

Learning aligned EEG representations with subject-specific encoders

Bruna J. Lopes^{1,2,3}, Gabriel Schwartz¹, Sylvain Chevallier², Raphael Y. de Camargo^{1,4}, and Bruno Aristimunha^{2,4,5,6}

¹University of São Paulo, São Paulo, Brazil

²Université Paris-Saclay, Inria TAU team, LISN-CNRS, France

³Institut de neuromodulation, GHU Paris, psychiatrie et neurosciences, centre hospitalier Sainte-Anne, pôle hospitalo-universitaire 15, Université Paris Cité, Paris, France

⁴Federal University of ABC (UFABC), Santo André, Brazil

⁵Yneuro, Paris, France

⁶Swartz Center for Computational Neuroscience (SCCN), University of California San Diego, La Jolla, USA

ABSTRACT

Cross-subject EEG decoding promises more training data, but it also exposes neural networks to strong inter-subject distribution shifts. We study whether task supervision and architecture alone can learn subject-aligned representations. We replace a shared EEG encoder with subject-specific encoders followed by a common classifier, and compare this hybrid model with standard EEGNet, AttentionBaseNet, and CTNet baselines with Euclidean Alignment (EA) on four motor-imagery datasets. EA improves shared encoders by recentering subject covariances, but the hybrid encoder largely internalises this role: validation-loss curves and latent-distance analyses change little when EA is removed. Subject-specific heads increase class distinctiveness and place each subject close to its own latent manifold, improving most subjects while leaving a method-sensitive subset. These results support subject-specific encoders as a learned alignment mechanism for EEG decoding and identify head selection for unseen subjects as the remaining bottleneck.

1 Introduction

Brain decoding is an exciting yet challenging application of machine learning that aims at translating recordings of brain activity into their corresponding mental states^{1,2}. These systems have broad applications, including emotion recognition, sleep staging, and brain-computer interfaces (BCIs). In the latter, the objective is often to decode imagined movements from brain signals, typically represented using electroencephalography (EEG) because of its portability, high temporal resolution, and relatively low cost, making it well-suited for practical use³.

Traditional EEG decoding pipelines divide this process into preprocessing, feature extraction, and classification stages, relying heavily on task-specific engineering decisions rather than data-driven representation extraction^{4–6}. Separating feature extraction from the main classification objective may discard task-relevant information before the predictive objective can shape the representation, and could benefit from a more automatic approach. Deep learning techniques are currently driving a shift in this direction⁷. Initially established in computer vision, deep learning is an automatic feature extractor that learns discriminative representations directly from raw data⁸. Applied to EEG, end-to-end architectures jointly optimize representation learning and classification, producing compact, task-relevant embeddings without handcrafted features.

Despite this progress, deep learning performance scales with the volume of available training data⁹. This collides with a primary bottleneck in EEG-based brain decoding: data scarcity¹⁰. Because each subject contributes few EEG trials¹¹, reliable models usually need data from multiple subjects or datasets.

Aggregating EEG data, however, introduces new challenges. Different datasets frequently use distinct recording hardware, montages, and sampling frequencies, making standardization difficult¹². EEG activity patterns themselves depend on the neuroanatomy and cognitive state of the individual at the instant of the recording, making them highly subject-specific^{2,13,14}. Because deep neural networks typically assume that all training and test samples are drawn

from the same underlying distribution, models trained using multiple subjects often suffer from *negative transfer*¹⁵, failing to learn a unified mapping because the training data lie on different manifolds. The field therefore faces a trade-off: multi-subject data is needed for scale, but pooling subjects can damage transfer.

Domain adaptation methods aim to reduce these shifts, with some acting at the preprocessing level^{16–18}. For instance, Riemannian geometry-based techniques^{17–19} apply linear and non-linear transformations on the covariance representation of samples, but operate in covariance space and do not naturally extend to raw time-series inputs. Euclidean Alignment¹⁶ is a technique designed for EEG time series, but it still consists of a simple rigid recentering of the data²⁰, which, while effective, may be insufficient to capture the full complexity of inter-subject variability.

End-to-end training also allows deep domain adaptation (DDA) methods to place alignment inside the learning objective^{21–26}. Zhong et al.²⁴ introduce a CORAL-based loss for minimizing statistical differences, while other methods use an adversarial objective to make representations domain-invariant^{22,23}. These methods, however, assume that target data is available during training, which is not always feasible to obtain. The work of Wei et al.²¹ also explores Deep domain adaptation, developing a Separate-Common-Separate (SCSN) modification to Shallow FBCSP Net²⁷ to extract common features across domain subjects. This work, however, focuses solely on training adaptation, not extending its analysis to a novel target distribution. Closest to our motivation,²⁵ align subjects in the model’s latent space rather than at the input; we instead factorise this latent alignment into per-subject encoders trained end-to-end.

Building on these ideas, we ask whether task supervision alone can induce cross-subject alignment when the architecture explicitly factorizes subject-specific transformations. Inspired by the standard EEG processing pipeline, we decompose the common *feature extractor–classifier* deep learning model structure by introducing subject-specific encoder heads that project each individual’s EEG into a shared latent space, which is then passed to a common classification module. The aim is to learn a subject-tailored filter for each training subject.

This structure imposes a strong inductive bias: each head projects the subject into an individual manifold, while the joint classification objective creates pressure to align those projections for a unified classification space. Beyond cross-subject decoding performance, we aim to study how this architectural and task pressure reshapes the geometry of the latent space and the conditional class distributions, and we examine how such a structure can be adapted at test time for unseen subjects, along with the limitations of the design. We find that EA yields per-subject accuracy gains across backbones, that the subject-specific encoders largely reproduce EA’s recentering from task supervision alone, and that they sharpen class distinctiveness for source subjects while leaving head selection for unseen subjects as the main bottleneck.

2 Methodology

2.1 EEG decoding notation

Let $\mathcal{S} = \{(\mathbf{X}_i, y_i, s_i)\}_{i=1}^N$ denote the sampled EEG dataset. Each EEG trial $\mathbf{X}_i \in \mathbb{R}^{C \times T}$ is a multichannel epoch recorded from C electrodes over T time samples, $y_i \in \mathcal{Y}$ is its class label, and $s_i \in \{1, \dots, K\}$ is its subject index. For subject k , we write the subject-specific sample as

$$\mathcal{S}_k = \{(\mathbf{X}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}.$$

All experiments are binary motor-imagery or motor-execution decoding problems with $\mathcal{Y} = \{0, 1\}$; the concrete event pair is dataset-dependent and is specified in the dataset paragraph below.

EEG decoding can be formulated as a supervised classification problem: learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ that associates brain activity with the corresponding mental state²⁸.

Within a deep learning framework, this mapping is approximated by a parametric model f_θ . Given a training split \mathcal{S}_{tr} , its parameters are optimized by minimizing the empirical risk

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{|\mathcal{S}_{\text{tr}}|} \sum_{(\mathbf{X}_i, y_i) \in \mathcal{S}_{\text{tr}}} \ell(f_\theta(\mathbf{X}_i), y_i), \quad (1)$$

where ℓ is the supervised classification loss. We use this objective before adding subject-specific parameters.

End-to-end EEG classifiers often mimic signal-processing pipelines by separating representation extraction from classification^{29–31}. We write the encoder as

$$e_\phi : \mathcal{X} \rightarrow \mathcal{Z}, \quad (2)$$

which maps a raw EEG trial to a latent representation $z_i = e_\phi(\mathbf{X}_i) \in \mathcal{Z}$. The classifier is

$$h_\psi : \mathcal{Z} \rightarrow \mathcal{Y}, \quad (3)$$

and the full model is

$$f_\theta = h_\psi \circ e_\phi. \quad (4)$$

During training, feature extraction and classification are optimized together, so the objective can preserve task-relevant structure that a manual feature pipeline may discard³².

2.2 Subjects as domains

Following the definition in³³, a *domain* consists of a feature space and a marginal distribution over that space, while a *task* consists of a label space and the predictive function to be learned. We denote by \mathcal{S}_k the finite sample set of subject k and the domain by

$$\Omega_k = (\mathcal{X}, P_k(\mathbf{X})).$$

The associated subject-level prediction task is described by $\tau_k = (\mathcal{Y}, q_k(y | \mathbf{X}))$.

In transfer-learning terms, each leave-one-subject-out fold selects one target subject t and treats all other subjects as sources. We denote the subject index set by $\mathcal{I} = \{1, \dots, K\}$ and the source set for target t by $\mathcal{I}_{-t} = \mathcal{I} \setminus \{t\}$. The source data are $\mathcal{S}_{-t} = \bigcup_{k \in \mathcal{I}_{-t}} \mathcal{S}_k$, while the target data \mathcal{S}_t are held out from model pre-training. The goal is to use the labelled source domains $\{\Omega_k : k \in \mathcal{I}_{-t}\}$ to learn a model that transfers to Ω_t with only a small labelled calibration set from the target subject.

Traditional EEG decoding often assumes that training and testing samples are drawn from the same domain³⁴. EEG signals violate this assumption because physiology, electrode placement, and non-stationarity vary across subjects³⁵.

In a single study, \mathcal{X} and \mathcal{Y} are typically shared across subjects, but the distributions differ. We distinguish two discrepancies:

- **Covariate shift:** the marginal distributions differ, $P_i(\mathbf{X}) \neq P_k(\mathbf{X})$ for subjects $i \neq k$, even when the task remains fixed.
- **Concept shift:** the conditional label distributions differ, $q_i(y | \mathbf{X}) \neq q_k(y | \mathbf{X})$, so the same observed pattern may map to different labels across subjects.

Under these conditions, a single shared encoder e assumes that one input transformation fits every subject. This may fail when inter-subject variability is large.

Figure 1 shows the scale of this gap: the between-subject shift is much larger than the within-subject, between-session shift, even though the task and recording protocol are the same.

