SELF-SUPERVISED EEG PRETRAINING AMPLIFIES INDIVIDUAL DIFFERENCES IN NEURAL REPRESENTATIONS

Laurine Marty
Machine Learning Research Unit & Neuromodulation Institute
TU Wien
Vienna, Austria & Paris, France
laurine.marty@tuwien.ac.at

Michael Pereira
Grenoble Institut Neurosciences
Univ. Grenoble Alpes
Inserm, Grenoble, France.
michael.pereira@univ-grenoble-alpes.fr

Guillaume Bellec
Machine Learning Research Unit
TU Wien
Vienna, Austria
guillaume.bellec@tuwien.ac.at

Joao Barbosa
Neuromodulation Insitute & NeuroSpin
Inserm, Paris, France
palerma@gmail.com

ABSTRACT

Self-supervised learning can be leveraged to build generalized representations from massive and heterogeneous datasets. In the case of neural data, trained models have been shown to generalize across domains, thus sometimes called foundation models. Generalization is typically measured with downstream decoding of behavior, such as movements. Here, we replicate these benchmarks and provide a complementary evaluation to existing benchmarks for NeuroGPT, a foundation model of electroencephalography (EEG). Using simultaneous recording of EEG and BOLD signals over 18 human subjects, we test (1) the extent to which the model captures subject-specific idiosyncrasies, and (2) whether the model EEG embeddings improve the prediction of simultaneously recorded BOLD signals. We find that NeuroGPT embeddings reliably amplify individual idiosyncrasies and improve BOLD prediction relative to raw EEG and randomly initialized models. These results suggest that self-supervised pre-training enriches, rather than homogenizes, subject-level representational structure that can be leveraged to align different data modalities.

1 INTRODUCTION

Large models trained on massive and heterogeneous brain data via self-supervised learning offer a promising new lens to study neural representations (Yuan et al., 2024; Jiang et al., 2024; Chen et al., 2024; Wang et al., 2023; Thomas et al., 2022; Ferrante et al., 2024; Cui et al., 2024; Jiang et al., 2024). These models are sometimes referred as foundation models (Wang & Chen, 2025; Wiggins & Tejani, 2022) and they have shown improvements across behavioral decoding tasks (e.g., motor imagery (Cui et al., 2024) for the NeuroGPT model), speech decoding (Défossez et al., 2023; d'Ascoli et al., 2025) and clinical diagnostic applications (Jiang et al., 2024).

Here, we set out to investigate how *NeuroGPT* (Cui et al., 2024), an EEG foundation model trained with self-supervised learning on 15,000 subjects (Reed et al., 2022), transforms neural representations within a dataset of 20 subjects not included in pretraining (Pereira et al., 2020). In line with previous findings (Cui et al., 2024), we found that *NeuroGPT* embeddings improve decoding of different behavioral conditions (trial type, Table 1). Intuitively, the explanation for this improvement could be that self-supervised pre-training would facilitate generalization by "averaging out" individual variability. This would lead to weakening of the ability to decode subject-specific neural states. Conversely, if pre-training preserves or enriches individual structure, it may enable new forms of personalized alignment, control, and cross-system comparison.

Motivated by this tension, we ask: *Does NeuroGPT preserve subject-specific structure in neural representations, or does it push representations toward a shared canonical geometry?* Here, we tackled this question by comparing decoding subject identity from raw EEG data, EEG data projected on pretrained and non pretrained embeddings of *NeuroGPT*. To evaluate how models treat individual differences in neural representations, we used Euclidean and Procrustes distances to quantify differences in representational geometry across subjects (Williams et al., 2021; Barbosa et al., 2025).

