What Do EEG Foundation Models Capture from Human Brain Signals?

Ling Tang[1,2*] **Qian Chen**[2,3*] **Jilin Mei**[2,4] **Houshi Xu**[1]
Quanshi Zhang[1] **Jing Shao**[2] **Na Zou**[2,5] **Xia Hu**[2] **Dongrui Liu**[2†]
[1]Shanghai Jiao Tong University [2]Shanghai Artificial Intelligence Laboratory
[3]Tongji University [4]Fudan University [5]University of Houston

Abstract

Electroencephalography (EEG) foundation models learn directly from brain signals, yet their predictions on clinical tasks remain a black box. We compare their internal representations with hand-crafted features developed over decades of EEG research to ask what these models encode, which features they use, and how much of their predictive advantage those features explain. We evaluate three models on five tasks using 72 features. On validation data, linear probes predicted feature values from model activity in 89.4% of the 1,080 combinations of tasks, models, and features. In 68.7% of all combinations, erasing activity associated with the feature caused a significant performance drop that exceeded the drop under a matched random intervention. These effects spanned all eight feature families. Spectral features had the largest sum of positive performance drops across individual erasures, even after adjustment for redundancy. In the Closure analysis, a linear classifier trained on features supported by erasure recovered, on average, 79.4% of the models' advantage over a baseline trained on the same number of random inputs.

1 Introduction

Clinicians interpret electroencephalography (EEG) through recognizable waveforms and rhythms. Epileptiform discharges help neurologists assess suspected epilepsy [Smith, 2005], while sleep spindles and K-complexes identify stage N2 [Younes and Hanly, 2016]. Under propofol, loss of consciousness is accompanied by stronger slow activity and frontal alpha rhythms [Purdon et al., 2013]. After subarachnoid hemorrhage, a sustained fall in the alpha-to-delta power ratio can signal delayed cerebral ischemia [Claassen et al., 2004].

Decades of EEG research have produced hand-crafted descriptors of waveforms, spectra, and changes in frequency content over time [Hjorth, 1970, Singh and Krishnan, 2023], as well as measures of complexity, coupling, and spatial organization [Bandt and Pompe, 2002, Tort et al., 2010, Stam et al., 2007]. Drawing on this literature, including work on aperiodic spectra and microstates [Donoghue et al., 2020, Michel and Koenig, 2018], we organize the features tested here into eight families: time-domain morphology (T), spectral power and shape (F), time-frequency envelope dynamics (TF), complexity (C), cross-frequency coupling (X), cross-channel relations (R), aperiodic spectral properties (A), and microstate dynamics (M).

In recent years, EEG foundation models [Kostas et al., 2021, Jiang et al., 2024, Zhou et al., 2025, Wang et al., 2025, 2024] have begun to change how EEG signals are analyzed. Pretrained on large

*Equal contribution. Work done during an internship at Shanghai Artificial Intelligence Laboratory.
†Correspondence to: liudongrui@pjlab.org.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

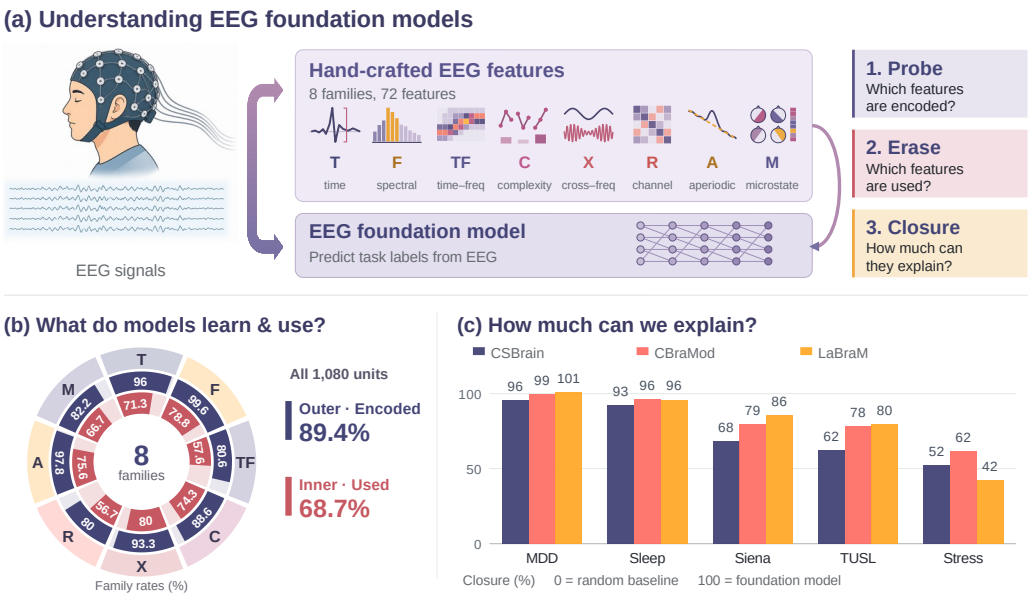

Figure 1: **(a)** We probe for EEG features, erase their associated activity, and predict task labels from features supported by erasure. **(b)** Within each family sector, the outer and inner arcs show the percentages of units with encoding and intervention support, respectively. **(c)** Closure measures how much of the foundation model's advantage the classifier recovers relative to a baseline with random inputs (100% denotes parity).

unlabeled EEG corpora through self-supervised learning, these models learn representations directly from the signal and can be fine-tuned for tasks such as sleep staging and seizure detection.

An EEG model can correctly flag a seizure without telling the clinician which part of the recording drove its decision. This opacity makes it difficult to judge whether a prediction will hold for the next patient. In other medical settings, pneumonia classifiers have exploited cues about the hospital where an image was acquired [Zech et al., 2018], and uncertainty about what a risk model had learned kept it from clinical testing [Caruana et al., 2015]. For EEG foundation models, the corresponding question is which familiar properties of the brain signal actually contribute to a decision.

To open this black box, we formalize the comparison between model decisions and hand-crafted EEG features as three questions. ***What does the model learn?*** Which hand-crafted features does the model encode, and at which layer? ***What does the model use?*** Of the encoded features, which ones does the model rely on for task decisions? ***How much can be explained?*** How much of the foundation model's advantage can a logistic regression classifier recover using these features, relative to a baseline trained on equally many random inputs?

We answer each question with a distinct interpretability method (Figure 1). Layer-wise ridge probes predict each feature from the model's activations and locate the layer at which it is most readable [Belinkov, 2022, Vig et al., 2020]. At that layer, an orthogonal cross-covariance eraser removes target-correlated components from the frozen model's representation [Belrose et al., 2023, Ravfogel et al., 2020]. We call an encoded feature representation-causal (RC), or confirmed, when the drop in test performance is significant and exceeds the drop under the random subspace control. Finally, a logistic regression classifier predicts task labels from the confirmed features. Closure measures how much of the foundation model's advantage this classifier recovers relative to a baseline trained on equally many random inputs. We apply these methods across three foundation models, five tasks, and 72 features, giving 1,080 combinations of dataset, model, and feature.

Across these units, we find four patterns. ***(i) Encoding.*** 965 units (89.4%) meet the validation encoding criterion, covering all 72 features. ***(ii) Usage.*** 742 units (68.7%) are representation-causal and 223 (20.6%) are encoded-only: they meet the encoding criterion but lack RC support. At the feature level, strong task-level support means RC support from all three models on one task. The 58 universal candidates have strong support on at least two tasks; 8 task-specific features have it on

exactly one; 6 model-specific features have RC support but no strongly supported task. ***(iii) Family effects.*** Spectral features have the largest mean sum of positive score drops after individual erasures, even after adjustment for redundancy; all eight families receive intervention support. ***(iv) Recovery.*** Confirmed features recover 79.4% of the model's advantage over the baseline with random inputs on average, with task means of $\approx 99\%$ for major depressive disorder classification (MDD) and $\approx 52\%$ for stress detection (Stress).

2 Related work

We review three areas of research: hand-crafted EEG features in clinical and neuroscience research, EEG foundation models, and interpretability methods for EEG classifiers.

Hand-crafted EEG features in clinical and neuroscience research. Decades of clinical and neuroscience work have produced a catalog of EEG features. Time-domain morphology is captured by the Hjorth parameters [Hjorth, 1970]; frequency-domain content by band power across the canonical $\delta, \theta, \alpha, \beta, \gamma$ rhythms [Welch, 1967, Buzsáki and Draguhn, 2004]. Cross-frequency interactions have been characterized as mechanisms for neural communication [Canolty et al., 2006, Canolty and Knight, 2010, Jensen and Colgin, 2007, Lisman and Jensen, 2013], with the Tort modulation index [Tort et al., 2010] the standard measure. Inter-channel relations are formalized through the phase-locking value [Lachaux et al., 1999] and the phase-lag index (PLI) [Stam et al., 2007], signal complexity through permutation entropy [Bandt and Pompe, 2002] and sample entropy [Richman and Moorman, 2000], and surrogate-data tests assess statistical structure [Theiler et al., 1992, Schreiber and Schmitz, 2000, Prichard and Theiler, 1994]. Aperiodic spectral parameterization separates broadband background properties from oscillatory peaks [Donoghue et al., 2020], while microstate analysis describes recurring scalp-map configurations and their temporal organization [Michel and Koenig, 2018]. We use these descriptors and the computational proxies defined in Appendix A to examine their relationship with learned representations.

EEG foundation models. Self-supervised pretraining on unlabeled EEG corpora has produced a family of foundation models [Kostas et al., 2021, Wang et al., 2023, Yang et al., 2023, Zhang et al., 2023, Jiang et al., 2024, Wang et al., 2024, Foumani et al., 2024, Cui et al., 2024, Jiang et al., 2025, Zhou et al., 2025, Wang et al., 2025], spanning contrastive pretraining, masked patch reconstruction, long-context transformers, and cross-scale tokenization. These works report strong downstream performance, but which signal content their representations rely on has remained largely open.

Interpretability for EEG models. Existing work mainly applies post-hoc attribution to individual classifiers: Shapley additive explanations (SHAP) and layer-wise relevance propagation (LRP) for clinical classifiers [Sylvester et al., 2024, Nam et al., 2023, Yi et al., 2025], gradient-weighted class activation mapping (Grad-CAM) and SHAP for sleep staging and schizophrenia diagnosis [Dutt et al., 2023, Saadatinia and Salimi-Badr, 2024, Almadhor et al., 2025], surrogate-based channel importance and ictal attribution [Lee et al., 2025, Sánchez-Hernández et al., 2024], frequency-band perturbation [Nahmias and Kontson, 2020], and concept-activation vectors for transformers [Gjølbye et al., 2023]. Our study asks how established EEG descriptors relate to model representations and task performance across multiple models and tasks.

3 Method

We use three interpretability methods to trace the relationship between EEG features and model predictions. Section 3.1 introduces the models, tasks, and feature targets shared by all three analyses. To identify what a model encodes, we fit a linear probe at each layer to predict each feature from the model's activations (Section 3.2). For each encoded feature, we then erase the associated component at its peak layer and compare the change in task performance with a random-subspace control to test whether the model uses that feature (Section 3.3). Finally, for each combination of dataset and model, we train a logistic regression classifier on the features supported by erasure and measure how much of the foundation model's advantage it recovers relative to a baseline with random inputs (Section 3.4).

3.1 Setup and notation

We work with three pretrained EEG foundation models, $m \in \mathcal{M} = \{\text{CSBrain}, \text{CBraMod}, \text{LaBraM}\}$, on five EEG tasks, $d \in \mathcal{D} = \{\text{MDD}, \text{Stress}, \text{Sleep}, \text{TUSL}, \text{Siena}\}$ (Figure 1). We call each pair (d, m) of a dataset and a model a *cell*; adding a feature q gives one *unit*. Each foundation model has L_m analyzable backbone layers (encoder layers for CSBrain and CBraMod, transformer blocks for LaBraM); the activation of model m at layer l on an EEG epoch $x \in \mathbb{R}^{C \times T}$ (C channels, T time samples) is $h_{m,l}(x) \in \mathbb{R}^{d_{m,l}}$. The task-fine-tuned model with its linear classifier head is f_m, and $M(f_m, d)$ is its task metric on d (area under the receiver operating characteristic curve, ROC-AUC, for binary tasks MDD, Stress, and Siena; class-averaged F1, macro-F1, for Sleep and TUSL).