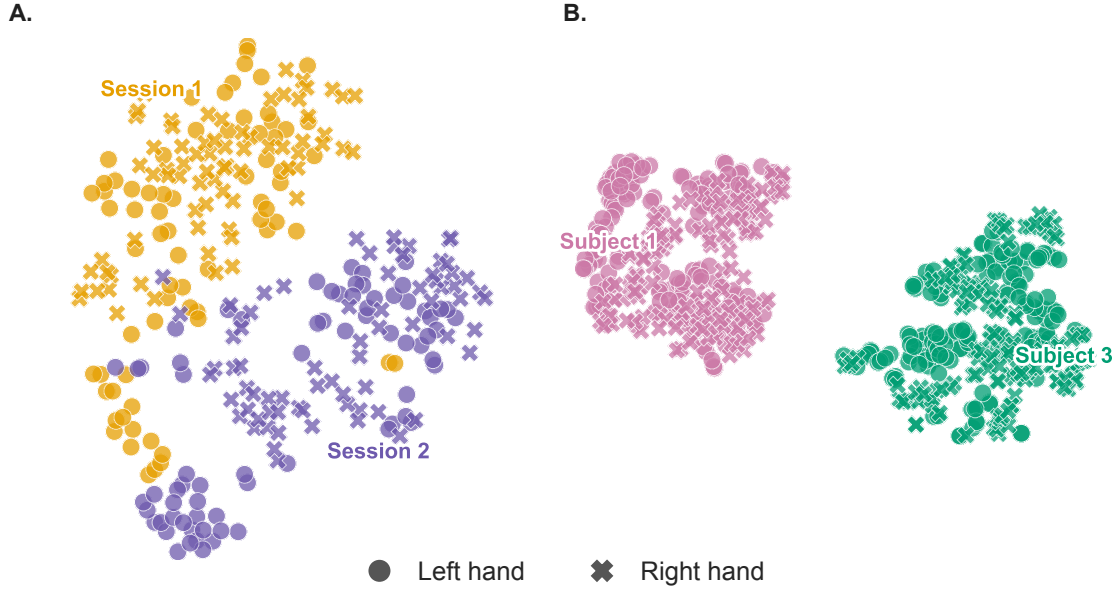


Figure 1. The between-subject domain gap is larger than the within-subject one. Two-dimensional projection of raw spatial-covariance features; marker shape encodes class (circle = left hand, cross = right hand). The two panels use different colour pairs because they compare different things. **(A)** Within-subject: the two sessions of one subject, coloured by the source/target role (orange source session 1 \rightarrow purple target session 2). The sessions overlap heavily; the within-subject covariate shift, while non-zero, is small. **(B)** Between-subject: two different subjects (subject 1 and subject 3), each shown in its own subject colour. The two clouds are clearly separated, even though both subjects performed the same motor-imagery task with the same recording montage. This between-subject separation is the source of the negative-transfer problem motor-imagery decoders face when trained on pooled multi-subject data, and motivates both the alignment and architectural strategies developed in Sections 2.2–3.6.

2.3 Euclidean Alignment

One strategy to mitigate inter-subject shifts is to apply transformations that align the distributions of EEG signals across domains. A widely used approach in EEG-based deep learning pipelines is *Euclidean Alignment* (EA)¹⁶.

EA aims to reduce covariate shift by recentering each subject with respect to its mean covariance matrix³⁶. Let $\mathbf{X}_i^{(k)} \in \mathbb{R}^{C \times T}$ denote the i -th EEG trial recorded from subject k . Each trial is represented by its spatial covariance matrix

$$\Sigma_i^{(k)} = \frac{1}{T} \mathbf{X}_i^{(k)} \mathbf{X}_i^{(k)\top}.$$

The average covariance matrix for subject k is the Euclidean mean

$$\bar{\Sigma}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \Sigma_i^{(k)}. \quad (5)$$

EA then whitens each trial using the inverse square root of this mean covariance matrix:

$$\tilde{\mathbf{X}}_i^{(k)} = (\bar{\Sigma}^{(k)})^{-1/2} \mathbf{X}_i^{(k)}. \quad (6)$$

The transformation is estimated independently for each subject and does not use class labels. In the experiments below, EA is therefore a subject-level preprocessing condition: the *EA* runs use $\tilde{\mathbf{X}}_i^{(k)}$, while the *NoEA* runs use the same preprocessed trials before this whitening step.

Covariance matrices belong to the space of symmetric positive definite (SPD) matrices^{4,37}. Under the affine-invariant Riemannian metric (AIRM) ρ , EA recentres each subject’s covariance distribution around the identity:

$$\rho\left(\Sigma_i^{(k)}, \bar{\Sigma}^{(k)}\right) = \rho\left(\left(\bar{\Sigma}^{(k)}\right)^{-1/2} \Sigma_i^{(k)} \left(\bar{\Sigma}^{(k)}\right)^{-1/2}, I_C\right). \quad (7)$$

Equation 7 is written in covariance space, but the same whitening matrix can be applied directly to the EEG trials. EA can therefore be inserted before a deep EEG model.

2.4 Multi-head architecture

Although Euclidean Alignment reduces inter-subject variability, it only recenters the covariances, and may leave other subject-specific patterns unchanged¹⁹.

To address this, we explore whether the architecture itself can learn subject-specific transformations during end-to-end training. Inspired by the standard *feature extractor-classifier* model structure commonly used in EEG decoding^{30,38}, we replace the shared feature extraction encoder with a set of subject-specific encoders, while keeping a single shared classifier:

$$f_{\theta}^{(k)}(\mathbf{X}) = h_{\psi}\left(e_{\phi_k}^{(k)}(\mathbf{X})\right),$$

where $e_{\phi_k}^{(k)}$ is the encoder specialized for subject k , and h_{ψ} is the classifier shared across all subjects. Figure 2 shows a diagram of the proposed structure. For a target fold t , the trainable hybrid parameter set is

$$\Theta_{-t} = \{\psi, \{\phi_k : k \in \mathcal{I}_{-t}\}\}.$$

The shared baseline is recovered by replacing the encoder bank with a single encoder e_{ϕ} used for every source subject.

Each encoder can learn nonlinear subject-specific transformations of the raw EEG. The shared classifier keeps those representations compatible with one decision boundary, so the classification loss can act as an implicit alignment signal.

Using this framework, we investigate four main questions:

1. Can subject-specific encoders capture characteristic subject-dependent patterns in EEG signals?
2. Does the multi-head structure produce latent representations that are better aligned across subjects?
3. How do these representational changes affect cross-subject decoding performance?
4. How does the proposed structure compare to the original backbone architectures?

2.5 Training and target-subject adaptation

Model training is divided into two stages: pre-training and fine-tuning.

During pre-training, each optimization step samples one mini-batch \mathcal{B}_k per source subject $k \in \mathcal{I}_{-t}$ and routes it through the corresponding subject-specific encoder, as shown in Figure 2. The training objective for one target fold is

$$\min_{\Theta_{-t}} \frac{1}{|\mathcal{I}_{-t}|} \sum_{k \in \mathcal{I}_{-t}} \frac{1}{|\mathcal{B}_k|} \sum_{(\mathbf{x}, y) \in \mathcal{B}_k} \ell\left(h_{\psi}\left(e_{\phi_k}^{(k)}(\mathbf{x})\right), y\right). \quad (8)$$

Gradients from all subject losses accumulate in the shared classifier, while each encoder receives gradients only from its own subject. We update the model only after all subject batches have been processed. The delayed update balances the shared classifier gradient across subjects while each encoder specializes in one subject.

Source training yields one shared classifier and one encoder per source subject. The encoders can be reused for their own subjects or transferred to a new subject.

For a held-out target subject, we initialize a target encoder from one of the pretrained source encoders and combine it with the shared classifier. Let $\mathcal{C}_t \subset \mathcal{S}_t$ be the first 48 labelled calibration trials of the target subject and let $\mathcal{T}_t = \mathcal{S}_t \setminus \mathcal{C}_t$ be the held-out target evaluation trials. We select the source head

$$k^*(t) = \arg \max_{k \in \mathcal{I}_t} \text{BA}_{\mathcal{C}_t} \left(h_{\psi} \circ e_{\phi_k}^{(k)} \right), \quad (9)$$

where $\text{BA}_{\mathcal{C}_t}$ is balanced accuracy on the calibration set. The selected encoder initializes the target branch by copying its weights, $\phi_t \leftarrow \phi_{k^*(t)}$. This target branch is then fine-tuned on \mathcal{C}_t and evaluated on \mathcal{T}_t . This is a low-calibration setting: head selection uses only target calibration data, while the oracle head reported in the results is an upper bound.

2.6 Models and data splitting

We use *EEGNet*³⁰, *AttentionBaseNet (AttnBNet)*³⁸, and *CTNet*³⁹ as backbones because all three are compact EEG decoders used for motor imagery, spanning a purely convolutional model, an attention-augmented convolutional model, and a convolutional-transformer model.

EEGNet is a CNN architecture for EEG classification. It uses depthwise and separable convolutions to mimic traditional EEG feature extraction, such as spatial filtering and filter-bank construction, before the final classification layer.

AttentionBaseNet is a compact attention-based backbone for motor-imagery EEG decoding that combines learned temporal filters with an attention module before classification.

CTNet is a convolutional-transformer backbone that uses an *EEGNet*-style convolutional stem to extract local spatial-temporal features, followed by a multi-head self-attention encoder that models global dependencies before classification.

We compare each backbone in its original shared-encoder form and in the hybrid extension, both with and without EA. This separates preprocessing alignment (EA versus NoEA) from architectural alignment (shared encoder versus subject-specific encoders). For the mechanism analyses, we also include a center-loss hybrid variant when available.

Figure 2 illustrates the use of the trained subject-specific heads to initialize a new head for a target subject before fine-tuning.

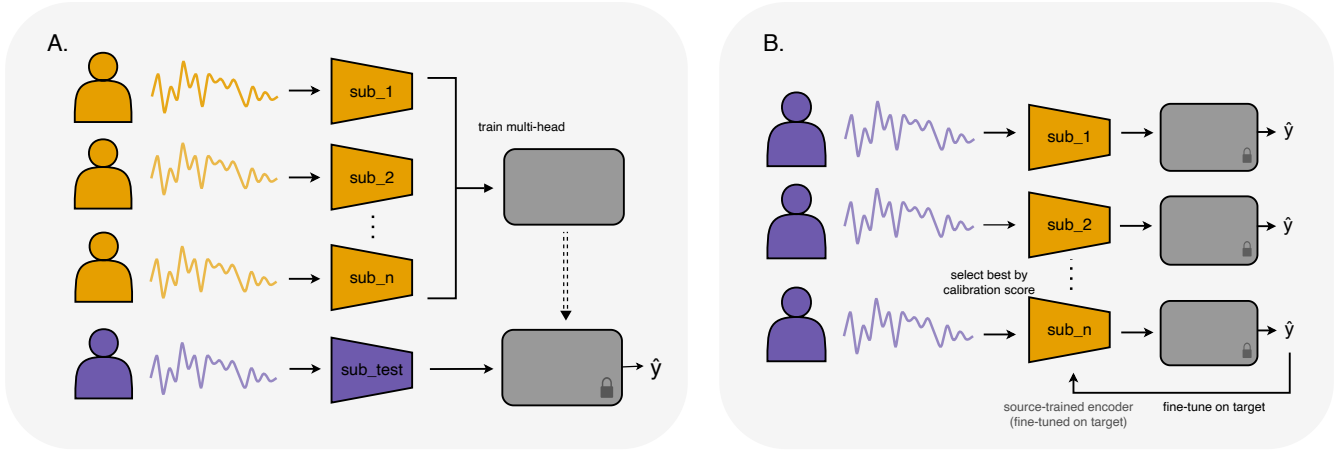


Figure 2. Overview of the proposed adaptation strategy for unseen subjects. After training the multi-head model, each subject-specific encoder is combined with the shared classifier to form an individual initialization candidate. These candidate models are evaluated using calibration data from the target subject, and the best-performing initialization is selected for further fine-tuning.

