Brainprints: identifying individuals from magnetoencephalograms

Shenghao Wu[1,2], Aaditya Ramdas[2,3], and Leila Wehbe[1,2,*]

[1]Neuroscience Institute, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Department of Statistics and Data Science, Carnegie Mellon University

* correspondence should be addressed to: lwehbe@cmu.edu

Word count: 5118
Number of figures: 7
Supplementary Figures & Tables: 8

Abstract

Magnetoencephalography (MEG) is used to study a wide variety of cognitive processes. Increasingly, researchers are adopting principles of open science and releasing their MEG data. While essential for reproducibility, sharing MEG data has unforeseen privacy risks. Individual differences may make a participant identifiable from their anonymized recordings. However, our ability to identify individuals based on these individual differences has not yet been assessed. Here, we propose interpretable MEG features to characterize individual difference. We term these features brainprints (brain fingerprints). We show through several datasets that brainprints accurately identify individuals. Furthermore, we identify consistent brainprints components that are important for identification. We study the dependence of identifiability on the amount of data available. We also relate identifiability to the level of preprocessing and the experimental task. Our findings reveal specific aspects of individual variability in MEG. They also raise concerns about unregulated sharing of brain data, even if anonymized.

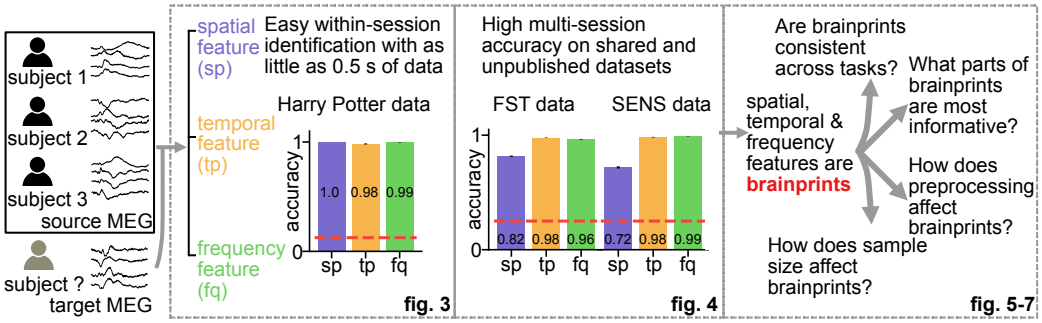

Figure 1: **Graphical abstract.** Identifying which subject a segment of MEG data belongs to is strikingly easy when other data from the same session is available for every subject. We propose three types of interpretable features that can be used to identify individuals across sessions with high accuracy. Identifiability of individuals is influenced by factors such as resting state vs. task state, components of each feature, the sample size and the level of preprocessing. Our results reveal aspects of individual variability in MEG signals and highlight privacy risks associated with MEG data sharing.

Introduction

The open science movement is a result of the increasing awareness of the importance of sharing data and code to promote scientific reproducibility [1]. Public repositories enable researchers to share their neuroimaging data (fMRI, EEG, MEG, etc.) while making sure to censor out individual information [2]. However, data anonymization does not always preserve privacy [3]. Combining different types of information using methods such as record linkage approaches [4] may cause serious privacy violations. This problem is exacerbated when multiple datasets that happen to contain the same individual are available, which is rather common in neuroimaging (e.g. [5]). Hence it is natural to ask if anonymized individuals can be identified from neuroimaging datasets and if so to what degree. Specifically, we ask: do individuals have a *brainprint*, a brain-activity analog of a fingerprint? If there is evidence for a brainprint, then researchers may be warned about how easily individual information can be inferred, and it may cause them (and the field) to act with more caution when publishing neuroimaging data online. For instance, it may pave the way for the adoption of more sophisticated data-release mechanisms like differential privacy [6] and homomorphic encryption [7].

Assume there are two multi-individual neuroimaging datasets with overlapping participants: a "source" dataset and a "target" dataset. The question of interest is: can we accurately decide which individual in the source dataset corresponds to the individual in the target set? In other words, is there individual *identifiability* between the two datasets? The aforementioned question could arise naturally in practice: it is very common for university labs to recruit their own lab members for preliminary studies; these are anonymously released with an associated publication. Assume that one year later, lab member A relocates to city B, and privately volunteers for a study by a public hospital that tracks the effect of a drug (or some intervention) on patients in early stages of early-onset Alzheimer's, while collecting MEG data. If this data is also anonymously released at a future point, brainprints could plausibly be used to detect a common participant, thus identifying that A has Alzheimer's because only one member of the lab moved to city B. This would already be a gross unintended violation of privacy, but one can further imagine that an insurance company uses this to prove that a condition was pre-existing at the time of the first scan (before the individual themselves knew), or use it to decide individual-level pricing.

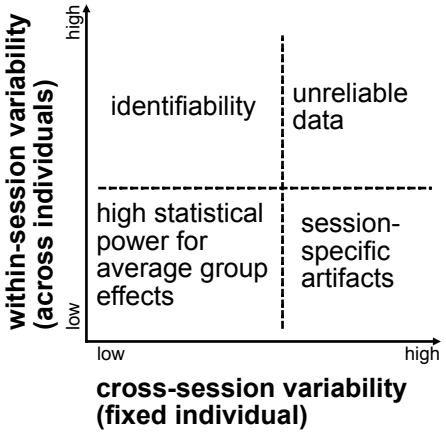

Figure 2: **Individual identifiability is a function of individual and session variability in neuroimaging**. Consider repeating an experiment in multiple sessions for a group of individuals. Cross-session variability refers to the change in the recorded data for the same individual across sessions, while within-session variability refers to differences in a single session's recorded data across individuals (keeping all other variables, including the stimulus, unchanged). The ideal conditions for the scientific discovery of an effect shared by the group is low within-session and low cross-session variability. However, the combination of low within-session and high cross-session variability indicates an artifact or a confound in the experiment design (e.g., each month, one session is recorded for all individuals and the instrument has a drift over time). High within-session variability paired with low cross-session variability leads to individual identifiability with the individual's data acting like a stable signature that differentiates them from others. Finally, high within-session and cross-session variability leads to unreliable data.

If high individual identifiability exists even if the source and target set were recorded in separate sessions for each individual, there might be essential differences in the patterns of the data among individuals which is preserved across scanning sessions. Namely, individual identifiability might be related to variability in brain structure or function (or other individual characteristics such as head size). In multi-individual, multi-session neuroimaging data, there exists "within-session" variability across individuals in the same session and "cross-session" variability of the same individual cross sessions [8]. For simplicity, consider the four scenarios in Figure 2. Low variability in both within-session (individuals are similar) and cross-session (an individual's data is consistent across session) is likely to promote statistical power for detecting average group effects with fixed sample size, thereby facilitating

reproducibility [9, 10]. High cross-session and low within-session variability (e.g. individual 1's data in session 1 is very different from their data for session 2, but somehow very similar to individual 2's data in session 1) may indicate session-specific artifacts (e.g. the scanner was faulty during the recording of session 1 for all individuals). High cross- and within-session variability makes data unreliable. Finally, high within-session (individuals are different from each other) and low cross-session variability (individuals are similar to themselves) leads to individual identifiability. Individual identifiability in turn indicates *consistent* individual differences, which in themselves are an important topic of scientific enquiry [8, 11]. Understanding sources of consistent variability can help learn the underpinnings of disease or more generally to map the relationship of brain structure and activity to individual behavioral characteristics.

Similar individual identification problems have been studied using EEG and fMRI for the purpose of biometric authentication and to investigate individual differences [12, 13, 14, 15, 16, 17, 18]. The term 'brainprint' has also been previously used to represent brain-specific information, such as morphology and event related potential biometrics [19, 20, 21], that can be used to identify individuals. Individual identification with MEG data, however, has not been fully explored. Due to availability of MEG datasets, only single-trial MEG data has been studied for person identification [22]. Other MEG studies focusing on variability of individual data [8, 11] may not make connections with individual identifiability.

In this paper, we argue that individuals can be easily identified with MEG data. We measure identifiability as identification accuracy with three interpretable MEG features on multiple public and private MEG datasets. We show that identifiability is not a product of environmental artifacts and specific features have a consistent performance between task and resting state data. We further dissect the contribution of each features into sub-features to understand what may be leading to the high identifiability. Factors such as the amount of data and level of preprocessing are also shown to have influence on identifiability. Our analysis not only confirms the worrisome potential of privacy being compromised by released MEG data via extracting simple features but also leverage the interpretability of the features to explain the underlying mechanism for the high identifiability, thereby relating it to individual variability.

Results

We organize our results from simple datasets to more complicated ones in our context, to a closer investigation of the methodology itself. We first use machine learning tools as well as interpretable features to show that identification is easy when the MEG sessions were collected on a single visit. We then show that the proposed features also achieve high accuracy on datasets of multiple visits to the scanner, and some feature is even consistent on datasets between different tasks. We finally show which components of each feature is important for individual identification, and that sample size and level of preprocessing will also affect identification accuracy.

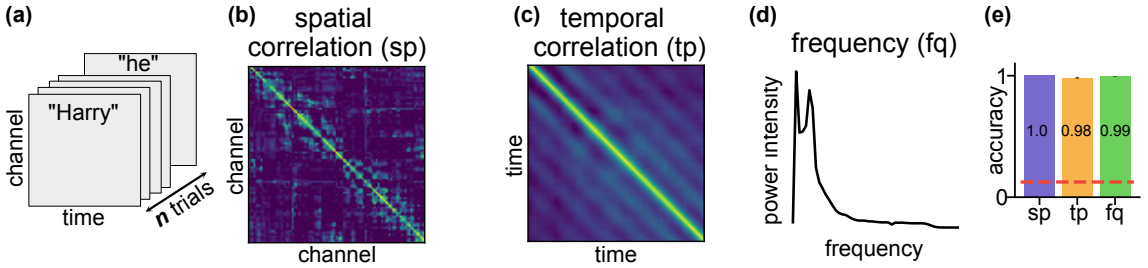

Figure 3: **High within-session identification accuracy on HP data with three interpretable features.** **(a)** Shape of the HP data before featurization. The HP data consists of participants reading a book chapter one word at a time for 0.5s each. The data is resampled to have the dimension $[102 \text{ channels}, 100 \text{ time points}, n \text{ trials}]$ where each trial corresponds to one word and n to the number of words. **(b)** The spatial correlation feature **sp** is a 102×102 Pearson's correlation coefficient matrix computed across the time points and trials. **(c)** The temporal correlation feature **tp** is a 100×100 Pearson's correlation matrix computed across the channels and trials. **(d)** The frequency feature **fq** is a vector in $\mathbb{R}^{51}$ where 51 is the number of frequency bands. The power at each band was averaged across channels and trials. **(e)** Identification accuracy with the three features. The accuracy was averaged across 100 identification runs of 8 individuals. The red dashed line represents the chance level ($= 0.125$). The error bars are the SE across individuals and identification runs and are invisible due to their small values.