The feature lexicon $\mathcal{Q}$ contains 72 hand-crafted EEG features in the eight families defined in Section 1. The six families T, F, TF, C, X, and R contribute 328 columns; the aperiodic and microstate families contribute 51, for 379 in total (Appendix A). We write $z_q(x) \in \mathbb{R}^{p_q}$ for the p_q-dimensional target vector of feature q. A feature can contribute multiple columns; a row corresponds to one sample (one 30 s epoch for Sleep). This gives $5 \times 3 \times 72 = 1{,}080$ combinations of dataset, model, and feature.

All datasets are split into training, validation, and test partitions. The training partition supplies all standardization, probe fitting, eraser estimation, residual probe fitting, and surrogate-classifier training; the validation partition supplies regularization strength, the encoding peak layer, and the validation-side encoded flag; the test partition supplies held-out probe scores, task metrics, and the intervention evidence used by the representation-causal criterion below.

3.2 What does the model learn?

To test whether a feature is linearly readable from the model's representation, we predict its value with a linear function of the activation, following the probing-classifier paradigm of language and vision interpretability [Conneau et al., 2018, Belinkov, 2022]. We fit a ridge regression probe at each backbone layer because most hand-crafted EEG features are continuous-valued. Comparing layers locates the encoding peak.

For each triplet (d, m, q) and each layer $l \in \{1, \ldots, L_m\}$, we fit a probe

$$P_{d,m,q,l} \,:\, h_{m,l}(x) \;\mapsto\; z_q(x), \tag{1}$$

on the training partition with regularization chosen on validation. The probe strength on split $s \in \{\text{val}, \text{test}\}$ is the coefficient of determination clipped at zero,

$$R^{(s)}_{d,m,q,l} \;=\; \max\!\big(0,\; R^2_s\big(P_{d,m,q,l}(h_{m,l}(x)),\; z_q(x)\big)\big), \tag{2}$$

averaged across output dimensions when $p_q > 1$. The encoding peak layer is selected on the validation split, $l^*_{d,m,q} = \arg\max_l R^{(\text{val})}_{d,m,q,l}$, and the test-split probe strength $R^{(\text{test})}_{d,m,q,l^*}$ is reported as held-out evidence. We call a unit *encoded* when its validation probe meets the score, control-margin, and non-dominance criteria in Appendix D.5; only encoded units enter the erasure stage. A separate *test-encoded* flag applies the analogous criterion on the test split as a diagnostic; it does not enter the encoded counts or gate erasure. Appendix C.3 compares ridge and nonlinear probes on nine Sleep units.

3.3 What does the model use?

Linear readability is a property of the activation, not of the network's downstream computation [Belinkov, 2022, Vig et al., 2020]. To assess reliance, we use an orthogonal cross-covariance eraser within the concept-erasure family [Belrose et al., 2023, Ravfogel et al., 2020, 2022a,b]. We center the activation and target on training data, compute their cross-covariance, and remove its retained left-singular-vector subspace. This is an orthogonal erasure choice, related to the LEAst-squares Concept Erasure (LEACE) framework's `orth` option; it does not use the activation-covariance whitening of minimum-distortion LEACE. Appendix E states the distinction and the numerical rank rule; Appendix C.1 reports the operator comparison.

At the encoding peak $l^* \equiv l^*_{d,m,q}$, write $h \equiv h_{m,l^*}(x)$, $\mu_h = \mathbb{E}[h]$, $\mu_z = \mathbb{E}[z_q]$, and $\Sigma_{hz} = \mathbb{E}[(h - \mu_h)(z_q - \mu_z)^\top]$. We center activations and targets on the training partition; we do not estimate

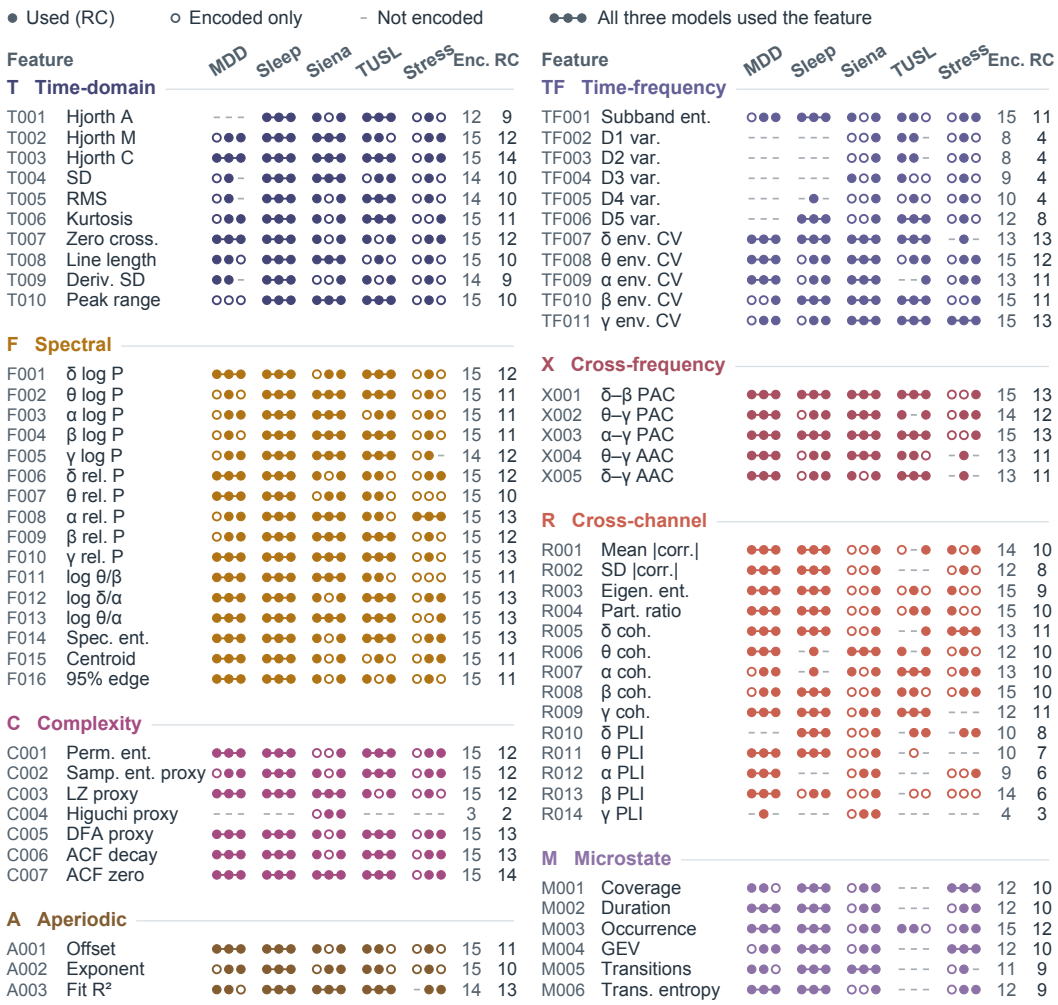

Figure 2: Encoding and intervention support for 72 EEG features across five tasks. Each row is one feature. Within each task, the three positions correspond to CSBrain, CBraMod, and LaBraM from left to right. Color identifies the feature family. A filled circle denotes intervention-supported use (RC), an open circle denotes encoding alone, and a gray dash denotes no detected encoding. A line joins the three dots when all three models show RC support on that task. Enc. and RC count the combinations of dataset and model with encoding and RC support, respectively, out of 15. Feature IDs follow Appendix A.

or use the activation covariance Σ_{hh}, so the eraser removes the Euclidean orthogonal projection onto the column span of Σ_{hz} rather than the whitened LEACE projector. We form $\Sigma_{hz} = USV^\top$ by singular value decomposition (SVD) and let U_r collect the left singular vectors whose singular values exceed a numerical rank threshold (or, if none do, the leading p_q vectors). The eraser is

$$\mathcal{E}_{d,m,q}(h) \;=\; h - \Pi_q(h - \mu_h), \qquad \Pi_q \;=\; U_r U_r^\top, \tag{3}$$

where Π_q projects onto the retained target-correlated subspace; $\mathcal{E}$ removes this component. We construct an edited model $\tilde{f}_{m,q}$ that runs the standard forward pass up to l^*, replaces $h_{m,l^*}(x)$ with $\mathcal{E}_{d,m,q}(h_{m,l^*}(x))$, and continues with the unchanged remainder of the network and classifier head. No model parameters are modified. The drop in test-split task metric on the edited model gives the erasure drop,

$$\Delta_{\text{erase}}(d, m, q) \;=\; M(f_m, d) - M(\tilde{f}_{m,q}, d). \tag{4}$$

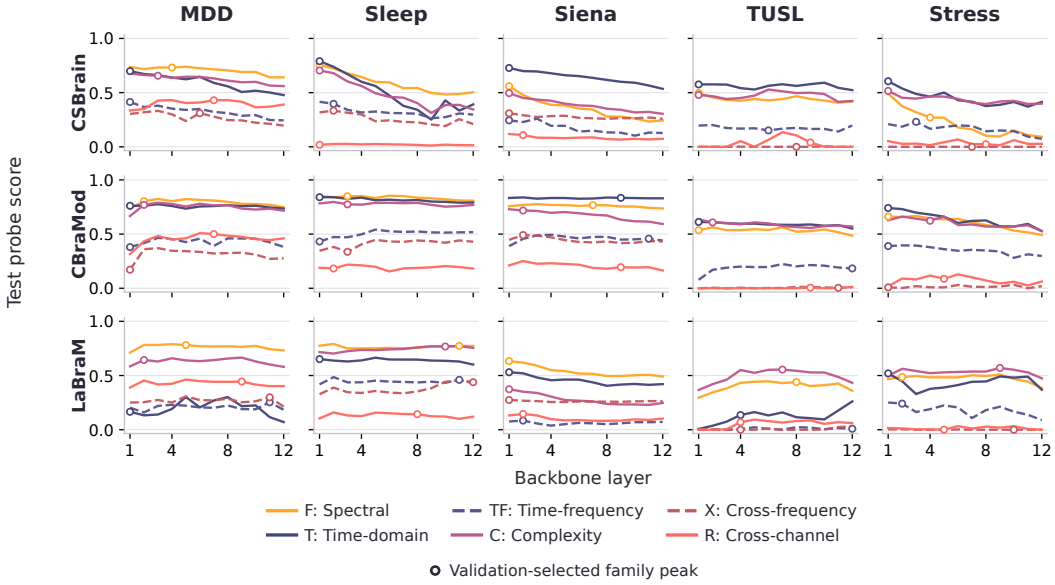

Figure 3: Encoding profiles for six feature families across backbone layers. Curves show median test probe scores (per-output R^2 clipped at zero, then averaged) within families T, F, TF, C, X, and R over all 12 layers. Open circles mark the layer with the largest family-median validation score. Rows are models and columns are tasks; all panels share the same probe-score scale. Individual-feature erasure sites are selected separately from family-median peaks.

We use the same representation-causal (RC) criterion across all eight families in the 72-feature lexicon. To test whether a drop is specific to the target, we compare erasure with removal of a random orthogonal subspace matched to the target expansion dimension p_q, which can differ from the retained eraser rank [Ravfogel et al., 2020, 2022a]. A feature q receives RC support on (d, m) when four conditions hold: it meets the validation encoding criterion, the test-set bootstrap lower bound for Δ_{erase} exceeds zero, the Benjamini–Hochberg (BH) adjusted q-value controlling the false discovery rate (FDR) within the cell is below 0.05, and $\Delta_{\text{erase}} - \Delta_{\text{erase}}^{(\text{rand})} > 0$. Separate diagnostics use shuffled pairings of inputs and targets, independent Gaussian targets of the same dimension, and a ridge probe refit on the erased activations. Appendix D.6 describes these checks and their coverage. RC support describes the frozen model's response to erasure.

