FROM REPRESENTATION TO COMPETENCE: WHAT IS MISSING IN EEG FOUNDATION MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) foundation models are pretrained for broad transfer, yet their downstream performance remains uneven. Adding physiological descriptors improves prediction from frozen classifier features, and the same descriptor families are recoverable across the evaluated backbones. How much additional task performance can be recovered from these representations has remained unclear. At several internal locations, readout of the descriptor families is stronger than at the classifier input. A low rank residual adapter is therefore learned within the backbone from task labels, with the pretrained weights and classifier kept fixed. The preferred update depth varies across models and tasks, and task validation yields performance close to that of the better fixed placement. Across four backbones and four tasks, adaptation improves performance in most settings with only a few thousand trainable parameters. At the same budget, internal updates outperform updates at the structured final output in most paired comparisons. These results support placing a small residual adapter inside the backbone at a depth selected by task validation.

1 INTRODUCTION

Electroencephalography (EEG) foundation models are pretrained on heterogeneous recordings to support broad transfer (Kuruppu et al., 2026). However, their downstream performance varies across tasks and depends strongly on how the pretrained model is adapted (Yang et al., 2026b). With the pretrained backbone frozen, a reference classifier is fitted on the classifier input features (POST). Internal representations may support stronger predictions than this classifier delivers. We refer to the gap between task-relevant structure accessible within the pretrained model and the task performance obtained from its frozen output as the *representation-to-competence gap*.

Physiological knowledge can improve EEG representation learning (Wagh et al., 2021), and corresponding audits also have recovered spectral and spatial structure from pretrained hidden states (Tang et al., 2026; Wang et al., 2026; Kommineni et al., 2026). For downstream transfer, the central issue is *how much structure with predictive value beyond the frozen output remains accessible inside the model*. Physiological descriptors allow this issue to be examined through both their incremental predictive value and their readout from internal representations. This motivates examining whether adaptation is more effective before the final output.

This question concerns update location, which determines how much pretrained computation follows an update. An internal update is processed by the remaining backbone before reaching the classifier, whereas a final-output update acts after these transformations. We learn a small low rank residual from task labels while keeping the pretrained weights and reference classifier fixed. Internal and final-output residuals have the same trainable parameter budget within each backbone. Residuals at candidate internal depths are trained independently, and task validation selects between them.

Across four backbones (Jiang et al., 2024; Wang et al., 2025a; Zhou et al., 2025; Xiao et al., 2025) and four clinical EEG tasks (Van Dijk et al., 2022; Miltiadous et al., 2023; Kim et al., 2023; Shah et al., 2018), this study makes three contributions:

- Physiological descriptors reveal predictive information beyond frozen output features. At least one family adds predictive value in every combination of model and task, and some descriptors improve performance on two datasets across all four backbones. These findings

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

reveal a recurring phenomenon in which measured physiological information can improve prediction beyond frozen outputs alone.

- Physiological structure is broadly accessible within pretrained EEG models. Readout is positive in most combinations of model, task, and descriptor family, and several families show stronger readout at internal sites than at the classifier input. Paired readout after adaptation also changes differently across representation sites and reader types, suggesting a direction for studying how physiological structure evolves during adaptation.
- A residual with only a few thousand trainable parameters improves task performance through internal adaptation. Under the same parameter budget, internal updates outperform updates at the final output in nearly every matched comparison. The preferred internal depth varies across models and tasks, and task validation remains close to the better fixed placement. These results show that further exploiting the representations within pretrained EEG backbones can translate into greater model competence.

2 Related Work

EEG representations and transfer. Deep neural networks support EEG decoding (Schirrmeister et al., 2017; Lawhern et al., 2018), and these models are usually trained for a single task on a single labeled dataset. Their transfer is complicated by heterogeneity in clinical recordings (Obeid & Picone, 2016) and by within-subject, between-subject, and cross-dataset variability (Saha & Baumert, 2020; Xu et al., 2020). Foundation models instead pretrain on large collections of unlabeled recordings and learn representations for reuse across tasks, through different architectures and pretraining objectives (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2024; 2025a; Zhou et al., 2025; Xiao et al., 2025). These objectives range from contrastive learning in BERT-inspired Neural Data Representations (BENDR) (Kostas et al., 2021) to masked token prediction in the Large Brain Model (LaBraM) (Jiang et al., 2024) and signal reconstruction in the criss-cross brain foundation model (CBraMod) (Wang et al., 2025a). Beyond pretraining, Yang et al. (2026b) show that downstream performance also depends on how the pretrained model is adapted. The value of pretraining therefore depends on how its representations are used as well as how they are learned.

Physiological readout and representation analysis. Self-supervised objectives informed by EEG physiology can improve downstream classification (Wagh et al., 2021). Physiological audits examine how such structure is represented after pretraining. Tang et al. (2026) distinguish readable physiological structure from structure used in prediction. Their closure analysis measures how much of a model's predictive advantage over a matched random-feature baseline can be reproduced from those features. Wang et al. (2026) combine performance evaluation with neurophysiological input tests, while Kommineni et al. (2026) identify spectral biases associated with reconstruction pretraining. Sparse autoencoders and concept steering further relate internal features to physiological and clinical concepts (Lehn-Schiøler et al., 2026). The interpretation of these readouts depends on probe capacity and appropriate controls (Belinkov, 2022; Hewitt & Liang, 2019).

Adapters and update placement. Adapters and low rank updates enable adaptation with few trainable parameters (Houlsby et al., 2019; Hu et al., 2022). Kumar et al. (2022) show that fine-tuning can distort pretrained features and yield worse out-of-distribution performance than linear probing. The best layers to update also depend on the type of distribution shift (Lee et al., 2023). In language models, Yang et al. (2026a) show that injection depth changes how subsequent layers transform learned task vectors. In EEG, NeuroTTT evaluates low rank adaptation (LoRA) across transformer blocks as a supervised adaptation baseline (Wang et al., 2025b), while Ma et al. (2026) use prototypes to condition adapters in later layers.

3 Method

Four representative EEG foundation models as LaBraM, CBraMod, the cross-scale spatiotemporal brain foundation model (CSBrain), and BrainOmni are considered in this study. We first measure the incremental task value of EEG descriptors beyond frozen POST features (Figure 1(a)), along with their readout across representation sites (Figure 1(b)). We then compare update locations under

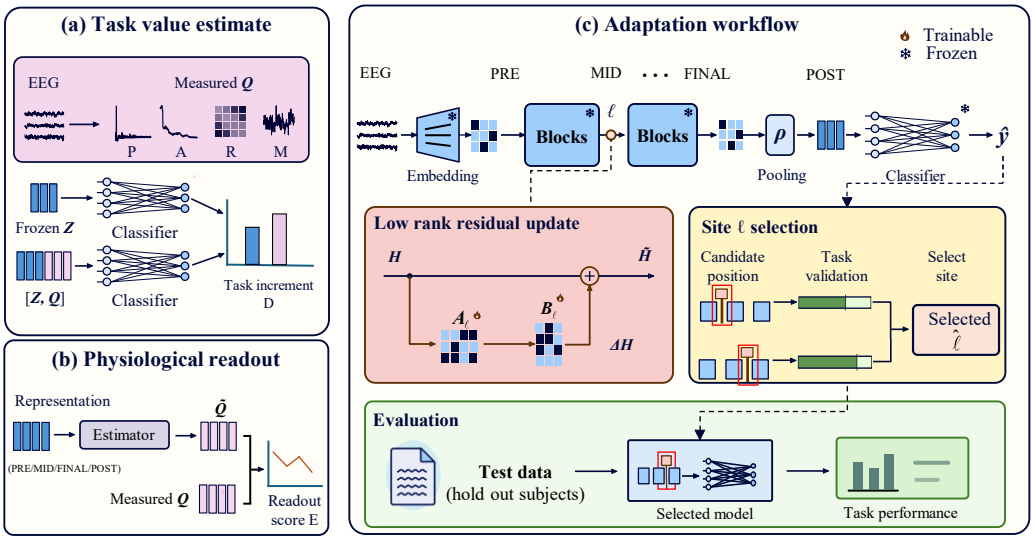

Figure 1: **From physiological measurement to internal adaptation of a frozen EEG backbone.** (a) Incremental task value D compares a classifier fitted on frozen POST feature matrix $\boldsymbol{Z}$ with one fitted on $[\boldsymbol{Z}, \boldsymbol{Q}]$, where $\boldsymbol{Q}$ contains the measured descriptors of one family and brackets denote concatenation along the feature dimension. (b) Physiological readout E measures how well the same descriptors can be predicted from representations at PRE, MID, FINAL, and POST. (c) A low rank residual is inserted at site ℓ while the pretrained backbone and reference classifier remain fixed. In the adaptive depth comparison, PRE and MID candidates are trained independently and selected by task validation, which is independent of test subjects.

a fixed parameter budget and use task validation to select the depth of internal updates. The low rank internal residual is trained with classification loss and a displacement penalty, while the pretrained weights and reference classifier remain frozen (Figure 1(c)). Appendix A.1 lists the detail of model–task settings covered by each experiment.

3.1 Representations and Physiological Measurements

For backbone m, the pretrained network f_θ maps an EEG window $\boldsymbol{X}_i$ to the structured final output (FINAL) $\boldsymbol{S}_i = f_\theta(\boldsymbol{X}_i)$, where θ denotes the pretrained parameters and i indexes windows. Depending on the backbone, $\boldsymbol{S}_i$ retains channel or token dimensions in addition to the embedding dimension. A backbone-specific reduction ρ_m maps this output to the feature vector at POST as $\boldsymbol{z}_i = \rho_m(\boldsymbol{S}_i) \in \mathbb{R}^{d_{\mathrm{POST}}}$, which is supplied to the classifier. Here d_{POST} is the classifier input dimension. Individual feature and descriptor vectors are written as column vectors; sample matrices collect their transposes as rows. The reference logistic classifier g_{ϕ_0} is fitted on the frozen training features and their task labels, with ϕ_0 denoting its fitted parameters. It produces three class logits $\boldsymbol{\eta}_i = g_{\phi_0}(\boldsymbol{z}_i) \in \mathbb{R}^3$. The corresponding class probabilities are $\mathrm{softmax}(\boldsymbol{\eta}_i)$, and the predicted class is $\widehat{y}_i = \arg\max_{c\in\{0,1,2\}} \eta_{ic}$, where η_{ic} is the logit for class c. Details are given in Appendix A.2.

We compute four physiological descriptor families directly from the same EEG windows used to obtain the model representations, using an extraction procedure that does not depend on the backbone. The families summarize spectral residuals (P), smoothed spectral structure (A), channel correlation geometry (the relational family, R), and multiscale temporal and time–frequency statistics (the morphology family, M). P and A are smoothing-based proxies for periodic and aperiodic structure. For window $\boldsymbol{X}_i$, let $\boldsymbol{q}_i^a \in \mathbb{R}^{32}$ denote the standardized descriptor vector of family $a \in \{\mathrm{P}, \mathrm{A}, \mathrm{R}, \mathrm{M}\}$. These vectors are obtained using principal component analysis (PCA) and standardization fitted on the training data, as detailed in Appendix A.4.

