KERNELIZED TASK-SUFFICIENCY REGULARIZATION FOR EEG FOUNDATION MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting EEG foundation models to new tasks and individuals is essential, especially when task labels are scarce. Yet strong predictive performance does not reveal whether adaptation captures the task-relevant information available from pretraining. A residual-information audit on held-out subjects shows that frozen pretrained features can improve adapted logits in several settings, including after full fine-tuning. Motivated by this finding, we propose *ReSuff*, which encourages adapted representations to capture label information available in a fixed, compact view of pretrained features. Using label residuals defined by conditional class probabilities, we derive a joint Gaussian-kernel criterion that vanishes exactly at task sufficiency relative to this view at the population level. During training, ReSuff fits a kernel predictor on support subjects and evaluates residual moments on disjoint query subjects, while class-conditioned subject-adversarial learning addresses intersubject variation. Across three limited-label, subject-independent EEG benchmarks and three pretrained backbones, ReSuff improves balanced accuracy over the matched task-focused adaptation baseline in all nine backbone and dataset combinations, with an average gain of 2.3 percentage points. On REVE, it further improves the two-task mean balanced accuracy of four adaptation strategies by 1.5–2.2 points. These results show ReSuff is a practical plug-in to improve EEG foundation model adaptation without increasing inference cost.

1 INTRODUCTION

EEG foundation models learn transferable representations from large and heterogeneous collections of neural recordings (Jiang et al., 2024; El Ouahidi et al., 2025; Wang et al., 2025; Kostas et al., 2021). Realizing their potential requires adaptation to downstream tasks with limited labels, while accounting for differences across individuals and recording conditions. This is important for applications ranging from motor imagery decoding to sleep staging. Existing approaches such as full fine-tuning, parameter-efficient updates, and task-guided modules, aim to convert pretrained representations into accurate task predictors.

There is evidence that pretrained EEG representations contain reusable task information. Representation analyses have identified conventional EEG features encoded by foundation models and found that some contribute to downstream prediction (Tang et al., 2026). What remains less clear is whether downstream adaptation makes full use of the available information. In a preliminary residual-information audit, we find that features extracted from the frozen pretrained encoder can improve adapted logits on held-out subjects, reducing negative log-likelihood in several settings. Such gains persist after full fine-tuning in some cases and are more pronounced for task-guided adaptation. Because this audit examines predictions, it does not establish what information is missing from the adapted representation itself. This motivates our method: can downstream training push the representation to capture more label information from pretraining?

We formalize this question through *task sufficiency*. An adapted representation is task-sufficient relative to a fixed, compact view of the frozen pretrained features if that view provides no additional label information once the adapted representation is known. This gives downstream adaptation a goal beyond accurate prediction: the representation should account for the task evidence accessible through the pretrained reference. The residual audit motivates this goal at the level of predictions; task sufficiency addresses it at the level of representations.

To make task sufficiency a training objective, we propose *ReSuff*. We first estimate the conditional mean of the one-hot label given the adapted representation and use the unexplained component as a

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.

label residual. A joint Gaussian-kernel criterion then measures whether the residual retains a systematic relationship with the frozen reference and the adapted representation. With the true conditional label mean, this criterion vanishes exactly at task sufficiency relative to the reference view. In practice, ReSuff fits a kernel predictor on support subjects and evaluates residual moments on disjoint query subjects, with subject-balanced weights. Class-conditioned subject-adversarial training further discourages subject-specific features. Both auxiliary objectives are used only in training and can guide full fine-tuning or parameter-efficient adaptation without changing the inference architecture.

We evaluate ReSuff on three limited-label, subject-independent EEG benchmarks using REVE, LaBraM, and CBraMod. Compared with the matched task-focused adaptation baseline, ReSuff improves balanced accuracy in all nine backbone and dataset combinations, with an average gain of 2.3 percentage points. On REVE, adding ReSuff to four adaptation strategies increases their mean balanced accuracy across two benchmarks by 1.5–2.2 points. These results show ReSuff improves downstream performance across backbones and adaptation regimes with no added inference cost.

Our main contributions are as follows: (1) We use a residual-information audit to identify frozen pretrained features that improve adapted predictions, and formulate task sufficiency relative to a fixed reference view as an objective for EEG foundation model adaptation. (2) We derive a joint Gaussian-kernel criterion that characterizes task sufficiency through conditional label residuals at the population level, and develop a subject-disjoint estimator for using it during training. (3) We demonstrate performance gains across three EEG foundation models and show that ReSuff can be added to multiple adaptation strategies without changing their inference architecture.[1]

2 Preliminaries

2.1 Problem Formulation

Downstream adaptation determines which pretrained EEG features enter the representation used for prediction. We ask whether a fixed view of the original pretrained features still contains label information beyond that adapted representation.

Let (X, Y) follow the subject-balanced distribution of a downstream EEG classification task, where X is an EEG input and $Y \in \{1, \dots, C\}$ is its class label. Subject balancing assigns equal weight to each subject and, within a subject, to each trial. We write $\mathbf{Y} \in \{0, 1\}^C$ for the one-hot encoding of Y, used below to define class-wise label residuals. Let F_0 be a frozen copy of the pretrained encoder, and let $H_0 = F_0(X)$ denote its selected hidden features. A fixed summarization map ρ yields the compact reference view $U = \rho(H_0)$, whose construction is specified in Sec. 3.2. Both F_0 and ρ remain fixed during downstream training.

Let f_θ denote the adaptation model built on the pretrained encoder and g_ω its classification head. The task representation is the input to that head:

$$T_\theta = f_\theta(X), \qquad Z = g_\omega(T_\theta), \tag{1}$$

where $Z \in \mathbb{R}^C$ contains the class logits. The parameters θ are those updated by the adaptation strategy, and ω denotes the head parameters. Thus, θ includes the encoder parameters under full fine-tuning but only the trainable adaptation parameters under parameter-efficient adaptation.

