stride S larger than 1. By using a stride of 2, one reduces the temporal dimension by 2. Indeed it boils down to taking every other time sample from the output of the convolved time series. Given that the LSTM module is by nature sequential, this reduces the number of time steps it has to consider when learning, which accelerates both training and evaluation. Further, there is evidence that LSTMs can only pass information over a limited number of time steps [57]. In practice, we use d_h output channels for the convolutional encoder.

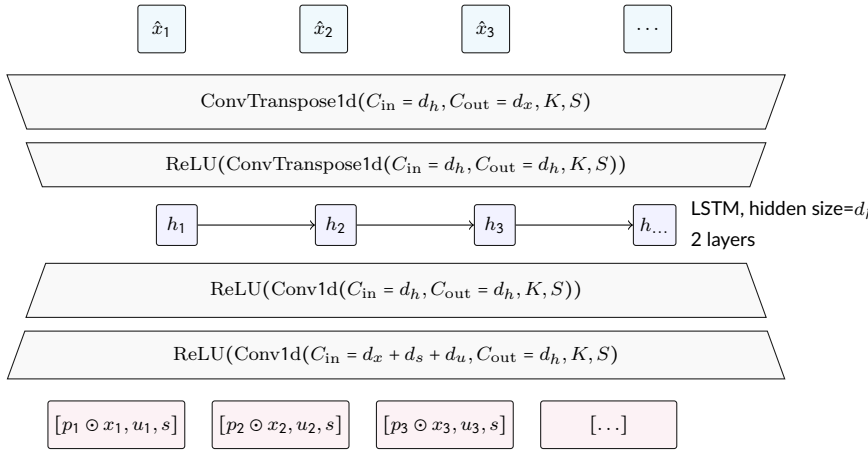

FIGURE 1 Representation of the Deep Recurrent Encoder (DRE) model used to predict MEG activity. The masked MEG $p_t \odot x_t$ enters the network from the bottom, along with the control representation u_t and the subject embedding s. The encoder transforms the input with convolutions and ReLU nonlinearities. Then, the LSTM models the sequence of hidden states $\boldsymbol{h}_t$, which are converted back to the MEG activity estimate $\hat{x}_t$. $\mathrm{Conv1d}(C_{\mathrm{in}}, C_{\mathrm{out}}, K, S)$ represents a convolution over time with C_{in} input channels, C_{out} output channels, a kernel size K, and a stride S. Similarly, $\mathrm{ConvTransposed1d}(C_{\mathrm{in}}, C_{\mathrm{out}}, K, S)$ represents a transposed convolution over time.

In summary, our DRE model generalizes TRF and RTRF models by using nonlinearities both in the dynamics of the brain response h_t and in its measurement x_t (1). It is done respectively with LSTM cells and a convolutional auto-encoder. Importantly, the DRE is equipped with a subject embedding allowing us to learn a joint model for the group of subjects.

2.4 | Optimization losses

The dynamics for the above three models (TRF, RTRF, DRE) are given by different expressions of f_θ as well as the mappings between x and h via C for TRF and RTRF, or c and e for DRE.

At test time, the models aim to accurately forecast MEG data from initial steps combined with subsequent stimuli. Consequently, one should train the models in the same setup. This boils down to minimizing a “multi-step-ahead" ℓ_2

prediction error:

$$\begin{cases} \text{minimize}_{\ \theta_1,\theta_2,\theta_3} & \sum_t \|x_t - \hat{x}_t\|_2^2 \\ \text{s.t.} & \boldsymbol{x}_{t:t-\tau_x+1} = d_{\theta_2}(h_t) \\ & h_t = f_{\theta_1}(\boldsymbol{h}_{t-1:t-\tau_h}, e_{\theta_3}(\tilde{\boldsymbol{u}}_{t:t-\tau_u})) \end{cases}$$

This “multi-step-ahead" minimization requires unrolling the recurrent expression of h_t over the preceding time steps, which the LSTM-based DRE model is able to do [58]. The DRE model takes as input the observed MEG at the beginning of the sequence, and must predict the future MEG measurements using the (augmented) stimulus $\tilde{u}_t$. Note that the mapping to and from the latent space, e_{θ_3} and d_{θ_2}, are learned *jointly* with the dynamical operator f_{θ_1}. Furthermore, the DRE has reduced temporal receptive fields, thus the computational load is lightened and allows for a low or high-dimensional latent space.

Given that the RTRF model is linear and can suffer from numerical instabilities (see above), it is trained with the “one-step-ahead” version of the predictive loss with squared ℓ_2 regularization:

$$\begin{cases} \text{minimize}_{\ \theta} & \sum_t \|x_t - \hat{x}_t\|_2^2 + \lambda\|\theta\|_2^2 \\ \text{s.t.} & \hat{x}_t = Ch_t \\ & h_t = f_\theta(\boldsymbol{h}_{t-1:t-\tau_h}, \boldsymbol{u}_{t:t-\tau_u}) \\ & \boldsymbol{h}_{t-1:t-\tau_h} = C^\dagger \boldsymbol{x}_{t-1:t-\tau_h} \end{cases}$$

TRF models are also trained with this “one-step-ahead” loss. As mentioned above, the linear models (TRF) require a larger receptive field than the nonlinear DRE. Large receptive fields induce a computational burden, because each time lag comes with a spatial dimension of size d_h or d_u. To tackle this issue, C is chosen to reduce this spatial dimension. In practice, we choose to learn C *separately* from the dynamics to simplify the training procedure of the linear models. Given a participant, we fit a Principal Component Analysis (PCA) with 40 components on their averaged (evoked) MEG data: the resulting PCA map is taken to be the matrix $C \in \mathbb{R}^{d_x \times 40}$. The resulting latent space will thus explain most of the variance of the original recording. Indeed, when training the TRF model on all 270 MEG sensors with no hidden state ($6.4 \pm 0.22\%$, MEAN and SEM across subjects) or using a 40-component PCA ($6.43 \pm 0.17\%$), we obtained similar performances. The pseudo-inverse $C^\dagger$ required to compute the previous latent state h_{t-1} is also obtained from the PCA model. Note that dimensionality reduction via linear demixing is a standard preprocessing step in MEG analysis [34, 59, 60].

2.5 | Model Evaluation

Evaluation Method

Following seminal works (*e.g.* by Kay et al. [61], Güçlü and Gerven [62]), models are evaluated using the Pearson Correlation R (between -1 and 1) between the model prediction $\hat{x}$ and the true MEG signals x for each channel and each time sample (after the initial state) independently[4]. When comparing the overall performance of the models, we average over all time steps after the stimulus onset, and over all MEG sensors for each subject independently.

