Blackboxes Meet Blackboxes Across the World: A Multilingual Analysis of LLMs' and Human Linguistic Information Processing

David Frühbuß and **Andrei Blahovici** and **Joris Galema** and **Ron Kremer** and **Francesco Tinner**
University of Amsterdam

Abstract

As large models reach human-like capabilities especially in the domain of natural languages it becomes increasingly interesting to understand whether these models process information the same way that humans do. In the past, brain representations have been compared to the representations of large language models. In this work we extend this line of work by comparing brain and model representations across three languages. For the brain representations, we use fMRI scans of the *Le Petit Prince* (Li et al., 2021) dataset. We obtain language model representations from a multilingual XLM-R model and three monolingual models. We use Representational Similarity Analysis (RSA) to compare the representations across models, subject and languages. We find that the multilingual models have a distinct two step structure in their representation and are more similar to the brain representations than the representation of monolingual models.

1 Introduction

Neuroscientists have been studying the mechanisms underlying language processing in the human brain for some time, with the aim of understanding how the words and sentences are represented. Recent developments in functional magnetic resonance imaging (fMRI) allow for a noninvasive detection of brain activation patterns during cognitive tasks, which facilitates the collection of data drastically (Sutterer and Tranel, 2017; Wildgruber et al., 2006; Newman et al., 2001; Wu et al., 2012). Analyzing brain activations of humans performing linguistic activities could allow to better understand the processes and mechanisms of human language processing (Newman et al., 2001; Friederici et al., 2003; Kuperberg et al., 2003).

Advancements of artificial neural networks and computational methods, enabled neural approaches to surpass traditional approaches in the domain of Natural Language Processing (NLP) tasks. This has been attributed in part to the brain-like structure of neural networks. The language processing capabilities of deep neural networks are being investigated, but it is not yet known how closely they resemble human cognition. A method that allows to estimate the relatedness between the activations of a neural network and the fMRI activations is Representational Similarity Analysis (RSA) (Abnar et al., 2019; Hale et al., 2022).

Research in this area has - up to this point - mainly utilized data in English. However, the first multilingual fMRI dataset, *Le Petit Prince* (Li et al., 2021), represents a new advancement in this field of study. This dataset offers a chance to compare the activations of contextualized, multilingual neural language models with brain activation data from human participants listening to the same story in their own native language (English, French, and Chinese).

In this paper, we extend Abnar et al.'s work to the multilingual setting and answer three main questions: (1) Does the structure of the representations in large language models (LLMs) change between multilingual and monolingual models? (2) How similar are the brain representations between subjects and between languages? (3) Are the representations of multilingual models more similar to human brain representations than the representations of a monolingual model?

Regarding (1), as LLMs have been shown to develop better representations with more data, and because each language serves the purpose of transferring information but archives this using individual syntax and vocabulary. We hypothesize that training on data of different languages allows multilingual models to develop more general representations compared to monolingual models.
For (2), we reason that it has been shown that there is large variability in brain representations between different humans. We want to investigate the extend of the differences in the representations be-

tween humans listening to the same story in different languages. Our hypothesis is that the similarity between subjects listening to different languages should mirror the patterns observed between models that process different languages.

Lastly, considering (3), as the human brain is inherently multi-modal, and thus receives information about every object and concept from multiple sources, it is likely that the brain utilizes general purpose representations and computations, especially for high level cognitive processes, such as language processing. We hypothesize therefore that human brain representations are more similar and therefore comparable to multilingual models than monolingual models.

2 Dataset and Models

2.1 Le Petit Prince Dataset

The *Le Petit Prince Dataset* (Li et al., 2021) contains fMRI images obtained from 49 English speakers, 35 Chinese speakers and 28 French speakers which were obtained while they listened to the audio book *The Little Prince* in their native language. We use the first of 9 sections for our analysis, which includes for each subject a timeseries of fMRI images, recorded in 2 second intervals with close to 500.000 voxels each.

The text data of the audio book is provided with the dataset, containing detailed information about the onset and offset of every word. Further details are in Appendix A.

2.2 Neural Language Models

For our analysis of language model representations, we use (1) a multilingual language model, XLM-R (**xlm-roberta-base**), and (2) a monolingual language model for each of our 3 languages. The XLM-R model is trained on 100 languages, including English, French and Chinese. The monolingual models are trained on only one language respectively. For Chinese, we use **bert-base-chinese**, for French **Geotrend/bert-base-fr-cased** and for English **bert-base-uncased**.

3 Method

Our code is publicly available on GitHub.

3.1 Representational Similarity Analysis (RSA)

Representational Similarity Analysis is a way to compare representations from different representational spaces (Kriegeskorte et al., 2008) (Laakso and Cottrell, 2000; Abnar et al., 2019). Instead of directly mapping one representation to the other, one first calculate a self similarity matrix across a timeseries of representations from each representational space. The representational similarity score between the resulting RSA matrices is then obtained by calculating the correlation, in our case Pearson correlation, between the two.

3.2 Extracting Representation Timeseries

We separate the text data into sentence blocks of two sentences per block. For the model representation we pass each block into the model and obtain the hidden states at each layer. For the FMRI scans we already have a timeseries, but we need to align the timesteps to the sentence blocks. We align by averaging all FMRI scans that are recorded within the time window of one text block. For the FMRI scans we further apply Principle Component Analysis to capture the main features the scans. During the analysis we recognised the importance including later principle components. Therefore we upsample our timeseries to 786 timesteps before taking the PCA of the FMRI representations. This allows us to keep 786 principle components.

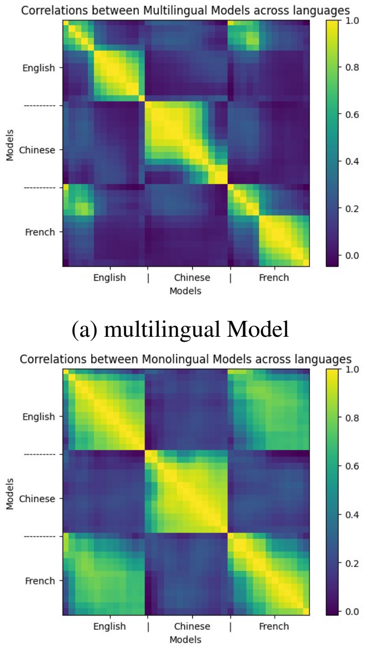
Next we compute the RSA matrices for each of our models' layers and our each of our subjects by taking the pairwise cosine distance between every timestep in their respective timeseries. After removing subjects with artefacts in their RSA matrices, we are left with 88 subjects of which 23 are Chinese speakers, 43 English speakers and 22 French speakers. We obtain similarity scores between two RSA matrices by taking the pearson correlation of the upper trinagel of the matrix, without the 5 main diagonals.