Definition 1 (Task sufficiency). The representation T_θ is *task-sufficient relative to* U if observing U does not change the conditional label distribution once T_θ is known:

$$P(Y = c \mid U, T_\theta) = P(Y = c \mid T_\theta) \quad \text{almost surely, for every } c \in \{1, \dots, C\}. \tag{2}$$

Equivalently, $I(Y; U \mid T_\theta) = 0$. This condition concerns the information in the representation, independently of the fitted classification head. A model can make accurate predictions while U still contains complementary label information beyond T_θ.

This gives an ideal constrained formulation for downstream adaptation:

$$\min_{\theta, \omega} \ \mathbb{E}\big[\ell\big(g_\omega(T_\theta), Y\big)\big] \qquad \text{s.t.} \qquad I(Y; U \mid T_\theta) = 0, \tag{3}$$

where ℓ is the classification loss. Directly enforcing the conditional mutual information constraint is difficult. Sec. 3 instead develops a label-residual kernel criterion that characterizes sufficiency relative to U at the population level and provides a practical training regularizer.

[1] Due to the page limit, a comprehensive discussion of related work is provided in Appendix A.

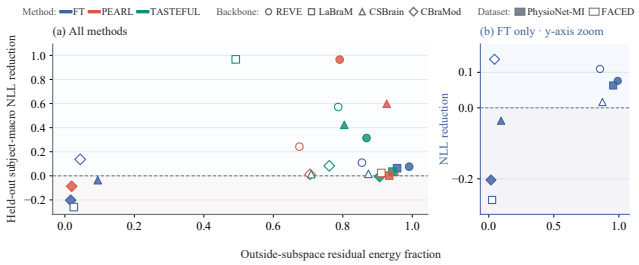


Figure 1: Residual-information analysis across adaptation methods, backbones, and datasets. The right panel enlarges the FT results using a tighter vertical scale.

2.2 Empirical Analysis

We first examine whether frozen pretrained features can improve the predictions of existing adaptation methods. This prediction-level audit motivates the representation-level question in Definition 1, without directly testing the sufficiency condition. For each method, SPLINCE uses its logits to identify the prediction subspace already utilized by the adapted model. It then constructs a method-specific correction from frozen pretrained features, removing the component aligned with that subspace while retaining information useful for correcting training-subject predictions. We evaluate whether adding this correction to the logits reduces subject-macro-averaged negative log-likelihood (NLL) on held-out subjects, using within-subject permutations as a control.

Fig. 1 plots the fraction of residual correction energy outside the existing prediction subspace against the held-out NLL reduction. Most full fine-tuning (FT) results show small or negative gains, although positive gains remain in some settings. TASTEFUL shows substantial gains in several backbone–dataset combinations, including LaBraM on FACED (Chen et al., 2023). PEARL is more variable but also exhibits a pronounced gain with REVE on PhysioNet-MI (Schalk et al., 2004; Goldberger et al., 2000). These results indicate that frozen pretrained features can provide complementary evidence beyond adapted logits, with the size of that contribution depending on the adaptation method and task. Because the audit conditions on predictions rather than the full task representation, it does not establish a violation of task sufficiency. It instead motivates ReSuff's representation-level objective: encouraging adaptation to capture label information that remains accessible through the frozen reference.

3 Method

We introduce *ReSuff* to encourage the adapted representation T_θ to capture label information available in the fixed pretrained reference view U. As illustrated in Fig. 2, the classification loss trains T_θ for label prediction but does not assess whether U remains informative once T_θ is known. ReSuff addresses this question through label residuals: it removes the conditional class probabilities explained by T_θ from the one-hot label, then measures whether the residual retains a systematic relationship with (U, T_θ). A joint Gaussian-kernel criterion characterizes task sufficiency relative to U at the population level. During training, we estimate the conditional label mean on support subjects and evaluate residual moments on disjoint query subjects. We optimize the resulting regularizer alongside the classification loss and a class-conditioned subject-adversarial objective that discourages subject-specific features. Population expectations in this section are taken under a subject-balanced downstream distribution, which gives equal weight to subjects.

3.1 Label-Residual Characterization of Task Sufficiency

Task sufficiency requires that the frozen reference view U provide no additional label information once the adapted representation T_θ is known. To characterize this requirement, we define the conditional mean of the one-hot label $\mathbf{Y}$ and the corresponding label residual:

$$m_\theta(T_\theta) = \mathbb{E}[\mathbf{Y} \mid T_\theta], \qquad \mathbf{r}_\theta = \mathbf{Y} - m_\theta(T_\theta). \tag{4}$$

For each class c, the component $m_{\theta,c}(T_\theta) = P(Y = c \mid T_\theta)$ is its true conditional probability given the representation; $r_{\theta,c}$ denotes the corresponding residual component. Defining residuals using these population probabilities separates information missing from T_θ from errors caused by fitting a particular classification head.

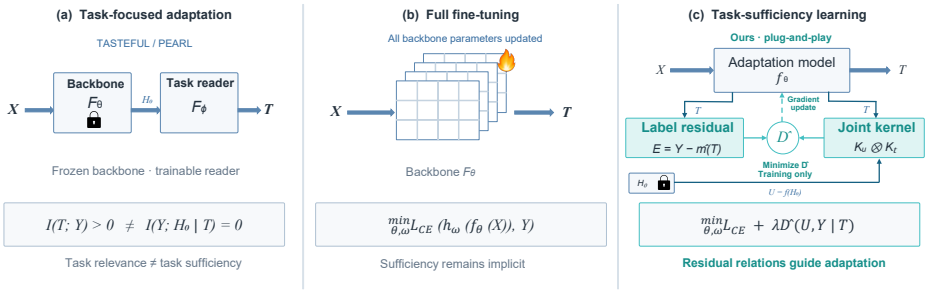


Figure 2: Overview of task-sufficiency regularization for EEG foundation model adaptation.

