Generalizing Brain Decoding Across Subjects with Deep Learning

Richard Csaky[1,2,5], **Mats W.J. van Es**[1,2], **Oiwi Parker Jones**[2,3,4], and **Mark Woolrich**[1,2]

[1]Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford
[2]Wellcome Centre for Integrative Neuroimaging, University of Oxford
[3]Department of Engineering Science, University of Oxford
[4]Jesus College, University of Oxford
[5]Christ Church, University of Oxford
{richard.csaky@psych, mats.vanes@psych, oiwi.parkerjones@eng, mark.woolrich@ohba}.ox.ac.uk

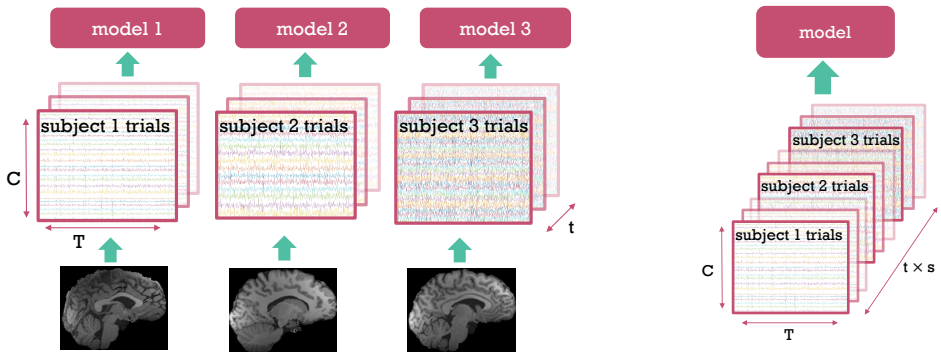
Abstract

Decoding experimental variables from brain imaging data is gaining popularity, with applications in brain-computer interfaces and the study of neural representations. Decoding is typically subject-specific and does not generalise well over subjects. Here, we investigate ways to achieve cross-subject decoding. We used magnetoencephalography (MEG) data where 15 subjects viewed 118 different images, with 30 examples per image. Training on the entire 1s window following the presentation of each image, we experimented with an adaptation of the WaveNet architecture for classification. We also investigated the use of subject embedding to aid learning of subject variability in the group model. We show that deep learning and subject embedding are crucial to closing the performance gap between subject and group-level models. Importantly group models outperform subject models when tested on an unseen subject with little available data. The potential of such group modelling is even higher with bigger datasets. Furthermore, we demonstrate the use of permutation feature importance to gain insight into the spatio-temporal and spectral information encoded in the models, enabling better physiological interpretation. All experimental code is available at `https://github.com/ricsinaruto/MEG-group-decode`.

1 Introduction

In recent years, decoding has gained in popularity in neuroscience (Kay et al., 2008), specifically decoding external variables (e.g. stimulus category) from internal states (i.e. brain activity). Such analyses can be useful for brain-computer interface (BCI) applications (Willett et al., 2021) or for gaining neuroscientific insights (Guggenmos et al., 2018; Kay et al., 2008). Analysing deep learning methods on such data is also beneficial for the machine learning community. Namely, the small, noisy, high-dimensional datasets test the limits of popular architectures on real data and demand research into new methods. Applications of decoding to brain recordings typically fit separate (often linear) models on a per dataset, per subject basis (Guggenmos et al., 2018). This has the benefit that the decoding is tuned to the dataset/subject, but has the drawback that it is unable to leverage knowledge that could be transferred across datasets/subjects. Practical drawbacks of subject-specific (subject-level) models include increased computational load, a higher chance of overfitting, and the inability to adapt to new subjects. We aim to leverage data from multiple subjects and train a shared model that can generalise across subjects (group-level). A conceptual visualisation of subject-level (SL) and group-level (GL) models is given in Figure 1.

Preprint. Under review.



(a) Subject-level models.

(b) Group-level model.

Figure 1: Comparison of subject and group-level modelling. (a) A separate model is trained on the trials (examples) of each subject. (b) A single, shared model is trained on the trials of all subjects. Each trial is **C** x **T** (channels x timesteps) dimensional. Each of the **s** subjects has **t** trials.

Magnetoencephalography (MEG) scanners measure the magnetic fields induced by electrical activity in the brain. MEG is one of the main non-invasive brain recording methodologies, alongside electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI). MEG scanners produce a time series of magnetic fields usually sampled at 1000 Hz across hundreds of sensors (channels) providing full coverage of the brain. Among other methods, neuroimaging data can be analysed in terms of decoding (Du et al., 2019), often involving machine learning. For MEG specifically (the focus of this paper) input examples (trials) are **C** x **T** (channels x timesteps/time samples) dimensional. Most decoding research focuses on SL models. This is because between-subject variability of neuroimaging data limits the application of a single, shared model across subjects (Olivetti et al., 2014; Li et al., 2021), which we will refer to as naive group modelling. This variability has multiple sources, such as different anatomical structures, different positions in the scanner, signal-to-noise ratio, etc. (Saha and Baumert, 2020). We propose a general architecture capable of jointly decoding multiple subjects with the help of subject embeddings.

We make the following contributions using a MEG dataset with visual task (Cichy et al., 2016): 1. A deep learning-based group model using subject embeddings is introduced, which improves significantly over naive group modelling. 2. Insight is provided into how nonlinearity and subject embedding helps group modelling. 3. We show how GL models generalise much better to new subjects than SL models. 4. We show how neuroscientific insights can be gained from a deep learning-based group model. 5. Analysis of model weights reveals how meaningful spatio-temporal and spectral information is encoded.

2 Related Work

Decoding can be applied to most tasks/modalities, such as images (Cichy et al., 2016), phonemes (Mugler et al., 2014), words (Cooney et al., 2019b; Hultén et al., 2021), sentences (Dash et al., 2020a), and motor movements like imagined handwriting (Willett et al., 2021), jaw movements (Dash et al., 2020b), or finger movements (Elango et al., 2017). Chaibub Neto et al. (2019) make a strong argument about the methodological issues with SL modelling, which should ideally be avoided in the case of clinical applications. Recently some transfer learning approaches have been proposed to deal with the between-subject variability problem. These frameworks consist of applying a model trained on one subject to a different (target) subject (Elango et al., 2017; Dash et al., 2019; Cooney et al., 2019a; Olivetti et al., 2014; Halme and Parkkonen, 2018; Li et al., 2021). Some approaches use learnable affine transformations between subjects (Elango et al., 2017), while others finetune the whole model on the target subjects (Cooney et al., 2019a; Dash et al., 2019). However, these approaches are limited since they do not use a shared model across multiple subjects, and they offer a marginal improvement over the naive group modelling approach. In this work, we aim to significantly improve on this using a general framework and model capable of decoding multiple subjects.

