PREDICTIVE STRUCTURE IS DISTRIBUTED ACROSS DEPTH IN EEG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate layers often outperform final layers when foundation models are used as frozen feature extractors, but the source of this advantage is unclear. We study this question in EEG foundation models by separating three factors: how representations are read out, whether intermediate layers retain predictive features that complement the final layer, and whether useful depth structure transfers across datasets. Across five models and three tasks, improving the final-layer readout consistently recovers substantial predictive performance, yet intermediate layers remain advantageous in most settings. Residual analyses further show that intermediate representations contain predictive variation that complements final-layer features, including cases where the intermediate layer is weaker on its own. This pattern persists with a nonlinear cross-layer mapping. We also find that layers selected on one dataset remain useful on another, and that a compact set of source-derived task directions retains predictive value after transfer. Together, these results show that intermediate-layer utility cannot be reduced to either poor final-layer pooling or better standalone representations. Instead, predictive structure is distributed across depth, and its usefulness depends on how representations are accessed and combined.

1 INTRODUCTION

Electroencephalography foundation models (EEG-FMs) have recently emerged as a powerful paradigm for neural decoding, leveraging large-scale pre-training to learn rich, transferable brain representations across diverse paradigms and recording configurations. However, when evaluating these frozen encoders for downstream tasks, probes on intermediate layers frequently outperform those on the final layer. This ubiquitous phenomenon prompts a fundamental question: does the final layer inherently lose task-relevant information, or is its capability simply masked by how hidden tokens are aggregated? In the context of EEG FMs, the answer is tightly coupled with the readout mechanism. Standard global averaging over spatial channels and temporal steps deliberately discards structured dynamics, whereas channel- or dispersion-preserving summaries retain them. Consequently, a weak final-layer baseline may be less a limitation of the representation itself than a symptom of a suboptimal readout.

While recent work highlights that layer selection and pooling strategies profoundly impact frozen representation decoding across vision and biosignal domains (Evci et al., 2022; Sirca et al., 2026), downstream evaluations remain acutely sensitive to choices in probing protocols and classification heads (Kontras et al., 2026; Wang et al., 2026; Lahiri et al., 2026). These insights leave two critical questions unresolved: Do intermediate layers still offer distinct representational utility once the final-layer readout is fully optimized? And can these layer-specific gains be reliably transferred across datasets? Furthermore, evaluating layers in isolation overlooks their synergy: intermediate features that appear weaker on their own can still offer crucial, complementary value when paired with the final layer.

We structure our investigation around three hypotheses, as illustrated in Figure 1. First, we hypothesize that readout mismatch accounts for a substantial portion of the final layer's apparent performance gap (**H1**). Second, intermediate layers retain predictive variation that improves standalone performance or complements final-layer features through cross-fitted residual augmentation (**H2**). Third, the informative residual directions learned from one source dataset represent transferable task knowledge

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.

that generalizes to another dataset for the same task (**H3**). We systematically test these hypotheses by isolating readout selection from depth selection, evaluating residual feature dynamics, and validating cross-dataset transfer via subspace intervention.

By evaluating five representative foundation models across motor imagery and two event-related potential (ERP) paradigms, we first reproduce the widespread empirical finding that intermediate layers often outperform final-layer representations. Selecting the readout improves final-layer prediction in every condition, and selecting the layer yields further gains in most conditions. To capture feature utility beyond simple depth rankings, we project intermediate representations onto final-layer features and augment linear probes with cross-fitted residuals across participant groups. This residual augmentation reliably enhances prediction compared to permutation controls. This phenomenon appears even in regimes where intermediate representations exhibit inferior standalone performance.

Finally, source-selected layers and residual directions remain useful in within-task cross-dataset transfer for sleep staging and motor imagery. In sleep staging, removing these source task directions drops target classifier performance significantly more than matched controls, though refitting the classification head recovers most of the accuracy. Together, these results separate readout artifacts from genuine depth utility, show that intermediate layers provide complementary value rather than mere redundancy, and link cross-dataset transfer directly to the task directions utilized by downstream classifiers.

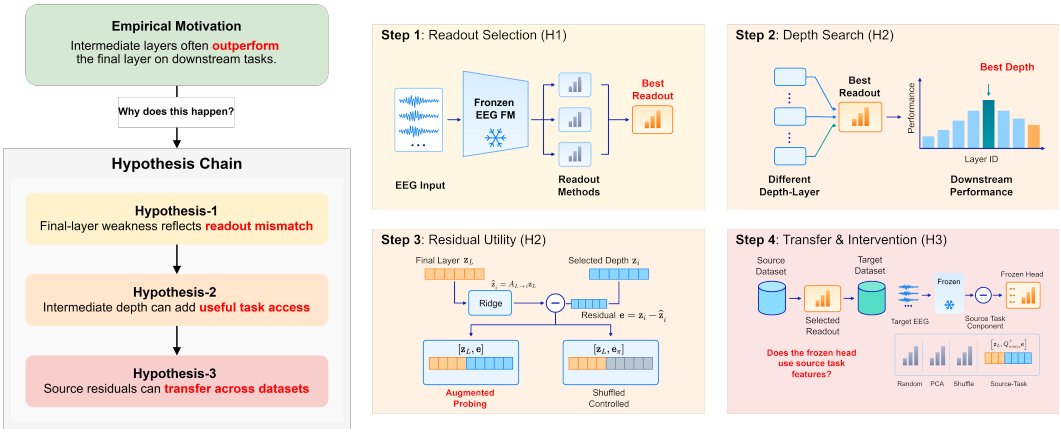


Figure 1: From intermediate-layer performance to transfer and intervention. **Readout selection (H1)**: Evaluate different readout methods on final-layer features to fix output mismatch. **Depth search (H2)**: Identify the optimal layer using development data, and measure its **complementary predictive variation**. **Transfer and intervention (H3)**: Apply selected layers to target datasets and evaluate representation manipulation (e.g., removing source-task components during sleep staging).

2 Related Work

2.1 EEG foundation models and their evalutions

Early work on EEG pretraining laid the foundation with two main approaches: BENDR introduced contrastive self-supervised learning (Kostas et al., 2021), while MAEEG focused on reconstructing masked EEG features (Chien et al., 2022). Later models expanded on these ideas. BIOT introduced support for varying channel setups (Yang et al., 2023), and both EEGPT and EEG2Rep trained on masked signals by using learning objectives directly in the representation space (Wang et al., 2024; Foumani et al., 2024).

The five models we examine differ primarily in how they organize and format EEG features. LaBraM uses spectrum prediction to learn discrete codes across different datasets (Jiang et al., 2024). CBraMod separates spatial and temporal information using a cross-attention mechanism paired with masked patch reconstruction (Wang et al., 2025). REVE combines masked autoencoding with positional embeddings to handle flexible electrode setups and varying signal lengths (Ouahidi et al., 2025).

LUNA uses learned queries to compress channels into a fixed set of hidden tokens, decoupling feature representations from the physical layout of the electrodes (Döner et al., 2025). Finally, CSBrain divides signals across temporal windows and brain regions using structured sparse attention (Zhou et al., 2025). Because these designs preserve different aspects of the input signal, both the choice of intermediate layer and how its outputs are summarized directly affect what information reaches the downstream classifier.

Standard evaluations usually test these models either by training a simple linear classifier on top of frozen features or by fine-tuning the whole network on a new task. Recent benchmarks have standardized probing on frozen encoders (Kontras et al., 2026) and compared how models handle different adaptation strategies and signal perturbations (Wang et al., 2026). Classifier architecture choice also introduces variability, though most prior comparisons focus on full fine-tuning scenarios (Lahiri et al., 2026). Instead of adapting the entire network, we keep the encoder fixed and focus on how feature extraction varies across both layer depth and readout design, benchmarked against the standard final-layer baseline.

2.2 Extracting information from intermediate layers

It is well established that feature usefulness varies with model depth and depends heavily on how closely the pretraining task aligns with the downstream target (Yosinski et al., 2014). Training independent linear classifiers at each layer offers a straightforward way to trace how class separability evolves across a model (Alain & Bengio, 2016). For EEG models specifically, preserving the underlying token layout often improves downstream accuracy: simply flattening tokens frequently works better than averaging them, and intermediate blocks can contain rich, useful representations (Sirca et al., 2026). Combining features from multiple layers can similarly boost performance on new tasks (Evci et al., 2022), a trend also observed in time-series and audio models (Feofanov et al., 2026; Rauch et al., 2026). These insights highlight why we need to evaluate not just which layer to extract features from, but also how those features are summarized. Building on this, we isolate what intermediate layers provide beyond the final layer, comparing this added value against randomized feature controls of identical size.

Separating a practical feature summary from a genuinely rich representation is also a key challenge in language modeling. For instance, pooling attention across layers can improve sentence embeddings (Oh et al., 2022). Conversely, certain dominant feature dimensions can artificially skew similarity metrics without reflecting actual model behavior (Timkey & van Schijndel, 2021), often driven by frequency artifacts in the inputs (Puccetti et al., 2022; Baggetto & Fresno-Fernández, 2022). These observations demonstrate that conclusions about representation geometry or feature quality are highly sensitive to the choice of readout.

2.3 Evaluating feature utility and representations

A classifier's accuracy reflects not only the quality of the underlying representation, but also the classifier's own capacity to learn. Prior studies use synthetic control tasks or measure the learning effort required to extract information to avoid overestimating representation quality (Hewitt & Liang, 2019; Voita & Titov, 2020). Merely being able to decode a feature does not mean the model actually relies on it to make predictions (Belinkov, 2022). To test whether features are functionally important, we can selectively remove specific information from the representation and measure the impact on downstream task performance (Elazar et al., 2021; Ravfogel et al., 2020; Belrose et al., 2023). In EEG research, feature removal has typically focused on explicit, predefined physiological properties (Tang et al., 2026). Here, we adopt a different approach: we isolate the unique components in intermediate layers—beyond those captured by final-layer predictions—and test how modifying them impacts target task performance.

3 Methods

3.1 Representations and Readout Families

Let $H_\ell(x)$ denote the hidden representations at layer $\ell \in \{1, \ldots, L\}$ of a frozen foundation model encoder given an EEG input x, where L represents the final layer. Downstream classifiers access

these representations through a readout function $r(\cdot)$, yielding a fixed-dimensional feature vector $z_{\ell,r}(x) = r(H_\ell(x))$.

