DANCE: DOMAIN-AWARE NETWORK-COUPLING ENCODER FOR COORDINATE-FREE INTRACRANIAL EEG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The brain is a network of interacting regions whose coupling changes with behavioural state. Existing foundation models for intracranial EEG (iEEG) either ignore this structure or encode the static spatial location of each electrode, leaving the interactions between electrodes to be learned implicitly. We propose DANCE, a Domain-Aware Network-Coupling Encoder, which places functional connectivity in the pretraining objective: a channel-agnostic encoder is trained to align its embeddings with measured connectivity, estimated over short intervals to reflect the dynamic coupling of the network as brain state changes. We evaluate DANCE on two distinct datasets. On Brain Treebank, a public language-decoding benchmark, DANCE matches the strongest published model, improves speech-decoding ROC-AUC by 16% on the subjects with the most electrodes and the fewest labels, and loses no performance when the evaluated patient is excluded from pretraining. On a chronic ambulatory cohort recorded by an implanted device from four channels, DANCE outperforms state-of-the-art models adapted to the cohort, reaching 0.845 macro-F1 on sleep stage classification and 0.967 ROC-AUC on seizure detection. Strong performance on two datasets that differ in recording setting, electrode configuration and task demonstrates the broad applicability of DANCE and its potential for wider use in the analysis of brain activity.

1 INTRODUCTION

The brain is organised as a network of interacting regions, and the strength of their interaction changes with behavioural state (Bassett & Sporns, 2017; Hutchison et al., 2013). These state-dependent interactions support cognitive function (Bressler & Menon, 2010) and are disrupted in neurological disorders such as epilepsy (Stam, 2014; Kramer & Cash, 2012). Foundation models for intracranial EEG (iEEG) learn representations of neural activity from unlabelled recordings and enable downstream clinical and neuroscientific tasks where labelled data are scarce. Such models have progressed from channel-agnostic encoders (Wang et al., 2023) to models that encode the static spatial location of each electrode (Chau et al., 2025; Oganesian et al., 2025), an addition to which much of their improvement has been attributed. Location, however, is fixed at implantation: it records where activity is measured but not how the sites interact, and it is this interaction that changes with brain state. Our goal is a foundation model for iEEG whose representations are grounded in functional connectivity, the strength of coupling between activity recorded at different brain sites. Because the interaction between brain regions is central to the study of any brain activity or disorder, such a representation can be applied broadly across recording settings and tasks.

Existing models supply spatial structure as an input, as three-dimensional coordinates (Chau et al., 2025) or atlas parcel identities (Oganesian et al., 2025), and leave the interactions between electrodes to be learned implicitly, through a proxy task such as reconstructing masked signal. A measured connectivity target supplies these interactions directly: it provides a supervisory signal for every pair of electrodes, and because connectivity shifts with behavioural and vigilance state (Hutchison et al., 2013), it carries state-dependent structure that characterises the network in a way that neither a single site nor a static location can. Using functional connectivity in this way raises three requirements. First, the model must accommodate any electrode configuration, because the number and placement of electrodes differ across patients according to clinical need. Second, the interaction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

between electrodes must be quantified with a measure that reflects physiological coupling and can be estimated reliably from the recording. Third, the resulting inductive bias must be dynamic and track the changing brain state.

We propose the Domain-Aware Network-Coupling Encoder (DANCE), which addresses these requirements by placing functional connectivity in the learning objective. Functional connectivity is central to DANCE because brain activity is shaped by interactions between regions that change with brain state. The encoder is channel-agnostic: each channel is encoded independently, so a single model accommodates any number of electrodes in arbitrary order, and for a downstream task the per-channel embeddings are pooled into one vector read by a task head. Functional connectivity is quantified as the amplitude-envelope correlation of every pair of electrodes, a standard estimator of functional brain networks (Hipp et al., 2012). The objective regresses the similarity between the embeddings of every two electrodes onto their measured connectivity, so that the geometry of the representation aligns with the network. Connectivity is computed over consecutive blocks, so the target is not a single matrix per recording but a sequence of matrices that changes as the network state changes, and the objective captures the dynamics of connectivity rather than a static average.

We evaluate DANCE on two distinct datasets: Brain Treebank (Wang et al., 2024), a public language-decoding benchmark of hours-long hospital recordings sampled at 2048 Hz from 72 to 205 localised electrodes spanning multiple brain regions, and a chronic ambulatory cohort with drug-resistant epilepsy, recorded over weeks by an implanted device at 250 Hz from four channels, with sleep stage, interictal epileptiform discharge (IED) rate and seizure labels. On Brain Treebank, DANCE matches BaRISTA (Oganesian et al., 2025), the strongest published model, and improves speech-decoding ROC-AUC by 16% on the two subjects with the most electrodes and the fewest labelled windows. Excluding the evaluated patient from pretraining costs no performance, so the gain reflects structure that transfers across patients rather than exposure to the patient's own recordings. Controlled comparisons suggest that a dynamic connectivity target outperforms a static one, and that adding the parcel embedding and cross-channel attention on which BaRISTA relies does not improve DANCE, consistent with the connectivity target already supplying the information these components provide. On the chronic cohort, DANCE outperforms state-of-the-art models adapted to this cohort, reaches 0.845 macro-F1 on sleep stage classification and 0.967 ROC-AUC on seizure detection. Across two datasets that differ in recording setting, sampling rate and task, the same architecture and objective apply unchanged, demonstrating the broad applicability of DANCE.

Our contributions are:

- **Functional connectivity as the pretraining target.** DANCE encodes each electrode independently and learns the interactions between electrodes through the objective alone, by aligning its embedding geometry with measured, dynamic functional connectivity. To our knowledge, no prior iEEG model uses connectivity as the pretraining objective.
- **Largest gains where data are scarcest, and transfer to unseen patients.** DANCE matches the strongest model on Brain Treebank, with largest gains on subjects with many electrodes and few labels, and transfers to patients excluded from pretraining without loss of performance.
- **Applicability across tasks, electrode configurations and recording settings.** Being agnostic to electrode configuration and capturing brain network dynamics, DANCE achieves competitive performance on both language decoding from hundreds of hospital electrodes and clinical classification from a four-channel implanted device, indicating potential for broader application.

2 Related Work

Self-supervised models for intracranial EEG. Existing models differ in how they treat the relation between electrodes. BrainBERT (Wang et al., 2023), LaBraM (Jiang et al., 2024) and Du-IN (Zheng et al., 2024) encode each channel independently, which accommodates any montage but leaves inter-electrode structure out of the representation. Brant (Zhang et al., 2023; Yuan et al., 2024) and DIVER-1 (Han et al., 2025) add attention across channels without any description of where the channels are, so the relation between electrodes is learned implicitly from the pretext task. PopT (Chau et al., 2025) and BaRISTA (Oganesian et al., 2025) supply spatial structure explicitly, as three-dimensional coordinates and as Destrieux parcel identities, and BaRISTA reports the strongest published results on Brain Treebank, attributing much of the improvement to this information. In

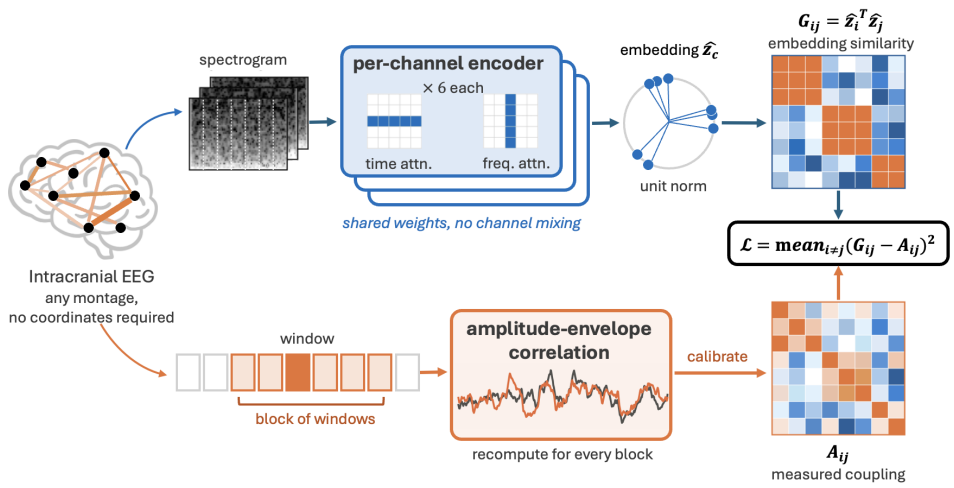


Figure 1: **DANCE pretraining.** Top: each channel's spectrogram is encoded by the same transformer (temporal attention with rotary embeddings, then cross-frequency attention), with no mixing across channels, and its outputs are projected to a unit-norm embedding $\hat{z}_c$. Bottom: amplitude-envelope correlation is measured over a block of consecutive windows, calibrated to the cosine range. The loss regresses the embedding similarities G_{ij} onto the measured coupling A_{ij}.

each case the interaction between electrodes is absent, learned implicitly, or approximated by a location that is fixed at implantation, whereas the interaction itself changes with brain state (Hutchison et al., 2013). DANCE keeps the per-channel encoder of the first group and supplies the interaction directly, as a measured, dynamic target in the objective; on Brain Treebank the anatomical components of BaRISTA add nothing once this target is in place (§5.3).

Connectivity in learned models of brain signals. Functional connectivity (Hipp et al., 2012; Lachaux et al., 1999) has entered learned models as an *input*. Graph neural networks for EEG build a channel graph from signal correlation (Tang et al., 2022), MBrain (Cai et al., 2023) learns such a graph and uses it to weight message passing, and Brain-JEPA (Dong et al., 2024) turns fMRI connectivity gradients into positional encodings. In these models connectivity is a fixed structure per recording, which the network may weight arbitrarily and which must be recomputed at inference. DANCE makes connectivity the *target*: the loss requires embedding similarity to match measured coupling, recomputed over short blocks so that it tracks brain state, and the condition is testable on held-out recordings and on unseen patients (Appendix M); the encoder itself consumes only the spectrogram.

Relational objectives and domain-guided pretraining. Our loss shares its form with relational (Park et al., 2019) and similarity-preserving (Tung & Mori, 2019) distillation, which match a student's pairwise-similarity matrix to a teacher's; we replace the teacher with a physical measurement and index pairs by simultaneously recorded channels rather than by batch samples (§3.4). Among attempts to build neuroscientific priors into the pretext task itself, closest is Wagh et al. (2021), whose scalp-EEG objectives derive from spatial similarity of activity, behavioural state and subject age. Such priors are categorical and fixed for a recording; ours is continuous, defined on every pair of electrodes, and recomputed as coupling changes. To our knowledge, no prior model of intracranial EEG uses functional connectivity as its pretraining objective.

3 Method

DANCE places the prior over brain-network organisation in the learning *objective*: functional connectivity is estimated from the signal itself, and the encoder is trained so that the similarity of two channel embeddings matches their measured coupling. Figure 1 provides an overview of the proposed framework.

3.1 PROBLEM SETTING

A recording session provides a multichannel signal $X \in \mathbb{R}^{C\times N}$ sampled at f_s, where the channel count C differs between sessions and between corpora. The same encoder is applied to every channel independently, so one set of weights is pretrained across all sessions within the corpora, and the inter-channel structure enters through the objective (§3.4). Each session is partitioned into non-overlapping windows of fixed duration, indexed by w, and $x_c^{(w)}$ denotes channel c in window w. Corpus-specific constants (f_s, window length, spectrogram size) are listed in Appendix H.