3.4 How much can be explained?

The third question measures predictive recovery from RC features using a surrogate classifier, related to concept bottleneck models [Koh et al., 2020]. For each cell (d, m), let $\mathcal{Q}^*_{d,m} = \{\, q \in \mathcal{Q} : q \text{ is representation-causal on } (d, m) \,\}$, and train a logistic regression classifier $c_{d,m}$ on the concatenation $\big[z_q(x)\big]_{q \in \mathcal{Q}^*_{d,m}}$ with a fixed L_2 penalty (Appendix E). As a reference, B_{rand} is the same classifier on i.i.d. Gaussian features matched in dimension to the concatenated RC feature vector, controlling input dimension when comparing the selected features with random inputs. The Closure ratio is

$$\text{Closure}(d, m) \;=\; \frac{M(c_{d,m}, d) - M(B_{\text{rand}}, d)}{M(f_m, d) - M(B_{\text{rand}}, d)}, \tag{5}$$

the fraction of f_m's advantage over B_{rand} that the confirmed features recover; here B_{rand} is fitted per cell so that its column count matches $\sum_{q \in \mathcal{Q}^*_{d,m}} p_q$, and Closure is undefined when the denominator is non-finite or its absolute value is at most 10^{-12} (Appendix D.10). We denote $c_{d,m}$ by B_{rep} in comparisons; table entries under B_{rep} and B_{rand} are their task scores, and FM denotes $M(f_m, d)$. Because $\mathcal{Q}^*_{d,m}$ is defined using test-side intervention confirmation, Closure is a descriptive recovery analysis of a test-informed feature set. We report this ratio as a percentage, with 100% denoting

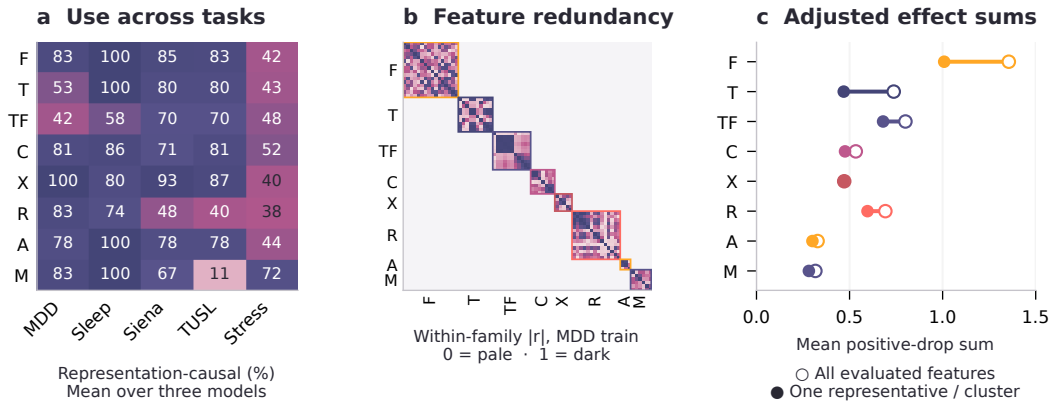

Figure 4: Family support and redundancy. **(a)** Representation-causal rate by task and family, averaged over the three models. **(b)** Within-family absolute correlations of training-set feature summaries on MDD; cross-family entries are masked. **(c)** Mean sums of positive drops across 15 cells before (open) and after (filled) retaining one largest-drop feature per greedy correlation cluster.

parity with the foundation model. Closure can exceed one when the surrogate outperforms the foundation model.

4 Experiments

4.1 Setup

Datasets and models. We evaluate the models on five EEG clinical tasks, *i.e.*, MDD on Mumtaz, Stress on Mental Arithmetic, ISRUC-Sleep, classification of background, seizure, and slowing events on the Temple University Hospital EEG Slowing Corpus (TUSL), and Siena seizure detection, scored by ROC-AUC for binary tasks and macro-F1 for multiclass. The three foundation models are CSBrain [Zhou et al., 2025], CBraMod [Wang et al., 2025], and LaBraM [Jiang et al., 2024]; for each (d, m) cell we load the released fine-tuned checkpoint with a linear classifier head and freeze the backbone. The lexicon is the catalog of 72 features in eight families (Appendix A), which together yield a vector of 379 dimensions.

Splits and statistics. Training rows fit probes, erasers, residual probes, and surrogate classifiers; validation selects probe regularization and encoding peaks; test rows supply reported metrics. We use the provided train/validation/test partitions. Paired bootstraps resample analysis rows (samples for MDD, Stress, TUSL, and Siena; 30 s epochs for Sleep), and the 72-feature analysis applies BH correction within each cell at $q < 0.05$. Implementation details are described in Appendix E.

4.2 Foundation models encode most of the lexicon

We asked how consistently the models encoded each feature across tasks and where along the network the feature was easiest to read out. We counted the units that met the encoding criterion and located each unit's peak using its validation probe scores (Section 3.2).

Coverage across feature families. Of the 1,080 units, 965 (89.4%) met the encoding criterion, and all 72 features passed in at least one model on one task. Spectral features passed in 239 of 240 units, time-domain features in 144 of 150, and complexity features in 93 of 105. Encoding also extended to the aperiodic and microstate families, which together passed in 118 of 135 units. Figure 3 shows the layer profiles for the time, spectral, time-frequency, complexity, cross-frequency, and cross-channel families. All 328 columns in these six families passed the non-finite checks; the Higuchi slope proxy (C004) had low variance on four tasks and remains flagged.

Depth of peak encoding. Among encoded units, the median peak was at layer 1 of 12 for CSBrain, layer 3 for CBraMod, and layer 5 for LaBraM. For CSBrain, 250 of 319 units peaked within the first four layers, compared with 146 of 318 for LaBraM. Depth also depended on the task: within

Table 1: Linear Closure using the lexicon of 72 features. RC counts test-supported representation-causal features. B_{rep} uses that feature set, B_{rand} matches its input dimension, and FM is the foundation model's metric. Closure reports the ratio in Equation 5 as a percentage.

Model	Task	RC	B_{rep}	B_{rand}	FM	Closure (%)
CSBrain	MDD	44	0.978	0.516	0.998	95.9
	Sleep	56	0.748	0.191	0.792	92.6
	Siena	45	0.817	0.519	0.955	68.4
	TUSL	49	0.699	0.315	0.932	62.2
	Stress	8	0.670	0.535	0.794	52.1
CBraMod	MDD	62	0.984	0.481	0.987	99.4
	Sleep	66	0.757	0.179	0.779	96.4
	Siena	39	0.832	0.503	0.918	79.4
	TUSL	51	0.689	0.406	0.766	78.5
	Stress	56	0.744	0.517	0.884	61.9
LaBraM	MDD	53	0.988	0.497	0.984	100.9
	Sleep	63	0.755	0.187	0.778	96.0
	Siena	72	0.828	0.462	0.889	85.6
	TUSL	43	0.622	0.315	0.700	79.8
	Stress	35	0.670	0.575	0.800	42.3
				Mean Closure across 15 cells:		**79.4**

LaBraM, the median was layer 11 on Sleep and layer 1 on Siena. We therefore used the peak for each individual feature, model, and task when choosing where to intervene.

4.3 Intervention support varies across tasks and models

We next assess intervention-supported use with the controlled erasure criterion in Section 3.3.

From encoding to use. Of the 1,080 triplets, 742 (68.7%) pass the representation-causal criterion, 223 (20.6%) remain encoded-only, and 115 (10.6%) are not-encoded. Of the 118 encoded units in the aperiodic and microstate families, 94 also passed the erasure test (Appendix B, Table 3).

Erasure effects across tasks. On TUSL/LaBraM, several time-frequency features (TF006 detail variance at level 5 and TF007–TF008 envelope coefficients of variation (CV) in δ/θ) and the related cross-frequency features (X005, X002) approach a metric-saturation drop of $\Delta_{\text{erase}} \approx 0.529$. On MDD/CBraMod, microstate and aperiodic descriptors have the largest family-level median erasure effects (0.095 and 0.086), followed by cross-channel relations (0.081) and spectral features (0.076). On Sleep, the largest family median is spectral on CSBrain (0.043), aperiodic on CBraMod (0.315), and complexity on LaBraM (0.210).

Large individual effects also recur across tasks. Of the 15 features with the largest cross-cell maximum erasure (Appendix B, Table 4), most have RC support in all five tasks, and 7 tie at the maximum drop 0.529. Every cell, including each Stress model, retains a non-empty $\mathcal{Q}^*_{d,m}$.

Agreement across models. Figure 2 reports encoding and intervention-support counts for each feature across the 15 cells. At the feature level, *strong task-level support* means that all three architectures satisfy the representation-causal criterion on the same task. Under this definition, 58 features are *universal candidates* (strong support in at least two tasks), 8 are *task-specific* (exactly one), and 6 are *model-specific* (no strongly supported task but at least one supported cell). For example, the aperiodic exponent was encoded in all 15 cells and passed the erasure test in ten. Only on Sleep did all three models pass, so this feature meets our criterion for task-specific support.

Redundancy within families. At the family level, spectral features have the largest mean sum of positive drops across the 15 cells (Figure 4c). Summing individual effects can count correlated signal properties more than once. We therefore cluster training-set feature summaries within each family at $|r| \geq 0.8$ and retain the largest-drop representative per cluster. The mean spectral sum decreases from 1.356 to 1.008, but remains the largest; all eight families retain positive adjusted sums.

4.4 Predictive recovery varies across tasks

We quantify predictive recovery with Closure (Equation 5), comparing a logistic regression surrogate trained on the confirmed features with the foundation model and a baseline trained on the same number of random inputs.

Recovery across tasks. Across the 15 cells, confirmed features recover 79.4% of the foundation model's advantage over B_{rand} on average (Table 1). Mean Closure is 98.7% on MDD, 95.0% on Sleep, 77.8% on Siena, 73.5% on TUSL, and 52.1% on Stress. MDD/LaBraM reaches 100.9%, indicating slightly higher surrogate performance on this test set. Stress/CSBrain retains eight RC features and reaches 52.1%, whereas CBraMod and LaBraM reach 61.9% and 42.3%.

Effect of expanding the lexicon. Adding the nine aperiodic and microstate descriptors raised mean Closure only slightly, from 79.34% to 79.43%. This average masks opposite changes on two tasks: mean Closure rose from 66.1% to 73.5% on TUSL and fell from 55.9% to 52.1% on Stress. In the expanded analysis, each random baseline was fitted to the input dimension of the corresponding confirmed feature set (Table 11).

Nonlinear readout on Stress. To assess how much the choice of classifier matters on Stress, we selected each model's confirmed features from the original lexicon of 63 features and compared logistic regression with a two-layer multilayer perceptron on those same inputs (Table 10). Closure rose from 68.0% to 70.0% for CSBrain, from 54.7% to 71.0% for CBraMod, and from 45.1% to 93.7% for LaBraM. For LaBraM in particular, a nonlinear classifier recovered much more of the model's advantage from the same features.

4.5 Robustness and sensitivity

Erasing two features together. On CSBrain/Sleep, erasing relative delta power (F006) alone reduced macro-F1 by 4.77 percentage points. Erasing the spectral centroid (F015) alone cost 3.15 points, but erasing both together cost 4.15 points, less than erasing F006 alone. The pattern differed for the complexity feature C003: its individual drop was 2.54 points, and its joint drop with F006 was 7.30, close to the sum of the two individual drops (Table 7). The sum closely approximated the joint drop for one pair and nearly doubled it for the other.