We train both shared and hybrid models with leave-one-subject-out (LOSO) splits. For each target subject t , no trial from \mathcal{S}_t is used during pre-training. Within each source subject, we split trials into 80% training and 20% validation while preserving the subject assignment. Model checkpoints are selected based on source validation performance; target calibration trials are used only after pre-training for head selection and fine-tuning.

2.7 Datasets

For our experiments, we used dataset IIa from BCI Competition 2014⁴⁰ (*BNCI2014-001*), dataset I from BCI Competition 2015⁴¹ (*BNCI2015-001*), High-Gamma²⁷ (*Schirrmester2017*), and the motor-imagery dataset from Yi et al.⁴² (*Weibo2014*), all available through MOABB⁴³.

BNCI2014-001 is a motor imagery (MI) dataset with four tasks (left hand, right hand, both feet, and tongue) and data from 9 healthy subjects, sampled on 22 channels and with a frequency of 250 Hz. The recordings are divided into two sessions on different days, with each session comprising six runs of 12 trials per class. We use the binary left-hand versus right-hand subset. *BNCI2015-001* is included as a second BCI Competition ERD/BCI benchmark. The MOABB version used here contains 12 subjects and two events, right hand and feet; we therefore use that binary task directly.

Schirrmester2017 contains motor execution (ME) data with four classes (left hand, right hand, both feet, and rest) from 14 healthy subjects. It consists of two sessions with around 880 per subject in total, with variation across subjects. It is divided into training and testing sessions and was originally sampled using 128 channels at 512 Hz. We use the binary left-hand versus right-hand subset. Subject 1 has only 240 trials, so we excluded it and limited the remaining subjects to the first 408 trials to keep batches balanced across subjects.

Weibo2014 contains data from 10 subjects and seven motor imagery conditions in MOABB: left hand, right hand, both hands, feet, compound hand-foot imagery, and rest. We use the same binary left-hand versus right-hand subset used for the other left/right datasets.

Apart from the *Schirrmester2017* trial balancing, we used standard motor-imagery preprocessing^{43,44}: 8–32 Hz band-pass filtering with overlap-add, resampling to 250 Hz, and selecting the 22 *BNCI2014-001* channels⁴⁰ for *Schirrmester2017* and *Weibo2014*.

2.8 Evaluation metrics and representation diagnostics

The primary decoding metric is balanced accuracy on the held-out target trials,

$$\text{BA} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \quad (10)$$

which weights the two classes equally and makes the score comparable across datasets and subjects. For the hybrid model, `+hybrid` denotes the head selected by Eq. 9, whereas `+hybrid*` denotes the oracle head selected by the target test score. The oracle is reported only to quantify how much performance remains available to a better non-oracle head-selection rule. We summarise paired across-subject comparisons with the Wilcoxon signed-rank test, report Wilson score intervals on win-rates, and use the matched-pairs rank-biserial correlation r as an effect size; within each family of comparisons, p-values are Holm-corrected for multiple testing.

To analyse what the models learn beyond accuracy, we compute diagnostics on trial-level covariance matrices. For a representation map g (raw signal, shared encoder, or a subject-specific encoder), the covariance of a trial i is

$$\Gamma_i^g = \frac{1}{T_g} g(\mathbf{X}_i) g(\mathbf{X}_i)^\top,$$

where T_g is the temporal length of that representation. Class distinctiveness⁴⁵ extends the Fisher criterion to covariance matrices using the same affine-invariant Riemannian metric ρ on the SPD manifold. For two classes $\mathcal{K}_0, \mathcal{K}_1$ with Riemannian means M_0, M_1 and intra-class dispersions $\sigma(\mathcal{K}_0), \sigma(\mathcal{K}_1)$,

$$\text{ClassDis}(\mathcal{K}_0, \mathcal{K}_1) = \frac{\rho(M_0, M_1)}{\frac{1}{2}(\sigma(\mathcal{K}_0) + \sigma(\mathcal{K}_1))}, \quad \sigma(\mathcal{K}_c) = \frac{1}{|\mathcal{K}_c|} \sum_{\Sigma \in \mathcal{K}_c} \rho(\Sigma, M_c). \quad (11)$$

Large values indicate classes whose centroids are far apart and whose internal spread is small.

3 Results

3.1 Euclidean Alignment recenters subjects and translates to per-subject accuracy gains

We begin by separating the mechanism through which EA acts from the effect it has on real decoding. Figure 3 reports both. Panel A simulates two subjects whose covariances sit at different points on the SPD manifold; EA’s whitening collapses their distributions to the same center. Panel B then carries the same intervention to the empirical data: for the shared-encoder baseline on three backbones, we report the per-subject paired NoEA \rightarrow EA test-accuracy delta. EA wins on the majority of subjects on every backbone, with median improvements of +6.6 to +7.8 percentage points (Wilcoxon signed-rank $p < 10^{-5}$ on every backbone, Holm-corrected; rank-biserial $r = 0.80\text{--}0.91$). The mechanism in A and the per-subject effect in B answer two distinct objections (*does EA do something geometrically meaningful?* and *does it transfer to subject-level accuracy?*) and form the baseline against which all later architectural changes are compared.

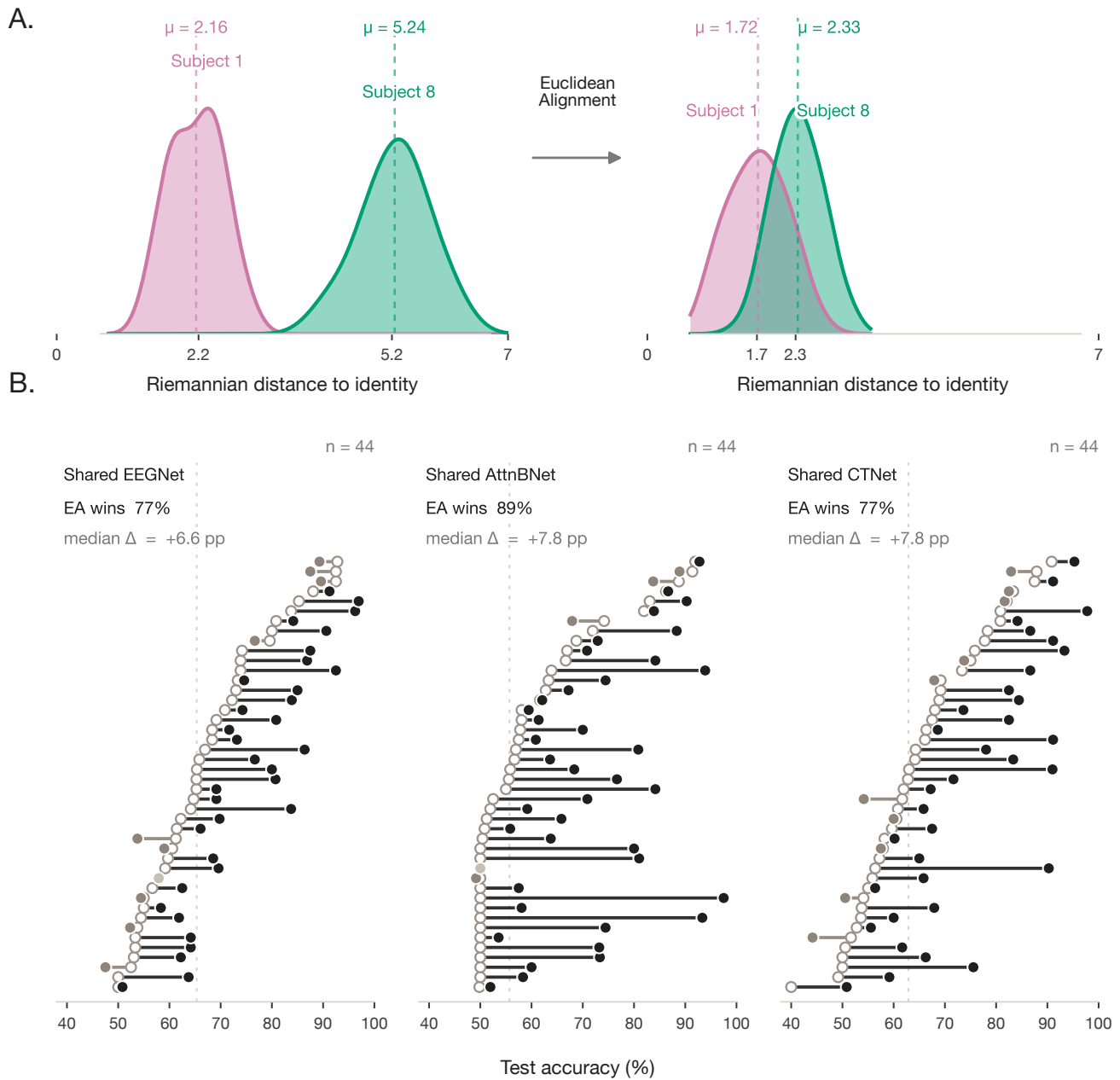
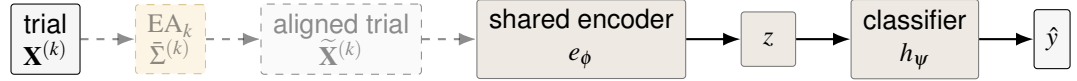


Figure 3. Euclidean Alignment recenters each subject's covariance onto a shared reference, and this translates to per-subject accuracy gains. (A) Mechanism. Riemannian distance-to-identity for two simulated subjects ($n = 200$ trials each). Without EA, the subjects sit in separated distance regimes ($\mu_{S1} = 2.16$, $\mu_{S8} = 5.24$). EA whitens each subject's trials with the inverse square root of its own Riemannian mean; both distributions collapse onto a shared regime around the identity ($\mu_{S1} = 1.72$, $\mu_{S8} = 2.33$). **(B) Effect.** Per-subject NoEA \rightarrow EA test accuracy for the shared-encoder baseline on three backbones (EEGNet, AttnBNet, CTNet). Each row is one subject; the open grey circle marks NoEA and the filled circle marks EA, joined by a segment drawn in dark ink where EA improves accuracy ($\Delta_{EA-NoEA} > 0$) and in grey where it does not. Rows are sorted by NoEA accuracy. EA wins on a majority of subjects on every backbone (77%, 89%, 77%; Wilson 95% CIs 63–87, 76–95, 63–87), with median improvements of +6.6, +7.8, and +7.8 percentage points.