3.2 TOKENISATION AND ENCODER

Spectrogram. Following BrainBERT (Wang et al., 2023), each channel-window is mapped to a dB-magnitude spectrogram $S_c^{(w)} \in \mathbb{R}^{F\times T}$ by a short-time Fourier transform with F linearly spaced frequency bins and T frames. Each (channel, band) series is standardised, $\tilde{S}_c[f,t] = (S_c[f,t] - \mu_{c,f})/\sigma_{c,f}$, to remove the $1/f$ gradient so that no band dominates the input scale; during pretraining $\mu_{c,f}, \sigma_{c,f}$ are the statistics of the session and downstream those of the training split. Each band's series is then unfolded into P overlapping patches of length p and stride s, each linearly projected to $\mathbb{R}^d$, and a learned `[CLS]` token is prepended, giving $F(P{+}1)$ tokens per channel.

Spectral profile. Per-band standardisation destroys the across-band power *profile*—the relative distribution of energy over frequency, a marker of physiological state. We adopt the remedy of Brant (Zhang et al., 2023; Yuan et al., 2024): the log band-power vector of each channel, computed from the unstandardised spectrogram, is normalised across bands, projected and added to every token of that channel. The normalisation across bands makes this a *relative* profile: it carries spectral shape, not absolute amplitude, which is dominated by electrode impedance and contact placement.

Encoder. Tokens pass through a factorised transformer with two stages. (i) *Temporal attention*: n_ℓ pre-norm layers over the $P{+}1$ time positions, run independently for each (channel, band). Temporal position is supplied by rotary embeddings (Su et al., 2024) inside the attention, so the model encodes relative time and remains valid under a change of T. (ii) *Cross-frequency attention*: a learned band embedding $\beta_f \in \mathbb{R}^d$ is added, after which n_ℓ layers attend over the F bands independently for each (channel, time position). Bands are not translation-invariant—4 and 8 Hz are not related as 100 and 104 Hz are—so this axis uses an absolute learned embedding rather than a rotary one.

Channel embedding. Let $h_{c,f} \in \mathbb{R}^d$ be the encoder output at the `[CLS]` position of band f of channel c. The objective acts on a per-channel embedding obtained by concatenating over bands and projecting through a two-layer MLP $g_\theta : \mathbb{R}^{Fd} \to \mathbb{R}^d$,

$$z_c^{(w)} \;=\; g_\theta\big([\,h_{c,1};\,\ldots;\,h_{c,F}\,]\big), \tag{1}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation.

3.3 FUNCTIONAL-CONNECTIVITY TARGET

We use amplitude-envelope correlation (AEC) (Hipp et al., 2012), a standard connectivity estimator in electrophysiology, as the functional-connectivity measure for the relational target. It computes the Pearson correlation over time between the log-amplitude envelopes of two channels within a frequency band. Per-band correlations are averaged over the physiological band range to give $M_{ij}(b) \in [-1,1]$ for every channel pair (time-frequency decomposition, band grid and band selection in Appendix C). We compute M over *blocks* of K consecutive windows; writing $b(w)$ for the block containing window w, all windows of a block share one target. The block is long enough for a stable estimate and short enough that the target varies with brain state across a session (Appendix C gives the choice of K and quantifies that variation). While raw AEC is overwhelmingly positive (on a representative session, 91% of pairs), which would instruct the model to make nearly all pairs similar, we standardise on the off-diagonal entries pooled over all blocks of the session and squash into the cosine range,

$$A_{ij}(b) \;=\; \tanh\!\left(\frac{M_{ij}(b)-\mu}{2\,\sigma}\right) \in (-1,1), \qquad \mu,\sigma \;=\; \text{mean, std}\,\big\{M_{ij}(b') : i\neq j,\ \forall b'\big\}, \tag{2}$$

so that $A_{ij} > 0$ means i and j are more coupled than the session-typical pair; the $\tanh$ preserves the ordering of coupling strengths where clipping would not (alternatives are compared in Appendix C).

3.4 THE RELATIONAL OBJECTIVE

Let $\hat{z}_c = z_c / \|z_c\|$, with $\|\cdot\|$ the Euclidean norm, and let $G^{(w)}_{ij} = \hat{z}^{(w)\top}_i \hat{z}^{(w)}_j$ be the Gram matrix of the normalised channel embeddings in window w. DANCE regresses it onto the target of the enclosing block:

$$\mathcal{L}_{\mathrm{rel}} \;=\; \frac{1}{|\mathcal{V}|} \sum_{(w,i,j)\,\in\,\mathcal{V}} \Big(G^{(w)}_{ij} \;-\; A_{ij}\big(b(w)\big) \Big)^2, \tag{3}$$

where $\mathcal{V}$ contains the off-diagonal pairs of real channels of every window that belongs to a block; padded channels, which arise only from batching sessions of different C, are excluded.

Because A is recomputed for every block, a fixed pair receives different targets as its coupling changes with brain state, which is the dynamic prior that a static spatial embedding cannot express. The loss is one batched $C \times C$ product per window. Its form is that of relational distillation (Park et al., 2019; Tung & Mori, 2019), with a physical measurement of the recorded system in place of a teacher network. Because the objective states an equality, it is directly checkable: on held-out windows we report $\rho = \mathrm{corr}_{(i,j)}(G_{ij}, A_{ij})$, the correlation between embedding similarity and measured connectivity, which distinguishes a model that has internalised the functional geometry from one that has merely reduced the loss by collapsing the embedding scale (§5).

3.5 DOWNSTREAM TRANSFER

For a downstream task, g_θ and the target are discarded and a classifier is attached to the encoder's `[CLS]` outputs. The band tokens of each channel are pooled by learned attention into one channel vector; the channel vectors are concatenated and reduced by a low-rank linear map, followed by a two-layer MLP classifier. The concatenation keeps a weight for every (channel, feature) pair, so the classifier can read different features from different channels; readouts that average over channels, or learn one scalar per channel, discard this structure and cap every pretraining objective alike, and the low-rank factorisation retains it at a fraction of the parameters (Appendix B).

4 EXPERIMENTS

We evaluate on two intracranial corpora that differ in scale, localisation and task: a public, densely sampled benchmark on which anatomy-based models have been evaluated, and a clinical cohort recorded by an implanted device.

4.1 DATASETS AND TASKS

Brain Treebank (Wang et al., 2024). Ten subjects recorded at 2048 Hz during naturalistic film viewing. We adopt the preprocessing of BaRISTA (Oganesian et al., 2025), artefactual contact exclusion, 60 Hz notch filtering and Laplacian re-referencing, and segment the recordings into non-overlapping 3 s windows. Of the corpus's 26 sessions (≈ 56 h), 17 are used for pretraining and one held-out session from each of seven subjects for evaluation (91 to 205 channels per subject); the remaining two sessions are excluded following BaRISTA's protocol. We evaluate the two binary decoding tasks of prior work (Wang et al., 2023; Oganesian et al., 2025), *speech vs. non-speech* and *sentence onset*, scored by ROC-AUC.

Chronic ambulatory iEEG. Four patients with drug-resistant mesial temporal lobe epilepsy, recorded with a chronically implanted sensing device from four bipolar channels spanning the bilateral hippocampi and anterior thalamic nuclei at 250 Hz; this corpus is not publicly available. The first ten recording days of each patient (707 h) are used for pretraining, and separate later days (351 h of labelled recording) for evaluation. Windows are 10 s and splits are chronological (60/20/20) within patient. Electrode localisation is not available to us for this cohort. PopT, which requires it, is not evaluated here; BaRISTA is adapted to the cohort as described in §4.2. Three tasks are evaluated. (i) *Sleep stage classification* is a three-way classification of wakefulness, slow-wave sleep (N3)

and rapid-eye-movement (REM) sleep, scored by macro-F1. (ii) *Interictal epileptiform discharge (IED) rate* is the number of discharges per electrode in each window, fit by Poisson regression and scored by the mean squared error between the predicted rate and the true count over windows and electrodes. (iii) *Seizure detection* separates ictal from interictal windows in the two patients with annotated seizures. Per-patient values are in Appendix Table 8; Appendices D and E give split sizes and artefact handling.

4.2 Baselines, controls and protocol

Baselines. We re-run BaRISTA (Oganesian et al., 2025) with its released checkpoint and code on the same splits and seeds as our models. Brant (Zhang et al., 2023) and PopT (Chau et al., 2025) are quoted from Oganesian et al. (2025), whose preprocessing and segmentation we share. On the chronic cohort, BaRISTA's released weights do not apply (they are trained at 2048 Hz on localised electrodes) and its pretraining code is not released, so we reimplemented its pretraining, pretrained it on the same windows as DANCE with parcel identities given by the implant targets, and fine-tuned it on the same splits, seeds and selection rule. As a reference for the effect of pretraining, every model is compared with the same encoder trained from random initialisation. A linear model on band power, an ℓ_2-regularised regression on the encoder's input with its strength chosen on the validation split, measures how much of each task is linearly extractable from spectral power.

Comparisons. To test whether the connectivity target itself drives transfer, we pretrain the same encoder under two alternative objectives. *Masked latent reconstruction*, the DATA2VEC-style objective (Baevski et al., 2022) used by BaRISTA, predicts the representations of masked time spans and frequency bands (Appendix F), learning structure within each electrode but no relation between electrodes. *Region InfoNCE* contrasts each channel with a partner from its own functional region against 32 channels from other regions, under a three-way grouping derived from the localisation table (auditory and language, limbic, remainder); it uses inter-electrode information, but as a binary relation over sampled pairs rather than the measured coupling of every pair. To test which property of the target transfers, we replace amplitude-envelope correlation by phase-locking value, and the block-wise target by a static one that assigns every block the recording's average connectivity. To test whether anatomical information adds to the functional target, we add the two components through which BaRISTA encodes spatial structure, cross-channel attention and a parcel embedding (Appendix G). Comparisons that require localisation are run on Brain Treebank only; on the chronic cohort, DANCE is compared with random initialisation, the linear model, masked latent reconstruction and the adapted BaRISTA.

Generalisation and label efficiency. To measure transfer to patients never seen in pretraining, we repeat pretraining once per evaluation patient with all of that patient's recordings excluded, then fine-tune on the excluded patient (leave-one-patient-out), on both corpora (Table 1, Appendix P and Appendix Table 9). To measure label efficiency, we fine-tune on 25% and 50% subsets of each subject's training windows at a fixed number of optimiser steps (Figure 2b and Appendix L).

Fine-tuning and evaluation. Every model is read out through the rank-8 map of Eq. 5. It is fine-tuned in two stages, a linear probe and then the top two transformer blocks, with schedule points set as fractions of the total optimiser steps. Each reported value is a mean over five fine-tuning seeds per subject. Model selection uses the validation split, reported values come from the held-out test split, and normalisation statistics come from the training split alone. Tables report the mean of the per-subject means with the standard error across subjects, and comparisons use two-sided Wilcoxon signed-rank tests paired over subjects. Hyperparameters are in Appendix H.