Probe sensitivity. On Sleep, replacing the linear probe with a multilayer perceptron increased the test probe score in six of nine comparisons; one score was unchanged at the reported precision and two were lower (Table 9). Both decreases occurred for the same phase-amplitude coupling feature, X001: its test R^2 fell from 0.257 to 0.223 on CBraMod and from 0.223 to 0.106 on LaBraM. On CSBrain the same feature improved from 0.287 to 0.370.

5 Discussion and conclusion

We asked which familiar EEG features foundation models encode, which they use to make predictions, and how much of their performance those features can explain. We tested 72 features across three models and five tasks. Linear probes could read out a feature in 89.4% of the 1,080 combinations of model, task, and feature. When we erased the activity associated with each encoded feature, performance fell significantly and more than under a random intervention in 68.7% of all combinations. A logistic regression classifier trained on the features supported by erasure recovered, on average, 79.4% of the models' advantage over a baseline trained on random inputs of the same dimension. Recovery ranged from 98.7% on MDD to 52.1% on Stress. Thus, familiar EEG features accounted for nearly all of the performance advantage on MDD, while the same linear readout left a larger gap on Stress.

Limitations. Our probes measure linear readability, and erasing one feature can also remove information it shares with others. The resulting performance drop therefore reflects sensitivity to the erased activity rather than a unique contribution from that feature. Closure describes recovery on the evaluated tasks: the test interventions helped select the features used by the classifier. Resampling EEG epochs can also understate uncertainty when epochs come from the same patient. Extending these findings to clinical decisions would require validation in independent patient cohorts.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 92370115 and 62276165. We thank Shanghai Artificial Intelligence Laboratory for its support. The authors declare no competing interests.

References

Ahmad Almadhor, Stephen Ojo, Thomas I. Nathaniel, Shtwai Alsubai, Abdullah Alharthi, Abdullah Al Hejaili, and Gabriel Avelino Sampedro. An interpretable XAI deep EEG model for schizophrenia diagnosis using feature selection and attention mechanisms. *Frontiers in Oncology*, 15:1630291, 2025. doi: 10.3389/fonc.2025.1630291.

Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88(17):174102, 2002. doi: 10.1103/PhysRevLett.88.174102.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL `https://aclanthology.org/2022.cl-1.7/`.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL `https://arxiv.org/abs/2306.03819`.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *Science*, 304 (5679):1926–1929, 2004. doi: 10.1126/science.1099745.

Ryan T. Canolty and Robert T. Knight. The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11):506–515, 2010. doi: 10.1016/j.tics.2010.09.001.

Ryan T. Canolty, Erik Edwards, Sarang S. Dalal, Maryam Soltani, Srikantan S. Nagarajan, Heidi E. Kirsch, Mitchel S. Berger, Nicholas M. Barbaro, and Robert T. Knight. High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793):1626–1628, 2006. doi: 10.1126/science.1128115.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. doi: 10.1145/2783258.2788613.

Jan Claassen, Lawrence J. Hirsch, Kurt T. Kreiter, Evelyn Y. Du, E. Sander Connolly, Ronald G. Emerson, and Stephan A. Mayer. Quantitative continuous EEG for detecting delayed cerebral ischemia in patients with poor-grade subarachnoid hemorrhage. *Clinical Neurophysiology*, 115: 2699–2710, 2004. doi: 10.1016/j.clinph.2004.06.017.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $\$\&!\#*$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2126–2136, 2018. doi: 10.18653/v1/P18-1198. URL `https://aclanthology.org/P18-1198/`.

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)*, 2024. URL `https://arxiv.org/abs/2311.03764`.

Thomas Donoghue, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H. Lara, Joni D. Wallis, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23(12):1655–1665, 2020. doi: 10.1038/s41593-020-00744-x.

Micheal Dutt, Surender Redhu, Morten Goodwin, and Christian W. Omlin. SleepXAI: An explainable deep learning approach for multi-class sleep stage identification. *Applied Intelligence*, 53 (13):16830–16843, 2023. doi: 10.1007/s10489-022-04357-8.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024. URL https://arxiv.org/abs/2402.17772.

Anders Gjølbye, William Lehn-Schiøler, Áshildur Jónsdóttir, Bergdís Arnardóttir, and Lars Kai Hansen. Concept-based explainability for an EEG transformer model, 2023.

Bo Hjorth. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29(3):306–310, 1970. doi: 10.1016/0013-4694(70)90143-4.

Ole Jensen and Laura L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, 11(7):267–269, 2007. doi: 10.1016/j.tics.2007.05.003.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Io9yFt7XH7.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://www.frontiersin.org/articles/10.3389/fnhum.2021.653659/full.

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. doi: 10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C.

Choel-Hui Lee, Daesun Ahn, Hakseung Kim, Eun Jin Ha, Jung-Bin Kim, and Dong-Joo Kim. NeuroXAI: Adaptive, robust, explainable surrogate framework for determination of channel importance in EEG application. *Expert Systems with Applications*, 261:125364, 2025. doi: 10.1016/j.eswa.2024.125364.

John E. Lisman and Ole Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013. doi: 10.1016/j.neuron.2013.03.007.

Christoph M. Michel and Thomas Koenig. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. *NeuroImage*, 180:577–593, 2018. doi: 10.1016/j.neuroimage.2017.11.062.

David M. Nahmias and Kimberly L. Kontson. Easy perturbation EEG algorithm for spectral importance (easyPEASI): A simple method to identify important spectral features of EEG in deep learning models. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2398–2406, 2020. doi: 10.1145/3394486.3403289.

Hyeonyeong Nam, Jun-Mo Kim, WooHyeok Choi, Soyeon Bak, and Tae-Eui Kam. The effects of layer-wise relevance propagation-based feature selection for EEG classification: A comparative study on multiple datasets. *Frontiers in Human Neuroscience*, 17:1205881, 2023. doi: 10.3389/fnhum.2023.1205881.

Dean Prichard and James Theiler. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters*, 73(7):951–954, 1994. doi: 10.1103/PhysRevLett.73.951.

Patrick L. Purdon, Eric T. Pierce, Eran A. Mukamel, Michael J. Prerau, John L. Walsh, Kin Foon K. Wong, Andres F. Salazar-Gomez, Priscilla G. Harrell, Aaron L. Sampson, Aylin Cimenser, Shi-Nung Ching, Nancy J. Kopell, Casie Tavares-Stoeckel, Kathleen Habeeb, Rebecca Merhar, and Emery N. Brown. Electroencephalogram signatures of loss and recovery of consciousness from propofol. *Proceedings of the National Academy of Sciences*, 110:E1142–E1151, 2013. doi: 10.1073/pnas.1221180110.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 18400–18421, 2022a. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6034–6055, 2022b. doi: 10.18653/v1/2022.emnlp-main.405. URL https://aclanthology.org/2022.emnlp-main.405/.

Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000. doi: 10.1152/ajpheart.2000.278.6.H2039.

Mehrshad Saadatinia and Armin Salimi-Badr. An explainable deep learning-based method for schizophrenia diagnosis using generative data-augmentation. *IEEE Access*, 12:98379–98392, 2024. doi: 10.1109/ACCESS.2024.3428847.

Sergio E. Sánchez-Hernández, Sulema Torres-Ramos, Israel Román-Godínez, and Ricardo A. Salido-Ruiz. Evaluation of the relation between ictal EEG features and XAI explanations. *Brain Sciences*, 14(4):306, 2024. doi: 10.3390/brainsci14040306.

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3–4):346–382, 2000. doi: 10.1016/S0167-2789(00)00043-9.

Anupreet Kaur Singh and Sridhar Krishnan. Trends in EEG signal feature extraction applications. *Frontiers in Artificial Intelligence*, 5:1072801, 2023. doi: 10.3389/frai.2022.1072801.

S. J. M. Smith. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(Suppl 2):ii2–ii7, 2005. doi: 10.1136/jnnp.2005.069245.

Cornelis J. Stam, Guido Nolte, and Andreas Daffertshofer. Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. *Human Brain Mapping*, 28(11):1178–1193, 2007. doi: 10.1002/hbm.20346.

Sophia Sylvester, Merle Sagehorn, Thomas Gruber, Martin Atzmüller, and Benjamin Schöne. SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods. *Behavior Research Methods*, 56(6):6067–6081, 2024. doi: 10.3758/s13428-023-02335-7.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1–4):77–94, 1992. doi: 10.1016/0167-2789(92)90102-S.

Adriano B. L. Tort, Robert Komorowski, Howard Eichenbaum, and Nancy Kopell. Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of Neurophysiology*, 104(2):1195–1210, 2010. doi: 10.1152/jn.00106.2010.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

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 *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=xmcYx_reUn6.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://openreview.net/forum?id=lvS2b8CjG5.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://openreview.net/forum?id=c2LZyTyddi.

Eun-Gyoung Yi, Miseon Shim, Hyeon-Ho Hwang, Sunhae Jeon, Han-Jeong Hwang, and Seung-Hwan Lee. Layer-wise relevance propagation approach for diagnosis of drug-naïve men with major depressive disorder using resting-state electroencephalography. *Depression and Anxiety*, 2025:5512539, 2025. doi: 10.1155/da/5512539.

Magdy Younes and Patrick J. Hanly. Minimizing interrater variability in staging sleep by use of computer-derived features. *Journal of Clinical Sleep Medicine*, 12:1347–1356, 2016. doi: 10.5664/jcsm.6186.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15:e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

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 36 (NeurIPS)*, 2023. URL https://openreview.net/forum?id=DDkl9vaJyE.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://openreview.net/forum?id=agcXjEHmyW.

A Hand-crafted EEG feature lexicon

The lexicon comprises 72 features in eight families. Families T, F, TF, C, X, and R contain 49 per-channel descriptors and 14 cross-channel descriptors, contributing 328 dimensions. Three aperiodic and six microstate descriptors contribute 51 further dimensions, giving 379 in total. Identifiers such as F013, A001, and M001 link the registry to all result tables.

Notation. Let $x \in \mathbb{R}^T$ denote one channel of an EEG epoch with T samples and sampling rate f_s, $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ its mean, and $\dot{x}_t = x_{t+1} - x_t$ its first finite difference. Let $S(f) = |X(f)|^2$ denote the one-sided fast Fourier transform (FFT) bin energy of x, where $X(f)$ is the discrete Fourier transform of the entire epoch (no windowing, no segment averaging, no overlap); we use this single-window FFT-bin energy as a proxy for spectral power throughout [Welch, 1967]. Let $\mathcal{H}[x]$ denote the analytic signal of x (with the original signal as its real part and its Hilbert transform as its imaginary part), and write $A(t) = |\mathcal{H}[x](t)|$ for its amplitude envelope and $\phi(t) = \arg \mathcal{H}[x](t)$ for its instantaneous phase. The five canonical bands are $\delta = [0.5, 4)$, $\theta = [4, 8)$, $\alpha = [8, 13)$, $\beta = [13, 30)$, $\gamma = [30, 45)$ Hz, indexed $b \in \{1, \ldots, 5\}$, and are clipped at each task's preprocessing low-pass and Nyquist limit (so the gamma band on ISRUC-Sleep is effectively $[30, 35)$ Hz). Let x_b denote x band-pass filtered to band b via an FFT-domain rectangular band mask. For multi-channel features, $X \in \mathbb{R}^{N \times T}$ is an N-channel epoch and ρ_{ij} is the Pearson correlation between channels i and j (computed with population normalization).