4 Results

4.1 Multi-lingual and mono lingual models

Figure 1 shows the Representational Similarity of monolingual and multilingual models with themselves and eachother.

First observe in the RSA of the multilingual model 1a the high correlation between lower layers and high correlation between higher layers regarding the same language. The figure show a clear distinctive layer for each language where the layers before this layer are similar to each other and the layers after this layer are similar to each other. This could be due to a distinction between lower layers



(a) multilingual Model

(b) monolingual models

Figure 1: Representational Similarity between hidden layers of 1a: multilingual XLM-R, 1b: monolingual language models

being related to syntactic information processing as suggested in (Tenney et al., 2019)

Moreover, English and French show similarities in exactly these layers. English and French are syntactically similar languages which further suggests that these layers are related to syntax.

Another explanation for this could be the fact that English and French share a large part of their vocabulary which might have resulted in similar hidden representations for early layers since they could have been processing the same words.

Chinese does not show this similarity with French and English which could be due to it having no overlap with their vocabulary. Chinese has it's dividing layer a lot later than English and French which could be due to the larger vocabulary size of Chinese.

However, Chinese is also syntactically different from both English and French and has a richer syntax. The fact that Chinese does not show these similarities could likewise support the statement that these layers are similar due to syntactic similarities.

Regarding the Representational similarity of the monolingual models, we again see high similarity between English and French again, this time across all layers. However, the pattern of lower layer similarity and higher layer similarity appears now to much less extent. Moreover, These dividing layers are not the same as in XLM-R. English and Chinese show the division in a very early layer whereas French shows it later.

Table 1 shows mean correlations across layers of XLM-R for each language pair. Table 2 shows mean correlations across layers of XLM-R for each language pair.

Table 1: Mean correlations across XLM-R layers for language pairs

	Model English	Model Chinese	Model French
Model English	0.495	0.099	0.177
Model Chinese	0.099	0.715	0.075
Model French	0.177	0.075	0.528

4.2 Analysing Brain Representations Across Languages

Figure 2 shows Representational Similarity of subjects with each other. Observe that the similarities between subjects of the same language are higher than between subjects with different languages. Table 3 shows mean correlations across participants.

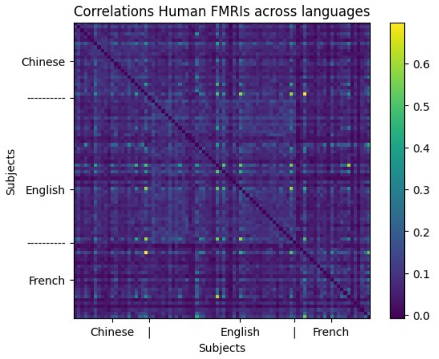


Figure 2: Representational Similarity between subjects across languages

Table 2: Mean correlations across monolingual model layers for language pairs

	Model English	Model Chinese	Model French
Model English	0.527	0.127	0.215
Model Chinese	0.127	0.705	0.094
Model French	0.215	0.094	0.565

Table 3: Mean correlation across subjects for language pairs

	Subj. English	Subj. Chinese	Subj. French
Subj. English	0.095	0.083	0.066
Subj. Chinese	0.083	0.097	0.068
Subj. French	0.066	0.068	0.080

Table 4: Mean correlation across subjects and layers of XLM-R for language pairs

	Subj. English	Subj. Chinese	Subj. French
Model English	0.091	0.079	0.059
Model Chinese	0.093	0.090	0.071
Model French	0.084	0.095	0.068

4.3 Comparing Brain and Model Representations Across Languages

Figure 3 shows the Representational Similarity between subjects and the multilingual XLM-R layers. Observe for the early layers of the French model, a higher correlation pattern can be seen across participants of all languages.

Moreover, observe that there is a Chinese subject who shows higher correlation with layers of all the languages, an English subject who shows notably low correlation with the layers that correspond to the lower/higher layer division discussed in 4.1 and an English subject showing high correlation with layers of English.

Table 4 shows mean correlations over layers and subjects of language pairs.

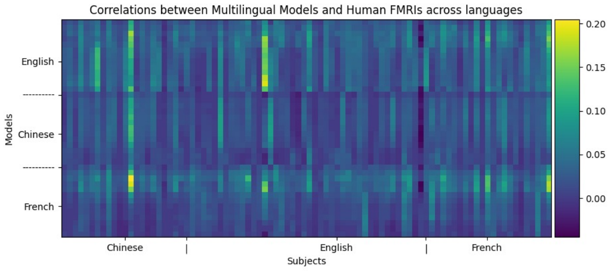


Figure 3: Representational Similarity between all subjects and layers of multilingual XLM-R

Figure 4 shows the Representational Similarity between subjects and the monolingual models.

For some subjects a similar pattern of high similarity can be seen as in 3. However the higher similarigty corresponding to earlier, possibly syntactic, layers is now less apparent.

Table 5 shows mean correlations over layers and subjects of language pairs

Table 5: Mean correlation across subjects and layers of Monolingual models for language pairs

	Subj. English	Subj. Chinese	Subj. French
Model English	0.064	0.069	0.069
Model Chinese	0.047	0.046	0.055
Model French	0.053	0.052	0.060

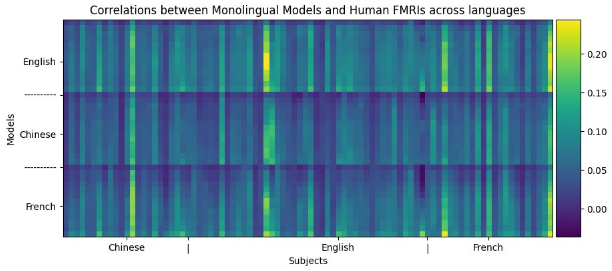


Figure 4: Representational Similarity between all subjects and layers of monlingual models

5 Related Work

5.1 Encoding of Linguistic Information in the Brain

The way linguistic information is being encoded in language models has been studied in multiple works. (Tenney et al., 2019) focus on the pretrained text encoder model BERT, and investigate how linguistic information is encoded within the model. They discover that BERT represents the various steps of the traditional NLP pipeline in an interpretable way that can be localized within the network. These steps include POS-tagging, parsing, NER, semantic roles, and coreference resolution, and they can be found in the expected sequence of a NLP pipeline. Thus, syntactic information seems to be represented in the early layers (Tenney et al., 2019). Additionally, the researchers observe that BERT can dynamically adjust this pipeline, modifying lower-level decisions based on higher-level representations to disambiguate information.