Our results reveal a clear answer: We found that self-supervised learning *amplifies* recognizable idiosyncratic geometries present in raw EEG. In contrast, randomly initialized models compress them. In addition, we show that NeuroGPT embeddings can be leveraged to improve predictions of simultaneously recorded BOLD activity of individual subjects. Notably, this prediction was stronger when performing within subject (rather than across) predictions. This result further suggests that NeuroGPT preserves individual differences. In summary, our main technical contributions are:

- We analyze the representational geometry of NeuroGPT (a foundational model trained with self-supervision) embeddings and show that self-supervised learning *amplifies* individual differences in neural representations. Consistently, pretraining on large datasets improves the k-nearest neighbors classification of subject identity from a dataset not included in pretraining.
- We show that NeuroGPT embeddings improve the prediction of the fMRI signals from simultaneously recorded EEG signals. This is confirmed both using a single EEG-to-fMRI decoder or a subject-specific decoder. We find that subject specific decoders have higher predictive performance, supporting that NeuroGPT embeddings carry subject-specific cross modality information.
- We rule out the trivial explanation that improved performance arises from projecting the data into a (random) higher-dimensional space by comparing pretrained models to their randomly initialized models.

2 METHODS

Dataset. We leveraged a previously published dataset (Fig. 1a). For detailed explanation, we refer the reader to the original publication (Pereira et al., 2020). The dataset consists of simultaneous EEG and fMRI recordings collected from 18 healthy adult participants performing a speeded perceptual decision-making task involving brief numerosity judgments. This dataset also contains EEG-only recordings from 2 additional subjects. On each trial, subjects viewed two arrays of dots flashed for 60 ms and either reported which side contained more dots (active trial type) or observed a computer-generated response matched in accuracy and response time (observation trial type). The dataset includes between 68 and 201 trials per subject, equally split between active and observational conditions.

NeuroGPT. We leveraged a previously published model (Cui et al., 2024). For a detailed explanation, we refer the reader to the original publication. NeuroGPT is a self-supervised EEG foundation model composed of a spatiotemporal EEG encoder and a decoder-only GPT architecture. The model is pre-trained on the large-scale TUH EEG corpus (Obeid & Picone, 2016) using a masked-chunk reconstruction task. The encoder integrates convolutional layers with self-attention to extract denoised, temporally structured EEG features, while the GPT module models long-range dependencies across chunks. Pre-training on EEG recordings from 14,987 different subjects leads the encoder to learn generalizable representations (Cui et al., 2024), which we used without task-specific fine-tuning in our analyses of representational geometry. Here, we remove the decoding head and perform all our analyzes in the latent domain.

Measuring differences in representational geometry. To quantify how different models transform subject-specific neural geometry, we used shape metrics (Williams et al., 2021), recently shown to capture inter-individual differences in neural geometry (Barbosa et al., 2025). Briefly, shape metrics define a distance between neural population representations through explicit geometric alignment transformations and provide a principled metric space for comparing multiple subjects, sessions, or models. For each subject i and a given representation type (raw EEG, NeuroGPT with random initialization, NeuroGPT after pre-training), we computed a representational matrix