[4] Figure 5 in the Appendix 5 reports the same evaluations, using a different metric: the explained variance quantified by the coefficient of determination R^2 of the brain response by the model predictions.

Feature Importance

To investigate what a model actually learns, we use Permutation Feature Importance [63] which measures the drop in prediction performance when the jth input feature u^j is shuffled:

$$\Delta R_j = R - R_j \quad , \tag{8}$$

By tracking ΔR over time and across MEG channels, we can locate in time and space the contribution of a particular feature (e.g. word length) to the brain response.

Experiment

The model weights are optimized with the training and validation sets, while the penalization λ for the linear models (TRF and RTRF) is optimized with a grid search over five values distributed logarithmically between 10^{-3} and 10^{3}. Training of the DRE is performed with ADAM [64] using a learning rate of 10^{-4} and PyTorch's default parameters [65] for the running averages of the gradient and its square. The training is stopped when the error on the validation set increases. In practice, the DRE and the DRE-PCA were trained over approximately 20 and 80 epochs, respectively.

Statistics

Each subject score is obtained using the model prediction on held-out trials, using the learned subject embeddings as the "alignment function", similar to Chen et al. [66], Zhang et al. [67], Haxby et al. [68], Bazeille et al. [42], Richard et al. [43]. To test the reliability of our effects (e.g. prediction performance, feature importance, model comparison), we assess confidence intervals and p-values across subjects using a non-parametric Wilcoxon rank test across subjects. When applicable, we correct these estimates for multiple comparisons using a false discovery rate (FDR) across time samples and channels. Note that subjects can be treated as independent observations to derive meaningful p-values since the statistics are based on held-out data independent from the training set.

Noise Ceiling

Noise ceilings are typically estimated using batches of repeated conditions [69, 70, 62], to evaluate the maximal amount of explainable variance. This involves multiple presentations of the same stimulus characterized by a given feature set. In our case, however, sentences are presented only once to each subject. Further, we cannot control one of the variables input to our encoding models: namely, the baseline brain activity.

2.6 | Data

Experimental design

We analyze 99 subjects from the Mother Of Unification Studies (MOUS) dataset [71] who performed a one-hour reading task while being recorded with a 273-channel CTF MEG scanner. The task consisted of reading approximately 2,700 words flashed on a screen in rapid series. Words were presented sequentially in groups of 10 to 15, with a mean inter-stimulus interval of 0.78 seconds (min: 300ms, max: 1,400ms). Sequences were either random word lists or actual sentences (50% each). For this study, both conditions were used. However, this study does not investigate the differences obtained across these two conditions. Out of the original 102 subjects, 3 were discarded from the study because we could not reliably parse their stimulus channels.

Stimulus preprocessing

We focus on four well-known features associated with reading, namely word length (i.e., the number of letters in a word), word frequency in natural language (as derived by the wordfreq Python package [72], and measured on a logarithmic scale), and a binary indicator for the first and last words of the sequence. At a given time t, each stimulus $u_t \in \mathbb{R}^4$ is therefore encoded with four values, fed to the models as a square function that is non-zero for the duration of the stimulus. Each feature is standardized to have zero mean and unit variance. Word length is expected to elicit an early (from t=100 ms) visual response in posterior MEG sensors, whereas word frequency is expected to elicit a late (from t=400 ms) left-lateralized response in anterior sensors. In the present task, word length and word frequency are correlated R=-0.48.

MEG Preprocessing

As we are primarily interested in evoked responses [73], we band-pass filtered between 1 and 30 Hz and downsampled the data to 120 Hz using the MNE software [74] with default settings: i.e. a FIR filter with a Hamming window, a lower transition bandwidth of 1 Hz with -6 dB attenuation at 0.50 Hz and a 7.50 Hz upper transition bandwidth with an attenuation of -6 dB at 33.75 Hz.

To limit the interference of large artefacts on model training, we use Scikit-learn's RobustScaler with default settings [75] to normalize each sensor using the 25th and 75th quantiles. Following this step, most of the MEG signals will have a scale around 1. Since we observed a few large scale outliers, we chose to reject any segment of 3 seconds that contains amplitudes higher than 16 in absolute value (fewer than 0.8% of the time samples).

These continuous data are then segmented between 0.5 s before and 2 s after word onset, yielding a three-dimensional MEG tensor per subject: words, sensors, and time samples relative to word onset. For each subject, we form a training, validation, and test set using respectively 70%, 10%, and 20% of these segments, ensuring that two segments from different sets do not originate from the same word sequence to avoid information leakage. This corresponds to 191K, 27K, and 53K segments used for the train, validation, and test sets, respectively. Each segment has a spatial dimension of 273 sensors and a temporal dimension of 300 time points (2.5 s sampled at 120 Hz).

For clarity, some figures use global field power (GFP) to summarize effects over time. GFP refers to the standard deviation across MEG channels of an average evoked response.

2.7 | Model training

Model hyper parameters

We compare the three models introduced in Section 2.3 over $n_s = 99$ subjects. For the TRF, we use a lag on the control of $\tau_u = 250$ time steps (about 2 s). This corresponds to the duration of the signal after the stimulus onset. For the RTRF, we use $\tau_u = \tau_h = 40$ time steps. This lag is close to the minimum inter-word duration of 300 ms, and corresponds to the part of the initial MEG (i.e. 333 ms out of 500 ms, at 120 Hz) that is passed to the model to predict the 2.5 s MEG sequence during the evaluation.

For the DRE model, we use a subject embedding of dimension $d_s = 16$, a latent state of dimension $d_h = 512$, a kernel size $K = 4$, and a stride $S = 2$. The subject embeddings are initialized as Gaussian vectors with zero mean and unit variance, while the weights of the convolutional layers and the LSTM are initialized using the default "Kaiming" initialization [76]. Like its linear counterpart, the DRE is given the first 333 ms of the MEG signal to predict the complete 2.5 s of a training sample.

Ablation study

To investigate the importance of the different components of the DRE model, we implement an ablation study by fitting the model with all but one of its components. To this end, we compare the DRE to i) the DRE without using the 333 ms of pre-stimulus initial MEG data (DRE NO-INIT), ii) the DRE trained in the 40-dimensional PCA space used for the linear models (DRE PCA), iii) the DRE devoid of a subject embedding (DRE NO-SUBJECT), and to iv) the DRE devoid of the convolutional auto-encoder (DRE NO-CONV).

2.8 | Code availability

The code developed for the present study is available at https://github.com/facebookresearch/deepmeg-recurrent-encoder.