In the context of metaphoric sentence processing, (Djokic et al., 2020) further investigate the similarities between brain activations, and representations obtained using language models and other types of semantic models. Their research aims at investigating whether semantic models are able to decode brain activity associated with literal or metaphoric sentences / text segments. They conclude that the differences between literal and metaphoric sentence processing align with the brain's processing of the concepts concrete versus abstract. Concrete concepts rely heavier on the sensorimotor areas, while abstract concepts rely more on brain regions related to language processing (Djokic et al., 2020).

5.2 Link Between Brain Activations and Language Model Representations

Abnar et al. extend this analysis to other modalities, and try to compare the language model's internal representations of linguistic information to the activation patterns visible in fMRI scans, and ultimately aim at making comparisons between the internal representations of artificial neural networks, and the human brain. They do so by comparing the representations generated by language models such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2018), and others, with brain data comprised of fMRI brain imaging. Using a novel approach ReStA, which is a variant of the popular algorithm RSA, that is widely used in the cognitive neuroscience field, they investigate the amount of correlations between these two types of signals. Specifically their focus lays on the per-layer representations of the language model. The title's term "blackbox" refers to the black box nature of SOTA language models (Abnar et al., 2019).

5.3 Multilingual fMRI Data

While Abnar et al. study the correlation between the two types of activations purely based on English data, (Li et al., 2021) provides the multilingual corpus - *Le Petit Prince: A multilingual fMRI corpus using ecological stimuli* - consisting of audio transcripts and corresponding fMRI scans. This work represents a significant advancement in the field of neuroimaging and NLP research, as it allows to explore language processing in a multilingual setting. The goal of this research project is to provide insight into the brain processes underlying language production and comprehension across the languages French, Chinese, and English. For more details about corpus' content see 2.1.

6 Discussion

Regarding the first research question, we see differences in similarity for the monolingual models and multilingual models. Both show a dividing layer which divides lower layers which show similarity and higher layers that show similarity allthough the monolingual models show this to less extent and in different layers.
Regarding the second research question, we see a similarity between subjects in the same language and we see more similarity between French and English than with Chinese.

Regarding the third research question, we see that multilingual model representations are more similar to human representations and we see a clear distinction between lower layers and higher layers of the model.

Further research is needed to test our hypotheses on additional languages that (1) share syntactic properties but differ in large parts of the vocabulary and (2) on syntactically similar languages to Chinese that use a different alphabet. Our RSA method could be further optimized such that only voxels from specific brain regions are considered. Also, we see potential in changing the the method of feature extraction from the fMRI data. For future works, methods such as convolutions, the motion of brain activation, or brain region weighting could be investigated. Lastly, we see potential in filtering out sentences that exhibit interesting patterns on fMRI, and investigate whether such sentences, e.g., passages accompanied by hand movements, not only show visible activations of the motor skill regions on fMRI data but also on the LLM's activations.

Acknowledgements

We would like to thank Rochelle Choenni and Marianne de Heer Kloots for providing us with insights and guidance throughout this project.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Vesna G. Djokic, Jean Maillard, Luana Bulat, and Ekaterina Shutova. 2020. Decoding brain activity associated with literal and metaphoric sentence comprehension using distributional semantic models. *Transactions of the Association for Computational Linguistics*, 8:231–246.

Angela D Friederici, Shirley-Ann Rüschemeyer, Anja Hahne, and Christian J Fiebach. 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex*, 13(2):170–177.

John T. Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8(1):427–446.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Gina R Kuperberg, Phillip J Holcomb, Tatiana Sitnikova, Douglas Greve, Anders M Dale, and David Caplan. 2003. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, 15(2):272–293.

Aarre Laakso and Garrison Cottrell. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1):47–76.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2021. Le petit prince: A multilingual fmri corpus using ecological stimuli. *bioRxiv*.

Aaron J Newman, Roumyana Pancheva, Kaori Ozawa, Helen J Neville, and Michael T Ullman. 2001. An event-related fmri study of syntactic and semantic violations. *Journal of psycholinguistic research*, 30:339–364.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Matthew J Sutterer and Daniel Tranel. 2017. Neuropsychology and cognitive neuroscience in the fmri era: A recapitulation of localizationist and connectionist views. *Neuropsychology*, 31(8):972.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline.

D. Wildgruber, H. Ackermann, B. Kreifelts, and T. Ethofer. 2006. Cerebral processing of linguistic and emotional prosody: fmri studies. In S. Anders, G. Ende, M. Junghofer, J. Kissler, and D. Wildgruber, editors, *Understanding Emotions*, volume 156 of *Progress in Brain Research*, pages 249–268. Elsevier.

Chiao-Yi Wu, Moon-Ho Ringo Ho, and Shen-Hsing Annabel Chen. 2012. A meta-analysis of fmri studies on chinese orthographic, phonological, and semantic processing. *Neuroimage*, 63(1):381–391.

A Dataset

A.1 Speech Annotations

We extract sentence boundaries from the automatically generated syntactic trees and use these markers to extract sentence chunks of two successive sentences from the same section.

The reason we do not take one sentence as input for the model is that we hypothesize that the sentences and their semantic content are not perfectly aligned across languages. Therefore we broaden this window as to avoid possible misalignment of sentences across languages. Using a window size too large is also not a feasible approach, as limitations in the language model's capacity apply and also, as we average over multiple fMRI scans to obtain a final representation of brain activations we suspect that interesting activation patterns would be marginalized.

A.2 fMRI Images

We work with the preprocessed fMRI derivatives, available in the *Le Petit Price dataset*. Each section of the story, corresponds to a time series containing 282 to 309 fMRI scans (depending on the language). Each scan consists of close to 500'000 voxels.

We obtain the fMRI images corresponding to the sentence chunks discussed in A.1 by taking into account the onset and offset from the audio transcripts and account for the BOLD signal[1] response's time lag by adding a delay of five seconds (Li et al., 2021). Then, we take the mean of the fMRI images corresponding to a text chunk to obtain the fMRI data to be analysed with respect to that sentence chunk. We then flatten each scan and perform Principle Component Analysis to obtain a representation of ..(size).. for each text chunk.

A.3 Artefacts

When computing Representational Similarities between fMRI representations, we came across several participants who showed unusually high self similarities within their RSA matrices.

These participants were removed from further analysis. We have added an image of the meaned fMRI image and the self similarity matrix for each of these participants in Appendix B.