$$X_i \in \mathbb{R}^{N \times M},$$

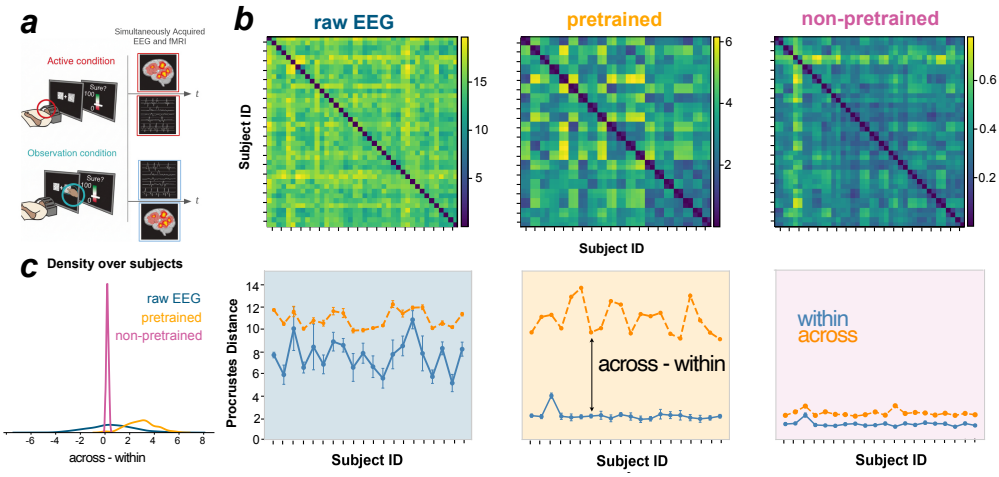

Figure 1: ***Pretraining amplifies subject-specific geometry in EEG representations.*** **a.** Illustration of the EEG-fMRI dataset **b. Top**, Procrustes distance matrices (D^{Proc} in Methods) for each representation type. **Bottom**, average within- and across-subject distances (Δ in methods) for each representation type. For Within-subject distances, Error-bars are standard deviations across different independent folds attributions. For Across subject distances, Error-bars are standard-deviations of mean distances across different independent folds attribution. **c.** Histogram of within - across distances (Δ in methods).

where N is the number of EEG channels and M the number of condition-averaged trials. Here, we used $M = 8$ (2 (active vs observation) x 2 (left vs right stimulus) x 2 (time bins of 2.0 sec)). Each matrix X_i is treated as an embedding of M shared measurement points in an N-dimensional neural space. Before computing distances, all matrices were mean-centered across conditions. To compare the three representations, we employ a geometrical comparison with different distance metrics: the Euclidian distance and the Procrustes distance. We define the two distances as follows:

$$d_{\text{Eucl}}(X_i, X_j) = \|X_i - X_j\|_F \qquad \text{and} \qquad d_{\text{Proc}}(X_i, X_j) = \min_{Q \in O(N)} \|QX_i - X_j\|_F,$$

where $O(N)$ is the group of orthogonal matrices. The Procrustes distance aligns representational manifolds up to rotation and reflection, making it invariant to arbitrary basis rotations induced by encoder architectures while remaining sensitive to changes in manifold shape. Like the Euclidean distance, the Procrustes distance satisfies symmetry and the triangle inequality, allowing for straightforward downstream applications such as k-nearest neighbors decoding (Table 2). A substantial difference however, is that the Procrustes distance is invariant to simple orthogonal transformation like a rotation of the channels, thereby avoiding seemingly high distances for equivalent representations. Finally, note that Procrustes distances include the the Euclidean distance when $Q = I$, so Euclidean distances are expected to capture neural differences at most equally well as, but not better than, Procrustes distances.

Comparing within subject and across subject distances. For each representation type, we constructed distance matrices Δ^{Eucl} and Δ^{Proc} to compare the EEG recording representation across subjects. The distance matrices are defined as follows. To ensure that self–self comparisons did not artificially inflate separability, we creates two feature matrices X and X' per subject using non-overlapping sets of trials. We define intermediate distance matrices D across subjects of shape $2K \times 2K$ ($K = 20$ is the number of subjects) where the even indices are computed using the feature matrix X and the odd indices are computed using the non-overlapping trial selection in X'. So for a given subject index s there is a the corresponding index s' which corresponds to the same subject on the non-overlapping trials. Finally, to assess whether a subject is more similar to itself than to others, we compute for two set of trials of different subject indices s and t the difference $\Delta^{\text{Proc}}_{st} = D^{\text{Proc}}_{s,t} - D^{\text{Proc}}_{s,s'}$. If $\Delta^{\text{Proc}}_{st} > 0$, the subject s is closer to itself s (across independent folds) than to subject t, meaning that the representation clusters features from the same subject together. In this case, we interpret than the representation geometry captures subject-specific idiosyncrasies. We compute similarly the matrices $\Delta^{\text{Eucl}}_{st}$ and $\Delta^{\text{Proc}}_{st}$ across all subject pairs for each representation

	Raw	Pretrained	Non-pretrained
Trial type	0.82 ± 0.01	$\mathbf{0.87 \pm 0.01}$	0.76 ± 0.002
Subject ID	0.43 ± 0.02	$\mathbf{0.51 \pm 0.02}$	0.16 ± 0.007

Table 1: ***Pretraining improves decoding trial type and subject identity.*** Decoding accuracy using the raw EEG data, pre-trained and frozen NeuroGPT embeddings, or randomly initialized NeuroGPT embeddings. We train train a multi-layer perceptron and the mean and standard deviation are computed across 50 cross-validated folds.