To evaluate structural feature retention, we define a primary family of fixed readouts $\mathcal{R}$ for channel-time tensors. This includes global mean pooling, global mean plus temporal standard deviation, channel-wise mean pooling, channel-wise mean with temporal standard deviation, and temporal binning. For architectures lacking a physical channel dimension (e.g., LUNA), applicable global and temporal summaries are used. Across all experiments, the foundation model encoder remains strictly frozen. Linear probes are trained on fixed feature summaries, whereas learned attention readouts fit query parameters on downstream training data. Detailed layer-to-readout mappings are provided in Appendix A.

3.2 Disentangling Readout Mismatch from Depth Utility

To isolate the predictive utility of depth from readout choice, we employ a sequential selection protocol on development data. Let $\widehat{P}_{\mathrm{dev}}(\ell, r)$ denote the validation performance under candidate-specific regularization. Within each outer fold, we first select the optimal final-layer readout r_L, and subsequently identify the optimal depth ℓ_{depth} under that chosen readout:

$$r_L = \arg\max_{r \in \mathcal{R}} \widehat{P}_{\mathrm{dev}}(L, r), \qquad \ell_{\mathrm{depth}} = \arg\max_{\ell} \widehat{P}_{\mathrm{dev}}(\ell, r_L). \tag{1}$$

We also determine the globally optimal depth-readout pair $(\ell_*, r_*) = \arg\max_{(\ell,r)} \widehat{P}_{\mathrm{dev}}(\ell, r)$. Evaluating these fitted candidates on held-out participants yields three core performance contrasts:

$$\Delta_{\mathrm{readout}} = P_{\mathrm{test}}(L, r_L) - P_{\mathrm{test}}(L, r_{\mathrm{mean}}), \tag{2}$$

$$\Delta_{\mathrm{depth}} = P_{\mathrm{test}}(\ell_{\mathrm{depth}}, r_L) - P_{\mathrm{test}}(L, r_L), \tag{3}$$

$$\Delta_{\mathrm{joint}} = P_{\mathrm{test}}(\ell_*, r_*) - P_{\mathrm{test}}(L, r_L). \tag{4}$$

Here, $\Delta_{\mathrm{readout}}$ quantifies the performance restored to the final layer by replacing naive global mean pooling (r_{mean}) with an optimized readout. Our primary metric, Δ_{depth}, isolates intrinsic depth utility by varying layer depth while holding the final-layer readout r_L fixed. Feature standardization, regularization parameters, and probe heads are fitted independently at each layer to account for layerwise distributional shifts.

3.3 Quantifying Complementary Feature Utility via Residuals

We test whether intermediate features contain predictive variation not recovered by a fitted mapping from final-layer features in the PCA200 representation.

Within each outer fold, intermediate features z_m (for $m = \ell_{\mathrm{depth}}$) and final-layer features z_L are standardized using statistics estimated from the outer-training split and independently reduced to 200 dimensions by Principal Component Analysis (PCA). PCA is fitted only on the outer-training data and then applied unchanged to held-out participants, yielding $\tilde{z}_m, \tilde{z}_L \in \mathbb{R}^{200}$. We fit a linear mapping $\widehat{A}$ from final to intermediate features using ridge regression, selecting regularization hyperparameters via inner cross-validation. To prevent data leakage during probe training, we compute cross-fitted residuals across participant folds:

$$e^{\mathrm{cf}}_{m|L} = \tilde{z}_m - \widehat{A}_{-k}\tilde{z}_L, \tag{5}$$

where $\widehat{A}_{-k}$ is estimated excluding participant fold k. Test residuals are computed using the full outer-training mapping $\widehat{A}_{\mathrm{tr}}$.

We then augment the final-layer probe with these residual features, comparing its performance against a trial-shuffled residual control:

$$E_{\mathrm{resid}} = P\left([\tilde{z}_L, e^{\mathrm{cf}}_{m|L}]\right) - \frac{1}{S}\sum_{s=1}^{S} P\left([\tilde{z}_L, \Pi_s e^{\mathrm{cf}}_{m|L}]\right), \tag{6}$$

where Π_s denotes a trialwise permutation operator applied within each participant across $S = 20$ seeds. Because both augmented probes operate on identical feature dimensions, E_{resid} isolates the specific predictive value provided by the trialwise alignment of intermediate residuals.

3.4 Cross-Dataset Transfer and Subspace Interventions

To evaluate the transferability of layer depth and residual dynamics, we freeze the source-selected configuration (ℓ_s, r_s) and evaluate it on target datasets, re-fitting only downstream classification heads and standardizations.

To measure whether source residual dynamics represent transferable task structure, we construct a task-relevant subspace from source residuals and remove these directions from target representations. Let Q_s denote an orthonormal basis capturing the mean residual directions across task classes in the source dataset. On the target domain, we remove these source task directions from intermediate features z_m:

$$z_m^{(-Q)} = z_m - \gamma Q_s Q_s^\top (z_m - \mu_{\text{target}}), \tag{7}$$

where μ_{target} is the target training mean and $\gamma = 1$ corresponds to complete subspace removal. We evaluate performance drops relative to rank-matched controls (random orthogonal subspaces, residual PCA components, and label-shuffled controls) with matched displacement energy after target standardization (Appendix D.3). Finally, we measure functional recovery by re-fitting downstream classification heads on the modified representations $z_m^{(-Q)}$.

3.5 Experimental Setup and Evaluation Metrics

We evaluate five representative EEG foundation models (CBraMod, REVE, LaBraM, LUNA, and CSBrain) on motor imagery (OpenBMI; Lee et al., 2019), two ERP tasks (ERP CORE N170 and N400; Kappenman et al., 2020), and emotion recognition (SEED-V; Liu et al., 2022). Within-task cross-dataset transfer uses HMC (Alvarez-Estevez & Rijsman, 2020) to ISRUC-1 (Khalighi et al., 2016) for sleep staging, and OpenBMI to BCI Competition IV-2a (Leeb & Brunner, 2008) for motor imagery.

Evaluation strictly follows participant-disjoint outer folds to prevent selection bias (Varma & Simon, 2006; Cawley & Talbot, 2010). Performance is measured using participant-balanced accuracy (BA) for motor imagery and SEED-V, participant-balanced AUROC for ERP paradigms, and macro-F1 for sleep staging. Statistical significance is assessed via paired 95% bootstrap confidence intervals resampled at the participant level across all conditions. All implementation and pre-processing details are provided in Appendices A–B.

4 Results

4.1 Depth utility after final-layer readout selection

We first separate readout mismatch from depth effects (H1). Selecting the readout at the final layer improves prediction over global mean pooling in all 15 model–task conditions, with positive paired confidence intervals (Figure 2A; Table S2). The effect is particularly large on OpenBMI for CBraMod and LaBraM, showing that weak mean-pooled final-layer performance can substantially underestimate the information available at the final layer.

After fixing this final-layer-selected readout, choosing the layer provides an additional improvement in 12 of 15 conditions (Figure 2B). The size of this depth benefit varies considerably across models and tasks. In particular, models with similarly weak mean-pooled final-layer performance can show very different residual gains from moving to an intermediate layer. Readout mismatch and depth utility therefore capture distinct effects.

Joint layer-readout selection chooses the same readout as final-layer selection in every condition, giving $\Delta_{\text{joint}} = \Delta_{\text{depth}}$. The selected readout is also stable across all five outer folds within each condition, whereas the preferred layer can vary (Table S5). Several conditions show little additional depth gain, indicating that intermediate layers are not uniformly preferable once readout mismatch has been accounted for.

The same pattern holds in the four conditions evaluated with learned readouts, where depth gains retain positive confidence intervals (Table S3). Reduced-label and SEED-V results appear in Tables S4 and S2. Full layer-wise curves and test-selected oracle peaks appear in Figure S1 and Table S17.

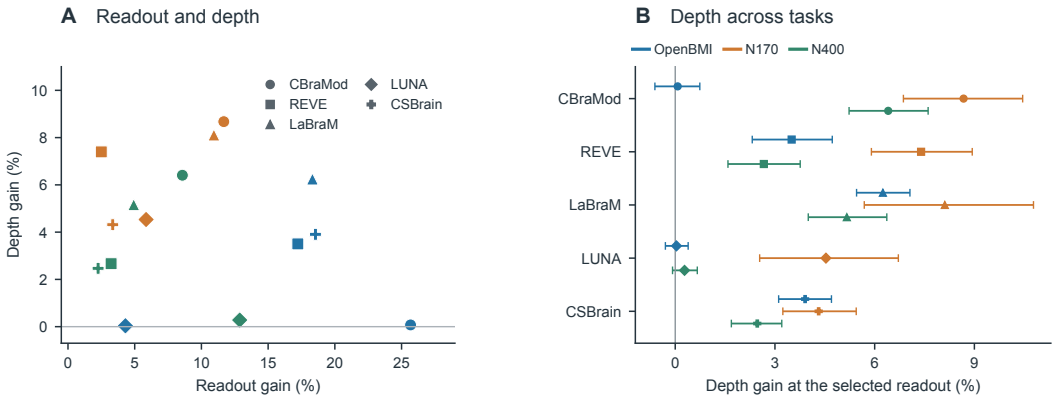


Figure 2: Readout and depth gains across five models and three tasks. **A:** Final-layer readout gain versus additional depth gain, with one point per model–task condition. **B:** Depth gain using the final-layer-selected readout, with paired 95% bootstrap intervals. Colors identify tasks and symbols identify models. Tables S1–S2 report the corresponding scores and contrasts.

4.2 Intermediate features add predictive value beyond the final layer

H2 asks whether intermediate representations contain predictive variation that is not already captured by the final layer. In the PCA200 analysis, cross-fitted residual augmentation improves final-layer prediction in 13 of 15 conditions. Relative to shuffled residuals matched in distribution, covariance, and scale, the residual contribution E_{resid} also has a positive paired confidence interval in 13 of 15 conditions (Figure 3A; Table S9). Thus, preserving the correspondence between intermediate residuals and individual trials matters beyond simply adding features with similar marginal statistics.

The effect is especially clear for CBraMod/N170, where intermediate residuals provide a large augmentation benefit while shuffled residuals do not. More generally, residual-only decoding remains above chance in all 15 conditions (Figure 3B). These two analyses answer different questions: residual-only decoding measures information remaining outside the final representation, whereas augmentation tests whether that information improves prediction when the final representation is already available.