3 | RESULTS

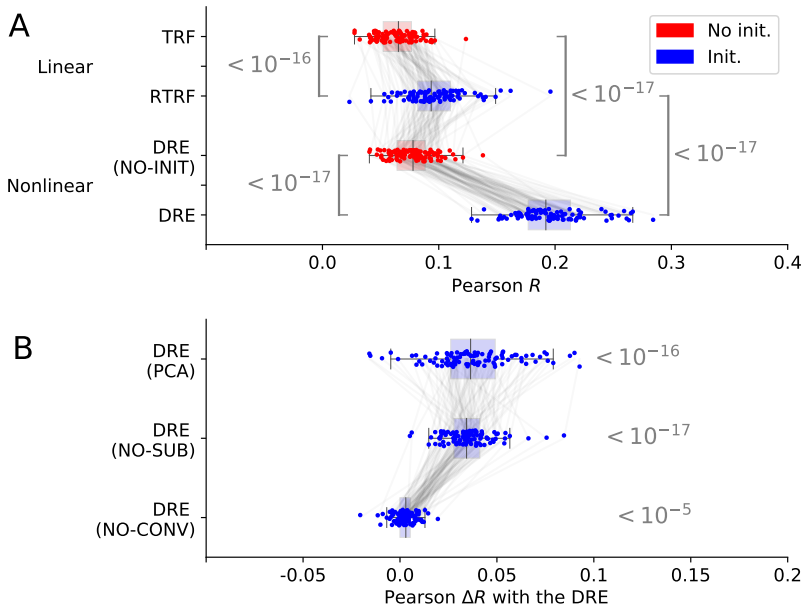

FIGURE 2 *A: Model Comparison*. Predictive performance of different models over 99 subjects (dots) (chance=0). Boxplots indicate median and interquartile range, while gray lines follow each subject. The gray brackets indicate the p-value obtained with a pair-wise Wilcoxon comparison across subjects. Initialization and nonlinearity increase predictive performance. *B: Ablation Study of our model (DRE)*. Pearson R mean obtained by removing a component from (Convolutional Auto-encoder or Subject-embedding) or adding an element (PCA embedding composed with the Convolutional Auto-encoder) to the DRE architecture and retraining the model, and the resulting p-values obtained by comparing the resulting scores to those of the normal DRE. Convolutions marginally impact performance of the DRE (they are essentially used for speed), but training over all MEG sensors (vs. PCA space) and using subject embeddings designed for realignment are responsible for an increase of approximately 4%.

We first evaluate the DRE's ability to predict brain responses to written words presented in rapid serial visual presentation and measured with MEG, and compare these brain predictions to those of linear encoding models (TRF, RTRF). Then, we show with ablation experiments which elements of the DRE help address the challenges of rich dynamics, inter-subject, and inter-trial variability. Finally, we show how feature importance helps address the third challenge introduced above, namely: identifying the relationship between brain responses and stimulus features.

Modeling rich MEG dynamics: model comparison.

DRE outperforms the baseline TRF and RTRF models with up to a three-fold improvement (Figure 2). To provide a fair comparison between the models, we also compare TRF to a NO-INIT DRE, i.e. to a DRE architecture that ignores the pre-stimulus MEG activity. The results show that DRE NO-INIT consistently outperforms TRF (mean correlation score R = 0.077 on average across all subjects, time samples, and all channels; standard error of the mean across subjects: ± 0.002 for DRE NO-INIT vs. R = 0.064 ± 0.002 for TRF). This difference is strongly significant ($p < 10^{-17}$) under a Wilcoxon test across subjects. Similarly, DRE (R=0.20 ± 0.003) significantly ($p < 10^{-17}$) outperforms RTRF (R=0.10 ± 0.003), when both of these models are given as input the pre-stimulus MEG activity. To verify that this gain is not trivially accounted for by the limited dimensionality of RTRF (trained with 40-dimensional Principal Components because of computational limitations), we trained DRE with the *same* PCA-reduced data as RTRF. The results confirm that DRE obtains a higher performance (R=0.16 ± 0.003, $p < 10^{-16}$) than RTRF. Overall, these results suggest that DRE better models the rich M/EEG dynamics than linear models.

Subject embeddings efficiently capture inter-individual variability

To evaluate the importance of the subject-embedding layer in capturing inter-individual variability, we trained the DRE without a subject embedding layer (DRE NO-SUB). The comparison between DRE and DRE NO-SUB reveals a clear difference (ΔR = 0.038, $p < 10^{-17}$). This result shows that the subject embedding layer efficiently re-aligns subjects' brain responses to model the dynamics specific to each – or shared across – subject(s).

Recurrence efficiently captures inter-trial variability.

Brain responses to sensory input are known to vary with ongoing brain activity[29, 77]. Recurrent models (RTRF, DRE) are thus well suited to capture such phenomenon: initialized with 333 ms of pre-stimulus MEG, they can use basal brain activity to modulate the post-stimulus MEG predictions. Our results confirm this prediction: TRF is outperformed by RTRF (0.10 ± 0.003, $p < 10^{-16}$) with an average performance increment of ΔR = 0.03, and DRE (0.20 ± 0.003) outperforms DRE NO-INIT (0.077 ± 0.002%, $p < 10^{-17}$) with an average performance increment of ΔR = 0.12.

DRE's recurrence specifically captures alpha-dependent evoked responses.

To further explore how DRE learns inter-trial variability, we investigate a well-known interaction between evoked responses and pre-stimulus activity. Specifically, brain responses to sensory input are known to be modulated by pre-stimulus oscillatory activity in the "alpha" frequency range (8 - 13 Hz) [29]. To test whether this phenomenon can be detected in the present dataset, we compared the average evoked responses to words for "high pre-stimulus alpha" versus "low alpha" trials, using a median split for each subject separately. The results (Figure 3) show an effect of up to 50×10^{-12}T in fronto-temporal channels, peaking around 400 ms after word onset. Critically, while the single-trial predictions of DRE capture this phenomenon, neither TRF nor RTRF learn to modulate their evoked responses depending on the alpha power (Figure 3B. Bottom).