	Raw	Pretrained	Non-pretrained
k-NN with d_{Eucl}	0.42 ± 0.07	$\mathbf{1.00 \pm 0.00}$	0.87 ± 0.05
k-NN with d_{Proc}	0.70 ± 0.08	$\mathbf{1.00 \pm 0.00}$	0.93 ± 0.03

Table 2: ***Pretraining amplifies the distances across subjects.*** Classification accuracy of the subject ID using k-nearest neighbor (k-NN) on the raw EEG data, pre-trained and frozen NeuroGPT embeddings, or randomly initialized NeuroGPT embeddings. We report results using the Euclidean and Procrustes distance. The mean and standard deviation are computed across 50 cross-validated folds. In comparison with Table 1, the features are averaged over time.

type (raw EEG, pre-trained NeuroGPT and not pre-trained NeuroGPT). These matrices are used in the next section us to test whether each representation preserves, suppresses, or amplifies inter-individual structure.

Quantifying EEG-fMRI cross-modality alignment. To assess whether different EEG representations capture neural information that generalizes across modalities, we quantified the alignment between EEG-derived features and simultaneously recorded fMRI BOLD activity using a predictive framework. For each subject, we constructed trial-level feature vectors by averaging EEG activity (raw or embedded) within the same temporal windows used for representational analyses. These EEG features were used to predict corresponding fMRI responses using ridge regression.

Model performance was evaluated using Pearson correlation between predicted and observed BOLD responses on held-out trials. We performed two complementary analyses. First, in a *within-subject* setting, separate models were trained and tested independently for each subject using cross-validation splits of trials. Second, in a *cross-subject* setting, we trained a single model across multiple subjects and evaluated generalization using leave-one-subject-out cross-validation.

We deliberately restricted the cross-modal mapping to a linear model. This design choice ensures that any nonlinear transformations arise from the EEG representations themselves—particularly from the pretrained foundation. While more expressive nonlinear models (e.g., deep neural networks) could potentially increase predictive performance, they would also confound interpretation by introducing additional sources of nonlinear structure. Our approach therefore provides a conservative estimate of alignment that reflects the intrinsic compatibility between EEG and fMRI representations.

3 RESULTS

Pretrained embeddings improve decoding of trial type and subject identity. In our dataset Pereira et al. (2020), the trial type is either active (e.g. the subjects makes a decision) versus observational (e.g. the subject observes somebody else's decision). Representations extracted from the pretrained NeuroGPT encoder yielded significantly higher decoding accuracy (using a two-layer MLP) of both trial type (see Dataset) and subject identity, compared to raw EEG or randomly initialized models (Table 1). These results indicate that the pre-trained latent structure of NeuroGPT amplifies the difference between an active choice and passive observations of a comparable choice. Importantly, with our small sample size, the gap between 0.93 and 1.00 could reflect just a few borderline cases flipped. In what follows, we show that this is not the case. This finding encourages a deeper investigation that we report below.

Pretraining amplifies inter-individual differences. We use Procrustes distances (d_{Proc}) to reveal systematic cross-subject differences in representational geometry (Fig. 1b). Raw EEG exhibited

moderate structure: subjects were more similar to themselves than to others, but the distribution of detectable individual idiosyncrasies (Δ_{st}; see Methods) was broad (Fig. 1c), reflecting substantial overlap across subjects. Increasing or decreasing the number of conditions did not change Δ_{st} noticeably (Fig. A1). The NeuroGPT without self-supervised pre-training (randomly initialized weights) decreased average individual idiosyncrasies, but substantially reduced variability (Fig. 1c). This reduction increased subject separability with k-nearest neighbor decoding, relative to raw EEG (Table 2). However, the distances across subject appear to be redundant between the raw EEG and the random NeuroGPT embeddings: pearson correlations between raw EEG distances and those derived from randomly initialized embeddings were high ($r = 0.31, p = 5.8e-19$), consistent with the behavior of a random nonlinear projection that largely preserves coarse geometric relationships.