Family T: time-domain morphology (10 features, per-channel)

The Hjorth descriptors [Hjorth, 1970] summarize amplitude and slope statistics directly in the time domain. All variances and standard deviations in this family use the population estimator $(1/T)$; the kurtosis is the bias-corrected Pearson kurtosis (`scipy.stats.kurtosis` with `fisher=False, bias=False`); and the line-length feature stores the mean absolute first difference, consistent with the formula below.

$$\text{T001 (Hjorth activity)} = \mathrm{Var}(x) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2, \tag{6}$$

$$\text{T002 (Hjorth mobility)} = \sqrt{\mathrm{Var}(\dot{x}) \,/\, \mathrm{Var}(x)}, \tag{7}$$

$$\text{T003 (Hjorth complexity)} = \text{T002}(\dot{x}) \,/\, \text{T002}(x), \tag{8}$$

$$\text{T004 (standard deviation)} = \sqrt{\mathrm{Var}(x)}, \tag{9}$$

$$\text{T005 (root-mean-square)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}, \tag{10}$$

$$\text{T006 (kurtosis)} = 3 + \frac{(T^2-1)K - 3(T-1)^2}{(T-2)(T-3)}, \tag{11}$$

$$\text{T007 (zero-crossing rate)} = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbf{1}[x_t \cdot x_{t+1} < 0], \tag{12}$$

$$\text{T008 (line length)} = \frac{1}{T-1}\sum_{t=1}^{T-1} |x_{t+1} - x_t|, \tag{13}$$

$$\text{T009 (derivative std.)} = \sqrt{\mathrm{Var}(\dot{x})}, \tag{14}$$

$$\text{T010 (peak-to-peak amplitude)} = \max_t x_t - \min_t x_t. \tag{15}$$

Here $K = \left[T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^4\right] / \mathrm{Var}(x)^2$ is the uncorrected moment ratio. The T006 expression applies to nonconstant epochs with $T > 3$; the implementation uses `nan_policy="omit"` and replaces non-finite outputs with zero.

Family F: spectral power and shape (16 features, per-channel)

Let $P_b = \sum_{f \in B_b} S(f)$ denote the FFT-bin energy summed within band b and $P_{\text{total}} = \sum_{b=1}^{5} P_b$ the FFT-bin energy summed over the union of canonical bands. The band-ratio features are stored as log ratios; the spectral entropy, centroid, and edge are computed over the canonical-band FFT bins only.

$$\text{F00}b \text{ (log band energy, } b = 1, \ldots, 5) = \log P_b, \tag{16}$$

$$\text{F00}(5{+}b) \text{ (relative band energy)} = P_b \,/\, P_{\text{total}}, \quad b = 1, \ldots, 5, \tag{17}$$

$$\text{F011 (log theta–beta ratio)} = \log\bigl(P_\theta \,/\, P_\beta\bigr), \tag{18}$$

$$\text{F012 (log delta–alpha ratio)} = \log\bigl(P_\delta \,/\, P_\alpha\bigr), \tag{19}$$

$$\text{F013 (log theta–alpha ratio)} = \log\bigl(P_\theta \,/\, P_\alpha\bigr), \tag{20}$$

$$\text{F014 (normalized spectral entropy)} = -\frac{1}{\log N_F} \sum_{f \in B_{\text{tot}}} p(f) \log p(f), \quad p(f) = S(f)/P_{\text{total}}, \tag{21}$$

$$\text{F015 (spectral centroid)} = \sum_{f \in B_{\text{tot}}} f \cdot p(f), \tag{22}$$

$$\text{F016 (spectral edge 95\%)} = \min\left\{ f \in B_{\text{tot}} : \sum_{f' \le f} p(f') \ge 0.95 \right\}, \tag{23}$$

where B_{tot} is the union of the five canonical-band frequency bins and $N_F = |B_{\text{tot}}|$ is its bin count.

Family TF: time-frequency envelope dynamics (11 features, per-channel)

Let d_k denote the detail coefficients at level k of a 5-level Haar-style dyadic decomposition of x (recursive even/odd $\pm$ averaging, $1/\sqrt{2}$ normalization; not Daubechies-4), let a_5 denote the final approximation, and let $\bar{E}_k = \text{mean}_n(d_{k,n}^2)$ for $k = 1, \ldots, 5$ and $\bar{E}_6 = \text{mean}_n(a_{5,n}^2)$ be the per-coefficient mean energies, with $\bar{E} = \sum_{j=1}^{6} \bar{E}_j$.

$$\text{TF001 (normalized subband entropy)} = -\frac{1}{\log 6} \sum_{j=1}^{6} (\bar{E}_j/\bar{E}) \log(\bar{E}_j/\bar{E}), \tag{24}$$

$$\text{TF00}(1{+}k) \text{ (detail variance, level } k) = \text{Var}(d_k), \quad k = 1, \ldots, 5, \tag{25}$$

$$\text{TF00}(6{+}b) \text{ (envelope CV in band } b) = \text{std}(A_b) \,/\, \text{mean}(A_b), \quad b = 1, \ldots, 5, \tag{26}$$

where $A_b(t) = |\mathcal{H}[x_b](t)|$ is the amplitude envelope of x after an FFT-domain rectangular band mask in band b, followed by the analytic transform; all variances and standard deviations in this family use population normalization.

Family C: signal complexity (7 features, per-channel)

Several entries approximate standard complexity measures to facilitate computation on GPUs.

$$\text{C001 (normalized order-3 permutation entropy [Bandt and Pompe, 2002])} = -\frac{1}{\log 6} \sum_{\pi \in \mathfrak{S}_3} p(\pi) \log p(\pi), \tag{27}$$

with embedding dimension $m = 3$ and lag $\tau = 1$, $p(\pi)$ the empirical frequency of ordinal pattern π over the $T - 2$ embeddings of x, and a deterministic tie-handling rule.

$$\text{C002 (sample-entropy proxy [Richman and Moorman, 2000])} = \log\bigl(1 + r^{-1}\,\overline{|\ddot{x}|}\bigr) - \log\bigl(1 + r^{-1}\,\overline{|\dot{x}|}\bigr), \tag{28}$$

with $r = 0.2 \cdot \text{std}(x)$, $\dot{x}$ the first finite difference, $\ddot{x}$ the second finite difference, and $\overline{|\cdot|}$ the temporal mean magnitude. This is a derivative-irregularity proxy.

$$\text{C003 (Lempel–Ziv (LZ)-style irregularity proxy)} = \frac{1}{2}\,\eta + \frac{1}{2}\,H_2, \tag{29}$$

where $b_t = \mathbf{1}[x_t > \mathrm{median}(x)]$, $\eta = \mathrm{mean}_t\, \mathbf{1}[b_{t+1} \neq b_t]$ is the binarized transition rate, and H_2 is the normalized Shannon entropy of the four two-bit pair states (b_t, b_{t+1}). This is a transition-rate / two-bit-entropy proxy.

$$\begin{aligned} &\text{C004 (Higuchi-style lag-difference slope proxy)} \\ &\quad = \mathrm{clip}_{[0,3]}\big(\text{OLS slope of } \log L^*(k) \text{ vs. } \log(1/k)\big), \end{aligned} \tag{30}$$

where OLS denotes ordinary least squares and $L^*(k) = \mathrm{mean}_t\, |x_{t+k} - x_t|$ for $k = 1, \ldots, k_{\max}$ with $k_{\max} = \min(8, \lfloor T/4 \rfloor)$. This is a lag-difference slope proxy.

$$\text{C005 (DFA-style exponent proxy)} = \mathrm{clip}_{[-1,2]}\big(\text{OLS slope of } \log F(s) \text{ vs. } \log s\big), \tag{31}$$

where DFA denotes detrended fluctuation analysis, $Y_t = \sum_{u=1}^{t}(x_u - \bar{x})$, $F(s)$ is the root-mean-square residual after per-window linear detrending of Y on the fixed dyadic window set $s \in \{16, 32, 64, 128, 256, 512\} \cap \{w : 2w \leq T\}$, and the slope is fit by ordinary least squares.

$$\text{C006 (normalized } 1/e\text{-decay ACF lag)} = \tau_{1/e} \,/\, \tau_{\max}, \tag{32}$$

$$\text{C007 (normalized first-zero ACF lag)} = \tau_0 \,/\, \tau_{\max}, \tag{33}$$

where ACF denotes autocorrelation function, $\rho(\tau)$ is the FFT-based autocorrelation, $\tau_{1/e} = \min\{\tau > 0 : \rho(\tau) < 1/e\}$, $\tau_0 = \min\{\tau > 0 : \rho(\tau) \leq 0\}$ (each defaulting to $\tau_{\max}$ when no qualifying lag exists), and $\tau_{\max} = \min(200, T - 1)$ is the maximum lag in samples.

Family X: cross-frequency coupling (5 features, per-channel)

The Tort modulation index [Tort et al., 2010] measures phase-amplitude coupling (PAC) between low-frequency band p and high-frequency band q. We compute x_p and x_q via FFT-domain rectangular band masks (Notation), then form $\phi_p(t) = \arg \mathcal{H}[x_p](t)$ and $A_q(t) = |\mathcal{H}[x_q](t)|$. Partition the phase circle into $N = 18$ equal bins, and let $\bar{A}_q(j)$ denote the mean of $A_q(t)$ over the times at which $\phi_p(t)$ falls in bin j. Define the normalized distribution $P(j) = \bar{A}_q(j) / \sum_{j'} \bar{A}_q(j')$. Below, MI denotes modulation index and D_{KL} is the Kullback–Leibler divergence from the uniform distribution U.

$$\mathrm{MI}_{p \to q} = \frac{D_{\mathrm{KL}}\big(P \,\|\, U\big)}{\log N}, \quad U(j) = 1/N, \tag{34}$$

$$\mathrm{X001} = \mathrm{MI}_{\delta \to \beta}, \quad \mathrm{X002} = \mathrm{MI}_{\theta \to \gamma}, \quad \mathrm{X003} = \mathrm{MI}_{\alpha \to \gamma}. \tag{35}$$

Amplitude–amplitude coupling (AAC) between bands p and q is the Pearson correlation of their FFT-bandpass-derived envelopes, clipped to $[-1, 1]$,

$$\mathrm{AAC}_{p,q} = \mathrm{clip}_{[-1,1]}\big(\mathrm{Corr}(A_p(t),\, A_q(t))\big), \tag{36}$$

$$\mathrm{X004} = \mathrm{AAC}_{\theta,\gamma}, \quad \mathrm{X005} = \mathrm{AAC}_{\delta,\gamma}. \tag{37}$$

The general role of cross-frequency coupling as a neural communication channel is reviewed in [Canolty et al., 2006, Canolty and Knight, 2010, Jensen and Colgin, 2007, Lisman and Jensen, 2013].

Family R: cross-channel relations (14 features, global)

Let ρ_{ij} be the population-normalized Pearson correlation between channels i and j, and let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of the signed correlation matrix ρ itself (clipped at a small positive ϵ to absorb numerical noise, then normalized to sum to one); R003 and R004 are computed from these eigenvalues of the signed correlation matrix rather than from the absolute correlation matrix $|\rho|$; coherence uses five pair-summary statistics, whereas PLI uses the mean over channel pairs

(Appendix D.3).

$$\text{R001 (corr. abs. mean)} = \frac{2}{N(N-1)} \sum_{i<j} |\rho_{ij}|, \tag{38}$$

$$\text{R002 (corr. abs. std)} = \mathrm{std}\big(\{|\rho_{ij}|\}_{i<j}\big), \tag{39}$$

$$\text{R003 (normalized eigenvalue entropy)} = -\frac{1}{\log N} \sum_{k=1}^{N} \lambda_k \log \lambda_k, \tag{40}$$

$$\text{R004 (participation ratio)} = \big(\sum_k \lambda_k\big)^2 \,/\, \sum_k \lambda_k^2. \tag{41}$$

The mean coherence-proxy in band b is the FFT-bin magnitude coherence $C_{ij}(b) = |S_{ij}(b)|/\sqrt{S_{ii}(b)\,S_{jj}(b)}$ (clipped to $[0,1]$), where $S_{ij}(b)$ is the FFT-bin cross-spectrum averaged over the bins in band b and $S_{ii}(b)$ is the corresponding auto-spectrum; this is a magnitude coherence proxy, not the Welch magnitude-squared coherence of Welch [1967]:

$$\text{R}(004{+}b) \text{ (magnitude coherence proxy in band } b) = \frac{2}{N(N-1)} \sum_{i<j} C_{ij}(b), \quad b = 1, \ldots, 5. \tag{42}$$

The mean phase-lag index in band b [Stam et al., 2007] reduces common-source bias by ignoring zero-lag synchrony. Let $\Delta\phi_{ij}^{(b)}(t) = \phi_{b,i}(t) - \phi_{b,j}(t)$ be the instantaneous phase difference between channels i and j in band b after FFT-domain rectangular bandpassing and analytic-signal phase extraction; following common practice we implement the sign as $\mathrm{sign}\big(\sin(\Delta\phi)\big)$:

$$\text{R}(009{+}b) \text{ (PLI in band } b) = \frac{2}{N(N-1)} \sum_{i<j} \Big| \frac{1}{T} \sum_{t=1}^{T} \mathrm{sign}\big(\sin \Delta\phi_{ij}^{(b)}(t)\big) \Big|, \quad b = 1, \ldots, 5. \tag{43}$$

Coherence values are aggregated across the $N(N-1)/2$ unordered channel pairs by the five-statistic rule of Appendix D.3 (mean, std, median, 75%-quantile, top-$\lceil 10\% \rceil$ mean), giving five columns per band. PLI contributes one column per band: its mean over the same channel pairs. The phase-locking value [Lachaux et al., 1999] and surrogate-data testing methodology [Theiler et al., 1992, Schreiber and Schmitz, 2000, Prichard and Theiler, 1994] are the standard companion tools to this family but are not included in the feature lexicon.