Proposition 1 (Residual characterization of task sufficiency). For a fixed adaptation mapping f_θ and a categorical label Y,

$$I(Y; U \mid T_\theta) = 0 \quad \Longleftrightarrow \quad \mathbb{E}[\mathbf{r}_\theta \mid U, T_\theta] = \mathbf{0}. \tag{5}$$

Indeed, for each class c, the conditional mean of $r_{\theta,c}$ given (U, T_θ) equals $P(Y = c \mid U, T_\theta) - P(Y = c \mid T_\theta)$. The proposition therefore requires equality of the conditional class probabilities. An individual residual may be nonzero even when the condition holds; what must vanish is its conditional mean given the reference and task representations.

This condition is equivalent to a collection of unconditional moments. Let a be any bounded measurable real-valued function of (U, T_θ). Then task sufficiency holds if and only if

$$\mathbb{E}[r_{\theta,c}\, a(U, T_\theta)] = 0 \qquad \text{for every class } c \text{ and every such function } a. \tag{6}$$

The next subsection uses a joint kernel to aggregate these residual moments into a scalar criterion.

3.2 Joint-Kernel Sufficiency Regularization

Proposition 1 characterizes task sufficiency through residual moments against every bounded measurable test function. Checking these moments individually is infeasible. We instead construct a kernel criterion that aggregates residual relationships over a rich function space.

Frozen reference view. Direct kernel computation on multilayer pretrained features can be costly, and layer outputs may have different numbers of representation positions. We therefore use the fixed compact view $U = \rho(H_0)$ introduced in Sec. 2.1. For each of L selected layers, let $H_0^{(\ell)} \in \mathbb{R}^{n_\ell \times d_\ell}$ denote its hidden features, where n_ℓ is the number of positions and d_ℓ is the feature dimension. We summarize each feature dimension by its mean and standard deviation across positions:

$$U^{(\ell)} = \left[\mathrm{mean}_{\mathrm{pos}}\, H_0^{(\ell)}\,;\, \mathrm{std}_{\mathrm{pos}}\, H_0^{(\ell)}\right], \qquad U = \left(U^{(1)}, \ldots, U^{(L)}\right). \tag{7}$$

The semicolon denotes vector concatenation. If a selected layer already produces a pooled vector, we use it directly. The reference encoder and ρ remain fixed, and gradients do not propagate through U. Accordingly, the criterion below assesses sufficiency relative to this compact view, not to every feature in H_0.

Joint-kernel construction. A residual relationship with U may differ across values of T_θ and cancel when tested using functions of U alone. We therefore use a kernel on the joint space of (U, T_θ). For values u, u' of U and t, t' of T_θ, define

$$k_U(u, u') = \exp\left(-\frac{1}{2L}\sum_{\ell=1}^{L}\frac{\|u^{(\ell)} - u'^{(\ell)}\|_2^2}{\sigma_{U,\ell}^2}\right), \tag{8}$$

$$k_T(t, t') = \exp\left(-\frac{\|t - t'\|_2^2}{2\sigma_T^2}\right), \quad k\big((u,t),(u',t')\big) = k_U(u,u')\, k_T(t,t'). \tag{9}$$

Here, $\sigma_{U,\ell}$ and σ_T are fixed, positive, finite bandwidths. The layer-specific bandwidths scale distances within each reference layer, while the factor $1/L$ averages their contributions. Their product with k_T forms a Gaussian kernel on the combined reference and task representation.

Population regularizer. Let $\mathcal{H}$ be the reproducing kernel Hilbert space of k, with feature map $\phi(u,t) = k((u,t),\cdot)$. We define the class-wise residual moment criterion

$$\mathcal{D}(\theta) = \sum_{c=1}^{C} \|\mathbb{E}[r_{\theta,c}\,\phi(U,T_\theta)]\|_{\mathcal{H}}^2 . \tag{10}$$

The criterion is nonnegative and aggregates residual relationships with the joint representation. By the reproducing property, each norm equals the largest absolute residual moment over functions in the unit ball of $\mathcal{H}$ (Muandet et al., 2020).

Proposition 2 (Joint-kernel characterization of task sufficiency). Fix f_θ and use the population conditional mean in Eq. 4. If U and T_θ are finite-dimensional, then for the positive, finite bandwidths above,

$$\mathcal{D}(\theta) = 0 \quad \Longleftrightarrow \quad \mathbb{E}[\mathbf{r}_\theta \mid U, T_\theta] = \mathbf{0} \quad \Longleftrightarrow \quad I(Y; U \mid T_\theta) = 0. \tag{11}$$

The product Gaussian kernel distinguishes finite signed measures on the joint space (Sriperumbudur et al., 2011). A zero residual embedding therefore implies that the moments in Eq. 6 vanish for every class. This exact equivalence concerns the population criterion with the true conditional label mean; its finite-sample estimate is a training surrogate.

3.3 Subject-Disjoint Label-Residual Estimation

Kernel ridge regression (KRR) (Saunders et al., 1998) may fit the labels of its training examples closely, making residual moments evaluated on those same examples artificially small. We therefore divide each training batch by subject into disjoint support and query sets, S and Q. All samples from a subject belong to one set: S is used to estimate the conditional label mean, and Q is used to evaluate the residual moments. To prevent subjects with more trials from dominating either step, we weight samples within each set. For $A \in \{S, Q\}$, let $\mathcal{S}_A$ be the subjects represented in A, let s_i be the subject of sample i, and let n_{A,s_i} be that subject's number of samples in A. Define

$$w_i^A = \frac{1}{|\mathcal{S}_A|\, n_{A,s_i}}, \qquad W_A = \operatorname{diag}\big((w_i^A)_{i\in A}\big). \tag{12}$$

These weights sum to one within each set and assign equal total weight to each represented subject, matching the subject-balanced population distribution.