Incremental task value. For backbone model m, task t, and descriptor family a, let $\boldsymbol{Z}_{m,t}$ denote the matrix of POST features and $\boldsymbol{Q}_{t,a}$ the corresponding matrix of measured descriptors. Within each data partition, their ith rows are $\boldsymbol{z}_i^\top$ and $(\boldsymbol{q}_i^a)^\top$, respectively, and refer to the same window. Fold and partition indices are omitted for clarity. We fit a reference logistic classifier on the training rows of $\boldsymbol{Z}_{m,t}$ and a separate logistic classifier on the training rows of $[\boldsymbol{Z}_{m,t}, \boldsymbol{Q}_{t,a}]$, with concatenation along the feature dimension, so family a contributes 32 additional input coordinates. Both classifiers use the same data splits and scoring procedure, and the difference between their scores defines incremental task value:

$$D_{m,t,a,k} = \mathcal{M}_{t,k}([\boldsymbol{Z}_{m,t}, \boldsymbol{Q}_{t,a}]) - \mathcal{M}_{t,k}(\boldsymbol{Z}_{m,t}). \tag{1}$$

Here k indexes macro-averaged F1 (Macro-F1) or balanced accuracy (BAcc), and $\mathcal{M}_{t,k}$ denotes the score, expressed as a percentage, of a classifier fitted on the features in its argument and evaluated on the held-out windows of task t. Positive $D_{m,t,a,k}$ means that supplying family a raises that score above the score obtained from POST features alone. The main results report held-out values, with experimental coverage specified in Appendix A.1. Additional comparisons on development folds appear in Appendix B.1.

Physiological readout. To measure how well the descriptor families can be predicted from the pretrained representations, we evaluate readout at several sites along the backbone. The sampled sites are an early embedding (PRE), a middle layer (MID), the structured final output (FINAL), and the classifier input (POST). At a backbone site $\ell \in \{\mathrm{PRE}, \mathrm{MID}, \mathrm{FINAL}\}$, the state is $\boldsymbol{H}_i^{(\ell)} = f_\theta^{(\le \ell)}(\boldsymbol{X}_i)$, where $f_\theta^{(\le \ell)}$ is the backbone up to that site and $f_\theta^{(>\ell)}$ is the remaining backbone. At FINAL, $\boldsymbol{H}_i^{(\mathrm{FINAL})} = \boldsymbol{S}_i$; POST uses $\boldsymbol{z}_i$ after the reduction ρ_m. Appendix A.3 gives the tensors and reductions for each backbone. Ridge and multilayer perceptron (MLP) readers use averaged channel or token features at PRE, MID, and FINAL, and the classifier features directly at POST.

Each reader predicts the 32 standardized coordinates of one descriptor family. For each seed, we compare its predictions with a constant baseline that returns the descriptor mean estimated from the training data of each fold, without using the model representation. Readout E measures the proportional reduction in held-out squared prediction error relative to this baseline. For a fixed model, task, family a, site, reader, and seed, we suppress these indices. For held-out window i in fold f, let $\boldsymbol{q}_{f,i}$ denote the corresponding vector $\boldsymbol{q}_i^a$ computed using that fold's descriptor transforms, $\widehat{\boldsymbol{q}}_{f,i}$ the reader prediction and $\bar{\boldsymbol{q}}_f^{\mathrm{train}}$ the training mean of those targets. The readout score is

$$E = 1 - \frac{\sum_f \sum_i \|\boldsymbol{q}_{f,i} - \widehat{\boldsymbol{q}}_{f,i}\|_2^2}{\sum_f \sum_i \left\|\boldsymbol{q}_{f,i} - \bar{\boldsymbol{q}}_f^{\mathrm{train}}\right\|_2^2}. \tag{2}$$

The sums run over the held-out windows in each evaluation fold, and the squared Euclidean norm sums all 32 descriptor coordinates. For CAUEEG, the evaluation consists of the single official test partition. We average MLP E across three initialization seeds. Ridge fitting is deterministic. Positive E indicates lower squared prediction error than the training-mean baseline, with $E = 1$ for perfect prediction. $E = 0$ matches the baseline, and negative E indicates higher error. D measures the incremental task value of a descriptor family, whereas E measures its accessibility from a pretrained representation. A family is counted as linearly encoded when ridge readout is positive at any sampled site, and as having nonlinear readout only when MLP readout is positive while ridge readout is nonpositive at every site. Appendix A.5 gives the full procedure.

3.2 Internal adaptation

To test whether internal representations can support stronger task prediction, A small low rank residual is applied at a single site while keeping the pretrained weights and the previously fitted reference classifier fixed. For an embedding vector $\boldsymbol{h} \in \mathbb{R}^{d_\ell}$ at site ℓ, the residual update is

$$\widetilde{\boldsymbol{h}} = \boldsymbol{h} + \boldsymbol{B}_\ell \boldsymbol{A}_\ell \boldsymbol{h}, \tag{3}$$

where d_ℓ is the embedding dimension at site ℓ, $\boldsymbol{A}_\ell \in \mathbb{R}^{r \times d_\ell}$, and $\boldsymbol{B}_\ell \in \mathbb{R}^{d_\ell \times r}$, with $r = 8$. As $\boldsymbol{B}_\ell$ is initialized to zero, the initial adapted map matches the reference map.

The same two matrices are shared across all channel or token positions at site ℓ. Applying the residual to $\boldsymbol{H}_i^{(\ell)}$ gives the adapted state $\widetilde{\boldsymbol{H}}_i^{(\ell)}$. This state then passes through the remaining backbone and the reduction, giving the adapted logits

$$\boldsymbol{\eta}_i^{\mathrm{int}} = g_{\phi_0}\left(\rho_m\left(f_\theta^{(>\ell)}\left(\widetilde{\boldsymbol{H}}_i^{(\ell)}\right)\right)\right). \tag{4}$$

The softmax and prediction rules of the reference classifier apply unchanged. Gradients pass through the fixed downstream computation, while only $\boldsymbol{A}_\ell$ and $\boldsymbol{B}_\ell$ are updated. We optimize the unweighted mean task cross-entropy (CE) $\mathcal{L}_{\mathrm{CE}}$ with a penalty on relative hidden state displacement:

$$\mathcal{L}_{\mathrm{internal}} = \mathcal{L}_{\mathrm{CE}} + \left[\max\left(0, \frac{\|\Delta\boldsymbol{H}\|_{\mathrm{F}}}{\|\boldsymbol{H}\|_{\mathrm{F}} + 10^{-12}} - 0.1\right)\right]^2. \tag{5}$$

Here $\boldsymbol{H}$ and $\widetilde{\boldsymbol{H}}$ stack the reference and adapted hidden states of the batch at site ℓ, whose index is omitted, and $\Delta\boldsymbol{H} = \widetilde{\boldsymbol{H}} - \boldsymbol{H}$ is their displacement under adaptation. The Frobenius norms cover all entries of these tensors, and the penalty activates when the relative displacement exceeds 0.1.

Location and depth. To determine where a fixed adaptation budget is most effective, we compare residual updates at different representation sites. Each setting includes the reference classifier, classifier adaptation, a residual at FINAL before the reduction ρ_m, and a residual at an assigned representation site. Classifier adaptation updates the fitted affine weights and intercept. The internal and FINAL residuals have the same number of trainable parameters within each backbone. The assigned site is fixed before adaptation for each model–task setting, as specified in Appendix A.7. To examine depth within the backbone, we separately compare fixed PRE, fixed MID, and task validation selection between independently trained PRE and MID candidates. Training and epoch selection follow Appendix A.6.

Function class of FINAL adaptation. To characterize how update location changes the downstream computation, we examine the transformations that follow the residual. For CBraMod and CSBrain, the reduction from FINAL to POST is linear, so a residual at FINAL produces logits that remain affine functions of the reference POST features. These predictors therefore lie within the affine function class available to classifier adaptation. An internal residual is followed by the remaining backbone layers, which transform the adapted hidden state before it reaches POST. The two placements therefore use the same residual parameter budget while entering different downstream computations. Appendix A.8 gives the matrix derivation and describes the input-dependent reductions of LaBraM and BrainOmni.

3.3 Paired representation analysis

To examine how internal adaptation changes physiological readout along the remaining backbone, we compare CBraMod representations before and after adaptation on TD-BRAIN and ADFTD. PRE and MID candidates are trained independently with Equation 5, and task validation selects the depth within each fold. Appendix A.7 gives the complete selection rule. We track relational descriptors on TD-BRAIN and smoothed spectral structure on ADFTD. Ridge and MLP readers are refitted separately on the reference and adapted representations at two locations: immediately after the adapter, which we call the proximal site, and at POST. The paired measurements therefore compare readout at the adapted representation with readout after processing by the remaining backbone. Appendix E.1 gives the paired fitting details and additional measurements.

4 Experiments

4.1 Experimental setup

We evaluate four backbones on three-class tasks from TD-BRAIN (Van Dijk et al., 2022), ADFTD (Miltiadous et al., 2023), CAUEEG (Kim et al., 2023), and the Temple University Hospital EEG Seizure Corpus (TUSZ) (Shah et al., 2018), concerning psychiatric and memory complaints, dementia diagnoses, cognitive status, and seizure types, respectively. These tasks exhibit different pathological biases. On TD-BRAIN, ADFTD, and TUSZ, we use five-fold cross-validation grouped

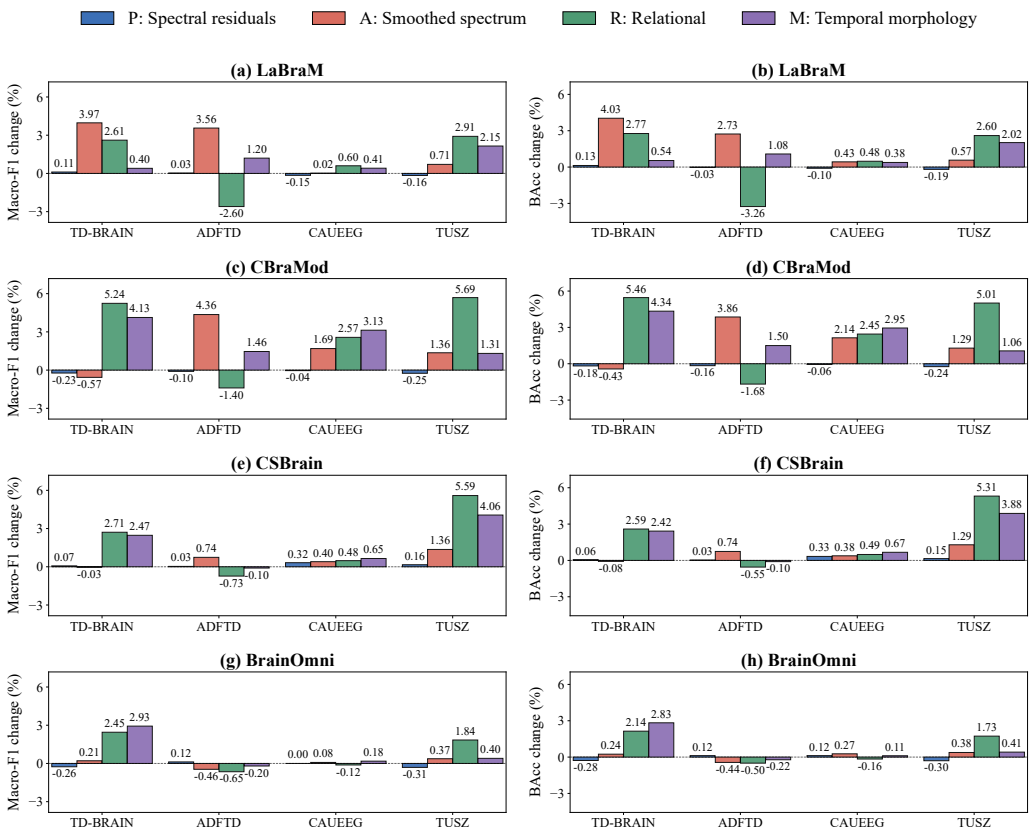

Figure 2: **Measured physiological descriptors add task value beyond frozen POST features.** Rows (a,b), (c,d), (e,f), and (g,h) show LaBraM, CBraMod, CSBrain, and BrainOmni across the four tasks. Bars show Macro-F1 (left) and balanced accuracy (right) differences (%) between descriptor-augmented and POST-only classifiers.

by subject, while CAUEEG retains its official partitions. We compute Macro-F1 and balanced accuracy from each seed's pooled held-out windows, then average across three seeds. Appendix A.2 gives dataset processing, class counts, reference fitting, and hardware.