Family A: aperiodic spectral properties (3 features)

A001 is the fitted aperiodic offset, A002 the aperiodic exponent, and A003 the goodness of fit (R^2) of the spectral parameterization. Each contributes six channel-summary dimensions using the existing aggregation convention. These descriptors distinguish the broadband spectral background from the original band-power catalog.

Family M: microstate dynamics (6 features)

Microstate descriptors use four templates fitted on training data. M001 reports time coverage, M002 mean segment duration, M003 occurrence rate, M004 global explained variance (GEV), M005 the transition-probability matrix, and M006 transition entropy. The first four have four dimensions each, M005 has $4 \times 4 = 16$, and M006 has one. Table 2 lists their dimensions.

B Robustness and failure record

We retain features that fail quality-control or encoding checks in the reported summaries.

Feature-quality control. The C004 Higuchi-style slope proxy triggers the low-variance flag on four of the five datasets (MDD, Sleep, Stress, TUSL); Siena does not trigger that flag. We retained C004 in the analyses and summaries.

Probe failures. Of the 1,080 combinations of dataset, model, and feature, 115 fail the encoding criterion of Appendix D.5 on validation; these are flagged *not-encoded* and excluded from the erasure

Table 2: Aperiodic and microstate descriptors and their output dimensions (51 in total).

ID	Descriptor	Dimensions
A001	Aperiodic offset	6
A002	Aperiodic exponent	6
A003	Aperiodic fit R^2	6
M001	Microstate coverage	4
M002	Mean duration	4
M003	Occurrence rate	4
M004	Global explained variance	4
M005	Transition probability	16
M006	Transition entropy	1

stage. The remaining 965 triplets enter the erasure stage and are reported as either *representation-causal* (742) or *encoded-only* (223); the test-encoded field of Appendix D.5 is recorded for transparency but does not gate the erasure stage. Table 3 reports the per-family per-cell breakdown referenced from Section 4.3.

Top features by maximum erasure. Table 4 lists the 15 features with the largest cross-cell maximum Δ_{erase} referenced from Section 4.3; seven features tie at the saturated drop of 0.529, and most of the top-15 are representation-causal in all five tasks.

Closure data flow. Standardization and classifier fitting use training rows, and probe hyperparameters and peak layers use validation rows. The representation-causal feature set is subsequently determined by test-side confirmation. The surrogate classifier therefore evaluates recovery from a test-informed feature set; a separately held-out partition or nested feature selection would be needed to measure unbiased generalization of that selection procedure.

C Supplementary comparisons

We examine sensitivity to the erasure operator, target resolution, readout, lexicon coverage, and bootstrap budget. The erasure-operator and Stress nonlinear-readout comparisons use the lexicon of 63 features in six families.

C.1 Orthogonal erasure and covariance-whitened LEACE

On ISRUC-Sleep, each operator erases T003, F010, or X001 at its validation-selected peak, followed by a refitted ridge probe on erased activations. Table 5 reports signed test R^2, rather than the zero-clipped main probe score. Both operators reduce all nine reported residual scores to zero or below at the displayed precision. Across the 945 units evaluated with the six-family lexicon, orthogonal erasure yields 648 RC units and covariance-whitened LEACE 634, with Dice overlap 0.849 (twice the intersection size divided by the sum of the two RC-set sizes) (Table 6).

C.2 Joint erasure and target resolution

On CSBrain/ISRUC-Sleep, F006 alone reduces macro-F1 by 4.77 percentage points from a 79.23% baseline. Table 7 reports joint erasure with each of five partners, using concatenated targets. The joint drop for F015 is smaller than the drop for F006 alone, while C003 is close to the sum of the individual drops. Estimated joint subspaces and nonlinear downstream predictions need not produce monotone or additive metric changes.

The main Sleep analysis uses six EEG input channels. In the separate resolution comparison (Table 8), the targets comprise six channel summary statistics, 19 channel values, or $19 \times 30 = 570$ patch values. The patch eraser is limited to rank six. For all three features, patch targets produced larger macro-F1 drops than summary targets.

Table 3: Family erasure summary for the lexicon of 72 features. Each entry gives the RC count and median Δ_{erase} over evaluated features; not-encoded entries are excluded from the median. A and M denote aperiodic and microstate descriptors.

CSBrain					
Family	MDD	Sleep	Siena	TUSL	Stress
F	10 (0.004)	16 (0.043)	14 (0.014)	14 (0.071)	1 (-0.012)
T	4 (0.003)	10 (0.028)	9 (0.015)	8 (0.048)	0 (-0.019)
TF	3 (0.002)	4 (0.016)	7 (0.012)	8 (0.066)	1 (-0.003)
C	5 (0.006)	6 (0.035)	5 (0.010)	6 (0.067)	0 (-0.006)
X	5 (0.005)	2 (0.005)	5 (0.010)	5 (0.066)	0 (-0.001)
R	10 (0.006)	9 (0.010)	2 (0.001)	4 (0.047)	4 (0.018)
A	2 (0.005)	3 (0.039)	2 (0.012)	3 (0.048)	0 (-0.011)
M	5 (0.003)	6 (0.031)	1 (0.001)	1 (0.086)	2 (0.023)
CBraMod					
Family	MDD	Sleep	Siena	TUSL	Stress
F	16 (0.076)	16 (0.304)	11 (0.013)	15 (0.197)	13 (0.097)
T	8 (0.026)	10 (0.261)	5 (0.008)	8 (0.142)	9 (0.098)
TF	5 (0.054)	8 (0.269)	5 (0.003)	9 (0.190)	10 (0.097)
C	6 (0.074)	6 (0.296)	3 (0.010)	5 (0.196)	6 (0.111)
X	5 (0.064)	5 (0.250)	4 (0.029)	4 (0.208)	3 (0.059)
R	13 (0.081)	12 (0.129)	4 (0.000)	6 (0.126)	6 (0.076)
A	3 (0.086)	3 (0.315)	2 (0.177)	3 (0.234)	3 (0.101)
M	6 (0.095)	6 (0.264)	5 (0.037)	1 (0.170)	6 (0.104)
LaBraM					
Family	MDD	Sleep	Siena	TUSL	Stress
F	14 (0.031)	16 (0.183)	16 (0.025)	11 (0.097)	6 (0.008)
T	4 (0.022)	10 (0.080)	10 (0.024)	8 (0.100)	4 (0.020)
TF	6 (0.238)	7 (0.144)	11 (0.028)	6 (0.121)	5 (0.020)
C	6 (0.060)	6 (0.210)	7 (0.025)	6 (0.136)	5 (0.038)
X	5 (0.264)	5 (0.103)	5 (0.030)	4 (0.150)	3 (0.045)
R	12 (0.032)	10 (0.072)	14 (0.021)	7 (0.082)	6 (0.010)
A	2 (0.217)	3 (0.175)	3 (0.029)	1 (0.084)	1 (0.025)
M	4 (0.011)	6 (0.108)	6 (0.022)	0 (0.032)	5 (0.040)

C.3 Nonlinear probes and Closure

Table 9 compares ridge with a validation-selected two-layer multilayer perceptron (MLP) on three targets across all models on Sleep. Six MLP scores are higher, one is equal at the reported precision, and two PAC scores are lower.

For Stress, we compare linear and nonlinear readouts on the same RC sets from the lexicon of 63 features in six families (Table 10). Recovery rises most for LaBraM. This supports nonlinear combinations as one source of linear-readout residuals.

C.4 Sensitivity to lexicon coverage

Including aperiodic and microstate descriptors contributes 135 combinations of dataset, model, and feature, of which 118 are encoded and 94 are RC. Encoding and RC labels for the shared 63 features remain unchanged. Table 11 compares linear Closure with six-family and eight-family lexicons; both the confirmed feature set and the baseline using random inputs of the same dimension can change. Table 12 compares alternative feature blocks within the six-family lexicon.

Table 4: Top-15 of 72 features by maximum Δ_{erase} across 15 cells. RC counts cells with representation-causal support for that feature; Tasks lists tasks with at least one such cell. Ties are ordered by identifier.

Feature	Family	Max	RC	Tasks supported
A003 (aperiodic fit R^2)	A	0.529	13	MDD, Sleep, Siena, TUSL, Stress
F010 (γ rel. energy)	F	0.529	13	MDD, Sleep, Siena, TUSL, Stress
TF006 (detail var. L5)	TF	0.529	8	Sleep, Siena, TUSL, Stress
TF007 (δ env. CV)	TF	0.529	13	MDD, Sleep, Siena, TUSL, Stress
TF008 (θ env. CV)	TF	0.529	12	MDD, Sleep, Siena, TUSL, Stress
X002 (PAC θ–γ)	X	0.529	12	MDD, Sleep, Siena, TUSL, Stress
X005 (AAC δ–γ)	X	0.529	11	MDD, Sleep, Siena, TUSL, Stress
T006 (kurtosis)	T	0.513	11	MDD, Sleep, Siena, TUSL, Stress
TF005 (detail var. L4)	TF	0.513	4	Sleep, Siena, TUSL, Stress
T009 (derivative std.)	T	0.505	9	MDD, Sleep, Siena, TUSL, Stress
X001 (PAC δ–β)	X	0.503	13	MDD, Sleep, Siena, TUSL, Stress
T001 (Hjorth activity)	T	0.497	9	Sleep, Siena, TUSL, Stress
R006 (θ coherence proxy)	R	0.480	10	MDD, Sleep, Siena, TUSL, Stress
F008 (α rel. energy)	F	0.479	13	MDD, Sleep, Siena, TUSL, Stress
T010 (peak-to-peak)	T	0.458	10	Sleep, Siena, TUSL, Stress

Table 5: Residual ridge test R^2 on Sleep after each eraser. Before denotes the unedited representation; Orth denotes orthogonal cross-covariance erasure, and LEACE denotes covariance-whitened erasure.

Model	Feature	Before	LEACE	Orth
CSBrain	T003	0.7944	-0.0165	-0.0165
	F010	0.7023	-0.0012	-0.0012
	X001	0.2874	-0.0255	-0.0255
CBraMod	T003	0.7651	-0.0165	-0.0165
	F010	0.7737	-0.0012	-0.0012
	X001	0.2570	-0.0253	-0.0255
LaBraM	T003	0.6896	-0.0165	-0.0165
	F010	0.7408	-0.0012	-0.0012
	X001	0.2229	-0.0254	-0.0255

C.5 Sensitivity to bootstrap budget

Table 13 summarizes bootstrap-budget sensitivity by the median across 15 cells. The Jaccard index (intersection size divided by union size) compares the reported positive-effect sets (features with positive erasure effects) against the $B = 2048$ reference; B denotes the number of bootstrap resamples. These sets are distinct from RC sets, which also require encoding and control/FDR criteria. Monte Carlo standard error (MCSE) quantifies simulation uncertainty; here it is reported for the estimate of mean erasure drop.