For sample i, write $T_i = f_\theta(x_i)$ and $U_i = \rho(F_0(x_i))$. Let Y_A contain the one-hot label vectors $\mathbf{y}_i^\top$ as rows for $i \in A$, and define $(K_T^{AB})_{ij} = k_T(T_i, T_j)$ for $i \in A$ and $j \in B$. Weighted KRR obtains the coefficient matrix B_S by solving

$$\Big(W_S^{1/2} K_T^{SS} W_S^{1/2} + \lambda_{\mathrm{KRR}} I\Big)\, B_S = W_S^{1/2} Y_S, \tag{13}$$

where $\lambda_{\mathrm{KRR}} > 0$ is the ridge coefficient. Its query predictions and the resulting residual matrix are

$$\widehat{M}_Q = K_T^{QS} W_S^{1/2} B_S, \qquad E_Q = Y_Q - \widehat{M}_Q. \tag{14}$$

Each row of $\widehat{M}_Q$ estimates the conditional label mean for a query example, and each row of E_Q is its estimated label residual. Query labels are used to compute E_Q but do not enter the KRR fit.

Let $(K_U^{QQ})_{ij} = k_U(U_i, U_j)$ for $i, j \in Q$, and let $\odot$ denote elementwise matrix multiplication. Substituting the query residuals into the joint-kernel criterion gives

$$\widehat{\mathcal{D}}(\theta) = \operatorname{tr}\Big[E_Q^\top W_Q \big(K_U^{QQ} \odot K_T^{QQ}\big) W_Q E_Q\Big]. \tag{15}$$

The normalized weights account for averaging over query subjects, so no additional factor is needed in this expression. The KRR solve and the task-kernel matrices are differentiable with respect to the trainable representation (Bertinetto et al., 2019), allowing the empirical regularizer to guide adaptation through both residual estimation and moment evaluation.

3.4 Joint Optimization and Inference

The kernel regularizer encourages T_θ to account for label information available in the frozen reference. The subject loss is

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{CE}}(g_\omega(T_\theta), Y) + \lambda\, \widehat{\mathcal{D}}(\theta), \tag{16}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the classification cross-entropy, $\lambda \geq 0$ controls the kernel regularizer. The true class label is used by the discriminator only during training. At inference, the frozen reference, KRR estimator, and subject discriminator are discarded; prediction uses $g_\omega(f_\theta(X))$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate the proposed method on three public EEG datasets: PhysioNet-MI (Schalk et al., 2004; Goldberger et al., 2000), Mental Arithmetic (MA) (Zyma et al., 2019), and ISRUC (Khalighi et al., 2016), corresponding to motor imagery, mental arithmetic state classification, and sleep staging, respectively. These tasks differ in their neural activity patterns and temporal scales, and span binary and multiclass classification settings. Together, they provide complementary settings for evaluating the method under different classification demands. For all datasets, the training, validation, and test splits are subject-disjoint.

Baselines. We use REVE, LaBraM, and CBraMod as pretrained EEG foundation models. Under a unified data split and evaluation protocol, we compare linear probing (LP), full fine-tuning (FT), LoRA, Adapter, and the task-oriented adaptation methods TASTEFUL and PEARL. To assess the applicability of ReSuffacross adaptation methods, we incorporate it as a representation regularizer during training where applicable, and compare each method with itsReSuff -augmented counterpart. Within each pair, we keep the pretrained backbone, data split, random seed, adaptation architecture, prediction head, and optimization settings fixed. For parameter-efficient methods, we also match the trainable parameter budget, including the LoRA rank or Adapter bottleneck dimension. Thus, differences within each pair primarily reflect the effect of ReSuffon task representation learning rather than changes in model capacity or experimental protocol.

Evaluation metrics. Following prior evaluations of EEG foundation models, we report balanced accuracy (BA) (Brodersen et al., 2010) as the primary metric for all datasets. For the multiclass PhysioNet-MI and ISRUC tasks, we additionally report Cohen's κ(Cohen, 1960); for the binary MA task, we report the area under the precision–recall curve (AUPRC) to assess positive-class performance under class imbalance(Saito & Rehmsmeier, 2015). All results are reported as the mean and standard deviation over three random seeds.

Implementation details. We use publicly available pretrained weights for REVE, LaBraM, and CBraMod, following the input processing pipeline of each backbone. In the main experiments, we use TASTEFUL for task adaptation and define T as the representation passed to the final prediction head. ReSuff regularizes the learning of this representation without directly regularizing the classification head. For each backbone–dataset pair, we select the TASTEFUL parameter budget based on the baseline's validation performance and keep the budget and all other training settings fixed when adding ReSuff. To compute ReSuff, we partition the subjects in the current batch into disjoint support and query sets, assigning 67% of the subjects to the support set. Kernel ridge regression estimates the conditional label mean given T on the support set. Label residuals on the query set, together with the task-representation kernel and the frozen-reference-feature kernel, define the regularizer. The ridge coefficient is set to 0.01. RBF kernel bandwidths are estimated from the median off-diagonal pairwise distance (Garreau et al., 2017), with a lower bound of 10^{-6}. The regularization weights for PhysioNet-MI, MA, and ISRUC are 0.5, 0.02, and 0.01, respectively, with a five-epoch warm-up. Gradients from the regularizer propagate to trainable parameters only through T. ReSuffis is used during training and incurs no additional computation at inference.

4.2 EXPERIMENTAL RESULTS

Overall performance. ReSuff achieves mean BA values of 0.730, 0.685, and 0.659 on REVE, LaBraM, and CBraMod, respectively, ranking first numerically among frozen-backbone adaptation methods for each backbone. Relative to TASTEFUL with the same adaptation architecture, these correspond to improvements of approximately 2.2%, 2.1%, and 6.3% in mean BA. Both BA and Cohen's κ improve across all nine backbone–dataset combinations. On REVE and LaBraM, ReSuff also achieves mean BA comparable to full fine-tuning.