In sharp contrast, pretrained NeuroGPT embeddings produced a substantial expansion of detectable representational geometry idiosyncrasies (Fig. 1c). Inspection of self- versus cross-subject distances showed that this expansion arose primarily from reduced self-distances (Fig. 1b, bottom), yielding perfect subject decodability in k-NN decoding (Table 2). Notably, correlations between raw EEG distances and pretrained distances were virtually zero ($r = 0.0003, p = 0.9$), indicating that pretraining induces a qualitatively new representational organization that amplify relevant dimensions, rather than a random transformation of the raw geometry. Finally, as a baseline comparison, we also repeated the aforementioned analyses using Euclidean distances (d_{Eucl}). As expected, d_{Proc} lead to better k-NN decoding (Table 2). Surprisingly, d_{Eucl} in the pretrained embeddings also led to perfect subject ID decoding. Together, these results demonstrate that pretraining does not homogenize neural representations. Instead, it accentuates subject-specific idiosyncrasies in representational geometry.

Pretrained embeddings enhance EEG–fMRI cross-modal alignment. We next tested whether pretrained embeddings could be leveraged to improve alignment between simultaneously recorded EEG and fMRI (Fig. 2). As a simple proof of concept, we tested this with cross-validated ridge regression between modalities at the single-trial, time-averaged level and independently for each subject (Fig. 2a). Raw EEG showed no detectable cross-modal correspondence: regressing EEG features onto fMRI BOLD responses yielded near-zero predictive performance (p=0.6). Randomly initialized embeddings performed similarly (p=0.74, paired t-test), indicating that simple nonlinear transformations of the EEG do not improve cross-modal alignment. In contrast, pretrained Neu-

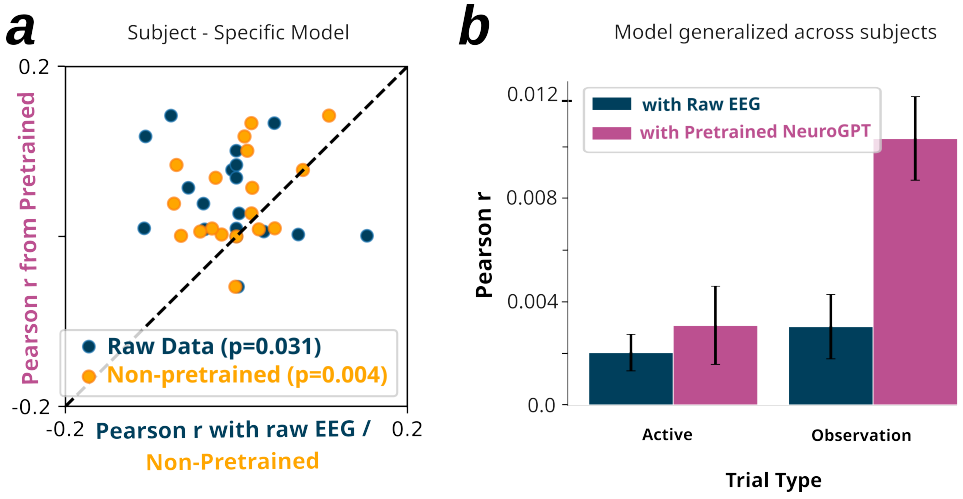

Figure 2: ***Pretraining enables significantly better prediction of BOLD signals from EEG.*** **a.** Predictive accuracy (measured with Pearson r) of BOLD activity from EEG for each subject. The coordinate of each orange dot is the prediction from non-pretrained embeddings (x-axis) and the ones from pretrained embeddings of each subject (y-axis). Similarly for blue dots, but from raw EEG. Thus, each dot located above the diagonal represent higher predictions from pretrained embeddings. Predictions from pretrained embeddings are higher than both raw EEG or non-pretrained embeddings. **b.** Similar analyses as in b, but training only one regression across all subjects and separating by trial type.

roGPT embeddings produced detectable cross-modal predictability (p=0.002), significantly higher than raw EEG or non-pretrained embeddings (Fig. 2a). These within-subject regressions show reliable positive correlations between predicted and observed BOLD activity, indicating that pretrained features capture neural components shared across imaging modalities. Next, we tested whether NeuroGPT embeddings would improve generalization across subject. With this aim, we performed leave-one-subject out cross modal prediction (Fig. 2b). Cross-subject regressions also showed above-chance prediction, albeit weaker than within-subject performance, suggesting that the pretrained latent space supports a degree of modality-general representation that extends beyond individual idiosyncracies. Notably, this effect was strongest for the observation condition, motivating future work to clarify why. These findings indicate that self-supervised EEG pretraining not only enhances within-modality structure but also facilitates alignment across modalities.