5 Results

5.1 Comparison with prior models on Brain Treebank

Table 1 compares DANCE with prior models on the same splits and seeds. DANCE matches BaRISTA on both tasks, 0.869 against 0.851 on speech and 0.860 against 0.866 on sentence onset, with neither difference significant. Relative to random initialisation, the gain of DANCE on speech

Table 1: **Comparison on Brain Treebank.** Test ROC-AUC, mean ± s.e.m. across 7 subjects (5 seeds each). BaRISTA is re-run with its released checkpoint and code on the same splits and seeds; Brant and PopT are quoted from Oganesian et al. (2025). The linear model is an ℓ_2-regularised logistic regression on band power. The last row excludes the evaluated patient from pretraining. Best in bold, second underlined. †Requires electrode localisation.

Model	Sentence Onset	Speech / Non-Speech
Brant (Zhang et al., 2023) *(quoted)*	$0.767_{\pm 0.017}$	$0.691_{\pm 0.017}$
PopT + BrainBERT† (Chau et al., 2025) *(quoted)*	$0.795_{\pm 0.014}$	$0.775_{\pm 0.016}$
BaRISTA† (Oganesian et al., 2025), our reproduction	$0.866_{\pm 0.029}$	$0.851_{\pm 0.045}$
Linear model on band power	$0.846_{\pm 0.024}$	$0.860_{\pm 0.033}$
Random initialisation	$0.850_{\pm 0.028}$	$0.845_{\pm 0.037}$
DANCE (ours)	$0.860_{\pm 0.030}$	$\underline{0.869_{\pm 0.031}}$
DANCE, evaluated patient excluded from pretraining	$\underline{0.862_{\pm 0.023}}$	$0.875_{\pm 0.034}$

Table 2: **Controlled comparisons on Brain Treebank.** Test ROC-AUC, mean ± s.e.m. across 7 subjects, with the encoder, readout and fine-tuning recipe fixed. Δ (p): difference from random initialisation, two-sided Wilcoxon signed-rank test over subjects. Further ablation rows are in Appendix Table 6; all pairwise contrasts in Appendix Table 7. Best in bold, second underlined.

	Sentence Onset		Speech / Non-Speech	
Model	AUC	Δ (p)	AUC	Δ (p)
DANCE (random init.)	$0.850_{\pm 0.028}$	—	$0.845_{\pm 0.037}$	—
Masked latent reconstruction	$\underline{0.854_{\pm 0.027}}$	$+0.005$ (0.938)	$0.859_{\pm 0.035}$	$+0.014$ (0.031)
Region InfoNCE	$0.849_{\pm 0.029}$	-0.001 (0.938)	$\underline{0.864_{\pm 0.030}}$	$+0.019$ (0.297)
DANCE (relational, dynamic AEC)	$0.860_{\pm 0.030}$	$+0.010$ (0.688)	$0.869_{\pm 0.031}$	$+0.024$ (0.031)
Dynamic PLV	$0.841_{\pm 0.031}$	-0.008 (0.688)	$0.843_{\pm 0.040}$	-0.002 (0.688)
Static AEC (session average)	$0.846_{\pm 0.029}$	-0.004 (0.688)	$0.851_{\pm 0.038}$	$+0.006$ (0.938)
DANCE w/ anatomical components	$0.844_{\pm 0.031}$	-0.005 (0.688)	$0.851_{\pm 0.041}$	$+0.006$ (0.578)

is significant (+0.024, $p = 0.031$, six of seven subjects), whereas on sentence onset no model separates from it. The last row evaluates DANCE on patients unseen in pretraining (leave-one-patient-out; Appendix P). Excluding the evaluated patient from pretraining costs nothing: speech AUC is 0.875, +0.030 over random initialisation ($p = 0.031$, six of seven subjects), and sentence onset shows the same pattern at a smaller effect (0.862; +0.012, $p = 0.22$). The representation therefore transfers to a patient whose montage and anatomy it has never encountered, and for four of seven subjects the held-out-patient model exceeds BaRISTA.

Figure 2a and Appendix Table 5 show where the models differ. For several subjects the tasks are close to linearly separable: the linear model reaches 0.860 on speech and 0.846 on sentence onset, and on `sub_1` and `sub_3`, which have the fewest contacts, every model approaches this floor. DANCE stands out on `sub_7` and `sub_10`, which have the most contacts (205 and 173, against a median of 135) and among the fewest labelled windows per contact (3 and 5, against a median of 10): 0.724 against 0.626 for BaRISTA, and 0.925 against 0.798. These are the subjects on which state-dependent connectivity has the most to add over a static anatomical label. Label efficiency is evaluated in Appendix L by fine-tuning on 25% and 50% of each subject's training windows (Figure 2b). The gain over random initialisation is largest when labels are scarcest, +0.039 at 25% against +0.024 at 100%, though it is not monotone.

5.2 Chronic clinical cohort

Table 3 evaluates DANCE on the chronic cohort against random initialisation, the linear model on band power, masked latent reconstruction and the adapted BaRISTA; per-patient values are in Appendix Table 8. On sleep stage classification, DANCE is the best model in every patient, reaching 0.845 macro-F1, +0.022 over random initialisation and +0.067 over BaRISTA. Random initialisation performs no better than the linear model, so the untrained encoder recovers only what band power provides. Masked reconstruction, which learns temporal and spectral structure within each channel, adds only marginally. BaRISTA falls below random initialisation, a consequence of its design at

Table 3: **Chronic ambulatory cohort.** Mean ± s.e.m. across patients ($n = 4$; seizure detection $n = 2$), 5 seeds each; per-patient values in Appendix Table 8. Sleep staging: macro-F1; IED rate: mean squared error (lower is better); seizure detection: ROC-AUC. BaRISTA is adapted to this cohort as described in §4.2. Best in bold, second underlined.

Pretraining objective	Sleep staging (macro-F1)	IED rate (MSE ↓)	Seizure detection (AUC)
Linear model	$0.823_{\pm 0.036}$	$3.579_{\pm 1.064}$	$0.955_{\pm 0.036}$
DANCE (random init.)	$0.824_{\pm 0.046}$	$3.809_{\pm 1.106}$	$0.875_{\pm 0.081}$
Masked latent recon.	$\underline{0.827}_{\pm 0.050}$	$\underline{3.469}_{\pm 1.048}$	$\underline{0.961}_{\pm 0.004}$
BaRISTA (Oganesian et al., 2025)	$0.778_{\pm 0.033}$	$3.592_{\pm 1.147}$	$0.929_{\pm 0.061}$
DANCE (relational, dynamic AEC)	$0.845_{\pm 0.041}$	$3.390_{\pm 1.000}$	$0.967_{\pm 0.017}$

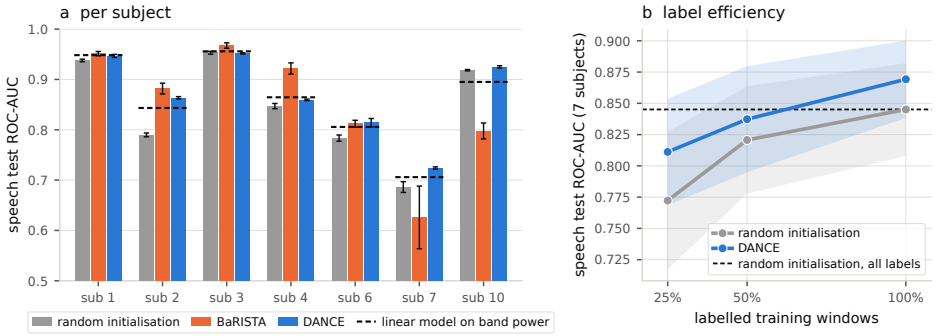


Figure 2: **Speech decoding on Brain Treebank.** (a) Test ROC-AUC per subject for random initialisation, BaRISTA and DANCE (bars: mean over five seeds; whiskers: s.e.m.), with the linear model on band power as a dashed line. Bars start at chance (0.5). (b) Test ROC-AUC against the fraction of labelled training windows at a fixed number of optimiser steps. Lines and shading show the mean and s.e.m. over the seven subjects; the dashed line is random initialisation trained on all labels.

four channels: its parcel embedding distinguishes only two identities and its cross-channel attention has only four channels to relate. DANCE, trained with the functional-connectivity objective, instead learns the coupling between regions that distinguishes sleep stages. On IED rate, DANCE has the lowest error on average, lowering it by 0.42 MSE relative to random initialisation, and is comparable with BaRISTA. On seizure detection, DANCE reaches 0.967 ROC-AUC, while the linear model reaches 0.955, indicating that this task is largely solved by band power alone and leaves little room for pretraining to contribute. Notably, these gains are obtained with only four channels, where the connectivity target contains six electrode pairs per window rather than the thousands available on Brain Treebank, indicating that the objective remains informative with minimal spatial coverage. DANCE also generalises to patients excluded from pretraining on this cohort: it improves on random initialisation on both sleep staging and IED rate, at a cost of 0.007 macro-F1 relative to pretraining on the patient and no cost in IED error (Appendix Table 9).

5.3 Controlled comparisons: objective, connectivity measure and anatomy

Table 2 isolates the contribution of each design choice: the learning objective, the functional-connectivity measure used as the target, and the anatomical components added to the encoder, with all other settings held fixed.

Learning objective. We compare the relational objective with masked latent reconstruction, the objective family used by BaRISTA, which learns temporal and spectral structure within each channel, and with region InfoNCE (van den Oord et al., 2018), a contrastive objective that uses inter-electrode information, but as a binary same-region decision over sampled pairs rather than a measured coupling for every pair. On speech, the relational objective gives the largest and most consistent gain over random initialisation (+0.024, $p = 0.031$, six of seven subjects). Region InfoNCE gains a similar amount on average (+0.019) but on only four subjects ($p = 0.30$): a binary objective pushes every cross-region pair apart equally, although most such pairs are positively coupled on these recordings (§3.3), whereas the relational objective uses each pair's measured coupling.

Masked reconstruction improves consistently but by less (+0.014, $p = 0.031$), as it carries no inter-electrode information. The differences between the objectives themselves lie within the across-subject variability (Appendix Table 7); the evidence is their consistent ordering across subjects. On sentence onset no objective separates from random initialisation, which already reaches 0.85 ROC-AUC: the linear model on band power reaches 0.846, so the task is largely solved by spectral power and leaves little for pretraining to add.

Functional-connectivity measure. We compare three targets for the relational objective: the dynamic AEC target of DANCE, computed per block; a static AEC target that assigns every block the recording's average connectivity; and a dynamic PLV target. The static target removes most of the gain (+0.006 over random initialisation, against +0.024 for the dynamic target): a session average is a fixed map of the montage, closer to an anatomical description than to the state-dependent coupling the downstream tasks depend on. The PLV target removes the gain entirely (-0.002 against random initialisation, -0.027 against the AEC target). Although it is recomputed per block, phase locking in this corpus is almost entirely static, with 88 to 91% of its variance in the fixed across-pair structure (Appendix C), so it behaves as a static target. Both results point the same way: what transfers is the state-dependent variation of coupling. We verify separately that the objective makes embedding similarities track the measured target, on held-out windows and on patients excluded from pretraining (Appendix M).

Anatomical components. To determine whether the anatomical information used by BaRISTA is complementary to, or redundant with, the connectivity target, we add its two components to DANCE: cross-channel attention, through which electrodes interact, and a parcel embedding. Together they do not improve DANCE; speech AUC falls by 0.018 and sentence-onset AUC by 0.016. Appendix Table 6 separates the two: the decrease is due to cross-channel attention, and the parcel embedding added on top of it leaves performance unchanged, whereas the same embedding improves masked reconstruction, a model with no inter-electrode prior, by +0.014 on both tasks. Anatomical information therefore benefits a model that lacks an inter-electrode prior and adds nothing to one that already learns functional connectivity from its objective, indicating that the connectivity target captures the information the atlas provides; the control's improvement is consistent across both tasks but not significant for 7 subjects.