Importantly, standalone intermediate performance is not required for complementarity. CBraMod/OpenBMI still benefits from residual augmentation despite comparatively weak intermediate-only performance under the same pipeline. Conversely, LUNA/N400 shows no useful augmentation effect, and the evidence for LUNA/OpenBMI is inconclusive. Complementary information is therefore common but not universal across model–task pairs.

Under the same balanced probe, cross-fitting changes augmentation scores by at most 0.40%; differences from direct concatenation are at most 1.36% (Table S10). Additional in-sample results appear in Table S11.

The residual effect also persists beyond the linear mapping. Replacing ridge regression with a two-layer nonlinear mapper reduces residual decoding in several conditions, showing that part of the intermediate representation can be recovered through a more flexible transformation of the final features. Nevertheless, residual decoding remains above chance in all 15 conditions, indicating that label-predictive variation remains after nonlinear reconstruction (Table S13).

The forward–reverse analysis shows a similar pattern. Under the corresponding in-sample protocol, forward residual utility exceeds reverse utility in 12 of 15 conditions (Table S12), with the larger differences generally occurring where intermediate layers also show stronger depth gains. The forward-reverse difference broadly follows the layerwise depth contrast, providing a complementary view of how predictive information is distributed across representations. Random-projection controls retain positive asymmetry in five of six conditions, while the four unreduced ERP conditions have negative augmentation gains (Table S14).

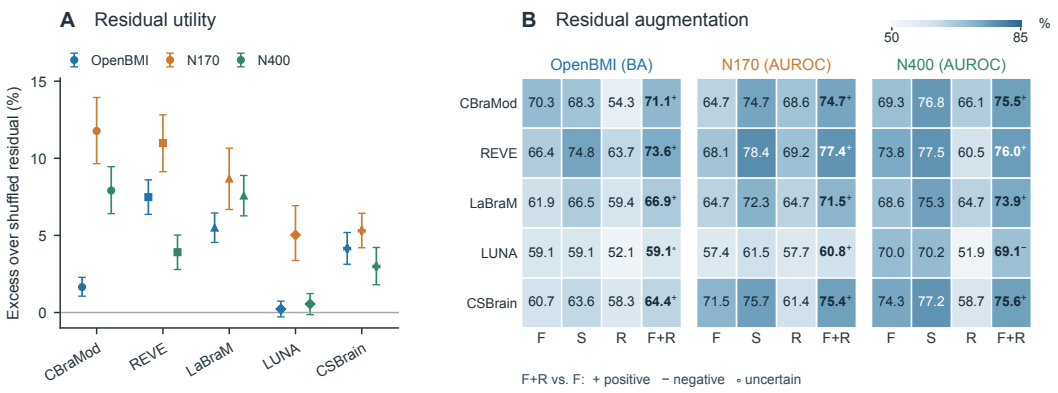


Figure 3: Residual prediction across five models and three tasks. **A:** Cross-fitted residual augmentation relative to shuffled-residual controls, with paired 95% participant-bootstrap intervals. **B:** Performance from final (F), selected intermediate (S), residual (R), and combined final-plus-residual (F+R) representations. Superscripts on F+R indicate whether its paired gain over F is positive, negative, or inconclusive.

4.3 Reusing source-selected layers across datasets

We next test whether depth choices selected on one dataset remain useful on another (H3). Across both HMC-to-ISRUC sleep staging and OpenBMI-to-BCIC-IV-2a motor imagery, the source-selected intermediate layer outperforms the final layer on the target dataset (Table 1), with positive paired bootstrap intervals in all four model–transfer conditions.

Table 1: Interface reuse across two task families. Scores and gains are in percent with paired 95% bootstrap intervals. Source interfaces are fixed and target probes refit; MI sign-flip tests are in Appendix C.

Target strategy / contrast	CBraMod	REVE
Sleep: HMC → ISRUC ($n = 98$, macro-F1)		
Final layer / global mean	67.35	70.59
Final layer / source readout	70.70	71.66
Source-selected layer / same readout	71.73	72.79
Depth gain at fixed readout	**1.03 [0.47, 1.61]**	**1.13 [0.49, 1.80]**
Motor imagery: OpenBMI → BCIC-IV-2a ($n = 9$, BA)		
Final layer / global mean	54.63	63.27
Final layer / source readout	66.98	77.55
Source-selected layer / same readout	71.99	79.82
Depth gain at fixed readout	**5.02 [3.40, 6.40]**	**2.28 [0.42, 4.24]**
Exact p	0.0078	0.0781

Depth and readout are selected only on the source dataset and then kept fixed during transfer; only the downstream probe is refit on target data. The same source-selected readout is used for both intermediate- and final-layer features. The transferred depth advantage therefore reflects a reusable layer choice rather than target-side layer selection.

The pattern is observed in both transfer settings, with larger gains in some model–task pairs than others. In the HMC-to-ISRUC experiments, improving the final-layer readout already raises target performance, and the source-selected intermediate layer provides a further gain under that same readout.

Participant-level effects for motor imagery are more consistent for CBraMod than for REVE. Tables S6–S7 report auxiliary metrics, participant effects, and an exploratory MI frozen-head removal experiment with larger drops for task than random directions.

4.4 SOURCE RESIDUAL DIRECTIONS REMAIN USEFUL AFTER TRANSFER

We next examine which intermediate features contribute to the transferred advantage. Source residuals are used to define a low-dimensional task subspace, which is then applied to target representations. Removing these directions with the target scaler and classifier frozen produces a pronounced performance drop for both CBraMod and REVE (Figure 4A; Tables S15–S16). The drop is consistently larger than for random, residual-PCA, and shuffled-label control directions.

When the scaler and classifier are refit after removal, performance largely recovers and the gap to the matched controls becomes much smaller. The source-derived directions therefore account for features used strongly by the transferred frozen head, while alternative predictive structure remains available after refitting.

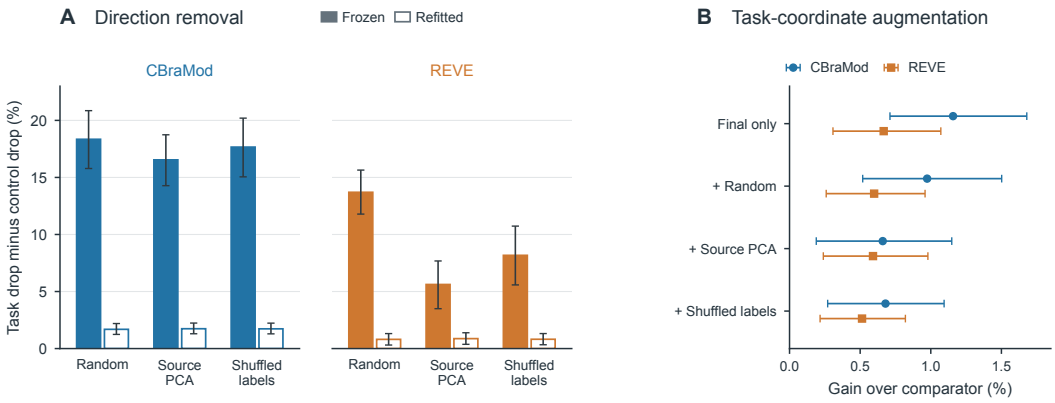


Figure 4: Source residual interventions under HMC-to-ISRUC transfer. **A:** Additional performance drop from removing source task directions relative to matched controls, with the target head frozen or refit. **B:** Gain from augmenting final-layer features with source task coordinates relative to final-only prediction and matched controls. Intervals are paired 95% whole-participant bootstrap intervals.

The augmentation experiment gives a complementary result. Adding four source-derived task coordinates to final-layer features improves target prediction for both models and exceeds the corresponding control augmentations (Figure 4B). For CBraMod, these coordinates recover approximately the selected-layer performance; for REVE, they recover part of the difference.

A controlled rotation illustrates when source directions cease to help. Rotating the target signal away from the source task direction reduces source-interface BA from 87.55% to 72.45%, near the mean-only baseline, while full-representation BA stays near 87.2–87.3%. Thus predictive information can remain available even as the benefit of reusing a source direction disappears (Appendix C.2, Table S8). These results link the reusable depth advantage to a compact set of source-derived predictive directions that remain informative after transfer.

5 DISCUSSION

Our results suggest that the apparent weakness of final-layer EEG representations has more than one source. In some cases, much of the gap disappears once the representation is read out appropriately. At the same time, intermediate layers retain predictive variation that remains useful even after the final-layer readout has been improved. These effects are therefore better viewed as two separate properties of a representation: how accessible its information is through a particular readout, and how that information changes across depth.

This distinction changes how frozen foundation models should be evaluated. A single pooling rule can make differences between layers partly reflect the readout rather than the representation itself, so layer comparisons are most informative once the readout has been established. Under this setting, intermediate layers still improve prediction in most conditions. Their value also extends beyond selecting a better standalone layer: residual features from intermediate representations can improve prediction when combined with final-layer features. This suggests that depth can preserve

complementary predictive structure even when an intermediate layer is not the strongest representation on its own. Evaluating frozen models therefore requires considering both layer-wise performance and how information from different depths can be combined.

The residual experiments provide a more direct view of this complementarity. Intermediate residuals remain predictive after subtracting the fitted prediction from final-layer PCA200 features, and combining them with final-layer features improves prediction across most conditions. A nonlinear reconstruction reduces this residual signal in several cases but does not remove it. This pattern is consistent with depth progressively reorganizing task-relevant structure rather than simply preserving the same information in increasingly better form. Final layers can make some factors easier to access while attenuating or transforming others that remain available at earlier depths.

The transfer experiments further show that useful depth structure is not confined to the dataset on which it was selected. Layers chosen on a source dataset remain advantageous after transfer to a second dataset for both sleep staging and motor imagery. The intervention results sharpen this observation: a small set of task-relevant directions derived from source residuals remains useful on the target representation. Removing these directions strongly affects a frozen target classifier, whereas adding the corresponding coordinates to final-layer features improves target prediction. This suggests that part of the intermediate-layer advantage is carried by compact predictive structure that is reused across datasets rather than by a purely dataset-specific preference for a particular layer.