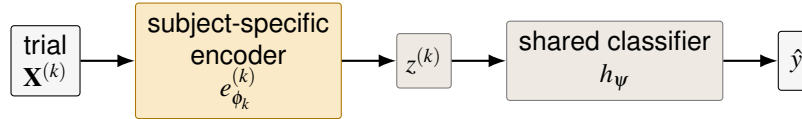
3.2 From preprocessing to architecture: making alignment learnable inside the model

Figure 3 confirms that EA preprocessing yields per-subject accuracy gains, but EA is a fixed transformation applied separately to inputs. A natural follow-up question is whether subject-specific transformations can be *learned* end-to-end as part of the model. We pursue this by replacing the shared feature encoder with the multi-head architecture of Section 2.2: each subject’s trials are routed through a dedicated encoder $e_{\phi_k}^{(k)}$ before reaching the shared classifier h_{ψ} . The schematic in Figure 4 shows the decomposition $f_{\theta}^{(k)}(\mathbf{X}) = h_{\psi}(e_{\phi_k}^{(k)}(\mathbf{X}))$ and the mechanism by how the per-subject encoders could learn some similar transformation by specializing on filtering individual patterns.

Fixed preprocessing



Learned subject-specific alignment



The hybrid model removes the explicit EA_k block and lets $e_{\phi_k}^{(k)}$ learn the subject-conditioned transformation needed by the shared classifier.

Figure 4. The hybrid architecture replaces EA’s fixed preprocessing step with a learned per-subject encoder.

Given a trial \mathbf{X} from subject k , a subject-specific encoder $e_{\phi_k}^{(k)}$ maps the signal into a shared latent space, and a single classifier h_{ψ} produces the prediction. This factorization lets the encoder learn a subject-conditioned transformation that plays a role similar to $EA_k(\mathbf{X})$ through the supervised loss.

3.3 Hybrid gains over the plain EA baseline are dataset- and backbone-dependent

Before turning to the mechanism, we report the cross-dataset performance results. Figure 5 compares each plain backbone trained under Euclidean Alignment (EEGNet, AttnBNet, or CTNet³⁹ + EA) against its hybrid extension on four motor-imagery datasets, and against an oracle ceiling +hybrid* that selects each subject’s best test-set head (an upper bound, not a deployable policy). The vertical guide within each backbone block marks that block’s plain-baseline EA median, so Δ values read directly as the hybrid lift over the same backbone trained without subject-specific heads.

We can observe that the lift is not uniform across datasets or backbones. BNCI2015-001 is the one dataset where the hybrid model wins on all three backbones ($\Delta = +5.3, +7.4, +4.6$ pp for EEGNet, AttnBNet, CTNet), and the oracle ceiling lifts each further (to $+14.6, +9.2, +12.2$ pp). The other three datasets are mixed. On Weibo2014 the EEGNet and AttnBNet hybrids gain ($+8.6$ and $+3.3$ pp) but the CTNet hybrid collapses (-7.2 pp) with an oracle that stays below baseline (-3.7 pp), indicating a genuine failure of the CTNet hybrid representation on this dataset rather than a head-selection artefact. On Schirrmeyer2017 the hybrid model is roughly flat on every backbone (-2.5 to $+1.4$ pp), and the oracle only recovers to about baseline. On BNCI2014-001 the EEGNet and CTNet hybrids underperform their plain baselines (-7.1 and -4.2 pp) while AttnBNet gains ($+5.4$ pp); for both losing backbones the oracle nearly closes the gap ($+1.2$ and $+2.1$ pp), implicating validation-driven head selection rather than the representation itself. The per-cell Δ values are a descriptive map of where the lift concentrates rather than independent confirmatory tests; the confirmatory comparison pools subjects within each backbone across the four datasets (one paired test per backbone, Holm-corrected across the three). This split is significant only for AttnBNet, whose hybrid improves over its EA baseline (median $+1.9$ pp, Wilcoxon $p = 0.014$, Holm $p = 0.042$); the EEGNet hybrid is indistinguishable from baseline ($+1.2$ pp, $p = 0.26$) and the CTNet hybrid trends downward (-4.2 pp, $p = 0.06$). The remaining sections decompose where the lift comes from (class distinctiveness,

head-subject specificity, latent geometry) and where it breaks (subject heterogeneity, head-selection failure modes, and backbone-specific collapse).

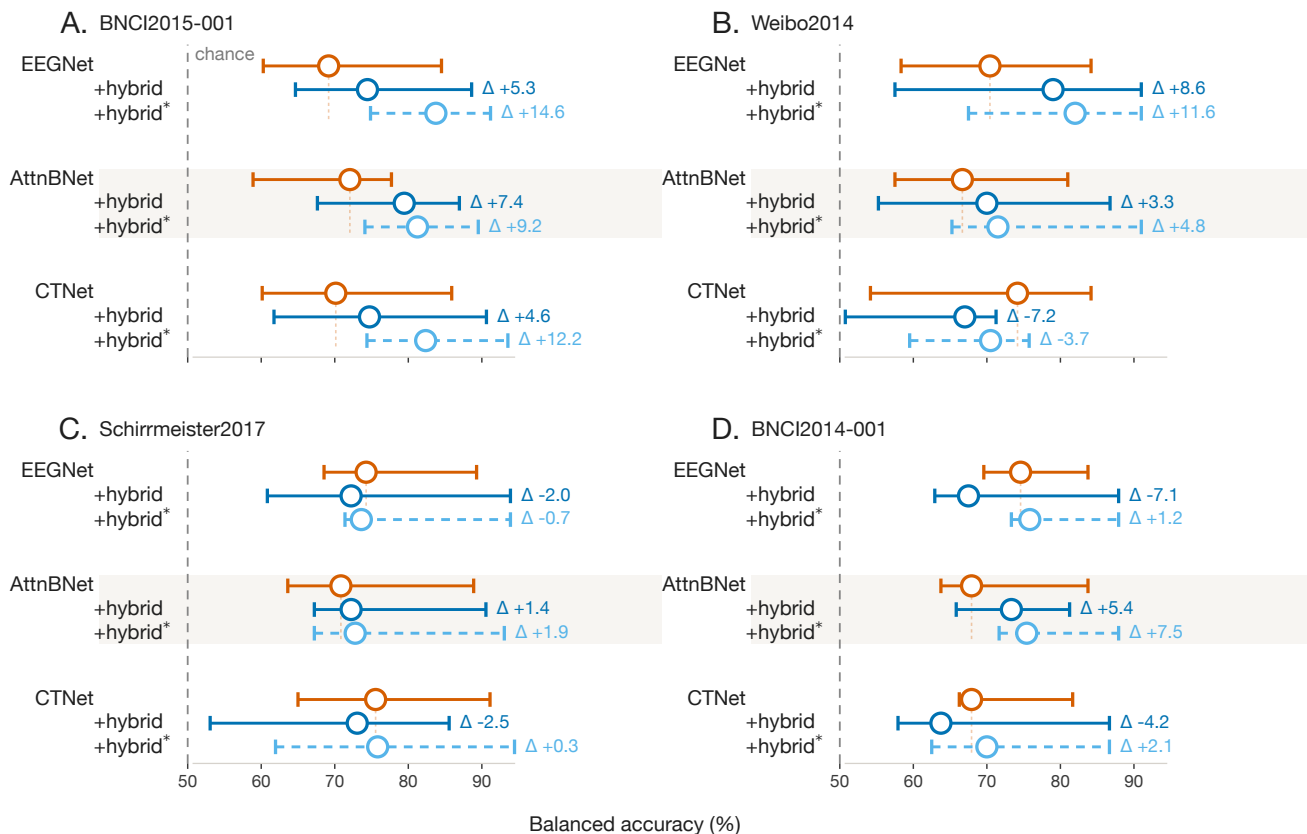


Figure 5. Per-subject balanced accuracy of the proposed +hybrid extension and its oracle ceiling +hybrid*, compared to plain baselines, across three backbones (EEGNet, AttnBNet, CTNet) on four motor-imagery datasets under Euclidean Alignment. Open circles mark the subject median; horizontal bars span the inter-quartile range; the dashed line marks chance accuracy. The thin vermilion vertical guide within each backbone block marks that block’s plain-baseline EA median; the Δ annotation on each +hybrid (and +hybrid*) row reports median balanced accuracy minus that baseline median, both measured under EA. The oracle variant selects each subject’s best test-set head and therefore serves only as an upper bound for any validation-based head selection. Method labels appear on the left column of each row and apply to the right-column panel in the same row. Datasets are arranged in reading order (top-left to bottom-right) by the average +hybrid – baseline gap across the three backbones, so the largest hybrid wins appear in the top-left panel.

3.4 Class distinctiveness and the role of subject-specific encoders

A model can only classify what its representations separate. Classification accuracy alone confounds two distinct competences: how well the encoder produces *distinct* class patterns, and how well the downstream classifier exploits them⁴⁵. To isolate the first, we use the class-distinctiveness diagnostic introduced in Eq. 11, a classifier-independent quantity computed directly on the trial-level covariances of the learned representations.

Two properties make this metric well-suited to our diagnostic. First, ρ is affine-invariant, so ClassDis does not depend on the linear classifier head; it scores the representation, not the decision boundary. Second, the metric is computed inside each subject’s own trials, so the geometry of the latent space is isolated from any cross-subject alignment effect (which Figure 3 already accounts for separately).

We compute ClassDis on three representations of the same trials (Figure 6B): the raw spatial covariance; the latent covariance after the shared encoder; and the latent covariance after the hybrid encoder. Since the analysis of

source and target separation would reveal different things about the model, such as how much each encoder learned about their subject vs how much this can generalize to different subjects, we divided the analysis into source and target distinctiveness. For the target on the hybrid model, we didn't use the fine-tuned encoders as we wanted to assess the generalization on the pre-trained versions.