Transfer learning is popular in the wider machine learning field. Parallels can be drawn with domain adaptation (Long et al., 2015), or transferring knowledge from large to small datasets within the

same domain (Wang et al., 2019; Zhuang et al., 2020). Natural language processing (NLP) datasets often contain data from widely different sources, but due to the sheer dataset size, models trained on the joint data achieve good results (Brown et al., 2020; Devlin et al., 2019). As discussed before, this naive concatenation of subjects does not work well on small neuroimaging datasets. Perhaps the most relevant parallels can be drawn with dialogue modelling work, which models inter-speaker differences using speaker embeddings (Li et al., 2016; Zhang et al., 2018). Such embeddings are similarly useful in speech applications (Saito et al., 2019; Mridha et al., 2021).

3 Methods

3.1 Data

In this work, a task-MEG dataset is used where 15 subjects view 118 different images, with each image viewed 30 times (Cichy et al., 2016). This dataset has been collected with appropriate consent from participants and ethical review by Cichy et al. (2016). The data is publicly available[1], however, we obtained the continuous raw MEG data directly from the authors to be able to run our preprocessing pipeline. Raw data is bandpass filtered between 0.1 and 125 Hz and line noise is removed with notch filters at multiples of 50 Hz. After downsampling to 250 Hz, 1.024-second epochs are extracted starting 100 ms before stimulus presentation. This resulted in 306 x 256 dimensional trials (channels x timesteps) of preprocessed data from the 306 MEG sensors. Whitening is used to remove covariance between channels for SL models, whereas for GL models a standardisation is performed per channel. We do multiclass decoding, predicting a separate probability for each of the 118 classes (images).

3.2 Models

Our deep learning model, WaveNet Classifier, is inspired by previous approaches to apply WaveNet (van den Oord et al., 2016) for classification (Zhang et al., 2020). The model consists of 2 parts: the convolutional block, intended to act as a feature extractor; and the fully-connected block, which is designed for classification (Figure 2). The convolutional block uses a stack of 1D dilated convolutional layers, which include dropout and the inverse hyperbolic sine activation function. The dilated convolutions in WaveNet are effective for modelling time series data as successive layers extracts complementary frequency content of the input (Borovykh et al., 2018). Since the dilation factor is doubled in successive layers, the receptive field of the convolutional block is 2^{num_layers}. This large receptive field is achieved with far fewer parameters than in standard convolutional neural networks. Given there is no pooling and a convolution stride of 1, the output of each layer preserves the temporal dimensionality[2]. At the end of the convolutional block, we downsample temporally by the size of the receptive field. In the model with 6 convolutional layers, this means that the initial input of size 256 is downsampled by a factor of 64, resulting in 4 values per channel. Next, this output is flattened and fed into a fully-connected block. The final output is a logit vector corresponding to the 118 classes.

For SL modelling the Wavenet Classifier contains 3 convolutional layers, whereas for group modelling it has 6. This is further motivated in Section 4. Our improvement of the naive group model includes subject embeddings which are introduced as a way of dealing with between-subject variability. Similar to word embeddings in NLP, each subject has a corresponding dense vector (Mikolov et al., 2013). This vector is concatenated with the channel dimension of the input trial across all timesteps. Thus, in the embedding-aided GL model input trials are (**C**+**E**) x **T** dimensional, where **E** is the embedding size. Subject embeddings are learned together with other model weights using backpropagation. We reason that the model aided by embeddings can learn general features across subjects, with the capability of adapting its internal representations for each subject.

3.3 Model analysis

In this section, we describe several approaches to uncover the information encoded in the WaveNet Classifier. In *Kernel FIR Analysis*, we investigate the frequency characteristics of the convolutional kernels. Random noise is fed into a trained model, and the power spectral density of the output of specific kernels is computed to assess their finite impulse response (FIR) properties. Permutation

[1] `http://userpage.fu-berlin.de/rmcichy/fusion_project_page/main.html`

[2] Except the amount that gets chopped off because of the kernel size itself, since we do not use padding.

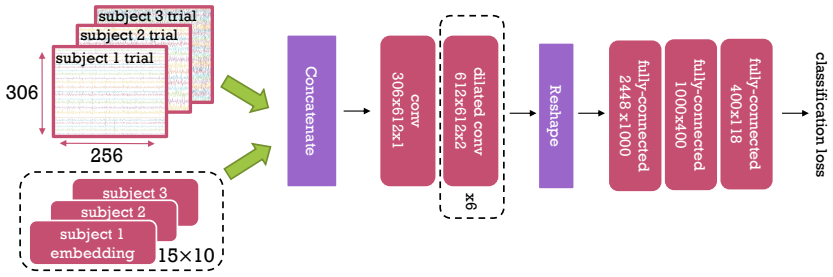


Figure 2: Group-level WaveNet Classifier with subject embeddings. Dashed boxes represent parts of the model which differ between SL and GL versions of our architecture. Red boxes represent learnable parameters. For convolutional layers *input channels* x *output channels* x *kernel size* is shown. For fully-connected layers *input neurons* x *output neurons* is shown. The embedding layer dimensionality is given as **s** x **E**, where **s** is number of subjects, and **E** is the embedding size. Embeddings are concatenated with input trials to provide information about which trial is coming from which subject.

Feature Importance (PFI) is a powerful method to assess which features contribute the most toward model performance (Altmann et al., 2010; Chehab et al., 2021). For MEG data we can assess both spatial and temporal information by permuting across timesteps (for each channel) and across channels (for each timestep), respectively. We call these temporal and spatial PFI respectively. A decrease from the original accuracy (with unpermuted inputs) indicates that stimulus-related information is present in the MEG data in certain time periods or sensors. Alternatively, when looking at individual kernels our feature importance measure is the difference between the kernel's output using the original and permuted inputs. We reason that a more important feature will cause a higher output deviation. For assessing the frequency sensitivity of individual kernels, we introduce spectral PFI. First, the data in each channel of each trial is Fourier transformed, and the Fourier coefficients are shuffled across channels for each frequency (or frequency band). Then, the inverse Fourier transform is computed, obtaining a trial with disrupted information in specific frequency bands.

4 Results

4.1 Group models using subject embedding achieve similar performance to SL models

Train and validation splits with a 4:1 ratio, respectively, were constructed for each subject and class. The first 20% of the continuous MEG data is used to extract validation trials, with the remaining used for training trials. Both SL and GL models are trained and evaluated on the same splits. For each model, an extra training is conducted wherein the (linear) identity function is used as an activation function to assess how nonlinearity, which is the bedrock of deep learning, influences the results. Linear and nonlinear models are trained for 500 and 2000 epochs (full passes of the training data), respectively. We observed that nonlinear models' validation accuracy tends to plateau, whereas linear models overfit after 500 epochs, hence the different number of epochs (see Section 4.2 for details). Dropout was set to 0.4 and 0.7, and a batch size of 590 and 59 was used for GL and SL models, respectively. The Adam optimiser (Kingma and Ba, 2015) was used with a learning rate of 0.0001 for GL, and 0.00005 for SL models. We compute paired samples T-tests for comparisons of interest, where the samples are the subject-level mean accuracies over validation trials. Training of a single SL and GL model took 5-15 minutes and 4 hours on an NVIDIA A100 GPU, respectively. We used PyTorch for training (Paszke et al., 2019); MNE for preprocessing MEG data (Gramfort et al., 2013); scikit-learn (Pedregosa et al., 2011), SciPy (Virtanen et al., 2020), and NumPy (Harris et al., 2020) for analysis; and pandas (Wes McKinney, 2010), seaborn (Waskom, 2021), and matplotlib (Hunter, 2007) for visualisation. All packages use some form of the BSD license.