4.2 Measured descriptors provide incremental task value

We compare logistic classifiers fitted on frozen POST features with classifiers augmented by one measured descriptor family. Each augmented classifier is compared with the same POST-only reference within a model–task setting. Relational descriptors improve both held-out metrics on TD-BRAIN and TUSZ for all four backbones (Figure 2). For CBraMod on TD-BRAIN, the gains are 5.24% in Macro-F1 and 5.46% in balanced accuracy. On TUSZ, the same family improves CSBrain by 5.59%/5.31% despite its different classifier reduction.

The descriptor families that improve prediction vary across backbones and tasks with different physiological emphasis. Smoothed spectral descriptors improve both metrics on ADFTD for LaBraM, CBraMod, and CSBrain. CBraMod gains 4.36%/3.86% on ADFTD, whereas the same family changes its TD-BRAIN scores by $-0.57\%/-0.43\%$. LaBraM gains 3.97%/4.03% on TD-BRAIN and 3.56%/2.73% on ADFTD. Temporal morphology improves both metrics in 14 of 16 settings, with mean gains of 1.54%/1.49%. Its exceptions are CSBrain and BrainOmni on ADFTD, at $-0.10\%/-0.10\%$ and $-0.20\%/-0.22\%$. Across the four families, all 16 model–task settings have at least one positive difference on both metrics. Matched noise changes only $-0.04\%/-0.02\%$ on average, while measured descriptors exceed noise on both metrics in 45 of 64 combinations (Ap-

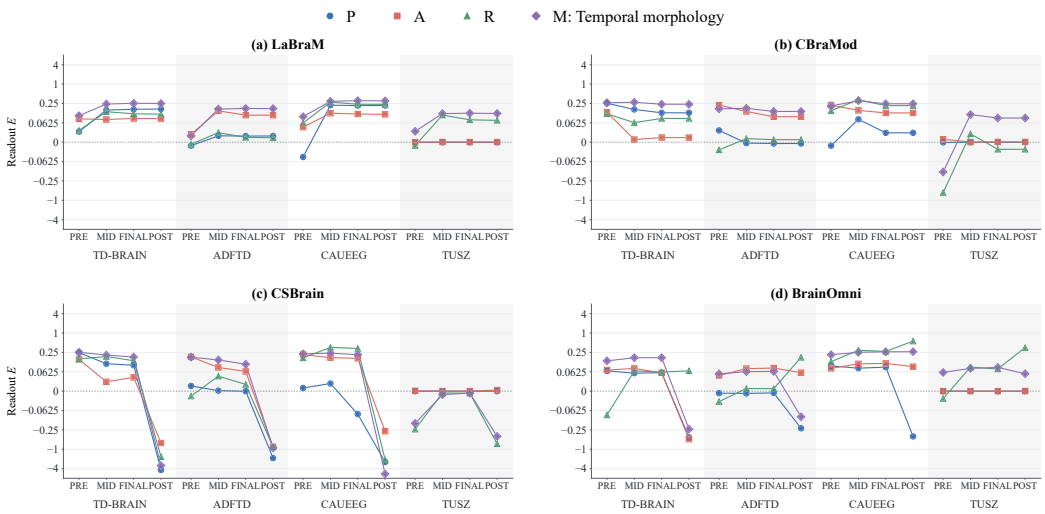

Figure 3: **Linear physiological readout varies across representation sites.** Panels (a)–(d) correspond to LaBraM, CBraMod, CSBrain, and BrainOmni. Within each task, colored curves show ridge readout E for P, A, R, and M across PRE, MID, FINAL, and POST.

pendix B.2). Measured descriptors therefore supply task-relevant structure that is missing from the frozen POST features of EEG foundation models, and adding the corresponding family recovers task performance.

4.3 Physiological structure across the pretrained backbone

The descriptors that add task value D at POST (Section 4.2) are next traced through the pretrained backbone with readout E. Ridge and MLP readout are compared at PRE, MID, FINAL, and POST for all four descriptor families. Ridge and MLP readers recover descriptors in 62 of 64 model–task–family combinations. Ridge readout is positive in 57 combinations, and MLP adds five cases with no positive ridge readout. The two combinations without positive readout are R and M in CSBrain on TUSZ. Both have positive incremental task value under augmentation, so descriptor task value can remain even when these readers do not recover the descriptor from the sampled representations. Figure 3 shows the ridge profiles across all four sampled sites; the MLP profiles and full classification of readable families appear in Appendix C.

The families with incremental task value can be followed across the backbone, and access to them varies with the family and the task. In CBraMod on ADFTD, smoothed spectral readout falls from PRE to POST, at 0.2229/0.1416/0.0969/0.0969 across PRE/MID/FINAL/POST (Figure 3). Although these descriptors remain readable at POST, adding their measured values to POST features still improves Macro-F1 and balanced accuracy by 4.36% and 3.86%, respectively (Section 4.2). For CBraMod on CAUEEG, relational readout instead peaks at MID. For CSBrain, several families have negative POST ridge scores while readout from the averaged channel or token features at PRE, MID, or FINAL is positive, so the reduction ρ_m and the reader's input both shape the observed profile.

Combined with the incremental task value, these profiles locate the representation-to-competence gap. Descriptors that add task value are readable in nearly every combination, and several families are read out more strongly at internal sites than at the classifier input, as in CBraMod on ADFTD and CAUEEG. Predictions from frozen POST features therefore make incomplete use of this structure.

4.4 Adaptation location and depth

The internal residual at the assigned site (Section 3.2) is compared with a FINAL residual under the same parameter budget and with classifier adaptation. Both residuals are trained with task cross-entropy and the displacement penalty in Equation 5, while the pretrained weights and the previously fitted reference classifier remain fixed.

Table 1: Classification performance by update location across four backbones and four tasks. Scores are pooled-window Macro-F1 (F1) and balanced accuracy (BAcc), averaged across three seeds. Assigned sites are MID except the two † FINAL rows, which coincide with final-output adaptation and are excluded from internal-versus-FINAL comparisons. Table 13 reports seed standard deviations.

						Equal-budget residuals			
		Reference (%)		Classifier (%)		FINAL (%)		Assigned site (%)	
Model	Task	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc
LaBraM	TD-BRAIN	63.54	62.20	63.65	62.30	63.42	62.16	**64.03**	**62.93**
LaBraM	ADFTD	45.43	46.44	45.38	46.42	46.09	46.89	**47.70**	**48.01**
LaBraM	CAUEEG	51.50	52.67	**51.89**	**53.28**	51.33	52.35	51.37	52.45
LaBraM	TUSZ	44.58	44.57	44.45	44.47	45.06	45.15	**45.49**	**45.77**
CBraMod	TD-BRAIN	60.00	58.65	60.01	58.64	60.01	58.70	**60.61**	**59.47**
CBraMod	ADFTD	44.05	44.77	44.06	44.76	44.02	44.74	**47.51**	**47.67**
CBraMod	CAUEEG	49.69	51.13	49.10	50.65	50.11	51.40	**52.24**	**53.50**
CBraMod	TUSZ	42.43	42.92	42.44	42.95	42.69	43.14	**43.22**	**43.48**
CSBrain	TD-BRAIN†	52.82	52.75	52.83	**52.78**	**52.85**	52.77	**52.85**	52.77
CSBrain	ADFTD	48.79	48.71	48.62	48.54	48.73	48.65	**48.93**	**48.85**
CSBrain	CAUEEG†	43.97	44.11	43.48	43.57	**44.01**	**44.17**	**44.01**	**44.17**
CSBrain	TUSZ	37.42	37.50	**38.00**	**38.08**	34.94	36.73	36.41	37.19
BrainOmni	TD-BRAIN	55.80	55.70	**56.24**	**56.11**	56.06	55.80	56.06	55.88
BrainOmni	ADFTD	52.68	52.56	52.87	52.77	52.73	52.64	**53.00**	**52.87**
BrainOmni	CAUEEG	50.15	50.88	49.48	50.43	50.35	50.91	**50.88**	**51.58**
BrainOmni	TUSZ	38.90	38.93	**39.88**	**39.90**	39.03	39.05	39.27	39.31

Among the 14 settings assigned to an internal site, internal updates outperform FINAL on both metrics in 13 settings. In BrainOmni on TD-BRAIN, Macro-F1 is tied at the reported precision and balanced accuracy is higher (Table 1). Internal updates also outperform classifier adaptation on both metrics in 10 of these 14 settings. This comparison highlights the role of the computation after the adapter. For CBraMod and CSBrain, FINAL adaptation keeps the logits affine in the reference POST features, within the function class already available to classifier adaptation. An internal update instead changes the states processed by the remaining backbone layers (Appendix A.8). Those layers retain their pretrained weights but receive inputs adjusted through task supervision. Internal updates improve both metrics over the reference in 12 of the 14 settings; the exceptions are LaBraM on CAUEEG and CSBrain on TUSZ. Because the reference classifier remains fixed, these gains come from changes in the features it receives.

Choosing the internal depth. The preferred internal depth varies across tasks and backbones (Table 2). CBraMod favors PRE on TD-BRAIN and MID on ADFTD, while LaBraM shows the reverse ordering. Earlier placement is therefore not consistently better. Selection by task validation matches or exceeds both fixed policies on TD-BRAIN in both backbones and stays close to the better fixed policy on ADFTD. Across the four settings, its largest shortfall from the better fixed policy is smaller than that of always choosing PRE or always choosing MID. Appendix D.1 gives the fold choices. Together, the location and depth results favor placing the adapter inside the backbone and choosing its depth by task validation. When adapting to a new dataset, the preferred internal depth is not known in advance. Validation provides a practical way to choose between PRE and MID for downstream tasks, rather than assuming that one fixed depth will work best across settings.

4.5 Task gains and paired physiological readout

To understand how the internal adapter changes representations along the remaining backbone, we compare physiological readout before and after adaptation at the adapter output and at POST. The adapter is trained with task cross-entropy and the displacement penalty, so this analysis follows physiological changes arising during task adaptation. We examine relational descriptors on TD-BRAIN and smoothed spectral descriptors on ADFTD, both with positive incremental task value D and readout E in the pretrained comparison.

Table 2: Classification performance by internal depth in the independent depth comparison. Scores are pooled-window Macro-F1 (F1) and balanced accuracy (BAcc), averaged across three seeds. Task validation selects between independently trained PRE and MID candidates within each fold.

		Fixed PRE (%)		Fixed MID (%)		Validation selection (%)	
Model	Task	F1	BAcc	F1	BAcc	F1	BAcc
CBraMod	TD-BRAIN	**62.46**	**61.62**	60.61	59.47	**62.46**	**61.62**
CBraMod	ADFTD	45.90	46.60	**47.51**	**47.67**	47.07	47.40
LaBraM	TD-BRAIN	63.78	62.51	64.03	62.93	**64.08**	**62.95**
LaBraM	ADFTD	**47.86**	**48.34**	47.70	48.01	47.70	48.01

Table 3: Task gains and paired CBraMod readout before and after task-only adaptation. Proximal E before $\rightarrow$ after and POST ΔE are dimensionless seed means, with $\Delta E = E_{\text{adapted}} - E_{\text{reference}}$. Readers are fitted in each state at each site. TD-BRAIN uses R and ADFTD uses A; the ADFTD rows summarize different evaluation folds.