From the figure, we observe that the shared encoder slightly increases the distinctiveness of each source subject's embeddings relative to the raw signal (34 vs 28 under EA; 34 vs 24 without). For the source subjects, both aligned and non-aligned models have approximately the same average class distinctiveness, which could make sense, since EA is simply a translation and shouldn't affect class separation; this is the role of feature extraction to select features that better separate them. For held-out target subjects, however, that value decreases to 16 under EA vs 14 without, which could indicate the generalization of the embedding extraction capability of the encoder is still limited.

The hybrid encoder, however, sharply increases it for source subjects routed through their own head (67 under EA, 39 without), which could indicate a high level of specialization. The difference between the hybrid EA and the regime without EA could be interpreted as additional improvement transformations: the model does not have to deal with aligning covariances, and thus can devote its capacity to non-linear transformations that further improve class separability. For held-out target subjects, distinctiveness falls, yet remains higher than under the shared encoder in both conditions, indicating that subject-specific encoders retain some transferable structure even when applied out of domain.

Two cautions apply to this diagnosis. ClassDis scores the trial-level spatial covariance of the representation, so any separability that lives only in a non-linear classifier head downstream of the covariance read is invisible to it. The metric also requires SPD covariances; if an encoder produces rank-deficient features (e.g., when an output channel is constant across trials), the Riemannian distance is undefined, and we restricted the analysis to representations where this did not occur.

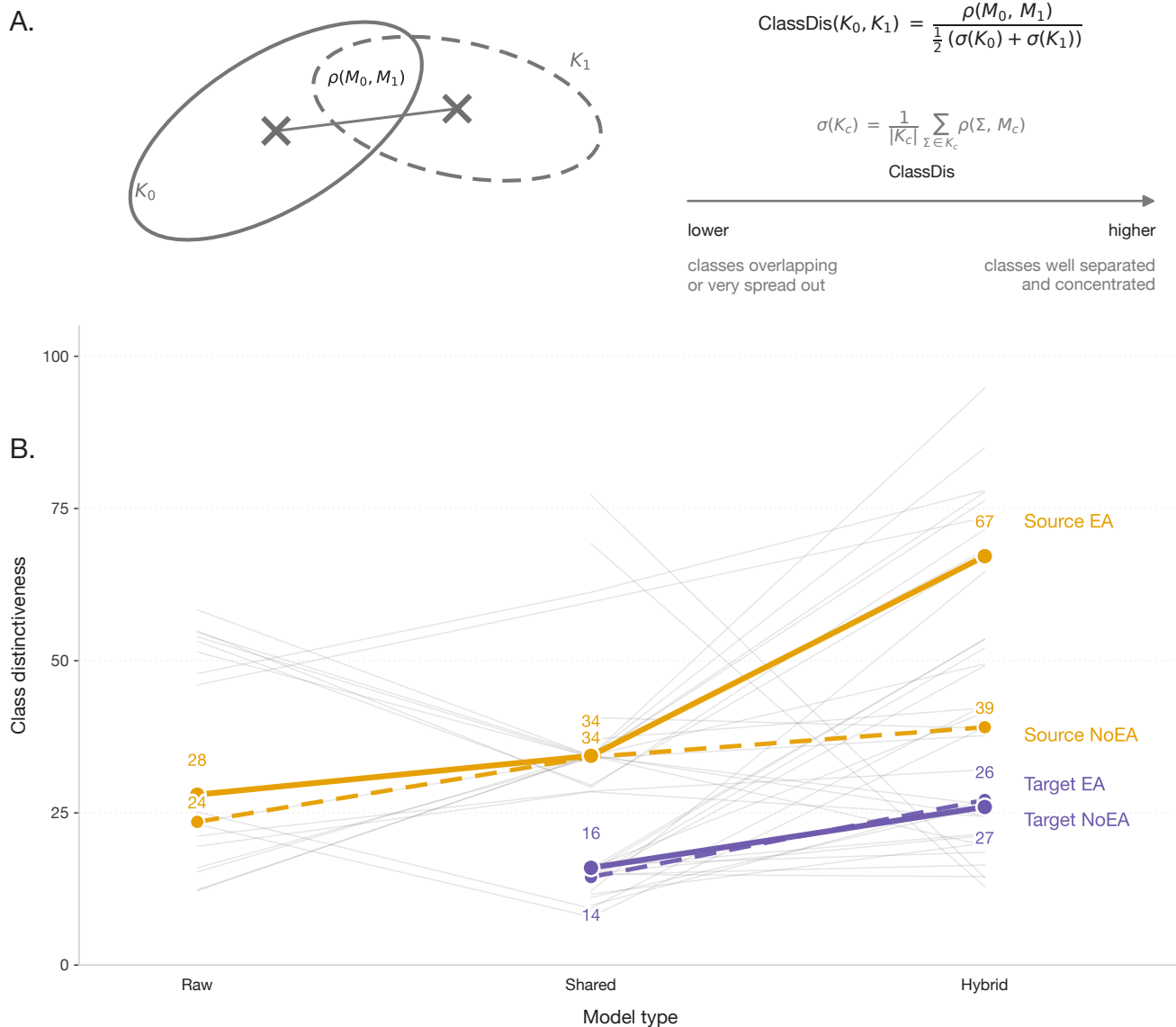


Figure 6. Subject-specific encoders sharply lift class distinctiveness for source subjects, with a smaller gain that still generalises to held-out target subjects. (A) The metric compares inter-class Riemannian distance to average intra-class dispersion. (B) Median per-subject distinctiveness across the pipeline (Raw, Shared, Hybrid) for source subjects (orange) and held-out target subjects (purple), with Euclidean Alignment (solid) and without (dashed); faint lines are individual subjects. The hybrid encoder lifts source distinctiveness well above the raw signal and the shared encoder (28 → 67 under EA); for unseen target subjects the gain is smaller (16 → 26) but stays above the shared-encoder level, indicating the subject-specific encoders retain transferable structure out of domain.

This metric measures an encoder-related quantity. Thus, an important question is whether *ClassDis acts as a diagnostic that relates to classifier behaviour, or is solely a geometric quantity that drifts from what the classifier cares about?* Figure 7 plots per-pair *ClassDis* against balanced accuracy under both alignments. The two quantities are tightly coupled in each condition (Pearson $r = 0.84$ without EA, $r = 0.80$ with EA), with a near-linear relationship from chance level to near-ceiling. *ClassDis* therefore acts as a classifier-independent proxy for decoding accuracy under both alignments, which justifies using it for the EA-versus-NoEA comparisons that follow.

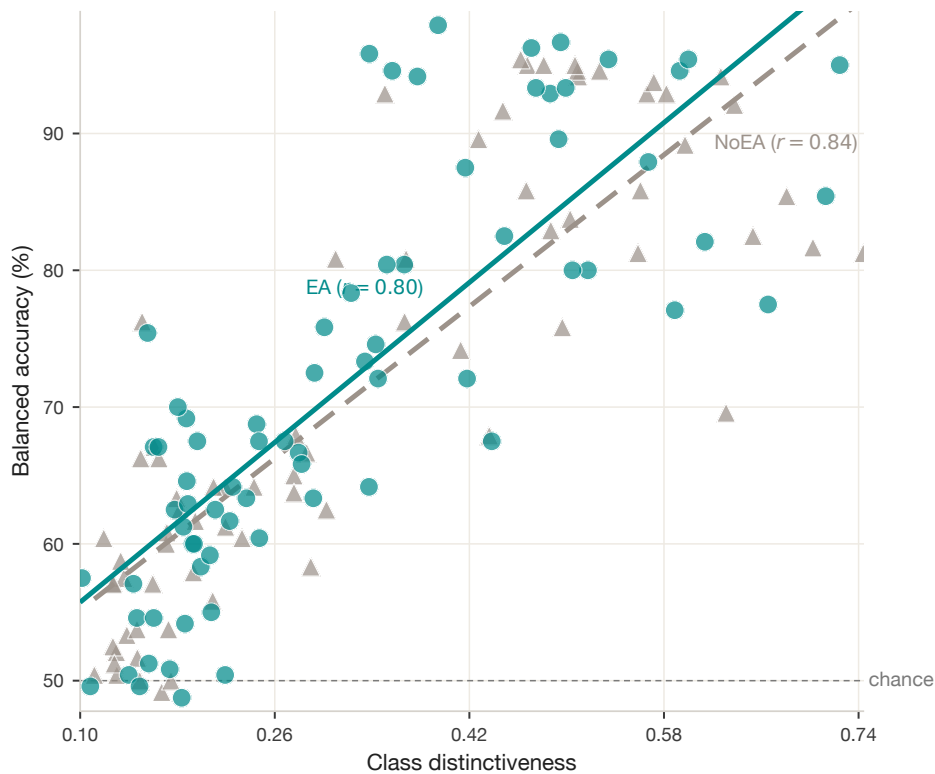


Figure 7. Class distinctiveness predicts downstream classification accuracy, under both alignments. Each point is one (target subject, encoder head) pair (NoEA: grey triangles; EA: teal circles); lines are least-squares fits (grey dashed = without EA, teal solid = with EA). The strong positive correlation under both conditions (Pearson $r = 0.84$ without EA, $r = 0.80$ with EA) shows that the affine-invariant, classifier-free *ClassDis* score tracks balanced accuracy rather than being an artefact of the alignment step.

3.5 Each hybrid head learns its subject’s manifold

Class distinctiveness (Section 3.4) confirms that the hybrid encoder produces more separable representations, especially for their own subjects, but it does not yet show that each head genuinely *prefers* them. The control we report in Figure 6B already pointed in that direction: passing a subject’s trials through a different subject’s head sharply decreases class distinctiveness values. We now make the head-subject geometry explicit on the manifold itself.

For each target subject we compute two normalized Riemannian distances on the trial-level latent covariances: the distance from the target’s trials to their own head’s centroid (*target to head*), and the distance from the same trials to every other head’s centroid (*target to others*). If the heads are truly subject-specific, the former should be small and the latter large; if the heads converged to a common transformation, the two would be indistinguishable.