Validation accuracies for all models are shown in Figure 3. Interestingly, at the subject level, linear models performed slightly better than nonlinear (4% increase, $p = 5.7e-4$). We think that both the limit in data size and noise levels in the data contribute to the subpar performance of nonlinear models compared to simpler linear models. This is an important lesson for applying deep learning on such data. Another common feature of MEG datasets is the large between-subject variability, with individual subjects' accuracy ranging from 5% to 88%. As expected, a naive application of either

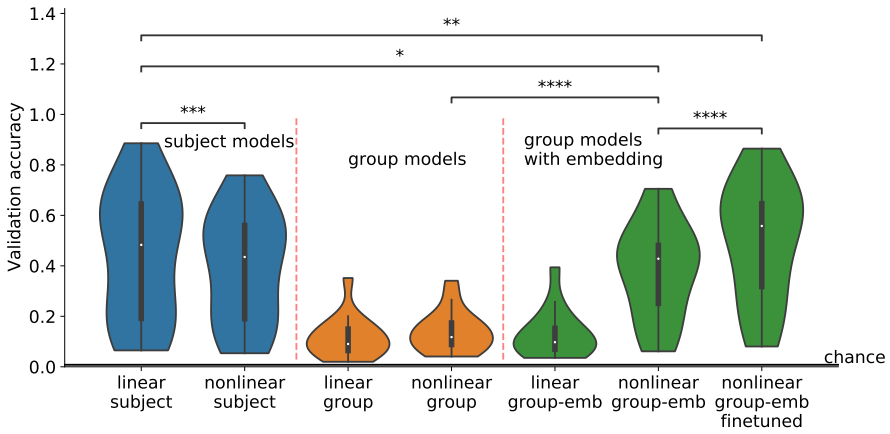


Figure 3: SL and GL models evaluated on the validation set of each subject. Paired samples T-tests are shown for comparisons of interest ($* = p < 5e{-}2, ** = p < 1e{-}2, *** = p < 1e{-}3, **** = p < 1e-4$). The `nonlinear group-emb finetuned` model is finetuned separately on each subject, initialized with the `nonlinear group-emb` model. Chance level is $1/118$.

the linear or nonlinear WaveNet Classifier to the group modelling problem results in much worse performance than SL models (30% decrease). As discussed, inferring variability between subjects implicitly is a difficult task, especially in a dataset with few subjects and high variability. Adding subject embeddings to the nonlinear model improves performance by 24% ($p = 1.9e-6$), with no increase for the linear model. This shows that leveraging subject embeddings can narrow the gap with SL models (6% difference, $p = 1.3e-2$), but only when using nonlinear activation functions.

We also finetuned the embedding-aided GL model on the training data of each subject separately. This results in SL models initialised with the group model. Interestingly this finetuning approach improves over training SL models from scratch (5% increase, $p = 1.1e-3$). This shows that using representations learned at the group level is useful for SL modelling. The variance of `nonlinear group-emb` (0.19) and `nonlinear group-emb finetuned` (0.24) is lower than the SL models (0.26). When looking at individual subjects in the case of `nonlinear group-emb`, 4 subjects with generally low accuracies (15-30%) had higher accuracies than SL models (even though the mean across subjects is lower). In the case of `nonlinear group-emb finetuned`, only 2 high-accuracy subjects were slightly worse than SL models, and generally low/mid-accuracy subjects gained more accuracy than high-accuracy subjects. This shows that the group model is reducing between-subject variability and this effect is to some extent even preserved in the finetuned SL models.

4.2 Insights into subject embeddings and other modelling choices

Number of layers In nonlinear SL models performance improves as we use fewer convolutional layers. This effect plateaus at 2-3 layers. In nonlinear GL models we noticed the opposite relationship between the number of layers and accuracy. Since more convolutional layers result in higher temporal downsampling, GL models have more weights in the convolutional block and fewer weights in the fully-connected block compared to SL models. This shows that SL models must rely more on the fully-connected block as they are unable to extract good features, and GL models rely more on the convolutional block to learn shared features across subjects. The `nonlinear group-emb finetuned` models use the same number of layers and nonlinearity as the group model, but achieve higher accuracy than SL models. This shows that when initialised well (with a group model trained on multiple subjects) even SL models can benefit from nonlinearity and more convolutional layers.

Subject embeddings We tried different approaches to gain insight into how subject embeddings help the group model. A clustering or 2D projection of the embedding space such as PCA or t-SNE (Van der Maaten and Hinton, 2008) did not show any clusters. This can be a consequence of only having 15 subjects, compared to thousands of words in the case of word embeddings where such visualisations work well (Liu et al., 2017). To assess whether the embeddings just encode which subjects are good or bad, we transformed the embeddings with PCA and correlated all components

with the accuracies across subjects. We found no significant correlations; thus embeddings do not appear to encode information about SL accuracy. To assess how much embeddings contribute to a trained model, we tried both setting the embeddings to zero and shuffling them. This reduced accuracy to 10% (from 38%). Thus, embeddings encode crucial information to aid decoding, but the model is still better than chance without them. In addition to our general results with **E**=10, we trained group models with **E**=3 and **E**=14, achieving 20% and 38% accuracy, respectively. Considering there are 15 subjects, compressing the embedding representations too much is not possible. This is in line with the clustering analysis above and again can be due to having few subjects (further insights in Appendix).

Linearity For linear models, validation losses (cross-entropy) and accuracies were negatively correlated, i.e. loss decreases while accuracy increases, and eventually both suggested overfitting. Since nonlinear models are more expressive, they overfitted sooner according to the loss, but accuracy kept improving until it reached a plateau, never overfitting. Analysing the loss distribution across validation examples (for nonlinear models) shows that even during overfitting most examples' loss keeps decreasing with a few high-loss outliers disproportionately influencing the mean. Since accuracy is binary, outliers are diminished, explaining the apparent difference in learning behaviour. For linear models, this unintuitive behaviour was not observed probably due to inherent model simplicity. Nonlinearity is crucial to leverage subject embeddings (Section 4.1). Limiting the nonlinear activation to the first layer resulted in a subpar performance similar to that of a linear model. This indicates that nonlinearity is needed within multiple layers to benefit from subject embeddings.