Within-session identification is surprisingly easy. To measure identifiability, we consider the test accuracy of a classifier trained to identify participants from their MEG recording. We first focus on within session identifiability. In this context, we assume that each participant undergoes one session. A classifier is trained on a subset

of the session, in which each trial is labeled with the identity of the participant it corresponds to. In our framework, we refer to the training set as the source set. Then, on held-out test data, the classifier predicts which participant is associated with each test trial. We refer to the test set as the target set. As an example, we investigated individual identifiability on a MEG dataset of *eight* participants during a reading task. Participants were asked to read a chapter of Harry Potter [23] while each word was presented for 0.5s on a screen. The Harry Potter (HP) data is a single-session dataset: the data for each individual were collected on a single visit of the MEG scanner. Hence the source and target set are non-overlapping subsets of that single session. We trained a random forest classifier [24] using the MEG recording of all channels at a randomly selected time point, a flattened vector representing the snapshot of the topographic map (topomap) of the brain activity. Under this setting, we are asking if there is any individual-specific information contained in the topomap, the basic element of MEG recording. We split the dataset into 10 non-overlapping folds and use one as the target (testing) set and the other nine as the source (training) set. This 10-fold cross-validation scheme yielded a high identification accuracy (0.94) while the chance accuracy is only 0.125. This surprisingly high accuracy on merely $0.05s$ of MEG data suggests the existence of strong patterns detected by the random forest classifier. This strong pattern may be contained on the transient spatial distribution of an individual's MEG activity and is strongly distinctive of an individual. This high accuracy with the limited amount of information used suggests that within-session identification is a strikingly easy task.

Interpretable MEG features yield high identification accuracy. The random forest classifier may not enclose enough information to explain the high identifiability of the HP data because of the black-box nature of the algorithm. The topomap mainly contains the spatial information: how heterogeneous the amplitude of the signal is across channels at a certain time point. High identifiability may also be attained using temporal and frequency information. We proposed three interpretable features for individual identification to further disseminate the individual-specific information. These features are interpretable because they bear biological meanings and hence can be used to interpret the high identification accuracy. The three features were based on n randomly selected trials (words) and have the shape $[102 \text{ channels}, 100 \text{ time points}, n \text{ trials}]$

(Figure 3(a)). **sp** (Figure 3(b)) is the spatial correlation between different sensors which may be related to individual-specific correlated activities between areas of the brain or the anatomy of the individual (e.g. brain size) [8, 25]. **tp** (Figure 3(c)) is the temporal correlation between the time points into a trial. A high value in the **tp** matrix indicates highly synchronous brain signals between two temporal points, which might be related to participant specific stimulus processing latencies. A relevant study shows that the temporal change of brain activities in auditory steady-state responses are different between individuals [26]. **fq** (Figure 3(d)) represents the distribution of the power intensity of signal frequency. Individual differences might also manifest as differences in the power distribution along frequency bands [27, 22].

We used the 1-Nearest Neighbor (1NN) identification procedure, similar to Finn et al. [15], to test if the three features are *brainprints* for the within-session identification task. For a given feature such as **sp**, the feature is computed for each participant on the source set using n trials. Target set features are also computed (but unlabeled) with the same number of trials. The 1NN classifier simply assigns each target feature to the participant with the closest source feature (we use correlation to measure distance). The matching process is repeated for 100 runs to account for the variance of the feature on the sampled trials. The simplicity of this 1NN classifier optimizes the interpretability of the result. With $n = 300$ trials all three features achieve near-perfect identification accuracy (Figure 3 (e), the accuracy for **sp**, **tp** and **fq** is respectively $1 \pm 0, 0.9825 \pm 0.0046, 0.995 \pm 0.0025, \text{mean} \pm \text{SE}$, $p < 0.0002$, see Supplement F for how we computed the p-values). In fact, the high identifiability can be attained with as few as $n = 100$ trials (Supplement C). The high identifiability with **sp**, **tp** and **fq** suggests they are *brainprints*, at least for identifying individuals within a session. Therefore, multiple features capturing different aspects of the MEG activity can be used for identifying individuals.

Cross-session identification confirms the existence of brainprints. The high within-session identification accuracy suggests **sp, tp**, and **fq** are individual-specific within a session. Artifacts such as environmental noise and equipment configurations, however, might be the main contributing factor to within-session identification accuracy. Hence, we examined the consistency of the three features when the same type of task data was collected from each individual on multiple sessions. This setting tests if the features are preserved over time, i.e. if they are indeed *brainprints* and not

mere artifacts. If the identifiability is significantly lower on multi-session datasets, the high identifiability on the HP data may be a mere result of session-specific artifacts, since the recording session for each individual is performed on different days. If high cross-session identifiability is observed, **sp, tp**, and **fq** can be considered genuine brainprints because they are unique to individual and invariant between sessions. This would also suggest low cross-session and high within-session variability (Figure 2).

We tested the three features on two multi-session datasets: FST [28], a four-session dataset where four individuals were shown pictures of familiar and unfamiliar faces and SEN, a three-session dataset where four individuals were shown sentences. Since each individual has recordings conducted on different days, we can set the target and source data to be from different sessions (Figure 4 (a)), to test the role of environmental artifacts and further confirm the existence of the brainprints. In addition to identification accuracy, we used a relaxed version, the *rank accuracy*, which captures more information in a failure case where an individual is mis-identified. Rank accuracy captures the rank of the correct assignment out of all possible assignments; it is 1 if the target feature of each individual have the largest similarity to the source features for that individual, and is $\frac{1}{K}$ if the similarity is the smallest. The chance rank accuracy is $\frac{K+1}{2K}$.

Both **tp** and **fq** achieved almost perfect average identification and rank accuracy on both FST and SEN data whereas **sp** achieved lower but still well above-chance accuracy (Figure 4 (c),(f)). The high cross-session identification accuracy of **sp, tp**, and **fq** confirms that it is reasonable to call them brainprints for individual identification in MEG. The lower identification accuracy for **sp** was due to low accuracy on a two of the individuals (Figure 4 (d),(g)) in both datasets. However, identification accuracy of these individuals is not consistently low across all session pairs (Figure 4 (b),(e)) indicating that **sp** only perform worse for these subjects between certain sessions.

For SEN data, the MEG recording of two subjects were taken on the same day for session 1 and 2. Since the identification accuracy of **sp** corresponding to these two pair of sessions (1 vs 2 and 2 vs 1) did not yield higher accuracy than the average (the mean identification accuracy between these two session pairs is 0.655, lower than 0.72, the mean across all cross-session pairs), the accuracy for **sp** was not inflated due to this issue with duplicated recording times. In line with the results on the HP dataset, **sp, tp**, and **fq** are the brainprints that are consistent even between recording sessions with **tp, fq** leading to higher identifiability.

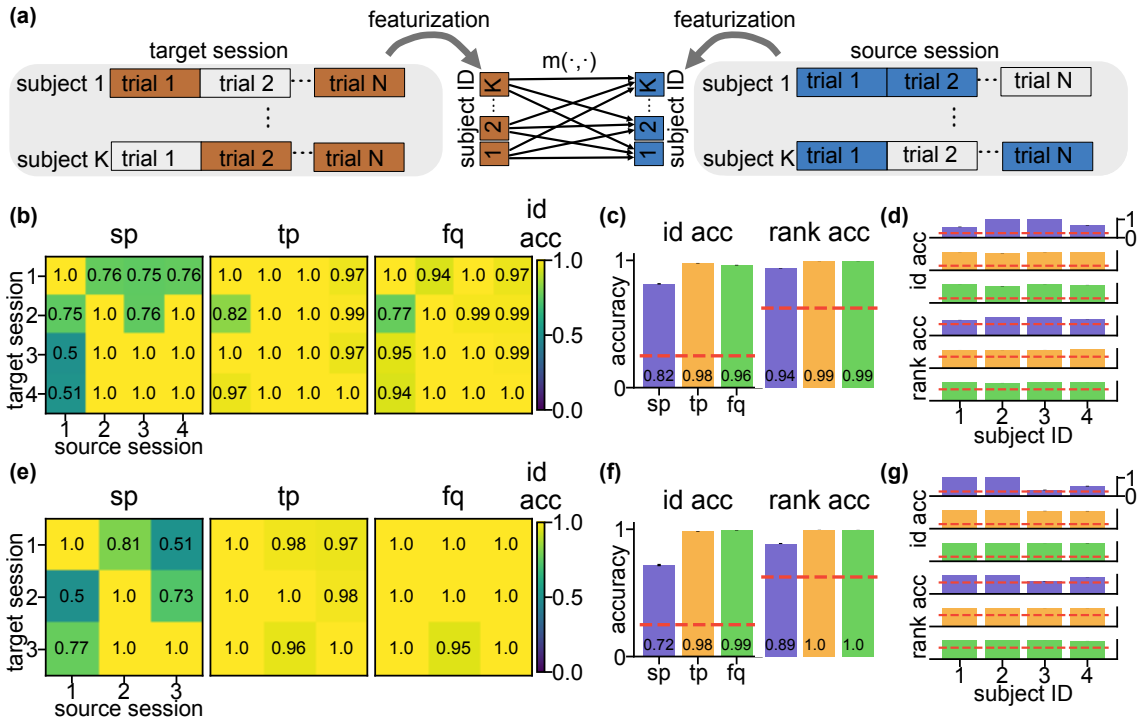

Figure 4: **Cross-session identification on FST and SEN data confirms existence of brainprints.** **(a)** Schema of the cross-session identification task. For one identification run, the features of each individual are computed using randomly sampled trials ($N = 300$) from both the source and target session. Target session features are then classified by selecting the individual with the largest similarity score in the source session. **(b)** Heat maps of the cross-session identification accuracy using the three features on FST data. Each grid represents the average accuracy across 4 individuals and 100 identification runs. The within-session accuracy (diagonal entries) are computed using the same source-target splitting procedure as on the Harry Potter data to avoid data leakage. **(c)** Average cross-session identification accuracy and rank accuracy for each feature on FST data. Within-session accuracy (diagonal entries in **(b)**) were excluded in computation. Error bars are the SE across cross-sessions ($N = 12$), individuals ($N = 4$), and identification runs ($N = 100$) and are invisible due to small values. Red dashed lines are the chance level for the identification accuracy ($= 0.25$) and rank accuracy ($= 0.625$). **(d)** Identification and rank accuracy on FST data by individual. Within-session accuracy were excluded in computation. Error bars are the SE across cross-sessions ($N = 12$) and and identification runs ($N = 100$) and are invisible due to small values. The red dashed lines are the same as in **(c)**. **(e)-(g),** same as **(b)-(d)** but on SEN data with the same number of individuals and identification runs ($N = 4$ and $N = 100$) but different number of cross-sessions ($N = 6$). The high identification accuracy with the three features on multi-session datasets confirms these features can be brainprints for individual identification.