	Task gain (%)		Proximal E		POST ΔE	
Task / depth	F1	BAcc	Ridge	MLP	Ridge	MLP
TD-BRAIN PRE	+2.46	+2.97	0.1191→0.1194	0.0597→0.0637	−0.0024	+0.0019
ADFTD PRE	+2.37	+2.40	0.2056→0.2056	0.1226→0.1178	+0.0100	+0.0003
ADFTD MID	+4.08	+3.79	0.1707→0.1707	0.1062→0.0933	+0.0091	+0.0050

For CBraMod, the paired analysis examines PRE adaptation on TD-BRAIN and PRE and MID adaptation on ADFTD. Adaptation improves both task metrics over the matched reference in every seed for all three comparisons (Table 3). The two ADFTD comparisons use different held-out folds and their corresponding references. In ADFTD, ridge readout of smoothed spectral descriptors changes little at the adapter output and increases at POST in all three seeds for both comparisons. TD-BRAIN follows a different pattern, with little change in ridge readout at either site. MLP readout varies in direction across comparisons at the adapter output and rises slightly on average at POST, with mixed seed directions (Table 3). Further paired measurements appear in Table 14.

These findings show that the effect of internal adaptation extends beyond the representation directly modified by the adapter, through the remaining frozen layers. Examining both locations therefore reveals changes that an analysis of the adapter output alone would miss. The different ridge and MLP patterns also show that better downstream prediction accompanies changes in how physiological structure can be read out, rather than a uniform increase in every readout score.

5 Conclusion

In this study, we examine how pretrained EEG foundation models are adapted across downstream tasks, and what their internal representations encode. Across the evaluated backbones and tasks, physiological descriptors provide predictive value beyond frozen output representations. These descriptors are also broadly recoverable by readout across network depths, with stronger readout at some internal sites than at the classifier input. Together, these findings point to a representation-to-competence gap between task-relevant structure accessible within pretrained models and the task performance obtained from frozen outputs. Guided by this observation, internal residual adaptation improves classification in most evaluated settings and outperforms final output updates under the same parameter budget in nearly every matched comparison. The better adaptation depth changes with the task and backbone, and selection by task validation stays close to the better fixed choice.

In the adapted CBraMod models on ADFTD, ridge readout of these descriptors rises at the classifier input while it changes during the backbone and at the adapter output. Future work should therefore investigate how physiological structure evolves through the backbone and how these changes influence classification decisions. Understanding these relationships can in turn guide more effective adaptation across EEG tasks. More broadly, making fuller use of existing pretrained models deserves attention alongside expanding the scale of pretraining.

Artificial intelligence use statement

Generative AI tools were used to implement parts of the methods in code, to polish and revise the manuscript text for readability, and to search for and identify related work. All AI-generated code was reviewed and checked by the authors before it was used in the reported experiments, and related work identified with AI assistance was checked against the original publications. Synthetic data generation and qualitative or thematic data analysis are not applicable to this work. The authors take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This work is a secondary analysis of previously collected, de-identified EEG recordings obtained from public repositories under their respective data use agreements. The original studies obtained ethical approval and informed consent; no new human-subject data were collected, and no additional ethical approval was required.

Reproducibility statement

The adaptation methods are specified in Section 3, and the evaluation procedure in Section 4.1. Probe and paired readout specifications are given in Appendices A.5 and E.1; dataset preprocessing and data partitions are described in Appendix A.2.

References

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

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large Brain Model for learning generic representations with tremendous EEG data in BCI. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf.

Min-jae Kim, Young Chul Youn, and Joonki Paik. Deep learning-based EEG analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset. *NeuroImage*, 272: 120054, 2023.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based EEG foundation models, 2026. URL https://arxiv.org/abs/2605.26434.

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.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2):021001, 2026.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

William Lehn-Schiøler, Magnus Ruud Kjær, Rahul Thapa, Magnus Guldberg Pedersen, Anton Mosquera Storgaard, Nick Williams, Radu Gatej, Tue Lehn-Schiøler, Andreas Brink-Kjær, Sadasivan Puthusserypady, Sándor Beniczky, James Zou, and Lars Kai Hansen. Mechanistic interpretability of EEG foundation models via sparse autoencoders, 2026. URL https://arxiv.org/abs/2605.13930.

Jingying Ma, Feng Wu, Yucheng Xing, Qika Lin, Tianyu Liu, Chenyu Liu, Ziyu Jia, and Mengling Feng. SCOPE: Structured prototype-guided adaptation for EEG foundation models with limited labels, 2026. URL https://arxiv.org/abs/2602.17251.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp EEG recordings of Alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6):95, 2023.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Simanto Saha and Mathias Baumert. Intra- and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in Computational Neuroscience*, 13:87, 2020.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.

Ling Tang, Qian Chen, Jilin Mei, Houshi Xu, Quanshi Zhang, Jing Shao, Na Zou, Xia Hu, and Dongrui Liu. What do EEG foundation models capture from human brain signals?, 2026. URL https://arxiv.org/abs/2605.11410.

Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9(1):333, 2022.

Neeraj Wagh, Jionghao Wei, Samarth Rawal, Brent Berry, Leland Barnard, Benjamin Brinkmann, Gregory Worrell, David Jones, and Yogatheesan Varatharajah. Domain-guided self-supervision of EEG data improves downstream classification performance and generalizability. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer (eds.), *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 130–142. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/wagh21a.html.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 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 *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025a.

Suli Wang, Yangshen Deng, Zhenghua Bao, Xinyu Zhan, and Yiqun Duan. NeuroTTT: Bridging pretraining-downstream task misalignment in EEG foundation models via test-time training, 2025b. URL https://arxiv.org/abs/2509.26301.

Xianheng Wang, Yige Yang, and Damien Coyle. EEG-FM-Audit: A systematic evaluation and analysis pipeline for EEG foundation models, 2026. URL https://arxiv.org/abs/2605.26910.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 41179–41212. Curran Associates, Inc., 2025. doi: 10.52202/085713-1375. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3aef4a18c2646b9d11d9ab4d9bec72c8-Paper-Conference.pdf.

Lichao Xu, Minpeng Xu, Yufeng Ke, Xingwei An, Shuang Liu, and Dong Ming. Cross-dataset variability problem in EEG decoding with deep learning. *Frontiers in Human Neuroscience*, 14: 103, 2020.

Haolin Yang, Hakaze Cho, Kaize Ding, and Naoya Inoue. Task vectors, learned not extracted: Performance gains and mechanistic insights. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=RGEbVZgf4E.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? Comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=5Xwm8e6vbh.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 87150–87195. Curran Associates, Inc., 2025. doi: 10.52202/085713-2915. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7e199ad8ae40eb19b2980f61f659cb07-Paper-Conference.pdf.

A EXPERIMENTAL AND METHODOLOGICAL DETAILS

A.1 EXPERIMENTAL COVERAGE

The descriptor and update-location analyses cover four backbones and four tasks. Descriptor diagnostics measure incremental predictive value, specificity relative to matched noise, and readout across representation sites. They include all 16 model–task settings for each of the four families, giving 64 combinations in total. The location comparison evaluates classifier, FINAL, and assigned-site updates in all 16 settings.

The independent depth comparison evaluates fixed PRE, fixed MID, and task-validation selection in LaBraM and CBraMod on TD-BRAIN and ADFTD. Paired representation and auxiliary-supervision analyses use CBraMod on these two tasks. Paired readout is summarized by task and selected depth. Appendix A.7 specifies the selection procedures.

A.2 DATA, PREPROCESSING, AND EVALUATION

Cohorts and splits. Table 4 lists the task labels and retained cohorts. In the class order listed there, the retained window counts are (9,201, 5,730, 3,431) for TD-BRAIN, (14,506, 8,258, 12,010) for ADFTD, (18,359, 15,665, 11,941) for CAUEEG, and (44,427, 19,086, 14,512) for TUSZ.

Table 4: Task labels and cohort sizes, reported as distinct subjects or patients, recordings, and retained EEG windows. Class labels are major depressive disorder (MDD), attention-deficit/hyperactivity disorder (ADHD), subjective memory complaints (SMC), Alzheimer's disease (AD), frontotemporal dementia (FTD), mild cognitive impairment (MCI), focal nonspecific seizures (fnsz), generalized nonspecific seizures (gnsz), and complex partial seizures (cpsz).

Task	Classes in label order	Participants	Recordings	Windows
TD-BRAIN	MDD, ADHD, SMC	589	589	18362
ADFTD	AD, FTD, Healthy	88	88	34774
CAUEEG	Normal, MCI, Dementia	1115	1115	45965
TUSZ	fnsz, gnsz, cpsz	215	997	78025

We use five-fold cross-validation grouped by subject for TD-BRAIN and ADFTD and by patient for TUSZ. All recordings and windows from an individual are assigned to the same outer fold, so training and test individuals are disjoint in every split. CAUEEG retains its official training, validation, and test sets, containing 945, 82, and 88 participants and 38,818, 3,545, and 3,602 windows, respectively. We form its five development folds within the official training data. Task seeds are 41, 42, and 43.

Descriptor preprocessing uses the corresponding training partitions. On TD-BRAIN, spectral transforms are fitted on training partitions from an 18,441-window cohort, whereas relational and morphology transforms use training partitions from the retained 18,362-window cohort. All four families are evaluated on the same 18,362 windows. The evaluation sets contain 18,362, 34,774, 3,602, and 78,025 windows per seed, respectively.

Preprocessing. Table 5 summarizes the EEG properties after dataset preprocessing. TD-BRAIN and TUSZ inputs additionally receive 50 Hz and 60 Hz notch filtering, respectively.

Backbone inputs. Each backbone then applies its own input transformation. LaBraM and CBraMod receive ADFTD and TUSZ after polyphase resampling from 250 Hz to 200 Hz. LaBraM normalizes each channel by the training mean and standard deviation, with a 10^{-6} floor on the scale. CBraMod splits the input into one-second patches of 200 samples, four per four-second window and ten for CAUEEG. BrainOmni resamples to 256 Hz and uses 26 sensors for TD-BRAIN and 19 for the other tasks. For TUSZ, its four-second intervals are extracted from European Data Format (EDF) recordings using 19 referential sensors. Sensor coordinates follow the standard 10–20 montage. Its EEG normalization subtracts the channel mean at each time, then divides by the population standard deviation over the EEG group plus 10^{-5}. Centered coordinates are scaled by $\sqrt{3\,\mathrm{mean}_u \|\boldsymbol{p}_u - \bar{\boldsymbol{p}}\|_2^2}$,

Table 5: EEG properties after dataset preprocessing and before backbone-specific input preparation. A dash denotes an unspecified parameter.

Task	Hz	Channels	Window (s)	Stride (s)	Band (Hz)
TD-BRAIN	200	26	4	—	0.5–95
ADFTD	250	19	4	2	0.5–45
CAUEEG	200	19	10	10	0.5–45
TUSZ	250	18	4	2	0.5–70

Table 6: Sampled representation sites and feature dimensions. Block numbers start at one; each site denotes the corresponding layer output.

Model	PRE	MID	FINAL	Internal readout	POST
LaBraM	Patch embedding	Block 6	Final normalization	200	200
CBraMod	Patch embedding	Block 6	Block 12	200	200
BrainOmni	Input projection	Block 6	Block 11	256	4096
CSBrain	Patch embedding	Block 6	Block 12	200	Task dependent

where $\boldsymbol{p}_u$ is the coordinate of sensor u and $\bar{\boldsymbol{p}}$ is the mean sensor coordinate. The input is divided into 512-sample segments separated by 384 samples; an incomplete final segment is padded with zeros. CSBrain runs at 200 Hz over 19 channels for the first three tasks and 16 bipolar derivations for TUSZ. Amplitudes are in microvolts divided by 100, and ADFTD and TUSZ are resampled from 250 Hz. Input durations remain four or ten seconds. We use the released LaBraM-base, CBraMod, and BrainOmni tiny models, and the official CSBrain model pretrained on the Temple University Hospital EEG Corpus (TUEG).