4.3 Group models generalise much better to new subjects than SL models

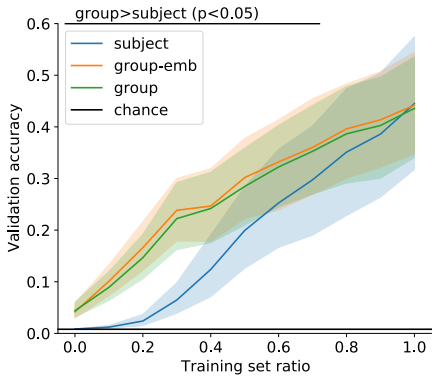
While having a shared model across multiple subjects is certainly useful, a prominent use case in neuroscience is when data is collected for a new subject, with some other subjects' data already available. In such cases, it would be useful to apply a model to the new subject with limited training. Here we evaluated how a GL model trained on the training data of 14 subjects performs on the validation data of the left-out subject. To analyse the effect of training set size, the GL model is further trained (finetuned) across different training set partitions of the left-out subject. This is repeated across all subjects, so the results without finetuning can also be interpreted as cross-validation across subjects. We compared the nonlinear group and group-embedding models with an SL model trained only on the left-out subject. When finetuning the group-embedding model, the left-out subject's embedding was initialised randomly.

Figure 4a shows that initialising with a GL model is much better than training on the left-out subject from scratch with little training available. This is significant ($p < 0.05$, corrected for multiple comparisons), up to the case when 70% of the training set is used. Group models are above chance (5%) even when no training is performed on the left-out subject. This shows that GL models can extract features that are generalisable to new subjects, which could be especially useful for BCI applications. Unsurprisingly, the embedding-aided group model is similar to the naive group model. Embeddings truly help when we have a single model over multiple subjects, whereas here a separate model is trained for each new subject. This could change when looking at datasets with more subjects, where between-subject similarities could be reflected in the embeddings.

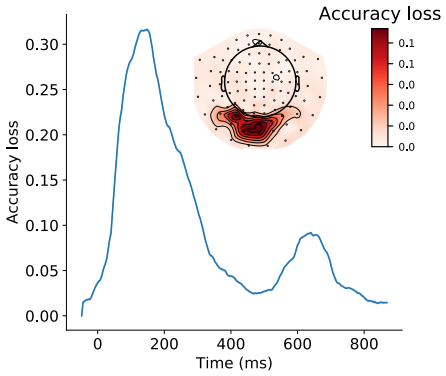
Models are similar when all training data is used. We think this effect is due to the limited dataset size (30 examples per class). In such low data regimes, the validation set (6 examples per class) is probably not representative of the full data distribution. Thus, we think that all models performed the same because of the ratio of training and validation data, and the low number of examples. We expect that with more trials, the gap between group initialisation and training from scratch would continue. Results should be interpreted relatively rather than absolutely, i.e. the models converge when 100% of training data is used (with this specific train/validation ratio), rather than the models converge when 24 training examples per class are used. This reasoning would only change with orders of magnitude more training and validation data, as is the case in many machine learning datasets.

4.4 Neuroscientific spatio-temporal insights are gained from a deep learning based GL model

An established critique of deep learning models applied to neuroimaging data is the lack of interpretable insight they provide about the underlying neural processes that drive the decoding. To gain such neuroscientific insights it is useful to assess the time and space-resolved informa-



(a) Generalisation and finetuning on left-out subjects. The horizontal axis shows the amount of training data used for finetuning. `subject` is trained from scratch, while `group-emb` and `group` is initialised with the nonlinear GL model with and without embeddings, respectively. Shading shows the 95% confidence interval of the accuracy (vertical axis) across left-out subjects.



(b) Temporal (line) and spatial (sensor space map) PFI for the `nonlinear group-emb` model. For temporal PFI accuracy loss (vertical axis) is plotted with respect to time since image presentation (horizontal axis). Shading shows the 95% confidence interval. Due to small variability across repetitions, this is not visible. For spatial PFI, darker red shading is equivalent to higher accuracy loss.

Figure 4: Generalization performance (a), Temporal and Spatial PFI (b).

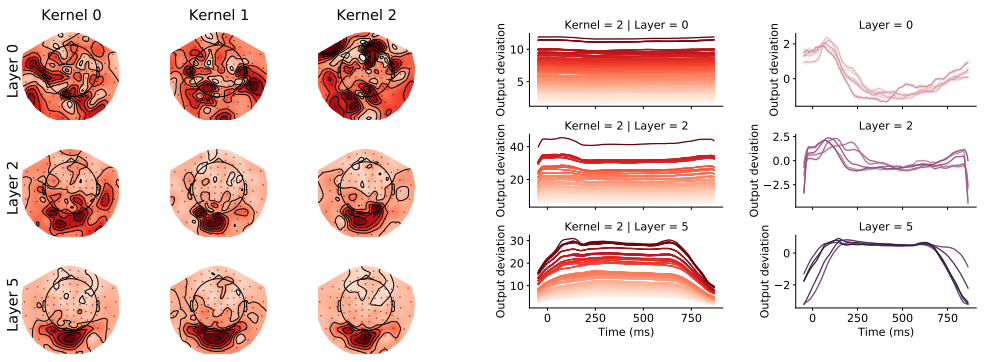
tion/discriminability within trials. As mentioned in Section 3.3, permutation feature importance (PFI) is a suitable, model-agnostic measure. Figure 4b shows the temporal and spatial PFI of the nonlinear group-embedding model. To make the results robust and smooth, the shuffling for temporal PFI was applied to 100 ms windows, and magnetometers and gradiometers in the same location were shuffled together for spatial PFI.

Time windows or channels with higher accuracy loss than others are interpreted as containing more information about the neural discriminability of images. This indicates when and where information processing related to the presented images is happening in the brain. Temporal PFI shows a large peak around 150 ms which is often observed in neuroscientific decoding studies employing sliding window analysis to assess temporal information content (Higgins et al., 2022). After this, information content rapidly decreases, with a second, smaller peak around 650 ms. This could signal a brain response following the end of image presentation at 500 ms. Spatial PFI shows that the most important channels are in the back of the head at sensors over visual areas, which is expected for a visual task.