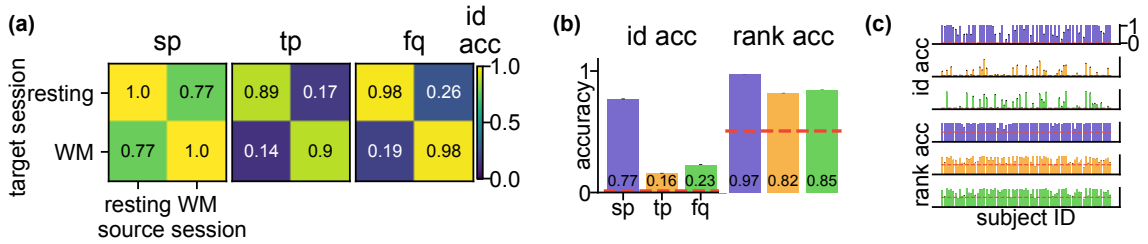

Figure 5: **Consistent sp for cross-task identification on Human Connectome Project data. (a)** Heat maps of the cross-task identification accuracy using the three features on HP data. Both resting and working memory (WM) data were recorded on the same day. Each grid represents the average accuracy across 77 individuals and 100 identification runs. The within-task accuracy (diagonal entries) are computed using the same source-target splitting procedure as on the Harry Potter data to avoid data leakage. **(b)** Average cross-task identification accuracy and rank accuracy for each feature on HCP data. Within-task accuracy (resting vs. resting, WM vs. WM) are excluded in computation. Error bars are the SE across cross-task sessions ($N = 2$), individuals ($N = 77$), and identification runs ($N = 100$) and are invisible due to small values. The red dashed lines are the chance level for the identification accuracy ($= \frac{1}{77}$) and rank accuracy ($= \frac{39}{77}$). **(c)** Identification (upper three rows) and rank (lower three rows) accuracy on HP data by individual. Within-task accuracy are excluded in computation. Error bars are the SE across cross-task sessions ($N = 2$) and identification runs ($N = 100$) and are invisible due to small values. The red dashed lines are the same as in **(b)**. These results indicate that **sp** is consistent even when performing different tasks (resting vs WM) in the source and target session.

Spatial brainprints are consistent across resting-state and tasks. The high performance and interpretability of the brainprints make it enticing to study the factors and the underlying mechanism for identification. We looked at the performance of these features between two sessions of different types collected on the same day to test their consistency between different brain states. We compared the features using the Human Connectome Project (HCP) MEG data [5] between a resting-state session in which individuals (N=77) rest and do not perform a task and a task-MEG session where these same individuals view images and perform a working memory task.

Consistent with the cross-session results in Figure 4, **sp** yielded a high identification accuracy (Figure 4 (b), 0.77 ± 0.0034, mean $\pm$ SE, $p < 0.0002$), well above the 0.013 random baseline. This suggests that the spatial fingerprint is consistent between different brain states which confirms a similar finding in fMRI [15]. The by-individual identification accuracy (Figure 5(c)) shows that there was a small subset of individuals

whose accuracy is below random, which may be due to the lack of head position correction in the HCP collection protocol. **tp** and **fq** did not perform as well as **sp**, suggesting that the temporal rhythm and frequency involved might be different between resting-state and task [29, 30].

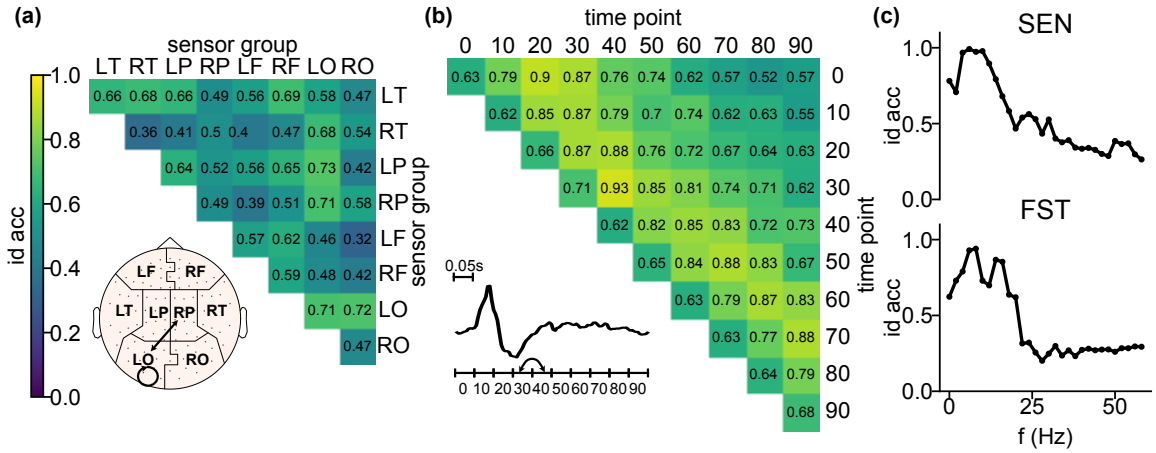

Figure 6: **Identification accuracy of components of the features.** See Supplement D for **a,b** on FST data. **(a)** Identification accuracy of the sub-features of **sp** on SEN data. Each grid represents the identification accuracy using the corresponding entries of **sp** averaged across cross-sessions ($N = 6$), individuals ($N = 4$), and identification runs ($N = 100$). Inset is the plot of the sensor group layout and edges correspond to the sensor group pair with over 0.7 accuracy for both FST and SEN. The topomap is plotted using the python MNE package [31]. **(b)** Identification accuracy of the sub-features of **tp** on SEN data. Each grid represents the identification accuracy using the corresponding entries of **tp** averaged across the same dimensions as in **a**. Inset is an example MEG signal of one individual averaged across channels ($N = 102$) and trials ($N = 1000$). Arrows correspond to the entries of the heatmap with over 0.9 accuracy for both FST and SEN. **(c)** Identification accuracy of the sub-features of **fq** on SEN (upper plot) and FST (lower plot) data. Each dot represents the identification accuracy using the corresponding entries of **tp** averaged across cross-sessions ($N = 6$ for SEN and 12 for FST), individuals ($N = 4$), and identification runs ($N = 100$). Accuracy values of f larger than 60 Hz were truncated since the curve became flat. Error bars are SE across cross-sessions, individuals, and identification runs and are invisible due to small values. The curve peaks at $f = 6$ Hz for SEN and $f = 8$Hz for FST. The accuracy of some components of a feature is consistently higher than the rest on both datasets, indicating that some parts of a certain feature may be more important in identifying individuals.

The rank accuracy of **tp** and **fq** (Figure 5(b), 0.82 ± 0.0017 and 0.85 ± 0.0016, mean $\pm$ SE, $p < 0.0002$ for all) are much higher than the baseline (0.506). The majority of

the individuals also have higher rank accuracy than baseline for **tp** and **fq** (Figure 5(c)). The higher rank accuracy suggests that **tp** and **fq** may still contain individual-specific information but are not strong enough to achieve a high identification accuracy. Since the individuals perform different tasks on the source and target session, the rank accuracy indicates the potential consistent brainprint the generalizes beyond the task. It is noticeable that for the HCP dataset, the recording sessions of one individual were recorded on the same day. Hence one may exercise caution when extend the conclusions to cross-session datasets.

Not every part of a brainprint is equally important. What contributes to the high identifiability of the three brainprints? Understanding the relative contribution of the components of brainprints could help understand individual identifiability and variability. We divided the three brainprints into sub-features and looked at their identification accuracy to see which components contain the most individual-specific information. **sp** was divided into correlations between groups of sensors. **tp** was divided into correlations between time intervals. **fq** was divided into frequencies within a sliding window. We use the SEN and FST dataset to focus on cross-session patterns.

For both SEN and FST, the correlations between sensors within Left Occipital (LO) and between LO and Right Parietal (RP) yielded high accuracy (Figure 6 (a), inset, and Supplement D). LO is involved in visual processing [32] and RP is involved in sensory integration [33], both of which are functions recruited by the experimental task. Due to the nature of the sampled signal and the physical properties of the skull, each MEG sensor samples coarsely from the brain, making it hard to say whether MEG spatial correlation effectively corresponds to functional connectivity, especially for nearby sensors [8]. However, the fact that correlations between faraway groups of sensors, for example, LT and RT, still have good accuracy suggesting it may be due to actual functional correlation between these areas, but it could still be the case that it is the difference in skull shapes that contributes to the high **sp** accuracy.

For both SEN and FST, the super-diagonal of the heat map for temporal sub-features (Figure 6 b and Supplemental Section D) had high accuracy. The super-diagonal entries correspond to the cross-correlation of the MEG signal between two consecutive segments of 0.05 s. Hence the rhythm of the signal within a short segment of time contributes to identifiability, which can also be seen from the banded structure of **tp** (Figure 2(c)). Moreover, the correlations between fourth and fifth 0.05 s yield

considerably high accuracy on both datasets **tp** (Figure 6(b) inset). These time periods overlap with the time we expect the brain is processing word and picture stimuli [34].

The power intensity of frequencies between 4 and 13 Hz yielded the highest accuracy on both SEN and FST data (Figure 6 c), the peak is 6 Hz for SEN adn 8 Hz for FST. These peaks roughly corresponds to the Theta and Alpha frequency band which are related to the resting state, memory, and mental coordination [35]. The accuracy is also moderately high on part of Beta band (14-31 Hz) where attention and concerntration are recruited[35].

Identifiability changes with data size and preprocessing. The last dimensions that we investigate is the dependence of individual identification on the amount of available data and on the level of data preprocessing.

We look at the identification accuracy using the three brainprints while increasing the sample size n. The identification accuracy increases with the amount of data used for computing **sp**, **fp**, and **fq** (Figure 7(a)) as the sampling variance becomes smaller. In general, with 50s of data, the brainprints perform well on cross-session identification of the same task. **sp** becomes reasonably accurate on the HCP dataset with 100 trials corresponding to 250s of recording, possibly because more trials are required to accurately compute features that are distinguishable within a larger pool of individuals. For FST and SEN, the identification accuracy of **sp** saturates at fewer number of trials than **tp** and **fq**. It is possible that **sp** requires fewer trials to be estimated robustly.

Preprocessing may also affect identification accuracy. We compared the difference in the identification and rank accuracy between the raw and preprocessed data (Figure 7 b,c). The changes in accuracy were all statistically significant (Figure 7(b,c), $p < 10^{-26}$, two-sided paired t-test) when the raw data was preprocessed for all the three features (Figure 7(b)). For both FST and SEN, preprocessing yielded better accuracy for **tp** and **fq**. However, for **sp**, the results point in opposite directions: preprocessing increases identifiability for FST and decreases it for SEN. There was one difference in the preprocessing pipeline for both datasets: FST preprocessing did not include head position correction due to a lack of head position recordings. Head position correction might be changing the signal in non-homogenous ways thereby undermining the identifiability with **sp**.