Reference classifier. For each setting, we standardize the training features and fit a logistic reference classifier with inverse regularization strength set to 1, the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver, at most 3000 iterations, an intercept, and no class weighting. Its forward map is

$$g_{\phi_0}(\boldsymbol{z}) = \boldsymbol{W}\big((\boldsymbol{z} - \boldsymbol{\mu}_z) \oslash \boldsymbol{s}_z\big) + \boldsymbol{b},$$

where $\boldsymbol{\mu}_z, \boldsymbol{s}_z \in \mathbb{R}^{d_{\mathrm{POST}}}$ are the fitted training feature mean and scale, $\boldsymbol{W} \in \mathbb{R}^{3\times d_{\mathrm{POST}}}$ is the coefficient matrix, and $\boldsymbol{b} \in \mathbb{R}^3$ is the intercept. The symbol $\oslash$ denotes elementwise division. These fitted quantities and the class order initialize adaptation.

Evaluation metrics. Within each seed, we pool the applicable evaluation windows across folds and compute Macro-F1 and balanced accuracy on that pool. Reported scores average the three seed values. Macro-F1 averages over the three predefined classes, assigning zero to an undefined class F1. Balanced accuracy is the unadjusted mean recall over classes present in the pooled ground truth. We report differences between conditions after averaging the three seed scores, separately for each metric, model, and task. Scores and their differences use the percentage scale defined in Section 4.1.

Hardware. Adaptation runs on an NVIDIA A5000 graphics processing unit (GPU; 24 GB).

A.3 Representation sites

Table 6 locates PRE, MID, FINAL, and POST in each backbone. PRE samples the embedding module, and POST is the vector supplied to the classifier. FINAL adaptation updates the structured FINAL representation before its reduction to POST.

For each EEG window, LaBraM represents each token by 200 coordinates. PRE contains 104, 76, 190, and 72 tokens on TD-BRAIN, ADFTD, CAUEEG, and TUSZ, respectively. MID and FINAL include an additional class token. The classifier input excludes that token, averages the remaining patch tokens, and applies the fixed final normalization.

CBraMod represents each channel and time patch by a 200-dimensional vector. The numbers of channels and patches are (26,4), (19,4), (19,10), and (18,4) for TD-BRAIN, ADFTD, CAUEEG, and TUSZ, respectively. Its classifier input averages over both channels and patches. BrainOmni

produces 16×24×256 features per four-second window and 16×56×256 per ten-second CAUEEG window. Normalized averaging over tokens followed by vectorization gives 4096 classifier features.

CSBrain also uses 200-dimensional embeddings for each channel and time segment. Its representations have dimensions 19×4×200 on TD-BRAIN and ADFTD, 19×10×200 on CAUEEG, and 16×4×200 on TUSZ. Vectorizing all embeddings gives classifier inputs of 15,200, 38,000, and 12,800 dimensions, respectively.

Internal readers average over channel or token positions while retaining the embedding coordinates. Where validity weights are supplied, these determine the weighted mean; otherwise, all finite positions are weighted equally. POST readers use the classifier input directly.

A.4 Physiological descriptor definitions

The four descriptor families quantify spectral, interchannel, and temporal structure in EEG windows sampled at 200 Hz. ADFTD and TUSZ are resampled from 250 Hz using polyphase filtering. Let $\boldsymbol{X} \in \mathbb{R}^{C\times T}$ contain C channels and T samples, and let $\boldsymbol{x}_u \in \mathbb{R}^T$ denote the trace of channel u. Each family is reduced to 32 standardized principal-component scores using transforms fitted on training data.

Spectral residuals and smoothed structure (P and A). For each channel, we calculate the log-power spectrum from the discrete Fourier transform $\mathcal{F}$,

$$s_{u,b} = \log\Big(1 + |\mathcal{F}(\boldsymbol{x}_u)_b|^2\Big),$$

where b indexes the nonnegative Fourier frequencies. We retain the 1–45 Hz range and represent the spectrum on the common grid 1,1.25,…,45 Hz, forming $\boldsymbol{L} \in \mathbb{R}^{C\times 177}$. Four-second windows have a frequency spacing of 0.25 Hz. For the ten-second CAUEEG windows, whose spacing is 0.1 Hz, the log-power values are linearly interpolated onto the common grid.

A centered 17-bin moving average along frequency defines $\boldsymbol{L}_{\mathrm{smooth}}$, with the nearest boundary value extended beyond each endpoint. Family A contains the smoothed values, and family P contains the residual $\boldsymbol{L} - \boldsymbol{L}_{\mathrm{smooth}}$. Each matrix is flattened by channel and then frequency, giving $177C$ features per family. The two families provide smoothing-based descriptions of aperiodic structure and periodic deviations.

Interchannel relations (R). Channel traces are centered and scaled by their sample standard deviations, with a lower bound of 10^{-6} on the scale. The resulting matrix $\boldsymbol{X}_{\mathrm{std}}$ defines

$$\boldsymbol{G} = \frac{\boldsymbol{X}_{\mathrm{std}}\boldsymbol{X}_{\mathrm{std}}^\top}{T-1}, \qquad \boldsymbol{G}_{\mathrm{reg}} = \frac{\boldsymbol{G}+\boldsymbol{G}^\top}{2} + 10^{-4}\frac{\mathrm{tr}(\boldsymbol{G})}{C}\boldsymbol{I}_C,$$

where $\boldsymbol{I}_C$ is the C-dimensional identity matrix. The diagonal term regularizes the correlation matrix before we compute its matrix logarithm by symmetric eigendecomposition. After symmetrizing $\log \boldsymbol{G}_{\mathrm{reg}}$, we vectorize its upper triangle in row order and multiply off-diagonal entries by $\sqrt{2}$. This scaling preserves the Frobenius norm of the symmetric matrix in the resulting $C(C+1)/2$-dimensional vector.

Temporal morphology (M). Morphology descriptors summarize the complete EEG window: four seconds for TD-BRAIN, ADFTD, and TUSZ, and ten seconds for CAUEEG. Each channel is robustly normalized using its median and median absolute deviation (MAD). Hjorth mobility and complexity characterize local temporal variation in segments of 0.5, 1, and 2 seconds. Their median and interquartile range (IQR) across channel–segment estimates provide 12 features. Lagged line length measures absolute changes between samples separated by a time lag. For each combination of these three segment lengths and lags of 0.02, 0.08, and 0.32 seconds, the median across channel–segment estimates provides one feature, giving nine further features. Segments overlap by 50%, with the final segment reaching the window endpoint.

Burst descriptors capture intermittent activity in the 4–30 Hz band. For each channel, bursts are detected from the Hilbert envelope using its median plus $3 \times 1.4826 \times \mathrm{MAD}$ as the threshold, a minimum duration of 0.05 second, and a merge gap of 0.025 second. The channel medians of burst

occupancy, event duration, and event density provide three features. Burst summaries are zero when no events are detected. Occupancy denotes the fraction of the window occupied by bursts, and event density denotes the number of bursts per unit time.

Time–frequency features use Morlet wavelet coefficients $w_{u,\nu}(\tau)$ for channel u, frequency ν, and time τ. For each frequency in {2,4,8,16,30} Hz, we calculate

$$e_{u,\nu} = \log\bigl(1 + \bigl\langle |w_{u,\nu}(\tau)|^2 \bigr\rangle_\tau\bigr)\,,$$

where $\langle\cdot\rangle_\tau$ averages over the window. The median and IQR of $e_{u,\nu}$ across channels give two features at each frequency. Together, the temporal, burst, and wavelet summaries form 34 raw features.

Dimensionality reduction. Randomized PCA, initialized with seed 42 and without whitening, retains 32 components per family. Before PCA, raw P, A, and R features are standardized on CAUEEG and TUSZ, and raw M features are standardized on all four datasets. Component scores are standardized after PCA. The transforms are fitted on the official training set for CAUEEG and on the training partition of each fold for the other tasks; CAUEEG development partitions use the same training-set transforms. Appendix A.2 describes the fitting cohorts.

A.5 Probe fitting and readout scores

Readers predict the 32 standardized coordinates of one descriptor family. Input standardization uses the probe training data, and target standardization uses the corresponding outer training partition or official training set. Constant target coordinates use a scale of one. Each window receives equal weight in fitting and evaluation.

The linear reader is multivariate ridge regression with an intercept and regularization strength 1. We solve the regularized least-squares problem by Cholesky decomposition. The same deterministic fitting procedure is used across models, tasks, descriptor families, and representation sites.

The MLP reader has a 64-unit hidden layer with a Gaussian error linear unit (GELU) activation and a 32-unit output layer, with biases in both layers. We train it by adaptive moment estimation with decoupled weight decay (AdamW) at a constant learning rate of 10^{-3}, weight decay 10^{-4}, moment decay coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and numerical stability constant $\epsilon = 10^{-8}$. Each update uses the full inner training set, without gradient clipping. Training stops after at most 100 epochs or ten epochs without improvement in validation mean squared error (MSE). We retain the model from the epoch with the lowest validation MSE, choosing the earlier epoch in a tie.

For early stopping, the reader training subjects or patients are randomly partitioned into five folds using seed 42. The first fold supplies validation data, and the other four supply fitting data. Input normalization uses the fitting portion, while target normalization uses the outer or official training statistics described above. We evaluate three MLP initializations with seeds 41, 42, and 43. In a layer with d_{in} inputs, weights and biases are sampled from $\mathcal{U}(-1/\sqrt{d_{\mathrm{in}}}, 1/\sqrt{d_{\mathrm{in}}})$, where $\mathcal{U}$ denotes the uniform distribution.

For a fixed model, task, and descriptor family, ℓ indexes the representation site, p the probe type (ridge or MLP), and s the seed. Within evaluation fold f, i indexes windows and $j = 1, \ldots, 32$ indexes standardized descriptor coordinates. Let $q_{f,ij}$ be the jth coordinate of $\boldsymbol{q}_{f,i}$ in Equation 2, $\widehat{q}^{(\ell,p)}_{f,s,ij}$ its prediction at site ℓ from reader p with seed s, and $\bar{q}^{\mathrm{train}}_{f,j}$ the corresponding training mean. Model, task, and family indices are suppressed. The readout score and the mean across MLP initialization seeds are

$$E_{\ell,p,s} = 1 - \frac{\sum_f \sum_{i,j} (q_{f,ij} - \widehat{q}^{(\ell,p)}_{f,s,ij})^2}{\sum_f \sum_{i,j} (q_{f,ij} - \bar{q}^{\mathrm{train}}_{f,j})^2}, \qquad \bar{E}_{\ell,\mathrm{MLP}} = \frac{1}{3} \sum_{s\in\{41,42,43\}} E_{\ell,\mathrm{MLP},s}.$$

The denominator is the held-out squared error of the constant training-mean predictor. Target coordinates are centered using the corresponding training statistics, so $\bar{q}^{\mathrm{train}}_{f,j} = 0$ in standardized coordinates and the denominator equals $\sum_f \sum_{i,j} q^2_{f,ij}$, up to numerical precision. This equivalence does not require the held-out targets to have zero mean. The sums pool the five outer test folds; CAUEEG uses its official test set. Positive E indicates that the reader predicts the descriptors more accurately

Table 7: Trainable parameter counts. Internal and output residuals each have $2rd_\ell$ parameters; the three-class affine classifier has $3(d_{\mathrm{POST}} + 1)$. Equal parameter budgets apply to the two residual locations.