These findings argue against treating the final hidden state as the canonical representation of an EEG foundation model. Depth, readout, and downstream adaptation jointly determine which predictive structure becomes accessible. For frozen-model evaluation, this motivates reporting not only final-layer performance but also the interface used to summarize that layer and the extent to which earlier representations contribute complementary information. For model design, it also suggests a practical alternative to replacing the final layer altogether: intermediate features can be exposed through lightweight auxiliary coordinates or multi-depth readouts when they provide complementary signal.

Limitations. Our analyses cover several complementary ways of probing depth-dependent representations, including fixed and learned readouts, linear and nonlinear cross-layer mappings, and cross-dataset transfer. The current experiments use lightweight mappings and focus the intervention analysis on a subset of the transfer settings. Future work could extend this framework to more expressive cross-layer transformations, broader transfer settings, and models with explicit multi-depth objectives. These extensions would help determine how consistently complementary predictive structure emerges across architectures, tasks, and training objectives.

6 Conclusion

We show that intermediate-layer advantages in frozen EEG foundation models arise from more than a single source. Better readouts recover substantial performance at the final layer, yet intermediate representations still provide additional predictive value, both as standalone features and as complements to final-layer representations. These depth-dependent features can also remain useful across datasets, with source-derived predictive directions retaining value after transfer. Together, the results suggest that predictive structure is distributed across model depth rather than concentrated in a single canonical representation. Evaluating frozen foundation models should therefore consider not only which layer performs best, but also how representations are read out, combined, and reused across datasets within a task.

AI Use Statement

Generative AI tools were used to assist with language polishing and the generation of portions of the experimental code. All AI-assisted code was manually reviewed, verified, and, where necessary, modified by the authors before use. The authors take full responsibility for the manuscript, implementation, and reported results.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2016. URL https://openreview.net/forum?id=ryF7rTqgl.

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 2020. doi: 10.1371/journal.pone.0256111.

Alejandro Fuster Baggetto and Víctor Fresno-Fernández. Is anisotropy really the cause of BERT embeddings not being semantic? In *Conference on Empirical Methods in Natural Language Processing*, pp. 4271–4281, 2022. doi: 10.18653/v1/2022.findings-emnlp.314. URL https://aclanthology.org/2022.findings-emnlp.314/.

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

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36*, volume 36, pp. 66044–66063. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2884. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. doi: 10.5555/1756006.1859921. URL https://jmlr.org/papers/v11/cawley10a.html.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked auto-encoder for EEG representation learning. In *NeurIPS Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=kttuLV59ZuJ.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. *Advances in Neural Information Processing Systems 38*, pp. 78772–78798, 2025. doi: 10.52202/085713-2377. URL https://arxiv.org/abs/2510.22257.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C. Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6009–6033, 2022. URL https://proceedings.mlr.press/v162/evci22a.html.

Vasilii Feofanov, Songkang Wen, Jianfeng Zhang, Lujia Pan, and Ievgen Redko. MantisV2: Closing the zero-shot gap in time series classification with synthetic data and test-time strategies. *arXiv.org*, 2026. doi: 10.48550/arxiv.2602.17868. URL https://arxiv.org/abs/2602.17868.

Navid Mohammadi Foumani, Geoff Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Knowledge Discovery and Data Mining*, pp. 5544–5555. ACM, 2024. doi: 10.1145/3637528.3671600. URL https://arxiv.org/abs/2402.17772.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2405.18765. URL https://openreview.net/forum?id=0zTpTRVtrP.

Emily S. Kappenman, Jaclyn L. Farrens, Wendy Zhang, Andrew X. Stewart, and Steven J. Luck. ERP CORE: An open resource for human event-related potential research. *NeuroImage*, 225:117465, 2020. doi: 10.31234/osf.io/4azqm.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. doi: 10.1016/j.cmpb.2015.10.013.

Konstantinos Kontras, Trui Osselaer, Stylianos G. Mouslech, Angeliki-Ilektra Karaiskou, Guido Gagliardi, Thomas Strypsteen, Mohammad Hossein Badiei, Anku Rani, Maarten Vanmarcke, Miguel Bhagubai, et al. NeuroAtlas: Benchmarking foundation models for clinical EEG and brain-computer interfaces. *arXiv (Cornell University)*, 2026. doi: 10.48550/arXiv.2605.14698. URL https://arxiv.org/abs/2605.14698.

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

Jeet Bandhu Lahiri, Parshva Runwal, Arvasu Kulkarni, Mahir Jain, Aditya Ray Mishra, Siddharth Panwar, and Sandeep Singh. PRISM: Exploring heterogeneous pretrained EEG foundation model transfer to clinical differential diagnosis. *arXiv (Cornell University)*, 2026. doi: 10.48550/arXiv.2603.02268. URL https://arxiv.org/abs/2603.02268.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. doi: 10.1093/gigascience/giz002.

R. Leeb and Clemens Brunner. BCI Competition 2008—Graz data set A. Official BCI Competition IV dataset 2a description, 2008. URL https://www.bbci.de/competition/iv/desc_2a.pdf.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Dongsuk Oh, Yejin Kim, Hodong Lee, H. Howie Huang, and Heuiseok Lim. Don't judge a language model by its last layer: Contrastive learning with layer-wise attention pooling. In *International Conference on Computational Linguistics*, pp. 4585–4592, 2022. doi: 10.48550/arXiv.2209.05972. URL https://aclanthology.org/2022.coling-1.405/.

Y. El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, B. Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems 38*, pp. 25583–25619, 2025. doi: 10.52202/085713-0760. URL https://arxiv.org/abs/2510.21585.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and F. Dell'orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1304. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.93. URL https://aclanthology.org/2022.findings-emnlp.93/.

Lukas Rauch, René Heinrich, Houtan Ghaffari, Lukas Miklautz, Ilyass Moummad, Bernhard Sick, and Christoph Scholz. Unmute the patch tokens: Rethinking probing in multi-label audio classification. In *arXiv (Cornell University)*. Cornell University, 2026. doi: 10.48550/arxiv.2509.24901. URL https://arxiv.org/abs/2509.24901.

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

Urban Sirca, Maryam Alimardani, Stefanos P. Zafeiriou, and Konstantinos Barmpas. Beyond accuracy: Robustness, interpretability and expressiveness of EEG foundation models. *arXiv.org*, 2026. doi: 10.48550/arXiv.2605.17562. URL https://arxiv.org/abs/2605.17562.

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? *arXiv.org*, 2026. doi: 10.48550/arXiv.2605.11410. URL https://arxiv.org/abs/2605.11410.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006. doi: 10.1186/1471-2105-7-91. URL https://link.springer.com/article/10.1186/1471-2105-7-91.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Conference on Empirical Methods in Natural Language Processing*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

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 *Neural Information Processing Systems*, volume 37, pp. 39249–39280. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

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*, 2025. URL https://arxiv.org/abs/2412.07236.

Xianheng Wang, Yige Yang, and Damien Coyle. EEG-FM-Audit: A systematic evaluation and analysis pipeline for EEG foundation models. *arXiv.org*, 2026. doi: 10.48550/arXiv.2605.26910. URL https://arxiv.org/abs/2605.26910.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Neural Information Processing Systems*, volume 36, pp. 78240–78260. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-3420. URL https://papers.nips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Neural Information Processing Systems*, volume 27, 2014. URL https://arxiv.org/abs/1411.1792.

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. *Advances in Neural Information Processing Systems 38*, pp. 96859–96904, 2025. doi: 10.52202/085713-2915. URL https://arxiv.org/abs/2506.23075.

A Experimental details

A.1 Datasets, readouts, and preprocessing

All data splits are participant-disjoint, and all sessions from the same participant remain in the same split. ERP CORE N170 and N400 share the same participant cohort. SEED-V probabilities are averaged within video before participant-equal scoring. For HMC-to-ISRUC transfer, we use the 103-participant HMC development subset as the source and 98 ISRUC participants as the target after excluding S008 and S040 for channel/reference incompatibility.

Readouts operate directly on frozen encoder tokens. For $H \in \mathbb{R}^{C\times T\times d}$, global mean averages channels and time, while global mean+std additionally summarizes token dispersion. Channel mean averages over time while preserving channel identity, and channel mean+std additionally retains temporal dispersion within each channel. The structure readout concatenates channel summaries from four fixed temporal bins when the tokenization permits this. LUNA has no physical channel axis and therefore uses the applicable global and temporal summaries. Each readout is standardized using training-fold statistics before fitting the downstream classifier.

The sleep interface uses four mastoid-referenced leads. Each 30-s epoch is represented by six 5-s encoder outputs, with temporal dispersion computed within segment before segment averaging.

A.2 Model checkpoints and evaluation protocol

We evaluate CBraMod, REVE-base, LaBraM-base, LUNA-base, and CSBrain with frozen encoders throughout. REVE's released pretraining inventory includes OpenBMI.

All layer–readout candidates use the same outer and development folds. Regularization is selected on development data for each candidate, and standardization and dimensionality reduction are fitted only on the corresponding training data. Participant-level contrasts use paired participant bootstrap intervals. ISRUC macro-F1 is computed from pooled participant confusion counts within each bootstrap resample.

B Additional readout and depth results

B.1 Complete decomposition and learned readouts

Tables S1 and S2 provide the full scores and contrasts underlying Figure 2, including SEED-V.

Table S1: Absolute scores underlying Figure 2: $100\times$ BA (OpenBMI) or AUROC (N170/N400). Depth selection retains the final-layer-selected readout. Paired effect estimates and intervals are reported separately in Table S2.

Task	Model	Final mean	Final r_L	Depth / r_L
OpenBMI	CBraMod	50.42	76.09	76.17
OpenBMI	REVE	59.73	76.95	80.46
OpenBMI	LaBraM	51.42	69.74	75.99
OpenBMI	LUNA	55.37	59.67	59.71
OpenBMI	CSBrain	52.65	71.20	75.10
N170	CBraMod	57.10	68.77	77.45
N170	REVE	70.91	73.41	80.81
N170	LaBraM	56.07	67.01	75.12
N170	LUNA	52.40	58.25	62.79
N170	CSBrain	71.42	74.78	79.09
N400	CBraMod	62.61	71.18	77.59
N400	REVE	72.58	75.81	78.48
N400	LaBraM	65.95	70.87	76.04
N400	LUNA	58.70	71.57	71.85
N400	CSBrain	73.26	75.52	77.99

Table S2: Complete primary effects (percent, paired 95% intervals). Readout correction, fixed-readout depth utility, and joint interface utility remain separate quantities.