4.5 Model weights encode meaningful spatio-temporal and spectral information

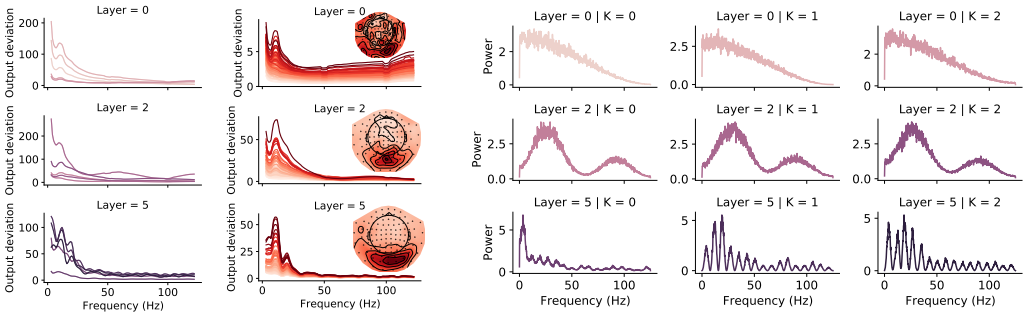
Deep learning faces difficulties when interpretability comes into question (Murdoch et al., 2019). Here we show that neuroscientifically interpretable spatial, temporal, and spectral information can be gained by analysing a trained model's weights All visualisations are from 3 convolutional layers of the `nonlinear group-emb` model, with all 6 layers shown in the Appendix. Kernels within a layer seem to have similar temporal sensitivity, even though only 5 are shown from over 1e5 total kernels (Figure 5c). Output deviations are standardised to compare temporal PFI across kernels with different output magnitudes. In early layers sensitivity peaks around 100 ms (as in Section 4b), then rapidly decreases, eventually climbing up again slowly. Kernels in the last layer have a flatter response to temporal disruptions, with sensitivity falling sharply at the start and end of the input trial.

Kernels in early layers have somewhat random spatial sensitivity (Figure 5a). In deeper layers, this gets narrowed down to channels over the visual cortex, with some differences between individual kernels within a layer. This sensitivity is similar to the spatial features that were shown to be most informative for classification performance (see Section 4.4). Figure 5b shows the temporal profile of the spatial PFI. This is achieved by focusing the shuffling to 100 ms time windows and 4-channel neighbourhoods (3 closest channels for each channel), repeated across all timesteps and channels. Spatial sensitivity does not seem to change with time, i.e. the same channels are always the most



(a) Spatial (b) Channel-wise temporal (c) Temporal

Figure 5: Spatio-temporal insights can be obtained using PFI. Spatial (a), channel-wise temporal (b), and temporal (c) PFI across `nonlinear group-emb` kernels within 3 layers (rows). For spatial PFI kernels are plotted separately, whereas for temporal PFI 5 kernels (lines) are plotted together. Channel-wise temporal PFI shows the temporal PFI of each channel for Kernel 2. Channel colouring is matched to the corresponding spatial PFI map, and darker reds mean higher output deviation. For temporal PFI output deviation is normalised. The horizontal axis shows the time elapsed since image presentation for both temporal PFI types. 95% confidence interval is shown with shading.



(a) Spectral PFI (b) Chn-wise spectral PFI (c) Kernel FIR analysis

Figure 6: Frequency sensitivity of kernels via spectral PFI (a), channel-wise spectral PFI (b), and frequency characteristics via kernel FIR analysis (c), from 3 layers (rows). Kernels are plotted together (lines) for spectral PFI, and in separate columns for kernel FIR analysis (normalised). Each channel-wise spectral PFI plot is for 1 kernel, where lines show the spectral PFI of corresponding channels in the sensor space map. 95% confidence interval is shown with shading for spectral PFI. Due to small variability across permutations, this is barely visible.

important across time. Generally, only channels with high sensitivity (visual area) have the temporal profile observed in temporal PFI, with other channels showing a flatter sensitivity.

Spectral PFI measures the change in kernel output to perturbations in specific frequency bands (Figure 6a). Band-width was set to 5 Hz to get a smooth frequency profile. Across all layers and kernels, the profile has a 1/f (frequency) shape with a clear peak at 10 Hz. These are both common features of the MEG signal (Demanuele et al., 2007; Drewes et al., 2022). This indicates that the spectral sensitivity of kernels coincides with the power spectra of the input examples. The spectral PFI of 4-channel neighbourhoods is shown in Figure 6b. Kernels are sensitive to the same channels (in the visual area) across all frequencies, with these channels having larger 10 Hz peaks.

Kernel FIR analysis shows the power spectra of kernels' outputs when input examples are Gaussian noise (Figure 6c). The subject embedding was set to a subject with average accuracy. The power spectra were normalised to make visual comparison across kernels easier. Since the WaveNet architecture uses dilated filters with only 2 values per filter, early layers show broad filtering characteristics, but already in layer 2 more emphasis is put on lower frequencies. As the model gets deeper, the filters (kernels) become more tuned to specific frequencies, generally below 20Hz. This is in line with

the spectral properties of MEG data as discussed above. Both spectral PFI and kernel FIR analysis shows that there is significant variability between the spectral information encoded by various kernels. From the analysis presented in this section, we can conclude that kernels are sensitive to interpretable temporal, spatial, and spectral features of the MEG data.

5 Discussion

We have proposed a deep learning-based GL model which improves significantly over naive GL modelling, achieving similar performance to SL models. We have shown how subject embeddings and nonlinearity are crucial for this. In addition, we have demonstrated that we can use PFI to obtain insights into which time points and channels contributed to the decoding and to obtain meaningful information encoded in convolutional kernels. We also found that when we used the GL model to initialise an SL model, it was able to outperform a randomly initialised SL model. However, we found that in this context, using subject embeddings did not improve performance. It may be that the outcome is different when applying our method to larger neuroimaging datasets with more subjects. Further research is needed into deep learning models capable of implicitly learning inter-subject variability. An important question is whether scaling up models on large datasets would achieve this goal, a phenomenon often observed in machine learning.

While our methods are general, it remains to be seen if similar results would be observed in other electrophysiological modalities (e.g. EEG), or even other task modalities. While PFI has shown impressive results, further work is needed to strengthen neuroscientific claims. PFI can easily be run on a per-subject or per-condition/class basis, which could provide further insight. This is left for future work, as well as applying PFI to the fully-connected block. Note that due to computational constraints we have not carried out a full analysis of different random seeds and cross-validation. From our experience with experiments in this work, we believe our main claims would not be affected by such small differences due to the reported big effect sizes and statistical significance.

Brain decoding studies have exciting future applications, however, as these technologies improve the inherent potential of negative societal effects increases. For example, malicious actors could build on it to extract personal information from brain data, violating privacy. In practice, such issues are mostly mitigated by needing ethical approval and consent for the collection of brain data and limiting access. This paper focuses on advancing noninvasive GL decoding, which is more anonymous than SL decoding, with less potential for harmful applications (Chaibub Neto et al., 2019).

Acknowledgments and Disclosure of Funding

RC is supported by a Wellcome Centre Integrative Neuroimaging Studentship. MVE's research is supported by the Wellcome Truest (215573/Z/19/Z). MW's research is supported by the NIHR Oxford Health Biomedical Research Centre, the Wellcome Trust (106183/Z/14/Z, 215573/Z/19/Z), the New Therapeutics in Alzheimer's Diseases (NTAD) study supported by UK MRC and the Dementia Platform UK (RG94383/RG89702) and the EU-project euSNN (MSCA-ITN H2020-860563). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Borovykh, A., Bohte, S., and Oosterlee, C. W. (2018). Dilated convolutional neural networks for time series forecasting. *Journal of Computational Finance, Forthcoming*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).

Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chaibub Neto, E., Pratap, A., Perumal, T. M., Tummalacherla, M., Snyder, P., Bot, B. M., Trister, A. D., Friend, S. H., Mangravite, L., and Omberg, L. (2019). Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ digital medicine*, 2(1):1–6.

Chehab, O., Defossez, A., Loiseau, J.-C., Gramfort, A., and King, J.-R. (2021). Deep Recurrent Encoder: A scalable end-to-end network to model brain signals. *arXiv preprint arXiv:2103.02339*.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):1–13.

Cooney, C., Folli, R., and Coyle, D. (2019a). Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*, pages 1311–1316. IEEE.

Cooney, C., Korik, A., Raffaella, F., and Coyle, D. (2019b). Classification of imagined spoken word-pairs using convolutional neural networks. In *The 8th Graz BCI Conference, 2019*, pages 338–343. Verlag der Technischen Universitat Graz.

Dash, D., Ferrari, P., and Wang, J. (2020a). Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Frontiers in Neuroscience*, 14:290.

Dash, D., Ferrari, P., and Wang, J. (2020b). Decoding Speech Evoked Jaw Motion from Non-invasive Neuromagnetic Oscillations. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Dash, D., Wisler, A., Ferrari, P., and Wang, J. (2019). Towards a Speaker Independent Speech-BCI Using Speaker Adaptation. In *INTERSPEECH*, pages 864–868.

Demanuele, C., James, C. J., and Sonuga-Barke, E. J. (2007). Distinguishing low frequency oscillations within the 1/f spectral behaviour of electromagnetic brain signals. *Behavioral and Brain Functions*, 3(1):1–14.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Drewes, J., Muschter, E., Zhu, W., and Melcher, D. (2022). Individual resting-state alpha peak frequency and within-trial changes in alpha peak frequency both predict visual dual-pulse segregation performance. *Cerebral Cortex*.

Du, C., Li, J., Huang, L., and He, H. (2019). Brain encoding and decoding in fMRI with bidirectional deep generative models. *Engineering*, 5(5):948–953.

Elango, V., Patel, A. N., Miller, K. J., and Gilja, V. (2017). Sequence transfer learning for neural decoding. *bioRxiv*, page 210732.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, page 267.

Guggenmos, M., Sterzer, P., and Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *Neuroimage*, 173:434–447.

Halme, H.-L. and Parkkonen, L. (2018). Across-subject offline decoding of motor imagery from MEG and EEG. *Scientific reports*, 8(1):1–12.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Higgins, C. J., van Es, M. W., Quinn, A. J., Vidaurre, D., and Woolrich, M. W. (2022). The relationship between frequency content and representational dynamics in the decoding of neurophysiological data. *bioRxiv*.

Hultén, A., van Vliet, M., Kivisaari, S., Lammi, L., Lindh-Knuutila, T., Faisal, A., and Salmelin, R. (2021). The neural representation of abstract words may arise through grounding word meaning in language itself. *Human brain mapping*, 42(15):4973–4984.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185):352–355.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016). A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003. Association for Computational Linguistics.

Li, J., Pan, J., Wang, F., and Yu, Z. (2021). Inter-Subject MEG Decoding for Visual Information with Hybrid Gated Recurrent Network. *Applied Sciences*, 11(3):1215.

Liu, S., Bremer, P.-T., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., and Pascucci, V. (2017). Visual exploration of semantic relationships in neural word embeddings. *IEEE transactions on visualization and computer graphics*, 24(1):553–562.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mridha, M. F., Ohi, A. Q., Monowar, M. M., Hamid, M., Islam, M., Watanobe, Y., et al. (2021). U-vectors: Generating clusterable speaker embedding from unlabeled data. *Applied Sciences*, 11(21):10079.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., Shih, J. J., Krusienski, D. J., and Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, 11(3):035015.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Olivetti, E., Kia, S. M., and Avesani, P. (2014). MEG decoding across subjects. In *2014 International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4. IEEE.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saha, S. and Baumert, M. (2020). Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience*, page 87.

Saito, Y., Takamichi, S., and Saruwatari, H. (2019). DNN-based speaker embedding using subjective inter-speaker similarity for multi-speaker modeling in speech synthesis. *arXiv preprint arXiv:1907.08294*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, K., Gao, X., Zhao, Y., Li, X., Dou, D., and Xu, C.-Z. (2019). Pay attention to features, transfer learn faster CNNs. In *International conference on learning representations*.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., and Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Zhang, X., Wang, J., Cheng, N., and Xiao, J. (2020). MDCNN-SID: Multi-scale Dilated Convolution Network for Singer Identification. *arXiv preprint arXiv:2004.04371*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

A Appendix

To gain further insight into the learned subject embeddings we computed accuracy on each subject's validation data using other subjects' embeddings. In the resulting subject-by-subject confusion matrix the value in the *i*-th row and *j*-th column shows how well the embedding of subject *i* can be replaced with the embedding of subject *j* (Figure 7). After division with the original accuracies the metric shows how much accuracy can be retained when swapping subject embeddings. Some subjects' embedding cannot be replaced by others (e.g. subject 3), and some subjects' embedding can be more easily replaced (e.g. subject 12). Conversely, some subjects' embeddings are more general as they can replace many others (e.g. subject 14), and some are less general (e.g. subject 2). We tried clustering this matrix, and looked at correlation with both embedding distance and subject accuracy, however no meaningful results were found.

	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
V0		0.16	0.09	0.08	0.26	0.21	0.13	0.12	0.13	0.18	0.12	0.10	0.20	0.11	0.19
V1	0.13		0.04	0.11	0.11	0.09	0.10	0.13	0.11	0.18	0.12	0.10	0.13	0.03	0.15
V2	0.09	0.05		0.14	0.06	0.10	0.08	0.10	0.10	0.11	0.13	0.10	0.10	0.08	0.14
V3	0.04	0.04	0.07		0.03	0.06	0.03	0.05	0.05	0.06	0.06	0.06	0.03	0.03	0.07
V4	0.36	0.16	0.12	0.07		0.17	0.33	0.24	0.13	0.25	0.16	0.16	0.23	0.15	0.35
V5	0.15	0.08	0.08	0.11	0.11		0.09	0.16	0.18	0.14	0.08	0.08	0.27	0.11	0.11
V6	0.14	0.11	0.05	0.05	0.23	0.08		0.09	0.09	0.15	0.08	0.12	0.09	0.10	0.28
V7	0.23	0.16	0.17	0.13	0.26	0.29	0.21		0.24	0.21	0.26	0.33	0.25	0.23	0.29
V8	0.15	0.10	0.05	0.07	0.11	0.11	0.07	0.17		0.13	0.14	0.13	0.18	0.07	0.10
V9	0.25	0.24	0.12	0.12	0.19	0.15	0.19	0.20	0.16		0.27	0.16	0.20	0.03	0.32
V10	0.29	0.13	0.11	0.10	0.16	0.15	0.15	0.24	0.24	0.25		0.15	0.25	0.06	0.23
V11	0.11	0.15	0.05	0.10	0.18	0.12	0.10	0.21	0.13	0.19	0.10		0.15	0.18	0.18
V12	0.60	0.30	0.23	0.19	0.40	0.49	0.30	0.37	0.47	0.42	0.49	0.42		0.28	0.28
V13	0.10	0.06	0.05	0.05	0.14	0.07	0.10	0.10	0.07	0.04	0.06	0.14	0.13		0.09
V14	0.19	0.10	0.05	0.06	0.30	0.10	0.29	0.15	0.10	0.20	0.12	0.17	0.18	0.17	