[1] The blood-oxygen-level-dependent (BOLD) signal. This is the signal that is being detected using fMRI scanners.

A.4 Sentiment Analysis

We extract sentiment labels and confidence scores by applying the huggingface sentiment analysis pipeline on the text chunks. Thus, we associate each text chunk to one of the labels (positive, neutral, or negative) with a corresponding softmax-score. For the languages English and French, we use **cardiffnlp/twitter-xlm-roberta-base-sentiment** model fine-tuned for sentiment analysis on $\approx$ 198 million tweets in eight languages. As Chinese is not one of these languages, we use another network for the sentiment extraction in Chinese: **c2-roberta-base-finetuned-dianping-chinese**. Both models use the XLM-R-base language model.

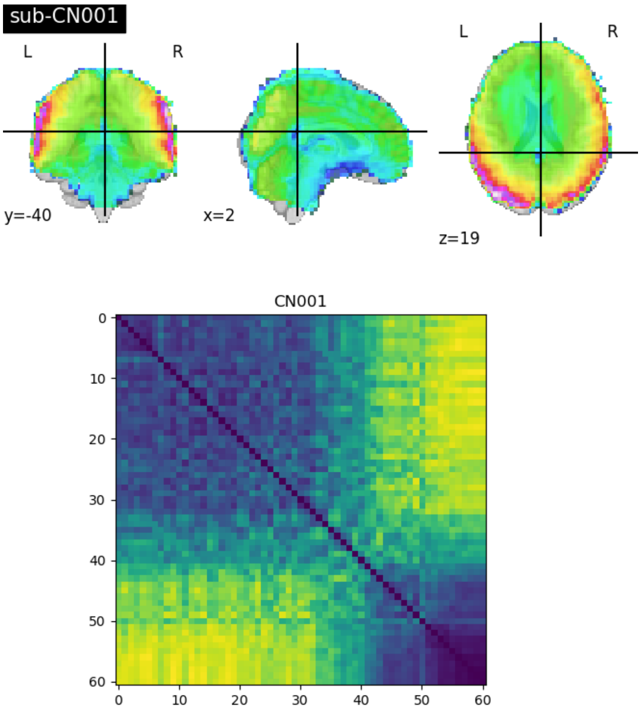


Figure 5: Artefact

B Artefacts