This interaction between pre-stimulus alpha activity and evoked responses varies with the content of words, and more specifically, with their frequency in natural language: a factorial split between "high alpha" *versus* "low alpha"

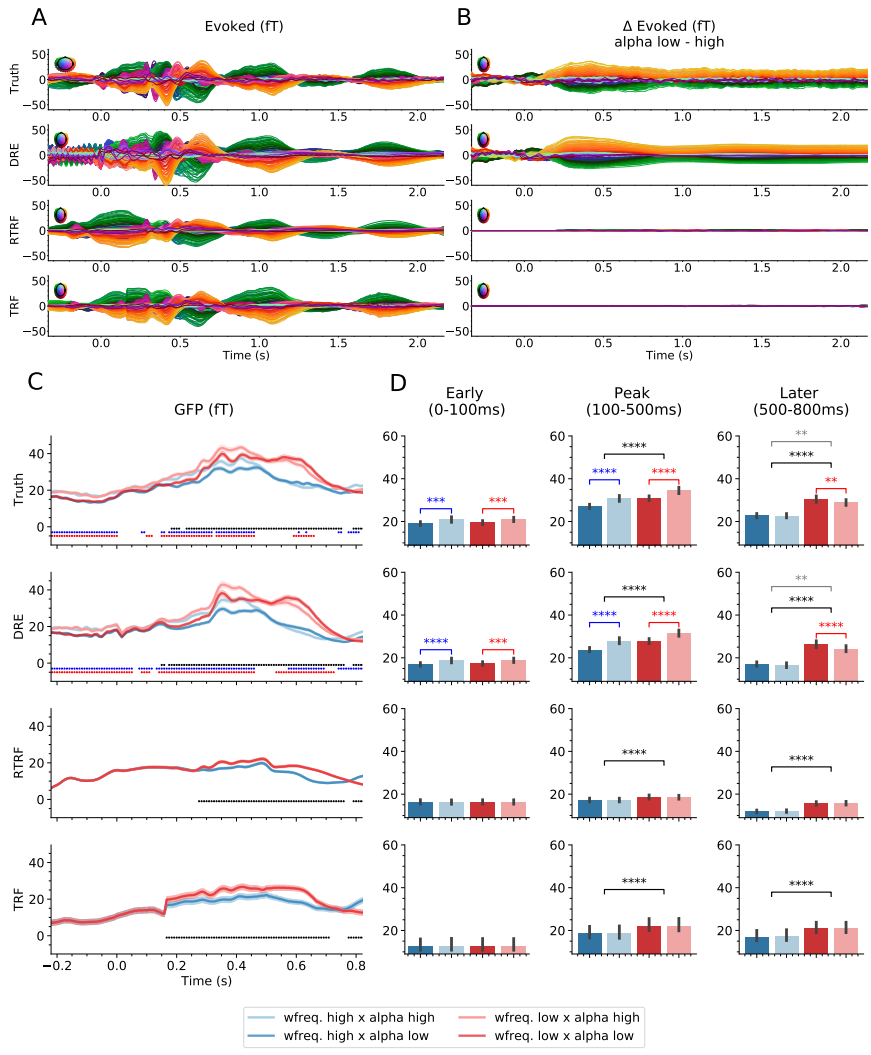

FIGURE 3 **A:** Brain response obtained for all three models (TRF, RTRF, DRE) and the actual data (Truth), averaged over the test set for all 99 subjects. Qualitatively, all models learn the mean MEG response, although with a variable precision. **B:** Difference of average brain response between trials with high *vs* low pre-stimulus alpha power (8-13 Hz). This analysis shows that DRE modulates the predicted brain response as a function of pre-stimulus alpha power. **C:** Global Field Power (GFP) of the brain response, as a function of pre-stimulus alpha power and word frequency. Qualitatively, all models capture the stimulus-modulation of the brain response, but only the DRE captures the alpha-modulation, which varies with the latency. **D:** Global Field Power (GFP) of the brain response averaged before the peak (0-100ms), during the ramp up of the peak (100-500ms) and after the peak of the evoked response (500-800ms). The number of stars illustrates the level of significance (**: 10^{-2}, ***: 10^{-3}, or ****: 10^{-4}). The marginal pre-stimulus alpha effect (light vs. dark) is followed by the marginal word frequency effect (black) and then an interaction (grey) between pre-stimulus alpha and word frequency (red vs. blue). Only DRE captures the same effects as in the actual data.

trials and "high word frequency" *versus* "low word frequency" trials resulted in both main and interaction effects (Figure 3C-D). Specifically, the comparison between these 2x2 conditions reveals three main phases. First, a main effect of alpha can be observed before the evoked response (light vs. dark lines in Figure 3C, $p < 10^{-3}$). Second, the main effect of word frequency starts to become significant from $\approx$ 200 ms (blue vs. red lines, $p < 10^{-4}$). Finally, the main effect of alpha starts to fade away after $\approx$ 500 ms ($p > 10^{-2}$ for high-frequency words), but its interaction with the stimulus continues to be significant ($p < 10^{-2}$). Critically, DRE learns these interactions between pre-stimulus alpha power and stimulus responses, while the linear models do not.

Feature importance helps interpreting the links between brain responses and stimulus features

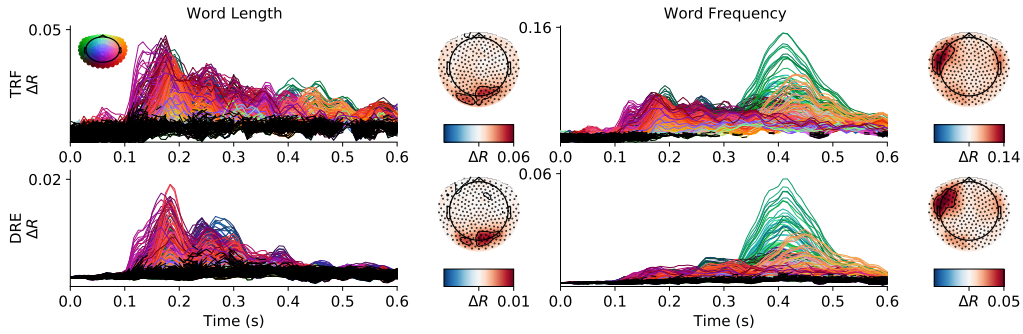

FIGURE 4 Permutation importance (ΔR) of word length (left column) and word frequency (right column), as a function of spatial location (color-coded by channel position, see top-left rainbow topography) and time relative to word onset for the two main encoding models (rows). The amplitudes of each line represent the mean across words and subjects for each channel. An FDR-corrected Wilcoxon test across subjects assesses the significance for each channel and time samples independently. Non-significant effects (p-value higher than 5%) are displayed in black. Overall, this feature importance analysis confirms that early visual and late frontal responses are modulated by word length and word frequency, respectively, as expected.

Interpreting nonlinear and/or high-dimensional models is notoriously challenging [78]. This issue poses strong limitations on the application of deep learning to neural recordings, where interpretability remains a major goal [3, 79]. While DRE faces the same types of issues as any deep neural network, we show below that a simple feature importance analysis of the predictions of this model (as opposed to its parameters) yields results that are consistent with those obtained by linear models, and with those described in neuroscientific literature (cf. Section 2.5).