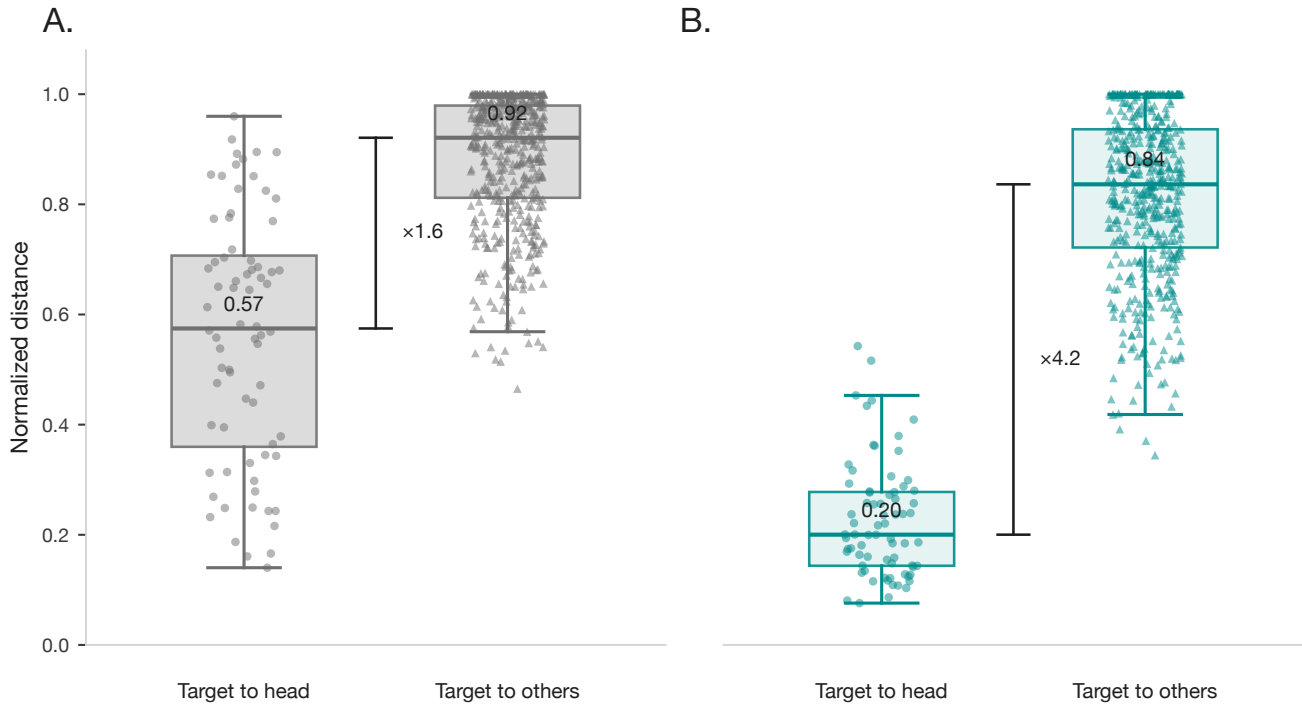


Figure 8. Each subject’s trials sit closer to their matched head’s manifold than to other heads’, and the gap widens under Euclidean Alignment. Normalized Riemannian distance from target-subject trials to per-head encoder centroids in the hybrid model’s latent covariance space, pooled over the leave-one-subject-out folds. In each panel the left box (circle points) is the distance to the matched head’s own subject (*target-to-head*); the right box (triangle points) is the distance to the other heads’ subjects (*target-to-others*). **Left (without EA).** The two distributions overlap, with a median gap of $\approx 1.6\times$; the panel is drawn in neutral grey. **Right (with EA).** Target-to-head drops well below target-to-others, a median gap of $\approx 4.2\times$; the panel is drawn in teal to mark the aligned condition. Box edges are the inter-quartile range; whiskers extend to 1.5 IQR; points are subject-head pairs.

Separation is present under both alignment regimes but is far larger after EA ($\approx 4.2\times$ versus $\approx 1.6\times$). In both conditions the matched head is significantly closer than the others (one-sided Mann–Whitney U over 72 matched versus 504 mismatched subject-head pairs: without EA $p = 1.8 \times 10^{-28}$, rank-biserial $r = 0.80$; with EA $p = 5.6 \times 10^{-43}$, $r = 1.00$, i.e. near-perfect separation). Once EA places the source and target distributions on a common centre, each subject’s trials are filtered most cleanly by their own head, because the encoder is trained to select the features most relevant for that subject; a mismatched head projects the same trials farther away.

The separation between target-to-head and target-to-others distances confirms that subject-specificity is not a side-effect of capacity or of having multiple heads: it is the *content* of what each head learned. This explains the class-distinctiveness decrease under the mismatched head situation in Figure 6B, and sets up the next question: if each head encodes its own subject, what does the resulting cross-subject geometry look like at the architecture level?

3.6 The hybrid encoder internalizes EA’s alignment role

Two independent measurements point in the same direction: the hybrid architecture’s per-subject encoders take over the functional role that EA plays externally on the shared model, reaching a similar effect through learned feature transformations rather than an explicit covariance recentring. We report the evidence from two angles: training dynamics (validation-loss convergence with and without EA) and latent geometry (pairwise inter-subject distance in the learned representation).

Training dynamics. Figure 9 reports the validation-loss trajectories for both architectures with and without EA, smoothed by a 7-epoch rolling median over 9 per-subject runs and truncated at the largest epoch where at least half

of the runs are still alive. For the shared encoder, EA is essential: applying it cuts the final median validation loss by 33% (from 0.50 to 0.34), and removing it leaves the loss plateauing well above. For the hybrid encoder, EA has little effect on the final loss: with or without EA it converges to essentially the same level (median 0.40 with EA versus 0.38 without, a +0.02 shift), well below the shared non-aligned curve (0.50). The without-EA curve nearly coincides with the aligned one. That the same architecture requires EA under a shared encoder but not once each subject has its own encoder indicates that the per-subject encoders take over EA’s role, learning from task supervision alone the *common* representation EA otherwise imposes externally. The latent geometry below corroborates this.

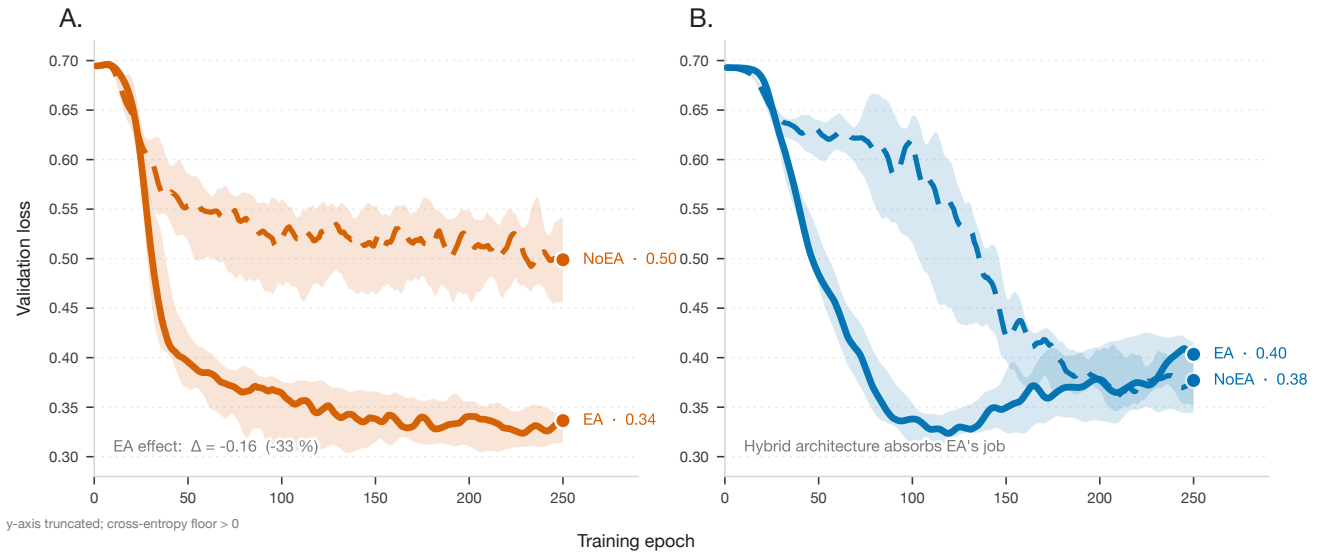


Figure 9. The hybrid architecture takes over the role of Euclidean Alignment. Validation loss per training epoch, median across 9 per-subject runs (shaded band = inter-quartile range, smoothed with a 7-epoch rolling window). **Left.** Shared model. Solid vermilion = EA; dashed vermilion = NoEA. EA reduces the final median loss by 33% ($\Delta = -0.16$); without EA the loss never converges to the same level. **Right.** Hybrid model. Solid blue = EA; dashed blue = NoEA. The hybrid reaches essentially the same final loss with or without EA (NoEA \approx 0.38 versus EA 0.40, $\Delta = +0.02$); the two curves nearly coincide, so the per-subject encoders take over EA’s role. Either way the hybrid reaches the regime the shared encoder attains only with external EA.

Latent geometry. By projecting the encoder’s latent features for both the aligned shared and the non-aligned hybrid architectures using t-SNE, we can observe the impact of the training on the subjects’ latent distributions. For the shared encoder, all distributions are initially projected in the same manifold, and learning makes the classes more well separated. For the hybrid encoders, however, the representations start with a strong inductive bias of separation, since all training subjects are projected differently. Figure 10 makes this contrast explicit, tracking both architectures from random initialisation to the trained model.



Figure 10. Training splits the held-out subject by class. Two-dimensional t-SNE of the 496-dimensional encoder features for the held-out target subject (*Target*, subject 8 from *BNCI2014* dataset, encoded through one source head) and the eight source subjects (*Source*), under Euclidean Alignment. Rows: the hybrid encoder bank (top) and the shared encoder (bottom). Columns: random initialisation (left, the untrained control) and after training (right). Color marks domain (Source / Target), and shape marks class (circle / cross); target points are drawn larger because they are the held-out focus. Each panel is an *independent* t-SNE embedding (perplexity = 50), so only within-panel grouping is interpretable, not distances, positions, or cluster sizes across panels. At random initialization, the hybrid encoder already places the target in its own cluster, separate from the source pool; after training the target splits by class, which cluster at the periphery of the source pool, tracking the matching source class rather than dissolving into it. The shared encoder begins with all subjects overlapping in one manifold, and training mainly sharpens the class boundary within it.

Neither architecture receives an explicit class-separation or domain-alignment loss, so the only pressure on the latent space is the shared classifier acting on top of the encoders. Therefore, the figure highlights that, even with a strong initial bias, the hybrid model is able to reorganize the latent space using task supervision alone, making distributions closer and better separated. Although still not as aligned as the representations from the shared model, this shows that the objective is sufficient to make the subject-specific heads specialize and filter task-relevant features, while improving cross-subject alignment.