The baseline results show that the utility of pretrained information depends on both the backbone and the adaptation method. Linear probing achieves 0.795 BA on MA with REVE but only 0.500 with LaBraM. Increasing adaptation capacity does not necessarily close the gap. On PhysioNet-MI with LaBraM, LoRA reaches at most 0.506 BA as its rank varies from 1 to 32, while Adapter reaches at most 0.529 as its bottleneck dimension varies from 4 to 126; ReSuff achieves 0.559. On MA with CBraMod, the highest BA values across the tested budgets are 0.634 for LoRA and 0.619 for Adapter, compared with 0.678 for ReSuff.

Table 1: Comparison under the limited-label setting. Results are reported as mean ± standard deviation over three seeds. Avg. BA is the unweighted average of the dataset-wise mean BA values. Bold and underlined values indicate the best and second-best performance, respectively, among frozen-backbone adaptation methods within each backbone. Full fine-tuning is excluded from ranking.

Backbone	Method	PhysioNet-MI		MA		ISRUC		Avg. BA
		BA	Kappa	BA	Kappa	BA	Kappa	
Non-pretrained	EEGNet	.582±.011	.442±.014	.596±.026	.184±.047	.720±.032	.658±.064	.633
	EEGConformer	.581±.009	.441±.011	.627±.038	.254±.126	.776±.010	.706±.020	.661
	ShallowConvNet	.566±.010	.422±.014	.603±.037	.191±.138	.735±.009	.704±.011	.635
REVE	Full fine-tuning	.623±.007	.498±.010	.772±.013	.464±.091	.786±.009	.756±.021	.727
	Linear probing	.573±.006	.431±.008	**.795±.030**	**.600±.062**	.681±.010	.658±.010	.683
	LoRA	.609±.003	.479±.003	.633±.022	.315±.083	.775±.011	.743±.011	.672
	Adapter	.634±.016	.511±.021	.575±.073	.150±.130	.776±.009	.741±.011	.662
	PEARL	.579±.015	.439±.020	.583±.072	.222±.192	.677±.004	.651±.006	.613
	TASTEFUL	.630±.015	.507±.021	.731±.054	.473±.055	.780±.006	.745±.014	.714
	tPoE-EIB	.391±.077	.188±.103	.677±.089	.334±.157	.662±.058	.568±.104	.577
	ReSuff	**.641±.009**	**.522±.012**	.764±.035	.547±.038	**.786±.007**	**.749±.014**	**.730**
LaBraM	Full fine-tuning	.612±.018	.483±.024	.635±.053	.246±.071	.782±.012	.754±.002	.676
	Linear probing	.284±.005	.046±.006	.500±.000	.000±.000	.652±.002	.619±.002	.479
	LoRA	.506±.018	.341±.023	.525±.011	.067±.029	.763±.007	.740±.007	.598
	Adapter	.529±.020	.372±.026	.714±.033	.392±.056	**.779±.009**	**.751±.007**	.674
	PEARL	.502±.010	.336±.013	.612±.025	.278±.048	.719±.006	.691±.009	.611
	TASTEFUL	.548±.025	.398±.033	.701±.044	.362±.078	.765±.004	.743±.006	.671
	tPoE-EIB	.250±.000	.000±.000	.500±.000	.000±.000	.704±.003	.658±.007	.485
	ReSuff	**.559±.027**	**.413±.036**	**.727±.020**	**.420±.046**	.770±.004	.746±.011	**.685**
CBraMod	Full fine-tuning	.614±.010	.485±.013	.648±.052	.328±.127	.768±.003	.739±.006	.677
	Linear probing	.540±.017	.386±.023	.605±.017	.233±.069	.396±.059	.286±.088	.514
	LoRA	.552±.015	.403±.020	.620±.014	.258±.034	.656±.029	.547±.055	.609
	Adapter	**.577±.012**	**.437±.016**	.619±.061	.252±.148	.711±.012	.715±.009	.636
	PEARL	.560±.002	.414±.002	.654±.059	**.355±.148**	.760±.005	.728±.001	.658
	TASTEFUL	.505±.017	.341±.022	.613±.067	.177±.093	.741±.016	.710±.012	.620
	tPoE-EIB	.250±.001	.000±.000	.500±.000	.000±.000	.724±.017	.673±.021	.491
	ReSuff	.536±.010	.381±.014	**.678±.061**	.288±.085	**.762±.020**	**.742±.016**	**.659**

MA denotes Mental Arithmetic. Full fine-tuning updates the entire backbone, whereas the remaining pretrained methods keep the original backbone parameters frozen.

Task-oriented methods also vary across settings. TASTEFUL and PEARL perform above chance in all evaluated combinations, whereas tPoE-EIB achieves 0.662–0.724 BA on ISRUC but falls to chance-level BA on PhysioNet-MI and MA with LaBraM and CBraMod. No adaptation method dominates every backbone and task. Against this variation, ReSuff's consistent gains over TASTEFUL suggest that encouraging task representations to capture label-relevant information still available in pretrained features can improve downstream adaptation.

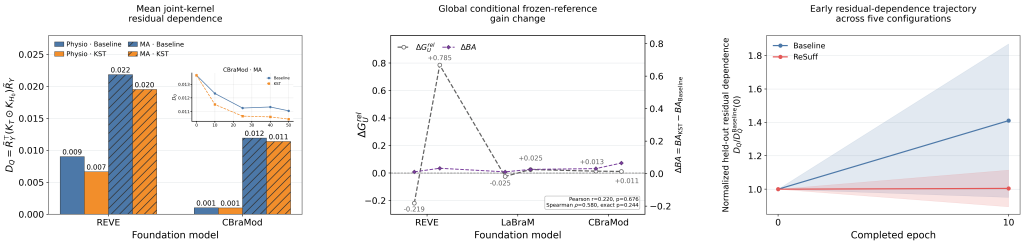
Compatibility across adaptation methods. Table 2 examines the compatibility of ReSuff with different adaptation strategies. Adding ReSuff improves mean BA for all four methods. Even under full fine-tuning and TASTEFUL, which explicitly learns task representations, mean BA increases from 0.698 to 0.716 and from 0.681 to 0.703, respectively; on Mental Arithmetic, their Cohen's κ values increase from 0.464 to 0.549 and from 0.473 to 0.547. These results suggest that updating all model parameters or introducing a task-oriented architecture does not necessarily exhaust the task-relevant information available in pretrained features. However, LoRA shows slight declines on PhysioNet-MI, indicating that the residual-dependence regularizer does not benefit every combination. Overall, the varying responses across tasks and methods suggest that improving how task representations use pretrained information is a useful objective alongside the choice of parameter-update strategy.