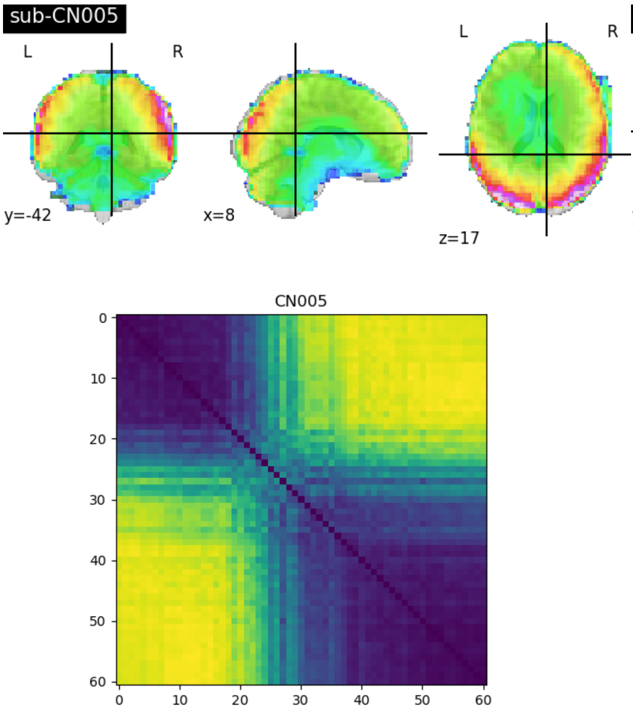


Figure 6: Artefact

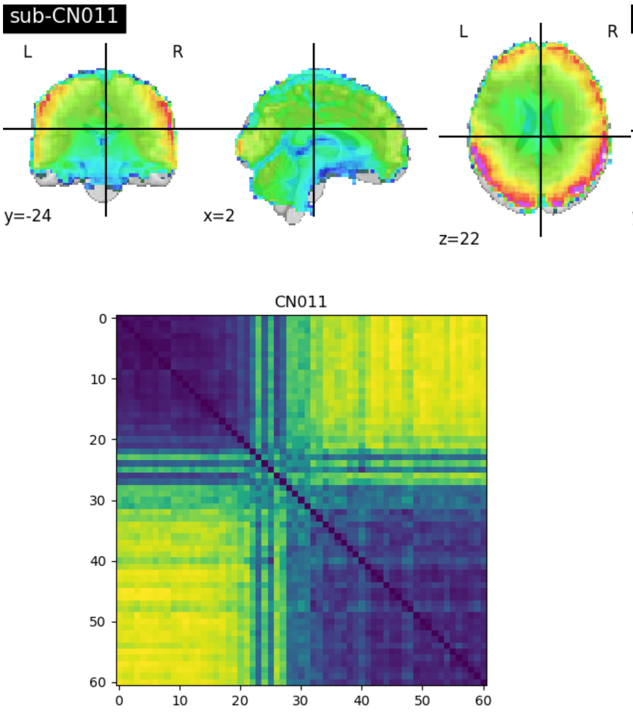


Figure 7: Artefact

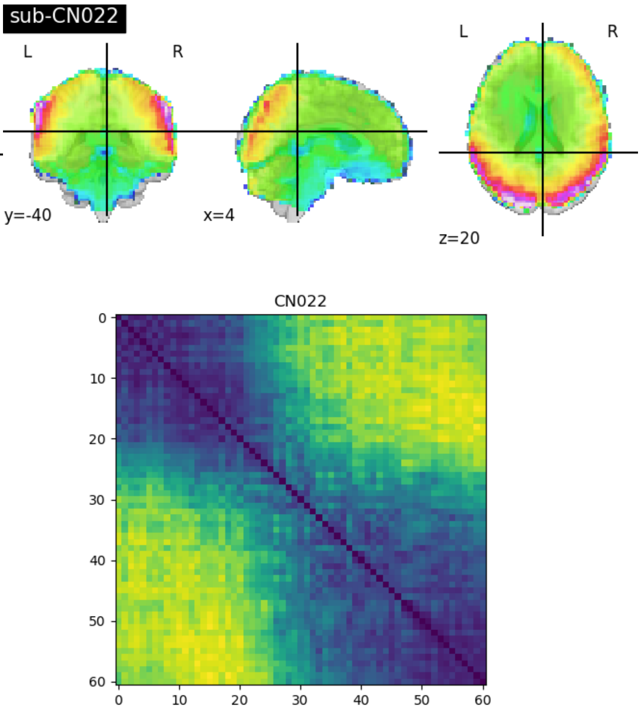


Figure 8: Artefact

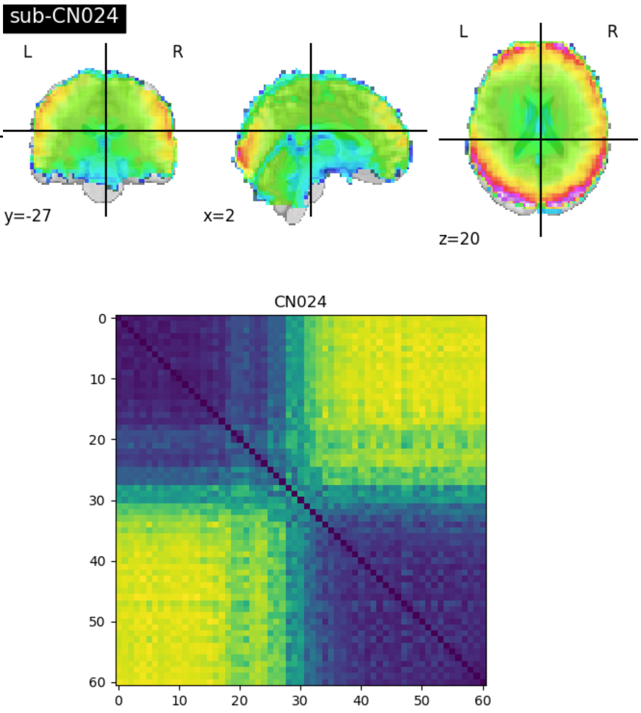


Figure 9: Artefact

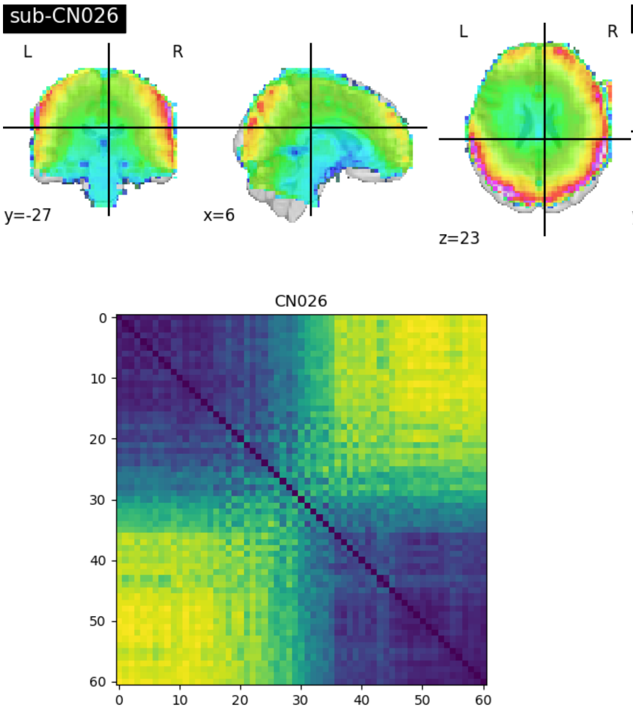


Figure 10: Artefact

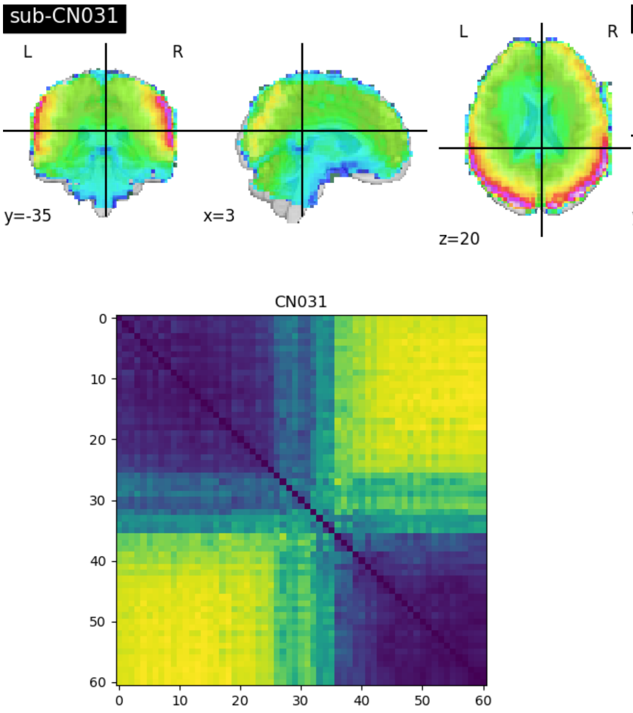


Figure 11: Artefact

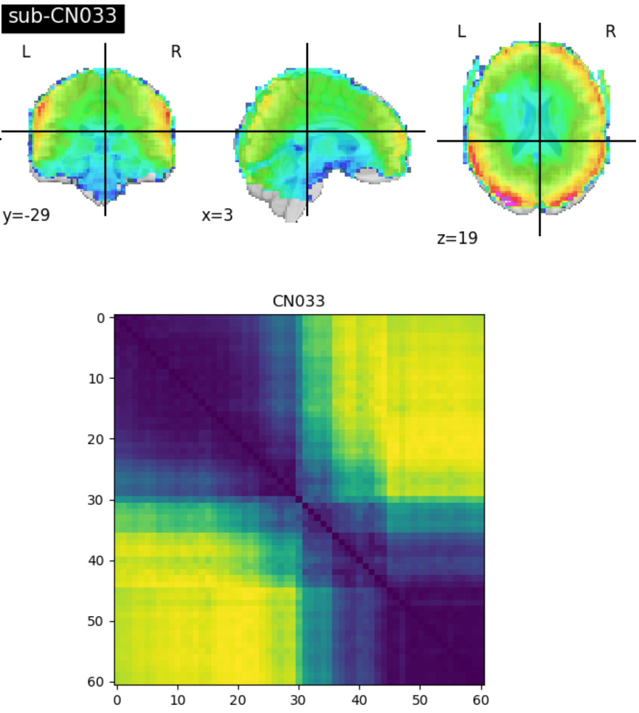