D Detailed experimental protocol

This appendix specifies data preparation, feature extraction, probing, erasure, and surrogate evaluation. The split roles, including the use of test-confirmed labels for Closure feature selection, are stated in Appendix D.2.

Table 6: RC counts and Dice overlap for the lexicon of 63 features in six families (945 units).

Model	Orth	LEACE	Dice
CSBrain	177	203	0.800
CBraMod	236	224	0.848
LaBraM	235	207	0.891
Total	648	634	0.849

Table 7: Individual-partner and joint erasure with F006 on CSBrain/Sleep. Values are macro-F1 drops in percentage points.

Partner	Alone	With F006
F011 (log θ/β ratio)	2.79	5.95
F015 (spectral centroid)	3.15	4.15
T007 (zero crossing)	3.13	5.51
C003 (LZ-style proxy)	2.54	7.30
R009 (γ coherence proxy)	5.02	6.20

D.1 Inputs and analysis rows

The procedure takes three fixed inputs. The first input is the five preprocessed EEG datasets, each split into train, validation, and test partitions. For MDD, Stress, TUSL, and Siena, each preprocessed sample is one analysis row, and the partition row counts are $(4891, 1041, 1211)$ on MDD, $(1343, 172, 192)$ on Stress, $(109, 86, 105)$ on TUSL, and $(29671, 7426, 14252)$ on Siena. ISRUC-Sleep is input to the foundation models as 30 s-epoch sequences, $(3559, 468, 435)$ per partition, but each sequence is unrolled into its constituent 30 s epochs for analysis, yielding $(71180, 9360, 8700)$ analysis rows; every downstream feature, representation, erasure, and metric on Sleep is indexed by epoch row. The second input is the foundation model checkpoints; backbone weights are frozen throughout, and predictions use the task-specific classifier stored with each checkpoint. The third input is the predefined lexicon of 72 features of Appendix A, with 379 expansion columns under the aggregation rule of Appendix D.3.

D.2 Splits and seeding

Roles of the three partitions are fixed at the start and never reassigned. The training partition fits all standardization parameters, ridge probes, orthogonal cross-covariance erasers, and surrogate classifiers. The validation partition selects probe regularization, the encoding peak layer, and supplies validation-side bootstrap diagnostics. The test partition supplies headline intervention metrics and confirmation. Test-confirmed labels subsequently determine the Closure feature set; thus Closure selection is test-informed, although classifier parameters are fit only on training rows. In the six-family implementation, stochastic procedures use a global seed (`4311`); several draws use deterministic offsets based on identifiers such as the task, model, feature, and operation.

D.3 Feature matrix construction

For each partition of each task, every analysis row is treated as a channel-by-time EEG epoch with sampling rate $f_s = 200$ Hz. The MDD, Stress, TUSL, and Siena epochs are interpreted up to a low-pass of 75 Hz; ISRUC-Sleep is interpreted up to 35 Hz, so the gamma-band quantities of Family F, TF, X, and R on Sleep cap at 35 Hz. Per-task signal scaling matches the convention of the preprocessing pipeline that produced each dataset; feature vectors and model activations are paired by the EEG epoch from which they were computed.

We compute the 72 registered features. For the subset of 63 features in six families, per-channel values contribute six summary statistics (mean, standard deviation, median, both quartiles, and maximum); pairwise coherence values contribute five (mean, standard deviation, median, upper quartile, and mean of the largest $\lceil 10\% \rceil$ pair values); PLI is averaged over channel pairs, and each scalar

Table 8: Target-resolution comparison on CSBrain/Sleep; macro-F1 drops in percentage points. Parentheses identify target dimensions; the per-patch erasure rank is capped at six.

Feature	Summary (6)	Channels (19)	Patches (570, rank 6)
T003	3.938	5.498	4.435
F010	3.626	3.601	4.467
X001	4.114	5.070	5.572

Table 9: Sleep probe test R^2 for ridge and a two-layer MLP.

Model	Feature	Ridge	MLP
CSBrain	T003	0.794	0.880
	F010	0.702	0.856
	X001	0.287	0.370
CBraMod	T003	0.765	0.855
	F010	0.774	0.774
	X001	0.257	0.223
LaBraM	T003	0.690	0.840
	F010	0.741	0.820
	X001	0.223	0.106

global descriptor contributes one column. These rules yield 328 columns. The aperiodic descriptors contribute 18 channel-summary columns, and the four-template microstate descriptors contribute 33 columns as listed in Table 2. The full registry therefore has 379 columns, with a fixed order across tasks.

Standardization parameters are estimated on the training partition alone. For each expansion column, we use median and interquartile range (IQR) robust scaling when $\mathrm{IQR} > 10^{-8}$, fall back to mean and standard deviation when $\mathrm{IQR} \leq 10^{-8}$, and force the scale to 1 when $\mathrm{std} \leq 10^{-8}$ or non-finite. Validation and test apply the parameters estimated on the training set without re-estimation. After standardization, every per-task feature matrix is row-aligned with the partition's analysis rows and column-aligned with the registry's 379 expansion columns. A quality-control pass records non-finite ratios, low-variance flags ($\mathrm{IQR} \leq 10^{-8}$ and $\mathrm{std} \leq 10^{-8}$ on training), and any deviation from the registry order; flagged columns are retained in the feature matrix, but they propagate to the failure record of Appendix B.

D.4 Layer-wise activations

For each task, model, and partition, the foundation model is evaluated in inference mode, with dropout disabled and parameters frozen. Inputs are fed in the same row order as the feature matrix; for Sleep, sequence-level outputs are unrolled back into epochs so the row order matches Appendix D.1. At each of the $L_m = 12$ encoder blocks, we record the block output and reduce all non-feature axes (channel, patch, token, sequence) by mean pooling, retaining the 200-dimensional hidden axis. Each layer activation is therefore an analysis-row-by-200 matrix, row-aligned with the feature matrix; the 12 layer matrices per cell form the input to the probe and erasure stages.

D.5 Encoding criterion

For each (d, m, q, l) quadruple, the ridge probe $P_{d,m,q,l}$ takes the layer activation from the training set as input and the corresponding column block of the standardized feature matrix as multi-output target. Rows with non-finite input or target are filtered out. Activations are standardized by the mean and standard deviation estimated on training data; targets keep the standardization of Appendix D.3. The probe is fit at four candidate regularization strengths $\lambda \in \{0.1, 1.0, 10.0, 100.0\}$, each scored by validation nonnegative R^2 (per-output-dimension R^2 clipped at zero, then averaged); the highest-scoring λ is selected as the probe at that quadruple.

Table 10: Stress Closure under linear and two-layer MLP readouts, using RC sets from the lexicon of 63 features in six families.

Model	Linear (%)	MLP (%)
CSBrain	68.0	70.0
CBraMod	54.7	71.0
LaBraM	45.1	93.7

Table 11: Linear recovery with six-family (63-feature) and eight-family (72-feature) lexicons. RC denotes the number of representation-causal features.

Model	Task	RC_{63}	RC_{72}	$Closure_{63}$ (%)	$Closure_{72}$ (%)
CSBrain	MDD	37	44	97.4	95.9
	Sleep	47	56	91.5	92.6
	Siena	42	45	68.8	68.4
	TUSL	45	49	52.7	62.2
	Stress	6	8	68.0	52.1
CBraMod	MDD	53	62	99.5	99.4
	Sleep	57	66	95.9	96.4
	Siena	32	39	87.4	79.4
	TUSL	47	51	68.5	78.5
	Stress	47	56	54.7	61.9
LaBraM	MDD	47	53	101.2	100.9
	Sleep	54	63	95.1	96.0
	Siena	63	72	87.2	85.6
	TUSL	42	43	77.0	79.8
	Stress	29	35	45.1	42.3

Each feature is also probed twice under matched controls. The shuffled-target probe uses the same activations and the same target column block but with the row pairing permuted on the training partition. The Gaussian control uses independent standard normal training targets of the same shape. Both control probes use the same regularization sweep and are evaluated against the original feature values on validation and test rows.

The peak layer is the layer with the highest validation probe score, selected before any test number is examined; the test probe score at this layer is reported as a held-out diagnostic. A feature is *encoded* on (d, m) if its validation probe score reaches 0.04 and exceeds each control score by at least 0.01. For a multidimensional feature, the largest nonnegative per-output R^2 must contribute at most 0.90 of the sum across outputs. The implementation also records the gap between the two best layers. Its stability flag accepts either a gap of at least 0.002 or a validation score of at least 0.04, so it imposes no additional restriction on encoded units. *Test-encoded* is recorded as a separate field with the same threshold and margin applied to the test partition; encoded is the gating field for erasure, so test-side numbers do not determine the encoding peak layer or eligibility for erasure. The later Closure feature set uses test-confirmed RC labels, as specified in Appendix D.2.

D.6 Orthogonal cross-covariance eraser construction and control interventions

For each encoded feature, we re-fit an orthogonal cross-covariance eraser at its peak layer using the activations from the training set. Activations and targets are centered on the training partition; we do not estimate or use the activation auto-covariance Σ_{hh}, so the eraser removes the Euclidean orthogonal projection onto the column span of the empirical cross-covariance Σ_{hz} (the orthogonal geometry, distinct from the covariance-whitened minimum-distortion LEACE operator of Belrose et al. [2023]; see Appendix E for the explicit relation). The cross-covariance Σ_{hz} is decomposed by singular value decomposition, and we retain the directions whose singular values exceed $\sigma > 10^{-4} \cdot \max(\sigma_{\max}, 1)$; if no direction meets this threshold, the top components up to the target column

Table 12: Closure baselines for the lexicon of 63 features in six families. B_0: F001–F005; all: full six-family lexicon; enc: encoded; rep: RC; fam: families with RC support; rand: Gaussian features with the same input dimension as the confirmed set. These headings abbreviate B_{all}, B_{enc}, B_{rep}, B_{fam}, and B_{rand}; entries are task scores, and FM is $M(f_m, d)$.

Model	Task	B_0	all	enc	rep	fam	rand	FM	Closure (%)
CSBrain	MDD	0.981	0.985	0.986	0.984	0.985	0.477	0.998	97.4
	Sleep	0.661	0.752	0.750	0.741	0.752	0.191	0.792	91.5
	Siena	0.881	0.843	0.840	0.817	0.843	0.512	0.955	68.8
	TUSL	0.693	0.630	0.642	0.618	0.630	0.268	0.932	52.7
	Stress	0.686	0.706	0.699	0.693	0.764	0.481	0.794	68.0
CBraMod	MDD	0.981	0.985	0.986	0.985	0.985	0.523	0.987	99.5
	Sleep	0.661	0.752	0.754	0.754	0.752	0.179	0.779	95.9
	Siena	0.881	0.843	0.840	0.864	0.843	0.488	0.918	87.4
	TUSL	0.693	0.630	0.637	0.597	0.630	0.231	0.766	68.5
	Stress	0.686	0.706	0.721	0.723	0.706	0.530	0.884	54.7
LaBraM	MDD	0.981	0.985	0.986	0.989	0.985	0.516	0.984	101.2
	Sleep	0.661	0.752	0.750	0.750	0.752	0.190	0.778	95.1
	Siena	0.881	0.843	0.840	0.840	0.843	0.504	0.889	87.2
	TUSL	0.693	0.630	0.643	0.630	0.630	0.398	0.700	77.0
	Stress	0.686	0.706	0.694	0.644	0.706	0.515	0.800	45.1

Table 13: Sensitivity to bootstrap budget: median positive-effect set Jaccard and Monte Carlo standard error across 15 cells.