4.3 Mechanistic Analysis of ReSuff

Why condition on the frozen reference? A reduction in the magnitude of label residuals alone does not establish that the adapted representation has absorbed information from the pretrained reference. To distinguish ReSuff's conditional-dependence objective from a generic pressure to fit labels, we compare it with *query-MSE*, which penalizes the squared label residual on query subjects without testing its relationship to (U, T). Both methods use the same adaptation architecture. As shown in Table 3, query-MSE improves balanced accuracy on CBraMod but reduces it on REVE

Table 2: Effect of adding ReSuff to different adaptation methods on REVE. Results are reported as mean ± standard deviation over three seeds. BA denotes balanced accuracy; AUPRC denotes the area under the precision–recall curve. Avg. BA is the unweighted average of the two mean BA values. Bold indicates the higher mean within each method pair.

Method	PhysioNet-MI			Mental Arithmetic			Avg. BA
	BA	Kappa	AUPRC	BA	Kappa	AUPRC	
Full fine-tuning	0.623 ± 0.007	0.498 ± 0.010	0.698 ± 0.007	0.772 ± 0.013	0.464 ± 0.091	0.749 ± 0.044	0.698
+ ReSuff	$\mathbf{0.635} \pm 0.055$	$\mathbf{0.514} \pm 0.074$	$\mathbf{0.705} \pm 0.067$	$\mathbf{0.796} \pm 0.020$	$\mathbf{0.549} \pm 0.095$	$\mathbf{0.782} \pm 0.011$	**0.716**
Adapter	0.634 ± 0.016	0.511 ± 0.021	0.707 ± 0.025	0.575 ± 0.073	0.150 ± 0.130	0.456 ± 0.196	0.605
+ ReSuff	$\mathbf{0.652} \pm 0.016$	$\mathbf{0.536} \pm 0.021$	$\mathbf{0.726} \pm 0.005$	$\mathbf{0.587} \pm 0.076$	$\mathbf{0.194} \pm 0.173$	$\mathbf{0.471} \pm 0.235$	**0.620**
LoRA	$\mathbf{0.609} \pm 0.003$	$\mathbf{0.479} \pm 0.003$	$\mathbf{0.667} \pm 0.005$	0.633 ± 0.022	0.315 ± 0.083	0.551 ± 0.132	0.621
+ ReSuff	0.608 ± 0.009	0.478 ± 0.012	0.664 ± 0.001	$\mathbf{0.669} \pm 0.114$	$\mathbf{0.330} \pm 0.234$	$\mathbf{0.584} \pm 0.143$	**0.639**
TASTEFUL	0.630 ± 0.015	0.507 ± 0.021	0.669 ± 0.067	0.731 ± 0.054	0.473 ± 0.055	0.695 ± 0.121	0.681
+ ReSuff	$\mathbf{0.641} \pm 0.009$	$\mathbf{0.522} \pm 0.012$	$\mathbf{0.683} \pm 0.061$	$\mathbf{0.764} \pm 0.035$	$\mathbf{0.547} \pm 0.038$	$\mathbf{0.740} \pm 0.065$	**0.703**



(a) Held-out D_Q. (b) ΔG_U and BA. (c) Early D_Q (norm.).

Figure 3: Complementary mechanistic diagnostics. (a) ReSuff directly targets lower query–subject residual dependence. (b) Change in incremental frozen-reference prediction gain vs. change in native balanced accuracy. (c) Normalized dependence over the first ten epochs across five configurations. Because these diagnostics address different questions, they may not improve together.

and LaBraM. ReSuff achieves the highest balanced accuracy and Cohen's κ on all three backbones, although query-MSE yields a slightly higher AUPRC on REVE. These results suggest that minimizing residual magnitude alone does not account for ReSuff's gains; targeting residual dependence on the pretrained reference provides an additional benefit.

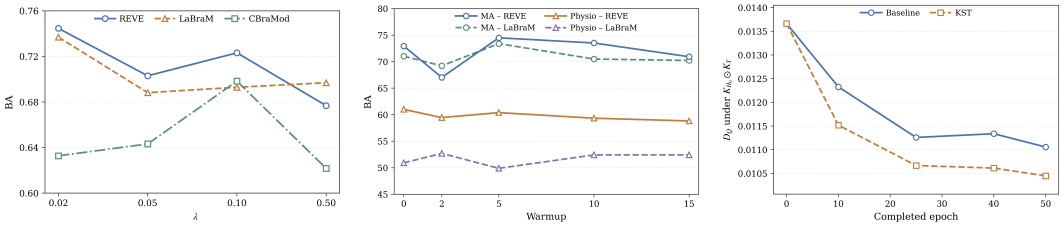
Held-out residual dependence. We evaluate the joint-kernel residual moment D_Q on query subjects excluded from the conditional-mean fit. As shown in Fig. 3a, ReSuff reduces this statistic on both REVE tasks and yields a smaller reduction for CBraMod on Mental Arithmetic. For CBraMod on PhysioNet-MI, the baseline value is already close to 0.001, leaving limited measurable room for improvement. The inset further shows that the CBraMod–Mental Arithmetic difference is not confined to a single checkpoint. Overall, the reductions on query subjects support ReSuff's intended effect of diminishing label-residual dependence on the frozen reference, with the observable gain depending on the baseline level.