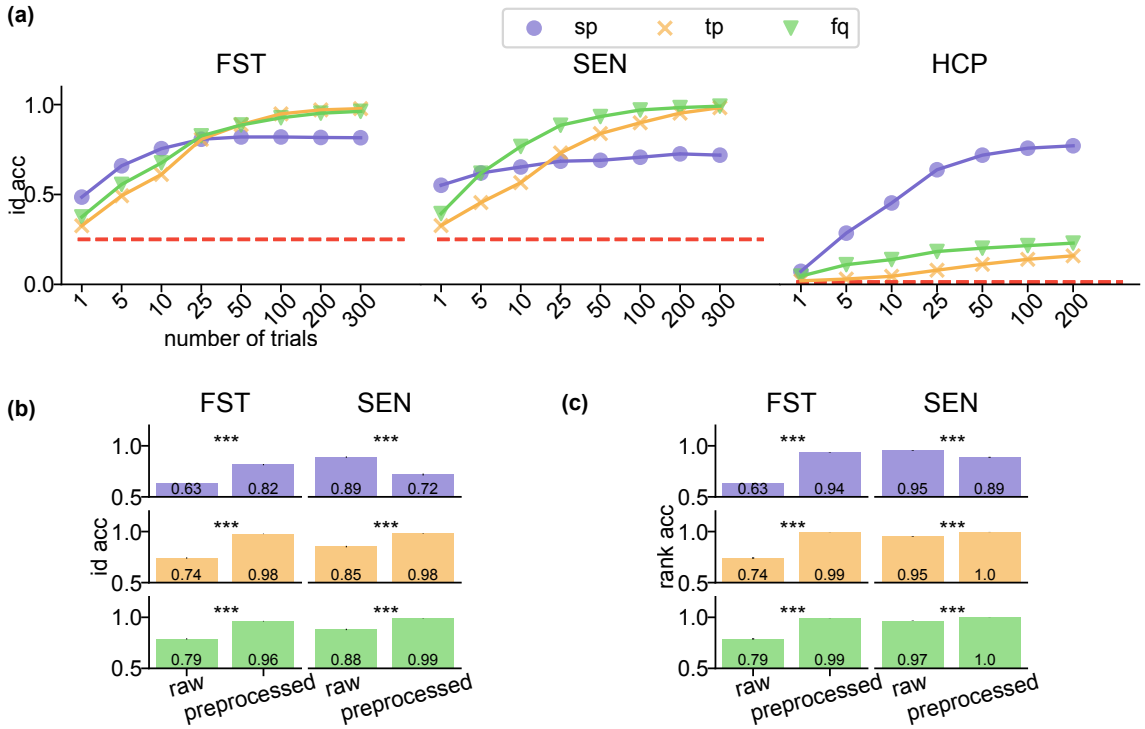

Figure 7: **Factors affecting identification accuracy. (a)** Identification accuracy with respect to the number of trials (sample size) used for the featurization of FST, SEN, and HCP data. Each dot represents the identification accuracy averaged across individuals, identification runs, and cross-sessions (or cross-task sessions) excluding the within-session or within-task results. Error bars are the SE across the corresponding cross-sessions (or cross-task sessions), individuals, and identification runs of each dataset and are invisible due to small values. **(b)-(c)** Identification **(b)** and rank **(c)** accuracy of the three features computed on raw and fully preprocessed FST and SEN data. The same color represents the same feature as in **(a)**. For **(b)**, the identification accuracy across sessions ($N = 12$ for FST and $= 6$ for SEN), individuals ($N = 4$), and identification runs ($N = 100$) are put into one vector (of $N = 4800$ entries for FST and 2400 entries for SEN) for each feature and level of preprocessing. The heights of the bar plots are the mean of the corresponding vector and the error bars are its SE and are invisible due to small values. A two-sided paired t-test is performed on the binary vectors of the same feature and dataset between the raw and preprocessed data. The p-values for all pairs are less than 0.001. For **(c)**, the rank accuracy were put into one vector in the same way as in **(b)**. The heights of the bar plots are the mean of the corresponding vector and the error bars are its s.d. A two-sided paired t-test is performed on the vectors of the same feature and dataset between the raw and preprocessed data. The p-values for all pairs are less than 0.001.

Discussion

An individual can be identified with a number of differential characteristics, including their real 'fingerprints'. Existing studies have suggested the existence a fingerprint in brain signals (e.g. [14, 15]). In this paper, we argued that such brainprints also exist in MEG data and, in fact, there are multiple of them that capture different information from the MEG data. We showed that these brainprints are likely not by-products of environmental artifacts and may pertain to the underlying brain response to stimuli. These analyses, apart from adding to the existing evidence of the brainprints, may bear alarming meanings in privacy issues and provoke thoughts on how scientific conclusions based on multiple individuals have to be examined carefully given these consistent individual-specific features.

In this section, we first discuss the implications of these results in detail, following the same order of the previous section. We then mention limitations and potential improvement to our analysis of brainprints.

Within-session identifiability. Using the HP data, we showed that both random forest classification with topomaps and 1NN classification with certain interpretable features can be used to correctly identify individuals when the data is collected on a single session. The high accuracy based on merely 0.5s of data for **sp** and 25s for **tp** and **fq** is striking since small amounts of data usually leads to inaccurate estimates of these features, unless the underlying patterns are strong. The easy task of identifying individuals on single-session dataset points to strong individual-specific patterns which may or may not be brain-activity related.

Uniqueness of brainprints. The three features we proposed may not be the only characteristics of MEG data that can be used for individual identification. However, these features represent fundamental aspects of MEG data (and even time series in general) hence they may be a vital first step to understand brainprints. Specifically, we propose the temporal feature, **tp**, because of the high temporal-resolution of MEG data. This feature may have not been used for other types of neuroimaging datasets, suggesting that different features may be informative depending on the nature of dataset of interest.

Cross-session identifiability. The high cross-session identification accuracy using **sp** confirms it is a brainprint, and supports the previous literature on the similar features in fMRI and EEG [15, 36]. The higher accuracy by **sp**, **tp** and **fq** suggest that multiple aspects of the individual activity captured by MEG may be used for identification. The generally lower accuracy from **sp** might be the result of the change of alignment of sensors for each individual. However, since necessary steps have taken in the preprocessing pipeline to align the sensors (Supplement A) and each MEG sensor measures brain activity from a non-trivially large area, it remains unclear if the issue is the alignment. Another interpretation of this result is that the temporal and frequency information is more consistent for an individual across time and the spatial information may slowly evolve over time (e.g., when the individual slowly moves during the recording).

Some source-target session pairs have lower identification accuracy than others for **sp** (Figure 4 (b),(e)) and the identification accuracy is not necessarily reciprocal, for example, 0.76 vs 1 (mean, session 3 as source, session 2 as target vs session 2 as source, session 3 as target) in FST. The lower accuracy of **sp** of on specific source-target sessions of specific individuals suggests that the identifiability of **sp** may not be uniform over time and individuals.

The three highly identifiable features on FST and SEN represent an alarming message for experimentalists to consider before releasing MEG data. The existence of brainprints are also examples of certain functions of the MEG data with high cross-individual variability preserved across sessions, which has been widely discussed on various types of neuroimaging data [8, 37, 11, 38, 39]. For example, the high accuracy with **tp** suggests the existence of individual variability in their temporal response to the same stimuli. Understanding brainprints will facilitate the understanding of the underlying anatomical and functional variability between individuals.

Cross-task identifiability. The consistent performance of **sp** on the HCP data is in line with a previous study on fMRI of overlapping individuals that the spatial connectome is preserved between tasks [15]. The rank accuracy of **tp** and **fq** on HCP data indicates the potential of these two features to be consistent within individuals (Figure 5(c)) because the majority of individuals still have higher than chance rank accuracy than identification accuracy. The current underperformance of these two features, as expected, is likely due to the different temporal dynamics between the

resting and task data. This difference may be eliminated by removing the trial part from the task MEG, focusing on inter-trial intervals or baseline periods, and hence boost the identification accuracy of **tp** and **fq**. More complicated matching method may be proposed to further boost the performance of these two brainprints. The within-task identification accuracy (Figure 4 (a)), on the other hand, is still high for all features. With the large pool of participants, the high accuracy confirms the strong individual-specific information contained in the three features within a certain task.

Interpretability of brainprints. For the three brainprints, higher accuracy seems to be associated with the components of features with more stimuli-driven activity: the occipital lobe, the time around the stimulus, and frequency bands the with highest power intensities (Supplement E). Indeed, MEG signal is most sensitive to transient, coordinated firings of many neurons that happen after stimulus onset. This commonality indicates the possibility that higher accuracy is related to event-related signals, which in turn suggests that identifiability might be caused by different individuals responding differently to the stimulus. This dependence on stimulus may explain the low accuracy with **tp** and **fq** on HCP data and also suggest that the identifiability originates from brain-related activities instead of session- and individual-specific artifacts.

However, these accuracy patterns of specific components of a feature could also be explained by a signal-to-noise ratio argument: regions, time-points, or frequencies related to stimulus processing correspond to parts of the underlying brain signal with higher amplitudes (while the ambient noise amplitude is constant). It might be that the increase in signal magnitude make the (spatial, temporal or frequency) activity patterns that are specific to a individual more detectable by increasing their amplitude relative to the ambient noise, even if these patterns are not inherently related to stimulus processing and are just consistent features of a individual's brain activity.

Sample size and level of preprocessing. **sp** accuracy tends to saturate with fewer number of trials than the other two featuers on FST and SEN data but with more trials on HCP data (Figure 7(a)). This difference is likely due to the difference in the maximum accuracy a feature can attain: in HCP data, **tp** and **fq** has much lower maximum accuracy and will reach the peak with smaller number of trials. In FST and SEN data, the spatial pattern may require fewer trials to estimate accurately,

as compared to the temporal and frequency features.

The artifact removal and temporal filtering in the preprocessing pipeline might have prevented session-specific noise from contaminating individual-specific features, resulting in higher accuracy for **tp** and **fq**. The seemingly contradictory accuracy on **sp** does not justify our results: identifiability using **sp** increases after prepossessing when not performing head position correction but decreases when performing it. On the one hand, it is expected that head position correction would improve identifiability by recentering each individual's data to the same position in each session. On the other hand, head position correction may remove individual-specific information such as the head shape, causing the decrease in the accuracy of **sp**. Future work and analysis of additional datasets are required to investigate this result. The difference in the accuracy between raw and preprocessed data suggests, for example, encrypting the data with session-specific noise may lower identification accuracy.

Limitations. Due to the availability of the multi-session MEG data, more experiments are needed to generalize the conclusions of this paper to a larger population and more types of tasks. For example, the cross-session identifiability results depend on 4 subjects and may suffer from high variance. A larger population (with multiple sessions for each participant) may benefit the interpretation of brainprints to eventually attribute the high identifiability of certain components of features to the underlying brain mechanism.

Throughout the paper, we assumed both the target and source datasets had the same pool of participants in the scope of this paper. If we don't know if one individual from the target set is included in the source set, other classification methods which allow for abstaining from classification (e.g. [40, 41]) may be used to account for the case when no label in the source set can be assigned to the individual. This situation is an example of a more realistic identification problem because an individual's participation in multiple MEG studies is usually unknown to the public.

Future solutions. More complicated features can be proposed which combine the spatial, temporal, and frequency information to improve identifiability. For example, functional connectivity at different frequency bands has been used to identify twins from other participants [17]. New feature similarity function that focuses on the structure of the correlation matrices may also be used to improve accuracy [42].

Metric learning techniques [43] can also be used to learn the similarity function in a surpervised manner which may boost the performance with sufficient amount of labeled MEG data.