Metric	$B = 128$	$B = 512$	$B = 2048$
Positive-effect set Jaccard	0.975	1.000	1.000
MCSE of mean drop	0.00069	0.00035	0.00017

count are retained so the eraser does not collapse to identity. The eraser is the orthogonal projector that removes the retained directions from the centered activation and adds the mean back. The edited model replaces the peak-layer activation by its erased counterpart and continues with the unchanged remainder of the network and classifier head; no model parameters are modified. Predictions are recomputed on validation and test; ROC-AUC is computed for the binary tasks (MDD, Stress, Siena), and macro-F1 is computed for the multiclass tasks (Sleep, TUSL).

For each feature we additionally fit three control erasers. The *random-subspace* eraser removes a random orthonormal subspace of dimension equal to the target expansion dimension p_q (which need not equal the retained rank of the real eraser), ignoring the target. The *shuffled-target* eraser permutes the row pairing of the target on the training set before re-running the same cross-covariance SVD pipeline. The *Gaussian-target* eraser replaces the target by a standard normal of matching column count and re-runs the pipeline. All three controls share the activation cache and the bootstrap row index of the real eraser, so paired contrasts are well-defined; only the random-subspace control fixes its dimension a priori, while the shuffled-target and Gaussian controls inherit whatever rank their own SVD threshold selects.

After erasure, a residual ridge probe is re-fit on the erased activations to predict the same feature. The residual is judged on the validation partition: it must satisfy $R_{\text{resid}} < \max(0.02,\ 0.35 \cdot \max(R_{\text{probe}},\ 0.04))$, where R_{probe} is the validation probe score at the peak layer. The residual probe score is recorded for each evaluated unit to expose cases where the cross-covariance eraser failed to remove the target's linear component; this diagnostic is recorded for the subset of 63 features in six families. The held-out test-supported representation-causal status of Appendix D.7 does not depend on the residual threshold.

D.7 Representation-causal criterion

For each encoded feature, the bootstrap of Appendix E produces paired replicates of Δ_{erase} on the held-out test partition. The same replicates are used for the contrasts of interest: Δ_{erase} against zero and against the random-subspace eraser. A feature qualifies as *representation-causal* on (d, m) when four conditions all hold: (i) the validation-side encoded flag is true, (ii) the lower endpoint of the 95% test-side bootstrap interval for Δ_{erase} is positive, (iii) the test p-value computed with add-one smoothing survives Benjamini–Hochberg FDR at $q < 0.05$ in the 72-feature analysis within the same (d, m) panel, and (iv) the real-feature drop exceeds the random subspace control of matching target dimension described in Appendix D.6, $\Delta_{\mathrm{erase}} - \Delta_{\mathrm{erase}}^{(\mathrm{rand})} > 0$. The shuffled-target and Gaussian-target null-target erasers and the residual ridge probe of Appendix D.6 are reported as diagnostics alongside the main status, but they do not gate it; validation-side erasure replicates and split-consistency flags are likewise saved as diagnostic fields.

D.8 Cross-task taxonomy

Once each unit has a final status, the cross-task taxonomy aggregates over the 15 (d, m) cells of each feature. For each task, we say the feature has *strong task-level support* when all three architectures (CSBrain, CBraMod, LaBraM) qualify as representation-causal on that task ($\mathrm{MS}(d, q)$ counts supporting models, so $\mathrm{MS}(d, q) = 3$). A feature with strong task-level support on at least two tasks is a *universal candidate*. A feature with strong task-level support on exactly one task is *task-specific*. A feature with no strong task-level support but at least one representation-causal unit is *model-specific*. A feature with no RC support but at least one encoded unit is *encoded-only*; the rest are *not-encoded*. We additionally record the task-specificity index $\mathrm{TSI} \in [0, 1]$, the share of the largest task-averaged positive Δ_{erase} in the total positive task-averaged effect across the five tasks (negative effects clipped at zero before averaging), but it is not used to split the categories. The category labels feed Figure 2 and the discussion in Section 4.3.

D.9 Family analysis and redundancy control

The family-level summary records, per task and model, the encoded rate, the RC rate, and the sum of positive drops within each family. Family sizes are $|T| = 10$, $|F| = 16$, $|\mathrm{TF}| = 11$, $|C| = 7$, $|X| = 5$, $|R| = 14$, $|A| = 3$, and $|M| = 6$. This sum includes positive Δ_{erase} over all evaluated features in the family, including encoded-only features; skipped entries contribute zero. Because family members can describe near-duplicate signal structure, we also report a sum of positive drops adjusted for redundancy: for each family, multidimensional expansions are first averaged into a single representative value per row on the training partition, the absolute Pearson correlation matrix between representatives is computed, and a deterministic greedy procedure takes the first remaining feature in registry order and groups it with remaining features whose correlation to that feature is at least 0.80, then repeats. Each cluster contributes only its highest-Δ_{erase} feature to the adjusted sum, which is computed over clusters. The adjusted family sums are shown in Figure 4c.

D.10 Closure feature blocks

The main analysis with 72 features fits B_{rep} on the confirmed set and B_{rand} on a standard-normal matrix with the same number of input columns; FM is the frozen foundation model's metric. The baseline comparison for the 63 features in six families includes B_0 (F001–F005, 30 columns), B_{all} (all 328 columns), B_{enc} (encoded features), and B_{fam} (all features in families with at least one RC feature). Table 12 reports the task score of each block and Closure for the confirmed features.

For each block, columns of the training set are standardized by mean and standard deviation, with non-finite entries imputed to zero and columns of standard deviation below 10^{-6} set to scale 1. The classifier is the multinomial logistic regression of Appendix E, fit on the training partition with class weights set to the reciprocals of class frequencies in the training set. Predictions on the test set are scored as $\mathrm{ROC\text{-}AUC}$ for the binary tasks (MDD, Stress, Siena) and macro-F1 for the multiclass tasks (Sleep, TUSL). The Closure ratio of Equation 5 compares the task scores of B_{rep} and B_{rand} with the foundation model's score FM; cells with a non-finite denominator or $|M(f_m, d) - M(B_{\mathrm{rand}}, d)| \leq 10^{-12}$ are flagged as undefined and not used as headline numbers.

D.11 Separate interpretation of analysis outputs

We report feature availability, validation encoding, test-supported intervention status, cross-task taxonomy, family aggregates, and surrogate recovery separately. The criteria and outputs are defined in Appendices D.3–D.10. The failure record retains features that do not meet the relevant criteria.

E Reproducibility details

Foundation models. CSBrain [Zhou et al., 2025], CBraMod [Wang et al., 2025], and LaBraM [Jiang et al., 2024] are loaded from their public checkpoints. We evaluate task-specific fine-tuned checkpoints with model parameters fixed. CSBrain and CBraMod expose $L_m = 12$ analyzable encoder blocks; LaBraM (`labram_base_patch200_200`) exposes 12 backbone blocks. For each block we collect a sample-level activation by mean-pooling all non-feature axes (channel, patch, token, sequence) and retaining the model's hidden axis, so every layer activation is aligned row-by-row with the feature matrix.

Probe and eraser estimation. Ridge probes are fit on activations from the training set with regularization λ selected on validation over the grid $\{0.1, 1.0, 10.0, 100.0\}$; the selected λ maximizes the validation nonnegative R^2 (per-output-dimension R^2 clipped at zero, then averaged). The eraser uses Σ_{hz} truncated to its left singular components above $\sigma > 10^{-4} \cdot \max(\sigma_{\max}, 1)$; if no component crosses the threshold, the top components up to the target expansion dimension are retained so the eraser does not collapse to identity. Sample mean μ_h and cross-covariance Σ_{hz} are computed on the training partition only and held fixed for all downstream stages; the activation auto-covariance Σ_{hh} is *not* estimated or used. In column-vector notation, the orthogonal choice is

$$E_{\mathrm{orth}}(h) = h - U_r U_r^\top (h - \mu_h). \tag{44}$$

For comparison, write $W = (\Sigma_{hh}^{1/2})^\dagger$ for whitening on the covariance support and let $\widetilde{U}_r$ span the retained left singular space of $W\Sigma_{hz}$. The unregularized covariance-whitened LEACE form is

$$E_{\mathrm{LEACE}}(h) = h - W^\dagger \widetilde{U}_r \widetilde{U}_r^\top W(h - \mu_h). \tag{45}$$

The orthogonal operator corresponds to the projection geometry offered by the reference implementation's `orth` option; our rank threshold and fallback are specified above. With the complete empirical cross-covariance range retained, orthogonal projection makes that cross-covariance zero. Numerical truncation may leave residual correlations, and linear erasure does not guarantee nonlinear removal. The two geometries coincide for identity activation covariance with matched rank choices; in general they induce different edits. The surrogate classifier is a class-weighted multinomial logistic regression with fixed L_2 penalty 10^{-4}, fit by limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization at learning rate 0.8 for at most 120 iterations; class weights are set to inverse class frequencies on the training partition.

Bootstrap and FDR. Confidence intervals and paired tests resample analysis rows, with shared indices across the real-feature and control effects. Add-one smoothing gives $\hat{p} = (1 + |\{b : \delta_b \leq 0\}|)/(1 + B_{\mathrm{eff}})$, where B_{eff} is the valid-replicate count. BH-FDR applies within each model–task panel (72 features for the eight-family lexicon and 63 for the six-family subset). The six-family implementation adjusts the finite test p-values within each panel. Appendix C.5 compares resampling budgets of 128, 512, and 2,048.

Compute. The experiments with the lexicon of 63 features in six families were estimated to require approximately one to two GPU-days on a single NVIDIA A100 80GB.

Code and data. The five EEG datasets are available through their respective access procedures. The public BrainPEC implementation covers the 63 features in the six-family lexicon, including feature extraction, probing, erasure, and Closure (`https://github.com/Kian-Chen/BrainPEC`). To extend the lexicon, a feature extractor must return values aligned to the EEG samples in each split. To evaluate another model, the analysis requires access to its frozen activations and a way to intervene at a specified layer.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] Justification: The abstract and Sections 4–5 report the same 965 encoded and 742 representation-causal units and mean Closure of 79.4%. Interpretations are bounded to the tested lexicon, models, tasks, and intervention protocol.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] Justification: Section 5 discusses linear probes, shared feature information, test-informed feature selection, and uncertainty from repeated observations within patients.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A] Justification: No new theorem is claimed. The paper applies established probing, concept-erasure, and classification tools; Appendix E distinguishes orthogonal and covariance-whitened erasure.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No] Justification: Appendices A and D describe the feature definitions, data splits, and evaluation criteria.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No] Justification: The datasets have public access procedures, and the code for the 63 features in six families is available at `https://github.com/Kian-Chen/BrainPEC`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No] Justification: Appendices D and E describe probe selection, erasure, and classifier fitting. The backbone checkpoints are fixed during these analyses.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] Justification: The representation-causal criterion uses a positive paired-bootstrap lower confidence bound, BH-adjusted $q < 0.05$, and an effect exceeding the random-subspace control. Appendix E specifies analysis-row resampling; figure aggregates are descriptive, not confidence intervals.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No] Justification: Appendix E reports one NVIDIA A100 80GB and one to two GPU-days for the experiments with the six-family lexicon.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] Justification: The study analyzes existing EEG datasets and pretrained models, without new participant recruitment. The limitations and risks of interpreting model interventions clinically are discussed in Section 5.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] Justification: Section 5 describes the potential benefit of distinguishing encoding from model reliance and the risk of treating intervention evidence as biological causality or clinical validation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A] Justification: No new generative model or raw participant dataset is introduced. Any release of derived artifacts must respect the underlying dataset access restrictions.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No] Justification: Section 4.1 identifies the datasets and model sources, and Appendix E describes data access.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No] Justification: Appendix A documents the feature registry.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A] Justification: There is no new crowdsourcing or human-subject data collection; the study uses existing EEG datasets.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A] Justification: No new recruitment, data collection, or participant recontact is conducted. Source datasets are used under their original collection protocols and access conditions.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A] Justification: LLMs are not an original or non-standard component of the reported methodology. Writing, editing, and formatting assistance does not require a declaration under the conference policy.