Figure 7: Subject embedding confusion matrix. Columns (E0-E14) refer to subject embedding indices and rows (V0-V14) refer to subject validation sets. Greener shading (higher values) shows subjects with higher retained accuracy when their embeddings are swapped.

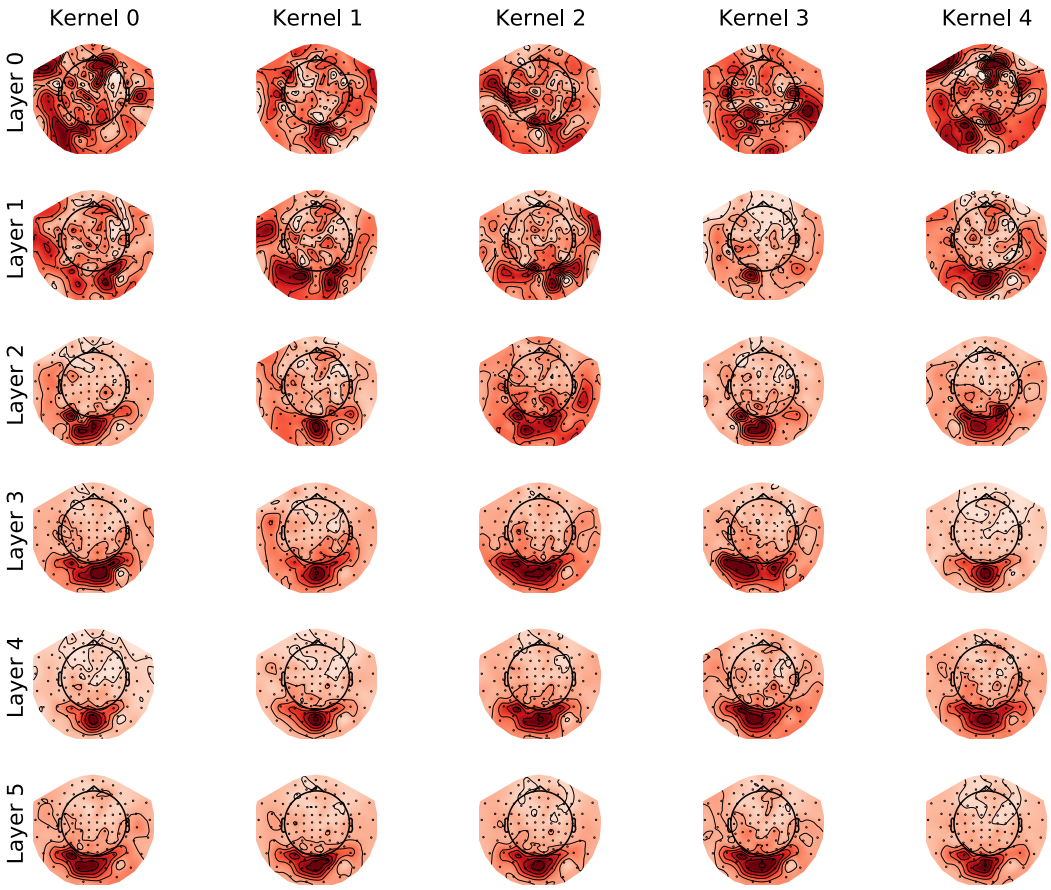
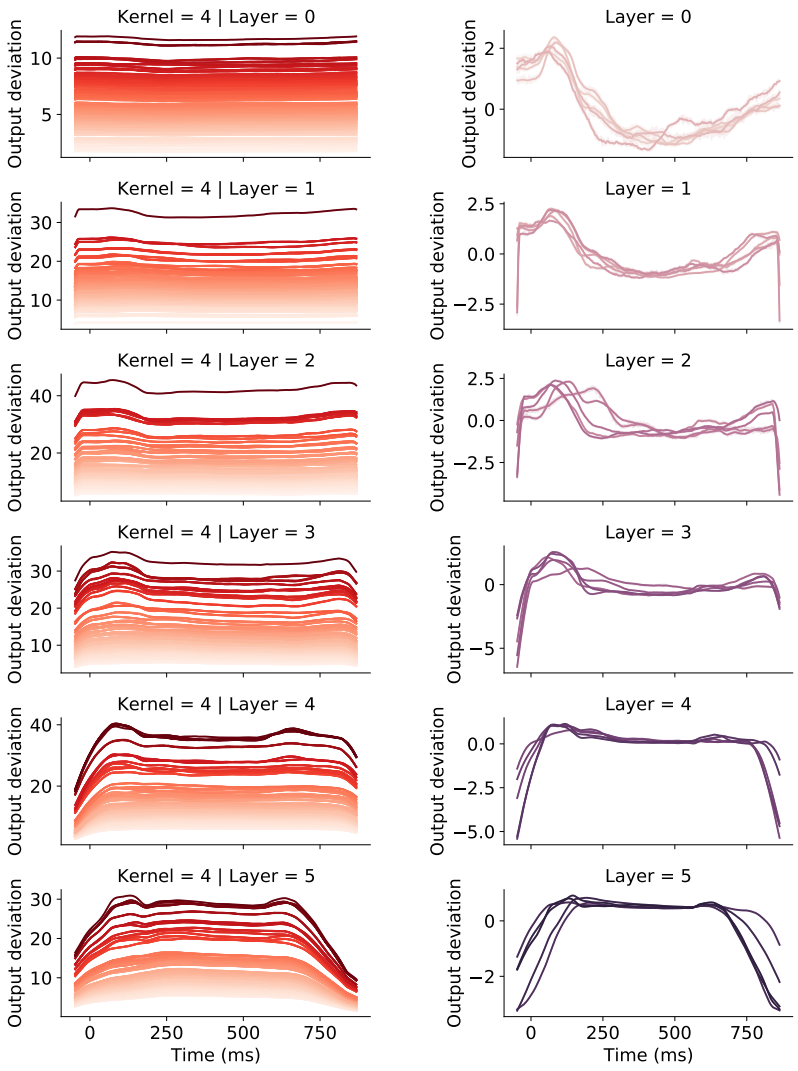