Figure 12: Artefact

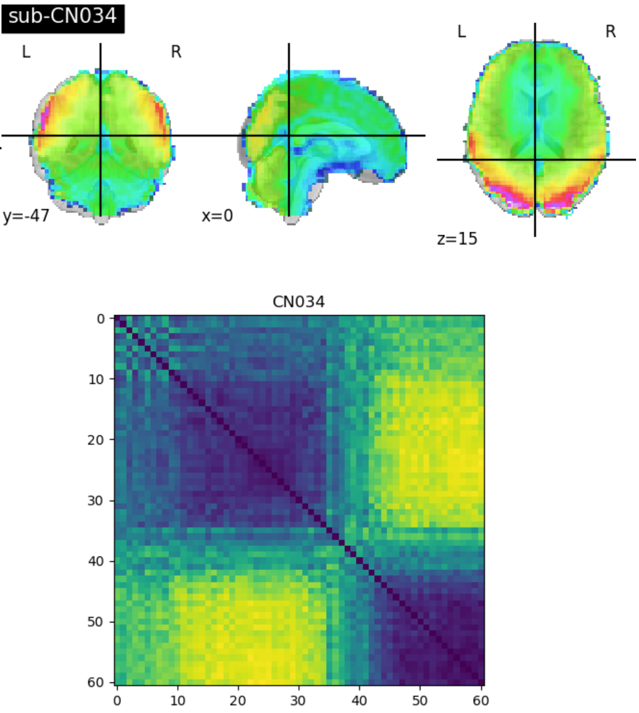


Figure 13: Artefact

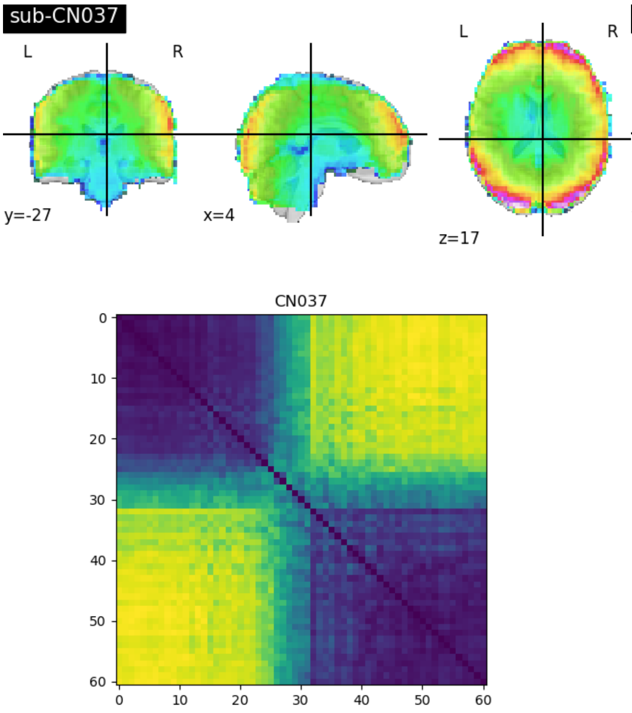


Figure 14: Artefact

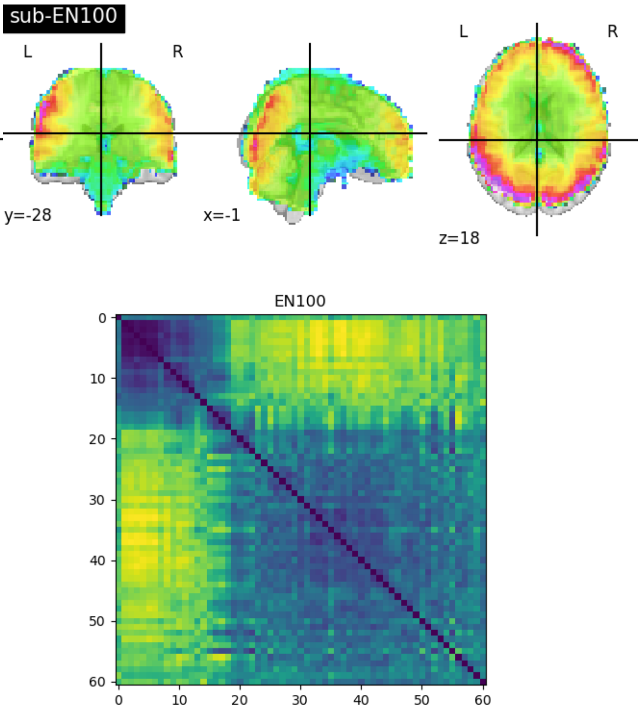


Figure 16: Artefact

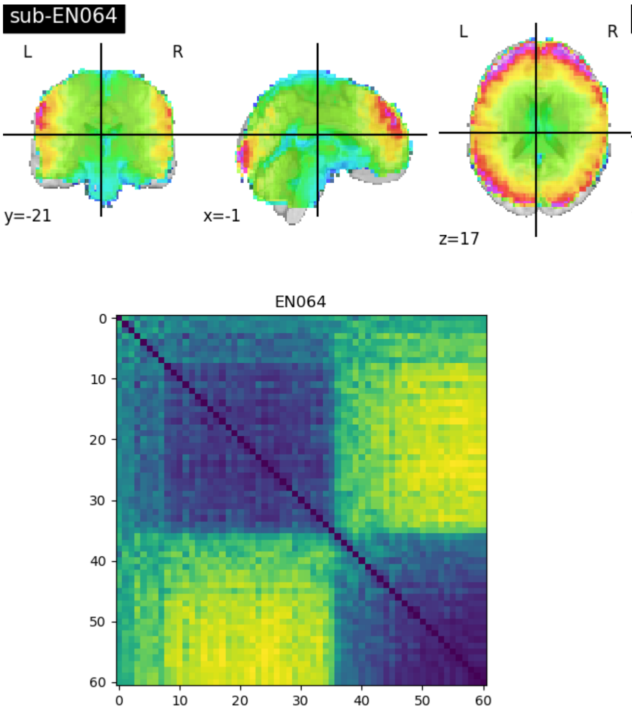


Figure 15: Artefact

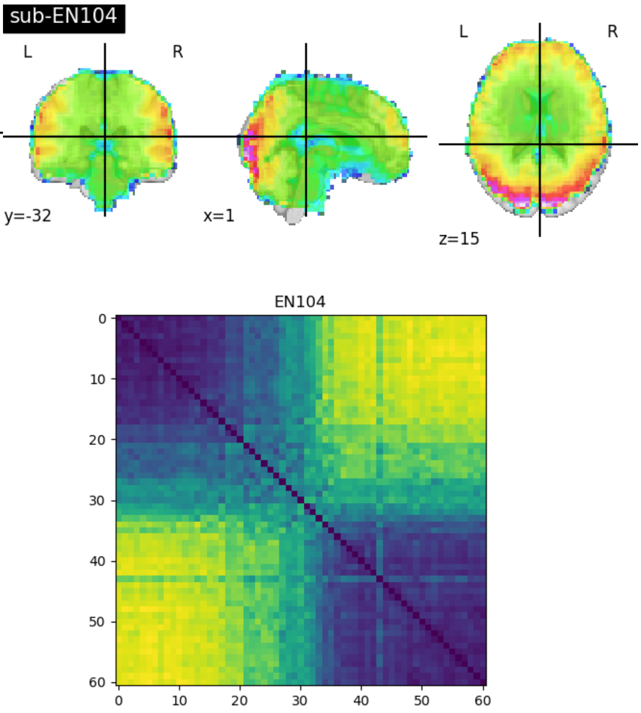


Figure 17: Artefact

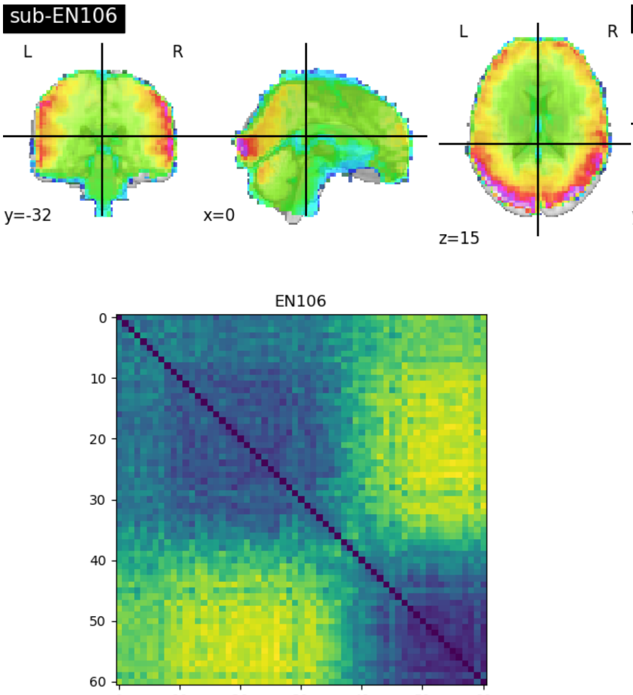


Figure 18: Artefact