5.4 Limitations and future work

The proposed framework has limitations. Connectivity is estimated separately in each frequency band but averaged across bands into a single target, so band-specific coupling is not represented; future work could extend the objective to preserve it. The model's generalisability can be further evaluated on additional cohorts, on recording modalities such as scalp EEG, and on downstream tasks in other neurological conditions such as Parkinson's disease. Finally, DANCE is pretrained separately on each dataset. Pretraining a single model on multiple datasets spanning recording conditions, electrode configurations and patient populations, and evaluating it on entirely unseen datasets, would yield a more robust foundation model of brain activity.

6 Conclusion

We introduced DANCE, a foundation model for iEEG whose representations are grounded in functional connectivity. A channel-agnostic encoder is pretrained to align its embeddings with the dynamic coupling between electrodes, measured from the recording itself. On Brain Treebank, DANCE matches the strongest published model, with its largest gains on subjects with many electrodes and few labels, and it transfers to patients excluded from pretraining. On a chronic cohort recorded by an implanted device, it outperforms an adapted state-of-the-art model on sleep staging in every patient. Controlled comparisons point to the dynamic component of connectivity as the source of the gain, and anatomical components add nothing once connectivity is the target. Its strong performance on two datasets that differ in recording setting, electrode configuration and task demonstrates the broad applicability of a connectivity-based objective, and suggests its potential for the study of brain activity more widely.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, we disclose our use of generative AI tools. The only generative AI tool used in this work was Anthropic's Claude (Claude Opus 4.8, Opus 5.0, Opus 5.5, Fable 5.0 and Fable 5.1), used through the Claude Code coding assistant.

Tasks with required disclosure. We did not use generative AI to propose or refine hypotheses, to develop the theoretical model or conceptual framework, to design the research methodology or experiments or to give feedback on them, to implement methods, to generate synthetic datasets, to clean or reformat datasets, to interpret results, or for translation. Formulating mathematical claims, providing critical ingredients for proofs, writing proofs, and supporting qualitative or thematic data analysis do not apply to this work.

Tasks with recommended disclosure. We used Claude to help polish the writing of the paper; to help draft the appendix and the AI use and ethics statements; to help locate and fix bugs in author-written code; and to look up the ICLR author policies and the public terms of the Brain Treebank dataset. We did not use generative AI to create or modify scientific figures or images; to suggest the structure of the paper, a title or keywords; to brainstorm, find research topics or identify gaps; to identify, search for or summarise literature; to suggest experimental parameters; or to format references. Writing survey or interview questions and transcribing recordings do not apply.

Review of AI-assisted work. The authors reviewed and revised all AI-assisted text, checking the appendix against the run outputs, and reviewed and tested every AI-assisted code change and analysis script before it was used for any reported result. The submission contains no hidden text or instructions directed at AI-assisted review tools. The authors take full responsibility for the content of this paper, including any text, code or claims produced with the aid of generative AI.

ETHICS STATEMENT

This work is a secondary, offline analysis of previously collected intracranial recordings from people with epilepsy. Brain Treebank (Wang et al., 2024) is a public dataset whose participants include minors; its authors report IRB approval, informed consent and electrode placement dictated by clinical considerations alone, and we use it under its CC BY 4.0 licence. The chronic-cohort recordings were collected in a previously conducted clinical research study; to preserve anonymity during review, we withhold the recording site and system, which will be named at camera-ready. We report no demographic information, make no attempt to re-identify any participant, and release no chronic-cohort data. Our language tasks classify speech versus non-speech windows and detect sentence onsets; we do not decode speech content. No model is validated for clinical use.

REPRODUCIBILITY STATEMENT

The encoder, objective and connectivity target are specified in Section 3 and Appendix C; hyperparameters are listed in Appendix H and data splits in Appendix D. The anonymous supplementary material contains a condensed implementation of the encoder, the objective, the connectivity target and the training loops. Brain Treebank is publicly available (Wang et al., 2024); the chronic-cohort recordings are not.

REFERENCES

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, pp. 1298–1312. PMLR, 2022.

Danielle S. Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017. doi: 10.1038/nn.4502.

Steven L. Bressler and Vinod Menon. Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences*, 14(6):277–290, 2010. doi: 10.1016/j.tics.2010.04.004.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. MBrain: A multi-channel self-supervised learning framework for brain signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 130–141, 2023. doi: 10.1145/3580305.3599426.

Geeling Chau, Christopher Wang, Sabera J Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=FVuqJt3c4L.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Danny Dongyeop Han, Yonghyeon Gwon, Ahhyun Lucy Lee, Taeyang Lee, Seong Jin Lee, Jubin Choi, Sebin Lee, Jihyun Bang, Seungju Lee, David Keetae Park, Shinjae Yoo, Chun Kee Chung, and Jiook Cha. DIVER-1: Scaling intracranial EEG foundation models for transferable representations. *arXiv preprint arXiv:2512.19097*, 2025.

Joerg F Hipp, David J Hawellek, Maurizio Corbetta, Markus Siegel, and Andreas K Engel. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature Neuroscience*, 15 (6):884–890, 2012.

R. Matthew Hutchison, Thilo Womelsdorf, Elizabeth A. Allen, Peter A. Bandettini, Vince D. Calhoun, Maurizio Corbetta, et al. Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, 80:360–378, 2013.

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Mark A. Kramer and Sydney S. Cash. Epilepsy as a disorder of cortical network organization. *The Neuroscientist*, 18(4):360–372, 2012. doi: 10.1177/1073858411422754.

Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208, 1999.

Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureşan. Time-frequency super-resolution with superlets. *Nature Communications*, 12(1):337, 2021.

Lucine L. Oganesian, Saba Hashemi, and Maryam M. Shanechi. BaRISTA: Brain scale informed spatiotemporal representation of human intracranial neural activity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/pdf?id=LDjBDk3Czb.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

Cornelis J. Stam. Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695, 2014. doi: 10.1038/nrn3801.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Siyi Tang, Jared A. Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L. Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=k9bx1EfHI_-.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019.

Table 4: **Per-subject results, Sentence Onset.** Test ROC-AUC, mean ± s.e.m. over 5 finetuning seeds $\{0, \ldots, 4\}$; the last column is the across-subject mean ± s.e.m. ($n = 7$) reported in the main text. `sub_1` and `sub_3` are near-saturated (random init ≥ 0.93) and cannot separate methods. In each column the best value is bold and the second best underlined.

Arm	sub_1	sub_2	sub_3	sub_4	sub_6	sub_7	sub_10	Mean
Random initialisation	0.926±0.004	0.859±0.004	0.939±0.001	0.786±0.005	0.829±0.014	0.731±0.004	0.877±0.004	0.850±0.028
Masked latent reconstruction	0.918±0.001	0.860±0.003	0.941±0.001	0.798±0.005	0.875±0.010	0.729±0.009	0.861±0.002	0.854±0.027
Region InfoNCE	0.928±0.001	0.827±0.005	0.938±0.003	0.836±0.002	0.841±0.004	0.706±0.011	0.866±0.002	0.849±0.029
DANCE (ours)	0.910±0.002	0.857±0.004	0.936±0.001	0.851±0.003	0.874±0.002	0.694±0.004	0.895±0.002	0.860±0.030
DANCE + cross-channel attention	0.918±0.003	0.843±0.007	0.932±0.002	0.814±0.007	0.845±0.010	0.695±0.011	0.872±0.005	0.846±0.030
DANCE + cross-channel attention + parcel embedding	0.930±0.002	0.820±0.012	0.936±0.003	0.812±0.003	0.862±0.006	0.695±0.007	0.856±0.002	0.844±0.031
Masked latent reconstruction + cross-channel attention	0.920±0.001	0.837±0.009	0.924±0.003	0.681±0.015	0.826±0.020	0.712±0.008	0.860±0.005	0.823±0.036
Masked latent reconstruction + cross-channel attention + parcel embedding	0.916±0.001	0.839±0.007	0.935±0.001	0.757±0.005	0.817±0.019	0.724±0.008	0.866±0.002	0.836±0.029
BaRISTA (Oganesian et al., 2025), our reproduction	0.922±0.003	0.812±0.006	0.960±0.005	0.902±0.007	0.851±0.031	0.732±0.034	0.881±0.016	0.866±0.029
DANCE, dynamic PLV target	0.907±0.002	0.843±0.003	0.919±0.001	0.827±0.003	0.855±0.007	0.673±0.012	0.864±0.002	0.841±0.031
DANCE, evaluated patient excluded from pretraining	0.930±0.002	0.864±0.003	0.936±0.004	0.833±0.004	0.834±0.012	0.762±0.006	0.871±0.004	0.862±0.023

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent Berry, Leland Barnard, Benjamin Brinkmann, Gregory Worrell, David Jones, and Yogatheesan Varatharajah. Domain-guided self-supervision of EEG data improves downstream classification performance and generalizability. In *Proceedings of Machine Learning for Health (ML4H)*, volume 158 of *Proceedings of Machine Learning Research*, pp. 130–142. PMLR, 2021. URL https://proceedings.mlr.press/v158/wagh21a.html.

Changtian Wang, Adam Yaari, Anushree Singh, Vishnu Subramaniam, Dana Rosenfarb, Pratik Misra, Joseph R Madsen, Steven Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Zhizhang Yuan, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. Brant-2: Foundation model for brain signals. *arXiv preprint arXiv:2402.10251*, 2024.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=DDkl9vaJyE.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-IN: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 79996–80033, 2024.

A Per-subject results

Main-text tables report the across-subject mean; the tables below give every subject separately, as mean ± s.e.m. over the five finetuning seeds. We report these in full because several of our aggregate differences are small, and a reader should be able to see whether a mean is carried by a consistent shift or by one or two subjects. Note in particular that `sub_1` and `sub_3` are near-saturated and that `sub_7` is the hardest subject for every model including BaRISTA.

B The channel-readout ceiling

This appendix documents a methodological finding that materially affected our own results and, we believe, affects published comparisons of iEEG pretraining objectives more generally.

A downstream classifier must reduce a set of C per-channel embeddings to a single vector. A common choice, and the one we used initially, learns one scalar per channel: $o = \sum_c a_c \bar{h}_c$ with

Table 5: **Per-subject results, Speech / Non-Speech.** Test ROC-AUC, mean ± s.e.m. over 5 fine-tuning seeds $\{0, \ldots, 4\}$; the last column is the across-subject mean ± s.e.m. ($n = 7$) reported in the main text. `sub_1` and `sub_3` are near-saturated (random init ≥ 0.93) and cannot separate methods. In each column the best value is bold and the second best underlined.