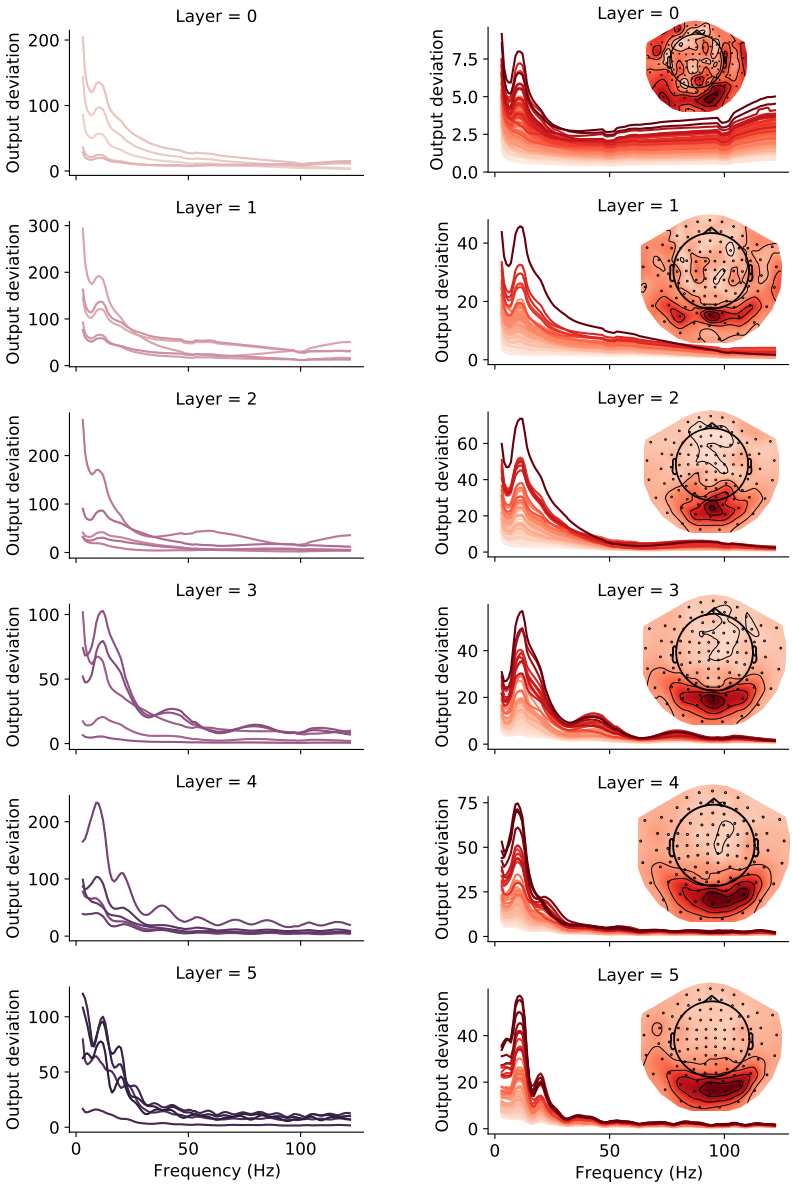
Figure 8: Spatial PFI across 6 layers (rows), with 5 kernels per row. Darker reds mean higher output deviation.



(a) Chn Temporal PFI

(b) Temporal PFI

Figure 9: Channelwise temporal PFI (a), and temporal PFI (b) across kernels of the `nonlinear group-emb` model in 6 layers (rows). For temporal PFI 5 kernels (lines) are plotted together. Channelwise temporal PFI shows the temporal PFI of each channel for Kernel 5. Channel coloring is matched to the corresponding spatial PFI map, and darker reds mean higher output deviation. For temporal PFI output deviation is normalized. The horizontal axis shows time elapsed since image presentation for both temporal PFI types. 95% confidence interval is shown with shading.



(a) Spectral PFI

(b) Chn Spectral PFI

Figure 10: Frequency sensitivity of kernels via spectral PFI (a), channelwise spectral PFI (b) from 6 layers (rows). Kernels are plotted together (lines) for spectral PFI. Each channelwise spectral PFI plot is for 1 kernel, where lines show the spectral PFI of corresponding channels in the topomap. 95% confidence interval is shown with shading for spectral PFI. Due to small variability across permutations this is barely visible.

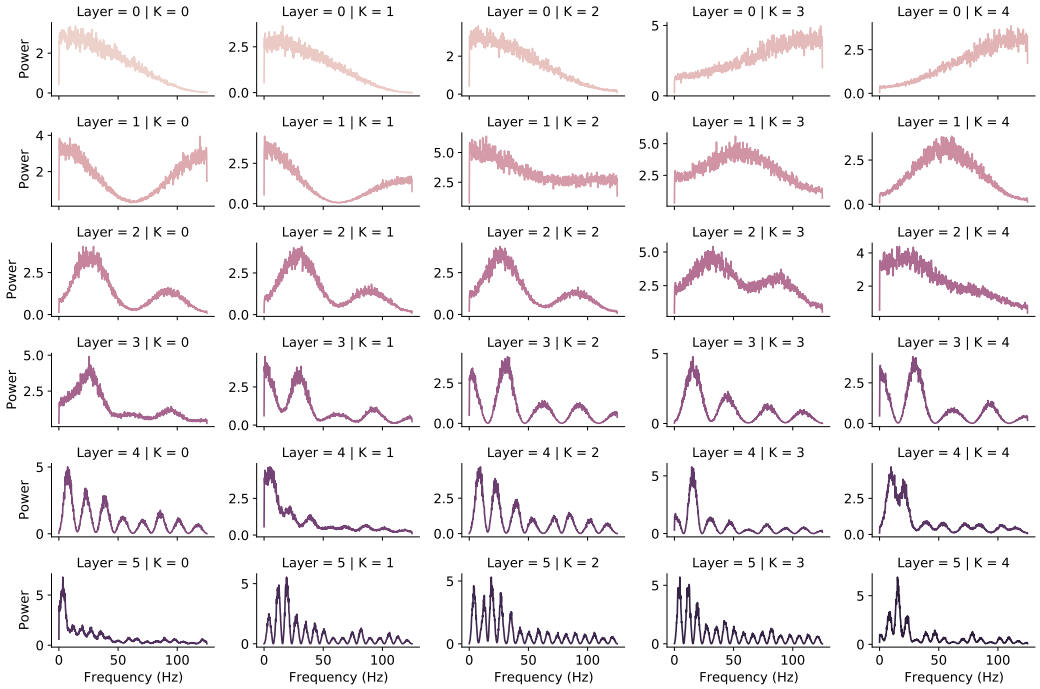


Figure 11: Frequency characteristics of 5 kernels across 6 layers (rows) via kernel FIR analysis. The power spectra is normalized.