Feature importance quantifies the loss of prediction performance ΔR when a unique feature is shuffled across words as compared to a non-shuffled prediction. Here, we focus our analysis on word length and word frequency, as these two features have been repeatedly associated with early sensory responses in the visual cortex and late lexical responses in the temporal cortex, respectively [80, 81]. As expected, the feature importance for word length in Figure 4 peaked around 150 ms in posterior MEG channels, whereas the feature importance of word frequency peaked around 400 ms in fronto-temporal MEG channels, for both the TRF and the DRE models. Furthermore, we recover a second phenomenon known in the literature: the *lateralization* of lexical processing in the brain. Indeed, Figure 4 shows, for the word frequency, an effect similar in shape across hemispheres, but significantly higher in amplitude for the left hemisphere (e.g. $p < 10^{-10}$ in the frontal region, $p < 10^{-12}$ in the temporal region, for the DRE).

These results suggest that, in spite of being high dimensional and nonlinear, DRE can be interpreted similarly to

linear models in the present context.

4 | DISCUSSION

The present study demonstrates that DRE outperforms several of its linear counterparts to predict MEG time series. In particular, it addresses the three challenges introduced above. First, the complex, nonlinear and non-stationary dynamics can be efficiently modeled by deep convolutional LSTM layers. Second, inter-trial and inter-individual variability can be addressed with recurrence (i.e. MEG-INIT) and subject embeddings, respectively. Finally, the relationship between stimulus features and brain responses can be interpreted in light of a permutation-based feature importance analyses.

Overall, the present study shows that the gain in prediction performance obtained by deep learning algorithms may not necessarily come at the price of interpretability. Indeed, we show here that DRE can be probed *a posteriori* to reveal how evoked responses relate to each stimulus feature and/or to pre-stimulus brain activity. This feature importance supplements ongoing efforts to open black-box models of brain activity. For example, Güçlü and Gerven [62] used a recurrent neural network to predict fMRI recordings, and quantified the impact of stimulus features by correlating them with the model's hidden state. Similarly, Keshishian et al. [5], analyzed the activations of a deep convolutional network with TRF to show how they captured "dynamical receptive fields". In both of these cases, however, these post-hoc analyses are based on (i) linear assumptions and (ii) the inner activations of the model. By contrast, the permutation feature importance used here focuses on probabilistic dependencies between input features and the models' predictions, which generalize linear dependencies measured by correlation [63]. This approach can thus be applied to any black-box predictive model.

Deep neural networks have not yet emerged as the go-to method for neuroimaging [82, 83]. Nevertheless, several studies have successfully used deep learning architectures to model brain signals [84, 85, 86, 87, 45]. In particular, deep nets trained on natural images [88, 89, 90], sounds [5], or text [91] are being increasingly used as an *encoding* model for predicting neural responses to sensory stimulation [6, 92, 7, 8]. Conversely, deep nets have also been trained to *decode* sensory-motor signals from neural activity, successfully reconstructing text [93] from Electrocorticography (ECoG) recordings, or images [94] from functional magnetic resonance imaging (fMRI). Despite these successes, we would like to argue that what possibly limits a wider impact of deep learning in human cognitive neuroscience is a combination of factors including: (i) low signal-to-noise ratio, (ii) small datasets, and (iii) a lack of high temporal resolution, where nonlinear dynamics may be the most prominent. The present experimental results make a step in this direction, and could thus open an avenue towards leveraging the many existing shorter naturalistic stimulus datasets collected on many subjects. This could be an alternative to making new long recordings of many hours of data from a handful of subjects [95].

While the DRE's architecture may be efficient at handling the dynamical structure of brain data, the dynamics assessed in this study are driven by specific linguistic features (i.e. word-length and word frequency). By contrast, recent ECoG and MEG studies have used more complex word features, represented as activations of a deep network pretrained on visual or language tasks [96, 97, 98], and then predict the brain response in a way that is unaware of the dynamics (using a linear classifier for each time sample independently). Given the successes independently observed with these two approaches, a natural extension of this work would be to combine the two and learn to map complex stimuli to brain responses using (1) rich representations for the stimuli (such as the activations of a pretrained deep network), followed by (2) a rich dynamical model such as the DRE. It is however worth pointing out that this approach would naturally lead to more high-dimensional parameter spaces, which would require larger datasets to

limit potential overfitting.

The present work is based on a deterministic and predictive framework using deep learning. Other complementary approaches such as Hidden Markov Models (HMMs) and Gaussian Processes (GPs) have also been proposed to model brain data in a probabilistic framework. Such approaches have been exploited to explore the spatio-temporal statistics of fMRI or MEG data [99, 100], but also in an encoding context [101, 102]. In particular, [99] combine GPs and dynamical system modeling to account for MEG responses to tactile input and shows that it captures meaningful modulations of oscillatory activity. This approach may offer a promising avenue to further clarify the interaction between baseline alpha oscillations and visual responses captured by DRE (see Figure 3). Similarly, [102] show that Gaussian modeling efficiently learns to predict fMRI responses to visual stimuli, and, importantly, can be inverted to achieve zero-shot decoding of individual characters. By contrast, our encoding approach would necessitate additional fine-tuning to transfer DRE to a novel decoding task. Overall, a key advantage of probabilistic models like GP is the ability to quantify uncertainty in the predictions, which in the present forecasting scenario would likely increase when looking at late latencies. While the proposed approach does not offer this possibility, the present study benefits from the highly-optimized ecosystem of deep learning, which allows us to efficiently deal with the large size of raw MEG data (273 MEG channels sampled at 1,200 Hz and recorded for 60 min in 99 subjects).

It is worth noting that because the losses used to train the models in this paper are MSEs evaluated in the time domain, the TRF, RTRF, and DRE are solely trained and evaluated on their ability to predict the *amplitude* of the neural signal *at each time sample*. Consequently, the objective solely focuses on “evoked" activity, i.e. neural signals that are strictly phase-locked to stimulus onset or to past brain signals [73]. A quick time-frequency decomposition of the models suggest that none of them capture “induced" activity, e.g. changes in the amplitude of neural oscillations with a non-deterministic phase. A fundamentally distinct loss would be necessary to capture such non phase-locked neural dynamics.

As for many other scientific disciplines, deep neural networks will undoubtedly complement – if not shift – the myriad of analytical pipelines developed over the years toward standardized end-to-end modeling. While such methodological development may improve our ability to predict how the brain responds to various exogenous stimuli, the present attempt already highlights the many challenges that this approach entails. Nevertheless, the present results hopefully clearly demonstrate that deep networks are a very relevant technology to capture complex neural dynamics collected non-invasively by MEG and certainly EEG.