Arm	sub_1	sub_2	sub_3	sub_4	sub_6	sub_7	sub_10	Mean
Random initialisation	$0.938_{\pm 0.003}$	$0.790_{\pm 0.004}$	$0.953_{\pm 0.003}$	$0.847_{\pm 0.005}$	$0.783_{\pm 0.006}$	$0.686_{\pm 0.011}$	$0.918_{\pm 0.001}$	$0.845_{\pm 0.037}$
Masked latent reconstruction	$0.946_{\pm 0.001}$	$0.801_{\pm 0.007}$	$0.951_{\pm 0.002}$	$0.861_{\pm 0.002}$	$0.826_{\pm 0.009}$	$0.698_{\pm 0.012}$	$\underline{0.930_{\pm 0.003}}$	$0.859_{\pm 0.035}$
Region InfoNCE	$0.948_{\pm 0.002}$	$0.842_{\pm 0.002}$	$0.948_{\pm 0.001}$	$0.846_{\pm 0.002}$	$\underline{0.846_{\pm 0.003}}$	$\underline{0.717_{\pm 0.001}}$	$0.903_{\pm 0.003}$	$0.864_{\pm 0.030}$
DANCE (ours)	$0.947_{\pm 0.003}$	$0.864_{\pm 0.002}$	$0.952_{\pm 0.001}$	$0.859_{\pm 0.001}$	$0.814_{\pm 0.008}$	$0.724_{\pm 0.002}$	$0.925_{\pm 0.002}$	$\underline{0.869_{\pm 0.031}}$
DANCE + cross-channel attention	$0.952_{\pm 0.001}$	$0.822_{\pm 0.003}$	$0.945_{\pm 0.001}$	$0.828_{\pm 0.004}$	$0.782_{\pm 0.005}$	$0.706_{\pm 0.014}$	$0.899_{\pm 0.002}$	$0.848_{\pm 0.034}$
DANCE + cross-channel attention + parcel embedding	$0.953_{\pm 0.002}$	$0.803_{\pm 0.004}$	$0.944_{\pm 0.001}$	$0.844_{\pm 0.006}$	$0.842_{\pm 0.003}$	$0.643_{\pm 0.005}$	$0.927_{\pm 0.001}$	$0.851_{\pm 0.041}$
Masked latent reconstruction + cross-channel attention	$0.926_{\pm 0.001}$	$0.783_{\pm 0.005}$	$0.946_{\pm 0.001}$	$0.801_{\pm 0.003}$	$0.773_{\pm 0.007}$	$0.671_{\pm 0.008}$	$0.892_{\pm 0.006}$	$0.827_{\pm 0.037}$
Masked latent reconstruction + cross-channel attention + parcel embedding	$0.932_{\pm 0.001}$	$0.794_{\pm 0.009}$	$\underline{0.954_{\pm 0.001}}$	$0.831_{\pm 0.002}$	$0.751_{\pm 0.006}$	$0.711_{\pm 0.012}$	$0.917_{\pm 0.003}$	$0.842_{\pm 0.036}$
BaRISTA (Oganesian et al., 2025), our reproduction	$0.951_{\pm 0.004}$	$0.882_{\pm 0.011}$	$0.967_{\pm 0.005}$	$0.922_{\pm 0.011}$	$0.812_{\pm 0.007}$	$0.626_{\pm 0.062}$	$0.798_{\pm 0.016}$	$0.851_{\pm 0.045}$
DANCE, dynamic PLV target	$0.937_{\pm 0.001}$	$0.839_{\pm 0.002}$	$0.940_{\pm 0.001}$	$0.828_{\pm 0.003}$	$0.835_{\pm 0.003}$	$0.626_{\pm 0.006}$	$0.895_{\pm 0.006}$	$0.843_{\pm 0.040}$
DANCE, static AEC target (session average)	$0.955_{\pm 0.001}$	$0.857_{\pm 0.002}$	$0.932_{\pm 0.001}$	$0.836_{\pm 0.004}$	$0.842_{\pm 0.003}$	$0.646_{\pm 0.007}$	$0.888_{\pm 0.003}$	$0.851_{\pm 0.038}$
DANCE, evaluated patient excluded from pretraining	$\underline{0.954_{\pm 0.001}}$	$\underline{0.864_{\pm 0.005}}$	$0.953_{\pm 0.001}$	$\underline{0.870_{\pm 0.003}}$	$0.847_{\pm 0.004}$	$0.700_{\pm 0.006}$	$0.934_{\pm 0.002}$	$0.875_{\pm 0.034}$

Table 6: **Anatomical components in the encoder, both objectives.** Test ROC-AUC (± s.e.m. across 7 subjects) with the per-channel encoder, with cross-channel attention added, and with cross-channel attention plus a Destrieux parcel embedding, for masked latent reconstruction (the positive control: a model with no inter-electrode prior) and for DANCE. Paired contrasts between consecutive rows are in Table 7. In each column the best value is bold and the second best underlined.

Objective	Encoder	Sentence Onset	Speech / Non-Speech
Masked reconstruction	per-channel	$\underline{0.854_{\pm 0.027}}$	$\underline{0.859_{\pm 0.035}}$
	+ cross-channel attention	$0.823_{\pm 0.036}$	$0.827_{\pm 0.037}$
	+ cross-channel attention + parcel embedding	$0.836_{\pm 0.029}$	$0.842_{\pm 0.036}$
DANCE	per-channel	$0.860_{\pm 0.030}$	$0.869_{\pm 0.031}$
	+ cross-channel attention	$0.846_{\pm 0.030}$	$0.848_{\pm 0.034}$
	+ cross-channel attention + parcel embedding	$0.844_{\pm 0.031}$	$0.851_{\pm 0.041}$

$a \in \mathbb{R}^C$. This map is *rank-one in the channel mixing*: each channel contributes its embedding scaled by a single number, so the readout cannot select different feature subspaces from different electrodes. For a spatially heterogeneous montage, where the informative subspace genuinely differs between an auditory contact and a hippocampal one, this is a severe restriction. We instead use the full map over concatenated channel embeddings,

$$o \;=\; W_{\text{agg}}\,[\,\bar{h}_1;\,\bar{h}_2;\,\ldots;\,\bar{h}_C\,], \qquad W_{\text{agg}} \in \mathbb{R}^{d \times Cd}, \tag{4}$$

where $\bar{h}_c$ is a learned attention pooling over the tokens of channel c.

Empirically it acts as a ceiling that binds every arm simultaneously and compresses the differences between them. Every objective ranking we had measured under the rank-one readout was invalidated and had to be re-measured; the numbers in this paper are all post-correction, and the superseded measurements are not reported.

Two consequences. First, the honest magnitude of the pretraining effect, once the ceiling is removed, is smaller than it appeared beforehand: a substantial part of what we had been measuring was the readout's limitation rather than the representation's quality.

Second, and more consequential, **the readout changes the measured pretraining effect through the control arm**. Comparing three readouts on identical weights and data across all seven subjects (five seeds each, speech), the pretraining Δ is $+0.028$ ($p = 0.039$) under the full map ($k = d$), $+0.024$ ($p = 0.048$) under the rank-8 map, and $+0.009$ ($p = 0.44$) under a per-token head that preserves the frequency axis. The rank-k map factorises Eq. 4 through k dimensions,

$$o \;=\; W_{\uparrow} W_{\downarrow}\,\big[\,\bar{h}_1;\,\ldots;\,\bar{h}_C\,\big], \qquad W_{\downarrow} \in \mathbb{R}^{k \times Cd},\; W_{\uparrow} \in \mathbb{R}^{d \times k}, \tag{5}$$

with $k = 8$ throughout the paper. The full and rank-8 maps are indistinguishable on both tasks (paired difference $+0.001$, $p = 0.41$ on speech; -0.002, $p = 0.46$ on sentence onset) while the latter has $8\times$ fewer parameters, which is why it is the readout of record. The per-token head acts by *raising the random-initialisation arm* ($0.842 \to 0.860$) while leaving the pretrained arm unchanged ($0.871 \to 0.869$): it nulls the measured effect through the control, not through the pretrained model. Parameter count does not explain this—the three heads span two orders of magnitude without tracking Δ—whereas whether the frequency axis is pooled *before* channel aggregation does. We tested and refuted one explanation on a single subject (injected log-band-power: ablating it under the per-token head costs -0.010 versus -0.059 under the full map); the cause is unknown and we do not

Table 7: **All pairwise contrasts.** Mean difference in test ROC-AUC, paired over the 7 subjects (subject means over 5 seeds), with two-sided Wilcoxon signed-rank p and the number of subjects on which the difference is positive. At $n = 7$ the smallest attainable p is 0.016, so these tests detect consistent sign, not effect magnitude.

Task	Contrast	Δ AUC	p	subj. improved
Sentence Onset	relational − random	+0.0099	0.688	3/7
Sentence Onset	infonce − random	−0.0008	0.938	3/7
Sentence Onset	maskboth − random	+0.0049	0.938	4/7
Sentence Onset	relational − infonce	+0.0107	0.297	4/7
Sentence Onset	relational − maskboth	+0.0050	0.812	2/7
Sentence Onset	relational − BaRISTA	−0.0060	0.688	3/7
Sentence Onset	rel_chanattn − relational	−0.0138	0.109	2/7
Sentence Onset	mb_chanattn − maskboth	−0.0317	0.047	1/7
Sentence Onset	rel_parcel − rel_chanattn	−0.0013	0.938	3/7
Sentence Onset	mb_parcel − mb_chanattn	+0.0135	0.219	5/7
Sentence Onset	relplv − random	−0.0084	0.688	2/7
Sentence Onset	relational − relplv	+0.0184	0.016	7/7
Sentence Onset	lopo − random	+0.0119	0.219	5/7
Sentence Onset	lopo − relational	+0.0020	1.000	4/7
Speech / Non-Speech	relational − random	+0.0241	0.031	6/7
Speech / Non-Speech	infonce − random	+0.0192	0.297	4/7
Speech / Non-Speech	maskboth − random	+0.0138	0.031	6/7
Speech / Non-Speech	relational − infonce	+0.0050	0.375	5/7
Speech / Non-Speech	relational − maskboth	+0.0103	0.812	4/7
Speech / Non-Speech	relational − BaRISTA	+0.0181	1.000	3/7
Speech / Non-Speech	rel_chanattn − relational	−0.0215	0.031	1/7
Speech / Non-Speech	mb_chanattn − maskboth	−0.0316	0.016	0/7
Speech / Non-Speech	rel_parcel − rel_chanattn	+0.0033	0.812	4/7
Speech / Non-Speech	mb_parcel − mb_chanattn	+0.0142	0.109	6/7
Speech / Non-Speech	relplv − random	−0.0024	0.688	2/7
Speech / Non-Speech	relational − relplv	+0.0265	0.078	6/7
Speech / Non-Speech	static − random	+0.0056	0.938	3/7
Speech / Non-Speech	relational − static	+0.0185	0.297	5/7
Speech / Non-Speech	lopo − random	+0.0296	0.031	6/7
Speech / Non-Speech	lopo − relational	+0.0054	0.219	6/7

Table 8: **Chronic cohort, per patient.** Each cell is the mean over 5 fine-tuning seeds; the across-patient mean ± s.e.m. is the value reported in Table 3; the last column is the difference from random initialisation with the number of patients improved (for MSE, a negative difference is an improvement). M5 recorded no seizures over the monitoring period and M9's fall outside the days used here. In each column the best value (for MSE the lowest) is bold and the second best underlined.