Synthesis. Two independent measurements converge on the same conclusion: validation-loss convergence (Figure 9) and inter-subject latent representation (Figure 10) both show that the hybrid encoder *internalizes a role similar to EA for the shared encoder*. The shared encoder requires EA both to converge and to bring subject distributions together; removing EA collapses both. The hybrid encoder achieves similar outcomes with or without EA, consistent with each subject’s encoder performing filtering and recentering of a functional effect similar to EA. We do not claim that the learned encoders reproduce EA’s recentering: their mechanism differs, and the convergence and latent-geometry effects above need not act at the covariance level on which EA operates. What the two measurements rule out is the alternative explanation that the hybrid simply adds capacity.

3.7 Subject-level heterogeneity and its implications

The summary measurements so far treat subjects as exchangeable units. Subject-level performance, however, is far from uniform, and the variability has more than one mode. Some subjects are consistently hard for every method; others swing widely depending on which alignment and architecture they encounter. Disentangling these two regimes matters because they call for different mitigations: a consistently hard subject suggests irreducible signal limitations, while a method-noisy subject suggests the model bank still has room to grow.

Figure 11 plots each subject as a single point with $x =$ mean test accuracy and $y =$ standard deviation across the 18 method \times alignment \times head-selection combinations available per dataset (the standard deviation is our proxy for the per-subject prediction entropy the underlying experiments do not record directly). The two callouts in each panel name the most extreme subject along each axis. BNCI2014-001’s subject 2 (56.9 ± 4.1) is consistently hard; BNCI2015-001’s subject 3 (72.7 ± 18.3) is the most method-noisy; Weibo2014’s subject 4 is consistently hard (52.6 ± 5.0) and its subject 2 (76.5 ± 13.7) the most method-noisy. Schirrmeister2017’s subjects are comparatively uniform, its noisiest reaching only $\sigma \approx 9$.

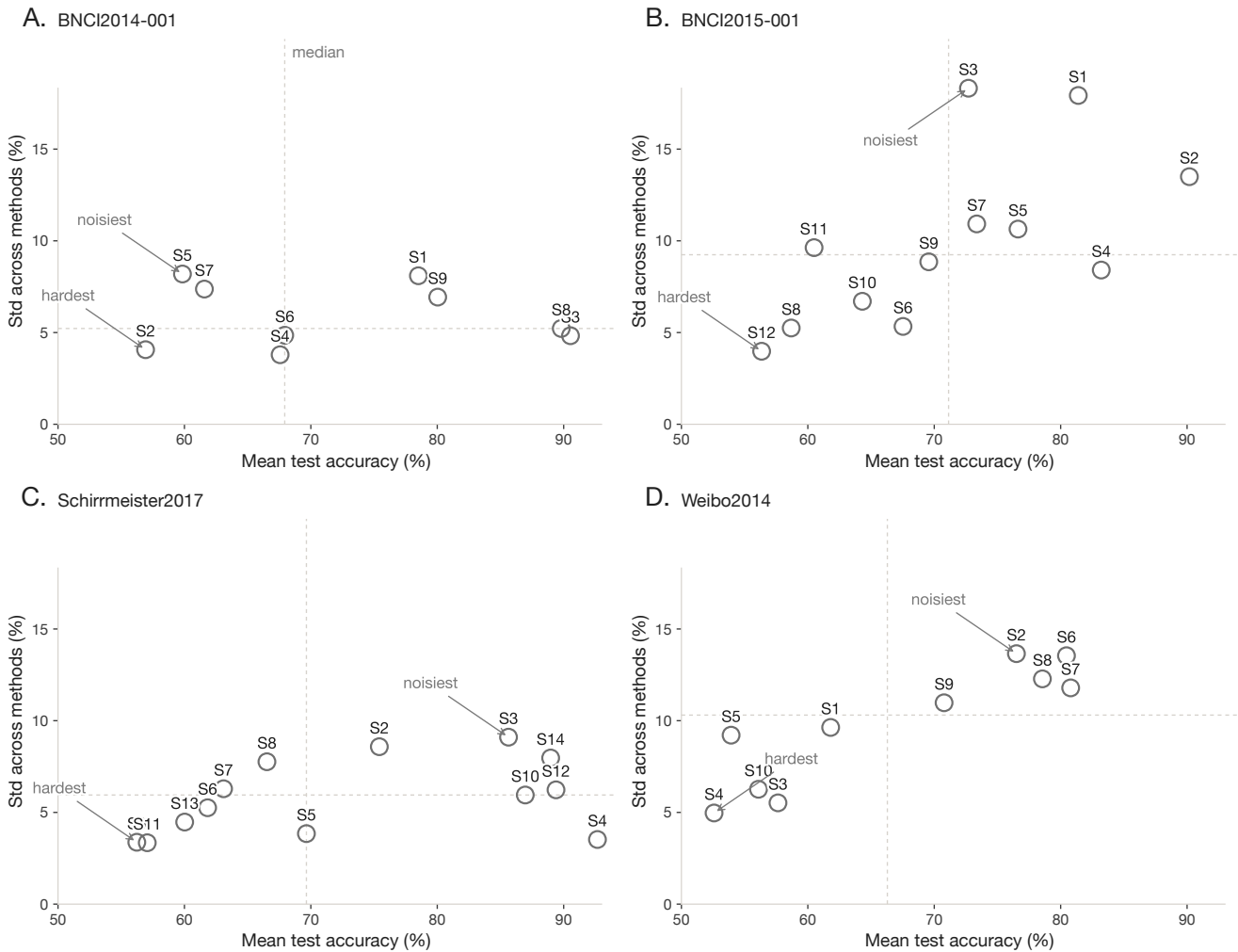


Figure 11. Subjects split into a consistently-hard mode and a method-noisy mode, with the mix changing per dataset. Each point is one subject; x = mean test accuracy across all 18 method \times alignment \times head-selection combinations; y = standard deviation of that same population. Dashed grey lines mark per-panel medians; the resulting quadrants read as: lower-left = consistently hard, lower-right = consistently easy, upper-half = method-noisy. The callouts label the lowest-mean and highest-std subjects per dataset. Subject behaviour is heterogeneous and at least two-dimensional: method-noisy subjects sit anywhere along x , including near-ceiling accuracy.

The two regimes motivate different next steps to test. For consistently-hard subjects, none of the architectural variants we evaluated closed the gap, suggesting that improvements likely require either subject-specific calibration data or signal-quality interventions outside the current pipeline. For method-noisy subjects, the per-subject spread across configurations implies that some configuration is closer to the achievable upper bound; whether a method-ensembling or per-subject head-selection policy can recover that upper bound at test time without oracle access is the question those subjects pose for future work.

4 Discussion and limitations

The experiments show that cross-subject EEG decoding is limited less by classifier capacity than by how each subject is represented before the shared decision rule. EA improves the shared encoder because it removes a dominant covariance-level shift before learning begins, which otherwise would cause negative transfer and strongly limit the capacity of the classifier to find a class-discriminant decision boundary. The hybrid architecture targets the same bottleneck inside the model: each subject-specific encoder learns a transformation that makes its trials more compatible with a common classifier, as shown by the loss and latent representation analyses. The class distinctiveness and relative distances show additional benefits of applying EA on top of the hybrid model, such as bigger class separation on training subjects and easier target adaptation.

This does not mean that the learned encoders completely reproduce the EA pointwise. Rather, they appear to be working under a similar mechanism, induced by the pressure of the architecture and the objective: bringing representations closer while keeping classes well separated. The class-distinctiveness and head-specificity analyses further show that the gains are not explained by adding generic capacity; instead, they depend on routing a subject through an encoder whose learned manifold matches that subject.

Comparing the evolution of the latent representations shows that the sole pressure of this learning mechanism wins out over the inductive bias of subject separation, making the representations closer in space, although still not as unified as on the aligned shared baseline, and the classes better separated.

Although the model architecture and supervised objective together can be driven to learn subject-specific preprocessing transformations, this study has limitations in both analysis and deployment.

First, the individual encoders learn subject-specific characteristics, obtaining individualized weight representations of the training distributions. However, this specialization limits its applicability to test-time inference on a new subject. For unseen subjects, the model must select or adapt an initialization from limited calibration data, and the subject-profile analysis shows that this problem is heterogeneous: some subjects remain consistently hard, while others are sensitive to method choice. Future work should therefore focus on non-oracle head selection, calibration-efficient adaptation, and diagnostics that identify whether a target subject needs more data or a better-matched initialization.

Although this work focused on cross-subject decoding, the same head specialization could improve temporal robustness in a cross-session setting, as studied by Wei et al.²¹. The representation visualizations show a spread and separated latent space for the training subjects, meaning that performance could largely benefit from that. Later work should investigate the benefits of using this type of architecture on additional data from training subjects.

Related ideas appear in recent non-invasive *speech* decoding, where a shared encoder is equipped with an early, per-subject *linear* layer, a channel-wise transform that maps each subject into a common latent space while the rest of the network stays shared^{46–48}. Ablating this layer markedly degrades decoding⁴⁶, confirming that subject-dependent effects must be modelled explicitly. Our design takes the *opposite* parameterisation: rather than a single linear block on top of a shared encoder, we learn a full per-subject *encoder* feeding one shared classifier, and the loss and latent-geometry analyses (Figures 9 and 10) show that these encoders reach the same latent organisation with or without EA, taking over its alignment role through learned feature transformations rather than the explicit covariance recentring EA applies. This is an alignment role those subject layers assume but never test. This expressiveness comes at a cost: full per-subject encoders grow far faster than a per-subject matrix as the number of training subjects increases, which is not scalable to more diverse datasets and motivates parameter-efficient alternatives such as subject-conditioned low-rank adapters²⁶. It also leaves the same open problem these speech models do not

address: none evaluates a held-out subject, which is precisely the deployment bottleneck that our head-selection and calibration analysis targets.

Two limitations bound these conclusions. First, deploying the model to a new subject still depends on selecting or adapting a head from limited calibration data, a problem the subject-profile analysis shows is heterogeneous across subjects. Second, dedicating one encoder per subject does not scale to large or diverse subject pools, which motivates the parameter-efficient alternatives discussed above.