Model	Tasks	Internal/output residual	Classifier
LaBraM	All four	3200	603
CBraMod	All four	3200	603
BrainOmni	All four	4096	12291
CSBrain	TD-BRAIN, ADFTD	3200	45603
CSBrain	CAUEEG	3200	114003
CSBrain	TUSZ	3200	38403

Table 8: Adaptation batch sizes. Each batch produces one optimizer update, with no gradient accumulation.

Model	TD-BRAIN	ADFTD	CAUEEG	TUSZ
LaBraM	48	48	24	48
CBraMod	256	512	48	96
BrainOmni	8	8	8	8
CSBrain	8	8	8	8

than the training-mean baseline. MLP scores are averaged across three initialization seeds; the deterministic ridge fit gives one score, denoted $E_{\ell,\mathrm{ridge}}$. In the results, E refers to this ridge score or to $\bar{E}_{\ell,\mathrm{MLP}}$, according to the reader.

A.6 Adapter training

The residual in Equation 3 acts on each embedding vector. After arranging the batch's embedding vectors as rows of $\boldsymbol{H}$, its matrix form is $\Delta\boldsymbol{H} = (\boldsymbol{H}\boldsymbol{A}_\ell^\top)\boldsymbol{B}_\ell^\top$. The rank parameter is $r = 8$, and the LoRA scaling coefficient $\alpha = 8$ gives $\alpha/r = 1$. Entries of $\boldsymbol{A}_\ell$ are sampled from $\mathcal{U}(-1/\sqrt{d_\ell}, 1/\sqrt{d_\ell})$, where d_ℓ is the embedding dimension at site ℓ. We initialize $\boldsymbol{B}_\ell$ to zero, so adaptation starts from the pretrained representation. Internal and FINAL adaptation train only these two matrices, with the reference classifier and backbone weights fixed. Classifier adaptation trains the reference coefficient matrix and intercept, with feature normalization held fixed. Table 7 lists the trainable parameter counts, using the embedding dimension d_ℓ and POST dimension d_{POST} defined in Section 3. Task-only adaptation uses Equation 5, with unweighted mean cross-entropy and a coefficient of one on the displacement penalty. The Frobenius norms include all windows and embedding positions in the batch. Physiological adapter training uses the same task loss and displacement penalty, with the descriptor loss specified in Appendix F.1. Classifier adaptation uses cross-entropy alone. Gradients pass through the fixed remaining backbone and reference classifier to the adapter matrices.

Adaptation optimization. All location conditions use AdamW with a constant learning rate of 10^{-4}, weight decay of 10^{-4}, and global gradient norm clipping at 1. Training lasts at most 30 epochs, with early stopping after five epochs without validation improvement. We select the model from the epoch with the highest pooled-window validation Macro-F1, using balanced accuracy and then the earlier epoch to resolve ties. Conditions within each model–task setting share the same shuffled batch order at each epoch. Each batch produces one optimizer update, including the final partial batch. Table 8 gives the batch sizes.

A.7 Update locations and depth selection

The location comparison uses one fixed site per model–task setting, shared across folds and seeds. The assigned site is MID in all settings except CSBrain on TD-BRAIN and CAUEEG, where it is FINAL (Table 1). The separate depth comparison selects between independently trained PRE and MID candidates using task validation within each fold.

Selecting the internal depth. The separate depth comparison evaluates PRE and MID in LaBraM and CBraMod on TD-BRAIN and ADFTD. Candidates at the two depths are trained independently. Within each fold, task-only validation scores averaged across seeds select the depth, with Macro-F1 first, balanced accuracy second, and PRE resolving numerical ties, defined as score differences no larger than 10^{-12}. This comparison selects depth from task performance; the location study uses the fixed sites specified above. Table 12 lists the selected depths by fold; Table 2 reports their task scores.

A.8 Function class of FINAL adaptation

Proposition. For CBraMod's structural mean and CSBrain's row-wise vectorization, a FINAL residual preserves affine logits on the reference POST vector $\boldsymbol{z}$. Let $\ell = \mathrm{FINAL}$ and $\boldsymbol{U}_\ell = \boldsymbol{I}_{d_\ell} + \boldsymbol{B}_\ell \boldsymbol{A}_\ell$ denote the transform applied to each embedding column vector. Here $\boldsymbol{I}_{d_\ell}$ is the d_ℓ-dimensional identity matrix. For CBraMod let $\boldsymbol{T} = \boldsymbol{U}_\ell$, and for CSBrain let $\boldsymbol{T} = \boldsymbol{I}_J \otimes \boldsymbol{U}_\ell$, where J is the number of positions in the structured output and $\otimes$ denotes the Kronecker product. Thus $\boldsymbol{T} \in \mathbb{R}^{d_{\mathrm{POST}} \times d_{\mathrm{POST}}}$ maps the reference POST vector to the adapted POST vector. Define the frozen scaling matrix $\boldsymbol{\Lambda}_z = \mathrm{diag}(1/\boldsymbol{s}_z)$, with the reciprocal taken elementwise and diag placing its argument on the diagonal. Then

$$\begin{aligned}\boldsymbol{\eta}^{\mathrm{FINAL}}(\boldsymbol{z}) &= \boldsymbol{W}\boldsymbol{\Lambda}_z(\boldsymbol{T}\boldsymbol{z} - \boldsymbol{\mu}_z) + \boldsymbol{b} \\ &= \boldsymbol{W}\boldsymbol{\Lambda}_z\boldsymbol{T}\boldsymbol{z} + \boldsymbol{b}_{\mathrm{eff}}, \qquad \boldsymbol{b}_{\mathrm{eff}} = \boldsymbol{b} - \boldsymbol{W}\boldsymbol{\Lambda}_z\boldsymbol{\mu}_z.\end{aligned}$$

Every such predictor belongs to the affine function class available to classifier adaptation. Here $\boldsymbol{\eta}^{\mathrm{FINAL}}$ is the three-class logit vector and $\boldsymbol{b}_{\mathrm{eff}}$ is the effective intercept.

Proof. Arrange the J embedding vectors of one FINAL output as rows of $\boldsymbol{S} \in \mathbb{R}^{J \times d_\ell}$, omitting the window index. The residual gives $\widetilde{\boldsymbol{S}} = \boldsymbol{S}\boldsymbol{U}_\ell^\top$. For CBraMod, averaging these rows and expressing the result as a column gives $\rho_m(\widetilde{\boldsymbol{S}}) = \boldsymbol{U}_\ell \rho_m(\boldsymbol{S})$. For CSBrain, stacking the transposed rows gives $\rho_m(\widetilde{\boldsymbol{S}}) = (\boldsymbol{I}_J \otimes \boldsymbol{U}_\ell)\rho_m(\boldsymbol{S})$. Substitution into the frozen standardized classifier gives the displayed logits. Since $\boldsymbol{\Lambda}_z$ is invertible, classifier adaptation can represent them with

$$\begin{aligned}\boldsymbol{W}_* &= \boldsymbol{W}\boldsymbol{\Lambda}_z\boldsymbol{T}\boldsymbol{\Lambda}_z^{-1}, \\ \boldsymbol{b}_* &= \boldsymbol{b} + \boldsymbol{W}\boldsymbol{\Lambda}_z(\boldsymbol{T} - \boldsymbol{I}_{d_{\mathrm{POST}}})\boldsymbol{\mu}_z,\end{aligned}$$

where $\boldsymbol{W}_*$ and $\boldsymbol{b}_*$ are the corresponding classifier parameters.

Although $\mathrm{rank}(\boldsymbol{U}_\ell - \boldsymbol{I}_{d_\ell}) \le r$, the shared residual induces a vectorized CSBrain update of rank $\mathrm{rank}(\boldsymbol{T} - \boldsymbol{I}_{Jd_\ell}) = J\,\mathrm{rank}(\boldsymbol{U}_\ell - \boldsymbol{I}_{d_\ell}) \le Jr$.

LaBraM and BrainOmni instead include input-dependent normalization in their reductions. For example, let $\boldsymbol{v} \in \mathbb{R}^{d_\ell}$ be an embedding vector and let LN denote layer normalization. Then $\mathrm{LN}(\boldsymbol{U}_\ell \boldsymbol{v})$ cannot generally be expressed as a fixed affine transformation of $\mathrm{LN}(\boldsymbol{v})$. The effect of a FINAL residual therefore depends on the normalization of the adapted features, and the affine equivalence established above does not hold in general. In all four backbones, an internal residual is followed by the remaining pretrained layers.

B Predictive value and specificity of physiological descriptors

B.1 Spectral and relational task value on development folds

Table 9 reports the incremental task value of spectral and relational descriptors on development folds. Every model–task setting has at least one family with positive development values on both metrics. R improves both metrics in 13 of 16 settings and lowers both on ADFTD for LaBraM, CSBrain, and BrainOmni. A gives the largest development values, 7.03%/6.63% for CBraMod and 6.14%/5.69% for LaBraM on ADFTD.

B.2 Specificity under matched standardized-noise augmentation

The noise control measures the effect of adding 32 classifier inputs without descriptor content. We replace each descriptor family with standard-normal noise and standardize it using the training-partition mean and standard deviation. Separate noise matrices are generated for the training, validation, and test partitions of each dataset and fold. For a given window, noise coordinates are shared

Table 9: Incremental predictive value of spectral and relational descriptors on development folds. Each entry gives Macro-F1 / balanced accuracy differences (%), averaged over five development folds and three seeds.

Model	Task	P	A	R
LaBraM	TD-BRAIN	+0.13 / +0.13	**+5.38 / +5.45**	+2.24 / +2.44
LaBraM	ADFTD	+0.06 / +0.02	**+6.14 / +5.69**	−0.99 / −0.81
LaBraM	CAUEEG	+0.08 / +0.05	−0.07 / −0.12	**+1.53 / +1.43**
LaBraM	TUSZ	−0.41 / −0.34	−0.67 / −0.69	**+2.59 / +2.22**
CBraMod	TD-BRAIN	+0.06 / +0.12	+0.64 / +0.64	**+3.94 / +4.08**
CBraMod	ADFTD	+0.26 / +0.16	**+7.03 / +6.63**	+1.60 / +1.40
CBraMod	CAUEEG	+0.19 / +0.16	+0.21 / +0.08	**+1.72 / +1.48**
CBraMod	TUSZ	−0.07 / −0.12	−0.60 / −0.72	**+3.78 / +2.90**
CSBrain	TD-BRAIN	+0.29 / +0.31	+0.02 / −0.01	**+2.87 / +2.78**
CSBrain	ADFTD	+0.07 / +0.06	**+0.36 / +0.26**	−4.37 / −4.51
CSBrain	CAUEEG	−0.21 / −0.22	−0.02 / −0.03	**+0.42 / +0.43**
CSBrain	TUSZ	+0.02 / −0.07	+0.31 / +0.13	**+3.90 / +3.97**
BrainOmni	TD-BRAIN	−0.27 / −0.26	+0.48 / +0.48	**+2.07 / +1.61**
BrainOmni	ADFTD	−0.16 / −0.16	**+0.55 / +0.70**	−1.33 / −1.56
BrainOmni	CAUEEG	−0.23 / −0.25	−0.36 / −0.37	**+0.18 / +0.20**
BrainOmni	TUSZ	−0.08 / −0.08	+0.04 / +0.08	**+1.61 / +1.70**

Table 10: Matched standardized-noise comparison. Entries give pooled held-out Macro-F1 / balanced accuracy differences (%). Noise reports noise-augmented minus POST-only scores; family columns report descriptor-augmented minus noise-augmented scores (Equation 6).