Acknowledgements

Experiments on MEG data were made possible thanks to MNE [103, 74], as well as the scientific Python ecosystem: Matplotlib [104], Scikit-learn [75], Numpy [105], Scipy [106] and PyTorch [65].

This work was supported by the French ANR-20-CHIA-0016 and the European Research Council Starting Grant SLAB ERC-StG-676943 to AG, and by the French ANR-17-EURE-0017 and the Fyssen Foundation to JRK for his work at PSL.

The authors would like to thank Nicholas Tolley and Hubert Banville for their feedback and suggestions on the early versions of this manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 2007;164(1):177–190.

[2] Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage 2011;56(2):400–410.

[3] King JR, Gwilliams L, Holdgraf C, Sassenhagen J, Barachant A, Engemann D, et al. Encoding and Decoding Framework to Uncover the Algorithms of Cognition. The Cognitive Neurosciences VI, MIT Press 2020;6:691–702.

[4] Mahto S, Vo VA, Turek J, Huth AG. Multi-timescale representation learning in LSTM Language Models. ArXiv 2021;abs/2009.12727.

[5] Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. eLife 2020 jun;9:e53445.

[6] Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience 2016;19(3):356.

[7] Caucheteux C, King JR. Brains and algorithms partially converge in natural language processing. Communications Biology 2022;5(1):1–10.

[8] Millet J, King JR. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. arXiv preprint arXiv:210301032 2021;.

[9] Petreska B, Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, et al. Dynamical segmentation of single trials from population neural data. In: Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.; 2011. p. 756–764.

[10] Duncker L, Bohner G, Boussard J, Sahani M. Learning interpretable continuous-time models of latent stochastic dynamical systems. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research PMLR; 2019. p. 1726–1734.

[11] Maheswaranathan N, Williams A, Golub M, Ganguli S, Sussillo D. Universality and individuality in neural dynamics across large populations of recurrent networks. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 32 Curran Associates, Inc.; 2019. .

[12] Marblestone AH, Wayne G, Kording KP. Towards an integration of deep learning and neuroscience. bioRxiv 2016;.

[13] Seeliger K, Fritsche M, Güçlü U, Schoenmakers S, Schoffelen JM, Bosch SE, et al. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. NeuroImage 2018;180:253–266.

[14] Seeliger K, Ambrogioni L, Güçlütürk Y, van den Bulk LM, Güçlü U, van Gerven MA. End-to-end neural system identification with neural information flow. PLOS Computational Biology 2021;17(2):e1008558.

[15] Aertsen AMHJ, Johannesma PIM. The Spectro-Temporal Receptive Field. Biological Cybernetics 1981 Nov;42(2):133–143.

[16] Smith NJ, Kutas M. Regression-based estimation of ERP waveforms: I. The rERP framework. Psychophysiology 2015;52(2):157–168.

[17] Smith NJ, Kutas M. Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. Psychophysiology 2015;52(2):169–181.

[18] Holdgraf CR, Rieger JW, Micheli C, Martin S, Knight RT, Theunissen FE. Encoding and Decoding Models in Cognitive Electrophysiology. Frontiers in Systems Neuroscience 2017;11:61.

[19] Crosse MJ, Di Liberto GM, Bednar A, Lalor EC. The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience 2016;10:604.

[20] Gonçalves NR, Whelan R, Foxe JJ, Lalor EC. Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. NeuroImage 2014;97:196 – 205.

[21] Martin S, Brunner P, Holdgraf C, Heinze HJ, Crone NE, Rieger J, et al. Decoding spectrotemporal features of overt and covert speech from the human cortex. Frontiers in Neuroengineering 2014;7:14.

[22] Lalor E, Foxe J. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience 2009;31(1):189–193.

[23] Ding N, Simon JZ. Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. Journal of Neuroscience 2013;33(13):5728–5735.

[24] Sassenhagen J. How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. Language, Cognition and Neuroscience 2019;34(4):474–490.

[25] Poline JB, Brett M. The general linear model and fMRI: Does love last forever? NeuroImage 2012;62(2):871–880. 20 YEARS OF fMRI.

[26] Cromwell JB, Terraza M. Multivariate tests for time series models. Sage; 1994.

[27] Noguchi Y, Inui K, Kakigi R. Temporal Dynamics of Neural Adaptation Effect in the Human Visual Ventral Stream. Journal of Neuroscience 2004;24(28):6283–6290.

[28] Foxe J, Snyder A. The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. Frontiers in Psychology 2011;2:154.

[29] VanRullen R. Perceptual cycles. Trends in cognitive sciences 2016;20(10):723–735.

[30] Baillet S. Magnetoencephalography for brain electrophysiology and imaging. Nature Neuroscience 2017;20(3):327–339.

[31] Iturrate I, Chavarriaga R, Montesano L, Minguez J, Millán J. Latency correction of event-related potentials between different experimental protocols. Journal of neural engineering 2014;11(3):036005.

[32] Huth AG, De Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 2016;532(7600):453–458.

[33] Vigário R, Jousmäki V, Hämäläinen M, Hari R, Oja E. Independent Component Analysis for Identification of Artifacts in Magnetoencephalographic Recordings. In: Jordan M, Kearns M, Solla S, editors. Advances in Neural Information Processing Systems, vol. 10 MIT Press; 1998. p. 229–235.

[34] Uusitalo M, Ilmoniemi R. Signal-space projection method for separating MEG or EEG into components. Medical and biological engineering and computing 1997 04;35(2):135–40.

[35] Barachant A, Bonnet S, Congedo M, Jutten C. Multiclass brain–computer interface classification by Riemannian geometry. IEEE Transactions on Biomedical Engineering 2011;59(4):920–928.

[36] Sabbagh D, Ablin P, Varoquaux G, Gramfort A, Engemann DA. Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states. NeuroImage 2020;222:116893.

[37] Parra LC, Spence CD, Gerson AD, Sajda P. Recipes for the linear analysis of EEG. NeuroImage 2005;28(2):326 – 341.

[38] de Cheveigné A, Di Liberto GM, Arzounian D, Wong DD, Hjortkjær J, Fuglsang S, et al. Multiway canonical correlation analysis of brain data. NeuroImage 2019;186:728–740.

[39] Xu H, Lorbert A, Ramadge PJ, Guntupalli JS, Haxby JV. Regularized hyperalignment of multi-set fMRI data. In: 2012 IEEE Statistical Signal Processing Workshop (SSP) IEEE; 2012. p. 229–232.

[40] Lim M, Ales J, Cottereau B, Hastie T, Norcia A. Sparse EEG/MEG source estimation via a group lasso. PLOS ONE 2017 06;12:e0176835.

[41] King JR, Charton F, Lopez-Paz D, Oquab M. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. NeuroImage 2020;220:117028.