References

1. Xu, L., Xu, M., Jung, T.-P. *et al.* Review of brain encoding and decoding mechanisms for eeg-based brain-computer interface. *Cogn. Neurodynamics* **15**, 569–584, DOI: [10.1007/s11571-021-09676-z](https://doi.org/10.1007/s11571-021-09676-z) (2021).
2. Guetschel, P. & Tangermann, M. Transfer learning between motor imagery datasets using deep learning-validation of framework and comparison of datasets. *arXiv preprint arXiv:2311.16109* (2023).
3. Roy, Y. *et al.* Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* **16**, 051001 (2019).
4. Barachant, A., Bonnet, S., Congedo, M. & Jutten, C. Multiclass Brain-Computer Interface Classification by Riemannian Geometry. *IEEE Transactions on Biomed. Eng.* **59**, 920–928, DOI: [10.1109/TBME.2011.2172210](https://doi.org/10.1109/TBME.2011.2172210) (2012).
5. Ang, K. K., Chin, Z. Y., Wang, C., Guan, C. & Zhang, H. Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. *Front. Neurosci.* **6**, DOI: [10.3389/fnins.2012.00039](https://doi.org/10.3389/fnins.2012.00039) (2012).
6. Lotte, F., Bougrain, L. & Clerc, M. Electroencephalography (EEG)-based Brain-Computer Interfaces. In *Wiley Encyclopedia of Electrical and Electronics Engineering*, 44, DOI: [10.1002/047134608X.W8278](https://doi.org/10.1002/047134608X.W8278) (Wiley, 2015).
7. Ludwig, S. *et al.* Eegminer: discovering interpretable features of brain activity with learnable filters. *J. Neural Eng.* **21**, DOI: [10.1088/1741-2552/ad44d7](https://doi.org/10.1088/1741-2552/ad44d7) (2024).
8. Banville, H. *et al.* Self-supervised representation learning from electroencephalography signals. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6, DOI: [10.1109/MLSP.2019.8918693](https://doi.org/10.1109/MLSP.2019.8918693) (2019).
9. Banville, H., Benchetrit, Y., d’Ascoli, S., Rapin, J. & King, J.-R. Scaling laws for decoding images from brain activity (2025). [2501.15322](https://arxiv.org/abs/2501.15322).
10. Aristimunha, B. *et al.* Evaluating the structure of cognitive tasks with transfer learning. *arXiv preprint arXiv:2308.02408* (2023).
11. Han, J., Wei, X. & Faisal, A. A. Eeg decoding for datasets with heterogeneous electrode configurations using transfer learning graph neural networks. *J. Neural Eng.* **20**, 066027, DOI: [10.1088/1741-2552/ad09ff](https://doi.org/10.1088/1741-2552/ad09ff) (2023).
12. Mellot, A., Collas, A., Chevallier, S., Engemann, D. A. & Gramfort, A. Physics-informed and Unsupervised Riemannian Domain Adaptation for Machine Learning on Heterogeneous EEG Datasets. *ArXiv* **abs/2403.15415** (2024).
13. Khazem, S., Chevallier, S., Barthélemy, Q., Haroun, K. & Noûs, C. Minimizing subject-dependent calibration for BCI with Riemannian transfer learning. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 523–526 (IEEE, 2021).
14. Marty, L., Pereira, M., Bellec, G. & Barbosa, J. Self-supervised EEG pretraining amplifies individual differences in neural representations. In *ICLR 2026 Workshop on Representational Alignment (Re⁴-Align)* (2026).
15. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowl. Data Eng.* **22**, 1345–1359 (2010).
16. He, H. & Wu, D. Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomed. Eng.* **67**, 399–410 (2020).

17. Mellot, A., Collas, A., Coelho Rodrigues, P. L., Engemann, D. & Gramfort, A. Harmonizing and aligning M/EEG datasets with covariance-based techniques to enhance predictive regression modeling. *Imaging Neuroscience* **1**, 1–26, DOI: [10.1162/imag_a_00040](https://doi.org/10.1162/imag_a_00040) (2023).
18. Bleuzé, A., Mattout, J. & Congedo, M. Tangent space alignment: Transfer learning for Brain-Computer Interface. *Front. Hum. Neurosci.* 16–1049985 (2022).
19. Rodrigues, P., Jutten, C. & Congedo, M. Riemannian Procrustes Analysis: Transfer learning for brain-computer interfaces. *IEEE Transactions on Biomed. Eng.* **PP** (2018).
20. Junqueira, B., Aristimunha, B., Chevallier, S. & de Camargo, R. Y. A systematic evaluation of euclidean alignment with deep learning for eeg decoding. *J. Neural Eng.* DOI: [10.1088/1741-2552/ad4f18](https://doi.org/10.1088/1741-2552/ad4f18) (2024).
21. Wei, X., Ortega, P. & Faisal, A. A. Inter-subject Deep Transfer Learning for Motor Imagery EEG Decoding. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 21–24 (2021).
22. Phunruangsakao, C., Achancaray, D. & Hayashibe, M. Deep adversarial domain adaptation with few-shot learning for motor-imagery brain-computer interface. *IEEE Access* **10**, DOI: [10.1109/ACCESS.2022.3178100](https://doi.org/10.1109/ACCESS.2022.3178100) (2022).
23. Zhou, Y. *et al.* Cross-subject cognitive workload recognition based on eeg and deep domain adaptation. *IEEE Transactions on Instrumentation Meas.* **72** (2023).
24. Zhong, X.-C. *et al.* A deep domain adaptation framework with correlation alignment for eeg-based motor imagery classification. *Biomed. Signal Process. Control.* **83**, 104658, DOI: [10.1016/j.bspc.2023.104658](https://doi.org/10.1016/j.bspc.2023.104658) (2023).
25. Bakas, S. *et al.* Latent alignment in deep learning models for eeg decoding. *J. Neural Eng.* **22**, DOI: [10.1088/1741-2552/adb336](https://doi.org/10.1088/1741-2552/adb336) (2025).
26. Klein, T., Minakowski, P. & Sager, S. Mitigating subject dependency in EEG decoding with subject-specific low-rank adapters. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body* (2025).
27. Schirrmester, R. T. *et al.* Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **38**, 5391–5420 (2017).
28. King, J.-R. *et al.* Encoding and decoding framework to uncover the algorithms of cognition, DOI: [10.7551/mitpress/11442.003.0076](https://doi.org/10.7551/mitpress/11442.003.0076) (2020).
29. Wimpff, M., Gizzi, L., Zerfowski, J. & Yang, B. Eeg motor imagery decoding: A framework for comparative analysis with channel attention mechanisms. *J. Neural Eng.* (2024).
30. Lawhern, V. J. *et al.* EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **15**, 056013 (2018).
31. Chen, X., Teng, X., Chen, H., Pan, Y. & Geyer, P. Toward reliable signals decoding for electroencephalogram: A benchmark study to eegnex. *Biomed. Signal Process. Control.* **87**, 105475, DOI: <https://doi.org/10.1016/j.bspc.2023.105475> (2024).
32. Lian, S. & Li, Z. An end-to-end multi-task motor imagery eeg classification neural network based on dynamic fusion of spectral-temporal features. *Comput. Biol. Medicine* **178**, 108727, DOI: <https://doi.org/10.1016/j.combiomed.2024.108727> (2024).
33. Zhuang, F. *et al.* A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76, DOI: [10.1109/JPROC.2020.3004555](https://doi.org/10.1109/JPROC.2020.3004555) (2021).
34. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowl. Data Eng.* **22**, 1345–1359, DOI: [10.1109/TKDE.2009.191](https://doi.org/10.1109/TKDE.2009.191) (2010).
35. Lotte, F. *et al.* A review of classification algorithms for EEG-based brain-computer interfaces: A 10-year update. *J. Neural Eng.* **15** (2018).

36. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. & Smola, A. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J. & Hoffman, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 19 (MIT Press, 2006).
37. Yger, F., Berar, M. & Lotte, F. Riemannian Approaches in Brain-Computer Interfaces: A Review. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **25**, 1753–1762 (2017).
38. Wimpff, M., Gizzi, L., Zerfowski, J. & Yang, B. Eeg motor imagery decoding: a framework for comparative analysis with channel attention mechanisms. *J. Neural Eng.* **21**, 036020, DOI: [10.1088/1741-2552/ad48b9](https://doi.org/10.1088/1741-2552/ad48b9) (2024).
39. Zhao, W., Jiang, X., Zhang, B., Xiao, S. & Weng, S. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Sci. Reports* **14**, 20237, DOI: [10.1038/s41598-024-71118-7](https://doi.org/10.1038/s41598-024-71118-7) (2024).
40. Tangermann, M. *et al.* Review of the BCI Competition IV. *Front. Neurosci.* **6** (2012).
41. Faller, J., Vidaurre, C., Solis-Escalante, T., Neuper, C. & Scherer, R. Autocalibration and recurrent adaptation: Towards a plug and play online erd-bci. *IEEE Transactions on Neural Syst. Rehabil. Eng.* **20**, 313–319, DOI: [10.1109/TNSRE.2012.2189584](https://doi.org/10.1109/TNSRE.2012.2189584) (2012).
42. Yi, W. *et al.* Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLoS ONE* **9**, e114853, DOI: [10.1371/journal.pone.0114853](https://doi.org/10.1371/journal.pone.0114853) (2014).
43. Aristimunha, B. *et al.* Mother of all BCI Benchmarks (2023). Version 1.0.0, doi:10.5281/zenodo.10034223.
44. Nam, C. S., Nijholt, A. & Lotte, F. *Brain–computer interfaces handbook: technological and theoretical advances* (CRC Press, 2018).
45. Lotte, F. & Jeunet, C. Defining and quantifying users’ mental imagery-based BCI skills: a first step. *J. Neural Eng.* **15**, 046030, DOI: [10.1088/1741-2552/aac577](https://doi.org/10.1088/1741-2552/aac577) (2018).
46. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5**, DOI: [10.1038/s42256-023-00714-5](https://doi.org/10.1038/s42256-023-00714-5) (2023).
47. d’Ascoli, S. *et al.* Decoding individual words from non-invasive brain recordings across 723 participants. *arXiv preprint arXiv:2412.17829* (2024).
48. Lévy, J. *et al.* Brain-to-text decoding: A non-invasive approach via typing. *arXiv preprint arXiv:2502.17480* (2025).

Data availability

The datasets analysed in this study are publicly available and were accessed through MOABB⁴³: BNCI2014-001⁴⁰, BNCI2015-001⁴¹, High-Gamma²⁷, and Weibo2014⁴².

Competing interests

Bruno Aristimunha has been associated with Yneuro since November 2025. The other authors declare no competing interests.