On the other hand, given the high identification accuracy with brainprints in this study, privacy-preserving algorithms need to be proposed to account for this privacy issue. Federated learning method [44] may be a promising framework as data collected from multiple sessions and sites can be analyzed together without revealing critical information of each specific dataset.

Methods

Within- vs cross- session. We call a pair of *source* and *target* sets "within-session" if, for each individual, both datasets were collected in the same visit to the scanner. For example, two blocks of a resting-state recording of a participant collected on the same day are within-session. If the two datasets are collected on different days for each individual, they are "cross-session". For example, a resting state recording on day 1 and another resting-state recording on day 2 are cross-session. Individuals with within-session data may be easier to identify since the source and target data were collected under almost the same environment.

Within-session data. Individuals were asked to read a chapter of Harry Potter [23] while each word was presented for 0.5 s on a screen. There were 306 sensors at 102 locations where each location has one magnetometer and two planar gradiometers whose signal was averaged. The sampling frequency of the data was 1000 Hz which was further downsampled to 200 Hz. Details about the preprocessing of all the datasets in this paper can be found in Supplement A. The data was parsed into trials where each trial corresponds to the MEG recording when an individual was reading a word. Specifically, the trials of individual k is $\{X_i^k \in \mathbb{R}^{102\times 100}\}_{i=1}^{I_k}$ where I_k is the number of trials for individual k, 102 represents the number of spatial channels, and 100 represents the number of temporal points in the trial. Since the recording of each individual was collected in one session, we simply split the data into a target and source dataset for the within-session identification task.

Cross-session data. We considered the following two datasets which have recordings on multiple days:

1- FST data [28], shared online:[1] individuals saw faces with each face appearing on the screen. Each trial lasted 0.5 s. There were 4 individuals and 4 sessions. The sampling frequency was 1000 Hz and was downsampled to 200 Hz. Intervals between consecutive sessions were several days.

2- SEN data (unpublished anonymized citation): individuals read sentences. Each

[1] `https://figshare.com/articles/FST_raw_data/4233107`

trial lasted 0.5 s. There were 4 individuals and 3 sessions. The sampling frequency was 1000 Hz and was downsampled to 200 Hz. Intervals between consecutive sessions ranged from days to weeks. In this dataset, two sessions for two individuals were recorded at the same day.

The shape of one trial of the two datasets is 102 channels by 100 time points, the same as the Harry Potter data. We used 300 trials to create features for each run of identification. For the within-session identification (diagonal entries of Fig 3 (b) (c)), we split the recording for each individual into non-overlapping source and target set before featurization.

Task vs resting data. We looked at the Human Connectome Project data[2][5] . There were two sessions, one resting-state recording and one working-memory (WM) task recording where the stimuli were images for the participants to remember. Each trial of the WM corresponded to the 2.5 s of the recording after the onset of the stimulus. The two datasets had 77 individuals in common and we only looked at these individuals. There were 146 channels and the signal was downsampled to 200 Hz. The two sessions were collected on the same day with a break of several hours. We used 200 trials for featurization for each run of identification due to fewer number of total trials as compared to the aforementioned datasets.

Random forest identification with raw features. We trained a random forest classifier with 256 estimators by first concatenating all the trials of all the individuals along the time dimension, resulting in $\mathbf{X} \in \mathbb{R}^{102\times N}$ where $N = \sum_{k=1}^{8} 100 I_k$ is the total number of time points of all the individuals. The training data is $\{X(:, i) \in \mathbb{R}^{102}\}_{i=1}^{N}$ is a falttened vector with 102 entires corresponding to the signal across all channels at one time point, and the training label is $y_i \in \{1, 2, \cdots, 8\}$. Data was z-scored by channel separately on training and testing data.

Interpretable MEG features. Let $X \in \mathbb{R}^{102\times 100\times n}$ represent the recording used for featurization, with 102 channels, 100 time points, and n randomly sampled trials.

[2] `https://www.humanconnectome.org/study/hcp-young-adult`

The three features were defined as follows:

1- Spatial correlation (**sp**): Pearson correlation between channels averaged over time. X was reshaped into $\mathbb{R}^{102\times 100n}$ before the correlations between rows of the reshaped matrix were computed.

2- Temporal correlation (**tp**): Pearson correlation between time points averaged over channels. X was reshaped into $\mathbb{R}^{100\times 102n}$ before the correlations between rows of the reshaped matrix were computed.

3- Frequency (**fq**): power spectrum averaged over channels. Power spectrum of $X(i,:,j)$ was computed using a Tukey window with shape parameter of 0.25 and window size of 100 time points for $i = 1,\cdots,102, j = 1,\cdots,n$. The final power spectrum was obtained by averaging across i, j.

Identification using 1NN. We performed $R = 100$ identification runs. In identification run r, we randomly split the Harry Potter dataset into non-overlapping source and target set, z-scored the source and target by channel separately, and computed the feature $x^{\alpha}_{i,r,F}$ averaged over $n = 300$ randomly sampled trials using data $\alpha \in \{\text{target, source}\}$ for individual i and $F \in \{\textbf{sp}, \textbf{tp}, \textbf{fq}\}$. The features from the target to the source set were matched with a labeling with replacement protocol :

$$\hat{y}(x^{\text{target}}_{i,r,F}) \;=\; \underset{j\in\{1,2,\cdots,K\}}{\arg\max}\; m(x^{\text{target}}_{i,r,F}, x^{\text{source}}_{j,r,F})$$

where $K = 8$ is the total number of individuals and $m(\cdot,\cdot)$ is the similarity function measuring the similarity between the two features. We used Pearson correlation as our similarity function. The identification accuracy for individual i and feature F is $\frac{1}{R}\sum_{r=1}^{R}\mathbb{1}_{\hat{y}(x^{\text{target}}_{i,r,F})=i}$. The averaged identification accuracy for feature F is $\frac{1}{KR}\sum_{i=1}^{K}\sum_{r=1}^{R}\mathbb{1}_{\hat{y}(x^{\text{target}}_{i,r,F})=i}$. The random baseline is $\frac{1}{K}$.

When the source set and target set were from the same session, we split the dataset into non-overlapping sets as we did in the within-session identification. We didn't split data when the source and target data are from different sessions since there is no potential data leakage. We z-scored the data by channel on the source and target separately.

Rank accuracy. The rank accuracy of individual i on one run of identification (suppressing notations of feature F and run r) is defined as $\frac{1}{K}\text{rank}(m(x_i^{\text{target}}, x_i^{\text{source}}))$ where K is the number of individuals, $\text{rank}(m(x_i^{\text{target}}, x_i^{\text{source}}))$ is over $\{m(x_i^{\text{target}}, x_j^{\text{source}}), j = 1, 2, \cdots, K\}$. The rank accuracy equals to 1 if the feature of the same individual has the largest similarity between the source and target sets among all K individuals, and is $\frac{1}{K}$ if the similarity is the smallest. The rank accuracy captures more information in a failure case where an individual is mis-identified. The random baseline for the rank accuracy is $\frac{K+1}{2K}$.

Sub-features. Each feature was decomposed as follows:

1- **sp**: The sensors were partitioned into 8 subgroups according to the map in Figure 1 of [45]: Left Frontal (LF), Right Frontal (RF), Left Temporal (LT), Right Temporal (RT), Left Parietal (LP), Right Parietal (RP), Left Occipital (LO), Right Occipital (LO). Each subfeature was the rows and columns of the spatial correlation matrix corresponding to the sensors in one of the eight groups: let $\Sigma_s \in \mathbb{R}^{102\times102}$ be the spatial correlation matrix, then the subfeature corresponding to the correlation between RT and LT, for example, is $\Sigma_s(ind_{RT}, ind_{LT})$ where ind_{RT} is the set of channel indices in the RT group and ind_{LT} corresponds to the LT group.

2- **tp**: The 100 temporal points were divided into 10 consecutive segments containing 10 time points. Each subfeature was the rows and columns of the temporal correlation matrix corresponding to one of the ten segments: let $\Sigma_t \in \mathbb{R}^{100\times100}$ be the spatial correlation matrix, then subfeature corresponding to the correlation between the first and second time segment, for example, is $\Sigma_t(1:10, 11:20)$.

3- **fq**: Each subfeature was the segment of the frequency feature vector corresponding to $[f, f+10]$ Hz where $f \in \{0, 2, \cdots, 90\}$ Hz.

Raw vs preprocessed data. In Figure 6 (b), we compare the identification accuracy between the raw and preprocessed data for FST and SEN dataset. The details of the full preprocessing pipeline is included in the supplement. In FST dataset, for a given feature and dataset, there were 4800 binary matching results (12 cross-session comparisons $\times$ 4 individuals $\times$ 100 identification runs) where each one corresponds to the result of deciding which individual from the source session matches the one individual from the target session. A Pearson's χ^2 test was performed to determine

if there is a significant difference in the identification accuracy between the raw and preprocessed data. In SEN dataset, for a given feature and dataset, there were 2400 binary matching results (6 cross-session comparisons $\times$ 4 individuals $\times$ 100 identification runs) instead.

References

[1] Thomas Landrain, Morgan Meyer, Ariel Martin Perez, and Remi Sussan. Do-it-yourself biology: challenges and promises for an open science and technology movement. *Systems and synthetic biology*, 7(3):115–126, 2013.

[2] Krzysztof Gorgolewski, Oscar Esteban, Gunnar Schaefer, Brian Wandell, and Russell Poldrack. Openneuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for human brain mapping. Vancouver, Canada*, 1677(2), 2017.

[3] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[4] Halbert L Dunn. Record linkage. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416, 1946.

[5] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[6] Jaewoo Lee and Chris Clifton. Differential identifiability. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1041–1049, 2012.

[7] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178, 2009.

[8] Giles L Colclough, Mark W Woolrich, PK Tewarie, Matthew J Brookes, Andrew J Quinn, and Stephen M Smith. How reliable are meg resting-state connectivity metrics? *Neuroimage*, 138:284–293, 2016.

[9] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

[10] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews neuroscience*, 18(2):115, 2017.

[11] Vincent Wens, Mathieu Bourguignon, Serge Goldman, Brice Marty, Marc Op De Beeck, Catherine Clumeck, Alison Mary, Philippe Peigneux, Patrick Van Bogaert, Matthew J Brookes, et al. Inter-and intra-subject variability of neuromagnetic resting state networks. *Brain topography*, 27(5):620–634, 2014.

[12] M Poulos, M Rangoussi, N Alexandris, and A Evangelou. Person identification from the eeg using nonlinear signal classification. *Methods of information in Medicine*, 41(01):64–75, 2002.

[13] Lan Ma, James W Minett, Thierry Blu, and William SY Wang. Resting state eeg-based biometrics for individual identification using convolutional neural networks. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2848–2851. IEEE, 2015.

[14] Xiang Zhang, Lina Yao, Salil S Kanhere, Yunhao Liu, Tao Gu, and Kaixuan Chen. Mindid: Person identification from brain waves through attention-based recurrent neural network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–23, 2018.

[15] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664, 2015.

[16] Oscar Miranda-Dominguez, Brian D Mills, Samuel D Carpenter, Kathleen A Grant, Christopher D Kroenke, Joel T Nigg, and Damien A Fair. Connectotyping: model based fingerprinting of the functional connectome. *PloS one*, 9(11), 2014.