Task	Model	Δ_{readout}	Δ_{depth}	Δ_{joint}
OpenBMI	CBraMod	25.67 [23.19, 28.17]	0.07 [-0.61, 0.74]	0.07 [-0.61, 0.74]
OpenBMI	REVE	17.23 [15.54, 18.99]	3.50 [2.32, 4.72]	3.50 [2.32, 4.72]
OpenBMI	LaBraM	18.32 [16.07, 20.52]	6.25 [5.46, 7.06]	6.25 [5.46, 7.06]
OpenBMI	LUNA	4.30 [2.72, 5.93]	0.04 [-0.30, 0.39]	0.04 [-0.30, 0.39]
OpenBMI	CSBrain	18.55 [16.40, 20.80]	3.91 [3.11, 4.70]	3.91 [3.11, 4.70]
N170	CBraMod	11.68 [9.28, 14.07]	8.67 [6.86, 10.45]	8.67 [6.86, 10.45]
N170	REVE	2.50 [1.10, 3.95]	7.40 [5.90, 8.94]	7.40 [5.90, 8.94]
N170	LaBraM	10.94 [8.35, 13.48]	8.11 [5.69, 10.78]	8.11 [5.69, 10.78]
N170	LUNA	5.85 [4.00, 7.82]	4.53 [2.54, 6.71]	4.53 [2.54, 6.71]
N170	CSBrain	3.35 [1.61, 5.11]	4.32 [3.24, 5.45]	4.32 [3.24, 5.45]
N400	CBraMod	8.57 [6.69, 10.39]	6.41 [5.23, 7.61]	6.41 [5.23, 7.61]
N400	REVE	3.23 [2.10, 4.36]	2.67 [1.59, 3.76]	2.67 [1.59, 3.76]
N400	LaBraM	4.93 [3.52, 6.34]	5.16 [4.00, 6.37]	5.16 [4.00, 6.37]
N400	LUNA	12.87 [10.87, 14.93]	0.28 [-0.08, 0.66]	0.28 [-0.08, 0.66]
N400	CSBrain	2.26 [1.23, 3.30]	2.47 [1.69, 3.21]	2.47 [1.69, 3.21]
SEED-V	CBraMod	5.00 [1.11, 8.75]	0.28 [-2.78, 3.61]	1.53 [-1.11, 4.44]
SEED-V	REVE	-1.67 [-5.69, 2.08]	0.69 [-1.94, 3.19]	1.53 [-3.61, 6.81]
SEED-V	LaBraM	3.47 [0.28, 6.67]	-2.50 [-5.97, 0.83]	-2.36 [-6.11, 1.53]
SEED-V	LUNA	1.25 [-1.11, 3.61]	-0.56 [-4.31, 2.92]	-0.28 [-4.31, 3.47]
SEED-V	CSBrain	0.28 [-2.22, 3.06]	1.11 [-1.53, 4.03]	-0.69 [-2.50, 1.11]

The learned-readout controls use the same frozen token representations and participant-disjoint folds. A global attention readout uses one learned query over all tokens, while the channel-preserving variant applies a shared temporal query within each channel and concatenates the resulting channel features. Readout parameters are fitted separately at each layer. Table S3 reports the corresponding final-layer and selected-layer results.

Table S3: Learned-family controls. Absolute scores are percentages; selected-minus-final gains are expressed in percent with paired 95% intervals. Seeds are averaged within participant before bootstrap.

Family	Model	Task	Final	Selected	Gain [95% interval]
Global query	CBraMod	N170	62.96	72.16	9.20 [6.14, 12.36]
Global query	CBraMod	OPENBMI	56.61	60.59	3.98 [2.55, 5.49]
Global query	REVE	N170	69.74	73.92	4.18 [1.86, 6.22]
Global query	REVE	OPENBMI	65.91	70.88	4.97 [3.69, 6.40]
Channel-preserving	CBraMod	N170	68.38	77.15	8.77 [7.05, 10.48]
Channel-preserving	CBraMod	OPENBMI	71.32	74.90	3.57 [2.58, 4.52]
Channel-preserving	REVE	N170	72.64	80.50	7.86 [6.22, 9.56]
Channel-preserving	REVE	OPENBMI	77.05	79.86	2.81 [1.48, 4.17]

B.2 Supervision and selection stability

The reduced-label analysis uses 10%, 25%, 50%, and 100% of the outer-training data, with stratified subsampling by participant and class while keeping validation and test participants fixed. Table S4 reports the resulting depth contrasts.

Table S4: Full minus 10% fitting-label effects (percent, paired intervals). Validation pools stay fixed. The second quantity is joint interface utility.

Task	Model	Change in Δ_{readout}	Change in Δ_{joint}
OPENBMI	CBraMod	8.82 [7.46, 10.14]	0.38 [-0.30, 1.08]
OPENBMI	LaBraM	8.77 [7.49, 10.06]	1.27 [0.30, 2.27]
OPENBMI	CSBrain	8.43 [7.00, 9.93]	1.91 [0.85, 2.95]
N170	CBraMod	7.43 [5.36, 9.63]	-1.38 [-3.31, 0.51]
N170	LaBraM	4.24 [2.32, 6.23]	2.96 [1.41, 4.71]
N170	CSBrain	1.54 [-0.47, 3.77]	0.62 [-0.92, 2.13]

Table S5: Layer selection across five outer folds. Columns list jointly selected layers and the mean development-score margin over the runner-up (percent). Readout identities and their stability are summarized in the text.

Task	Model	Foldwise selected layer	Joint margin
N170	CBraMod	H08,H02,H04,H02,H04	0.227
N170	CSBrain	H01,H01,H02,H02,H02	0.219
N170	LaBraM	H01,H01,H02,H01,H02	0.251
N170	LUNA	H01,H03,H01,H01,H01	0.368
N170	REVE	H07,H07,H08,H08,H08	0.078
N400	CBraMod	H01,H05,H02,H03,H04	0.184
N400	CSBrain	H03,H04,H05,H04,H04	0.198
N400	LaBraM	H04,H04,H07,H04,H04	0.109
N400	LUNA	H07,H08,H07,H07,H07	0.338
N400	REVE	H10,H10,H09,H09,H09	0.168
OPENBMI	CBraMod	H08,H12,H09,H07,H04	0.169
OPENBMI	CSBrain	H03,H03,H03,H03,H03	0.528
OPENBMI	LaBraM	H07,H04,H06,H04,H07	0.226
OPENBMI	LUNA	H08,H08,H08,H07,H08	0.213
OPENBMI	REVE	H11,H11,H11,H10,H11	0.194
SEEDV	CBraMod	H09,H10,H01,H08,H09	0.350
SEEDV	CSBrain	H08,H04,H03,H01,H10	0.427
SEEDV	LaBraM	H12,H01,H09,H05,H05	0.450
SEEDV	LUNA	H05,H03,H02,H01,H02	0.490
SEEDV	REVE	H11,H20,H14,H20,H10	0.698

Readout selection is stable across the five outer folds for OpenBMI, N170, and N400, whereas the selected layer can vary across folds (Table S5). SEED-V shows greater variation in the selected readout.

C Cross-dataset reuse details

C.1 Motor-imagery transfer protocol

The OpenBMI-to-BCIC-IV-2a experiment uses the shared 21-channel montage Fz, FC3, FC1, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2, and POz with a common-average reference. Features are re-extracted under this common input protocol.

Channel mean plus temporal standard deviation is fixed as the source readout. Source development data select CBraMod H01 and REVE H11, which are then transferred unchanged to BCIC-IV-2a; H12 and H22 serve as their final-layer comparators. Target evaluation uses nine leave-one-participant-out outer folds, with the scaler, regularization, and classifier refitted within each target fold. Table S6 gives the full balanced-accuracy and macro-F1 results.

Table S6: OpenBMI → BCIC-IV-2a left/right interface reuse on nine target participants. Scores and additive contrasts are in percent with paired target-participant 95% intervals. The fixed source readout is channel mean plus temporal standard deviation. Source-selected/final layers are H01/H12 for CBraMod and H11/H22 for REVE. Target scalers and classifiers are refitted. BA is the main metric; macro-F1 is auxiliary.

Target strategy / contrast	CBraMod	REVE
BA (primary)		
Final layer / global mean	54.63 [51.89, 57.60]	63.27 [58.02, 69.48]
Final layer / fixed source readout	66.98 [61.73, 72.72]	77.55 [72.80, 82.48]
Source-selected layer / same readout	71.99 [67.75, 76.66]	79.82 [75.00, 84.72]
Depth gain at fixed source readout	5.02 [3.40, 6.40]	2.28 [0.42, 4.24]
Macro-F1 (auxiliary)		
Final layer / global mean	53.13 [50.27, 56.03]	61.99 [56.28, 68.61]
Final layer / fixed source readout	66.09 [60.51, 72.22]	77.15 [72.27, 82.24]
Source-selected layer / same readout	71.44 [67.03, 76.35]	79.52 [74.61, 84.57]
Depth gain at fixed source readout	5.35 [3.57, 6.90]	2.37 [0.39, 4.46]

Table S7: Motor-imagery participant consistency and frozen-head dependence. Upper block: source-selected minus final BA at the same readout (rounded percent). Lower block: source-selected-layer task-removal results; brackets are paired bootstrap 95% intervals. Counts and p concern excess over random. A dash denotes an unavailable value.

Participant / contrast	CBraMod H01	REVE H11
A01	+4.86	-0.69
A02	+6.94	+2.43
A03	+3.82	+0.69
A04	+4.51	+3.82
A05	+5.56	+2.43
A06	+7.99	+7.64
A07	+6.94	-2.43
A08	-0.35	+5.90
A09	+4.86	+0.69

Frozen-head contrast	CBraMod H01	REVE H11
Task-removal BA drop	5.36	5.32
Task minus matched random	5.30 [2.14, 8.16]	5.45 [0.44, 10.58]
Task minus matched PCA	5.29	–
Task minus matched shuffled	5.21	–
Positive participants; exploratory p	8/9; 0.0156	7/9; 0.0820

For participant-level depth effects d_i, we additionally report the exhaustive two-sided sign-flip statistic

$$p_{\mathrm{sf}} = 2^{-9} \sum_{s \in \{-1,1\}^9} \mathbb{I}\left\{ \left| \sum_{i=1}^{9} s_i d_i \right| \geq \left| \sum_{i=1}^{9} d_i \right| \right\}. \tag{8}$$

This calculation treats participant-effect signs as exchangeable under the null and is reported alongside the paired bootstrap intervals.