Pretraining objective	M1	M4	M5	M9	Mean	Δ vs. random
Sleep staging (macro-F1)						
Linear model on band power	0.924	0.789	0.823	<u>0.757</u>	$0.823_{\pm 0.036}$	—
Random init	0.945	0.804	0.825	0.721	$0.824_{\pm 0.046}$	—
Masked latent recon.	<u>0.957</u>	<u>0.805</u>	<u>0.833</u>	0.714	<u>$0.827_{\pm 0.050}$</u>	+0.004 (3/4)
BaRISTA (Oganesian et al., 2025), adapted	0.878	0.740	0.749	0.745	$0.778_{\pm 0.033}$	−0.046 (1/4)
DANCE (ours)	0.960	0.806	0.848	0.768	$0.845_{\pm 0.041}$	+0.022 (4/4)
IED rate (MSE, lower is better)						
Linear model on band power	5.333	4.730	0.552	3.701	$3.579_{\pm 1.064}$	—
Random init	5.236	5.202	0.564	4.235	$3.809_{\pm 1.106}$	—
Masked latent recon.	4.519	5.339	0.541	3.478	<u>$3.469_{\pm 1.048}$</u>	−0.340 (3/4)
BaRISTA (Oganesian et al., 2025), adapted	5.869	<u>4.671</u>	0.534	3.294	$3.592_{\pm 1.147}$	−0.217 (3/4)
DANCE (ours)	<u>4.939</u>	4.605	<u>0.539</u>	<u>3.477</u>	$3.390_{\pm 1.000}$	−0.419 (4/4)
Seizure (AUC)						
Linear model on band power	0.919	0.991	—	—	$0.955_{\pm 0.036}$	—
Random init	<u>0.956</u>	0.795	—	—	$0.875_{\pm 0.081}$	—
Masked latent recon.	0.957	0.964	—	—	<u>$0.961_{\pm 0.004}$</u>	+0.085 (2/2)
BaRISTA (Oganesian et al., 2025), adapted	0.868	<u>0.990</u>	—	—	$0.929_{\pm 0.061}$	+0.054 (1/2)
DANCE (ours)	0.950	0.983	—	—	$0.967_{\pm 0.017}$	+0.092 (1/2)

speculate. We state the readout alongside every Δ in this paper. We are not aware of another iEEG representation-learning paper that does so.

Table 9: **Chronic cohort, patient held out of pretraining.** Each patient is fine-tuned from an encoder pretrained on the other three patients' recordings only, with the recipe, splits and seeds of Table 3; *transductive* is the encoder of Table 3, whose pretraining included the patient's unlabelled recordings. Each cell is the mean over 5 fine-tuning seeds; the last column is the difference from random initialisation with the number of patients improved. With four patients a sign test cannot reach $p < 0.05$ ($4/4$ gives one-sided $p = 0.0625$), so no p-values are given. The held-out encoders were pretrained for 11.8k–12.5k optimiser steps on three patients' recordings, the transductive encoder for 7.0k on all four, so the held-out arm is not disadvantaged by training length.

Encoder	M1	M4	M5	M9	Mean	Δ vs. random
Sleep staging (macro-F1)						
Random init	0.945	0.804	0.825	0.721	$0.824_{\pm 0.046}$	—
DANCE, transductive	0.960	0.806	0.848	0.768	$0.845_{\pm 0.041}$	+0.022 (4/4)
DANCE, patient held out	0.954	0.800	0.834	0.767	$0.839_{\pm 0.041}$	+0.015 (3/4)
held out − transductive	−0.005	−0.007	−0.013	−0.001	−0.007	(0/4 better)
IED rate (MSE, lower is better)						
Random init	5.236	5.202	0.564	4.235	$3.809_{\pm 1.106}$	—
DANCE, transductive	4.939	4.605	0.539	3.477	$3.390_{\pm 1.000}$	−0.419 (4/4)
DANCE, patient held out	4.945	4.640	0.536	3.377	$3.375_{\pm 1.005}$	−0.434 (4/4)
held out − transductive	+0.007	+0.035	−0.002	−0.099	−0.015	(2/4 better)

This is not a claim about other published models. We initially believed the rank-one limitation was shared by prior iEEG readouts. It is not: measured on our own frozen features, BaRISTA's head shape scores 0.869 on `sub_2` and 0.894 on `sub_4`, *above* what `flatten` achieves. Their head pools the temporal-patch axis while keeping per-token weights, whereas our `linear` averaged the frequency axis away. The finding here concerns our own pipeline and the general principle that readout choice is not neutral; it is not a defect we attribute to prior work.

C Functional-connectivity target construction

Estimator. Within block b, let $e_{c,f}[\tau]$ be the log-magnitude envelope of channel c in band f, obtained from a superlet time-frequency decomposition (Moca et al., 2021). The per-band amplitude-envelope correlation is the Pearson correlation of the envelopes over the block,

$$\mathrm{AEC}_{ij}^{(f)}(b) = \frac{\sum_{\tau \in b}(e_{i,f}[\tau] - \bar{e}_{i,f})(e_{j,f}[\tau] - \bar{e}_{j,f})}{\sqrt{\sum_{\tau \in b}(e_{i,f}[\tau] - \bar{e}_{i,f})^2}\sqrt{\sum_{\tau \in b}(e_{j,f}[\tau] - \bar{e}_{j,f})^2}} \in [-1, 1], \tag{6}$$

and $M_{ij}(b) = |\mathcal{B}|^{-1} \sum_{f \in \mathcal{B}} \mathrm{AEC}_{ij}^{(f)}(b)$ over the band subset $\mathcal{B} = \{1, \ldots, 35\}$ (0.5–93 Hz). Note that the target is computed from a superlet decomposition while the encoder consumes an STFT: the superlet supplies the adaptive time-frequency resolution and the phase estimates needed for the PLV variant we ablate below, and the objective does not require the two decompositions to agree.

Why a block rather than a window. A correlation estimated over a single 3 s window is too noisy to regress onto, while a session-level average is static by construction and reduces the objective to a fixed spatial prior. The block length must therefore sit between the two. As a guide to the state timescale of this corpus we measured the lag-autocorrelation of the band-power features and found it to decay over approximately 6–9 s. We set the block to $K = 7$ windows (21 s): a few multiples of that timescale, long enough that the correlation over $7 \times T$ envelope samples is stable, and short enough that the target still differs across a session. Shorter blocks (≈ 10 s) are viable and more dynamic; longer ones (≥ 30 s) risk averaging across state transitions. We did not tune K downstream.

Why the target must be calibrated. The raw estimator is unusable as a cosine target. On a representative session the band-averaged off-diagonal AEC is 91% positive, instructing the model to make nearly every channel pair mutually similar; the corresponding PLV target is 99% *negative*, instructing near-universal repulsion, which is additionally unsatisfiable—one cannot place many points in d dimensions at mutual cosine -0.7, so the loss would be minimised by saturation rather than by matching structure. Comparing AEC against PLV without calibration would therefore compare an

attract-everything target against a repel-everything one, with the coupling signal buried under that offset. After the calibration of Eq. 2 the two land comparably (AEC: median -0.107, 59% negative; PLV: median -0.169, 69% negative) and differ only in the coupling metric, which is what the comparison is supposed to isolate.

Why band averaging must be plain, not standardised. Standardising each band before averaging is band-fair but unbounded: measured targets left the cosine range entirely ($[-2.37, +9.81]$ for PLV, $[-1.10, +4.84]$ for AEC), placing most of the target beyond what a cosine can reach. We use the plain band mean for the regression target, and reserve the standardised reduction for the scale-invariant percentile-based negative weighting used by the InfoNCE control.

Why the five highest bands are dropped. The target's superlet grid has 40 logarithmically spaced bands over 0.5–200 Hz; the five highest cover 108–200 Hz. Per-band phase-locking decays smoothly with frequency from 0.352 at 0.5 Hz to 0.084 at 68 Hz, as physiology predicts, then *rises* monotonically from 0.113 to 0.308 across the five highest bands while becoming uncorrelated with the band mean ($r = -0.07$). Phase locking does not increase with frequency, so we read the rise as an artefact of the decomposition at the top of its range and restrict the target to the 35 lowest bands (0.5–93 Hz). The chronic cohort's grid stops at 75 Hz and all of its bands are used.

How much of the target is dynamic. Between 66% and 75% of AEC variance (and 88–91% of PLV's) is stable across-pair structure; the remainder varies from block to block, and on a representative session that remainder is 31% of the variance of the calibrated target. A model could therefore satisfy much of the plain relational target by learning a fixed per-recording connectivity map. The comparison that measures what the block-to-block variation adds is a *static* target, which assigns every block that recording's average connectivity under the same calibration statistics, so that the two arms differ only in the presence of the state-varying component; the result is in Table 2.

D Dataset statistics and splits

Brain Treebank. The pretraining corpus is 17 sessions ($\approx$ 37 h of recording; $39{,}126$ windows of 3 s after segmentation). The seven evaluation sessions are disjoint from it; they come from the same subjects (see §4, transductivity) but never from the same session. Table 10 gives per-subject contact counts and split sizes. Splits are $80/10/10$; we verified that they reproduce the split sizes of Oganesian et al. (2025) in all 14 task $\times$ subject cells, and we report ours so the correspondence can be checked. Note the range of scale: `sub_7` has the most contacts (205) and the fewest training windows (624), which is why it is the hardest subject for every model in this paper.

Table 10: **Brain Treebank evaluation sessions.** Contacts retained after preprocessing, and windows per split for each task.

		Speech / Non-Speech			Sentence Onset		
Subject	Contacts	train	val	test	train	val	test
`sub_1`	91	2156	269	269	1488	186	186
`sub_2`	100	1470	183	183	1036	129	129
`sub_3`	91	1072	133	133	1066	133	133
`sub_4`	151	1468	183	183	1022	127	127
`sub_6`	135	638	79	79	480	60	60
`sub_7`	205	624	78	78	650	81	81
`sub_10`	173	916	114	114	944	117	117

Chronic cohort. Four patients, ten recording days each, 707 hours in total ($254{,}587$ windows of 10 s) used for pretraining. The downstream splits are chronological within patient (Table 11). Two properties of this cohort drive choices elsewhere in the paper. First, the class prior *shifts* between chronological splits: N3 falls from 11.1% of `M1`'s training epochs to 5.0% of its test epochs, and from 15.7% to 5.6% for `M9`. Evaluation therefore re-tilts the balanced posterior onto the validation prior rather than the uniform one (§H); scoring against a uniform prior on a skewed test split penalises exactly the rare stage that macro-F1 weights equally. Second, the seizure set is small in the

way that matters: its test splits contain a single pre-ictal event per patient, so we report no seizure decoding number (§4.1).

Table 11: **Chronic-cohort splits.** Windows per split, and sleep-stage counts as AWAKE / N3 / REM. Seizure windows are balanced by construction across inter-ictal / pre-ictal / ictal.

		Sleep staging (AWAKE / N3 / REM)			Seizure
Patient	Pretrain h	train	val	test	train/val/test
M1	197.5	10130/2130/7020	3732/549/2145	4545/321/1560	1395/441/747
M4	207.3	9822/1547/8425	3004/580/3014	3455/440/2703	1386/129/264
M5	209.4	8870/2018/5054	2270/694/2349	2518/791/2004	—
M9	93.0	11657/2843/3663	4649/517/888	4830/336/888	—

E PREPROCESSING AND ARTEFACT HANDLING

In the chronic cohort, 0.85% of windows show near-zero log-power simultaneously on all four channels—telemetry dropout, appearing as isolated windows (median run length 1). These are approximately nine interquartile ranges below the median, far more extreme than any physiological event, and are trivially predictable; left in place they constitute a shortcut for any reconstruction objective. We exclude them from pretraining, from the connectivity estimate, and from evaluation. We deliberately do *not* apply per-window amplitude normalisation to suppress ictal power: measured across all splits, the seizure median sits 0.54 IQR above the interictal median and seizures constitute 0.57% of the pretraining set, so the logarithmic transform already provides the needed compression, and per-window normalisation would destroy the across-band power profile that the encoder consumes (§3.2).