4 Discussion

Our findings demonstrate that NeuroGPT, an EEG foundation models trained with self-supervision, does not collapse individual variability into a shared canonical space; instead, it selectively amplifies subject-specific structure in neural geometry, presumably by creating a richer, higher-dimensional space where idiosyncrasies become more separable. This stands in contrast to our initial intuition that large-scale pretraining on heterogeneous datasets would enforce representational homogenization and suggests instead that exposure to diverse neural data enables the model to carve out higher-dimensional, idiosyncrasy-preserving manifolds. At the same time, the improved alignment between EEG and fMRI reveals that the model extracts components that generalize across modalities, providing a bridge between electrophysiological and hemodynamic signals. Throughout this study, we systematically compared pretrained models against their randomly initialized counterparts, ruling out the trivial explanation that the observed gains arise merely from operating in a higher-dimensional embedding space. Finally, we used here minimally processed EEG (e.g. without filtering in specific bands), but future work should compare the amplification of NeuroGPT with standard preprocessing pipelines. All together, our results suggest that representation alignment across modalities may benefit from pre-trained foundation models, not because they enforce uniformity, but because they enhance structured variability.

References

Joao Barbosa, Amin Nejatbakhsh, Lyndon Duong, Sarah E Harvey, Scott L Brincat, Markus Siegel, Earl K Miller, and Alex H Williams. Quantifying differences in neural population activity with shape metrics. *bioRxiv*, pp. 2025–01, 2025.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. *arXiv preprint arXiv:2401.10278*, 2024.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, 2025.

Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Towards neural foundation models for vision: Aligning eeg, meg and fmri representations to perform decoding, encoding and modality conversion. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Michael Pereira, Nathan Faivre, Iñaki Iturrate, Marco Wirthlin, Luana Serafini, Stéphanie Martin, Arnaud Desvachez, Olaf Blanke, Dimitri Van De Ville, and José del R Millán. Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences*, 117(15):8382–8390, 2020.

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2584–2594, 2022.

Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in neural information processing systems*, 35:21255–21269, 2022.

Eric Y Wang, Paul G Fahey, Kayla Ponder, Zhuokun Ding, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, et al. Towards a foundation model of the mouse visual cortex. *bioRxiv*, pp. 2023–03, 2023.

Ran Wang and Zhe Sage Chen. Large-scale foundation models and generative ai for bigdata neuroscience. *Neuroscience Research*, 215:3–14, 2025.

Walter F Wiggins and Ali S Tejani. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119, 2022.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Zhizhang Yuan, Daoze Zhang, Junru Chen, Gefei Gu, and Yang Yang. Brant-2: Foundation model for brain signals. *CoRR*, 2024.

A Appendix

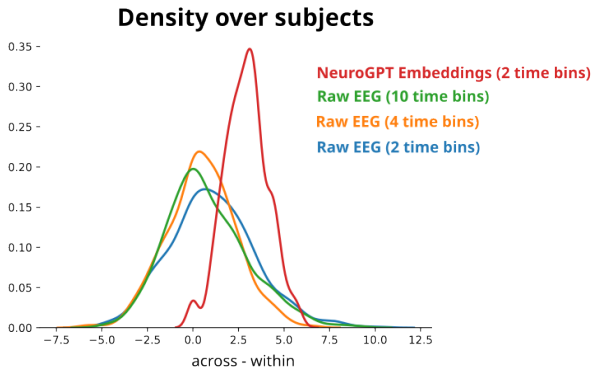

Figure A1: ***No impact of the number of time-bins of raw data on the distribution of within-across distances.*** Changing the number of trial-averaged conditions M by increasing number of time bins does not affect the distribution of within-across distances (Δ). The different distributions correspond to the same recordings but binned in 2, 4, or 10 time bins, corresponding to 2, 1, or 0.4 seconds bins of raw EEG data and thus M = 8, 16 or 40. For comparison we also show Δ for M = 8 when using NeuroGPT embeddings, as in Fig 1c.