Model	Task	Noise	P	A	R	M
LaBraM	TD-BRAIN	−0.19 / −0.18	+0.30 / +0.31	**+4.16 / +4.21**	+2.80 / +2.95	+0.59 / +0.72
LaBraM	ADFTD	+0.03 / −0.01	−0.00 / −0.02	**+3.53 / +2.74**	−2.63 / −3.25	+1.17 / +1.09
LaBraM	CAUEEG	+0.00 / +0.12	−0.15 / −0.22	+0.02 / +0.31	**+0.60 / +0.36**	+0.41 / +0.26
LaBraM	TUSZ	+0.02 / +0.01	−0.18 / −0.20	+0.69 / +0.56	**+2.89 / +2.59**	+2.13 / +2.01
CBraMod	TD-BRAIN	−0.24 / −0.21	+0.01 / +0.03	−0.33 / −0.22	**+5.48 / +5.67**	+4.37 / +4.55
CBraMod	ADFTD	−0.05 / −0.05	−0.05 / −0.11	**+4.41 / +3.91**	−1.35 / −1.63	+1.51 / +1.55
CBraMod	CAUEEG	−0.25 / −0.17	+0.21 / +0.11	+1.94 / +2.31	+2.82 / +2.62	**+3.38 / +3.12**
CBraMod	TUSZ	−0.01 / −0.00	−0.24 / −0.24	+1.37 / +1.29	**+5.70 / +5.01**	+1.32 / +1.06
CSBrain	TD-BRAIN	−0.02 / −0.08	+0.09 / +0.14	−0.01 / +0.00	**+2.73 / +2.67**	+2.49 / +2.50
CSBrain	ADFTD	+0.10 / +0.09	−0.07 / −0.06	**+0.64 / +0.65**	−0.83 / −0.64	−0.20 / −0.19
CSBrain	CAUEEG	+0.44 / +0.47	−0.12 / −0.14	−0.04 / −0.09	+0.04 / +0.02	**+0.21 / +0.20**
CSBrain	TUSZ	−0.08 / −0.07	+0.24 / +0.22	+1.44 / +1.36	**+5.67 / +5.38**	+4.14 / +3.95
BrainOmni	TD-BRAIN	+0.01 / +0.03	−0.27 / −0.31	+0.20 / +0.21	+2.44 / +2.11	**+2.92 / +2.80**
BrainOmni	ADFTD	+0.02 / +0.01	**+0.10 / +0.11**	−0.48 / −0.45	−0.67 / −0.51	−0.22 / −0.23
BrainOmni	CAUEEG	−0.33 / −0.32	+0.33 / +0.44	+0.41 / **+0.59**	+0.21 / +0.16	**+0.51** / +0.43
BrainOmni	TUSZ	−0.02 / −0.03	−0.29 / −0.27	+0.39 / +0.41	**+1.86 / +1.76**	+0.42 / +0.44

across backbones, descriptor families, and classifier seeds. Evaluation pools held-out windows, using the official test set for CAUEEG. The noise-augmented classifier uses the same concatenation, logistic classifier, regularization, splits, and scoring procedure as the descriptor comparison in Equation 1. With $\boldsymbol{N}_t$ denoting the 32-column noise matrix for task t, with rows aligned to $\boldsymbol{Z}_{m,t}$ and fold and partition indices omitted,

$$\begin{aligned} D^{\mathrm{noise}}_{m,t,k} &= \mathcal{M}_{t,k}([\boldsymbol{Z}_{m,t}, \boldsymbol{N}_t]) - \mathcal{M}_{t,k}(\boldsymbol{Z}_{m,t}), \\ D^{\mathrm{Q-N}}_{m,t,a,k} &= \mathcal{M}_{t,k}([\boldsymbol{Z}_{m,t}, \boldsymbol{Q}_{t,a}]) - \mathcal{M}_{t,k}([\boldsymbol{Z}_{m,t}, \boldsymbol{N}_t]). \end{aligned} \tag{6}$$

Positive $D^{\mathrm{Q-N}}_{m,t,a,k}$ means that family a raises the score above that obtained with matched noise.

Figure 2 reports $D = D^{\mathrm{noise}} + D^{\mathrm{Q-N}}$, using the contrasts at the precision reported in Table 10. Within each model–task setting in this control, descriptor and noise augmentation are compared

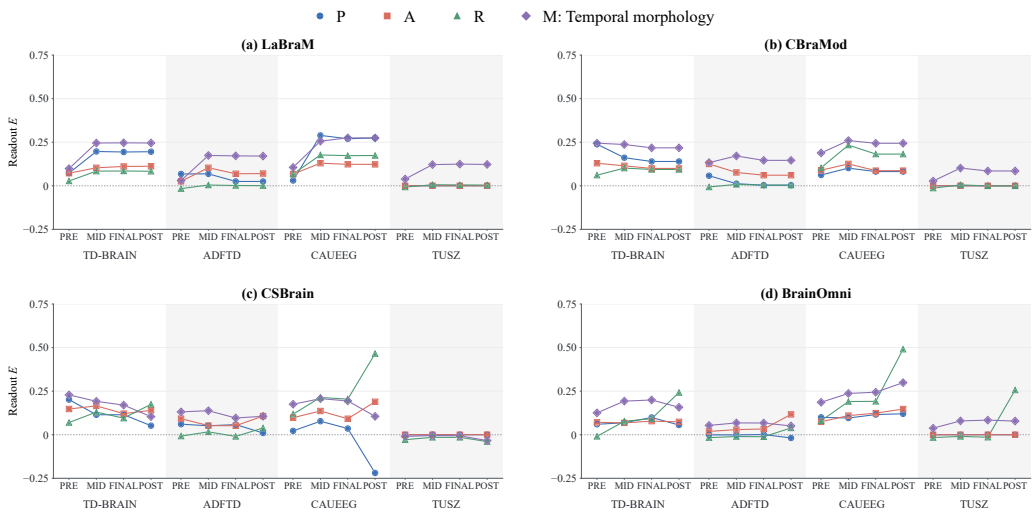

Figure 4: **Nonlinear physiological readout across representation sites.** Panels (a)–(d) show MLP readout for LaBraM, CBraMod, CSBrain, and BrainOmni across the four tasks. Markers distinguish P, A, R, and M, and lines connect PRE, MID, FINAL, and POST within each task.

with the same POST-only reference. Adaptation comparisons use their own reference classifier. Matched noise changes Macro-F1 by -0.33% to $+0.44\%$ and balanced accuracy by -0.32% to $+0.47\%$, with means of -0.04% and -0.02%, respectively. Measured descriptors exceed matched noise on both metrics in 45 of 64 model–task–family combinations, including R in 12 of 16 settings and M in 14 of 16. For M, the mean advantages over noise are 1.57% in Macro-F1 and 1.52% in balanced accuracy. The matched comparison supports the predictive contribution of descriptor content beyond the effect of adding 32 classifier inputs.

C Physiological accessibility across pretrained representations

C.1 Nonlinear readout profiles

Figure 4 examines whether a nonlinear reader recovers descriptor information beyond that accessible to ridge regression. It uses the same families, representation sites, and score E as the ridge analysis in Figure 3, allowing the two readout profiles to be compared directly.

C.2 Readout coverage across models, tasks, and descriptor families

Ridge and MLP readers give positive readout in 62 of 64 model–task–family combinations (Table 11). Ridge is positive at one or more sites in 57 combinations. MLP adds five cases with no positive ridge readout: P on TUSZ for all four backbones and P on ADFTD for BrainOmni. The two combinations without positive readout under either reader are R and M for CSBrain on TUSZ. Both have positive held-out task value under the corresponding descriptor augmentation. M is linearly readable in 15 of 16 settings. Readout varies with site, with several families more accessible at PRE or MID than at POST. For M, MLP readout at MID exceeds POST in 12 of 16 settings, with a median difference of 0.0171. FINAL denotes the structured backbone output before reduction to POST. In CBraMod, the structural averaging used by FINAL readers produces the same features as POST, giving identical readout values at these two sites.

Table 11: Readout coverage across model–task–family combinations. Linear means positive ridge E at any sampled site; MLP only means positive MLP E with no positive ridge E.

Family	Settings	Linear	MLP only	Either reader
P	16	11	5	16
A	16	16	0	16
R	16	15	0	15
M	16	15	0	15
Total	64	57	5	62

Table 12: Task-validation-selected depths in folds 0–4. Paired CBraMod readout is summarized by task and selected depth in Appendix E.1.

Model	Task	0	1	2	3	4
CBraMod	TD-BRAIN	PRE	PRE	PRE	PRE	PRE
CBraMod	ADFTD	PRE	PRE	PRE	MID	MID
LaBraM	TD-BRAIN	PRE	MID	MID	MID	PRE
LaBraM	ADFTD	MID	MID	MID	MID	MID

D Internal adaptation across locations and depths

D.1 Validation-selected internal depths

Table 12 gives the fold choices underlying Table 2. CBraMod selects PRE in every TD-BRAIN fold and both PRE and MID on ADFTD, while LaBraM selects MID in every ADFTD fold and both depths on TD-BRAIN. Task validation therefore adapts the placement to each model, task, and fold. For each model and task setting, metric k, and depth policy π, the shortfall from the better fixed depth is $\max_{\ell \in \{\mathrm{PRE},\mathrm{MID}\}} \bar{\mathcal{M}}_{k,\ell} - \bar{\mathcal{M}}_{k,\pi}$, where $\bar{\mathcal{M}}_{k,\ell}$ is the mean score for metric k at fixed depth ℓ across the three task seeds, and $\bar{\mathcal{M}}_{k,\pi}$ is the corresponding mean under policy π. Model and task indices are omitted. This difference is negative when the policy outperforms both fixed depths. The difference is computed from unrounded scores.

E Representation changes after internal adaptation

E.1 Matched proximal and POST readout

Paired CBraMod readout is summarized by task and validation-selected depth (Table 12). The TD-BRAIN PRE, ADFTD PRE, and ADFTD MID comparisons contain 18,362, 21,513, and 13,261 evaluation windows per seed, respectively. Targets are the 32 R coordinates for TD-BRAIN and 32 A coordinates for ADFTD. The two ADFTD comparisons cover different held-out folds. Task scores and readout pool evaluation windows within each comparison and task seed before averaging seeds 41, 42, and 43.

Proximal reader architectures, normalization, optimization, and scoring follow Appendix A.5. Readers are refitted separately on reference and adapted features at both sites. Proximal features are the 200-dimensional structural mean immediately after the residual; POST features are the classifier input after the remaining backbone. Table 3 reports task gains and proximal readout before and after adaptation. Table 14 adds absolute task scores, shuffled-target readout, and POST changes, including whether the direction of each change is consistent across seeds.

E.2 Variation across task seeds

Table 13 reports sample standard deviations across task seeds 41, 42, and 43. Each seed contributes one score from its pooled evaluation windows, with the evaluation partitions held fixed across seeds. Reference and classifier-adaptation standard deviations round to zero at the reported precision in

Table 13: Sample standard deviation (SD) across the three task seeds (%). Each seed contributes one pooled-window score. Cells give Macro-F1 / balanced accuracy; conditions and assigned sites follow Table 1.