[17] M Demuru, AA Gouw, A Hillebrand, CJ Stam, BW Van Dijk, P Scheltens, BM Tijms, E Konijnenberg, M Ten Kate, A Den Braber, et al. Functional and effective whole brain connectivity using magnetoencephalography to identify monozygotic twin pairs. *Scientific reports*, 7(1):1–11, 2017.

[18] Andrew James Anderson, Kelsey McDermott, Brian Rooks, Kathi L Heffner, David Dodell-Feder, and Feng V Lin. Decoding individual identity from brain activity elicited in imagining common experiences. *Nature communications*, 11(1):1–14, 2020.

[19] Christian Wachinger, Polina Golland, William Kremen, Bruce Fischl, Martin Reuter, Alzheimer's Disease Neuroimaging Initiative, et al. Brainprint: A discriminative characterization of brain morphology. *NeuroImage*, 109:232–248, 2015.

[20] Blair C Armstrong, Maria V Ruiz-Blondet, Negin Khalifian, Kenneth J Kurtz, Zhanpeng Jin, and Sarah Laszlo. Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for erp biometrics. *Neurocomputing*, 166:59–67, 2015.

[21] Christian Wachinger, Polina Golland, and Martin Reuter. Brainprint: Identifying subjects by their brain. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 41–48. Springer, 2014.

[22] Debadatta Dash, Paul Ferrari, and Jun Wang. Spatial and spectral fingerprint in the brain: Speaker identification from single trial meg signals. In *INTERSPEECH*, pages 1203–1207, 2019.

[23] JK Rowling. *of Book: Harry Potter and the Sorcerer's Stone*. Year: Scholastic Press/1997, 1997.

[24] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[25] Sofie S Meyer, James Bonaiuto, Mark Lim, Holly Rossiter, Sheena Waters, David Bradbury, Sven Bestmann, Matthew Brookes, Martina F Callaghan, Nikolaus Weiskopf, et al. Flexible head-casts for high spatial precision meg. *Journal of neuroscience methods*, 276:38–45, 2017.

[26] H-RM Tan, Joachim Gross, and PJ Uhlhaas. Meg—measured auditory steady-state oscillations show high test–retest reliability: A sensor and source-space analysis. *Neuroimage*, 122:417–426, 2015.

[27] Douglas Owen Cheyne and Paul Ferrari. Meg studies of motor cortex gamma oscillations: evidence for a gamma "fingerprint" in the brain? *Frontiers in human neuroscience*, 7:575, 2013.

[28] Mark D Vida, Adrian Nestor, David C Plaut, and Marlene Behrmann. Spatiotemporal dynamics of similarity-based neural representations of facial identity. *Proceedings of the National Academy of Sciences*, 114(2):388–393, 2017.

[29] Sheng Zhang and Chiang-Shan R Li. Task-related, low-frequency task-residual, and resting state activity in the default mode network brain regions. *Frontiers in psychology*, 3:172, 2012.

[30] Omer Grigg and Cheryl L Grady. Task-related effects on the temporal and spatial dynamics of resting-state functional connectivity in the default network. *PloS one*, 5(10), 2010.

[31] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.

[32] Tony Ro, Bruno Breitmeyer, Philip Burton, Neel S Singhal, and David Lane. Feedback contributions to visual awareness in human occipital cortex. *Current biology*, 13(12):1038–1041, 2003.

[33] Oliver Jakobs, Robert Langner, Svenja Caspers, Christian Roski, Edna C Cieslik, Karl Zilles, Angela R Laird, Peter T Fox, and Simon B Eickhoff. Across-study and within-subject functional connectivity of a right temporo-parietal junction subregion involved in stimulus–context integration. *Neuroimage*, 60(4):2389–2398, 2012.

[34] Riitta Salmelin. Clinical neurophysiology of language: the meg approach. *Clinical Neurophysiology*, 118(2):237–254, 2007.

[35] Pravat K Mandal, Anwesha Banerjee, Manjari Tripathi, and Ankita Sharma. A comprehensive review of magnetoencephalography (meg) studies for brain functionality in healthy aging and alzheimer's disease (ad). *Frontiers in Computational Neuroscience*, 12:60, 2018.

[36] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[37] Pilar Garcés, María Carmen Martín-Buro, and Fernando Maestú. Quantifying the test-retest reliability of magnetoencephalography resting-state functional connectivity. *Brain connectivity*, 6(6):448–460, 2016.

[38] Kristina T Legget, Allison K Hild, Sarah E Steinmetz, Steven T Simon, and Donald C Rojas. Meg and eeg demonstrate similar test-retest reliability of the 40 hz auditory steady-state response. *International Journal of Psychophysiology*, 114:16–23, 2017.

[39] María Carmen Martín-Buro, Pilar Garcés, and Fernando Maestú. Test-retest reliability of resting-state magnetoencephalography power in sensor and source space. *Human brain mapping*, 37(1):179–190, 2016.

[40] Corinna Cortes, Giulia De Salvo, and Mehryar Mohri. Boosting with abstention. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1668–1676, 2016.

[41] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeffrey Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Knows when it doesn't know: Deep abstaining classifiers. 2018.

[42] Manasij Venkatesh, Joseph Jaja, and Luiz Pessoa. Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *NeuroImage*, 207:116398, 2020.

[43] Brian Kulis et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012.

[44] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[45] Yegang Hu, Chunli Yin, Jicong Zhang, and Yuping Wang. Partial least square aided beamforming algorithm in magnetoencephalography source imaging. *Frontiers in neuroscience*, 12:616, 2018.

[46] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open source scientific tools for python. 2001.

[47] Samu Taulu and Matti Kajola. Presentation of electromagnetic multichannel data: the signal space separation method. *Journal of Applied Physics*, 97(12):124905, 2005.

[48] Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in meg measurements. *Physics in Medicine & Biology*, 51(7):1759, 2006.

[49] Mikko A Uusitalo and Risto J Ilmoniemi. Signal-space projection method for separating meg or eeg into components. *Medical and Biological Engineering and Computing*, 35(2):135–140, 1997.

Supplementary Material

A. Data preprocessing

Here we list the preprocessing steps applied to the four types of datasets: Harry Potter (HP), SEN, FST, and Human Connectome Project (HCP). A summary is listed in table 1. For all datasets, we used an order 8 Chebyshev type I anti-aliasing filter in Python Scipy package[46] for downsampling. For any within-session identification task, data was z-scored within its corresponding type of dataset (target vs source). Some steps of preprocessing were performed using the python MNE package [31].

1- **HP/SEN**: The 306-channel Elekta Neuromag system was used for the recording. Source-space separation (SSS) along with Maxwell filtering and their temporal extension (tSSS) [47, 48] were used for bad channel correction, head position correction, and electromagnetic artifacts removal. Empty room artifacts were removed. $1 \sim 150$ Hz bandpass filter and 60 & 120 Hz notch filter were used to remove line noise. Heartbeats and eyeblinks artifacts were removed with signal-space projection (SSP) [49]. The data was downsampled to 200 Hz and z-scored by channel within each individual and session.

2- **FST** (preprocessing pipeline was included in the source code): The 306-channel Elekta Neuromag system was used for the recording. Source-space separation (SSS) along with Maxwell filtering and their temporal extension (tSSS) were used for bad channel correction and electromagnetic artifacts removal. Empty room artifacts were removed. We didn't perform head position correction since there was no head position data. $1 \sim 150$ Hz Bandpass filter and 60 & 120 Hz Notch filter were used to remove line noise. Heartbeats and eyeblinks artifacts were also removed with SSP. The data was downsampled to 200 Hz and z-scored by channel within each individual and session.

3-**HCP**: Both resting and WM datasets were already preprocessed and downloaded from the HCP database[3]. The details of the preprocessing pipeline can be found at `https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf`. MAGNES 3600 (4D Neuroimaging, San Diego, CA) system was used for the recording. For WM data, we looked at the TIM partition which corresponds to $-1.5 \sim 2.5$ s relative to

[3] `https://www.humanconnectome.org/study/hcp-young-adult`

the onset of the image. For both resting and WM data, the sampling frequency of the preprocessed data is 508.63 Hz, and 2 s of data were selected from each trial. This corresponds to the whole 1018 time points in the resting data and $[763 : 1780]$-th time point for the WM data (corresponding to $0 \sim 2$ s relative to the onset of the image). The 2 s data was then downsampled to 101.73 Hz. Data was z-scored by channel within each individual and each data type (resting and WM). We looked at the 146 channels which were marked "good" among all the 77 overlapping individual between resting and WM.

Table 1: Summary of the preprocessing stpes for HP, SEN, FST, and HCP data

Steps	**HP/SEN**	**FST**	**HCP**
bad data	corrected	corrected	removed
head position	corrected	not corrected	not corrected[4]
electromagnetic artifacts	removed using SSS	removed using SSS	removed with bad data
empty room artifacts	removed	removed	removed[5]
band filtering	$1 \sim 150$ Hz	$1 \sim 150$ Hz	$1.3 \sim 150$ Hz
notch filtering	60 & 120 Hz	60 & 120 Hz	$59 - 61$&$119 - 121$ Hz
ECG (heartbeat) artifacts	removed with SSP	removed with SSP	removed with ICA
EOG (eyeblink) artifacts	removed with SSP	removed with SSP	removed with ICA
downsampling	200 Hz	200 Hz	101.73 Hz
z-scoring	by channel within individual, session	same	same
shape of a trial $[channels, timepoints]$	$[102, 100]$	$[102, 100]$	$[146, 204]$

[4]No continuous recording of head position was available in HCP data

[5]page 68 of `https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP_S1200_Release_Reference_Manual.pdf`