[42] Bazeille T, Dupre E, Richard H, Poline JB, Thirion B. An empirical evaluation of functional alignment using inter-subject decoding. Neuroimage 2020;.

[43] Richard H, Gresele L, Hyvarinen A, Thirion B, Gramfort A, Ablin P. Modeling Shared responses in Neuroimaging Studies through MultiView ICA. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems, vol. 33 Curran Associates, Inc.; 2020. p. 19149–19162.

[44] Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, Nichols TE. Guidelines for reporting an fMRI study. Neuroimage 2008;40(2):409–414.

[45] Benjamin AS, Fernandes HL, Tomlinson T, Ramkumar P, VerSteeg C, Chowdhury RH, et al. Modern Machine Learning as a Benchmark for Fitting Neural Responses. Frontiers in Computational Neuroscience 2018;12:56.

[46] Ivanova AA, Schrimpf M, Anzellotti S, Zaslavsky N, Fedorenko E, Isik L. Is it that simple? Linear mapping models in cognitive neuroscience. bioRxiv 2021;.

[47] James G. An introduction to statistical learning : with applications in R. New York, NY: Springer; 2013.

[48] Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of modern Physics 1993;65(2):413.

[49] Wilson HR, Cowan JD. Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical journal 1972;12(1):1–24.

[50] Friston KJ, Harrison L, Penny W. Dynamic causal modelling. Neuroimage 2003;19(4):1273–1302.

[51] Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. Siam Review 2018;60(2):223–311.

[52] Billings SA. Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. Wiley; 2013.

[53] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation 1997;9(8):1735–1780.

[54] Défossez A, Zeghidour N, Usunier N, Bottou L, Bach F. SING: Symbol-to-Instrument Neural Generator. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31 Curran Associates, Inc.; 2018.p. 9041–9051.

[55] Masci J, Meier U, Ciresan DC, Schmidhuber J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In: ICANN; 2011. .

[56] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–507.

[57] Koutnik J, Greff K, Gomez F, Schmidhuber J. A Clockwork RNN. In: Xing EP, Jebara T, editors. Proceedings of the 31st International Conference on Machine Learning, vol. 32 of Proceedings of Machine Learning Research Bejing, China: PMLR; 2014. p. 1863–1871.

[58] Bengio S, Vinyals O, Jaitly N, Shazeer N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 28 Curran Associates, Inc.; 2015. p. 1171–1179.

[59] Jung TP, Makeig S, McKeown MJ, Bell AJ, Lee TW, Sejnowski TJ. Imaging brain dynamics using independent component analysis. Proceedings of the IEEE 2001;89(7):1107–1122.

[60] de Cheveigné A, Wong DD, Di Liberto GM, Hjortkjaer J, Slaney M, Lalor E. Decoding the auditory brain with canonical component analysis. NeuroImage 2018;172:206–216.

[61] Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature 2008;452:352.

[62] Güçlü U, Gerven MV. Modeling the Dynamics of Human Brain Activity with Recurrent Neural Networks. Frontiers in Computational Neuroscience 2017;11.

[63] Breiman L. Random Forests. Machine Learning 2001 Oct;45(1):5–32.

[64] Kingma D, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations 2014 12;.

[65] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32 Curran Associates, Inc.; 2019.p. 8024–8035.

[66] Chen PHC, Chen J, Yeshurun Y, Hasson U, Haxby J, Ramadge PJ. A Reduced-Dimension fMRI Shared Response Model. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 28 Curran Associates, Inc.; 2015. .

[67] Zhang H, Chen PH, Ramadge P. Transfer learning on fMRI datasets. In: International Conference on Artificial Intelligence and Statistics PMLR; 2018. p. 595–603.

[68] Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, et al. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. Neuron 2011;72.

[69] Sahani M, Linden JF. How Linear are Auditory Cortical Responses? In: NIPS; 2002. .

[70] Kay KN, Winawer JA, Mezer AA, Wandell BA. Compressive spatial summation in human visual cortex. Journal of neurophysiology 2013;110 2:481–94.

[71] Schoffelen JM, Oostenveld R, Lam N, Udden J, Hulten A, Hagoort P. A 204-subject multimodal neuroimaging dataset to study language processing. Scientific Data 2019 12;6.

[72] Speer R, Chin J, Lin A, Jewett S, Nathan L. LuminosoInsight/wordfreq: v2.2; 2018, https://github.com/LuminosoInsight/wordfreq.

[73] Tallon-Baudry C, Bertrand O. Oscillatory gamma activity in humans and its role in object representation. Trends in Cognitive Sciences 1999;3(4):151–162.

[74] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MEG and EEG data analysis with MNE-Python. Frontiers in neuroscience 2013;7:267.

[75] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research 2011;12:2825–2830.

[76] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 IEEE International Conference on Computer Vision (ICCV) 2015;p. 1026–1034.

[77] Haegens S, Golumbic EZ. Rhythmic facilitation of sensory processing: A critical review. Neuroscience & Biobehavioral Reviews 2018;86:150–165.

[78] Lipton ZC. The mythos of model interpretability. Queue 2018;16(3):31–57.

[79] Ivanova A, Schrimpf M, Isik L, Anzellotti S, Zaslavsky N, Fedorenko E. Is it that simple? The use of linear models in cognitive neuroscience; 2020, workshop Proposal.

[80] Federmeier K, Kutas M. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. Journal of Memory and Language 1999 11;41(4):469–495.

[81] Fedorenko E, Scott TL, Brunner P, Coon WG, Pritchett B, Schalk G, et al. Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences 2016;113(41):E6256–E6262.

[82] He T, Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage 2020;206:116276.

[83] Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. Journal of neural engineering 2019;16(5):051001.

[84] Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nature Neuroscience 2019;22(11):1761–1770.

[85] Güçlü U, van Gerven M. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience 2015 07;35:10005–10014.

[86] Kuzovkin I, Vicente R, Petton M, Lachaux JP, Baciu M, Kahane P, et al. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Communications Biology 2018 12;1.

[87] Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias A, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Computational Biology 2019;15.

[88] Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage 2017;152:184–194.

[89] Klindt D, Ecker AS, Euler T, Bethge M. Neural system identification for large populations separating "what" and "where". In: Advances in Neural Information Processing Systems 31; 2017. .

[90] Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, et al. Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences 2021;118(3).

[91] Jain S, Huth A. Incorporating Context into Language Encoding Models for fMRI. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada; 2018. p. 6629–6638.

[92] Kriegeskorte N, Golan T. Neural network models and deep learning. Current Biology 2019;29(7):R231–R236.

[93] Sun P, Anumanchipalli GK, Chang E. Brain2Char: A Deep Architecture for Decoding Text from Brain Recordings. Journal of Neural Engineering 2020;.