Model	Task	Reference	Classifier	Final output	Assigned site
LaBraM	TD-BRAIN	0.000 / 0.000	0.000 / 0.000	0.014 / 0.100	0.128 / 0.092
LaBraM	ADFTD	0.000 / 0.000	0.000 / 0.000	0.095 / 0.074	0.314 / 0.238
LaBraM	CAUEEG	0.000 / 0.000	0.000 / 0.000	0.731 / 0.217	0.208 / 0.168
LaBraM	TUSZ	0.000 / 0.000	0.000 / 0.000	0.463 / 0.380	0.323 / 0.313
CBraMod	TD-BRAIN	0.000 / 0.000	0.000 / 0.000	0.052 / 0.048	0.282 / 0.345
CBraMod	ADFTD	0.000 / 0.000	0.000 / 0.000	0.028 / 0.028	0.558 / 0.505
CBraMod	CAUEEG	0.000 / 0.000	0.000 / 0.000	0.329 / 0.254	0.147 / 0.207
CBraMod	TUSZ	0.000 / 0.000	0.000 / 0.000	0.028 / 0.042	0.271 / 0.361
CSBrain	TD-BRAIN	0.000 / 0.000	0.000 / 0.000	0.043 / 0.047	0.043 / 0.047
CSBrain	ADFTD	0.005 / 0.004	0.018 / 0.018	0.015 / 0.017	0.061 / 0.058
CSBrain	CAUEEG	0.000 / 0.000	0.000 / 0.000	0.041 / 0.066	0.041 / 0.066
CSBrain	TUSZ	0.000 / 0.000	0.000 / 0.000	0.694 / 1.524	0.774 / 0.685
BrainOmni	TD-BRAIN	0.000 / 0.000	0.000 / 0.000	0.064 / 0.047	0.165 / 0.194
BrainOmni	ADFTD	0.008 / 0.009	0.128 / 0.131	0.065 / 0.075	0.037 / 0.046
BrainOmni	CAUEEG	0.000 / 0.000	0.000 / 0.000	0.203 / 0.201	0.263 / 0.183
BrainOmni	TUSZ	0.000 / 0.000	0.000 / 0.000	0.012 / 0.019	0.276 / 0.260

Table 14: Task performance and paired CBraMod readout. Task scores are Macro-F1 / balanced accuracy (%); readout pairs are ridge / MLP. Shuffled E is measured proximally; POST ΔE is adapted minus reference. All entries average three task seeds. Both task metrics improve in every seed. † marks POST increases in all three seeds; the direction of the other POST changes varies across seeds.

	Task scores (%)		Readout	
Task / depth	Reference	Task-only	Shuffled E	POST ΔE
TD-BRAIN PRE	60.00 / 58.65	**62.46 / 61.62**	−0.0639/−0.0182	−0.00237/+0.00191
ADFTD PRE	44.29 / 45.35	**46.66 / 47.75**	−0.0326/−0.0153	+0.0100†/+0.000258
ADFTD MID	43.44 / 43.73	**47.52 / 47.52**	−0.0549/−0.0205	+0.00908†/+0.00501

14 of 16 settings. Residual conditions have SDs below 0.8% in every setting except the FINAL balanced accuracy of CSBrain on TUSZ, at 1.524%.

Within each dataset and fold, the same reference proximal MLP reader is used for all task seeds. The remaining proximal readers are fitted separately for each adaptation state and target condition, with their initialization fixed for each dataset, fold, and task seed.

Internal adaptation improves both task metrics over the matched reference in every seed for all three comparisons. The accompanying readout changes differ between sites and readers. On ADFTD, ridge readout of A changes little at the proximal site and increases at POST in all three seeds for both comparisons. MLP readout at the proximal site rises on TD-BRAIN and falls in both ADFTD comparisons, and at POST it rises slightly on average with mixed seed directions. On TD-BRAIN, the task gain of 2.46%/2.97% occurs with little change in ridge readout of R at either site.

E.3 Target specificity of paired physiological readout

Shuffled-target readers are fitted on the adapted proximal features and scored as in Appendix A.5. After adaptation, correct-target readers give positive proximal E in every comparison under both readers, from 0.0637 to 0.2056 (Table 3). Shuffled-target readers fitted on the same features give negative E in every comparison under both readers, from -0.0639 to -0.0153 (Table 14). The paired readout scores therefore reflect the correspondence between the adapted representations and the measured descriptors.

F Physiological supervision during internal adaptation

Measured descriptors add predictive value beyond POST features (Appendix B) and are readable inside the backbone (Appendix C). This appendix examines how this physiological information can

Table 15: Depths selected separately for each auxiliary objective in folds 0–4.

Task	Objective	0	1	2	3	4
TD-BRAIN	Target family	PRE	PRE	PRE	MID	PRE
TD-BRAIN	Alternative family	PRE	PRE	MID	PRE	MID
TD-BRAIN	Shuffled target	MID	PRE	PRE	MID	PRE
ADFTD	Target family	PRE	PRE	PRE	PRE	PRE
ADFTD	Alternative family	PRE	PRE	MID	PRE	PRE
ADFTD	Shuffled target	PRE	PRE	MID	PRE	PRE

be used during internal adaptation, with descriptor supervision added to the task objective. Task-only adaptation provides the comparison baseline; training follows Appendix A.6.

F.1 Auxiliary objective and calibration

CBraMod is adapted on TD-BRAIN and ADFTD with task-only, target-family, alternative-family, or shuffled-target objectives, whose targets are defined in Appendix F.2. Each condition trains 3,200 residual parameters with the backbone and reference classifier fixed. Optimizer, initialization, batches, and model selection follow Appendix A.6. The task-only baseline also provides the paired readout in Appendix E.1. For a batch of n examples, physiological supervision adds a weighted MSE to Equation 5:

$$\lambda_a \mathcal{L}^{(a)}_{\text{phys}} = \frac{\lambda_a}{32n} \sum_{i=1}^{n} \sum_{j=1}^{32} \left(\left[p_a \left(\text{mean}_{\text{pos}}(\widetilde{\boldsymbol{H}}_i^{(\ell)}) \right) \right]_j - q_{ij}^{a,\text{sup}} \right)^2,$$

where λ_a is the auxiliary loss weight and $p_a : \mathbb{R}^{200} \to \mathbb{R}^{32}$ is a frozen ridge reader, including its fitted input normalization. The operator mean_{pos} averages the channel or token positions and returns a column vector. Here $\widetilde{\boldsymbol{H}}_i^{(\ell)}$ is the adapted hidden state, and $q_{ij}^{a,\text{sup}}$ is coordinate j of the descriptor target for the supervision condition, standardized using the reader's training statistics. Correct-target conditions use the descriptor of window i; shuffled supervision uses the reassigned descriptor specified in Appendix F.2. The superscript sup distinguishes these supervision targets from the readout targets; task and fold indices are omitted.

Supervision readers. A separate ridge reader maps 200 structurally averaged features to 32 descriptor coordinates for each task, fold, and family (R or A). Fitting follows Appendix A.5, using all training windows to estimate the reader and its normalization. Target coordinates are scaled by their training population standard deviations, with a scale of one for constant coordinates. Validation windows are excluded from fitting.

Five-fold out-of-fold readout within the training data selects the reader fitting site: PRE in TD-BRAIN folds 0–3, MID in TD-BRAIN fold 4, and PRE throughout ADFTD. Each fitted reader is shared across task seeds and the PRE and MID adapter candidates. Its coefficients, intercept, and input normalization remain fixed during adaptation, while gradients propagate through the reader to the adapter.

Calibration. The auxiliary weight balances the initial gradient magnitudes of the task and descriptor losses. It is calibrated separately for each depth, objective, fold, and seed using the first training batch before any update (256 windows for TD-BRAIN and 512 for ADFTD). With gradients taken over both adapter matrices, we define

$$G_\ell(\mathcal{L}) = \sqrt{\|\nabla_{\boldsymbol{A}_\ell} \mathcal{L}\|_{\text{F}}^2 + \|\nabla_{\boldsymbol{B}_\ell} \mathcal{L}\|_{\text{F}}^2},$$

$$\lambda_a = \text{clip}_{[0.1,10]} \left(\frac{G_\ell(\mathcal{L}_{\text{CE}})}{G_\ell(\mathcal{L}^{(a)}_{\text{phys}}) + 10^{-12}} \right).$$

Calibration compares mean cross-entropy with the unweighted physiological MSE; the displacement penalty is excluded. The denominator offset prevents division by zero, and clipping limits the auxiliary weight to [0.1,10]. For each auxiliary objective, the resulting λ_a remains fixed throughout training.

Table 16: CBraMod classification with physiological auxiliary supervision under the shared-depth policy. Each entry gives pooled-window Macro-F1 / balanced accuracy (%), averaged across three seeds. All objectives use the depths selected by task-only validation.

Condition	TD-BRAIN	ADFTD
Reference	60.00 / 58.65	44.05 / 44.77
Task-only	**62.46 / 61.62**	**47.07 / 47.40**
Task + target family	61.17 / 60.33	46.04 / 46.20
Task + alternative family	60.39 / 59.12	45.89 / 46.22
Task + shuffled target	59.94 / 58.75	45.92 / 46.07

Table 17: Task-only advantage over physiological auxiliary supervision under shared and objective-specific depth selection. Each entry gives Macro-F1 / balanced accuracy differences (%). Differences are computed from unrounded seed means and rounded independently of the absolute scores in Table 16.

Task	Comparator	Shared depth	Objective-specific depth
TD-BRAIN	Target family	+1.28 / +1.28	+2.19 / +2.31
TD-BRAIN	Alternative family	+2.07 / +2.50	+1.70 / +2.05
TD-BRAIN	Shuffled target	+2.52 / +2.87	+3.01 / +3.28
ADFTD	Target family	+1.03 / +1.20	+1.63 / +1.25
ADFTD	Alternative family	+1.18 / +1.18	+1.97 / +2.01
ADFTD	Shuffled target	+1.15 / +1.33	+1.92 / +1.80

Depth policies. The shared-depth comparison uses the task-only choices in Table 12 for every objective, so all objectives are evaluated at the same depth within each fold. The objective-specific comparison selects PRE or MID separately for each objective from mean validation scores across seeds. Selection prioritizes Macro-F1, then balanced accuracy, with PRE resolving numerical ties within 10^{-12}. Table 15 lists these choices.

F.2 Target correspondence in auxiliary supervision

The three physiological objectives differ in the targets supplied to the frozen reader. Target-family supervision uses the family tracked in the paired analysis, R on TD-BRAIN and A on ADFTD. Alternative-family supervision uses A on TD-BRAIN and R on ADFTD. Shuffled supervision keeps the target-family reader and hierarchically reassigns descriptor targets over subjects, recordings, and windows, using the same procedure for training and validation. When subjects have unequal recording counts or recordings have unequal window counts, reassignment can reuse source windows.

F.3 Downstream performance

The shared-depth comparison measures how the training objective affects performance at matched update locations (Table 16). Allowing each objective to select its own depth tests whether the comparison depends on that location choice (Table 17). Scores pool the five held-out folds within each seed and average seeds 41, 42, and 43.

Target-family supervision improves both metrics over the reference on both tasks. Under the shared-depth policy, it raises Macro-F1/balanced accuracy from 60.00%/58.65% to 61.17%/60.33% on TD-BRAIN and from 44.05%/44.77% to 46.04%/46.20% on ADFTD. It exceeds shuffled-target supervision on both metrics in both tasks and under both depth policies. Against the alternative family, it is higher on both metrics on TD-BRAIN under shared depth and on ADFTD under objective-specific depth, and lower on both metrics on TD-BRAIN under objective-specific depth. On ADFTD under shared depth, it is higher in Macro-F1 and 0.02% lower in balanced accuracy.

Task-only adaptation gives the highest scores. All twelve comparisons with physiological objectives favor task-only training on both metrics, with advantages from 1.03% to 3.28% (Table 17).

Descriptor supervision with the correct correspondence thus retains task value and outperforms shuffled correspondence, and direct task supervision gives the strongest downstream signal. Together

with the predictive-value, readout, and location analyses, these results support using physiological descriptors to locate task-relevant structure in pretrained representations and optimizing the internal residual directly for the downstream task.