Incremental gain from the frozen reference. To assess whether the task representation still leaves usable evidence in the frozen reference, we fit a predictor q_T using the adapted representation and an augmented predictor q_{TU} that also receives U, under the same held-out-subject protocol. Their NLL difference, $G_U = \mathrm{NLL}(q_T) - \mathrm{NLL}(q_{TU})$, measures the additional predictive gain from U. As shown in Fig. 3b, most configurations exhibit a positive relative reduction in this gain, and some also improve validation BA. This provides prediction-level evidence consistent with ReSuff encouraging greater use of pretrained task information in the adapted representation. However, the overall change was not so significant, motivating our subsequent analysis of favorable and unfavorable cases.

Task Representation Learning and Predictive Outcomes. EEG downstream tasks are affected by intersubject variability, recording conditions, noise levels, and task-specific factors. Consequently, even when employing full fine-tuning (FT)—as shown in Table 2—performance gains can still be achieved in some settings. Against this variability, ReSuff's effect is not confined to a single backbone or dataset: across multiple configurations, adding ReSuff leads to a more favorable learning trajectory early in training. As shown in Fig. 3c, NLL decreases markedly. The early separation in residual-dependence trajectories further suggests that ReSuff helps prevent a continued increase in

Table 3: Query-MSE control on Mental Arithmetic across three EEG foundation models. Results are mean ± standard deviation over three seeds. Bold indicates the highest mean.

	REVE			LaBraM			CBraMod		
Method	BA	Cohen's κ	AUPRC	BA	Cohen's κ	AUPRC	BA	Cohen's κ	AUPRC
TASTEFUL	0.731 ± 0.054	0.473 ± 0.055	0.695 ± 0.121	0.701 ± 0.044	0.362 ± 0.078	0.651 ± 0.076	0.613 ± 0.067	0.177 ± 0.093	0.392 ± 0.074
query-MSE	0.704 ± 0.072	0.465 ± 0.129	$\mathbf{0.752 \pm 0.072}$	0.676 ± 0.028	0.308 ± 0.009	0.579 ± 0.068	0.645 ± 0.060	0.229 ± 0.079	0.393 ± 0.059
ReSuff	$\mathbf{0.764 \pm 0.035}$	$\mathbf{0.547 \pm 0.038}$	0.740 ± 0.065	$\mathbf{0.727 \pm 0.020}$	$\mathbf{0.420 \pm 0.046}$	$\mathbf{0.665 \pm 0.034}$	$\mathbf{0.678 \pm 0.061}$	$\mathbf{0.288 \pm 0.085}$	$\mathbf{0.453 \pm 0.038}$



(a) Validation BA versus λ.

(b) Validation BA versus warm-up length.

(c) Illustrative residual-moment trajectory.

Figure 4: Hyperparameter sensitivity and training dynamics. Validation balanced accuracy varies with the regularization weight and warm-up duration; the displayed trajectory illustrates the effect of ReSuff on the held-out joint-kernel residual moment.

detectable associations between label residuals and the frozen reference features, rather than acting only at the selected final checkpoint.

To examine when this learning trajectory translates into predictive gains, we analyze two contrasting cases. For REVE on PhysioNet-MI, samples identified by the baseline as carriers of residual reference evidence exhibit greater targeted residual absorption, accompanied by improved native-head accuracy; the complementary group does not share this gain. This pattern is consistent with reference-associated task evidence becoming more useful to the adapted representation for those samples. For CBraMod on PhysioNet-MI, the directional residual change is small, and native-head performance does not improve at the plotted cached checkpoints. Together, these cases suggest that constraining residual dependence can guide task learning, while predictive gains depend on whether the resulting change helps the model use task-relevant evidence in the samples being evaluated.

4.4 Sensitivity to Regularization Weight and Warm-up

Regularization weight and warm-up. The regularization weight λ controls the contribution of the estimated residual moment relative to the classification loss. The response depends on the backbone: REVE and LaBraM achieve their highest plotted BA near 0.02, whereas CBraMod favors an intermediate value near 0.10; increasing the weight to 0.50 generally lowers BA in this sweep. Thus, reducing the estimated residual moment more aggressively does not necessarily improve held-out classification. We select the weight separately for each dataset, using 0.50 for PhysioNet-MI and 0.01 for ISRUC in the main experiments. Figure 4(b) examines the warm-up duration before ReSuff reaches its full regularization weight. Applying the full weight too early may interfere with the initial formation of a task-discriminative representation, whereas a prolonged warm-up delays the use of residual evidence. A five-epoch warm-up provides a favorable balance in the displayed validation results and is therefore used throughout the main experiments.

5 Conclusion

We studied whether downstream EEG adaptation uses label information available in a fixed view of pretrained features. ReSuff converts this question into a label-residual joint-kernel regularizer, estimated across disjoint support and query subjects, and can be added to several adaptation strategies without changing their inference architecture. Across three backbones and three limited-label, subject-independent benchmarks, it improves BA over the matched TASTEFUL baseline in all nine combinations. A query-MSE control, held-out residual moments, prediction-level conditional-gain analysis, and case studies jointly indicate that the gains cannot be explained solely by penalizing residual magnitude. Their alignment varies across backbones and checkpoints, however, and finite-sample diagnostics do not establish exact task sufficiency. The results support conditional residual regularization as a practical way to strengthen the use of pretrained EEG information, while leaving reliable cross-subject estimation of remaining task evidence as an open challenge.

AI USE STATEMENT

We used generative AI tools only to improve the language of the manuscript. All manuscript text was reviewed and approved by the human authors. We take responsibility for the final content of this paper.

ETHICS STATEMENT

This work uses previously collected EEG datasets and involves no new recruitment or recording of participants. Because EEG data may contain sensitive information, the use and distribution of data and derived artifacts must follow the applicable dataset terms. We use subject-disjoint evaluation to assess transfer to unseen participants; the reported results do not establish reliable performance across all populations or recording conditions. The method is intended for research and has not been validated for clinical decision-making.

REPRODUCIBILITY STATEMENT

The method section defines the ReSuff objective, kernel estimator, and procedure for constructing support and query sets with disjoint subjects. The experimental section describes the datasets, backbones, baselines, evaluation metrics, and model-selection protocol. The supplementary material details data preprocessing and splits, implementation settings hyperparameters, random seeds, and the procedures used for the diagnostic analyses.

things that can be referenced in the reproducibility statement.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019.