B. Sensor layout for FST, SEN, and HP data

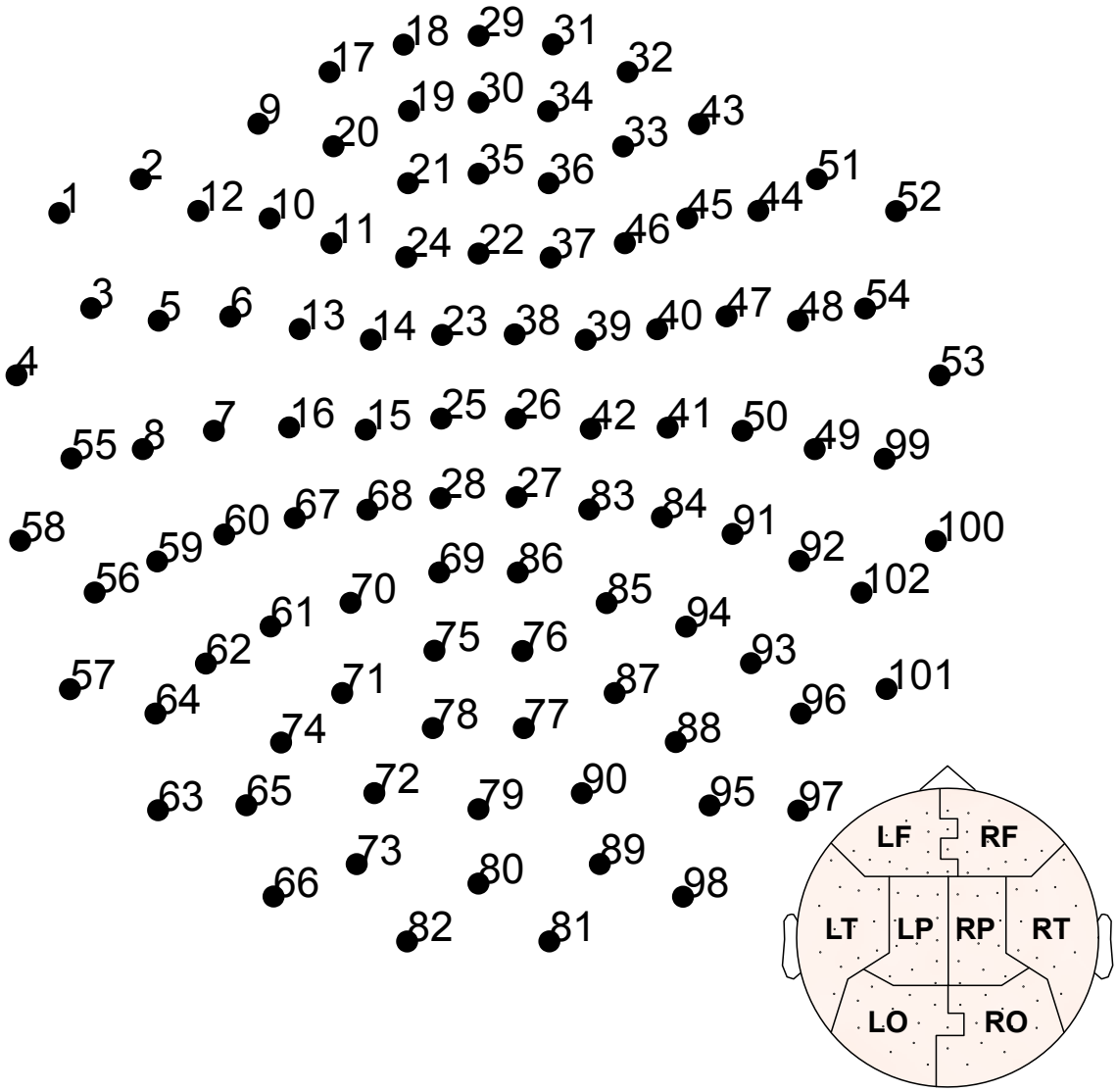

Figure 8: **Layout of the sensors for FST, HP, and SEN data (306-channel Elekta Neuromag system)**. Channel numbers are consistent with the channel index in Figure 11. Inset is the partitioning of the sensors same as Figure 6 (a) of the main text.

C. Identification accuracy vs. sample size for Harry Potter data

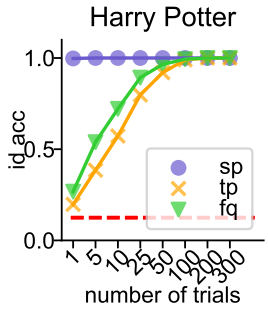

Figure 9: **Identification accuracy of sp,tp, and fq on the Harry Potter data.** Each dot was averaged across individuals (8) and identification runs (100). Error bars are the SE across individuals and identification runs and are invisible due to small values. Each trial is 0.5s in length. The trends for **tp** and **fq** are similar to that of the cross-session data (SEN and FST). **sp** requires as few as one trial to achieve a perfect accuracy. This indicates strong spatial patterns in the HP data which are specific to each individual. This is expected since HP does not have more than one session, and the identification accuracy for **sp** may be lower if there are multiple sessions in HP data, similar to what we have observed on FST and SEN data.

D. Identification accuracy with components of brainprints for FST data

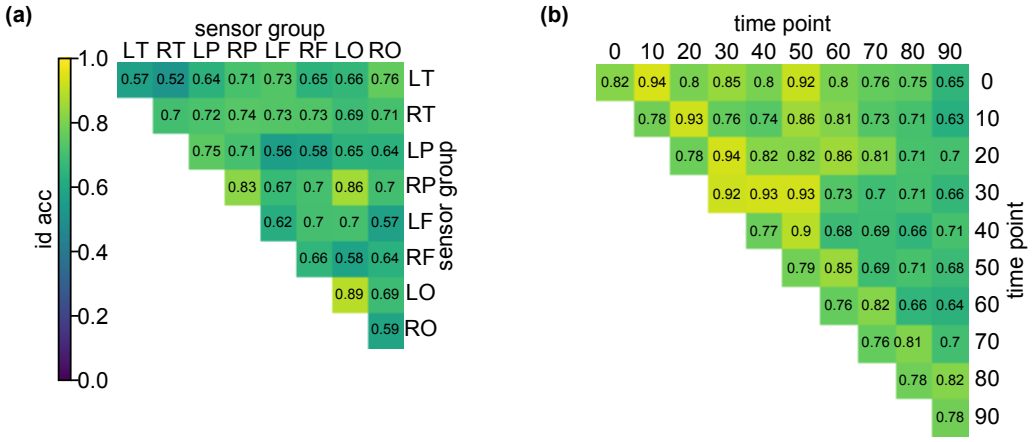

Figure 10: **Identification accuracy with sub features for a: sp and b: tp in FST data (similar to Figure 6 in the main text).** For both FST and SEN, the within-LO and LO-RP correlations yield high identification accuracy. Similarly, for both FST and SEN, the super-diagonal and the correlations between the fourth and fifth 0.05 s yield high accuracy. The consistency of the results on the two datasets suggest that our conclusions in Section 4.2 are not due to experiment-specific artifacts.

E. Example brainprints of FST data

Note: we have emphasized the importance of preserving individual privacy throughout the paper. Since the FST dataset is published online and our way of computing brainprints (as either discussed in the main text or the source code) will eventually be publicly available, showing individual brainprints will not reveal new information about the individuals. Hence we decided to include the following examples of brainprints to show more intuition behind the high identification accuracy of the three brainprints. (.Figures in the next page).

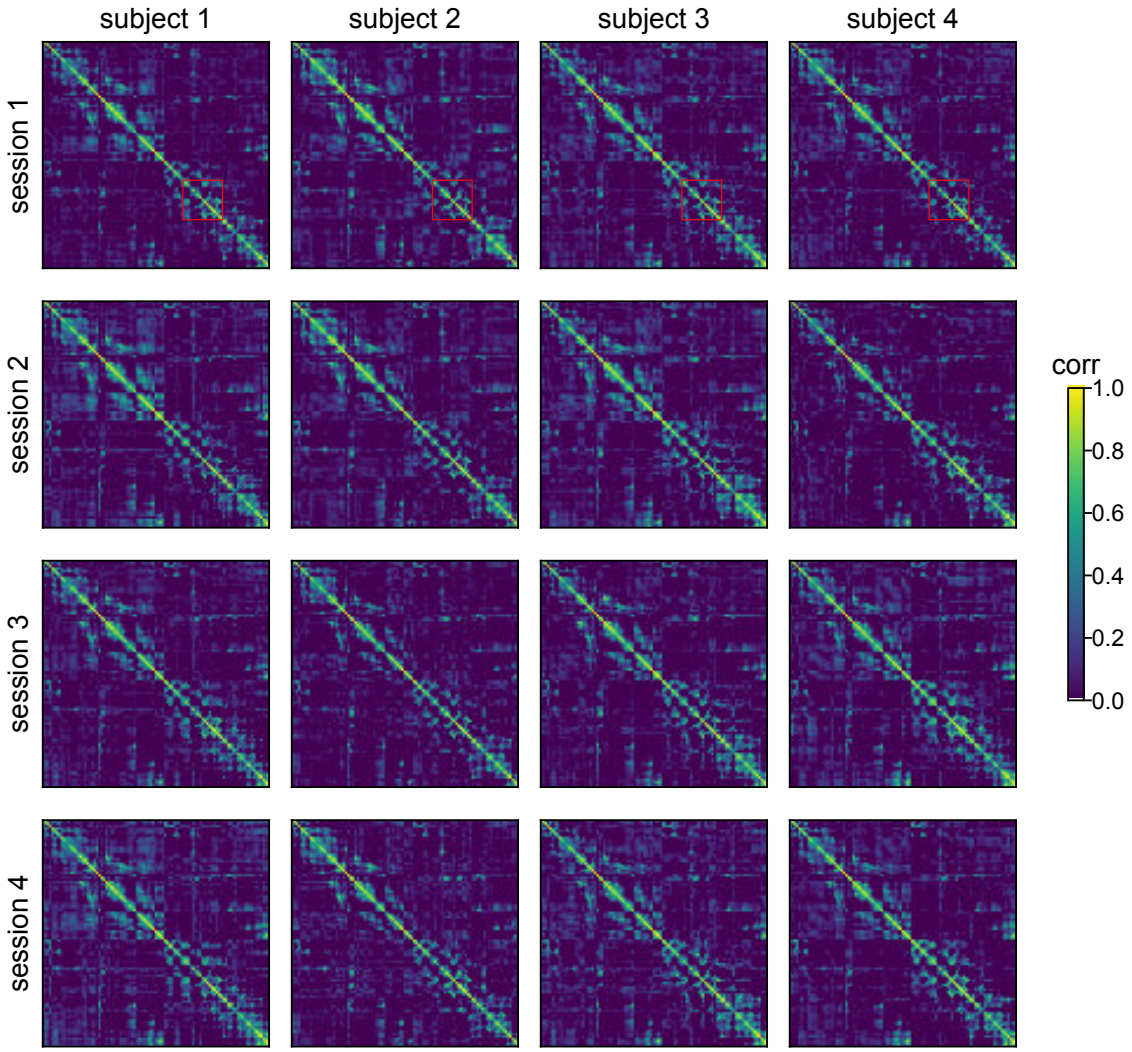

Figure 11: **Example sp (spatial connectivity) of FST data.** Each heatmap represents a 102×102 spatial correlation matrix. For better illustration we clipped the correlation into [0,1]. The general patterns of the correlation matrices are similar to each other. Some subsets of the heatmap, for example, the bottom-right corner, the top-left corner, and the red rectangle areas are more consistent within a individual and different between individuals. This suggests that only the interactions among a subset of sensors are individual-specific. The red rectangle areas, in particular, roughly correspond to the correlations within the left occipital (LO) lobe which yields the highest identification accuracy on both FST and SEN data (see Figure 6(a) and Figure 10). More complicated comparison algorithms may be proposed to focus on these specific subsets to improve the identification accuracy.