F CONTROL OBJECTIVES

Region InfoNCE. With $\mathcal{P}(i)$ a same-region positive and $\mathcal{N}(i)$ a set of 32 different-region negatives drawn under the region assignment of Oganesian et al. (2025), and $\hat{z}$ the normalised channel embedding of Eq. 1,

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{|\mathcal{V}|}\sum_{i}\log\frac{\exp\left(\hat{z}_i^{\top}\hat{z}_{\mathcal{P}(i)}/\tau\right)}{\exp\left(\hat{z}_i^{\top}\hat{z}_{\mathcal{P}(i)}/\tau\right)+\sum_{k\in\mathcal{N}(i)}\exp\left(\hat{z}_i^{\top}\hat{z}_k/\tau\right)}, \qquad \tau = 0.07. \tag{7}$$

This is the binary-relation counterpart of Eq. 3: every pair is either pulled or pushed, with a single global temperature in place of the measured coupling.

Masked latent reconstruction. A DATA2VEC-style objective (Baevski et al., 2022): an exponential-moving-average teacher supplies target token embeddings at masked positions and the student predicts them from the unmasked context. Masking is structured—contiguous time spans and contiguous frequency-band blocks, ratio 0.4—because i.i.d. patch masking of a spectrogram is solvable by two-dimensional interpolation and transfers below random initialisation (Appendix J). This is the objective family used by BaRISTA (Oganesian et al., 2025).

G ARCHITECTURAL VARIANTS

Two encoder stages are used only for ablation and are not part of DANCE.

Cross-channel attention. Two additional layers attend over the C electrodes within each (band, time-position) slice, with a key-padding mask for variable channel counts and no positional or coordinate encoding, so the stage remains permutation-equivariant. We place it after the temporal and spectral stages, and use fewer layers than those stages because this axis is the most expensive.

Anatomical parcel encoding. A learned embedding $e_{\mathrm{parcel}(c)} \in \mathbb{R}^d$ is added to every token of channel c, keyed on that electrode's hemisphere-specific Destrieux parcel (120 parcels plus a padding index) and shared across subjects, matching the spatial granularity of BaRISTA (Oganesian et al., 2025). The two variants are complementary by construction: parcel encoding supplies the *identity* of each electrode, while cross-channel attention is the *mechanism* through which identities can interact, so the former is close to inert without the latter. This is why the 2×2 of Table 6 is run with cross-channel attention enabled throughout. Parcel encoding requires localisation and is therefore unavailable in precisely the setting DANCE targets.

H HYPERPARAMETERS

Corpus-specific constants. The method is stated for a generic corpus; the values used are given in Table 12.

Table 12: Corpus-specific constants of the tokeniser and the connectivity target. $P = \lfloor (T-p)/s \rfloor + 1$ with patch length $p = 16$ and stride $s = 8$ in both corpora. The target's band grid is that of the superlet decomposition and is distinct from the encoder's STFT bins.

	Brain Treebank	Chronic cohort
Sampling rate f_s	2048 Hz	250 Hz
Window length	3 s	10 s
STFT window / hop (samples)	400 / 50	100 / 10
Encoder bins F (linear)	40, 0–200 Hz (5.12 Hz)	30, 2.5–75 Hz (2.5 Hz)
Frames T	124	251
Patches P	14	30
Target bands (superlet, log-spaced)	40, 0.5–200 Hz	30, 0.5–75 Hz
Bands entering M	lowest 35 (0.5–93 Hz)	all
Block K (windows / seconds)	7 / 21 s	2 / 20 s
Minimum run kept (windows)	3	2

Encoder. $d = 64$; $n_\ell = 6$ temporal and 6 spectral transformer layers; 4 attention heads; feed-forward width 256. 0.61 M parameters in the transferred encoder, 0.78 M including the pretraining projection head g_θ. The two attention stages run *sequentially*—all temporal layers, then all spectral layers—rather than interleaved. This regroups the token axis twice instead of once per layer, which keeps the computation graph static under variable C; the cost is that temporal features are formed first and spectral mixing acts on top of them rather than the two being refined jointly.

Pretraining. Adam, learning rate 2×10^{-4}, weight decay 0.1, gradient clipping at 1.0, cosine schedule with warmup. Batch size 2 with 16 accumulation steps across 4 GPUs (effective 32 windows). 30,300 optimiser steps. Validation and checkpoint selection every 100 steps. Data loading runs in the main process: forked dataloader workers aborted runs at validation boundaries on our cluster, at a throughput cost of roughly 10%.

Fine-tuning. Head learning rate 10^{-3}, backbone 3×10^{-5}, weight decay 0.1, batch size 8 with 8 accumulation steps, 30 epochs. A linear-probe stage precedes unfreezing of the top two blocks; warmup and unfreeze points are fractions of total optimiser steps so that subjects with order-of-magnitude different session lengths are treated equivalently. Spectrograms are standardised per (channel, band); during pretraining the statistics are those of the session, and downstream the validation and test splits use the training split's statistics. Model selection is on validation; all reported numbers are test.

Controls. The random-initialisation arm passes through the identical two-stage schedule including the linear-probe stage. This matters: an earlier version of our code skipped the probe stage when no checkpoint was loaded, which silently gave the random arm a different recipe from every pretrained arm and made the pretraining effect look larger than it is.

I COMPUTE

All experiments ran on NVIDIA A40 GPUs. A single pretraining run uses 4 GPUs for 32–48 hours (130–190 GPU-hours); the reported relational model took 38.6 hours (154 GPU-hours) for $30{,}300$ optimiser steps. Fine-tuning runs on a single GPU with a median wall-clock of 19.9 minutes; the 403 fine-tuning runs behind the tables in this paper total approximately 139 GPU-hours.

As the checklist asks whether the project consumed more compute than the reported experiments: it did, substantially. Across development we trained 27 distinct pretraining configurations for approximately $4{,}100$ GPU-hours, of which the arms reported here are a subset; the remainder were exploratory objectives, masking variants and combination schedules, several of which are reported as negative results in §J. Superseded fine-tuning runs conducted under the rank-one readout (§B) are not included in the figure above and were discarded.

J NEGATIVE RESULTS

We report these because they cost real compute and constrain the design space, and because a reader deciding whether to build on this objective benefits more from them than from another ablation that worked.

Combining masked reconstruction with the relational objective does not help. Trained jointly from scratch, the best combination reached $+0.041$ over random initialisation against $+0.077$ for the relational objective alone (panel measurement, pre-readout-correction scale). The two objectives conflict directly: the cosine similarity between their gradients on the shared encoder is -0.71. A naive loss weight does not balance them, because gradient magnitudes differ by roughly $57\times$; we implemented an online recalibration in gradient-contribution space and the combination remained a null.

Masked reconstruction alone transfers below random initialisation. Measured on three subjects at -0.084 relative to a readout-matched control. It becomes useful only in combination, which is itself a null—see above.

I.i.d. patch masking is solvable and useless. Independent per-patch masking of a spectrogram can be inverted by two-dimensional interpolation, and transfers below random initialisation. Structured masking—contiguous time spans and contiguous frequency bands—is required for the objective to be non-trivial.

The power-injection branch bypasses the mask. Because the injected log-band-power vector is computed from the clean spectrogram and added to every token, a fully masked band remains reconstructible from the injected profile alone. This is near-fatal for band masking specifically and inert for i.i.d. and span masking. We replace fully masked bands with the mean of surviving bands. We record this because it is the kind of leak that silently makes a pretext task trivial and is invisible in the loss curve.

K WHERE THE CLASSIFIER READS FROM

The rank-k readout of Eq. 5 enters each channel through one d-column block of $W_{\downarrow}$, so the norm of that block is a direct measure of how much the fine-tuned classifier draws on the electrode, the analogue of the per-parcel projection weights BaRISTA reports. DANCE receives no coordinates, so we map these norms to Destrieux regions only afterwards, for display. Figure 3 ranks the regions for the speech task. The classifier concentrates on lateral superior temporal cortex, planum polare, the transverse temporal sulcus and precentral cortex, which is the anatomy of speech perception and production. The control matters for how this is read: the identical readout fitted on the random-initialisation encoder concentrates on nearly the same regions (lateral superior temporal: 0.79 against 0.78), so the anatomy is recovered by the task-supervised head from the data and cannot be credited to pretraining. Where the two differ, DANCE places more weight on hippocampus (0.71 against 0.56) and insular cortex (0.69 against 0.60), sites whose coupling to auditory cortex is

state-dependent, which is consistent with the target's emphasis but is a description, not a test. The readout's attention over frequency bands carries no such structure: for both arms it stays within one percent of uniform (not shown).

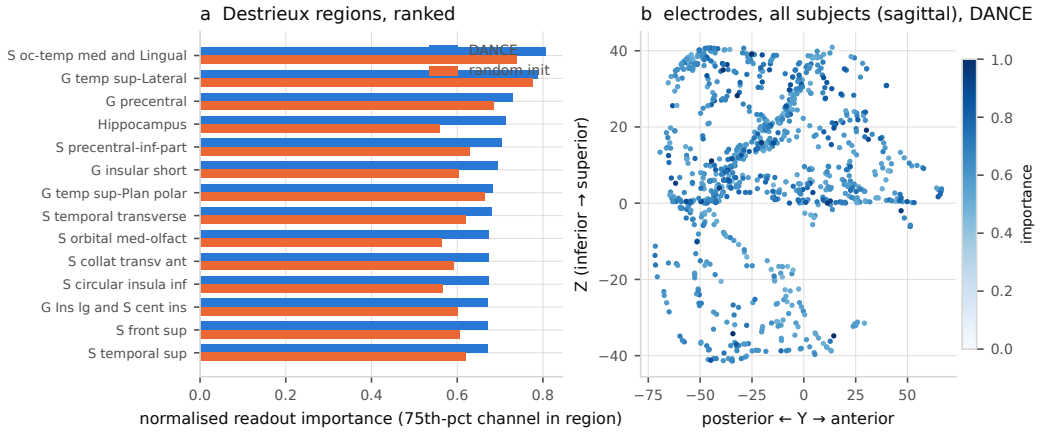


Figure 3: **Readout importance by region, speech.** (a) Destrieux regions ranked by the normalised norm of their channels' readout blocks (75th percentile within region, averaged over seeds and over the subjects that have the region), for DANCE and for the same readout fitted on the random-initialisation encoder. (b) Every electrode of the seven subjects in sagittal projection, coloured by its DANCE importance; coordinates are used for display only.

L Label efficiency

Pretraining is often argued to reduce the labels a downstream task needs. We fine-tune on a fixed, label-stratified 25% or 50% subset of each subject's training windows with the number of optimiser steps held constant, so that the amount of labelled data is the only variable (speech, rank-8 readout, seven subjects, five seeds; Figure 4). The pretraining gain is largest when labels are scarcest, $+0.039$ at 25% against $+0.024$ at 100% (six of seven subjects at both), but it is not monotone: at 50% it falls to $+0.017$ (four of seven), because `sub_7`, the subject with the most contacts and fewest windows, flips from $+0.085$ to -0.016 between the two. Only the 100% contrast is significant under the paired Wilcoxon test ($p = 0.031$; $p = 0.078$ at 25%). The stronger claim, that a pretrained model with a quarter of the labels matches random initialisation trained on all of them, holds on two subjects (`sub_2`, `sub_6`) and fails on the other five, so we do not make it: pretraining widens the margin when labels are scarce but does not substitute for four times as many.