C.2 Controlled rotation of the target task

For balanced labels $y \in \{-1, +1\}$ and orthonormal directions $u, v, w, q \in \mathbb{R}^{12}$, we generate

$$x_\theta = y\{au + b(\cos\theta\, v + \sin\theta\, w)\} + \epsilon + \eta q, \tag{9}$$

where $a = 0.6$, $b = \sigma = 1$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and $\eta \sim \mathcal{N}(0, \tau^2)$ with $\tau \in \{0, 3\}$. The source uses $\theta = 0$; the target rotates the task signal from v toward w. Full-representation Bayes BA remains 87.82%, while the source-coordinate advantage vanishes at 90°.

Each condition averages 50 repetitions with 1,000 source, 500 target-training, and 5,000 target-test observations (seed 2026092006). The source task direction is estimated from class-mean differences in the coordinates orthogonal to u. Each compact representation appends one source task, PCA, random, or label-shuffled coordinate to $u^\top x$; full prediction uses all coordinates. Target classifiers use standardization and balanced logistic regression with $C = 1$, L-BFGS, a 1,000-iteration limit, and tolerance 10^{-8}. Table S8 reports fitted scores. With strong nuisance, source PCA follows task-orthogonal variance even before rotation, whereas the source task direction remains useful when aligned.

D Residual analyses and source-residual interventions

D.1 Cross-fitted residual analysis

Intermediate and final representations are standardized and independently reduced to 200 principal components using transformations fitted on the outer-training fold. Ridge regularization for the final-to-intermediate map is selected on participant-grouped inner splits. Training residuals are then generated by participant-grouped cross-fitting, while outer-test residuals use a map refitted on the full outer-training data.

Table S8: All rotation cells: mean balanced accuracy (percent) across 50 repetitions. Repeat-percentile ranges are retained in the numerical record.

τ	θ	Full	Mean	Task	PCA	Random	Shuffled
0	0	87.25	72.54	87.55	87.45	74.64	76.30
0	30	87.29	72.61	85.27	85.20	74.69	75.16
0	60	87.32	72.69	78.30	78.32	74.70	75.43
0	90	87.24	72.52	72.45	72.51	75.00	74.60
3	0	87.30	72.69	87.09	72.59	74.72	74.35
3	30	87.32	72.68	84.95	72.61	73.82	74.03
3	60	87.06	72.59	77.68	72.53	74.11	73.24
3	90	87.19	72.50	72.44	72.44	74.14	73.08

The shuffled control permutes residual rows within participant, separately in training and test data, and refits the augmented probe for each of 20 permutations. This preserves the within-participant residual distribution while breaking trial correspondence with the final-layer features. The cross-fitted, shuffled, in-sample, and direct-concatenation comparisons in Tables S9–S10 use balanced logistic probes; Table S11 reports in-sample results with unweighted probes.

Table S9: Cross-fitted forward utility and shuffled residuals of the same dimension (percent; paired 95% intervals). Permutation scores are averaged over 20 seeds within participant before resampling. Both Δ_{resid} and E_{resid} have positive intervals in 13/15 conditions.

Task	Model	Δ_{resid}	$\overline{\Delta}_{\text{perm}}$	E_{resid}
OpenBMI	CBraMod	0.84 [0.19, 1.53]	-0.81 [-1.13, -0.50]	1.65 [1.06, 2.28]
OpenBMI	REVE	7.25 [6.07, 8.45]	-0.23 [-0.54, 0.08]	7.48 [6.36, 8.60]
OpenBMI	LaBraM	5.00 [4.03, 5.99]	-0.51 [-0.86, -0.16]	5.50 [4.54, 6.45]
OpenBMI	LUNA	-0.01 [-0.53, 0.52]	-0.23 [-0.46, -0.00]	0.22 [-0.28, 0.74]
OpenBMI	CSBrain	3.74 [2.71, 4.80]	-0.41 [-0.83, 0.02]	4.15 [3.13, 5.19]
N170	CBraMod	10.03 [7.92, 12.17]	-1.74 [-2.16, -1.32]	11.77 [9.65, 13.94]
N170	REVE	9.24 [7.34, 11.11]	-1.75 [-2.06, -1.46]	10.99 [9.13, 12.82]
N170	LaBraM	6.78 [4.84, 8.70]	-1.90 [-2.31, -1.48]	8.69 [6.68, 10.66]
N170	LUNA	3.44 [1.66, 5.49]	-1.58 [-2.11, -1.08]	5.03 [3.38, 6.93]
N170	CSBrain	3.97 [2.86, 5.10]	-1.33 [-1.62, -1.04]	5.30 [4.20, 6.43]
N400	CBraMod	6.18 [4.70, 7.68]	-1.73 [-2.02, -1.45]	7.91 [6.41, 9.46]
N400	REVE	2.25 [1.13, 3.31]	-1.66 [-1.95, -1.37]	3.91 [2.79, 5.02]
N400	LaBraM	5.26 [3.91, 6.58]	-2.32 [-2.69, -1.94]	7.58 [6.27, 8.89]
N400	LUNA	-0.97 [-1.71, -0.25]	-1.52 [-1.91, -1.15]	0.55 [-0.14, 1.23]
N400	CSBrain	1.24 [0.07, 2.46]	-1.74 [-1.98, -1.50]	2.99 [1.80, 4.22]

D.2 Reverse, nonlinear, and projection controls

The reverse analysis predicts final-layer features from the selected intermediate layer using the same 200-dimensional representations. Let J_f and J_r denote the forward and reverse augmented-probe scores. Their relation to the single-layer depth contrast is

$$A_{\text{access}} = \{P(\tilde{z}_m) - P(\tilde{z}_L)\} + (J_f - J_r). \tag{10}$$

For the nonlinear control, the linear ridge map is replaced by a 200–256–200 ReLU network trained to reconstruct the intermediate representation from final-layer features. Reconstruction training uses Adam with early stopping on inner-fold reconstruction loss.

We also repeat the directional analysis after projecting each layer to 200 dimensions with independent Gaussian random projections. The mapping and downstream probes are then refitted in the projected space.

D.3 Source-residual intervention

For HMC-to-ISRUC transfer, the selected/final layers are H06/H12 for CBraMod and H14/H22 for REVE, using channel means and temporal standard deviations. Source standardization and 200-dimensional PCA are fitted on HMC and applied unchanged to ISRUC.

Table S10: Cross-fitted representation scores and residual comparisons (percent). Top: scores underlying Figure 3B, with marginal 95% participant-bootstrap intervals. F: final; S: selected; R: residual; J: final plus residual. Bottom: N is mean permuted augmentation; I is in-sample augmentation and C is direct concatenation, all under the same balanced probe. Only $I - C$ has a single fixed training-map reparameterization.

Task	Model	F	S	R	J
OpenBMI	CBraMod	70.28 [67.89, 72.68]	68.30 [65.94, 70.81]	54.27 [52.98, 55.62]	71.12 [68.84, 73.50]
OpenBMI	REVE	66.39 [64.12, 68.81]	74.80 [72.30, 77.32]	63.66 [61.53, 65.91]	73.65 [71.13, 76.22]
OpenBMI	LaBraM	61.87 [59.64, 64.25]	66.45 [64.01, 68.96]	59.44 [57.67, 61.36]	66.87 [64.57, 69.29]
OpenBMI	LUNA	59.06 [57.08, 61.11]	59.09 [57.17, 61.13]	52.07 [51.01, 53.23]	59.05 [57.04, 61.11]
OpenBMI	CSBrain	60.70 [58.35, 63.25]	63.56 [61.25, 66.00]	58.26 [57.00, 59.54]	64.44 [61.93, 67.07]
N170	CBraMod	64.71 [62.14, 67.32]	74.74 [71.36, 78.12]	68.56 [65.07, 72.18]	74.74 [71.55, 77.93]
N170	REVE	68.12 [65.42, 70.83]	78.40 [74.98, 81.75]	69.16 [66.17, 72.22]	77.35 [74.11, 80.55]
N170	LaBraM	64.72 [62.11, 67.31]	72.29 [68.80, 75.73]	64.69 [61.38, 68.04]	71.50 [68.28, 74.76]
N170	LUNA	57.37 [55.68, 59.06]	61.48 [59.56, 63.50]	57.71 [55.74, 59.85]	60.81 [58.95, 62.71]
N170	CSBrain	71.47 [67.94, 74.99]	75.69 [72.03, 79.33]	61.40 [59.18, 63.62]	75.44 [72.07, 78.73]
N400	CBraMod	69.33 [67.08, 71.64]	76.79 [74.20, 79.39]	66.14 [63.75, 68.62]	75.51 [72.93, 78.16]
N400	REVE	73.78 [71.16, 76.44]	77.47 [75.10, 79.87]	60.46 [58.34, 62.63]	76.03 [73.33, 78.74]
N400	LaBraM	68.61 [66.33, 70.93]	75.25 [72.94, 77.54]	64.66 [62.56, 66.94]	73.87 [71.45, 76.29]
N400	LUNA	70.03 [67.55, 72.55]	70.17 [67.62, 72.69]	51.90 [50.04, 53.66]	69.06 [66.69, 71.44]
N400	CSBrain	74.31 [72.06, 76.57]	77.25 [74.91, 79.60]	58.69 [56.64, 60.72]	75.55 [73.26, 77.91]

Task	Model	N	$J - I$	$J - C$	$I - C$
OpenBMI	CBraMod	69.47	-0.18	-0.24	-0.06
OpenBMI	REVE	66.16	-0.38	1.27	1.65
OpenBMI	LaBraM	61.37	-0.16	-0.01	0.16
OpenBMI	LUNA	58.83	0.10	0.09	-0.02
OpenBMI	CSBrain	60.29	-0.07	-0.04	0.03
N170	CBraMod	62.97	-0.21	-0.19	0.02
N170	REVE	66.36	-0.40	-0.82	-0.41
N170	LaBraM	62.82	-0.04	-0.58	-0.54
N170	LUNA	55.79	0.15	0.06	-0.10
N170	CSBrain	70.14	0.14	-0.23	-0.37
N400	CBraMod	67.60	-0.01	-0.35	-0.34
N400	REVE	72.12	0.13	-1.01	-1.14
N400	LaBraM	66.29	0.05	-0.59	-0.63
N400	LUNA	68.51	-0.05	-1.36	-1.31
N400	CSBrain	72.57	0.07	-1.21	-1.28

Table S11: In-sample residual prediction. Scores are percentages; $U_{m|L}$ is an additive contrast in percent with a paired 95% interval. The four representations have 200, 200, 200, and 400 inputs.