Kay Henning Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, pp. 3121–3124. IEEE, 2010.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10:740, 2023.

Heejae Choi, Hyunseo Kim, and Eun Yi Kim. PEARL: Prompt-based EEG adaptation via resource-efficient learning for generalizable brain-computer interface. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien 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. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

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 *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 2790–2799. PMLR, 2019.

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.

Jaehyun Jeon, Seungwoo Jeong, Yeajin Shon, and Heung Il Suk. Parameter-efficient transfer learning for EEG foundation models via task-relevant feature focusing. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 1068–1077, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://arxiv.org/abs/2405.18765`.

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.

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.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pp. 41–50. PMLR, 2020.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10:e0118432, 2015.

Craig Saunders, Alex Gammerman, and Vladimir Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 515–521. Morgan Kaufmann, 1998.

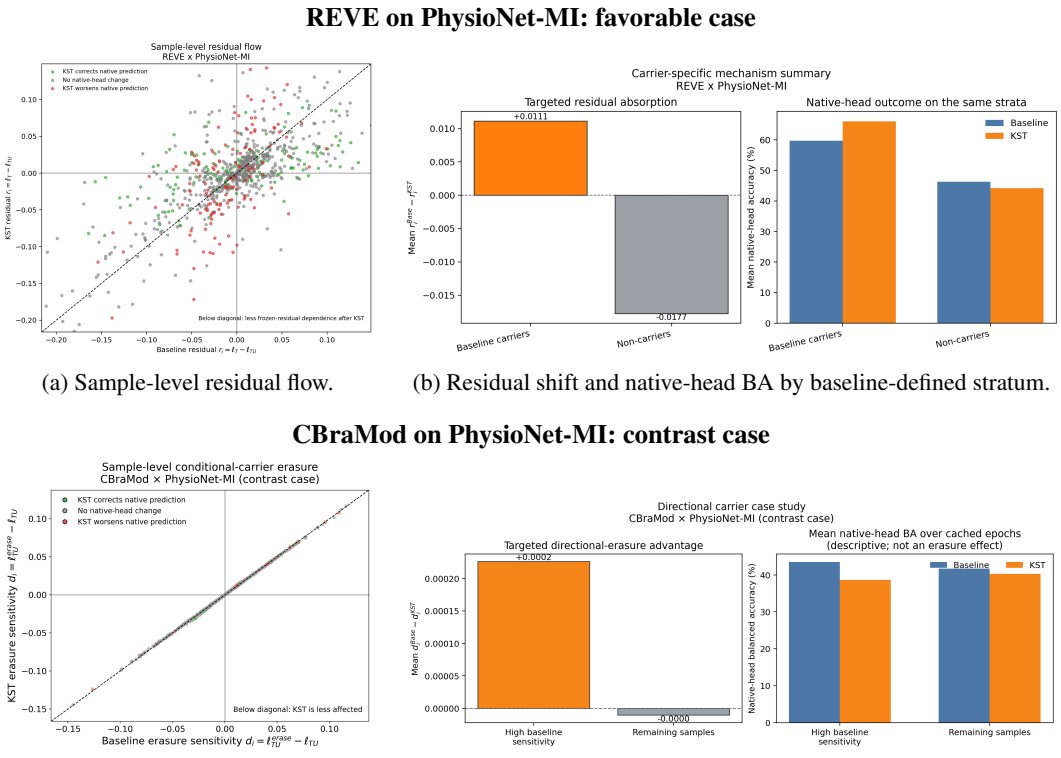
Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

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 preprint arXiv:2605.11410*, 2026.

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.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.



(a) Sample-level residual flow.

(b) Residual shift and native-head BA by baseline-defined stratum.

(c) Sample-level directional erasure.

(d) Directional change and BA across cached epochs.

Figure 5: Illustrative favorable and contrast cases on PhysioNet-MI. For REVE (top), the residual shift among baseline-defined carriers coincides with improved native-head BA, whereas the complementary stratum does not share this gain. For CBraMod (bottom), the diagnostic shift is small and native-head BA does not improve across the displayed cached epochs. The strata and epoch-wise comparisons are descriptive; the latter is distinct from evaluation of the selected final checkpoint.

A Related Work

EEG foundation models. EEG foundation models learn representations intended to transfer across recording setups, subjects, and downstream tasks. LaBraM converts EEG channel patches into discrete neural codes and predicts the codes of masked patches during pretraining (Jiang et al., 2024). CBraMod models spatial and temporal dependencies separately and learns through masked patch reconstruction (Wang et al., 2025). Recording heterogeneity has also motivated architectures that accommodate different electrode configurations. REVE supports varying electrode arrangements and signal lengths through positional encoding and was pretrained on over 60,000 hours of EEG from 92 datasets and approximately 25,000 subjects (El Ouahidi et al., 2025). LUNA maps signals with different electrode configurations into a shared latent representation (Döner et al., 2025). These models provide different sources of transferable features, whose value for a target task ultimately depends on downstream adaptation.

EEG foundation model adaptation. Downstream adaptation must learn task-specific predictions from limited labels while generalizing across subjects (Kumar et al., 2022). Parameter-efficient methods limit trainable updates through adapters, low-rank matrices, or learned prompts (Houlsby et al., 2019; Hu et al., 2022; Lester et al., 2021). EEG-specific approaches further tailor adaptation to the task: TASTEFUL focuses on task-relevant pretrained features (Jeon et al., 2025), while PEARL uses prompt-based adaptation (Choi et al., 2026). These methods improve how pretrained features are selected or updated, but predictive performance alone cannot establish whether the resulting representation captures the label information available from a frozen reference. Kernel conditional-moment methods provide tools for characterizing when a residual has zero conditional mean (Muandet et al., 2020). ReSuff uses this principle as a training objective, encouraging the adapted representation to account