M The learned geometry tracks measured coupling

The objective asserts that embedding similarity equals measured coupling, so the trained encoder can be checked against it on windows it was not fit to. Figure 5 evaluates ρ, the per-window Pearson correlation between $\cos(\hat{z}_i, \hat{z}_j)$ and the calibrated target A_{ij} over all channel pairs. On the 392 windows held out of the pretraining gradient, ρ averages 0.63 for DANCE (median 0.66) against 0.17 for the same encoder at random initialisation, and the two distributions are almost disjoint: the fifth percentile of the trained encoder (0.47) lies above the 95th percentile of the control (0.30). The untrained encoder scores low for a visible reason: it maps every channel pair to a cosine between 0.8 and 1, whereas the trained encoder spreads pairs across the target's range (panel a). These held-out windows come from recordings the encoder saw during pretraining. The same measurement for the leave-one-patient-out encoders, each evaluated on the patient it never saw, gives $\rho = 0.36$ on average (range 0.26–0.48 over the seven patients) against 0.14 for random initialisation (range 0.03–0.26), with the trained encoder above the control on every patient (panel c). The geometry transfers to a montage the encoder has never seen, at about half the strength it reaches on seen recordings, which is the same ordering as the downstream result.

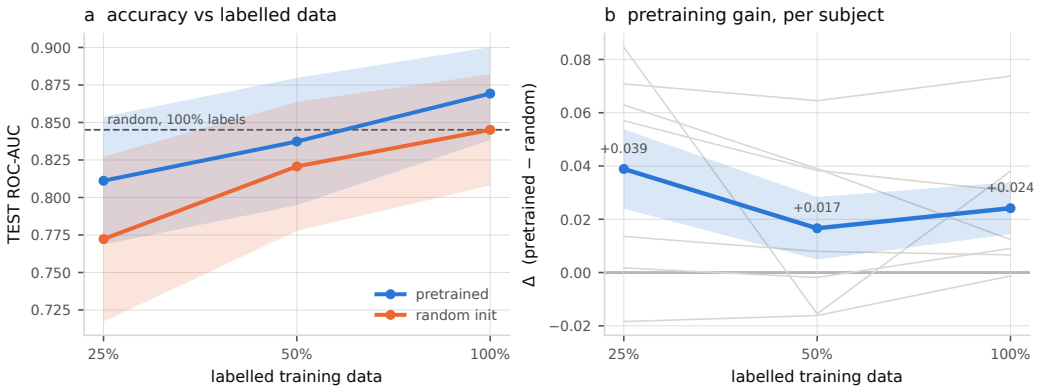


Figure 4: **Label efficiency.** (a) Speech test ROC-AUC against the fraction of labelled training windows, mean and s.e.m. over seven subjects, at a fixed optimiser-step budget; the dashed line is random initialisation at 100%. (b) The pretraining gain per subject (grey) and its mean.

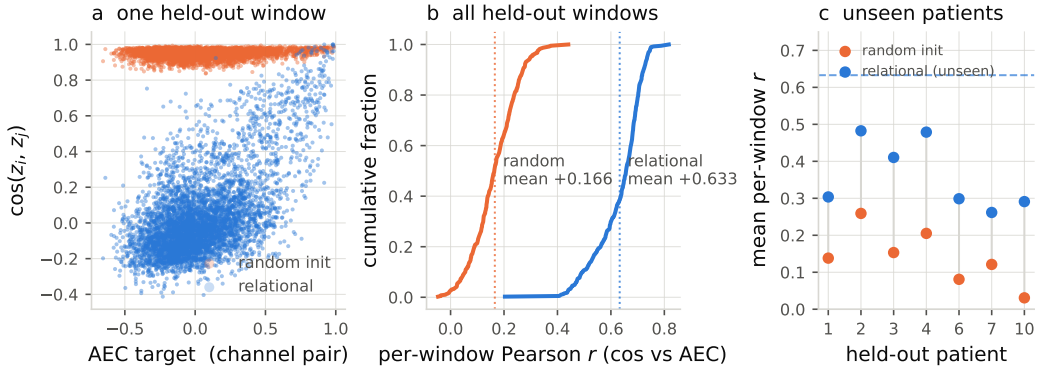


Figure 5: **Embedding similarity tracks measured coupling on held-out windows.** (a) One held-out window: cosine similarity of every channel-pair embedding against the calibrated AEC target, for DANCE and for the same encoder at random initialisation. (b) Per-window Pearson ρ over all held-out windows, as cumulative distributions; dotted lines are the means (0.63 and 0.17). Held-out windows are excluded from the pretraining gradient but come from recordings seen during pretraining. (c) Unseen patients: mean ρ of each leave-one-patient-out encoder on the patient it never saw, against random initialisation on the same windows; the dashed line is the mean over seen recordings from (b).

N EMBEDDING SIMILARITY ACROSS BRAIN STATES IN THE CHRONIC COHORT

Figure 6 reads the learned geometry against the two clinical state labels of the chronic cohort, descriptively: for each of the six electrode pairs, the mean cosine similarity of the two embeddings within each state, next to the within-window amplitude-envelope correlation of the same windows. The embedding similarity of a pair moves with state in the same direction as its measured coupling, most visibly for the thalamo-thalamic pair (`lA--rA`) and in the ictal state. No decoding claim is attached to this figure: the sleep labels are a clinician-reviewed automated hypnogram, and the seizure panel pools all windows of every split (2,583 for `M1`, 1,779 for `M4`) from 19 and 8 seizures respectively, so its ictal and pre-ictal points summarise a handful of events.

Figure 7 shows the same embeddings without aggregation: for each patient and task, a two-dimensional t-SNE projection of the pretrained encoder's electrode embeddings, one point per window and contact, coloured by state. The encoder is the pretrained model before any fine-tuning, so the structure visible here is what the connectivity target alone induces. Sleep stages occupy partly distinct regions in every patient and ictal windows separate from interictal ones in `M4`; no state is

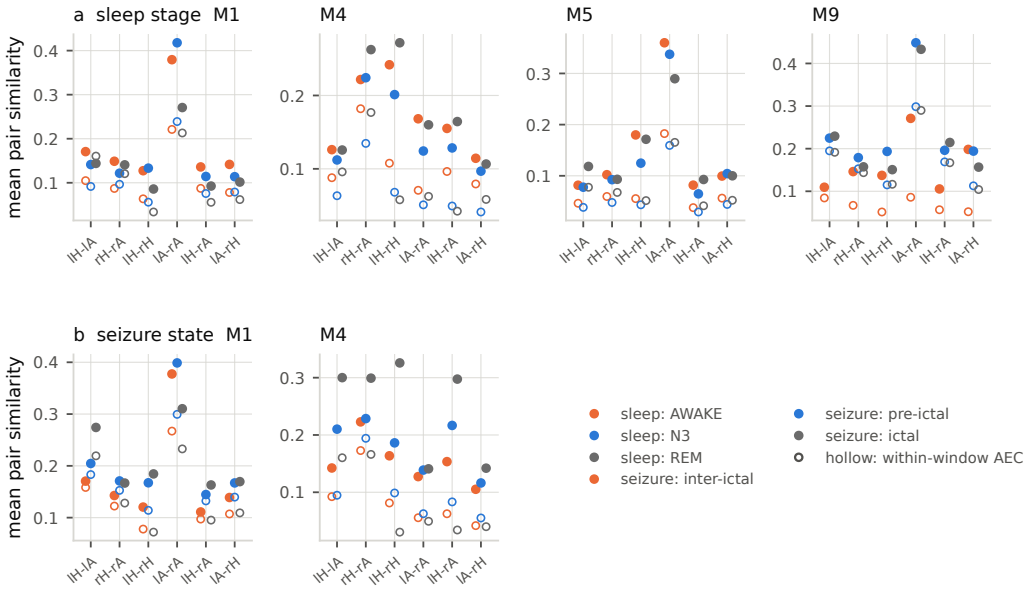


Figure 6: **Pair similarity by brain state, chronic cohort.** Mean cosine similarity between the embeddings of each electrode pair (filled) and the within-window amplitude-envelope correlation of the same windows (hollow), per patient, by (a) sleep stage and (b) seizure state. Contacts: `lH`, `rH` left and right hippocampus; `lA`, `rA` left and right anterior thalamus. Descriptive only; see text for label provenance and event counts.

separable, and the projection is shown to convey the organisation of the space rather than as evidence of decoding, which the fine-tuned results in Table 3 provide.

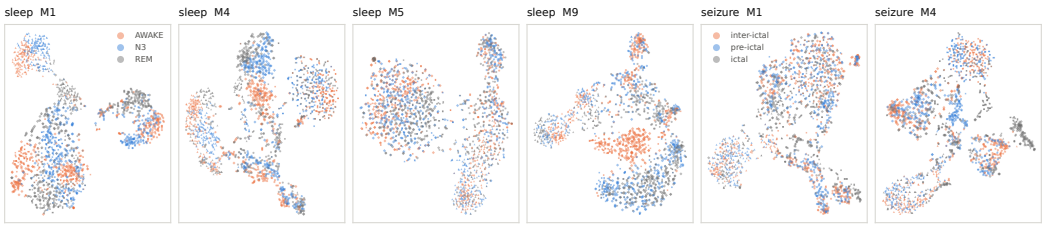


Figure 7: **Pretrained embeddings by brain state, chronic cohort.** t-SNE of the pretrained encoder's electrode embeddings (no fine-tuning), one panel per patient and task, one point per window and contact (marker shape), coloured by sleep stage or seizure state; up to 400 windows per state are drawn. Descriptive only.

O SEIZURE DETECTION ON THE CHRONIC COHORT

Curated seizure annotations exist for two of the four patients (§4.1); their ictal-vs-interictal ROC-AUC per arm is in Tables 3 and 8. On `M1` all three arms are at $\approx$ 0.95 and the task does not separate them; on `M4` the random-initialisation arm is unstable across seeds (0.795 ± 0.081) and both pretrained arms remove the instability (0.964 and 0.983). With two patients we draw no inference beyond that.

P HELD-OUT-SUBJECT GENERALISATION

The pretraining corpus contains other sessions of every evaluation subject, so the main comparison is transductive. To measure transfer to a patient the encoder has never seen, we pretrain seven relational encoders, each with one evaluation subject's sessions removed entirely, and fine-tune each on

the subject it excluded with the recipe and seeds of the main sweep (Table 1, last row; per-subject values in Tables 5 and 4; contrasts in Table 7). On speech the held-out-patient encoder scores 0.875: $+0.030$ over random initialisation ($p = 0.031$, six of seven) and $+0.005$ over the transductive encoder ($p = 0.22$). On sentence onset it scores 0.862: $+0.012$ over random initialisation ($p = 0.22$, five of seven) and $+0.002$ over the transductive encoder. Excluding the patient therefore costs nothing detectable on either task; what transfers is not the patient's identity. Two caveats. The held-out folds were trained for 33 epochs against 100 for the transductive encoder, so a shorter transductive control would be needed to bound the comparison from above; and the fold that excludes `sub_2` removes five of the seventeen pretraining sessions, so its encoder saw a third less data than the others. The corresponding mechanism check, the correlation between embedding similarity and measured coupling on the unseen patient, is reported in §M and Figure 5c.