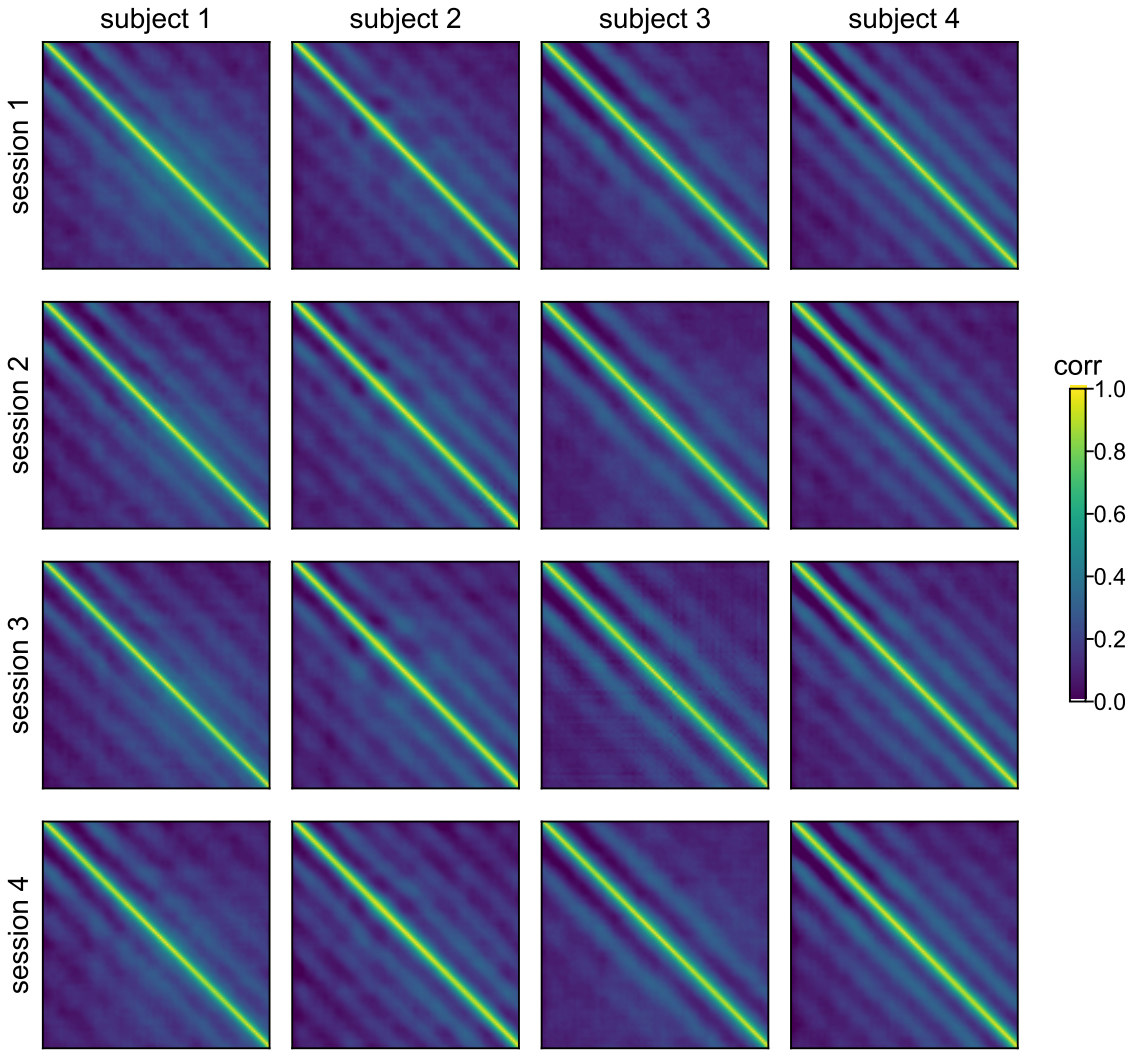

Figure 12: **Example tp (temporal connectivity) of FST data.** Each heatmap represents a 100×100 temporal correlation matrix. For better illustration we clipped the correlation into [0,1]. The banded structure of the matrices are preserved for the same individual across sessions, and are different between individuals in terms of the number of bands and the relative locations of the bands. The banded structure indicates that there are stronger correlations of the signal with itself at certain lags. In other words, looking at the auto-correlation of the signal or even cross-correlation between different channels may reveal interesting results about the temporal dynamics of the brain activities. The individual-specific band structures also confirm the findings in Figure 6 (b) that correlations of the signal with itself at certain lags are best able to identify individuals.

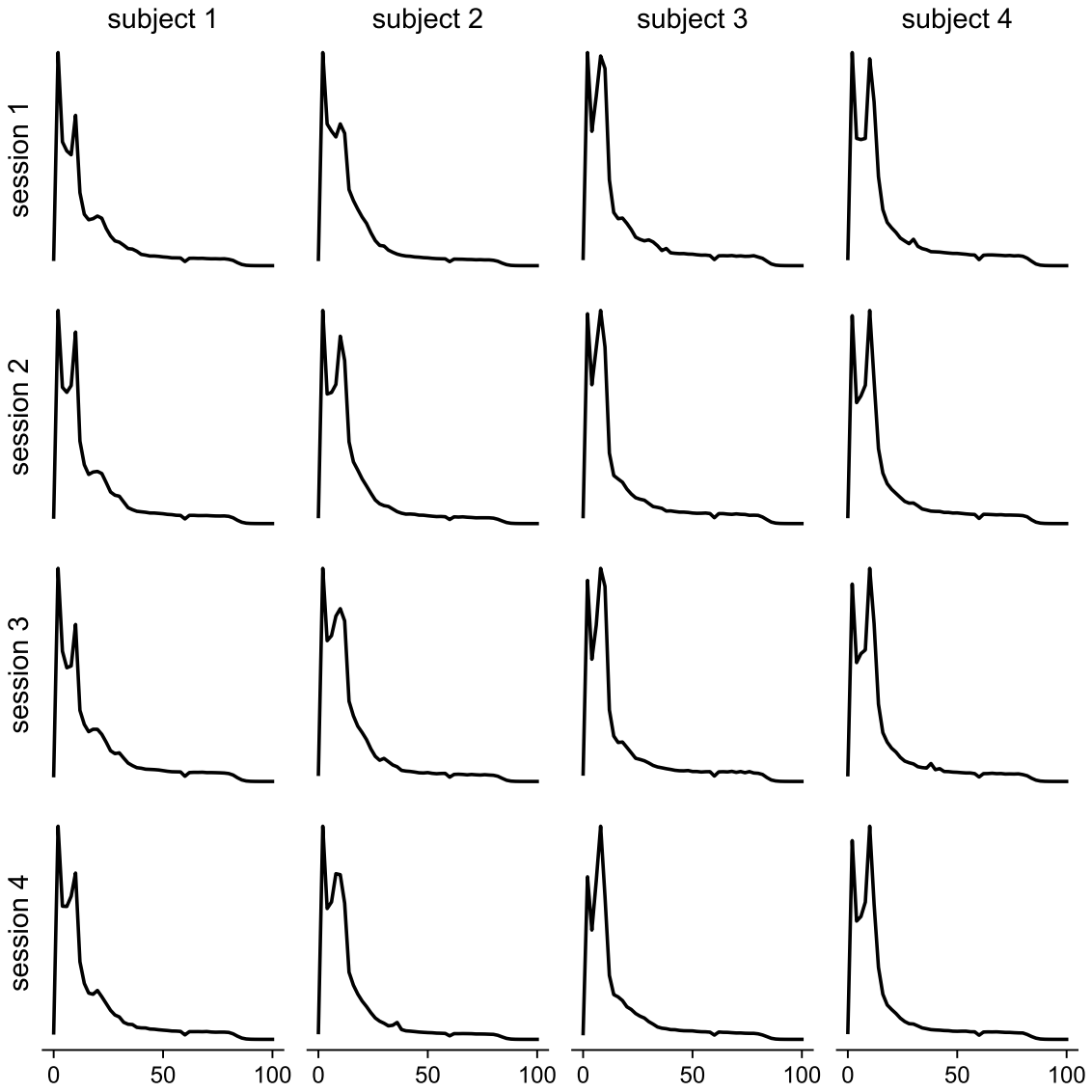

Figure 13: **Example fq (frequency) of FST data.** Each plot represents spectrum (averaged across channels) vs. frequencies (Hz), where the range of frequencies is $[0, 100]$ with a 2 Hz increment. For all individuals, there are two peaks in the power spectrum. The two peaks correspond to around 5 and 10 Hz. The relative height of the two peaks as well as the shape of the curve near the two peaks are consistently unique to an individual across sessions and different across individuals. There are also small peaks near 20 Hz for some individuals. These frequencies with higher amplitudes seem to align with the results shown in Figure 6 (c) where the frequency band near 10 Hz yields the highest identification accuracy. Hence the components of **fq** associated with more stimuli-driven activity or larger signal-to-noise ratio seem to yield better results.

F. Statistical significance of the results

The identification and rank accuracy were averaged across subjects, identification runs, and session pairs. The reported accuracies, being so large, are both statistically and practically significant, and are nearly impossible to attribute to random chance, but accurately quantifying the uncertainty is challenging in our setup. Since featurization for each session of each subject was done before the matching, there is some weak dependence on the accuracy between subjects, session, and identification runs. This dependence makes it hard to analytically obtain a p-value for the accuracy. One numerical alternative is to permute the original recording within each session across subjects before performing matching, but this is computationally expensive as it involves loading and computing large chunks of data 1000s of times. Hence we provide below a (natural, but approximate) permutation-based method for a p-value to test the null that the match is a random guess.

Let $\mathbf{y}^i$ denote the true labels of session i. Note that $\mathbf{y}^i = [1, 2, 3, 4]^T$ for any session. The permutation test is performed as follows:

Algorithm 1: Null distribution for the identification/rank accuracy

```
N_null ← {}: samples for the null distribution
T: number of permutation runs
for t ← 1 to T do
    y_t^i ← permute(y^i), ∀i
    Re-compute the average accuracy, a_t, using {y_t^i}_i
    N_null = N_null ∪ {a_t}
end
return N_null
```

To calculate the p-value, we simply compute $p = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{a \leq a_t}$, where a is the observed average accuracy of a feature across subjects, sessions, and identification runs. Algorithm 1 permutes the labels for each session independently but the permutation remains unchanged for the same source-target pair across identification runs. We summarize the p-values for the identification and rank accuracy of three features on the FST, SEN, and HCP data using $T = 4999$ permutation runs. For all the p-values, since we have not encountered any a_t that exceeds the accuracy number, their values are simply $\frac{1}{T+1} = 0.0002$. We emphasize that even though these p-values are technically only approximate due to some weak dependence, the fact that we did not

see a single permutation which achieved a higher accuracy than ours should convince even rigorous skeptics that it is nearly impossible to explain away our accuracies to chance.

Table 2: Statistical significance for the accuracy numbers

Data	**Feature**	**id acc**	**p-val: id**	**rank acc**	**p-val: rank**
FST	**sp**	0.816	0.0002	0.936	0.0002
FST	**tp**	0.978	0.0002	0.994	0.0002
FST	**fq**	0.962	0.0002	0.991	0.0002
SEN	**sp**	0.719	0.0002	0.888	0.0002
SEN	**tp**	0.983	0.0002	0.996	0.0002
SEN	**fq**	0.991	0.0002	0.998	0.0002
HCP	**sp**	0.771	0.0002	0.974	0.0002
HCP	**tp**	0.159	0.0002	0.819	0.0002
HCP	**fq**	0.229	0.0002	0.845	0.0002