Task	Model	$P(\tilde{z}_L)$	$P(\tilde{z}_m)$	$P(e_{m\|L})$	$P([\tilde{z}_L, e])$	$U_{m\|L}$
OpenBMI	CBraMod	70.16	68.27	54.45	71.22	1.06 [0.45, 1.69]
OpenBMI	REVE	66.26	74.75	65.29	74.01	7.74 [6.26, 9.30]
OpenBMI	LaBraM	61.85	66.44	60.07	67.05	5.20 [4.25, 6.14]
OpenBMI	LUNA	59.04	59.05	51.42	58.97	-0.07 [-0.52, 0.37]
OpenBMI	CSBrain	60.68	63.56	58.23	64.44	3.76 [2.73, 4.79]
N170	CBraMod	64.72	74.74	69.28	74.96	10.24 [8.06, 12.48]
N170	REVE	68.12	78.39	69.92	77.75	9.63 [7.50, 11.74]
N170	LaBraM	64.72	72.29	64.90	71.73	7.01 [5.00, 8.92]
N170	LUNA	57.37	61.47	57.89	60.66	3.29 [1.43, 5.29]
N170	CSBrain	71.48	75.70	62.08	75.29	3.81 [2.65, 4.98]
N400	CBraMod	69.35	76.79	66.35	75.60	6.25 [4.73, 7.78]
N400	REVE	73.78	77.50	61.28	76.00	2.22 [0.99, 3.45]
N400	LaBraM	68.63	75.25	64.64	73.92	5.29 [3.92, 6.64]
N400	LUNA	70.03	70.18	52.21	69.17	-0.87 [-1.63, -0.14]
N400	CSBrain	74.30	77.24	59.15	75.55	1.25 [0.01, 2.55]

Table S12: Forward and reverse conditional utility (percent; paired 95% intervals), using in-sample training residuals. Positive A_{access} means adding the intermediate residual to final features helps more than the reverse addition. Twelve intervals are positive; the condition-average asymmetry is 4.76%.

Task	Model	$U_{m\mid L}$	$U_{L\mid m}$	A_{access}
OpenBMI	CBraMod	1.06 [0.45, 1.69]	2.80 [1.80, 3.93]	-1.74 [-2.99, -0.59]
OpenBMI	REVE	7.74 [6.26, 9.30]	-0.81 [-1.87, 0.21]	8.55 [6.75, 10.43]
OpenBMI	LaBraM	5.20 [4.25, 6.14]	0.29 [-0.41, 1.00]	4.90 [3.88, 5.96]
OpenBMI	LUNA	-0.07 [-0.52, 0.37]	-0.07 [-0.61, 0.48]	0.00 [-0.46, 0.39]
OpenBMI	CSBrain	3.76 [2.73, 4.79]	1.12 [0.24, 2.06]	2.64 [1.63, 3.66]
N170	CBraMod	10.24 [8.06, 12.48]	0.19 [-0.74, 1.22]	10.05 [7.41, 12.61]
N170	REVE	9.63 [7.50, 11.74]	-0.57 [-1.37, 0.28]	10.21 [8.00, 12.41]
N170	LaBraM	7.01 [5.00, 8.92]	-0.47 [-1.58, 0.62]	7.48 [4.78, 10.12]
N170	LUNA	3.29 [1.43, 5.29]	-0.76 [-1.87, 0.35]	4.05 [1.74, 6.49]
N170	CSBrain	3.81 [2.65, 4.98]	-0.40 [-1.35, 0.57]	4.21 [2.88, 5.55]
N400	CBraMod	6.25 [4.73, 7.78]	-1.33 [-2.08, -0.59]	7.57 [6.20, 8.92]
N400	REVE	2.22 [0.99, 3.45]	-1.44 [-2.14, -0.74]	3.66 [2.46, 4.86]
N400	LaBraM	5.29 [3.92, 6.64]	-1.36 [-2.18, -0.60]	6.65 [5.19, 8.14]
N400	LUNA	-0.87 [-1.63, -0.14]	-0.96 [-1.56, -0.41]	0.10 [-0.54, 0.75]
N400	CSBrain	1.25 [0.01, 2.55]	-1.76 [-2.52, -1.00]	3.01 [1.73, 4.34]

Table S13: Forward nonlinear-mapper sensitivity: participant-equal scores in percent with marginal bootstrap 95% intervals. All MLP residual lower bounds exceed chance (50%). The mapper has architecture 200–256–200.

Task	Model	Linear residual	MLP residual	Final + MLP residual
OpenBMI	CBraMod	54.45 [53.28, 55.67]	56.43 [55.26, 57.62]	70.67 [68.51, 72.90]
OpenBMI	REVE	65.29 [62.96, 67.81]	66.01 [63.90, 68.29]	72.64 [70.32, 75.05]
OpenBMI	LaBraM	60.07 [58.27, 62.05]	60.06 [58.26, 61.93]	66.33 [64.09, 68.68]
OpenBMI	LUNA	51.42 [50.47, 52.43]	52.24 [51.20, 53.35]	56.42 [54.81, 58.13]
OpenBMI	CSBrain	58.23 [56.97, 59.54]	57.89 [56.57, 59.22]	63.86 [61.49, 66.34]
N170	CBraMod	69.28 [65.86, 72.83]	69.32 [66.22, 72.50]	73.89 [70.79, 77.00]
N170	REVE	69.92 [66.78, 73.10]	69.50 [66.61, 72.36]	76.55 [73.32, 79.74]
N170	LaBraM	64.90 [61.55, 68.33]	64.98 [61.57, 68.43]	70.88 [67.52, 74.21]
N170	LUNA	57.89 [55.83, 60.16]	57.54 [55.65, 59.60]	60.35 [58.51, 62.23]
N170	CSBrain	62.08 [60.00, 64.18]	64.14 [61.85, 66.44]	75.00 [71.51, 78.35]
N400	CBraMod	66.35 [63.96, 68.92]	67.18 [64.67, 69.79]	74.99 [72.36, 77.67]
N400	REVE	61.28 [59.19, 63.47]	61.85 [59.65, 63.99]	74.83 [72.32, 77.35]
N400	LaBraM	64.64 [62.62, 66.86]	64.17 [61.85, 66.59]	72.51 [70.00, 75.01]
N400	LUNA	52.21 [50.30, 54.06]	54.07 [52.39, 55.77]	67.98 [65.65, 70.34]
N400	CSBrain	59.15 [56.97, 61.34]	59.40 [57.28, 61.50]	73.41 [71.04, 75.83]

Table S14: Projection and unreduced sensitivity (percent; paired 95% intervals). RP200 uses separate Gaussian projections for each layer and balanced probes; five of six asymmetry intervals are positive.

Mode	Task	Model	Forward	Reverse	Asymmetry
native	N170	CBraMod	-6.18 [-9.54, -3.15]	-10.15 [-12.17, -8.17]	3.98 [0.86, 6.96]
native	N170	REVE	-14.35 [-17.57, -11.15]	-16.78 [-19.51, -14.06]	2.43 [-1.10, 6.12]
native	N400	CBraMod	-9.81 [-12.04, -7.59]	-7.81 [-9.72, -5.95]	-2.01 [-4.70, 0.63]
native	N400	REVE	-18.32 [-21.32, -15.41]	-15.18 [-16.83, -13.45]	-3.14 [-6.35, -0.06]
rp200	N170	CBraMod	9.63 [7.19, 12.16]	-0.06 [-1.08, 0.98]	9.69 [6.74, 12.76]
rp200	N170	REVE	7.52 [5.37, 9.54]	0.33 [-0.54, 1.27]	7.19 [4.39, 9.78]
rp200	N400	CBraMod	6.83 [5.18, 8.44]	0.34 [-0.53, 1.29]	6.49 [4.31, 8.51]
rp200	N400	REVE	2.88 [1.65, 4.16]	0.13 [-0.62, 0.89]	2.76 [1.22, 4.34]
rp200	OpenBMI	CBraMod	2.53 [1.50, 3.55]	2.31 [1.37, 3.26]	0.22 [-1.17, 1.61]
rp200	OpenBMI	REVE	3.79 [2.72, 4.82]	2.29 [1.31, 3.23]	1.51 [0.22, 2.74]

A ridge map is fitted on HMC from final-layer to selected-layer features,

$$W_s = (Z_{L,s}^\top Z_{L,s} + \lambda I)^{-1} Z_{L,s}^\top Z_{m,s}, \qquad E_s = Z_{m,s} - Z_{L,s} W_s, \tag{11}$$

with λ selected by participant-grouped source validation. Class prototypes are formed from the source residuals using participant-equal class means. Centering the five sleep-stage prototypes and taking their right singular vectors gives the four-dimensional task basis Q_s.

Removal controls use a rank-four random basis, the leading four source-residual principal components, and a basis obtained from within-participant shuffled source labels. Control perturbations are scaled to match the task-basis displacement after the target scaler. For target-training center μ_t and scaler standard deviations D, define

$$\delta_Q(z) = QQ^\top(z - \mu_t), \tag{12}$$

$$E(Q) = \left[\frac{1}{N}\sum_z \|D^{-1}\delta_Q(z)\|_2^2\right]^{1/2}, \tag{13}$$

$$\gamma_Q = E(Q_s)/E(Q), \qquad \gamma_{Q_s} = 1. \tag{14}$$

The frozen-head intervention applies $z - \gamma_Q \delta_Q(z)$ while keeping the target scaler and classifier fixed. For the refitted condition, the scaler, regularization, and classifier are fitted again after intervention.

For augmentation, target residuals are projected onto the source task basis and the four coordinates $Q_s^\top e_t$ are appended to the 200 final-layer features. Random, source-PCA, and shuffled-label augmentations use the same coordinate count and fitting procedure.