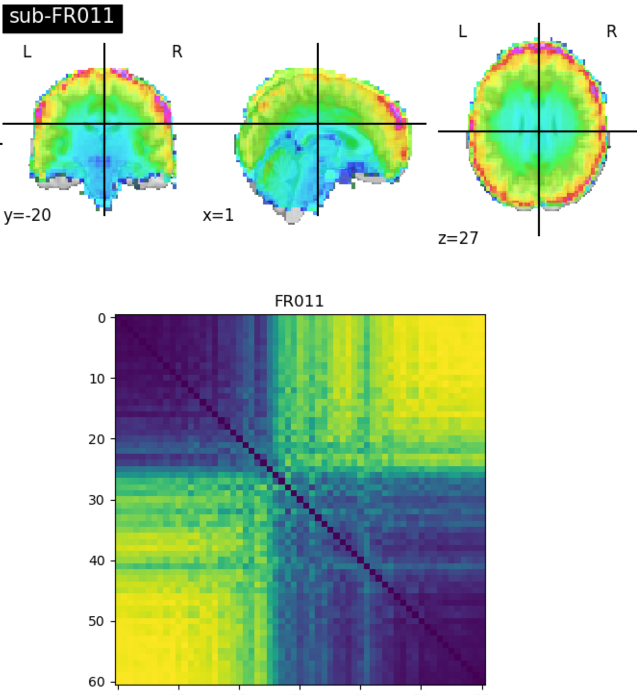


Figure 20: Artefact

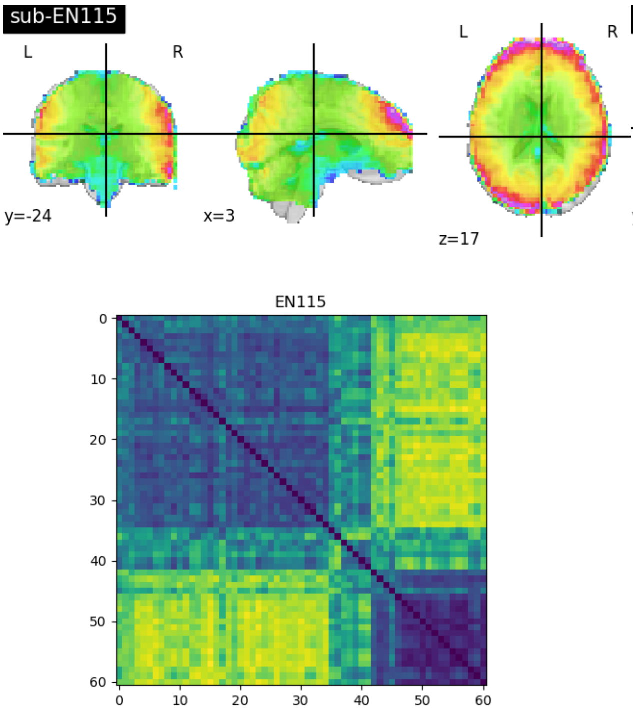


Figure 19: Artefact

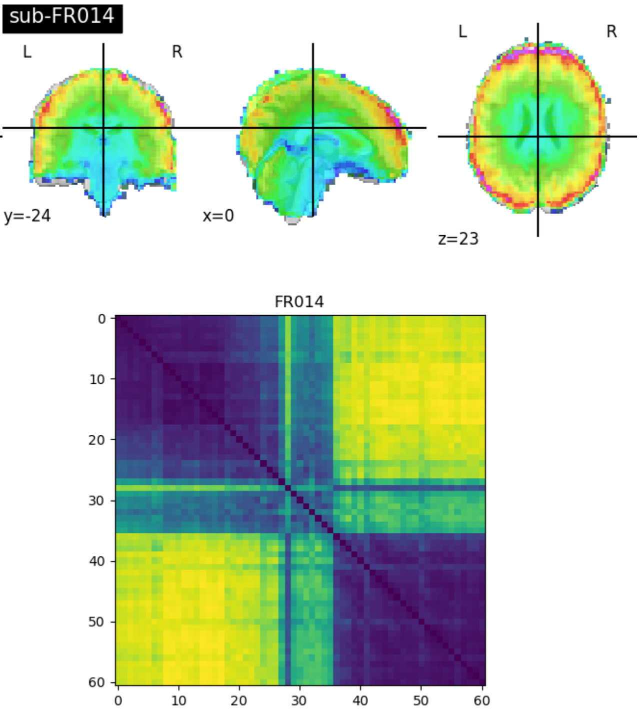


Figure 21: Artefact

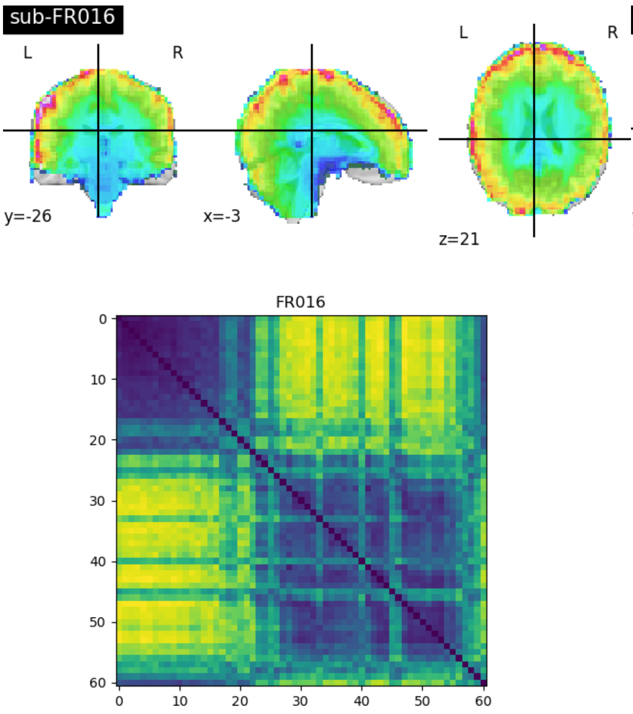


Figure 22: Artefact

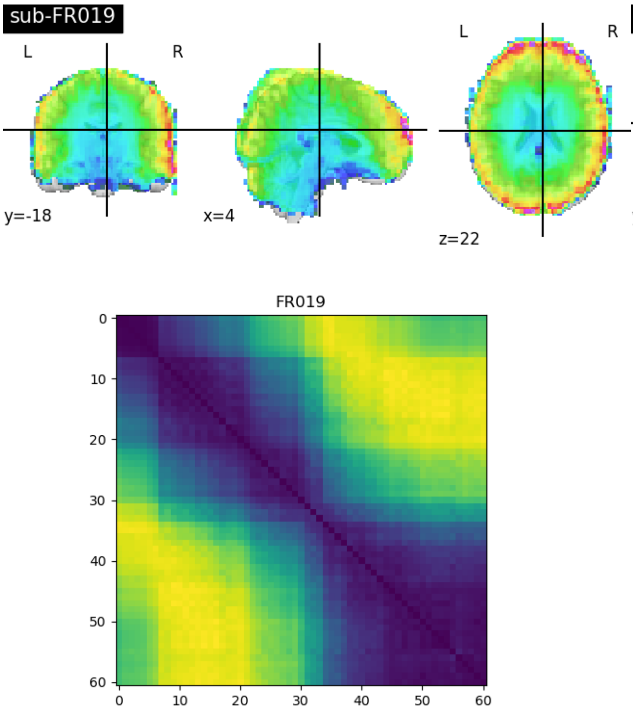


Figure 23: Artefact

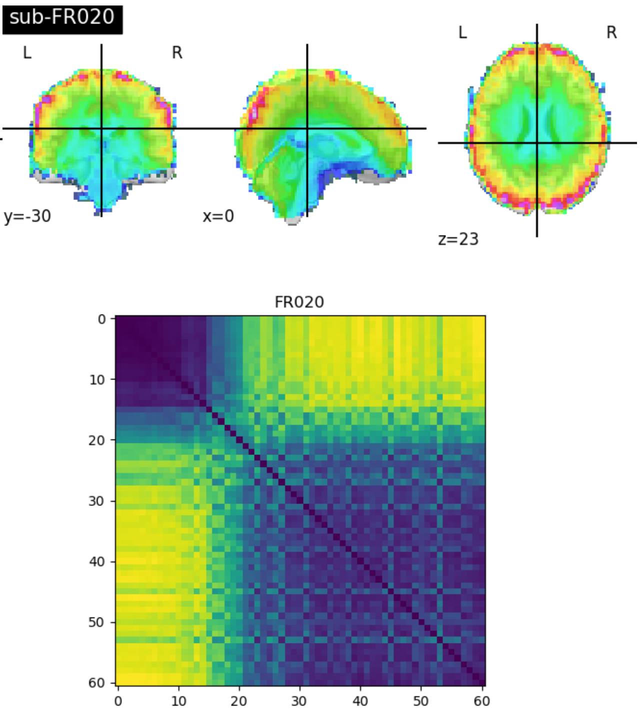


Figure 24: Artefact

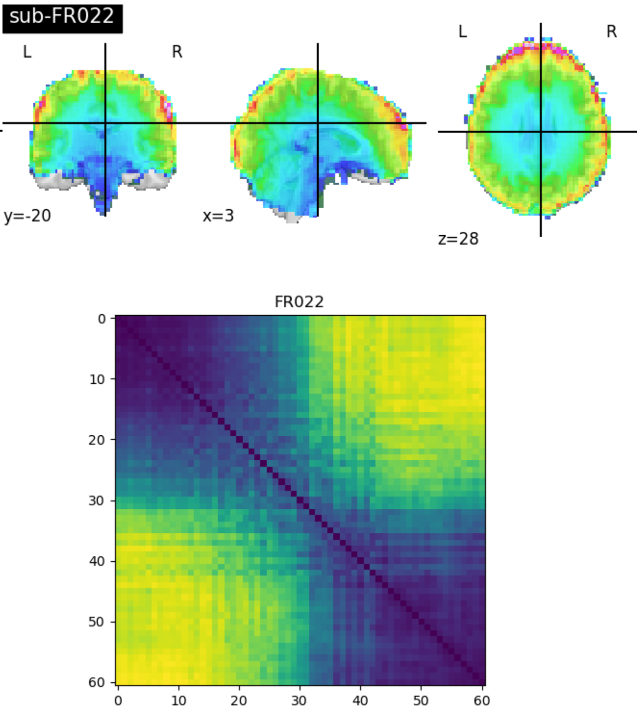


Figure 25: Artefact