[94] Güçlütürk Y, Güçlü U, Seeliger K, Bosch S, van Lier R, van Gerven M. Deep adversarial neural decoding; 2017, arXiv Preprint 1705.07109.

[95] Pinho AL, Amadon A, Ruest T, Fabre M, Dohmatob E, Denghien I, et al. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. Sci Data 2018;5:180105.

[96] Caucheteux C, King JR. Language processing in brains and deep neural networks: computational convergence and its limits. bioRxiv 2020;.

[97] Schrimpf M, Blank I, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, et al. Artificial Neural Networks Accurately Predict Language Processing in the Brain. bioRxiv 2020;.

[98] Toneva M, Wehbe L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 32 Curran Associates, Inc.; 2019. .

[99] Ambrogioni L, van Gerven MAJ, Maris E. Dynamic decomposition of spatiotemporal neural signals. PLOS Computational Biology 2017 05;13(5):1–37.

[100] Vidaurre D, Quinn AJ, Baker AP, Dupret D, Tejero-Cantero A, Woolrich MW. Spectrally resolved fast transient brain states in electrophysiological data. NeuroImage 2016;126:81–95.

[101] van Gerven MAJ. A primer on encoding models in sensory neuroscience. Journal of Mathematical Psychology 2017;76:172–183. Model-based Cognitive Neuroscience.

[102] Schoenmakers S, Barth M, Heskes T, van Gerven M. Linear reconstruction of perceived images from human brain activity. NeuroImage 2013;83:951–961.

[103] Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. NeuroImage 2013 10;86.

[104] Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering 2007;9(3):90–95.

[105] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020 Sep;585(7825):357–362.

[106] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020;17:261–272.

5 | APPENDIX

Figure 5 reports an alternative scoring of the main models using the R^2 score instead of the Pearson correlation R. It is interpretable as explained variance is a quantity upper bounded by 1, as opposed to MSE which depends on the input scaling of the data.

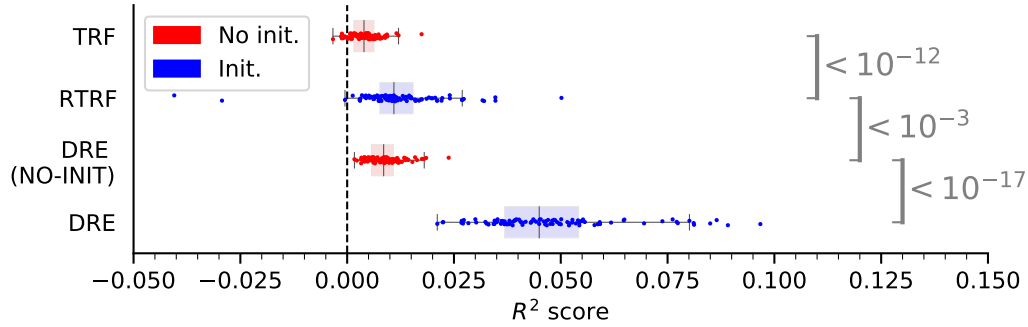

FIGURE 5 Explained Variance captured by coefficient of determination (R^2) for each model. Boxplots indicate median and interquartile range, while each point corresponds to a single subject. The gray brackets indicate the p-value obtained with a pairwise Wilcoxon comparison across subjects. Initialization and non-linearity increase predictive performance.

Due to the very small SNR of single trial and unaveraged MEG recordings, we obtain low values as expected. Indeed, some variance in the signal is explained by noise only, whose amplitude is on a scale 10 times larger that of the evoked response. However, the ordering of model performance and the conclusions are left unchanged.

Note on convolution and computational efficiency

The introduction of convolutional layers is here mainly motivated by computational efficiency: convolutional layers reduced training time on an NVIDIA V100 GPU from 2.6 h for a DRE devoid of convolutional layers down to 1.4 h for our DRE. The performance between these two models is relatively similar, although with a slight benefit in favour of DRE:(NO-CONV: $0.194 \pm 0.003\%$; DRE: $0.197 \pm 0.003\%$, $p < 10^{-5}$).

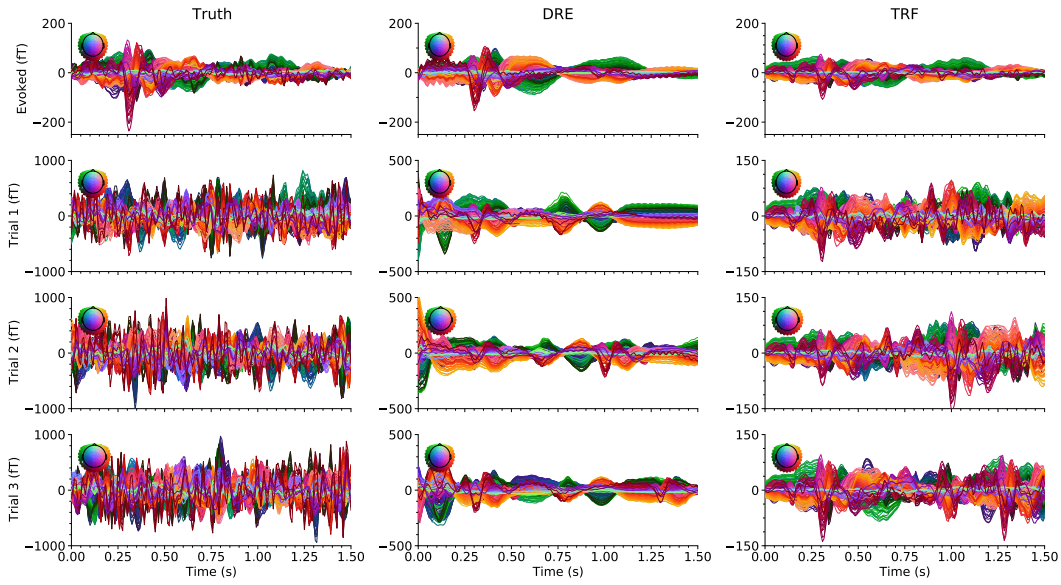

FIGURE 6 True and predicted brain response (averaged “Evoked" (top row) or single-trial) to visual stimuli. These qualitative results illustrate that individual MEG responses vary along many different aspects, including the amplitude, latency and overall shape of the evoked-related fields. These variations are particularly challenging for TRF to learn. The predictions of TRF illustrated here also highlight that this model appears to modulate the MEG response but typically outputs predictions with reduced amplitudes – an effect likely due to a reduction-to-the-mean phenomenon: *i.e.* when the model fails to learn the MEG dynamics, it converges its coefficients towards 0.