Table S15: HMC-to-ISRUC intervention scores: pooled five-class macro-F1 in percent, with marginal 95% whole-participant bootstrap intervals ($n = 98$). Source-only PCA200 coordinates and target-fitted probes are used. Table 1 reports native-dimensional transfer scores. Random and shuffled-label controls each evaluate one orthonormalized average basis.

Representation / comparison	CBraMod	REVE
Selected layer / frozen baseline	69.09 [67.01, 70.99]	71.45 [69.26, 73.42]
Selected layer / refit baseline	69.09 [67.01, 70.99]	71.45 [69.26, 73.42]
Final layer	68.10 [65.78, 70.18]	69.89 [67.56, 72.02]
Frozen removal / task	46.33 [44.07, 48.37]	54.76 [52.45, 56.93]
Frozen removal / random	64.67 [61.90, 67.16]	68.46 [66.10, 70.60]
Frozen removal / source PCA	62.87 [60.19, 65.26]	60.38 [57.02, 63.48]
Frozen removal / shuffled labels	63.99 [61.01, 66.69]	62.93 [59.33, 66.28]
Refitted removal / task	67.34 [65.21, 69.27]	70.63 [68.64, 72.49]
Refitted removal / random	69.02 [66.93, 70.94]	71.44 [69.22, 73.43]
Refitted removal / source PCA	69.07 [66.98, 70.99]	71.50 [69.30, 73.47]
Refitted removal / shuffled labels	69.06 [66.99, 70.97]	71.45 [69.24, 73.44]
Augmentation / task	69.26 [66.97, 71.30]	70.55 [68.27, 72.63]
Augmentation / random	68.29 [65.97, 70.37]	69.95 [67.64, 72.07]
Augmentation / source PCA	68.60 [66.39, 70.57]	69.96 [67.69, 72.08]
Augmentation / shuffled labels	68.58 [66.33, 70.59]	70.04 [67.72, 72.15]

All intervention intervals use paired whole-participant bootstrap resampling on the 98 ISRUC participants, with macro-F1 computed from pooled participant confusion counts.

E ALL-LAYER RESULTS ACROSS READOUTS

Figure S1 shows the complete layer–readout score grid for the five models. The dark curve is the best observed readout at each layer, and its peak gives the descriptive test-set envelope. The development-selected comparisons in the main text remain the primary estimates.

Table S17 reports the envelope peak, the best final-layer score, and their difference. SEED-V curves use the exported trial-level balanced accuracy, whereas the main SEED-V evaluation uses video-level balanced accuracy.

Table S16: Paired HMC-to-ISRUC contrasts in percent, with 95% whole-participant bootstrap intervals. Positive removal contrasts mean larger drops under task-direction removal. Positive augmentation contrasts favor task coordinates. Frozen removal matches probe-space displacement; augmentation controls match rank and fitting procedure. The final post-hoc contrast uses the same participant draws.

Representation / comparison	CBraMod	REVE
Frozen drop: task minus random	18.34 [15.78, 20.85]	13.70 [11.79, 15.64]
Frozen drop: task minus source PCA	16.54 [14.28, 18.74]	5.62 [3.50, 7.68]
Frozen drop: task minus shuffled labels	17.66 [15.05, 20.20]	8.17 [5.58, 10.73]
Refitted drop: task minus random	1.69 [1.24, 2.19]	0.81 [0.31, 1.31]
Refitted drop: task minus source PCA	1.74 [1.30, 2.24]	0.87 [0.37, 1.39]
Refitted drop: task minus shuffled labels	1.73 [1.29, 2.23]	0.82 [0.33, 1.32]
Task augmentation minus final only	1.16 [0.71, 1.68]	0.67 [0.31, 1.07]
Task augmentation minus random	0.97 [0.52, 1.50]	0.60 [0.26, 0.96]
Task augmentation minus source PCA	0.66 [0.19, 1.15]	0.59 [0.24, 0.98]
Task augmentation minus shuffled labels	0.68 [0.27, 1.09]	0.51 [0.22, 0.82]
Augmentation minus selected (post hoc)	0.17 [-0.67, 1.02]	-0.90 [-1.50, -0.30]

Table S17: Test-selected peaks across layers and readouts. Scores and gains are in percent. GM: global mean; GS: global mean plus standard deviation; CM: channel mean; CS: channel mean plus temporal standard deviation; S: four temporal bins. The 95% intervals resample participants and reselect the maximum in each draw; they describe the oracle statistic and retain test-selection optimism. Final best selects the readout at the final layer using the same test scores.

Task	Model	Layer / readout	Peak [95% interval]	Final best	Gain
OpenBMI / BA	CBraMod	H04 / S	76.71 [74.37, 79.24]	76.09	+0.62
	REVE	H11 / S	80.40 [77.89, 83.10]	76.95	+3.45
	LaBraM	H05 / S	76.39 [74.10, 78.84]	69.74	+6.64
	LUNA	H07 / S	59.89 [58.02, 62.01]	59.67	+0.22
	CSBrain	H03 / S	75.10 [72.87, 77.38]	71.20	+3.91
N170 / AUROC	CBraMod	H04 / CM	78.27 [75.19, 81.66]	68.77	+9.50
	REVE	H08 / CM	80.98 [77.94, 83.97]	73.41	+7.57
	LaBraM	H01 / CM	75.58 [72.15, 79.19]	67.01	+8.57
	LUNA	H01 / S	63.23 [61.27, 65.62]	58.25	+4.98
	CSBrain	H02 / CM	79.22 [75.74, 82.74]	74.78	+4.44
N400 / AUROC	CBraMod	H03 / CM	77.95 [75.75, 80.48]	71.18	+6.77
	REVE	H10 / CM	78.63 [76.29, 81.06]	75.81	+2.82
	LaBraM	H04 / CM	75.95 [73.63, 78.28]	70.87	+5.08
	LUNA	H07 / S	72.12 [69.85, 74.52]	71.57	+0.55
	CSBrain	H04 / CM	78.27 [75.97, 80.62]	75.52	+2.75
SEED-V / trial BA	CBraMod	H04 / CS	34.72 [31.11, 41.11]	32.22	+2.50
	REVE	H21 / GS	36.94 [32.78, 44.03]	35.42	+1.53
	LaBraM	H09 / CM	35.28 [31.94, 42.36]	34.03	+1.25
	LUNA	H03 / GS	33.47 [29.58, 40.14]	33.33	+0.14
	CSBrain	H07 / S	33.06 [29.44, 37.92]	29.17	+3.89

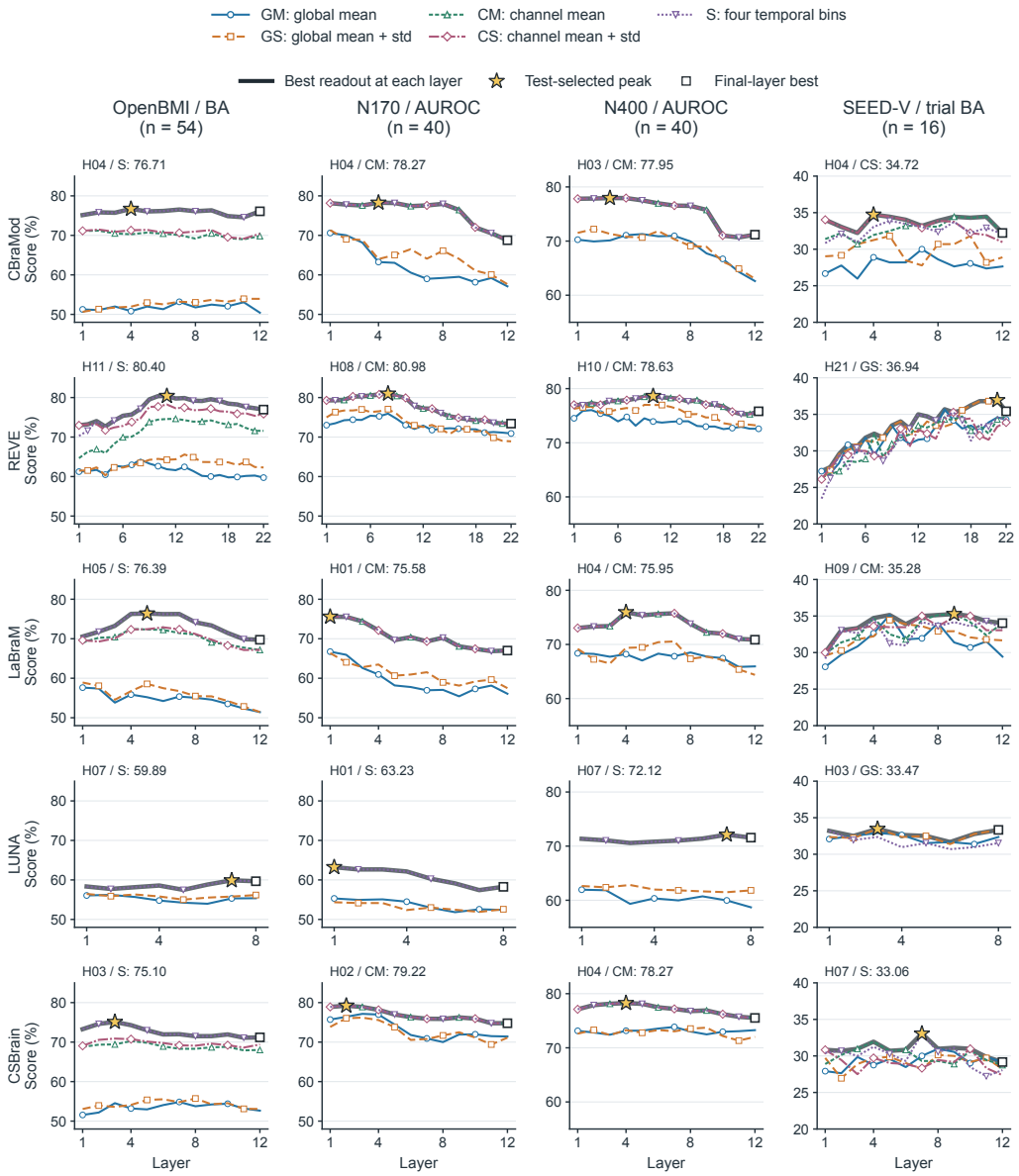


Figure S1: All-layer performance across readouts. Rows identify models and columns identify tasks. Colored curves show the individual readouts; the dark curve is the best observed readout at each layer. Stars mark the envelope peak and open squares the best final-layer score.